Conceptual Revision in Action

Ethan Landes and Kevin Reuter

Draft as of 23 October, 2023

Abstract

Conceptual engineering is the practice of revising concepts to improve how people talk and think. Its ability to improve talk and thought ultimately hinges on the successful dissemination of desired conceptual changes. Unfortunately, the field has been slow to develop methods to directly test what barriers stand in the way of propagation and what methods will most effectively propagate desired conceptual change. In order to test such questions, this paper introduces the masked time-lagged method. The masked time-lagged method tests people's conceptual understanding at two different points in time without their knowledge of being tested, allowing us to measure conceptual revision in action. Using a masked time-lagged design on a content internalist framework, we attempted to revise PLANET and DINOSAUR in online participants to match experts' concepts. We successfully revised PLANET but not DINOSAUR, demonstrating some of the difficulties conceptual engineers face. Nonetheless, this paper provides conceptual engineers, regardless of framework, with the tools to tackle questions related to implementation empirically and head-on.

1 Introduction

Conceptual engineering is the practice of improving concepts so that we can think and talk better.[1] Rather than merely trying to understand existing conceptual or linguistic structures, conceptual engineers aim to be conceptually or linguistically creative, revisionary, and innovative in order to address shortcomings in our current conceptual schema. Using the four-part process of Isaac et al. (2022), conceptual engineering involves *describing* people's existing concepts, *evaluating* whether they fall short in some way, *improving* upon them by designing a new or revised concept, and *implementing* the product by spreading it to the right people.

What exactly one should take the target of conceptual engineering to be is an open and contentious issue in the conceptual engineering literature (Belleri, 2021; Isaac et al., 2022). Irrespective of this unresolved question, however, the first three steps of conceptual engineering involve processes that are well-understood by academics. Cognitive science, developmental science, experimental philosophy, armchair philosophy, and other fields have long been in the business of describing people's concepts, such as CAUSE, TRUTH and RESPONSIBILITY, using methods like thought experiments, laboratory demonstrations, corpus linguistics, etc. Similarly, philosophy and other fields have long been in the business of the second step of conceptual engineering, evaluation, when they use thought experiments, arguments, formalizations, and other tools to determine whether concepts lead to fallacies or other unwanted consequences. Improving the concept, the third step of conceptual engineering, has a less storied history, but recent philosophers have drawn lessons from design and engineering and have developed improved concepts themselves (Haslanger, 2000; Napolitano & Reuter, 2021; Reuter & Brun, 2022; Scharp, 2013).

[1] For the rest of this section, we use "concept" in the broadest possible sense to include any potential target of conceptual engineering. Starting in Section 2, we use "concept" specifically in the content internalist sense to refer to token cognitive entities.

This leaves the final step, the implementation of engineered concepts. Here, conceptual engineers face two types of empirical questions:

- Global Implementation Questions: What factors *in general* affect the success or failure of propagating concepts?
- Local Implementation Questions: What factors affect the success or failure of revising or replacing *specific* concepts?

Conceptual engineers have hitherto tried to answer global implementation questions by either drawing parallels from historical examples of conceptual propagation and linguistic change or by drawing lessons from cognitive science. The historical case studies start from a straightforward premise: linguistic change and collective conceptual change have happened in the past, and lessons can be learned from those case studies (Koslow, 2022; Landes, 2023; Thomasson, 2021). For example, Landes (2023) examines the problems faced by public health officials when SOCIAL DISTANCING was propagated worldwide in early 2020, arguing the label "social distancing" potentially hindered accurate conceptual propagation.

Others have tried to make ground on global implementation questions by looking towards cognitive science and arguing that cognitive features make some implementations easier than others. Machery (2021) combines experimental philosophy work on the defectiveness of INNATE (Griffiths & Machery, 2008; Machery et al., 2019) with cognitive science work on cultural evolution (Buskell, 2017; Scott-Phillips et al., 2018; Sperber, 1996). Some concepts, such as INNATE, will not easily be revised, Machery argues, because we are naturally attracted to features of the concept. Fischer (2020) instead focuses on what work on the cognitive structure of polysemy can teach about conceptual revision. Drawing on work showing certain types of polysemy can lead to conflation (Fischer & Engelhardt, 2019; Fischer et al., 2015; Giora, 2003), Fischer argues that conceptual engineers should avoid novel senses of words that require us to suppress features of more dominant senses.

Local implementation questions have gained less attention. While changes have been proposed by conceptual engineers, very little effort has been made to actually propagate them. Conceptual engineers have come up with various proposed designed changes, such as splitting TRUTH (Scharp, 2013), ameliorating WOMAN (Haslanger, 2000), and introducing "conspiratorial explanation" as a neutral counterpart to "conspiracy theory" (Napolitano & Reuter, 2021). These have, however, merely remained proposals. In contrast, the examples we do have of successfully propagated designed cognitive and linguistic devices — the sex/gender distinction (Muehlenhard & Peterson, 2011), the revision of PLANET (IAU, 2006), and the relabeling of "battle fatigue" to "Post-Traumatic Stress Disorder" (Jones, 2013) — have all come from outside of the academic discipline of conceptual engineering. At the time of writing, self-identifying conceptual engineers have largely limited their efforts at propagating their designed changes to presenting the changes in academic articles, and they have been slow to research what sorts of propagation efforts are required to get their designed changes to stick.

Ultimately both global and local implementation questions require an empirical approach, regardless of the conceptual engineering framework ([Author ms]). Whatever the target of conceptual engineering is, something needs to be propagated, whether a cognitive disposition, causal historical chain, speaker meaning, etc. However, since these targets differ significantly, one might assume that each framework presents its own set of empirical inquiries. For instance, what factors influence the success of the spread of a linguistic norm may well differ from the factors that impact the uptake of adjusted categorization patterns. Nevertheless, the purpose of this paper is to demonstrate that our proposed empirical approach is both comprehensive and foundational, enabling the examination of global and local implementation questions independently of one's preferred conceptual engineering framework.

By employing this broad methodology, we aim to demonstrate that both global and local implementation questions can be directly tested. In the empirical parts of this paper, we will

adopt a content internalist framework of conceptual engineering (e.g., Fischer, 2020; Machery, 2017; Pollock, 2021) to introduce the masked time-lagged method (Section 2) as well as the hypotheses and results of our study (Section 3). Content internalist frameworks take concepts to be token cognitive representations that possess structure and influence cognitive processes such as categorization and inference (Machery, 2009; Margolis & Laurence, 2007). We adopt this framework primarily due to its longstanding dominance within the field of cognitive science. After we present our findings, we tackle possible objections and discuss how and why this framework can be expanded and applied to other conceptual engineering frameworks (Section 4). In particular, we discuss frameworks that target speaker meaning (Pinder, 2020, 2021) and semantic externalist frameworks (Cappelen, 2018; Sterken, 2020).

2 A New Methodology for Testing Conceptual Change: Masked Time-Lagged Studies

How can conceptual engineers directly test specific hypotheses raised by the global and local propagation questions? Developmental psychology may prove as a source of inspiration, particularly in relationship to global implementation questions. Like conceptual engineers, many developmental psychologists are interested in how concepts change, albeit typically during natural development and education as opposed to in response to conceptual engineering (e.g., Carey, 2011; Poling & Evans, 2004; Shtulman & Calabi, 2013; Spelke & Tsivkin, 2001). Here, the methods of developmental psychology have uncovered lessons relevant to conceptual engineering. For one, naive folk concepts do not morph into scientific concepts, instead the folk concepts appear to be suppressed by the novel scientific concepts (Shtulman & Valcarcel, 2012). For another, depending on the concept, conceptual change can take months or even years, and such complex conceptual changes may only be present in a fraction of the adult population (Carey, 2011).

There are multiple reasons why the methods of developmental psychology are of limited use to conceptual engineers hoping to answer global and local questions about implementation. First, developmental psychologists primarily study how children and young adults undergo conceptual change, focusing on their ability to learn new concepts and adapt to scientific theories. In contrast, conceptual engineers often aim to implement conceptual change in adults who have held specific concepts for long periods of time. Second, existing empirical studies often examine conceptual change in traditional educational settings such as classrooms. However, conceptual engineers may want to target adult populations and change their views through non-formal means of information dissemination. Third, practical considerations favor survey-based designs over classroom-based and other in-person designs. In general, survey-based designs are easier and cheaper to run than in-person designs. At the same time, due to contingent historical facts, empirically-driven analytic philosophers typically do not currently have the logistics or expertise in place to run, for example, in-person demonstration-based or game-based studies – a common design to test for conceptual change in infants and children. The shift to adults and survey-based designs create issues, however, as surveys on adults are notoriously susceptible to noise caused by participants either misreading the intention of the experimenter or relying on subtle pragmatic cues to interpret questions in unwanted ways (Conrad et al., 2014; Cullen, 2010; Schwarz, 1995). Moreover, for reasons discussed shortly, we believe tests of conceptual revision are particularly susceptible to such survey pragmatics.

