

EXPERIMENTING ON CONTEXTUALISM

Nat Hansen, Emmanuel Chemla

March 1, 2012

To appear in Mind & Language

Abstract

In this paper we refine the design of *context shifting experiments*, which play a central role in contextualist debates, and we subject a large number of scenarios involving different types of expressions of interest to contextualists, including ‘know’ and color adjectives like ‘green’, to experimental investigation. Our experiment (i) reveals an effect of changing contexts on the evaluation of uses of the sentences that we examine, thereby overturning the absence of results reported in previous experimental studies (so-called *null results*), (ii) uncovers evidence for a ‘truth bias’ in favor of positive over negative sentences, and (iii) reveals previously unnoticed distinctions between the strength of the contextual effects displayed by scenarios involving knowledge ascriptions and for scenarios concerning color and other miscellaneous scenarios.

Word count: 15,202

1 Introduction

1.1 Overview

This paper concerns the central method of generating evidence in support of contextualist theories, what we call *context shifting experiments*. We begin by explaining the standard design of context shifting experiments, which are used in both quantitative surveys and more traditional thought experiments to show how context affects the content of natural language expressions (§1.2). We discuss some recent experimental studies that have tried and failed to find evidence that confirms contextualist predictions about the results of context shifting experiments (§1.3), and consider the criticisms of those studies made by DeRose (2011) (§1.4). We show that DeRose’s criticisms are incomplete, and we argue that the design of context shifting experiments he proposes is itself subject to some of the

We wish to thank Aidan Gray, Chauncey Maher, Eliot Michaelson, Daniel Rothschild, Tim Sundell, members of the philosophy department at Umeå University, and the organizers and participants of the conference on ‘Meaning, Context and Implicit Content’ in Cerisy, Normandy, in June 2011 for comments on this material. Thanks to two anonymous reviewers for this journal for providing very helpful remarks. And special thanks to François Recanati and Marie Guillot for proposing and organizing the colloquium on experimental work in semantics and pragmatics that prompted our collaboration on this project. The research leading to these results has received funding from the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 229 441 - CCC.

Address for correspondence: Nat Hansen, Institutionen för idé- och samhällsstudier, Umeå Universitet, SE-901 87, Umeå, Sweden.

Email: nathaniel.hansen@gmail.com

same problems as the studies he criticizes. We propose a refined approach to the design of context shifting experiments that addresses these problems and which allows us to investigate the effect of context on both positive and negative sentences. This aspect of our design allows us to control for several forms of bias, including a particular form of ‘truth bias’ that favors positive over negative sentences (§2). We then deploy our improved design in an experiment that tests a large number of scenarios involving different types of expressions of interest to contextualists, including the verb ‘know’ and color adjectives like ‘green’ (§3). Our experiment (i) reveals an effect of changing contexts on the evaluation of sentences in all scenarios we examined, thereby overturning the absence of results reported in previous experimental studies (so-called *null results*) and (ii) reveals previously unnoticed distinctions between the strength of the contextual effects we observed for scenarios involving knowledge ascriptions and for scenarios concerning color, as well as other miscellaneous scenarios (§4). We conclude by discussing the importance of basic features of experimental design for both quantitative surveys and thought experiments, and consider possible objections to our approach (§5).

1.2 Context Shifting Experiments

Many expressions in natural language shift their content in different contexts.¹ Uncontroversial examples of context sensitive expressions include first person pronouns such as ‘I’ and adverbs such as ‘here’ and ‘now’, which shift their contents in different contexts, depending on who is speaking, and where and when the utterance takes place, respectively. The scope of context sensitivity and how best to explain it are controversial topics. Sometimes the controversy is intensified when it concerns whether philosophically significant expressions like ‘know’ or ‘wrong’ are context-sensitive, and acknowledging the context-sensitivity of these expressions is alleged to help resolve classic problems in epistemology or ethics.

There are different techniques that can be used to generate evidence that particular expressions are context sensitive, but perhaps the most widely used involves constructing *context shifting arguments* (Cappelen and Lepore 2005). It is helpful to think of a context shifting argument as consisting of two parts: (i) a *context shifting experiment*, which elicits intuitions about uses of an expression *e* in different imagined contexts, and (ii) an argument that the best way to explain the intuitions generated in response to the experiment involves semantic features of *e*.

The following story, due to Charles Travis (1997), illustrates the standard structure of context shifting experiments. The story involves the leaves of a Japanese maple that have been painted green, a context (C1) in which someone is decorating, a second context (C2) in which a botanist is looking for leaves to use in a study of green leaf chemistry, and two utterances of the target sentence ‘The leaves are green’, one in each context:

A story. Pia’s Japanese maple is full of russet leaves. [She paints them green ‘for a decora-

¹Hansen (forthcoming) also investigates the design of context shifting experiments. That investigation relies on existing experimental data about the reliability of judgments about affirmative and negative sentences from Wason (1961), while the present paper generates and analyzes new experimental data. The earlier paper shares some of the material presented in this section.

tion’].² Returning, she reports, ‘That’s better. The leaves are green now’. She speaks truth. A botanist friend then phones, seeking green leaves for a study of green-leaf chemistry. ‘The leaves (on my tree) are green’, Pia says. ‘You can have those’. But now Pia speaks falsehood.

Travis’s intuitions about the painted leaves scenario are represented in Table 1.

	C1 <i>Decorator</i>	C2 <i>Botanist</i>
‘The leaves are green’	TRUE	FALSE

Table 1: Travis’s Intuitions about the *Painted Leaves* Scenario

There is an extensive debate about how best to explain the intuitions elicited by context shifting experiments like Travis’s painted leaves scenario. (Competing explanations of the painted leaves scenario can be found in Hansen 2011, Kennedy and McNally 2010, Predelli 2005, Rothschild and Segal 2009, Sainsbury 2001 and Szabó 2000, 2001). However, attention has recently turned to examining the methods by which the intuitions are elicited by context shifting experiments in the first place. Experimental surveys have failed to reproduce contextualists’ fundamental intuitions about certain prominent context shifting experiments (see Buckwalter 2010, DeRose 2011, and Schaffer and Knobe 2011 for discussion). Our own effort is part of this line of empirical research with a methodological focus. We will not enter into debates about which explanation of the data is best. Instead, we are interested in the soundness of the data that the theoretical debate is based on, and the methods used to generate and analyze that data.

