



ELSEVIER

Cognitive Science 92 (2002) 1–42

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. Implications for the roles of intuitive theories in models of concepts and cognition are discussed. © 2002 Published by Cognitive Science Society, Inc.

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).

ARTICLE IN PRESS

31 1969). Theories seem to tell us what features to emphasize in learning new concepts as well
32 as highlighting the relevant dimensions of similarity (Murphy, 2002). Intuitive theories have
33 also been heavily emphasized in accounts of the cognitive development of children (Gelman &
34 Koenig, 2002) and even of infants (Spelke, Breinlinger, Macomber, & Jacobson, 1992).

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

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

46 Folk theories, we claim, are even more fragmentary and skeletal, but laypeople, unlike some
47 scientists, usually remain unaware of the incompleteness of their theories (Ahn & Kalish,
48 2000; Dunbar, 1995; diSessa, 1983). Laypeople rarely have to offer full explanations for most
49 of the phenomena that they think they understand. Unlike many teachers, writers, and other
50 professional “explainers,” laypeople rarely have cause to doubt their naïve intuitions. They
51 believe that they can explain the world they live in fairly well. They are novices in two re-
52 spects. First, they are novice “scientists”—their knowledge of most phenomena is not very
53 deep. Second, they are novice epistemologists—their sense of the properties of knowledge
54 itself (including how it is stored) is poor and potentially misleading.

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

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

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

75 exhaustive, VCR-like storage of everything one sees (Levin, Momen, Drivdahl, & Simons,
76 2000).

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

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

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

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

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

117 In the studies that follow we explore illusions of understanding for explanations, contrast-
118 ing the exaggerated sense of knowing explanations with better calibrations for other sorts of

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

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

133 To clarify the distinctive nature of our proposal it is useful to briefly consider prior research
134 on overconfidence. Relevant research in the judgment and decision making tradition has used
135 the disparity between people’s average confidence levels for their answers to almanac questions
136 and the proportion of correct answers to argue that people are overconfident (Fischhoff, 1982;
137 Lichtenstein & Fischhoff, 1977; Yates, Lee, & Shinotsuka, 1996; Yates, Lee, & Bush, 1997).
138 This tradition, however, does not focus on how illusions of knowing might differ across kinds of
139 knowledge. Lumping diverse kinds of knowledge into a hypothetical “general knowledge” and
140 looking for an overall overconfidence effects may well obscure large differences in calibration
141 between knowledge types.

142 The cognitive psychology literature on text comprehension also suggests overconfidence
143 about one’s knowledge. People are often poor at detecting when they have failed to understand
144 a piece of text, both as adults (Glenberg & Epstein, 1985; Glenberg, Wilkinson, & Epstein,
145 1982; Lin & Zabrocky, 1998) and as children (Markman, 1977; Markman, 1979). In contrast,
146 the current studies are concerned with people’s ability to assess the knowledge they have before
147 coming into the lab, rather than things learned in the course of an experiment. The implications
148 of our research are different: they tell us less about how people learn when reading, and more
149 about individuals’ intuitive theories about how knowledge is stored and about the mismatch
150 between what they think they already know and what they really know.

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

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

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

167 Overconfidence also exists in areas that have little to do with knowledge. Participants have
168 been shown to be overconfident about their future performance on motor tasks (e.g., [West &](#)
169 [Stanovich, 1997](#)), their abilities compared to other people's abilities (e.g., [Kruger & Dunning,](#)
170 [1999](#)), and about their competence to perform a broad range of tasks ([Bjork, 1998](#)).

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

176 A powerful illusion of depth distinctive to theory-like knowledge would have major im-
177 plications for the alleged roles of theories in conceptual structure. One possibility is that the
178 illusion represents a cognitive bias like naïve essentialism ([Medin & Ortony, 1989](#)) and reflects
179 a tendency on the part of laypeople and cognitive scientists to assume that intuitive theories are
180 a powerful component of our knowledge systems even when there is little data to support that
181 conjecture. The essentialist bias may liberally create placeholders that reserve mental locations
182 for essences, in anticipation that the locations will be filled later ([Murphy, 2002; Gelman &](#)
183 [Koenig, 2002](#)). We may create such placeholders even for concepts that lack essences ([Medin,](#)
184 [1989](#)). Indeed, an excessive essentialism may have hindered scientific thought in cases where
185 fixed essences are inconsistent with a new theory, such as evolution by natural selection ([Hull,](#)
186 [1965](#)).

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

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

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

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

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

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

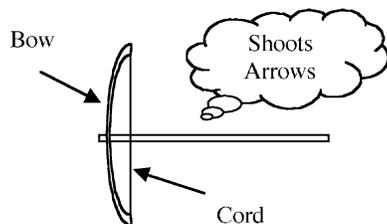
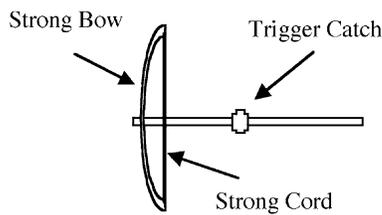
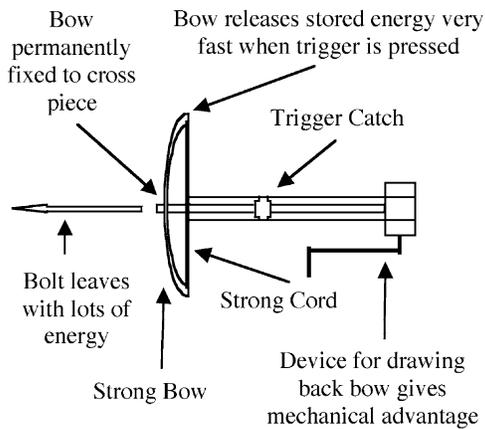
225 2. An illusion of explanatory depth with devices

226 2.1. Study 1: Documenting the illusion

227 2.1.1. Methods

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

238 Sixteen graduate students from various departments at Yale University participated in the
239 study. In Phase 1 of the study, the participants learned to use a 7-point scale that indicated
240 whether they had a deep, partial, or shallow understanding of a device or phenomenon. The
241 point “1” was the lowest level of understanding and “7” the highest. Participants considered
242 two training examples, one for which most adults had a good mechanistic understanding, a
243 cross bow, and one for which most do not, a global positioning satellite receiver. The explana-
244 tion of the scale included text and diagrams illustrating the different levels of understanding
245 for the two examples (see Fig. 1). All participants indicated that they clearly grasped the
246 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.

247 In Phase 2, participants studied a list of 48 items (shown in Appendix A) and rated each
 248 for their level of understanding on the 7-point scale described during the training. Participants
 249 were asked to rate the list without pausing excessively on any item, and took approximately
 250 10 min to rate the entire list (rating "T1").

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

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

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

263 In Phase 5, the participants read a brief expert description of the phenomenon and then
 264 re-rated their prior level of understanding relative to that description (rating “T4”). The expert
 265 explanations were taken from a CD-ROM by David McCauley titled “*The Way Things Work*
 266 *2.0.*” (Kindersley, 1996). The explanations contained diagrams and text, and ranged in length
 267 (including illustrations) from half a page to several pages.

268 Finally, as a manipulation check, participants were asked how well they understood the phe-
 269 nomenon after having read the expert explanation (rating “T5”). The sequence of the procedures
 270 is diagrammed in Fig. 2.

