Adrian Ziółkowski

University of Warsaw

adrian.a.ziolkowski@uw.edu.pl

The Stability of Philosophical Intuitions: Failed Replications of Swain et al. (2008)[1]

DRAFT (please do not cite)

Forthcoming in *Episteme*

ABSTRACT

In their widely cited article, Swain et al. (2008) report data that, purportedly, demonstrates instability of folk epistemic intuitions regarding the famous Truetemp case authored by Keith Lehrer. What they found is a typical example of priming, where presenting one stimulus before presenting another stimulus affects the way the latter is perceived or evaluated. In their experiment, laypersons were less likely to attribute knowledge in the Truetemp case when they first read a scenario describing a clear case of knowledge, and more likely to ascribe knowledge when they first read a vignette describing a clear case of nonknowledge. We tried to replicate Swain et al. findings in three experiments: one devised in Polish, and the other two conducted in English. We found no priming effect for knowledge ratings regarding the Truetemp case – laypersons were similarly likely to attribute knowledge in all three investigated conditions (primed with a clear case of knowledge, primed with a clear case of nonknowledge, and not primed). These three failed

[1] I am grateful to Joshua Alexander and Jonathan Weinberg, two authors of the replicated study, for their valuable remarks and suggestions. I would like to thank my colleagues from KogniLab, experimental philosophy lab at the Institute of Philosophy, University of Warsaw, Mieszko Tałasiewicz and Bartosz Maćkiewicz for discussion and technical support. I would like to express my gratitude to an Anonymous Referee for helpful comments which led to important improvements to the manuscript. And finally, I would like to thank Nika Sidorowicz, a cognitive science undergraduate student at the University of Warsaw, who carried out one of the studies discussed in this paper (Replication 1) under my supervision. Although Nika Sidorowicz did not coauthor this paper, she actively participated in the research project, and for that reason the pronoun "we" is used instead of "I" throughout the rest of the paper.

replication attempts are not decisive as to whether the priming effect in question occurs, nevertheless, the collected data puts Swain et al. conclusions about instability of epistemic intuitions in jeopardy.

Introduction

Experimental philosophers, who conduct systematic empirical research on philosophical intuitions, regularly report data that seem to be at odds with philosophical consensus in many different areas of philosophical investigations. In philosophy of action, Joshua Knobe (2003) famously found that when laypersons judge whether a certain effect is brought about intentionally, their answers might depend on whether its consequences are good or bad, which is not predicted by most prominent accounts of intentionality. In philosophy of language, Machery et al. (2004) discovered that semantic intuitions expressed by subjects from China tend to diverge from philosophers' verdicts regarding the famous Gödel case authored by Saul Kripke – while most philosophers share intuitions consistent with Kripke's causal-historical account, the majority of Chinese respondents give answers in line with the descriptivist theory of proper names[2]. In epistemology, it turned out that laypersons are happy to attribute knowledge to protagonists of different Gettier-style scenarios, which many philosophers take to be clear cases of nonknowledge (Starmans, Friedman, 2012; Colaco et al. 2014; Ziółkowski, 2016). Data on such discrepancies between philosophers' and folk intuitions provide the crucial premise for the argumentation against the use of intuitions in philosophy put forward by the representatives of negative experimental philosophy, also referred to as "experimental restrictionism" (Nadelhoffer, Nahmias, 2007). Since certain intuitions are not universal, as experimental restrictionists argue, the consensus among philosophers should not be used as evidence in favor or against philosophical theses. Another argumentative strategy leading to a similar conclusion points at instability of intuitions, because, if intuitions are shaky, they are not reliable and should not be trusted.

In their *The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp*, Swain, Alexander, and Weinberg (2008) use the latter strategy to argue against the use of intuitions in epistemology. They present empirical data which suggests that laypersons' judgements

[2] Both the results reported by Knobe (2003) and Machery et al. (2004) were replicated a number of times.

regarding one of the crucial epistemological thought experiments called “Truetemp” (Lehrer, 1990) vary depending on whether, and what, other thought experiments are presented beforehand. In other words, they found that folk intuitions regarding the Truetemp case are susceptible to priming effects, which means they are unstable and unreliable.

In this paper, we investigate the robustness of the priming effect reported by Swain et al. (2008) by running three replication attempts aimed at finding similar results – one on Polish native speakers, and the other two on English native speakers. In the first section, we present the original results in detail and discuss the conclusions put forward by Swain et al. The second section provides a brief rationale for attempting at reproducing the original findings. In the two following sections, we describe the methodology of our three experiments and the results they yielded. In the final section, we sum up our findings and discuss their importance for the discussion about the alleged instability of epistemic intuitions. Contrary to the original experiment, our data indicate that the existence of the priming effect reported by Swain et al. is highly dubious, and therefore, their arguments against the reliability of epistemic intuitions might be based on a false premise.

1. The purported instability of epistemic intuitions: Swain et al. (2008) original findings

As mentioned above, Swain et al. (2008) study is a classic example of an experiment utilizing the priming paradigm. Two of the vignettes used in their survey that played the role of priming stimuli were not based on any famous thought experiments, since one was expected to be a clear case of knowledge, the other – an obvious case of nonknowledge. Besides the above-mentioned Lehrer’s (1990) Truetemp case and two vignettes functioning as primes, their study also investigated another important thought experiment from the field of epistemology – the Fake Barns case provided by Carl Ginet (Goldman, 1976).

Both the Truetemp and the Fake Barns cases were originally designed as purported counterexamples to certain analyses of the notion of knowledge. The scenario put forward by Carl Ginet (but introduced to the literature by Goldman) poses a problem for Goldman’s causal theory of knowledge, which replaces one of the necessary conditions for knowing included in the classic tripartite analysis (i.e. agent *A* knows that *p* iff *A* has a justified, true belief that *p*) with a causal condition. Instead of requiring the belief in question to be justified, Goldman argued that an agent knows that *p* only if their belief is caused by the fact that *p*. The Fake Barns case describes a situation in which an agent forms a certain true belief which is obviously caused by its truth (it is based on perception, a clearly causal mechanism), but, nevertheless, we would not attribute knowledge to the agent in question (or at least Goldman claims so)[3]. Lehrer's thought experiment was supposed to undermine reliabilism, according to which knowledge is a justified, true belief that resulted from a reliable (i.e. successful most of the time when used) cognitive mechanism. The scenario describes an agent who forms a true belief that is based on a highly reliable cognitive process, but – as Lehrer argues – intuitively, that belief does not seem to be an instance of knowledge. The version of the scenario used in Swain et al. (2008) study is presented below:

> **Truetemp**
>
> *One day Charles was knocked out by a falling rock; as a result his brain was "rewired" so that he is always right whenever he estimates the temperature where he is. Charles is unaware that his brain has been altered in this way. A few weeks later, this brain rewiring leads him to believe that it is 71 degrees in his room. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is 71 degrees.*

After being presented with that story, each subject was asked to indicate the extent they agree or disagree with the following statement: "Charles knows that it is 71 degrees in his room." by choosing the answer from a 5-point scale ranging from "strongly agree" to "strongly disagree" (with a midpoint labeled as "neutral").

According to some experimental philosophers (including most of experimental restrictionists), the two above-mentioned thought experiments can be used as evidence in epistemological debates about knowledge only if: (1) they universally elicit epistemic intuitions similar to those shared by their authors; (2) the intuitions evoked by these thought experiments are stable and insensitive to philosophically irrelevant factors. Swain et al. (2008) focus on (2), and they assume that susceptibility of epistemic intuitions to priming would be an instance of instability caused by philosophically unimportant features. The story that played the role of a positive prime in their experiment (the clear case of knowledge) is quoted below:

[3] We will not quote the vignette used in the study here, as Swain et al. (2008) did not discover any priming effect for the Fake Barns case, so it is of little interest for us.

Chemist

Karen is a distinguished professor of chemistry. This morning, she read an article in a leading scientific journal that mixing two common floor disinfectants, Cleano Plus and Washaway, will create a poisonous gas that is deadly to humans. In fact, the article is correct: mixing the two products does create a poisonous gas. At noon, Karen sees a janitor mixing Cleano Plus and Washaway and yells to him, "Get away! Mixing those two products creates a poisonous gas!"

