Title
The Disparity Between What We Know and How We Communicate

Author
Roeder, Scott

Publication Date
2016

Peer reviewed|Thesis/dissertation

The Disparity Between What We Know and How We Communicate

By

Scott Roeder

A dissertation submitted in partial satisfaction of the

Requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Clayton R. Critcher, Chair
Professor Leif D. Nelson
Professor Don A. Moore
Professor Iris Mauss

Summer 2016

Abstract

The Disparity Between What We Know and How We Communicate

by

Scott Roeder

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Clayton R. Critcher, Chair

Research has demonstrated that people systematically overrate their knowledge, intelligence, and skills in various domains. Confronting people with evidence of their miscalibration, however, causes them to reassess these claims. For example, simply asking people to explain how a sewing machine works leads them to subsequently report understanding it less—a bias called the *illusion of explanatory depth* (Rozenblit & Keil, 2002). While previous work argues that this process is domain-bound, we demonstrate in several experiments that the bias to inflate subjective knowledge is attenuated not only by explanations of the focal item itself but also by explanations of other, entirely different things, implying the existence of a more parsimonious, *domain-agnostic* process for this bias. We then show that the illusion of explanatory depth holds for difficult, but not easy, explanations.

Acknowledgements

I have always struggled to find motivation to see things though to the end. Prior to meeting my wife, despite my many years of schooling, and sports, I was never able to find even an ounce of it. I was an unruly and obstinate child, and my parents were faultless in this. They tried everything. They set goals for me, pushed me, complimented me when deserved, and even bribed me with everything within their means. When I failed to follow through, they justifiably punished me. Nothing worked. If it were not for my wife and best friend Janaia and my two small children, Alexis and Callahan, who have become my motivation, I would never have started down this path. This dissertation and the work that I have done over these past six years has been for and because of them. Without them I surely would not have been able to abide the stress, long days, sleepless nights, depression and feelings of inadequacy. It could be argued that Janaia, more than me, deserves credit for this accomplishment. Her boundless optimism keeps me from giving up, and I will forever be indebted to her for it.

My parents, John and Kim, are as close to perfect examples of parents as one could imagine, and I credit them for teaching me how to think critically. Nightly dinner conversations at the Roeders' were often controversial, deep and/or political. One could not have an opinion without backing it up or defending it. I came to enjoy the (friendly) arguing and I have managed to maintain the thick skin I developed during these years. It has come in handy many times in class discussions, the research process and especially during weekly meetings with my advisors.

To my friends and family reading this: I worked closely with two incredible academic mentors in Leif Nelson and Clayton Critcher. Clayton joined the faculty at Haas the same year I joined the Ph.D. program (2010), and is so hardworking and productive that he could have had tenure within 3 years. He is brilliant and cares deeply about doing good research. His intensely theoretical and methodical approach to research is lauded among the students and faculty who know him. I was extremely fortunate to have had him on my committee and to work with him over these past years at Berkeley.

Leif, like Clayton, is prodigious. When he is present in the room he makes everyone around him smarter. He is eloquent, friendly, and a very likable person—something that I have found to be rare among academics. When I came to him with ideas he humored me and talked me through them until we found something that we thought might work (or that would be fun to investigate). He encouraged me at every step to think deeper and work harder, but also to enjoy it. His research strategies have rubbed off on me more than anyone I've worked with and I am better for it.

Berkeley has changed me in many ways. I have come to love the school, the city, and the people I met during my time here, all elements that deserve some credit for any success I have had or will have as an academic.

Introduction

Though I had always been interested in the egocentric biases literature, it wasn't until I attended at small conference at Yale that I felt engaged enough to attempt to contribute to it. During this specific conference, Steven Sloman, a cognitive psychologist at Brown University, gave a talk on the "Illusion of Explanatory Depth" (Rozenblit & Keil, 2002) as it relates to the CRT, or Cognitive Reflection Task (Frederick, 2005). He and Phil Fernbach (University of Colorado) found that people have different thresholds for understanding how products work and, subsequently, different thresholds for payment given their subjective level of understanding. Some were satisfied with superficial explanations for how the product worked and paid a lot, while others were more willing to pay top dollar the more the product was explained. These results were surprising to me not so much because of the downstream consequences—That is, differences in willingness to pay—But instead because of the sensitivity and apparent robustness of the instrument. Researchers who study the IOED regularly find surprisingly extreme differences between Time 1 and Time 2 estimates on this task. In Sloman and Fernbach's experiments, for example, people would drastically attenuate how much they claimed to understand a product after simply having had to provide an explanation for how it works.

I returned home from this conference with an embiggened sense of purpose. First, to find out whether or not the effect is truly replicable, and second, to discover its boundaries. After meeting with Leif Nelson, one of my academic mentors, we decided to attempt to answer these questions. What follows in this dissertation is the result of this inquiry. First I will present results from several studies showing that the Illusion of Explanatory Depth is replicable, but also far broader than the narrow constraints from which it was formerly bound in the literature. I will then suggest some future avenues for my own (and others') research based on some exploratory results I found on this topic during my Ph.D.

Some of this work is found in an ongoing working paper that has been submitted for review at *Psychological Science*. This research has implications for the egocentric biases literature as well as how people may or may not re-evaluate subjective knowledge in light of conflicting or disconfirming information. We find that difficult explanations reduce the sense of subjective knowledge both within and *across* domains. That is, contrary to what the former *domain-diagnostic* account would predict, people do not have infinite silos of mechanistic knowledge, updated one-by-one as their level of understanding is challenged. Instead, they appear to hold a somewhat broader view for how they understand things (i.e. their level of overall mastery). When that broad view is challenged, they update not their level of mastery for a given subject, but instead all subjects. We find that this effect is moderated by explanation difficulty: The more difficult the explanation, the more pronounced the illusion of depth. In concept, previous accounts have treated subjective knowledge like a sponge retaining water; if you press on one spot, some water is lost, but the bulk is retained. Alternatively, we suggest that the sense of mastery is more like a balloon; a single prick and all is lost.

The Disparity Between What We Know and How We Communicate

A person needs to know her limitations. Navigating life decisions requires knowledge and understanding of course, but also requires a self-assessment of personal knowledge and understanding. In order to repair a parachute, a person must necessarily understand how a parachute works and how to fix it, but in order to then confidently jump out of a plane, that same person must also correctly recognize their own mastery (or lack thereof). That extra level of understanding – metacognitive skill – has been widely

considered within cognitive and social psychology, with a repeated observation that people are not always so great at assessing their personal knowledge and mastery. Already there has been wide consideration of the relationship between actual relative skill and self-assessed relative skill (e.g., Kruger & Dunning, 1999; Burson, Larrick & Klayman, 2006; Moore & Healy, 2008). People who are very funny seldom realize how much life they bring to a party, and people who are very unfunny do not realize how much they drain from it. Whether explained through personality, an absence of metacognitive ability, or simply regressive prediction, there is ample evidence that people are imperfect at knowing how their own skills stack up in the world.

But even on non-relative estimation (e.g., the precise number of correct folds in a parachute), accurate self-assessments appear mixed with a critical dependence on lay theories of knowledge. Because people can imagine the touch and feel of a parachute, and can observe the obvious deployment and success of parachutes in general, their lay theories seem good enough to say, "I understand how parachutes work."
That is, right up until they are asked to actually explain how a parachute works. Our subjective understanding for how things work is driven by incomplete folk theories; theories which, though incomplete, have a veneer of completeness. In this way, people appear to have an illusion of explanatory depth (or IOED, Rozenblit & Keil, 2002). In a prototypical example, people will cheerfully report a good understanding for how a sewing machine works. But, when subsequently asked to actually explain how a needle can pass through one side of the cloth and yet create a thread looped through both sides, people witness their own impoverished explanation and meaningfully downwardly revise their self-assessment. An external evaluation from an expert seamstress is not needed, the illusion is both built and shattered by the novice himself.

This now well-documented error gives us insight into the very nature of self-knowledge. The original authors suggest that this bias arises due to what can be characterized as a feeling of knowing or mastery. Several factors are suggested to contribute to this feeling. First, people seem capable of holding both an intuitive theory about the way something works even while simultaneously holding only a skeletal representation of its actual mechanistic complexity. Second, because people overestimate the accuracy of their recall of scenes and objects (Levin, Momen, Drivdahl and Simons, 2000), their intuitive theories seem to fully cover their truly skeletal knowledge. Finally, people might more generally confuse higher and lower levels of analysis—that is, because someone thinks they understand a higher-level function of something (e.g. that a toaster gets hot), they might then falsely intuit that they have a similar understanding of the lower level functions (e.g. how a toaster's filaments convert electricity into heat). Overall these cognitive errors give people the mistaken impression (or feeling) that they understand something more than they really do. In addition, because people rarely need to explain the intricacies of certain complex phenomena, they are rarely faced with the incompleteness of their mental blueprints.

