Journal of Experimental Psychology: General
2016, Vol. 145, No. 5, 573–588

0096-3445/16/$12.00 http://dx.doi.org/10.1037/xge0000161

Reflecting on Explanatory Ability: A Mechanism for Detecting Gaps in Causal Knowledge

Dan R. Johnson, Meredith P. Murphy, and Riley M. Messer
Washington and Lee University

People frequently overestimate their understanding—with a particularly large blind-spot for gaps in their causal knowledge. We introduce a metacognitive approach to reducing overestimation, termed *reflecting on explanatory ability* (REA), which is briefly thinking about how well one could explain something in a mechanistic, step-by-step, causally connected manner. Nine experiments demonstrated that engaging in REA just before estimating one's understanding substantially reduced overestimation. Moreover, REA reduced overestimation with nearly the same potency as generating full explanations, but did so 20 times faster (although only for high complexity objects). REA substantially reduced overestimation by inducing participants to quickly evaluate an object's inherent causal complexity (Experiments 4–7). REA reduced overestimation by also fostering step-by-step, causally connected processing (Experiments 2 and 3). Alternative explanations for REA's effects were ruled out including a general conservatism account (Experiments 4 and 5) and a covert explanation account (Experiment 8). REA's overestimation-reduction effect generalized beyond objects (Experiments 1–8) to sociopolitical policies (Experiment 9). REA efficiently detects gaps in our causal knowledge with implications for improving self-directed learning, enhancing self-insight into vocational and academic abilities, and even reducing extremist attitudes.

Keywords: causal knowledge, explanation, metacognition, overestimation, reflection

Supplemental materials: http://dx.doi.org/10.1037/xge0000161.supp

"True wisdom is knowing what you don't know"
—Socrates (Dialogues of Plato, Raffaello Sanzio da Urbino, 1509/1986)

We frequently overestimate our understanding of the world around us—with a particularly large blind-spot for gaps in our causal knowledge. We overestimate our understanding of the often hidden causal mechanisms operating in daily life, like those underlying how a bicycle works (Lawson, 2006; Rozenblit & Keil, 2002), or the consequences of implementing a merit-based teacher pay policy (Fernbach, Rogers, Fox, & Sloman, 2013), or even the steps involved in generating an argument to support your position (Fisher & Keil, 2014). Most research suggests that we need to go through the time-intensive process of explanation *generation* to detect gaps in our causal knowledge (Fernbach et al., 2013; Fisher & Keil, 2014; Rozenblit & Keil, 2002; Walker, Lombrozo, Legare, & Gopnik, 2014). The current studies explore a novel question—can brief explanation *reflection* detect gaps in causal knowledge? Can reflecting on our ability to explain something, in a step-by-step, causally connected manner, help us gain insight into our knowledge? Put another way, can *reflecting on our explanatory ability* (REA) help us detect gaps in our causal knowledge and consequently reduce overestimation?

Research from diverse literatures strongly supports explanation generation's effectiveness in detecting causal knowledge gaps, from text comprehension and learning (Ainsworth & Burcham, 2007; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; McNamara & Magliano, 2009) to political extremism (Fernbach et al., 2013) to inductive reasoning in children (Walker et al., 2014). However, explanation generation is prohibitively labor- and time-intensive, which is why we are not likely to rely on it to gauge our understanding (e.g., Dunlosky et al., 2013; Karpicke, Butler, & Roediger, 2009; Rozenblit & Keil, 2002). Instead, we employ folk theories, that is, causally impoverished, overgeneralized understandings to navigate the world (Gelman & Legare, 2011; Keil, 2012). Applying folk theories leads us to overestimate how well we understand how things work (Rozenblit & Keil, 2002). For example, termed the illusion of explanatory depth, studies show individuals substantially overestimate their understanding of complex material and that generating a detailed explanation of the material's underlying mechanisms reveals this illusion and reduces overestimation (Alter, Oppenheimer, & Zemla, 2010; Fernbach et al., 2013; Lawson, 2006; Rozenblit & Keil, 2002).

Need for Improving Self-Insight Into Knowledge and Abilities

The ubiquitous nature of our tendency to overestimate our knowledge and abilities is supported by abundant literature (Alter et al., 2010; Atir, Rosenzweig, & Dunning, 2015; Bjork, Dunlosky, & Kornell, 2013; Fernbach et al., 2013; Fisher, Goddu, & Keil, 2015; Fisher & Keil, 2014; Kelemen et al., 2013; Moore & Healy,

This article was published Online First March 21, 2016.

Dan R. Johnson, Meredith P. Murphy, and Riley M. Messer, Psychology Department, Washington and Lee University.

We thank Washington and Lee University for funding this project with the Lenfest Grant and Summer Scholar Research Fund.

Correspondence concerning this article should be addressed to Dan R. Johnson, Psychology Department, Washington and Lee University, Lexington, VA 24450. E-mail: johnsondr@wlu.edu

2008; Rozenblit & Keil, 2002; Zell & Krizan, 2014). For example, in a recent metasynthesis, Zell and Krizan (2014) showed that people overestimate numerous ecologically meaningful abilities, including vocational ability, academic ability, medical skills, athletic ability, and intellectual ability. The growing use of the Internet also seems to be increasing people's tendency to overestimate their knowledge (Fisher et al., 2015). Overestimating one's knowledge while learning new material can hinder the application of successful study strategies (Bjork et al., 2013). One study even demonstrated that individuals with high self-perceived expertise claimed impossible knowledge (Atir et al., 2015). Notably, two independent, large literature reviews suggested people overestimate their knowledge and abilities more when they are self-estimating in domains reliant on complex causal knowledge (e.g., integrative tasks requiring many steps and/or multiple skills; Moore & Healy, 2008; Zell & Krizan, 2014).

People's tendency to overestimate their knowledge has important implications for the formation and modification of sociopolitical attitudes. A recent article showed that people tended to overestimate how well they could support their positions on important sociopolitical issues, like capital punishment (Fisher & Keil, 2014). Indeed, overestimating one's knowledge of such controversial issues may partially underlie the formation of extremist attitudes (Fernbach et al., 2013). Fortunately, recent work highlights the power explanation generation has to reduce overestimation of one's knowledge in the sociopolitical domain (Fisher & Keil, 2014) and it even reduced extremist attitudes (Fernbach et al., 2013).

However, given the labor- and time-intensive nature of explanation generation, there is a need for efficient and effective tools to help individuals calibrate their self-estimated knowledge and abilities to objectively rated knowledge and abilities. Developing such a tool was the primary goal of the current set of experiments. Next, evidence will be reviewed to determine whether reflection alone can serve as an efficient and effective overestimation-reduction tool.

Explanation Theory

Explanation theory purports causal relationships and mental simulations are core to explanation (Keil, 2006; Lombrozo, 2006, 2012; Sloman & Lagnado, 2015). Generating explanations guide judgment by allowing people to mentally simulate underlying mechanisms, often constrained to a narrative, step-by-step structure (Sloman & Lagnado, 2015). The critical role given to mental simulation of step-by-step thinking points to a novel prediction—perhaps *thinking* about explanation is sufficient to reap benefits.

Preliminary support for this idea comes from a study of the illusion of explanatory depth where participants who rated their understanding of objects with a more concrete mindset significantly reduced their overestimation (Alter et al., 2010). Alter et al. (2010) manipulated the framing of how participants rated their understanding of how objects, like a zipper, work. Abstract framing asked participants to judge how well they understood how objects work, whereas concrete framing asked participants to judge how well they understood how the parts of an object enable it to work. This subtle change in framing significantly reduced overestimation and suggests that simply thinking or reflecting about one's understanding differently may be sufficient to reduce overestimation. However, although it seems likely that the concrete framing manipulation induced step-by-step thinking, the mechanism underlying how this step-by-step thinking reduced overestimation remains unknown.

While Alter et al.'s (2010) results point to reflection's utility, both explanation theory and other findings suggest that reflection alone may be ineffective at reducing overestimation. There is general agreement that when evaluating understanding, one will either focus on the purpose of an object, that is, take a teleological explanatory stance or take a mechanistic explanatory stance and focus on the mechanistic underpinnings that enable an object to work (Keil, 2006; Lombrozo, 2012). For example, when considering their understanding of a vacuum cleaner, individuals may consider its main function, to clean, or how the motor creates a pressure differential for suction. When asked to evaluate understanding, children, adults, and even professional scientists are biased to take a teleological explanatory stance at the expense of a mechanistic explanatory stance (Kelemen, 1999; Kelemen & Rosset, 2009; Kelemen, Rottman, & Seston, 2013). This teleological explanatory bias leads people to make errors in causal reasoning and overestimate their understanding (Kelemen, 1999; Kelemen & Rosset, 2009; Kelemen et al., 2013). Studies show that even when participants are encouraged to carefully reflect about their causal reasoning decisions, their errors persist (Kelemen et al., 2013; Kelemen & Rosset, 2009). These results suggest that careful, but *unguided reflection* does little to combat overestimation. However, Alter et al.'s (2010) work suggests *guided reflection* or thinking in a concrete manner about understanding may effectively reduce overestimation.

Current Studies

The current experiments take a novel metacognitive approach to help individuals discover gaps in their casual knowledge. Before being asked to evaluate their understanding of how a common object works, participants engaged in a brief, guided reflection period where they considered how well they could explain how an object works—a process we termed reflecting on explanatory ability (REA). We propose that reflecting in a step-by-step, causally connected manner should induce a strong mechanistic explanatory stance and therefore allow participants to detect gaps in their causal knowledge. In addition, we propose REA operates by inducing participants to assess the overall causal complexity of the object and then anchor their estimate of understanding on this assessment. Challenging prior literature that suggests overestimation is a tenacious metacognitive bias that requires full explanation generation to combat, we propose REA-guided reflection will reduce overestimation.

