ELSEVIER

Cognitive Science 26 (2002) 521–562

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

The misunderstood limits of folk science: an illusion of explanatory depth

Leonid Rozenblit*, Frank Keil

Department of Psychology, Yale University, 2 Hillhouse Avenue, P.O. Box 208205, New Haven, CT 06520-8205, USA

Received 20 August 2001; received in revised form 26 April 2002; accepted 3 May 2002

Abstract

People feel they understand complex phenomena with far greater precision, coherence, and depth than they really do; they are subject to an illusion—an illusion of explanatory depth. The illusion is far stronger for explanatory knowledge than many other kinds of knowledge, such as that for facts, procedures or narratives. The illusion for explanatory knowledge is most robust where the environment supports real-time explanations with visible mechanisms. We demonstrate the illusion of depth with explanatory knowledge in Studies 1–6. Then we show differences in overconfidence about knowledge across different knowledge domains in Studies 7–10. Finally, we explore the mechanisms behind the initial confidence and behind overconfidence in Studies 11 and 12, and discuss the implications of our findings for the roles of intuitive theories in concepts and cognition.
© 2002 Leonid Rozenblit. Published by Cognitive Science Society, Inc. All rights reserved.

Keywords: Concepts; Epistemology; Meta-cognition; Knowledge; Overconfidence

1. Introduction

Intuitive or lay theories are thought to influence almost every facet of everyday cognition. People appeal to explanatory relations to guide their inferences in categorization, diagnosis, induction, and many other cognitive tasks, and across such diverse areas as biology, physical mechanics, and psychology (Gopnik & Wellman, 1994; Keil, 1998; Murphy & Medin, 1985; Murphy, 2000). Individuals will, for example, discount high correlations that do not conform to an intuitive causal model but overemphasize weak correlations that do (Chapman & Chapman,

* Corresponding author. Tel.: +1-203-432-6763; fax: +1-203-432-4623.
E-mail addresses: leonid.rozenblit@yale.edu (L. Rozenblit), frank.keil@yale.edu (F. Keil).

0364-0213/02/$ – see front matter © 2002 Leonid Rozenblit. Published by Cognitive Science Society, Inc. All rights reserved.
PII: S0364-0213(02)00078-2

1969). Theories seem to tell us what features to emphasize in learning new concepts as well as highlighting the relevant dimensions of similarity (Murphy, 2002). Intuitive theories have also been heavily emphasized in accounts of the cognitive development of children (Gelman & Koenig, 2002) and even of infants (Spelke, Breinliinger, Macomber, & Jacobson, 1992).

Concepts seem to be embedded within larger sets of explanatory relations that are essential to understanding the structure of the concepts themselves, how they are learned, and how they change over time. But even as theories have become more central to the study of concepts, it is also now evident that folk theories are rarely complete or exhaustive explanations in a domain (Wilson & Keil, 1998). Indeed, even the theories used daily to guide scientific research are now considered to be incomplete, or at least less formally logical than classical views assumed them to be (Boyd, 1991; Salmon, 1989, 1998). Science-in-practice is often driven by hunches and vague impressions.

The incompleteness of everyday theories should not surprise most scientists. We frequently discover that a theory that seems crystal clear and complete in our head suddenly develops gaping holes and inconsistencies when we try to set it down on paper.

Folk theories, we claim, are even more fragmentary and skeletal, but laypeople, unlike some scientists, usually remain unaware of the incompleteness of their theories (Ahn & Kalish, 2000; Dunbar, 1995; diSessa, 1983). Laypeople rarely have to offer full explanations for most of the phenomena that they think they understand. Unlike many teachers, writers, and other professional "explainers," laypeople rarely have cause to doubt their naïve intuitions. They believe that they can explain the world they live in fairly well. They are novices in two respects. First, they are novice "scientists"—their knowledge of most phenomena is not very deep. Second, they are novice epistemologists—their sense of the properties of knowledge itself (including how it is stored) is poor and potentially misleading.

We argue here that people's limited knowledge and their misleading intuitive epistemology combine to create an illusion of explanatory depth (IOED). Most people feel they understand the world with far greater detail, coherence, and depth than they really do. The illusion for explanatory knowledge–knowledge that involves complex causal patterns—is separate from, and additive with, people's general overconfidence about their knowledge and skills. We therefore propose that knowledge of complex causal relations is particularly susceptible to illusions of understanding.

There are several features of explanatory, theory-like knowledge that may converge to convince people they have vivid, blueprint-like senses of how things work, even when their actual knowledge is skeletal and incomplete. One factor concerns a confusion between what is represented in the head with what can be recovered from a display in real time. When people succeed at solving problems with devices they may underestimate how much of their understanding lies in relations that are apparent in the object as opposed to being mentally represented. We expect the representation/recovery confusion to be less important with other kinds of knowledge, e.g., facts, narratives, or procedures.

This confusion of environmental support with internal representation is related to a confusion that has been noted in the "change blindness" literature: people grossly overestimate their ability to remember what they have observed in a scene. This phenomenon, termed "change blindness blindness," presumably occurs because people are mistaken about how visual information is stored—they confuse their ability to acquire details by re-sampling a live scene with

exhaustive, VCR-like storage of everything one sees (Levin, Momen, Drivdahl, & Simons, 2000).

The confusion of environmental support for detailed representation might be expected to be strongest for phenomena that have perceptually vivid mechanisms. If we can see many of the "working parts" of a system, we may assume that the mechanisms can be easily internalized. But there is far more complexity in the interactions of the parts than is immediately apparent. Furthermore, as suggested by the change blindness literature, we may assume we remember vividly things we have seen as vivid.

A second feature leading to the IOED may be a confusion of higher with lower levels of analysis. Most complex artificial and natural systems are hierarchical in terms of explanations of their natures. In explaining a car one might describe the function of a unit, such as the brakes, in general terms, and then turn to describing the functions of subcomponents, such as pistons and brake pads, which in turn can be broken down even further. The iterative nature of explanations of this sort (Miyake, 1986) may lead to an illusion of understanding when a person gains insight into a high level function and, with that rush of insight, falsely assumes an understanding of further levels down in the hierarchy of causal mechanisms. This effect can easily happen for many natural and artificial systems with complex causal structures, especially those that have "stable subassemblies." The concept of stable subassemblies was developed by Simon (1996) as a way of describing units in the hierarchical structure of complex systems that are sufficiently internally stable that they can be conceived of as an operational unit.

Confusion between higher and lower levels of analysis may be related to the confusion of environmental support with representation, especially for perceptually vivid mechanisms which may trigger a sense of understanding at higher levels. For example, functional sub-assemblies that are easy to visualize and mentally animate may lead to strong (but mistaken) feelings of understanding at a high level of analysis, and thereby induce inaccurate feelings of comprehension about the lower levels.

A third feature of explanations leading to the illusion is related to the second: because explanations have complex hierarchical structure they have indeterminate end states. Therefore, self-testing one's knowledge of explanations is difficult. In contrast, determining how well one knows, e.g., a fact, can be trivially simple. Do you know the capital of England? If you can produce "London," then "yes." Similarly, to assess whether one knows a procedure one can envision a clear end state (e.g., a baked cake, a successful log-on to the Internet) and then work backwards to see if one knows how to get to that state. Errors of omission are still possible but they are constrained by knowledge of the end state. But with explanatory understanding one usually has little idea of what the final explanation will look like and the end state is largely indeterminate from the posing of the question.

A fourth feature is the rarity of production: we rarely give explanations and therefore have little information on past successes and failures. By contrast, we often retrieve facts, tell narratives of events, or perform procedures; hence it is often easy to assess our average level of knowledge in these cases by inspections of past performance. Although each of these four features may be present to some extent with other kinds of knowledge, such as facts or procedures, we claim they converge most strongly with explanations, producing a powerful illusion of knowing.

In the studies that follow we explore illusions of understanding for explanations, contrasting the exaggerated sense of knowing explanations with better calibrations for other sorts of

knowledge. We predict that the miscalibration for explanations, which we call "the illusion of explanatory depth," will be consistently larger than for many other domains of knowledge and that it will be linked to the distinctive features of the explanations just described. This emphasis on variations in overconfidence as a function of the kind of knowledge involved is a departure from more traditional accounts of overconfidence that focus on domain general effects.

In this paper, we examine differences in overconfidence between types of knowledge, with a special emphasis on the illusion of deep understanding that arises with explanatory knowledge. We use a novel method for measuring overconfidence in these studies: participants' self-rating of long term knowledge. Essentially we ask participants to indicate how surprised they are by how much or how little explanatory knowledge they can produce. Our method is also based on experimentally manipulating participant's perceptions of how much they know, rather than on comparing people's performance on a test with some normative standard. Because we compare across kinds of knowledge and find large differences in the magnitude of the effect, we are able to consider what specific properties contribute to an especially strong illusion.

To clarify the distinctive nature of our proposal it is useful to briefly consider prior research on overconfidence. Relevant research in the judgment and decision making tradition has used the disparity between people's average confidence levels for their answers to almanac questions, and the proportion of correct answers, to argue that people are overconfident (Fischhoff, 1982; Lichtenstein & Fischhoff, 1977; Yates, Lee, & Shinotsuka, 1996; Yates, Lee, & Bush, 1997; for critique of method see, e.g., Bjoerkman, 1994; Gigerenzer, 1996; Gigerenzer, Hoffrage, & Kleinboelting, 1991; Juslin, 1993a, 1993b, 1994). This tradition, however, does not focus on how illusions of knowing might differ across kinds of knowledge. Lumping diverse kinds of knowledge into a hypothetical "general knowledge" and looking for a general overconfidence effects may well obscure large differences in calibration between knowledge types.

The cognitive psychology literature on text comprehension also suggests overconfidence about one's knowledge. People are often poor at detecting when they have failed to understand a piece of text, both as adults (Glenberg & Epstein, 1985; Glenberg, Wilkinson, & Epstein, 1982; Lin & Zabrucky, 1998) and as children (Markman, 1977; Markman, 1979). In contrast, the current studies are concerned with people's ability to assess the knowledge they have before coming into the lab, rather than things learned in the course of an experiment. The implications of our research are different: they tell us less about how people learn when reading, and more about individuals' intuitive theories about how knowledge is stored and about the mismatch between what they think they already know and what they really know.

Another area of research has focused on meta-cognition and feelings of knowing (FOK) (Koriat, 1995; Metcalfe, Schwartz, & Joaquim, 1993). One recent analysis considers the two main models for FOK to be the cue familiarity model and the accessibility model (Koriat & Levy-Sadot, 2001). The accessibility model claims that the ease of accessing information prompted by the target drives FOKs. The cue familiarity model claims that FOK judgments are elicited by the familiarity of the cues themselves.

FOK judgment literature tends to focus on fact retrieval, especially those cases where a person feels they know an item, but cannot recall it. The IOED procedure instead asks about the depth of knowledge, granting to participants some degree of knowledge. Moreover, the IOED focuses on much larger knowledge structures than facts. Nonetheless, both the familiarity of items in the probe question and the accessibility of related information might be factors in

overconfidence about knowledge. The distinctive nature of explanations, as opposed to other kinds of knowledge, might be understood as being caused by some version of an accessibility bias, in which easy access to some visualized components may help cause an illusion of knowing. In the studies that follow, the influence of familiarity will also be examined.

Overconfidence also exists in areas that have little to do with knowledge. Participants have been shown to be overconfident about their future performance on motor tasks (e.g., West & Stanovich, 1997), their abilities compared to other people's abilities (e.g., Kruger & Dunning, 1999), and about their competence to perform a broad range of tasks (Bjork, 1998).

We argue that the illusion of depth seen with explanatory knowledge is a separate phenomenon from this general, self-image-related overconfidence, and that the illusion's magnitude varies depending on the structural properties of the knowledge in question. Thus, we will proceed by demonstrating that people are more overconfident about knowing explanations than they are about knowing other things.

A powerful illusion of depth distinctive to theory-like knowledge would have major implications for the alleged roles of theories in conceptual structure. One possibility is that the illusion represents a cognitive bias like naïve essentialism (Medin & Ortony, 1989) and reflects a tendency on the part of laypeople and cognitive scientists to assume that intuitive theories are a powerful component of our knowledge systems even when there is little data to support that conjecture. The essentialist bias may liberally create placeholders that reserve mental locations for essences, in anticipation that the locations will be filled later (Murphy, 2002; Gelman & Koenig, 2002). We may create such placeholders even for concepts that lack essences (Medin, 1989). Indeed, an excessive essentialism may have hindered scientific thought in cases where fixed essences are inconsistent with a new theory, such as evolution by natural selection (Hull, 1965).

The essentialist bias assumes there is an essence beyond or beneath the observable. An analogous "theory bias" would assume there is a rich network of causal relations that gives rise to surface phenomena, a network that corresponds to a form of theoretical understanding. Those biases, however, would not on their own cause illusions of knowing or understanding. One might think that a class of things has an essence or is explainable by an underlying network of causal relations without thinking one actually knows either. One might then readily defer to experts (e.g., Putnam, 1975), but not be inclined to overestimate how much one knows.

An additional step is needed to conclude that an essentialist bias should lead to overconfidence about knowledge. For example, feelings of certainty about the existence of essences and hidden mechanisms may foster the conviction that one must have substantial knowledge of essences and mechanisms. Thus, an attempt to account for the strong intuition that essences and hidden mechanisms exist guides one to attribute to oneself knowledge of those hidden properties. This attribution would, in turn, lead people to believe that theoretical relations structure their concepts, when in fact they have no real knowledge of those relations. Indeed, the relations might not even exist. The concepts-as-theories idea might therefore be untrue for most real-world concepts.