In the recent investigation of how data is generated in the contextualist debate, attention has focused on context shifting experiments that involve epistemologically significant expressions, like ‘know’. In particular, versions of DeRose’s (1992, 2009) well known ‘bank’ scenario have received the most attention. DeRose’s bank scenario has an interestingly different design than the standard design of context shifting experiments: rather than asking for intuitions about uses of a single sentence in two different contexts, he asks for intuitions about the use of a sentence in one context, and the *negation* of the sentence in another context. In DeRose’s bank scenario, for example, we are first asked to evaluate the truth value of an utterance of ‘I know the bank will be open on Saturday’ in a low stakes context C1 where no possibilities of error are mentioned (‘Low’), and then we are asked to evaluate the truth value of an utterance of ‘I don’t know the bank will be open on Saturday’ in a high stakes context C2 where a possibility of error is mentioned (‘High’). Here is DeRose’s bank scenario:

Bank Case A. My wife and I are driving home on a Friday afternoon. We plan to stop at the bank on the way home to deposit our paychecks. But as we drive past the bank, we notice that

²This additional remark is from the version of the thought experiment that appears in Travis (1994, p. 172).

the lines inside are very long, as they often are on Friday afternoons. Although we generally like to deposit our paychecks as soon as possible, it is not especially important in this case that they be deposited right away, so I suggest that we drive straight home and deposit our paychecks on Saturday morning. My wife says, ‘Maybe the bank won’t be open tomorrow. Lots of banks are closed on Saturdays’. I reply, ‘No, I know it’ll be open. I was just there two weeks ago on Saturday. It’s open until noon.’ [The bank is open on Saturday.]

Bank Case B. My wife and I drive past the bank on a Friday afternoon, as in Case A, and notice the long lines. I again suggest that we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon. But in this case, we have just written a very large and very important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a *very* bad situation. And, of course, the bank is not open on Sunday. My wife reminds me of these facts. She then says, ‘Do you know the bank will be open tomorrow?’ Remaining as confident as I was before that the bank will be open then, still, I reply, ‘Well, no, I don’t know. I’d better go in and make sure’. [The bank is open on Saturday.]

DeRose offers the following intuitions about his bank scenario:

[...] It seems to me that (1) when I claim to know that the bank will be open on Saturday in Case A, I am saying something true. But it also seems that (2) I am saying something *true* in Case B when I say that I *don’t* know that the bank will be open on Saturday.

DeRose’s intuitions about his bank scenario are represented in Table 2. As with Travis’s

	C1 <i>Low</i>	C2 <i>High</i>
‘I know... ...the bank will be open on Saturday’	TRUE	
‘I don’t know... ...the bank will be open on Saturday’		TRUE

Table 2: DeRose’s Intuitions about the *Bank* Scenario

painted leaves scenario, there has been a great deal of debate over how best to explain the intuitions elicited by DeRose’s bank scenario.³ Because our topic is the design of the context shifting experiments that provide the empirical foundation for those debates, we will not comment on any of those competing explanations here.

1.3 Consensus Lost

While there is widespread disagreement about how best to explain the data generated by context shifting experiments, for a long time there has been equally widespread agreement about the data itself. DeRose (2011, p. 82) summarizes this situation as follows:

³See the papers collected in Part 1 of Preyer and Peter (2005) for a sample of the relevant debates.

A fairly extensive and robust debate has been raging in the professional philosophy journals for a while now, with almost all participants being at least largely in agreement about what the key intuitions are that should somehow be addressed, but disagreeing about how best to handle them.

But, as DeRose observes, the consensus about data has recently been challenged by experimental philosophers (see Schaffer and Knobe 2010 and the papers discussed therein, especially Buckwalter 2010). In existing studies, the intuitions reported by ordinary speakers in response to context shifting experiments have not confirmed the expectations and intuitions reported by contextualists.

If we consider the standard design of a context shifting experiment, depicted in Table 1, a contextualist should predict that the responses of ordinary speakers generally line up with the intuitions reported by contextualists themselves (indicated on Table 1 by TRUE and FALSE). In a quantitative study of the responses of ordinary speakers, in which many responses are analyzed, a contextualist should predict responses to the standard design of context shifting experiments that look something like the pattern represented in Table 3 (p. 5), where the length of the bar represents, e.g., the proportion of those participants who respond to the use of the sentence with the response TRUE, or an average of some scores given on a scale for which the highest end is ‘truth’.

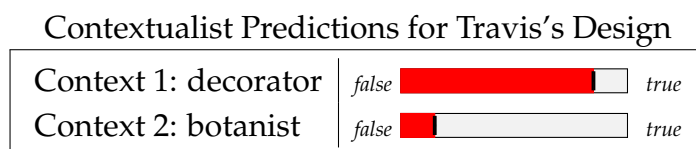


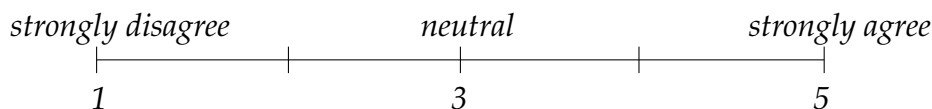
Table 3: Contextualist Prediction for Travis’s Design. Responses that are expected to be ‘true’ are represented as long red bars reaching towards the right end, while responses that are expected to be ‘false’ are represented with short red bars. Note that this is just a dummy chart; it does not report any actual results. And note also that we represent a slight deviance from the pure contextualist prediction (100% in the decorator context and 0% in the botanist context), because various factors contribute to the production of noise in a quantitative survey.

Buckwalter (2010) designed an experiment that evaluated the responses of ordinary speakers to the use of the sentence ‘I know the bank will be open on Saturday’. Buckwalter’s experiment used the design of standard context shifting experiments (employed by Travis in the painted leaves scenario, and depicted in Table 1). As we will explain in detail below, Buckwalter did not find the pattern of different responses to the two contexts that is represented in Table 3, and he argues that his results pose a challenge to contextualism about knowledge ascriptions.