The evidence for the IOED is both broad and deep. The effect, originally demonstrated with machines and electronics, extends to the understanding of politics and policies (Alter, Oppenheimer & Zemla, 2010). That work, in particular, highlighted the role of construal level (Trope & Liberman, 2003) in the assessment of self-knowledge. The same insight can influence the policies the people endorse and the political positions they hold (Fernbach, Rogers, Fox, & Sloman, 2013). When asked to explain a particular policy, such as imposing unilateral sanctions on Iran, people recognize their ignorance and scale back the extremity of their position. Fernbach and colleagues were further able to broadly segment people into two groups—those who are satisfied with surface level explanations and those who are dissatisfied (Fernbach, Sloman, St. Louis, & Shube, 2013). Related work shows that a similar effect exists in the domain of argument

justification (Fisher & Keil, 2014). The authors find that even for less mechanistic topics, such as abortion, people over-predict the quality of and their ability to justify their positions before outlining their argument than after. As shown by Fernbach et al. (2013), these individual differences are likely correlated with a general propensity to give intuitive, rather than reasoned, answers on the cognitive reflection task (CRT, Frederick, 2005) while dissatisfaction with surface-level explanations is unrelated to the desire to think more generally, as assessed by the need for cognition scale (NFC, Haugtvedt, Petty, and Cacioppo, 1992). Across many domains then, people seem to express initial knowledge that is both exaggerated and vulnerable, as merely trying to explain would seem to bring it back down to size.

It appears that, when asked to explain something, the feeling of knowledge is replaced with a recognition of ignorance. The central conceptualization of this phenomenon has focused on the details of the object itself. A skeletal knowledge of a sewing machine can feel complete only until it is articulated. At that point, even a novice can distinguish skeletal knowledge from complete knowledge. By this account, the insight has to date been domain-diagnostic: the act of explaining a sewing machine directly informs someone's subjective assessment of sewing-machine mastery. As we summarize below, the domain-diagnostic account offers a parsimonious explanation for existing research into the IOED. It is not, however, the only possible explanation.
An alternative account need not be focused on the object itself, but rather on knowledge and explanation more generally. We refer to this as the domain-agnostic account. This account suggests that the act of explaining a sewing machine, rather than informing the skeletal nature of knowledge for that particular device, instead informs the skeletal nature of explanations more generally. These two accounts lead to starkly different predictions. Whereas the domain diagnostic account suggests the inflated sense of sewing machine knowledge will (only) be undone by an explanation of sewing machines, the domain agnostic account suggests that explaining something else entirely (e.g., a carburetor or gumball machine, perhaps) will operate just as pointedly on sewing-machine knowledge. As we will demonstrate, explanations of topics far removed from the specific domain can still have a precise and significant influence within that domain. Moreover, we establish that this domain-agnostic account offers parsimony by both explaining existing findings and testing new predictions. The alternative account paints a new portrait of human self-knowledge. People certainly have folk theories of how things work, but notably, disrupting one theory would appear to call all of them into question.

Experiment 1

The present experiment was designed test whether the proposed *domain-agnostic* account also exhibits an illusion of explanatory depth. To explore this, we had participants take the IOED task with slight variation: They either explained the same devices they rated at Time 1 (Condition 1) or explained other, unrated (i.e. nonfocal) devices (Condition 2).

Method

University of California, Berkeley undergraduates (N = 175) participated in this experiment (online) for course credit and were told that they would "rate how well [they] feel [they] understand how different things work." Sample size was determined by adherence to Simmons, Nelson and Simonsohn's (2011) suggestion of at least 20 participants per cell, and was run for the final two weeks of the semester until no more participants were available. No data analyses were conducted until the full set was collected. We decided a priori to eliminate participants who either did not provide explanations, provided nonsense explanations (i.e. gibberish), had missing data, or were duplicate responses (based on IP address).

As dictated by the IOED task, participants were first trained on the 7-point scale. Instructions for this training were taken verbatim from the original Rozenblit and Keil (2002) materials and, using a crossbow as an example, include how they should interpret each scale point. They read that "a score of 7 would feature all the elements of the description (e.g., what the parts are, their function, how they interact), a 4 would require knowledge of some of the basics but not all the intricacies of the description, and a score of 1 would reflect an absence of knowledge about how the object worked." After reading this they were randomized to one of two conditions. Instructions were identical between conditions and followed the traditional IOED procedure: At Time 1 (T1) participants rate their understanding of three devices (sewing machine, bicycle lock, zipper) on the 7-point scale before being asked to imagine that they had "just met a person who did not understand how these three items work" and to "write as complete an explanation of how each item works as [they] can manage." After generating these explanations, participants were asked to rate again how well they feel they understand the devices (T2). While Condition 1 mirrored the traditional IOED procedure, Condition 2 instead required that participants generate explanations for three entirely different devices (i.e. piano keys, a transistor, and a manual clutch) instead of a sewing machine, bicycle lock and zipper (see Figure 1).

Results

Twenty-four participants were excluded from analysis based on the criteria outlined in the Method section leaving 151 participants for analyses. Twenty-one participants were excluded for not providing any explanation at all, and three were excluded for typing gibberish. Analyses including these participants did not change the results.

The standard IOED task asks participants to estimate their level of understanding for something (Time 1) on a 7-point scale before having to explain how it works. They are then asked again to rate their level of understanding on the same scale (Time 2). To analyze the data, a comparison metric is typically computed by subtracting Time 2 scores from Time 1. An average IOED score greater than '0' would therefore imply a decrease in reported understanding between time points. Overall results from this task are robust—participants routinely identify significantly higher numbers on the 7-point scale pre-explanation than post-explanation. The *domain-diagnostic* account predicts that self-reported level of understanding for a sewing machine should decrease only after having to explain how a sewing machine works, and not after having to explain something else. We predicted, in line with the *domain agnostic* account, that we would see understanding scores greater at Time 1 than Time 2 for both conditions.

Participants did not differ in terms of Time 1 knowledge across conditions, $F(1, 149) = .529$, $p > .250$. Analyses for this experiment were conducted with T-tests. We first examined responses for the same domain condition where participants had to explain the same devices they rated at Time 1. In line with previous research supporting the *domain diagnostic* account, all three devices individually showed strong IOED effects: Levels of understanding for a sewing machine ($t(78) = 3.83$, $p < .001$), bike lock ($t(78) = 4.18$, $p < .001$) and zipper ($t(78) = 8.10$, $p < .001$) all decreased significantly from Time 1 to Time 2. Recall that in different domain condition, as shown in Figure 1, participants were instead asked to explain other, un-rated devices from those seen at Time 1. In support of the *domain agnostic* account, levels of understanding for a sewing machine ($t(71) = 3.38$, $p = .001$), bike lock ($t(71) = 4.62$, $p < .001$) and zipper ($t(71) = 5.67$, $p < .001$) all similarly decreased after explaining how piano keys, a transistor, and manual clutch work. Combining across items, we correctly predicted that there would be a significant decrease in understanding ratings from T1 to T2 in both the *domain diagnostic* [$t(78) =$

7.57, $p < .001$] and *domain agnostic* conditions [$t(71) = 5.89$, $p < .001$] (See Figure 2). Somewhat notably however, the difference between T1 and T2 scores is smaller in the *domain agnostic* condition, $F(1, 149) = 3.98$, $p = .048$. We hesitate to speculate on the reason for this anomaly since the finding is both statistically precarious and does not replicate reliably in future studies. In fact, as we will show, the direction of this disparity occasionally reverses.

Taken together the results were largely consistent with the *domain agnostic* account. Nevertheless, before generalizing to the IOED or self-knowledge more broadly, it is worthwhile to generalize to the literature as much as possible. To that end, we now turn to a consideration of political policies, a domain previously investigated heavily (e.g., Alter et al., 2010, Fernbach et al., 2013).

Experiment 2

Experiment 2 tests whether results from the previous experiment could be achieved in a different domain—understanding of political policies. We followed the procedure outlined by Fernbach et al. (2013) with slight modification to suit the current investigation.

Method

We sought 600 Mechanical Turk participants (100 per explanation), and the experiment ran for one week. 602 participants took part in total. Subjects first saw a set of IOED scale training instructions similar to Experiment 1. Wording for these instructions was taken verbatim from Fernbach et al. (2003). They describe how the participants' goal is to rate "how well [they] feel [they] understand different political issues," and that a 1 on the scale should denote a "vague understanding," while a 7 indicates "thorough understanding." Participants then read an example describing the level of understanding that someone with a Level 1, Level 4, or Level 7 would have for immigration policy. As is then dictated by standard IOED procedure, participants used the scale to rate six phenomena also taken verbatim from Fernbach et al's original materials. Subjects rated the extent to which they understand the impact of imposing unilateral sanctions on Iran for their nuclear program, raising the retirement age for social security, transitioning to a single-payer health care system, establishing a cap-and-trade system for carbon emissions, the institution of a national flat tax, and implementing merit-based pay for teachers. Fernbach et al. do not report whether or not these six items were counter-balanced in their paper, but since the IOED is a repeated measures design, we added this to control for order effects.