Alternatively, the IOED might arise from ways in which people construct skeletal but effective causal interpretations. They may wrongly attribute far too much fidelity and detail to their mental representations because the sparse renderings do have some efficacy and do provide a rush of insight. By this second view, concepts may be strongly influenced by theories and the

IOED is one sign of that influence. These intuitive theories, however, are quite different from how they might seem at first. Thus, the very usefulness of highly sparse causal schema may create an illusion of knowing much more. In addition, even a very small amount of relevant explanatory understanding can have strong effects on category learning, perhaps leading people to think they have a much richer understanding than they do (Murphy & Kaplan, 2000).

In light of current tensions concerning the relative roles of similarity, perceptual information and theory-like information in models of concepts (Goldstone & Johansen, 2002), a further understanding of this illusion, and its implications for the concepts-as-theories view, is critical. We proceed first by documenting the illusion and its relative specificity to explanatory understanding. We then explore reasons for the illusion and potential consequences for the roles of intuitive theories in cognitive science.

We explore the illusion in a series of 12 studies. The first four studies document the illusion's existence with knowledge about devices across several populations. Studies 5 and 6 test the robustness of the illusion. Studies 7–10 show its distinctive nature by tracking the magnitude of the illusion across several knowledge domains. The final two studies examine factors that influence the extent of the illusion. Stimuli and instructions used in the following 12 studies are available in the Supplemental Materials section through the Cognitive Science on-line Annex at http://www.elsevier.com/gej-ng/10/15/15/show/.

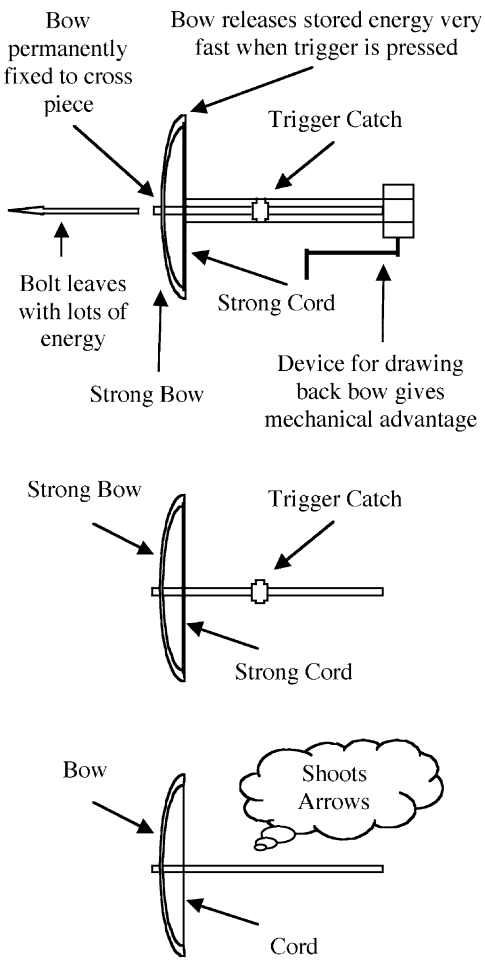
2. An illusion of explanatory depth with devices

2.1. Study 1: Documenting the illusion

2.1.1. Methods

We constructed a stimulus set using 40 distracter items and eight test items. The items were phenomena selected to represent a range of complexity and familiarity. We selected the items from a variety of books and web-resources that described "How things work." The test items (and most of the distracter items) were selected from a CD-ROM by David McCauley titled "*The Way Things Work 2.0*." (Kindersley, 1996). We selected the test items whose explanations did not depend on any specialized knowledge, because we needed all participants to understand the explanations at a later stage in the study. The test items were a speedometer, a zipper, a piano key, a flush toilet, a cylinder lock, a helicopter, a quartz watch, and a sewing machine. The complete stimulus packet can be found with the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/.

Sixteen graduate students from various departments at Yale University participated in the study. In Phase 1 of the study, the participants learned to use a 7-point scale that indicated whether they had a deep, partial, or shallow understanding of a device or phenomenon. The point "1" was the lowest level of understanding and "7" the highest. Participants considered two training examples, one for which most adults had a good mechanistic understanding, a cross bow, and one for which most do not, a global positioning satellite receiver. The explanation of the scale included text and diagrams illustrating the different levels of understanding for the two examples (see Fig. 1). All participants indicated that they clearly grasped the instructions.



Level 7 Knowledge: diagram and text excerpt

[… 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.

Level 4 Knowledge: diagram and text excerpt

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. If this person were to draw a diagram of a crossbow it might look like this.

Level 1 Knowledge: diagram and text excerpt

Some people might know even less. For example, someone might really only know what a crossbow looks like and what it does -- shoots arrows. That person's understanding might be best represented by the following diagram, where the lack of important parts and labels indicate they really don't have any idea about the details.

Fig. 1. Three illustrations included with the training instructions in Studies 1 and 2 for the crossbow example. The excerpts from the verbal descriptions of each level of knowledge are shown to the right of each diagram.

In Phase 2, participants studied a list of 48 items (shown in Appendix A) and rated each for their level of understanding on the 7-point scale described during the training. Participants were asked to rate the list without pausing excessively on any item, and took approximately 10 min to rate the entire list (rating "T1").

In Phase 3, we asked each participant to write a detailed, step-by-step causal explanation of each of the four test phenomena. Only four test phenomena were used in each item-set to keep the total time for the experiment under 1 h. (Eight participants were asked questions

about a speedometer, a zipper, a piano key, and a flush toilet. Eight were asked about a cylinder lock, a helicopter, a quartz watch, and a sewing machine.) After the participants provided an explanation, they re-rated how well they understood that phenomenon (rating "T2").

In Phase 4 of the study, the participants answered a "diagnostic" question about each of the four phenomena that required critical knowledge about the mechanism for an accurate response. For example, they were asked to explain, step-by-step, how one could pick a cylinder lock. For the helicopter they were asked to explain, step-by-step, how a helicopter changes from hovering to forward flight. They were then asked to re-rate their understanding in light of their answer to the diagnostic question (rating "T3").

In Phase 5, the participants read a brief expert description of the phenomenon and then re-rated their prior level of understanding relative to that description (rating "T4"). The expert explanations were taken from a CD-ROM by David McCauley titled "*The Way Things Work 2.0*." (Kindersley, 1996). The explanations contained diagrams and text, and ranged in length (including illustrations) from half a page to several pages.

Finally, as a manipulation check, participants were asked how well they understood the phenomenon after having read the expert explanation (rating "T5"). The sequence of the procedures is diagrammed in Fig. 2.

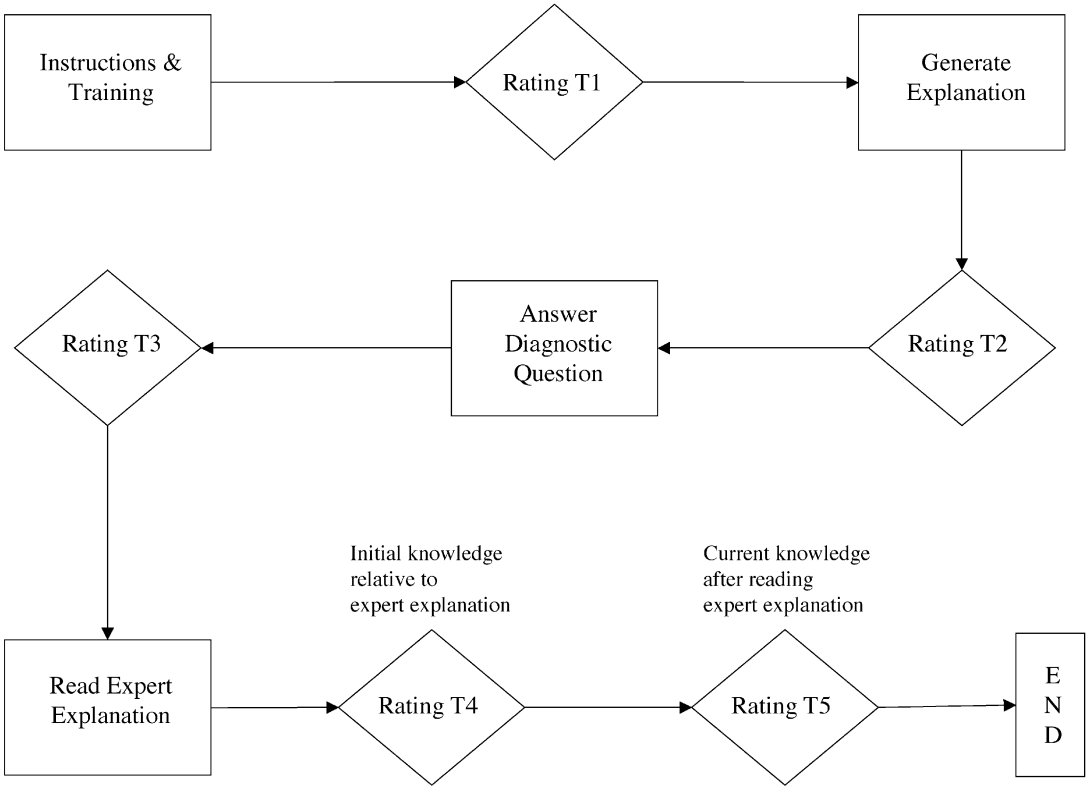


Fig. 2. Diagram of the procedure used in Studies 1–4.

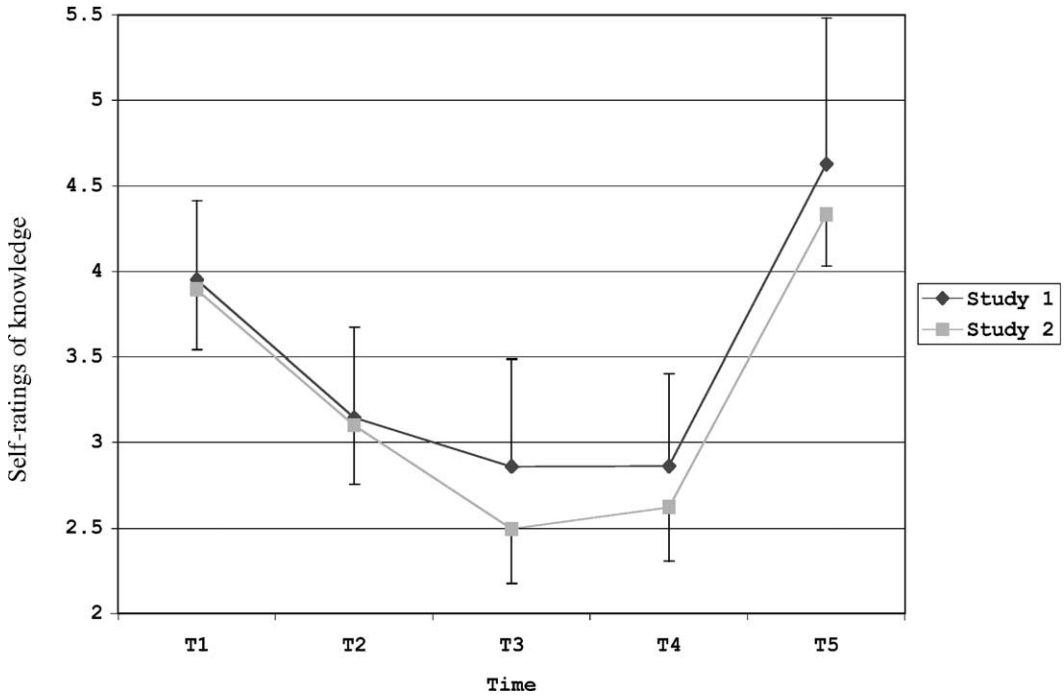


Fig. 3. Self-ratings of knowledge averaged across all items and subjects by time for Studies 1 and 2. The *x*-axis shows the sequence of self-ratings from Time 1 to Time 5. T1 is the initial self-rating, prior to any attempt to produce an explanation. T2 is the rating immediately after the first effort to explain. T3 is the rating after an attempt to answer a diagnostic question. T4 is the re-rating of one's initial knowledge provided after reading an expert explanation. T5 is the rating of one's current knowledge acquired as a result of reading an expert explanation, and is essentially a manipulation check. Self-ratings of knowledge in both Studies decrease as the result of efforts to explain.

2.1.2. Results

Nearly all participants showed drops in estimates of what they knew when confronted with having to provide a real explanation, answer a diagnostic question, and compare their understanding to an expert description, as seen in Fig. 3. The darker line in Fig. 3 shows the means for all items for T1–T5. A repeated measures ANOVA showed time was significant, $F(4, 56) = 16.195$, $p < .001$, $\eta^2 = .536$. The planned linear contrasts for Time 1 versus Time 2, Time 1 versus Time 3, and Time 1 versus Time 4 were all significant at $p < .002$. The contrasts for Time 2 versus Time 3, and Time 2 versus Time 4 were not significant.

There was also a clear rise in rated understanding as a result of learning from the expert description, suggesting that participants were not just continuously losing confidence in what they knew. The linear contrasts for Time 5 versus Times 4, 3, and 2 were all significant at $p < .003$.

There were differences between the estimates of knowledge for each of the eight items presented to the two groups of participants but no overall differences between the means for the two item sets given to each group. That is, in a repeated measures ANOVA with time (T1–T4) as a within-subject factor, and SET as a between-subject factor, neither SET nor SET × TIME were significant.