Buckwalter (p. 401) asked subjects to perform the following task with regard to versions of the bank scenario in which uses of ‘I know the bank will be open on Saturday’ are evaluated in low-stakes or high-stakes contexts, and (separately) contexts with or without mentioned possibilities of error:

On a scale of 1 to 5, circle how much you agree or disagree that [DeRose’s] assertion, ‘I know the bank will be open on Saturday’ is true.

His prompt was accompanied by a scale with the following structure :



Buckwalter's survey found no statistically significant difference between the number of participants who 'agree' with the assertion (that is, those who circle either 4 or 5 on the scale shown above) when it concerned the low-stakes and high-stakes contexts, or the contexts with or without mentioned possibilities of error, though the means in all contexts were substantially above the midpoint. In Table 4 (p. 6), Buckwalter's results are compared with what he takes to be the contextualist prediction for the high and low stakes contexts and for contexts in which there is no mention of a possibility of error (a 'low standard' context) and those in which there is a mentioned possibility of error (a 'high standard' context).

Buckwalter (2010)'s results

	Contextualist prediction	Buckwalter's results (% of 4 and 5 responses)	
		<i>Stakes</i>	<i>Error</i>
Low			
High			

Table 4: Buckwalter (2010)'s results compared to the contextualist predictions (repeated from Table 3).

What is the significance of Buckwalter's finding? Buckwalter says '[I]n the particular bank cases tested we have reason to doubt the contextualist hypothesis' (p. 403), where the contextualist hypothesis is the prediction that ordinary speakers will generally have intuitions that correspond with the contextualists' intuitions about the sentences used in 'Low' and 'High' contexts. Indeed, at first glance it might appear that the contextualist prediction about responses to the bank scenarios (using the standard design of context shifting experiments) is disconfirmed by Buckwalter's finding: The contextualist predicts that there will be a significant change between evaluations of uses of 'I know the bank will be open on Saturday' across the 'Low' and 'High' (stakes and standards) contexts, whereas Buckwalter found no such change in evaluations.

But what Buckwalter has in fact found is a *null result*: he did not find a statistically significant difference between evaluations of 'I know the bank is open on Saturday' in 'High' and 'Low' (stakes and standards) contexts. Null results are generally considered to be inconclusive, rather than as showing that two variables are unrelated. Roughly speaking, that is because there are many reasons why a study may fail to uncover a relation between variables even when the relation does in fact obtain. One may be relying on instruments that do not have the necessary degree of resolution to detect the relevant relation, for example. And every experimental result is noisy to some degree. An absence

of difference cannot establish that the difference does not exist, unless one also proves the counterfactual claim that the experiment would have been sufficiently powerful to detect it.

An example may help clarify what is going on behind the scenes with a null-result. Consider the following experimental data: a coin thrown 100 times came up heads 53 times and tails 47 times. Should we conclude that the coin is fair or not? The answer is that it is hard to tell. The application of a standard statistical test to that data would produce the same (inconclusive) result. Standard statistical results take the following form: Under the assumption that the coin is fair, the probability of finding that data (or any more extremely unbalanced data) is $p = .62$.⁴ The phrase 'under the assumption that the coin is fair' in the previous sentence introduces the so-called *null-hypothesis*, which is necessary to compute probabilities (if we know that a coin is fair, we can compute the odds of any outcome). From this .62 probability, we can only infer that the outcome is compatible with the coin being fair.

A p -value below .05 is conventionally taken as evidence against the hypothesis that the coin is fair. Indeed, such a p -value would indicate that the result would have been very unlikely if the hypothesis (that the coin is fair) was correct. In other words, it indicates that the result and the hypothesis are incompatible. It is wrong to believe that a p -value above this 5% conventional threshold is evidence *for* the hypothesis that the coin is fair, because such a result merely indicates that the data is compatible with the hypothesis, and thus does not lead to any strong conclusion about the fairness of the coin.

There may seem to be an arbitrary asymmetry between proving that the coin is fair and proving that it is not. But the asymmetry is not arbitrary, and in fact it is essential to conducting the statistical analysis of the data. There are two reasons that conspire to make the asymmetry: (i) low p -values lead to the rejection of a null-hypothesis, while high p -values do not lead to validation of a null-hypothesis, and (ii) null-hypotheses must be designed so that probabilities can be computed (while it is possible to compute probabilities if we know that the coin is fair, it is not possible to compute probabilities if we know that the coin is not fair). Standard statistical tests have this asymmetrical form. In particular, this is the case with tests used to investigate possible differences between two conditions, such as the tests reported in Buckwalter's study. Failure to find a significant statistical difference between two conditions cannot be used as evidence for the sameness of the two conditions.

1.4 DeRose's Replies and His Recommended Experimental Design

Keith DeRose (2011) replies to Buckwalter's study differently. He does not question the significance of Buckwalter's (null) result. Rather, he argues that the design of Buckwalter's experiment is flawed in two respects and that the results generated by the flawed experiment could not threaten (DeRose's particular variety of) contextualism. DeRose then spells out his favored design for context shifting experiments, whether they are conducted as thought experiments or quantitative surveys, and whether they are meant to

⁴This is the result of a two-tailed binomial test.

generate evidence for his version of contextualism or competing versions.

DeRose's first criticism of Buckwalter's experiment is that it is a mistake to separate *stakes* from *possibilities of error* when testing contextualism. Unlike competing theories like subject-sensitive invariantism (Stanley 2005) or contrastivism (Schaffer 2004), DeRose's generic version of contextualism is not committed to any predictions about which of those two contextual factors is responsible for shifting truth conditions, or whether it is some interaction between the two contextual factors that accounts for the observed variation in truth conditions (DeRose, 2011, pp. 89–90). Since Buckwalter does not test a situation in which both the stakes are high and a relevant possibility of error is mentioned, he has not shown that ordinary speakers' intuitions about the bank case diverge from those reported by DeRose.