To preview the following nine experiments, all experiments demonstrated that REA substantially reduced overestimation compared to unguided reflection. Experiments 2 and 3 showed that performing step-by-step, causally connected processing during the REA reflection period was critical to reducing overestimation and simultaneously ruled out a comparative ignorance effect as an alternative account (Fox & Tversky, 1995; Fox & Weber, 2002). Experiments 4 through 8 strongly supported that the underlying mechanism driving REA's overestimation-reduction effect is that it induces participants to perform an assessment of an object's causal complexity and then anchor their understanding estimate on an object's inherent complexity. Experiments 2 and 6 through 8 ruled out differences in scale interpretation or framing as alternative accounts (e.g., Alter et al., 2010). Experiments 4 and 5 ruled out a conservatism account and a sensitivity to fluency account of REA's overestimation-reduction effect. Experiment 8 ruled out a covert explanation account of REA's effect. Finally, Experiment 9 generalized REA's overestimation-reduction effect to the sociopolitical domain and demonstrated REA reduced extremist attitudes (Fernbach et al., 2013). In support of REA serving as an efficient and effective overestimation-reduction tool, REA reduced overestimation with comparable potency to explanation generation, but did so 20 times faster (although only for high complexity objects).

Experiment 1: Reflection on Explanatory Ability (REA)

The primary goal of Experiment 1 was to investigate REA's ability to reduce overestimation of understanding. Participants rated their understanding of how a common object worked. However, before estimating their understanding, participants either engaged in unguided but careful reflection about how the object works, reflected on their ability to explain how the object works, or typed out their explanation of how the object works. It is hypothesized that reflecting on one's ability to explain will induce step-by-step, causally connected processing that reveals gaps in causal knowledge, thereby lowering participants' self-estimated understanding ratings. In addition, given both reflection conditions were equally engaged in careful reflection, any REA effects should be due to a step-by-step reflection mode, instead of a generalized caution-inducing or more deliberative reflection mode. Given that this represents the introduction of REA, it is unclear how it will perform compared with explanation generation.

Method

Participants. A sample of 189 participants (64% female, 36% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; M = 34.05, range 18–68). See supplemental material available online for methods used to determine sample sizes for all experiments.

Design and procedure. Participants were randomly assigned to the unguided reflection (n = 63), REA (n = 65), or explanation generation (n = 61) condition in a fully between-subjects design. Using Qualtrics software, participants were first given instructions about what they were supposed to do before rating their understanding of how an objects works. These instructions varied by condition, where participants in the unguided reflection condition were asked to "*carefully reflect on your understanding of how the object works,*" participants in the REA condition were asked to "*carefully reflect on your ability to explain to an expert, in a step-by-step, causally-connected manner, with no gaps in your story how the object works,*" and participants in the explanation generation condition were asked to "*type out your full explanation as if you were explaining to an expert, in a step-by-step, causally-connected manner, with no gaps in your story how the object works.*" In addition, participants were told they would be rating their understanding on a 1 (*shallow understanding*) to 7 (*deep understanding*) scale. Then, all participants were warned that on the next page, they will be asked to type out the instructions in their own words without copying and pasting and that we may not be able to pay them unless they did this successfully (all participants were subsequently paid regardless of their responses). Next, they were given a text box in which to type the instructions and on the next page they were given the same instructions again in case they felt they needed a reminder after typing the instructions. This instruction testing process was utilized because the instructions constituted the primary manipulation and MTurkers tend to work quickly through studies (Rand, 2012). Note that all remaining experiments used this instruction testing process to ensure participants knew how to reflect and what scale they needed to use to rate their understanding.

After the instruction testing process, all participants were asked to reflect according to their prompt (i.e., above instructions) for 15 s before rating their understanding of how a vacuum cleaner works on the 1 to 7 scale. A vacuum cleaner was selected because it has sufficient complexity and visible parts to foster overestimation of understanding (Rozenblit & Keil, 2002). In the unguided reflection and REA conditions, the computer software controlled the timing of the reflection period, so that participants could not move on until 15 s passed. Immediately after the 15-s reflection period, the next screen appeared where participants gave their understanding rating in a self-paced manner. Participants in the explanation generation condition typed out their full explanations of how a vacuum cleaner works before giving their understanding rating. Both typing and rating periods were self-paced. Finally, all participants completed demographic questions. Note that all remaining experiments ended with demographics.

Results

Confidence intervals around Cohen's *d*. To inform the magnitude of the differences in head to head comparisons between REA and other conditions and to determine which effect sizes given below substantively differ in magnitude, 95% confidence intervals (CI) were computed around Cohen's d (Cumming, 2012; Fritz, Morris, & Richler, 2012), where nonoverlapping CIs are significantly different effect sizes. All CIs reported in future studies are 95% confidence intervals (CI) around Cohen's d.

Understanding ratings. A one-way analysis of variance (ANOVA) on understanding ratings revealed a main effect for condition, $F(2, 186) = 17.55$, $p < .001$, $\eta^2 = .16$. Follow-up t tests with a Bonferroni-adjusted criterion of .017 (total $\alpha = .05$) indicated that compared to unguided reflection ($M = 6.11$, $SD = 1.17$), both REA ($M = 5.20$, $SD = 1.46$), $t(126) = 3.90$, $p < .001$, $d = 0.69$, CI [0.50 – 0.88], and explanation generation ($M = 4.62$, $SD = 1.58$), $t(122) = 5.98$, $p < .001$, $d = 1.07$, CI [0.85 – 1.29] significantly reduced overestimation of understanding. Explanation generation marginally reduced overestimation more than REA, $t(124) = 2.13$, $p = .035$, $d = 0.38$, CI [0.20 – 0.56]. Note that compared with unguided reflection, REA's power to reduce overestimation was moderate to large ($d = 0.69$), and explanation generation's ($d = 1.07$) was large, with overlapping confidence intervals around Cohen's d. These results indicate while explanation generation does edge out REA in its ability to reduce overestimation, they have comparable effects.

Response time of understanding ratings. It is important to note that there were significant differences between conditions in how long participants took to give their understanding rating, where the explanation generation condition took the longest, followed by the REA condition, and then unguided condition (see Table S1). However, all prior analyses were repeated using an Analysis of Covariance (ANCOVA) where response duration served as the covariate and all results maintained the same patterns of statistical significance and effect size and therefore are not reported. In addition, correlations between understanding ratings and their response times were run in all conditions and none were statistically significant.

Discussion

Supporting the predictions, REA substantially decreased overestimation of understanding. Both unguided reflection and REA conditions were engaged in careful reflection for identical durations. Therefore, REA's effects cannot be attributed to participants simply becoming more cautious during a more deliberative reflection mode. Instead, REA likely engaged step-by-step, causally connected processing that allowed participants to detect gaps in their causal knowledge. Remarkably, engaging in a brief 15-s REA reflection period reduced overestimation to levels comparable to those achieved by writing out a full explanation.

Experiment 2: REA and Causally Connected Processing

Experiment 2 tested whether step-by-step, causally connected processing is critical to REA's effects. In addition, the role of using the word *expert* in the reflection prompt was also investigated. Considering one's knowledge in the context of expert knowledge could simply make REA participants more conservative by inducing a state of comparative ignorance (Fox & Tversky, 1995; Fox & Weber, 2002).

In addition, to provide better comparability to previous literature on the illusion of explanatory depth (Rozenblit & Keil, 2002), explicit instructions and concrete examples were given so that each value on the understanding rating scale represented a particular level of knowledge. These scale instructions should reduce ambiguity in interpretation of how to use the understanding scale. For example, some participants may have thought they were to rate their understanding based on their knowledge of how to use the object, as opposed to how the parts of the object work together to make it function.

Method

Participants. A sample of 117 participants (70% female, 30% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; M = 36.98, range 18–78).

Design and procedure. To provide comparability to previous literature on the illusion of explanatory depth, participants were first given identical scale training instructions from Rozenblit and Keil (2002). The scale training provided example text of "1," "4," and "7" levels of knowledge about how a crossbow works and explicitly stated that these values represent the, "1," "4," and "7" on the 1 through 7 understanding rating scale. In addition, sample diagrams were provided that represented what a participant with a "1," "4," and "7" level of knowledge could draw.

Next, participants were randomly assigned to the unguided reflection (n = 58) or REA-noncausal (n = 59) condition. Participants in the unguided reflection condition were told to "*carefully reflect on your understanding of how the object works,*" before giving their initial understanding rating. Participants in REA-noncausal condition were told to "*carefully reflect on your ability to explain to an expert how the object works*," before giving their initial understanding rating. Participants reflected for 15 s according to their respective prompts (computer-paced), before giving their understanding rating (self-paced) of each object. They did this serially for eight objects in the following order: piano key, smoke detector, VCR, ear buds, gas stove, treadmill, Polaroid camera, and power drill. To reduce ambiguity in which version of these objects to rate, a prototypical picture of the object was presented during the 15 s reflection period (see supplemental material available online for pictures). Comparing these two conditions will test the role of the word *expert* in the REA reflection prompt.

To test whether step-by-step, causally connected processing is critical to REA's effects, a within-subjects condition was added. Next, participants in both conditions were given a REA-causal reflection prompt and asked to, "*carefully reflect on your ability to explain to an expert, in a step-by-step, causally-connected manner, with no gaps in your story, how the object works,*" for 15 s (computer-paced) before giving a second understanding rating (self-paced) for each object. They did this serially for the same eight objects in the same order used for their initial understanding ratings. If step-by-step, causally connected processing is crucial to REA then participants in the REA-noncausal condition should additionally reduce their understanding ratings after receiving the REA-causal prompt.

