Filozofia Nauki (*The Philosophy of Science*)
ISSN 1230-6894 e-ISSN 2657-5868
2021, vol. 29(2) [114]: 157-188
DOI: 10.14394/filnau.2021.0013

ADRIAN ZIÓŁKOWSKI*

THE CONTEXT-SENSITIVITY OF COLOR ADJECTIVES AND FOLK INTUITIONS**

Abstract

In this paper, I report new empirical data on folk semantic intuitions concerning color adjectives in so-called context-shifting experiments. Contextualists present such experiments — that is, they describe different conversational contexts in which a given sentence is uttered — in order to argue that context can shape meaning and truth conditions to such a degree that competent speakers would give opposite truth evaluations of the same sentence in different contexts. The initial findings of Hansen and Chemla (2013) suggest that laypersons' semantic judgments are sensitive to context in the same way that is predicted by contextualists. In this paper, I focus on context-shifting experiments that involve color adjectives; also, I present experiments that are a partial replication and methodological extension of Hansen and Chemla's study. One aim of my study was to corroborate these authors' findings using a bigger sample (total N = 1128), but the main goal was to test the stability of results in different methodological variants of empirical adaptations of context-shifting experiments. This part of the study addresses the issues pointed out in my earlier paper (Ziółkowski 2017), where I argued that certain experimental settings (within-subjects) might bring data that is more favorable to contextualism than other settings (between-subjects). My study compares three different experimental settings: within-subjects (with randomized order of context presentation), between-subjects (where participants evaluating different contexts are distinct groups), and "contrastive design" (where both contexts are presented side by side on the same screen). My results are highly consistent across the methodological variants I employed, but while they show some of the effects expected by contextualists, it is disputable whether they bring strong support to contextualism with respect to color adjectives.

* Faculty of Philosophy, University of Warsaw, Krakowskie Przedmieście 3, 00-927 Warsaw, Poland, e-mail: adrian.a.ziolkowski@uw.edu.pl, ORCID: https://orcid.org/0000-0003-3151-0196.

** I would like to thank Nat Hansen, Tomasz Zyglewicz, and Joanna Odrowąż-Sypniewska for their comments on an early draft of this paper. I am also thankful to the participants of the 1st European Experimental Philosophy Conference (June 2021), where I presented a talk based on this project, for a helpful discussion. I am also grateful to the anonymous reviewers, whose comments led to substantial improvements in the manuscript. And lastly, I would like to express my gratitude to Bartosz Maćkiewicz, my colleague from KogniLab, the experimental philosophy lab based at the University of Warsaw, for technical support during data collection.

Keywords: contextualism, context-sensitivity, context-shifting experiments, color adjectives, folk semantic intuitions

In this paper, I report new empirical data on folk semantic intuitions concerning color adjectives in so-called context-shifting experiments. Contextualists present such experiments — that is, they describe different conversational contexts in which a given sentence is uttered — in order to argue that context can shape meaning and truth conditions to such a degree that we would give opposite truth evaluations to the same sentence in different contexts. Since context-shifting experiments can easily be adapted in a systematic empirical study using the methods of experimental philosophy, contextualists should expect that their claims would be reflected in laypersons' semantic judgments (hereafter, I will refer to these expectations as "contextualist predictions"). I present studies that empirically test contextualist predictions regarding the purported context-sensitivity of color adjectives. My experiments address the issues raised in an earlier study (Ziółkowski 2017), where I noticed that certain experimental settings (within-subjects) might bring data more favorable to contextualism than other settings (between-subjects). In order to establish whether the results of different empirical adaptations of context-shifting experiments are stable and consistent, I ran experiments that used the same materials but employed three different experimental settings: within-subjects (with randomized order of context presentation), between-subjects (where participants evaluating different contexts were distinct groups), and "contrastive design" (where both contexts were presented side by side on the same screen).

Section 1 briefly sketches the theoretical background of the empirical work I carried out. It discusses the main premises of contextualism and the empirically testable predictions that follow from these premises. It also presents some previous experimental findings that are relevant to my project, in particular the experiments carried out by Nat Hansen and Emmanuel Chemla (Hansen, Chemla 2013; cf. Ziółkowski 2017), which directly inspired my studies. Section 2 presents my research objectives and hypotheses. Briefly, the idea was to empirically test contextualist predictions concerning color adjectives in a variety of cases, and investigate whether different methodological variants of context-shifting experiments yield different results. Section 3 is a detailed description of Experiment 1. First, I present the methods and experimental procedure, then I report the results and statistical analyses, which is followed by discussion and evaluation of the research hypotheses. Section 4 is devoted to a detailed presentation of Experiment 2, its results, and the final conclusions that can be drawn from both experiments. My results are highly consistent across the methodological variants, but while they show some of the effects expected by contextualists, it is disputable whether they bring strong support to contextualist predictions regarding color adjectives.

1. THEORETICAL AND EMPIRICAL BACKGROUND OF THE STUDY

1.1. CONTEXTUALISM, CONTEXT-SENSITIVITY OF LINGUISTIC ITEMS, AND CONTEXT-SHIFTING EXPERIMENTS

Contextualists such as Keith DeRose (e.g., 1992, 1999), Charles Travis (e.g., 1997), or François Recanati (e.g., 2004, 2010) often make claims about the intuitive truth conditions of utterances and indicate the strong dependence of truth conditions on the pragmatic factors that are introduced via the context of utterance. Proponents of contextualism believe that the vast majority (if not all) of natural-language terms are context-sensitive; that is, the meaning they convey might differ radically from one context to another. This in turn means that, depending on the occasion on which the utterance in question is made, two utterances of one and the same sentence (provided it contains context-sensitive terms as constituents) might have radically different truth conditions. Contextualists' line of argumentation is often termed "context-shifting experiments," since it consists in presenting pairs of hypothetical cases describing different contexts of an utterance of a certain sentence (using experimental terminology, one might say that context is an independent variable here that is subjected to manipulation by the experimenter). According to contextualists, these contexts affect the truth conditions of an utterance *U* to such a degree that a competent speaker would judge that *U* is true in context C1 and false (or at least not true) in context C2, even though important factual details remain constant across contexts (which would mean that differences in truth values are due to differences in meaning or content). I will illustrate this kind of argumentation with an example that involves color adjectives and shows their purported context-sensitivity, which is the main focus of the empirical studies I will present later. The scenario, called Leaves, was adapted for the purposes of experimental studies by Hansen and Chemla (2013); it was inspired by the hypothetical cases discussed by Travis (1997). Consider the following story:

> LEAVES — ACCEPTANCE CONTEXT
>
> *Pia has a Japanese maple tree in her backyard that has russet (reddish brown) leaves. She paints the leaves of the tree green. A friend of Pia's who is making decorations for a play asks if Pia has any green leaves she can use in her stage set. "The leaves on my tree are green," Pia says.*

According to Travis, we should expect competent speakers to accept as true the utterance "The leaves on my tree are green" in this conversational context. To make things simple, I will use Gricean terminology to explain why and to show the important differences between this and the following context.[1] Paul Grice (1975) claims that every conversation is a form of collaboration between interlocutors that is governed by certain implicit rules (tacitly assumed by the participants of the linguistic interaction) and whose aim is to reach a certain goal that the interlocutors agree on (in some cases the goals differ, as in the case of deception, but these occur less often). In the case of the vignette presented above, it seems clear that the purpose of the collaboration is to find leaves that look green enough to be used as props in a stage set. Green-painted but naturally russet leaves might well serve this purpose, so we may conclude that the truth conditions of Pia's utterance are met. (Grice would instead say that what we should be concerned with in this case is not the utterance itself but the conversational implicature it carries, but he would probably also judge the latter to be true.) Since the empirical prediction here is that subjects would accept Pia's utterance as true, I will call such contexts "acceptance contexts" (Ziółkowski 2017). Now compare the acceptance variant of the Leaves scenario with the vignette below:

> LEAVES — REJECTION CONTEXT
>
> *Pia has a Japanese maple tree in her backyard that has russet (reddish brown) leaves. She paints the leaves of the tree green. A friend of Pia's who is conducting a study of green-leaf chemistry asks if Pia has any green leaves she can use in her study. "The leaves on my tree are green," Pia says.*

As one might expect, the contextualist prediction for this variant of the story is that competent speakers will not accept Pia's utterance as true. By analogy to the explanation provided above, the purpose of the conversation here is to find leaves that would be suitable for a study of green-leaf chemistry.

[1] I am aware that Grice was not a contextualist, and his account is a form of invariantism (i.e., minimalism). Nevertheless, I find the pragmatic explanations provided by his framework useful here for the sake of illustration.

Since it does not seem that russet leaves that are painted green can serve this purpose, we should say that the truth conditions of Pia's utterance are not met in this context.[2] Since contextualists predict these types of contexts to elicit negative truth evaluations, I will call such contexts "rejection contexts." I will use the term "scenario" to refer to a content-matched pair of contexts (acceptance and rejection contexts) that together constitute a context-shifting experiment.