Ultimately, one should try to determine the respective contribution of stakes and relevant possibilities of error to intuitions about knowledge ascriptions. DeRose's first criticism does not rule out the fact that Buckwalter's results have a significance for contextualist debates, but only that they do not threaten DeRose's particular, generic version of contextualism which does not tease apart the effects of stakes and mentioned possibilities of error. In short, Buckwalter's results do not address the form of contextualism that makes the weakest explanatory claim and therefore do not offer the strongest challenge to contextualism.

The second criticism of Buckwalter's design made by DeRose concerns the *polarity* of the sentences used in the imagined scenarios. As in the standard design of context shifting experiments, Buckwalter asks participants how much they agree or disagree that DeRose's assertion 'I know the bank will be open on Saturday' is true in contexts that vary in terms of stakes and mentioned possibilities of error. That is, Buckwalter asks participants to evaluate uses of a sentence of *positive* polarity in different contexts. DeRose thinks that this aspect of Buckwalter's design (and hence also the standard design of context shifting experiments) is flawed.

Why is it flawed? DeRose (2011, p. 88) says that 'there is pressure on us as interpreters of the ascription ["I know that the bank will be open on Saturday"] to understand it as having a content that makes it true, due to the operation of what David Lewis calls a "rule of accommodation"'. According to DeRose, the rule of accommodation puts pressure on participants to find the ascription 'I know the bank will be open on Saturday' true in both the 'Low' and 'High' contexts, so the contrast between intuitions about 'Low' and 'High' contexts that contextualists expect to find would be reduced or eliminated. Participants would tend to find uses of the positive sentence true in *both* 'Low' and 'High' contexts. The rule of accommodation would therefore obscure the effect of context on truth conditions in the standard design of context shifting experiments.

With the rule of accommodation in mind, DeRose recommends a different design for context shifting experiments. The schematic representation of the bank scenario given in Table 2 (p. 4) captures DeRose's basic idea: instead of evaluating uses of a single positive sentence in different contexts, one should evaluate a use of a sentence with *positive* polarity in the 'Low' context, and a use of a sentence with *negative* polarity in the 'High'

context. DeRose’s intuitions about those uses of the positive and negative sentences are that both say something true. A contextualist employing DeRose’s design should predict that responses from ordinary speakers would come out as represented in Table 5 (p. 9).



Positive sentence – Context 1: Low	false		true
Negative sentence – Context 2: High	false		true

Table 5: Contextualist predictions for DeRose’s design. See visual conventions in Table 3.

We will refer to the different possible combinations of sentence polarity and context as ‘cells’ in the context shifting experiment: the *Positive–Low* cell, the *Positive–High* cell, the *Negative–Low* cell, and the *Negative–High* cell. Intuitions are produced in response to particular cells (combinations of uses of sentences with positive or negative polarity and particular contexts).

		C1 <i>Low</i>	C2 <i>High</i>
<i>Positive</i>	‘I know... ...the bank will be open on Saturday’	TRUE	TRUE
<i>Negative</i>	‘I don’t know... ...the bank will be open on Saturday’	?	TRUE

Table 6: DeRose’s Intuitions in the *Bank Scenario* for three ‘cells’, i.e. 3 combinations of context (Low or High) and polarity of the target sentence (positive or negative).

DeRose’s intuitions about three cells in the bank scenario are indicated in Table 6. The first row in Table 6 represents the standard design of context shifting experiments employed by Buckwalter to test the bank scenario. DeRose’s remarks about the rule of accommodation indicate that he thinks speakers will find uses of the positive sentence true in both ‘Low’ and ‘High’ contexts. The *diagonal* composed of the cells *Positive–Low* and *Negative–High* (in bold in Table 6) represents DeRose’s recommended design for context shifting experiments.

Two features of DeRose’s design are problematic: First, anyone who adopted DeRose’s design to use in a quantitative survey would be aiming to produce a particular *null result*—contextualists using this design would hope to find no significant difference between participants’ evaluations of the *Positive–Low* and *Negative–High* cells (they would expect responses to both cells to be true). But as we discussed above, there are many practical reasons why a particular design could fail to detect a difference that in fact exists, e.g., the resolution of the instruments one is using may be insufficient to detect the relevant relation. Absence of evidence is not evidence of absence. For this reason, the sound practice is to design experiments that aim to show the existence of some difference, and to remain cautious about drawing conclusions if that difference fails to show

up in one's results. Furthermore, if one explains Buckwalter's 'flat' *true-true* result—his finding that there is no statistically significant difference between evaluations in 'High' and 'Low' (stakes and standards) contexts—by appealing to the rule of accommodation, then this rule of accommodation may very well explain any other similar *true-true* flat result.

Second, whereas the standard design employed by Buckwalter holds the target sentence fixed and varies the context in which the sentence is used, DeRose's design simultaneously varies *both* the target sentence used *and* the context in which the sentence is used. That will make it difficult to identify whether it is the change in context or the polarity of the sentence used that is responsible for the intuitions elicited by each cell.

One cell in Table 6 is conspicuously empty: the *Negative-Low* cell. When we developed this study, we weren't aware of anyone (other than us) who had reflected on what the intuitive response to this cell would be and on what its significance would be for the debate over contextualism.⁵ We think that context shifting experiments that elicit responses to all four cells are an important improvement to contextualist experimental methodology. In the sections that follow, we will show how the data for this neglected cell can help clarify experimental results potentially affected by the rule of accommodation.

2 Designing Context Shifting Experiments

We can make substantial improvements to the existing design of context shifting experiments as they are used in both quantitative surveys and thought experiments.

2.1 Testing All Four Cells

DeRose's disagreement with Buckwalter concerns which cells in context shifting experiments are the most productive to test. But it is important to test *all* of the cells, including the previously neglected *Negative-Low* cell. There are a couple of reasons for preferring this inclusive approach.