In short, an experimental method is needed that (a) targets adult populations, (b) employs non-formal means of information dissemination, (c) uses a survey-based design, and (d) minimizes survey pragmatics.

One of the central goals of the present paper is to propose a novel method that satisfies all four desiderata, the masked time-lagged design. Our design tests people's conceptual understanding at two different points in time without their knowledge of being tested, allowing us to measure conceptual revision in action. During an initial phase, participants receive an intervention that

attempts to change their conceptual content. Then, during a second phase that has no apparent connection to the first, participants' conceptual content is measured. By employing this method, we can observe and analyze the process of conceptual revision as it unfolds over time.

Masking and time-lagging are required to solve two methodological challenges faced by anyone wanting to test for genuine conceptual change. Starting with time-lagging, time-lagging is required to test whether any measured change occurs over the timescale that interests conceptual engineers. Conceptual engineers aim to bring about long-term linguistic or conceptual changes, and it is not enough that people learn novel content for a few minutes and then either forget it or never deploy it again. Accordingly, any design testing conceptual propagation will have to take steps to avoid collecting the contents of short-lived ad hoc concepts or information stored in working memory. Testing revision at separate points in time avoids this issue. It is, however, an open question how much of a time lag is needed to check for long-term uptake. The experiment reported below operated at the scale of a few hours to a few days, but the methods described could easily be scaled up to order of weeks or months.

In contrast to time-lagging, masking is required to manipulate survey pragmatics and participant mind-reading that may get in the way of measuring conceptual change. Pragmatic effects are particularly troublesome on invariantist frameworks, like those popular in the conceptual engineering literature, where conceptual content is the default information retrieved in every context or outside of context (Fischer, 2020; Machery, 2009, 2017). On such accounts, *any* context is a source of noise. However, survey pragmatics is not just a problem for such invariantist pictures of conceptual content. While it is tempting to think of participants as passively following the instructions, the reality is that participants are best thought of as being in conversation with the experiment or experimenter (Conrad et al., 2014; Schwarz, 1995). As participants work their way through a study, being helpful interlocutors, they use clues in wording and design to try to determine what exactly the experimenter is interested in and answer accordingly. Such Gricean mechanisms have, for example, been found to cause order effects and affect how participants interpret scales (Cullen, 2010; Schwarz, 1995).

At least three mechanisms, avoided by masking, risk artificially raising the rate of measured conceptual propagation among test groups. First, if the test questions were not masked, it would obvious to participants what answers the study is trying to cause. The stimuli are carefully sourced and argued (see Section 3.3), suggesting that not only do the experimenters believe what is being presented (which is true, we do), but that the experimenters want participants to believe it too. This may artificially increase measured rates of revision when participants answer in a way they think is helpful. Second, an explicit connection to the pre-test during the post-test phase will raise the salience of the intervention in participants' minds. While we want participants to remember the content of the intervention, remembering merely because they recognize the connection between the pre-test and post-test is not genuine change. Third, and most crucially, masking is required to allow for meaningful comparisons between the control and test groups. Testing conceptual revision requires measuring whether specific content changes over time. In many cases, this will only be portion of a participants' conceptual content.[2] Any measure of change, especially open-ended questions like those employed here (see Section 3.2), will therefore include many non-relevant answers. The intervention will raise the salience of specific content. Thus participants in the test group in non-masked designs will, through pragmatic mechanisms, interpret the content being measured as part of the question under discussion during the "conversation" of the study. This will lower the validity of between-subject comparisons, as participants in the test group will be driven by pragmatic mechanisms to be more likely to answer according to the content being measured than the control group will.

[2] This is only true when working with concepts whose content are not necessary and sufficient conditions. However, few, if any, (internalist) concepts have content structured this way (Laurence & Margolis, 1999).

3 Empirical Study

In this section, we discuss the selection of concepts, introduce our methods and hypotheses, and present the results of the masked time-lagged study. Hypotheses and methodology were pre-registered with the Open Science Framework. The different surveys, including all the stimuli, that were administered to the participants are available on this online repository.

3.1 Selection of Concepts

In order to demonstrate how conceptual uptake can be directly tested, we selected two concepts, DINOSAUR and PLANET. We aimed to revise them to be in line with recent scientific discoveries. According to common folk wisdom, dinosaurs are extinct and Pluto is a planet. However, in the last few decades scientists have generally come to the consensus that (a) birds are an existing form of dinosaur, and so dinosaurs are not extinct and (b) Pluto is not a planet because, unlike planets, it has not cleared its orbit of debris. This means many folk have a concept DINOSAUR that incorrectly excludes birds and a concept PLANET that incorrectly includes Pluto. Compared to other potential concepts, we chose DINOSAUR and PLANET because they have four key features:

First, they are real cases backed by relevant experts. It would be unethical to attempt to use our position of authority to propagate concepts that are not in the participants' epistemic interests (Shields, 2021, 2023). However, the shift in content of DINOSAUR and PLANET have been widely adopted and advocated for by paleontologists and astronomers, respectively (Brusatte, 2017; NASA, 2023).[3] Therefore, the choice of DINOSAUR and PLANET allows the study of propagation of content that both a majority of relevant experts and we as experimenters think is the best or correct content.

Second, these are scientific cases. Conceptual change will likely require buy-in by participants. Participants need to believe that the intervention is truthful and requires the cognitive effort to appreciate the stimuli to the extent that is needed to change concepts. Accordingly, the concepts DINOSAUR and PLANET allow us to write stimuli in a way that piggybacks on the prestige and social stature of science, by, for example, using real NASA-generated images of the solar system.[4]

Third, these are live cases as people are still learning and/or adjusting to the decision by the IAU that PLANET excludes Pluto and the recent series of discoveries by paleontologists that birds are a type of dinosaur. Therefore, a large percentage of the participant pool does not yet have concepts with the content we hope to revise.[5]

Fourth, these concepts allow for easy measurement of the conceptual content under question. Piloting found that for many people Pluto has the salient feature of being a planet and dinosaurs have the salient feature of being extinct, both of which are consistent with the old concepts but inconsistent with the new concepts. Therefore, we were able to construct both multiple choice and short answer questions that allowed us to measure the conceptual content under question.

[3] Note that there are nonetheless a few prominent opponents to the current IAU definition of planet (Chang, 2022).

[4] This is not to say testing non-scientific conceptual change is not possible, but to the extent that buy-in is required by participants (which is itself an open empirical question), other concepts will require different ways to establish legitimacy.

[5] By our best estimate from the control conditions (see Figures 2 and 4), at the time of data collection in December 2022, somewhere between 60% to 80% of the US and UK Prolific population had the old concept of PLANET and somewhere between 70% to 100% had the old concept of DINOSAUR. This estimate was reached by using the Categorization measure as one bound and the higher of either the Completion or SFP measure as the other bound. The higher of the Completion and SFP measure was used because we believe that these, if anything, underestimate the rate of content among the population (see Section 3.2).

3.2 Selection of Measures

In order to demonstrate the strengths and weaknesses of different ways of measuring conceptual change, we deployed three measures: Completion tasks, Semantic Feature Production (SFP) tasks, and Categorization tasks. Specifically, the measures were as follows:

Completion: Participants were asked to complete the sentence: "Dinosaurs are . . . " Or "Pluto is . . . ", such that a true sentence would be stated.

Semantic Feature Production: Participants were asked to state three characteristics that came to mind when they thought about dinosaurs or Pluto.

Categorization: Participants were asked to categorize whether four items were dinosaurs or planets (the test item being a seagull or Pluto).

Completion tasks are suggested by Fischer (2020) as a way to test for conceptual content on certain invariant accounts of conceptual content. Invariantist accounts take conceptual content to be stable over time (compare to variantists like Casasanto (2015)). Fischer has in mind, in particular, invariantist accounts that take content to be the belief-like states stored in long-term memory that are retrieved by default – that is, conceptual content is the information retrieved quickly outside of context (Machery, 2009).[6] Fischer argues that single-word priming tasks such as "x is ___" are well-suited to get at such content due to their low context and open-ended nature.