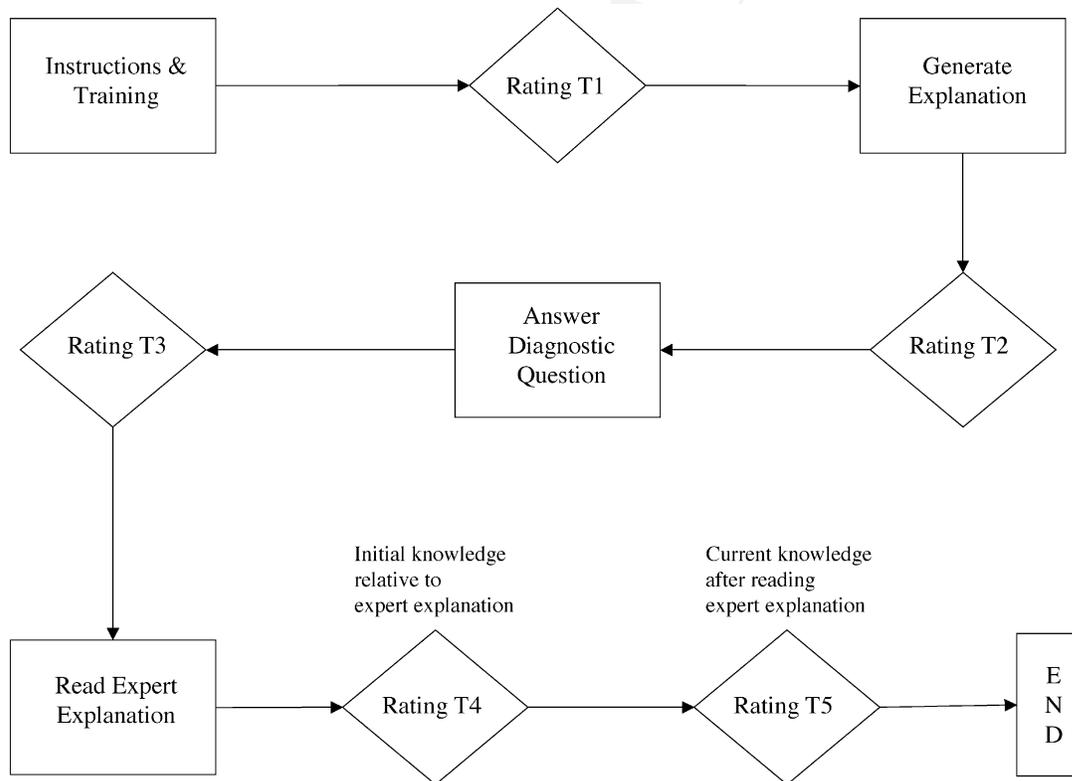


Fig. 2. Diagram of the procedure used in Studies 1–4.

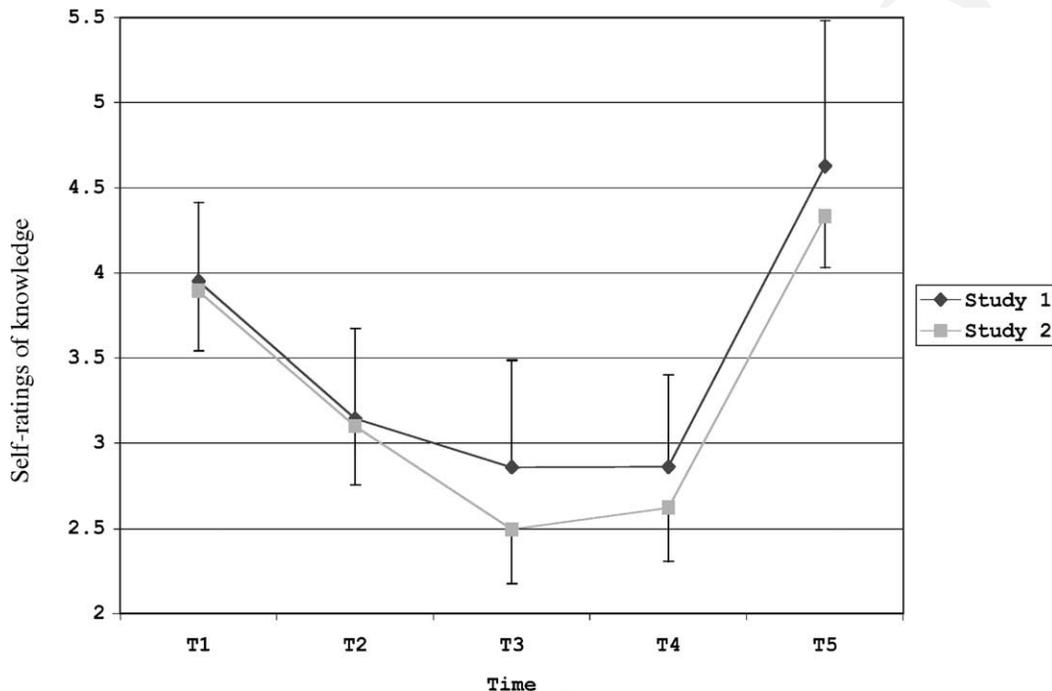


Fig. 3. Self-ratings of knowledge averaged across all items and subjects by time for Studies 1 and 2. The *x*-axis shows the sequence of self-ratings from Time 1 to Time 5. T1 is the initial self-rating, prior to any attempt to produce an explanation. T2 is the rating immediately after the first effort 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.

271 2.1.2. Results

272 Nearly all participants showed drops in estimates of what they knew when confronted
 273 with having to provide a real explanation, answer a diagnostic question, and compare their
 274 understanding to an expert description, as seen in Fig. 3. The darker line in Fig. 3 shows the
 275 means for all items for T1–T5. A repeated measures ANOVA showed time was significant,
 276 $F(4, 56) = 16.195, p < .001, \eta^2 = .536$. The planned linear contrasts for Time 1 versus
 277 Time 2, Time 1 versus Time 3, and Time 1 versus Time 4 were all significant at $p < .002$. The
 278 contrasts for Time 2 versus Time 3, and Time 2 versus Time 4 were not significant.

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

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

287 2.1.3. Discussion

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

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

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

311 2.2. Study 2: Replicating the illusion in a larger, younger group

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

316 2.2.1. Methods and results

317 Thirty-three Yale undergraduates received the same stimuli and the same series of five
318 questions as in Study 1. As shown by the lighter line in Fig. 3, the undergraduate partici-
319 pants exhibited the same overall pattern of results as the graduate students in Study 1. The
320 repeated measures ANOVA showed time was significant, $F(4, 124) = 38.9, p < .001, \eta^2 =$
321 $.555$. Planned linear contrasts for Time 1 versus Times 2, 3 and 4 were all significant at
322 $p < .001$.

323 A direct comparison of Study 1 with Study 2 using a repeated measures ANOVA with time
324 as a within-subject factor and study as a between-subject factor showed no differences between
325 the two results. That is, the TIME \times STUDY interaction was not significant, $F(4, 188) = .462,$
326 $p = .902, \eta^2 = .006$. Of course, time remained highly significant in the combined analysis,
327 $F(4, 188) = .44.11, p < .001, \eta^2 = .448$.

328 2.2.2. *Discussion*

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

333 2.3. *Study 3: Replicating the illusion in a less selective university*

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

338 2.3.1. *Methods and results*

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

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

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

353 While the main effect of institutional affiliations was not significant ($F(1, 63) = .750, p =$
354 $.390, \eta^2 = .012$), the interaction of affiliation by time was highly significant, $F(4, 252) =$
355 $3.874, p = .005, \eta^2 = .058$. As Fig. 4 suggests, the interaction was driven primarily by the
356 later steps in procedure, i.e., the changes between Times 3 and 5. The direction of the difference
357 between groups was, furthermore, opposite of what the “elitist arrogance” hypothesis would
358 predict. The effect was larger in students from a regional university, primarily because their
359 initial ratings of their knowledge were nearly a point higher than that of students from a selective
360 university (see Fig. 4).