2.1.3. Discussion

Attempts to explain the various phenomena led to a decrease in self-rated understanding. Attempts to answer the diagnostic questions led to a further (non-significant) decrease. Finally, comparing one's earlier understanding with an expert explanation (at T4) did not lead to any change in the estimate of one's knowledge after the first three steps. That is, participants felt their knowledge self-ratings after the diagnostic questions (at T3) were accurately low even after they read the expert explanations—there was no change from T3 to T4. However, participants' estimates of their knowledge were not permanently depressed. Participants indicated (at T5) that their knowledge increased dramatically *as a result* of reading the expert explanations.

The scores at T4 and T5 help rule out certain classes of explanations for the initial drop. First, if the diagnostic questions were in some sense "unfair" or nit-picky, people's ratings of how much they know at T4 would have gone up once they read the expert explanations. After all, if the experts did not think a piece of knowledge is critical why should the participant? Second, if our procedure was shaking people's confidence in a general way, their knowledge ratings should have remained depressed at T5. Finally, if people's confidence simply dropped as a function of time, we would have expected drops at both T4 and T5.

Debriefing also revealed an interesting pattern in the subjective experience of participants: many participants reported genuine surprise and new humility at how much less they knew than they originally thought. Debriefing also suggested a robustness to the effect. Several participants remarked that "if only" they had received the other item-set, they would have done much better even though the overall levels of performance on the two stimulus sets were identical.

2.2. Study 2: Replicating the illusion in a larger, younger group

Study 2 replicated Study 1 with a larger group of participants, using Yale undergraduate instead of graduate students. Conceivably, graduate study leads to an intellectual arrogance and the illusion of explanatory competence might be less in undergraduates who are still awed by what they do not know.[1]

2.2.1. Methods and results

Thirty-three Yale undergraduates received the same stimuli and the same series of five questions as in Study 1. As shown by the lighter line in Fig. 3, the undergraduate participants exhibited the same overall pattern of results as the graduate students in Study 1. The repeated measures ANOVA showed time was significant, $F(4, 124) = 38.9$, $p < .001$, $\eta^2 = .555$. Planned linear contrasts for Time 1 versus Times 2, 3 and 4 were all significant at $p < .001$.

A direct comparison of Study 1 with Study 2 using a repeated measures ANOVA with time as a within-subject factor and study as a between-subject factor showed no differences between the two results. That is, the TIME × STUDY interaction was not significant, $F(4, 188) = .462$, $p = .902$, $\eta^2 = .006$. Of course, time remained highly significant in the combined analysis, $F(4, 188) = 44.11$, $p < .001$, $\eta^2 = .448$.

2.2.2. Discussion

As with the graduate student participants, attempts to explain devices lowered participants' estimates of knowledge. Again, many participants expressed surprise at their own levels of ignorance. Moreover, if anything, the direction of the effect was for the illusion to be somewhat (but not significantly) stronger with undergraduates than with graduate students.

2.3. Study 3: Replicating the illusion in a less selective university

Arguably the results of the first two studies might reflect an unusual population—graduate and undergraduate students at elite institutions might suffer from fatal overconfidence in how much they know. To control for that possibility, we replicated Studies 1 and 2 outside an Ivy League campus.

2.3.1. Methods and results

Sixteen graduate and undergraduate students at a regional and nonselective state university campus participated in the study. One index of the difference in selectivity can be seen in the mean math + verbal SAT scores for Yale undergraduates (approximately 1,500) compared to those at the regional university (approximately 960). The students received identical stimuli and instructions as participants in Studies 1 and 2.

The results were essentially similar to those obtained in Studies 1 and 2. The ratings of knowledge decreased significantly over time, in the now familiar pattern, as shown in Fig. 4. In a repeated measures ANOVA on the regional university data with time as a within-subjects factor, time was highly significant, $F(3, 42) = 23.557$, $p < .001$, $\eta^2 = .627$.

To test the "elitist arrogance" hypothesis directly, we compared the students at a selective university (Yale) to students at the regional university using a repeated measures ANOVA with institutional affiliation as the between-subject factor, and time as a within-subject factor. To maximize power to detect differences we combined graduate and undergraduate participants from Studies 1 and 2 into a single comparison group ($n = 49$).

While the main effect of institutional affiliations was not significant ($F(1, 63) = .750$, $p = .390$, $\eta^2 = .012$), the interaction of affiliation by time was highly significant, $F(4, 252) = 3.874$, $p = .005$, $\eta^2 = .058$. As Fig. 4 suggests, the interaction was driven primarily by the later steps in procedure, i.e., the changes between Times 3 and 5. The direction of the difference between groups was, furthermore, opposite of what the "elitist arrogance" hypothesis would predict. The effect was larger in students from a regional university, primarily because their initial ratings of their knowledge were nearly a point higher than that of students from a selective university (see Fig. 4).

2.3.2. Discussion

The study with students at a regional university replicated the results of the Studies 1 and 2. If anything, the IOED was larger among students at the regional university.

2.4. Study 4: Replicating the illusion with a different set of devices

So far, the results have been consistent across several populations. However, the results were obtained with a single set of stimuli, leaving open the possibility that a peculiar selection of

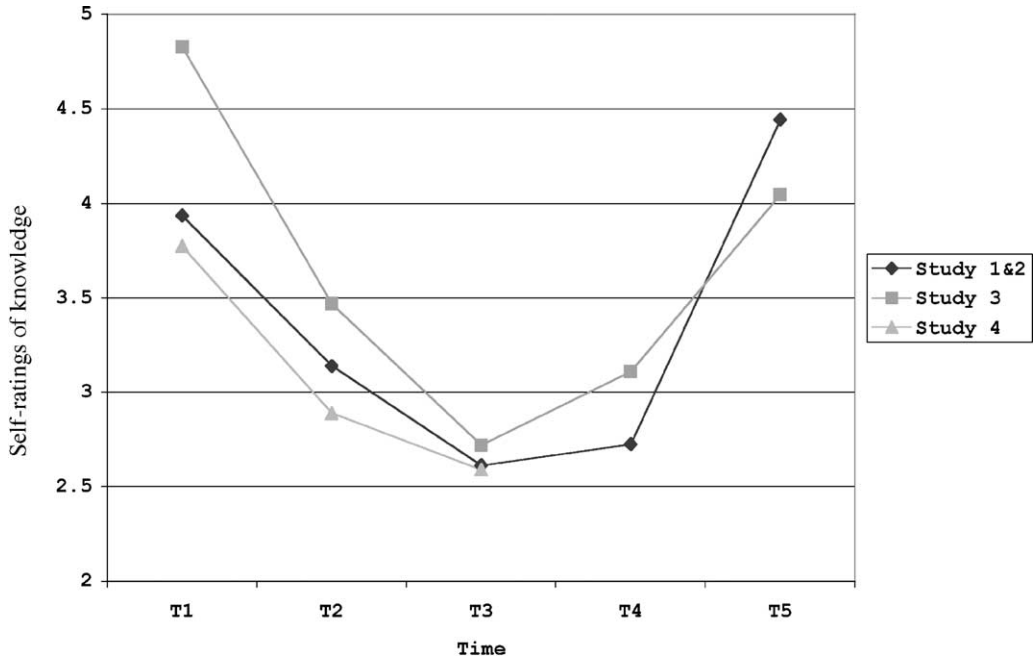


Fig. 4. Self-ratings of knowledge averaged across all items and subjects by time comparing Studies 1 and 2 (Yale students with original sub-set of items) with Study 3 (regional university students with original items) and Study 4 (Yale students with new items). The *x*-axis shows the sequence of self-ratings from Time 1 to Time 5 (T1–T5) as explained in Fig. 3.

test items is driving the effect. We addressed that possibility by using four alternative sets of test items, with identical instructions and procedures as Studies 1–3.

2.4.1. Methods

Four additional sets of items were selected from the list of 48 initial-ratings stimuli shown in Appendix A. Each set consisted of eight items. Diagnostic questions were constructed, and expert explanations were selected for each of the 32 new test items.

The procedures were identical to those used in Studies 1–3 until Time 3, the diagnostic question. To keep the total time for the experiment manageable, the expert explanations and the related ratings were omitted from this study. By omitting the last two ratings we could keep the time of the experiment slightly under 1 h, even with eight test items in each test item-set instead of four. Participants were provided with copies of the expert explanations of each test item after the debriefing.

Thirty-two Yale undergraduates participated in the study, eight participants for each test item-set. Participants in each of the four item-set conditions received a different set of eight test items after rating their knowledge of the initial set of 48 items.

2.4.2. Results

The pattern of results with the new stimuli was identical to the patterns we saw in Studies 1–3, and is shown in Fig. 4. A repeated measures ANOVA that considered time and item

as within-subject factors and item-set as a between-subject factor showed time to be highly significant, $F(2, 42) = 22.695$, $p < .001$, $\eta^2 = .519$.

We also directly compared the results of Study 4 with those of Studies 1 and 2, using a repeated measures ANOVA with time as a within-subject and study as a between-subject variable. To maximize power, the data from Studies 1 and 2 were again combined into a single comparison group, $n = 49$.

The analysis showed no differences between the two results. That is, the TIME × STUDY interaction was not significant, $F(2, 158) = .562$, $p = .571$, $\eta^2 = .007$. Of course, TIME remained highly significant in the combined analysis, $F(2, 158) = 69.36$, $p < .001$, $\eta^2 = .468$.

2.4.3. Discussion

There were large and significant drops in participants' estimates of their knowledge as a result of trying to explain the devices. The drops were identical in magnitude and direction to those seen with the original set of eight test-item devices in Studies 1–3. The identical average findings across different stimulus sets suggest the described effect should generalize to other items of similar type.

Is it possible that the passage of time alone is enough to produce the observed drops in knowledge-ratings? Two kinds of data make that account unlikely. First, we saw a substantial increase in self-ratings of knowledge at T5, after reading the expert explanations. Second (as described in Studies 7–10), we found no decrease in knowledge ratings over time with procedures or narratives, and a significantly smaller drop with facts.

2.5. Study 5: Measuring the illusion by rating participants' explanations

Thus far, we have demonstrated the IOED with devices using variants on a single method. This leaves open the possibility that we are looking at an artifact of our testing procedure. While subsequent studies (7–12) with different types of knowledge help rule out that possibility, we thought it useful to include an alternative measure of the effect in a separate study. This study also tests one set of alternative explanations for the drop in confidence: that decreased self-ratings are caused by becoming more cautious or modest as a result of feeling "challenged" by the experimenter rather than by an accurate recalibration of understanding as a result of an effort to explain.

In this study, an independent and naïve group of participants rated all the explanations offered by participants in Study 2. If the proffered explanations were indeed shallow, then independent raters should detect the shallowness. Further, if participants in Study 2 were miscalibrated in the direction of overconfidence at T1, as we claim, independent ratings of the explanations should be more closely related to self-ratings given after T1 than at T1 itself.

2.5.1. Methods

Twelve naïve Yale undergraduates were recruited to participate. The undergraduates were selected from the same participant pool as the participants in Study 2 to ensure that the two populations were comparable.

Each of the 12 new raters received the exact same training on the 7-point knowledge scale as that used with the Study 2 participants. After the raters completed the training, they rated the

quality of explanations provided by the original participants on the 7-point scale. In the second part of the study, the raters read the expert explanations for each phenomenon, and then re-rated the original participant's explanation of the phenomenon in light of the expert explanations. So, the raters judged each explanation provided by the Study 2 participants twice: once before reading the expert explanation, and once after.

2.5.2. Stimuli

All the explanations offered by participants in Study 2 were put into a FileMaker Pro database. Since Study 2 had 33 participants, each rater rated all 33 explanations from T2, once before reading the expert explanations (IR.1-T2), and once after (IR.2-T2). The data base was set up in such a way that only one explanation appeared on the screen at a time. The order of explanations was randomized for each rater. Navigation buttons led the participant through the experiment in the correct sequence.

2.5.3. Results

The independent ratings were closer to self-ratings at T2 and later than to self-ratings at T1. We first looked at the average rating of the Time 2 explanation by the 12 raters, before they read the expert explanation: Independent Rating 1 at Time 2 (IR.1-T2). IR.1-T2 reliability across all 12 raters was high, $\alpha = .946$.

A repeated measures ANOVA was first used to compare the means of T1–T4 in Study 2 with IR.1-T2 to control for experiment-wise error. Each item was treated as a "subject/row," with the average rating for the item across all participants entered into each cell. Time was treated as a within-subject factor, with variables T1–T4, and IR.1-T2 entered into separate columns.

The overall ANOVA was highly significant, $F(4, 524) = 33.913$, $p < .001$, $\eta^2 = .206$. A set of simple contrasts comparing IR.1-T2 with the participant ratings at T1 through T4 shows that the independent ratings of accuracy are significantly lower than those provided by participants at T1 and T2 ($p < .001$, and $p = .005$, respectively), but are not significantly different from those provided by the participants at T3 and T4. These results are summarized in Table 1.

The second variable of interest was the average rating of the T2 explanations across all 12 subjects after the raters read the expert explanation: Independent Rating 2 at T2 (IR.2-T2). IR.2-T2 reliability across all 12 raters was high, $\alpha = .931$. These independent ratings were also closer to the later, than to the initial, self-ratings.

Table 1
Study 5: First-order correlations and means for self- and independent ratings of knowledge

	T1	T2	T3	T4	IR.1-T2	Mean	*SE*
T1	–					3.89	.18
T2	.64	–				3.10	.176
T3	.53	.82	–			2.49	.162
T4	.57	.82	.85	–		2.62	.165
IR.1-T2	.43	.65	.56	.52	–	2.72	.111
IR.2-T2	.45	.64	.56	.56	.93	2.44	.096

Note. All correlations are significant at the .001 level (two-tailed).

