Pain Judgments and T-Tests[1]

Justin Sytsma

Abstract: What is pain? Perhaps surprisingly the standard answer to this question among philosophers does not derive from research in biology or other sciences, but from claims about common sense and thought experiments intended to draw out our intuitions about the nature of pain. This raises a number of issues, among them the question of whether philosophers' claims about the commonsense conception of pain are accurate. In this chapter, I'll explore some of the empirical research that has been done on this question in recent years, focusing on the claim that common sense tells us that there can be no unfelt pains. In doing so, I'll walk through several sets of studies, introducing the empirical research process and illustrating the use of one type of statistical tool—t-tests.

1. Introduction

The present chapter has two main goals. The first is to introduce you to a powerful family of statistical tests—t-tests. To do this, I'll walk you through three main case studies exploring the question of whether lay people tend to believe that there can be unfelt pains. Each of these will introduce one main type of t-test—*one sample t-tests*, *independent samples t-tests*, and *paired samples t-tests*. Coupled with further studies bearing on the ordinary conception of pain, I hope that these studies will make it plausible to you that the ordinary view is rather different than many philosophers have supposed. This is the second goal of this chapter.

[1] Penultimate draft of Chapter 2 in *Experimental Philosophy for Beginners: A Gentle Introduction to Methods and Tools* by Stephan Kornmesser, Alexander Max Bauer, Mark Alfano, Aurélien Allard, Lucien Baumgartner, Florian Cova, Paul Engelhardt, Eugen Fischer, Henrike Meyer, Kevin Reuter, Justin Sytsma, Kyle Thompson, and Marc Wyszynski.

Here is how I will proceed. In the next section I lay out the standard philosophical view about pains, the assumptions that have been made about the ordinary conception of pain, and the challenge that has been raised by some experimental philosophers. I begin to explore this challenge in more depth in Section 3, laying out the primary case study for this chapter—Study 3 in Sytsma (2010a)—and illustrating the use of one sample t-tests. To do this I detail the empirical research process I follow (§3.1), present the background motivation for the study (§3.2), and then detail the research question driving it (§3.3), the research design (§3.4), how I constructed the instrument (§3.5) and conducted the study (§3.6), and finally how I analysed (§3.7) and interpreted (§3.8) the results. In Section 4, I then consider some potential worries about this study, detailing a selection of further studies that address them, including the second and third main case studies, which illustrate independent samples t-tests (§4.1) and paired samples t-tests (§4.4) respectively. Finally, in Section 5, I consider a worry that often arises in using t-tests: when we run multiple tests we *might* need to apply a correction.

2. Philosophical Background: The Standard View of Pain in Philosophy

The standard view among philosophers is that pains belong to the mind, not the body. When you cut your finger, for example, this view contends that the sharp pain you feel in your finger is in fact a property of your mind, not your finger. Put another way, the standard view holds that there is no distinction to be drawn between *pain* and *feeling pain*. More carefully, the standard view is that pains are properties of conscious mental states. Consciousness is a notoriously tricky notion, with the term being used to pick out a number of different phenomena (Block 1995). Philosophers are most often interested in a specific sense of consciousness, however—what is typically termed ‘phenomenal consciousness’. Researchers are not in complete agreement about

how to understand this notion, but the standard idea is that phenomenally conscious mental states are those states for which there is 'something it is like' (Nagel 1974) to be in them, where this is meant to pick out a diverse range of states that are thought to have a distinctive 'feel'. These feels are associated with the mental states, typically being taken to be properties of them, and understood in this way they are known as *phenomenal qualities* (or *qualia* for short).

Phenomenally conscious mental states and their associated qualia are typically drawn out by listing examples, detailing ordinary perceptual, bodily, or emotional experiences and assuming that the distinctive qualities we're aware of in these episodes are phenomenal qualities. For instance, Michael Tye (2021) opens his *Stanford Encyclopedia of Philosophy* entry on 'Qualia' in just this way:

> I run my fingers over sandpaper, smell a skunk, feel a sharp pain in my finger, seem to see bright purple, become extremely angry. In each of these cases, I am the subject of a mental state with a very distinctive subjective character. There is something it is *like* for me to undergo each state, some phenomenology that it has.

Focusing on pains, the idea is that pains are properties of mental states and that what makes the mental states states of that type is just those felt properties—the way that they feel to the person who has them. As such, according to the standard view in philosophy there is no appearance–reality distinction to be drawn for pains. In other words, to have the appearance of pain—for someone to experience pain—is for that person to have a pain. And, likewise, for someone to have a pain is for them to experience pain. Thus, the standard view endorses the following two conditionals:

> If a person has a pain, then she feels that pain.
> If a person feels a pain, then she has that pain.

These conditionals have a number of implications. Most important for present purposes is that the first conditional excludes the possibility of there being unfelt pains: Since having a pain implies feeling pain, there could be no pain that is not felt.

The idea that there can be no unfelt pains has a long and venerable history in philosophy. More than two-hundred years ago, Thomas Reid (1785, 1.1.12) asserted that 'pain, when it is not felt, has no existence'. In the previous century, Saul Kripke (1980, 152) expressed the underlying view succinctly when he wrote that pain 'is picked out by the property of pain itself, by its immediate phenomenological quality'. And many contemporary philosophers of pain have continued this tradition, often clearly noting that they take the standard view to follow from our commonsense conception of pain. For instance, Murat Aydede (2005a, x) asserts that 'it is part of the commonsense conception' that pains 'can't exist without someone's feeling them'. Indeed, he holds that 'there is an air of paradox when someone talks about unfelt pains', noting that 'one is naturally tempted to say that if a pain is not being felt by its owner then it does not exist' (2005b, 4). Similarly, Christopher Hill (2009, 169–170) expresses the standard view for a range of bodily sensations, holding that the way we talk about them 'presupposes that the appearance of a bodily sensation is linked indissolubly to the sensation itself', and asserting that 'this is true, in particular, of our thought and talk about pain'.

But should we accept the standard view of pain, including that there can be no unfelt pains? As we've just seen, this view is typically defended not by explicit philosophical arguments or empirical data, but by appeal to how we are said to commonly think and talk about pain. While this is often left implicit, the idea seems to be that the commonsense view should have default status—that we should assume the commonsense view in our philosophical discussions until or unless we have good reason to abandon it. Relatedly, the standard view is

sometimes supported by appeals to intuition, laying out a hypothetical scenario involving pain and claiming that what *we* want to say about it coincides with the standard view. For example, in his *Stanford Encyclopedia of Philosophy* entry on 'Pain', Aydede (2009) offers the following thought experiment in support of the claim that there can be no unfelt pains, contrasting the standard view with an alternative that treats pains as properties of body parts:

> Suppose that we do in fact attribute a *physical* condition, call it *PC*, when we attribute pain to a body part, and that *PC* is the perceptual object of such experiences. So, for instance, John's current excruciating experience (call this *E*) is caused by and represents a physical condition in his thigh. From this it would follow that
>
> > (a) John would not have any pain if he had *E*, but no *PC* in his thigh (as in the case of, for instance, phantom limb pains and centrally generated chronic pains such as sciatica),
>
> and, conversely,
>
> > (b) John would have pain if he had *PC* but no *E* (as would be the case, for instance, if he had taken absolutely effective painkillers or his thigh had been anesthetized).
>
> But these statements are intuitively incorrect. They appear to clash with our ordinary or dominant concept of pain, which seems to track the experience rather than the physical condition.

Similarly, Hill (2009, 171) appeals to intuitions about a thought experiment to support the standard view over the alternative picture:

> If we were fully committed to the picture, we would be prepared to consider it epistemically possible that an injured soldier actually has a severe pain, despite his professions to the contrary, but that there is something wrong with the mechanisms in his brain that support attention, and that this is preventing the pain from penetrating the threshold of consciousness. When I have asked informants to assess the likelihood of this scenario, however, they have all been inclined to dismiss it as absurd.

In such passages, Aydede and Hill do not merely make claims about their own intuitions about the scenarios they lay out, but instead treat their intuitions as being general and, therefore, take them to tell us about the commonsense conception of pain.

But are such claims about pain intuitions and the commonsense conception of pain accurate? Is it true, for instance, that common sense rules out the possibility of unfelt pain? Across an extended series of papers, Kevin Reuter and I have argued that it is not, supporting this contention with a diverse array of empirical studies that suggest that lay people (i.e., non-philosophers) do not tend to have the intuitions that advocates of the standard view claim. Indeed, we find evidence that people often treat pains as properties of body parts and, doing so, happily countenance the possibility of having unfelt pains.[2] This includes background work in experimental philosophy of mind[3] arguing that lay people do not tend to share the philosophical concept of phenomenal consciousness in the first place (and so don't treat pains as phenomenal qualities),[4] work using tools from corpus linguistics that indicates that the way people ordinarily talk about pains involves an appearance-reality distinction,[5] and work using the types of questionnaire methods that I'll focus on in this chapter. The latter includes three papers using questionnaire methods to test whether people believe that unfelt pains are possible—Sytsma (2010a), Sytsma and Reuter (2017), as well as Reuter and Sytsma (2020)—which I'll focus on in the present chapter.

[2] Earlier work in this area focused on the shortcomings of the standard view, drawing out that a prominent thread in ordinary thinking about pains treated them as properties of body parts (e.g., Reuter et al. 2014, Reuter et al. 2019, Kim et al. 2016, Reuter 2017). More recent work in this area has focused more on simply understanding ordinary thinking about pain, often emphasizing that this is complicated and showing that people sometimes also treating pains as mental states. Disagreement remains on the extent of bodily versus mental aspects in people's thinking about pain and how they relate (e.g., Borg et al. 2020, Liu 2020, Liu 2023, Salomons et al. 2021, Coninx et al. 2023, Goldberg et al. forthcoming).

[3] See Sytsma (2014), Sytsma and Buckwalter (2016, Part II.C), and Phelan (forthcoming) for introductions to this area of experimental philosophy.

[4] For a short overview, see Machery and Sytsma (2011), for more extended review see Sytsma (2010b, 2016), Gonnerman (2018). For a few recent studies dealing with experimental philosophy of consciousness, see Díaz (2021), Fischer and Sytsma (2021), and Gregory et al. (2022).

[5] Corpus linguistics collects and analyses pre-existing 'real world' data on the use of words (McEnery and Wilson 2002, McCarthy and O'Keefe 2010). Philosophers have increasingly called on such methods, ranging from simple web searches, to more balanced corpora, to sophisticated computational approaches. See Chapter 5 of this volume for an extended illustration, and Bluhm (2016), Sytsma et al. (2019), Caton (2020), Ulatowski et al. (2020), as well as Fischer and Sytsma (forthcoming) for further examples and discussion. These tools are employed in assessing the standard view of pain in Reuter (2011) as well as Sytsma and Reuter (2017). See Sytsma and Fischer (forthcoming) for a recent study applying corpus methods to related an issue in experimental philosophy of consciousness.

This focus represents a divergence from the typical chapter in this volume, which illustrates a type of method or analysis in experimental philosophy by walking readers through a single-case study. While there is much to be said for this approach, it is not feasible for the present chapter. The reason is that t-tests are not a single type of test, but a family of tests, with the different members of this family being applied in different circumstances. Thus, to give a reasonable introduction to t-tests and when each type applies, I'll need to walk you through multiple studies. There are three main types of t-tests that you are likely to encounter in the literature or want to apply in your own reasearch—*one sample t-tests*, *independent samples t-tests*, and *paired samples t-tests*. Illustrating these three types of t-tests forms the heart of the present chapter and I'll present a primary case study concerning the possibility of unfelt pains for each. I begin in the next section by detailing Study 3 from Sytsma (2010a), using this to illustrate the basic research process and introducing a first use of t-tests in statistical analysis—*one sample t-tests*.

3. Illustrating One Sample T-tests

This section will provide the most detailed case study in the chapter. The goal will be to use the third study from Sytsma (2010a) to illustrate the first, and simplest, type of t-test that we'll discuss. To do this, I'll first introduce the general *research process* in §3.1. In §3.2, I'll discuss the *philosophical background* for our target study, including the first two studies from Sytsma (2010a), and connect this to the general background provided in the previous section. The remaining sections will then walk us through the study: §3.3 lays out the *research question* motivating our target study, §3.4 details the *research design*, §3.5 the *instrument* used, §3.6 *conducting the study*, §3.7 the *analysis of the results*, and §3.8 the *interpretation* of the findings.

3.1 The Empirical Research Process

In a previous text with Jonathan Livengood (Sytsma and Livengood 2015) we detailed a four-stage process for conducting empirical research in experimental philosophy:

1. Formulate a research **QUESTION**
2. Develop a **PLAN** to address your research question
3. **CONDUCT** the study laid out in your plan
4. **ANALYZE** the results of the study you conducted

There are multiple components to each of these stages, including that the **PLAN** stage involves determining the *design* for your study and constructing an *instrument* corresponding to this design. The resulting process corresponds closely with the plan for the present text, with the individual chapters aiming to illustrate how to develope of a *research question* (Stage 1), construct a corresponding *study design* and *instrument* (Stage 2), *conduct a study* (Stage 3), and *analyze* and *interpret* the results of that study (Stage 4).

This process is quite general and can be used for most empirical research, not just work in philosophy. The philosophical focus of the research, however, will shape how the process is applied. This is most clear with regard to the first stage. While discussions of developing a research question in the typical text on experimental methodology will start with formulating a hypothesis, experimental *philosophers* should start a step earlier: philosophy generally begins with formulating arguments, and x-phi is no exception to this rule. In my opinion, the first step in developing a solid research question in experimental philosophy is to formulate a philosophical argument with an empirical premise, a premise that—with suitable clarification and specification—can be tested in your study (or studies). This clarification and specification converts your premise into a testable hypothesis. With this in hand, it is important to think through the general strategy you will employ in testing your hypothesis and analyzing the results.

This involves thinking about the goals of your study and the type of claim your hypothesis makes.

There are three basic types of claims you might make in your hypothesis—an *estimation* claim, a *comparison* claim, or a *relation* claim. *Estimation claims* are about putting a number on a feature of a population that you're interested in. For instance, we might want to estimate the percentage of people who agree with the claim that there can be no unfelt pains. Very often, for philosophical purposes we're not so much interested in the exact number, though, but instead more concerned with how it sits relative to another number. The claim that common sense rules out the occurrence of unfelt pains, for example, plausibly entails that such a belief should be *common*—that it should be the majority belief—but not that exactly 72% (or 86%, or 92%, or whatever) of people will hold this belief. The claim that a majority of people hold a given belief is a *comparison claim*. In this case, it compares one number that we'll try to assess in our research to a fixed point (50%). In other cases, however, we'll want to compare two numbers that we'll try to assess. For instance, we might predict that the proportion of philosophers who believe that there cannot be unfelt pains is greater than the proportion of lay people who believe this. Finally, *relation claims* are about how multiple features are associated or how one changes relative to the other. For example, rather than simply comparing the proportion of philosophers who deny unfelt pains to the proportion of lay people, we might want to assess level of training in philosophy and belief in the possibility of unfelt pains. We might predict, for instance, that there will be an inverse relationship—that belief in unfelt pains will go down as training in philosophy goes up.

I will focus on comparison claims in this chapter, as the statistical tools it introduces concern certain types of comparison claims: t-tests are applicable when we're comparing a

number we've assessed using a continuous measure, or a suitable approximation of it, either to a fixed point (*one sample t-tests*) or to another such number (*independent samples t-tests*, *paired samples t-tests*, *partially paired samples t-tests*), and if some other assumptions hold. I'll return to this below. For now, the key thing is that the type of claim your hypothesis makes will matter not just for the plan you develop for testing your hypothesis, but how you analyze your results after conducting the study.