The contextualist predictions concerning truth evaluations in context-shifting experiments are empirically testable, and it is obvious that context-shifting experiments could easily be adapted for a full-blown experimental setting. Experimental philosophy (also referred to as "x-phi"), which uses tools borrowed from social sciences to investigate folk philosophical intuitions, provides an appropriate framework for testing contextualist predictions.

Before we proceed, however, it is important to express some reservations about the importance of context-shifting experiments for contextualism. Contextualism is not a unified theory; rather, it is an instance of family resemblance, to use Wittgenstein's (1953) terminology. Philosophers who subscribe to contextualism offer theories of language that are different in many respects and put forward different arguments in support of their claims. Context-shifting experiments are only one of the argumentative strategies used by contextualists, and it is obvious that the argumentative weight of this strategy will not be equivalent for every proponent of contextualism. For this reason, the importance of x-phi studies employing context-shifting experiments might vary depending on which contextualist theory we consider. What we would need here is a detailed classification of contextualist views, a list of the testable predictions that follow from each kind of contextualism, and an assessment of the importance of context-shifting experiments for these views.[3] Although I believe this task is well worth pursuing, it goes beyond the scope of this study, which focuses on the analysis of folk intuitions elicited by context-shifting experiments (even more narrowly, those that involve color adjectives), and, as will be explained in the next section, on the investigation of some of the methodological twists and turns of experimental philosophy.

[2] In fact, the situation is slightly more complex here. It seems that the expression "green leaves" is intended in its technical (biological) sense — i.e., leaves that can perform photosynthesis thanks to the chlorophyll they contain. Although the leaves on Japanese maple trees look reddish-brown, they do contain chlorophyll, so they *are* green leaves in this technical sense. It is hard to say whether the author of this case was aware of this, but let us assume that the leaves in question are still inappropriate for a green-leaf chemistry study due to contamination by paint.

[3] I am grateful to an anonymous reviewer for this suggestion.

It is also worth noting that the criteria of the demarcation between contextualism and minimalism (invariantism) are still subject to debate (see, e.g., Recanati 2003, Borg 2007). Some philosophers who count themselves in the minimalist (invariantist) camp are classified as contextualists by others,[4] and vice versa. The claim that the vast majority of natural-language expressions are context-sensitive is not the only defining feature of contextualism that is discussed in the literature, and some of these other features cannot be easily investigated in x-phi studies. This includes, for example, the rejection of the minimalist thesis termed "propositionalism," according to which every sentence expresses a complete (minimal) proposition regardless of the context of utterance; or the contextualists' claim that the impact of context on truth conditions of utterances is not triggered by syntax (*bottom-up*) but is entirely different in its nature (*top-down*). I will refrain from such considerations here.

Regardless of the reservations addressed above, it seems that a systematic investigation of folk intuitions elicited by context-shifting experiments is a worthwhile scientific endeavor that can contribute to the debate between contextualists and minimalists (invariantists).

1.2. FOLK SEMANTIC INTUITIONS IN CONTEXT-SHIFTING EXPERIMENTS: PREVIOUS EMPIRICAL FINDINGS

Contextualism is a philosophical view that received much attention from previous x-phi projects. The majority of studies on the topic concerned epistemic contextualism (e.g., DeRose 1992) and sought evidence for the context-dependence of knowledge attributions. According to epistemic contextualism, the truth conditions of a knowledge attribution, "*A* knows that *p*" (where *A* is an agent, and *p* is a proposition) might vary from context to context due to some pragmatic factors, such as what is at stake for the agent (DeRose 1992) or *A*'s practical interest (Stanley 2005).[5] If being wrong about the truth value of *p* does not matter much to the agent (the acceptance context in my terminology), then the standards for knowledge possession are lower (to put it roughly, it is easier to know something when it matters less). On the other hand, if there were a lot at stake, and erroneous beliefs about *p* were practically problematic for the agent (the rejection context), then the standards for

[4] For example, Kent Bach (2006) calls himself a minimalist, but Emma Borg classifies his view as moderate contextualism (because Bach rejects propositionalism).

[5] Jason Stanley calls his view "Interest Relative Invariantism" and does not take it to be an instance of contextualism. Nevertheless, many theoreticians involved in the dispute between invariantism (minimalism) and contextualism tend to classify Stanley's account as contextualist.

knowledge possession would go up. Many initial x-phi studies that explored this issue did not observe the predicted impact of context on knowledge ascriptions (Buckwalter 2010, May et al. 2010, Feltz, Zarpentine 2010). Although some experiments found evidence that was seemingly in favor of contextualism (Pinillos 2012, Sripada, Stanley 2012), recent large-scale studies did not confirm contextualists' hopes (Buckwalter, Schaffer 2015, Francis, Beaman, Hansen 2019,[6] Rose et al. 2019). The total body of empirical evidence suggests that epistemic contextualism is not supported by folk intuitions.

However, the case might not be entirely lost for contextualists. Hansen and Chemla (2013) investigated a number of different context-shifting experiments, including four knowledge scenarios,[7] four scenarios involving color adjectives (including the Leaves scenario presented above and three scenarios that can be found in the Appendix), and two scenarios they called "miscellaneous," since they could not be classified as a certain type. They used a within-subjects design: every participant assessed all ten scenarios and an additional control scenario; the variants of the scenarios were presented in a randomized order.[8] Every vignette was actually presented in four versions, as the researchers manipulated not only the context (acceptance/rejection) but also the valence of the target utterance (e.g., "The leaves on my tree are green" or "The leaves on my tree are not green").[9] Thus, every participant in their

[6] It is important to note that although Francis, Beaman, and Hansen (2019) did not find the contextualist effect on knowledge attributions in "standard" x-phi adaptations of context-shifting experiments (designs similar to studies of Buckwalter, May et al., or Feltz and Zarpentine), and did not manage to replicate Sripada and Stanley's (2012) results, they did find the effect of stakes when using what they call "the evidence-seeking design" (similar to that employed by Pinillos 2012). However, in their valuable discussion of the results they provide reasons to doubt whether the effects that emerge in the evidence-seeking designs are in fact effects of stakes *on* knowledge (these might be effects on something else entirely and affect knowledge ascriptions only indirectly).

[7] Each was a variation of DeRose's (1992) Bank Case, where the contextual shift manipulates what is at stake for the purported knower.

[8] To be more precise, Hansen and Chemla (2013) used a sophisticated randomization method they called "block design," which minimized the risk that two contextual variants of the same scenario would be presented directly one after another.

[9] This manipulation was introduced to empirically test some issues pointed out by DeRose (2011), who worries that the results of many x-phi studies concerning epistemic contextualism might be distorted due to the rule of accommodation (a term coined by Lewis 1979). The rule of accommodation is a pragmatic phenomenon that encourages hearers of an utterance to find an interpretation of it which might be taken to be true. In particular, subjects might feel the pressure to interpret the statement "I know that the bank will be open on Saturday" as true even in the rejection context. DeRose stresses that the true prediction of his version of epistemic contextualism is that laypersons will accept a positive statement in the acceptance context and a negative one (e.g., "I don't know that the

study had to make 44 truth evaluations in total. Subjects' answers were measured on a graphic scale (a continuous slider that was recorded to a 100-point scale for the sake of analysis), which made it possible to detect even subtle differences in judgments. Hansen and Chemla (2013) found systematic support for contextualist predictions: due to the shift in context, their subjects were less likely to judge that certain utterances were true. The size of the effect in the case of knowledge scenarios was very small (and, in fact, it showed only in the composite score, where the ratings of each of the four knowledge scenarios were aggregated), but for the other two types of cases the contextualist effect was well-pronounced. In the case of some scenarios, the difference between the average ratings in the acceptance and rejection contexts was even qualitative in nature (a shift from positive to negative truth evaluations).

Unfortunately, Hansen and Chemla's (2013) study suffers from some weaknesses. One is its small sample size (N = 39), which results in a relatively low statistical power to accurately estimate the effect sizes. Even though their experiment employed a within-subjects design, and in some analyses they used composite scores with which they aggregated subjects' responses for each scenario type (knowledge, colors, and "miscellaneous"), one would still like to see a larger study before drawing strong conclusions from the data. One might also worry about the workload assigned to each participant, as it is difficult for a subject filling out an online survey to pay full attention when going through 44 vignettes. On the other hand, if the participants did pay attention, there might be another distorting factor at play: even though the presentation of vignettes was randomized, after assessing a number of cases some subjects might have easily learned what the experimental manipulation was; therefore, they could have "learned" a certain pattern of responses in the early stages of the experiment and later copied this pattern in further parts of the experimental procedure. These concerns could easily be dismissed if we found that similar contextualist effects occur when each scenario is evaluated by separate groups of respondents or in a between-subjects design. Thus, I decided to replicate the part of Hansen and Chemla's study concerning color adjectives but using a bigger sample (N = 1128) and different experimental settings in order to corroborate their findings.

