Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# Surface-Level Knowledge, Crowd-Level Confidence: Two Biases in One Mindset

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

In

Psychology

At Massey University, Manawatu, New Zealand

Tobias Colville

2025

## Abstract

The present research investigated whether prompting individuals to generate detailed explanations of complex topics recalibrates not only self-assessed knowledge but also perceptions of others' knowledge and social consensus. Two studies combined an explanation task with consensus estimation measures across mechanical and political domains. In Study 1 (mechanical devices), 110 participants rated their own understanding, the understanding of others, estimated how many others shared their level of knowledge, produced causal explanations, and then re-rated these judgments. Explanatory prompting reliably reduced self-rated knowledge and modestly reduced perceived consensus. In Study 2 (political policies), the same procedure (N = 123) was implemented with a revised consensus measure. Although overestimation of consensus persisted, explanation-induced recalibration was weaker and did not reliably influence consensus estimates.

Across both studies, participants consistently overestimated how many others shared their level of understanding, demonstrating robust consensus overestimation even when knowledge ratings were modest. However, the link between explanation-induced recalibration and consensus judgments appeared domain-sensitive: mechanical topics showed partial coupling between self-revision and consensus revision, whereas political topics did not. These findings suggest that explanatory recalibration does not automatically generalize to social projection, particularly in identity-relevant domains. Interventions designed to improve epistemic humility through explanation may therefore have limited impact on perceptions of shared understanding unless social consensus assumptions are addressed directly.

**Acknowledgements**

Firstly, I would like to thank my supervisor, Dr. Stephen Hill, for the immense amount of support that he provided during the creation of this thesis. I really couldn't have done it without your wealth of knowledge and experience. Your ability to gently take my ideas (and sometimes confused ramblings) and provide structure and direction was much needed and greatly appreciated. Your approach to supervision made me feel capable and respected, and turned feelings of doubt into feelings of determination.

To my family, thank you for sticking by me while I completed this project. Thank you for believing in me and giving me the confidence and motivation to keep studying, your unwavering faith means the world to me. Thank you for being my biggest supporters.

And to my friends, thank you for listening to me go on about this thesis for a whole year, I know it's not the most exciting topic of conversation. Thank you for all the good times, and providing some much needed distraction from this thing. I'm extremely fortunate to have such genuinely good people around me.

**Table of Contents**

**Surface Level Knowledge, Crowd-Level Confidence: Two Biases in One Mindset.....6**

The Illusion of Explanatory Depth Overview and Original Research.....8
Background information, Underlying Mechanism and Domains.....11
Similarities and Differences Between Domains.....15
Further Research on Underlying Mechanism.....17
The False Consensus Effect Overview and Original Research.....18
Mechanisms, Moderators, and Extensions of the False Consensus Effect.....21
Purpose of linking IOED and FCE.....24
Research Informing Design of Current Study.....26
Hypotheses and Research Questions.....28

**Experiment 1 - Devices.....30**

Method.....30
*Participants*.....*30*
*Materials and Procedure*.....*31*
*Adapting FCE Measures to a Knowledge Judgement Context*.....*32*
Results.....33
*Illusion of Explanatory Depth for Self*.....*33*
*Illusion of Explanatory Depth for Others*.....*34*
*False Consensus Effect*.....*35*
*Change in Consensus*.....*37*
*Change in Consensus Accuracy*.....*39*
Discussion.....40
*Illusion of Explanatory Depth for Self*.....*40*
*Illusion of Explanatory Depth for Others*.....*43*
*False Consensus Effect*.....*44*
*Change in Consensus*.....*46*
*Change in Consensus Accuracy*.....*47*

**Experiment 2 – Policies.....50**

Method.....50
*Participants*.....*50*
*Materials and Procedure*.....*51*
*Modification of Consensus Measure to Reduce Extreme Responses*.....*52*
Results.....53
*Illusion of Explanatory Depth for Self*.....*53*
*Illusion of Explanatory Depth for Others*.....*54*
*False Consensus Effect*.....*55*
*Change in Consensus*.....*56*
*Change in Consensus Accuracy*.....*57*
Discussion.....58

*Illusion of Explanatory Depth for Self* ..... 58
*Illusion of Explanatory Depth for Others* ..... 61
*False Consensus Effect* ..... 63
*Change in Consensus* ..... 64
*Change in Consensus Accuracy* ..... 64

**General Discussion** ..... **66**

Summary of Key Findings ..... 66
Interpreting the IOED Across Domains ..... 69
Understanding FCE in Knowledge Judgements ..... 70
Why the IOED-Style Explanation Task did not Reliably Influence FCE ..... 73
Limitations ..... 75
Future Research ..... 77
Broader Implications ..... 77

**Conclusion** ..... **79**

**References** ..... **82**

**Appendix A – Study 1 Survey** ..... **89**

**Appendix B – Study 1 Descriptive Statistics** ..... **98**

**Appendix C – Study 2 Survey** ..... **101**

**Appendix D – Study 2 Descriptive Statistics** ..... **121**

# Surface Level Knowledge, Crowd-Level Confidence: Two Biases in One Mindset

Most people believe that they have a more complete understanding of the world around them than they do. This isn't an accident. Our minds have evolved to convey to us a version of the world that makes sense and allows us to operate more efficiently within it, even if this version of the world is not entirely accurate (Haselton et al., 2009). Over the years, cognitive science has lifted the veil on many of the shortcuts, heuristics, and biases that our minds use to streamline our lives and reduce cognitive load. Exposing these shortcuts has provided a perspective on human cognition that paints the picture of a mind filled with gaps in knowledge, illusory beliefs, and opinions with unstable foundations. The implications of this perspective are far-reaching and have the power to cause considerable change to our lives without us even knowing it.

Take a seemingly low-stakes example like that from Daniel Kahneman's *Thinking Fast, and Slow* (2011).

"*Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.*

*Is Steve more likely to be a librarian or a farmer?*"

Most people's immediate assessment is that Steve is more likely to be a librarian. This decision takes into account that most librarians that we come across, and indeed most depictions of librarians that we can bring to mind, portray those with such an occupation in the same way that our Steve is described. On the contrary, most farmers that we come across, and again, most depictions of farmers, might be described differently.

However, few people consider that there are many times more male farmers than there are male librarians. In the United States for example, male farmers outnumber male librarians

about twenty to one (Kahneman, 2011). So, we can conclude that despite his stereotypically librarian-esque description, Steve is much more likely to be a farmer.

In this example, people rely on stereotyping, a cognitive shortcut that allows us to use previously formed generalisations to quickly draw conclusions. Along with stereotypes, people also fall prey to the *availability heuristic.* This is when people tend to rely heavily on the information available to them, Steve's description, rather than taking a step back and drawing on information outside of that provided in the question, like statistics of occupation rates.

Although this example seems relatively harmless, "why care if we mix up a librarian with a farmer?", the implications of these cognitive biases extend to more than just cases of guessing occupations. For instance, what if we had a similarly detailed description about an alleged perpetrator of a violent crime? Would a jury fall prey to the availability heuristic and stereotyping too? The literature would argue that they would (Curley et al., 2022; Greene & Ellis, 2007).

Although it helps to paint a thought-provoking example, this thesis is not about the availability heuristic or stereotyping. It is about two other cognitive biases that are perhaps lesser known and more likely to be overlooked, the Illusion of Explanatory Depth (IOED) and the False Consensus Effect (FCE). These phenomena are rooted in the idea that people tend to make incorrect assumptions about themselves and others. The IOED refers to the tendency for people to overestimate their own understanding until they attempt to provide an explanation and the FCE refers to the tendency of people to overestimate how similar other people are to them. When a person falls prey to these biases they make an inaccurate assumption about the world around them, and these inaccurate assumptions can lead to incorrect judgements and actions.

Although these biases differ in terms of what type of overconfidence they produce, there is a rationale to explore a possible link between them. When applied to a social context, it is simple to reach the conclusion that being both overconfident in your own understanding while also being overconfident about how many people agree with you could create a situation where opinions that differ from your own are automatically dismissed. The social ramifications of this as well as other reasons for linking the IOED and FCE will be explored further in later sections.

**The Illusion of Explanatory Depth Overview and Original Research**

"*The wise man knows that he knows nothing*" is an often-used appropriation of Socrates' words about his search for meaning after the Oracle of Delphi proclaimed him the wisest man in Athens. This quotation emphasises the importance of the Socratic method and champions humility in knowledge. For those who understand the limitations of their knowledge understand what they can and can't explain.

The quote highlights a central part of the illusion of explanatory depth and the study of cognitive biases in general. Fundamentally, the IOED is a phenomenon that causes people to discover that they know less than they thought they did, finding where the limits of their knowledge lay. The term *Illusion of Explanatory Depth* was first put to paper in the article *The Shadows and Shallows of Explanation* by Wilson and Keil (1998). They described a tendency for people to go through life operating on explanations that are deceivingly shallow, yet complex enough to carry them through everyday interactions. They described a theoretical abyss between our ability to provide these shallow explanations and the knowledge and ability needed to provide richer, more theoretically sound explanations like those which come from experts. They posit that people tend to confuse their ability to provide shallow explanations that 'get the job done' with the ability to provide explanations that are complete and theoretically sound.

This idea was explored further in Rozenblit and Keil's (2002) article. Within, the authors conducted a series of twelve experiments that focused on highlighting and replicating this effect in different conditions. As the studies in this paper progress, the central method is distilled to three steps:

1. Ask the participant how much they know about or understand an object or topic.
2. Ask the participant to provide a detailed explanation of said object or topic.
3. Ask the participant again how much they know about or understand the object or topic.

They repeatedly found that participants' second rating of their own knowledge was lower than their first rating. From this result, the authors concluded that people tend to overestimate their knowledge, and this overestimation is only revealed to them once they attempt to provide an explanation. The act of producing an explanation forces people to explore where the limits of their knowledge lie, and in most cases, find that those boundaries are much closer than they thought.

This original study tested this model in multiple domains, including knowledge of mechanical devices, natural phenomena, facts, procedures, and narratives. Interestingly, the effect size was significantly larger when concerning mechanical devices and natural phenomena than when concerning procedural knowledge or knowledge of facts. This led the researchers to believe that this phenomenon is domain specific, which helped them to argue that general overconfidence is not to blame. If the effect was due to general overconfidence, it would likely occur with similar magnitude when asked about any domain. Instead, because it occurred more in certain domains of knowledge, it cannot be explained by the general sense of overconfidence that is present in many people, although they do believe it is additive to the effect (Kruger & Dunning, 1999; Yates et al., 1996).

Instead of overconfidence, Rozenblit and Keil argue that the illusion of explanatory depth occurs because people mistake familiarity, shallow causal schemas, and socially distributed knowledge for genuine mechanistic understanding. Instead of just being overconfident, people are deceived into an illusory perception of their own understanding.

In their paper, the IOED is linked to the specific task of explaining complex systems with causal chains. That is, systems that have a clear chain of events/interactions that each lead to the next step in the chain. Knowledge of facts does not fit into this model because knowledge of a fact is a binary measure, one either knows the fact or does not. This differs from knowledge of causally complex systems because one's understanding of such a system exists on a spectrum. Facts also lend themselves more easily to self-testing than causal explanations for the same reason. It is much simpler to know whether you got a fact right or wrong than it is an explanation of a complex system. Moreover, this self-testing of factual knowledge is likely to happen much more frequently than self-testing of explanations. For example, in conversation, knowledge of facts comes up fairly often. Someone might ask you what the actor on screen's name is or what country the city of Vienna is in, whereas being asked to give a detailed explanation of how something like an air conditioner works is much less common for most people.

On the other hand, knowledge of procedures like baking a cake or changing a tyre provide a much more useful analogue for explanations of causally complex systems. In procedures and explanations alike, one step leads to another – a causal chain is present and can be recited if the person knows the steps. However, when Rozenblit and Keil (2002) tested this domain of knowledge they found no decrease in self-ratings. In fact, they found a slight (non-significant) increase.

They proposed three possible explanations for this difference. First, participants may be less likely to misjudge the appropriate level of analysis needed when explaining a procedure.

Second, they may be less prone to confusing 'environmental support' with genuine procedural knowledge than they are for devices or natural phenomena. Third, the end states of procedures are likely to be more apparent than the end states of causal explanations. Together, these differences lend more credibility to the theory that the IOED is not due to general overconfidence, is domain-specific, and is best illustrated with knowledge of complex systems that require deep, theoretically sound knowledge of the underlying processes that make them work rather than other types of knowledge.

Further research has attempted to use this model to find other domains that the IOED is present in as well as to provide theoretical explanations and identify factors that influence the magnitude of the effect.

**Background information, Underlying Mechanism and Domains**

Since Rozenblit and Keil's (2002) paper was published, this area of research has developed, and many researchers have set out to expand our understanding of the IOED. Different branches of the literature have aimed to answer questions around the theoretical underpinnings and mechanisms behind the effect, the effect in novel knowledge domains, competing theories, as well as introducing other interesting and sometimes contradictory findings.

One of the principal mechanisms that has emerged for the IOED, expanding on the original authors' ideas, is the idea of 'environmental support' and transactive memory systems (TMS). In the original study, the authors theorise that the effect is partly caused by people's confusion between what they can infer from an object in front of them with information stored in their memory. Because of this they suggested that the effect would be more pernicious in cases where components of the device are more easily visualized (Rozenblit & Keil, 2002). For example, someone explaining how a zipper works may be more susceptible to the IOED because the zipper and its component parts are all easily visualized. This leads to a more

profound IOED because the participant confuses the readily available information that is stored within the object with information in their memory.

This is a similar concept to the idea of a one-way transactive memory system. A transactive memory system is a system where information is shared between multiple sources (Wegner, 1987). Within a TMS one relies on an external source to retrieve information which can lead people to overestimate their own understanding, mistaking accessible knowledge for personally held knowledge. In this example the zipper itself is the external source of knowledge which the participant relies on. Fisher et al. (2015) explored this idea more deeply by conducting a series of experiments where participants were found to conflate their own knowledge with knowledge found on the internet. Participants were asked to rate their ability to answer an explanatory question, then asked to confirm their explanation by using the internet, and then asked to rate their explanatory ability again in regard to six other, unrelated questions. Those who used the internet to answer the question subsequently produced increased self-ratings for their ability to answer the unrelated questions. This finding suggests that people use the internet as part of a TMS, similarly to how the original authors saw the object itself as part of a TMS. It is this conflation of knowledge that may be one of the foundations of the IOED.

Rabb et al. (2019) takes these ideas further by creating a framework for what they refer to as a collective knowledge system. They suggest that individuals possess three types of knowledge. First, Type I knowledge – this is detailed, intricate knowledge. Most people possess Type I knowledge for a few domains like those that they have a specific interest in or is needed for their occupation. Second, Type II knowledge – this is much more coarse, superficial knowledge. People possess this type of knowledge for most domains, things that they lack a deep understanding for. And last, Type III knowledge – this is a kind of metaknowledge, markers for where expert (Type I) knowledge can be found. For example, a

mechanical engineer that loves cooking might have Type 1 knowledge in the domains of engineering and cooking, Type 2 knowledge for most other things, and Type 3 knowledge for where they can access information that fills the gaps. In today's world, the internet is an enormous source of Type I knowledge and because it is so easily accessible, it inflates individuals' perceptions of their own knowledge (Fisher et al., 2015; Hamilton & Yao, 2018). People tend to mistake Type III knowledge with Type I knowledge, which could contribute heavily to the IOED (Rabb et al., 2019).

Another branch of the literature has delved into the mediating effect that individual differences may have. As with many things, some people are just more predisposed to some things than others, and the IOED doesn't seem to be any different. Fernbach et al. (2013) explored this idea by asking participants to complete the Cognitive Reflection Test (CRT) and an IOED task and compared their scores. They found a correlation between these two scores, those that did poorly on the CRT demonstrated a more profound IOED. This makes sense considering there is some similarity to the two tasks. To perform well on the CRT, you must be able to disregard your initial, intuitive thoughts in favour of more deliberate and rational thoughts. One of the questions for example:

*A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?* (Frederick, 2005).

For most people the intuitive thought is that the ball must cost ten cents. However, this is incorrect, the ball must cost five cents for the bat to cost a dollar more than the ball. This test is meant to separate those that operate on a more intuitive, quick system of thinking rather than a slower and more deliberate system of thinking. It follows then that scores on the CRT correlate with scores on the IOED task because part of the IOED may be that participants make a quick, intuitive judgement of their knowledge of the topic without taking the time to think deliberately about their ability to produce an explanation. This system of thinking may

tie the two tasks together, as well as point to a basis for individual differences to have a mediating effect on the IOED. That is, some people may just be more prone to exhibiting it due to the way that they think.

Alter et al. (2010) take a similar position on this idea. However, instead of discussing individual differences as a contributing factor to the IOED they posit that an individual's adopted *construal style* may be to blame. Construal style exists on a continuum between abstract and concrete. When two people are asked the same question, "what do you know about ballpoint pens?", for example, those who adopt a concrete construal style will produce a different answer from those who adopt an abstract construal style. While the abstract respondent might explain that a ballpoint pen can be used to write and convey information, the concrete respondent might explain that the ball at the tip of the pen allows ink to be transferred from the ink cartridge to the paper by rolling. Obviously, these answers are completely different, but both valid answers to the question.

Alter et al. (2010) conducted six studies showing that construal level plays a significant role in the IOED. Individuals who naturally adopted a more concrete construal style exhibited a smaller IOED than those who tended toward abstract construal styles. When construal level was experimentally manipulated, the same pattern emerged: participants induced to think concretely showed a reduced IOED compared with those induced to think abstractly. These findings suggest that people often rely on abstract, high-level representations when judging their understanding of concrete mechanisms, leading them to overestimate how well they could explain them. At first glance, this appears inconsistent with Study 6 of Rozenblit and Keil (2002), in which participants were explicitly warned that they would later be asked to explain the items they were rating. Although this instruction should, in principle, prompt more concrete thinking, the IOED nevertheless persisted. This discrepancy highlights that prompting individuals to think concretely may be insufficient on its own; concrete construal

may need to be actively engaged rather than merely suggested to meaningfully reduce the IOED.

**Similarities and Differences Between Domains**

One of the central observations in Rozenblit and Keil's (2002) original work was that the IOED appeared most reliably in domains involving mechanistic systems and natural phenomena. However, subsequent research has demonstrated that the IOED is far more pervasive than initially proposed. A growing body of evidence shows that people also overestimate their understanding in a wide range of abstract and socially important domains, including mental disorders, historical events, political issues, conspiracy beliefs, and other forms of conceptual knowledge (Alter et al., 2010; Fernbach et al., 2013; Gaviria et al., 2021; Vitriol & Marsh, 2018; Voelkel et al., 2018; Zeveney & Marsh, 2016). This broader literature indicates that the IOED is not confined to concrete mechanical systems but reflects a more general metacognitive bias that emerges across diverse topics whenever individuals rely on shallow or gist-based representations of complex information.