The next step in the research process concerns developing a plan to test your hypothesis, starting with formulating a design for your study. Deciding on a design involves a number of factors, including the specifics of the hypothesis you're looking to test and the type of claim it makes (*estimation*, *comparison*, or *relation*), among others. In turn, the design you arrive at will specify a number of important details about your study, including the type of study it is (i.e., a *true experiment* versus a *quasi-experiment* or a *descriptive study*) and the variables you will be manipulating and measuring in your study. The type of study you conduct and the types of variables in it are connected. There are two basic types of variables—the things that are varied in your study (known as *independent variables* or *predictor variables*) and the things that are measured (*dependent variables* or *response variables*). Every study will have at least one response variable. In the studies we'll be looking at, these correspond with the test questions that we ask participants. Not every study will have predictor variables, however: in some studies there is just one condition—nothing is varied and every participant gets participant gets treated the same. These are *descriptive studies*. The case study we'll focus on in this section is an example of such a study. In other studies something is varied, either by nature (*quasi-experiments*) or by the researcher (*true experiments*). Studies looking at demographic differences in philosophical intuitions, such as Machery et al. (2004), are a classic example of the former.

The study from Sytsma and Snater (2023a) discussed in Section 5 is a good example of the latter. While a general theoretical exploration of these differences and how they inform study design is beyond the scope of the present text—this requires quite lengthy texts on their own to cover—the process will be illustrated through the array of case studies presented, including those given in this chapter.[6]

For studies involving the types of questionnaire methods we'll be focusing on here—i.e., studies where you're asking participants to answer one or more questions—the next step will be to construct the instrument you will use. Again, questionnaire design is a large topic and one that we will largely illustrate via examples.[7] A few preparatory remarks will help with understanding the process, however. Most often in experimental philosophy questionnaires center on presenting participants with a short framing text, or *vignette*, followed by one or more questions about that text, often employing a fixed scale as we'll illustrate below. But these instruments are almost always comprised of more than this, generally also including a consent form that introduces the researchers and project, instructions that guide participants in completing the questionnaire, and various demographic questions (e.g., asking for the participant's age, gender, and so on). You might also want to include check questions to test that participants are putting in sufficient effort (*attention checks*) and/or understand what is going on in the questionnaire (*comprehension checks*), as is illustrated in Section 4.

To construct an effective questionnaire, I recommend thinking of it as a conversation between the researchers and the participants—a conversation that is shaped, in part, by each of

[6] For a more extended if still quite brief general introduction study design directed at experimental philosophers, see Chapters 7 and 8 of Sytsma and Livengood (2015). For an excellent introduction to research methods and design in psychology, see Goodwin and Goodwin (2016). For a more advanced treatment, see Shadish et al. (2001).
[7] See Chapter 11 of Sytsma and Livengood (2015) for a more extended discussion. See Sudman et al. (1996) and Schuman and Presser (1996) for excellent book-length treatments.

the components of the questionnaire and how you phrase them. The key point here is to be on the lookout for how your questionnaire might lead the conversation astray, potentially generating responses from participants that don't actually reflect the judgments you wanted to measure. This is not an easy task, however, and several potential pittfalls will be illustrated in the case studies presented below.

Once you've designed your study and constructed the instrument, the next step is to carry it out. This involves getting ethics approval, piloting and refining the instrument, and determining how you'll recruit participants. Piloting is basically to take your study out on a test run, conducting a preliminary version on a small number of participants. Often this might involve using a modified version of your instrument, typically including additional open-ended questions that ask participants to explain their answers to the central questions you're interested in. The goal here is to identify problems with your study design and conversational pitfalls in your instrument before committing full resources to this study. Pilot testing is an important part of good research practice, but is often overlooked by new practitioners. If your pilot study reveals issues with your design or instrument, you'll want to make modifictions and then pilot again before finally running your full study.

Once you've run the study, you'll need to analyze and interpret the results. I'll illustrate this in the examples below, but first it is important to say a little bit about *why* we perform a statistical analysis in the first place. It might seem that all that is needed at this stage is to describe the basic details of your results, perhaps simply noting how participants responded to the questions you asked. But statistical analysis goes beyond giving such basic details: it involves drawing *inferences* on the basis of your data and offering a *justification* for those inferences. In conducting empirical research, our goal is to use observation to answer questions

about the world. But many of the questions we want to answer cannot be adequately addressed just through simple observation either because we cannot exhaustively observe what we're interested in or because we're interested in something that cannot be straightforwardly observed (or both). For instance, suppose we want to know something *relatively* straightforward, such as whether a majority of people believe that unfelt pains are impossible. People's beliefs are not something that we can directly observe (as of yet) and figuring out how to assess belief with regard to an abstract question like the possibility of unfelt pains is no easy task. Setting this aside, however, we're still left with the issue that we aren't in a position to survey *all* people. Heck, we generally won't be in a position to interact with more than an extremely small fraction of the population we're interested in! So how do we answer our original question, moving from the relatively small number of participants we received responses from to conclusions about the wider population they're part of? We employ *statistical inference*.

The goal of statistical inference is to make an *educated* guess about things that we have not yet observed on the basis of things that we have observed. More technically, we infer something about features of a *population* (what we call 'parameters') from observations of corresponding features of a *sample* drawn from that population (what we call 'statistics'). In doing so, we reason that since *most* samples drawn from a population will have features that are similar to the features of the population they're drawn from, and since similarity is symmetric, we should expect the population to have similar features to the sample. Of course, the population of interest is unlikely to have exactly the same features as any given sample. As such, we wouldn't be justified in simply asserting that the population has the same features as the sample. What we could reasonably say, however, is that the features of the population (the parameters) are *probably and approximately* the same as the features of the sample (the statistics). In other

words, we infer parameters from statistics, while recognizing the hedge that this is only probably and approximately the case. A key part of our statistical analysis—including using tools like t-tests—is then to flesh out this 'probably and approximately'. Let's see how this all works by considering a preliminary example drawn from Sytsma (2010a), which provides important background for the case study that we'll walk through in the remainder of this section.

3.2 Background

My work on the commonsense conception of pain grew out of more general work in experimental philosophy of mind investigating whether non-philosophers tend to have a concept that is suitably similar to the philosophical concept of phenomenal consciousness introduced above. In Sytsma (2010a), I note that it is common for philosophers of mind to make assumptions about folk psychology—assumptions about our ordinary, pre-theoretical thinking about the mind—in discussions of phenomenal consciousness. This includes that both realists (e.g., Chalmers 1995) and skeptics (e.g., Dennett 1991) about phenomenal consciousness take the concept to be a part of folk psychology, assuming that the existence of qualia is pretheoretically obvious. Claims about folk psychology are empirical claims, however, and it is quite possible for scholarly training to skew our perspective on ordinary, pre-theoretical thinking. Focusing on whether lay people tend to have a concept of phenomenal consciousness, I noted that empirical work was beginning to be done on the question and I surveyed conflicting findings from Knobe and Prinz (2008), who argue in favor of the claim, and Sytsma and Machery (2010), who argue against.[8]

[8] A number of criticisms have been raised against each of these works. For responses to Knobe and Prinz (2008), see Sytsma and Machery (2009), Huebner (2010), Arico (2010), Strickland and Suben (2012), Phelan et al. (2013). The most prominent criticism of Sytsma and Machery (2010) has been the *ambiguity objection* (Sytsma 2016), which has been raised by Huebner (2010), Peressini (2013), Fiala et al. (2013), and Chalmers (2018). For recent responses, see Sytsma and Ozdemir (2019), Ozdemir (2022), Sytsma (n. d.). Objections to our explanation of our results have been

Sytsma and Machery (2010) present evidence that lay people, in contrast with philosophers, treat two prototypical examples of supposed phenomenally conscious mental states—seeing red and feeling pain—quite differently. In our main study we gave participants either a description of a normal human or a simple non-humanoid robot performing behaviorally analogous tasks expected to elicit attributions of one or the other of these mental states for the human, then asked the participants whether the entity (human or robot) had the mental state at issue (saw red, felt pain). We found that while philosophers tended to treat both states similarly, denying that the robot either saw red or felt pain, lay people tended to treat them differently, denying that the robot felt pain but affirming that it saw red.[9] Based on these results we argued that if lay people were employing the concept of phenomenal consciousness in responding to these questions, then they should have treated the two states similarly, just as the philosophers did. But they did not. We took this to suggest that the lay participants were *not* generally employing the concept of phenomenal consciousness.

Building off of the arguments given in Sytsma (2009, 2010c), in my (2010a) I further explored one explanation for the pattern of findings for lay people found in Sytsma and Machery (2010). I hypothesized that this pattern reflects that lay people tend to hold a *naïve view* of both colors and pains: rather than treat colors or pains as qualities of mental states, they conceive of them as qualities of objects outside of the mind/brain. Focusing on pain, the idea is that people tend to deny that the simple robot in our study feels pain because they conceive of pains as being

raised by Talbot (2012), Buckwalter and Phelan (2013), as well as the studies discussed below suggesting that participants' responses reflect that they tend to hold a naïve view of colors and pains.

[9] In general, we shouldn't accept a conclusion based on just one set of results. Rather, our credence in the results should be tempered and should rise as they are replicated—as similar results are found in subsequent studies, especially studies by other researchers. This includes both exact or approximate replications, which attempt to repeat a study as closely as possible, and conceptual replications that test the same hypothesis in another way (see Cova et al. 2021). The key result from Sytsma and Machery (2010) has been replicated a number of times, including in Sytsma and Machery (2012), Sytsma (2012), Sytsma (2013), Sytsma and Ozdemir (2019), Cova et al. (2021), and Ozdemir (2022).

instantiated in injured body parts, but doubt that the robot has the right sort of body parts to support pains. On this view, while soft and fleshy body parts can instantiate pains, hard and metallic body parts cannot. This hypothesis not only explains the pattern of results in Sytsma and Machery (2010), however, but also suggests against the claim that common sense supports the standard view of pain among philosophers.

In my first study in Sytsma (2010a), I asked participants a set of questions about how they understand colors. The results were consistent with lay people tending to hold a naïve view, with a majority of participants answering that colors are properties of external objects, denying that they are mental or mind-dependent, and denying that spectrum inversion is possible (as we would expect if they treated colors as belonging to the objects seen rather than to perceivers). Study 2 extended these findings to pains, in addition to colors. Results were comparable, with a majority of participants seemingly embracing a naïve view for both colors and pains, treating these as properties of things outside the mind/brain and denying that they are mental or mind-dependent. Further, this study included a question about unfelt pain: 'Do you think that there is still pain in a badly injured leg even when the person is not aware of it?' I hypothesized that if people tend to hold a naïve view, taking pains to be properties of injured body parts, then the presence (or absence) of the pain would not depend on whether the person *felt* that pain. And, indeed, a majority of participants answered this question affirmatively, suggesting that they hold that unfelt pains are possible. The next two studies explored this finding further. Let's walk through the process of designing, conducting, and analyzing Study 3 from this paper in more detail.

3.3 Research Question

Recall the research process laid out above. The first step is to formulate a research question. I indicated that for research in experimental philosophy, it is best to first start with formulating an argument with an empirical premise. We've just surveyed the broader dialectic that Sytsma (2010a) fits into. Focusing on just Study 3, however, we can lay out a rather straightforward argument. We've seen that according to the standard view of pain in philosophy there can be no unfelt pains, and this is often supported by appeal to the (supposed) dictates of common sense. If common sense allows for unfelt pains, however, then the standard view would not enjoy this support. The key empirical premise here is that common sense allows for unfelt pain.

This premise requires some clarification and specification before we can test it, however. Most importantly we need to determine what such claims about common sense amount to. As detailed in Section 1, claims that common sense precludes unfelt pains in the literature are often laid out in terms of our intuitions about hypothetical cases involving someone being injured but not feeling pain, with the suggestion that people will generally have the intuition that in such cases there is no pain. This suggests a general strategy for testing: Give participants a vignette describing such a case and then ask them whether the injured person had a pain despite their not feeling it. Since we're not varying the vignette or the question, this would be a *descriptive study*: it doesn't have any *predictor variables* and has just one *response variable* (the question we ask about whether the case involves an unfelt pain). Our predictions for this study concern the response variable: The standard view predicts that the majority of participants will give a negative response to this question; in contrast, if people tend to hold a naïve view of pain, then we would expect the opposite—that the majority will give an affirmative response. Each of these predictiosn make a *comparison claim*: in making the prediction that the majority will give an

affirmative response, for example, we're predicting that the number of affirmative answers will be *greater than* 50%.

3.4 Research Design

Having developed a research question, the next step is to put together a research plan, starting with deciding on a study design. The hypothesis formed in the first stage will help guide the design of your study. In the present case, my hypothesis concerns how people will tend to respond to a question about a simple scenario: Will people tend to judge that an injured person has a pain even if they don't feel it? As such, there is no need to compare responses between samples drawn from different populations (as in a *quasi-experiment*) or between participants assigned to different conditions (as in a *true experiment*). As we just noted, a *descriptive study* is sufficient here. This means that there is no need to worry about predictor variables (*independent variables*) for the present study, and given that it involves just one question of interest the study can be restricted to a single response variable (*dependent variable*). In other words, for this study I simply want to give participants a description of a scenario about an injured person who doesn't feel pain and ask them a question about whether that person has a pain. As such, I just need to develop a single instrument that each participant will receive.

3.5 Constructing the Instrument

As noted above, a full instrument will typically include an introduction, instructions to participants, and demographic questions, in addition to the *philosophical probe* that we're most concerned with—the vignette and questions that test our hypothesis. The full instrument I

developed is available in the supplemental materials **[Sytsma_2010a_STUDY_3.pdf]**.[10] Here I'll focus on the philoshiphical probe. For this study, I chose to describe a common scenario—an injured person being distracted and not noticing a pain. My aim was to describe this rather directly, keeping the text to a minimum. This has the potential benefit of maintaining participants' attention (which can wain with longer probes or questionnaires) and avoiding extra verbiage that might bias their responses one way or another. After the scenario, I then wanted to ask participants whether the subject of the story still had the pain or whether there was no pain during this period. At a first pass, the response options here would appear to be binary: either there is a pain or there isn't. But I wanted to allow for participants to register that they weren't sure about the answer, as well as degree of certainty in a response. As such, I asked participants to answer using the partially anchored 7-point scale shown below (Figure 1). Here the end points are anchored with text descriptions, as is the midpoint of the scale. Of course, a number of alternative decisions could have been made in designing this study, including the vignette used, the framing of the question, and the response options. I will return to some of these choices below in detailing subsequent studies.

[10] Standardly, this should include questions about the participants age and gender. (If I were to run this study now, I would include an option for "non-binary" in the gender question, and I encourage you to do the same in your studies.) Given the concern with common sense, I also wanted to check whether participants had training in philosophy or areas where they might have been taught scientific accounts of nociception. Finally, given concerns about the relevant population for the claims about the commonsense conception of pain, discussed below, I asked about participants' native language.

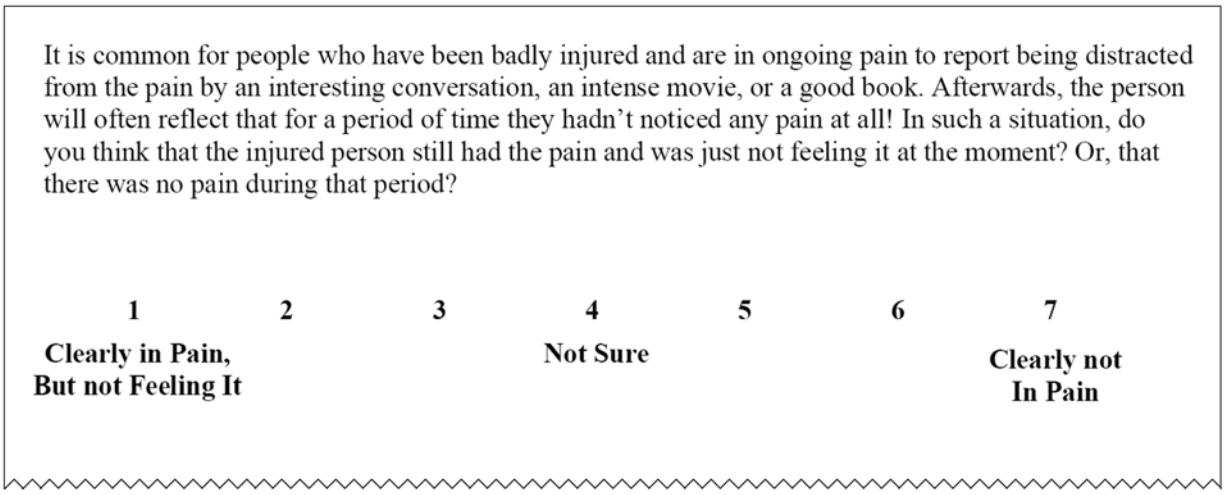

Figure 1: Vignette and scale for Study 3 in Sytsma (2010a, 124).