After rating the six items, rather than requiring participants to explain all of the items, we randomly inserted one of the six and asked them to explain it. Doing this allowed us to systematically compare "matched" items with "unmatched" items. That is, we compared IOED scores for each participant's matched items and unmatched items, controlling for within-participant variance in responding. Results from Experiment 1 would predict that, as is the case for devices, matched and unmatched policies are both subject to the illusion of explanatory depth.

Results

89 participants met our exclusion criteria and were removed from the final data leaving 513 for analyses. 77 were eliminated for not providing any explanations, 10 for having a duplicate IP address, and two for providing nonsense explanations. It appears that many participants started but did not complete the survey, which accounts for the high number of non-explanations.

This experiment required a different analysis plan since, unlike Experiment 1, each participant rated all six items but only explained one of them. We were interested in whether or not systematic differences would emerge based on which item they explained. As a reminder, the *domain agnostic* account predicts that a decrease in levels of understanding T1-T2 should emerge for all items, regardless of policy explained. The *domain diagnostic* account on the other hand would predict that there should only be a decrease in understanding T1-T2 for matched-policies; That is, participants' ratings for how well they understand imposing unilateral sanctions on Iran should only decrease when they have to explain how that particular policy might work, but not when they have to instead explain how to institute merit-based pay for teachers, for example. To this end, we employed a hierarchical linear model with random intercepts. This model allows one to control for person specific variance in responding between time points in a repeated-measures design. Dummy variables were constructed for each policy such that 1 = Explained the same policy ("matched") and 0 = Explained anything else ("unmatched"). Participants demonstrated a decreased level of understand across time points for all six policies, regardless of policy explained (See Figures 3a to 3f).

The HLM analysis further revealed a significant main effect of rating from Time 1 to Time 2 across all six models (i.e. policies), with betas ranging from ß = -.45 to ß = -.25 (all *ps* < .001). For two of the six policies (i.e. Cap and Trade and Retirement Age for Social Security) there was a larger drop in understanding in the domain agnostic condition (Z = -5.27 and Z = -5.81, respectively). For the four remaining policies (i.e. Sanctions on Iran, Single-Payer Health Care, National Flat Tax and Merit-Based Pay), Z's ranged from Z = -5.26 to Z = -7.59). There were no significant interactions (See Table S1 in the Supplemental Materials for full summary). This experiment offers further support for the *domain-agnostic* account. In sum, participants showed attenuated levels of understanding for each policy from T1 to T2 independent of whether or not they explained matched vs. unmatched policies.

Taken together, the first two studies show strong support for a domain agnostic account of the IOED. Nevertheless, they leave open some alternative interpretations. Perhaps, for example, the act of explanation more broadly challenges a sense of accomplishment and mastery. That is, it operates much the same as simply failing at a task or being told that you are incompetent. Such an explanation, whether operating through self-esteem threat (i.e., people feel bad about themselves more generally) or somewhat rational updating (i.e., people assume that if they are bad at one thing, maybe that means they are bad at other things), would be different than the account we have articulated. We believe that it is the act of explaining, rather than the act of *poorly* explaining, that matters. Accordingly, we would predict that even if people are called upon to explain something for which an explanation can be articulated with perfection, the experience would still puncture the inflated sense of self-knowledge.
To that end, we next manipulated whether people would be asked to explain something incredibly simple: how to boil an egg. As we show below, everyone expresses near certainty about how to boil an egg and can describe the procedure with precise accuracy. They are masterful in both their beliefs and in their execution. Nevertheless, people asked to explain how to boil an egg subsequently appear to report possessing less knowledge of sewing machine operations.

Experiment 3

Could having to explain something completely extra-categorical similarly contaminate reported levels of understanding of the focal device? With a design similar to Experiment 2, we asked participants to explain one of the following: How a helicopter flies, how an official is elected to the Nigerian House of Representatives, or how to boil

an egg. Having both "easy" and "difficult" items (egg boiling and Nigerian elections, respectively) allow us to rule out the possibility that any IOED effect is simply an artifact of making participants feel incompetent during the process of explanation as well as allow us to further confirm the *domain-agnostic* account for the illusion of explanatory depth. Given results from the previous experiments, we expected to see a statistically significant decrease in reported understanding of how a helicopter flies at Time 2 after explaining both the Nigerian electoral process and how to boil an egg.

Method

In line with the previous two experiments, we decided our exclusion criteria a priori. As in the other studies, participants were first trained on the 7-point IOED scale before rating their level of understanding for three phenomena: How a helicopter flies (i.e. "changes from hovering to forward flight"), "how an official is elected to the Nigerian House of Representatives", and "how to boil an egg." They were then randomly assigned to explain only one of them, after which they again rated their level of understanding for all three.

Results

We again sought at least 100 participants per cell. 455 subjects from Mechanical Turk took part in this experiment. A total of 133 Participants either had incomplete data or met our exclusion criteria leaving 322 for final analyses. 131 participants were excluded for providing no explanations, one for having a duplicate IP, and 1 for having incomplete data. It again appears that many participants quit when they noticed that they would have to provide a written explanation.

We first checked Time 1 ratings (1-7) to assess initial knowledge levels for each item. Participants were similarly (un)knowledgeable about helicopter flight ($M = 2.71$, $SD = 1.61$) and Nigerian elections ($M = 1.58$, $SD = 1.22$), and substantially more knowledgeable about how to boil an egg ($M = 5.98$, $SD = 1.42$). IOED results were again analyzed with T-tests. While all items showed decreased levels of understanding from Time 1 to Time 2, magnitudes varied. Level of understanding for how a helicopter flies decreased from T1-T2 both when explaining how Nigerian elections work [$M = .18$, $SD = .89$, $t(118) = 2.26$, $p = .026$] or how to boil an egg [$M = .21$, $SD = .82$, $t(112) = 2.76$, $p = .007$] (See Figure 4).

However, despite robust past results in support of the *domain diagnostic* account, understanding for how a helicopter flies, while directionally supportive, did not decrease significantly from T1 to T2 ($M = .11$, $SD = 1.0$), $t(89) = 1.01$, $p > .250$, under standard IOED conditions. That is, explanations for how a helicopter flies did not act on levels of understanding for the helicopter item itself. We are unsure how to interpret this except with reference to the other results in this experiment noted above: It appears as though having to explain either the Nigerian electoral process or how to boil an egg *better* predicts a decrease in understanding for how a helicopter flies than having to actually explain that process itself.

Experiment 4

Given the somewhat surprising results from Experiment 3 with regards to the power of egg boiling explanations on levels of helicopter knowledge, we sought to replicate it with a larger sample as well as conduct some additional exploratory tests. A different possible alternative focuses not on changes in true feelings of knowledge, but instead on the measure of knowledge itself. Perhaps merely being asked to explain something is enough to rescale the meaning of the measure of subjective knowledge. It may be the case that the person who knows that a sewing machine uses a needle, bobbin,

and presser bar, naturally assesses their knowledge as a five on a seven point scale. But after being asked to explain something (such as how a sewing machine works or how to boil an egg), the same person might interpret the scale differently and assess the same knowledge to represent only a four. Subjective knowledge has not changed in this scenario, but the meaning of the scale has. Similar to how anchors distort the meaning of measurable units (Frederick and Mochon, 2012), the explanation manipulation itself might change the meaning of a unit change on the scale. In other words, the illusion of explanatory depth may indeed tap into the cognitive processes suggested by our findings and those of previous researchers. However, it may be the case that the act of explanation operates on the measurement of the feeling of knowledge but not on the feeling itself.

Method

Using a preregistered confirmatory design, we set out to collect 1500 participants using a large online panel with whom we were already working on an unrelated large-scale project. The panel overshot by 1417, collecting 2917 responses in total. After excluding participants who exited the survey prematurely, had duplicate IP addresses or generated gibberish explanations, we were left with 2251 responses for analysis. In total, the survey ran for 22 days. Respondents were paid $0.50 each.

As in previous experiments, participants were told that they would be asked about "how things work" before being instructed on the meaning of and how to use the 7-point IOED scale. Next, they were randomly assigned to one of 6 conditions.

1. **Basic IOED task replication attempt.**

2. **Egg (Replication attempt):** From Experiment 3, above. Participants were asked how well they understood how a sewing machine works before attempting to explain how to boil an egg. They then again rated their level of sewing machine understanding.