Among these domains, political topics stand out as a particularly consequential and widespread context in which the IOED is observed. People tend to know less about political topics than they think they do (Alscher et al., 2025). On top of this, Fernbach et al. (2013) claims that political extremism is supported by this overconfidence and more specifically, the IOED because people do not realise how little they know about policies until they are asked to explain them. They posit that this illusion of knowledge leads to a more radical stance on issues and this heightened radicalisation leads to a more polarized society. To explore this hypothesis, they asked participants to rate their understanding of political topics as well as their political position on the topic, the participants were then asked to provide a causal explanation of the policy, and then asked to rerate both their position and their understanding of the policy. This exercise resulted in a lower rating of understanding as well as more

moderate stances on the policies, providing evidence that the IOED exists in political topics, and also may contribute to political extremism and polarisation.

An important qualification about politically charged domains is highlighted by Vitriol and Marsh (2021). In their study examining how an IOED-style explanation prompt influenced conspiratorial thinking about COVID-19 policies, they found that the intervention sometimes *increased* conspiracy endorsement, particularly among individuals already predisposed to conspiratorial beliefs. Rather than uniformly reducing overconfidence, the explanation task appeared to amplify motivated interpretations for certain participants, leading them to rationalise or reinforce their prior beliefs. This finding adds meaningful nuance to the idea that IOED interventions can be used to reduce political extremity or polarisation. Although explanation prompts often reveal knowledge gaps, their effects are not universally corrective and may, in some individuals, produce the opposite of the intended outcome.

Continuing this theme that the IOED could be used as an intervention to decrease polarisation and radicalization, Meyers et al. (2020) offer useful insight into how people respond to expert advice. Under normal circumstances, they found that people do not adjust their beliefs much more after hearing from experts than after hearing from laypeople, which is surprising given that experts clearly have more relevant knowledge. However, Meyers and colleagues found that this changes once the IOED is brought to light. When people realise that their understanding is weaker than they thought, they become more open to expert input and less swayed by lay opinion. Interestingly, this shift occurs even when the IOED is revealed in an unrelated topic, suggesting that recognising one's own knowledge gaps can have a broader impact on how people evaluate new information.

Zeveney and Marsh (2016) go one step further by examining how the IOED is affected by consideration of others' knowledge. Their research highlights that judgements of personal understanding are not solely based on internal reflection. They found that manipulating what

participants believed about how much experts understand about a topic influenced how the IOED manifested. When participants believed that experts possessed a deep knowledge of the topic, their ratings of their own knowledge showed a more pronounced decrease after the explanation task compared to when they were led to believe that experts possessed only shallow knowledge.

**Further Research on Underlying Mechanism**

Two further studies sought to clarify ideas around the precise mechanism of the IOED. Roeder et al. (2015) explored the effect of cross-domain explanations on the IOED, and Johnson et al. (2016) explored how simply reflecting on one's explanatory ability mediated the IOED. Both of these studies looked to modify the traditional IOED paradigm to provide deeper understanding of the mechanism responsible.

It appears that the IOED persists even when asking participants to explain a topic from a completely different domain than that of which they are rating their understanding (Roeder et al., 2015). In the basic IOED paradigm, the participant would rate their understanding of an object/topic, explain said object/topic, and then rerate their understanding of the object/topic. In Roeder et al. (2015) this second step was altered so that participants would be asked to explain a different object/topic from which they were rating their understanding, and particularly from a different domain of knowledge. For example, ask a participant to rate their understanding of a zipper (mechanical object), then ask them to provide an explanation of how a rainbow is formed (natural phenomenon), then ask them to rerate their understanding of the zipper. This approach allowed the authors to produce an argument against the domain-bound hypothesis and push future research toward a domain-agnostic view.

Similarly, Johnson et al. (2016) explored the idea of replacing the explanation phase of the IOED paradigm with an exercise they called 'reflecting on explanatory ability' (REA). Instead of asking participants to explain how the zipper works, they simply asked them to

reflect on their ability to explain the object in a step-by-step, causally connected manner. Even without producing any actual explanations the participants experienced a similar decrease in perceived knowledge ratings. This aligns with Roeder et al. (2015) by showing that the act of actually explaining the object/topic may not be the only intervention that can cause people to realise the shallowness of their understanding.

**The False Consensus Effect Overview and Original Research**

When Anaïs Nin wrote "we don't see things as they are, we see them as we are" in *The Seduction of the* Minotaur (1961), she touched on an idea that is central to the false consensus effect. People do not simply interpret the world through a neutral lens. Their perceptions are shaped by their own beliefs, preferences, experiences, and emotional states. These subjective filters guide the quick judgements we make about why events occur, what is reasonable, and how others ought to think or behave. However, the false consensus effect arises when individuals go a step further and assume that their own interpretations are widely shared. Because their personal viewpoint feels natural and self-evident, they treat it as typical rather than idiosyncratic. As a result, two people exposed to the same information may reach very different conclusions, yet each may believe their view reflects common sense and represents what most others think as well.

Originally put to paper by Ross, House, and Greene (1977), the FCE is a pervasive effect that describes a tendency for the average person to overestimate how common their own thoughts, feelings, behaviours, and opinions are in the general population. For example, if you preferred brown bread to white bread you would be likely to overestimate the proportion of other people who also prefer brown bread to white. Or, for a more impactful example, if you believed that marijuana should be decriminalised, you would be likely to overestimate the proportion of people who agreed with you. This overestimation of the similarity between ourselves and others creates a 'false consensus' where we tend to believe that we hold the

same opinion as the majority of people or at least overestimate the popularity of our opinions. It is easy to see from this example why the FCE is an important phenomenon to study, as it affects the way that we think about society and particularly about contentious issues.

In the original study by Ross et al. (1977) they conducted four studies that each had the following underlying structure:

1. Ask the participant a binary question, i.e., if they would or would not carry out a task, or if they prefer x or y.
2. Ask the participant what percentage of other people would choose the same option as themselves.
3. Compare the actual proportion of people who chose the option with the participant's estimate.

Repeatedly, they found that participants overestimated those who would choose the same option as themselves. For example, if 30 percent of people indicated that they preferred brown bread to white, those who were part of this group might predict that 50 percent of people prefer brown bread to white – an overestimate of consensus.

They attributed this effect to several mechanisms:

1. Selective exposure to similar others
2. Self-anchoring in judgment
3. A desire to validate one's own beliefs.

Selective exposure to similar others refers to the tendency for people to compare themselves to those around them – their friends, family, and colleagues. It is not a stretch to suggest that the people that we tend to interact with more often (friends, family, colleagues) are more likely to make similar choices, hold similar opinions, and perform similar behaviours to ourselves (McPherson, 2001). This is due not only to a preference to interact with those who are similar to us, but also to the formative influences that our individual circumstances,

culture, and early childhood contribute to (de Klepper et al., 2010). This exposure to similar others may disproportionately influence our beliefs about the rest of the population and thus produce a perception of false consensus.

In a similar vein, self-anchoring in judgement is caused by the availability of our own opinion to us. When we are asked what proportion of the population prefers white to brown bread our first thought is to refer to our own preference because it is immediately available. In an absence of representative data, your own opinion is usually a decent first guess, however this strategy works much better in tight-knit communities where opinions are more likely to be similar than when referring to a larger population. This over-reliance on our own opinion is an egocentric bias that contributes to the FCE.

In addition to cognitive mechanisms, motivational processes also contribute to the false consensus effect. Research suggests that perceiving one's attitudes and behaviours as widely shared provides social validation and supports self-esteem, making such beliefs psychologically advantageous (Marks & Miller, 1987). When individuals assume their views are common, those views appear more normative and socially acceptable, which can reinforce a sense of belonging and reduce the threat of social isolation. From this perspective, overestimating the prevalence of one's own opinions is not merely a perceptual error but may reflect an adaptive tendency to maintain positive self-regard and social connectedness, both of which are central to functioning in social environments. This tendency also provides room for confirmation bias and interpretation bias to have an influence. We tend to recall conversations where people agree with us more readily than the opposite and when presented with an ambiguous statement, we are more likely to interpret it as aligned with our own stance than the contrary (Nickerson, 2024).

Further research has gone on to replicate the effect in different populations, modify the original method, develop theories around the mechanisms responsible, as well as cultivate new explanations.

**Mechanisms, Moderators, and Extensions of the False Consensus Effect**

Since Ross et al. (1977) published their paper and coined the FCE, a wealth of research has extended on their findings. For example, ten years on from their study, Marks and Miller (1987) published a meta-analysis that examined the previous decade of research on the FCE. Their analysis focused on theories of the mechanisms behind the FCE and grouped these theories into four general theoretical perspectives: a) selective exposure and cognitive availability, (b) salience and focus of attention, (c) logical information processing, and (d) motivational processes. These categories mirror the findings of the original study, bar (c), which results from finding a change in effect strength depending on if participant choices are caused by situational factors or attributed to personal dispositions. Choices dependent on personal dispositions showed a decreased estimation of consensus in comparison, which indicates that logical reasoning affects the FCE.

Furthermore, multiple studies around the same time began to look at what can moderate the size of the effect. Familiarity with the topic, social desirability, and group membership all play a part in altering effect size (Goethals, et al., 1991; Sherman et al., 1984), but increasingly it has been found that methodological considerations such as the number of estimates participants are asked to make, and the sequence of these estimates also plays a large part in moderating effect size (Mullen et al., 1985).

Another avenue explored is the impact of culture. As this effect is dependent on comparison between the participant and their societal peers, it would be expected that differences may crop up between participants from different cultures. One of the most often-studied divides in culture is that between individualist and collectivist cultures. Thus, a 2019 study out of South

Korea by Choi & Cha provides an interesting perspective. They ran studies with a similar paradigm to Ross et al. (1977) and found that evidence that the FCE is stronger in South Koreans than North Americans, particularly in the domains of political expectations and personal problems. They postulate that this difference exists because of the difference in societal expectations for social harmony and group identity in collectivist cultures. This finding also builds on a similar finding from Park (2012) that hypothesizes that need for uniqueness may be a driver behind differences in FCE between the two cultures. With individualist cultures displaying a higher need for uniqueness, thus decreasing the FCE. Gilovich (1990) investigates another perspective, the effect that construal style can have on people's estimates of consensus. This was touched on previously with regard to the IOED and could potentially provide a link between the mechanisms of the two effects. Gilovich set out to investigate the role that construal styles play in the FCE by comparing questions that allowed for different levels of subjective interpretation. In those that allowed for more subjective interpretation, overestimates of consensus were larger than for those with less room for subjective interpretation. To confirm this finding, construal style was experimentally manipulated, and the same trend was present. This leads to the conclusion that the magnitude of the FCE may depend on if participants adopt a concrete or abstract construal while they make their estimates of consensus, similar to findings reported by Alter et al. (2010) for the IOED.

Other research shows just how robust the FCE is. Krueger and Clement (1994) developed a series of studies that display a level of robustness that borders on illogical and challenges the Bayesian perspective that would predict that participants change their estimates of consensus based on statistical information. The third study in their paper demonstrated this particularly vividly. In this study, participants were given explicit statistical information about how common each response actually was before estimating the percentage of people who would

agree with them. Remarkably, even with this information at hand, participants continued to overestimate the prevalence of their own positions. Knowing the true base rates did not meaningfully reduce the bias. Instead, people appeared to rely primarily on their own responses as an anchor and adjusted insufficiently toward the provided data. Krueger and Clement argued that this pattern challenges the idea that consensus judgments operate in a Bayesian manner, since participants effectively ignored the very information that should have improved their estimates. Their findings highlight how deeply rooted the tendency to project one's own perspective is, and why consensus beliefs remain biased even when corrective evidence is directly available.

Like the IOED, this effect is particularly relevant when positioned in the context of political topics. Van Boven et al. (2018) showed that people routinely overestimate how much members of the opposing political party disagree with them on major policy issues, including climate policy. These misperceptions persist even when there is clear evidence of substantial bipartisan agreement. In essence, individuals assume greater division than actually exists, projecting their own stance—and the stance they believe their group holds—onto the broader population. This mirrors the findings observed by Krueger and Clement (1994) and demonstrates how egocentric projection can distort perceptions of political ideas. Such misjudgements can create unnecessary barriers to cooperation, reduce support for broadly endorsed policies, and reinforce the sense that political conflict is deeper and more intractable than it truly is. By illustrating how false assumptions about consensus shape political attitudes and behaviour, the work of Van Boven and colleagues underscores the real-world significance of understanding and addressing consensus-related cognitive biases.

To address these troubling implications of the FCE, some studies have set out to find ways to debias individuals. Morewedge et al. (2015) developed a game-based intervention that significantly decreased multiple cognitive biases bases, including biases related to social

judgement, even multiple weeks later. This intervention was centred around teaching people to notice when they are thinking in a biased way. Interventions like this could be used to decrease the FCE (and possibly the IOED) which in turn may play a part in decreasing political polarisation.

**Purpose of linking IOED and FCE**

Bringing the IOED and the FCE together provides a fuller picture of how people form and justify their political beliefs, and why these beliefs can become resistant to correction. Although traditionally studied as separate cognitive biases, there are several reasons to consider how they may interact. First, the two biases may be driven by similar underlying mechanisms such as egocentric anchoring, construal style, or motivational processes. If recognising the limits of one's explanatory understanding reduces perceived consensus this would indicate a shared metacognitive process. In this sense, breaking the IOED could help break the false belief that "most people think like me," revealing a potential cognitive link between the two.

Social motivations may also connect the IOED and FCE. False consensus is often interpreted as an attempt to maintain social belonging or to view one's own position as socially normative (Marks & Miller, 1987; Ross et al., 1970). Similarly, overestimating one's own understanding can serve a socially desirable function. Knowledge, especially about political issues, is treated as a marker of competence and sophistication, so overstating one's understanding may help people maintain a socially favourable self-image (Gaviria & Corredor, 2021). Both biases, then, can be shaped by motivations to appear informed and aligned with valued groups.

Linking the two biases also opens the door to interventions that influence both biases concurrently. If an intervention can increase accuracy in how people evaluate their own knowledge, it may simultaneously reduce overconfidence about how widely shared their

views are. This “two-for-one” effect is especially relevant in political contexts, where misjudging one’s own knowledge can reinforce misjudging the knowledge or beliefs of others.

The political implications of this combined miscalibration are significant. The false consensus effect can create an illusion of majority support, leading people to believe that any “reasonable” person should share their views. This is reinforced by selective exposure to like-minded individuals, especially via social media, online communities, and homogenous social circles, which amplify echo chambers. This illusion of majority support combined with an illusion of understanding works in a way that shifts those with opposing viewpoints into the “other” category. Anyone who holds opposing beliefs is perceived as a minority opinion holder, and the perception that they are not as informed follows. Van Boven et al. (2018) show that people routinely overestimate the extremity of opposing views, contributing to affective polarization—seeing political opponents as morally or intellectually inferior, rather than simply holding different perspectives. Combined IOED and FCE could be a contributing factor.

When the FCE is paired with the IOED, the risks increase. People may not only assume that many others agree with them, but also that those who agree are informed and rational, while those who disagree must be ignorant or acting in bad faith. Overconfidence in one’s own political understanding can therefore magnify the perceived righteousness of one’s position and intensify dismissiveness toward opposing viewpoints. This dynamic fosters a toxic environment for democratic discourse, where individuals become less willing to engage with disagreement or to consider the possibility that others’ perspectives might be valid.

The combination of IOED and FCE also has practical consequences. It can make individuals more susceptible to misinformation, more confident in interpreting complex political claims they do not fully understand, and more likely to act upon this misinformation. This

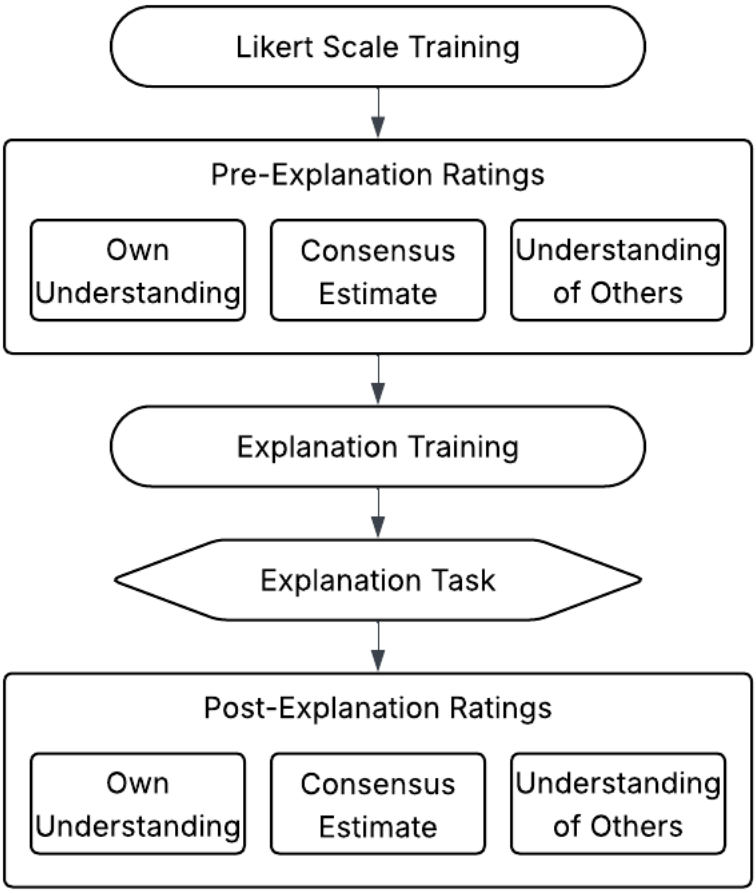
susceptibility to misinformation may come from a decreased tendency to fact-check claims, for they feel like they already understand the claim, or it feels intuitively correct due to their IOED. Or, through the lens of the FCE, people may assume that claims that they agree with (even if erroneous) are commonly held beliefs. Understanding how these two biases interact therefore offers important insights into why political misunderstandings spread so easily and why correcting them is often so challenging.

**Research Informing Design of Current Study**

The current research takes design elements from both the original IOED paradigm as laid out by Rozenblit and Keil (2002) as well as the original FCE paradigm by Ross et al. (1977). The resulting 'IOED-FCE hybrid paradigm' is summarised in Figure 1.

**Figure 1**

*The IOED-FCE Hybrid Paradigm*




*Note.* Pre-Explanation Ratings were collected for all items at the same time. Explanation Tasks and Post-Explanation Ratings were completed separately for each item.