Semantic Feature Production (SFP) tasks, sometimes called feature listing tasks, are designed to capture the salient characteristics that participants associate with a given entity. Early examples of such tasks can be found in works by Hampton (1979) and Barsalou (1983), while a comprehensive discussion can be found in Machery (2009). McRae et al. (2005) conducted a study wherein participants were requested to provide salient features associated with hundreds of concepts. Salient features are those that stand out in our mental representation of a particular category compared to other properties. For instance, the feature "dangerous" is considered salient in the concept shark, despite the fact that sharks are not necessarily or typically dangerous creatures. Salient features of concepts might be particularly hard to change in people's representation of kinds. Consequently, Semantic Feature Production tasks offer a way to measure how salient features change over time.[7]

Categorization tasks are multiple-choice questions asking participants to select which of several options are members of a given category or kind. Unlike the other two measures, the Categorization task is less likely to under-count the conceptual content of interest because every participant must give a response that bears on the content being studied. Moreover, the Categorization task directly targets what is arguably one of the main functions of concepts –

[6] Fischer's account is a modification of Machery's. Machery instead takes conceptual content to be information retrieved *regardless of context*. Unfortunately, truly context-free data gathering is impossible due to the pragmatics of survey design, as participants will inevitably form beliefs (whether accurate or inaccurate) about the intentions of the experimenter and answer accordingly (Cullen, 2010). Nonetheless, the hope is that with sufficient masking, zero-context responses can be approximated.

[7] The SFP and Completion task are both open-ended. While this means they potentially provide a rich data set that may reveal unexpected results, the open-endedness is also a weakness. For one, not all conceptual revisions will involve changing salient features or the most obvious answer to a no-context prime. This is the case with planet. The target change of content for planet is the addition of the condition that planets have to clear their orbit of debris. However, while piloting we found the most salient properties of planethood are being big and orbiting the sun. As the study below demonstrates, this is not fatal, as designs can work around this by finding a nearby term that is affected by the targeted conceptual content. The second weakness of the SFP and Completion task is that they under-count the conceptual content of interest. For example, "clearing orbit of debris" will not be someone's only salient property of planet, so it will not always end up as an answer in free-response questions measuring salient properties.Therefore, the SFP and Completion tasks may not be an accurate estimate of the level of prevalence of some content among some population, although they are still useful to test how content changes in response to interventions.

sorting objects into categories. One of the main reasons concepts (again, understood as token cognitive entities) are of interest to conceptual engineers is because they determine how we sort things in the world (Isaac, 2021; Machery, 2017; Margolis & Laurence, 2007), which can in turn influence inferences and decision-making. The Categorization task in particular, however, risks introducing unwanted context by providing implicit contrast cases in the other options.

3.3 Stimuli

In order to try to change concepts while also offering initial answers to local implementation questions about DINOSAUR and PLANET, two sets of stimuli were written for each concept. The first set of stimuli were minimal, text-only interventions. The text-only interventions were written to explain the conceptual changes behind either DINOSAUR or PLANET in around 200 words and to be readable by participants in one to two minutes. The DINOSAUR text-only intervention quoted passages from a Scientific American article about the discovery that birds are dinosaurs whereas the PLANET text-only intervention quoted NASA's version of the IAU definition of planet and quoted a BBC explanation for why Pluto is not a planet. Here is a snippet from the text-only intervention for dinosaur (the full texts can be accessed in the online repository) :

> "The feathered dinosaurs of Liaoning clinched it: birds really did evolve from dinosaurs. But that statement is perhaps a little misleading because it suggests that the two groups are totally different things. In truth, birds are dinosaurs—they are one of the many subgroups that can trace their heritage back to the common ancestor of dinosaurs and therefore every bit as dinosaurian as Triceratops or Brontosaurus. You can think of it this way: birds are dinosaurs in the same way that bats are a type of mammal that can fly."

The second set of interventions were meant to be more in-depth, multimodal and take 2 to 3 minutes to read. They used pictures in addition to text to explain conceptual change and also included quotes from prominent figures stating that birds are dinosaurs or that Pluto is not a planet. The longer DINOSAUR stimulus's argumentative structure was to try to show that birds and dinosaurs really are of the same kind. The pictures included recreations of prehistoric Theropods, which are notably bird-like, as well as pictures of emus and shoebills, particularly prehistoric looking birds. The longer PLANET stimulus argued that there are two distinct kinds of things that are round and orbiting the sun, and Pluto is more like the non-planets than planets. Images compared the sizes and shapes of Pluto, the Earth, similarly sized objects beyond Neptune, and other celestial bodies, as well as the comparatively messy orbit of Pluto compared to the eight actual planets.

3.4 Methods & Hypotheses

As discussed above, the survey used a masked time-lagged design in order to check for genuine conceptual change. This was accomplished using Prolific's screening tools, as we opened the masked follow-up only to participants who took the appropriate pre-test. The follow-up was made available to the participants who took the intervention task from 2 hours to 72 hours after the intervention task was posted.The mean time between responses was 13.0 hours, and the median was 5.1 hours. Every effort was made to mask the connection, including posting the follow-up on a different Prolific account and making sure the look and feel of the two surveys differed significantly.

The survey was a between-subject design with the following procedure: 720 participants were given one of four interventions (test group). In the implementation experiment pre-test, members of the test group saw stimuli either about how birds are dinosaurs and thus dinosaurs are not extinct or stimuli about how Pluto is not a planet because it has not cleared its orbit

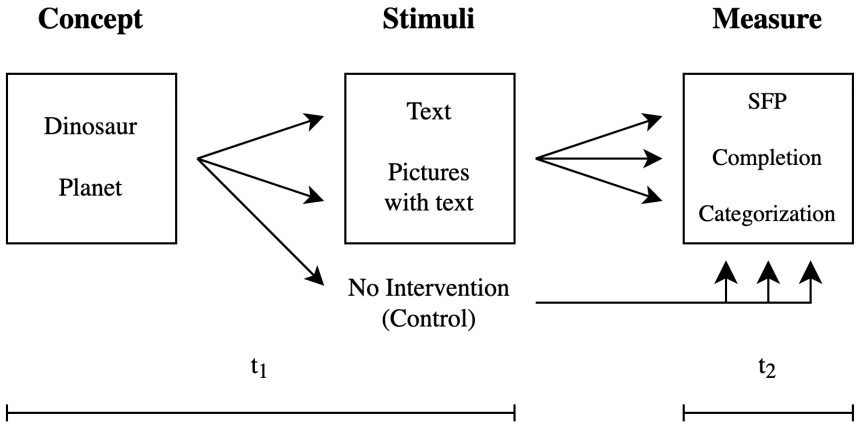

Figure 1: Diagram of the survey's 2 x 3 x 3 design. Participants were asked questions related to the intervention at t_1, but responses were not analyzed as part of this study. The measure step (t_2) was masked and completed between 1.5 and 73.3 hours after the initial intervention.

of debris. Participants saw one of two versions of these stimuli. Either they saw a roughly 200-word explanation quoting popular science writing (the Text condition) or they saw a longer explanation that employed several pictures (the Picture condition). Therefore, there were 4 intervention conditions: (a) Dinosaur-Text, (b) Dinosaur-Picture, (c) Planet-Text, (d) Planet-Picture. During the pre-test but after the intervention, participants were asked ersatz test questions in order to make it look like the study was completed at the end of the pre-test.[8] Participants in the control group were assigned to one of the concepts (DINOSAUR / PLANET) but were not given an intervention.

In the masked follow-up, participants in both the test group and control group were asked to respond to a post-experiment, in which we aimed to test whether we were successful in implementing conceptual change on DINOSAUR or PLANET in our test group. Every participant saw one of 3 measures in the follow-up (Figure 1). Thus, our design was 2 (Concept) x 3 (Intervention) x 3 (Measure).

Participants were randomly assigned to one of the three measures (see Section 3.2 above). The test groups and control group saw the same post-test, but only the test groups saw a pre-test intervention. The test question was hidden as one question among five and appeared fourth in a set of five questions. The exact questions they saw depended on the measure and can be accessed in the online repository.[9]

Conceptual engineers disagree about about how easy it is to implement changes (e.g., Cappelen, 2018; Jorem, 2021; Koch, 2021; Nimtz, 2021). Different frameworks will face different challenges as to how to implement concepts, as different targets will require different changes to take hold. Content internalism is not immune to these worries. Two of the content internalists discussed above—Fischer (2020) and Machery (2021)—both theorize certain conceptual changes would be very difficult to carry out due to cognitive factors. Therefore, for each of the three measures, our null hypothesis is that there would be no significant difference between control and test groups, both aggregating the two DINOSAUR and PLANET interventions together and

[8] While it would have increased our statistical power to use a within-subject design by asking participants the same questions before and after the intervention, we were worried that including the same test questions during both the pre-test and the post-test sessions would make masking the post-test nearly impossible.