Apart from simply exploring the impact of conversational context on folk semantic intuitions regarding color adjectives, in my studies I introduced another experimental manipulation to test the methodological objections ad-

bank will be open on Saturday") in the rejection context. However, Hansen and Chemla (2013) did not find evidence that the rule of accommodation biased the answers provided by their respondents, as the impact of the contextual shift they observed was similar in size for positive and negative statements.

dressed in my earlier study (Ziółkowski 2017). I argued that the within-subjects experimental setting used by Hansen and Chemla (2013) could artificially enhance the impact of the contextual shift on truth evaluations. Hansen (2014) also believes that choosing a within-subjects design makes a difference, but he argues that it is the preferred method for experimental adaptations of context-shifting scenarios. He claims that giving subjects the opportunity to contrast cases helps them focus on the important features of the vignettes (the shifts in context) and see their importance in shaping the contents of utterances. In other words, he believes that a within-subjects design makes the experimental context manipulation more apparent and effective; on the other hand, in a between-subjects design, subjects might overlook the importance of contextual features for truth conditions since they are not encouraged to consider different possible contexts of utterances. In my previous paper (Ziółkowski 2017), I disagreed and noticed that the purported boost in the sizes of contextual effects observed in a within-subjects design (compared to between-subjects) might have resulted from a different pragmatic mechanism that goes beyond the sheer influence of the context of utterance on truth conditions. I presented a hypothesis inspired by the Gricean framework, according to which contrasting cases in a within-subjects design might trigger specific conversational implicatures in the communication between the researchers and the participants of the study: when confronted with very similar stories accompanied by the exact same question, subjects might feel encouraged to look for an interpretation of the questions (or stories) that would result in more divergent responses between the acceptance and rejection contexts (for a detailed argumentation, see Ziółkowski 2017: 153-154). Such a phenomenon, however, is not what we are after when conducting context-shifting experiments. A within-subjects design might not only conflate the two different mechanisms but also distort the data. To substantiate my concerns, I reported some tentative data from an experiment in which I used two scenarios borrowed from the study by Hansen and Chemla (the “miscellaneous” cases). I found contextual effects in the predicted direction, but also observed order effects in the within-subjects design: participants contrasted their truth evaluation between contexts more for some orders of context presentation. As a result, the effect sizes I observed in the within-subjects design were somewhat larger than those obtained for the between-subjects design part of my study. According to the argumentation I provided (Ziółkowski 2017), since this “boost” can be explained within the minimalist (invariantist) Gricean framework, it does not necessarily lend more empirical support to contextualist predictions.

As a result of this debate, which is rich in empirical predictions, I decided to compare the possible methodological variants of empirical adaptations of context-shifting experiments, and test each scenario in three different experimental settings. Experiment 1 mimics the methodology used in (Ziółkowski 2017): it compares between-subjects (where participants evaluating different contexts are distinct groups) with within-subjects (with a counterbalanced order of context presentation) and allows order effects to be detected. However, one might easily notice that Experiment 1 is not sufficient to test Hansen's (2014) methodological predictions, because a within-subjects design with a counterbalanced order of presentation of contexts does not accurately reflect what he had in mind when he used the expression "contrasting cases." The reasons for this are twofold. First of all, since the procedure involves viewing one context after another on separate screens, no direct comparison of contexts is possible. Second, as the order of presentation of contexts is counterbalanced, when making their truth evaluations, some respondents have a chance to compare the rejection context with the acceptance context, but not the other way around (A-R order), while others can only compare the acceptance context with the rejection context (R-A order), without being able to see the difference between contexts when assessing the rejection context.

Fortunately, the idea of contrasting contexts can be approached differently. In order to establish whether more explicit ways of contrasting cases bring different results in context-shifting experiments than the regular within-subjects design, in Experiment 2 I adopt a method of scenario presentation that I will call "contrastive design," which is another variant of a within-subjects design. It consists in presenting both contextual variants of a given scenario to subjects simultaneously (vignettes viewed next to each other on the same screen of the survey). If contrasting contexts makes the importance of context for truth conditions more vivid to participants and encourages judgments along the lines of contextualist predictions, then we should, as Hansen (2014) suggests, observe more pronounced contextualist effects with the contrastive design in Experiment 2 than with the between-subjects design in Experiment 1.

I believe that resolving the methodological issue discussed above is important in terms of progress in experimental philosophy, particularly in x-phi studies concerning contextualism. If we find the expected differences between the within- and between-subjects designs, we will need to continue the debate as to which design is more appropriate for investigating contextualist predictions about context-shifting experiments — an issue that probably cannot be resolved empirically, because it requires philosophical argumentation. On the other hand, if it turns out that all the methodological variants of experimental adaptations of context-shifting scenarios bring similar results, there will be no reason to favor one over another.

In the following sections, I spell out the research hypotheses and present a detailed description of my studies, their results, and the conclusions that can be drawn from the new empirical data.

2. RESEARCH OBJECTIVES AND HYPOTHESES

The first aim of the study was to empirically test contextualist predictions regarding the context-sensitivity of color adjectives. Since I used materials borrowed from the experiment conducted by Hansen and Chemla (2013) — including the Leaves scenario presented above and three other vignettes that can be found in the Appendix — but did not use the exact same experimental setting, we can say that my study was a conceptual replication of their experiment. After all, both these studies tested the same hypothesis:

(H1) Laypersons are more likely to agree with the target utterance in the acceptance contexts than in the rejection contexts across all the scenarios included in the study and for each scenario considered separately.

Besides this substantive philosophical hypothesis, I also investigated the methodological claims raised by Hansen (2014) and Ziółkowski (2017); see the section above for details. I expected that the experimental setting that allows subjects to compare contexts would bring more support to contextualist predictions:

(H2) The differences in folk truth evaluations between the acceptance and rejection contexts are larger when subjects are given the opportunity to contrast the contexts (within-subjects design) than when separate groups evaluate each context (between-subjects design).

Moreover, I linked the second hypothesis to my earlier claims (Ziółkowski 2017) about order effects in context-shifting experiments: I expected that if H2 is borne out, it will be due to the impact of context ordering on subjects' judgments:

(H3) In the within-subjects design with a counterbalanced order of context presentation, participants contrast their answers when given such an opportunity: they agree more with the target utter-

> ance in the acceptance context when it is preceded by the rejection context (in comparison to the condition in which the acceptance context is presented first); they agree less with the target utterance in the rejection context when it is primed with the acceptance context (in comparison to the condition in which the rejection context is not primed).

Before I present the methods and results of the experiments, I would like to state an important reservation. Although H2 is explicitly endorsed by Hansen (2014), and H3 is a prediction following from my own claims and initial empirical findings (Ziółkowski 2017), these two methodological hypotheses are *not* empirical predictions that are entailed by (most) contextualist accounts. Many proponents of contextualism do not discuss different possible empirical adaptations of context-shifting experiments nor do they make predictions about them. But if they did, they might expect H2 to turn out false, and they even provide theoretical reasons for such a prediction.[10] For example, when DeRose (1999, 2011) discusses his version of epistemic contextualism, he does consider a situation in which different contexts of utterance of the same knowledge attribution are contrasted with each other, and he takes it to be a situation in which an entirely different context might affect semantic intuitions. Via contrasting contexts, the within-subjects adaptations of context-shifting experiments might result in the creation of a new context (let us call it "the respondent's context") that is not reducible to the features of the acceptance and rejection contexts considered separately (which we might call the "speaker's context"). In their respondent's context, the participants of a within-subjects x-phi adaptation of a context-shifting experiment might feel inclined to confer the same interpretation on both the utterances, regardless of the differences between the acceptance and rejection contexts that are nevertheless present in the speaker's context. (Similarly, when a subject considers both the claim "I have hands" and "I am a handless brain-in-a-vat," DeRose would predict that the subject might be inclined to judge both as true or both as false.) Therefore, if my experiments find no support for H2 and H3, this does not have to be bad news for contextualists (some of them might even welcome such a result). I would like to stress here that I am aware that only H1 is crucial for contextualists, while H2 and H3 are important for the methodology of x-phi adaptations of context-shifting experiments.

[10] I am grateful to an anonymous reviewer for noticing this.