It was a 2 (between-subjects; unguided reflection vs. REA-noncausal reflection) by 2 (within-subjects; pre-REA-causal reflection vs. post-REA-causal reflection) mixed design. See Figure 1 for a depiction of the design.

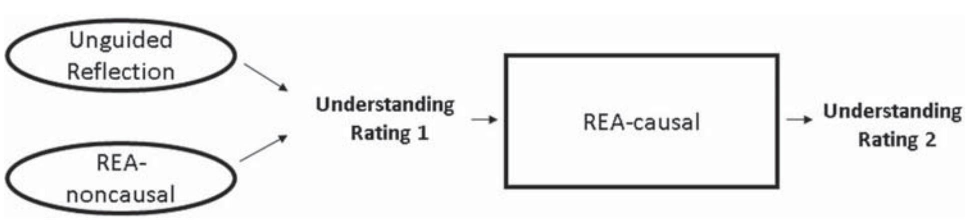

Figure 1. Design of Experiment 2.

Results

Understanding ratings. A 2 (unguided vs. REA-noncausal) × 2 (pre-REA-causal vs. post-REA-causal) mixed ANOVA was performed on understanding ratings averaged across all eight objects. It revealed that ratings significantly dropped from pre- to post-REA-causal reflection, $F(1, 115) = 121.44$, $p < .001$, $\eta_p^2 = .51$, but this effect was qualified by an interaction between condition and pre- to post-REA-causal reflection, $F(1, 115) = 21.40$, $p < .001$, $\eta_p^2 = .16$. Follow-up *t* tests indicated participants in the REA-noncausal reflection condition ($M = 3.98$, $SD = 1.23$) had significantly lower understanding ratings compared to participants in the unguided reflection condition ($M = 4.77$, $SD = 1.15$), $t(115) = 3.58$, $p < .001$, $d = 0.66$, CI [0.46, 0.86], with no difference between conditions after the REA-causal reflection period, $p = .98$ (see Figure 2). To ensure the effects were not attributable to any specific object or subset of objects, a 3-way mixed ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 3-way interaction (see Table S2 for understanding ratings for object means).

To determine whether adding REA-causal guidance to the reflection prompt increases REA's power to reduce overestimation above and beyond simply using the word expert, follow-up *t* tests compared understanding ratings from before to after REA-causal reflection. In the REA-noncausal condition, REA-causal reflection significantly reduced understanding ratings from before ($M = 3.98$, $SD = 1.23$) to after it ($M = 3.44$, $SD = 1.53$), $t(58) = 4.60$, $p < .001$, $d = 0.60$, CI [0.32, 0.88]. Given the unguided reflection condition had not yet engaged in REA, for them, REA-causal reflection dramatically reduced understanding ratings from before ($M = 4.77$, $SD = 1.15$) to after it ($M = 3.45$, $SD = 1.43$), $t(57) = 10.88$, $p < .001$, $d = 1.43$, CI [1.06, 1.79]. These results indicated that additional guidance in REA that explicitly instructs participants to reflect in a step-by-step, causally connected manner potentiates REA's power to reduce overestimation of understanding.

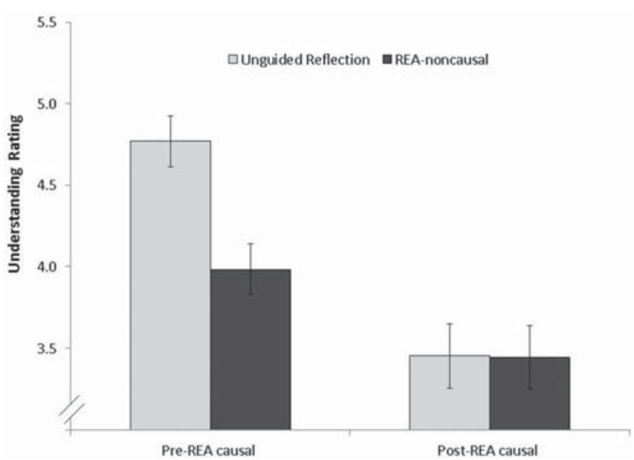

Figure 2. Experiment 2 understanding ratings, depicting REA-causal's ability to reduce overestimation. Specifically the significant drop for the REA-noncausal condition (black bars) from before to after REA-causal reflection demonstrates that step-by-step, causally connected instructions are critical to REA's effect. Standard error bars of the mean are shown.

Response time of understanding ratings. It is important to note that there were no significant differences between conditions for any of the eight objects in how long participants took to give their initial understanding rating (see Table S1 for rating time means). In addition, correlations between initial ratings and their response times were run in both conditions and none were statistically significant.

Discussion

Explicitly instructing participants to use step-by-step, causally connected thinking reduced overestimation more than having participants reflect on their knowledge in comparison to an expert. Ruling out a comparative ignorance explanatory account (Fox & Tversky, 1995; Fox & Weber, 2002), REA engaged a step-by-step processing mode that allowed participants to detect gaps in their causal knowledge. It is worth reiterating that participants following the REA-noncausal and REA-causal prompts were both engaged in explanation reflection where an expert was the target, so this comparison isolated the role of step-by-step, causally connected processing.

In addition, explicitly defining the levels of knowledge required to select a particular number on the understanding rating scale provided comparability to the illusion of explanatory depth literature (Rozenblit & Keil, 2002). The results provide converging evidence that reflection alone can reduce overestimation—as opposed to needing to generate a full explanation. In addition, the scale training instructions should have minimized differences in scale interpretation. However, differences in scale interpretation across REA and unguided reflection conditions remain possible. Scale interpretation and framing are directly addressed in Experiment 6.

In sum, these results indicate that one mechanism underlying REA's ability to reduce overestimation is that it induces a mechanistic explanatory stance that allows one to detect gaps in their causal knowledge (Keil, 2006; Lombrozo, 2012). However, given the within-subjects nature of the design, alternative explanations remain possible. All participants rated their understanding of the same objects twice. Participants may have felt compelled to decrease their understanding ratings for the second ratings because it was expected of them. To rule out demand characteristics, Experiment 3 will utilize a fully between-subjects design.

Experiment 3: REA and Causally Connected Processing (Between-Subjects)

Method

A sample of 177 participants (59% female, 41% male) was recruited from Amazon's Mechanical Turk, where each participant received a \$0.50 payment for participation (age; $M = 38.01$, range 18–78).

Experiment 3 was identical to Experiment 2 except for the following changes. There was no within-subjects condition, and instead participants were randomly assigned to the unguided reflection ($n = 58$), REA-noncausal ($n = 60$), or REA-causal ($n = 59$) condition in a fully between-subjects design. The understanding rating period was changed from self-paced to computer-paced, where all participants were allotted 5 s to give their rating of each object. Finally, instead of eight objects, the following four objects were rated in the following order: printer, treadmill, power drill, and gas stove.

Results

A one-way ANOVA on understanding ratings revealed a main effect for condition, $F(2, 174) = 10.53$, $p < .001$, $\eta^2 = .11$. Follow-up *t* tests with a Bonferroni-adjusted criterion of .017 (total $\alpha = .05$) indicated that compared to unguided reflection ($M = 4.96$, $SD = 1.06$), REA-noncausal reflection ($M = 4.65$, $SD = 1.20$), did not significantly reduce overestimation ($p = .151$). However, compared to unguided reflection, REA-causal reflection ($M = 3.99$, $SD = 1.24$) substantially reduced overestimation, $t(115) = 4.54$, $p < .001$, $d = 0.84$, CI [0.63 – 1.05]. In addition, compared to REA-noncausal reflection, REA-causal reflection significantly reduced overestimation, $t(117) = 2.98$, $p = .003$, $d = 0.54$, CI [0.35 – 0.73]. To ensure the effects were not attributable to any specific object or subset of objects, a 2-way ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 2-way interaction (see Table S2 for understanding ratings for object means).

Discussion

These results replicated Experiment 2's findings and provide converging evidence that step-by-step, causally connected instruction is critical to REA's effects. REA appears to induce a mechanistic explanatory stance that helps participants detect gaps in their causal knowledge.

Experiment 4: Object Complexity and a Causal Complexity Assessment Account of REA

The primary goal of Experiment 4 was to investigate another potential mechanism through which REA reduces overestimation. We propose a *causal complexity assessment* account of REA's overestimation-reduction effect. If given only a short time to reflect on how well one could explain how something works, it would be advantageous to first get a sense of how many steps are involved, that is, assess the overall causal complexity. That way one could anchor on the overall causal complexity before judging how well one could explain something. For example, if REA participants quickly assess that a vacuum cleaner is highly complex, then they could deduce they will not likely be able to explain a high percentage of the steps involved in making it work. In contrast, without REA, unguided participants should not have the benefit of recognizing an object's inherent complexity and consequently overestimate their understanding—as already demonstrated in the illusion of explanatory depth (Rozenblit & Keil, 2002).

If REA reveals gaps in participants' causal knowledge by inducing a quick assessment of causal complexity, then it should reduce overestimation to a greater degree for objects high in complexity. This is because there are more opportunities to miss the hidden causal mechanisms operating in these objects. Supporting this idea, Rozenblit and Keil (2002) demonstrated that participants overestimated their understanding more for objects higher in complexity. In contrast, after REA participants assess the overall complexity of low complexity objects, they will have no reason to reduce their overestimation because they believe there are few total causal steps to explain. Consequently, REA participants should exhibit similar understanding ratings to unguided reflection participants for low complexity objects, but substantially lower understanding ratings for high complexity objects.