[9] While we did not explain why we were asking the questions we were, when asked, many participants guessed the survey purpose involved testing common knowledge or common associations.

for each of the four interventions separately.[10]

3.5 Results

Of 720 participants in the experimental conditions, 560 participants (78% of the pre-test) completed the post-test, receiving one of three tasks. Of these, 12 participants were excluded for correctly guessing the survey purpose at the end of the post-test, for a final total of 548 (76% of the pre-test). 361 other participants in the control condition also took one of the six (2 Concepts x 3 Measures) post-test tasks without any pre-test intervention. The experimenters coded results based on whether participants answered according to the "incorrect" folk concept (see online repository). Thus, any participant who wrote "extinct" (or a synonym) for dinosaur or "planet" (or a synonym) for Pluto were counted. Analysis was between-subject, and test scores for each measure were analyzed against the control group. The results for all conditions are summarized in Table 1 below.

		Dinosaur			
		Test		Control	
Measure		Percentage	CI	Percentage	CI
SFP	Text	58%	14%		
	Picture	50%	15%		
	Total	54%	10%	47%	13%
Completion	Text	59%	15%		
	Picture	62%	15%		
	Total	60%	10%	74%	11%
Categorization	Text	49%	15%		
	Picture	56%	15%		
	Total	52%	11%	98%	3%
		Planet			
		Test		Control	
SFP	Text	62%	13%		
	Picture	48%	15%		
	Total	55%	10%	77%	10%
Completion	Text	27%	13%		
	Picture	15%	10%		
	Total	21%	8%	43%	13%
Categorization	Text	10%	8%		
	Picture	24%	12%		
	Total	17%	7%	61%	12%

Table 1: Statistics of the Empirical Study. The percentages indicate the amount of participants who answered according to the "incorrect" folk concept. Confidence intervals are 95 %.

We had very different results for DINOSAUR and PLANET. Starting with DINOSAUR, combining both the text and picture interventions, there neither was a significant difference between the

[10] For instance, for the Semantic Feature Hypothesis (H1), our pre-registered hypothesis read: There is no significant difference between the test groups and the control group on how often responses referring to being extinct (in the Dinosaur case) or responses referring to planethood (in the Pluto case) are noted in the SFP task.

test and the control group in the Completion task ($\chi^2 = 2.88$, $p = 0.089$) nor in the SFP task ($\chi^2 = 0.77$, $p = 0.38$) (Figure 2). In the Categorization task, we did, however, see a significant difference between the test and control group ($\chi^2 = 36.61$, $p < 0.001$).

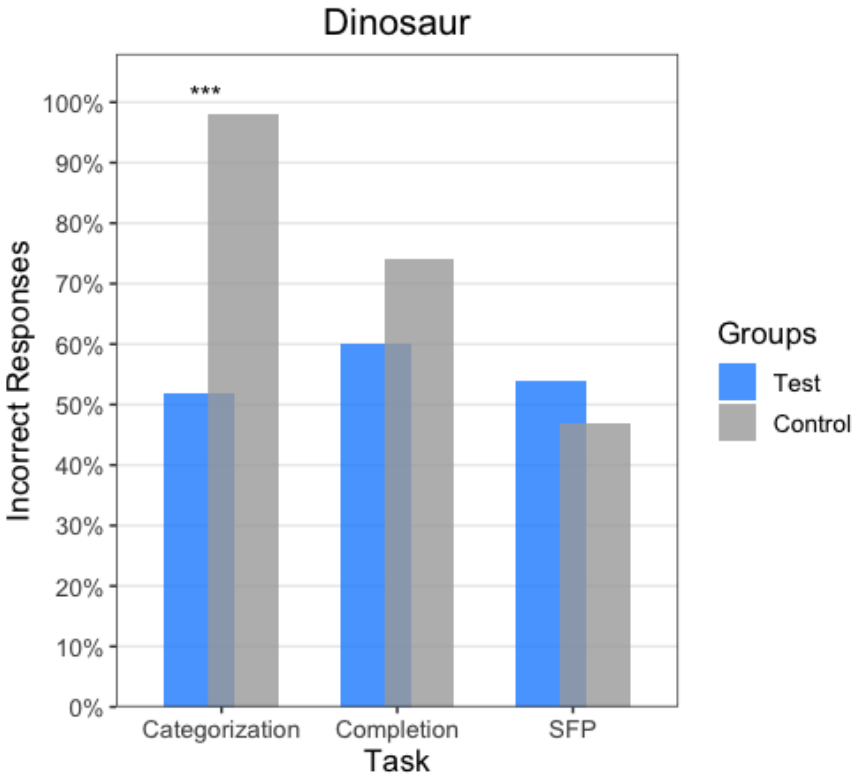

Figure 2: Results for DINOSAUR. The height of the bars show the number of incorrect responses for the three measurement tasks.

There was no difference between stimuli for DINOSAUR (Figure 3). Among those who saw the text intervention, the Categorization task ($\chi^2 = 35.63$, $p < 0.001$) was statistically significant, but the SFP ($\chi^2 = 1.27$, $p = 0.26$) and Completion tasks ($\chi^2 = 2.59$, $p = 0.11$) were not. Similarly, among those who saw the picture intervention, the Categorization task ($\chi^2 = 28.89$, $p < 0.001$) was statistically significant, but the SFP ($\chi^2 = 0.11$, $p = 0.74$) and Completion tasks ($\chi^2 = 1.70$, $p = 0.19$) were not. Exploratory analysis comparing the two stimuli found no significant difference between interventions in any of the three measures (Completion: $\chi^2 = 0.07$, $p = 0.79$; SFP: $\chi^2 = 0.53$, $p = 0.47$; Categorization: $\chi^2 = 0.44$, $p = 0.51$).

In PLANET, results were a bit more promising for conceptual engineering (see Figure 4). Combining both interventions, the test group was far less likely to say Pluto was a planet than the control group in the Completion task ($\chi^2 = 8.67$, $p = 0.003$), the SFP task ($\chi^2 = 8.05$, $p = 0.005$), and the Categorization task ($\chi^2 = 33.48$, $p < 0.001$). This means that all three null-hypotheses can be rejected as stated in the preregistration.

Collectively, the two sets of PLANET stimuli lowered incorrect responses on all three measures, suggesting conceptual revision occurred. However, this effect was uneven between the two interventions (Figure 5). The responses of participants who saw the picture stimulus were all statistically significant (Completion: $\chi^2 = 10.14$, $p = 0.001$; SFP: $\chi^2 = 9.98$, $p = 0.002$; Categorization: $\chi^2 = 14.97$, $p < 0.001$), whereas among participants who saw the text stimulus, only the Categorization task was significant (Completion: $\chi^2 = 2.93$, $p = 0.087$; SFP: $\chi^2 =$

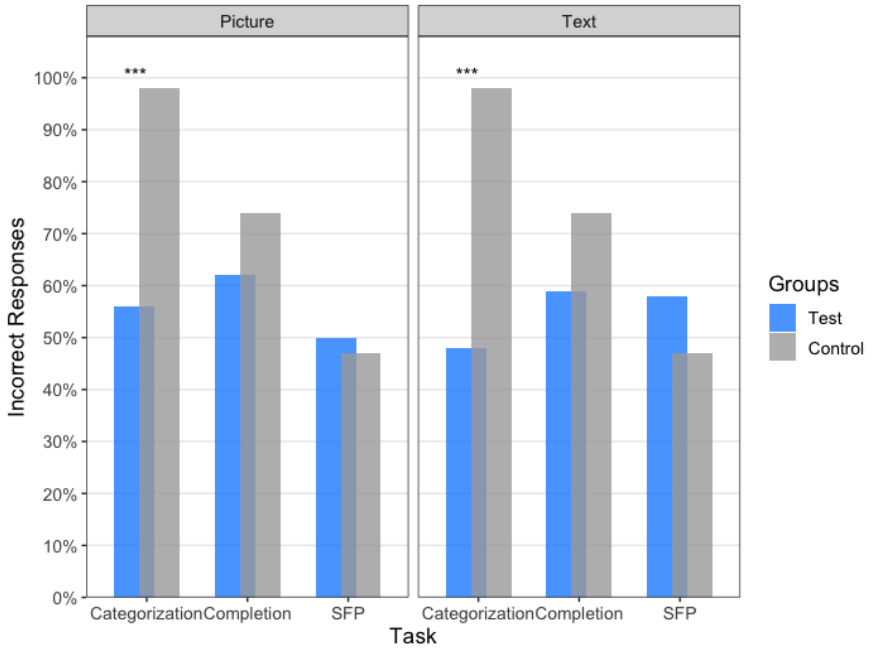

Figure 3: Results for DINOSAUR broken down by stimulus. The height of the bars show the number of incorrect responses for the three measurement tasks.