3.6 Conducting the Study

The next step in the research process is to conduct the study. As noted above, this involves getting ethics approval and, especially for more involved studies, piloting them. Given the simple nature of the present study, I didn't formally pilot it, but rather asked friends and colleagues for their input. In conducting a study, a key decision point is figuring out how you'll recruit participants. This involves determining what the relevant population is and figuring out how to obtain a sample from that population. Often the relevant population will be somewhat unclear. When philosophers claim that common sense tells us that there can be no unfelt pains, it is plausible that this is thought to be something common to all people with a concept of pain. But we might also suspect that at least some aspects of common sense will be culturally variable. Further, we've seen that claims about the common sense understanding of pain are often coupled with claims about how we talk about pain. And, of course, pain language will vary between groups of language users. As such, in all but the last of the studies I'll detail in this chapter, the target population is narrower, being comprised of English speakers in North America. A question

therefore remains about how widely these results generalize, although some cross-cultural work on ordinary conceptions of pain has been done (e.g., Kim et al. 2016, Sytsma and Reuter 2017).

Even for a narrower population like English speakers in North America, however, it is impossible to exhaustively survey members of the population and we'll need to content ourselves with testing just a small sample of the larger population. Ideally, we would sample randomly from the population, such that each individual was equally likely to be chosen for our sample. In practice, however, researchers are seldom, if ever, in a position to solicit a truly random sample: we simply don't have equal access to each member of a population. Instead, we do what we can, aiming to use a recruitment method that we hope will produce a reasonably unbiased sample with regard to our research question. Typically, this will involve *convenience sampling*: we sample from the people we have access to—those people who are convenient for the researcher. While this is non-ideal, concerns can be at least partially alleviated by using different recruitment methods, as I will illustrate below.

For the present study, I used a participant pool that was convenient to me—students in introductory classes at the university I was attending. To do this I talked with the instructors for two courses I had not previously surveyed and got permission to administer the questionnaire at the start of one of their classes. The instrument was printed out on paper and handed out to students.[11] Using classroom samples like this means that there will be some variability in how many participants complete your questionnaire, based on how many students attend class that day and choose to complete the survey. In the present case, this generated 54 responses (excluding one person who took the survey in both classes). Ideally, however, we would first

[11] It is more common today to use online samples, as will be illustrated in studies described in subsequent sections. To do this you'll need to create a web-based version of the instrument. Most often this is done using survey software such as Qualtrics or Lime Survey, as described in Chapter 1.

estimate the sample size that we need. We'll return to this process in §4.2 after we've finished with this first case study.

One disadvantage of in-class studies, as opposed to studies conducted online, is that after collecting the completed questionnaires you'll need to enter the responses into a digital form. I did this by hand, looking through each questionnaire and entering the responses for each in a row of a spreadsheet. (A reduced version of this spreadsheet is available in **[Sytsma_2010a_STUDY_3.csv]**, which removes unnecessary demographic details to further protect anonymity.) Entering data by hand can be a slow process, especially for large studies, and potentially creates an extra source of human error. This is one reason that many now prefer to use online samples.

Classroom samples will also tend to be less varied than the larger population in a number of ways, including that they will tend to be younger (the average age of my sample was 19.6 years) and more likely to have education in relevant areas. Indeed, I found that five participants had more than minimal training in philosophy or psychology.[12] The responses of these participants were removed based on criteria specified prior to running the study. Ideally, such criteria, as well as other important details of your studies (e.g., predictions and plans for statistical analysis) will be registered before conducting your study. Such *pre-registration* is now considered best practice and is becoming increasingly common. This can be done using websites like https://osf.io/, which also provides a repository for materials and data.

[12] Participants were counted as having more than minimal training in philosophy or psychology if they indicated that they had completed some graduate work in philosophy or psychology, had completed an undergraduate degree with a major in philosophy or psychology, or were completing an undergraduate degree with a major in philosophy or psychology.

3.7 Analysis

Basic data for Study 3 from Sytsma (2010a) is available in the supplemental materials as a comma delineated spreadsheet—**[Sytsma_2010a_STUDY_3.csv]**—with responses to the main test question shown in the column labeled 'RESPONSE'. To explore this data, I'll use the free statistical software package R, which you can download from http://cran.r-project.org. We'll walk through the very basics of what you need to start using R to run t-tests here in this chapter. That said, it is important to note that R is a full-featured programming environment, such that it would be impossible to give a detailed overview of how to use it here. Fortunately, there are many excellent resources for using R. For a brief introduction, you might start with Chapter 10 in my text—Sytsma and Livengood (2015)—which also includes pointers to other references for learning to use R. To help with learning R, the code used for each analysis in this chapter is available in the supplemental materials. For the present study, this is provided in **[Sytsma_2010a_STUDY_3.txt]**. This file can be read using any standard text editor and I'll reproduce it piece-by-piece below as we walk through the analysis. Once you've installed R using the link above, simply open the txt file and you can follow along with the analysis by either copying-and-pasting the relevant lines from the file or by typing the text in at the command prompt in the R console window (the red ">"), as shown in Figure 2.

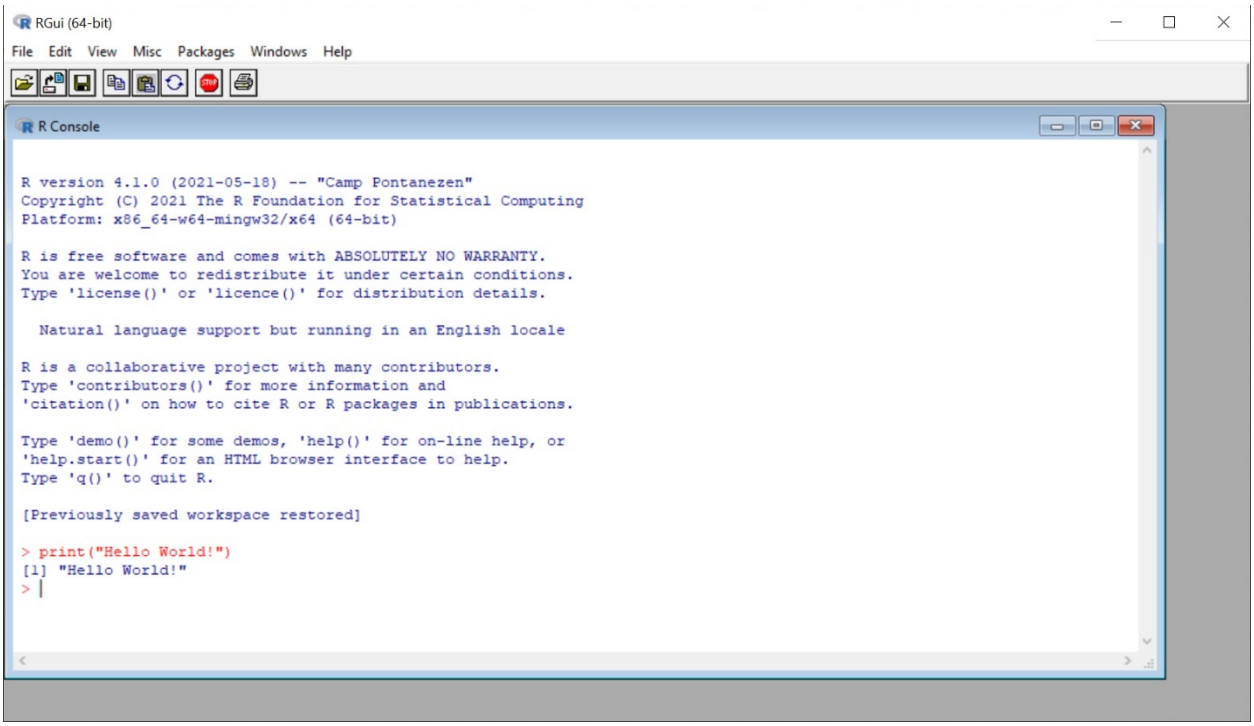

Figure 2: R console window with a standard first program when learning a new programming language (printing "Hello World!").

3.7.1 Getting Started and Loading the Data

Looking at the file, the first thing you'll notice is that it starts with the following two lines of text:

```
#install.packages("lsr")
library(lsr)
```

Together these lines of text will be used to install the **lsr** package and then load it from the corresponding library. A *package* is basically a collection of *functions* (and usually other stuff such as data sets) that someone has built and documented for use in R. Generally, you can find information about a given package on the CRAN website noted above.[13] A *library* is simply where the package is stored once it is installed. We'll return to the **lsr** package below when we

[13] For the **lsr** package this is available at https://cran.r-project.org/web/packages/lsr/

use the `cohensD()` function from it.[14] Functions will be distinguished by name followed by a pair of parentheses and are used to tell R to do some specific thing, such as to install a package (i.e., the `install.packages()` function) or load an installed package from a library (i.e., the `library()` function). Most of the time, however, we need to give R more guidance about what we want it to do. For example, we'll need to tell it which package we want it to install or load. We do this with *arguments*, which are specific bits of text that go inside the parentheses, such as `"lsr"` in `install.packages("lsr")` and `lsr` in `library(lsr)`. In each of these cases just a single argument is supplied to the functions, but as we'll see below we often want to tell a function multiple things. In such cases we'll do this by supplying multiple arguments, which will be separated by commas within the parentheses (such as `cohensD(D3$RESPONSE, mu=4)` which we'll discuss below).

So, each of the two lines of text at the start of our file calls a function, supplying R with a single argument for that function. If you simply paste these two lines into R, however, you'll probably get the following error:

```
Error in library(lsr) : there is no package called 'lsr'
```

This is because you need to *install* the package before you can load it into R… and the line of code that does this is *commented out*. Basically, the # mark tells R that the text following it is just a comment, not something it needs to pay attention to. So when you enter the first line of text from the file—`#install.packages("lsr")`—R will just ignore the function call and won't install the **lsr** package. I've added the # mark here because you'll only need to install the package once, so after the first time you run the code you'll want R to ignore this line.

[14] Details about many functions in R can be found through the RDocumentation website, including for the cohensD() function: https://www.rdocumentation.org/packages/lsr/versions/0.5.2/topics/cohensD

Try entering the first two lines of text again, but this time exclude the # mark from the first line:

```
install.packages("lsr")
library(lsr)
```

The first line will pop up a window asking you to select a 'mirror'; that is, a place to download the package from. You can simply leave it on the default location and select 'OK' or else choose a location that is close to you. R should now install the **lsr** package for you. Once that is done, you should be able to enter the `library(lsr)` command without error. This will load the package so we can use it later.

The next thing you'll see in the script are three lines starting with the # mark:

```
##########################################
# Load Data for Sytsma (2010a), Study 3 #
##########################################
```

This is another comment for the user—it tells the person reading the code what is going on… and tells R to ignore it. Here it tells us that the next line in the code will load data from the study we'll be looking at. Specifically, it uses the `read.table()` function to load the comma delineated spreadsheet noted above—**[Sytsma_2010a_STUDY_3.csv]**—from my desktop:

```
D3 = read.table("C:/USERS/jmsyt/Desktop/Sytsma_2010a_STUDY_3.csv",
header=TRUE, sep=",")
```

To use this yourself, you'll need to edit the *path* to point to where the file is on your own computer (i.e., change the bit that reads `C:/USERS/jmsyt/Desktop/` to indicate where the file is on your computer, which can found by right-clicking the file and selecting 'Properties' on a typical Windows PC or 'Get Info' on a typical Apple computer). Once you run this line, R will read the table from the spreadsheet and copy it into the variable `D3`. After loading the data, if you simply enter the variable name into R—if you type `D3` and hit enter—it should display the data from the table, as shown in Figure 3.

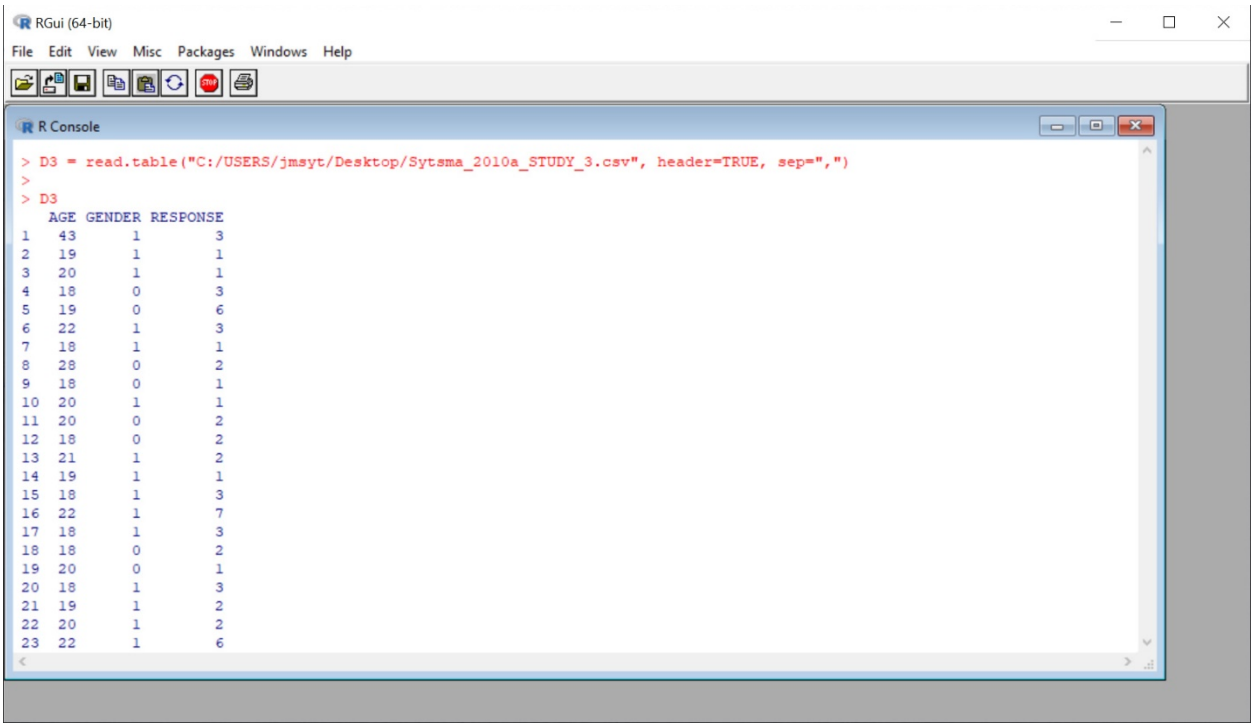

Figure 3: Displaying data from Sytsma (2010a), Study 3, in R.

3.7.2 Basic Visualization and Statistics

Having read in this data and stored it in `D3`, we can now do any number of things with it in R. We'll begin by using it to provide some basic details about the data, starting with looking at how many participants selected each response option (1 to 7) for the probe question. I began by generating a histogram using the `hist()` function in R. This gives a visual display of the number of participants selecting each answer choice. To do this I supplied two *arguments* to the function:

```
hist(D3$RESPONSE, breaks=BreakValues)
```

The first argument tells R that we want to use the values for `RESPONSE` from `D3` (the `$` symbol telling R to use that column from the table), while the second provides a list of values for how to divide up the lines on the histogram, which was provided using the `c()` function (the

concatenate function) in the previous line. This tells R to generate a histogram for the responses that centers the bars on the whole numbers from 1 to 7. To make this still more informative, I then added a dotted line to the histogram centered on the *mean*—the average response to the question—using the following command, which uses the `lines()` function, with calls to the `c()` and `mean()` functions in the arguments:

```
lines(x=c(mean(D3$RESPONSE),mean(D3$RESPONSE)), y=c(0,18),
type="l", col="red", lty="dashed")
```

Finally, I used the `nrow()` function, which counts the *number of rows* in our table that meet a certain criteria, to give an exact count for each response option:

```
nrow( D3[ D3$RESPONSE == 1, ] )
nrow( D3[ D3$RESPONSE == 2, ] )
nrow( D3[ D3$RESPONSE == 3, ] )
nrow( D3[ D3$RESPONSE == 4, ] )
nrow( D3[ D3$RESPONSE == 5, ] )
nrow( D3[ D3$RESPONSE == 6, ] )
nrow( D3[ D3$RESPONSE == 7, ] )
```

Specifically, I had R count the number of rows in the table stored in `D3` where `RESPONSE` was equal to 1, the number of rows where `RESPONSE` was equal to 2, and so on. The results have been added to the txt file after the command using the comment mark (#), as seen in Figure 4, which shows the console output and the histogram for this block of code.