361 2.3.2. *Discussion*

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

364 2.4. *Study 4: Replicating the illusion with a different set of devices*

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

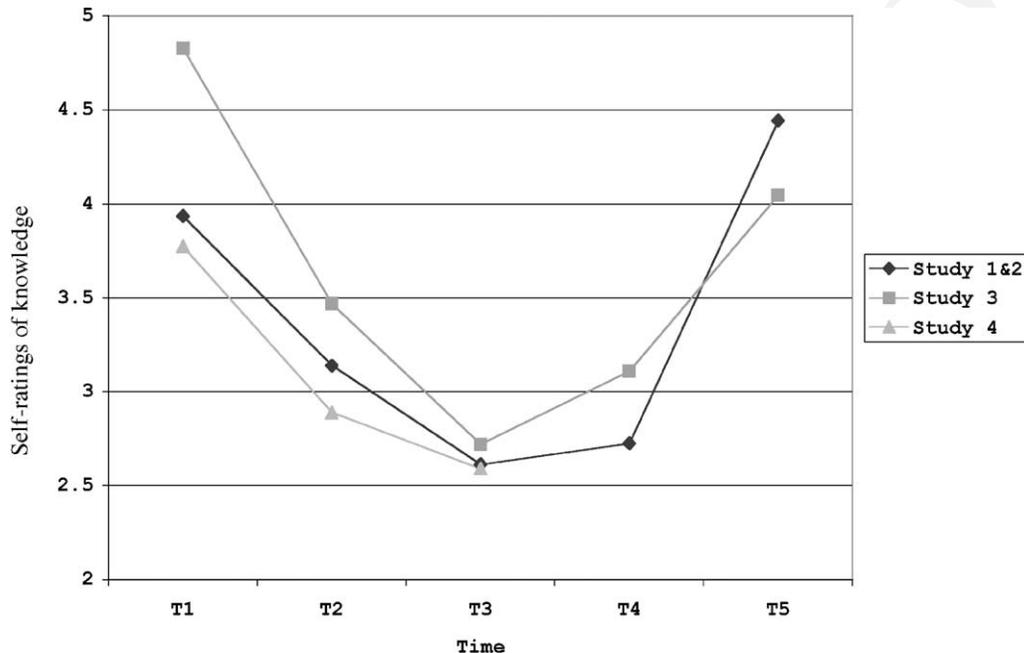


Fig. 4. Self-ratings of knowledge averaged across all items and subjects by time comparing Studies 1 and 2 (Yale students with original sub-set of items) with Study 3 (regional university students with original items) and Study 4 (Yale students with new items). The *x*-axis shows the sequence of self-ratings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

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

369 2.4.1. Methods

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

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

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

382 2.4.2. Results

383 The pattern of results with the new stimuli was identical to the patterns we saw in Studies
 384 1–3, and is shown in Fig. 4. A repeated measures ANOVA that considered time and item

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

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

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

394 2.4.3. Discussion

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

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

405 2.5. Study 5: Measuring the illusion by rating participants' explanations

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

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

419 2.5.1. Methods

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

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

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

430 2.5.2. Stimuli

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

437 2.5.3. Results

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

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

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

451 The second variable of interest was the average rating of the T2 explanations across all 12
 452 subjects after the raters read the expert explanation: Independent Rating 2 at T2 (IR.2-T2).
 453 IR.2-T2 reliability across all 12 raters was high, $\alpha = .931$. These independent ratings were
 454 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).

455 A repeated measures ANOVA was again used to control for experiment-wise error when
456 comparing the means of T1–T4 with IR.2-T2. The overall ANOVA was highly significant,
457 $F(4, 524) = 39.938, p < .001, \eta^2 = .234$. A set of simple contrasts comparing IR.2-T2
458 with the participant self-ratings at T1–T4 shows that the independent ratings of accuracy are
459 significantly lower than those provided by participants at T1 and T2 ($p < .001$ for both), but
460 are not significantly different from those provided by the participants at T3 and T4. These
461 results are also shown in [Table 1](#).

462 All correlations between self- and independent ratings were significant at the $\alpha = .01$ level.
463 The pattern of correlations was informative, and is shown in [Table 1](#). IR.1-T2 was most highly
464 correlated with T2 ($r = .648$), and with T3 ($r = .557$), a bit less with T4 ($r = .523$) and
465 least with T1 ($r = .428$). IR.2-T2 was most highly correlated with T2 ($r = .644$), and with T4
466 ($r = .563$), a bit less with T3 ($r = .555$) and least with T1 ($r = .449$).

467 We tested the differences between the non-independent correlations following the proce-
468 dure endorsed by [Steigler \(1980\)](#). For both, IR.1-T2 and IR.2-T2, correlations with the T2
469 self-ratings were significantly larger than correlations with T1 self-rating ($p < .001$).

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

475 2.5.4. Discussion

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

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

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

498 2.6. Study 6: Reducing the illusion with explicit instructions

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

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

516 2.6.1. Methods

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

522 2.6.2. Results

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

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

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

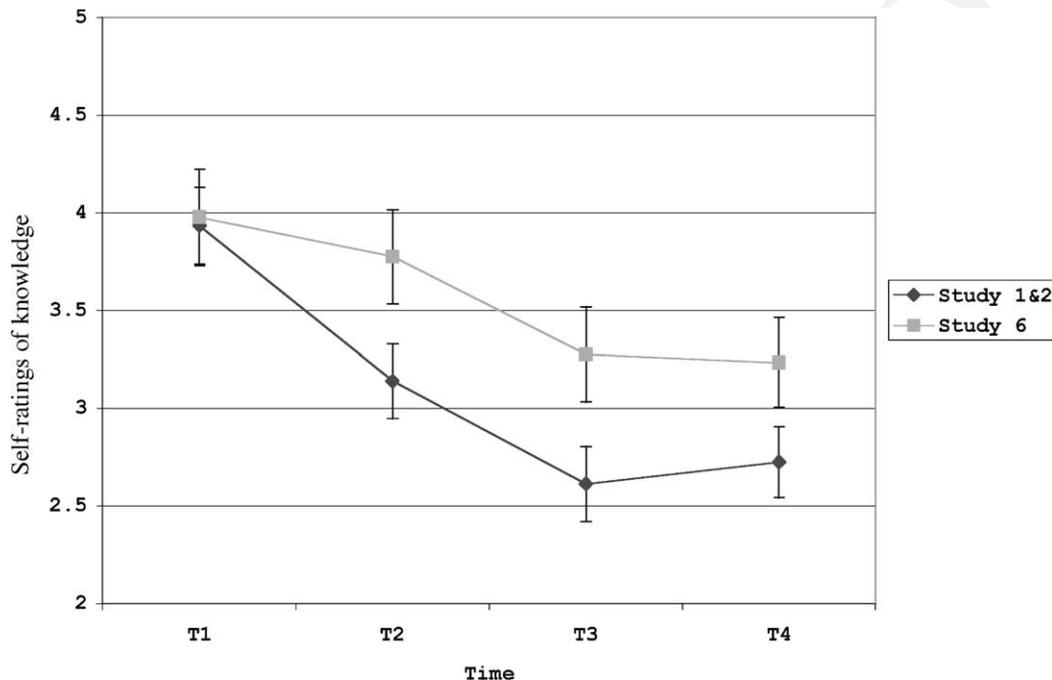


Fig. 5. Studies 1 and 2 (Yale students with devices) compared to Study 6 (Explicit warning). The x-axis shows the sequence of self-ratings from Time 1 to Time 4. T1–T4 as explained in Fig. 3.

539 $F(3, 234) = 4.119, p = .007, \eta^2 = .050$, and time remained highly significant, $F(3, 234) =$
 540 $44.924, p < .001, \eta^2 = .365$. As Fig. 5 shows, the drop in the explicit study was significant,
 541 but not as large as in the earlier studies.

542 2.6.3. Discussion

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

546 One unusual feature of the data deserves mention. We might have expected the manipulation
 547 to result in the lower initial (T1) ratings of knowledge, but it did not. Instead, the magnitude of
 548 the effect was reduced because T2 and T3 ratings were higher in this study compared to Studies
 549 1–4. This pattern is somewhat unexpected. One possibility is that the new instruction changed
 550 the way participants used the rating scales. For example, hearing the explicit instructions may
 551 have caused participants to try to be more consistent with their subsequent ratings because they
 552 had less justification for being surprised at the drop in performance.