Within this paradigm elements from both original studies are blended to create a new structure that preserves key parts of the constituent paradigms (IOED and FCE). First participants are trained on the use of the multi-point scale (taken from Rozenblit & Keil, 2002), then score their own understanding of the topic, other people's understanding of the topic, as well as produce an estimate of the proportion of people who would rate their understanding the same as they did (the latter functions as the consensus estimate measure).

These variables are the “pre-explanation ratings” seen in Figure 1. Participants are then trained to produce an explanation of the topic (Also from Rozenblit & Keil, 2002), and then proceed to do so. Post-explanation ratings of the same variables (own understanding, others’ understanding, and consensus estimate) are then collected.

**Hypotheses and Research Questions**

The following studies were designed around two primary hypotheses and a number of exploratory research questions.

Based on the IOED literature, it is expected that participant ratings of their own knowledge will decrease after being asked to produce an explanation (Fernbach et al., 2013; Rozenblit & Keil, 2002). This will be referred to as ‘the Self-IOED hypothesis’. If this happens, it indicates that participants realise the shallowness of their knowledge once asked to produce an explanation. This aligns with previous research, particularly because both of the domains in question (mechanical devices and political topics) have been replicated multiple times. There is, however, a chance that asking participants to consider the knowledge levels of others (by combining the FCE and IOED paradigms) will affect the outcome of this test because of this added variable. Unfortunately, it is not possible to test if this is the case given the within-subjects design of the study.

Based on the FCE literature, it is expected that participants will overestimate the proportion of people who rate their knowledge the same as they did. This will be referred to as ‘the FCE hypothesis’. If this happens it indicates that participants overestimate consensus on the topic of knowledge ratings. Although, the FCE has been replicated many times, this appears to be a novel domain in the FCE literature so this hypothesis is relatively exploratory. The existing research on the FCE tends to focus more on preferences, viewpoints, and opinions rather than understanding/knowledge. Whether or not the FCE extends to this domain is currently

unclear. Consensus ratings can be tested both before and after the explanation task, so it will also be possible to ascertain whether the explanation task affects the outcome.

Another question that this study will help to answer is if the IOED extends to participants' view of others' knowledge. This is the 'IOED for others' question. This question comes from the hybridisation of the IOED and FCE paradigms. If participants rate other peoples' knowledge lower after producing an explanation then it would indicate that the IOED effect (a decrease in knowledge rating after producing an explanation) does indeed extend to the perception of other peoples' knowledge. That is, does the same mechanism that tends to cause decreased knowledge ratings post-explanation for the self also cause participants to rate other people's knowledge lower as well?

An additional question raised by this research is whether individuals revise their estimates of consensus after the explanation task. For instance, if a participant estimates that 30% of people possess the same level of knowledge as themself, will this estimation change after the explanation task? If so, does perceived consensus increase or decrease, and on what basis? Further, a slightly different but closely related question asks if their estimation will become more or less *accurate* after the explanation task. This estimate of perceived understanding departs from traditional FCE research, which typically focuses on opinions and behaviours. Asking participants to estimate consensus in terms of perceived understanding therefore represents a relatively unexplored domain in the literature and offers a novel way of linking the IOED with FCE research.

# Experiment 1 - Devices

Study 1 was designed to provide a proof of concept and framework for testing the IOED and the FCE in concert and to identify any links between the two effects. To do this, we used one of the original domains from Rozenblit and Keil (2002), mechanical devices. The IOED has been replicated multiple times in this domain, making it a safe choice for replicating the IOED in this population. In contrast, and as mentioned in the previous section, it appears that the FCE has not been investigated for knowledge ratings as the current research sets out to do. This sets the stage for the findings to contain novel information about the mechanism of the FCE in this context, including if the FCE extends to this style of investigation at all.

## Method

### *Participants*

110 participants were recruited via Prolific (www.prolific.com). The only criteria these participants had to meet was that they currently reside in either Australia or New Zealand and that they are fluent in English. Participants were briefed on the study they were about to complete via an information and consent page and asked to provide consent to continue. Those that did not provide consent were automatically exited from the survey.

8 participants failed an attention check, but their answers were manually reviewed based on their engagement in the explanation questions and their data was subsequently included in data analysis. Of the 110 participants, most identified as white (75%), male (54%), and were between the ages of 19 to 87 (M = 40.41, SD = 13.9). 85 were residing in Australia, while 25 were in New Zealand. 5 of these participants missed one or more question in the survey, their data was excluded from analyses that relied on that data point but remained for the analyses that did not.

### *Materials and Procedure*

Data was collected via an online questionnaire created using Qualtrics (www.qualtrics.com) and distributed via Prolific. It was advertised to potential participants in their userbase that fit our screening criteria. The questionnaire was comprised of seventeen questions (including an attention check) that covered four items (a cylinder lock, sewing machine, helicopter, and a flush toilet). The mean time to complete was 19.4 minutes. Responses were collected between 28/06/25 and 01/09/2025. Participants were awarded 3GBP for successful completion. Responses were kept anonymous, apart from select demographic information provided by Prolific, and participants were free to opt-out at any time.

First, participants read a set of instructions (adapted from Rozenblit and Keil, 2002) on how to use the 7-point scale. After reading these instructions participants were asked to rate their own understanding and the understanding of others of each of the items on the 7-point scale. These ratings functioned as the 'pre-explanation' ratings for the IOED-type analyses. They were then asked to estimate the proportion of people who would choose the same number on the scale as themselves. This estimate provided the data for each participant's 'consensus estimate' rating. Next, the participants read another set of instructions that informed them that they would be expected to produce a causal explanation of how the item functions, and how to write this explanation in a causal manner, these instructions were also adapted from Rozenblit and Keil (2002). They were then asked to provide a causal explanation for the first item, then asked to re-rate their understanding, the understanding of others, and the proportion of people who would choose the same number as them, repeated for the remaining items. Participants then completed an attention check and were asked if they felt that their ratings changed, and if so, how and why. Figure 1 illustrates the flow of the questionnaire. The full questionnaire and consent form can be found in Appendix A.

### *Adapting FCE Measures to a Knowledge Judgement Context*

For this design to produce data that is usable for the intended analyses, participants will rate their and others' knowledge on a 7-point scale and estimate consensus by submitting an estimate of the proportion of the population that would answer the same as themselves ranging from 0 to 100 (e.g. 20 if the participant thinks that 20% of the population would answer the same as themselves). This departs from the majority of the existing literature on the FCE, where participants are instead presented with a binary choice. This introduces some challenges with replicating the effect as well as testing the hypotheses in the same way that the original FCE paper did.

The existing literature on FCE tends to give participants a choice between two options. Providing these two options gives the researchers a precise measure of the percentage of their sample that would choose each option. They can then compare this figure with each participant's estimates of the proportion of people who chose each option. This is a simple way to compare participants' estimates of consensus with the actual proportion of participants who chose each option. Because the present research will be using a multi-point scale rather than a binary, some challenges are introduced in comparing actual proportions with estimated proportions.

The knowledge estimate (IOED-style) questions could only be translated to a binary question in a couple of ways. One example would be asking "does the average person know more or less than yourself". However, asking the question this way takes away from the effectiveness with which we would be able to test the IOED hypotheses. It would provide data that is too coarse to make any meaningful inferences from and take away from any attempt to replicate the IOED. Therefore, the knowledge estimate questions remain answered by 7-point scale and the FCE analyses adapted to suit.

## Results

### *Illusion of Explanatory Depth for Self*

A paired samples t-test was conducted between participants' pre-explanation ratings of their own knowledge and their post-explanation ratings of their own knowledge.

The hypothesis that participants would decrease their ratings of their own knowledge after producing an explanation was supported in three of the four items. However, only two of these three items produced statistically significant results. Interestingly, in direct contrast to the self-IOED hypothesis, the cylinder lock item produced a strong and statistically significant effect in the opposite direction. Mean differences, effects sizes and *p*-values can be found in Table 1. Descriptive statistics can be found in Appendix B.

**Table 1**

*Differences Between Self-Rated Knowledge Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Lock | -6.98 | 109 | <.001 | -1.091 | 0.156 | -0.665 |
| Helicopter | 1.22 | 109 | 0.227 | 0.145 | 0.120 | 0.116 |
| Sewing Machine | 2.62 | 109 | 0.010 | 0.355 | 0.135 | 0.250 |
| Toilet | 3.76 | 109 | <.001 | 0.518 | 0.138 | 0.359 |

*Note*. A positive mean difference indicates a decrease in knowledge rating post-explanation.

### *Illusion of Explanatory Depth for Others*

Similar to the previous section, a paired samples t-test was conducted between participants' ratings of others' knowledge before and after producing their explanations for the items. The results of the group-IOED test are relatively similar to the self-IOED results. The same three items produced a decrease in ratings of others' understanding, while the cylinder lock item indicated the opposite effect. However, in this case all items produced results that are statistically significant ($\alpha = 0.05$). Notably, of the items that changed in the hypothesised direction (lower ratings post-explanation), there are clear differences in effect size when compared with the self-IOED results. For example, in the self-IOED analysis the helicopter item has an effect size of 0.116 compared with an effect size of 0.419 in the group-IOED analysis. Mean differences, effects sizes and *p*-values can be found in Table 2. Descriptive statistics can be found in Appendix B.

**Table 2**

*Differences Between Ratings of Others' Knowledge Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Lock | -5.04 | 108 | <.001 | -0.697 | 0.138 | -0.483 |
| Helicopter | 4.38 | 108 | <.001 | 0.486 | 0.111 | 0.419 |
| Sewing Machine | 2.05 | 107 | 0.043 | 0.231 | 0.113 | 0.197 |
| Toilet | 4.21 | 108 | <.001 | 0.486 | 0.116 | 0.403 |

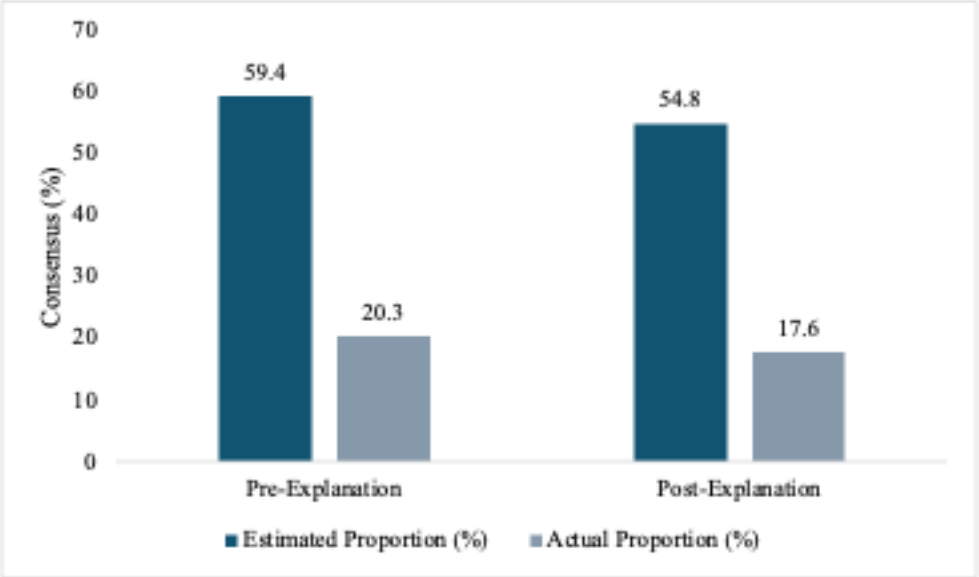
*Note.* A positive mean difference indicates a decreased knowledge rating post-explanation.

### *False Consensus Effect*

A paired samples t-test was conducted to compare ratings of perceived consensus (participants' estimates of the proportion of people who would choose the same number as themselves on the 7-point scale) and the proportion of participants who actually chose the same number as them. This was completed for all items both before and after explanations. The results of this test show a very large and statistically significant overestimate of consensus for all items both before and after explanations. Mean differences, effect sizes, and *p*-values can be found in Table 3. Descriptive statistics can be found in Appendix B.

**Figure 2**

*Average Consensus Estimate vs. Actual Proportion*




*Note.* Estimates and proportions within figure are an average across all four items.

**Table 3**

*Differences Between Estimated and Actual Consensus*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Lock (Pre) | 10.6 | 109 | <.001 | 29.5 | 2.77 | 1.01 |
| Helicopter (Pre) | 16.2 | 109 | <.001 | 43.1 | 2.66 | 1.54 |
| Sewing Machine (Pre) | 15.7 | 109 | <.001 | 36.6 | 2.33 | 1.50 |
| Toilet (Pre) | 20.7 | 109 | <.001 | 47.1 | 2.28 | 1.97 |
| Lock (Post) | 16.1 | 109 | <.001 | 36.5 | 2.27 | 1.53 |
| Helicopter (Post) | 16.5 | 109 | <.001 | 37.0 | 2.25 | 1.57 |
| Sewing Machine (Post) | 16.0 | 109 | <.001 | 34.2 | 2.14 | 1.52 |
| Toilet (Post) | 18.1 | 109 | <.001 | 41.0 | 2.26 | 1.73 |

*Note.* A positive mean difference indicates a larger consensus estimate than actual proportion.

### *Change in Consensus*

A paired samples t-test was used to compare participants' estimates of consensus before and after they produced explanations for each item. The aim of this test was to investigate whether the act of producing an explanation would influence participants' estimates of consensus.

The results of this test show that estimates of consensus decreased for all items post-explanation. However, only two items (toilet and helicopter) produced statistically significant results. Mean differences, effect sizes, and p-values can be found in Table 4. Descriptive statistics can be found in Appendix B.

**Table 4**

*Differences Between Consensus Estimates Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Lock | 1.546 | 109 | 0.125 | 3.63 | 2.35 | 0.1474 |
| Helicopter | 3.019 | 109 | 0.003 | 6.75 | 2.23 | 0.2879 |
| Sewing Machine | 0.827 | 109 | 0.410 | 1.84 | 2.22 | 0.0789 |
| Toilet | 3.080 | 109 | 0.003 | 6.05 | 1.97 | 0.2937 |

*Note*. A positive mean difference indicates a decrease in consensus estimate post-explanation.

### *Change in Consensus Accuracy*

Another paired samples t-test was used to investigate if participants provided *more accurate* estimates of consensus after producing explanations. This is a more precise indicator of whether producing an explanation reduces the FCE than simply looking at if consensus estimates increase or decrease. This was completed by computing the difference between estimates and actual proportions before and after explanations as absolute values and then comparing those pre- and post-explanation values.

Consensus estimates were more accurate post-explanation in 3 of the 4 items. However, only two (toilet and helicopter) were statistically significant. Mean differences, effect sizes, and *p*-values can be found in Table 5. Descriptive statistics can be found in Appendix B.

**Table 5**

*Differences Between Pre- and Post-Explanation Consensus Estimate Accuracy*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Lock | -0.586 | 109 | 0.559 | -1.34 | 2.29 | -0.0559 |
| Helicopter | 3.126 | 109 | 0.002 | 6.89 | 2.20 | 0.2981 |
| Sewing Machine | 1.201 | 109 | 0.232 | 2.53 | 2.10 | 0.1145 |
| Toilet | 3.060 | 109 | 0.003 | 6.20 | 2.03 | 0.2918 |

*Note.* A positive mean difference indicates a more accurate consensus estimate post-explanation.

## Discussion

This study was designed to trial the IOED-FCE hybrid paradigm, replicate the findings of Rozenblit and Keil's (2002) original paper on the IOED, extend those findings to the act of thinking about others' knowledge, examine the domain of knowledge ratings with the FCE paradigm, and investigate the relationships between the IOED and FCE. Within the results, we find answers to the two main hypotheses as well as a number of exploratory questions.

### *Illusion of Explanatory Depth for Self*

First, addressing the self-IOED hypothesis, the results show that in most cases people do tend to overestimate the depth of their own knowledge before producing an explanation, but it may depend heavily on what they are asked to explain. This is showcased clearly when comparing the mean difference of the cylinder lock item with the other three items. Analysing each item individually departs from much of the existing IOED literature which tends to examine change across several items simultaneously (Rozenblit & Keil, 2002). If this strategy was used in the present research, the cylinder lock outlier would likely affect the results in such a way that no mean decrease in post-explanation ratings would be reported ("cancelling out" the results of the other three items).

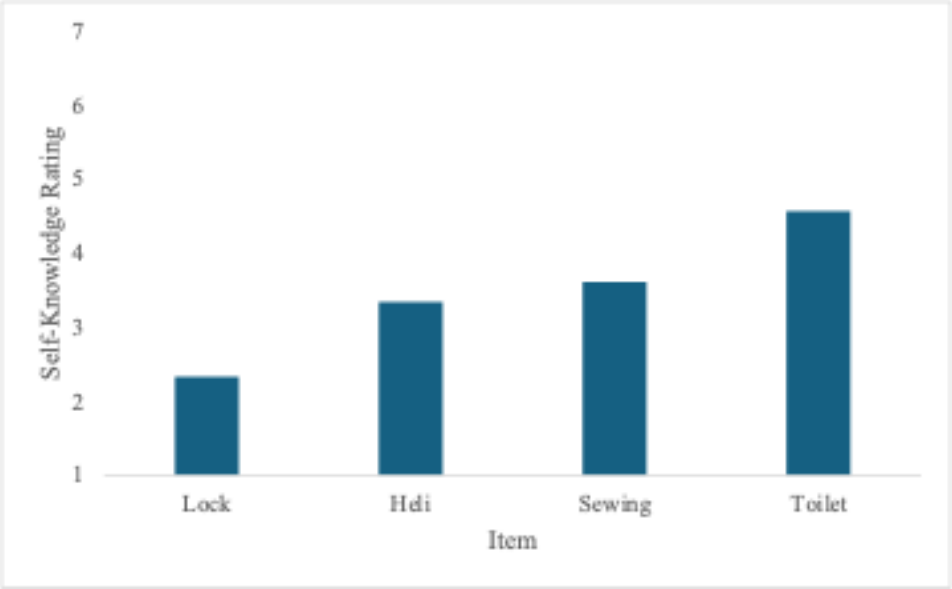
Explaining why this item was so different from the others is more complex. Three factors stand out:

1. The lock was the first item shown to participants, potentially producing a sequence effect;
2. The lock had low pre-explanation ratings compared to the other items, potentially introducing a floor effect, and;
3. The term used when referring to the lock, 'cylinder lock', may have resulted in some confusion for participants, because it is not common parlance like the other three items.