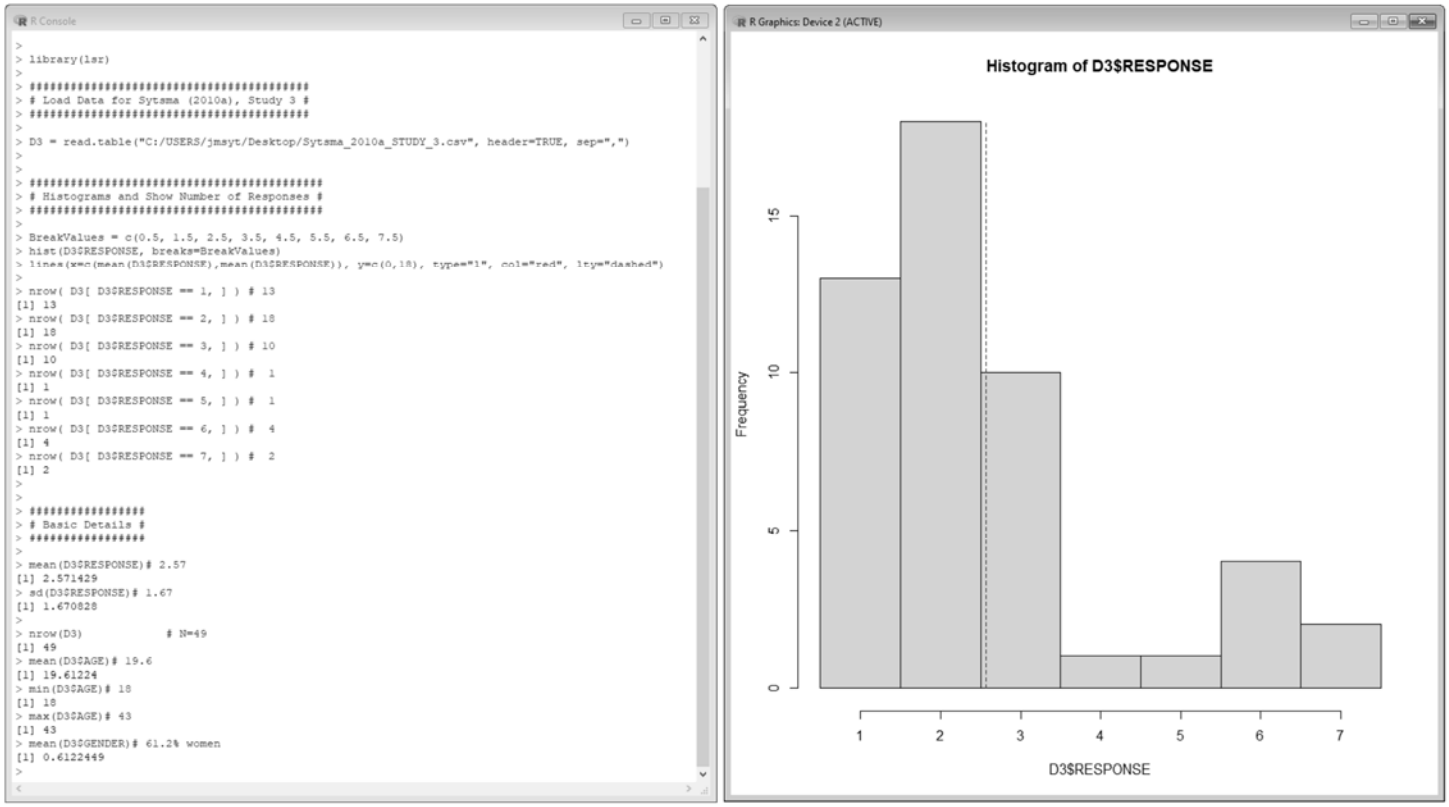

Figure 4: R output for initial script for analysis of Study 3 from Sytsma (2010a).

My next step was to calculate a few basic statistics for the responses to the probe question, including the mean and the *standard deviation*, which tells us about how the responses tend to spread out around that average, as well as the basic demographic questions I included in the sample data file:

```
mean(D3$RESPONSE) # 2.57
sd(D3$RESPONSE)   # 1.67
nrow(D3)          # N=49
mean(D3$AGE)      # 19.6
min(D3$AGE)       # 18
max(D3$AGE)       # 43
mean(D3$GENDER)   # 61.2% women
```

From the histogram we can readily see that a large majority of the responses to the test question were below the midpoint of 4, with 41 of 49 participants (83.7%) giving a response of 1, 2, or 3

on the 7-point scale. In other words, most of the participants leaned toward answering that the injured person in the vignette was 'clearly in pain, but not feeling it'. Not surprisingly, the mean response ($M = 2.57$) is also below the midpoint. This is in line with the prediction I made. As noted above, however, my concern is not specifically with *this set of 49 people*. Rather, the goal is to say something about *the population of interest* and I'm simply using the responses of these participants as an imperfect guide to what the larger population is like in this regard. As such, I need to do some statistical inference.

3.7.3 Null Hypothesis Significance Testing and the Basics of T-tests

There are different ways of doing statistical inference. The most common is *Null Hypothesis Significance Testing* (NHST).[15] In NHST we specify a null hypothesis that corresponds with finding no effect, then test how likely it is that we would have gotten data at least as extreme as the responses observed if the null hypothesis were true (and if other test assumptions hold). If this is suitably unlikely, then we reject the null hypothesis and take the results to be in line with our hypothesis.

In the present study, my hypothesis was that a majority of the population holds that unfelt pains are possible and so would respond that the subject of the vignette had a pain. One way of specifying this hypothesis (but not the only way), is to predict that the mean response would be below the midpoint on the scale. The relevant null hypothesis for this prediction is, then, that the mean response is at or above the midpoint on the scale. Clearly, the mean response for my *sample* was below the midpoint. But this *might* happen just by dumb luck even if a majority of

[15] One increasingly popular alternative is the use of Bayesian statistics, although this remains rare in experimental philosophy. A discussion of Bayesian statistics is well beyond the scope of the present chapter, but see Section 13.3 in Sytsma and Livengood (2015) for a Bayesian alternative to using t-tests and see Albert (2009) for a more general introduction to Bayesian statistics in R.

the people in the *population* believe that unfelt pains are impossible. We use a statistical test to put bounds on how likely this is. More carefully, we calculate a conditional probability for getting data at least as extreme (relative to the null hypothesis) as the data we actually got if the null hypothesis is true (and if various test assumptions hold). This conditional probability is called a *p-value* and if it is sufficiently small, we reject the null hypothesis. What counts as sufficiently small depends on the *significance level* specified, with 0.05 being conventional.[16]

The statistical tests I'll be focusing on here are *t-tests*. They are among the simplest and most commonly used statistical tests. Despite the way they are sometimes discussed, t-tests are actually a family of related procedures, including *one sample t-tests*, *independent samples t-tests*, and *paired samples t-tests*. What unites all of these tests is that the reference distribution for the comparison is a *t-distribution*. We needn't worry too much about what this means, here, but the t-distribution is closely related to the normal distribution (a standard bell curve). Indeed, the difference between the distributions becomes negligible as the degrees of freedom increases, where this is related to the sample size.[17] What is most important, for present purposes, is that this generates a key assumption that is at play when we use t-tests—that the feature we're interested in is normally distributed in the population. This means that if we plotted the histogram for the entire population, as we did above for our sample, the resulting histogram would approximate a bell curve. Of course, this isn't something that we'll typically know about

[16] Sometimes a significance level of 0.01 is used instead. See Benjamin et al. (2018) for an argument that we should lower this still further to 0.005.

[17] You can demonstrate this for yourself in R using the dnorm() and dt() functions to plot a normal curve and t-distributions, respectively. For example, the following code will compare the normal distribution to the t-distribution with degrees of freedom of 1, 5, 10, and 20:

```
curve(dt(x, df=20), from=-5, to=5, col="green",
     main="Distribution Comparison",
          ylab="Density")
curve(dt(x, df=10), from=-5, to=5, col="orange", add=TRUE)
curve(dt(x, df=5),  from=-5, to=5, col="red", add=TRUE)
curve(dt(x, df=1),  from=-5, to=5, col="purple", add=TRUE)
curve(dnorm(x), from=-5, to=5, col="black", add=TRUE)
```

our population. Nonetheless, the date from our sample can give us an indication of whether the assumption is warranted following the same logic as above (i.e., that features of the population are *probably and approximately* the same as features of the sample). The histogram for the present study, however, should give us some pause with regard to this assumption, since it doesn't obviously approximate a bell curve.

A second reason for pause is that t-tests assume that our data have *interval scale* such that they can be taken to approximate a *continuous distribution*. Interval scale means that the distance between a response of 2 and 3 on our scale, for example, is the same as between a response of 3 and 4, as opposed to these responses simply being rank-ordered as in finishing places in a race. The assumption of interval data is controversial for scales like those most commonly used in experimental philosophy, including the scale used in the present study. Nonetheless, while there are often reasons to doubt each of these two key assumptions (normal distribution and interval data), in practice t-tests are rather robust. As we proceed, though, I'll briefly detail alternative tests that do not make these assumptions. As we'll see, they lead to comparable conclusions in these case studies, and in my experience this is very often the case (hence the robustness).

3.7.5 Directionality and Conducting a One-sample T-test

Running t-tests is very easy in R. In fact, we can use the same function—the `t.test()` function[18]—for each of the three types of tests that we'll be focusing on in this chapter. (For the fourth type of t-test I noted above—*partially paired samples t-tests* discussed in Box 1—we'll need to use a different function, but such tests are quite rare and aren't likely to be something you'll need to worry about.) Recall from above that we said that t-tests are applicable when

[18] Once again, details for functions in R can be found through the RDocumentation website, including for the t.test() function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

we're comparing a number we've assessed using a (suitable approximation of a) continuous measure to either a fixed point or to another such number, and if other assumptions hold like those we just discussed. As such, we'll need to tell the `t.test()` function the two things we're comparing—within the parentheses we'll need to supply one argument pointing R to the data for the first number and a second argument for what that number is being compared to. If we're comparing to a fixed point, then the appropriate t-test will be a *one sample t-test* and we'll simply provide the number for that point as the second argument in the `t.test()` function. If we're comparing to another number we've assessed, however, then we'll want to run one of the remaining tests, which apply to comparisons between samples and vary based on the relationship between those samples—whether they are two different samples (*independent samples*), the same samples (*paired samples*), or a mixture of the two (*partially paired samples*).

What type of comparison are we interested in for the present study? Recall from our previous discussion that the null hypothesis for Study 3 in Sytsma (2010a) is that the mean response to the test question will be at or above the midpoint on the 7-point scale (i.e., comparing to 4). Of course, the midpoint on the scale is a fixed point, not a second number that we've assessed by collecting additional data. As such, we will want to use a *one sample t-test*. This is done in the first line of code in the next section of our txt document:

```
######################
# Statistical Tests #
######################

t.test(D3$RESPONSE, mu=4, alternative="less")
# t(48)=-5.99, p<.001
```

Here the `mu=4` argument is telling R that the relevant comparison for our response data (`D3$RESPONSE`) is to the fixed point of 4.[19]

[19] The Greek letter μ (or mu) is standardly used in statistics to represent the population mean.

What about the third argument in our function call (`alternative="less"`)? This specifies the *direction* of the test. Specifically, we could have been predicting any of three types of relationship here—either that the mean response would be *less* than 4, that it would be *greater* than 4, or that it would simply be different from 4, such that it is concerned with *both sides* (either less than *or* greater than 4). These are indicated by supplying the argument `alternative="less"`, `alternative="greater"`, or `alternative="two.sided"` respectively (note that if you leave out this argument, it will default to a *two-sided test*). I chose to do a *one-sided test* in this case because I had a directional hypothesis—I predicted that the mean would be *less* than 4—and a corresponding null hypothesis: the null hypothesis does not simply state that the population mean is the midpoint, such that we could reject the null hypothesis if the sample mean was suitably above *or* below 4, but that the mean is *at or above* the midpoint. This means that we can only reject the null hypothesis if the sample mean is suitably *below* the midpoint. As such, we can specify that the alternative hypothesis is that the mean is less than `mu=4` by adding the argument `alternative="less"` to the function call. In my experience, when you use a one sample t-test, it is quite likely that you'll have a directional hypothesis. But it might be that you simply predict that the mean is different from a specified value to make the test more conservative (it is easier to get a significant result using a one-tailed test than a two-tailed test). Indeed, it is common for researchers to report two-tailed tests even when they had a directional hypothesis. If this is desired, however, my preference would instead be to use a one-tailed test with a more stringent choice of significance level.

When we run our t-test, as specified above, R will provide an output that gives us a good bit of information about our statistical test:

```
    One Sample t-test

data:  D3$RESPONSE
t = -5.9851, df = 48, p-value = 1.321e-07
alternative hypothesis: true mean is less than 4
95 percent confidence interval:
     -Inf 2.971765
sample estimates:
mean of x
 2.571429
```

Perhaps most importantly, this output gives us the *p-value*, which is quite small indeed—$1.321e^{-07}$ or 0.000000132—and is obviously well below the conventional cut-off of 0.05. This means that we can say that the mean is *significantly* below the midpoint at the specified significance level (see also Chapter 1), although the latter specification of significance level is often left implicit. Other key information for reporting the test appears on the same line—the *t-value* (`t = -5.9851`) and the *degrees of freedom* (`df = 48`). Here is how I reported this in Sytsma (2010a, 124):

> $t(48) = -5.985$, $p < 0.001$ (one-tailed)

As illustrated here, for p-values below 0.001 (like 0.000000132) we typically just specify $p < .001$. In addition, it is important to specify whether you performed a one-tailed or two-tailed test, and in addition I would include the type of test in the text—e.g., that we performed a one-sample t-test comparing the mean response to the neutral point of 4.

Other useful information in the output is the 95% confidence interval. Testing a null hypothesis is intimately related to determining confidence intervals. Specifically, a null hypothesis is rejected at a specified significance level—standardly denoted α—when the value or range that the null hypothesis specifies is outside of the $(1 - \alpha) * 100\%$ confidence interval calculated for the sample. Using the 0.05 significance level, this is the 95% confidence interval

given in the output.[20] Since we used a one-tailed test, the lower bound is given as negative infinity in this output, indicating that no value we found would fall below the lower bound. As such, it is the upper bound of 2.971765 that matters for us here. Since 4 falls above this upper bound (2.971765 < 4), the null hypothesis can be rejected (at the 0.05 significance level). Often 95% confidence intervals will be included in bar graphs showing study results, as is shown in Figure 5 below. For this we would typically show the confidence interval for each side of the mean, however, which can be generated using a two-tailed test. This is easily done by modifying the above function call:

```
t.test(D3$RESPONSE, mu=4)
```

As expected, this produces a slightly different output, since we've removed the directionality of the test, including giving a smaller p-value and a confidence interval that is positive on boths ends:

```
        One Sample t-test

data:  D3$RESPONSE
t = -5.9851, df = 48, p-value = 2.643e-07
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 2.091511 3.051346
sample estimates:
mean of x
 2.571429
```

This new information can be reported by noting that the test showed a 95% confidence interval of [2.09, 3.05].