Also in this case, the respondents expressed their level of agreement with the claim "Karen knows that mixing these two products creates a poisonous gas." on an analogous 5-point scale. The negative prime (a clear case of nonknowledge) used in the experiment runs as follows:

Coinflip

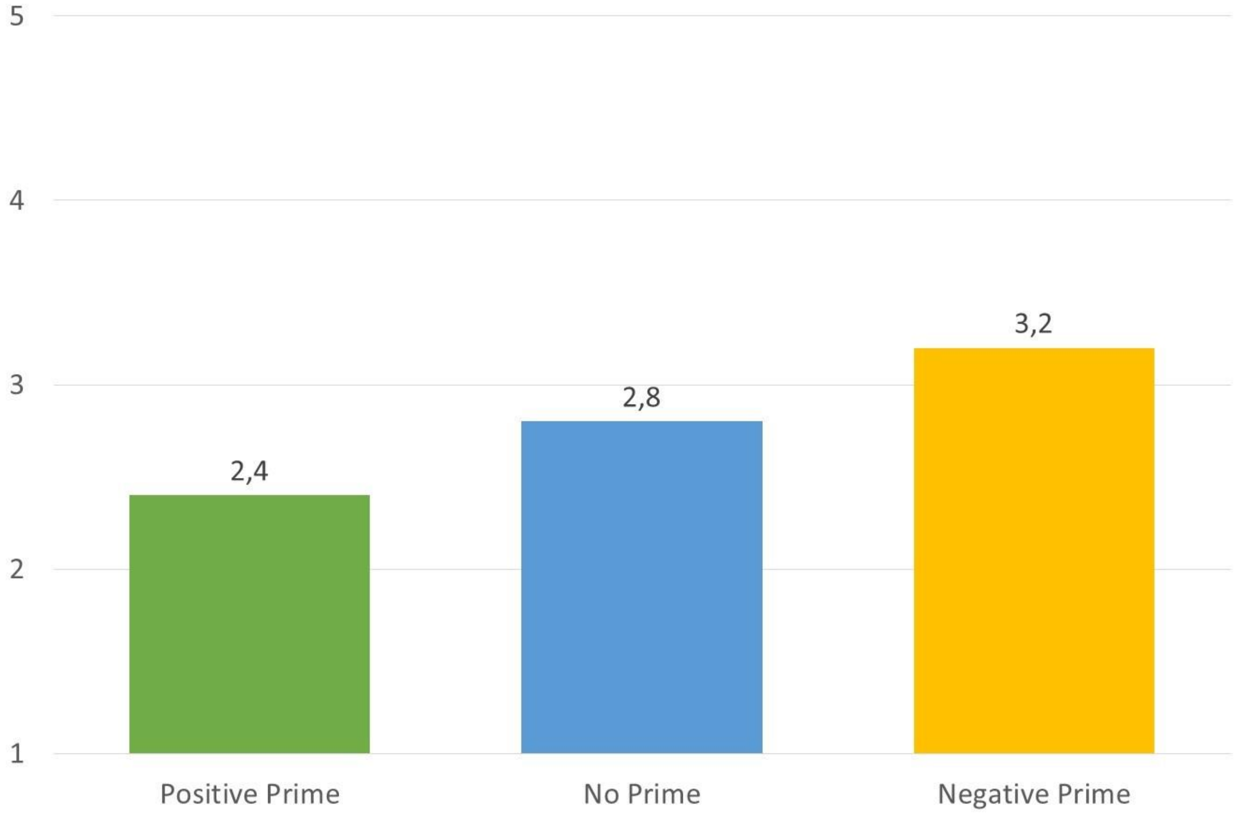
Dave likes to play a game with flipping a coin. He sometimes gets a "special feeling" that the next flip will come out heads. When he gets this "special feeling", he is right about half the time, and wrong about half the time. Just before the next flip, Dave gets that "special feeling", and the feeling leads him to believe that the coin will land heads. He flips the coin, and it does land heads.

Again, the researchers asked their subjects to indicate how much they agree with the statement: "Dave knew that the coin was going to land heads.".

Swain et al. (2008) expected that the primes will have a contrastive effect on the target thought experiment (Truetemp case): presenting the positive prime first will make laypersons to agree with the knowledge attribution in the target case to a lesser degree, while preceding the presentation of the target case with a negative prime should result with a higher tendency to agree with the crucial knowledge attribution.

They surveyed 220 subjects (undergraduate students) and divided them between eight experimental conditions, each defined by the order in which the four vignettes were presented (thus, every respondent assessed all four scenarios). We will not discuss their procedure in detail here, since some of the experimental conditions (orders of presentation) did not yield any important findings, but rather focus on the main discovery concerning the Truetemp case, which is supposed to support their claim about the instability of epistemic intuitions. Here, the data they collected confirmed their predictions – laypersons were less likely to ascribe knowledge to the protagonist of the Truetemp case when they first read a scenario describing a clear case of knowledge, more likely to attribute knowledge when they first read a vignette describing a clear case of lack of knowledge, with people that were not primed falling somewhere in the middle. This result is illustrated in the graph below, where "strongly disagree" is coded as "1", and "strongly agree" is coded as "5".

Graph 1. Average knowledge ratings regarding the Truetemp case in Swain et al. (2008) study.



Importantly, this main result comes from a notably smaller sample than the above-mentioned 220 subjects, since here the researchers only focused on 5 out of 8 experimental conditions – those in which the Truetemp case was not primed (three groups), the one where it was only preceded by the clear knowledge case (one group), and the one where it was only primed with the obvious case of nonknowledge (one group). Although Swain et. al. (2008) do not provide detailed information on the number of subjects assigned to each experimental condition, assuming that the distribution between groups was balanced, the sample size for their crucial discovery is around 137 subjects (with, on average, 27.5 respondents per group).

Unfortunately, many statistical details of this experiment remain unclear. First of all, it is worth noting that some statistics reported in the original article seem inconsistent, which, in turn, makes it highly likely that not all figures quoted above are accurate. The average knowledge ratings reported for three groups where the Truetemp case was presented first (no-prime condition) are, respectively, 2.64, 3.0, and 3.0, while the mean reported for these three groups lumped together is 2.8. This would be possible only if the numbers of subjects in these groups were not equal (if the 2.64 group was sufficiently bigger than at least one of the other two). If *n* in all three groups was balanced, the aggregated mean would rather be 2.9 (2.88, to be more precise), not 2.8. This leaves us with two options: at least one of the figures reported as mean knowledge rating for the Truetemp case is inaccurate, or there were notable differences in size between control groups investigated by Swain et al.

Moreover, the approach to statistical analysis employed by Swain et al. might seem non-standard (if not directly at odds with general methodological guidelines). First, they subjected their data to a one-way ANOVA analysis, which compared three experimental conditions: positive prime, negative prime, and control (no prime), where the data from three control groups were aggregated. Although the analysis yielded significant results, the *p*-value (.048) was dangerously close to the conventional threshold of statistical significance (.05). Swain et al. ran further analyses where they separately investigated the impact of each prime and assessed the differences between the primed and non-primed conditions. Even though they claim that such differences were found, the original findings were actually non-significant in this respect. Both pairwise comparisons between the primed groups and groups with no prime (when the three groups were aggregated) did not yield significant results (see Swain et al. 2008, footnotes 15. and 17.). Those pairwise comparisons which are reported as significant resulted from choosing one or another group out of those three where Truetemp was presented first. For comparison with 3.2 (negative prime), Swain and colleagues chose 2.64 (no prime); for comparison with 2.4 (positive prime), they chose 3,0 (no prime). Such a procedure seems arbitrary and might raise concerns about the reliability of resulting conclusions. We will now turn to discuss other studies that investigated the impact of priming on intuitions elicited by the Truetemp case, but in section 5. we will once more focus on the above-mentioned inconsistencies in data reporting and data analysis, and explain why they pose further problems.

A similar phenomenon to that observed by Swain et al. (2008) was found by Wright (2010). The participants of her experiment were presented with vignettes adopted directly from Swain et al. (2008) procedure. As in the case of the original study, Wright also recruited university undergraduates for her experiment. However, and most importantly, (i) the crucial knowledge question was formulated differently than in the original study (instead of measuring the level of agreement with a given knowledge attribution, she directly asked the subjects whether the protagonist of the scenario knew a certain proposition); (ii) the provided answer scale was different than in the original study (dichotomous "yes"-"no" scale vs. 5-point Likert scale). This resulted in further differences: instead of obtaining data on average agreement with the knowledge attribution (as in Swain et al.), here we obtain percentages of subjects ascribing knowledge; instead of using one-way ANOVA analysis and between-subject t-test comparisons (as in Swain et al.), here the data had to be analyzed using multiple chi-square tests. Wright observed a statistically significant order effect – subjects (total $N = 143$) were the most likely to attribute knowledge in the Truetemp case when it followed the Coinflip case (55%), slightly less likely when it was preceded by the Chemist case (40%), and the least likely to do so when it followed the Fake Barns case (26%). It is worth noting that the size of the difference between the two crucial conditions, Coinflip-Truetemp and Chemist-Truetemp, is considerably small[4]. Instead, Wright found a strong impact of the Fake Barns case on judgments regarding the Truetemp case, which was not observed in the original study by Swain et al. Moreover, the experiment conducted by Wright (2010) cannot be considered a proper replication of Swain et al. (2008) due to the above-mentioned methodological differences between these two studies.