3.41, $p = 0.064$; Categorization: $\chi^2 = 31.47$, $p < 0.001$). Exploratory analysis comparing the two stimuli found no significant difference between interventions for any of the three measures (Completion: $\chi^2 = 2.27$, $p = 0.13$; SFP: $\chi^2 = 1.18$, $p = 0.18$; Categorization: $\chi^2 = 3.82$, $p = 0.051$).

3.6 Summary of the Study

Having developed a novel methodology, referred to as the masked time-lagged design, our empirical study was directed toward investigating the propagation of conceptual change. By writing stimuli for two distinct concepts, DINOSAUR and PLANET, in two different formats, Text and Text plus Pictures, we measured the extent to which people's concepts align with the latest scientific discoveries. Our findings suggest that our efforts yielded only limited success in altering people's concept of DINOSAUR. Conversely, we achieved a higher degree of success in revising people's concept of PLANET.

While the Categorization task was significantly different between control and test groups among every intervention for both concepts, this was not the case for the Completion and Semantic feature production tasks. In those measures, after our intervention, people were far less likely to say Pluto was a planet across all three measures compared to our control, but they were just as likely to say that dinosaurs are extinct. It is worth noting that while the length and format of the stimuli appeared to influence participants' responses with respect to PLANET, no comparable effect was discerned for the concept of DINOSAUR.

4 General Discussion

As discussed in Section 2, masked time-lagged designs allow conceptual engineers to directly test questions related to propagation, and so the findings about DINOSAUR and PLANET offer direct,

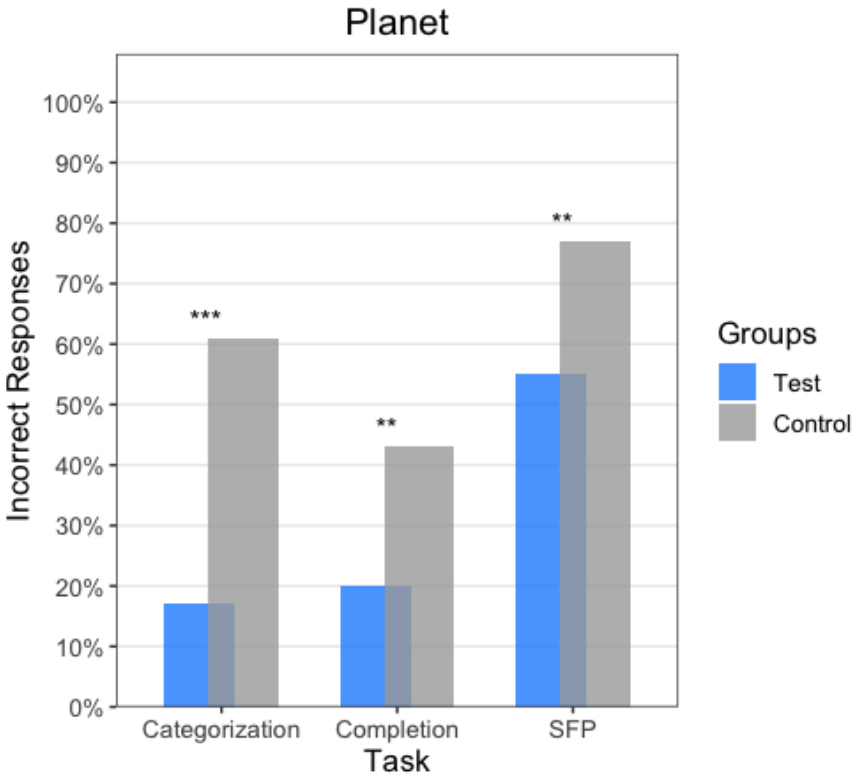

Figure 4: Results for PLANET. The height of the bars show the number of incorrect responses for the three measurement tasks.

albeit initial, insights into the factors that affect intentional conceptual revision. In Section 4.1, we will discuss how these findings offer first answers to global and local implementation questions. In Section 4.2, we address the objection that our data measure changes in beliefs as opposed to changes in concepts. In the final section, Section 4.3, we expand the discussion to other conceptual engineering frameworks. By using speaker meaning and semantic externalist frameworks as illustrations, we argue that masked time-lagged designs have utility beyond content internalist conceptual engineering frameworks.

4.1 Advancing Global and Local Implementation Questions

Philosophers have started to think about challenges and obstacles to implementing revised concepts and content (Nimtz, 2021; Pinder, 2017; Sterken, 2020). However, their approach is limited in two important respects. First, there is a growing recognition that the implementation challenge should be tackled empirically, but so far, the field has been slow to collect the necessary information. And those who have used empirical data, have not *directly tested* the way specific theories play out in the process of propagating concepts, meanings, or words (Fischer, 2020; Koslow, 2022; Landes, 2023; Machery, 2021). Second, conceptual engineers have not made a clear distinction between factors that exert influence on implementation at a global level (i.e., all concepts) and those that impact implementation at the local level (i.e., specific concepts). Undoubtedly, these two limitations are intertwined. In the absence of direct empirical data concerning the implementation of specific designed concepts, it will be difficult to disentangle what sort of factors affect the propagation of all concepts versus what factors are limited to

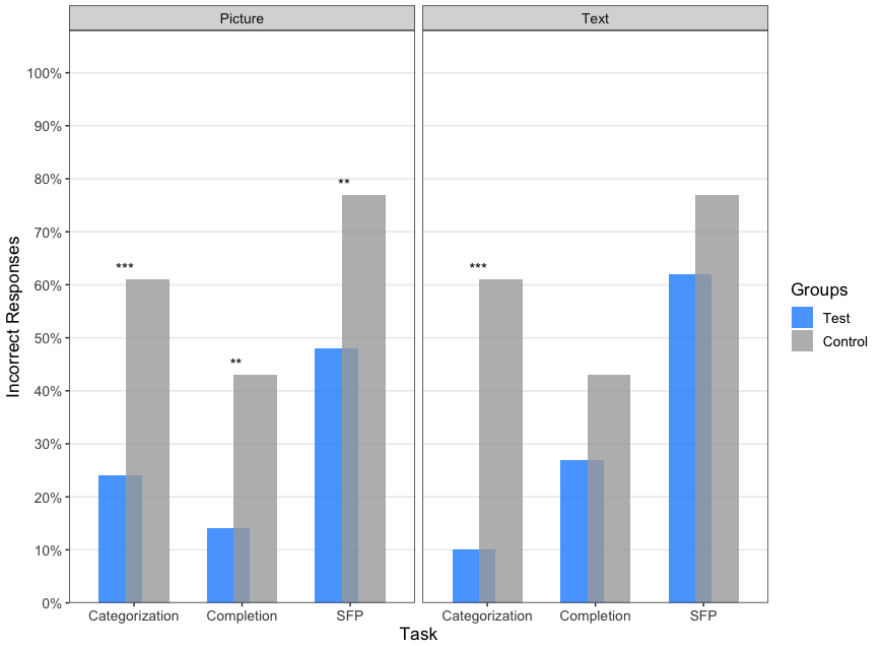

Figure 5: Results for PLANET broken down by stimulus. The height of the bars show the number of incorrect responses for the three measurement tasks.

specific concepts.

Indeed, care must be taken not to extrapolate too much from our data, as we have only examined two concepts. Moreover, PLANET and DINOSAUR share some commonalities that impose limitations on broader generalization. Perhaps most crucially, both PLANET and DINOSAUR are natural kind concepts. These concepts refer to classifications that mirror the inherent structure of the natural world, rather than being driven by the preferences of human beings (Bird & Tobin, 2023; Quine, 1969). The metaphysics of natural kinds is reflected in our own cognition, as people represent natural kinds as having a core essence responsible for observed superficial similarities (Gelman & Markman, 1987; Kornblith, 1997). In addition to natural kind concepts, there are also artificial kind concepts like TABLE (see Gelman, 2013; Rose & Nichols, 2019), abstract concepts like TRUTH, and social kind concepts like JANITOR, not to mention variations of each such as dual character concepts like ARTIST (Knobe et al., 2013; Reuter, 2019) and evaluative concepts like VANDAL (Eklund, 2011; Willemsen & Reuter, 2021). Differences between these kinds of concepts could potentially impact the feasibility of and relevant influences on conceptual propagation. Forthcoming studies will employ the masked time-lagged method to explore a greater range of concepts including artificial, abstract and social kinds. For now, we must confine our inferences to natural kind concepts only.

Nevertheless, similarities and differences between the two concepts, different stimuli, and three measures allow tentative conclusions about propagating conceptual change in general as well as specific conclusions about propagating revisions of both concepts. We can make more robust inferences than would otherwise be possible because the two concepts differ in significant and readily apparent ways. First, changes in DINOSAUR involve expanding the concept while changes in PLANET involve shrinking the concept (Liao & Hansen, 2023, see). Second, the change in PLANET is, at the time of data collection, far more in the cultural zeitgeist than the knowledge that birds are dinosaurs. Third, PLANET's intension changed as part of an explicit redefinition, while DINOSAUR's intension remains unchanged – roughly, the creatures that fall under the DINOSAUR evolutionary clade. Fourth, extinction is a salient property of dinosaurs

whereas clearing orbits of debris is not a salient feature of planets (and hence the slightly different measures in the Completion and SFP task).