A repeated measures ANOVA was again used to control for experiment-wise error when comparing the means of T1–T4 with IR.2-T2. The overall ANOVA was highly significant, $F(4, 524) = 39.938$, $p < .001$, $\eta^2 = .234$. A set of simple contrasts comparing IR.2-T2 with the participant self-ratings at T1–T4 shows that the independent ratings of accuracy are significantly lower than those provided by participants at T1 and T2 ($p < .001$ for both), but are not significantly different from those provided by the participants at T3 and T4. These results are also shown in Table 1.

All correlations between self- and independent ratings were significant at the $\alpha = .001$ level. The pattern of correlations was informative, and is shown in Table 1. IR.1-T2 was most highly correlated with T2 ($r = .648$), and with T3 ($r = .557$), a bit less with T4 ($r = .523$) and least with T1 ($r = .428$). IR.2-T2 was most highly correlated with T2 ($r = .644$), and with T4 ($r = .563$), a bit less with T3 ($r = .555$) and least with T1 ($r = .449$).

We tested the differences between the non-independent correlations following the procedure endorsed by Steigler (1980). For both, IR.1-T2 and IR.2-T2, correlations with the T2 self-ratings were significantly larger than correlations with T1 self-rating ($p < .001$).

Finally, we checked for possible non-linear relationships between self-ratings and independent ratings by including the squares of each variable in the correlation matrix. The pattern of the results (i.e., the ordinal relationships and the significance of differences between correlations) was unaffected by the transformation. The correlations and means for the self-ratings at T1–T4, and for the Study 5 independent ratings, are summarized in Table 1.

2.5.4. Discussion

The means of the independent ratings were much closer to the later than to the initial self-ratings of participant's knowledge. Similarly, the correlations between the independent and the self-ratings were higher for the later self-ratings. These findings support our interpretation of the drops in self-ratings shown in Studies 1–4: the participants are becoming more accurate in assessing their knowledge, not merely less optimistic or more conservative when confronted by the experimenter.

The IOED seems to reflect a genuine miscalibration in people's sense of how well they understand the workings of the world around them. Independent raters judged people's knowledge to be far more incomplete than the owners of that knowledge did prior to producing the explanations. The convergence of ratings between self and other in the later steps of the rating process confirms our interpretation of the findings by showing that, with additional information provided by the experimental manipulation, self- and independent rating tend to agree.

If the drops in self-ratings over time did not reflect an increasing awareness of the shallowness of one's knowledge and instead a different process, the pattern of linkage to ratings made by independent judges is likely to be different from that found here. For example, if the drops over time in Studies 1–4 reflected an increasing wariness in the face of being asked to support one's judgments and not a real change in evaluations of one's knowledge, independent ratings should be more closely linked to T1 ratings, rather than to later ratings, as they were in Study 5. Similarly, general drops over time in self-confidence or optimism about one's abilities should have led the independent ratings to be closer to the T1 self-rating than to the subsequent rating. The observed pattern suggest changes in self-ratings do, in fact, reflect accurate recalibration of one's knowledge as a result of an effort to explain.

2.6. Study 6: Reducing the illusion with explicit instructions

In the earlier studies, participants were given instructions and training on how to rate their knowledge, but they were not explicitly told, prior to the initial rating, that they would have to provide written explanations and answer diagnostic questions for some of the items they will rate. Suppose that we told participants at the outset that they would have to write out explanations and answer questions for some of the items? Participants might then be much more conservative, in anticipation of being tested, making the illusion disappear entirely because of the induced caution. However, if the illusion is robust, even the explicit warning might not be sufficient to do away with the illusion altogether. The warning should merely reduce the difference between self-ratings at T1 and T3. In Study 6, we tested those possibilities by providing participants with explicit descriptions of the testing procedure at the outset of the experiment.

Some of the following analyses (in this and in the following studies) compare data across different experiments. Cross-experimental comparisons are not a concern because the participants for all our experiments (unless explicitly manipulated) were recruited in similar ways and from similar populations. Since we worked with undergraduates, we were especially careful not to run any of our studies close to either the beginning or the end of the semester. Thus, the different studies are sufficiently similar to different conditions in a single experiment for equivalent treatment in analyses.

2.6.1. Methods

Thirty-one Yale undergraduates participated in the experiment. The stimuli were identical to those used in Studies 1–3, except that one paragraph was added to the instructions. The paragraph described the testing procedure, and warned the participants that they would have to provide written explanations and answer diagnostic questions for some of the items they were about to rate.

2.6.2. Results

First, we confirmed that the pattern of results with the new stimuli was analogous to the patterns we saw in Studies 1–4 (see Fig. 5). As before, the experimental manipulation led to a significant drop in self-ratings of knowledge from T1 to T3. A repeated measures ANOVA that considered time a within-subject factor and item-set as a between-subject factor showed time to be highly significant, $F(4, 116) = 44.11$, $p < .001$, $\eta^2 = .619$. The planned linear contrasts showed that the drop from Time 1 to Time 2 was not significant, but the drops from Time 1 to Time 3 and from Time 1 to Time 4 were both highly significant, $p < .001$.

We also directly compared the results of Study 6 with those of Studies 1 and 2, using a repeated measures ANOVA with time as a within-subject and study as a between-subject variable. To maximize power, the results from Studies 1 and 2 were again combined into a single comparison group, $n = 49$. Note that we considered only the first four ratings (T1–T4) in the subsequent analyses, since the 5th rating (T5) was a manipulation check, and was not relevant to the hypotheses being tested. However, including the 5th rating does not change the substance of the results.

The analysis showed significant differences between the two results: the drop was smaller in Study 6 then in Studies 1 and 2. That is, the TIME $\times$ STUDY interaction was significant,

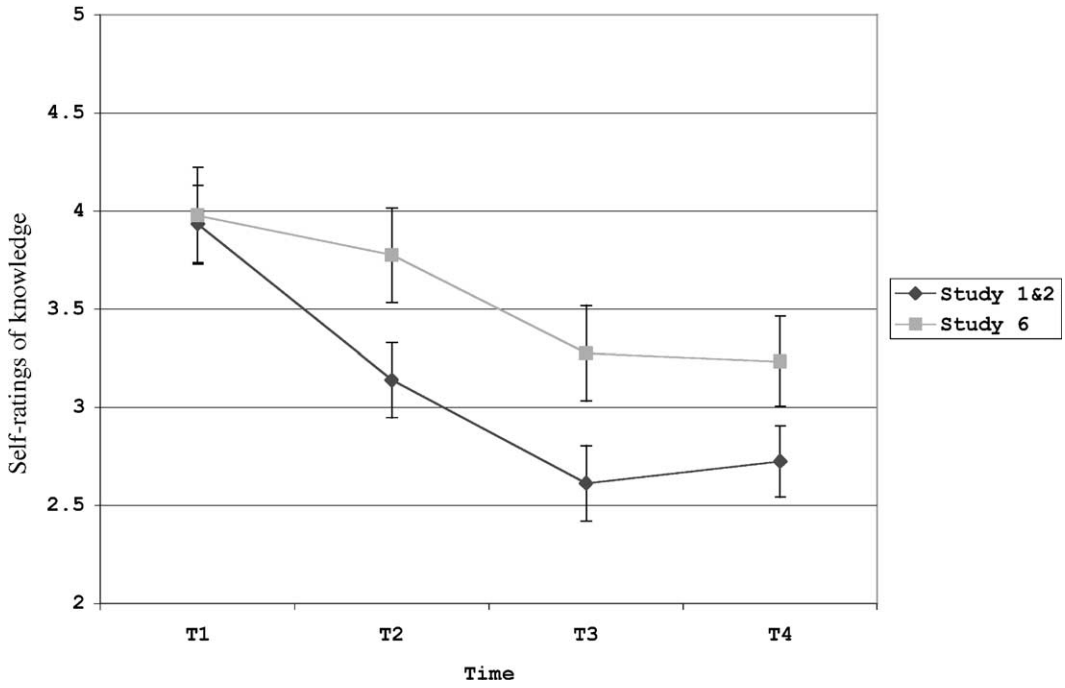


Fig. 5. Studies 1 and 2 (Yale students with devices) compared to Study 6 (Explicit warning). The *x*-axis shows the sequence of self-ratings from Time 1 to Time 4 (T1–T4) as explained in Fig. 3.

$F(3, 234) = 4.119$, $p = .007$, $\eta^2 = .050$, and time remained highly significant, $F(3, 234) = 44.924$, $p < .001$, $\eta^2 = .365$. As Fig. 5 shows, the drop in the explicit study was significant, but not as large as in the earlier studies.

2.6.3. Discussion

Offering an explicit warning about future testing reduced the drop from initial to subsequent ratings. Importantly, the drop was still significant—the illusion held even with the extreme warning.

One unusual feature of the data deserves mention. We might have expected the manipulation to result in the lower initial (T1) ratings of knowledge, but it did not. Instead, the magnitude of the effect was reduced because T2 and T3 ratings were higher in this study compared to Studies 1–4. This pattern is somewhat unexpected. One possibility is that the new instruction changed the way participants used the rating scales. For example, hearing the explicit instructions may have caused participants to try to be more consistent with their subsequent ratings because they had less justification for being surprised at their poor performance.

2.6.4. General discussion of Studies 1–6

So far, we have demonstrated that an IOED exists with devices, that it is robust across various populations, that it can be measured in alternative ways, and that it is resistant to changes in instructions that should reduce overestimation of one's knowledge. But two major questions

remain: How does the illusion vary in strength across domains? What factors influence the magnitude of the illusion both within and across domains? Our first step in exploring these questions was to study how well people estimate knowledge in other domains by considering knowledge of facts.

3. Calibration of comprehension across knowledge domains

Although the drop in ratings of knowledge might reflect a more general overconfidence in one's abilities, we predicted that the IOED is separate from, and additive with, general overconfidence effects. As described earlier, intuitive theories about causally complex phenomena may have several distinctive features that may promote overconfidence.

First, people may underestimate the environmental support for explanation. We can explain systems when they are in front of us by figuring out relations on the fly. But then, we may mistakenly assume that much of the information contained in the objects themselves is actually stored in our heads.

The confusion of environmental support for representation may be most pernicious in cases where the mechanism is perceptually vivid. For example, if components of a mechanical device are easy to visualize and mentally animate, they may cause a false sense of knowing the details of how the components interact. The more one is aware of discrete, easy to imagine parts of a system, the more one may be inclined to attribute deep causal knowledge of a system to oneself. Causally complex systems, on the average, may also have more perceptually salient components than procedures or facts. We might make one other prediction: among various causally complex systems those with more perceptually vivid component sub-processes should produce the greatest illusion of comprehension. Vivid perceptual components may mislead people into thinking they have vivid memories and thus a vivid representation of mechanism.

Second, people may confuse understanding relations at a higher level of description with understanding more mechanistic relations at a lower level. For example, with explanations of devices it might be easy to confuse knowing the function of with understanding the mechanism. Knowing what something does (e.g., a transmission changes the amount of power an engine sends to a car's wheels) may be confused with knowing how it does it (e.g., the complex interactions among gears).

The second factor is clearly related to the first, and the confusion between higher and lower levels of understanding may be most seductive when sub-processes are perceptually vivid. Similarly, we would expect the confusion of levels to be stronger when people have ready labels for components or sub-processes, and weaker when they do not. With devices, for example, people may confuse knowing labels for parts of a system with a full understanding of how the parts interact. If one knows the names of components, such as "hard-drive" "CPU" "RAM" and "PCI Bus," one may mistakenly conclude one also knows the causal relationships between the named components.

Like factor 1, factor 2 allows us to make two sorts of predictions: (1) about differences between knowledge domains (stronger illusion for explanations than for facts or procedures), and (2) about differences within the domain of explanations. We would predict that knowing

names of parts for a device would lead to a stronger illusion of understanding. We develop this point further in Study 12.

Third, explanations of complex causal systems have indeterminate end states leading to greater difficulty in self-testing of one's knowledge. Much time and introspection may be required for complex systems, and there may be no easy substitute for attempting to provide an explanation and then assessing the quality of the product. In contrast, determining whether one knows a procedure or a fact may be relatively straightforward. Where self-testing is less-than-trivial, people may use culturally prevalent theories about how much a person ought to know about different things (rather than any idiosyncratic information about their internal states) to make the initial estimates of their knowledge (see, e.g., Nisbett & Wilson, 1977).

Fourth, under normal conditions most of us rarely give explicit explanations conforming to our intuitive theories of everyday phenomena. Thus, we cannot usually check our memories of whether we have been successful in providing good explanations in the past.

The four subsequent studies (7–10) explore these factors by comparing people's performance in the explanations studies with performance in other knowledge domains. The final two studies (11 and 12) look at factors that may produce an illusion of depth with explanations more directly, by examining the variation between different items used in the earlier studies.

3.1. Study 7: Ruling out general overconfidence—factual knowledge

Our first task was to rule out general overconfidence as sufficient to account for our findings. In Study 7, we asked participants to estimate how well they know facts, keeping the rest of the design as similar as possible to Studies 1–4. Facts are unlike explanations in a number of ways. Participants could not confuse environmental support with internal representation, equate function with mechanism, mistake knowing labels with understanding causal relationships, or have too little experience with retrieving facts. However, if the results in Studies 1–4 are due to general overconfidence about how much people know, those results should also hold for facts.