[4] It remains unclear whether the difference in question (55% vs. 40%) passes the threshold for statistical significance, because Wright did not perform pairwise comparisons between conditions - only one chi-square test for all three conditions lumped together, where the significance might result from a very low ratio of knowledge attributions in the Fake Barns-Truetemp condition (26%) compared to the other two conditions.

Another experiment that investigated the impact of priming on intuitions regarding the Truetemp case was carried out by Weinberg et al. (2012)[5]. They used the same vignettes as in Swain et al. (2008) study and measured subjects' reactions with a 5-point Likert scale. Nevertheless, their experiment does not meet the requirements for a proper replication attempt, since it included only two experimental conditions, which is an important methodological divergence from the original study. Subjects were presented with all four stories adopted from the original experiment in the order that depended on the condition: it was either Chemist – Truetemp – Coinflip – Fake Barns (Condition 1) or Fake Barns – Coinflip – Truetemp – Chemist (Condition 2). Thus, while in Condition 1 the presentation of Truetemp was only preceded by Chemist, in Condition 2 it was preceded by both Fake Barns and Coinflip, which conflates the possible influence of these two cases on intuitions concerning the Truetemp scenario. Moreover, Weinberg et al. (2012) study did not include a condition in which the Truetemp case is presented first (without priming). For these reasons, regardless of the outcome, this experiment could neither strongly support nor undermine the findings discussed by Swain et al. (2008). However, Weinberg et al. (2012, p. 208) believe that the data they collected "nicely resembles the pattern of results reported in Swain et al.". The average knowledge rating on a 5-point Likert scale in Condition 1 was 2.53 (N = 64; SD = 1.07), while it seemed slightly higher in Condition 2, where it was 2.88 (N = 68; SD = 1.24). Yet, the conventional threshold for statistical significance ($p < 0.05$) was not exceeded – the observed p-value was only lower than 0.1[6], which led Weinberg et al. to conclude that the difference in question is "marginally significant" and that their subjects "are trending towards a pattern like that reported in Swain et al.". It is questionable whether this kind of difference still counts as statistical trend, but even if it does, the observed effect size is extremely small.

As already mentioned, basing on their results, Swain et al. (2008) conclude that epistemic intuitions are unstable and unreliable, which, in turn, means they have poor or no evidential value for philosophical argumentation. It seems to us that such a bold claim requires strong empirical grounding. It remains unclear whether the two other studies that aimed at finding a similar

[5] We are grateful to Jonathan Weinberg and Joshua Alexander for pointing that out.

[6] Here, unlike in the original study by Swain et al. (2008), Weinberg et al. (2012) used a non-parametric Mann-Whitney U test, not a between-subject t-test.

phenomenon we described above provide much (or any) support to that claim[7]. On the contrary – small sizes of the observed effects, together with considerably small numbers of participants assigned to experimental conditions, should rather raise our doubts and justify demands for further research on the issue. The next section provides more reasons why we should be interested in replicating Swain et al. (2008) findings.

2. Why replicate?

Experimental philosophy borrows its research techniques from social sciences, or – more specifically – experimental psychology and survey-based sociology. Thus, one should also expect that it might suffer from similar problems as these disciplines. One of the biggest issues for experimental psychology that is discussed in recent literature is poor replicability level of previous studies. For instance, Open Science Collaboration (2015) famously reported that only slightly over 1/3 of 100 replication attempts they conducted corroborated the original results, which suggests that more than a half of experimental effects reported in experimental psychology might be in fact non-existent. The same could be the case for experimental philosophy.

Some recent discoveries provide reason to believe that these expectations are not unexemplified. Probably the most spectacular case is the study by Weinberg and colleagues (2001) – one of the pioneering studiets of the experimental philosophy movement – which, purportedly, showed the existence of cross-cultural differences in epistemic intuitions. Unfortunately, further attempts at obtaining similar results did not confirm the original findings (see Nagel et al. 2013, Kim&Yuan 2015, Seyedsayamdost 2015a). Interestingly, it took more than a decade to make this discovery. During this time Weinberg et al. (2001) study was widely discussed – the participants of the discussion took Weinberg et al. findings as granted and spent a great deal of effort trying to somehow philosophically accommodate them (or simply explain them away). As it turns out, all these efforts were unnecessary. It is worth noting that many studies that fail to replicate belong to the negative branch of experimental philosophy – they point at systematic, undesired variation in intuitive judgments and question the evidential value of intuitions on these grounds. Another example of that sort is the study by Buckwalter&Stich (2013), which reported gender differences

[7] Joshua Alexander and Jonathan Weinberg (personal communication) believe that these experiments back up their conclusions about intuitional stability, but we dare to disagree.

in philosophical intuitions concerning many domains of philosophical inquiry. However, these findings were not corroborated by later studies (see Adelberg et al. 2015, Seyedsayamdost 2015b). Such results might raise worries about the level of replicability in experimental philosophy.

The recent big replication project within experimental philosophy, *The Xphi Replicability Project* coordinated by Cova and Strickland (2018), which provides data from 40 replication attempts, might bring some ease to these worries, since (depending on the chosen criteria) about three out of four replication attempts corroborated the initial findings. However, most importantly, the rate of successful replications depended on the kind of the effect reported in the original study: it was notably higher for studies which manipulated the content of stimuli presented to participants (90% of successful replications) compared to demographic effects (only 25%), such as the gender effect mentioned above, and contextual effects (also 25%)[8], such as framing (priming) effects similar to the phenomenon discussed by Swain et al. (2008). Moreover, contextual effects tend to yield low replication rates in psychology as well. Let us consider a famous case which played an important role in motivating further replication projects – the purported impact of the feeling of cleanliness one experiences on the severity of moral judgments one tends to formulate reported by Schnall and colleagues (2009). What they found is, for example, that when subjects are given the opportunity to wash their hands prior to assessing certain moral actions, they judge these actions as less wrong than participants who did not undergo the cleansing procedure. The study not only received much attention in academia, but also reached wider public, as it has been eagerly publicized in many influential magazines, such as *The Economist* or *The Huffington Post*. Unfortunately, later experiments which employed considerably larger samples did not confirm Schnall et al. findings (Johnson et al. 2014).

The results reported by Swain et al. (2008) and their bold conclusions were broadly discussed and frequently cited in the literature. However, so far no proper replication attempt of their original study has been carried out. Similar research projects – like those by Wright (2010) or Weinberg et al. (2012) described in the previous section – are far from being clearly in line with the original discovery. All this provides a rationale for running a replication (or multiple

[8] These two latter figures require cautious interpretation, since the number of replications concerning demographic and contextual effects included in *The Xphi Replicability Project* was rather small.

replications) of the main findings presented in Swain et al. (2008). In the following section, we describe the methods and procedures we adopted in our three replication attempts of *The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp*.

3. Methods and experimental procedure[9]

Since half of the experimental conditions in Swain et al. (2008) study were irrelevant to their main findings and the following argumentation they put forward, in our replication attempts we decided to only focus on the part of their experiment that yielded significant results. Thus, we did not include the Fake Barns case in our studies, but aimed directly at establishing whether the Truetemp case elicits different epistemic intuitions depending on whether (and how) it is primed.

Each replication attempt we carried out consisted of four experimental conditions: the Truetemp case was either preceded by a positive prime (Chemist-Truetemp), by a negative prime (Coinflip-Truetemp), or was presented first (Truetemp-Chemist and Truetemp-Coinflip)[10]. The subjects were randomly assigned to one of these four conditions and assessed two vignettes that were presented separately, without a possibility of going back in the survey and changing their previous judgments. The materials were directly adopted from the original Swain et al. (2008) study – the vignettes were exactly as presented in section 1. above, the crucial survey question was formulated in the same manner, and the participants expressed their intuitions on a 5-point Likert scale ranging from “strongly agree” (numbered as “5”) to “strongly disagree” (numbered as “1”) with the midpoint (“3”) labeled as “neutral”[11].

Each replication attempt was conducted as an online survey designed using an open-source software called LimeSurvey (www.limesurvey.org). In Replication 1., which was conducted in

[9] Customarily, if more than one study is presented, the methodological descriptions are provided separately for each experiment. However, since the methods employed in our three replications are (mostly) identical, we will summarize them in one section to avoid unnecessary repetitions.

[10] We decided to create two separate conditions in which the Truetemp case was presented first, in order to check whether that scenario might also work as a priming stimulus and affect how the other two vignettes are assessed.

[11] In Replication 1., which was conducted on Polish native speakers, we used Polish translations of the survey materials. These translations can be found in the Appendix.