3. EXPERIMENT 1 — BETWEEN-SUBJECTS AND WITHIN-SUBJECTS

3.1. METHODS AND PROCEDURE

In Experiment 1, I employed methods similar to those used in (Ziółkowski 2017) in the first study, where I attempted to compare the results of between- and within-subjects designs in context-shifting experiments. Each subject was randomly assigned to one of four scenarios (each was a pair of two contextual variants of the same story) and one of two possible orderings of contexts: acceptance-rejection (A-R) or rejection-acceptance (R-A). The contextual variants of the vignettes were presented one after another on separate screens so that participants had no opportunity to return to the previous pages of the survey and change their antecedent answers when confronted with the second variant. While in (Ziółkowski 2017) I investigated the two scenarios classified by Hansen and Chemla (2013) as "miscellaneous," in Experiment 1 I focused on four scenarios that involved color adjectives (also borrowed from Hansen and Chemla's study): Leaves (presented above in section 1), Walls, Kettle, and Apples. The contents of the last three vignettes can be found in the Appendix.

By assigning each subject to one scenario and to one ordering of contexts, we can compare the between- and within-subjects designs in one dataset and also detect possible order effects. If we look at judgments elicited by the contextual variants of the scenarios that were presented first, we obtain the between-subjects design (two distinctive groups of subjects assigned to conditions A-R or R-A). If we aggregate the data from both orderings and compare judgments for the acceptance and rejection contexts, we end up with the within-subjects design (counter-balanced for the order of presentation). Finally, we can compare judgments elicited by the same contextual variant of a given scenario in different experimental conditions (orders of presentation), which makes it possible to observe presumptive order effects and test whether the within-subjects design is more favorable to contextualist predictions than the between-subjects design, as suggested by Hansen (2014) and Ziółkowski (2017).

The experiment was designed with LimeSurvey (open-source software for online surveying; http://www.limesurvey.org) and was carried out online (it was posted on servers owned by KogniLab, the x-phi laboratory based at the University of Warsaw; http://www.kognilab.pl). The first part of the survey included demographic and screening questions. The participants were asked about their gender, age, and education. For screening purposes, the survey included questions about respondents' native language and their philosophical training. Subjects who were non-native English speakers or had an academic degree (BA, MA, or PhD) in philosophy were excluded from further analyses.

Each participant was presented with two contextual variants of one of the four scenarios mentioned above, one after another, depending on the randomly assigned experimental condition (A-R or R-A). The subjects had to answer all questions regarding the presented vignette before proceeding to the next vignette; after providing their answers, they had no chance to return to the previous pages of the survey and change their antecedent judgments. Each vignette (contextual variant of a given scenario) was accompanied by two questions: first, a comprehension check control question and then the crucial question about the truth value of the target utterance of the protagonist in the vignette. Comprehension questions were always simple true-or-false queries about some factual aspects of the presented vignette. For example, in the Leaves-Acceptance condition the question was "True or False: 'Pia's friend needs green leaves she can use in her stage set'." In the Leaves-Rejection condition, the question was "True or False: 'Pia's friend needs green leaves for a chemistry study'." Subjects who answered at least one of the two comprehension questions incorrectly ("false") were excluded from the final analysis. The target question about the truth value of the utterance made by the protagonist in the vignette had the same format in each experimental condition: [protagonist's name]'s claim "[target sentence]" is true. For instance, the question for the Leaves scenario was "Pia's claim 'The leaves on my tree are green' is true" (it was identical in both contextual variants, acceptance and rejection). The participants could express their intuitions on a 5-point Likert scale, ranging from "Disagree" (numbered "1" in the analysis) to "Agree" (numbered "5" in the analysis); only the end-points of the scale were labelled.

3.2. PARTICIPANTS

The subjects were recruited via ClickWorker (http://www.clickworker.com), which is a German-based internet service that offers access to a large community of internet users interested in completing simple paid tasks, including participating in academic research. Every respondent received small financial compensation for taking the survey.

In total, 1,052 participants completed the survey, but 320 were excluded from the analysis (for exclusion criteria, see the screening procedure described in section 3.1), which yields the final sample size N = 732. The following statistics are reported for the final filtered sample.

The average age of the respondents was 38.93 (SD = 12.42). 49.5% of participants identified themselves as female, 49.7% as male, and 0.8% chose the answer "other."

3.3. RESULTS

3.3.1. EXPERIMENT 1A: BETWEEN-SUBJECTS DESIGN

First, I will report the results of the analysis that focused on the judgments elicited by the first presented vignette, which can be identified with a full-blown between-subjects design. I subjected the data to a two-way 4x2 ANOVA analysis with Scenario (Apples, Kettle, Leaves, Walls) and Context (acceptance, rejection) as factors, and truth evaluation of the protagonist's utterance as the dependent variable.

The analysis revealed a significant main effect of Context: $F(1, 724) = 13.74$; $p < 0.001$; $\eta^2 = 0.019$. On average, the participants were more inclined to agree with the protagonist's target utterance in the acceptance context ($M = 3.97$; $SD = 1.46$) than in the rejection context ($M = 3.58$; $SD = 1.62$). Using the measure proposed by Jacob Cohen (1995), the size of the effect is $d = 0.25$, which could be classified as a small effect. Scenario also had a significant impact on subjects' judgments: $F(3, 724) = 38.58$; $p < 0.001$; $\eta^2 = 0.138$. Further post-hoc pairwise comparisons (Tukey's HSD) revealed the following pattern of differences: Kettle > Apples & Leaves; Walls > Apples & Leaves; Apples > Leaves. No significant differences between the Kettle and Walls scenarios were observed (these two scenarios received the highest average ratings). Additionally, a significant interaction between Scenario and Context emerged: $F(3, 724) = 12.6$; $p < 0.001$; $\eta^2 = 0.05$. Further inspection of the simple effects in the interaction found that this resulted from the fact that while for Kettle, Leaves, and Walls subjects were more likely to agree with the target utterance in the acceptance context than in the rejection context (which is along the lines of contextualist predictions), the direction of the differences was opposite for the Apples scenario, in which (contrary to expectations) participants were significantly more inclined to give a positive judgment in the rejection context than in the acceptance context. The results are illustrated in the figure below.

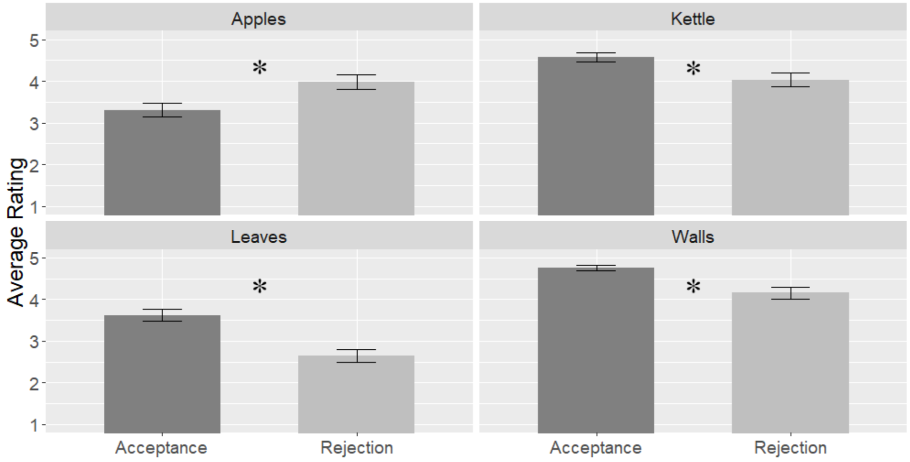

Figure 1. Average ratings for the truth-value question in each experimental condition in the between-subjects design (1 = "Disagree," 5 = "Agree"). Error bars represent standard error of mean. Significant differences between contexts are marked with an asterisk.

Table 1 summarizes the results we are most interested in: the contextualist effect for each scenario and its size. For Walls and Leaves, the observed contextualist effect can be classified as medium in size, according to conventional benchmarks proposed by Cohen (1995). In the case of Kettle and Apples, the effect sizes were small; but again, for Apples, the observed effect runs in the opposite direction to what was expected.

Scenario	Context	*M*	*SD*	Cohen's *d*	Effect Direction
Apples	Acceptance Rejection	3.30 3.98	1.74 1.53	**0.42**	Opposite to predictions
Kettle	Acceptance Rejection	4.58 4.03	0.88 1.32	**0.49**	As predicted
Leaves	Acceptance Rejection	3.62 2.65	1.49 1.62	**0.62**	As predicted
Walls	Acceptance Rejection	4.75 4.15	0.64 1.36	**0.56**	As predicted

Table 1. Truth evaluations for Scenario x Context in the between-subjects design: average ratings, standard deviations, effect sizes, and direction of the effect

3.3.2. EXPERIMENT 1B: WITHIN-SUBJECTS DESIGN

Now I will present the results of the within-subjects design that were obtained by aggregating the data from counter-balanced orders of presentation and comparing the two truth evaluations provided by each participant. Since here we look at repeated measures, I performed a 4x2 mixed-ANOVA analysis, where Scenario was the between-subjects factor, and Context was the within-subjects factor.