553 2.6.4. General discussion of Studies 1–6

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

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

561 3. Calibration of comprehension across knowledge domains

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

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

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

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

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

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

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

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

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

612 The four subsequent studies (7–10) explore these factors by comparing people's performance
613 in the explanations studies with performance in other knowledge domains. The final two studies
614 (11 and 12) look at factors that may produce an illusion of depth with explanations more directly,
615 by examining the variation between different items used in the earlier studies.

616 3.1. Study 7: Ruling out general overconfidence—factual knowledge

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

624 3.1.1. Methods

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

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

636 The phases of Study 3 are analogous to Phases 1, 2, 4, and 5 of the devices studies. Phase
637 3 of Study 1 had no analog because we could not ask diagnostic mechanism questions about
638 facts.

639 3.1.2. Results

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

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

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

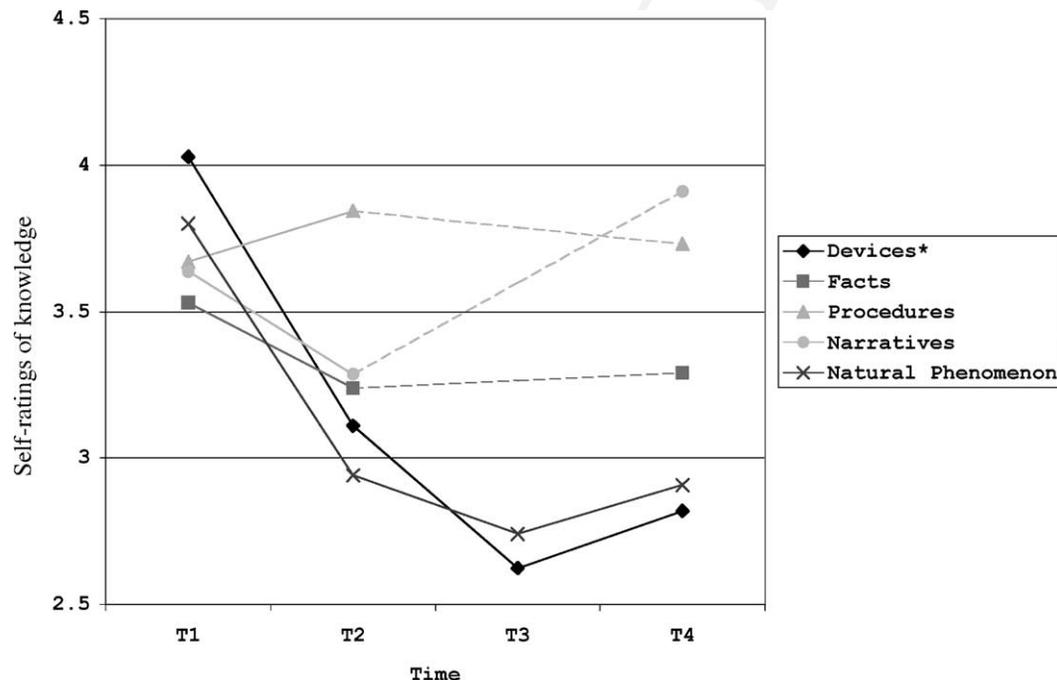


Fig. 6. Self-ratings of knowledge averaged across all items and subjects by time for Devices (Studies 1–4), Facts (Study 7), Procedures (Study 8), Narratives (Study 9), and Natural Phenomena (Study 10). The x -axis shows the sequence of self-ratings from Time 1 to Time 4. T1 is the initial self-rating, 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.)

654 showed that the drop from Time 1 to Time 2 and to Time 3 were significant, $p = .014$ and
655 $p = .001$, but the drop from Time 2 to Time 3 was not.

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

662 The analysis showed significant differences between the two results. That is, the TIME \times
663 DOMAIN interaction was significant, $F(1, 147) = 15.471$, $p < .001$, $\eta^2 = .095$. Time
664 remained highly significant for the combined data set, $F(1, 147) = 56.534$, $p < .001$, $\eta^2 =$
665 $.278$. As Fig. 6 shows, the drop from T1 to T2 in the Facts study was significant, but the
666 magnitude of the drop was also significantly smaller than the ratings drop in the devices
667 studies.

668 3.1.3. Discussion

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

674 Note also that the ratings for capitals were not bimodal (with some subjects simply picking
675 “1” and others picking “7” on the scale). Instead they were relatively normally distributed as
676 participants were often unsure, for example, which of several prominent cities in a country was
677 its capital.

678 3.2. Study 8: Ruling out general overconfidence—knowledge of procedures

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

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

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

700 3.2.1. *Methods*

701 We asked 37 participants in Study 8 about their knowledge of various real-world procedures,
702 such as how to bake chocolate chip cookies from scratch, how to tie a bow tie, or how to drive
703 from New Haven to New York City. As in Studies 1–4, participants were first trained on the
704 7-point scale using an example of a multi-step procedure. Then the participants rated how well
705 they understood each of the 10 procedures (shown in [Appendix B](#)).

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

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

721 3.2.2. *Results*

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

725 To compare results for procedures with those for devices, we used an overall repeated
726 measures ANOVA to compare data from Study 8 to data collapsed across Studies 1–4. Domain
727 was a between-subject factor with two levels (Procedures vs. Devices) and time was as a
728 within-subject factor with two levels (T1 and T2). The TIME \times DOMAIN interaction was
729 highly significant, $F(1, 132) = 40.934$, $p < .000$, $\eta^2 = .237$. The initial levels of confidence
730 were the same as for Studies 1–4, making ceiling and floor effects unlikely explanation for the
731 stability in knowledge ratings.

732 3.2.3. *Discussion*

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

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

739 3.3. Study 9: Calibration with knowledge of narratives (plots of movies)

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

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

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

759 3.3.1. Methods

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

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

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

777 3.3.2. *Results*

778 Unlike with devices, but like with procedures, there was no drop in self-rating as a result of
779 an effort to explain. A repeated measures ANOVA with time and item as within-subject factors
780 showed time to be highly significant but only if Time 5, the manipulation check, was included,
781 $F(3, 48) = 16.989$, $p < .001$, $\eta^2 = .515$. However, the planned linear contrasts showed no
782 significant differences between ratings at Times 1 and 2, or between ratings at Times 1 and 3.
783 Rating 3 was significantly different from Rating 2. But unlike the Devices studies, Rating 3
784 was significantly higher than Rating 2.

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

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

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

799 3.3.3. *Discussion*

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

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

814 3.4. *Study 10: Calibration with knowledge about natural phenomena*

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

818 cipants were overconfident but markedly less so than with explanatory knowledge of
819 devices.

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

830 3.4.1. *Methods*

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

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

845 3.4.2. *Results*

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

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

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

860 $p < .07$ (with the Greenhouse–Geisser correction for sphericity), $\eta^2 = .027$, suggested a
861 somewhat larger drop for devices.

862 3.4.3. Discussion

863 The pattern of results was similar to those found in Studies 1–4, as seen in Fig. 6. The
864 overconfidence values (T1–T2) across domains are also summarized in Table 7. The drop in
865 knowledge estimates over time was significant and much greater than that seen in Studies
866 7–9. There was also a suggestive pattern of a somewhat smaller drop for natural phenomena
867 at a marginal significance level of $p < .07$. The findings may, therefore, be taken to suggest
868 a function mechanism confusion as a small factor in overconfidence for devices: the lack of
869 functional descriptions in most accounts of non-biological natural phenomena is a small factor
870 that modestly decreases the magnitude of the illusion in comparison to devices.