It is important to note that the current design will also address whether a general *conservatism account* can explain REA's effects. It is possible that asking participants to reflect on their ability to explain how an object works to an expert in a step-by-step manner simply makes them more conservative. However, if REA only makes participants more conservative, they should apply this bias equally to low and high complexity objects. In contrast to the interaction predicted by the causal complexity assessment account, a general conservatism account predicts equivalent reductions in overestimation across low and high complexity objects for REA participants.

Method

Participants. A sample of 121 participants (62% female, 38% male) was recruited from Amazon's Mechanical Turk, where each

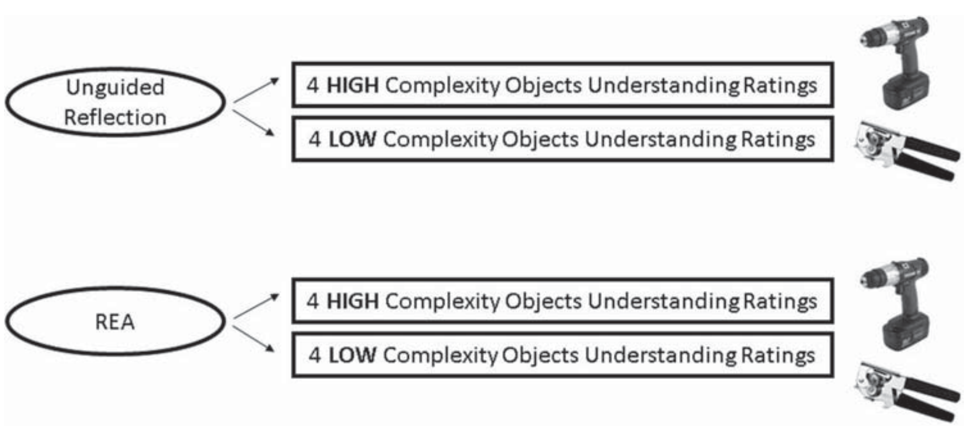

Figure 3. Design of Experiment 4. A power drill is pictured as an example of a high complexity object and a can opener is pictured as an example of a low complexity object.

participant received a \$0.50 payment for participation (age; M = 38.23, range 20–70).

Design and procedure. Participants were randomly assigned to reflect for 15 s following the unguided reflection prompt (n = 57) or REA-causal prompt (n = 64) before giving understanding ratings for each object on the 1 (*shallow understanding*) to 7 (*deep understanding*) scale. The computer software controlled the timing of the reflection period, so that participants could not move on until 15 s passed. Immediately after the 15-s reflection period, the next screen appeared and participants were given 5 s to give their understanding rating. All participants rated four low and four high complexity objects, making Experiment 4 a 2 (between-subjects; unguided reflection vs. REA-causal) × 2 (within-subjects; low vs. high complexity) mixed design. See Figure 3 for a depiction of the design.

Participants serially rated their understanding of eight objects in one of four randomly assigned counterbalanced orders. The low complexity objects were a spray bottle, can opener, umbrella, and candle. The high complexity objects were a printer, treadmill, power drill, and gas stove. The four orders were as follows: (a) spray bottle, can opener, umbrella, candle, printer, treadmill, power drill, and gas stove; (b) printer, treadmill, power drill, gas stove, spray bottle, can opener, umbrella, and candle; (c) spray bottle, printer, can opener, treadmill, umbrella, power drill, candle, and gas stove; (d) printer, spray bottle, treadmill, can opener, power drill, umbrella, gas stove, candle.

Results

A 2 (between-subjects; unguided reflection vs. REA) × 2 (within-subjects; low vs. high object complexity) mixed ANOVA was performed on understanding ratings. It revealed a main effect for reflection condition, where understanding ratings were significantly lower for REA compared to unguided reflection, $F(1, 119) = 24.34$, $p < .001$, $\eta_p^2 = .17$. It also revealed a main effect for object complexity, where understanding ratings were dramatically lower for objects high in complexity, $F(1, 119) = 263.78$, $p < .001$, $\eta_p^2 = .69$. But these main effects were qualified by an interaction between reflection condition and object complexity, $F(1, 119) = 18.77$, $p < .001$, $\eta_p^2 = .14$.

Supporting the causal complexity assessment account, follow-up t tests indicated that compared to unguided reflection, REA significantly reduced overestimation of understanding more for high complexity objects, $t(119) = 5.53$, $p < .001$, $d = 1.01$, CI [0.80, 1.23], than for low complexity objects, $t(119) = 2.76$, $p = .007$, $d = 0.50$, CI [0.31, 0.69], (see Figure 4). To ensure the effects were not due to any specific object or subset of objects, a 3-way ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 3-way interaction (see Table S2 for understanding ratings for object means). REA reduced overestimation more when there were greater opportunities to miss gaps in causal knowledge.

Discussion

Supporting the causal complexity assessment account, REA reduced overestimation more for objects high than low in complexity. REA appears to induce participants to quickly assess an object's inherent causal complexity and anchor their understanding ratings on its complexity. For objects high in complexity, REA participants, unlike unguided reflection participants, noted that there are many total causal steps underlying the objects' inner workings and deduced they must not understand how the object works particularly well. For low complexity objects, there are few causal steps for REA participants to detect, and therefore they did not reduce overestimation much compared to unguided reflection participants.

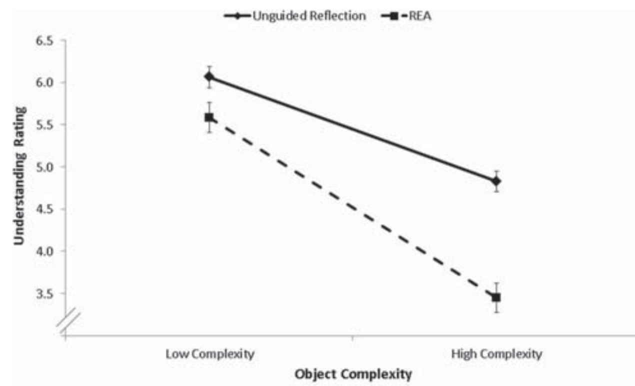

Figure 4. Experiment 4 understanding ratings depicting interaction between reflection condition and object complexity, driven by the fact REA more powerfully reduced overestimation for objects high in complexity, compared with unguided reflection. Standard error bars of the mean are shown.

These results also provide strong evidence against a conservatism account of REA's effects. A general conservatism account predicts REA should reduce overestimation equally across low and high complexity objects. In contrast, REA participants reduced overestimation substantially more for high complexity objects than low complexity objects.

However, it should be noted that the objects used in Experiment 4 were not pretested for perceived complexity or any other relevant object attribute. Therefore, it is difficult to conclude the observed results were due solely to object complexity. In particular, it is possible that high complexity objects are also lower in familiarity, that is, they are encountered less on an everyday basis. For example, one may encounter a candle more often than a power drill and therefore perceive the candle as more familiar. Familiarity is strongly related to feelings of fluency and fluency is a powerful cue on which people rely to judge their knowledge level (Bjork et al., 2013). Specifically, participants should overestimate their understanding of highly familiar objects more than objects lower in familiarity. The potentially lower familiarity of high complexity objects may reduce feelings of fluency when participants judge their understanding and alternatively explain why REA reduces overestimation. Put another way, REA may make people more sensitive to fluency cues instead of inducing a causal complexity assessment. REA may direct attention toward fluency so that participants reduce overestimation more for objects they perceive as both high in complexity and low in familiarity (Bjork et al., 2013). Experiment 5 addressed this *sensitivity to fluency account* as an alternative explanation for REA's effects.

Experiment 5: Object Complexity, Familiarity, and Explanation Generation

The primary goal of Experiment 5 was to match high and low complexity objects on familiarity and consequently, fluency cues. If the causal complexity assessment account can explain REA's effects, then the effects should persist after controlling for familiarity and manipulating object complexity alone. As in Experiment 4, REA participants should reduce overestimation more for high complexity objects than low complexity objects. An additional goal of Experiment 5 was to provide another head to head comparison of REA to explanation generation. REA is proposed to be an efficient alternative to explanation generation and therefore a replication (see Experiment 1) of a direct comparison was needed.

Method

Participants. A sample of 176 participants (67% female, 33% male) was recruited from Amazon's Mechanical Turk, where each participant received a \$0.50 payment for participation (age; $M = 38.39$, range 18–70).

Pilot study to match on familiarity. In a pilot study, a separate sample of participants rated 24 common objects on various dimensions including complexity and familiarity (see pilot study section of supplementary material). For Experiment 5, two objects high in complexity (vacuum cleaner, computer mouse), and two objects low in complexity (Velcro, reading glasses) were selected because they substantially differed on perceived complexity ($d = 2.65$), but were *matched on perceived familiarity* to rule it out as an alternative explanation.

Design and procedure. Participants were randomly assigned to the unguided reflection ($n = 58$), REA ($n = 60$), or explanation generation condition ($n = 58$). In the reflection conditions, participants were asked to reflect for 15 s following the unguided reflection or REA-causal prompt before giving understanding ratings for each object. The computer software controlled the timing of the reflection period, so that participants could not move on before 15 s passed. Immediately after the 15-s reflection period, the next screen appeared where participants gave their understanding rating in a self-paced manner. The explanation generation condition participants typed out an explanation about how each object worked (same prompt from Experiment 1) before giving their understanding rating and self-paced both activities. The two high and two low complexity objects were given in the following order to all participants: vacuum cleaner, Velcro, computer mouse, reading glasses. This makes Experiment 5 a 3 (between-subjects; unguided reflection vs. REA vs. explanation generation) × 2 (within-subjects; low vs. high complexity) mixed design.