3.1.1. Methods

We selected 48 representative countries from the list of all the world's nation states. The countries' capitals ranged from obscure to well-known. We selected the countries from the larger list through piloting so that about one-third of the capitals were obscure (e.g., Tajikistan), one-third intermediate (e.g., Brazil), and one-third well-known (e.g., England) to American undergraduates.

We asked 52 college undergraduates to rate how well they knew the capitals of each of the 48 countries, using a 7-point scale. The scale was explained in a manner analogous to the training instructions used in Studies 1–4. We then asked the participants to provide the names of each of the capitals of 24 of the 48 countries (the test items). The participants then re-rated their knowledge for the 24 test items. Finally, we told the participants the actual names of the capitals for the 24 test-item countries, and asked them to re-rate their knowledge once more.

The phases of Study 3 are analogous to Phases 1, 2, 4, and 5 of the devices studies. Phase 3 of Study 1 had no analog because we could not ask diagnostic mechanism questions about facts.

3.1.2. Results

In order to compare the Facts domain with the Devices domain we had to examine the first two ratings. To maximize power, data from Studies 1–4 were combined into a single comparison group, $n = 97$. Collapsing the data across studies was justified conceptually and statistically: Studies 1–4 measured the same thing, and the data showed no significant differences among the studies on the first two ratings (T1 and T2).

Note that the decision to include the regional university sample from Study 3 may be somewhat controversial, since that group did show a significant difference from the baseline on latter measures (T4–T5), although not on the first two ratings used in this analysis. However, excluding the regional university sample from the analysis does not change the substance of the results.

Self-ratings of knowledge did decrease from T1 to T2, but the drop was significantly smaller than with devices (see Fig. 6 and Table 7). A repeated measures ANOVA that considered time as a within-subject factor showed time to be significant, $F(2, 100) = 4.488$, $p < .023$ (with the Greenhouse–Geisser correction for sphericity), $\eta^2 = .082$. The planned linear contrasts

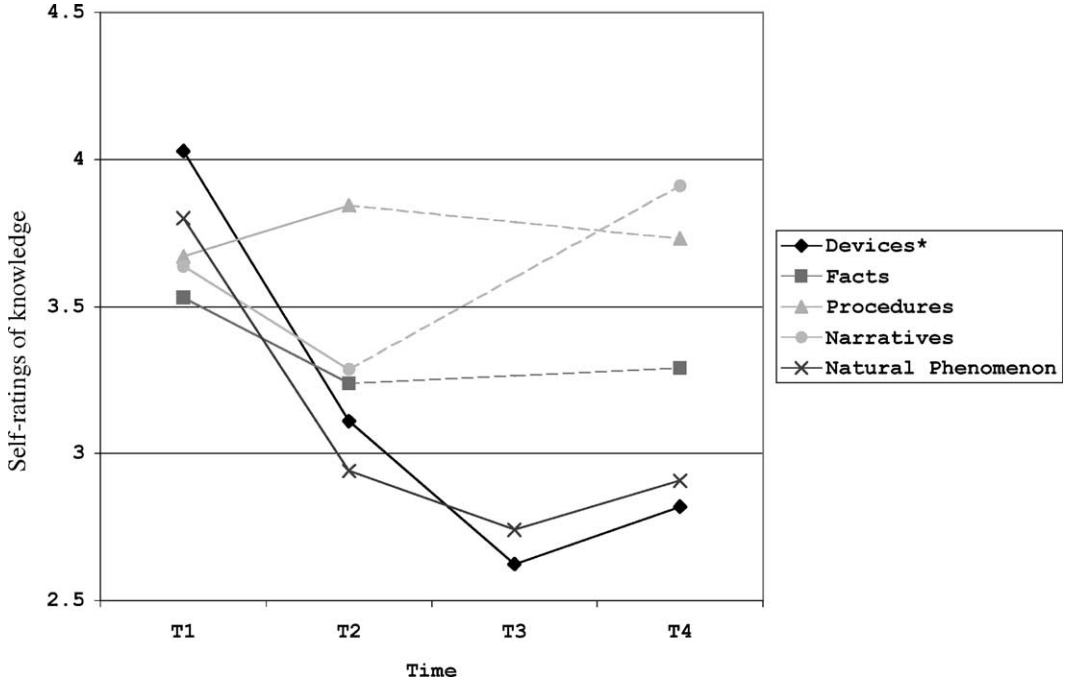


Fig. 6. Self-ratings of knowledge averaged across all items and subjects by time for Devices (Studies 1–4), Facts (Study 7), Procedures (Study 8), Narratives (Study 9), and Natural Phenomena (Study 10). The *x*-axis shows the sequence of self-ratings from Time 1 to Time 4. T1 is the initial self-rating, prior to any attempt to produce an explanation. T2 is the rating immediately after the first effort to explain. T3 is the rating after an attempt to answer a diagnostic question, and was measured only with Devices (Studies 1–4) and Natural Phenomena (Study 10). T4 is the re-rating of one's initial knowledge provided after reading an expert explanation. (Note: T1–T3 means for devices show combined data from Studies 1–4. The devices mean at T4 only shows data from Studies 1–3 because Study 4 did not include T4.)

showed that the drop from Time 1 to Time 2 and to Time 3 were significant, $p = .014$ and $p = .001$, but the drop from Time 2 to Time 3 was not.

We directly compared the results of Study 7 with those from the Devices studies, i.e., Studies 1–4, using a repeated measures ANOVA with time as a within-subject and study as a between-subject variable. To maximize power, the results from Studies 1–4 were again combined into a single comparison group, $n = 97$. Note that we considered only the first two ratings in the subsequent analyses, since only those ratings were completely equivalent across domains.

The analysis showed significant differences between the two results. That is, the TIME × DOMAIN interaction was significant, $F(1, 147) = 15.471$, $p < .001$, $\eta^2 = .095$. Time remained highly significant for the combined data set, $F(1, 147) = 56.534$, $p < .001$, $\eta^2 = .278$. As Fig. 6 shows, the drop from T1 to T2 in the Facts study was significant, but the magnitude of the drop was also significantly smaller than the ratings drop in the devices studies.

3.1.3. Discussion

The drop in the ratings of knowledge was significantly less with facts than with causally complex phenomena (Studies 1–4), as shown in Fig. 6. For easy reference, the overconfidence values (T1–T2) across all domains (Studies 1–4 and 7–10) are also summarized in the last row of Table 7. Equally important, the overall mean ratings in Study 7 were roughly the same as those for devices (i.e., there was no main effect for study), ruling out ceiling and floor effects.

Note also that the ratings for capitals were not bimodal (with some subjects simply picking "1" and others picking "7" on the scale). Instead they were relatively normally distributed as participants were often unsure, for example, which of several prominent cities in a country was its capital.

3.2. Study 8: Ruling out general overconfidence—knowledge of procedures

The smaller drop in confidence for facts than for explanations suggests that people do not inflate their feelings of understanding *equally* across all knowledge. However, it tells us little about why the differences exist. Facts about capitals lack components, or complex relationships, but they are also different from causal phenomena in other ways. For example, facts may be easier to self-test, and may have been successfully recalled more often than causal explanations.

In Study 8 we asked participants about a domain closer in complexity to causal explanations: knowledge of procedures. Knowledge of procedures does require remembering a sequence of steps ordered in time, often with considerable relational structure. That sort of knowledge however, typically differs from explanatory knowledge, in several ways. First, the function versus mechanism, or levels of analysis confusion, is much less likely to be present as most procedural knowledge consists in knowing the sequence of highest level functional units that will get the job done. Second, there is less opportunity to confuse environmental support with internal representation for procedures. Thus, in many cases, although the objects used in a procedure may be observable to a person, their relational patterns may not be recoverable from inspection in a manner that can be true for many complex systems. Knowing how to fold a flag, make an international phone call, or file one's income taxes may not be helped much by having

the relevant objects in view. Third, the end states and ways of self-testing procedural knowledge are likely to be more apparent. The criteria for successful execution of a procedure are usually easier to judge in advance than those for giving a full explanation. Finally, we engage in doing procedures more often than in providing explanations and can therefore examine whether or not we have done such a procedure in the past or at least observed someone else doing it.

3.2.1. Methods

We asked 37 participants in Study 8 about their knowledge of various real-world procedures, such as how to bake chocolate chip cookies from scratch, how to tie a bow tie, or how to drive from New Haven to New York City. As in Studies 1–4, participants were first trained on the 7-point scale using an example of a multi-step procedure. Then the participants rated how well they understood each of the 10 procedures (shown in Appendix B).

After the initial rating, participants were asked to write out the procedure (e.g., describe step-by-step how to bake chocolate chip cookies from scratch) and then to re-rate their knowledge. Unlike the participants in Studies 1–4, the participants in Study 8 did not have to provide the causal connections between the steps, since the kinds of procedures queried often lacked such connections. They merely had to provide the steps in the correct sequence and in as much detail as possible. (The relevant instruction for the "Write Explanations" phase in Study 8 can be found in Appendix B.) Finally, they were then given a written description of the procedure produced by an expert (we found the "expert" descriptions on well-established "how to" sites on the Internet, and edited them to suit our needs) and asked to re-rate their knowledge one final time.

As in Study 7, only the first two measurements (T1 and T2) were completely equivalent to their counterparts in the devices studies; the later measurements were analogous, but not equivalent, because we did not ask a "diagnostic question" at T3 with procedures, as we had with devices. The complete stimuli and instructions for Study 8 can be found in the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/.

3.2.2. Results

As shown in Fig. 6, there was no drop in the self-ratings across time, a significant difference from Studies 1–4. That is, a repeated measures ANOVA for Study 8, with time as a within-subjects factor was not significant, $F(2, 70) = 1.598$, $p = .210$, $\eta^2 = .044$.

To compare results for procedures with those for devices, we used an overall repeated measures ANOVA to compare data from Study 8 to data collapsed across Studies 1–4. Domain was a between-subject factor with two levels (Procedures vs. Devices) and time was as a within-subject factor with two levels (T1 and T2). The TIME $\times$ DOMAIN interaction was highly significant, $F(1, 132) = 40.934$, $p < .001$, $\eta^2 = .237$. The initial levels of confidence were the same as for Studies 1–4, making ceiling and floor effects unlikely explanation for the stability in knowledge ratings.

3.2.3. Discussion

A starkly different pattern of results appeared with procedures. Instead of a drop in knowledge estimates we saw a slight (non-significant) increase. In sharp contrast to Studies 1–4, none of the participants in debriefing expressed surprise at how little they knew; they knew

as much as they expected to know. People appear to be quite accurate in assessing their own knowledge of "cookbook" procedures in contrast to their knowledge of how things work.

3.3. Study 9: Calibration with knowledge of narratives (plots of movies)

The previous two studies showed significantly less overconfidence with procedures and facts than we saw with explanations in Studies 1–4. But both facts and procedures are unlike explanations in lacking a real narrative structure. Given that theories are sometimes characterized as explanatory stories, perhaps any sort of narrative structure will produce an illusion of deep understanding. Alternatively, causal explanations may have a rich, deeply embedded structure that is different from that governing most narratives. Narratives do have some degree of hierarchical structure (e.g., Mandler & Johnson, 1977) but not with the same level of depth as most complex systems. In addition, the sense of mechanism seems much less critical to narratives.

Ahn and Kalish (2000) have argued persuasively that a sense of mechanism is essential to everyday causal reasoning. Causal explanations of phenomena invoke people's intuitive theories of mechanism in a way that other types of narrative do not.

Because our intuitive theories about mechanisms may be shallow—far shallower than our meta-cognitions lead us to believe—they result in the IOED when we mistake shallow, sparse lay theories of mechanism for deep, rich ones. Thus, the centrality of mechanism to causal explanations suggests that illusions of knowing will be stronger for causal explanations than for other narratives. In addition, narratives are less likely to involve environmental support versus representation confusions and people are much more practiced at giving narrative descriptions than explanations. We tested this issue explicitly in Study 9.

3.3.1. Methods

We asked 39 participants about their knowledge of plots of some popular movies. Twenty-four of the participants were Yale undergraduates and 15 were undergraduate and graduate students at a regional university. (Note that because of unexpected time constraints only four of the regional university students completed all phases of the experiment; however, all 15 provided the first two data points relevant to inter-domain comparisons.)

The list of 20 popular movies (those most likely to have been seen by undergraduates) was selected through piloting. As in Studies 1–4, participants were first trained on the 7-point scale using an example of a movie plot description. The participants were then asked to mark which of the 20 movies on the popular movie list they had seen. Then the participants rated how well they understood the first five movies on the list that they had seen.

Next, participants were asked to write out a description of the plot for each of the five movies they had rated, and then to re-rate their knowledge of each movie. Finally, they were given a written "expert" description of the plot, written by professional reviewers, and asked to re-rate their knowledge once again. The "expert" descriptions were obtained from a movie-review web site, and were largely uniform as to level of detail and style. The description used during the training was an edited version of an expert description of the plot of *Good Will Hunting*.

3.3.2. Results

Unlike with devices, but like with procedures, there was no drop in self-rating as a result of an effort to explain. A repeated measures ANOVA with time and item as within-subject factors showed time to be highly significant but only if Time 5, the manipulation check, was included, $F(3, 48) = 16.989$, $p < .001$, $\eta^2 = .515$. However, the planned linear contrasts showed no significant differences between ratings at Times 1 and 2, or between ratings at Times 1 and 3. Rating 3 was significantly different from Rating 2. But unlike the Devices studies, Rating 3 was significantly higher than Rating 2.