The general pattern of results was similar to what was observed for the between-subjects design. First of all, Context had a significant impact on truth evaluations: $F(1, 728) = 76.09$; $p < 0.001$; $\eta^2 = 0.095$. When we look at data aggregated from all the scenarios, subjects were happier to agree with the target utterance in the acceptance context ($M = 4.02$; $SD = 1.41$) than in the rejection context ($M = 3{,}54$; $SD = 1{,}63$). The point-estimate of the effect size here is $d = 0.31$. A second main effect also emerged: subjects' ratings differed between scenarios: $F(3, 728) = 52.82$; $p < 0.001$; $\eta^2 = 0.179$. According to pairwise comparisons, subjects were least likely to give a positive judgment when confronted with the Leaves scenario (compared to the other three scenarios). Moreover, participants tended to agree more with the target utterance in Kettle and Walls than in Apples. Again, no significant differences between Kettle and Walls were found. Additionally, a significant interaction between the two factors (Scenario and Context) was observed: $F(3, 728) = 24.78$; $p < 0.001$; $\eta^2 = 0.093$. Here, the causes of the interaction were only slightly different from what was found for the between-subjects design. While Context affected subjects' judgments in the predicted way for Leaves and Walls, no impact of Context was found in the case of Apples and Kettle. Figure 2 illustrates these results.

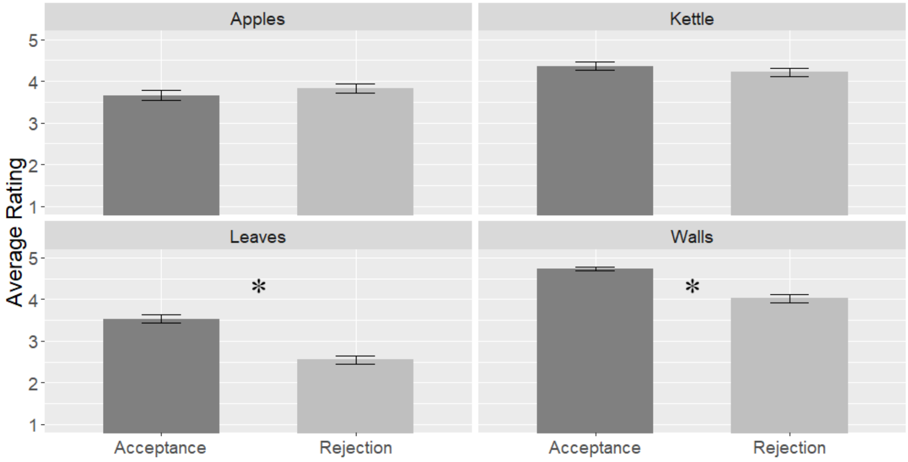

Figure 2. Average ratings for the truth-value question in each experimental condition in the within-subjects design (1 = "Disagree," 5 = "Agree"). Error bars represent standard error of mean. Significant differences between contexts are marked with an asterisk.

Detailed statistics regarding contextual effects for each investigated scenario can be found in Table 2. The results for Leaves and Walls were highly consistent with what was found in the between-subjects analysis (small- to medium-sized contextual effects). However, the picture is quite different when it comes to Apples and Kettle: contrary to what the between-subjects analysis found, no contextual effect was observed in the case of these two scenarios.

Scenario	Context	*M*	*SD*	Cohen's *d*	Effect Direction
Apples	Acceptance Rejection	3.66 3.82	1.64 1.56	**0.11**	Opposite to predictions (non-significant)
Kettle	Acceptance Rejection	4.36 4.21	1.05 1.17	**0.13**	As predicted (non-significant)
Leaves	Acceptance Rejection	3.53 2.55	1.53 1.64	**0.60**	As predicted
Walls	Acceptance Rejection	4.74 4.02	0.69 1.38	**0.47**	As predicted

Table 2. Truth evaluations for Scenario x Context in the within-subjects design: average ratings, standard deviations, effect sizes, and direction of the effect

3.3.3. ORDER EFFECTS

I will now address the methodological concerns I raised in my earlier paper (Ziółkowski 2017). As mentioned previously (section 1), I suggested that judgments elicited in within-subjects context-shifting experiments are prone to order effects. When investigating the purported order effects in Experiment 1, I will look at ratings for the acceptance and rejection contexts separately.

First, let us consider the truth evaluations in the acceptance context depending on the order of presentation, regardless of the scenario. If we aggregate the data collected for all four scenarios included in Experiment 1, we find no difference in subjects' tendency to agree with the target utterance between the acceptance-rejection condition (*M* = 3.97; *SD* = 1.46) and the rejection-acceptance condition (*M* = 4.06; *SD* = 1.36). Thus, on a larger scale, the order of context presentation did not affect subjects' judgments regarding the acceptance context. However, if we look at the data separately for each scenario, we can observe some interesting trends. While there was no detectable impact of order on judgments regarding the acceptance context in the case of Leaves and Walls scenarios, the order of presentation seemed to affect participants' judgments in the other two scenarios. Hansen (2014) and I (Ziółkowski 2017) suggested that subjects would be likely to contrast their answers when confronted with two conversational contexts one after another. This should result in more positive truth evaluations regarding the acceptance context when it is presented after the rejection context (R-A) compared to the condition in which it is presented first (A-R). This is what was found for the Apples scenario: the average rating in the acceptance context was significantly higher in the R-A condition (*M* = 4.14; *SD* = 1.36) than in the A-R condition (*M* = 3.30; *SD* = 1.74); t(186,28) = 3.70; *p* < 0.001; *d* = 0.54. Surprisingly, I also found an order effect in the case of the Kettle scenario, but here the direction was opposite to predictions: subjects tended to agree more with the target utterance in the acceptance context when it was presented first (*M* = 4.58; *SD* = 0.88) than when the rejection context preceded it (*M* = 4.13; *SD* = 1.18); t(110,66) = 2.40; *p* = 0.018; *d* = 0.43. Detailed results are summarized in Table 3.

Scenario	Acceptance First		Acceptance Second		Cohen's d
	M	SD	M	SD	
Apples	3.30	1.74	4.14	1.36	0.54
Kettle	4.58	0.88	4.13	1.18	0.43
Leaves	3.62	1.49	3.45	1.56	0.11
Walls	4.75	0.64	4.72	0.74	0.04

Table 3. Ratings for the truth-evaluation question in the acceptance context depending on the order of presentation

In the case of the rejection context, the analysis found no impact of the order of presentation on participants' judgments regarding the truth value of the utterance in question, both for the dataset as a whole (R-A order: *M* = 3.58; *SD* = 1.62. A-R order: *M* = 3.50; *SD* = 1.63) and for each scenario considered separately. Average ratings in each scenario and condition are presented in Table 4.

Therefore, contrary to the predictions based on my initial findings (Ziółkowski 2017), I did not observe that the order of presentation of conversational contexts strongly shaped subjects' truth evaluations in context-shifting experiments involving color adjectives.

Scenario	Rejection First		Rejection Second		Cohen's d
	M	SD	M	SD	
Apples	3.98	1.53	3.71	1.58	0.17
Kettle	4.03	1.32	4.38	0.99	0.30
Leaves	2.65	1.62	2.44	1.65	0.13
Walls	4.15	1.36	3.89	1.40	0.19

Table 4. Ratings for the truth-evaluation question in the rejection context depending on the order of presentation

3.4. EXPERIMENT 1 — DISCUSSION AND CONCLUSIONS

Let us now evaluate the research hypotheses in the light of the data obtained in Experiment 1. When it comes to the first hypothesis regarding the impact of conversational context on intuitive truth conditions, we might say that the data lends some support to contextualist predictions. If we look at all the ratings regardless of the scenario, our participants were less likely to agree with the target utterance in the rejection contexts than in the acceptance contexts for both the between- and within-subjects designs; this is along the lines of the arguments put forward by contextualists. Thus, we can also say that Experiment 1 at least to some extent corroborates previous findings regarding the context-sensitivity of color adjectives reported by Hansen and Chemla (2013). The results were highly consistent with the Leaves and Walls scenarios, where I observed contextualist effects similar in size (and in the same direction) in both the between- and within-subjects designs. The picture is considerably less clear for the Apples and Kettle scenarios, where we see a discrepancy between the between- and within-subjects designs (I will discuss this issue below when evaluating further hypotheses). It is worth noting, however, that the observed contextualist effects were, at best, medium in size according to the benchmarks proposed by Cohen (1995). Of course, these benchmarks are only "rules of thumb," and the decision whether a given observed effect size is theoretically important always depends on our theoretical interest and empirical predictions addressed *ex ante*. We have reasons to believe that most contextualist theoreticians would predict a larger difference between conversational contexts than what the experiment found: a shift in truth evaluations from mostly positive judgments in the acceptance condition to mostly negative judgments in the rejection condition.[11] This is not what can be seen in my data: while there is a difference in average truth evaluations between conversational contexts, it is not as extreme as many contextualists would like it to be. Therefore, it is an open question whether contextualist predictions regarding the context-sensitivity of color adjectives are empirically supported by my data.