871 To summarize, studies with devices and natural phenomena both show large drops in knowl-
872 edge estimates. Procedures and Narratives show no drop, while Geography Facts show only a
873 small drop. The results demonstrate large differences in knowledge calibration across knowl-
874 edge domains, casting serious doubt on the meaningfulness of “general overconfidence” about
875 knowledge. The studies also raise intriguing possibilities about the mechanism behind over-
876 confidence, which we address in the next few studies.

877 4. Exploring the causes behind the illusion

878 4.1. Study 11: Ruling out desirability as the explanation for inter-domain differences

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

888 4.1.1. Methods

889 Twenty-four Yale undergraduates participated in the study. The participants rated on a
890 7-point scale how embarrassed they thought they would be if they had to admit not having
891 a good knowledge or understanding of an item. The question was framed as rating “embar-
892 rassment over ignorance” because that phrasing seemed to most directly tap the participant’s
893 motivational experience in the IOED studies. The stimuli consisted of a combined list of all
894 test items used in the previous studies: devices, facts, procedures, narratives, and natural phe-
895 nomena. The participants were given instructions on a 7-point “embarrassment” scale, and
896 were asked: “For each item, please rate how embarrassed you think you would be if someone

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

900 4.1.2. Results

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

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

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

913 4.1.3. Discussion

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

918 The pattern of mean differences makes it unlikely that social desirability is the major factor
 919 behind the differences in overconfidence observed between knowledge types. Devices produced
 920 the most overconfidence but the lowest desirability. Movies produced the least overconfidence
 921 but the highest desirability. Natural Phenomena produced high overconfidence and intermediate
 922 desirability, while Geography Facts produced low overconfidence and intermediate desirability.
 923 The divergent patterns make it more plausible that factors other than desirability are driving
 924 the differences between knowledge types. If anything, high desirability may cause people to
 925 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.

926 4.2. *Study 12: Correlates of confidence and overconfidence*

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

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

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

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

949 4.2.1. *Methods 12a: Correlates of confidence*

950 To examine what factors might be related to a sense of deep knowledge, we asked a group of
951 54 participants (Yale undergraduates recruited in the same manner as in previous studies) to rate
952 all device items used in Studies 1–4. Forty items of the original 48 were rated on all the scales.
953 Eight items were excluded from some of the scales because they were not devices, but other
954 kinds of causally complex systems (e.g., the human liver, or the presidential election process).
955 Consequently, the rating scales developed for devices (e.g., number of parts, mechanical vs.
956 electrical) would have been confusing to the participants with those items. (Of course, these
957 eight non-device items were never used as test items in the devices studies.)

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

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

968 4.2.1.1. *Familiarity*. Please rate how familiar you are with the following devices/phenomena.

969 4.2.1.1.1. *Hidden versus visible parts*. Many devices, such as computers, hide almost all
970 of their important working parts inside an outer container. On the other hand, when you use a
971 device like a crossbow, you can clearly see almost all of the parts and how they work together.
972 For each of the following items, please estimate the proportion of parts that are visible during
973 the normal course of operation.

974 4.2.1.1.2. *Electrical versus mechanical devices*. Some devices can be easily classified as
975 either electrical or mechanical, whereas other devices may be some combination of both. For
976 each of the following devices, please rate the degree to which device operates by electrical or
977 mechanical principles.

978 4.2.1.1.3. *Names of parts*. How many different parts of the following items would you
979 guess you could identify by name? Please estimate whether you could name:

980 0, 1–2, 3–5, 6–10, 11–25, 26–50, 51–100, 101–250, or more than 250 parts.

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

987 Every device is made up of a number of different “parts.” For example, on one level, a computer
988 is made up of a monitor, mouse, CPU unit, and keyboard. On an even deeper level, the computer
989 is made up of a motherboard, power supply, RAM, hard drive, video card, etc. And every part that
990 was just named is obviously made up of even more parts. While it may be impossible to arrive at an
991 absolute number of parts making up a given device, it is safe to say that certain devices are made up
992 of more parts than others (i.e., a computer has more parts than a toaster.)

993 We are looking for your subjective estimate of how many parts make up the following items. When
994 you are trying to imagine how many parts something has, it may be helpful to think about how many
995 parts you can disassemble the device into without breaking anything. You can also imagine a blow-up
996 diagram of the item and estimate the number of parts you would see on that diagram.

997 These pictures of a combination lock and its constituent parts may help illustrate what we mean by
998 a “part.”

999 We showed the participants a picture of a combination lock, and the picture of the same
1000 lock disassembled into its components, and then asked them to estimate the number of parts
1001 the various items on our list had on a log-type scale. They were asked to “Please indicate how
1002 many parts you think the following items have:

1004 1–5, 6–10, 11–25, 26–50, 51–100, 101–250, 251–500, 501–1000,
1005 or more than 1000 parts.

1006 We used this non-linear scale to reduce variance at the higher end. During piloting with
1007 open-ended responses we found that people's estimates of the number of parts for some complex
1008 items varied by several orders of magnitude.

1009 How do these variables, (1) familiarity with the item, (2) the ratio of visible versus hidden
1010 parts, (3) the number of mechanical versus electrical parts, (4) the total number of parts, (5) the
1011 number of parts for which one knows names, and (6) the ratio of known to total parts, relate to our
1012 hypotheses? We predicted that three factors—underestimating environmental representational
1013 support, levels of analysis confusions, and mistaking knowing names of parts for knowing
1014 deeper causal roles—would drive the differences in initial confidence and in overconfidence.
1015 If these factors were sufficient to explain overconfidence, overall complexity, here measured
1016 by the estimate of the number of parts, should not influence the illusion, once other factors are
1017 controlled for.

1018 We expected three other scales to affect initial confidence for explanatory knowledge. Two
1019 scales were designed to capture different aspects of the amount of representational support an
1020 object provides. The number of visible versus hidden parts, and whether the device operates
1021 on electrical or mechanical principles, predict how easy it seems to discover the mechanism
1022 of operation in real time by simply examining the device closely. A device with mostly visible
1023 parts (e.g., a can opener or a bicycle derailleur) allows an observer to reconstruct the device's
1024 mechanism in real time by examining the device and may therefore create the strongest illusion
1025 of knowledge; therefore ratings of the ratio of visible to hidden parts should influence initial
1026 confidence. Similarly, if a device operates largely on mechanical principles it would be easier to
1027 figure out on-the-fly than something packed with circuit boards. We weren't sure how strongly
1028 the two variables would correlate, but expected both to influence initial confidence if they were
1029 not highly collinear.

1030 Two other measures, the number of known part names, and the (computed) ratio of known
1031 to total parts—were designed to test the hypotheses that subjects confuse knowing labels with
1032 understanding the mechanism. Knowing the names of many parts, or a large proportion of parts,
1033 in a device may mislead the observer into believing they understand the causal relationships
1034 between the parts. Knowing that my computer contains a hard drive, a CPU, an LCD screen,
1035 a SCSI Bus, and RAM, might be enough to give me a sense that I know how the computer
1036 works, even if I have very little understanding of how the parts interact.

1037 In informal piloting we asked participants to guess what factors would make a difference
1038 in overconfidence. Two factors emerged most often from naïve intuitions of participants: fa-
1039 miliarity and complexity (number of parts). Pilot subjects thought other participants would be
1040 more overconfident about familiar devices, and about complex ones. As mentioned above, we
1041 did not expect complexity to matter, once we controlled for other factors. However, we thought
1042 familiarity might be a possible factor, given its role in the FOK literature.

1043 As in phase one of Studies 1–4, only item labels were provided to the participants, with no
1044 additional descriptions. After being instructed on the use of a particular scale, a table containing
1045 all the item labels (e.g., telephone, fireplace, transistor) was presented to participants for each
1046 rating. Participants first rated the familiarity of each of the items, then estimated the ratio
1047 of visible versus hidden parts, then rated whether the devices operated on electrical versus
1048 mechanical principles, then estimated the number of parts, and finally estimated the number of
1049 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

| | <i>B</i> | <i>SE</i> | β | <i>t</i> | Significant |
|-----------------------------|----------|-----------|---------|----------|-------------|
| Constant | .376 | .278 | | 1.352 | .184 |
| Visible vs. hidden | .640 | .079 | .698 | 8.109 | 0 |
| Known names of parts | .163 | .037 | .321 | 4.394 | 0 |
| 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.