[20] (1 – 0.05) * 100% = 95%

3.7.6 Statistical Significance versus Effect Size

As noted above, the result of our test is significant, with the p-value ($1.321e^{-07}$) being less than the standard cut-off of 0.05. Simply referring to a result as 'significant', however, while common, is also potentially misleading. In ordinary language, describing something as significant often means that it is *important*, sometimes with a connotation that it is large (e.g., a sales manager for a company reporting that they expect significant additional sales in the next year). But this is not what we mean when we say that the result of a statistical test is significant. Here we need to distinguish between *statistical significance* and *effect size*. When we report a t-test and conclude that the result is significant, it is *statistical* significance we are reporting, and this simply means that we can reject the null hypothesis at the specified level. A result could be statistically significant, however, while the difference between the numbers we're comparing is very small (say, for example, a sample mean of 3.9 compared to the midpoint of 4). This is because statistical significance depends on sample size, such that if we had a large enough sample size in our study, even a very slight divergence of the sample mean from the comparison point (or between means as we'll see below) could be statistically significant. What we really want in addition to the p-value is a standardized indication of the size of the difference in our comparison. This is what effect size does. For a study like the present one, simply reporting the sample mean will give *some* indication of the effect size. Here, the mean response was 2.59 on a 7-point scale, placing it roughly 1.41 points below the midpoint. While this is likely adequate in the present situation, better practice is to also report a measure of the effect size that makes it easier for comparison.

For t-tests, the most common measure of effect size is Cohen's *d*. Cohen (1988) provides helpful rules of thumb for interpretation: $d = 0.2$ is a small effect, $d = 0.5$ a medium effect, and

$d = 0.8$ a large effect. Here you can think of a *small effect* as one that will be difficult to discern with the naked eye, a *medium effect* as one that can probably be discerned, and a *large effect* as one that can definitely be discerned. A handy example provided by Cohen is that the difference in height between 15-year-old and 16-year-old girls in the United States is a small effect, while the difference in height between 13-year-old and 18-year-old girls is a large effect. Cohen's *d* for the present study is readily calculated in R, here using a function from the **lsr** package that we installed and loaded at the beginning of the exercise:

```
cohensD(D3$RESPONSE, mu=4)
```

This outputs a value of 0.86, which indicates that the comparison between our sample mean and the midpoint of the scale shows a large effect size according to Cohen's rule of thumb.

3.7.7 A Non-parametric Alternative to a One Sample T-test

As discussed above, t-tests involve a number of assumptions—importantly including that the feature we're interested in is normally distributed in the population and that responses have interval scale—and often these assumptions are somewhat dubious for studies in experimental philosophy. This does not mean that you should avoid t-tests in your work, however. As noted, t-tests are generally quite robust to violations of these assumptions. Nonetheless, is is important to be mindful of such issues. One easy way to do this is to also run a comparable statistical test that doesn't make the same assumptions, such as a *non-parametric test*. Non-parametric tests do not make assumptions about the underlying distribution of the feature in the population or that your data has interval scale. There is never a free lunch, however, and these advantages of non-parametric tests have a cost: they come at the expense of some *statistical power*, meaning that you're less likely to get a significant result. One common non-parametric alternative to t-tests is

to use the Wilcoxon procedure, which is also easily run in R. For our present study this can be done with the following function call:

```
wilcox.test(D3$RESPONSE, mu=4, alternative="less")
```

This gives the following output (along with warnings that the p-value is an estimate, which we need not worry about here):

```
        Wilcoxon signed rank test with continuity correction

data:  D3$RESPONSE
V = 178, p-value = 9.672e-06
alternative hypothesis: true location is less than 4
```

As expected, the p-value is slightly lower than what we saw above for the corresponding t-test, but is overall fairly comparable: the result is highly significant on either measure. For thoroughness, you could report the Wilcoxon test alongside the t-test and Cohen's *d*, if desired, such as: $t(48) = -5.99$, $p < .001$ (one-tailed), $d = .86$; $V = 178$, $p < .001$ (one-tailed).

3.8 Interpretation

It is a commonly noted point that data is one thing, conclusions another, and that to draw a conclusion from a set of data involves interpretation. This is sometimes taken to suggest a divide between the objective, scientic study and the subjective, opiniated interpretation of it. But, as the above walkthrough hopefully draws out, there is no sharp divide like this in the research process. Indeed, the interests of the researcher are invariably present in forming a research question, in formulating and implementing a plan to address that question, and in analysing the results. This is not a bad thing. For instance, just focusing on the statistical analysis in the last section, given the data we collected there are any number of tests we might have run. For instance, we could have checked if there was a correlation between the responses of our participants to the test question and their age. This is certainly easy enough to do in R:

```
cor.test(D3$RESPONSE, D3$AGE)
# r=.063, p=.67
```

It is important to note, however, that our research question didn't specifically concern age and we made no predictions about the relation between responses and age. And absent such a prediction, there is no null hypothesis, such that the NHST framework makes no sense. This does not mean that there is anything wrong with running the correlation test, but it does bear on how we should think about the result: we should treat this as merely exploratory, such that if we had found a potentially interesting relation we'd want to confirm it with a new study directed at testing the relation and making the relevant prediction in advance of looking at the data. The basic reason is that if we look at enough comparisons for a given dataset, some of them are likely to be statistically significant just by chance. We'll elaborate on this point in Section 5 when we consider the question of correcting for multiple comparisons.

Nonetheless, while an element of interpretation is found throughout the research process, including the statistical analysis performed, once you've completed an analysis you'll want to describe what it means—you'll want to *interpret* those findings—typically focusing on drawing out the philosophically relevant conclusions. For the present study, recall that this project grew out of a wider set of work testing the common assertion in philosophy of mind that the concept of phenomenal consciousness is part of folk psychology, including the finding from Sytsma and Machery (2010) indicating that in contrast to philosophers, lay people treat two prototypical examples of supposed phenomenally consciousness mental states quite differently, happily ascribing *seeing red* to a simple robot while denying that the robot *feels pain*. Sytsma (2010a) explored one explanation for this finding, hypothesizing that lay people tend to hold a naïve view of both colors and pains. If people tend to hold such a view of pains, however, we would expect them to allow that there could be unfelt pains, in direct contradiction to the common justification

offered for the standard view in philosophy. The results of my second study gave an initial indication that people are open to the possibility of unfelt pains. My third study, then, attempted to test this more directly, generating my research question. This same research question guided the design and analysis of the study, and it in turn guides the interpretation of the results.

What we find is that the results from Study 3 in Sytsma (2010a) are in line with the prediction that lay Americans tend to allow for the possibility of unfelt pains. This in turn suggests against the standard view about the ordinary concept of pain in philosophy, while offering some support for the alternative hypothesis that the ordinary concept corresponds with a naïve view of pain. The evidence here is most direct for the explicit prediction, while the conclusion with regard to the opposed philosophical views are more tentative. Thus, as noted above, we shouldn't put too much weight on just a single study, or even a pair of studies, especially when it comes to rejecting or endorsing broad philosophical accounts. One reason is that this study involved a number of decisions points, including the vignette used, the question posed, and how participants were recruited. To make a truly compelling case for the claim that 'common sense' allows for unfelt pains, further studies are needed.

4. Further Studies, Further Tests

In this section, I'll detail a few of the many subsequent studies that have been conducted on the question of unfelt pain. I'll focus on two things in this discussion. First, we'll consider how follow-up studies can address potential concerns with previous studies. This will include varying the vignettes, answer choices, and recruitment strategy. Second, we'll illustrate two further types of t-tests—*independent samples t-tests* and *paired samples t-tests*—although we'll run through these in somewhat less detail, now that we're more familiar with the basics of statistical analysis

in R. We begin in §4.1 with the fourth study in Sytsma (2010a), which will be used to introduce *independent samples t-tests*. In §4.2, we'll use this study to discuss the relationship between *sample size*, *effect size*, and *power*, which is crucial for designing and interpreting effective empirical studies. In §4.3, we then turn to the first study in Sytsma and Reuter (2017), which uses a different type of vignette to assess judgments about unfelt pains and shifts from scale responses to binary answer choices; we'll use this as an opportunity to see how we can compare between these types of responses. Finally, in §4.4, we look at a study from Reuter and Sytsma (2020), which will be used to introduce *paired samples t-tests*.

4.1 Independent Samples T-tests

One issue that was raised in presenting the results of the study detailed in the previous section is that the vignette describes the situation in terms of an injured person being *distracted* from a pain. It is plausible, though, that you can only be distracted from something that exists, such that participants might have inferred from this wording that the pain was ongoing despite being unfelt. To address this criticism, in Study 4 of Sytsma (2010a), I revised the vignette to adjust the description, as shown in Figure 5, while keeping other details the same (including the question asked and recruitment method).

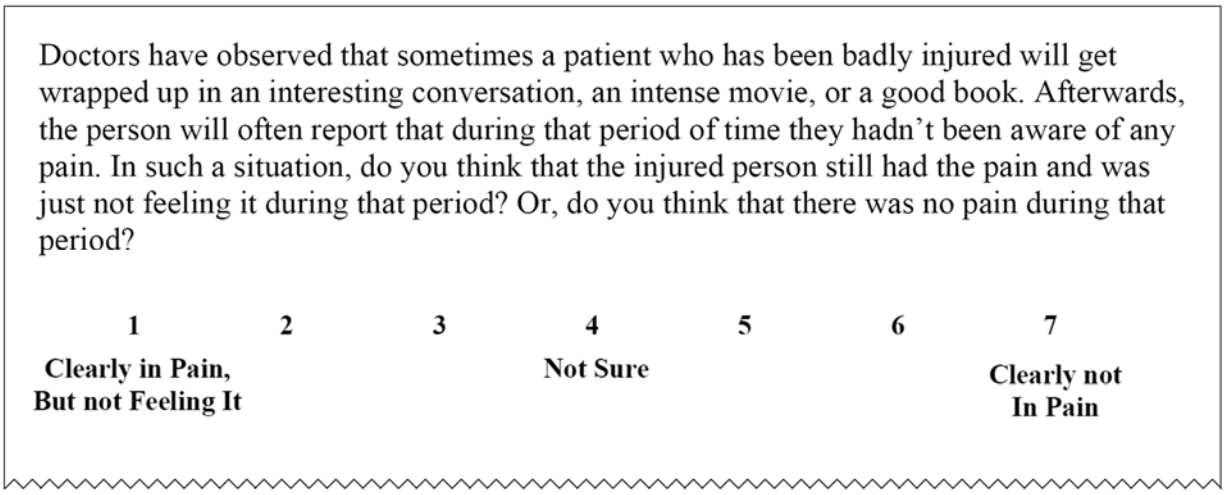

Figure 5: Vignette and scale for Study 4 in Sytsma (2010a, 125).

Data for this study is available in **[Sytsma_2010a_STUDY_4.csv]**. Let's begin by graphing the means for these two studies side-by-side, along with the histogram for Study 4, as shown in Figure 6. Code for producing these graphs and subsequent analyses can be found in **[Sytsma_2010a_STUDY_4.txt]**. After loading the data for each study (into `D3` and `D4` respectively), the code first produces a bar graph showing the means with 95% confidence intervals (as described in §3.7.5):

```
################
# Plot Studies #
################

# Run t-tests for confidence intervals
TD3 <- t.test(D3$RESPONSE, mu=4)
TD4 <- t.test(D4$RESPONSE, mu=4)

# Barplot with study means
barplot(height=c( mean(D3$RESPONSE), mean(D4$RESPONSE) ),
     width=0.5, ylim=c(1,7), xpd=FALSE,
     main="Unfelt Pain",
     names.arg=c("Study 3","Study 4"),ylab="Mean Response")

box(bty="l")

# Plot confidence intervals for Study 3
```

```
lines(x=c(0.35,0.35),y=c(TD3$conf.int[1],TD3$conf.int[2]),type="l")
lines(x=c(0.3,0.4),y=c(TD3$conf.int[1],TD3$conf.int[1]),type="l")
lines(x=c(0.3,0.4),y=c(TD3$conf.int[2],TD3$conf.int[2]),type="l")

# Plot confidence intervals for Study 4
lines(x=c(0.95,0.95),y=c(TD4$conf.int[1],TD4$conf.int[2]),type="l")
lines(x=c(0.9,1),y=c(TD4$conf.int[1],TD4$conf.int[1]),type="l")
lines(x=c(0.9,1),y=c(TD4$conf.int[2],TD4$conf.int[2]),type="l")
```

Based on the confidence intervals, we can tell that as in Study 3, the mean for Study 4 is significantly below the midpoint: the upper bar for the 95% confidence interval for each study in Figure 6 is well below the midpoint. This is telling us that our null hypothesis (that the mean is greater than or equal to 4) is outside of the range we expect the true value of the mean to fall with 95% probability (again given our test assumptions), indicating a probability of less than 5% that the null hypothesis is true (and where 5% can alternatively be expressed as 0.05). This inference can easily be confirmed by applying the same analysis we used before—starting with a one sample t-test, then checking the effect size, and finally confirming that the finding still holds when using a non-parametric test:

```
##################################
# Statistical Tests for Study 4 #
##################################

t.test(D4$RESPONSE, mu=4, alternative="less")
cohensD(D4$RESPONSE, mu=4 )
wilcox.test(D4$RESPONSE, mu=4, alternative="less")
```

These tests for Study 4 produce a comparable result to what we saw above for Study 3: $t(40) = -3.33$, $p < .001$ (one-tailed), $d = .52$; $V = 204$, $p = .0039$ (one-tailed). And, again, the results are in line with the hypothesis that a majority of the population holds that unfelt pains are possible.

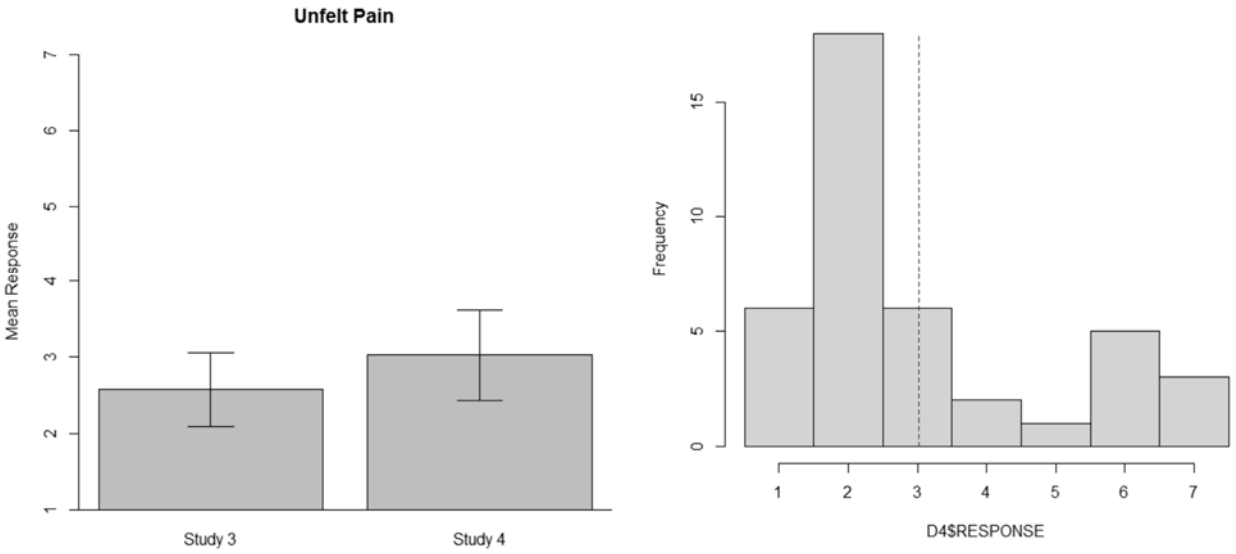

Figure 6: Bar graph for Studies 3 and 4 from Sytsma (2010a) on the left with 95% confidence intervals; histogram for Study 4 on the right.