Polish, we used a mixed sampling method – most of the respondents were university undergraduates who were provided the survey link by their lecturer, the rest were recruited using snowball sampling via social media. The participants of Replication 1. were volunteers and did not receive any pay for filling in the survey. In Replications 2. and 3. (both conducted in English) we employed the services of professional, Internet-based respondent panels: www.clickworker.com and www.prolific.ac respectively. In the case of these two replication attempts the participants received some small financial compensation for taking the survey.

Besides collecting data on crucial epistemic intuitions, we also included some demographic questions that were presented at the beginning of the survey. We asked our participants to indicate their gender, age and education. Moreover, two screening questions were shown in this part of the survey: subjects were asked whether they are English native speakers (or Polish native speakers in Replication 1.) and whether they studied philosophy before (and, if yes, whether they received a degree in philosophy). Respondents who admitted not being native speakers of the survey language or those who claimed to have a degree in philosophy (Bachelor's, Master's or PhD) were excluded from further analysis[12].

After filling in the demographic section, the participants were subsequently presented with two vignettes and answered questions regarding these vignettes. In Replications 2. and 3., where we recruited "professional" participants via respondent panels, the question regarding the level of agreement with the crucial knowledge attribution was accompanied by one comprehension question, which was devised in order to control whether the participants read the scenario carefully enough and understood its important aspects. In each case, the question had a simple true-false format and concerned some obvious factual features mentioned in the vignette. The comprehension check for the Truetemp case required the participants to recognize whether the claim "It is 71 degrees in Charles's room" is true or false. Similarly, for the Coinflip scenario, the claim that had to be evaluated was "The coin landed heads", and for the Chemist scenario it was "Mixing Cleano

[12] We believe that subjects who received a degree in philosophy most likely had contact with the Truetemp case. Since we are interested in untutored and untrained intuitions, we think that their answers should not be taken into consideration.

Plus and Washaway creates a poisonous gas". The respondents who failed to answer at least one of the comprehension questions correctly were excluded from further statistical analysis[13].

We decided to first run a replication study on Polish speakers, because we assumed that the data reported by Swain et al. (2008) proves the existence of the priming effect for epistemic intuitions among English speakers and we wanted to corroborate their findings by discovering a similar phenomenon in Polish. A failure in obtaining that result led us to conduct two additional replication attempts in English. The data collected in our three experiments is described in detail in the following sections.

4. The Data

4.1. Replication 1.[14]

248 respondents participated in this study. The answers from participants who did not complete the survey, reported having a degree in philosophy, or admitted not being a native Polish speaker were not included in further analysis, which gives a final sample size of 212. Out of those 212, 56 respondents were assigned to the Coinflip-Truetemp condition, 52 to the Chemist-Truetemp, 50 to the Truetemp-Coinflip, and 54 to the Truetemp-Chemist (which gives a total of 104 subjects who evaluated the Truetemp case without priming). 75% of the participants were female. Their age ranged from 18 to 55 years, but most of them were undergraduate students of management at the University of Warsaw, so the average age was considerably low (*M* = 24.44; *SD* = 5.47).

Firstly, it is worth noting that our participants reacted to clear cases of knowledge (Chemist) and nonknowledge (Coinflip) as predicted. They were happy to agree with the knowledge attribution regarding the Chemist case, regardless of the order of presentation – Chemist-Truetemp: *M* = 3.98, *SD* = 0.78; Truetemp-Chemist: *M* = 4.06, *SD* = 0.79. Also, the average agreement in the Coinflip case was clearly below the midpoint of the scale, without any differences between conditions – Coinflip-Truetemp: *M* = 2.09, *SD* = 1.01; Truetemp-Coinflip: *M* = 2.08; *SD* = 1.05.

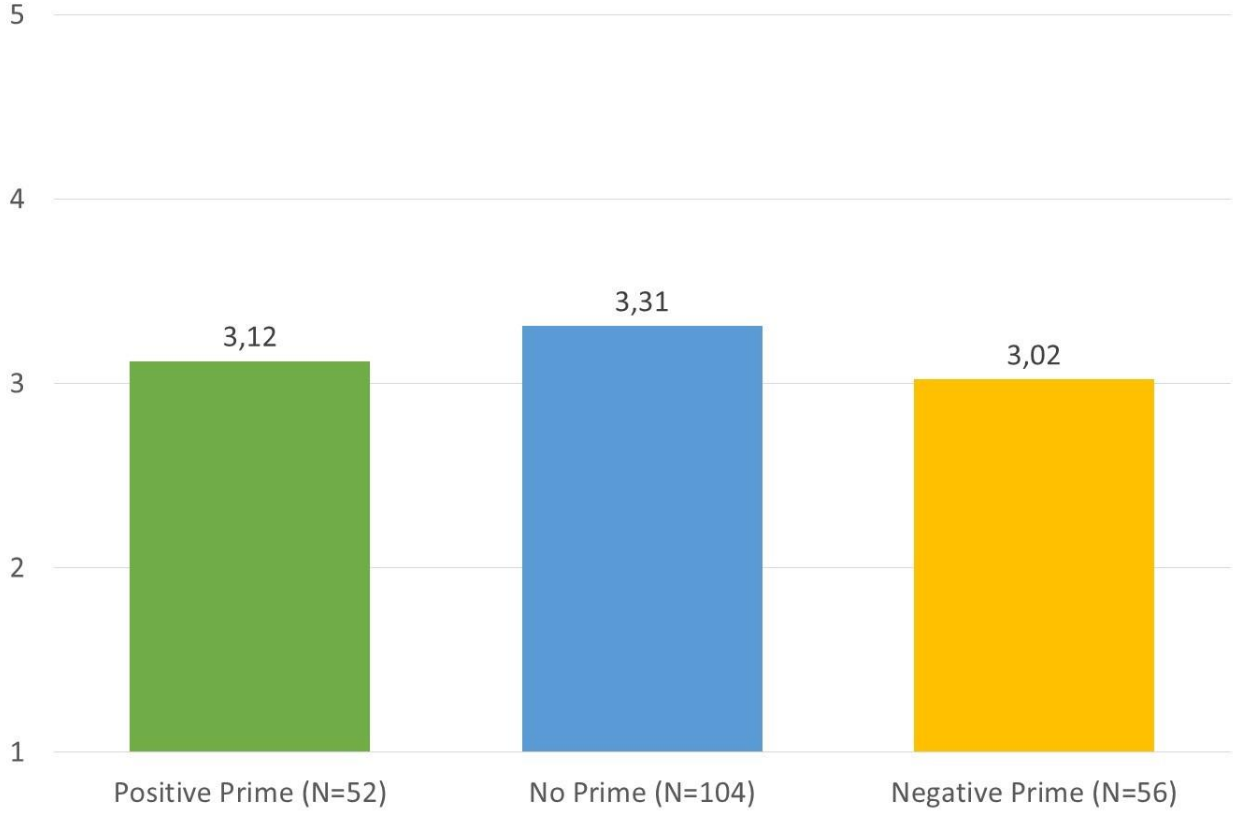
[13] It is worth noting that Swain et al. (2008) did not use a similar screening procedure. This issue will be discussed in detail later.

[14] This experiment was carried out by Nika Sidorowicz, a cognitive science undergraduate student, as a part of her BA thesis prepared under my supervision.

In order to analyze the data regarding the Truetemp case, similarly to Swain et al., we used a one-way ANOVA model, and between-subject t-tests to compare groups pairwise. The statistical analysis did not reveal any influence of the order of presentation on the level of agreement with the knowledge attribution for the Truetemp case[15]. Subjects were similarly likely to agree with the knowledge attribution when they first assessed a clear case of knowledge (Chemist-Truetemp, *M* = 3.12; *SD* = 1.13) and when they first read a description of a clear case of nonknowledge (Coinflip-Truetemp, *M* = 3.02; *SD* = 1.0). Moreover, both groups that included the priming procedure did not differ significantly from the condition in which the Truetemp case was evaluated first (two groups aggregated together: *M* = 3.31; *SD* = 1.1). In each condition, participants' answers clustered around the midpoint of the scale, but with a notable level of variance. A comparable number of participants tended to slightly disagree or slightly agree with the crucial knowledge attribution, with few choosing the extreme points of the scale across all four versions of the survey. The results are illustrated in the graph below.

[15] One-way ANOVA comparison (between four experimental conditions): $F(3,208) = 1.02$; $p = .384$; $\eta^2 = .014$. Between-subject t-test comparisons: negative prime-positive prime - $t(106) = -0.48$; $p = .635$; $g = -0.09$; positive prime-no prime - $t(154) = -1.02$; $p = .309$; $g = -0.17$; negative prime-no prime - $t(158) = -1.64$; $p = .103$; $g = -0.27$.

Graph 2. Average knowledge ratings regarding the Truetemp case in Replication 1.