1050 complete list of items (along with instructions) can be found with the Supplemental Materials
1051 section at <http://www.elsevier.com/gej-ng/10/15/15/show/>.

1052 4.2.2. Results 12a: Correlates of confidence

1053 Ratings provided by all 54 Study 12a participants were averaged for each of the 40 device
1054 items. These average ratings were then used to predict the average initial confidence for that
1055 item in Study 2. We used a stepwise multiple regression analysis to determine which variables
1056 were most related to initial high levels of confidence. As seen in Table 3, the visible/hidden
1057 parts estimate, the number of known parts, and known:total parts ratio are the central factors,
1058 accounting for 78% of variance in initial confidence (model adjusted- $R^2 = .777$). Familiarity
1059 was not a significant predictor of initial confidence once the other three variables were included.
1060 Complexity, as measured by the number of parts estimate, was similarly non-significant.

1061 First-order correlations between variables are shown in Table 4. The ratio of visible to hidden
1062 parts was highly correlated with the mechanical:electrical dimension, with known:total parts
1063 ratio, with the raw number of parts, and with familiarity. The high colinearity between the
1064 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

| | <i>B</i> | <i>SE</i> | β | <i>F</i> -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–T2) from a set of six independent variables measured in Study 12.

1065 4.2.3. *Methods 12b: Correlates of overconfidence*

1066 The previous study looked at what properties of the phenomena may predict initial confi-
 1067 dence levels. Although initial confidence and overconfidence were strongly correlated in our
 1068 study, our main interest is in the relative levels of overconfidence, and it would be desirable to
 1069 measure them directly. Unfortunately, our initial studies made this difficult since we only had
 1070 direct measures of the miscalibration, as opposed to measures of the initial confidence level, for
 1071 only a small sub-set of devices used as test items (eight). To solve the problem, we combined
 1072 the knowledge self-rating data from Studies 2 and 4 (because those studies had the most similar
 1073 populations—Yale undergraduates) to obtain measured levels of overconfidence for a much
 1074 larger set of devices. The two studies together gave us direct measures of overconfidence for
 1075 40 different devices. Because the participants in Study 12a rated all 40 items on the five rating
 1076 scales, we were able to use their ratings in this analysis as well.

1077 4.2.4. *Results 12b: Correlates of overconfidence*

1078 Overconfidence for each item was defined as the difference between the average T1 and T3
 1079 scores for an item across all subjects in Studies 2 and 4. The means or the ratings provided by
 1080 the 54 Study 12 participants were then used to predict overconfidence for each item. We again
 1081 used a stepwise multiple regression analysis to determine which variables were most related
 1082 to the degree of overconfidence. The final model is shown in Table 5.

1083 The visible/hidden parts ratio explained most of the variance in overconfidence, and adding
 1084 other predictors did not significantly improve model fit. This may be partly due to high co-
 1085 linearity between some of the measures: the visible/hidden ratio was highly correlated with
 1086 electrical/mechanical dimension, the number of parts, and the ratio of known to unknown parts.
 1087 Thus, a simple regression predicting overconfidence from the visible/hidden ratio was highly
 1088 significant (model adjusted- $R^2 = .472$, $p < .0001$, $\beta = .687$). The first-order correlations
 1089 between all variables are shown in Table 6. As in Study 12a, the high colinearity between some
 1090 predictor variables complicates interpretation of the step-wise regression.

1091 4.2.5. *Discussion*

1092 This study helps explain why the illusion of knowledge may be especially powerful in cases
 1093 involving explanatory understanding, especially in domains with compelling intuitive theories.
 1094 It is in those cases that several unique factors converge to give a powerful, but inaccurate, feeling
 1095 of knowing. For example, the prominence of visible, transparent mechanisms may fool people
 1096 into believing that they have understood, and have successfully represented, what they have
 1097 merely seen.

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

| | Overconfident (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).

1098 Study 12a suggest that three of the factors we mentioned earlier—levels of analysis confu-
1099 sion, label-mechanism confusion, and confusing environmental support with representation—
1100 contribute to initial feelings of knowing. Study 12b indicates that mistaking environmental
1101 support for representation may be the single most important factor in determining which ex-
1102 planations will cause the greatest illusion of understanding. The findings indirectly suggest that
1103 this factor may also be primarily responsible for the large differences between explanations
1104 and other knowledge types.

1105 We have not explored in these studies the low rate of producing explanations relative to
1106 other types of knowledge. Conceivably, people rarely explain and thus do not get sufficient
1107 feedback to learn that their explanations are poor. This leads to an interesting prediction we
1108 would like to explore in a future study: people who produce explanations frequently (e.g.,
1109 teachers, expository writers) should be less subject to the IOED than those who produce them
1110 rarely.

1111 5. General discussion

1112 The studies in this article first demonstrated a strong illusion of explanatory depth with
1113 one set of causally complex systems: devices. We then showed large differences in knowledge
1114 calibration across knowledge domains, with high levels of overconfidence for devices and
1115 natural phenomena, modest overconfidence for geography facts, and no overconfidence for
1116 knowledge of narratives or procedures. We then directly explored some of the factors that
1117 might influence overconfidence within the domain of devices, finding evidence that the degree
1118 of causal transparency for a system (as measured by the ratio of visible to hidden parts) may
1119 be a critical factor behind the strong illusion of explanatory depth. Neither familiarity, nor
1120 complexity made a difference. (The lack of a familiarity effect might suggest that FOK cue
1121 familiarity models cannot explain the illusion.) Finally, the ratio of known to total parts made
1122 a difference in initial confidence, but not in overconfidence, leaving it unclear whether the
1123 confusion of labels with mechanism contributes to the illusion.

1124 [Table 7](#) summarizes the factors we think are important to calibration across knowledge
1125 domains and indicates their hypothesized relative contributions. Procedures and results for all
1126 12 studies presented in this paper are summarized in [Table 8](#) for the reader's convenience.

1127 We have found that the ratio of visible to hidden parts is the best predictor of overconfi-
1128 dence for an item, implicating the representational support account of the illusion. Can the
1129 representational support also explain differences across domains? Why, for example, don't
1130 we get confused about representational support with procedures or narratives? It may be that
1131 phenomena that are easy to visualize, and especially those that are easy to mentally animate,
1132 trigger an especially strong illusion of understanding and these may be more common in the
1133 domain of causal explanations.

1134 People might use the following useful heuristic: perceptions are more accurate than proposi-
1135 tional inferences. When you imagine a can-opener cutting through the lid of a can, that mentally
1136 animated image feels a lot more like perception than like propositional reasoning or informal
1137 inference. Thus, it would be easy assume that you can derive the same kind of representational
1138 support from the mental movie that you could from observing a real phenomenon. Of course,

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.

1139 the mental movie is much more like Hollywood than it is like real life—it fails to respect reality
 1140 constraints. When we try to lean on the seductively glossy surface we find the façades of our
 1141 mental films are hollow card-board. That discovery, the revelation of the shallowness of our
 1142 mental representations for perceptually salient processes, may be what causes the surprise in
 1143 our participants.

1144 A better understanding of the mechanism behind the illusion should also enable us to find
 1145 cases where the illusion is reversed, where people are substantially underconfident in their
 1146 ability to explain. Presumably, in cases that are especially hard to mentally animate and have
 1147 mostly invisible parts but that are otherwise easy to propositionally infer causal steps for, we
 1148 might see underconfidence. Anecdotally, we have all had the experience where a gifted teacher
 1149 shows us that we already know why something is the case when we inspect the consequences
 1150 of our own beliefs.