The sequence of the items shown to participants in this study was not randomised. This meant that the cylinder lock item was shown to participants first in every case. This could have introduced a confounding sequence effect because participants are naïve to the structure of the questionnaire when they are asked to produce their first explanation, but once they produce their first explanation and go on to re-rate their knowledge, they know what to expect from the next item. If this is the case, it may be true that no matter what the first item is, it may produce a slightly different result (Meyers et al., 2013). This question is addressed in Study 2 by randomising the order that each item is shown to each participant with the aim of eliminating any potential sequence effect that could have influenced the results of Study 1. Possibly the most straightforward explanation for the lock's reverse-IOED (or 'illusion of explanatory shallowness' as described by Alter et al., 2010) is that participants initially rated their understanding much lower than the other items. Mean initial ratings for the lock were 2.34 on the 1-to-7 scale, compared to a mean of 3.84 for the other three items. It is easy to see that with such low initial ratings, there is less room on the scale for participants to subsequently decrease their ratings after producing their explanations. Participants who did decrease their ratings had less room on the scale to express this, and participants who increased their ratings had more room to express this. Because of these larger increases, the mean testing may have been skewed in this direction. These low initial ratings could also have been produced by a potential sequence effect introduced by not randomising the sequence of items shown to participants, as explained above.

**Figure 3**

*Mean Initial Knowledge Ratings*




Another, less technical explanation, is simply that the term ‘cylinder lock’ is not common parlance in New Zealand and Australia. In everyday life people simply refer to a cylinder lock as a lock. Participants may not have completely understood what was meant by the term and thus their initial ratings have this uncertainty built into them. Take these excerpts from participants’ explanations for example: “I don't actually know what a cylinder lock is - hence I gave it the lowest rating” “I don’t know what a cylinder lock is at all, so I don’t feel I can answer this one sorry”. It may also be the case that as a consequence of being asked to think more deeply and produce an explanation, they subsequently understand what is meant by the term cylinder lock and so their post-explanation ratings have this realisation built into them. This would explain the large jump from 2.34 pre-explanation to 3.43 post-explanation, the largest difference of means of any of the items.

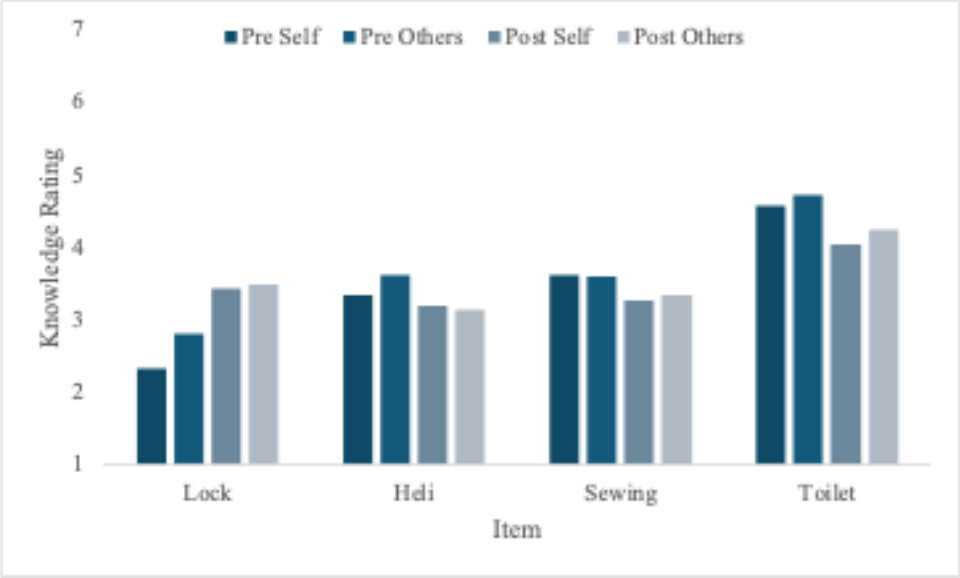
### *Illusion of Explanatory Depth for Others*

The group-IOED results effectively mirror the results for the self-IOED. The same three items show a decrease in knowledge ratings, while the cylinder lock item shows a large increase. This seems to indicate that participants base their assessments of other people's knowledge on their assessment of their own knowledge. This is in contrast to the possibility of knowledge assessments for others being distinct from assessments for the self.

This link between one's own knowledge and the knowledge of others points to an egocentric anchoring effect where participants start with their own experience as a point of reference and proceed to make assessments about others from that point (Epley & Gilovich, 2004). This idea was also proposed by FCE researchers as an explanation for why people overestimate consensus, which positions this mechanism as a possible link between the two effects (Krueger & Clement, 1994). Participants may begin their assessments from what is available to them internally (their own knowledge) then proceed to take into account other factors which result in their final assessment of others. Also, note that participants rated others' knowledge at around the same level as their own knowledge rather than higher or lower, which reinforces this idea.

**Figure 4**

*Knowledge Ratings by Item, Time, and Group*




### *False Consensus Effect*

The results show clear evidence that participants overestimated the proportion of others that share their level of knowledge for every item in the questionnaire both before and after the explanation task. This indicates that the FCE not only extends to the domain of knowledge ratings, but also that the effect can be investigated with topics that are not only binary, but exist on a continuum. In previous FCE research, binary questions like “do you prefer white or brown bread” were used to investigate this phenomenon (Ross et al., 1977), but the present research indicates that the scope of the effect can be widened to include measures of attitudes, opinions, and other topics that exist on a continuum. This not only widens the scope of the effect but allows increased nuance in investigations regarding it.

The degree to which participants overestimated consensus in their knowledge levels was surprisingly extreme. The mean overestimation of consensus across all items both pre- and

post-explanation was 38.13%. Compared to the existing FCE literature, this is an enormous overestimation. Mullen et al.'s (1985) meta-analysis of 115 studies reported an effect size of $Z = 0.3206$, compared to the average effect size of $Z = 0.711$ (converted from an average of $d = 1.546$ across items and times) found in the present study.

Within the domain of knowledge ratings, a landmark study is Kruger and Dunning's (1999) paper examining the overestimation of personal knowledge/understanding. The overestimation of consensus in knowledge ratings shown in this study seems to conflict with this finding. If people tend to overestimate their own knowledge as well as overestimate the proportion of people who are like themselves, then it follows that people have a skewed idea of the average knowledge levels of the population, leaving little room for considering those who have low knowledge levels.

One major consideration to take into account with the FCE hypothesis testing in this study is how the question was structured. To gauge participant's consensus estimates, the question they were asked was "What proportion of people would choose the same number as you on the 1-to-7 scale". It is possible that this style of question led to skewed results as it did not force participants to directly consider the proportion of participants that would choose *each* point on the scale. To elaborate, if participants' choices on the 1-to-7 scale were evenly distributed, each point would garner approximately 14% of the total participants. Considering that many participants responded that upwards of 70% of participants would choose the same number as themselves on the scale, it follows that they may not have considered the proportion of participants that would choose each other number on the scale. In fact, the average consensus estimate in this study (combined pre- and post-explanation) was 57.12%, an enormous overestimate. This was addressed in Study 2 by restructuring this question to invoke a need to consider the proportion of people who would choose each number on the scale rather than solely the number that the participant themselves chose.

### *Change in Consensus*

The notion that participants' ratings of consensus would change post-explanation appears to be supported by the results of Study 1 because all items showed a unidirectional change in consensus. After producing an explanation, participants' estimates of consensus decreased. This appears to be a novel finding that links the IOED with the FCE.

Although it may be novel, it is perhaps unsurprising when supported by the findings of Roeder and Nelson (2015) whose research supports the idea that even when participants explain an item/topic that is unrelated to the one that they are tasked with rating their understanding of, their knowledge ratings still decrease post-explanation. There is a possibility that the explanation task affects both participants' ratings of knowledge and their ratings of consensus concurrently. In this case, the consensus rating question could be seen as an analogue for the participant's own knowledge, or the knowledge of others, and is thus affected by some of the same mechanisms that produced the IOED for knowledge ratings. Even so, a hole in this theory presents itself when you consider that the lock item produced a strong 'illusion of explanatory shallowness', while also producing a decrease in consensus estimates. If the cross-domain explanation was affecting the consensus ratings in the same way it affected knowledge ratings, we might expect to see an increase in consensus estimate for the lock item, which was not found, although the result was not statistically significant in the negative direction. Instead of a direct correlation between perceived consensus and perceived knowledge (perceived knowledge and consensus estimates decrease in tandem) there may be a more nuanced link between the mechanisms of these two effects. In the case of the lock item, the explanation task producing an increase in perceived knowledge ratings while also producing a decrease in perceived consensus could instead be a result of participants realising that they know more than they thought they did (illusion of explanatory

shallowness) while also making the assumption that other people did not have this same realisation, thus leading to a decrease in perceived consensus.

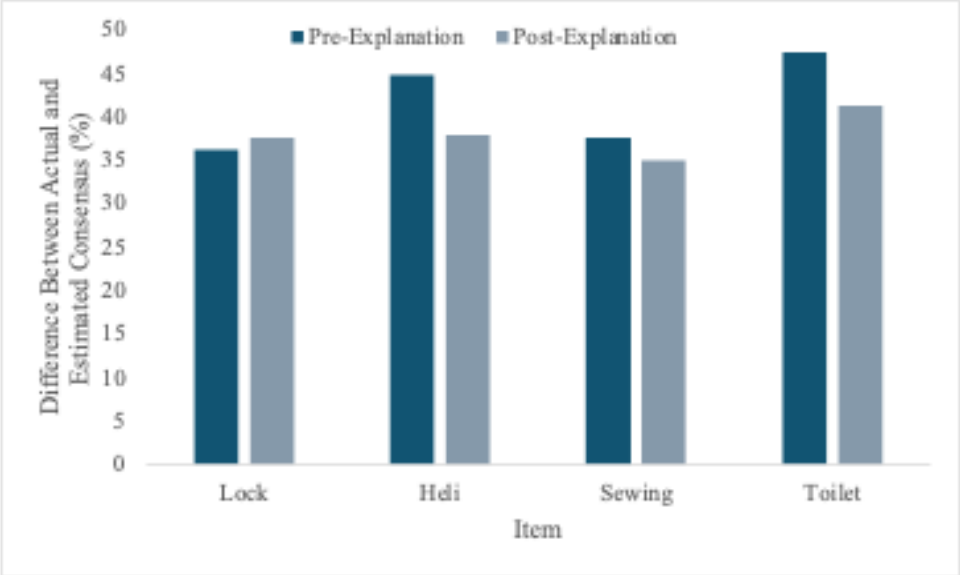
Instead, a major factor in the change in consensus results may be much simpler – cognitive recalibration and downward adjustment due to range restriction. The modest reduction in perceived consensus may reflect partial adjustment from an initially egocentric anchor rather than a deliberate re-estimation process. And, because initial consensus estimates were already high, subsequent change was constrained to downward adjustment, producing a small but relatively reliable decrease. This was briefly explored when discussing the 'illusion of explanatory shallowness' seen in the results of the lock item. Because the average pre-explanation consensus ratings were so high, even if the post-explanation consensus ratings were randomly selected, it would result in a mean decrease in consensus ratings simply because there is more 'room to move' below the pre-explanation rating point than above it.

### *Change in Consensus - Accuracy*

Rather than simply asking if the task of producing an explanation changes consensus estimates, a more meaningful question might be 'do consensus estimates become more accurate after producing an explanation?'. In this case, once again the lock item produced different results from the rest of the items. Even though all consensus estimates decreased post-explanation (which would indicate a more accurate consensus estimate), the difference between the perceived proportion of other participants who would choose the same point on the scale as themselves (participants' consensus estimates) and the actual proportion of participants who chose that point on the scale actually increased for the lock item, shown in Figure 5 below.

**Figure 5**

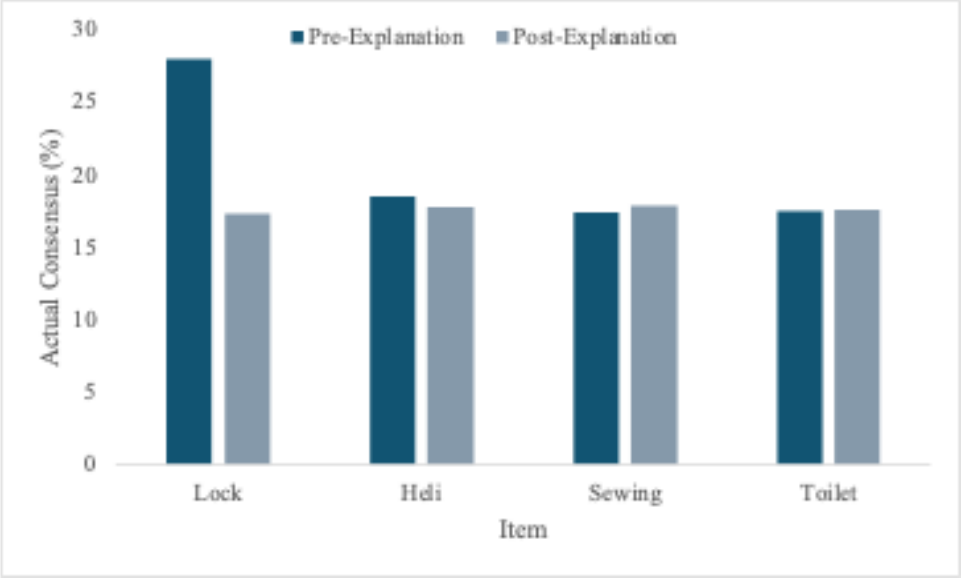
*Change in Difference between Actual Consensus and Estimated Consensus Pre- and Post-Explanation.*




This may not be a surprising result, given that the lock item has produced different results from the other three items throughout this study. However, the main factor that likely produced this result is less self-explanatory. The lock item showed a dramatic decrease in mean actual proportion (that is, the proportion of participants who chose the same point on the scale as themselves) after the explanation task compared to the other items, shown in Figure 6. This means that when calculating accuracy (by subtracting actual proportion from consensus estimate), the decrease in actual proportion resulted in a larger difference, resulting in less accurate estimates of consensus.

**Figure 6**

*Difference in Actual Proportion Pre- and Post-Explanation.*




This decrease in mean actual proportion is likely the result of a broadening of participants' knowledge ratings post-explanation. In essence, the distribution of pre-explanation ratings was skewed toward the lower end of the scale (with a mean value of 2.34), while post-explanation this distribution became more evenly distributed (with a mean value of 3.43). This in turn resulted in lower actual proportion figures and thus created the decrease in consensus estimate accuracy that we see here.

# Study 2 – Policies

Study 2 was designed to expand on the previous study in several ways. The main goal was to extend the IOED-FCE hybrid paradigm to investigate another domain that is well-researched in the IOED literature, social policies (Fernbach et al., 2013; Vitriol & Marsh, 2018). As well as this, it offered an opportunity to make adjustments to the methodology of Study 1 to improve validity of results, addressing some of the methodological shortcomings of the previous study.

## Method

### *Participants*

123 participants were recruited via Prolific (www.prolific.com). All participants were required to be fluent in English, and in contrast to Study 1, all participants resided in New Zealand to ensure topic relevancy. As with Study 1, participants were briefed on the study that they were about to complete via an information and consent page and asked to provide consent to continue. Those that did not provide consent were automatically exited from the survey.

17 participants failed an attention check, but their answers were manually reviewed based on their engagement in the explanation questions and their data was subsequently included in data analysis. 10 participants missed one or more question in the survey, their data was excluded from analyses that relied on that data point but remained for the analyses that did not. A further 11 participants responded with ratings over 89 for the consensus estimate questions, these answers were treated as outliers and their data was not used in the analyses that required these estimates.

Of the 123 participants, gender was evenly distributed between male and female (50.4% male), most identified as white (64%), and were between the ages of 19 to 78 (M = 38.94, SD = 13.59).

***Materials and Procedure***

Data was collected via an online questionnaire created using Qualtrics (www.qualtrics.com) and distributed via Prolific (www.prolific.com). It was advertised to potential participants in their userbase that fit our screening criteria. The questionnaire was comprised of twenty one questions (including an attention check) that covered four policies (recognising Palestine as a sovereign state, increasing minimum wage, enacting the Treaty Principles Bill, and imposing sanctions on Chinese imports). The mean time to complete was 36.8 minutes. Responses were collected between 24/09/25 and 11/10/2025. Participants were awarded 5.49GBP for successful completion. Responses were kept anonymous, apart from select demographic information provided by Prolific, and participants were free to opt-out at any time.

The questionnaire remained similar to that in Study 1 with some changes to question structure and minor changes to formatting. Participants first read the same set of instructions (adapted from Rozenblit et al., 2002) on how to use the multi-point scale, however the scale was changed from a 7-point scale to a 5-point scale. Participants were then asked to rate their own understanding and the understanding of others of each of the items on the scale. They were then asked to estimate the proportion of people who would choose each of the 5 points on the scale. They then read another set of instructions that informed them that they would be expected to produce an explanation of how the policy functions, and how to write this explanation in a causal manner, these instructions were adapted from Fernbach et al. (2013). They were then asked to provide a causal explanation for the first policy, then asked to re-rate their understanding, the understanding of others, and the proportion of people who would choose each point on the scale, and repeated this process for the remaining policies.

Participants then completed an attention check and were asked if they felt that their ratings changed, and if so, how and why.

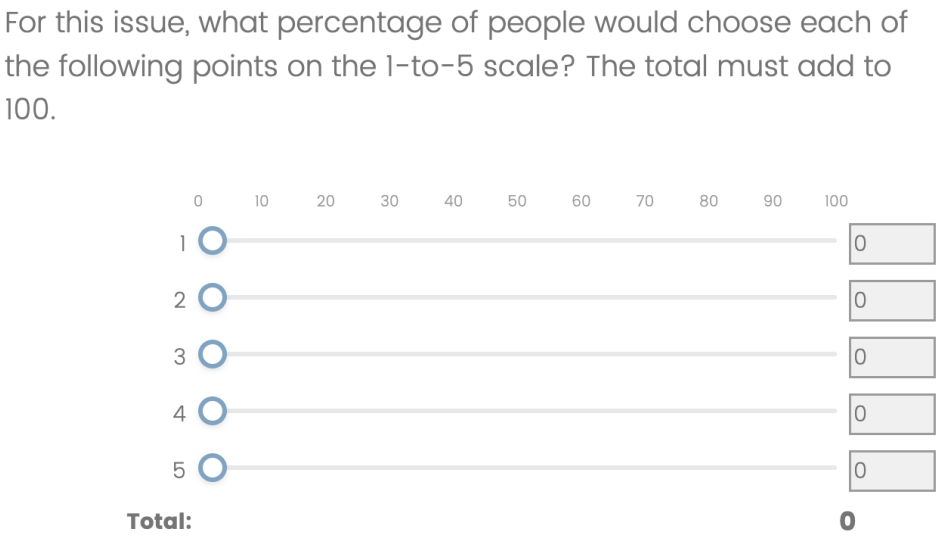
Additionally, the Generic Conspiracist Belief Scale – 5 (Kay & Slovic, 2023) was added to the end of the survey to serve other research interests not covered in this work. This is one reason for the increased time to complete and higher monetary compensation for Study 2 compared to Study 1. The full questionnaire and consent form can be found in Appendix C.