Starting with global questions about what factors affect how natural kind concepts, in general, are revised, the differences between DINOSAUR and PLANET offer some fruitful initial insights. If both concepts behaved like DINOSAUR did, our results would have suggested that participants generally exhibit a high degree of resistance to conceptual revision. In such a scenario, we could have inferred that the process of implementing conceptually revised content is significantly more challenging than initially expected. This would be consistent with a particularly pessimistic reading of Machery's attractor view (Machery, 2021) – in which psychological factors naturally attract us to certain conceptual content – where attractors are the norm, not the exception. In contrast, if DINOSAUR had been readily revised, it would have suggested even relatively surprising changes could be easily propagated in adults once the effort is put in to propagate the changes.

Consistent with empirical theorizing by philosophers (Fischer, 2020; Koslow, 2022; Machery, 2021), our findings present a significantly more complex picture about how conceptual revision, in general, works. Across all four interventions, we observed that individuals can be effectively instructed to categorize objects based on new classification schemes with relatively short interventions. Spending a few minutes informing individuals why Pluto is not a planet and why birds are dinosaurs brought about substantial changes in participants' classification patterns. Across all interventions, when presented with an image of Pluto and asked about its planetary status or shown a picture of a seagull and queried about its classification as a dinosaur, participants successfully categorized the objects according to the taught schema. Consequently, it appears that the process of revising individuals' mental representations in order to elicit accurate responses to questions regarding the superordinate categorization of Pluto and birds is relatively straightforward.

Hence, if the objective of implementation entails inducing individuals to modify their explicit classification scheme, we can offer encouraging findings. To some engineers, this might be enough. However, the aspirations of many engineers extend beyond merely altering individuals' classifications. Their ultimate goal lies in transforming people's reasoning patterns, enabling them to draw inferences based on their newfound knowledge and perceive the world through different perspectives. Accomplishing this likely necessitates the modification of individuals' associations and implicit reasoning processes.

Changing implicit associations and default information retrieval looks to be much more difficult. While in the Categorization task, we observed a substantial reduction of approximately 50% in incorrect responses for both PLANET and DINOSAUR, the Completion task and the Semantic Feature Production task were mixed. In the Completion task and the Semantic Feature Production task, we identified a decrease of 20% for PLANET, while insignificant changes were observed for DINOSAUR. Consequently, it appears that modifying individuals' explicit classification schema is significantly more feasible than altering salient features and what information is retrieved by default.

Turning now to local questions—that is, questions about what factors influence the revision of specific concepts—the contrast of PLANET and DINOSAUR offers at least one clue to why we were successful in revising PLANET. As discussed, the revision of PLANET appears to be better known than the revision of DINOSAUR among adults. Perhaps the increased awareness of PLANET played an important role in its success, suggesting conceptual revision is something that should be built up to instead of something that can be done in a single intervention (see Carey, 2011).

Looking at the difference between the two PLANET interventions, a few more guesses can be drawn about why we were successful at revising PLANET, especially in the longer intervention that included pictures.[11] Adding pictures and detail to our stimuli did not by itself have an

[11] Some caution is required here, as the difference between the two PLANET stimuli was smaller than 15% across

effect on conceptual revision, as evidenced by lack of revision of DINOSAUR among participants who saw the longer intervention containing images. However, length and different format appear to have played *some* role in uptake of the revised PLANET, suggesting that multimodal formats may be a useful tool for some propagation projects.

Nevertheless, it is important to exercise caution when drawing conclusions at either the global or local level, as our interpretations are based on a limited and initial data set from two natural kind concepts. In any case, let us summarize a few highly tentative conclusions about propagation that could serve as the basis for future work:

- **Global Level:**
    1. Short explanatory stimuli are capable of revising natural kind concepts.
    2. For natural kind concepts, changing classification patterns appear to be relatively straightforward.
    3. For natural kind concepts, it is more difficult to modify associations and information retrieved by default than classification patterns.
    4. Changes involving splitting and/or shrinking concepts may prove easier than changes involving combining and/or expanding concepts (see Fischer, 2020)
- **Local Level:**
    1. PLANET's ease of revision may have depended on previous exposure.
    2. Visual aids may have helped revise PLANET, but this appears to be a feature of PLANET, not visual aids' use in general.

Up until this point, our discussions and interpretations have assumed that our measures capture conceptual revision in PLANET. In the next section, we respond to the objection that in fact the measures capture something non-conceptual.

4.2 Objection: This Is Only Belief Revision

How confident can we be that we observed a conceptual change instead of a belief-based change? On content internalist frameworks, the distinction between beliefs and concepts can be quite subtle. Indeed, some of the conceptual engineers discussed above take concepts to be belief-like in that they are stable bodies of information (Fischer, 2020; Machery, 2017). On such an account, what makes concepts unique is that they are retrieved quickly, automatically, and independent of context, which means they affect inferences, categorizations, and other cognitive functions (Machery, 2017, 210-211). For example, someone may correctly believe that—due to the geography of French Guyana—France's longest border is with Brazil, but not have this information elicited quickly, automatically, and independently of context when they read the word "France". Instead, the information retrieved by default may involve Paris, the French flag, France's European borders, and/or facts about France's economy. This retrieved information, as it changes from person to person, is someone' concept of France on such an account, while the belief about France's border with Brazil is not.

Is there reason to think that the measures described above actually captured changes in concepts as opposed to changes in beliefs? In the experiment, the three measures followed two patterns. The Categorization task – a multiple choice question about scientific kinds – changed dramatically, regardless of stimuli or concept. The other two tasks, Completion and SFP, both of which were open-ended free responses, only significantly changed in response to one of the four stimuli. This suggests that the Categorization task is picking up on different, more plastic, phenomena than the SFP and Completion tasks. Categorization is taken to be one of the key

all three measures.

functions of concepts (Bloch-Mullins, 2018; Machery, 2009), and so we take the default reading of these findings to be related to concepts. However, one significant possibility, not ruled out here, is that the more plastic phenomena are beliefs as opposed to anything properly considered concepts. On such a reading of the results, participants are changing how they categorize Pluto and seagulls, not because of any change in conceptual content, but because of belief-level phenomena, such as the belief in the facts *Pluto is not a planet* and *birds are dinosaurs*. That is, the Categorization task, despite the intention behind its inclusion, is acting as more of a test of scientific knowledge rather than how they automatically classify objects.

However, beliefs are not the only thing that would explain the increased plasticity of the Categorization task over the other two tasks. Another possibility is that the Categorization measure is more sensitive to particular kinds of conceptual change than the SFP and Completion measure. For example, there are multiple terms that are polysemous in that they have a loose folk meaning and a more precise scientific meaning. In the everyday sense of "fruit", tomatoes are not fruit, but in the technical sense of "fruit" tomatoes are fruit (Engelhardt, 2019; Landes, 2021; Machery & Seppälä, 2011). Being able to understand the senses in which tomatoes are and are not fruits requires two distinct concepts, specifically fruit as a culinary/social kind and fruit as a botanical kind. A possible outcome of attempting to revise PLANET and DINOSAUR is that, like FRUIT, people end up with two concepts – folk and scientific counterparts to each other – which the Categorization task is for some reason sensitive to in a way the SFP or Completion tasks are not.

That said, even if we grant that the Categorization task is merely picking up on changes of belief, this skepticism does not extend to the other two measures. For one, the SFP task does not seem to be picking up on beliefs because the information it collects is not obviously propositional and so not obviously belief-based in the way the Categorization task may be. The SFP task asks participants to list what features come to mind related to some thing. Therefore, the SFP task measures salient properties of the concept as opposed to mere beliefs about the kinds (Machery, 2009; McRae et al., 2005). Because there is significant disagreement in cognitive science and philosophy of mind about how concepts are structured (see Bloch-Mullins, 2018; Casasanto, 2015; Quilty-Dunn, 2021; Vicente & Martínez Manrique, 2016), the SFP task might not measure core conceptual content, but the SFP task nonetheless appears to measure some prominent aspect of the concept. Moreover, because the results of the Completion Task closely follow the results of the SFP task in all interventions, the responses in the Completion and SFP tasks appear to have a common etiology.