1151 One conclusion that can be drawn from this research is that the well-established blanket
 1152 approach to overconfidence with “general knowledge” is almost certainly misleading. Large
 1153 inter-domain differences in calibration imply that structural properties of knowledge have a
 1154 powerful impact on the process of knowledge assessment. “General knowledge” is a chimera—
 1155 a mythological composite creature. Taking it seriously distracts from interesting questions
 1156 about how knowledge assessment works, and the theoretically important issues of how the
 1157 structural properties of knowledge influence calibration.

1158 Another conclusion of this research is that the IOED is quite robust with some kinds
 1159 of causally complex systems. Warning participants that they would be tested (in Study 6)
 1160 was insufficient to eliminate the illusion, and participants expressed substantial surprise at
 1161 the shallowness of their explanations across all studies with devices and natural pheno-
 1162 mena.

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) | Significantly smaller drop than w/devices—i.e., better calibrated |
| 8 | 37 Yale undergraduates | Same procedure, minus diagnostic Qs at T3 | No drop |
| 9 | 39 participants (24 Yale undergraduates; 15 regional) | Procedures Same procedure, minus diagnostic Qs at T3 | No drop |
| 10 | 31 Yale undergraduates | Narratives (movie plots) Same procedure, minus diagnostic Qs at T3 | Same as Devices |
| 11 | 24 Yale undergraduates | Natural Phenomena Rate 24, explain five times (10 between-subjects) | Desirability means pattern differently than overconfidence means |
| 12a | 54 Yale undergraduates | Desirability of knowledge across domains; used all test items from preceding studies 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. |

1163 As described in the introduction, in recent years, there has been considerable emphasis
1164 on the importance of intuitive theories in models of concepts, conceptual change, reasoning,
1165 and learning. How does the IOED bear on such claims? One possible conclusion from these
1166 studies is that the intuitive theories are ephemeral and are largely irrelevant to concepts and
1167 that our conviction that they do matter is driven by the illusion. That conclusion, however,
1168 leaves unanswered the large set of effects on categorization, induction, conceptual combi-
1169 nation, and conceptual change that do seem to override frequency information in ways that
1170 suggest influences of beliefs about causal relations and patterns. When people override typical-
1171 ity information, for example, the patterns they follow are in accord with their having intuitive
1172 theories of how and why the elements of those parents cohere as they do (Murphy & Medin,
1173 1985). Either something akin to theories is still at work or other factors cause effects that
1174 convincingly mimic the effects of theories. There is a recent surge of proposals of potential
1175 alternatives ranging from Bayesian models (Tenenbaum & Griffiths, 2001), to more powerful
1176 roles for similarity (Hampton, 2001) to an enhanced role for perceptual bases (Goldstone &
1177 Barsalou, 1998; Prinz, in press). The IOED could be taken as support for these alternatives that
1178 down play theory and favor other factors.

1179 A different interpretation, however, is that people do encode causal patterns in ways that
1180 capture theoretical relations, but do so in a highly sparse manner that, while skeletal, is still
1181 effective. One factor could involve degrees of coherence (Thagard, Eliasmith, Rusnock, &
1182 Shelley, in press). Sets of beliefs may cohere to the extent that they mutually support each
1183 other or conjointly provide insight to another relation or belief. Since a set of beliefs can
1184 have explanatory coherence without the full details of a mechanistic theory, they might pro-
1185 vide an intermediate level at which theory-like relations constrain concept acquisition and
1186 use. Preferences for some features or correlations over others might be guided by constraints
1187 arising from a coherentist bias while still allowing for major gaps in the details of knowl-
1188 edge. In short, if people have a drive for coherence and sense coherence when it emerges in
1189 a set of beliefs, they may confuse that sense of coherence with a more detailed understand-
1190 ing. Indeed, this may be a more accurate way of characterizing the levels-of-understanding
1191 confusion.

1192 In this view people have skeletal models of certain causal patternings that are much sparser
1193 than a fully detailed mechanistic account but which still work to create theory-like effects
1194 on concepts. For example, one might believe that color and surface markings are less likely
1195 to be causally central to understanding the nature of furniture and hand tools than animals
1196 and plants while overall shape might not be equally if not more important for furniture
1197 and hand tools (Keil, 1994; Medin & Shoben, 1988). That notion of differential causal po-
1198 tency would create a bias for certain feature clusters over others. One might also grasp the
1199 first member of a causal chain and give it a special status without knowing full details of
1200 the chain (Ahn, Kim, & Lassaline, 2000). These sorts of schematic understandings may
1201 play powerful roles in guiding concept acquisition and use and, in doing so, may impart a
1202 strong sense of theoretical efficacy that is then mistaken for a fuller blueprint-like mechanistic
1203 knowledge.

1204 Since it is impossible in most cases to fully grasp the causal chains that are responsible
1205 for, and exhaustively explain, the world around us, we have to learn to use much sparser
1206 representations of causal relations that are good enough to give us the necessary insights:

1207 insights that go beyond associative similarity but which at the same time are not overwhelming
1208 in terms of cognitive load. It may therefore be quite adaptive to have the illusion that we know
1209 more than we do so that we settle for what is enough. The illusion might be an essential governor
1210 on our drive to search for explanatory underpinnings; it terminates potentially inexhaustible
1211 searches for ever-deeper understanding by satiating the drive for more knowledge once some
1212 skeletal level of causal comprehension is reached.

1213 **Uncited references**

1214 Barrett, Abdi, Murphy, and Gallagher (1993), Bjoerkman (1994), Gigerenzer (1996),
1215 Gigerenzer, Hoffrage, and Kleinboelting (1991), Juslin (1993a, 1993b, 1994), Murphy and
1216 Allopenna (1994), Sloman, Love, and Ahn (1997), Solomon, Medin, and Lynch (1999),
1217 Thagard (2000), Winterfeldt and Edwards (1986), and Wisniewski and Medin (1994).

1218 **Note**

1219 1. The Graduate Arrogance theory was especially strongly advocated by the undergraduate
1220 research assistants in our lab.

1221 **Acknowledgments**

1222 The authors would like to thank Paul Bloom and Robert Sternberg for their helpful comments
1223 on early drafts of this paper. We would also like to thank our diligent and talented research
1224 assistants, especially Nicholas Noles and Emily McKee, for their help with the preparation
1225 of this manuscript. This research was funded by NIH Grant R01-HD23922 to Frank Keil.

1226 **Appendix A. Devices studies**

1227 *A.1. Devices studies stimuli*

1228 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 |

1229 **Appendix A.** (*Continued*)

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

1231 *A.2. Phase 3 instructions (write explanations) used in the devices studies*

1232 Now, we’d like to probe your knowledge in a little more detail, on some of the items.

1233 For each of the following, please describe all the details you know about the phenomena,

1234 going from the first step to the last, and providing the causal connection between the steps. That

1235 is, your explanation should state precisely how each step causes the next step in one continuous

1236 chain from start to finish. In other words, for each phenomenon, try to tell as complete a story

1237 as you can, with no gaps.

1238 If you find that your story does have gaps (i.e., you are not sure how the steps are connected)

1239 please write the word “GAP” in your description at that point, and then continue. Feel free to

1240 use labeled diagrams, or flow-charts to get your meaning across.

1241 When you are done, please re-rate your knowledge of the phenomenon on a 1–7 scale.

1242 **Appendix B. Procedures study**1243 *B.1. Procedures stimuli*

1244 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 |

1246 *B.2. Phase 3 instructions (write explanations) used in the procedures study*

1247 Now, we'd like to probe your knowledge in a little more detail on some of the items.

1248 For each of the following, please describe all the steps in the procedure that you know, going
1249 from the first step to the last. For each procedure, try to tell as complete a story as you can,
1250 with no gaps.

1251 If you find that your story does have gaps (i.e., you are not sure about some of the steps or
1252 how they are connected) please write the word "GAP" in your description at that point, and
1253 then continue. Feel free to use labeled diagrams, or flow-charts to get your meaning across.