By investigating all of the cells, our design embeds both Buckwalter's and DeRose's preferred designs. We will thus be able to ask whether the shift in context affects intuitions about the truth value of positive sentences, as in Buckwalter's design, and also evaluate DeRose's prediction that responses to the *Positive-Low* cell and the *Negative-High* cell will both tend to be true. But notice that contextualists should also predict that shifting the context from 'Low' to 'High' should affect negative sentences in the exact opposite way that it affects their positive counterparts. The negative sentence data will thus provide an immediate replication of the positive sentence part of the experiment. If everything goes as expected, the two results should go in opposite directions. That outcome would also show that the result obtained is not simply due to a greater tendency to find sentences true in context C1 than in context C2, but that the difference is tied to the actual sentences tested. This is a standard control precaution employed in experimental psychology, which guards against participants giving superficial, strategic responses.

⁵Daniel Rothschild has since brought it to our attention that Buckwalter (Ms.) reports the results of an unpublished study of the bank scenario that collects responses to all four cells.

Looking at the data from a different angle, we will also be able to evaluate and factor out some of the effect that the rule of accommodation may have on participants' responses. Indeed, if we find that a sentence and its negation are judged equally true (in the same context), this will be evidence in favor of a bias towards TRUE answers. In our results section, (§4), we will show how we can factor out the contribution of the rule of accommodation from the remaining genuine effect of changing contexts.

2.2 Block Design

Some philosophers (see, e.g., Neta and Phelan ms) have argued that whether or not participants are exposed to contrasting cases (between 'Low' and 'High' cells, for example) makes a significant difference to how subjects respond to those cells. Buckwalter's design does *not* allow any form of contrast, because participants were only asked about a single cell. But the original formulation of the 'bank' scenario *does* involve a contrast between *Positive-Low* and *Negative-High* contexts—those reading DeRose's original examples see both cells in succession. An improved design would make it possible to assess the effect of contrast by comparing intuitions at the beginning of the experiment that have not had the chance to be affected by contrast with intuitions that are reported later, when contrast has the opportunity to take effect.

We designed an experiment that makes such an assessment possible, using a multiple 'block' design that allowed us to isolate intuitions reported during the beginning of the experimental task that are not plausibly subject to any contrast effects and compare those intuitions with those reported later in the experiment, when contrast effects could conceivably be present. The implementation of this multiple 'block' design will be described in detail in the following Experimental Setup and Results sections (§3.4 and §4.3).

2.3 Comparing Knowledge, Color and Miscellaneous Scenarios

In addition to cases of knowledge ascription, which have received the most attention in the experimental literature on contextualism, our experiment presented participants with context shifting experiments involving *color adjectives* (like Travis's painted leaves case, described in §1.1 above) and other *miscellaneous* scenarios (involving sentences about *weight* attribution and the presence or absence of some relevant quantity of *milk* in a refrigerator). By gathering data about responses to these different kinds of expressions, it is possible to observe previously overlooked differences between responses to different kinds of context shifting experiments. The results of this comparison are discussed below (see §4.2.2).

3 Experimental setup

3.1 Participants

We recruited 40 participants over Amazon Mechanical Turk for \$2 each (see Sprouse 2011 for discussion of the reliability of the Mechanical Turk as a data gathering tool). One participant reported that he was a native speaker of Spanish and was excluded from subsequent analyses. The 39 participants included in the analyses reported to be native speakers of English.

3.2 Task

Participants were asked to read a series of stories. For each story, they were asked to assess the truth-value of some character's claim appearing in boldface, given the context offered in the rest of the story. They were instructed that their judgment may be subtle and were given the flexibility to provide their answers within a continuous range of options between FALSE and TRUE, by setting the right end of a red line between these two extreme anchors (see Fig. 1 and discussion in §5.2.1). Answers were coded as the percentage of the red line filled in red, 100% corresponding to an unambiguous TRUE response, and 0% to an unambiguous FALSE response.



Figure 1: Response Scale. Participants were offered the possibility to situate their responses within a range of possibilities between FALSE and TRUE, as above. Responses were coded as the percentage of the red line filled in red, 100% corresponding to an unambiguous TRUE response, and 0% to an unambiguous FALSE response. In the left example above, the answer would be around 5%, in the right example around 75%.

3.3 Material and Design

The short stories we presented were constructed from examples discussed in the contextualist literature. We altered these examples systematically to obtain the four cells we argued are needed for an optimal design (see details below). We will call a set of 4 such related stories a 'scenario'. The bank case discussed in the introduction provides an example of two cells of such a scenario, and the four stories extracted from the bank scenario are given explicitly in Fig. 2 (p. 14).

Our 10 main scenarios were inspired by context shifting experiments that target different types of expressions. There were 4 *knowledge* scenarios (involving a potential shift in intuitions about first-person *knowledge* ascriptions), 4 *color* scenarios (involving a potential shift in intuitions about statements concerning the *color* of some object), and 2 additional scenarios labeled as *miscellaneous*.⁶ We also added one *control* scenario, in which we varied the context in ways which should uncontroversially alter the truth value of the target statement in order to check that participants were performing the task competently. See Fig. 2 for an example of a scenario and appendix A for details about all the scenarios we used.

For each of these 11 scenarios, we constructed 4 short stories by manipulating two factors: *polarity* and *context*. The first factor, *polarity*, concerned the target sentence which

⁶Knowledge scenarios were based on DeRose's (1992, 2009) bank scenario, Feltz and Zarpentine's (2010) truck scenario, Fantl and McGrath's (2002) train scenario, and Pinillos's (forthcoming) spelling scenario. Color scenarios were based on Travis's (1994, 1997) painted leaves scenario, Travis's (1985a) black kettle scenario, Travis's (1989) beige walls scenario, and Bezuidenhout's (2002) red apple scenario. The miscellaneous scenarios were based on Travis's (1989) milk scenario, and Travis's (1985b) weighing 80 kilograms scenario. See the appendix for details of the scenarios used in the study.