Results

A 3 (unguided reflection vs. REA vs. explanation generation) × 2 (low vs. high object complexity) mixed ANOVA was performed on understanding ratings. It revealed a main effect for condition, $F(2, 173) = 23.52$, $p < .001$, $\eta_p^2 = .21$, and that understanding ratings were significantly lower for high complexity objects compared to low complexity objects, $F(1, 173) = 54.25$, $p < .001$, $\eta_p^2 = .24$. But these main effects were qualified by an interaction between condition and object complexity, $F(2, 173) = 3.92$, $p = .022$, $\eta_p^2 = .04$.

REA's effect as function of object complexity. Replicating Experiment 4, follow-up t tests using a Bonferroni-corrected criterion of .01 (adjusted for 5 comparisons, total $\alpha = .05$) indicated that REA reduced overestimation more for high complexity objects, $t(116) = 3.99$, $p < .001$, $d = 0.73$, CI [0.53, 0.93], than for low complexity objects, $t(116) = 2.05$, $p = .043$, $d = 0.38$, CI [0.19, 0.57], compared with unguided reflection (see Figure 5). To ensure the effects were not due to any specific object or subset of objects, a 3-way ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 3-way interaction (see Table S2 for understanding ratings for object means).

REA versus explanation generation. Explanation generation marginally reduced overestimation more than REA (after the Bonferroni-adjusted criterion) for high complexity objects, $t(116) = 2.15$, $p = .033$, $d = 0.40$, CI [0.21, 0.59] (see Figure 5). For low complexity objects, explanation generation reduced overestimation substantially more than REA, $t(116) = 4.18$, $p < .001$, $d = 0.76$, CI [0.55, 0.96]. Compared with unguided reflection, both REA, $t(116) = 3.99$, $p < .001$, $d = 0.73$, CI [0.53, 0.93], and explanation generation, $t(114) = 5.93$, $p < .001$, $d = 1.09$, CI [0.86, 1.32], reduced overestimation for high complexity objects. Given the overlapping CIs around Cohen's d (for high complexity objects) when REA and explanation generation were compared to unguided reflection, it can be concluded that REA reduces overestimation with comparable potency to explanation generation.

REA's efficiency. To compute the time saved using REA versus explanation generation, the total time it took participants to generate explanations for each of the four objects was recorded. For high complexity objects, generating an explanation ($M = 266$, $SD = 215$), took nearly 9 times longer to achieve comparable understanding ratings, compared to the 30 s of total REA reflection time. For low complexity objects ($M = 203$, $SD = 162$), explanation generation reduced overestimation substantially more than REA, but took nearly 7 times longer to do so.

Response time of understanding ratings. It is important to note that there were significant differences between conditions in how long participants took to give their understanding rating,

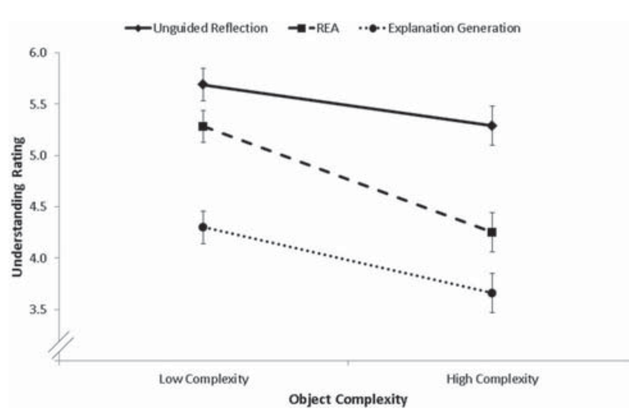

Figure 5. Experiment 5 understanding ratings depicting interaction between condition and object complexity, driven by the fact REA more powerfully reduced overestimation for objects high in complexity, whereas unguided reflection and explanation generation were relatively insensitive to object complexity. Standard error bars of the mean are shown.

where the explanation generation condition took the longest, followed by the REA condition, and then unguided condition—although only for vacuum cleaner and computer mouse. However, all prior analyses were repeated using an ANCOVA where response duration served as the covariate and all results maintained the same patterns of statistical significance and effect size and therefore are not reported (see Table S1 for mean rating times). In addition, correlations between understanding ratings and their response times were run in all conditions and none were statistically significant.

Discussion

Experiment 5 replicated the results of Experiment 4 and supported the hypothesis that REA detects gaps in causal knowledge by inducing a causal complexity assessment. Crucially, by matching high and low complexity objects on familiarity, Experiment 5 ruled out a sensitivity to fluency account of REA's effects. Compared with REA, explanation generation reduced overestimation more consistently for objects low and high in complexity. However, for high complexity objects, REA comparably reduced overestimation in about 10% of the time it took to generate full explanations. This provides striking evidence that REA may be used as an efficient alternative to detect gaps in causal knowledge.

Experiment 6: Direct Test of the Causal Complexity Assessment Account

There were three goals in Experiment 6 including to: (a) more directly test whether one of REA's underlying mechanisms is to induce a quick assessment of the total number of causal steps underlying how an object works, that is, perform a causal complexity assessment; (b) test whether REA does more than induce a concrete mindset (i.e., concrete construal, Alter et al., 2010); and (c) test whether REA's effect could be attributable to differences in scale interpretation.

To provide a more direct test of the causal complexity assessment account, participants first reflected using the unguided reflection or REA-causal prompt. Then, immediately after giving their initial understanding rating, participants in both conditions were asked to estimate how many total steps it would take to explain how the parts enable the object to work. Then, after giving this number, they were given an opportunity to revise their initial estimate of the percentage of steps they know that enable an object to work. If REA induces participants to quickly assess an object's complexity by roughly gauging the number of causal steps involved in explanation, then REA participants should not revise their rating because they've already performed the total steps estimation. In contrast, unguided participants should not consider the total number of steps involved until directly asked, and therefore they should reduce their understanding rating after this revelation.

To eliminate scale interpretation or framing as an alternative explanation of REA's effects, a new scale was created in Experiment 6. To reduce variability in interpretation, all participants in Experiment 2 were given examples of what 1, 4, and 7 levels of knowledge look like and then they were asked to rate their own knowledge in accordance with these examples. However, potentially important differences in scale interpretation between the REA participants and unguided reflection participants remain. REA participants may have assumed understanding was tantamount to how many mechanistic steps they could successfully explain, whereas unguided participants may have been thinking about how well they understood the function of the object or how to use it. There is evidence that when asked to evaluate understanding, children, adults, and even professional scientists are biased to think about function or teleology instead of mechanism (Kelemen, 1999; Kelemen & Rosset, 2009; Kelemen et al., 2013). To eliminate the scale interpretation confound, a new scale was created in Experiment 6 that promoted a concrete, more mechanistic mindset (Alter et al., 2010). Specifically, all participants rated their understanding by indicating what percentage of steps they know are involved in making an object work. Consequently, if REA participants reduce overestimation more than unguided participants for their initial understanding ratings, then the effect must be due to a mechanism other than scale interpretation or framing.

Method

Participants. A sample of 118 participants (64% female, 36% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; M = 38.26, range 19–72).

Design and procedure. First, all participants read instructions about how they will shortly be rating their understanding using a percentage of steps scale. Participants were told they would be asked to rate their understanding by, "*estimate what percentage of steps involved in how the parts enable the object to work do you know?*" with the following choices: 0% to 15%, 16% to 30%, 31% to 45%, 46% to 60%, 61% to 75%, 76% to 90%, 91% to 100%.[1] We intentionally used language that has previously been shown to induce a concrete mindset in all participants (Alter et al., 2010) to test whether REA's effect can be accounted for by scale framing alone. While we recognize the highest scale category has a slightly smaller range than the lower categories, we thought it more important to maintain a 7-point scale for comparability across the current studies and prior literature (e.g., Alter et al., 2010; Rozenblit & Keil, 2002). While participants used the percentage of steps scale, the computer software coded their responses on a 7-point scale and we report all results using the 7-point scale for comparability across studies and prior literature (Rozenblit & Keil, 2002).

Next, participants were randomly assigned to conditions and reflected for 5 s following the unguided reflection (n = 59) or REA-causal (n = 59) reflection prompt before giving their initial understanding rating of how a computer mouse works. The reflection time was reduced from 15 s to 5 s because we were testing whether REA induces participants to *quickly* assess the object's complexity. The computer software controlled the timing of the reflection period, so that participants could not move on until 5 s passed. Immediately after the 5-s reflection period, the next screen automatically appeared and participants were given only 5 s to give their understanding rating before the next instruction screen automatically appeared. Next, all participants were asked to esti-

[1] Thanks to a reviewer for recommending the use of a percentage of steps scale.

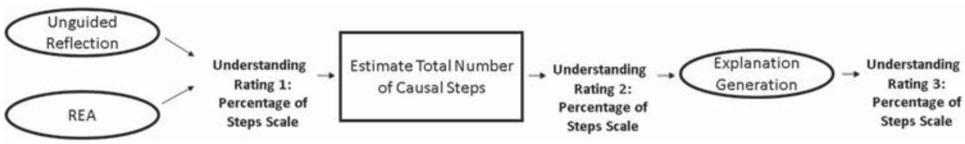

Figure 6. Design of Experiment 6.

mate, by giving an actual number, "*how many total steps it would take to explain how the parts enable the object to work.*" Then, they were given the opportunity to revise their estimate of understanding on the same percentage of steps scale and self-paced this revised rating. Finally, participants were asked to generate an explanation of how a computer mouse works and asked to give a third rating of their understanding in light of their explanation on the same percentage of steps scale. Typing the explanation and the final understanding rating were self-paced. See Figure 6 for a depiction of the design.