3. **Story (Exploratory)**: Participants were asked how well they understood how a sewing machine works and asked to write about "what [they] have done today." They then rated their level of sewing machine understanding at Time 2. This condition, similar to number two above, tested whether or not a simple description (rather than explanation) could similarly lead to an attenuation in subjective knowledge from Time 1 to Time 2.

4. **Explain-Clarify (Exploratory):** Participants were asked at Time 1 to estimate their level of sewing machine knowledge, with slight modification: They were given a minor clarification under the instructions that said "Note: By "understand" we mean "how well you can explain how it works." They then attempted to explain how a sewing machine works and re-rated their knowledge at Time 2. This condition was an initial test of our lay theory that people might be misunderstanding the scale at Time 1 (i.e. scale distortion). Our clarification would ensure that they knew they were rating their level of understanding for *how well they can explain it*, as opposed to comparing their level of understanding to some other person or their levels of knowledge for other things.

5. **Accountable (Exploratory)**: Standard task procedure was again interrupted by a minor clarification at Time 1, similar to the Explain-Clarify condition reported above. This time, the clarification read "Note: Please be aware that we will later ask you to justify your answer to this question." This condition was added to test another theory that we had discussed—Namely, that

participants felt no obligation to report accurate numbers on the scale since they feel no accountability. Telling participants that they will have to justify their answers later in the survey might act to curb this bias.

6. **Repeat Time 1 (Exploratory):** Participants were asked, after rating their understanding of a sewing machine and attempting to explain how the device works, to simply re-input their level of knowledge from Time 1 on the Time 2 scale. If a drop occurs from Time 1 to Time 2, this would surely be devastating to the IOED and results from all previous experiments that have employed it would subsequently thrown into question. The implication being that participants might be systematically selecting lower numbers at Time 2, even in the face of instructions to the contrary.

Results

Results from analyses are summarized in Figure 5. Data was analyzed in line with the previous experiments. As expected, the traditional illusion of explanatory depth manipulation successfully replicated, $M_{T1} = 4.45$, $M_{T2} = 4.22$, $t(384) = 3.37$, $p < .001$. Levels of knowledge for a sewing machine was unaffected by explanations of what people did during the day ($M_{T1} = 4.55$, $M_{T2} = 4.54$, $t(417) = .278$, *ns*), as well as when they were asked to simply repeat their Time 1 response at Time 2 ($M_{T1} = 4.49$, $M_{T2} = 4.45$, $t(347) = 1.48$, $p = .14$). Surprisingly, when the survey clarified that "by 'understand' we mean 'how well you can explain how it works,'" participants reported significantly lower understanding scores at Time 2, $M_{T1} = 4.40$, $M_{T2} = 4.06$, $t(340) = 5.34$, $p < .001$. Additionally, when participants were told that they would be held accountable for their Time 1 understanding estimates later in the survey, scores at Time 2 dropped, $M_{T1} = 4.14$, $M_{T2} = 4.02$, $t(356) = 2.22$, $p = .03$.

In this experiment, however, levels of understanding for how a sewing machine works was unaffected by explanations for how to boil an egg, $M_{T1} = 4.48$, $M_{T2} = 4.53$, $t(401) = -1.14$, $p > .250$, disconfirming the previous experiment's results. We therefore move forward with the assumption that the first, lower-powered results from Experiment 4 were a false positive. We plan to further investigate the results from the successful exploratory measures reported above (i.e. The Explain-Clarify and Accountable conditions) in future experiments.

Experiment 5

Disconfirming results from Experiment 4 in hand, we designed this experiment to attempt to confirm that A) Both the domain diagnostic account and the domain agnostic accounts indeed similarly predict differences between Time 1 and Time 2 responses when participants are asked to provide explanations, and investigate the possibility of B) That explanation difficulty might moderate this effect.

Method

Again using a preregistered confirmatory design, we set out to collect 1500 participants using a large online panel with whom we were already working on an unrelated large-scale project. The panel overshot by 1346, collecting 2846 responses in total. After excluding participants who exited the survey prematurely, had duplicate IP addresses or generated gibberish explanations, we were left with 2296 responses for analysis. In total, the survey ran for 24 days. Respondents were paid $0.50 each.

As in previous experiments, participants were told that they would be asked about "how things work" before being instructed on how to use the 7-point IOED scale. Next, they were randomly assigned to one of seven conditions. These conditions were constructed to confirm that A) Both the domain diagnostic account and the domain

agnostic accounts similarly predict differences between Time 1 and Time 2 responses, and B) That explanation difficulty moderates this effect. Conditions 1-3 escalated in explanation difficulty with regards to how well participants claim to understand how a cell phone works. In the "easy" condition (Condition 1), they were asked to explain "how to turn on a cell phone," in the "hard" condition (Condition 2), they were asked to explain "how a cell phone uses radio frequency to communicate with a cell tower," and in the "impossibly hard" condition (Condition 3), they were asked to explain "how a cell phone uses ultra-linear operation to eliminate intermodulation distortion." Conditions 4-6 similarly escalated in difficulty, but instead asked participants to explain various functions of a television set instead of a cell phone, in line with the domain agnostic procedure . We added a seventh condition, asking participants to explain how to put on socks, in a final attempt to replicate the earlier implausible egg-boiling finding. We will focus on conditions 1-6.

Our preregistered expectations were the following: First, that there would be no difference between Time 1 and Time 2 scores in either the domain diagnostic nor domain agnostic conditions when participants were asked to provide "easy" explanations. Second, we expected no differences between Time 1 and Time 2 scores in the "impossibly hard" condition, again regardless of domain. However, we did expect to see relatively large differences between Time 1 and Time 2 scores for both domains among participants in the "hard" explanation condition. We argued that both "easy" and "impossibly hard" explanations would not cause participants to self-audit. That is, these tasks might lead to the same cognitive strategy: Things that are easy to explain might make one think they know things well, while impossibly hard explanations might cause a person to throw their hands into the air and forego any attempt they might have made at an explanation. In both of these cases there is not likely to be a reassessment of knowledge. When asked to explain something reasonably "hard," however, participants might expect that they should know something of the subject but simply do not, which realization might elicit a self-audit and manifest in a downward revision in ratings from Time 1 to Time 2.

Results

Data was analyzed with T-tests and summarized in Figure 6. As predicted in our preregistration, and in accordance with the disconfirming results from the "Egg boiling" explanation experiment (i.e. Experiment 4), "easy" explanations did not lead to differences between Time 1-Time 2 scores for either *domain diagnostic* explanations, i.e. "how to turn on a cell phone, " $M_{T1} = 4.75$, $M_{T2} = 4.76$, $t(352) = -0.14$, *ns*) nor *domain agnostic* explanations, or "how to turn on a television," $M_{T1} = 4.79$, $M_{T2} = 4.74$, $t(315) = 0.92$, *ns*. However, participants did indeed report similarly low levels of understanding for a how a sewing machine works at Time 2 after hard explanations, both *domain diagnostic* ($M_{T1} = 4.71$, $M_{T2} = 3.73$, $t(302) = 11.07$, $p < .001$) and *agnostic* ($M_{T1} = 4.56$, $M_{T2} = 3.77$, $t(315) = 8.99$, $p < .001$). What's more, contrary to our preregistered expectations, participants showed even more extreme Time 1-Time 2 difference scores in the impossible explanation condition, both in the *domain diagnostic* ($M_{T1} = 4.61$, $M_{T2} = 2.91$, $t(290) = 15.49$, $p < .001$) and *domain agnostic* conditions ($M_{T1} = 4.66$, $M_{T2} = 3.54$, $t(352) = 13.55$, $p < .001$). Not surprisingly, when asked to explain how to put on socks, participants did not differ in terms of sewing machine knowledge between time points, $M_{T1} = 4.76$, $M_{T2} = 4.73$, $t(333) = .64$, *ns*.

Our preregistration expected to see no differences between time points in both easy and impossible conditions, regardless of domain (*diagnostic* vs. *agnostic*). These expectations were based on the assumption that both easy and impossible explanations would not lead to a self-audit. We are happy to have been wrong. It appears that the illusion of explanatory depth escalates in parallel with levels of explanation difficulty,

without regard for domain. This experiment confirms that the illusion of explanatory depth is much more pronounced than previously thought—That is, it is not that people lose faith in their level of understanding for something as they attempt explain it (i.e. as they consult an incomplete set of mental blueprints for that specific object), it's that, more broadly, as explanations become difficult, they broadly shatter illusions of understanding across the board.