### *Modification of Consensus Measure to Reduce Extreme Responses*

As mentioned in the previous section, Study 2 incorporates some slight modifications that were aimed to increase the validity of the findings. One of these modifications was a change to how the consensus estimate question was structured. In Study 1, participants grossly overestimated the proportion of other participants who would choose the same number as themselves for the knowledge rating questions. Because of this, Study 2 uses a different method of gathering consensus estimate data. Instead of just asking participants the proportion of people who chose the same number as themselves, we asked participants to estimate the proportion of other participants who would choose each number on the scale. The rationale behind this change was to attempt to induce participants to place their estimates of proportion in the context of the scale, and was adapted from Leviston et al. (2013). The updated consensus estimate question can be found in Figure 7.

**Figure 7**

*Updated Consensus Estimate Question.*




## Results

### *Illusion of Explanatory Depth for Self*

A paired samples t-test was conducted between participants' pre-explanation and post-explanation ratings of their own understanding of the four topics.

The hypothesis that participants' ratings would decrease after producing an explanation for the topic was partially supported. Only the Palestine and minimum wage topics showed a statistically significant decrease in mean ratings. The remaining two results were not statistically significant. The Treaty Principles Bill topic showed a small decrease, and the China sanctions topic showed a very small increase. Mean differences, effect sizes, and *p*-values can be found in Table 6. Descriptive statistics can be found in Appendix D.

**Table 6**

*Differences Between Self-Rated Knowledge Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Palestine | 3.696 | 109 | <.001 | 0.2545 | 0.0689 | 0.3524 |
| Minimum Wage | 4.157 | 110 | <.001 | 0.3243 | 0.0780 | 0.3946 |
| TPB | 1.190 | 107 | 0.236 | 0.1111 | 0.0933 | 0.1146 |
| China Sanction | -0.541 | 107 | 0.590 | -0.0463 | 0.0856 | -0.0520 |

*Note.* A positive mean difference indicates a decrease in knowledge rating post-explanation.

### *Illusion of Explanatory Depth for Others*

A paired samples t-test was conducted to compare participants' ratings of other people's knowledge before and after producing an explanation for the topic.

Two of the four topics, raising minimum wage and imposing sanctions on China produced statistically significant results. However, the results are in the opposite direction of one another. For the topic of imposing sanctions on China, participants' rated other people's knowledge higher after producing an explanation, while rating participant's knowledge lower post-explanation for the raising minimum wage item. For the other two topics, Palestine showed a small decrease and Treaty Principles Bill showed practically no change. These conflicting results do not support the idea that the IOED extends to thinking about other people's understanding in the domain of social policies. Mean differences, effect sizes, and *p*-values can be found in Table 7. Descriptive statistics can be found in Appendix D.

**Table 7**

*Differences Between Ratings of Others' Knowledge Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Palestine | 1.763 | 107 | 0.081 | 0.14815 | 0.0840 | 0.1696 |
| Minimum Wage | 2.447 | 110 | 0.016 | 0.17117 | 0.0700 | 0.2323 |
| TPB | -0.112 | 106 | 0.911 | -0.00935 | 0.0835 | -0.0108 |
| China Sanction | -2.334 | 106 | 0.021 | -0.17757 | 0.0761 | -0.2257 |

*Note.* A positive mean difference indicates a decrease in knowledge rating post-explanation.

### *False Consensus Effect*

A paired samples t-test was conducted between participants' estimates of the proportion of people who chose the same number on the 5-point scale as themselves and the actual proportion of participants who chose that number.

The hypothesis that participants would overestimate this proportion was supported both before and after the explanation task. All topics showed an overestimation of consensus in knowledge ratings, however only three of the four topics were statistically significant before and after explanation. Interestingly, the topic that was not statistically significant differed pre- and post-explanation. This indicates the possibility that producing an explanation may alter how consensus is perceived for specific topics. Mean differences, effect sizes, and *p*-values can be found in Table 8. Descriptive statistics can be found in Appendix D.

**Table 8**

*Differences Between Estimated and Actual Consensus*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Pre Palestine | -2.153 | 111 | 0.034 | -3.416 | 1.59 | -0.2034 |
| Pre Minimum Wage | -0.904 | 111 | 0.368 | -1.510 | 1.67 | -0.0854 |
| Pre TPB | -2.592 | 111 | 0.011 | -4.548 | 1.75 | -0.2450 |
| Pre China Sanction | -2.822 | 111 | 0.006 | -4.503 | 1.60 | -0.2667 |
| Post Palestine | -3.252 | 109 | 0.002 | -5.144 | 1.58 | -0.3100 |
| Post Minimum Wage | -4.037 | 110 | <.001 | -6.656 | 1.65 | -0.3832 |
| Post TPB | -0.113 | 107 | 0.910 | -0.162 | 1.43 | -0.0109 |
| Post China Sanction | -2.095 | 107 | 0.039 | -3.292 | 1.57 | -0.2015 |

*Note.* A negative mean difference indicates a larger consensus estimate than actual proportion.

### *Change in Consensus*

A paired samples t-test was conducted between participants' consensus estimates before and after producing an explanation for each topic. This was conducted to investigate whether producing an explanation for the topic affected how people perceived consensus.

None of the four topics resulted in a statistically significant difference in mean consensus estimate. Consensus estimates increased for some topics and decreased for others. Mean differences, effect sizes, and *p*-values can be found in Table 9. Descriptive statistics can be found in Appendix D.

**Table 9**

*Differences Between Consensus Estimates Pre- and Post-Explanation*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Palestine | -1.492 | 109 | 0.139 | -2.74 | 1.83 | -0.1422 |
| Minimum Wage | 1.248 | 110 | 0.215 | 2.26 | 1.81 | 0.1184 |
| TPB | 1.311 | 107 | 0.193 | 2.44 | 1.86 | 0.1261 |
| China Sanction | -0.997 | 107 | 0.321 | -1.54 | 1.54 | -0.0959 |

*Note.* A positive mean difference indicates a decrease in consensus estimate post-explanation.

### *Change in Consensus Accuracy*

A paired samples t-test was also conducted to test the accuracy of these estimates. This was completed by computing the difference between participants' consensus estimates and actual proportions before and after the explanation task.

As with the previous section, results were not statistically significant for any of the four topics. Noting the small mean differences, it cannot be reliably concluded that participants were more or less accurate after producing an explanation for the topics. Mean differences, effect sizes, and *p*-values can be found in Table 10. Descriptive statistics can be found in Appendix D.

**Table 10**

*Differences Between Pre- and Post-Explanation Consensus Estimate Accuracy*

| | Statistic (Student's t) | df | p | Mean difference | SE difference | Effect Size (Cohen's d) |
|---|---|---|---|---|---|---|
| Palestine | -1.1629 | 109 | 0.247 | -1.5700 | 1.35 | -0.11088 |
| Minimum Wage | -0.9660 | 110 | 0.336 | -1.3136 | 1.36 | -0.09169 |
| TPB | 1.6363 | 107 | 0.105 | 2.3059 | 1.41 | 0.15745 |
| China Sanction | -0.0257 | 107 | 0.980 | -0.0314 | 1.22 | -0.00247 |

*Note.* A positive mean difference indicates an increase in estimate accuracy.

## Discussion

This study was designed to further test the IOED-FCE hybrid paradigm, expand on the findings of the first study with some methodological changes, and extend the findings to a separate domain, social policies. The results provide answers, some clearer than others, to the two main hypotheses as well as some exploratory questions. The findings also prompt further questioning around the differences between cognition of policies as opposed to devices.

### *Illusion of Explanatory Depth for Self*

Although an IOED was not found for all four policies in Study 2, the general trend indicated that participants did decrease their ratings of their own understanding after producing an explanation. This was seen particularly with the Palestine and minimum wage topics, but a small (non-significant) decrease was also observed with the Treaty Principles Bill topic. The topic of introducing sanctions on Chinese imports showed practically no change.

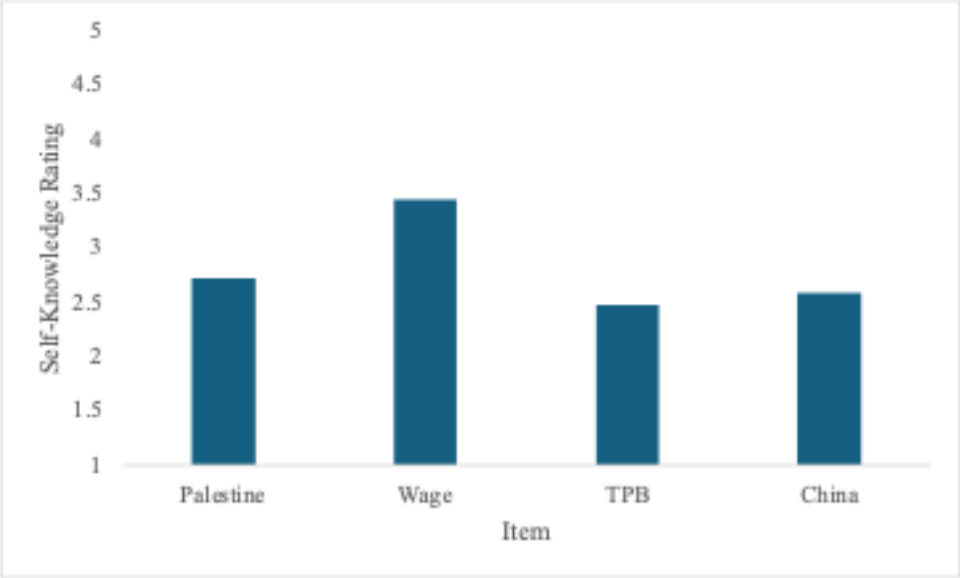
Unlike Study 1, there was no topic that produced an 'illusion of explanatory shallowness'. This may lend some credibility to the idea that the cylinder lock item was an outlier, either due to the reasons discussed prior or to other, unknown reasons. Because the sequence that

topics were shown to participants in this study were randomized the sequence effect should play no part. However, there may still be a similar effect in terms of common parlance. The idea of imposing sanctions on Chinese imports may be a relatively unfamiliar topic to New Zealanders, and even those in the rest of the world. Internationally, this was brought to light when the Trump administration enacted steep tariffs on many countries, including China. A recent polling report shows that only 45% of Americans know how tariffs work (IPSOS, 2024). If this familiarity with the concept of tariffs is consistent in the present sample, then a similar effect may be present as with the cylinder lock. There is little one can do to combat this lack of familiarity (aside from using more familiar topics) because any attempt to provide information about the term would interfere with participants pre-explanation ratings, as well as their ability to explain the topic independently.

Consistent with Study 1, the items that showed little or no decrease in knowledge ratings (TPB and China) also had the lowest pre-explanation knowledge ratings. This could have resulted in a floor effect (or reduction in subsequent rating range) similar to that which was present with the cylinder lock item in Study 1. However, mean ratings for the Palestine topic were not substantially higher than these two items, so it is difficult to conclude that a floor effect was the driving force behind the lack of IOED for these two items. Mean pre-explanation ratings for TPB and China were 2.47 and 2.59 respectively, while Palestine was 2.71 and minimum wage was 3.45. It is likely the case that examining this effect with more familiar topics that garner higher pre-explanation ratings would result in a more pronounced IOED.

**Figure 8**

*Pre-Explanation Self-Knowledge Ratings for Each Item*




*Note.* Pre-explanation ratings for TPB and China were the lowest, but not substantially lower than for the Palestine topic.

Another potential explanation for the weaker and less consistent IOED effects in Study 2 concerns the differences between mechanical systems and political issues. Much of the original IOED research used mechanical devices (e.g., locks or toilets), where understanding can be evaluated in terms of clear, step-by-step causal mechanisms (Rozenblit & Keil, 2002). In such cases, explanatory failure is more readily discoverable: attempting to describe how a device works readily exposes gaps in knowledge. Political topics, however, involve more diffuse, multi-level, and value-laden causal structures that are often not reducible to a single linear mechanism. Issues such as sanctions or treaty principles depend on institutional processes, economic trade-offs, and normative judgments, making them less amenable to purely mechanistic explanation. Although prompting explanation can reduce perceived understanding in political domains (Fernbach et al., 2013), the effect may depend on whether participants engage in genuine causal analysis rather than simply articulating general

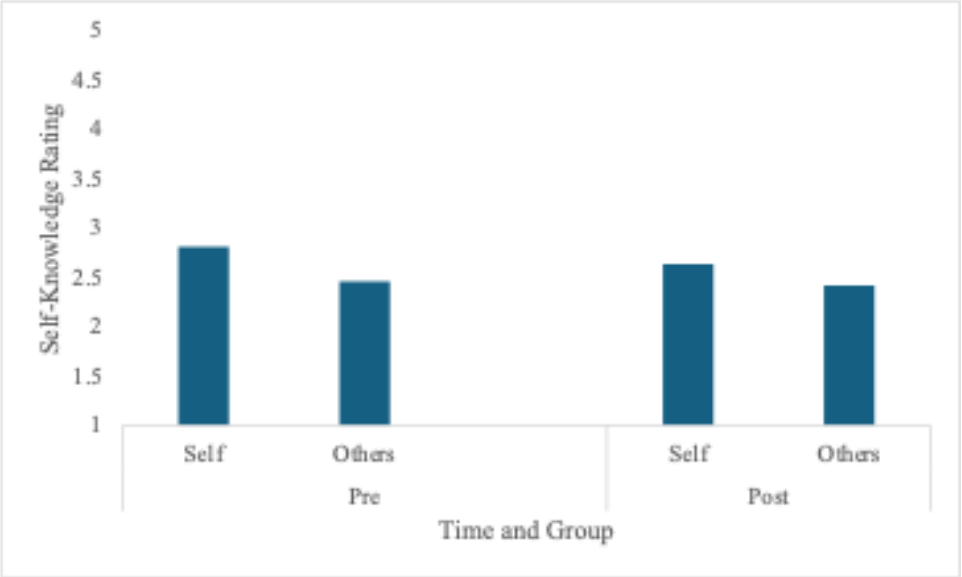
positions. If participants defaulted to surface-level rationales, the explanation task may have been less effective at exposing knowledge gaps, thereby reducing the observed IOED.

### *Illusion of Explanatory Depth for Others*

The ratings of others' knowledge followed the same pattern as the results for self-ratings in the sense that the ordering of effect sizes for each topic was consistent between the two tests. However, the magnitude of these differences is inconsistent. While there was a smaller IOED for the minimum wage topic, there was a more pronounced illusion of explanatory shallowness for the Chinese sanctions topic. It appears that participants shifted their ratings downward slightly across the board when compared to their ratings for self. This supports the idea that the IOED may extend to thinking about others' knowledge by way of egocentric anchoring, but the effect is decidedly less powerful (Epley & Gilovich, 2004). This is inconsistent with the findings of Study 1 where knowledge ratings for others were of a similar magnitude as self-ratings.

**Figure 9**

*Downward Shift of Knowledge Ratings Between Self and Others*




Participants consistently rated others' knowledge lower than their own both before and after explanation. This contrasts Study 1 where self and other ratings were very similar. One possible reason for this difference is that political policies may invite a stronger *above-average effect,* where people see themselves as more knowledgeable than the average person (Alicke & Govorun, 2005). This tendency is especially common in areas that are vague or open to interpretation. Political issues such as sanctions or treaty principles are broad and complex, and there is often no clear standard for what counts as fully understanding them. This ambiguity may make it easier for participants to view their own knowledge generously while assuming that others know less. In contrast, mechanical topics provide clearer benchmarks for understanding, which may limit the extent to which people can see themselves as especially knowledgeable. Research suggests that self-evaluations are more inflated when the topic is abstract or difficult to verify (Dunning, Meyerowitz, & Holzberg, 1989), which may help explain the stronger self–other gap observed for political items in Study 2.

The differences between the policy and device items may also relate to the role of social desirability in social comparison, as discussed by Moore (2007). Moore argues that above-average effects are strongest when a trait is seen as common and socially desirable, and can even reverse when a behaviour is rare or undesirable. In the context of political knowledge, being well-informed may be viewed as socially desirable, which could motivate participants to see themselves as more knowledgeable than others. Mechanical devices, by contrast, are less tied to identity or social value, reducing the incentive for self-enhancing comparisons. Thus, political topics may amplify socially desirable self-perceptions, contributing to the larger self–other gap observed in Study 2.

### *False Consensus Effect*

As with Study 1, participants overestimated consensus both before and after the explanation task for all items. This provides further evidence that the FCE can be explored by using non-binary measures, and also that it extends to the domain of knowledge ratings.

There was, however, a clear and substantial decrease in the level at which participants overestimated consensus compared to Study 1. One major reason for this is the change in question structure between the two studies. Study 1 used the question "what proportion of other people would choose the same number as you on the 1-to-7 scale" while Study 2 asked participants to estimate what proportion of people would choose *each* of the points on the scale. This change in question structure forced participants to not only estimate the proportion of people who are similar to them, but also the proportion of people who are not. This deeper level of thinking about consensus may have been a large driving factor behind the decreased overestimations of consensus. This change in question structure also precludes the results of this test from being directly compared to the results of Study 1, so conclusions of differences between devices and policies cannot be drawn directly.

However, given the increased social desirability of political knowledge as discussed in the previous section, one can imagine that consensus overestimation for social policies would be similar if not more pronounced if measured with the same question structure.

### *Change in Consensus*

Whether or not the explanation task changed participants' estimates of consensus is difficult to determine. None of the four political topics showed a statistically significant change, and of the mean differences that were observed, some increased while others decreased. This contrasts with Study 1, where consensus estimates for mechanical devices consistently decreased following the explanation task. For the device items, attempting a detailed mechanistic explanation appeared to reduce not only participants' confidence in their own understanding but also their belief that others shared the same level of knowledge as themselves. In other words, as participants recognized gaps in their own knowledge, they seemed to revise downward their assumptions about how common that knowledge was. The absence of a similar pattern for political topics suggests that the processes underlying consensus judgments may differ across domains. While explanation failure for mechanical systems may directly undermine both self-assessed knowledge and perceived normality, political issues are more socially embedded. Participants may treat consensus judgments about policies less as estimates of shared factual understanding and more as reflections of general opinion or group alignment. As a result, even if self-rated understanding decreases slightly after explanation, perceived consensus may remain relatively stable because it is anchored in broader social beliefs rather than purely in assessments of causal knowledge.