	'Low'	'High'
Positive	TRUE	FALSE
Negative	FALSE	TRUE

Positive	Low	
	High	
Negative	Low	
	High	

Table 7: Contextualism's Predicted Responses in table and then in pseudo-chart version.

was either positive or negative (e.g., 'I know that p' vs. 'I don't know that p').⁷ The second factor, context, concerned the rest of the story. Each scenario came in two different versions corresponding to two types of contexts: 'Low' and 'High'. If our context shifting experiments were to confirm contextualist predictions, participants should judge the target sentences differently in the 'High' and 'Low' contexts. In the knowledge scenarios, for instance, the difference between 'High' and 'Low' contexts consisted in manipulating sentences in the story that expressed different stakes and mentioned possibilities of error. The contextualist prediction is that the positive target sentence in a given scenario should be judged 'more true' in 'Low' than in 'High' contexts. The same distribution of responses should be expected for the Color, Miscellaneous and Control scenarios as well. The labels 'Low' and 'High' are applied to the color, miscellaneous, and control scenarios even though there is nothing in the contexts involved in those scenarios that corresponds directly to the stakes or mentioned possibilities of error in the knowledge ascription cases. In the non-knowledge ascription scenarios the labels track contextualist predictions for particular cells: *Positive-Low* should be judged 'more true' than *Positive-High*, and *Negative-Low* should be judged 'less true' than *Negative-High*.

To sum up, we constructed four different renderings of each of 11 scenarios. Contextualists predict a contrast between responses to the different cells that would follow the pattern schematized in Table 7. This pattern of results is also the one expected, independently of any contextualist commitments, for the control scenario.

3.4 Presentation of the Stories: Different Blocks

Each participant had to judge each cell of each scenario, resulting in 44 total judgments for each participant in the complete experiment. These items were organized in four consecutive blocks. Each block was constructed so as to contain only one cell of a given scenario, and so that each of the four cells (high/low, positive/negative) would not be exemplified by two different knowledge scenarios, two color scenarios or two miscellaneous scenarios in a given block. Within each block, the items were presented in random order to each participant and the different blocks were also shuffled.

This complex constraint on the presentation of the scenarios has two advantages. First, it maintains a relatively stable proportion of positive and negative sentences and true and false expected answers in any local part of the experiment. Second and most importantly,

⁷In the control scenario, there was no explicit negation. The sentences were 'You are quite tall!' (positive) and 'You are quite short!' (negative).

Sylvie and Bruno are driving home from work on a Friday afternoon. They plan to stop at the bank to deposit their paychecks, but as they drive past the bank they notice that the lines inside are very long.

{ Low: Although they generally like to deposit their paychecks as soon as possible, it is not especially important in this case that they be deposited right away.
 High: Bruno and Sylvie have just written a very large check, and if the money from their pay is not deposited by Monday, it will bounce, leaving them in a very bad situation with their creditors. And, of course, the bank is not open on Sunday.

Bruno suggests that they drive straight home and return to deposit their paychecks on Saturday morning. He remembers driving by last Saturday and seeing that it was open until noon.

{ Low: Sylvie says, 'Maybe the bank won't be open tomorrow. Lots of banks are closed on Saturdays. On the other hand, shops are often open on Saturdays in this neighborhood. ...'
 High: Sylvie reminds Bruno of how important it is to deposit the check before Monday and says, 'Banks are typically closed on Saturday. Maybe this bank won't be open tomorrow either. Banks can always change their hours, I remember that this bank used to have different hours. ...'

Do you know the bank will be open tomorrow?'

{ Positive: **Bruno replies, 'I know the bank will be open tomorrow'.**
 Negative: **Bruno replies, 'Well, no, I don't know the bank will be open tomorrow. I'd better go in and make sure'.**

It turns out that the bank is open on Saturday.

Figure 2: Example of a knowledge scenario, indicating all relevant differences between 'Low' v. 'High' contexts, and 'Positive' v. 'Negative' sentences. In this example, the first Low/High branching introduces the contrast between low and high stakes, while the second Low/High branching introduces the contrast in terms of mentioned possibility of error.

it was designed so that by extracting the results of the participants from the first block only, we would obtain results in which all scenarios in all conditions would be seen, but no single participant would have seen more than one cell of a given scenario. We report the results from the first block as 'local results' (see §4.3), in contrast with 'global results' that include results from all blocks. The block design was not transparent to participants; they saw only an apparently random sequence of stories, with stories only ever appearing one at a time on the screen.

4 Results

In this section, we analyze the data generated by the experiment, which leads to three main results.

- First, the control results are as expected, which suggests that participants are per-

		Block A	Block B	Block C	Block D
Knowledge scenarios:	\oplus -Low:	BANK	TRUCK	TRAIN	SPELLING
	\oplus -High:	SPELLING	TRAIN	TRUCK	BANK
	\ominus -Low:	TRAIN	SPELLING	BANK	TRUCK
	\ominus -High:	TRUCK	BANK	SPELLING	TRAIN
Color scenarios:	\oplus -Low:	LEAVES	KETTLE	WALLS	APPLES
	\oplus -High:	APPLES	LEAVES	KETTLE	WALLS
	\ominus -Low:	WALLS	APPLES	LEAVES	KETTLE
	\ominus -High:	KETTLES	WALLS	APPLES	LEAVES
Misc. scenarios:	\oplus -Low:	MILK	WEIGHT		
	\oplus -High:			MILK	WEIGHT
	\ominus -Low:		MILK	WEIGHT	
	\ominus -High:	WEIGHT			MILK
Control:		\oplus -Low	\oplus -High	\ominus -Low	\ominus -High

Figure 3: Block design, constructed (mostly) to test for contrast effects. This figure summarizes the constraint on the order of presentation of the different stories. In each box of four lines, the lines from top to bottom correspond to the cells *Positive-Low*, *Positive-High*, *Negative-Low*, *Negative-High*. Hence, in Block A, the BANK scenario, the LEAVES scenario and the MILK scenario appeared in their *Positive-Low* guise. There were two levels of randomization across participants. First, the order of the blocks was random: different participants received different blocks first, second, third and last. Second, the order of presentation of each story was randomized within each block (crossing the different *types* of scenarios). In visual terms, this means that columns were first shuffled around, and participants would first see all the stories appearing in the first column. The stories of this first column would be seen in a random order, the stories of the second column would be seen in a random order again, and so on for the other columns.

forming the task appropriately (§4.1).