1254 When you are done, please re-rate your knowledge of the procedures on a 1–7 scale in the
1255 space provided.

1256 **References**

- 1257 Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson
1258 (Eds.), *Cognition and explanation*. Boston, MA: MIT Press.
- 1259 Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality.
1260 *Cognitive Psychology*, *41*, 361–416.
- 1261 Barrett, S. E., Abdi, H., Murphy, G. L., & Gallagher, J. M. (1993). Theory-based correlations and their role in
1262 children's concepts. *Child Development*, *64*, 1595–1616.
- 1263 Bjoerkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Orga-
1264 nizational Behavior & Human Decision Processes*, *58*(3), 386–405.
- 1265 Bjork, R. A. (1998). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.),
1266 *Attention and performance XVII cognitive regulation of performance: Interaction of theory and application*
1267 (pp. 435–459). Cambridge, MA: MIT Press.
- 1268 Boyd, R. (1991). On the current status of scientific realism. In R. Boyd, P. Gaspar, & J. D. Trout (Eds.), *The
1269 philosophy of science* (pp. 195–222). Cambridge, MA: MIT Press.

- 1270 Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic
1271 signs. *Journal of Abnormal Psychology*, 74(3), 271–280.
- 1272 diSessa, A. D. R. D. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. L. Stevens (Eds.),
1273 *Mental models*. Hillsdale: Erlbaum.
- 1274 Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg
1275 & J. Davidson (Eds.), *The nature of insight*. Cambridge, MA: MIT Press.
- 1276 Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty:
1277 Heuristics and biases* (pp. 422–444). Cambridge, UK: Cambridge University Press.
- 1278 Gelman, S. A., & Koenig, M. A. (2002) Theory-based categorization in early childhood. In D. H. Rakison & L. M.
1279 Oakes (Eds.), *Early category and concept development: Making sense of the blooming buzzing confusion*. New
1280 York: Oxford University Press.
- 1281 Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological
1282 Review*, 103(3), 592–596.
- 1283 Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of
1284 confidence. *Psychological Review*, 98(4), 506–528.
- 1285 Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology:
1286 Learning, Memory, & Cognition*, 11(1–4), 702–718.
- 1287 Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment
1288 of comprehension. *Memory & Cognition*, 10(6), 597–602.
- 1289 Goldstone, R. L., & Barsalou, L. (1998). Reuniting perception and conception. *Cognition*, 65, 231–262.
- 1290 Goldstone, R. L., & Johansen, M. K. (2002) Conceptual development: From origins to asymptotes. In D. H. Rakison
1291 & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming buzzing confusion*.
1292 New York: Oxford University Press.
- 1293 Gopnik, A. A., & Wellman, H. M. (1994). The theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the
1294 mind: Domain specificity in cognition and culture* (pp. 257–293). Cambridge, UK: Cambridge University Press.
- 1295 Hampton, J. A. (2001). The role of similarity in natural categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity
1296 and categorization*. Oxford: Oxford University Press.
- 1297 Hull, D. L. (1965). The effect of essentialism on taxonomy: 2000 years of stasis. *British Journal of Philosophy of
1298 Science*, 15, 314–326.
- 1299 Juslin, P. (1993a). *An ecological model of realism of confidence in one's general knowledge*. Uppsala, Sweden:
1300 Uppsala Universitet, Acta Universitatis Upsaliensis.
- 1301 Juslin, P. (1993b). An explanation of the hard–easy effect in studies of realism of confidence of one's general
1302 knowledge. *European Journal of Cognitive Psychology*, 5(1), 55–71.
- 1303 Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of
1304 almanac items. *Organizational Behavior & Human Decision Processes*, 57(2), 226–246.
- 1305 Keil, F. C. (1994). Explanation-based constraints on the acquisition of word meaning. *Lingua*, 92, 169–196.
- 1306 Keil, F. C. (1998). Cognitive science and the origins of thought and knowledge. In R. M. Lerner (Ed.), *Theoretical
1307 models of human development* (5th ed., Vol. 1). New York: Wiley.
- 1308 Kindersley, D. (1996). *The Way Things Work 2.0*. [CD-ROM]: DK Multimedia.
- 1309 Koriat, A. (1995). Dissociating knowing and feeling of knowing: Further evidence for the accessibility model.
1310 *Journal of Experimental Psychology: General*, 124, 311–333.
- 1311 Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue familiarity and accessibility heuristics
1312 to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27, 34–53.
- 1313 Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incom-
1314 petence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134.
- 1315 Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The meta-cognitive
1316 error of overestimating change-detection ability. *Visual Cognition*, 7, 397–412.
- 1317 Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?
1318 *Organizational Behavior & Human Decision Processes*, 20(2), 159–183.
- 1319 Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and
1320 instruction. *Contemporary Educational Psychology*, 23(4), 345–391.

ARTICLE IN PRESS

42

L. Rozenblit, F. Keil / *Cognitive Science* 92 (2002) 1–42

- 1321 Mandler, J., & Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*,
1322 9, 111–151.
- 1323 Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development*,
1324 48(3), 986–992.
- 1325 Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsis-
1326 tencies. *Child Development*, 50(3), 643–655.
- 1327 Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, 44, 1469–1481.
- 1328 Medin, D. L., & A. Ortony. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and*
1329 *analogical reasoning* (pp. 179–195). Cambridge, UK: Cambridge University Press.
- 1330 Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in meta-cognition. *Journal of*
1331 *Experimental Psychology: Learning, Memory, and Cognition*, 19, 851–861.
- 1332 Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, 10,
1333 151–177.
- 1334 Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experi-*
1335 *mental Psychology: Learning, Memory, & Cognition*, 20(4), 904–919.
- 1336 Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning.
1337 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 53A, 962–982.
- 1338 Murphy, G. L., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(2),
1339 289–316.
- 1340 Murphy, G. L. (2000). Explanatory concepts, In R. A. Wilson & F. C. Keil (Eds.), *Explanation and cognition*.
1341 Cambridge MA: MIT Press.
- 1342 Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- 1343 Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes.
1344 *Psychological Review*, 84(3), 231–259.
- 1345 Prinz, J. (in press). *Furnishing the mind*. Cambridge, MA: Bradford Books, MIT Press.
- 1346 Putnam, H. (1975). The meaning of “meaning.” In K. Gunderson (Ed.), *Language, mind, and knowledge*
1347 (pp. 131–193). Minneapolis: University of Minnesota Press.
- 1348 Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.
- 1349 Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.
- 1350 Sloman, S. A., Love, B. C., & Ahn, W. (1997). Feature centrality and conceptual coherence. *Cognitive Science*, 22,
1351 189–228.
- 1352 Solomon, K. O., Medin, D. L., & Lynch, E. B. (1999). Concepts do more than categorize. *Trends in Cognitive*
1353 *Science*, 3(3), 99–105.
- 1354 Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*,
1355 99(4), 605.
- 1356 Steigler, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87, 245–251.
- 1357 Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, 1, 91–114.
- 1358 Thagard, P., Eliasmith, C., Rusnock, P., & Shelley, C. P. (in press). Knowledge and coherence. In R. Elio (Ed.),
1359 *Common sense, reasoning, and rationality* (Vol. 11). New York: Oxford University Press.
- 1360 West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differ-
1361 ences in performance estimation bias. *Psychonomic Bulletin & Review*, 4(3), 387–392.
- 1362 Wilson, R. A., & Keil, F. C. (1998). The shadows and shallows of explanation. *Minds & Machines*, 8(1), 137–159.
- 1363 Winterfeldt, D. V., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge
1364 University Press.
- 1365 Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive*
1366 *Science*, 18, 221–281.
- 1367 Yates, J. F., Lee, J. W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation.
1368 *Organizational Behavior & Human Decision Processes*, 65(2), 138–147.
- 1369 Yates, J. F., Lee, J.-W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response
1370 style, and reality. *Organizational Behavior & Human Decision Processes*, 70(2), 87–94.