The significant result for Study 4 suggests against the criticism of Study 3: it does not appear that use of the phrase 'distracted from' was a primary driver of the low mean for that study. Nonetheless, it is possible that this phrase does make *some* difference, even if participants still tend to judge that this is a case of unfelt pain even when the phrase is replaced. One way to test this is to directly compare the mean responses between the two studies. If 'distracted from' played a role in lowering mean responses, then we would predict that the mean for Study 4 ($M = 3.02$) would be higher than for Study 3 ($M = 2.57$). Just by looking at the means we can tell that Study 4 is higher, of course ($3.02 > 2.57$). The real question, though, is whether this difference is unlikely to simply be due to chance variation between the samples. To check this we need to run a statistical test.[21] Here what we want to do is to compare results from *two different samples*. Since we want to compare means between samples, and since these are

[21] Although, note that looking at Figure 6 and using the same logic as above, we can infer from the fact that the confidence interval for Study 4 includes the mean from Study 3 that this difference won't be significant at the 0.05 level.

different samples—the samples are *independent* of one another—what we want to use is an *independent samples t-test*.

We can conduct an independent samples t-test using the same `t.test()` function as before, but now we'll need to include an argument for the relevant table for each study (`D4` as well as `D3`) and an argument to indicate that the responses are from different participants—that the samples are *not* paired. As before, it is arguably appropriate to use a one-tailed test since the criticism makes a directional prediction. Here is the function call:

```
t.test(D3$RESPONSE, D4$RESPONSE, paired=FALSE, alternative="less")
```

And here is the output it produces:

```
        Welch Two Sample t-test

data:  D3$RESPONSE and D4$RESPONSE
t = -1.1981, df = 80.958, p-value = 0.1172
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1760992
sample estimates:
mean of x mean of y
 2.571429  3.024390
```

As expected, we find that the difference between the means is not significant at the 0.05 level. As such, we cannot reject the null hypothesis that the change to the vignette does not increase ratings.

You might have noticed that the test description in the output is different from our previous tests: not only did the number of samples noted change, as expected, but it is now described as a *Welch's t-test*. Recall from above that the label 't-test' describes a family of procedures. In fact, this family is even larger than I have indicated so far, including that there are multiple types of t-tests that we could run to compare between independent samples. The `t.test()` function in R defaults to a Welch's t-test for these. By contrast, the one sample t-

tests we ran above were *Student's t-tests*. We could have used a Student's t-test for the present comparison as well; we would simply need to set the `var.equal` argument in the function call to `TRUE`:

```
> t.test(D3$RESPONSE, D4$RESPONSE, paired=FALSE,
+ alternative="less", var.equal=TRUE)

        Two Sample t-test

data:  D3$RESPONSE and D4$RESPONSE
t = -1.2107, df = 88, p-value = 0.1146
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1689956
sample estimates:
mean of x mean of y
 2.571429  3.024390
```

As we can see, the p-value is comparable to what we found using Welch's t-test. Notice that the degrees of freedom differ between the two outputs. For the Student's t-test it is a whole number (88, which is the combined sample size for the two studies minus two, also see Chapter 1). For the Welch's t-test, however, it is 80.958. This reflects that the Student's t-test assumes that the values we're sampling from have equal variance—basically, that the values have the same spread with regard to the average—while the Welch's t-test does not make this assumption and applies an adjustment. We could perform a further test to see if our data supports the assumption of equal variance.[22] But, I wouldn't bother: there is a good reason that R performs a Welch's t-test by default; basically, whether the assumption of equal variance holds or not, a Welch's t-test will perform *at least* as well as a Student's t-test (see Delacre et al. 2017).

As with the one sample t-tests we performed above, we can also calculate the effect size for our comparison between Study 3 and Study 4:

[22] This is illustrated in **[Sytsma_2010a_STUDY_4.txt]** using *Levene's test*. Comparing Studies 3 and 4, this test gives a p-value of 0.51, meaning that we cannot reject the null hypothesis that the populations have equal variance.

```
cohensD(D3$RESPONSE, D4$RESPONSE)
```

This gives a Cohen's *d* of 0.26, which is a small effect size.[23] And, as before, we can address concerns based on the normality of our data and whether it is interval by performing a non-parametric alternative to the independent samples t-test using the same `wilcox.test()` function as above:

```
wilcox.test(D3$RESPONSE, D4$RESPONSE, paired=FALSE,
            alternative="less")
```

With independent samples, the function now performs a *Wilcoxon rank sum test*, as opposed to the signed rank test we saw before. As before, the non-parametric test gives a comparable p-value to that found for the t-test: $t(80.958) = -1.20$, $p = .12$, $d = .26$; $W = 863.5$, $p = .12$.

4.2 Sample Size, Effect Size, and Power

We've just seen that comparing the means between Studies 3 and 4 in Sytsma (2010a), we get a p-value above 0.05, which means that we cannot reject the null hypothesis at this significance level. Does this mean that we can conclude that the wording of the vignette does not matter? No. Being *unable to reject* the null hypothesis is not the same as *accepting* the null hypothesis. One issue here is that, as we noted above, p-values are tied to sample sizes. To illustrate, let's assume that we expect that the effect size we reported is roughly accurate and that the wording of the vignette does make a small difference, with the true effect size corresponding with a Cohen's *d* of 0.2. With this assumption in place, we can run some further tests to tell us how large our sample sizes would need to be for our study to be likely to *detect* an effect of that size (i.e., to get

[23] It might seem strange to calculate an effect size for a comparison that was not significant. Recall, however, that these are telling us two different things: roughly, statistical significance gives us a measure of how likely a difference is to be due to chance and depends on the sample size and the significance level we select, while effect size tells us how big that difference is. Thus, it can sometimes be useful to know what the effect size is if we were to accept a result, even if that result is non-significant at a given significance level.

a significant result). For instance, we could calculate how many participants we would need to sample to have an 80% chance to detect an effect with a Cohen's *d* of 0.2 using a t-test with a 0.05 significance level (and taking the test assumptions to hold). The chance to detect the effect is known as the *power* of the test, and is usually expressed as a probability. For an 80% chance, then, we'd want a power of 0.8, and for a 90% chance we'd want a power of 0.9.

We can calculate the sample size we would need for our study to have a given power using the **pwr** package in R.[24] Using the same sample size for each group, this can be calculated with the `pwr.t.test()` function. To do this we need to supply our expected effect size (*d* = 0.2), the significance level (default is 0.05), whether the test we're interested in is one- or two-tailed, and the power we want. Typical suggestions are to aim for a power between 0.8 and 0.9. The code document shows function calls for each of these options:

```
pwr.t.test(d=0.2, power=0.8, alternative="greater")  #n=309.8065
pwr.t.test(d=0.2, power=0.9, alternative="greater")  #n=428.8705
```

As we can see, these output a range of 310 to 429 participants. Note that this gives us the numbers we would need *per condition*. This means, for example, that to have a 90% chance of getting a significant result at the 0.05 level for an independent samples t-test checking that the mean response for the probe used in Study 4 is greater than the mean response for the probe used in Study 3, we would need roughly 858 participants (429 per condition)! Obviously, the studies in Sytsma (2010a) were a far cry short of this. Basically, the smaller the effect the more difficult it will be to detect it. Given this, we should be very cautious of concluding that there is no effect here—that whether the vignette uses 'distracted from' does not matter—as noted above. Rather,

[24] Remember to install (`install.packages("pwr")`) and load (`library(pwr)`) this package if you want to run the code for yourself!

we simply did not detect an effect (we did not get a significant result); but this is hardly surprising given the sample sizes and if we expect that the true effect might be rather small.

In fact, we can use the same **pwr** package to test the power these studies actually had to detect an effect of size $d = 0.2$. This calculates what is known as the *post hoc power* of the test:

```
pwr.t2n.test(d=0.2, n1=41, n2=49, alternative="greater")
```

In contrast to the previous test, here we use the `pwr.t2n.test()` function since the studies had different sample sizes, and we specify those sample sizes instead of the power: if we supply two out of three of *effect size*, *sample sizes*, and *power*, these functions will calculate the third. The result is that the pair of studies had just a 24% chance of detecting an effect of this size. Does this mean that the studies were *underpowered*? Not necessarily! This would depend on the effect size that we expect. To illustrate, let's suppose that what would be philosophically important here—what would vindicate the objection—is if there was a large effect size, say one of at least the size that we found above for comparing the mean in Study 3 to the midpoint ($d = 0.86$):

```
pwr.t2n.test(d=0.86, n1=41, n2=49, alternative="greater")
```

For detecting an effect of this size, we find that the two studies have a post hoc power of over 99%, meaning that we would be *very* likely to detect such an effect.

4.3 Binary Answer Choices and Dichotomization

While Study 4 from Sytsma (2010a) helps alleviate one concern with the vignette from Study 3, it is still roughly the same vignette. Perhaps there is something else about this story that tends to elicit responses from people that don't truly reflect their views about unfelt pains? Or perhaps the way the question was asked does this? Or perhaps the university students surveyed aren't representative of the wider population? Confidence in this finding is bolstered somewhat by the

results of the second study from Sytsma (2010a), discussed above, which asked about unfelt pains more directly, but it still behooves us to explore the hypothesis from further angles. One way to do this is to run additional studies that vary the vignettes used and the questions asked, as well as the recruitment method employed.

In Sytsma and Reuter (2017), we report on three studies concerning unfelt pains that adapt the vignette used in Study 5 in Sytsma (2010a). This study concerned the possibility of shared pains—that two people might feel one and the same pain if they were to share a body part, such as both being attached to the same hand. The standard view holds that the ordinary concept of pain precludes the possibility of shared pains in such cases (since pains are mental states it doesn't matter if a body part is shared), while the alternative predicts that people will tend to allow for this possibility (if pains are bodily states, then sharing a body part could lead to feeling the same pain). In the test condition of this study, participants were given a vignette describing two conjoined twins who share the lower part of their body. The twins run through a park, forcefully kick a rock hidden in the grass, and give behavioral indications of pain. Participants were then asked whether the twins felt *one and the same pain* or *two different pains*, answering on a 7-point scale. The mean response ($M = 3.29$) was significantly below the midpoint, indicating that participants tended to think that the twins felt one and the same pain.

In our first study in Sytsma and Reuter (2017), participants were given two different probes soliciting pain judgments, with the probes being given on separate pages and with the order of the two pages randomized. The first probe replicated Study 5 from Sytsma (2010a) that we just discussed: this study used the same vignette about conjoined twins, but adjusted the question to use a binary answer choice instead of a scale (participants answered by selecting either 'one and the same pain' or 'two different pains'). In addition, we added a comprehension

check question and used a different method for recruiting participants—we recruited them online using a *push strategy* rather than soliciting responses from students in class. A 'push strategy' involves recruiting participants who were not directly looking to participate in research by offering an alternative incentive.[25] What we found is that 68.5% (217/317) of participants who passed the comprehension check answered that the twins felt one and the same pain. Thus, we find a comparable result despite changing the question type and recruitment method, further suggesting against the standard view and in favor of the alternative naïve view.

More importantly for present purposes, the second probe concerned unfelt pains. Unlike Study 3 from Sytsma (2010a), however, this probe uses a vignette describing a pair of conjoined twins where just one takes a painkiller:

> Johnny and Tommy are conjoined twins that are joined at the torso. While they are distinct people, each with their own beliefs and desires, they share the lower half of their body. One day they accidentally dropped a heavy weight on their left foot. Johnny and Tommy both grimaced and shouted out 'Ouch!' They were then rushed to the hospital for treatment. Unfortunately, the nurse who checked them in was unfamiliar with conjoined twins. As a result, Johnny was given a pill for the pain while Tommy was left untreated. Ten minutes later, the doctor arrived to examine them. When she pushed on the injured foot, Tommy grimaced and shouted out 'Ouch!' while Johnny merely shrugged his shoulders and said it didn't hurt at all.

After reading the vignette, participants were asked to select which of two claims best reflected their view about it, with the choices presented in random order:

[25] In this case, participants were recruited through advertising for a free personality test on Google Ads, with the personality test being administered after the target questions. One notable benefit of using such a push strategy, in comparison to standard online recruitment methods in experimental philosophy (such as paid services like Amazon Mechanical Turk or Prolific Academic), is that participants are more likely to be 'experimentally naïve'—less likely to guess what the study is really about—and less likely to be motivated to provide the responses that they think the researchers are looking for (Haug 2018). Samples collected using the recruitment strategy employed here have been previously compared against samples collected with other methods in replication studies. And the present strategy has been consistently found to generate a diverse sample in terms of geography, socio-economic status, religiosity, political orientation, age, and education. Studies using this strategy have been previously reported in publications including, e.g., Livengood et al. (2010), Feltz and Cokely (2011), Murray et al. (2013), Machery et al. (2015), Livengood and Rose (2016), Livengood and Sytsma (2020), Fischer et al. (2021), Sytsma et al. (2012, 2015), and Sytsma (2010d, 2021, 2022), among many others.

> There was a pain in Johnny and Tommy's injured foot when the doctor pushed on it: While Tommy felt the pain in their foot, the painkiller prevented Johnny from feeling that pain.
>
> There was **not** a pain in Johnny and Tommy's injured foot when the doctor pushed on it: While the foot caused Tommy to feel pain, the painkiller prevented Johnny from feeling pain.

As with the shared pain probe, participants were also given a comprehension check question.[26] The order of the two probes was randomized, but responses did not vary noticeably based on which probe participants saw first. We found that 83.7% (251/300) of participants who passed the comprehension check selected the first answer for the present probe, indicating that they thought of this as a case of unfelt pain.

Results were comparable in our second study, which tweaked the answer choices for the unfelt pain question to emphasize that we meant the pain claims literally:

> There actually is a pain in the injured foot: while Tommy feels the pain in the foot, the painkiller prevents Johnny from feeling that pain.
>
> There is **not** actually a pain in the injured foot: while the foot causes Tommy to have the feeling of there being a pain in the foot, the painkiller prevents the foot from causing Johnny to have such a feeling.

This time 84.0% (110/131) selected the first option. Our third study further tweaked the answer choices, with the first option now including that 'the pain is literally in Johnny and Tommy's injured foot'. Again, a significant majority of participants selected the unfelt pain option (65.7%, 109/166).

In some ways, the use of binary response choices better matches the main hypotheses at issue for our research on unfelt pains: this work aims to test the contention that common sense denies that there can be unfelt pains, but this contention doesn't make any clear claims about

[26] Participants were asked, 'Did Johnny and Tommy drop a heavy weight on their left foot?' and answered by selecting either 'yes' or 'no'.

relative strength of belief or the extent to which people will be unsure about the issue. At the same time, using binary response choices like those given above forces participants to make a choice—indeed, these are sometimes called 'forced-choice questions'—and we might therefore worry that when participants aren't sure they will just answer randomly (see Sytsma and Livengood 2015, Section 9.2, for discussion). Using a scale like the one employed in the studies discussed in the previous section resolves this issue by giving participants the ability to register level of belief or to indicate uncertainty (by selecting the midpoint on the scale). The flip side of this is that the inclusion of such options might cause participants to be more cautious, leading to weaker responses than they would otherwise give.

Fortunately, when in doubt, you can always replicate your studies using different types of response choices and compare. One way to compare between binary and scale responses is to *dichotomize* the scale results. For instance, for Studies 3 and 4 from Sytsma (2010a) we can split the participants into groups based on whether they gave a response indicating that they judged the case to be one of unfelt pain (answering 1, 2, or 3 on the 7-point scale) or not (answering 4 or higher). Here is the code to dichotomize the results from Study 3, as shown in the **[Sytsma_2010a_STUDY_3.txt]** document:

```
#########################
# Dichotomize Responses #
#########################

# Percentage unfelt pain: 83.7%
nrow(D3[D3$RESPONSE==1|D3$RESPONSE==2|D3$RESPONSE==3, ])/nrow(D3)
```

Running this line we find that 83.7% (41/49) of participants in Study 3 judged there to be an unfelt pain. Calling on the statistical test discussed in Chapter 1 (χ^2 test) we can then test whether this proportion is significantly greater than 50% using the `prop.test()` function in R:

```
prop.test(x=41, n=49, p=0.5, alternative="greater")
```

Not surprisingly, we find that this proportion is significantly greater than 50%: $\chi^2 = 20.9$, $p < .001$ (one-tailed). Doing the same thing for Study 4, we find that 73.2% (30/41) of participants judged there to be an unfelt pain, which is again significantly greater than 50%: $\chi^2 = 7.90$, $p = .0025$ (one-tailed). Finally, we could compare either of these proportions to another, such as a proportion that was assessed directly using binary answer choices. To illustrate, let's compare the proportion of positive responses in Study 3 from Sytsma (2010a) to the proportion from Study 1 in Sytsma and Reuter (2017). We can do this using the same `prop.test()` function, although we'll now use the concatenate function `c()` to specify both the `x` value (positive count) and the `n` value (total count) for each proportion and we'll use a two-tailed test since we don't have reason to predict that either propotion would be greater than the other:

```
prop.test(x=c(41,251), n=c(49,300))
```

Given that these proportions are remarkably similar (`0.8367347` vs `0.8366667`), it is no surprise that the difference is not statistically significant. Indeed, we we get a p-value of 1.