It is worth noting that the within-subjects design has one merit that the between-subjects design does not possess (when it comes to assessing contextualist predictions for context-shifting experiments). While the between-subjects design only allows us to compare the distribution of answers provided by separate groups of respondents assigned to the acceptance or rejection contexts, in the within-subjects design we can look at pairs of judgments provided by individual subjects in each context, which grants us a new way of evaluating contextualist predictions.[12] In fact, this data is revealing: if we look at the whole sample from Experiment 1, it turns out that the majority of participants (61%) gave the exact same rating to the target utterance in the acceptance and rejection contexts. To put it differently, the majority of folk truth evaluations were not sensitive to the contextual shift, even when laypersons

[11] Travis (1997) is the most obvious example here; many philosophers later agreed with him.

[12] I am grateful to a reviewer for suggesting this approach.

had the opportunity to compare contexts that were presented one after another. Interestingly, nearly 10% of subjects gave a higher rating to the utterance in the rejection context than in the acceptance context, which is opposite to what contextualists expect, and only the remaining 29% gave ratings along the lines of contextualist predictions. However, even in this latter group, many subjects rated the utterance in the acceptance context only slightly higher than in the rejection context; only 14% of the subjects who participated in Experiment 1 gave responses that represent a true qualitative switch from a positive answer (categories 4 or 5) to a negative one (categories 1 or 2). Thus, we can say that the contextualist effects I found in my study were not only small but were also generated by a minority of participants. I will return to this issue in the final discussion.

Unlike the first hypothesis, for which the situation is disputable, the other two hypotheses were clearly not borne out by the experiment. When it comes to H2, although there were some discrepancies between the within- and between-subjects designs, their direction was opposite to what we expected. First of all, contrary to what the hypothesis predicted, the size of the contextual effect observed in the case of Leaves and Walls was quite similar in the between- and within-subjects designs. Second, for Apples and Kettle, the effect sizes turned out to be smaller (and the differences between contexts were non-significant) in the within-subjects design compared to the between-subjects design, which is the exact opposite of what I predicted. In other words, I did not find evidence in support of Hansen's (2014) and my own (Ziółkowski 2017) claims that contrasting cases with the help of a within-subjects design encourages laypersons to diversify their truth evaluations between the contrasted conversational contexts. The fact that we detected some effects in the between-subjects design that were not confirmed in the within-subjects design is somewhat surprising. It is hard to come up with a substantive interpretation of this result; we might presume that the between-subjects findings for the Apples and Kettle scenarios were false positives due to statistical noise, but such a claim would require further studies and a richer body of empirical evidence. Of course, it is possible that in the between-subjects analysis we discovered true effects that we failed to detect in the within-subjects part of the study, but this latter explanation seems less probable than the former (especially if we take into account the results of Experiment 2 that are presented in section 4.3).

When it comes to the third hypothesis, while I found some evidence that the order of presentation of conversational contexts might affect subjects' responses in context-shifting experiments, the observed order effects were not as prominent as expected. In the case of rejection contexts, the analysis revealed no order effects whatsoever: participants were as likely to agree with the target utterance when the rejection context was presented first as when it was preceded by the acceptance context. On the other hand, the judgments elicited by the acceptance context seemed to be influenced by the order of presentation, but only in the case of Apples and Kettle. Moreover, the direction of the effect fits the predictions only for the Apples scenario. Thus, I believe that my own earlier suggestions that the within-subjects variant of context-shifting experiments might favor contextualism due to order effects (Ziółkowski 2017) is not substantiated by the data collected in Experiment 1.

One last thing worth noting is that when we compare the influence of the two factors included in the analyses (Scenario and Context), it becomes clear that the contents of the scenario had a stronger impact on subjects' responses than manipulating the conversational context. Clearly, the four scenarios involving color adjectives that were borrowed from Hansen and Chemla's (2013) study are not on a par with respect to the semantic intuitions they elicit. While some scenarios, such as Kettle and Walls, rarely encourage subjects to make negative truth evaluations, others, such as Leaves and Apples, elicit diverse judgments, with some subjects leaning towards the positive and others towards the negative side of the scale. I will discuss this issue again in the final section.

4. EXPERIMENT 2 — CONTRASTIVE DESIGN

4.1. METHODS AND PROCEDURE

In the second experiment, I tried to further explore the possible impact of the experimental design on semantic judgments elicited by context-shifting experiments. In this study, I employed a contrastive design (participants assessed both contextual variants of a given scenario parallelly), which is yet another variant of the within-subjects design. Scenario remained a between-subjects factor (i.e., each participant was randomly assigned to one scenario).

Since all the questions concerning the vignettes were asked at the same time, the vignettes were labeled "Story 1" (acceptance context; viewed on the left-hand side of the screen) and "Story 2" (rejection context; viewed on the right-hand side of the screen). This made it possible to clearly refer to each vignette in the following questions. Both the comprehension check and the crucial question about the truth value started with the prefix "In Story 1, ..." or "In Story 2, ..." For example, the comprehension question in the Leaves acceptance condition was: "True or False: 'In Story 1, Pia's friend needs green leaves she can use in her stage set'." The target question about the truth value of the utterance was: "In Story 1, Pia's claim 'The leaves on my tree are green' is true." As was the case in Experiment 1, when answering the latter question, participants were offered a 5-point Likert scale ranging from "Disagree" (1) to "Agree" (5).

The demographic section preceded the main part of the survey. Here, participants were asked about their age, gender, education, philosophical training, and whether English was their first language. Again, like in Experiment 1, I excluded subjects who were not native English speakers, had an academic degree in philosophy, or answered at least one of the two comprehension questions incorrectly.

4.2. PARTICIPANTS

As was the case in Experiment 1, the respondents were recruited via ClickWorker (http://www.clickworker.com) and received a small sum of money for completing the survey.

Overall, 527 subjects participated in the second experiment. However, 131 respondents were excluded from further analyses (for exclusion criteria, see section 4.1). Thus, the final sample size is N = 396.

Out of those 396 participants, 51.5% identified themselves as female, 48.2% as male, and one subject (0.3%) chose the option "other." Their average age was 38.08 (*SD* = 11.56).

4.3. RESULTS

In order to analyze the data, I employed a 4x2 mixed-ANOVA model with Scenario (Apples, Kettle, Leaves, Walls) as a between-subjects factor and Context (acceptance, rejection) as a within-subjects factor (repeated measures).

Once again, both main effects were significant. Context had a noticeable impact on subjects' truth evaluations across all the scenarios: $F(1, 392) = 32.93$; $p < 0.001$; $\eta^2 = 0.077$. Subjects were less likely to agree with the target utterance in the rejection context ($M = 3.76$; $SD = 1.57$) than in the acceptance context ($M = 4.16$; $SD = 1.32$). The estimation of the effect size here is $d = 0.27$, which, again, is a rather small effect. However, there were also significant differences between judgments elicited by different scenarios: $F(3, 392) = 19.38$; $p < 0.001$; $\eta^2 = 0.129$. Post-hoc pairwise comparisons revealed that subjects were happier to agree with the target utterance in Kettle and Walls than in Apples and Leaves (no differences between Kettle and Walls were found). Additionally, participants were less likely to make a positive judgment when confronted with the Leaves scenario compared to the Apples scenario.

As was the case in Experiment 1, an interaction between the two factors emerged: $F(3, 392) = 21.36$; $p < 0.001$; $\eta^2 = 0.141$. This was due to the fact that while context had the predicted impact on truth evaluations for the Leaves and Walls scenarios (subjects disagreed more with the target utterance in the rejection context than in the acceptance context), no contextual effect was found in the case of Apples and Kettle. This result is consistent with the findings of the within-subjects design in Experiment 1, but it is inconsistent with what was obtained in Experiment 1 with the between-subjects design.

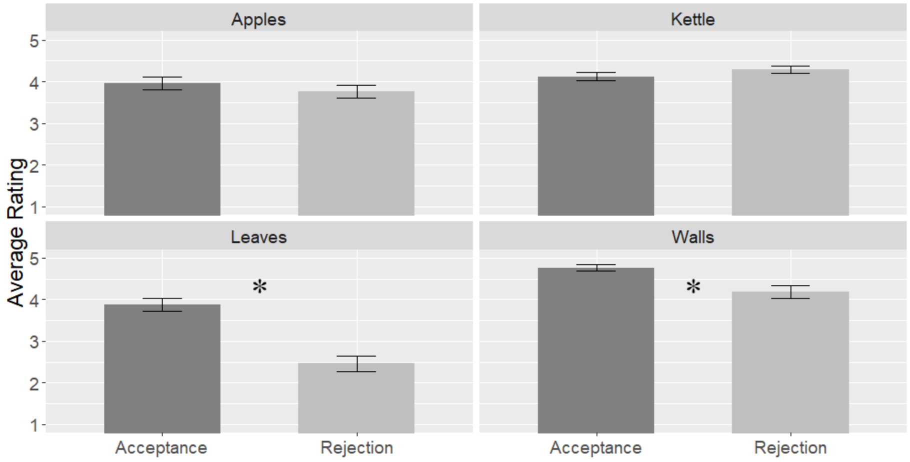

Figure 3. Average ratings for the truth-value question in each experimental condition in the contrastive design (1 = "Disagree," 5 = "Agree"). Error bars represent standard error of mean. Significant differences between contexts are marked with an asterisk.