Results

REA and a causal complexity assessment. A 2 (unguided vs. REA) × 3 (pretotal steps estimation vs. posttotal steps estimation vs. post explanation generation) mixed ANOVA was performed on understanding ratings for a computer mouse. It revealed a significant interaction, $F(2, 228) = 3.56$, $p = .03$, $\eta_p^2 = .03$. Supporting the primary prediction, follow-up Bonferroni corrected criterion t tests indicated that unguided participants significantly reduced their understanding ratings from before ($M = 4.71$, $SD = 1.91$) to after ($M = 4.09$, $SD = 1.80$) being asked to estimate the total steps needed to explain how a computer mouse works, $t(57) = 3.00$, $p = .004$, $d = 0.39$, CI [0.12, 0.66]. However, REA participants appeared to have already performed a causal complexity assessment, as their ratings did change from before ($M = 3.58$, $SD = 2.20$) to after ($M = 3.73$, $SD = 2.08$) being asked to estimate the total number of steps, $p = .513$ (see Figure 7). There was no difference between the unguided reflection ($M = 3.81$, $SD = 1.89$) and REA condition ($M = 3.71$, $SD = 1.90$) for post explanation understanding ratings. Finally, there was not a difference in how many actual steps unguided ($M = 8.15$, $SD = 12.62$) and REA ($M = 7.24$, $SD = 7.57$) participants estimated are needed to explain how a computer mouse works, $p = .656$.

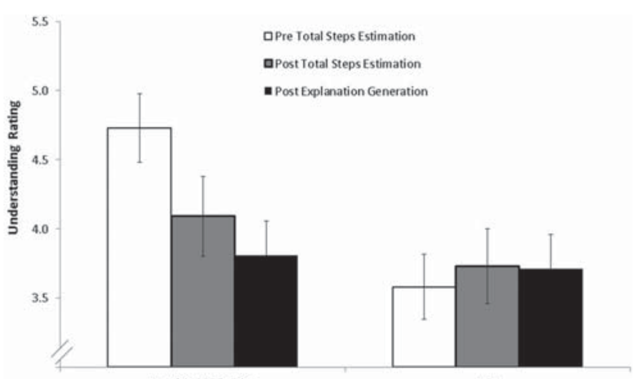

Figure 7. Experiment 6 understanding ratings showing that unguided reflection participants significantly reduced their understanding ratings (using the percentage of steps scale) after being directly asked to estimate the total number of steps required to explain how an object works. In contrast, it appeared that REA participants had already performed a causal complexity assessment, as they did not change their understanding ratings after being directly asked. Standard error bars of the mean are shown.

REA versus unguided reflection. To determine whether REA's effect is attributable to scale framing or interpretation, REA and unguided reflection's pretotal steps estimation ratings were compared, where all participants used the concrete, percentage of steps-based rating scale. REA ($M = 3.58$, $SD = 2.21$) significantly reduced overestimation compared to unguided reflection ($M = 4.73$, $SD = 1.90$), $t(116) = 3.04$, $p = .003$, $d = 0.56$, CI [0.37, 0.75] with an effect size comparable to previous studies—as indicated by overlapping CIs around Cohen's d when comparing the unguided reflection versus REA contrasts across experiments.

Discussion

Experiment 6 results provided more direct evidence for the causal complexity assessment account of REA's overestimation-reduction effect. REA appeared to induce participants to quickly (5 s or less) gauge the total number of causal steps needed to explain how an object works. This assessment revealed the normally hidden causal steps operating in an object and helped REA participants detect gaps in their causal knowledge. Without REA, unguided reflection participants failed to recognize the underlying complexity in an object and significantly overestimated their understanding until explicitly instructed to estimate the total number of steps required to explain how the parts of an object enable it to work. It is worth noting that there were no differences between the postexplanation ratings and initial ratings in the REA condition. This suggests REA can potentially completely eliminate the illusion of explanatory depth (Rozenblit & Keil, 2002).

The new scale participants used to rate their understanding in Experiment 6 explicitly defined understanding as the percentage of steps involved that enable an object to work a participant knows. Similar concrete scale framing has already been shown to substantially reduce overestimation (Alter et al., 2010). Given that REA's effect persisted using a concrete scale rules out scale interpretation or framing as alternative explanatory accounts.

However, given the within-subjects nature of the design, demand characteristics cannot be ruled out. Consequently, Experiment 7 is identical to Experiment 6 except it was a fully between-subjects design.

Experiment 7: Direct Test of the Causal Complexity Assessment Account (Between-Subjects)

Method

A sample of 222 participants (63% female, 37% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; $M = 37.11$, range 18–77).

Experiment 7 was identical to Experiment 6 except for the following changes. There was no within-subjects condition, and instead participants were randomly assigned to the unguided reflection ($n = 57$), REA-causal ($n = 55$), total steps ($n = 55$), or explanation generation ($n = 55$) condition in a fully between-subjects design. Participants in the unguided reflection and REA-causal conditions were given 5 s to reflect (computer-paced) before they were allotted 5 s to give their understanding rating. Participants in the total steps condition were asked to estimate, by giving an actual number, "*how many total steps it would take to explain how the parts enable the object to work*," before giving their understanding rating. The total steps estimation period was self-paced, but the understanding rating period was limited to 5 s. Participants in the explanation generation condition were asked to type out their full explanation following the same prompt from Experiment 1 before giving their understanding rating and both of these were self-paced. All participants rated their understanding of the following four objects in the following order on the same percentage of steps scale used in Experiment 6: printer, treadmill, power drill, and gas stove.

Results

Understanding ratings. A one-way ANOVA on understanding ratings revealed a main effect for condition, $F(3, 218) = 9.89$, $p < .001$, $\eta^2 = .12$. Follow-up Tukey t tests revealed that participants in the unguided reflection condition overestimated significantly more than all other conditions (see Table 1). No other pairwise comparisons were statistically significant (all $ps > .505$). Note the overlapping CIs around Cohen's d suggest all other conditions substantially reduced overestimation but with similar potency. To ensure the effects were not due to any specific object or subset of objects, a 2-way ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 2-way interaction (see Table S2 for understanding ratings for object means).

REA's efficiency. To compute the time saved using REA versus explanation generation, the total time it took participants to generate explanations for the four objects was recorded. For the high complexity objects ($M = 406$ s, $SD = 288$ s) used in Experiment 7, generating an explanation took more than 20 times longer to achieve comparable understanding ratings, compared to the 20 s of total REA reflection time.

Table 1
Descriptive and Inferential Statistics Comparing the Unguided, REA, Total Steps, and Explanation Generation Conditions on Understanding Ratings in Experiment 7

Condition	Unguided reflection	REA	Total steps	Explanation generation
M (SD)	4.79 (1.53)	3.83 (1.66)	3.41 (1.66)	3.36 (1.44)

Statistic	t-value	p-value	Cohen's d [95% CI]
Unguided vs.			
REA	3.22	.008	.60 [.40, .80]
Total steps	4.62	<.001	.86 [.64, 1.08]
Expl. generation	4.78	<.001	.96 [.73, 1.18]

Note. Unguided, $n = 57$; REA, $n = 55$; Total steps, $n = 55$; Explanation generation, $n = 55$.

Discussion

These results replicated Experiment 6's findings and provided additional support for the causal complexity assessment account of REA's effects. The fully between-subjects design ruled out demand characteristics as an alternative explanation of differences between any of the four conditions. Remarkably, REA reduced overestimation of understanding as effectively as generating full explanations, but did so 20 times faster.

Experiment 8: Causal Complexity Assessment or Covert Explanation

The purpose of Experiment 8 was to determine whether REA's overestimation-reduction effect can be accounted for, in part, by covert explanation. Given that REA involves reflecting on how well one could explain something, it is possible that participants begin the process of explanation covertly during the reflection period. This *covert explanation account* predicts that the more time participants get to reflect, the more covert explanation they complete. Consequently, they should have better insight into their actual level of knowledge, leading to greater reductions in overestimation.

However, data from the previous experiments are inconsistent with the covert explanation account. First, participants in Experiments 1 through 5 were given 15 s to reflect, whereas participants in Experiments 6 and 7 were given 5 s to reflect and yet REA's power to reduce overestimation was comparable across studies. Second, REA's effect varied as function of object complexity in Experiments 4 and 5, even though the time to covertly generate an explanation was equivalent across low and high complexity objects. Third, in Experiments 6 and 7, 5 s was not likely long enough to covertly generate many causal steps, so it seems implausible that covert explanation can fully account for REA's power to reduce overestimation.

The prior experiments provide only indirect evidence against the covert explanation account. To directly test the covert explanation account of REA's effect, participants were randomly assigned to reflect for five or 20 s prior to giving their understanding rating. If REA's effects are attributable to participants covertly generating partial explanations, then understanding ratings should be significantly lower when participants have more time to covertly generate additional causal steps. The causal complexity assessment account predicts no change as function of reflection time because REA participants quickly assess the number of causal steps involved early in the reflection period. Additionally, for the purpose of using REA in applied settings, there is value in determining how much reflection time is required to accrue benefits from REA.

Method

A sample of 228 participants (68% female, 32% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; $M = 37.24$, range 18–86).