General Discussion

People have an inflated, but fragile, sense of subjective knowledge. Inflated because, across many domains, people express more mastery than they possess, and fragile because that subjective sense can be undone by merely asking people to explain their knowledge. Previous work has suggested that people interpret personal difficulty with explanation to be diagnostic of lacking knowledge in a specific domain. Rozenblit and Keil (2002) in particular argue that this is a useful heuristic and a direct cause for the domain-bound account for the illusion of explanatory depth. They suggest, for example, that the mentally animated image that is acquired as one imagines using a can-opener cutting through the lid of a can feels so much like perception that it is confused with an understanding of the actual mechanistic complexity of that device. They argue that folk theories of knowledge such as this allow people to approach the world with a sufficient sense of mastery—That is , people want to spend their time being comfortable in their level of understanding of things. If a learning moment arises, however, such as when having to provide an explanation for how something works, they recognize their apparent deficiency, realize that they know less than they thought they did about the subject, update, and move on.

We suggest that this result should be construed much more broadly: explanation reduces the sense of subjective knowledge *across* domains. That is, contrary to what the *domain-diagnostic* account would predict, people do not have infinite silos of mechanistic knowledge, updated one-by-one as their level of understanding is challenged. Instead, they appear to hold a somewhat broader view for how they understand things (i.e. their level of overall mastery). When that broad view is challenged, they update not their level of mastery for a given subject, but instead all subjects. In concept, previous accounts have treated subjective knowledge like a sponge retaining water; if you press on one spot, some water is lost, but the bulk is retained. Alternatively, we suggest that the sense of mastery is more like a balloon; a single prick and all is lost. We also show that this general bias to re-evaluate subject understanding does have its boundaries. For things that are easy to explain, like how to turn on a cell phone, there is no tendency to downwardly revise one's knowledge after explaining it. For more difficult explanations, however, people appear to recognize their ignorance and self-correct.

Additional Exploratory Results

A key element in research development is discussing findings with other experts in the field. As we undertook this process, we noticed that many of our colleagues asked us whether or not we could answer the same few questions. The questions were:

1) What happens if you make it clear to participants that they are being asked to explain *exactly* how something works, as opposed to simply "how it works," which vague wording might increase the likelihood that people misinterpret the question and give higher Time 1 responses.

2) Does this egocentric bias extend to others? That is, do people think that other people will similarly claim to understand something less after attempting to explain how it works?

3) In the vein of question 1, what would happen if you change the key dependent variable from "how well do you understand" a given device/policy to "how well can you explain" a given device/policy. Would the illusion of explanatory depth extend to this frame?

Method

We constructed an experiment to respond to the questions presented above. To investigate the first question, we had participants randomly assigned to a condition wherein they took either the standard illusion of explanatory depth task or one where simply the word "exactly" was inserted at Time 1. If differences emerge between conditions under this relatively benign manipulation we might have some additional evidence for scale distortion (as described in Experiment 4 above). That is, participants may be interpreting the scale differently from what is suggested by theory, not realizing the full extent of the question until they proceed and are asked to write an explanation. When it is made clear to participants that the question is asking if they know "exactly" how something works might lead them to provide lower numbers at Time 1, dampening any large differences between time points. To investigate question 2 with regards to how people might respond for themselves vs. how they think others would react, we again had participants randomly assigned to one of two conditions. They either took part in the standard IOED task (the "self" condition) or completed one where they were asked how "someone else taking this survey right now" might respond (the "other" condition). Finally, to answer question 3, we had participants complete the standard IOED task with slight modification: Instead of being asked "how well they understand" how a sewing machine works, they were asked "how well they could explain" how it worked. For some conditions we varied whether or not participants received full instructions vs. no instructions at all. While this was initially meant as an effort to test whether or not the inclusion of instructions might lead to differences in the magnitude of the IOED, we failed to include adequate comparison conditions in all cases. We therefore reserve judgment on whether or not instructions matter for this task and will conduct an experiment in the future to test this possibility.

Results

We sought at least 100 participants per cell, conducting a six-condition within-subjects experiment. In total, 900 participants from Amazon's Mechanical Turk responded. After removing participants who either did not complete the survey or had duplicate responses (based on IP address), we were left with 657 responses for analysis. Overall results are found in Figure 7. Recall that our colleagues' first question was whether or not adding the word "exactly" at Time affects responses. Comparing conditions 2 and 6 showed that the inclusion of the word "exactly" does indeed significantly affect the magnitude of claimed understanding at Time 1, $F(1, 193) = 9.84$, $p = .002$ (see Figure 8). That is, participants who responded to the Standard IOED task showed greater decreased Time 2 estimates [$t(97) = 5.76$, $p < .001$] than those who were asked if they knew "exactly" how a sewing machine works, $t(116) = 1.73$, $p < .09$. Our colleagues' second question related to how participants might respond for themselves vs. others. Comparing conditions 4 and 6, we see that participants do not reliably differ in magnitude of the illusion of explanatory depth when they respond for themselves [$t(97) = 5.76$, $p < .001$] vs. respond for someone else [$t(105) = 3.54$, $p = .001$], $F(1, 202) = 2.24$, $p = .14$. The third and final question that we investigated in this experiment was whether or not changing the dependent variable from "how well do you understand" to "how well can you explain" could lead to a different pattern of responding. As seen in Figure 10,

using “how well can you explain” as a DV still leads to the same familiar attenuation pattern seen with the traditional DV, $t(128) = 2.94$, $p = .004$. In the interest of completeness, we also compared this condition to one wherein participants responded to this new DV, but instead answering for “someone else who might be taking this survey right now.” When we run this self vs. other comparison for this new DV, we see, surprisingly, that participants do not respond identically for themselves [$t(128) = 2.94$, $p = .004$] as they do for someone else [$t(109) = 0.76$, *ns*], $F(1, 237) = 6.15$, $p = .014$. Why this DV might lead to a different pattern of responding in self vs. other comparisons than one that uses the traditional DV will be a fascinating question for future research.

References

Alter A. L., Oppenheimer, D. M., Zemla, J. C. (2010). Missing the trees for the forest: a construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99(3)*, 436-51.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology,* 90(1), 60-77.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological science*, *24(6),* 939-946.

Fernbach, P. M., Sloman, S. A., St. Louis, R. & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research,* 39 (5), 1115-1131.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

Frederick, S. & Mochon, D. (2012). A scale distortion theory of anchoring. Journal of Experimental Psychology: General, 141(1), 124-133.

Haugtvedt, C. P., Petty, R. E., & Cacioppo, J. T. (1992). Need for cognition and advertising: Understanding the role of personality variables in consumer behavior. *Journal of Consumer Behavior*, 1, 239-260.

Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology,* 77(6), 1121-1134.

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

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review,* 115(2), 502-517.

Simmons, J. P. Nelson, L. D., Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science,* 22(11), 1359-1366.

Rozenblit, R., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *Vol 26(5)*, 521-562.

Trope, Y., & Liberman, N. (2010). Construal-Level Theory of Psychological Distance. *Psychological Review*, *117*(2), 440–463.

Figure 1.