4.3 Expanding the Method to Other Conceptual Engineering Frameworks

The experiment and resulting data has been discussed through the lens of content internalist conceptual engineering, which understands the target of conceptual engineering to be concepts and understands concepts to be token psychological entities. While this is currently a popular conceptual engineering framework, it is by no means the only one on the market. Other frameworks include those that propose that the goal of conceptual engineering is semantic meaning (Cappelen, 2018; Sterken, 2020), speaker meaning (Pinder, 2020, 2021), or conceptual content that is grounded in facts external to individuals (Haslanger, 2020; Sawyer, 2020; Scharp, 2013). In this section, we discuss the significance of the masked time-lagged method to other frameworks. We specifically focus on two prominent families of frameworks that focus on language instead of concepts, namely speaker meaning accounts and semantic externalist accounts.

The methods described above can easily be expanded to speaker meaning accounts of conceptual engineering. Speaker meaning accounts of conceptual engineering take the goal of conceptual engineering to be to change what people take themselves to mean by the words that they utter (Pinder, 2020, 2021). Speaker meaning conceptual engineers do not aim to change what a word means in some broad, interpersonal sense. Instead they target the intentions and

beliefs speakers have related to using a specific term (although on some frameworks, linguistic intentions, linguistic beliefs, and meaning go together). Thus, similar to content internalist frameworks, speaker meaning conceptual engineering places the target of conceptual engineering in token psychological states. The best way to test what speakers intend to mean by a term is to have them produce speech acts using the term and coding how the term is used. For that reason, the Completion task may be a suitable measure, although multi-sentence productions would provide richer sources of data about speaker intentions.

When it comes to applying the findings of our studies to semantic externalist frameworks of conceptual engineering, the extension of this study is not as straightforward because semantic externalists do not place the focus of their theories on token psychological states. Nonetheless, when we dig into the role token psychological states generally play in semantic externalist accounts of meaning, we can see that not only can masked time-lagged designs answer questions about how to spread true beliefs about meaning or reference changes that has already occurred, it can also answer questions about how best to use our limited ability to change externalist meaning or reference.

Externalists will not want to say that our experiment revised the concept PLANET or meaning of "planet". They will instead contend either that a) "planet" always excluded Pluto (Ball, 2020; Kripke, 1980, e.g.,), b) the concept or meaning was revised in 2006 by the International Astronomical Union, or c) revision of "planet" or PLANET occurred at some later date when, for example, the new linguistic norm became the dominant one among English speakers (Evans, 1973).[12] Even if semantic externalists contend our experimental data does not reveal anything about how to change meaning, the experiment does shed light on how conceptual engineers can help individuals become aware of changes in meaning. While on many of these frameworks, the meaning of people's utterances containing"planet" has changed since 2006, these sorts of changes can happen without people's awareness (see Pollock, 2021; Wikforss, 2015). Thus, externalists can still view the experiment as testing phenomena related to propagation. While the experiment did not change meaning, it propagated true beliefs about the meaning of "planet" in light of changes that have occurred.

The masked time-lagged design can do more than show how to spread true semantic beliefs related to revisions, however. Semantic externalists can use the masked time-lagged design proposed here to determine the most effective ways to change meaning—even if meaning is grounded externally to individual speakers. To see this, we need to focus on what externalists take ground semantic facts. For semantic externalists, the meaning of a word-type or utterance-type (that is, what a word means, in general) is determined by some combination of linguistic facts and non-linguistic facts. While they disagree about which linguistic and non-linguistic facts matter, many, if not most non-linguistic facts, such as the joints of natural kinds, are outside of conceptual engineer controls (Cappelen, 2018). This limits the scope of what sort of changes are possible. Nonetheless, many meaning-determining linguistic facts are within the scope of empirical methods (Koslow, 2022; Sterken, 2020; Thomasson, 2021), and a subset of those meaning-determining facts are things conceptual engineers have (limited) influence over. This is because many linguistic facts are grounded in, among other things, intentions and beliefs of individual speakers—that is, token psychological states (see Nimtz, 2021).

To illustrate how global and local questions about intentional semantic externalist meaning change can be tested, consider a version of an Evans-style metasemantics, where the reference of a word-token is determined by the dominant causal source of all the uses of that word-token (Evans, 1973; Leckie & Williams, 2019). On this view, my use of "cat" refers to cats because most of the people around me use "cat" in a way that traces back to cats. What things there are in the world to be a dominant causal source is by and large outside of conceptual engineers' control. Nonetheless, this Evan-style view locates part of the ground of reference in the collective uses of a community of speakers. The collective use of a community of speakers

[12] That said, the data in the control group raises doubts that it has in fact become the dominant use of "planet".

is, at rock bottom, a large number of token beliefs and practices about language. Therefore, changing meaning by changing psychological states is possible if enough psychological states change to disrupt the current dominant source (see Sterken, 2020). How to best use resources to disrupt the dominant source is something experimental methods can study. In fact, this will look similar to the speaker meaning account discussed above, as it will involve studying how people shift the way they speak in light of an intervention.[13]

The Evans-style view is merely illustrative of the larger constellation of semantic externalist theories. Semantic externalists generally take some sort of linguistic norm or practice to play a role in determining the meaning of word-types, such as the collective use of a causal historical chain (Kripke, 1980) or linguistic conventions (Lewis, 1969). Linguistic norms and practices are partially or wholly composed of linguistic beliefs and intentions, although the way linguistic beliefs and intentions combine to form norms may be extremely complicated (Lewis, 1969; Nimtz, 2021). This means that semantic externalists can study how to change meaning by studying how to bring about collective shifts in linguistic beliefs and intentions. Here we find ourselves in the realm of token psychological entities, and so semantic externalists can test what factors influence meaning or reference change using variations of the above empirical masked time-lagged design. Granted, no amount of experiments will make externalists semantically omnipotent – they will still be limited in what they can change by linguistic and non-linguistic facts outside of their control. This is not a problem unique to externalism, however. Internalist conceptual engineers will also be limited by features of our cognition that will prevent certain proposed changes from taking hold (Fischer, 2020; Machery, 2021). Nonetheless, the more we learn through experiments about how the grounds of conceptual content or semantic facts change, the better conceptual engineers will be at revising or replacing conceptual content or semantic facts.

5 Conclusion

Many philosophers have recently begun theorizing that the aim of philosophy should be to develop revised concepts and spread those revised concepts to the right people. Little is understood about how the propagation of such revised concepts could be spread or even measured. In this project, we demonstrated how a masked time-lagged design could be used to directly test the revision of participants' concepts. We specifically attempted to revise DINOSAUR and PLANET, finding tantalizingly mixed results. While our stimuli seems to have successfully revised PLANET in participants, the same cannot be straightforwardly said for DINOSAUR. Therefore, further work is needed to study the process of propagating conceptual changes for the ends of conceptual engineering now that an empirical framework is in place.

References

Ball, D. (2020). Relativism, Metasemantics, and the Future. *Inquiry: An Interdisciplinary Journal of Philosophy*, 41.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, *11*, 211–227.

Belleri, D. (2021). On pluralism and conceptual engineering: Introduction and overview. *Inquiry*, *0*(0), 1–19. doi: 10.1080/0020174X.2021.1983457

Bird, A., & Tobin, E. (2023). Natural Kinds. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University.

[13] This is assuming the framework is productivist. An interpretationalist semantic externalist could study how interpretations of statements change after intervention. See Simchen (2017).

Bloch-Mullins, C. L. (2018). Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts. *The British Journal for the Philosophy of Science*, *69*(3), 605–632. doi: 10.1093/bjps/axw039

Brusatte, S. (2017, January). *How Birds Evolved from Dinosaurs.* https://www.scientificamerican.com/article/how-birds-evolved-from-dinosaurs/. doi: 10.1038/scientificamerican0117-48

Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, *32*(3), 377–394.

Cappelen, H. (2018). *Fixing Language.* Oxford University Press. doi: 10.1093/oso/9780198814719.001.0001

Carey, S. (2011). *The origin of concepts* (1. iss. Oxford Univ. paperback ed.). Oxford: Oxford Univ. Press.

Casasanto, D. (2015). All concepts are ad hoc concepts. In *The conceptual mind: New directions in the study of the concepts* (pp. 543–566). MIT press.

Chang, K. (2022, January). Is Pluto a Planet? What's a Planet, Anyway? *The New York Times.*

Conrad, F. G., Schober, M. F., & Schwarz, N. (2014, September). Pragmatic Processes in Survey Interviewing. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology* (p. 0). Oxford University Press. doi: 10.1093/oxfordhb/9780199838639.013.005

Cullen, S. (2010). Survey-Driven Romanticism. *Review of Philosophy and Psychology*, *1*(2), 275–296.