### *Change in Consensus Accuracy*

As with Study 1, a further analysis examined whether the explanation task improved the *accuracy* of participants' consensus estimates, rather than simply shifting them up or down.

That is, we assessed whether post-explanation estimates were closer to the actual distribution of self-ratings within the sample. This analysis was largely inconclusive. Because there were no consistent directional changes in consensus estimates across topics, there was likewise no clear improvement in accuracy following explanation. In some cases, estimates moved slightly closer to the observed values, while in others they moved further away, and none of these changes were statistically reliable. This pattern mirrors the earlier findings: without a systematic reduction in perceived consensus, there was little basis for expecting a consistent improvement in calibration. Taken together, the results suggest that, for political topics, the explanation task did not meaningfully recalibrate participants' beliefs about how typical their own level of knowledge was relative to others.

# General Discussion

The central aim of the present research was to examine how people evaluate their own understanding, the understanding of others, their perceptions of consensus, and whether these judgement processes influence each other. By merging methodologies used to investigate the Illusion of Explanatory Depth (IOED; Rozenblit & Keil, 2002) and the False Consensus Effect (FCE; Ross et al., 1977), these studies sought to determine whether the act of producing an explanation would also lead to recalibrations in consensus estimation and perceptions of others' knowledge. The two studies conducted here contribute to ongoing debates about the cognitive basis of explanatory overconfidence, the generality of the FCE, and how individuals form assumptions about shared knowledge in both mechanical and political domains.

## Summary of Key Findings

The findings across both studies paint a complex picture. In Study 1, the IOED (for participants' own knowledge) was replicated relatively robustly with familiar mechanical devices, while in Study 2 this effect was shown to be less consistent with political topics. This pattern aligns with original suggestions by Rozenblit and Keil (2002) that the IOED is strongest for coherent mechanical systems where people hold "sparse but causally rich" mental models. When participants attempted to explain how a toilet or helicopter functions, their explanatory gaps were quickly revealed to them, resulting in decreased self-knowledge ratings. Mechanical devices typically have relatively clear causal chains and a definable end point, so an explanation can be evaluated in terms of whether it successfully traces how inputs produce outputs. Political topics, by contrast, often lack a single, bounded mechanism. Their causal pathways are diffuse, and intertwined with values, institutions, and competing interpretations, making it less obvious when an explanation is complete. As a result,

explanations of political topics, which are often represented as abstract principles rather than detailed causal systems, produced less reliable downward shifts in perceived understanding. Across both studies, the illusion of explanatory depth appeared to extend beyond judgments of one's own knowledge to judgments about others, but the strength and pattern of this extension differed by domain. In Study 1, the IOED *for others* closely mirrored the IOED *for self* for mechanical devices: the same items that showed decreases in self-rated understanding also showed decreases in ratings of others' knowledge, and overall levels were relatively similar. This suggests that participants relied heavily on their own knowledge as an anchor when estimating others' understanding, consistent with egocentric anchoring accounts (Epley & Gilovich, 2006). In Study 2, a similar ordering of effects was observed across political topics, indicating some shared underlying process, but the magnitude was less consistent and participants systematically rated others as less knowledgeable than themselves. This self–other gap points to additional influences beyond anchoring, such as above-average effects and social desirability concerns, which may be more pronounced for abstract, socially valued domains like politics (Alicke & Govorun, 2005). Together, these findings suggest that while egocentric anchoring may underlie judgments about others' knowledge in both domains, political topics introduce social-comparative dynamics that weaken the symmetry between self and other observed for mechanical devices.

Across both studies, clear evidence of the FCE emerged, with participants consistently overestimating the proportion of others who shared their level of knowledge. This extends the findings of past literature that demonstrate that people overestimate similarity of attitudes and behaviours (Mullen et al., 1985; Ross et al., 1977). In Study 1, this overestimation was particularly striking for mechanical devices, with very large mean differences and effect sizes relative to the broader FCE literature. These findings not only demonstrate that the FCE extends beyond binary choices to continuous knowledge scales, but also suggest that

consensus overestimation in the domain of perceived understanding can be substantial. However, the magnitude of the effect in Study 1 was likely amplified by the structure of the consensus question, which asked participants only to estimate how many others would select the same response as themselves. This format may have encouraged egocentric responding without requiring participants to consider the full distribution of possible answers. In Study 2, participants again overestimated consensus for political topics, indicating that the FCE generalises to this domain as well. However, the magnitude of overestimation was markedly reduced. This decrease is most plausibly explained by the revised question structure (adapted from Leviston et al., 2013), which required participants to estimate the proportion of people selecting each point on the scale, thereby encouraging a more distributional and less self-focused assessment. Because of this methodological change, direct comparisons between devices and policies should be made cautiously. Nonetheless, the persistence of overestimation across both domains suggests that egocentric anchoring in consensus judgments is robust, even if its magnitude is dependent on how consensus is elicited and to the social meaning attached to the topic.

Across the two studies, the relationship between explanation-induced recalibration and consensus estimation appears again to be domain-sensitive. In Study 1, the consistent post-explanation decreases in consensus estimates for mechanical devices suggest that attempting a mechanistic explanation may weaken both confidence in one's own understanding and the assumption that others share that understanding. This pattern is consistent with an egocentric anchoring account: if participants begin by using their own perceived knowledge as a reference point for estimating others, then any reduction in self-assessed understanding may naturally translate into lower perceived consensus (Epley & Gilovich, 2006). However, this link was not perfectly symmetrical, as shown by the cylinder lock item, indicating that the

connection between IOED and FCE is not purely mechanical and may depend on additional factors such as initial scale position or cognitive recalibration effects.

In Study 2, this coupling between explanation and consensus was not reliably observed. Political topics did not show consistent downward shifts in perceived consensus, even when modest decreases in self-rated knowledge occurred. This suggests that consensus judgments in political domains may be less tightly tethered to one's momentary sense of explanatory ability. Instead, they may draw more heavily on broader social beliefs about how common certain views or levels of knowledge are, which are less directly disrupted by a single explanatory attempt. Taken together, the findings imply that while IOED-style recalibration can influence consensus estimates under some conditions, particularly for mechanical systems, the two effects are only partially linked and may diverge when judgments concern socially embedded, identity-relevant topics.

**Interpreting the IOED Across Domains**

The domain-specific pattern of both IOED effects (for self and for others) provides a conflicting contribution to ongoing debates about the nature of the IOED. Although the effect has been extended to political topics, this was not as reliably observed in Study 2. Prior work on the topic and in the wider literature has emphasized that people rely on mental schemas and 'causal skeletons' that feel complete but lack sufficient detail to support a full explanation (Ahn & Kalish, 2000: Keil, 2006). These 'causal skeletons' are akin to Rabb et al.'s (2019) description of Type 2 knowledge discussed earlier. Mechanical devices, in particular, tend to activate such schemas because people interact with everyday devices regularly and form theories about their operation, even if these theories lack technical accuracy and detail. Alter et al. (2010) specifically posits that this lack of detailed knowledge is due to a difference in construal style, with more concrete thinking leading to more detailed knowledge and explanatory ability.

Social policies, on the other hand, are typically stored as semantic summaries rather than as mechanistic models (Baron, 2008; Furnham, 1988). Representations of social policies often involve institutional processes and normative or moral judgements rather than observable causal chains. This likely contributes to the smaller decrease in knowledge ratings observed in Study 2. When people hold only a superficial representation of a policy, their attempts to produce a causal explanation may not expose the same gaps simply because they already sense that their understanding is limited. In this case, the illusion of understanding may be less present because people are more accurately estimating their (lack of) knowledge before attempting to produce an explanation. This effect may be particularly pronounced for unfamiliar topics, where participants already begin with relatively low self-ratings of knowledge, thereby limiting the range available for further downward adjustment after the explanation task.

The decreased IOED for political topics may also reflect the influence of identity-protective cognition (Kahan, 2017). Research in political psychology shows the people often treat policy attitudes as expressions of group identity rather than knowledge-based positions (Bolsen & Druckman, 2015; Kahan, 2013). When the demand for an explanation of the topic threatens identity-consistent beliefs, individuals may resist revising their perceived understanding even when confronted with the gaps in their knowledge.

**Understanding FCE in Knowledge Judgements**

The FCE findings in both studies were consistent with the large body of social-psychological research demonstrating that people are anchored by their own internal perceptions when estimating what others think, know, or believe (Krueger & Clement, 1994; Nickerson, 1999). Even when participants rated their own knowledge as low, they still assumed that others' knowledge levels fell near their own ratings. This corroborates theories of egocentric-anchoring in judgements, where the self acts as an important cognitive reference point from

which judgements begin, followed by insufficient reasoning-based adjustments, which leads to inflated perceptions of similarity (Epley & Gilovich, 2006).

The strong FCE observed in Study 1 may be also partially explained by insufficiently calibrated base-rate reasoning (Bar-Hillel, 1980). People often underestimate variability in the population and assume distributions are narrower than they truly are (Krueger & Dunning, 1999). The redesigned consensus measure in Study 2 forced participants to consider the full distribution of responses, reducing the magnitude of the FCE but not eliminating it. This indicates that the FCE in knowledge judgements is both robust and sensitive to task framing, consistent with meta-analyses showing that question structure influences FCE magnitude (Marks & Miller, 1987).

The persistence of the FCE even after the explanation task provides evidence against the theory of metacognitive 'spillover' which proposes that calibrating one's own understanding automatically generalizes to social judgement. Instead, these findings support the notion of independent processes affecting the IOED and FCE.

**Reconsidering IOED for Others as a Consensus-Adjacent Factor**

An important implication of the present findings is that judgement of others' knowledge may be understood not only as an extension of the IOED, but also as a consensus-adjacent form of social judgement. The "other IOED" measure asks participants to estimate how much the average person knows in absolute terms, whereas the consensus measure asks how many others are similar to oneself. Despite this difference, both require participants to infer what others know before and after producing an explanation. They therefore reflect closely related, but distinct forms of social and metacognitive judgement.

Considering these measures together clarifies several patterns. In Study 1, ratings of others' knowledge largely mirrored ratings of the self. Mechanical items that showed a self-IOED also showed decreases in perceived knowledge of others, consistent with egocentric

anchoring (Epley & Gilovich, 2006). Once participants encountered limits in their own understanding, they revised their assumptions about others accordingly. However, this pattern did not fully extend to consensus estimation. In Study 2, "other IOED" ratings changed for some policy items even when consensus estimates did not. Participants lowered ratings of others' knowledge for minimum wage and marginally for Palestine, while increasing them for Chinese sanctions, yet no corresponding consensus changes reached significance. This divergence suggests that judgements about others' knowledge and judgements about similarity to the self are related but separate. Participants may revise their view of the average person's knowledge without revising their assumptions about consensus.

This distinction can be thought about in terms of absolute versus relative judgement. Consensus should not change simply because self- and other- knowledge ratings change. If both are decreased in tandem, the relative relationship between them remains constant, and perceived consensus should remain consistent. With this in mind, the absence of consensus change does not indicate a lack of recalibration, but rather that participants updated their beliefs in lockstep: "I know less than I thought, but so do others". Consensus should only shift when the perceived gap between self and others changes.

This is particularly relevant in the political domain. In Study 2, participants tended to rate their own understanding as higher than that of others, while still overestimating how many others shared their level of understanding. This shows a potentially contradictory stance. Participants appear to believe that they are simultaneously more knowledgeable than average, and that many others are broadly similar to them. One interpretation is that political reasoning appears to accommodate two narratives: epistemic superiority ("I understand more than others") and epistemic populism ("people like me are common"). Although not fully coherent, these beliefs may persist because political judgement are shaped not only by explanatory overconfidence, but also by identity-related factors (Kahan, 2013, 2017).

Taken together, the mismatch between "other IOED" and consensus findings suggests that explanation-induced recalibration does not operate on a single social judgement process. Instead, it appears to affect at least two partially distinct processes, one concerning perceived competence of others, and another concerning perceived similarity to self. Mechanical topics may align these processes more closely, whereas political topics appear to decouple them. More broadly, these findings suggest that epistemic humility does not necessarily translate to social humility. Recognising the limits of one's own understanding does not automatically alter beliefs about how widely one's perspective is shared.

**Why the IOED-Style Explanation Task did not Reliably Influence FCE**

The inconsistent interaction between the IOED task and the FCE findings highlights important theoretical differences. Many models of the IOED propose that encountering difficulty in producing an explanation surprises people with the lack of their own knowledge, leading to an internal shift in their own perceived understanding (Keil, 2003) However, adjusting one's own perceived understanding does not necessarily prompt a corresponding shift in beliefs about others. This could reflect a rift between self-knowledge and social projection processes.

An alternative explanation involves motivational factors in social judgment. In political domains, individuals may anchor consensus estimates not only on perceived understanding, but also on what they wish were true or what they believe their in-group values. Research on conspiracy theories and politically charged belief systems suggests that people selectively endorse and project beliefs that align with their social identity and desired worldview, even when objective evidence is weak or absent (Douglas et al., 2017). Similarly, work on expressive responding shows that individuals may provide responses that signal group commitment or ideological preference, rather than purely reflecting factual beliefs (Ross & Levy, 2023). Such motivated projection and expressive responding can override

metacognitive recalibration processes that might be triggered by an explanation task. As a result, participants may maintain or adjust their consensus estimates in ways that serve social or identity motives, rather than converging toward more accurate estimates of how many others share their views. This motivational anchoring may therefore limit the influence of IOED-style explanation tasks on false consensus effects in politically charged contexts.

Individual differences may also help explain variability in consensus estimates, particularly in politically charged domains. One relevant factor is *need for uniqueness*, the motivation to see oneself as distinct from others. People with higher need for uniqueness tend to perceive their attitudes, preferences, and knowledge as less widely shared (Stern, West, & Schmitt, 2014). Conversely, those lower in need for uniqueness are more inclined to assume similarity, yielding stronger false consensus effects. Park et al. (2012) similarly demonstrate that uniqueness motivations shape how people infer social norms and generalise from their own beliefs. In the present studies, participants with strong uniqueness motives may have been less willing to assume high levels of shared understanding, especially for political topics, whereas those with weaker uniqueness motives may have defaulted more strongly to egocentric projection. Although need for uniqueness was not measured here, the asymmetry observed in consensus ratings—particularly in Study 2—may partly reflect such motivational influences.

Methodological explanations are also viable for explaining the difference in consensus estimate change. Study 1 used a less structured consensus question than Study 2, which may have encouraged stronger egocentric anchoring. Research on anchoring and adjustment suggests that when individuals generate estimates without explicit structural guidance, they tend to rely heavily on readily available internal information—such as their own position—and adjust insufficiently from that anchor (Epley & Gilovich, 2004; Tversky & Kahneman, 1974). By asking participants only what proportion of others would choose the same number

as themselves, Study 1 may have promoted this self-based anchoring, inflating baseline consensus estimates and leaving greater room for downward adjustment following explanation.

In contrast, Study 2 required participants to estimate the proportion of respondents at each point on the scale, which likely prompted a more distributional and analytic mode of thinking from the outset (or a more concrete construal style). Work on cognitive reflection and deliberative processing suggests that increased task structure can reduce reliance on intuitive heuristics and egocentric projection (Frederick, 2005). If participants in Study 2 engaged in more deliberate estimation initially, their consensus judgments may have been better calibrated prior to the explanation task, thereby reducing the likelihood of further recalibration afterward. Under this account, the reduced post-explanation change in political topics may reflect differences elicited from question type rather than differences in underlying psychological processes.

**Limitations**

Several important limitations should be considered when interpreting the findings of the present research. First, methodological differences between Study 1 and Study 2 restrict direct comparisons between mechanical devices and political topics. Most notably, and as discussed prior, the structure of the consensus estimation question changed between studies. Study 1 asked participants only what proportion of others would select the same response as themselves, whereas Study 2 required participants to estimate the proportion selecting each point on the scale. Although this revision was intended to encourage more accurate, distributional thinking, it likely reduced egocentric anchoring at baseline. As a result, differences in post-explanation change between devices and political topics cannot be attributed solely to domain differences as they may also reflect differences attributable to question structure. This change in question structure also prompts questioning around which

of the two questions is the most precise analogue of how people think about consensus naturally. Perhaps the distribution-type question is not as similar to how people naturally come to consensus estimates as the original question in Study 1.

Second, variation in topic familiarity and subsequent baseline knowledge ratings complicates interpretation of IOED findings. This was demonstrated clearly with the cylinder lock item in Study 1, where participants were not familiar with the term used and subsequently produced an abnormal 'illusion of explanatory shallowness'. Incorporating items that participants are familiar with appears to be essential to the IOED effect, and more care should be taken to select items/topics that participants are familiar with. This also became apparent in Study 2 with the Chinese sanctions topic, where low initial knowledge ratings restricted the range available for downward adjustment, thus producing negligible IOED effects.

An additional limitation concerns the absence of coding for explanation quality and content. Although participants completed the explanation task, the depth, coherence, and accuracy of those explanations was not formally assessed. Prior research suggests that the magnitude of the IOED can depend on the extent to which an explanation attempt genuinely exposes gaps in causal understanding (Fernbach et al., 2013; Rozenblit & Keil, 2002). Individuals who produce more detailed or mechanistically structured explanations may experience stronger recalibration than those who provide brief or superficial responses. Without coding explanation quality or accounting for individual differences in explanatory depth, it is difficult to determine whether weaker effects—particularly in the political domain—reflect true domain differences or variability in how thoroughly participants engaged with the explanation task.

Taken together, these limitations suggest taking caution in drawing strong conclusions about domain differences in the relationship between the illusion of explanatory depth and the false consensus effect. Future research using consistent measurement formats, more tightly

controlled topic familiarity, and assessments of explanation quality would help clarify the mechanisms underlying these effects.

### Future Research

Future research should examine the interaction between the IOED and FCE using consistent measurement formats across domains and more tightly controlled levels of item familiarity. Employing identical consensus estimation procedures for both mechanical and political topics would allow clearer conclusions about whether domain differences are substantive or methodological. Coding explanation quality and content would also help to clarify whether deeper, more mechanistic explanations produce stronger recalibration effects for both self-assessed knowledge and consensus estimates. Additionally, although difficult to accomplish, incorporating objective knowledge measures would allow researchers to determine whether explanation tasks improve calibration accuracy or merely reduce confidence.

Further investigation of political topics is particularly warranted, as these domains appear to engage different or additional cognitive and motivational processes than mechanical systems. Understanding how explanatory overconfidence and consensus overestimation operate in politically charged contexts has important implications for misinformation, polarisation, and public policy understanding. Exploring the role of identity, social desirability, and motivated reasoning in moderating explanation-induced recalibration may provide a more complete account of when and why these effects interact.