4.4 Paired Samples T-tests

Finally, in Reuter and Sytsma (2020) we detail a large series of further studies testing whether common sense countenances unfelt pains. This includes a study replicating Study 4 from Sytsma (2010a) using the online push strategy from the previous study and binary response choices:

> Which of the following descriptions of this type of situation seems most appropriate to you?
>
> The injured person still had the pain and was just not feeling it during that period.
>
> The injured person had no pain during that period.

We found that 90.3% (28/31) of participants selected the first option, which is somewhat higher than the proportion from Study 4 noted above, although the difference is not significant ($\chi^2 = 2.31$, $p = .12$). Our paper also included four studies involving an injured patient taking a painkiller, as in the thought experiment from Aydede noted in Section 2. Unlike the painkiller studies just discussed from Sytsma and Reuter (2017), this time the vignettes did not involve conjoined twins and we used both scales and binary response options. In each case we again found that a significant majority of participants judged that the patient had a pain even though they didn't feel it while the painkiller was in effect. Another set of studies instead used vignettes describing a severely injured soldier who professes not to feel any pain, as in the thought experiment from Hill discussed in Section 1, with similar results. Thus, while Hill (2009, 171) states that when he 'asked informants to assess the likelihood of this scenario […] they have all been inclined to dismiss it as absurd', our results were quite different: In each of nine studies varying both the vignettes and the questions we found that a significant majority of participants responded that the injured soldier had a pain despite not feeling it.

Finally, Reuter and I report a series of seven studies that asked participants more direct questions about the possibility of unfelt pains. Perhaps most strikingly, in our 15th and 16th studies we asked the following four questions, with participants either answering by selecting 'yes' or 'no' (Study 15) or using a 7-point scale anchored at 1 with 'clearly no', at 4 with 'not sure', and at 7 with 'clearly yes' (Study 16):

> (1) Is it possible for a person to have a pain that they don't feel for a period of time?
> (2) Have you ever had a pain that you didn't feel for a period of time?
>
> (3) Is it possible for a person to have a pain that doesn't hurt for a period of time?
> (4) Have you ever had a pain that didn't hurt for a period of time?

In each case a significant majority of participants gave an affirmative answer to each question, giving a further indication that lay people tend to hold that unfelt pains are possible, and further that they tend to hold that they are actual. Focusing on Study 16, we can show this by running a similar analysis to what we saw before, using one sample t-tests to compare the means to the midpoint for each of the four questions. This is illustrated in **[RS_2020_STUDY_16.txt]** using the data provided in **[RS_2020_STUDY_16.csv]**. (We can also dichotomize and compare the proportion to 50% or to the binary responses from Study 15, as we did in §4.3, and as is illustrated in the code document.) Analyzing each question separately in this way, however, raises the *potential* issue of correcting for multiple comparisons, which I return to in the next section.

Our purpose in analyzing Study 16 was, once again, to test whether lay people tend to deny the possibility of unfelt pains as the standard view contends. And for this purpose the one sample t-tests just noted do the trick. But we could have asked other questions, here, motivated by different research interests. One possibility is comparing between the two sets of questions. I'll do that here for purposes of introducing a third type of t-test—*paired samples t-tests*.

One interesting facet of Study 16 from Reuter and Sytsma (2020) is that we didn't just ask participants about the *possibility* of unfelt pains, but their *actuality*. This is done using two different wordings, with Questions 1 and 2 forming a pair and Questions 3 and 4 forming a pair (the order of these pairs was counterbalanced in the studies). We might wonder whether responses differed within these pairs. Indeed, since more things are possible than are actual, we would predict that people would be more likely to affirm the possibility questions than the actuality questions (although, again, this isn't a prediction we specifically made in the actual paper). Making this prediction for illustrative purposes, we can then test it using a *paired*

samples t-test. A paired samples test is called for in this case because we're comparing mean responses, but these responses were given by the *same* participants—each participant in Study 16 answered all four of the questions—such that an independent samples t-test wouldn't be appropriate: these participants are not independent.

As before, we can run the paired samples comparison using the `t.test()` function in R. For this all we need to do is to change the `paired` argument from `FALSE` to `TRUE` in our function call, as illustrated here for the first pair of questions:

```
> t.test(D$RESPONSE_1, D$RESPONSE_2, paired=TRUE,
+ alternative="greater")

        Paired t-test

data:  D$RESPONSE_1 and D$RESPONSE_2
t = 3.26, df = 61, p-value = 0.0009122
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 0.3854137       Inf
sample estimates:
mean of the differences
              0.7903226
```

The output indicates that we've conducted a paired samples t-test and that the difference is significant (`p-value = 0.0009122`). This means that, as expected, people were more likely to judge that unfelt pains are possible than that they have actually had an unfelt pain. As before, we can also calculate the effect size using `cohensD()` and run a non-parametric comparison using `wilcox.test()`:

```
cohensD(D$RESPONSE_1, D$RESPONSE_2)
wilcox.test(D$RESPONSE_1, D$RESPONSE_2, paired=TRUE,
alternative="greater")
```

The combined result is: $t(61) = 3.26$, $p < .001$, $d = .46$; $V = 401.5$, $p = .0012$.[27] And we get comparable results comparing Questions 3 and 4: $t(61) = 2.79$, $p = .0035$, $d = .27$; $V = 223.5$, $p = .0039$.

5. Correcting for Multiple Comparisons

To close this chapter, it is important to consider a possible complication that I noted in the previous section: when conducting multiple statistical comparisons on the same set of data we need to be mindful of how this might affect the way we should interpret our p-values. This concern isn't specific to t-tests, but I most often see this type of issue arise in the x-phi literature for sequences of t-tests, such that it makes sense to address it in this context. First, in §5.1, I'll discuss why we sometimes need to correct for multiple comparisons and provide an example where I don't think such a correction is needed. In §5.2, I'll then discuss different types of correction, focusing on two methods—the *Bonferoni method* and the *Holm method*. Finally, in §5.3, I'll present a case where I do think a correction is needed and show you how to apply the Holm method to this case in R.

5.1 Why should we apply a correction?

For Study 16 in Reuter and Sytsma (2020) participants were each asked four questions about unfelt pains, and in our analysis we began by comparing the mean response for each to the midpoint using a sequence of four one sample t-tests, as discussed in the previous section. I noted above that this potentially raises a complication, however. Simply put, the worry is that if you test enough variables, it is quite likely that some will be significant just by chance even if the

[27] As before, we could also dichotomize these questions and compare proportions, although for paired data like this we will now want to run a McNemar's test, as illustrated in the code document.

null hypothesis is true. To illustrate, imagine that you're worried that a mint is producing coins that are biased toward coming up heads. Say that you test this by flipping a single coin 10 times. If it were to come up heads all 10 times, this would be some evidence for your worry. After all, such an outcome is quite unlikely if the coin is fair (roughly a probability of 0.00098). But what if you were to test 1000 coins this way? Now it would be more likely than not that at least one of the coins would come up heads 10 times in a row (roughly a 0.62 probability) even if all the coins are fair.[28] As such, finding that one of the coins came up heads 10 times would hardly be evidence that the mint is producing biased coins. Running t-tests on lots of variables without correcting for multiple comparisons faces a corresponding problem. In this analogy, each t-test corresponds with testing a different coin: the more tests we run, the more likely it is that we'll get one or more significant results just by chance even if the null hypothesis is true. Of course, this is possible even if we run just one test. This is why we report p-values, since they give us a sense of the likelihood of getting a result at least this extreme by chance. But the point is that running multiple tests will affect how we should think about those chances when considered as a group.

For the one sample t-tests in our analysis of Study 16 it is unclear that this is a serious worry, however. And, indeed, I'm inclined to think that it is not. The reason is that we were predicting a *pattern* of results across the four questions. In more detail, we varied two things across the four questions—*modality* and *phrasing*. With regard to modality, Questions 1 and 3 asked about the possibility of unfelt pain, while Questions 2 and 4 asked about their actuality for

[28] The probability that a single fair coin will come up heads on a single toss is 1/2. Let's write this P(H). The probability that it would come up heads 10 times in a row is then P(H) * P(H) * P(H) * P(H) * P(H) * P(H) * P(H) * P(H) * P(H) * P(H), which is $(1/2)^{10}$ or 1/1024. Let's call this P(10H). The probability that this would occur at least once in 1000 attempts is equivalent to the one minus the probability that it doesn't occur in any of the 1000 tests, which is equivalent to $(1 – P(10H))^{1000}$ or $(1023/1024)^{1000}$, which is roughly 0.3764. As such, $1 – (1 – P(10H))^{1000}$ is roughly 0.6236.

the participant. And while we expected that this would make a difference, for the reason detailed above, we nonetheless predicted that participants would still tend to affirm the actuality questions. With regard to phrasing, Questions 1 and 2 were phrased in terms a pain that wasn't felt, while Questions 3 and 4 were phrased in terms of a pain that didn't hurt. We didn't expect that the specific phrasing would make an important difference. As such, we predicted that participants would tend to affirm each of the four questions; and if they tended to deny any of the four, this would provide some evidence against the general hypothesis. Corrections for multiple comparisons, however, essentially make it *tougher to get significant results* at a given significance level. Given our prediction of the pattern of results across the questions, however, this doesn't seem warranted.

5.2 The Bonferroni Method and the Holm Method

Exactly when one should apply a correction for multiple comparisons is a difficult question and there is much disagreement on this score. Nevertheless, sometimes it is essential that we apply such a correction, as I'll illustrate in the next section. Further, when in doubt, I would recommend that you go ahead and apply a correction, as this will make your tests more stringent and hence render the results more convincing.

There are many different types of corrections that can be applied. Perhaps the most common is the *Bonferroni method*. The main positive of this method is that it is quite easy to use: we simply multiple the p-values of a sequence of tests by the number of tests performed (or, equivalently, keep the p-values the same but adjust the significance level by dividing it by the number of tests). Let's say that we decided that a correction for multiple comparisons is appropriate for the one sample t-tests performed for Study 16. We performed four tests, getting

p-values of $5.9e^{-13}$ (i.e., 0.00000000000059), 0.0019, $7.5e^{-8}$, and 0.0023, respectively. To apply the Bonferroni correction, we simply multiply each of these p-values by four (the number of tests performed), giving corrected values of $2.0e^{-12}$, 0.0074, $3.0e^{-7}$, and 0.0093. (Note that I've applied the correction to the full p-values given by R to minimize rounding errors.) As we can see, the results remain significant at the 0.05 level. The Bonferroni correction can also be applied using the `p.adjust()` function in R with the `method` argument set to `"bonferroni"`, as demonstrated in the code file.

While the Bonferroni method has the benefit of being simple, I wouldn't personally recommend this correction. The reason is the same as we saw above with regard to using a Student's t-test for comparing the means from independent samples: there is another method that is always at least as powerful. In the case of the Bonferroni method, there is is an extension—what is known as the *Holm method* or the *Holm–Bonferroni method*—that makes the same assumptions as the Bonferroni method and is always at least as powerful as it (Holm 1979). And while the Holm method is somewhat more complicated, it is equally easy to apply in R: all you need to do is to switch the value for the `method` argument in the `p.adjust()` function to `"holm"`. As shown in the code file, applying this method to the sequence of tests for Study 16 gives lower p-values than the Bonferroni method for three of the four questions: $2.0e^{-12}$, 0.0037, $2.3e^{-7}$, and 0.0037.

A second complication is that while I compared Questions 1 and 2 and Questions 3 and 4 separately above to illustrate the use of paired samples t-tests, this does not tell us the shared impact across the two sets of questions as the modality is varied. And if we also wanted to test the impact of the phrasing, we would then need to run two further t-tests using this method, now comparing Questions 1 and 3 and Questions 2 and 4. A better solution is to recognize that we are

crossing two variables (or 'factors') in our study, each taking on one of two values (or 'levels'): as noted above, we're varying the *phrasing* (feel, hurt) and the *modality* (possible, actual), with one question corresponding with each combination of values for these two variables. Recognizing this, we could test the impact of each variable across the questions, as well as their interaction, using an ANOVA (specifically, a two-way within-participants ANOVA). Tests like this will be discussed further in Chapter 3, so I won't try to explain them here, but two ways of performing the ANOVA in R are shown in the code document.[29] In line with our paired samples t-tests, we find that there is a significant main effect for *modality* ($p = .0012$). Further, we do not find a significant main effect for *phrasing* ($p = .27$) or for the interaction of these two factors ($p = .17$).

5.3 Illustrating the Holm Method

To conclude, I want to briefly discuss a final study where a correction for multiple comparisons is clearly called for. While this study does not involve judgments about unfelt pains, it does provide evidence for the underlying hypothesis about the commonsense conception of pain that motivated my predictions in the studies we discussed above.

In Sytsma and Snater (2023a), we conducted a global study in which participants answered a large number of test questions. Drawing on studies from Ozdemir (2022) as well as Fischer and Sytsma (2021), we either gave participants a vignette describing future scientists creating a physical duplicate of a person or creating an android duplicate of a person, then asking them whether they agreed or disagreed with each of 25 statements ascribing a mental capacity to

[29] For this, we need to restructure the data for Study 16: we need to add columns for each of our two variables, and we need to convert it to 'long form'—adding a column with a participant idea and repeating the data set so that each row shows the response for just one question. While this conversion can be done in R, to make things easier I've instead created a second spreadsheet with the converted data: **[RS_2020_STUDY_16_ANOVA.csv]**.

the resulting duplicate. Participants responded using a 7-point scale anchored at 1 with 'Disagree Strongly' and at 7 with 'Agree Strongly'. As part of the analysis detailed in the supplemental materials (Sytsma and Snater 2023b), we compared responses between the two conditions using a series of independent samples t-tests, as illustrated in **[SS_2023_STUDY_1.txt]** using the data provided in **[SS_2023_STUDY_1.csv]**. Without applying a correction for multiple comparisons, we found 10 significant differences at the 0.05 level. But unlike in Study 16 in Reuter and Sytsma (2020), we were not predicting a specific pattern of results across these tests and did not have specific predictions for all of these comparisons. Further, many of the significant results had negligible effect sizes. Given the large number of tests, it is quite likely that some of these significant results owed to chance—more than we should accept at the 0.05 significance level selected—and hence it was important for us to apply a correction for multiple comparisons. We did this using the Holm method introduced above. Applying the correction, we found that only three of the results remained significant. This included the question we asked about feeling pain, which we had a specific prediction for.

Recall that in Sytsma (2010a) I hypothesized that the reason lay people in Sytsma and Machery's (2010) first study tended to ascribe seeing red to a simple robot, but not feeling pain, is that people tend to hold a naïve view of both types of qualities. With regard to pains, I speculated that people tend to conceive of pains as being qualities of injured body parts, but that the entity needs the right sort of body parts to instantiate pains: they need soft and fleshy body parts, not hard and metallic ones. And, indeed, this hypothesis was directly tested in Sytsma (2012), where I found that giving the simple robot from Sytsma and Machery (2010) grasping arms made of bioengineered materials, instead of the original hard and metallic ones, notably increased ascriptions of feeling pain. Based on this, in Sytsma and Snater (2023a) we predicted

that participants would be significantly more likely to judge that the physical duplicate felt pain than that the android duplicate felt pain. And, indeed, this prediction was borne out, with participants being significantly more likely to agree with the statement 'the duplicate would feel pain when she is injured' in the physical duplicate condition compared to the android condition, even after correcting for multiple comparisons.