We also directly compared the results of Study 9 with those from the Devices studies, i.e., Studies 1–4, using a repeated measures ANOVA with time as a within-subject and domain as a between-subject variable. To maximize power, the results from Studies 1–4 were again combined into a single comparison group, $n = 97$. As in Studies 7 and 8, we considered only the first two ratings (T1 and T2) in the direct cross-domain analyses since only those ratings were completely equivalent across domains.

The analysis showed significant differences between the two results. That is, the TIME × DOMAIN interaction was significant, $F(1, 134) = 6.146$, $p < .014$, $\eta^2 = .044$. TIME remained highly significant, $F(1, 134) = 47.275$, $p < .001$, $\eta^2 = .261$.

A similar analysis showed a significant difference between narratives and procedures, but not between narratives and facts. A repeated measure ANOVA with time as a within-subject and domain as a between-subject variable showed the TIME × DOMAIN interaction was significant for narratives versus procedures, $F(2, 116) = 8.207$, $p = .001$, $\eta^2 = .124$, but not for Narratives versus Facts, $F(1, 89) = .496$, $p = .483$, $\eta^2 = .006$.

3.3.3. Discussion

As with knowledge of procedures, participants were relatively better calibrated about their knowledge of movie plots than they were about their knowledge of devices. The results are shown in Fig. 6. There was no drop in the self-ratings from Time 1 to Time 2, a significant difference from Studies 1–4. The initial levels of confidence were the same as for Studies 1–4, ruling out ceiling and floor effects as possible reasons for stability. To stray a bit from the data, we again noticed a striking difference in participants' experience during debriefing: in sharp contrast to Studies 1–4, none of the participants expressed surprise at how little they knew; they knew as much as they expected to know.

Why was there a significant increase in knowledge estimates from Time 2 to Time 3? In retrospect, we think the expert explanations provided before the T3 ratings were, on the average, somewhat less detailed than the edited expert explanation used in the instructions. Thus, participants accurately recalibrated their understanding relative to the explanations provided immediately before Time 3. The unexpected effect gives us additional confidence about the sensitivity of the IOED procedure.

3.4. Study 10: Calibration with knowledge about natural phenomena

The pattern of results so far indicates a special difficulty in calibrating one's explanatory knowledge about devices. In the studies with procedures and movies, the participants were well calibrated. In the study with factual knowledge about capitals, the

participants were overconfident but markedly less so than with explanatory knowledge of devices.

Are devices unique? Is there something about intentionally built complex causal chains that lures us into a premature feeling of comprehension? One possibility is that the designed nature of devices is a major cause for the illusion. With designed devices, multiple levels of function encourage function/mechanism confusions. With non-living natural phenomena, functional explanations are not usually appropriate (e.g., the tides don't happen as they do for a functional reason). An alternative is that causally complex systems create illusions of explanatory understanding, whether intentionally designed or not, because they allow individuals to encode explanations at several distinct levels of analysis that involve stable sub-assemblies. To test between these alternatives we ran one final study in the IOED series, this time using natural phenomena as stimuli.

3.4.1. Methods

Thirty-one Yale undergraduates participated in the study. The design of Study 10 was identical to that of Studies 1–4, except that instead of complex devices participants were asked about a set of 24 natural phenomena such as "how earthquakes occur," "why comets have tails," "how tides occur," and "how rainbows are formed." (The list of all phenomena, along with the complete instructions, is included in Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/.)

As in the first four studies, participants were first trained on a knowledge-rating scale. Then they provided initial ratings of the entire stimulus set. Next, participants produced written explanations for a sub-set of five phenomena, 10 total between-subjects (e.g., explain how tides occur). Then they answered a diagnostic question designed to probe for detailed understanding of each phenomenon (e.g., why are there two tides everyday?). Finally, participants read an expert explanation of each phenomenon, the explanations were obtained from science education sites on the Internet. As in the devices studies, participants re-rated their knowledge of the phenomena after each step.

3.4.2. Results

The results, shown in Fig. 6, were similar to those obtained in the devices studies. A repeated measures ANOVA with time as a within-subject factor and item-set as a between-subject factor showed time to be highly significant, $F(4, 112) = 73.698$, $p < .001$, $\eta^2 = .725$. The planned linear contrasts showed that ratings at Times 2, 3, 4, and 5, were all significantly different from rating at Time 1, with rating at Time 5 being significantly larger, and the rest being significantly smaller, as in the devices studies.

We also directly compared the results of Study 10 with those from the Devices studies, i.e., Studies 1–4, using a repeated measure ANOVA with time as a within-subject and domain as a between-subject variable. To maximize power, the results from Studies 1–4 were again combined into a single comparison group, $n = 97$. Note that we considered only the first four ratings in the subsequent analysis since only those ratings were relevant to the hypotheses tested.

The difference between the Devices and Natural Phenomena studies was not significant, but did show a trend towards significance: the TIME × DOMAIN interaction, $F(3, 279) = 2.612$,

$p < .07$ (with the Greenhouse–Geisser correction for sphericity), $\eta^2 = .027$, suggested a somewhat larger drop for devices.

3.4.3. Discussion

The pattern of results was similar to those found in Studies 1–4, as seen in Fig. 6. The overconfidence values (T1–T2) across domains are also summarized in Table 7. The drop in knowledge estimates over time was significant and much greater than that seen in Studies 7–9. There was also a suggestive pattern of a somewhat smaller drop for natural phenomena at a marginal significance level of $p < .07$. The findings may, therefore, be taken to suggest a function mechanism confusion as a small factor in overconfidence for devices: the lack of functional descriptions in most accounts of non-biological natural phenomena is a small factor that modestly decreases the magnitude of the illusion in comparison to devices.

To summarize, studies with devices and natural phenomena both show large drops in knowledge estimates. Procedures and Narratives show no drop, while Geography Facts show only a small drop. The results demonstrate large differences in knowledge calibration across knowledge domains, casting serious doubt on the meaningfulness of "general overconfidence" about knowledge. The studies also raise intriguing possibilities about the mechanism behind overconfidence, which we address in the next few studies.

4. Exploring the causes behind the illusion

4.1. Study 11: Ruling out desirability as the explanation for inter-domain differences

In Studies 7–10 we have found large differences in knowledge calibration across knowledge domains. We argue that these differences are systematic and result from the structural properties of how different types of knowledge are represented in the mind. However, it is possible that other factors influence calibration. One alternative explanation for cross-domain differences is that having detailed knowledge in some domains is more socially desirable than in others and that people, therefore, inflate estimates of knowledge in the more desirable domains. To test the desirability account, we asked another set of participants to rate how desirable it would be to have knowledge of each item used in the previous studies, or, more precisely, how undesirable it would be for them to have to admit ignorance of each item.

4.1.1. Methods

Twenty-four Yale undergraduates participated in the study. The participants rated on a 7-point scale how embarrassed they thought they would be if they had to admit not having a good knowledge or understanding of an item. The question was framed as rating "embarrassment over ignorance" because that phrasing seemed to most directly tap the participant's motivational experience in the IOED studies. The stimuli consisted of a combined list of all test items used in the previous studies: devices, facts, procedures, narratives, and natural phenomena. The participants were given instructions on a 7-point "embarrassment" scale, and were asked: "For each item, please rate how embarrassed you think you would be if someone

asked you to explain that item and it turned out that you did not have a good understanding or knowledge of that item." The complete list of items and the instructions can be found with the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/.

4.1.2. Results

We found significant differences in desirability ratings across domains. However, the differences did not predict the relative magnitudes of the illusion in each domain.

The main dependent variable was the "average embarrassment rating" for each of the 103 items, collapsed across all 24 participants. A one-way ANOVA was conducted on the embarrassment ratings with knowledge domain as the between-subjects factor. The ANOVA was highly significant, $F(4, 102) = 9.507$, $p < .001$, $\eta^2 = .280$. The means and confidence intervals are shown in Table 2.

Post hoc comparisons of means, using Scheffe's procedure, indicated that embarrassment for narratives was significantly higher than for devices and for facts ($p < .001$ and $p = .024$, respectively) but that facts and devices did not significantly differ from each other, nor did any of the other pairs of means (see Table 2). Importantly, embarrassment did not significantly correlate with overconfidence (mean T1–T2 for each domain), $r = -.185$, $p = .79$.

4.1.3. Discussion

As shown in Table 2, Devices produced the lowest embarrassment scores, and Narratives produced the highest. The means for the Procedures, Natural Phenomena, and Facts items were not significantly different from each other and were intermediate between the low Devices and the high Narratives means.

The pattern of mean differences makes it unlikely that social desirability is the major factor behind the differences in overconfidence observed between knowledge types. Devices produced the most overconfidence but the lowest desirability. Movies produced the least overconfidence but the highest desirability. Natural Phenomena produced high overconfidence and intermediate desirability, while Geography Facts produced low overconfidence and intermediate desirability. The divergent patterns make it more plausible that factors other than desirability are driving the differences between knowledge types. If anything, high desirability may cause people to more carefully assess their self-knowledge in a domain and, therefore, be more accurate.

Table 2
Study 11: Social desirability ratings by knowledge domain

Knowledge type	Mean	*SE*	95% CI lower bound	95% CI upper bound
Devices	2.73 a	.16	2.42	3.04
Facts	3.37 b	.20	2.97	3.76
Procedures	3.52 b	.31	2.90	4.14
Natural Phenomena	3.69 b	.31	3.07	4.32
Narratives	4.44 c	.24	3.96	4.91

Note. Larger numbers indicate participants would be more embarrassed not to know the answer to this type of question, indicating the knowledge is more socially desirable.

Means marked with the same letters are not significantly different from each other.

4.2. Study 12: Correlates of confidence and overconfidence

Studies 7–10 examined several factors that induce a stronger illusion of understanding for explanations than for other knowledge types and suggested some factors that might be responsible for the illusion. An alternative, and complementary, way to test the validity of the hypothesized factors is to look at individual items within a domain where the illusion is large.

We proposed in Section 1 several factors that might lead to an especially strong illusion of knowledge for explanations of causally complex phenomenon. Some of the most prominent include confusing environmentally supported information with mentally represented information and confusing higher and lower levels of understanding which may lead to confusing knowing easily visualized and labeled parts with deeper mechanistic understanding. We also wanted to test two common-sense explanations for the illusion: (1) whether familiarity with an item explains the magnitude of the illusion, and (2) whether sheer complexity of an item, as measured (for example) by the total number of parts, would predict the magnitude of the illusion.

Study 12 examines which factors predict an especially strong illusion of knowledge for explanations of devices by looking at differences in initial confidence and in overconfidence across different devices.

In order to study the factors, we had to operationalize them as answers on a rating scale. Five rating scales were developed to explore what factors contributed most to the illusion within the domain of devices. The detailed instructions and the rating scales can be found in the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/. We then asked a new set of participants to rate each of the device items used in Studies 1–4 on the five new rating scales.

4.2.1. Methods 12a: Correlates of confidence

To examine what factors might be related to a sense of deep knowledge, we asked a group of 54 participants (Yale undergraduates recruited in the same manner as in previous studies) to rate all device items used in Studies 1–4. Forty items of the original 48 were rated on all the scales. Eight items were excluded from some of the scales because they were not devices, but other kinds of causally complex systems (e.g., the human liver, or the presidential election process). Consequently, the rating scales developed for devices (e.g., number of parts, mechanical vs. electrical) would have been confusing to the participants with those items. (Of course, these eight non-device items were never used as test items in the devices studies.)

The 40 devices were rated on five scales: (1) familiarity with the item, (2) the ratio of visible versus hidden parts, (3) the number of mechanical versus electrical parts, (4) the total number of parts, and (5) the number of parts for which one knows names. We also computed the ratio of known parts to total number of parts (#5/#4) from the last two ratings.

The 54 participants were given detailed instructions on how to score each of the five attributes for all 40 items (from Studies 1–4) with different scales for each, depending on the kinds of attributes involved. The complete stimuli and instructions are available in the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/. The instructions for most scales were relatively brief. We present them here to help the reader better understand the rating task.

Familiarity. Please rate how familiar you are with the following devices/phenomena.

4.2.1.1. Hidden versus visible parts. Many devices, such as computers, hide almost all of their important working parts inside an outer container. On the other hand, when you use a device like a crossbow, you can clearly see almost all of the parts and how they work together. For each of the following items, please estimate the proportion of parts that are visible during the normal course of operation.

4.2.1.2. Electrical versus mechanical devices. Some devices can be easily classified as either electrical or mechanical, whereas other devices may be some combination of both. For each of the following devices, please rate the degree to which the device operates by electrical or mechanical principles.

4.2.1.3. Names of parts. How many different parts of the following items would you guess you could identify by name? Please estimate whether you could name:

0, 1–2, 3–5, 6–10, 11–25, 26–50, 51–100, 101–250, or more than 250 parts.

One set of instructions deserves detailed treatment. The "total number of parts" rating scale proved especially difficult to construct. In the early versions, the variance between-subjects was very large with estimates for number of parts in some devices ranging from 5 to millions. Asking participants to estimate the number of parts is potentially vague since participants may have very different ideas about what counts as a part. We found that, in order to obtain consistent ratings, we had to give explicit instructions in the following form:

> Every device is made up of a number of different "parts." For example, on one level, a computer is made up of a monitor, mouse, CPU unit, and keyboard. On an even deeper level, the computer is made up of a motherboard, power supply, RAM, hard drive, video card, etc. And every part that was just named is obviously made up of even more parts. While it may be impossible to arrive at an absolute number of parts making up a given device, it is safe to say that certain devices are made up of more parts than others (i.e., a computer has more parts than a toaster).
>
> We are looking for your subjective estimate of how many parts make up the following items. When you are trying to imagine how many parts something has, it may be helpful to think about how many parts you can disassemble the device into without breaking anything. You can also imagine a blow-up diagram of the item and estimate the number of parts you would see on that diagram.
>
> These pictures of a combination lock and its constituent parts may help illustrate what we mean by a "part".