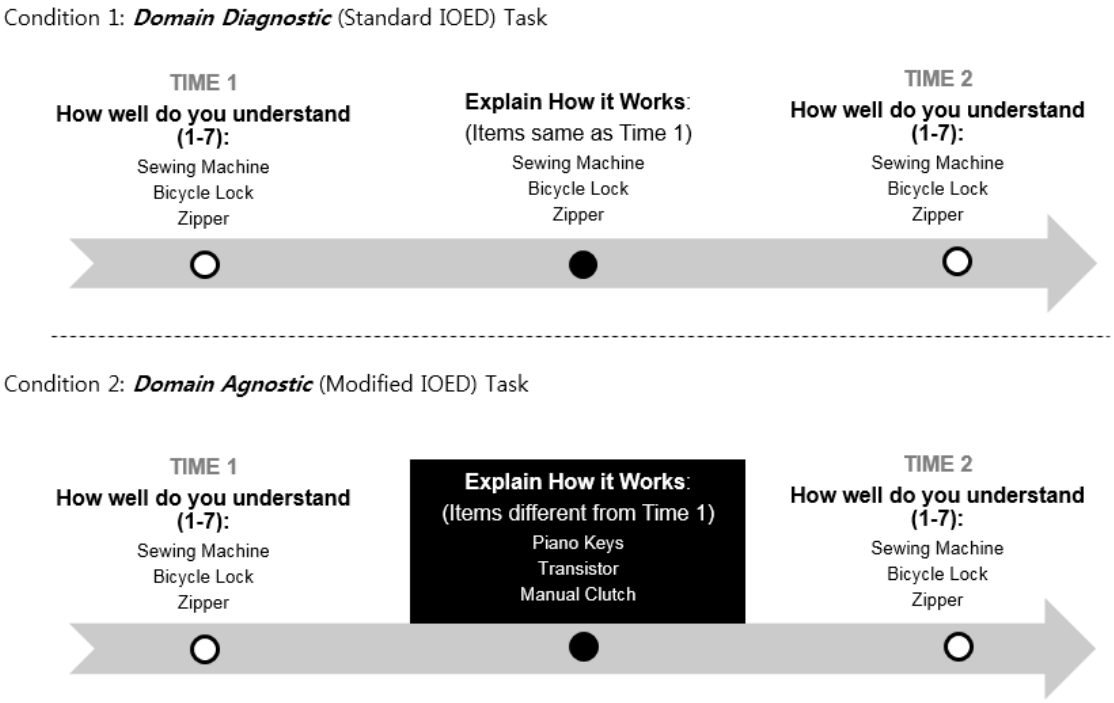
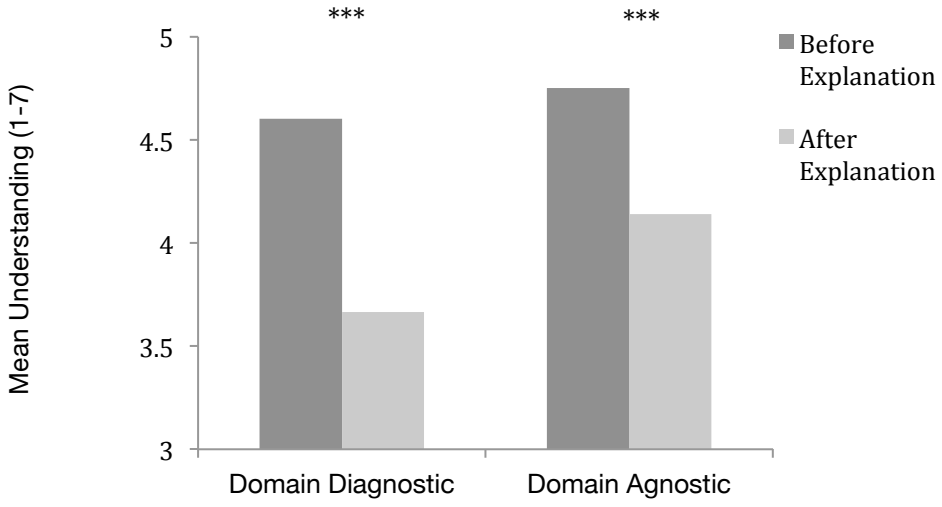
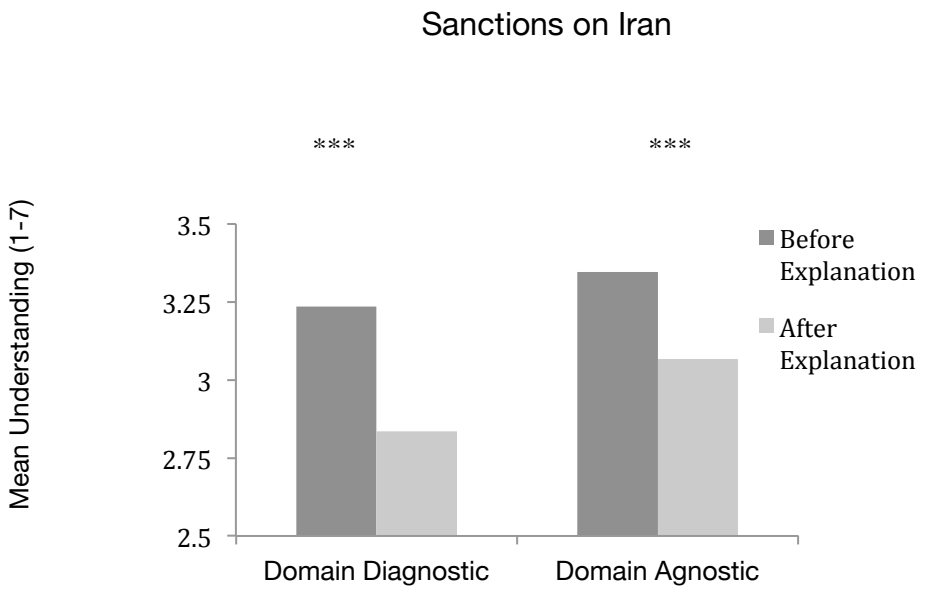

Condition 1: Domain Diagnostic (Standard IOED) Task
TIME 1
How well do you understand (1-7):
Sewing Machine
Bicycle Lock
Zipper
Explain How it Works:
(Items same as Time 1)
Sewing Machine
Bicycle Lock
Zipper
TIME 2
How well do you understand (1-7):
Sewing Machine
Bicycle Lock
Zipper
Condition 2: Domain Agnostic (Modified IOED) Task
TIME 1
How well do you understand (1-7):
Sewing Machine
Bicycle Lock
Zipper
Explain How it Works:
(Items different from Time 1)
Piano Keys
Transistor
Manual Clutch
TIME 2
How well do you understand (1-7):
Sewing Machine
Bicycle Lock
Zipper

Figure 2.



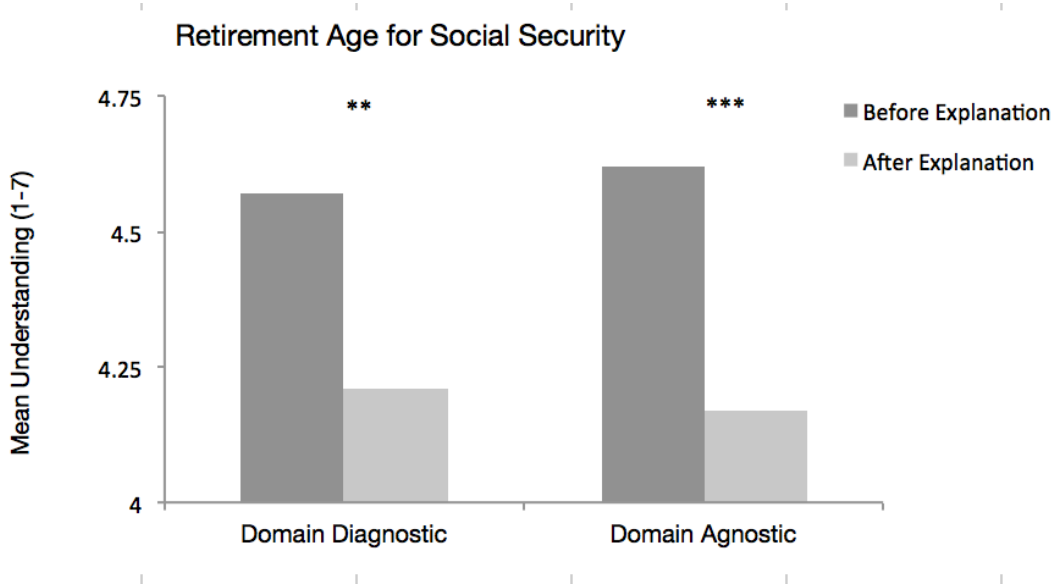
***$p < .001$

Figure 3a.



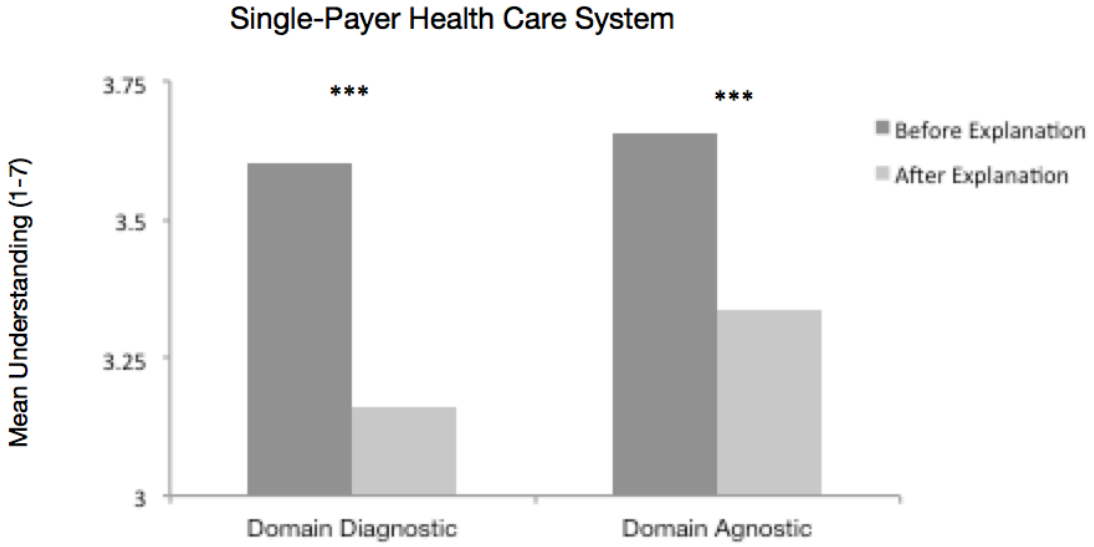
***$p < .001$

Figure 3b.