- Second, all our context shifting experiments give rise to statistically significant differences in the responses of participants to the uses of sentences in different contexts, although the strength of this effect is weaker for the knowledge scenarios than for the color and miscellaneous scenarios (§4.2).
- Third, we will focus on what we call ‘local’ results, corresponding to the first block of judgments in which all scenarios and conditions are exemplified, but in which no single participant sees the same scenario in two different conditions. We will show that in this first block without ‘contrast’, the contextualist effect disappears in the knowledge scenarios, although it remains strong in the other scenarios (§4.3).

4.1 The Control Scenario

Figure 4 (p. 16) shows the results for the control scenario. This scenario was included to ensure that participants were performing the judgment task appropriately. For example, the *Positive–Low* cell of the control scenario was the following:

Bill and Jane are at a huge speed dating party. Both Jane and Bill are very shy. Bill is 7 feet tall, but no one seems to notice him. Jane is a bit lonely and bored, but suddenly she faces Bill. She looks at him for a moment and suddenly says ‘**You are quite tall!**’

A response of ‘true’ is uncontroversially expected for this control story, independently of any contextualist or other theoretical commitments. The other control stories were equally uncontroversial. They were created by exchanging ‘tall’ with ‘short’ to obtain the negative cells, and ‘7 feet tall’ with ‘5 feet tall’ to obtain the ‘High’ contexts. Note that in the control scenario, there was no explicit negation in the ‘negative’ target sentences.⁸ And the titles ‘Low’ and ‘High’ for the contexts used in the control scenario do not indicate anything about stakes or mentioned possibilities of error—they merely serve as labels indicating what the predicted judgments about the use of the sentences in these contexts are. Like the knowledge, color, and miscellaneous scenarios, the prediction is that the ‘positive’ sentence will be judged true in the ‘Low’ context, and false in the ‘High’ context, and vice versa for the ‘negative’ sentence.

The results are as expected (see predictions in Table 7, p. 13). For example, we expected participants to judge the target sentence in the *Positive–Low* cell of the control as true, and this is what the long red line in the first line of Fig. 4 confirms. These results are worth examining in detail though because they provide a clear visual representation of what the results for the target cases should look like according to the contextualist predictions, and some acquaintance with the kind of analyses needed for those cases as well.

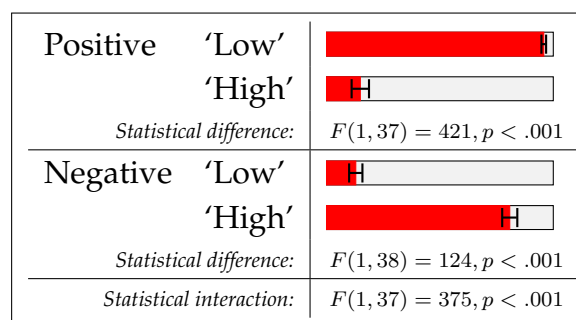


Figure 4: Mean results for the control scenario. Long red lines correspond to true responses, short red lines correspond to false responses. Concretely, the position of the right end of the red line corresponds to the average position of the responses given by the participants between the FALSE/left side and TRUE/right side anchors.

⁸One justification for the use of this ‘positive’ and ‘negative’ terminology in the control case is that ‘tall’ and ‘short’ are polar antonyms, with ‘tall’ being the positive member of the pair and ‘short’ the negative member. See Kennedy and McNally (2005) for a characterization of polar antonyms.

Focusing first on positive sentences (the two bars on top), one sees that the ‘Low’ context gives rise to a greater mean, i.e. TRUE responses, than the ‘High’ context. This difference is statistically significant: $F(1, 37) = 421, p < .001$.^{9,10} Importantly, the difference is reversed for negative sentences (the difference is also significant: $F(1, 38) = 124, p < .001$). In fact, the reversal itself can be assessed by a statistical test, an ANOVA, which reveals that there is a significant so-called ‘interaction’ between the two factors (context and polarity): $F(1, 37) = 375, p < .001$. This last result is the most important. It means that the differences found for positive and negative sentences are different. In other words, the two top rows receive high and low averages, the bottom rows show the opposite pattern: low and high averages. Visually, this corresponds to a ‘<’ shape: the top and bottom lines are judged high, reaching the right of the chart, while the two central lines are judged low, reaching the left of the chart.

This analysis also uncovers a significant main effect of polarity ($F(1, 37) = 10, p < .005$), meaning that, overall, positive sentences were judged ‘more true’ than negative sentences. This reveals a form of bias for the positive (‘tall’) sentences. The fact that we looked at the whole square of conditions (all four cells in the context shifting experiment) enables us to quantify this bias. More importantly, the type of analyses we rely on are designed so that such biases do not get in the way of the main effect we are interested in. If we had looked at two cells only, any difference or absence of difference between these two cells could have been attributed to a superficial bias. For instance, we may have found a preference for sentence S in context C1 over context C2, but this could have been the result of some orthogonal difference between the contexts independent of S. For example, one of the contexts might have been less exciting, or more difficult to memorize or parse, and such differences may introduce biases that could account for the preference.

But we also examined another sentence in these same contexts. For this other sentence, ‘not-S’, we expected and found the opposite preference: not-S is more acceptable in C2 than in C1. Such a full 2×2 pattern of results cannot be attributed to general features of the contexts. Instead, such a pattern has to be explained by the interaction between the context and the target sentence. Similarly, we could have looked at a design *à la* DeRose, and found a preference for S in C1 over not-S in C2. If we found such a result, it would be even harder to interpret because it could be driven either by a preference for C1 over C2, or by a preference for S over not-S (as with the bias for positive sentences we found throughout the experiment). But the 2×2 picture allows us to trace the effect we observe to the interaction of context and target sentence.

⁹The F -values we report in such formulae are an intermediate step towards the p -values, which receive a simpler interpretation. A p -value is an approximation of the probability to find a distribution of answers similar to the one obtained in the experiment in a situation where there would actually be no difference between the conditions. In other words, we hope to find low p -values, typically below .05. A low p -value says that the difference obtained is unlikely to be due to noise or chance, and is more likely due to some real effect that is likely to be replicated in another experiment (e.g., with other participants from the same population), and thus indicates an effect that calls for a substantial explanation.