We showed the participants a picture of a combination lock, and the picture of the same lock disassembled into its components, and then asked them to estimate the number of parts the various items on our list had on a log-type scale. They were asked to "Please indicate how many parts you think the following items have:

1–5, 6–10, 11–25, 26–50, 51–100, 101–250, 251–500, 501–1000, or more than 1000 parts.

We used this non-linear scale to reduce variance at the higher end. During piloting with open-ended responses we found that people's estimates of the number of parts for some complex items varied by several orders of magnitude.

How do these variables, (1) familiarity with the item, (2) the ratio of visible versus hidden parts, (3) the number of mechanical versus electrical parts, (4) the total number of parts, (5) the number of parts for which one knows names, and (6) the ratio of known to total parts, relate to our hypotheses? We predicted that three factors—underestimating environmental representational support, levels of analysis confusions, and mistaking knowing names of parts for knowing deeper causal roles—would drive the differences in initial confidence and in overconfidence. If these factors were sufficient to explain overconfidence, overall complexity, here measured by the estimate of the number of parts, should not influence the illusion, once other factors are controlled for.

We expected three other scales to affect initial confidence for explanatory knowledge. Two scales were designed to capture different aspects of the amount of representational support an object provides. The number of visible versus hidden parts, and whether the device operates on electrical or mechanical principles, predict how easy it seems to discover the mechanism of operation in real time by simply examining the device closely. A device with mostly visible parts (e.g., a can opener or a bicycle derailleur) allows an observer to reconstruct the device's mechanism in real time by examining the device and may therefore create the strongest illusion of knowledge; therefore ratings of the ratio of visible to hidden parts should influence initial confidence. Similarly, if a device operates largely on mechanical principles it would be easier to figure out on-the-fly than something packed with circuit boards. We weren't sure how strongly the two variables would correlate, but expected both to influence initial confidence if they were not highly collinear.

Two other measures, the number of known part names, and the (computed) ratio of known to total parts—were designed to test the hypotheses that subjects confuse knowing labels with understanding the mechanism. Knowing the names of many parts, or a large proportion of parts, in a device may mislead the observer into believing they understand the causal relationships between the parts. Knowing that my computer contains a hard drive, a CPU, an LCD screen, a SCSI Bus, and RAM, might be enough to give me a sense that I know how the computer works, even if I have very little understanding of how the parts interact.

In informal piloting we asked participants to guess what factors would make a difference in overconfidence. Two factors emerged most often from naïve intuitions of participants: familiarity and complexity (number of parts). Pilot subjects thought other participants would be more overconfident about familiar devices, and about complex ones. As mentioned above, we did not expect complexity to matter, once we controlled for other factors. However, we thought familiarity might be a possible factor, given its role in the FOK literature.

As in phase one of Studies 1–4, only item labels were provided to the participants, with no additional descriptions. After being instructed on the use of a particular scale, a table containing all the item labels (e.g., telephone, fireplace, transistor) was presented to participants for each rating. Participants first rated the familiarity of each of the items, then estimated the ratio of visible versus hidden parts, then rated whether the devices operated on electrical versus mechanical principles, then estimated the number of parts, and finally estimated the number of known part names. The order of the items in the tables was reversed for half the participants. The

Table 3
Study 12a: Final model regression coefficients

	B	*SE*	β	*t*	Significant
Constant	.376	.278		1.352	.184
Visible vs. hidden	.640	.079	.698	8.109	.001
Known names of parts	.163	.037	.321	4.394	.001
Known names:number of parts	1.611	.645	.215	2.497	.017

Note. Final model (adjusted-$R^2 = .777$) coefficients for the final model in a step-wise regression predicting initial confidence (T1) from a set of six independent variables measured in Study 12. The β values are the standardized regression coefficients, and can be interpreted as indicating the relative contribution of each dependent variable in explaining the variance in the independent variable.

complete list of items (along with instructions) can be found with the Supplemental Materials section at http://www.elsevier.com/gej-ng/10/15/15/show/.

4.2.2. Results 12a: Correlates of confidence

Ratings provided by all 54 Study 12a participants were averaged for each of the 40 device items. These average ratings were then used to predict the average initial confidence for that item in Study 2. We used a stepwise multiple regression analysis to determine which variables were most related to initial high levels of confidence. As seen in Table 3, the visible/hidden parts estimate, the number of known parts, and known:total parts ratio are the central factors, accounting for 78% of variance in initial confidence (model adjusted-$R^2 = .777$). Familiarity was not a significant predictor of initial confidence once the other three variables were included. Complexity, as measured by the number of parts estimate, was similarly non-significant.

First-order correlations between variables are shown in Table 4. The ratio of visible to hidden parts was highly correlated with the mechanical:electrical dimension, with known:total parts ratio, with the raw number of parts, and with familiarity. The high colinearity between the variables makes interpreting the step-wise regression less straightforward than desirable.

Table 4
Study 12a: First-order correlations between measured variables

	Initial confidence (T1)	Visible/ hidden	Mechanical/ electrical	Known parts	Number of parts	Known parts: number of parts
Visible/hidden	.815**	1				
Mechanical/electrical	.707**	.756**	1			
Known parts	.314*	.006	.073	1		
Number of parts	−.308*	−.506**	−.433**	.371*	1	
Known parts: number of parts	.570**	.531**	.577**	−.051	−.362*	1
Familiarity	.634**	.533**	.336*	.286	−.362*	.532**

* Correlation significant at the .05 level (two-tailed).
** Correlation significant at the .01 level (two-tailed).

Table 5
Study 12b: Final model regression coefficients

	B	*SE*	β	*F*-to-remove	Significant
Constant	−.975	.401	−.975	5.920	<.01
Visible vs. hidden	.639	.110	.687	33.921	<.001

Note. Final model (adjusted-$R^2 = .458$) coefficients for a step-wise regression predicting overconfidence (T1–T3) from a set of six independent variables measured in Study 12.

4.2.3. Methods 12b: Correlates of overconfidence

The previous study looked at what properties of the phenomena may predict initial confidence levels. Although initial confidence and overconfidence were strongly correlated in our study, our main interest is in the relative levels of overconfidence, and it would be desirable to measure them directly. Unfortunately, our initial studies made this difficult since we only had direct measures of the miscalibration, as opposed to measures of the initial confidence level, for a small sub-set of devices used as test items (eight). To solve the problem, we combined the knowledge self-rating data from Studies 2 and 4 (because those studies had the most similar populations—Yale undergraduates) to obtain measured levels of overconfidence for a much larger set of devices. The two studies together gave us direct measures of overconfidence for 40 different devices. Because the participants in Study 12a rated all 40 items on the five rating scales, we were able to use their ratings in this analysis as well.

4.2.4. Results 12b: Correlates of overconfidence

Overconfidence for each item was defined as the difference between the average T1 and T3 scores for an item across all subjects in Studies 2 and 4. The means of the ratings provided by the 54 Study 12 participants were then used to predict overconfidence for each item. We again used a stepwise multiple regression analysis to determine which variables were most related to the degree of overconfidence. The final model is shown in Table 5.

The visible/hidden parts ratio explained most of the variance in overconfidence, and adding other predictors did not significantly improve model fit. This may be partly due to high colinearity between some of the measures: the visible/hidden ratio was highly correlated with electrical/mechanical dimension, the number of parts, and the ratio of known to unknown parts. Thus, a simple regression predicting overconfidence from the visible/hidden ratio was highly significant (model adjusted-$R^2 = .472$, $p < .001$, $\beta = .687$). The first-order correlations between all variables are shown in Table 6. As in Study 12a, the high colinearity between some predictor variables complicates interpretation of the step-wise regression.

4.2.5. Discussion

This study helps explain why the illusion of knowledge may be especially powerful in cases involving explanatory understanding, especially in domains with compelling intuitive theories. It is in those cases that several unique factors converge to give a powerful, but inaccurate, feeling of knowing. For example, the prominence of visible, transparent mechanisms may fool people

Table 6
Study 12b: First-order correlations between measured variables

	Overconfidence (T1–T3)	Visible/hidden	Mechanical/ electrical	Known parts	Number of parts	Known parts: number of parts
Visible/hidden	.687**	1				
Mechanical/electrical	.565**	.766**	1			
Known parts	−.089	.036	.080	1		
Number of parts	−.378*	−.494**	−.425**	.381*	1	
Known parts:number parts	.493**	.523**	.578**	−.041*	−.363*	1
Familiarity	.426**	.505**	.339*	.336*	−.341*	.525**

* Correlation significant at the .05 level (two-tailed).
** Correlation significant at the .01 level (two-tailed).

into believing that they have understood, and have successfully represented, what they have merely seen.

Study 12a suggest that three of the factors we mentioned earlier—levels of analysis confusion, label-mechanism confusion, and confusing environmental support with representation—contribute to initial feelings of knowing. Study 12b indicates that mistaking environmental support for representation may be the single most important factor in determining which explanations will cause the greatest illusion of understanding. The findings indirectly suggest that this factor may also be primarily responsible for the large differences between explanations and other knowledge types.

We have not explored in these studies the low rate of producing explanations relative to other types of knowledge. Conceivably, people rarely explain and thus do not get sufficient feedback to learn that their explanations are poor. This leads to an interesting prediction we would like to explore in a future study: people who produce explanations frequently (e.g., teachers, expository writers) should be less subject to the IOED than those who produce them rarely.

5. General discussion

The studies in this article first demonstrated a strong illusion of explanatory depth with one set of causally complex systems: devices. We then showed large differences in knowledge calibration across knowledge domains, with high levels of overconfidence for devices and natural phenomena, modest overconfidence for geography facts, and no overconfidence for knowledge of narratives or procedures. We then directly explored some of the factors that might influence overconfidence within the domain of devices, finding evidence that the degree of causal transparency for a system (as measured by the ratio of visible to hidden parts) may be a critical factor behind the strong illusion of explanatory depth. Neither familiarity, nor complexity made a difference. (The lack of a familiarity effect might suggest that FOK cue familiarity models cannot explain the illusion.) Finally, the ratio of known to total parts made a difference in initial confidence, but not in overconfidence, leaving it unclear whether the confusion of labels with mechanism contributes to the illusion.

Table 7 summarizes the factors we think are important to calibration across knowledge domains and indicates their hypothesized relative contributions. Procedures and results for all 12 studies presented in this paper are summarized in Table 8 for the reader's convenience.

We have found that the ratio of visible to hidden parts is the best predictor of overconfidence for an item, implicating the representational support account of the illusion. Can the representational support also explain differences across domains? Why, for example, don't we get confused about representational support with procedures or narratives? It may be that phenomena that are easy to visualize, and especially those that are easy to mentally animate, trigger an especially strong illusion of understanding and these may be more common in the domain of causal explanations.

People might use the following useful heuristic: perceptions are more accurate than propositional inferences. When you imagine a can-opener cutting through the lid of a can, that mentally animated image feels a lot more like perception than like propositional reasoning or informal

Table 7
Factors proposed as responsible for increased overconfidence with explanatory knowledge relative to other knowledge domains

	Devices	Natural Phenomena	Narratives	Facts (Capitals)	Procedures
Ease of confusing environmental support with internal representation (e.g., vivid parts for interaction)	+++	+++	+		+
Ease of confusing levels of analysis (e.g., function w/mechanism)	+++	+++	+		
Ease of confusing labels with mechanism	++	+			
Indeterminate end state/difficulty of self-test	++	++	++		+
Lack of experience with production	++	++		++	
Desirability (from Study 11)	2.73	3.69	4.44	3.37	3.52
Overconfidence (T1–T2 from Studies 7–10)	.918	.860	.351	.291	−.173

Note. The number of plus signs indicate the estimated relative contribution of each factor to overconfidence. Several factors combine to produce extensive overconfidence with explanatory knowledge (e.g., explanations of Devices and Natural Phenomena). Desirability and overconfidence are measured variables from Studies 7–11.

inference. Thus, it would be easy to assume that you can derive the same kind of representational support from the mental movie that you could from observing a real phenomenon. Of course, the mental movie is much more like Hollywood than it is like real life—it fails to respect reality constraints. When we try to lean on the seductively glossy surface we find the façades of our mental films are hollow card-board. That discovery, the revelation of the shallowness of our mental representations for perceptually salient processes, may be what causes the surprise in our participants.

A better understanding of the mechanism behind the illusion should also enable us to find cases where the illusion is reversed, where people are substantially underconfident in their ability to explain. Presumably, in cases that are especially hard to mentally animate and have mostly invisible parts but that are otherwise easy to propositionally infer causal steps for, we might see underconfidence. Anecdotally, we have all had the experience where a gifted teacher shows us that we already know why something is the case when we inspect the consequences of our own beliefs.

One conclusion that can be drawn from this research is that the well-established blanket approach to overconfidence with "general knowledge" is almost certainly misleading. Large inter-domain differences in calibration imply that structural properties of knowledge have a powerful impact on the process of knowledge assessment. "General knowledge" is a chimera—a mythological composite creature. Taking it seriously distracts from interesting questions about how knowledge assessment works, and the theoretically important issues of how the structural properties of knowledge influence calibration.