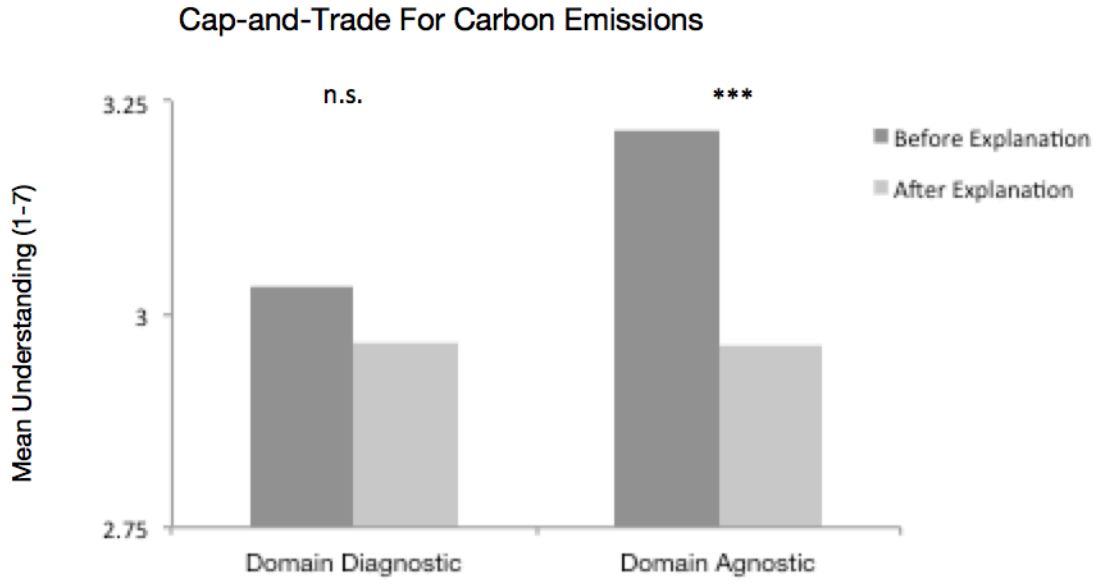
**$p < .03$
***$p < .001$

Figure 3c.



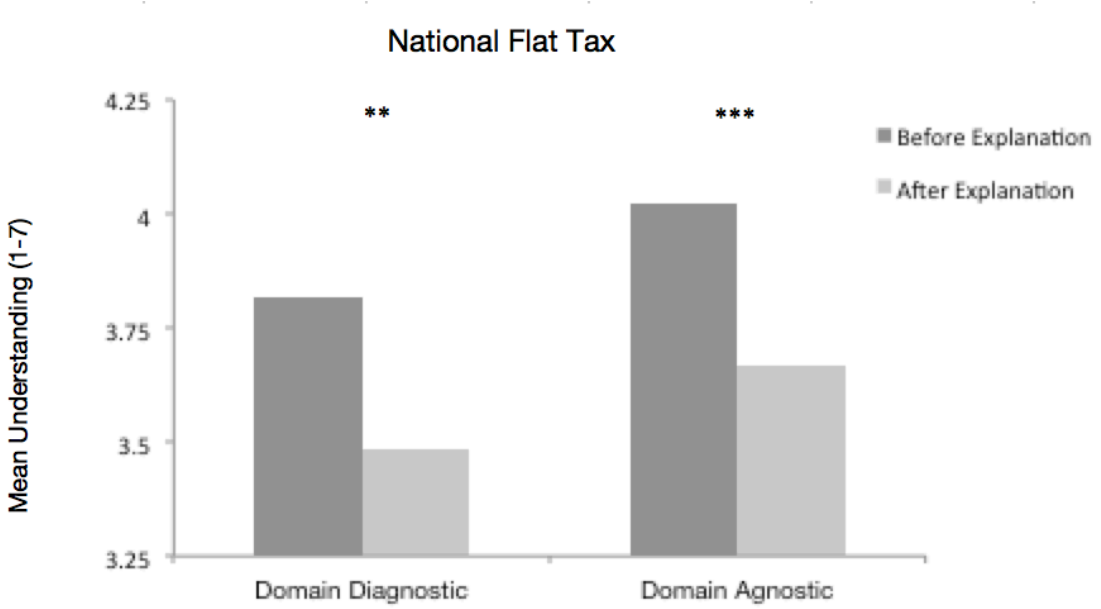
***$p < .001$

Figure 3d.



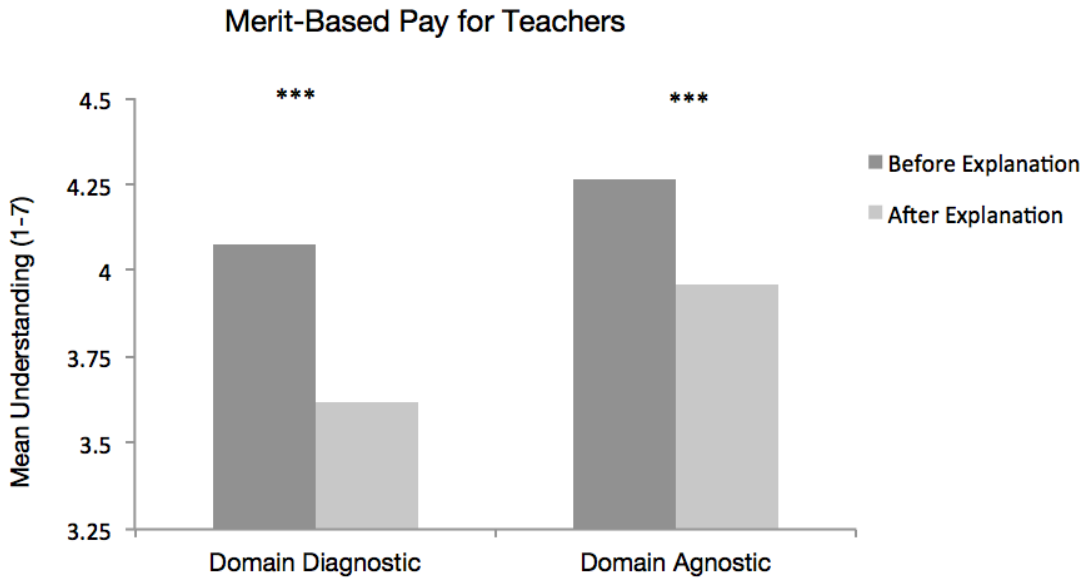
***$p < .001$

Figure 3e.



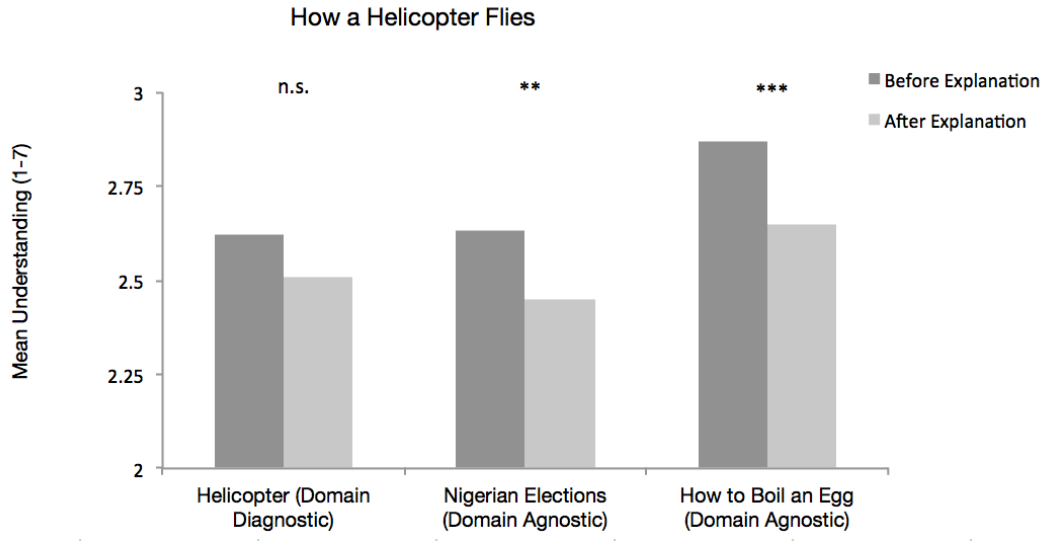
***$p < .001$

Figure 3f.