This first unsuccessful attempt at reproducing the results reported by Swain et al. (2008) led us to conduct further studies in English.

4.2. Replication 2.

For the purpose of the first replication study in English, 222 participants were recruited via www.clickworker.com. 28 subjects were excluded due to the fact that they either admitted not being English native speakers, failed to provide a correct answer to at least one comprehension question, or had a degree in philosophy. The final sample size was 194, with 63.4% of female participants. The average age was 36.81 (*SD* = 12.63). 50 subjects read the Coinflip-Truetemp survey version, 49 read the Chemist-Truetemp version, while 48 and 47 read the Truetemp-Coinflip and the Truetemp-Chemist versions respectively.

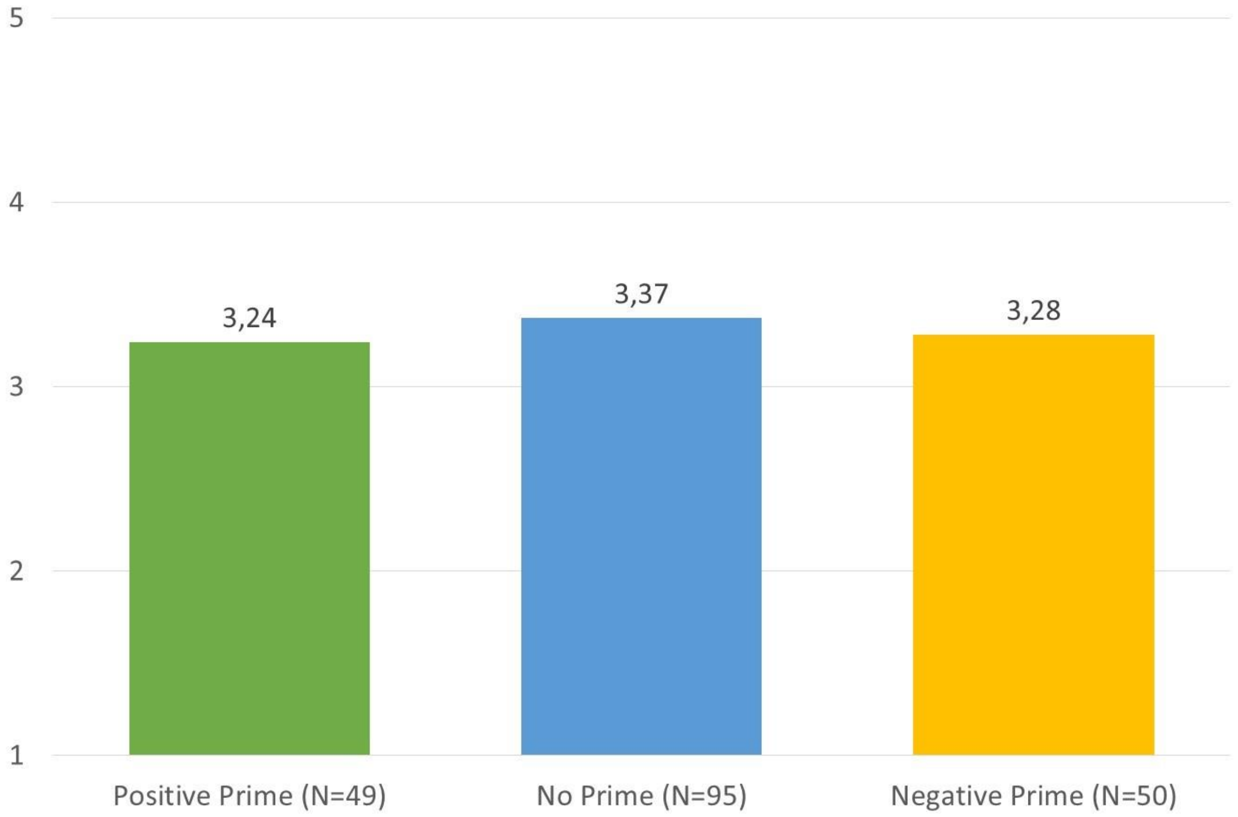
Our expectations regarding the Chemist and Coinflip cases were confirmed – while subjects tended to agree with the knowledge attribution in the former (Chemist-Truetemp: *M* =

4.47, *SD* = 0.89; Truetemp-Chemist: *M* = 4.23, *SD* = 0.96), they were likely to disagree with it in the latter (Coinflip-Truetemp: *M* = 2.08, *SD* = 1.07; Truetemp-Coinflip: *M* = 2.27, *SD* = 1.28), and the order of presentation had no impact on participants' reactions to these two vignettes.

Similarly as in the case of Replication 1., we subjected the data regarding the level of agreement with the knowledge attribution in the Truetemp case to one-way ANOVA analysis and t-test pairwise comparisons. The results were also similar to our first replication attempt – no statistically significant differences between experimental conditions were found[16]. Again, the average knowledge ratings were close to the midpoint of the scale, with a similar level of dispersity as in the case of Replication 1.: Chemist-Truetemp, *M* = 3.24; *SD* = 1.42; Coinflip-Truetemp, *M* = 3.28; *SD* = 1.18; Truetemp-Chemist and Truetemp-Coinflip aggregated together, *M* = 3.37; *SD* = 1.24. The results are summed up in the graph below.

[16] One-way ANOVA comparison (between four experimental conditions): $F(3,190) = 0.46$; $p = .708$; $\eta^2 = .007$. Between-subject t-test comparisons: negative prime-positive prime - $t(97) = 0.13$; $p = .894$; $g = 0.03$; positive prime-no prime - $t(142) = -0.54$; $p = .591$; $g = -0.09$; negative prime-no prime - $t(143) = -0.42$; $p = .678$; $g = -0.07$.

Graph 3. Average knowledge ratings regarding the Truetemp case in Replication 2.



The second replication attempt did not confirm the existence of the priming effect for epistemic intuitions regarding the Truetemp case.

4.3. Replication 3.

260 respondents recruited via www.prolific.ac participated in the third replication attempt (and second replication in English). The final sample size, however, was 231 subjects, as 29 participants either admitted not being native English speakers, had a degree in philosophy, or provided an incorrect answer to at least one comprehension question. 55% of those 231 subjects were female. The age of the respondents varied from 18 to 73 years (*M* = 33.8; *SD* = 12.04). The number of participants in each experimental condition was as follows: 62 – Coinflip-Truetemp; 57 – Chemist-Truetemp; 55 – Truetemp-Coinflip; 57 – Truetemp-Chemist.

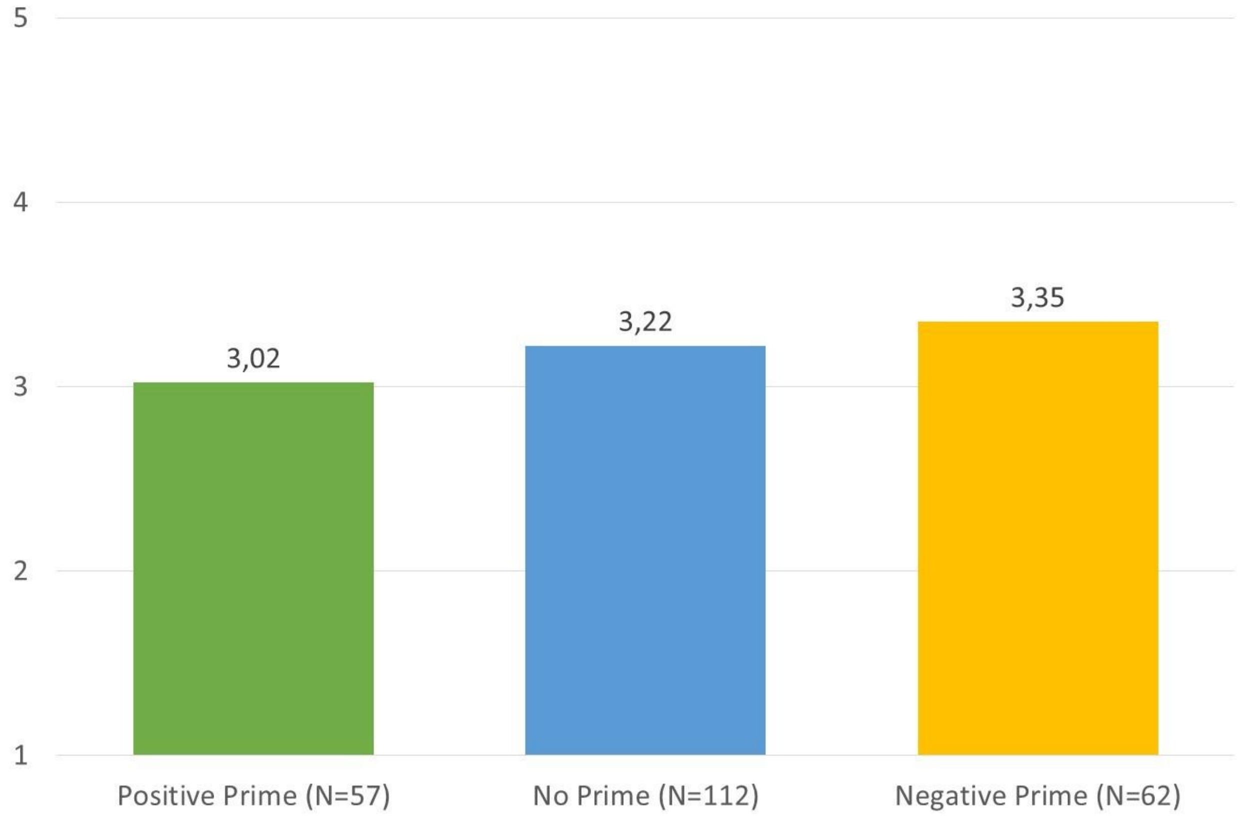
Similarly as in our previous replication attempts, the clear cases of knowledge and nonknowledge elicited intuitions that confirmed our expectations – the level of agreement for the Chemist case was very high (Chemist-Truetemp: *M* = 4.42, *SD* = 0.65; Truetemp-Chemist: *M* =