Eklund, M. (2011, March). What are Thick Concepts? *Canadian Journal of Philosophy*, *41*(1), 25–49. doi: 10.1353/cjp.2011.0007

Engelhardt, J. (2019, July). Linguistic labor and its division. *Philosophical Studies*, *176*(7), 1855–1871. doi: 10.1007/s11098-018-1099-2

Evans, G. (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume*, *47*(1), 187–208.

Fischer, E. (2020). Conceptual control: On the feasibility of conceptual engineering. *Inquiry*, 1–29.

Fischer, E., & Engelhardt, P. E. (2019, July). Lingering stereotypes: Salience bias in philosophical argument. *Mind & Language*, mila.12249. doi: 10.1111/mila.12249

Fischer, E., Engelhardt, P. E., & Herbelot, A. (2015). Intuitions and illusions: From explanation and experiment to assessment. In E. Fischer & J. Collins (Eds.), *Experimental Philosophy, Rationalism, and Naturalism. Rethinking Philosophical Method* (pp. 259–292). Routledge.

Gelman, S. A. (2013, September). Artifacts and Essentialism. *Review of Philosophy and Psychology*, *4*(3), 449–463. doi: 10.1007/s13164-013-0142-7

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child development*, 1532–1541.

Giora, R. (2003). *On our mind: Salience, context, and figurative language.* New York Oxford: Oxford University Press.

Griffiths, P., E., & Machery, E. (2008). Innateness, canalization, and 'biologicizing the mind'. *Philosophical Psychology*, *21*(3), 397–414.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of verbal learning and verbal behavior*, *18*(4), 441–461.

Haslanger, S. (2000). Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs*, *34*(1), 31–55. doi: 10.1111/0029-4624.00201

Haslanger, S. (2020). How not to change the subject. *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation*, 235–259.

IAU. (2006). *Definition of a Planet in the Solar System* (Tech. Rep. No. Resolution B5). Paris: International Astronomical Union.

Isaac, M. G. (2021). Which concept of concept for conceptual engineering? *Erkenntnis*, 1–25.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*.

Jones, J. A. (2013). From Nostalgia to Post-Traumatic Stress Disorder: A Mass Society Theory of Psychological Reactions to Combat. *Inquiries Journal*, *5*(02).

Jorem, S. (2021). Conceptual Engineering and the Implementation Problem. *Inquiry: An Interdisciplinary Journal of Philosophy*, *64*(1-2), 186–211. doi: 10.1080/0020174x.2020.1809514

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257.

Koch, S. (2021). The externalist challenge to conceptual engineering. *Synthese*, *198*, 327–348. doi: 10.1007/s11229-018-02007-6

Kornblith, H. (1997). *Inductive interference and its natural ground: An essay in naturalistic epistemology*. Cambridge, Mass.: MIT Press.

Koslow, A. (2022). Meaning change and changing meaning. *Synthese*, *200*(2), 1–26.

Kripke, S. (1980). *Naming and Necessity*. Oxford, UK ; Cambridge, USA: Blackwell Publishers.

Landes, E. (2021). *Philosophy and Philosophy: The Subject Matter and the Discipline*. Unpublished doctoral dissertation, University of St. Andrews.

Landes, E. (2023). How Language Teaches and Misleads: "Coronavirus" and "Social Distancing" as Case Studies. In M. G. Isaac, S. Koch, & K. Scharp (Eds.), *New Perspectives on Conceptual Engineering*.

Laurence, S., & Margolis, E. (1999). Concepts and Cognitive Science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core Readings* (pp. 3–81). MIT Press.

Leckie, G., & Williams, J. R. G. (2019). Words by convention. *Oxford Studies in Philosophy of Language*, *1*, 73–98.

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, Mass: Harvard University Press.

Liao, S.-y., & Hansen, N. (2023). 'Extremely Racist'and 'Incredibly Sexist': An Empirical Response to the Charge of Conceptual Inflation. *Journal of the American Philosophical Association*, *9*(1), 72–94.

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Machery, E. (2017). *Philosophy within its proper bounds* (First ed.). Oxford, United Kingdom: Oxford University Press.

Machery, E. (2021). A new challenge to conceptual engineering. *Inquiry*, *0*(0), 1–24. doi: 10.1080/0020174X.2021.1967190

Machery, E., Griffiths, P., Linquist, S., & Stotz, K. (2019). Scientists' concepts of innateness: Evolution or attraction? *Advances in experimental philosophy of science*, 172–201.

Machery, E., & Seppälä, S. (2011). Against hybrid theories of concepts. *Anthropology and Philosophy*, *10*, 99–126.

Margolis, E., & Laurence, S. (2007). The Ontology of Concepts—Abstract Objects or Mental Representations? *Noûs*, *41*(4), 561–593. doi: 10.1111/j.1468-0068.2007.00663.x

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

Muehlenhard, C. L., & Peterson, Z. D. (2011). Distinguishing Between Sex and Gender: History, Current Conceptualizations, and Implications. *Sex Roles*, *64*(11), 791–803. doi: 10.1007/s11199-011-9932-5

Napolitano, M. G., & Reuter, K. (2021). What is a Conspiracy Theory? *Erkenntnis*. doi: 10.1007/s10670-021-00441-6

NASA. (2023, July). *What is a Planet?* https://science.nasa.gov/solar-system/planets/what-is-a-planet/.

Nimtz, C. (2021). Engineering concepts by engineering social norms: Solving the implementation challenge. *Inquiry*, *0*(0), 1–28. doi: 10.1080/0020174X.2021.1956368

Pinder, M. (2017). Does Experimental Philosophy Have a Role to Play in Carnapian Explication? *Ratio*, *30*(4), 443–461. doi: 10.1111/rati.12164

Pinder, M. (2020). Conceptual engineering, speaker-meaning and philosophy. *Inquiry*, 1–15.

Pinder, M. (2021). Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning. *Mind*.

Poling, D. A., & Evans, E. M. (2004). Are dinosaurs the rule or the exception? Developing concepts of death and extinction. *Cognitive Development*, *19*, 363–383. doi: 10.1016/j.cogdev.2004.04.001

Pollock, J. (2021). Content internalism and conceptual engineering. *Synthese*, *198*(12), 11587–11605. doi: 10.1007/s11229-020-02815-9

Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language*, *36*(1), 158–185. doi: 10.1111/mila.12328

Quine, W. V. (1969). Natural kinds. In *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday* (pp. 5–23). Springer.

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, *14*(1), e12557. doi: 10.1111/phc3.12557

Reuter, K., & Brun, G. (2022). Empirical Studies on Truth and the Project of Re-engineering Truth. *Pacific Philosophical Quarterly*, *103*(3), 493–517. doi: 10.1111/papq.12370

Rose, D., & Nichols, S. (2019). Teleological Essentialism. *Cognitive Science*, *43*(4), e12725. doi: 10.1111/cogs.12725

Sawyer, S. (2020). Truth and Objectivity in Conceptual Engineering. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Scharp, K. (2013). *Replacing Truth*. Oxford, New York: Oxford University Press.

Schwarz, N. (1995). What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation. *International Statistical Review / Revue Internationale de Statistique*, *63*(2), 153–168. doi: 10.2307/1403610

Scott-Phillips, T., Blancke, S., & Heintz, C. (2018). Four misunderstandings about cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews*, *27*(4), 162–173.

Shields, M. (2021). Conceptual domination. *Synthese*, *199*(5-6), 15043–15067.

Shields, M. (2023, July). Conceptual Engineering, Conceptual Domination, and the Case of Conspiracy Theories. *Social Epistemology*, *37*(4), 464–480. doi: 10.1080/02691728.2023.2172696

Shtulman, A., & Calabi, P. (2013). Tuition vs. intuition: Effects of instruction on naive theories of evolution. *Merrill-Palmer Quarterly*, *59*(2), 141–167.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Simchen, O. (2017). *Semantics, metasemantics, aboutness* (First edition ed.). Oxford: Oxford University Press.

Spelke, E. S., & Tsivkin, S. (2001). Initial knowledge and conceptual change: Space and number. *Language acquisition and conceptual development*, 70–100.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Oxford Blackwell.

Sterken, R. K. (2020). Linguistic Interventions and Transformative Communicative Disruption. In H. Cappelen, D. Plunkett, & A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Thomasson, A. L. (2021). Conceptual engineering: When do we need it? How can we do it? *Inquiry*, *0*(0), 1–26. doi: 10.1080/0020174X.2021.2000118

Vicente, A., & Martínez Manrique, F. (2016). The Big Concepts Paper: A Defence of Hybridism. *The British Journal for the Philosophy of Science*, *67*(1), 59–88. doi: 10.1093/bjps/axu022

Wikforss, A. (2015). The insignificance of transparency. *Externalism, self-knowledge, and skepticism*, 2015.

Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*, *10*(2), 135–146. doi: 10.1002/tht3.488