Before engaging in reflection on how well participants understood how each object works, participants were randomly assigned to one of

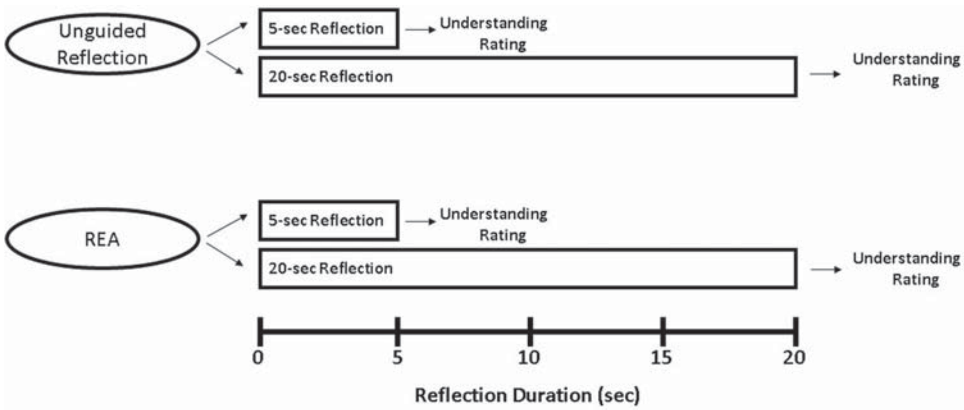

Figure 8. Design of Experiment 8.

four reflection conditions in a fully between-subjects design: (a) unguided reflection for 5 s (n = 56), (b) unguided reflection for 20 s (n = 55), (c) REA-causal reflection for 5 s (n = 59), or (d) REA-causal reflection for 20 s (n = 59). The computer software controlled the timing of the reflection periods, so that participants could not move on before their respective reflection periods passed. Immediately after the reflection period, the next screen automatically appeared and participants were given 5 s to give their understanding rating on the percentage of steps scale from Experiments 6 and 7 before the next instruction screen automatically appeared. All participants rated their understanding of the following four objects in the following order: printer, treadmill, power drill, and gas stove. See Figure 8 for a depiction of the design.

Results

A 2 (unguided reflection vs. REA-causal reflection) × 2 (5-s vs. 20-s reflection duration) between-subjects ANOVA was performed on understanding ratings averaged across the four objects. It revealed a main effect, where REA (M = 3.53, SD = 1.49) understanding ratings were significantly lower than unguided reflection ratings (M = 4.91, SD = 1.43), $F(1, 225) = 51.89$, $p <$ .001, $\eta_p^2 = .19$. In addition, there was a main effect of reflection duration, where understanding ratings were significantly lower in the 20-s condition (M = 4.00, SD = 1.59) than in the 5-s condition (M = 4.39, SD = 1.63), $F(1, 225) = 4.30$, $p = .039$, $\eta_p^2 = .019$. Although the interaction between reflection type and reflection duration was not significant (p = .173), it is worth noting that post hoc t tests revealed that participants in the unguided reflection condition significantly reduced their understanding ratings in the 20-s condition (M = 4.58, SD = 1.52), compared the 5-s reflection condition (M = 5.24, SD = 1.28), $t(109) = 2.48$, $p = .015$, d = 0.47, CI [0.27, 0.66]. In contrast, there was no difference for REA participants between understanding ratings in the 5-s condition (M = 3.59, SD = 1.52) and the 20-s reflection condition (M = 3.46, SD = 1.47), $p = .623$ (see Figure 9). To ensure the effects were not due to any specific object or subset of objects, a 3-way ANOVA was performed with object as an additional within-subjects factor and results revealed a nonsignificant 3-way interaction (see Table S2 for understanding ratings for object means).

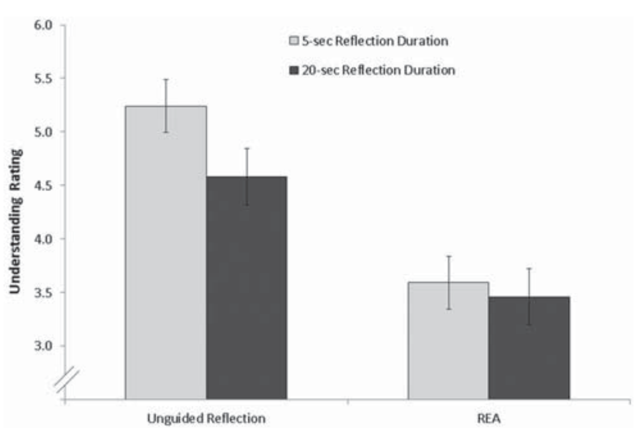

Figure 9. Experiment 8 understanding ratings showing that unguided reflection participants significantly reduced overestimation when given more time to reflect, whereas REA participants showed equivalent reductions in overestimation across reflection durations. Standard error bars of the mean are shown.

Discussion

Experiment 8's results favored the causal complexity assessment account over the covert explanation account of REA. According to the covert explanation account, REA should reduce overestimation more when participants have longer to covertly generate additional explanatory steps. In contrast, REA's power to reduce overestimation did not vary as a function of reflection time. Supporting the causal complexity assessment account, the results indicated REA participants estimate the total causal steps required to explain how the parts of an object enable it to work early in the reflection period, in 5 s or less.

Experiment 9: REA and the Socio-Political Domain

The purpose of Experiment 9 was to determine the generalizability of REA's effects. A recent study showed that reducing overestimation of understanding via explanation generation concomitantly reduced extremist political attitudes (Fernbach et al., 2013). The results suggested that generating an explanation of how

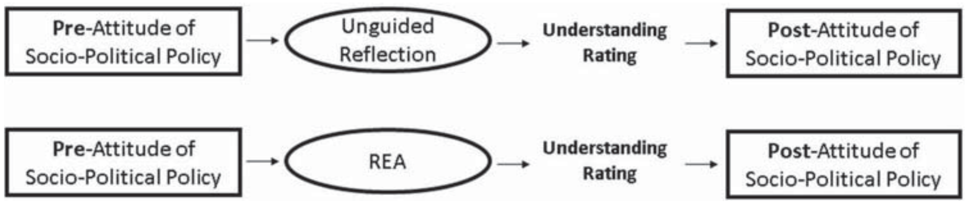

Figure 10. Design of Experiment 9.

a policy actually works highlighted participants' knowledge gaps and consequently moderated political attitudes. It is predicted REA will also reduce overestimation of understanding and moderate extremist political views by highlighting gaps in causal knowledge about complex policies.

Method

A sample of 224 participants (60% female, 40% male) was recruited from Amazon's Mechanical Turk, where each participant received a $0.50 payment for participation (age; $M = 34.94$, range 18–70).

Participants first serially rated their prereflection attitudes on implementing merit-based teacher pay and a cap-and-trade system for carbon emissions on a 1 (*strongly against*) to 7 (*strongly in favor*) scale (Fernbach et al., 2013) in a self-paced manner. Before giving their understanding rating on 1 (*vague understanding*) to 7 (*thorough understanding*) scale, participants were randomly assigned to reflect using the unguided reflection prompt ($n = 112$) or REA-causal prompt ($n = 112$). Unguided participants were told to, "*carefully reflect on how detailed your understanding of the issue is*," whereas REA participants were told to, "*carefully reflect on how well you could explain to an expert, in a step-by-step, causally-connected manner the details of the issue*." They serially reflected for 15-s (computer-paced) and then gave their understanding rating for each of the two issues in a self-paced manner. Finally, they rated their postreflection attitudes on merit-based teacher pay and a cap-and-trade system for carbon emissions in a self-paced manner. See Figure 10 for a depiction of the design.

Results

Understanding ratings. An independent *t* test on the average of the merit-based teacher pay and cap-and-trade understanding ratings indicated REA ($M = 3.58$, $SD = 1.44$) significantly reduced overestimation compared to unguided reflection ($M = 4.30$, $SD = 1.31$), $t(222) = 3.96$, $p < .001$, $d = 0.52$, CI [0.37, 0.67].

Extremity of political attitudes. Attitude ratings were transformed to reflect attitude extremity by subtracting 4 from the raw scores and taking the absolute value (Fernbach et al., 2013). Then, an attitude change score was created by subtracting prereflection extremity scores from postreflection extremity scores. An independent *t* test indicated REA ($M = .69$, $SD = .68$) reduced attitude extremity significantly more than unguided reflection ($M = .48$, $SD = .60$), $t(222) = 2.40$, $p = .017$, $d = 0.33$, CI [0.19, 0.47].

To test whether the reduction in overestimation mediated the drop in attitude extremity, a mediation analysis using Hayes' (2012) PROCESS Model was performed and indicated significant mediation with a CI not containing 0, CI [.001, .032], with effect size, $\kappa^2 = 0.03$ (Preacher & Kelley, 2011).

Discussion

REA successfully reduced overestimation of understanding for complex policies and simultaneously moderated sociopolitical attitudes. These results highlight the generalizability of REA's power to detect causal knowledge gaps—with applied implications for attitude change and reducing political extremism.

General Discussion

Reflecting on explanatory ability (REA) robustly exposed gaps in participants' causal knowledge about complex objects and sociopolitical policies. REA induced a strong mechanistic explanatory stance, that is, step-by-step, causally connected processing that was critical to reducing overestimation. Remarkably, REA's power to reduce overestimation was comparable to that of explanation generation (except for low complexity objects, discussed below). The mechanism underlying REA's reduction of overestimation is that it induces individuals to quickly assess the total causal steps involved in explaining how something works, that is, perform a causal complexity assessment. Alternative explanations for REA's effects were ruled out, including differences in rating scale interpretation or framing, general conservatism, sensitivity to fluency, and covert explanation. Finally, REA's overestimation-reduction effect generalized to the sociopolitical domain and REA even reduced extremist sociopolitical attitudes.

Explanation Theory

The results of the current studies suggest integrating explanation theory and metacognition could stimulate new paths of research. Tacit in most explanation theory is the assumption that generating explanations is required to guide judgment and improve learning (Lombrozo, 2006, 2012). In contrast, the current studies suggested metacognition offers an alternative fruitful approach. The roles of mental simulation and narrative structure have recently been recognized as critical in explanation theory and causal reasoning—perhaps a metacognitive focus will augment this new path (Sloman & Lagnado, 2015).