4.42, *SD* = 0.86), and considerably low for the Coinflip case (Coinflip-Truetemp: *M* = 1.97, *SD* = 1.07; Truetemp-Coinflip: *M* = 2.02, *SD* = 1.06). In accordance with the data collected in Replication 1. and 2., the intuitions regarding these two scenarios did not depend on the order of presentation.

Once again, in order to analyze the impact of the order of presentation on subjects' willingness to agree with the knowledge attribution in the Truetemp case, we used a one-way ANOVA model and between-subject t-tests. Although in the case of our third replication the mean knowledge ratings seem to fit the pattern predicted by Swain et al. (2008) – *i.e.* the condition with a negative prime yielded the highest average (*M* = 3.35; *SD* = 1.16), the condition with a positive prime brought the lowest (*M* = 3.02; *SD* = 1.25), with the average for conditions without priming falling somewhere in between (*M* = 3.22; *SD* = 1.19) – none of these small differences passes the threshold for statistical significance[17]. Both the ANOVA analysis and pairwise t-test comparisons return *p* values notably higher than 0.05. Therefore, one cannot conclude that the data confirms the existence of the priming effect for epistemic intuitions, since these small differences we observed in the sample might as well be due to chance.

[17] One-way ANOVA comparison (between four experimental conditions): $F(3,227) = 1.24$; $p = .296$; $\eta^2 = .016$. Between-subject t-test comparisons: negative prime-positive prime - $t(117) = 1.53$; $p = .129$; $g = 0.28$; positive prime-no prime - $t(167) = -1.05$; $p = .298$; $g = -0.17$; negative prime-no prime - $t(172) = 0.7$; $p = .482$; $g = 0.11$.

Graph 4. Average knowledge ratings regarding the Truetemp case in Replication 3.



To sum up, the results obtained in Replication 3. are not different from the findings of the two replication attempts described above in any important way. Here, also, the average level of subjects' agreement with the knowledge attribution in the Truetemp case was close to the midpoint of the scale regardless of the experimental condition (with a notable level of diversity in participants' judgments). Priming did not affect folk epistemic intuitions.

5. Conclusions and possible objections

The data collected in our three replication attempts are highly consistent and none of them confirmed the existence of the priming effect for epistemic intuitions regarding the Truetemp case. Contrary to what Swain et al. (2008) found in their original study, laypeople attribute knowledge to the protagonist of the Truetemp scenario at a similar degree, regardless of whether they first evaluate a clear case of knowledge, a clear case of nonknowledge, or analyze the Truetemp case first. To be more precise – folk judgments about that thought experiment are highly disperse and cluster around the midpoint of the provided scale, which may indicate uncertainty. Nevertheless, our data does not provide support to the claim that lay epistemic intuitions are unstable and can be easily changed by adopting a priming procedure. Therefore, it seems that Swain et al. (2008) argumentation against the use of epistemic intuitions in philosophy due to their alleged unreliability might rest on a false premise, as there is very little (if any) evidence that epistemic intuitions are unreliable in the way discussed by Swain et al. However, before stating the final conclusion, we will consider some possible objections to our studies.

Objection 1: Replication 1. is not a proper replication attempt

Replication attempts should be maximally similar to the original experiments that they aim to replicate. Since the original study by Swain et al. (2008) was conducted in English and Replication 1. described above was not, the latter should not be considered a legitimate replication attempt. The fact that Replication 1. did not corroborate the original results might suggest the existence of cross-cultural (or cross-linguistic) differences in epistemic intuitions between English and Polish speakers, rather than being a nonreplication of Swain et al. (2008). It might be that there is a priming effect for epistemic intuitions in English (the one discovered by Swain et al.), but an analogous effect in Polish does not exist.

Moreover, the research materials (vignettes and questions) were translated by the experimenters who conducted Replication 1., which leaves the door open to possible biases – for example, if the researchers expected not to replicate the original findings, they might have unwittingly formulate the translations in such a way that would favor a negative result.

Reply to Objection 1:

It is true that Replication 1. alone can be either interpreted as a nonreplication of Swain et al. (2008) or a study indicating cross-cultural (cross-linguistic) differences in epistemic intuitions between English and Polish speakers. However, the data collected in Replications 2. and 3. favors the former interpretation over the latter – since further experiments on English speakers yielded similar results to Replication 1. conducted in Polish, and also did not corroborate Swain et al. (2008) findings, probably there are no cross-cultural (cross-linguistic) differences here. Replication 1. together with Replications 2. and 3. suggest that the predicted priming effect for epistemic intuitions regarding the Truetemp scenario exists neither for Polish nor English speakers.

As for the second remark suggesting possible biases arising from the translation procedure – it could be troublesome if we indeed expected to obtain a nonreplication. But, on the contrary, at that point we actually hoped to find the effect reported by Swain et al. (2008) in Polish and support their English results. Thus, we were not biased against their hypotheses.

Objection 2: Replications 2. and 3. differ methodologically from the original study

Although both Replications 2. and 3. were conducted in English and used the survey materials (vignettes, questions) borrowed from the original study by Swain et al. (2008), they also adopt certain methodological solutions that were not present in the initial experiment: (a) the original study was a paper-pen survey, while both Replications 2. and 3. were conducted in the Internet; (b) participants for both Replications 2. and 3. were recruited online via professional respondent panels (www.clickworker.com, www.prolific.ac), whereas in the initial study, all subjects were university undergraduates; (c) both Replications 2. and 3. included a screening procedure (which excluded non-native English speakers, subjects who had a degree in philosophy, and those who failed to answer comprehension questions) that was not incorporated in the replicated experiment. Due to these discrepancies, the data collected in replication attempts 2. and 3. might not be adequate for a comparison with the original findings.

Reply to Objection 2:

It does not seem that the above-mentioned differences between Swain et al. (2008) study and Replications 2. and 3. importantly diminish the evidential value of the latter as far as the issue of stability of epistemic intuitions is concerned. In particular, if the priming effect reported by Swain et al. (2008) was robust[18], one should be able to observe it regardless of whether one uses a paper-pen or online-based survey. Likewise, if it appears in a sample comprised of university undergraduates, it should also emerge in a (more diverse) sample consisting of Internet users. The additional screening procedure mentioned above was needed due to the chosen sampling method – some Internet users who participate in surveys in exchange for financial compensation tend to do it mechanically and inattentively, filtering their answers out by a properly designed

[18] By this we mean statistical robustness, which involves replicability.

comprehension check might increase data quality[19]. The control questions included in Replication 2. and 3. are simple, obvious, and uncontroversial, while the rejection level that arose from their use was considerably low (around 10%). And most importantly, not including this screening procedure in data analysis does not change the outcome significantly – there is still no priming effect, if all, unfiltered data is analyzed.

Objection 3: the absence of evidence in not the evidence of absence

The results obtained in all three presented replication attempts are the so-called null results. In null hypothesis testing, attaining a statistically significant difference justifies rejecting the null hypothesis (which claims that there are no such differences), but, and most importantly, finding no statistically significant differences is not evidence in support of the null hypothesis (it just does not allow for its rejection). Therefore, even if the three replication attempts yielded no significant results and they do not support the claim about the instability of epistemic intuitions, one cannot conclude that they support the opposite claim about their stability. This would be an instance of a fallacy known as appeal to ignorance.