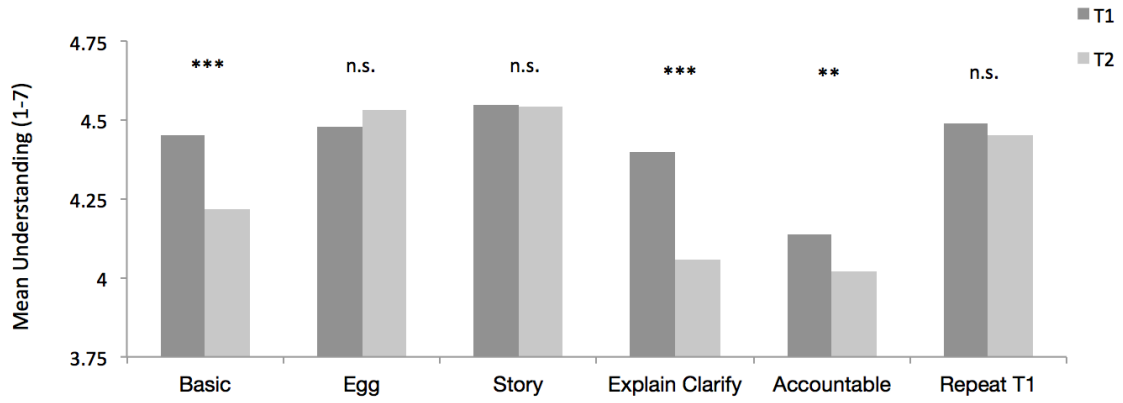
***$p < .001$

Figure 4.



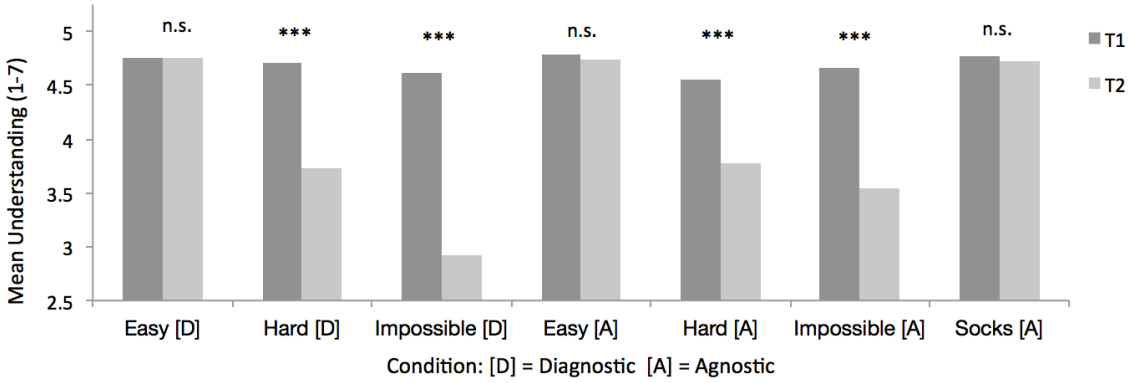
** $p < .03$
*** $p < .01$

Figure 5.



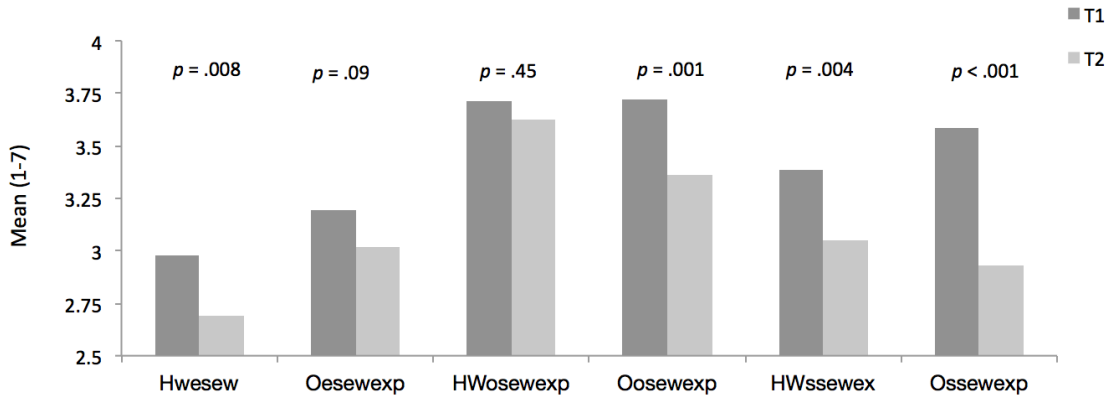
***$p < .001$
**$p < .03$

Figure 6.



***$p < .001$

Figure 7.



HWesew = Self, "How well can you explain **exactly**," no instructions
Oesewexp = Self, "How well do you understand **exactly**," instructions
HWosewexp = Other, "How well can others explain," no instructions
Oosewexp = Other, "How well do you understand," instructions
HWssewex = Self, "How well can you explain," no instructions
Ossewexp = Self, "How well do you understand," instructions

Figure 8.

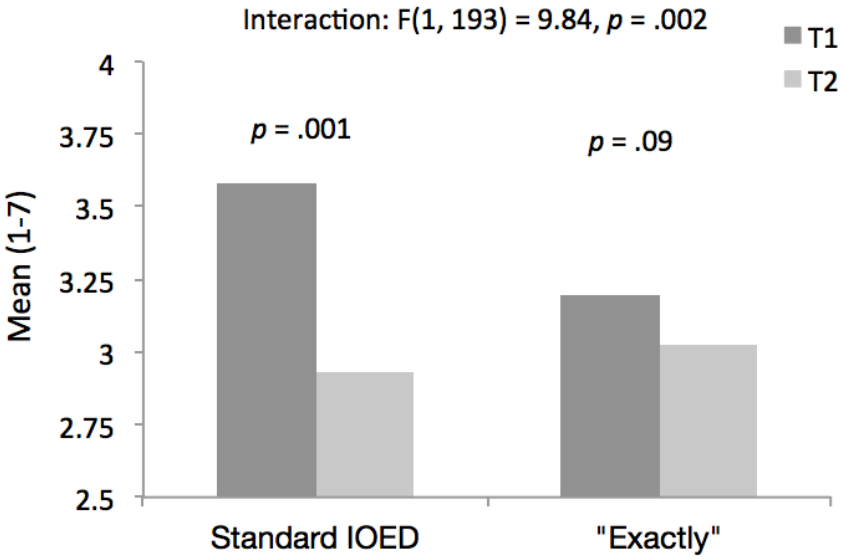
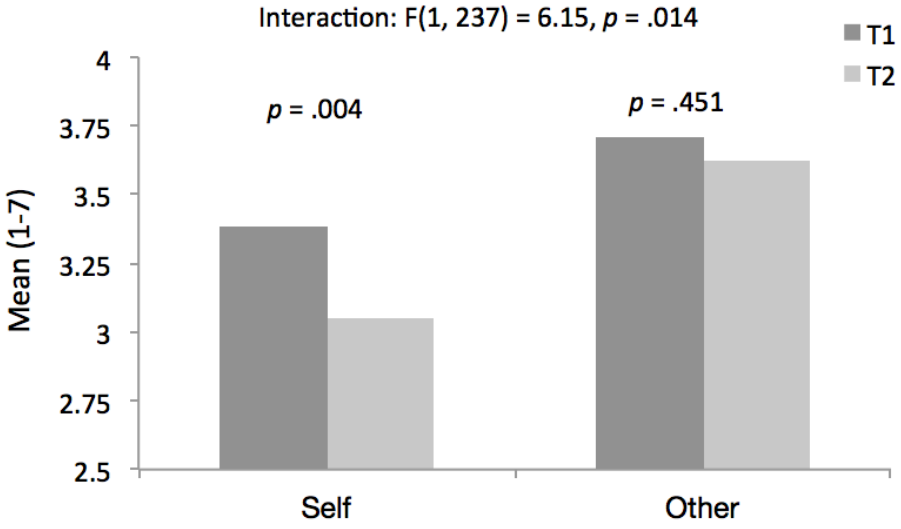

Interaction: F(1, 193) = 9.84, p = .002
T1
T2
4
3.75
3.5
3.25
3
2.75
2.5
Mean (1-7)
p = .001
p = .09
Standard IOED
"Exactly"

Figure 9.


Interaction: F(1, 202) = 2.239, p = .136
T1
T2
p < .001
p = .001
Mean (1-7)
4
3.75
3.5
3.25
3
2.75
2.5
Self
Other

Figure 10.


Interaction: F(1, 237) = 6.15, p = .014
T1
T2
p = .004
p = .451
Mean (1-7)
4
3.75
3.5
3.25
3
2.75
2.5
Self
Other