### Broader Implications

Previous research shows that when individuals are forced to explain policies in detail, their confidence in the depth of their knowledge decreases (Fernbach et al., 2013). But the present findings indicate that this calibration does not necessarily extend to beliefs about others' understanding. This distinction is important in light of van der Linden's (2015) gateway

belief model, which proposes that perceptions of consensus (among experts) can act as a central driver of attitude change. If perceived consensus acts as a gateway belief that shapes broader political attitudes and policy preferences, then interventions that only recalibrate self-assessed knowledge may have limited impact unless they also change assumptions about how widely shared this knowledge is. In other words, reducing overconfidence in one's own understanding may not meaningfully impact polarisation or position extremity if perceptions of consensus remain intact. These findings suggest real-world interventions aimed at reducing political extremism via explanation prompts (like those found in Fernbach et al., 2013) may be constrained unless they also directly address other (non-explanation related) misperceptions of consensus.

Many public issues, from climate policy to public health, require understanding complex causal systems and appreciating that one's own knowledge may be incomplete. IOED-driven overconfidence can reduce receptivity to expert explanations, while FCE-driven false agreement perceptions may reduce motivation to engage with diverse viewpoints (Vitriol & Marsh, 2018). Effective interventions may therefore need to target both personal metacognition (e.g., encouraging deeper explanation, self-questioning, or acknowledging uncertainty) and social metacognition (e.g., providing accurate information about opinion diversity, base rates, or population-level knowledge distributions). Without addressing both components, individuals may correct their own understanding but continue to assume that their social environment aligns with their views, limiting the impact of metacognitive interventions.

# Conclusion

The present research set out to examine whether two well-established cognitive biases—the Illusion of Explanatory Depth (Rozenblit & Keil, 2002) and the False Consensus Effect (Ross et al., 1977) are linked through shared metacognitive processes. By combining the IOED explanation task paradigm with consensus estimation tasks, these studies sought to determine whether recalibrating one's own perceived understanding would spill over into recalibrations of perceived shared understanding. The findings suggest a more nuanced relationship than initially anticipated.

First, the IOED was for the most part robustly replicated for familiar mechanical devices, consistent with the idea that people possess "sparse but causally rich" mental models that feel complete until probed (Keil, 2006; Rozenblit & Keil, 2002). When required to articulate step-by-step causal mechanisms, participants recognized explanatory gaps and revised their self-assessments downward. However, this pattern was less consistent for political topics, which are more abstract, socially embedded, and less defined by bounded causal chains (Baron, 2008; Furnham, 1988). These findings align with work suggesting that explanatory overconfidence is strongest for coherent mechanistic systems (Ahn & Kalish, 2000) and may be attenuated in domains shaped by identity-protective cognition (Kahan, 2013, 2017).

Second, strong evidence for the FCE emerged across both domains. Participants consistently overestimated how many others shared their level of knowledge, extending classic demonstrations of false consensus (Mullen et al., 1985; Nickerson, 1999) to continuous knowledge scales. Even when self-ratings were modest, participants assumed that others were similar, consistent with egocentric anchoring accounts (Epley & Gilovich, 2006). Although question structure influenced the magnitude of the FCE (Marks & Miller, 1987), the effect itself proved remarkably robust.

Crucially, the predicted interaction between IOED task and the FCE was only partially supported. For mechanical devices, explanation-induced recalibration was accompanied by modest decreases in perceived consensus, suggesting that self-knowledge may serve as an anchor for social projection (Epley & Gilovich, 2006). However, this coupling did not reliably emerge for political topics. Instead, considering the findings for "other IOED" alongside consensus estimates suggests that explanation tasks may influence different forms of social judgement in distinct ways. Judgements about others' knowledge (captured by the "other IOED") can shift following explanation even when perceived consensus does not. This indicates that individuals may revise their beliefs about what others know without correspondingly revising how similar others are to themselves.

This divergence can be understood in terms of absolute versus relative judgement. If participants downgrade both their own knowledge and their estimates of others' knowledge in parallel, perceived consensus may remain stable because the relative relationship between self and others has not changed. Under these conditions, the absence of consensus change does not imply an absence of recalibration, but rather that participants have updated their beliefs in lockstep. Consensus should only shift when explanation alters the perceived gap between self and others.

This distinction is particularly informative in the political domain. Participants tended to view themselves as more knowledgeable than others while simultaneously overestimating how many others shared their level of understanding. This pattern reflects a potentially contradictory stance, combining epistemic superiority with an assumption of widespread agreement. Such a configuration is consistent with the idea that political judgements are shaped not only by explanatory confidence, but also by identity, group norms, and motivated projection (Douglas et al., 2017; Ross & Levy, 2023). As a result, recalibrating one's own understanding may not be sufficient to alter broader assumptions about shared knowledge.

The broader implications are significant. While explanation prompts can reduce perceived understanding and, in some cases, political extremity (Fernbach et al., 2013), they may not automatically correct misperceptions of consensus. Given that perceived consensus can function as a "gateway belief" shaping other attitudes (van der Linden et al., 2015), interventions targeting overconfidence alone may have limited societal impact if assumptions about shared agreement remain unchanged. Moreover, overconfidence in explanatory knowledge has been linked to stronger endorsement of conspiratorial and politically extreme beliefs (Vitriol & Marsh, 2018), underscoring the importance of addressing both personal and social metacognition.

In sum, the present research demonstrates that explanatory overconfidence and consensus overestimation are robust but only partially intertwined. Mechanical systems reveal a clearer coupling between self-calibration and social projection, whereas political domains introduce motivational and identity-based factors that weaken this relationship. Understanding how these biases interact remains essential for designing interventions aimed at improving humility in knowledge, reducing polarization, and fostering more accurate perceptions of shared knowledge.

# References


Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. *Cognitive Psychology, 40*(2), 129–166. https://doi.org/10.1006/cogp.1999.0728

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451. https://doi.org/10.1037/a0020218

Alscher, P., Ludewig, U., Kleinkorres, R., & McElvany, N. (2025). When will they know what they don't know? Political knowledge and the infamous "Unskilled and Unaware" effect. *Contemporary Educational Psychology*, *81*, 102370. https://doi.org/10.1016/j.cedpsych.2025.102370

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge University Press.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication, 65*(5), 745–769. https://doi.org/10.1111/jcom.12171

Choi, I., & Cha, O. (2019). Cultural variations in the false consensus effect: Evidence from South Korea. *Frontiers in Psychology, 10*, Article 1238. https://doi.org/10.3389/fpsyg.2019.01238

Curley, L. J., Munro, G. D., & Lohnas, L. J. (2022). Juror decision-making and cognitive biases: A review. *Psychology, Public Policy, and Law, 28*(1), 1–19. https://doi.org/10.1037/law0000312

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science, 26*(6), 538–542. https://doi.org/10.1177/0963721417718261

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*(6), 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Psychological Science*, *15*(7), 463–469. https://doi.org/10.1111/j.0956-7976.2004.00696.x

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24*(6), 939–946. https://doi.org/10.1177/0956797612464058

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Furnham, A. (1988). *Lay theories: Everyday understanding of problems in the social sciences.* Pergamon Press.

Gaviria, S., Marsh, E. J., & Vitriol, J. A. (2021). Illusions of understanding and conspiracy beliefs. *Memory & Cognition, 49*(5), 927–940. https://doi.org/10.3758/s13421-021-01156-8

Gilovich, T. (1990). Differential construal and the false consensus effect. *Journal of Personality and Social Psychology*, *59*(4), 623–634. https://doi.org/10.1037/0022-3514.59.4.623

Goethals, G. R., Messick, D. M., & Allison, S. T. (1991). The uniqueness bias: Studies of constructive social inference. In S. T. Allison, J. M. Darley, & D. M. Messick (Eds.), *Social psychological applications to social issues* (Vol. 1, pp. 37–57). Springer.

Greene, E., & Ellis, L. (2007). Decisions in the legal system: Cognitive research and jury behavior. *Annual Review of Psychology*, *58*, 425–447. https://doi.org/10.1146/annurev.psych.58.110405.085542

Hamilton, K., & Yao, M. Z. (2018). The effect of Internet use on self-assessed knowledge: A review. *Computers in Human Behavior*, *85*, 129–141. https://doi.org/10.1016/j.chb.2018.03.018

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Social Cognition*, *27*(5), 733–763. https://doi.org/10.1521/soco.2009.27.5.733

Ipsos. (2024). *Public understanding of tariffs and trade policy* [Polling report]. https://www.ipsos.com

Johnson, S. G. B., Zhang, J., & Keil, F. C. (2016). Participants do not need to explain to realize their explanatory limitations. *Cognitive Science*, *40*(8), 2025–2050. https://doi.org/10.1111/cogs.12315

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kahan, D. M. (2017). Misconceptions, misinformation, and identity-protective cognition. In R. Y. Shapiro & L. R. Jacobs (Eds.), *The Oxford handbook of American public opinion and the media*. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kay, C. S., & Slovic, P. (2023). The generic conspiracist beliefs scale – 5: A short-form measure of conspiracist ideation. *Journal of Research in Personality*, *102*, 104315. https://doi.org/10.1016/j.jrp.2022.104315

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X

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

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*(4), 596–610. https://doi.org/10.1037/0022-3514.67.4.596

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. https://doi.org/10.1037/0022-3514.77.6.1121

Leviston, Z., Walker, I., & Price, J. (2013). Examining the role of social norms in climate change attitudes. *Climatic Change*, *119*(3), 521–533. https://doi.org/10.1007/s10584-013-0709-y

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: An empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72–90. https://doi.org/10.1037/0033-2909.102.1.72

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*, 415–444. https://www.jstor.org/stable/2678628

Meyers, E. A., Turpin, M. H., Białek, M., Fugelsang, J. A., & Koehler, D. J. (2020). Inducing feelings of ignorance makes people more receptive to expert (Economist) opinion. *Judgment and Decision Making*, *15*(6), 909–925.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of metacognitive awareness. *Journal of Experimental Child Psychology*, *87*(1), 1–32. https://doi.org/10.1016/j.jecp.2003.09.003

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C., Korris, J., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making via a single training intervention. *Judgment and Decision Making*, *10*(6), 509–526.

Mullen, B., Dovidio, J. F., Johnson, C., & Copper, C. (1985). In-group–out-group differences in social projection. *Journal of Experimental Social Psychology*, *21*(3), 262–283. https://doi.org/10.1016/0022-1031(85)90027-8

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Park, B. (2012). Uniqueness and social norms. *Personality and Social Psychology Review*, *16*(4), 298–322. https://doi.org/10.1177/1088868312440042

Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual and collective knowledge: The case for a dual-process model. *Cognition*, *189*, 21–35. https://doi.org/10.1016/j.cognition.2019.03.005

Roeder, S., & Nelson, L. D. (2015). *Folk Theories are Corrupted by Cross-Domain Explanations* (SSRN Scholarly Paper No. 2622301). Social Science Research Network. https://doi.org/10.2139/ssrn.2622301

Ross, R. M., & Levy, N. (2023). Expressive Responding in Support of Donald Trump: An Extended Replication of Schaffner and Luks (2018). *Collabra: Psychology*, *9*(1), 68054. https://doi.org/10.1525/collabra.68054

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Sherman, S. J., Presson, C. C., & Chassin, L. (1984). Mechanisms underlying the false consensus effect. *Journal of Personality and Social Psychology*, *47*(6), 1248–1261. https://doi.org/10.1037/0022-3514.47.6.1248

Stern, C., West, T. V., & Schmitt, P. G. (2014). The need to be unique and perceptions of normative self–other similarity. *Personality and Social Psychology Bulletin*, *40*(2), 269–283. https://doi.org/10.1177/0146167213510857

Van Boven, L., Judd, C. M., & Sherman, D. K. (2018). Political polarization projection: Social projection of partisan attitudes. *Nature Human Behaviour*, *2*(2), 109–118. https://doi.org/10.1038/s41562-017-0262-7

van der Linden, S., Leiserowitz, A., Feinberg, G., & Maibach, E. (2015). The scientific consensus on climate change as a gateway belief. *PLOS ONE*, *10*(2), e0118489. https://doi.org/10.1371/journal.pone.0118489

Vitriol, J. A., & Marsh, E. J. (2018). The illusion of explanatory depth and political extremism. *Political Psychology, 39*(3), 519–535. https://doi.org/10.1111/pops.12431

Vitriol, J. A., & Marsh, E. J. (2021). Illusions of understanding and conspiracy belief: Easy explanations deepen belief in falsehoods. *Journal of Applied Research in Memory and Cognition, 10*(4), 573–584. https://doi.org/10.1016/j.jarmac.2021.09.004

Voelkel, J. G., Brandt, M. J., & Colombo, M. (2018). Illusions of knowledge in political cognition. *Journal of Experimental Psychology: General, 147*(6), 897–911. https://doi.org/10.1037/xge0000464

Wilson, T. D., & Keil, F. C. (1998). The shadows and shallows of explanation. *Psychological Science, 9*(6), 464–470. https://doi.org/10.1111/1467-9280.00086

Yates, J. F., Lee, J.-W., & Shinotsuka, H. (1996). Beliefs about overconfidence and their cross-national variation. *Organizational Behavior and Human Decision Processes, 65*(2), 138–147. https://doi.org/10.1006/obhd.1996.0012

Zeveney, A., & Marsh, E. J. (2016). Illusions of knowledge and conceptual judgments. *Memory, 24*(4), 550–565. https://doi.org/10.1080/09658211.2015.1021253

# Appendix A – Study 1 Survey

Below is the complete survey issued to participants in Study 1. Please note that participants were not able to see some titles printed in bold, e.g. “Initial Ratings” or “Block 6”.

TE KUNENGA KI PŪREHUROA | MASSEY UNIVERSITY
UNIVERSITY OF NEW ZEALAND




```
<script> window.QUALTRICS_RESPONSE_ID = "${e://Field/ResponseID}";</script>
<script src="https://assets.prolific.com/assets/js/qualtrics/qualtrics.min.js"></script>
<script>window.tracker = MyTracker.getInstance("Wbm5M5YKINkz4a2fH-IvKHTqeXMHxG208u9J48e5PgjIkw5-NpJHsfHuNryvKDgicq4_5jFWyXxrFwuhky7BOVDYesj(
```


**Information Page and Consent Form**

Kia ora, thank you for your interest in this study. Below is some information for you to consider before you continue.

**Researcher introduction**
My name is Tobias Colville; I am a Masters student at the School of Psychology at Massey University in New Zealand. My supervisor for this is project Dr Stephen Hill.

**Project description**
This study is designed to investigate people's perception of their own understanding, the understanding of others, and how this perception can change.
To be eligible to participate in this study you need to be at least 18 years old, and be living in the Aotearoa New Zealand or Australia. We do not anticipate that completing this survey will cause you any discomfort or harm.

**Project procedures**
If you choose to participate, you will be asked to rate your own understanding, the understanding of others and be asked to explain certain topics.
The survey will include 11 questions in total, and take around 10-15 minutes. If you complete the survey, you will receive GBP£3.00 (approx. NZD$6.70/AUD$6.30) via Prolific.

**Data management**
After I have completed data analysis, the data from this study will be made publicly available on the Open Science Framework and stored indefinitely. This means that other researchers and members of the public will be able to access the data. Before sharing the data, I will remove any information that might indicate who you are (e.g., your Prolific ID).

**Participant rights**

You are under no obligation to accept my invitation to take part in this research. If you decide to participate, you have the right to decline to answer any particular question, or to stop answering questions at any time.

**Project contacts**

If you have any questions about this research you are welcome to contact me at [redacted]@massey.ac.nz or Dr Hill at S.R.Hill@massey.ac.nz.

**Ethics statement**

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research. If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz. (Ethics notification number: 4000030262).

Do you consent to partake in this study?

- ○ I consent
- ○ I do not consent

**Prolific ID**

What is your Prolific ID?
*Please note that this response should auto-fill with the correct ID*

${e://Field/PROLIFIC_PID}

**Likert Scale Training**

In this study, you'll rate how well you understand how different things work using a 1-to-7 scale.
Before you begin, here's what the scale represents:

A 7 means you understand how something works in detail—you could explain or draw a diagram of all the parts and processes involved (e.g., a detailed explanation of every step of the operation of a crossbow).

A 4 reflects moderate understanding—you grasp the main idea and some parts, but not everything (e.g., you know a crossbow uses a trigger and stores energy, but not how the mechanism works).

A 1 indicates very little understanding—you might know only what the object does, not how it works at all. (e.g., you know a crossbow shoots arrows/bolts, but don't know how).

The example of a crossbow shows that understanding exists on a continuum.
Your task is to judge where your knowledge lies on that continuum for each item.

Note: This scale is for non-expert knowledge. Experts are considered beyond a 7 on the scale.

**Instructions**

We're interested in how well people feel they understand a variety of phenomena including devices, natural phenomena and social phenomena. We're also interested in finding out what people think other people know about these phenomena.

You'll see 4 items. For each one, please rate your understanding and the understanding of others on a 1-to-7 scale, using the criteria provided on the previous screen.

Important Notes - PLEASE READ THIS!:
When rating your own understanding, focus only on how well you understand each item—not in reference to what you think others know.

When rating others' understanding, focus only on what you believe most people understand—not in reference to your own knowledge.

We are interested in your immediate assessments - there's no need to spend a lot of time deciding how to respond.

Simply go through the list and select a number from 1 (very little understanding) to 7 (very detailed understanding) for each item.

**Intial Ratings**

Please rate your knowledge of the following items

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| A cylinder lock | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A helicopter | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A sewing machine | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A flush toilet | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What percentage of the population would select the same point on the 1-to-7 scale as yourself?






How do you think the average person would rate their knowledge of the following items?

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| A cylinder lock | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A helicopter | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A sewing machine | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| A flush toilet | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**Explanations**

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

For each of the following, please describe all the details you know about how the item functions, going from the first step to the last, and providing the causal connection between the steps.

That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish.

In other words, for each phenomenon, try to tell as complete a story as you can, with no gaps.

If you find that your story does have gaps (that is, you are not sure how the steps are connected), please write the word "GAP" in your description at that point, and then continue.

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

Explain how a cylinder lock operates, from the point that a key is about to be inserted to when the door opens.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What percentage of the population would select the same point on the 1-to7 scale as you?




Explain how a helicopter flies, from the point of rest, through lift-off, to forward flight.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What percentage of the population would select the same point on the 1-to7 scale as you?




Explain how a sewing machine works, from the moment the needle goes in to make the first stitch to the moment when the needle goes in to make the second stitch.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What percentage of the population would select the same point on the 1-to7 scale as you?




Explain how a flush toilet operates, from the moment the flush mechanism is activated, to the moment the toilet is ready to be flushed again.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

What percentage of the population would select the same point on the 1-to7 scale as you?