Reply to Objection 3 – meta-analysis:

This objection is right and it is the exact reason why we do not claim our findings ultimately refute Swain et al. (2008) hypothesis about the instability of epistemic intuitions. It might be that their data reveals a genuine priming effect that our three replications simply failed to discover (for example, due to insufficient sensitivity of the used measures or even simply by chance). In fact,

[19] Jonathan Weinberg and Joshua Alexander (personal communication) express worries that people who participate in online surveys might be poorly engaged in the process and, in general, less attentive than university undergraduates. If that was the case, it could explain the absence of the priming effect in our replications. We agree that these worries are reasonable, but we believe that thanks to our comprehension check we control that unwanted factor. Also, since our respondents reacted to the Chemist and Coinflip cases as predicted (they ascribed knowledge in the former, but denied it in the latter), there is another reason to believe that they were engaged in the survey and attentive to the details presented in the vignettes.

the size of the priming effect reported by Swain et al. is rather small[20], which makes it "harder" to detect (assuming it exists) for purely statistical reasons. In other words, replications aiming at finding effects small in size are more likely to yield false negative results (Type II error) due to lower statistical power. Nevertheless, three unsuccessful attempts at obtaining results that would support the hypothesis about the instability of epistemic intuitions versus one study that confirms it (and two other experiments that are "trending" towards confirmation) make that hypothesis highly dubious[21].

In order to substantiate this claim, we decided to run some additional analyses which could lead to some "positive" conclusions about the data, besides reporting null results[22]. In such circumstances it would be natural to employ statistical power analysis and directly compare the results of the original study with the data collected in our three replication attempts. This approach would allow to evaluate (given the original effect size and sample sizes of our replication attempts) whether our studies were powerful enough to detect the effect in question or maybe the sample sizes were insufficient to ignore the probability of Type II error (*i.e.* false negative results). Unfortunately, this cannot be done due to incomplete data about the original findings. Swain et al. (2008) only report p-values (no test statistics), mean results for each experimental condition, and total sample size. They do not provide information about standard deviations and number of subjects assigned to each experimental condition, which is essential for calculating effect sizes[23].

[20] This observation is only an approximation, since the exact effect sizes cannot be computed for Swain et al. study due to insufficient data reporting. We discuss this issue in detail below.

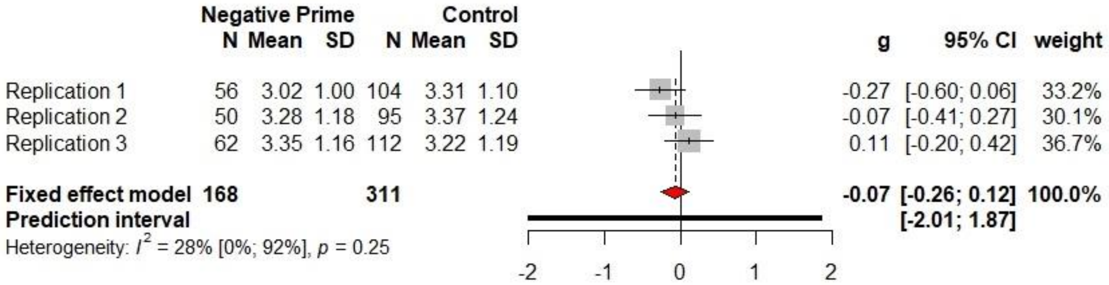
[21] Jonathan Weinberg and Joshua Alexander (personal communication) offer another explanation: since our replications were carried out ten years later than the original study, different results may track changes of epistemic intuitions over time. Since the Western world goes thru a period of "fake news", it might also affect the patterns of knowledge attribution. However, it is rather unlikely that intuitions concerning the Truetemp case are the only ones affected (note that our respondents were still strongly attributing knowledge in the Chemist case and denying it in the Coinflip case). Thus, this explanation does not seem very plausible to us.

[22] We are grateful to an Anonymous Referee who suggested this approach.

[23] The Authors of the original study were not able to provide these data on request, as they no longer have access to the original dataset. They attempted to recalculate standard deviations for

Moreover, some statistics reported in the original paper may seem inconsistent (see section 1. above for details) and it is unclear what exact figures – especially in the no-prime condition – should be taken into consideration for further analyses. For this reason, our additional analyses focus on data collected in our three replication attempts. We employ a meta-analytic approach (fixed effect model), which consists in aggregating the data obtained in our three studies and results in a combined effect size estimation. We present two such analyses regarding two main pairwise comparisons of interest: negative prime condition vs. control (no prime) condition and positive prime condition vs. control (no prime) condition. The graph below summarizes the results of our three replications as far as the former is concerned.

Graph 5. Meta-analysis of three replication attempts – negative prime vs. control (no prime).



	Negative Prime			Control					
	N	Mean	SD	N	Mean	SD	g	95% CI	weight
Replication 1	56	3.02	1.00	104	3.31	1.10	-0.27	[-0.60; 0.06]	33.2%
Replication 2	50	3.28	1.18	95	3.37	1.24	-0.07	[-0.41; 0.27]	30.1%
Replication 3	62	3.35	1.16	112	3.22	1.19	0.11	[-0.20; 0.42]	36.7%
Fixed effect model	**168**			**311**			**-0.07**	**[-0.26; 0.12]**	**100.0%**
Prediction interval								**[-2.01; 1.87]**	

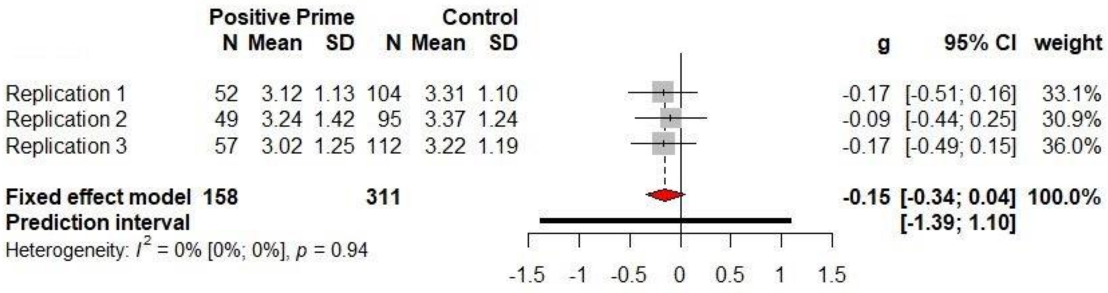
Heterogeneity: I^2 = 28% [0%; 92%], p = 0.25

As can be easily noted, the meta-analytic effect size (*g* = -0.07) falls below the conventional threshold for a small effect (Cohen, 1988), and the confidence interval of the effect size includes zero, which means our replications indicate that the negative prime effect might be non-existent. The situation is analogous for the positive prime effect, as illustrated in the following graph.

---

each experimental group in their study given average knowledge ratings, numbers of observations, and the upper bounds of CIs that can be found in one of their graphs (Swain et al., 2008: 143). However, the outcome figures are distorted in two ways, since the Authors of the original study: (i) do not know the exact distribution of subjects between conditions, and for the sake of calculations they assume it was equal and that there were 27.5 subjects in each experimental group; (ii) they establish the upper bounds of CI's by "visually deriving" them from the graph and were not able to estimate them past one decimal point.

Graph 6. Meta-analysis of three replication attempts – positive prime vs. control (no prime).

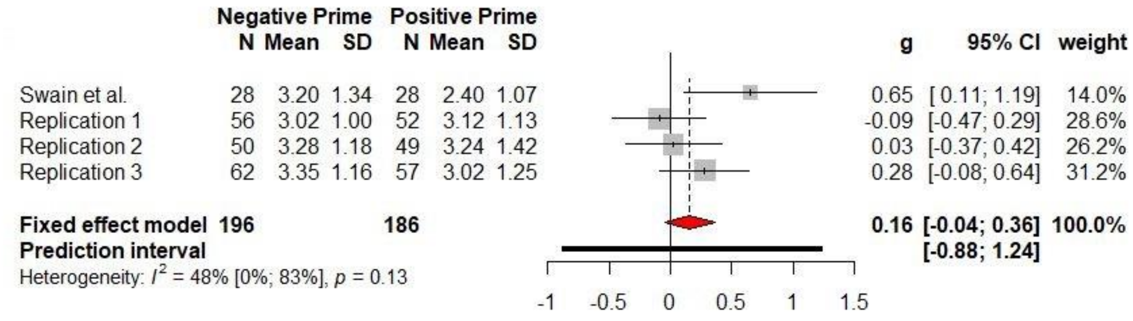


Again, the meta-analytic effect size is less than small (*g* = -0.15) and the confidence interval includes zero. It is worth noting that both these additional analyses involve samples much greater in size than the one investigated by Swain et al. (2008).