For example, REA's metacognitive mechanism for reducing overestimation revealed that the illusion of explanatory depth may not be as deep as previously thought (Fernbach et al., 2013; Rozenblit & Keil, 2002). Previous work suggested full verbal explanations or diagnostic questions are required in order for individuals to accurately probe their understanding, particularly for material high in causal complexity (Rozenblit & Keil, 2002). In contrast, the current experiments suggest the illusion can be substantially reduced with a relatively brief but guided reflection period (i.e., REA instructions)—especially for complex material. In addition, the mechanism through which explanation generation and REA reduce overestimation appears to be quite different. Whereas explanation generation highlights gaps in participants' causal knowledge by allowing them to produce what they know, REA requires no production and instead relies on a quick assessment of causal complexity (see Experiments 4–7). Further evidence for differing mechanisms comes from Experiments 4 and 5, where REA participants substantially reduced their overestimation only for high complexity objects, whereas participants who generated full explanations reduced their overestimation equally for low and high complexity objects. This suggests that REA's causal complexity assessment does not help as much for low complexity objects because REA participants perceive relatively few causal steps involved and therefore continue to overestimate their understanding. In addition, although REA has been highlighted here as an efficient and effective metacognitive tool, explanation generation remains invaluable. Explanation generation outperformed REA at reducing overestimation for low complexity objects. Also, explanation generation enhances learning in multitudinous ways researchers are still uncovering (Pyc, Agarwal, & Roediger, 2014).

These results are consistent with Alter et al.'s (2010) study on the illusion of explanatory depth. Alter et al. (2010) showed that the illusion can be reduced by inducing participants to think more concretely about how well they understand how an object works. REA also induced participants to think differently about how to judge their understanding of how an object works. However, it is worth reiterating that REA's effects were not a result of concrete thinking but instead were attributable to mechanistic, causally connected reflection and a causal complexity assessment.

The current studies' results also have implications for literature on explanatory stances (Keil, 2006; Lombrozo, 2012). There is general agreement that when asked to evaluate understanding, children, adults, and even professional scientists are biased to take a teleolgical explanatory stance, that is, to think about an object's purpose or teleology at the expense of a mechanistic explanatory stance (Kelemen, 1999; Kelemen & Rosset, 2009; Kelemen et al., 2013). In light of individuals' persistent teleological bias, the current findings raise an interesting possibility regarding the illusion of explanatory depth. The illusion was reduced, but not eliminated in Rozenblit and Keil's (2002) study when participants were explicitly forewarned they would need to explain how an objects works. It seems likely, given the results of the current studies, that an explicit warning was not enough to fully mitigate participants' teleologically biased explanatory stance. Instead, REA guidance is needed to explicitly induce a fully mechanistic explanatory stance and foster step-by-step, causally connected processing.

The persistence of this teleological bias across development, education, and even scientific expertise suggests it is an entrenched cognitive bias and difficult to combat (Kelemen et al., 2013). The current results challenge the conclusion it is difficult to combat, and suggest that a metacognitive approach to explanation can efficiently and effectively reduce overestimation of understanding.

Implications for Self-Regulated Learning and Self-Insight

REA could contribute to diverse literatures, like self-directed learning. Most literature suggests *unguided* reflection during learning (e.g., using judgments of learning) produces poor knowledge calibration and learning outcomes (Bjork et al., 2013). However, using *guided* reflection to direct learners to more diagnostic cues of understanding appears promising (Koriat & Bjork, 2006; McCabe & Soderstrom, 2011). Perhaps REA could redirect attention toward such diagnostic cues. According to cue utilization theory, learners are biased to rely on intrinsic cues, like familiarity, to guide their judgment of their current knowledge level (Koriat, 1997). Unfortunately, these cues lead learners astray (Bjork et al., 2013). Although speculative, REA may redirect attention away from intrinsic cues toward more diagnostic cues, like their own knowledge relative to the overall complexity of the material.

Relatedly, REA could help us gain insight into our abilities, from academic to vocational. A recent meta-analysis suggests that our biggest self-insight blind-spot is in high complexity domains (Moore & Healy, 2008; Zell & Krizan, 2014)—the very domain in which REA was most effective. REA's ability to improve self-insight even has implications for reducing extremist political attitudes (Experiment 9).

Metacognitive Accuracy of Explanations: An Infinite Regress Problem

An important future direction for studies of explanation will be to determine how absolute metacognitive accuracy (Nelson & Dunlosky, 1991) can be assessed for participant-generated explanations. The challenge is that explanations suffer from infinite regress, that is, they have indefinite end states—there is always another step in explaining how something works (e.g., mechanically, molecularly) (Keil, 2006; Lombrozo, 2012). Even participants who generated a full explanation of how something works estimated they knew between 30% and 60% of the total steps involved in how an object works (like a printer, Experiment 7). It is highly unlikely the general population knows roughly 50% of the total steps involved in how the parts work together to enable the printer to function. Consequently, any estimate of understanding above the lowest number on the percentage of steps scale (0%–15%) represents overestimation of understanding. Moreover, any reduction in understanding ratings represents an improvement in actual knowledge calibration and metacognitive accuracy. So, one could argue REA improved actual knowledge calibration in the current experiments.

However, to directly assess whether a participant's self-estimated knowledge is calibrated to their actual knowledge, there must be an objective "perfect" or complete explanation with which to compare. One solution is to have participants judge the accuracy of explanatory statements (Kelemen et al., 2013) or take a knowledge test, but this circumvents participant-generated explanations all together. A promising direction may be to have participants judge their knowledge compared with various groups—for example, peers, the general population, or an expert. Then, an assessment of a general population's average explanation quality, for example, could be quantified and used as a benchmark.

The Nature of Reflecting on Explanatory Ability

Although the results of the current experiments showed that REA induced a quick causal complexity assessment, future work could further explore the nature of this assessment. For example, it is unlikely that individuals engaged in REA generated a concrete number of the total steps involved in explaining how an object works. It is more likely they render a gross estimate of causal complexity; for example, they may use an internal scale from very simple to extremely intricate, requiring many steps to explain. Consequently, if they judge the object to be extremely intricate, they can easily deduce they will be able to explain a small percentage of the total steps involved in making an object work. Continued investigation into how REA allows individuals to quickly and accurately probe their causal knowledge has significant theoretical and applied implications.

Conclusion

The current findings highlight a fundamental aspect of human cognition. The results converged with broad trends in prior literature by confirming that unguided metacognition typically leads us astray when trying to gain insight into our knowledge and abilities (Bjork et al., 2013; Moore & Healy, 2008; Zell & Krizan, 2014). However, these studies offered rare insight and demonstrated how guided reflection that capitalizes on a mechanistic explanatory stance provides an accurate window into our knowledge. REA is a rare metacognitive tool in the arsenal to combat our proclivity to overestimate understanding. Perhaps REA can help us gain the wisdom to which Socrates was referring.

References

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17,* 286–303. http://dx.doi.org/10.1016/j.learninstruc.2007.02.004

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

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26,* 1295–1303. http://dx.doi.org/10.1177/0956797615588195

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823

Cumming, G. (2012). *Understanding the new statistics: Effects sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge/Taylor & Francis Group.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

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

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144,* 674–687. http://dx.doi.org/10.1037/xge0000070

Fisher, M., & Keil, F. C. (2014). The illusion of argument justification. *Journal of Experimental Psychology: General, 143,* 425–433. http://dx.doi.org/10.1037/a0032234

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics, 110,* 585–603. http://dx.doi.org/10.2307/2946693

Fox, C. R., & Weber, M. (2002). Ambiguity aversion, comparative ignorance, and decision context. *Organizational Behavior and Human Decision Processes, 88,* 476–498. http://dx.doi.org/10.1006/obhd.2001.2990

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18. http://dx.doi.org/10.1037/a0024338

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology, 40,* 379–398. http://dx.doi.org/10.1146/annurev-anthro-081309-145822

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling* [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17,* 471–479. http://dx.doi.org/10.1080/09658210802647009

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227–254. http://dx.doi.org/10.1146/annurev.psych.57.102904.190100

Keil, F. C. (2012). Running on empty? How folk science gets by with less. *Current Directions in Psychological Science, 21,* 329–334. http://dx.doi.org/10.1177/0963721412453721

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about object. *Trends in the Cognitive Sciences, 3,* 461–468. http://dx.doi.org/10.1016/S1364-6613(99)01402-3

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition, 111,* 138–143. http://dx.doi.org/10.1016/j.cognition.2009.01.001

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General, 142,* 1074–1083. http://dx.doi.org/10.1037/a0030399

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1133–1145.

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition, 34,* 1667–1675. http://dx.doi.org/10.3758/BF03195929

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10,* 464–470. http://dx.doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 260–276). New York, NY: Oxford University Press.

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKS). *Journal of Experimental Psychology: General, 140,* 605–621. http://dx.doi.org/10.1037/a0024014

McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). New York, NY: Routledge/Taylor & Francis Group.

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "Delayed-JOL Effect." *Psychological Science, 2,* 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–115. http://dx.doi.org/10.1037/a0022658

Pyc, M. A., Agarwal, P. K., & Roediger, H. L., III (2014). Test-enhanced learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (pp. 78–90). Washington, DC: Society for the Teaching of Psychology.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299,* 172–179. http://dx.doi.org/10.1016/j.jtbi.2011.03.004

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

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology, 66,* 223–247. http://dx.doi.org/10.1146/annurev-psych-010814-015135

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition, 133,* 343–357.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science, 9,* 111–125. http://dx.doi.org/10.1177/1745691613518075

Received August 7, 2015
Revision received February 9, 2016
Accepted February 11, 2016 ■