Box 1: Partially Paired Samples T-tests

In the main text we discuss three main types of t-tests: *one sample t-tests*, *independent samples t-tests*, and *paired samples t-tests*. These are distinguished by the number of conditions we are comparing (one for one sample t-tests, two for independent samples t-tests and paired samples t-tests) and whether the same participants make up the samples (no for independent samples t-tests, yes for paired samples t-tests). It is possible, however, that the answer could be *yes and no*: some of our participants could be the same across the two samples while others could be different. This is an unusual situation, and not one you're likely to run across. Indeed, while I include the test here for completeness, I've only run a partially paired samples t-test one time in my own work.

Specifically, a partially paired samples t-test was relevant to the analysis of Study 2 in Sytsma et al. (2022). One goal of this study was to test whether participants' judgments about a statement would differ if it was presented alone versus being presented alongside three other statements. In order to test this, we ran the study with both a *within-participants* condition (each participant giving judgments about all four statements) and *between-participants* conditions (each participant giving a judgment about just one of the four statements). Comparing judgments for each statement between conditions could be done using independent samples t-tests, as we saw in the main text. Doing so, we found no significant differences for any of the fourst

statements. A second prediction about this study concerned a comparison between two different statements. This could be done separately for the within-participants condition (using a paired samples t-test) and for the between-participants conditions (using an independent samples t-test). Given that the results were not significantly different between the types of conditions, though, there is reason to combine these conditions: it would allow us to conduct just one test that would have greater statistical power. This couldn't be done using the standard t-tests we've reviewed, however, since combining the data would mean that some participants were paired (having given judgments about both statements) and others were not (having given judgments about just one statement). Fortunately, this is exactly the (rare) type of situation in which a *partially paired samples t-test* is appropriate. Unfortunately, this type of test is uncommon enough, that I was unable to find a package in R that implements it. Instead, I adapted the code provided by Henriksen (2018) for the `t.test.partial()` function.

References

Albert, Jim (2009). *Bayesian Computation with R. 2nd ed.* New York: Springer.

Arico, Adam (2010). "Folk psychology, consciousness, and context effects." *Review of Philosophy and Psychology*, 1(3), 371–393.

Aydede, Murat (2005a). "Preface," in *Pain: New Papers on its Nature and the Methodology of its Study*, M. Aydede (ed.), Cambridge, MA: MIT Press: ix–xvii.

Aydede, Murat (2005b). "Introduction: A critical and quasi-historical essay on theories of pain," in *Pain: New Papers on its Nature and the Methodology of its Study*, M. Aydede (ed.), Cambridge, MA: MIT Press: 1–58.

Aydede, M. (2009). "Pain," in *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), ed. E. Zalta, http://plato.stanford.edu/archives/spr2013/entries/pain/

Benjamin, Daniel, James Berger, Magnus Johannesson, et al. (2018). "Redefine Statistical Significance," *Nature Human Behaviour*, 2: 6–10.

Bluhm, Roland (2016). "Corpus Analysis in Philosophy." In M. Hinton (ed.), *Evidence, Experiment and Argument in Linguistics and the Philosophy of Language*, pp. 91–109, Peter Lang.

Borg, Emma, Richard Harrison, James Stazicker, and Tim Salomons (2020). "Is the folk concept of pain polyeidic?" *Mind & Language*, 35(1): 29–47.

Buckwalter, Wesley and Mark Phelan (2013). "Function and feeling machines: a defense of the philosophical conception of subjective experience." *Philosophical Studies*, 166(2), 349–361.

Caton, Jacob (2020). "Using Linguistic Corpora as a Philosophical Tool." *Metaphilosophy*, 51(1): 51–70.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences. 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Coninx, Sabrina, Pascale Willemsen, and Kevin Reuter (2023). "Pain Linguistics: A Case for Pluralism." *The Philosophical Quarterly*, pqad048.

Cova, Florian et al. (2019). "Estimating the Reproducibility of Experimental Philosophy." *Review of Philosophy and Psychology*, 12: 9–44.

Chalmers, David (1995). "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies*, 2: 200–219.

Chalmers, David (2018). "The meta-problem of consciousness." *Journal of Consciousness Studies*, 25 (9-10): 6–61.

Delacre, M., D. Lakens, and C. Leys, C. (2017). “Why Psychologists Should by Default Use Welch’s t-test Instead of Student’s t-test.” *International Review of Social Psychology*, 30(1), 92–101.

Dennett, Daniel (1991). *Consciousness Explained*. New York: Little, Brown and Company.

Díaz, Rodrigo (2021). “Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness.” *Journal of Consciousness Studies*, 28(3-4): 55–75.

Feltz, Adam and Edward Cokely (2011). “Individual differences in theory-of-mind judgments: Order effects and side effects.” *Philosophical Psychology*, 24(3): 343-355.

Fiala, Brian, Adam Arico, and Shaun Nichols (2012). “You, Robot.” In E. Machery and E. O’Neill (eds.), *Current Controversies in Experimental Philosophy*, pp. 31–47, New York: Routledge.

Fischer, Eugen, Paul Engelhardt, and Justin Sytsma (2021). “Inappropriate stereotypical inferences? An adversarial collaboration in experimental ordinary language philosophy.” *Synthese*, 198(11): 10127–10168.

Fischer, Eugen and Justin Sytsma (2021). “Zombie Intuitions.” *Cognition*, 215: 104807.

Fischer, Eugen and Justin Sytsma (forthcoming). “Projects and Methods of Experimental Philosophy.” In A. Bauer and S. Kornmesser (eds.), *The Compact Compendium of Experimental Philosophy*, de Gruyter.

Goldberg, Benjamin, Kevin Reuter, and Justin Sytsma (forthcoming). “The History of the Concept of Pain: How the Experts Came to be Out of Touch with the Folk.” In K. Hens and A. De Block (eds.), *Advances in Experimental Philosophy of Medicine*, Bloomsbury

Gonnerman, Chad (2018). “Consciousness and Experimental Philosophy.” In R. Gennaro (ed.), *The Routledge Handbook of Consciousness*, pp. 463-477, New York: Routledge.

Goodwin, Kerri and C. James Goodwin (2016). *Research in Psychology: Methods and Design, 8th Edition.* Hoboken, NJ: John Wiley & Sons.

Gregory, Daniel, Malte Hendrickx, and Cameron Turner (2022). “Who knows what Mary knew? An experimental study.” *Philosophical Psychology*, 35(4): 522–545.

Henriksen, Askel Anker (2018). “T-test for partially paired data.” https://aksela.wordpress.com/2018/09/08/t-test-for-partially-paired-data/

Hill, Christopher (2009). *Consciousness*, Cambridge: Cambridge University Press.

Holm, Sture (1979). “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics*, 6(2), 65–70.

Huebner, Bryce (2010). “Commonsense concepts of phenomenal consciousness: does anyone care about functional zombies?” *Phenomenology and the Cognitive Sciences*, 9(1), 133–155.

Kim, Hyo-eun, Nina Poth, Kevin Reuter, and Justin Sytsma (2016). “Where is your pain? A Cross-cultural Comparison of the Concept of Pain in Americans and South Koreans,” *Studia Philosophica Estonica*, 9(1): 136–169.

Knobe, Joshua and Jesse Prinz (2008). “Intuitions about Consciousness: Experimental Studies,” *Phenomenology and the Cognitive Sciences*, 7: 67–83.

Kripke, Saul (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Livengood, Jonathan and David Rose (2016). “Experimental Philosophy and Causal Attribution.” In J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell, 434–449.

Livengood, Jonthan, Justin Sytsma, Adam Feltz, Richard Scheines, and Edouard Machery (2010). “Philosophical Temperament.” *Philosophical Psychology*, 23(3): 313–330.

Livengood, Jonathan and Justin Sytsma (2020). “Actual Causation and Compositionality.” *Philosophy of Science*, 87(1): 43–69.

Liu, Michelle (2020). “The intuitive invalidity of the pain-in-mouth argument.” *Analysis*, 80(3): 463–474.

Liu, Michelle (2023). “The Polysemy View of Pain.” *Mind & Language*, 38(1): 198–217.

Machery, Edouard and Justin Sytsma (2011). “Robot Pains and Corporate Feelings.” *The Philosophers’ Magazine*, 1st Quarter: 78–82.

Machery, Edouard, Justin Sytsma, and Max Deutsch (2015). “Speaker’s Reference and Cross-cultural Semantics.” In A. Bianchi (ed.), *On Reference*, Oxford University Press, 62–76.

Murray, Dylan, Justin Sytsma, and Jonathan Livengood (2013). “God Knows (But does God Believe?)” *Philosophical Studies*,166: 83–107.

Nagel, Thomas (1974). “What is it like to be a bat?” *The Philosophical Review*, 83: 435–450.

Ozdemir, Eyuphan (2022). *Empirical Evidence Against Phenomenal Theses*. PhD Dissertation, Victoria University of Wellington.

Peressini, Anthony (2013). “Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality.” *Philosophical Psychology*, 27(6): 862–889.

Phelan, Mark (forthcoming). “Experimental Philosophy of Mind.” In A. Bauer and S. Kornmesser (eds.), *The Compact Compendium of Experimental Philosophy*, de Gruyter.

Phelan, Mark, Adam Arico, and Shaun Nichols (2013). “Thinking things and feeling things: on an alleged discontinuity in the folk metaphysics of mind.” *Phenomenology and the Cognitive Sciences*, 12: 703–725.

Reid, Thomas (1785). *Essays on the Intellectual Powers of Man*. University Park, PA: Pennsylvania State University.

Reuter, Kevin (2011). “Distinguishing the appearance from the reality of pain.” *Journal of Consciousness Studies*, 18(9-10): 94–109.

Reuter, Kevin (2017). “The Developmental Challenge to the Paradox of Pain.” *Erkenntnis*, 82(2): 265–283.

Reuter, Kevin, Dustin Phillips, and Justin Sytsma (2014). “Hallucinating Pain,” in *Advances in Experimental Philosophy of Mind*, J. Sytsma (ed.), London: Bloomsbury.

Reuter, Kevin, Michael Sienhold, and Justin Sytsma (2019). “Putting Pain in its Proper Place.” *Analysis*, 79(1): 72–82.

Reuter, Kevin and Justin Sytsma (2020). “Unfelt Pain.” *Synthese*, 197: 1777–1801.

Salomons, Tim, Richard Harrison, Nat Hansen, James Stazicker, Astrid Sorensen, Paula Thomas, and Emma Borg (2021). “Is pain ‘all in your mind’? Examining the general public’s views of pain.” *Review of Philosophy and Psychology*, 1–16.

Schuman, Howard and Stanley Presser (1996). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage.

Shadish, William R., Thomas D. Cook, and Donald Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference. 2nd Edition.* Boston: Cengage Learning.

Strickland, Brent and Aysu Suben (2012). “Experimenter Philosophy: the Problem of Experimenter Bias in Experimental Philosophy.” *Review of Philosophy and Psychology*, 3: 457–467.

Sudman, Seymour, Norman Bradburn, and Norbert Schwarz (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Sytsma, Justin (2009). “Phenomenological Obviousness and the New Science of Consciousness,” *Philosophy of Science*, 76(5): 958–969.

Sytsma, Justin (2010a). “Dennett’s Theory of the Folk Theory of Consciousness,” *Journal of Consciousness Studies*, 17(3–4): 107–130.

Sytsma, Justin (2010b). “Folk Psychology and Phenomenal Consciousness,” *Philosophy Compass*, 5(8): 700–711.

Sytsma, Justin (2010c). *Phenomenal Consciousness as Scientific Phenomenon? A Critical Investigation of the New Science of Consciousness*. Ph.D. Dissertation. Pittsburgh, PA: University of Pittsburgh.

Sytsma, Justin (2010d). “The Proper Province of Philosophy: Conceptual Analysis and Empirical Investigation.” *Review of Philosophy and Psychology*, 1(3): 427–445.

Sytsma, Justin (2012). “Revisiting the Valence Account,” *Philosophical Topics*, 40(2): 179–198.

Sytsma, Justin (2013). “The Robots of the Dawn of Experimental Philosophy of Mind,” in *Current Controversies in Experimental Philosophy*, E. Machery and E. O’Neill (eds.), New York: Routledge.

Sytsma, Justin (2014). *Advances in Experimental Philosophy of Mind*, London: Bloomsbury.

Sytsma, Justin (2016). “Attributions of Consciousness.” In J. Sytsma and W. Buckwalter (eds.), A Companion to Experimental Philosophy, Blackwell.

Sytsma, Justin (2021). “Causation, Responsibility, and Typicality.” *Review of Philosophy and Psychology*, 12: 699–712.

Sytsma, Justin (2022). “Crossed Wires: Blaming Artifacts for Bad Outcomes.” *The Journal of Philosophy*, 119(9): 489–516.

Sytsma, Justin (n. d.). “Experiencers and the Ambiguity Objection.” http://philsci-archive.pitt.edu/15481/

Sytsma, Justin, Robert Bishop, and John Schwenkler (2022). “Has the side-effect effect been cancelled? (No, not yet.)” Synthese, 200: 395.

Sytsma, Justin, Roland Bluhm, Pascale Willemsen, and Kevin Reuter (2019). “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (eds.), *Methodological Advances in Experimental Philosophy*, pp. 209–238, Bloomsbury,

Sytsma, Justin and Wesley Buckwalter (2016). *A Companion to Experimental Philosophy*, Oxford: Blackwell.

Sytsma, Justin and Eugen Fischer (forthcoming). “‘Experience’, Ordinary and Philosophical: A Corpus Study.” *Synthese*.

Sytsma, Justin and Jonathan Livengood (2015). *The Theory and Practice of Experimental Philosophy*. Broadview.

Sytsma, Justin, Jonathan Livengood, and David Rose (2012). “Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions.” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43: 814–820.

Sytsma, Jonathan, Jonathan Livengood, Ryoji Sato, and Mineki Oguchi (2015). “Reference in the Land of the Rising Sun: A Cross-cultural Study on the Reference of Proper Names.” *Review of Philosophy and Psychology*, 6(2): 213–230.

Sytsma, Justin and Edouard Machery (2009). “How to Study Folk Intuitions about Phenomenal Consciousness,” *Philosophical Psychology*, 22: 21–35.

Sytsma, Justin and Edouard Machery (2010). “Two Conceptions of Subjective Experience,” *Philosophical Studies*, 151(2): 299–327.

Sytsma, Justin and Edouard Machery (2012). “On the Relevance of Folk Intuitions: A Reply to Talbot,” *Consciousness and Cognition*, 21(2): 654–660.

Sytsma, Justin and Eyuphan Ozdemir (2019). “No Problem: Evidence that the Concept of Phenomenal Consciousness is Not Widespread.” *Journal of Consciousness Studies*, 26(9-10): 241–256.

Sytsma, Justin and Kevin Reuter (2017). “Experimental Philosophy of Pain.” *Journal of Indian Council of Philosophical Research*, 34(3): 611–628.

Sytsma, Justin and Melissa Snater (2023a). “Consciousness, Phenomenal Consciousness, and Free Will.” In P. Henne and S. Murray (eds.), *Advances in Experimental Philosophy of Action*, Bloomsbury.

Sytsma, Justin and Melissa Snater (2023b). “Consciousness, Phenomenal Consciousness, and Free Will: Supplemental Materials.” http://philsci-archive.pitt.edu/19556/

Talbot, Brian (2012). “The irrelevance of folk intuitions to the ‘hard problem’ of consciousness.” *Consciousness and Cognition*, 21(2): 644–650.

Tye, Michael (2021). “Qualia.” In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/qualia/

Ulatowski, Joseph, Dan Weijers, and Justin Sytsma (2020). “Corpus Methods in Philosophy.” *The Brains Blog*: https://philosophyofbrains.com/2020/12/15/cognitive-science-ofphilosophy-symposium-corpus-analysis.aspx