Driven by curiosity, we also decided to run a simulation of a meta-analysis that could incorporate some of the original results (partly recreated by calculations performed by the Authors of the original study, see footnote 23. above). Unfortunately, we could not conduct an analysis including the control group, since the mean figures reported by Swain et al. for the control condition remain unclear (again, see section 1. above for details). Instead, we focused on the two experimental conditions where participants were subjected to the priming manipulation. This is where we should expect to find a significant effect, as in the original study the difference between mean ratings in negative- and positive-prime conditions classifies as a medium size effect (*g* = 0.65) according to benchmarks proposed by Cohen (1988). But even in this case, although we include the (approximation of) original findings in the analysis, the meta-analytic effect size (*g* = 0.16) still falls below the conventional benchmark for a small effect. Moreover, the confidence interval of the meta-analytic effect size includes zero. The data is summarized in the following graph. However, once more we need to stress that the SD figures and the number of participants assigned to each experimental group in the original study is only approximated and for this reason we should treat this analysis with some reservations[24].

[24] Moreover, it needs to be noted that the level of heterogeneity across samples was considerably large (I^2 = 48%), which indicates that the variance in effect sizes was partly due to systematic

Graph 7. Meta-analysis of three replication attempts and original findings – positive prime vs. negative prime.



	Negative Prime			Positive Prime					
	N	Mean	SD	N	Mean	SD	g	95% CI	weight
Swain et al.	28	3.20	1.34	28	2.40	1.07	0.65	[0.11; 1.19]	14.0%
Replication 1	56	3.02	1.00	52	3.12	1.13	-0.09	[-0.47; 0.29]	28.6%
Replication 2	50	3.28	1.18	49	3.24	1.42	0.03	[-0.37; 0.42]	26.2%
Replication 3	62	3.35	1.16	57	3.02	1.25	0.28	[-0.08; 0.64]	31.2%
Fixed effect model	**196**			**186**			**0.16**	**[-0.04; 0.36]**	**100.0%**
Prediction interval								**[-0.88; 1.24]**	

Heterogeneity: $I^2 = 48\%$ [0%; 83%], $p = 0.13$

The results yielded by all additional analyses we conducted suggest that the priming effect in question either does not exist, or – if it does – is very small, which puts its theoretical importance in question. Obviously, in order to settle the issue of susceptibility of epistemic intuitions regarding the Truetemp case to priming, we would have to run a series of further replication attempts (preferably in different labs and with different teams of researchers) and conduct another meta-analysis on all the collected data. Nonetheless, our findings put Swain et al. (2008) conclusions about instability of epistemic intuitions in jeopardy and shift the burden of proof on them – if they want to argue that epistemic intuitions are unreliable due to instability, they first need to provide more evidence that epistemic intuitions really are unstable.

differences between datasets included in the analysis. For this reason, the fixed effect model might not be fully appropriate in this case and another statistical approach would be preferable. However, since the analysis in question is only a simulation based on an approximation of the actual Swain et al. (2008) data, we will not elaborate on the issue here.

Appendix: Polish translations of survey materials

Truetemp

Pewnego dnia Karol został uderzony w głowę przez spadającą skałę. W wyniku tego wypadku jego mózg został tak "przeprogramowany", że teraz zawsze dobrze ocenia temperaturę miejsca, w którym się znajduje. Karol nie zdaje sobie sprawy ze zmiany, jaka zaszła w jego mózgu. Kilka tygodni po wypadku ta zmiana w mózgu powoduje, że Karol jest przekonany, że w pokoju, w którym siedzi, jest 21,5 °C. Oprócz jego przekonań nie ma żadnych innych przesłanek, aby tak twierdzić. W rzeczywistości jest 21,5 °C.

Proszę wskazać, w jakim stopniu zgadzasz się lub nie zgadzasz z następującym stwierdzeniem: "Karol wie, że w pokoju jest 21,5 °C ."

Chemist

Karen jest wybitnym profesorem chemii. Dziś rano przeczytała artykuł w czasopiśmie naukowym, że zmieszanie dwóch zwykłych środków dezynfekujących, Cleano Plus i Washaway, skutkuje wytworzeniem trującego gazu, który jest śmiertelny dla ludzi. W rzeczywistości artykuł jest poprawny: zmieszanie tych dwóch produktów stworzy trujący gaz. W południe Karen widzi woźnego, który miesza Cleano Plus i Washaway'a, i krzyczy do niego: "Odsuń się! Mieszanie tych dwóch produktów tworzy trujący gaz! "

Proszę wskazać, w jakim stopniu zgadzasz się lub nie zgadzasz z następującym stwierdzeniem: "Karen wie, że zmieszanie tych dwóch produktów skutkuje wytworzeniem się trującego gazu."

Coinflip

Dawid lubi grać w rzut monetą. Czasami ma "specjalne przeczucie", że moneta spadnie orłem do góry. Kiedy Dawid ma to "specjalne przeczucie", ma rację w około 50% przypadków i myli się w około 50% przypadków. Tuż przed kolejnym rzutem Dawid miał to "specjalne przeczucie" i wierzył, że moneta spadnie orłem do góry. Rzucił monetą i wypadł orzeł.

Proszę wskazać, w jakim stopniu zgadzasz się lub nie zgadzasz z następującym stwierdzeniem: "Dawid wiedział, że moneta spadnie orłem do góry."

References

Adleberg T., Thompson M. and Nahmias E. (2015). 'Do men and women have different philosophical intuitions? Further data.' Philosophical Psychology 28 (5):615-641.

Buckwalter W. and Stich, S. (2013). 'Gender and Philosophical Intuition.' In Joshua Knobe & Shaun Nichols (eds.), *Experimental Philosophy*, Vol.2. Oxford University Press. pp. 307-346.

Cohen, J. 1988. *Statistical power analysis for behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collaboration, Open Science. 2015. 'Estimating the Reproducibility of Psychological Science' *Science*, 349 (6251).

Colaço, D., Buckwalter, W., Stich, S., and Machery, E. 2014. 'Epistemic Intuitions in Fake-Barn Thought Experiments' *Episteme*, 11: 199–212.

Cova, F., Strickland, B., Abatista, A. et al. 2018. 'Estimating the Reproducibility of Experimental Philosophy' *Review of Philosophy and Psychology*, https://doi.org/10.1007/s13164-018-0400-9

Goldman, A. I. 1976. 'Discrimination and Perceptual Knowledge' *Journal of Philosophy*, 73: 771–91.

Johnson, D. J., Cheung, F., and Donnellan, M. B. 2014. 'Does Cleanliness Influence Moral Judgments? A Direct Replication of Schnall, Benton, and Harvey (2008)' *Social Psychology*, 45 (3): 209–15.

Kim M., and Yuan Y. 2015. 'No Cross-Cultural Differences in the Gettier Car Case Intuition: A Replication study of Weinberg et al. 2001' *Episteme*, 12 (3), 355–361.

Knobe, J. 2003. ‘Intentional Action and Side Effects in Ordinary Language’ *Analysis*, 63 (3): 190–4.

Lehrer, K. 1990. *Theory of Knowledge*. Routledge.

Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004. ‘Semantics, Cross-cultural Style’ *Cognition*, 92: B1–B12.

Nadelhoffer T., and Nahmias E. 2007. ‘The Past and Future of Experimental Philosophy’ *Philosophical Explorations*, 10: 123–49.

Nagel, J., San Juan, V., and Mar, R. 2013. ‘Lay Denial of Knowledge for Justified True Beliefs’ *Cognition*, 129 (3), 652–661.

Schnall, S., Benton, J., and Harvey, S. 2009. ‘With a Clean Conscience’ *Psychological science*, 19 (12), 1219-22.

Seyedsayamdost H. 2015a. ‘On Normativity and Epistemic Intuitions: Failure of Replication’ *Episteme*, 12 (1), 95–116.

Seyedsayamdost, H. 2015b. ‘On gender and philosophical intuition. Failure of replication and other negative results’ *Philosophical Psychology*, 28(5), 642-673.

Swain, S., Alexander, J., and Weinberg, J. M. 2008. ‘The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp’ *Philosophy and Phenomenological Research*, 76 (1): 138–55.

Starmans, C., and Friedman, O. 2012. ‘The Folk Conception of Knowledge’ *Cognition*, 124 (3): 272-83.

Weinberg, J. M., Alexander, J., Gonnerman, C., and Reuter, S. 2012. ‘Restrictionism and Reflection: Challenge Deflected, or Simply Redirected?’ *The Monist*, 95 (2): 200-22.

Weinberg, J. M., Nichols, S., & Stich, S. 2001. Normativity and epistemic intuitions. *Philosophical Topics*, 29 (1-2), 429–460.

Wright, J. C. 2010. ‘On intuitional stability: The clear, the strong, and the paradigmatic’ *Cognition*, 115 (3): 491–503.

Ziółkowski, A. 2016. Folk Intuitions and the No-Luck-Thesis. *Episteme*, 13 (3), 343–358.