In January 2013, an overview of the major discoveries in the natural sciences from the last several decades was published in the journal **Nature**. Some of these discoveries are harder to understand than others. This study is about how well you understand things. In order to study this, we need to know that

you have read all of the questions carefully. If you have, please ignore the question below and choose none of the above. We will then ask you just one more question after concluding this section.

How hard to understand did you find most of the phenomena?

- ○ Very hard
- ○ Moderately hard
- ○ Somewhat hard
- ○ Somewhat easy
- ○ Moderately easy
- ○ Very easy
- ○ None of the above

**Block 6**

Did your ratings change after you were asked to explain the device/phenomenon? Why?

Powered by Qualtrics

# Appendix B – Study 1 Descriptive Statistics

IOED for Self

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_lock_self | 110 | 2.34 | 2.00 | 1.67 | 0.159 |
| post_lock_self | 110 | 3.43 | 3.00 | 1.71 | 0.163 |
| pre_heli_self | 110 | 3.34 | 3.00 | 1.55 | 0.147 |
| post_heli_self | 110 | 3.19 | 3.00 | 1.58 | 0.150 |
| pre_sewing_self | 110 | 3.62 | 3.00 | 1.60 | 0.153 |
| post_sewing_self | 110 | 3.26 | 3.00 | 1.65 | 0.157 |
| pre_toilet_self | 110 | 4.56 | 5.00 | 1.59 | 0.151 |
| post_toilet_self | 110 | 4.05 | 4.00 | 1.60 | 0.152 |

IOED for Others

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_lock_other | 109 | 2.80 | 3.00 | 1.30 | 0.124 |
| post_lock_other | 109 | 3.50 | 3.00 | 1.27 | 0.122 |
| pre_heli_other | 109 | 3.63 | 3.00 | 1.45 | 0.139 |
| post_heli_other | 109 | 3.15 | 3.00 | 1.20 | 0.115 |
| pre_sewing_other | 108 | 3.57 | 4.00 | 1.33 | 0.128 |
| post_sewing_other | 108 | 3.34 | 3.00 | 1.31 | 0.126 |
| pre_toilet_other | 109 | 4.72 | 5.00 | 1.40 | 0.134 |
| post_toilet_other | 109 | 4.24 | 4.00 | 1.25 | 0.120 |

FCE

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_lock_prop | 110 | 57.4 | 61.0 | 24.51 | 2.337 |
| pre_lock_actualprop | 110 | 28.0 | 20.9 | 16.77 | 1.599 |
| pre_heli_prop | 110 | 61.6 | 66.5 | 25.51 | 2.432 |
| pre_heli_actualprop | 110 | 18.5 | 20.9 | 6.55 | 0.624 |
| pre_sewing_prop | 110 | 54.0 | 53.0 | 22.52 | 2.147 |
| pre_sewing_actualprop | 110 | 17.4 | 18.2 | 5.49 | 0.524 |
| pre_toilet_prop | 110 | 64.6 | 70.0 | 23.74 | 2.263 |
| pre_toilet_actualprop | 110 | 17.5 | 17.3 | 5.10 | 0.486 |
| post_lock_prop | 110 | 53.8 | 51.0 | 22.76 | 2.170 |
| post_lock_actualprop | 110 | 17.3 | 14.5 | 5.03 | 0.480 |
| post_heli_prop | 110 | 54.8 | 52.0 | 22.93 | 2.186 |
| post_heli_actualprop | 110 | 17.8 | 18.2 | 4.44 | 0.423 |
| post_sewing_prop | 110 | 52.1 | 50.0 | 21.52 | 2.051 |
| post_sewing_actualprop | 110 | 17.9 | 15.5 | 6.35 | 0.606 |
| post_toilet_prop | 110 | 58.6 | 63.5 | 22.78 | 2.172 |
| post_toilet_actualprop | 110 | 17.6 | 19.1 | 4.03 | 0.384 |

IOED-FCE

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_lock_prop | 110 | 57.4 | 61.0 | 24.5 | 2.34 |
| post_lock_prop | 110 | 53.8 | 51.0 | 22.8 | 2.17 |
| pre_heli_prop | 110 | 61.6 | 66.5 | 25.5 | 2.43 |
| post_heli_prop | 110 | 54.8 | 52.0 | 22.9 | 2.19 |
| pre_sewing_prop | 110 | 54.0 | 53.0 | 22.5 | 2.15 |
| post_sewing_prop | 110 | 52.1 | 50.0 | 21.5 | 2.05 |
| pre_toilet_prop | 110 | 64.6 | 70.0 | 23.7 | 2.26 |
| post_toilet_prop | 110 | 58.6 | 63.5 | 22.8 | 2.17 |

FCE Accuracy

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_lock_propdiff | 110 | 36.2 | 35.5 | 20.0 | 1.90 |
| post_lock_propdiff | 110 | 37.5 | 36.4 | 22.1 | 2.11 |
| pre_heli_propdiff | 110 | 44.8 | 46.0 | 25.0 | 2.39 |
| post_heli_propdiff | 110 | 37.9 | 36.5 | 22.1 | 2.11 |
| pre_sewing_propdiff | 110 | 37.6 | 37.0 | 23.0 | 2.19 |
| post_sewing_propdiff | 110 | 35.0 | 33.2 | 21.2 | 2.02 |
| pre_toilet_propdiff | 110 | 47.4 | 51.4 | 23.3 | 2.22 |
| post_toilet_propdiff | 110 | 41.2 | 45.3 | 23.4 | 2.23 |

# Appendix C – Study 2 Survey

Below is the complete survey issued to participants in Study 2. Please note that participants were not able to see some titles printed in bold, e.g. “explain_wage” or “Attention Check”.




**Information Page and Consent Form**

Kia ora, thank you for your interest in this study. Below is some information for you to consider before you continue.

**Researcher introduction**

My name is Tobias Colville; I am a Masters student at the School of Psychology at Massey University in New Zealand. My supervisor for this project is Dr Stephen Hill.

**Project description**

This study is designed to investigate people's perception of their own understanding, the understanding of others, and how this perception can change.
To be eligible to participate in this study you need to be at least 18 years old, and be living in the Aotearoa New Zealand. We do not anticipate that completing this survey will cause you any discomfort or harm.

**Project procedures**

If you choose to participate, you will be asked to rate your

own understanding, the understanding of others and be asked to explain certain topics.
The survey will include 28 questions in total, and take around 20 minutes. If you complete the survey, you will receive GBP£5.49 (approx. NZD$12.70) via Prolific.

**Data management**
After I have completed data analysis, the data from this study will be made publicly available on the Open Science Framework and stored indefinitely. This means that other researchers and members of the public will be able to access the data. Before sharing the data, I will remove any information that might indicate who you are (e.g., your Prolific ID).

**Participant rights**
You are under no obligation to accept my invitation to take part in this research. If you decide to participate, you have the right to decline to answer any particular question, or to stop answering questions at any time.

**Project contacts**
If you have any questions about this research you are welcome to contact me at [redacted]@massey.ac.nz or Dr Hill at S.R.Hill@massey.ac.nz.

**Ethics statement**
This project has been evaluated by peer review and judged to be low risk. Consequently it has not been

reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research. If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz. (Ethics notification number: 4000030262).

Do you consent to partake in this study?

- ○ I consent
- ○ I do not consent

**Prolific ID**

What is your Prolific ID?
*Please note that this response should auto-fill with the correct ID*

${e://Field/PROLIFIC_PID}

**Likert Scale Training**

In this study, you will rate how well you understand different political issues on a 5-point scale:

- 1 = very little understanding
- 5 = very detailed understanding

Before rating, review the examples below based on immigration policy and its effects on the economy. These show what different levels of understanding look like:

- **Level 5 Knowledge:** Very detailed knowledge. Understands multiple dimensions of impact, such as the balance between taxes paid and services received, effects on wages for both low- and high-skilled workers, changes in prices of goods and services, and benefits for some business owners. Also recognizes that while immigrants may compete with domestic workers, they can also expand the labor force by filling roles that otherwise wouldn't exist.

- **Level 3 Knowledge:** Moderate knowledge. Aware that immigration has both positive and negative economic effects. For example, immigrants may put pressure on social services and increase job competition, but they also supply affordable labor and often take jobs that local residents avoid.

• **Level 1 Knowledge:** Minimal knowledge. Knows only in a general sense that immigration affects the economy, such as by influencing job competition, but lacks deeper detail.

**Instructions**

We're interested in how well people feel they understand social issues. We're also interested in finding out what people think other people know about these issues.

You'll see 4 items. For each one, please rate your understanding and the understanding of others on a 1-to-5 scale, using the criteria provided on the previous screen.

**Important Notes - PLEASE READ THIS!:**
When rating your own understanding, focus only on how well you understand each item—not in reference to what you think others know.

When rating others' understanding, focus only on what you believe most people understand—not in reference to your own knowledge.

We are interested in your immediate assessments -

there's no need to spend a lot of time deciding how to respond.

Simply go through the list and select a number from 1 (very little understanding) to 5 (very detailed understanding) for each item.

**Intial Ratings**

Please rate your own understanding of the causal consequences of **recognising Palestine as a sovereign state and** how well you feel others understand the causal consequences of this issue.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.

| | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ○ | | | | | | | | | | 0 |




Please rate your own understanding of the causal consequences of **raising minimum wage and** how well you feel others understand the causal consequences of this issue.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.






Please rate your own understanding of the causal consequences of **enacting the Treaty Principles Bill and** how well you feel others understand the causal consequences of this issue.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.






Please rate your own understanding of the causal consequences of **imposing sanctions on Chinese imports and** how well you feel others understand the causal consequences of this issue**.**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.






**Total:** **0**

## Explanation Instructions

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

For each of the following topics, please describe all the details you know about the causal consequences of the topic, going from the first step to the last, and providing the causal connection between the steps.

That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish.

In other words, for each phenomenon, try to tell as complete a story as you can, with no gaps.

If you find that your story does have gaps (that is, you are not sure how the steps are connected), please write the word "GAP" in your description at that point, and then continue.

When you are done, please re-rate your knowledge of the phenomenon on a 1-to-5 scale.

**Palestine**

Please describe all the details you know about the causal consequences of **recognising Palestine as a sovereign state**, going from the first step to the last, and providing the causal connection between the steps. Your explanation should state precisely how each step causes the next step in one continuous chain from start to finish. In other words, try to tell as complete a story as you can, with no gaps. Please take your time, as we expect your best explanation.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.




**explain_wage**

Please describe all the details you know about the causal consequences of **raising minimum wage**, going from the first step to the last, and providing the causal connection between the steps. Your explanation should state precisely how each step causes the next step in one continuous chain from start to finish. In other words, try to

tell as complete a story as you can, with no gaps. Please take your time, as we expect your best explanation.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

What percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.






**explain_tpb**

Please describe all the details you know about the causal consequences of **enacting the Treaty Principles Bill**, going from the first step to the last, and providing the causal connection between the steps. That is, your explanation should state precisely how each step causes the next step in one continuous chain from start to finish. In other words, try to tell as complete a story as you can, with no gaps. Please take your time, as we expect your best explanation.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.




**explain_china**

Please describe all the details you know about the causal consequences of **imposing sanctions on Chinese goods**, going from the first step to the last, and providing the causal connection between the steps. That is, your explanation should state precisely how each step causes

the next step in one continuous chain from start to finish. In other words, try to tell as complete a story as you can, with no gaps. Please take your time, as we expect your best explanation.

Now, please rate your understanding of this item **and** how well you feel others understand this phenomenon.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Own understanding | ○ | ○ | ○ | ○ | ○ |
| Understanding of others | ○ | ○ | ○ | ○ | ○ |

For this issue, what percentage of people would choose each of the following points on the 1-to-5 scale? The total must add to 100.

| | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ○ | | | | | | | | | | | 0 |
| 2 | ○ | | | | | | | | | | | 0 |




**Change?**

Did your ratings change after you were asked to explain the topic? Why?

**Attention Check**

In January 2013, an overview of the major discoveries in the natural sciences from the last several decades was published in the journal **Nature**. Some of these discoveries are harder to understand than others. This study is about how well you understand things. In order to study this, we need to know that you have read all of the questions carefully. If you have, please ignore the question below and choose none of the above. We will

then ask you just a few more questions after concluding this section.

How hard to understand did you find most of the phenomena?

- ○ Very hard
- ○ Moderately hard
- ○ Somewhat hard
- ○ Somewhat easy
- ○ Moderately easy
- ○ Very easy
- ○ None of the above

**GCB5**

Finally, please indicate your level of agreement with each of the following five statements below.

The government permits or perpetrates acts of terrorism on its own soil, disguising its involvement.

- ○ Strongly agree
- ○ Moderately agree
- ○ Slightly agree
- ○ Slightly disagree

- ○ Moderately disagree
- ○ Strongly disagree

Evidence of alien contact is being concealed from the public.

- ○ Strongly agree
- ○ Moderately agree
- ○ Slightly agree
- ○ Slightly disagree
- ○ Moderately disagree
- ○ Strongly disagree

New and advanced technology which would harm current industry is being suppressed.

- ○ Strongly agree
- ○ Moderately agree
- ○ Slightly agree
- ○ Slightly disagree
- ○ Moderately disagree
- ○ Strongly disagree

Certain significant events have been the result of the activity of a small group who secretly manipulate world

events.

- ○ Strongly agree
- ○ Moderately agree
- ○ Slightly agree
- ○ Slightly disagree
- ○ Moderately disagree
- ○ Strongly disagree

Experiments involving new drugs or technologies are routinely carried out on the public without their knowledge or consent.

- ○ Strongly agree
- ○ Moderately agree
- ○ Slightly agree
- ○ Slightly disagree
- ○ Moderately disagree
- ○ Strongly disagree

Powered by Qualtrics

# Appendix D – Study 2 Descriptive Statistics

IOED for Self

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_rate_palestine_1 | 110 | 2.70 | 3.00 | 1.216 | 0.1159 |
| post_rate_palestine_1 | 110 | 2.45 | 2.00 | 1.275 | 0.1216 |
| pre_rate_wage_1 | 111 | 3.43 | 3.00 | 0.911 | 0.0864 |
| post_rate_wage_1 | 111 | 3.11 | 3.00 | 1.021 | 0.0969 |
| pre_rate_tpb_1 | 108 | 2.43 | 2.00 | 1.327 | 0.1277 |
| post_rate_tpb_1 | 108 | 2.31 | 2.00 | 1.173 | 0.1129 |
| pre_rate_china_1 | 108 | 2.61 | 3.00 | 1.159 | 0.1115 |
| post_rate_china_1 | 108 | 2.66 | 3.00 | 1.104 | 0.1062 |

IOED for Others

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_rate_palestine_2 | 108 | 2.37 | 2.00 | 0.882 | 0.0848 |
| post_rate_palestine_2 | 108 | 2.22 | 2.00 | 0.857 | 0.0825 |
| pre_rate_wage_2 | 111 | 2.89 | 3.00 | 0.767 | 0.0728 |
| post_rate_wage_2 | 111 | 2.72 | 3.00 | 0.800 | 0.0759 |
| pre_rate_tpb_2 | 107 | 2.27 | 2.00 | 0.886 | 0.0856 |
| post_rate_tpb_2 | 107 | 2.28 | 2.00 | 0.833 | 0.0805 |
| pre_rate_china_2 | 107 | 2.24 | 2.00 | 0.712 | 0.0688 |
| post_rate_china_2 | 107 | 2.42 | 2.00 | 0.765 | 0.0740 |

FCE

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_pal_actualprop | 112 | 22.4 | 26.8 | 5.58 | 0.527 |
| pre_palestine_cons_est | 112 | 25.8 | 22.0 | 15.95 | 1.507 |
| pre_wage_actualprop | 112 | 29.8 | 35.7 | 10.47 | 0.990 |
| pre_wage_cons_est | 112 | 31.3 | 29.0 | 15.85 | 1.498 |
| pre_tpb_actualprop | 112 | 23.2 | 27.7 | 7.39 | 0.698 |
| pre_tpb_cons_est | 112 | 27.7 | 25.0 | 17.33 | 1.637 |
| pre_china_actualprop | 112 | 24.1 | 29.5 | 7.79 | 0.736 |
| pre_china_cons_est | 112 | 28.6 | 26.0 | 15.33 | 1.449 |
| post_pal_actualprop | 110 | 23.4 | 22.7 | 6.97 | 0.665 |
| post_palestine_cons_est | 110 | 28.6 | 25.0 | 15.56 | 1.484 |
| post_wage_actualprop | 111 | 22.1 | 27.0 | 11.79 | 1.119 |
| post_wage_cons_est | 111 | 28.8 | 29.0 | 14.06 | 1.334 |
| post_tpb_actualprop | 108 | 25.2 | 29.6 | 8.02 | 0.772 |
| post_tpb_cons_est | 108 | 25.3 | 25.0 | 14.17 | 1.363 |
| post_china_actualprop | 108 | 26.8 | 30.6 | 11.07 | 1.065 |
| post_china_cons_est | 108 | 30.1 | 30.0 | 14.62 | 1.407 |

IOED-FCE

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_palestine_cons_est | 110 | 25.8 | 22.0 | 16.1 | 1.53 |
| post_palestine_cons_est | 110 | 28.6 | 25.0 | 15.6 | 1.48 |
| pre_wage_cons_est | 111 | 31.0 | 29.0 | 15.7 | 1.49 |
| post_wage_cons_est | 111 | 28.8 | 29.0 | 14.1 | 1.33 |
| pre_tpb_cons_est | 108 | 27.8 | 25.0 | 17.4 | 1.67 |
| post_tpb_cons_est | 108 | 25.3 | 25.0 | 14.2 | 1.36 |
| pre_china_cons_est | 108 | 28.6 | 26.0 | 15.3 | 1.48 |
| post_china_cons_est | 108 | 30.1 | 30.0 | 14.6 | 1.41 |

IOED-FCE Accuracy

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| pre_pal_prop_diff | 110 | 11.5 | 7.00 | 12.84 | 1.224 |
| post_pal_prop_diff | 110 | 13.1 | 10.77 | 11.37 | 1.084 |
| pre_wage_prop_diff | 111 | 13.7 | 11.61 | 10.36 | 0.984 |
| post_wage_prop_diff | 111 | 15.0 | 12.97 | 10.91 | 1.035 |
| pre_tpb_prop_diff | 108 | 13.6 | 10.36 | 13.38 | 1.288 |
| post_tbp_prop_diff | 108 | 11.3 | 9.63 | 9.57 | 0.921 |
| pre_china_prop_diff | 108 | 12.9 | 9.63 | 11.50 | 1.107 |
| post_china_prop_diff | 108 | 13.0 | 10.56 | 10.39 | 1.000 |