Below, I summarize the results regarding the sizes of the contextual effect separately for each scenario investigated in Experiment 2.

Scenario	Context	*M*	*SD*	Cohen's *d*	Effect Direction
Apples	Acceptance	3.96	1.61	**0.15**	As predicted (non-significant)
	Rejection	3.77	1.63		
Kettle	Acceptance	4.12	1.21	**0.14**	Opposite to predictions (non-significant)
	Rejection	4.29	1.05		
Leaves	Acceptance	3.87	1.35	**0.79**	As predicted
	Rejection	2.46	1.71		
Walls	Acceptance	4.76	0.73	**0.42**	As predicted
	Rejection	4.18	1.33		

Table 5. Truth evaluations for Scenario x Context in the contrastive design: average ratings, standard deviations, effect sizes, and direction of the effect

4.4. FINAL CONCLUSIONS

The results I obtained in Experiment 2 are surprisingly consistent with the data collected in the within-subjects part of Experiment 1. They are also largely in line with the findings of the between-subjects part of Experiment 1, because for the Leaves and Walls scenarios I found contextual effects in the predicted direction that were similar in size in all three methodological variants that were employed in the studies. The discrepancy I observed in the case of the Apples and Kettle scenarios, where significant contextualist effects were found in the between-subjects design but not in the other two experimental designs, is the opposite of what was predicted (I expected to find more pronounced contextual effects in the two within-subjects designs in comparison to the between-subjects design). Therefore, we cannot say that my experiments found much evidence in support of the hypothesis put forward by Hansen (2014) and Ziółkowski (2017), because contrasting cases does not make an important difference for empirical adaptations of context-shifting experiments.

When it comes to contextualist predictions, we might say that my data largely corroborates previous findings reported by Hansen and Chemla (2013) concerning the context-sensitivity of color adjectives. The semantic intuitions elicited by the Leaves and Walls scenarios fit the pattern of contextualist predictions in all three methodological variants of the study. However, the size of the influence of conversational context on truth evaluations in these cases is rather small, which leaves open the question of whether contextualism is empirically grounded (i.e., whether the strength of evidence is satisfactory).

Since the contrastive design used in Experiment 2 is a type of within-subjects design, we can again take a closer look at the pairs of judgments made by each subject in reaction to both the acceptance and rejection contexts, as was the case in Experiment 1b (see section 3.4). Once again, it turns out that the majority of participants (69%) assessed the target utterance identically regardless of the contextual manipulation. 8% gave a higher rating in the rejection context than in the acceptance context, and 22% of subjects provided answers that fitted the contextualist predictions. Unfortunately for contextualists, only 14% of participants of Experiment 2 exhibited the pattern of responses that fully fit contextualist predictions: they switched from a positive answer in the acceptance context to a negative answer in the rejection context. These results are strikingly similar to the data collected in Experiment 1b, as the contextualist effects I managed to observe also resulted from answers provided by a considerably small group of subjects.[13] The fact that I failed to detect contextualist effects in the case of the Apples and Kettle scenarios (at least in the within-subjects part of Experiment 1 and the contrastive design in Experiment 2) means that an explanation is very much needed.[14]

One might feel tempted to conclude that the data collected in my studies casts doubt on contextualism about the context-sensitivity of color adjectives, but I believe that this conclusion would be too hasty. Alternative explanations can easily be offered: one might doubt that the experimental manipulations with contextual features in the vignettes were apparent enough for the participants, or, to put it differently, whether the experimental manipulation with context was effective.[15] After all, it is possible that the vignettes used in

[13] Interestingly, in a recent study on the gradeability of color adjectives, Hansen and Chemla (2017) observed a similar phenomenon: the effects they found were driven by a small group of subjects and only a minority of participants exhibited contextualist intuitions.

[14] When we look closely at the data reported by Hansen and Chemla (2013), we can easily notice that the findings of my studies are surprising only in the case of Kettle. Although Hansen and Chemla did not conduct pairwise comparisons between contexts for each scenario separately (at least, they only present analyses of composite scores, where they aggregated the data from scenarios that belonged to the same type — e.g., "color cases"), they do present detailed data in a figure (see Hansen and Chemla, 2013: 305, Figure 6.). The ratings for the Apples scenario trended slightly in the predicted direction, but one might guess that the differences between contexts were non-significant. Therefore, after all, my findings are in line with what Hansen and Chemla observed. Why I could not find any contextualist effect in the case of Kettle for the within-subjects and contrastive designs, even though I found one in the between-subjects variant, remains unclear.

[15] An anonymous reviewer also suggested that the 5-point Likert scale used in my experiments might be somewhat confusing for the respondents since it is used for a true-false question, and this fact could distort the data. Although I admit it is possible, I believe that

the study were not the best possible illustrations of contextualist claims concerning the context-sensitivity of color adjectives, and maybe other scenarios would elicit intuitions more in line with contextualist predictions. The fact that my experiments found little support for these predictions does not mean that such support cannot be found in future studies: the body of empirical data is not yet very rich, so we cannot draw strong conclusions from it (this study and the experiment carried out by Hansen and Chemla are the only studies on this topic known to me). Clearly, we cannot say that my findings determined whether contextualist predictions concerning context-shifting experiments about color adjectives are supported by folk intuitions. Further studies are called for.

Let us now turn to the methodological part of my studies that was addressed in hypotheses 2 and 3. I believe that the data collected in my two experiments are good news for x-phi methodology, in particular x-phi adaptations of context-shifting experiments. It is satisfying to see that the results of the analyses conducted for the different experimental designs remain highly consistent with most of the scenarios investigated in my study. It seems that choosing a particular experimental design, whether it is within- or between-subjects, will not affect the results. However, I think we should pay more attention to one phenomenon that is often overlooked but is quite noticeable in my data — namely, pronounced differences between intuitions elicited by different scenarios. As already noticed in section 3.4, manipulating the contents of the scenario had a stronger impact on subjects' ratings than manipulating the conversational context. Some scenarios, such as Kettle and Walls, rarely encouraged negative judgments (even in the rejection context!) and nearly reached the ceiling effect. On the other hand, the responses elicited by the Leaves and Apples scenarios were noticeably more diverse, and were often negative in the acceptance context, where we did not expect negative judgments. This pattern of differences between scenarios was very consistent across the three methodological variants I employed in my experiments. Again, this is good news for x-phi methodology, but the differences between

if we used a dichotomous true-or-false scale with the same vignettes, the results would not be more favorable to contextualism. It is also worth noting that the Likert scale I used did not range from "true" to "false," but from "agree" to "disagree" (the subjects were asked about their level of agreement with a meta-linguistic statement that the utterance of the protagonist was true). This is a rather standard use of Likert scales, and I think it should not cause any confusion. Many previous x-phi studies regarding context-shifting experiments employed Likert scales. The reason for this was clearly that the researchers expected to discover rather subtle differences (as was also the case with my experiments). It seems unlikely that dichotomous scales will show a stronger effect, but of course it would be best if this issue were resolved empirically.

scenarios themselves seem to be problematic. If we believe that these four context-shifting experiments are reliable tools for measuring contextualist intuitions, we should expect them to be more homogenous in terms of the judgments they elicit.

I conducted a small-scale, informal, face-to-face survey with some laypersons and asked them how they perceive the scenarios used in my study, and how they would motivate their answers to survey questions. Interestingly, I noticed that many of my interlocutors had problems grasping the difference between contexts in the Apples scenario; they were also highly uncertain of their answers in this case. If the subjects in my quantitative studies had the same experience, this might explain some of the puzzling results I obtained for this scenario. When reacting to the Leaves scenario, some of my interlocutors were strongly opposed to the claim that a leaf painted green is *really* green: they seemed to exhibit some sort of essentialist intuition and thought that — as organic matter — a leaf has a *true* color. On the other hand, positive truth evaluations in reaction to the rejection context in Kettle and Walls (both kettles and walls are artifacts!) were sometimes motivated in the following way: "the object does not have this color if we understand it in a certain way, but it does have it in *some* way, so the utterance is true." On this basis, one might suspect that the difference between the sort of objects in question (specimens of natural kinds vs. artifacts) plays a significant role in context-shifting experiments concerning color adjectives. This, of course, is purely anecdotal evidence that cannot lead to any reliable conclusions, but I believe that experimental philosophy could benefit from the use of qualitative methods, such as interviews and focus groups. The data obtained in this way could lead to better understanding of the phenomena we observe in qualitative studies and could show us how to design research materials better so that their important philosophical features are comprehensible to laypersons participating in x-phi studies.