Table 8
Summary of experimental methods and results

Study	Participants	Stimuli/procedures	Results
1	16 Yale graduate students	Devices Rate 48, explain four times (eight total explained between-subjects); rate understanding five times	Self-ratings of knowledge decrease after explanation
2	33 Yale undergraduates	Same	Same
3	16 grad/undergraduates at a regional university	Same	Same
4	32 Yale undergraduates	New devices (four sets of eight items), same procedure	Same
5	12 Yale undergraduates used as raters	Read and rate all explanations provided by participants in Study 2, first before reading expert explanation, then after reading expert explanation.	Independent rating closer to self-ratings at T2 and later than to self-ratings at T1
6	31 Yale undergraduates	Same as Studies 1–3, but instructions said they would have to provide explanations and answer Qs	Still a significant drop in self-ratings, but not as big as in Studies 1–4
7	52 Yale undergraduates	Facts (capitals of countries) Same procedure, minus diagnostic Qs at T3	Significantly smaller drop than w/devices—i.e., better calibrated
8	37 Yale undergraduates	Procedures Same procedure, minus diagnostic Qs at T3	No drop
9	39 participants (24 Yale undergraduates; 15 regional)	Narratives (movie plots) Same procedure, minus diagnostic Qs at T3	No drop
10	31 Yale undergraduates	Natural Phenomena Rate 24, explain five times (10 between-subjects)	Same as Devices
11	24 Yale undergraduates	Desirability of knowledge across domains; used all test items from preceding studies	Desirability means pattern differently than overconfidence means
12a	54 Yale undergraduates	Rate 40 devices from Studies 1–3 on visible/hidden, number of parts familiarity, etc.; used to predict average initial confidence in Study 2	Visibility of internal parts, knowing names of parts, and the ratio of known to total parts predict initial confidence
12b	Used ratings from Study 12a	Used to predict overconfidence for 40 devices used in Studies 1–4	Visibility of internal parts predicts overconfidence.

Another conclusion of this research is that the IOED is quite robust with some kinds of causally complex systems. Warning participants that they would be tested (in Study 6) was insufficient to eliminate the illusion, and participants expressed substantial surprise at the shallowness of their explanations across all studies with devices and natural phenomena.

As described in the introduction, in recent years, there has been considerable emphasis on the importance of intuitive theories in models of concepts, conceptual change, reasoning, and learning (e.g., Barrett et al., 1993; Murphy & Allopenna, 1994; Sloman et al., 1997; Solomon et al., 1999). How does the IOED bear on such claims? One possible conclusion from these studies is that the intuitive theories are ephemeral and are largely irrelevant to concepts and that our conviction that they do matter is driven by the illusion. That conclusion, however, leaves unanswered the large set of effects on categorization, induction, conceptual combination, and conceptual change that do seem to override frequency information in ways that suggest influences of beliefs about causal relations and patterns. When people override typicality information, for example, the patterns they follow are in accord with their having intuitive theories of how and why the elements of those parents cohere as they do (Murphy & Medin, 1985). Either something akin to theories is still at work or other factors cause effects that convincingly mimic the effects of theories. There is a recent surge of proposals of potential alternatives ranging from Bayesian models (Tenenbaum & Griffiths, 2001), to more powerful roles for similarity (Hampton, 2001) to an enhanced role for perceptual bases of concepts (Goldstone & Barsalou, 1998; Prinz, in press). The IOED could be taken as support for these alternatives that down play theory and favor other factors.

A different interpretation, however, is that people do encode causal patterns in ways that capture theoretical relations, but do so in a highly sparse manner that, while skeletal, is still effective. One factor could involve degrees of coherence (Thagard, Eliasmith, Rusnock, & Shelley, in press). Sets of beliefs may cohere to the extent that they mutually support each other or conjointly provide insight to another relation or belief. Since a set of beliefs can have explanatory coherence without the full details of a mechanistic theory, they might provide an intermediate level at which theory-like relations constrain concept acquisition and use. Preferences for some features or correlations over others might be guided by constraints arising from a coherentist bias while still allowing for major gaps in the details of knowledge. In short, if people have a drive for coherence and sense coherence when it emerges in a set of beliefs, they may confuse that sense of coherence with a more detailed understanding. Indeed, this may be a more accurate way of characterizing the levels-of-understanding confusion.

In this view people have skeletal models of certain causal patternings that are much sparser than a fully detailed mechanistic account but which still work to create theory-like effects on concepts. For example, one might believe that color and surface markings are less likely to be causally central to understanding the nature of furniture and hand tools than animals and plants while overall shape might not be equally if not more important for furniture and hand tools (Keil, 1994; Medin & Shoben, 1988). That notion of differential causal potency would create a bias for certain feature clusters over others. One might also grasp the first member of a causal chain and give it a special status without knowing full details of the chain (Ahn, Kim, & Lassaline, 2000). These sorts of schematic understandings may

play powerful roles in guiding concept acquisition and use and, in doing so, may impart a strong sense of theoretical efficacy that is then mistaken for a fuller blueprint-like mechanistic knowledge.

Since it is impossible in most cases to fully grasp the causal chains that are responsible for, and exhaustively explain, the world around us, we have to learn to use much sparser representations of causal relations that are good enough to give us the necessary insights: insights that go beyond associative similarity but which at the same time are not overwhelming in terms of cognitive load. It may therefore be quite adaptive to have the illusion that we know more than we do so that we settle for what is enough. The illusion might be an essential governor on our drive to search for explanatory underpinnings; it terminates potentially inexhaustible searches for ever-deeper understanding by satiating the drive for more knowledge once some skeletal level of causal comprehension is reached.

Note

1. The Graduate Arrogance theory was especially strongly advocated by the undergraduate research assistants in our lab.

Acknowledgments

The authors would like to thank Paul Bloom and Robert Sternberg for their helpful comments on early drafts of this paper. We would also like to thank our diligent and talented research assistants, especially Nicholaus Noles and Emily McKee, for their help with the preparation of this manuscript. This research was funded by NIH Grant R01-HD23922 to Frank Keil.

Appendix A. Devices studies

A.1. Devices studies stimuli

Stimuli for Studies 1–4: 48 phenomena initially rated by participants

How a sewing machine works	How a flush toilet operates
How an LCD screen works	How a hydroelectric turbine changes water pressure into electricity
How a can opener works	How a car battery stores electricity
How a 35 mm camera (single-lens reflex camera) makes images on film	How a jet engine produces thrust
How a zipper works	How a self-winding watch runs without batteries

Appendix A. (*Continued*)

How a cellular phone works	How a microchip processes information
How a greenhouse works	How the U.S. Supreme Court determines the constitutionality of laws
How a fluorescent light works	How a photocopier makes copies
How a nuclear reactor produces electricity	How a car ignition system starts the engine
How a speedometer works	How the liver removes toxins from blood
How the heart pumps blood	How a car differential helps the car turn
How a water faucet controls water flow	How the presidential elections determine the next president
How a quartz watch keeps time	How steam central heating warms large buildings
How a VCR works	How a snare catches small animals
How a car's gearbox works	How an incinerator works
How a cylinder lock opens with a key	How a television creates pictures
How a helicopter flies	How a ball-point pen writes
How a radio receiver works	How an electric motor changes electricity into movement
How a telephone transmits sound through wires	How piano keys make sounds
How a fireplace works	How a spray-bottle sprays liquids
How a solid-fuel rocket produces thrust	How a manual clutch works
How the aqualung (Scuba-gear) regulates air-pressure	How an Ethernet network allows computers to share files
How a computer mouse controls the pointer on a computer screen	How a transistor works
How a scanner captures images	How the brain coordinates behavior

A.2. Phase 3 instructions (write explanations) used in the devices studies

Now, we'd like to probe your knowledge in a little more detail, on some of the items.

For each of the following, please describe all the details you know about the phenomena, 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, for each phenomenon, try to tell as complete a story as you can, with no gaps.

If you find that your story does have gaps (i.e., you are not sure how the steps are connected) please write the word "GAP" in your description at that point, and then continue. Feel free to use labeled diagrams, or flow-charts to get your meaning across.

When you are done, please re-rate your knowledge of the phenomenon on a 1–7 scale.

Appendix B. Procedures study

B.1. Procedures stimuli

Study 8 stimuli: 10 procedures

A correct procedure for how to drive from New Haven to New York City	The correct procedure for how to set a table
The correct procedure for how to tie a bow tie	A correct procedure for how to make pasta
The correct procedure for how to file your taxes	The correct procedure for how to tie a bow-tie
A correct procedure for how to drive from New Haven to Chicago	The correct procedure for how to make an international telephone call
A correct procedure for how to make scrambled eggs	A correct procedure for how to make chocolate chip cookies from scratch

B.2. Phase 3 instructions (write explanations) used in the procedures study

Now, we'd like to probe your knowledge in a little more detail on some of the items.

For each of the following, please describe all the steps in the procedure that you know, going from the first step to the last. For each procedure, try to tell as complete a story as you can, with no gaps.

If you find that your story does have gaps (i.e., you are not sure about some of the steps or how they are connected) please write the word "GAP" in your description at that point, and then continue. Feel free to use labeled diagrams, or flow-charts to get your meaning across.

When you are done, please re-rate your knowledge of the procedures on a 1–7 scale in the space provided.

References

Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Cognition and explanation*. Boston, MA: MIT Press.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.

Barrett, S. E., Abdi, H., Murphy, G. L., & Gallagher, J. M. (1993). Theory-based correlations and their role in children's concepts. *Child Development*, *64*, 1595–1616.

Bjoerkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior & Human Decision Processes*, *58*(3), 386–405.

Bjork, R. A. (1998). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Boyd, R. (1991). On the current status of scientific realism. In R. Boyd, P. Gaspar, & J. D. Trout (Eds.), *The philosophy of science* (pp. 195–222). Cambridge, MA: MIT Press.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*(3), 271–280.

diSessa, A. D. R. D. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale: Erlbaum.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *The nature of insight*. Cambridge, MA: MIT Press.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge University Press.

Gelman, S. A., & Koenig, M. A. (2002) Theory-based categorization in early childhood. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming buzzing confusion*. New York: Oxford University Press.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596.

Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *11*(1–4), 702–718.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, *10*(6), 597–602.

Goldstone, R. L., & Barsalou, L. (1998). Reuniting perception and conception. *Cognition*, *65*, 231–262.

Goldstone, R. L., & Johansen, M. K. (2002) Conceptual development: From origins to asymptotes. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming buzzing confusion*. New York: Oxford University Press.

Gopnik, A. A., & Wellman, H. M. (1994). The theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). Cambridge, UK: Cambridge University Press.

Hampton, J. A. (2001). The role of similarity in natural categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization*. Oxford: Oxford University Press.

Hull, D. L. (1965). The effect of essentialism on taxonomy: 2000 years of stasis. *British Journal of Philosophy of Science*, *15*, 314–326.

Juslin, P. (1993a). *An ecological model of realism of confidence in one's general knowledge*. Uppsala, Sweden: Uppsala Universitet, Acta Universitatis Upsaliensis.

Juslin, P. (1993b). An explanation of the hard–easy effect in studies of realism of confidence of one's general knowledge. *European Journal of Cognitive Psychology*, *5*(1), 55–71.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior & Human Decision Processes*, *57*(2), 226–246.

Keil, F. C. (1994). Explanation-based constraints on the acquisition of word meaning. *Lingua*, *92*, 169–196.

Keil, F. C. (1998). Cognitive science and the origins of thought and knowledge. In R. M. Lerner (Ed.), *Theoretical models of human development* (5th ed., Vol. 1). New York: Wiley.

Kindersley, D. (1996). *The Way Things Work 2.0*. [CD-ROM]: DK Multimedia.

Koriat, A. (1995). Dissociating knowing and feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, *124*, 311–333.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 34–53.

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

Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The meta-cognitive error of overestimating change–detection ability. *Visual Cognition*, *7*, 397–412.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Decision Processes*, *20*(2), 159–183.

Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*(4), 345–391.

Mandler, J., & Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, *9*, 111–151.

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development*, *48*(3), 986–992.

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, *50*(3), 643–655.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469–1481.

Medin, D. L., & Ortony A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, UK: Cambridge University Press.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*, 158–190.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in meta-cognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 851–861.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, *10*, 151–177.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*(4), 904–919.

Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *53A*, 962–982.

Murphy, G. L., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(2), 289–316.

Murphy, G. L. (2000). Explanatory concepts, In R. A. Wilson & F. C. Keil (Eds.), *Explanation and cognition*. Cambridge MA: MIT Press.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Prinz, J. (in press). *Furnishing the mind*. Cambridge, MA: Bradford Books, MIT Press.

Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge* (pp. 131–193). Minneapolis: University of Minnesota Press.

Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Salmon, W. (1998). Causality and Explanation, New York: Oxford University Press.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Sloman, S. A., Love, B. C., & Ahn, W. (1997). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189–228.

Solomon, K. O., Medin, D. L., & Lynch, E. B. (1999). Concepts do more than categorize. *Trends in Cognitive Science*, *3*(3), 99–105.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605.

Steigler, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Thagard, P., Eliasmith, C., Rusnock, P., & Shelley, C. P. (in press). Knowledge and coherence. In R. Elio (Ed.), *Common sense, reasoning, and rationality* (Vol. 11). New York: Oxford University Press.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*(3), 387–392.

Wilson, R. A., & Keil, F. C. (1998). The shadows and shallows of explanation. *Minds & Machines*, *8*(1), 137–159.

Yates, J. F., Lee, J. W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior & Human Decision Processes*, *65*(2), 138–147.

Yates, J. F., Lee, J.-W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and reality. *Organizational Behavior & Human Decision Processes*, *70*(2), 87–94.