APPENDIX — VIGNETTES AND SURVEY QUESTIONS

APPLES

ACCEPTANCE CONTEXT

Anne and her son Mark are sorting through a barrel of assorted apples to find those that have been afflicted with a horrible fungal disease. This fungus grows out from the core, and stains the flesh of the apple red. Mark slices each apple open, and puts the good ones in a cooking pot. The bad ones he hands to Anne. He cuts open a Granny Smith apple (with green skin) that has the disease. Anne asks, “Is that one red?” and Mark says, “Yes, this one is red.”

Rejection Context
Anne and her son Mark are investigating a horrible fungal disease that afflicts apples. This fungus grows out from the core and stains the flesh of the apple red. So far, all of the apples that have been discovered with the disease have been Granny Smiths (with green skin), and they’re interested in whether any apples with red skin have the disease. Mark cuts open another Granny Smith apple that has the fungal disease. Anne asks, “Is that one red?” and Mark says “Yes, this one is red.”

Comprehension Question (Acceptance): True or False: Mark puts the good apples in a cooking pot. [True / False]

Comprehension Question (Rejection): True or False: All of the apples that have been discovered with the disease had green skin. [True / False]

Target Question: *Mark’s claim “Yes, this one is red” is true.* [5-point Likert scale ranging from ‘Disagree’ to ‘Agree’]

KETTLE

Acceptance Context
Max fills his shiny new aluminum kettle with the makings of a stew, and sets it over the campfire. An hour later, he informs Clothilde that he has done this. “That was pretty stupid,” Clothilde replies, and rushes out to the fire. She returns holding a soot-blackened pot and says, “Look. The kettle is black.”

Rejection Context
Max and Clothilde are acquiring kitchen supplies. They want only black pots. An aluminum kettle (originally silver-colored) that has been blackened by soot has come to rest in the shop window into which they are now staring. Max says, “Look. There’s a nice kettle.” Clothilde looks closer and sees that the kettle is covered in soot. “Yes. The kettle is black,” she says.

Comprehension Question (Acceptance): True or False: Max put his aluminum kettle over the campfire. [True / False]

Comprehension Question (Rejection): True or False: Max and Clothilde are looking at a soot-covered kettle in the shop window. [True / False]

Target Question: *Clothilde’s claim “The kettle is black” is true.* [5-point Likert scale ranging from “Disagree” to “Agree”]

WALLS

ACCEPTANCE CONTEXT

Hugo and Odile have a new apartment. The walls of their apartment are painted beige, but are made of white plaster. Hugo and Odile are choosing a rug that will go with the walls of their new apartment. Odile points at an orange rug and says, "What do you think of this one?" Hugo says, "I don't like it. The walls in our apartment are beige."

REJECTION CONTEXT

Hugo and Odile have a new apartment. The walls of their apartment are painted beige, but are made of white plaster. When their building was built, two sorts of walls were put in: ones made of white plaster and ones made of beige plaster. It has recently been discovered that the walls made of beige plaster give off a poisonous gas so they are being demolished and replaced. The superintendent asks Hugo to find out what sorts of walls his are. After inspecting his walls, Hugo says, "The walls in our apartment are beige."

COMPREHENSION QUESTION (ACCEPTANCE): True or False: Hugo and Odile are choosing a rug that will go with the walls of their new apartment. [True / False]

COMPREHENSION QUESTION (REJECTION): True or False: Hugo is asked to find out what the color of plaster is in their apartment. [True / False]

TARGET QUESTION: *Hugo's claim, "The walls in our apartment are beige" is true.* [5-point Likert scale ranging from "Disagree" to "Agree"]

BIBLIOGRAPHY

Bach K. (2006), "The Excluded Middle: Semantic Minimalism without Minimal Propositions," *Philosophy and Phenomenological Research* 73(2), 435-442. https://doi.org/10.1111/j.1933-1592.2006.tb00626.x

Borg E. (2007), "Minimalism versus Contextualism in Semantics" [in:] *Context-Sensitivity and Semantic Minimalism: New Essays on Semantics and Pragmatics*, G. Preyer, G. Peter (eds.), Oxford: Oxford University Press, 546-571.

Buckwalter W. (2010), "Knowledge Isn't Closed on Saturdays," *Review of Philosophy and Psychology* 1: 395-406. https://doi.org/10.1007/s13164-010-0030-3

Buckwalter W., Schaffer J. (2015), "Knowledge, Stakes, and Mistakes," *Noûs* 49(2): 201-234. https://doi.org/10.1111/nous.12017

Cohen J. (1995), *Statistical Power Analysis for the Behavioral Sciences*, 2nd. ed., Hillsdale, N.J.: Erlbaum.

DeRose K. (1992), "Contextualism and Knowledge Attributions," *Philosophy and Phenomenological Research* 52: 913-929. https://doi.org/10.2307/2107917

DeRose K. (1999), "Contextualism: An Explanation and Defense" [in:] *The Blackwell Guide to Epistemology*, J. Greco, E. Sosa (eds.), Malden, MA: Blackwell, 187-205. https://doi.org/10.1111/b.9780631202912.1998.00011.x

DeRose K. (2011), "Contextualism, Contrastivism, and X-phi Surveys," *Philosophical Studies* 156(1): 81-110. https://doi.org/10.1007/s11098-011-9799-x

Feltz A., Zarpentine C. (2010), "Do You Know More When It Matters Less?," *Philosophical Psychology* 23(5): 683-706. https://doi.org/10.1080/09515089.2010.514572

Francis K., Beaman P., Hansen N. (2019), "Stakes, Scales, and Skepticism," *Ergo* 6(16): 427-487. https://doi.org/10.3998/ergo.12405314.0006.016

Grice H. P. (1975), "Logic and Conversation" [in:] *Syntax and Semantics*, vol. 3, P. Cole, J. Morgan (eds.), New York: Academic Press.

Hansen N. (2014), "Contrasting Cases" [in:] *Advances in Experimental Epistemology*, J. Beebe (ed.), London: Bloomsbury, 71-95.

Hansen N., Chemla E. (2013), "Experimenting on Contextualism," *Mind and Language* 28(3): 286-321. https://doi.org/10.1111/j.1468-0017.2013.12019.x

Hansen N., Chemla E. (2017), "Color Adjectives, Standards, and Thresholds: An Experimental Investigation," *Linguistics and Philosophy* 40: 239-278. https://doi.org/10.1007/s10988-016-9202-7

Lewis D. (1979), "Scorekeeping in a Language Game," *Journal of Philosophical Logic* 8: 339-359. https://doi.org/10.1007/BF00258436

May J., Sinnott-Armstrong W., Hull J. G., Zimmerman A. (2010), "Practical Interests, Relevant Alternatives, and Knowledge Attributions: An Empirical Study," *Review of Philosophy and Psychology* 1: 265-273. https://doi.org/10.1007/s13164-009-0014-3

Pinillos Á. (2012), "Knowledge, Experiments, and Practical Interests" [in:] *Knowledge Ascriptions*, J. Brown, M. Gerken (eds.), Oxford: Oxford University Press,. 192-219. https://doi.org/10.1093/acprof:oso/9780199693702.003.0009

Recanati F. (2003), "Literalism and Contextualism: Some Varieties" [in:] *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, G. Preyer, G. Peter (eds.), Oxford: Clarendon Press, 171-196.

Recanati F. (2004), *Literal Meaning*, Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511615382

Recanati F. (2010), *Truth-Conditional Pragmatics*, Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226993.001.0001

Rose D., Machery E., Stich S., ..., Zhu J. (2019), "Nothing at Stake in Knowledge," *Noûs* 53: 224-247. https://doi.org/10.1111/nous.12211

Sripada C. S., Stanley J. (2012), "Empirical Tests of Interest-Relative Invariantism," *Episteme* 9(1): 3-26. https://doi.org/10.1017/epi.2011.2

Stanley J. (2005), *Knowledge and Practical Interests*, Oxford: Oxford University Press. https://doi.org/10.1093/0199288038.001.0001

Travis C. (1997), "Pragmatics" [in:] *A Companion to the Philosophy of Language*, B. Hale, C. Wright (eds.), Oxford: Blackwell, 87-107.

Wittgenstein L. (1953), *The Philosophical Investigations*, Oxford: Blackwell.

Ziółkowski A. (2017), "Experimenting on Contextualism: Between-Subjects vs. Within-Subjects," *Teorema: International Journal of Philosophy* 36(3): 139-162.