¹⁰In the whole experiment, 16 data points were lost because of internet connection problems (.85% of the data). These include one of the conditions of this control scenario for one subject who was therefore lost for this within-subject analysis. Hence, ‘37’ degree of freedom in that particular case.

This first set of control results are unexciting, but they (i) offer a clear example of how to look at the data generated by our experiment on the contextualist scenarios and (ii) indicate that at some level or other participants were performing the task correctly, albeit with a bias towards finding some sentences ‘more true’ than others, independent of context.

4.2 Global Results

In this section, we report the results gathered in the experiment as a whole. Our analyses reveal clear contextual effects of various degrees of strength across all types of scenarios tested.

4.2.1 Contextual Effect. The mean results for the target scenarios are presented in Fig. 5. The relevant statistical figures are given along with the charts and show that, although the differences are smaller than for the control scenarios discussed above, they remain statistically significant both for the positive and the negative sentences. As discussed above, the most important result is that the interaction between polarity and context is statistically significant for all types of scenarios (knowledge, color and miscellaneous). This interaction corresponds to the ‘<’ shape that we see in the charts. This result shows that context reverses the favored sentence from positive to negative (independently of possible biases).

	Knowledge	Color	Miscellaneous
⊕ Low			
High			
Statistical diff.	$F(1, 38) = 24, p < .001$	$F(1, 38) = 41, p < .001$	$F(1, 38) = 55, p < .001$
⊖ Low			
High			
Statistical diff.	$F(1, 38) = 4.6, p = .05$	$F(1, 38) = 38, p < .001$	$F(1, 38) = 38, p < .001$
Stat. interaction	$F(1, 38) = 17, p < .001$	$F(1, 38) = 49, p < .001$	$F(1, 38) = 61, p < .001$

Figure 5: Mean results for the knowledge scenarios, the color scenarios and the miscellaneous scenarios.

In Fig. 6, we report the results for all individual scenarios. The interested reader can check that the results were more or less stable across the different versions of the knowledge, color and miscellaneous scenarios (that is, the patterns of results remain the same).

4.2.2 Effect: Variable Strength. Another striking fact that emerges from these results is that the contextual effect is weaker for the knowledge scenarios than for the color and miscellaneous scenarios. This is reflected by the fact that the interaction between types of scenarios (knowledge, color, miscellaneous), polarity and context is significant: $F(1, 76) = 26, p < .001$. The restricted interaction of types of scenarios and context is also significant both for the positive and negative sentences.

	Bank	Truck	Train	Spelling
⊕Low				
High				
⊖Low				
High				

Color scenarios

	Leaves	Kettle	Walls	Apples
⊕Low				
High				
⊖Low				
High				

Miscellaneous scenarios

	Milk	Weight
⊕Low		
High		
⊖Low		
High		

Figure 6: Mean results for each scenario.

4.2.3 Rule of Accommodation vs. ‘Truth Bias’ for Positive Sentences. Another effect we uncovered concerns different evaluations of positive and negative sentences. Positive sentences are overall judged higher (‘more true’) than negative sentences (main effect of polarity: $F(1, 38) = 4.2, p < .05$).¹¹ This effect suggests that participants’ answers

¹¹An anonymous reviewer mentioned an inherent asymmetry in responses to positive and negative sentences that may contribute to the difference we observe. Negative words can be used to signal agreement in response to negative sentences, as in the following discourse: A: ‘John is not at home’. B: ‘No (he’s not).’ This type of effect may contribute to a more superficial explanation of the negative bias we found for negative sentences: Responses to negative sentences may be ranked artificially *low* on the response scale because some people who *agree* with the negative sentence might signal their agreement with it by responding ‘no’ (‘false’). If that were the case, we would indeed see negative sentences rated lower on the scale, which we should not confuse with a sign that participants are actually rejecting the claim made by those sentences. We see two reasons why this ‘no’ as agreement fact cannot explain our data. First, this effect should be counterbalanced by a similar effect in the opposite direction, namely the fact that positive answers to negative sentences can be used to mark *disagreement*: A: ‘John is not at home’. B: ‘Yes (he is).’ Hence, the potential artificially low ranking of responses to negative sentences should be cancelled out by a similar artificial boost in responses to positive sentences, so neither of these effects would end up having visible consequences. Second, and more importantly, we did not use ‘yes’ and ‘no’ as response options, but rather ‘true’ and ‘false’ (anchoring the ends of the response scale). These responses do not lead to the same ambiguity, as shown by the (in)coherence of the following possible responses to ‘John is not at home’:

were influenced by issues that were orthogonal to our main question of whether specific changes in context affect participants' intuitions about the truth of what is said. However, our design allows us to isolate this effect from the effects of changing context. We did not uncover evidence in support of the rule of accommodation as formulated by DeRose, namely a general preference for a TRUE response, and it is not clear to us how one would discover such an effect, given that it is supposed to apply to all uses of sentences. We return to this issue in the general discussion of our results (§5.1.4).

4.3 Local Results

In this section, we report the results gathered from the first block of the experiment. The analyses reveal that some of the contextual effects are not present in this early part of the experiment, in which it is not plausible that participants are subject to contrast effects in their responses.

The mean results for the target scenarios in the first block (cf. §3.4) are presented in Fig. 7. The responses represented here were given by participants when they would see a scenario (e.g., the train scenario) for the first time, irrespective of which of the four possible cells of the scenario that appeared. Given our design, this data therefore includes one judgment for each of the four conditions (polarity \times context) for the knowledge and for the color scenarios, and one judgment for two of the four conditions for the miscellaneous scenarios. (See §3.4 and Fig. 3 for details.)

The relevant statistical figures are given along with the charts. They show that the effect of context is intact for the color and miscellaneous scenarios, but has now disappeared for the knowledge scenarios (we do not find the '<' -shaped pattern that we found in the global results for all scenarios; see Fig. 5 for '<' -shaped results). This replicates Buckwalter's (2010) null-result in an experimental setting that more closely resembles his original design, because it lacks a within-participant contrast between different conditions of the same scenario (although Buckwalter's design was more extreme in this respect and only included one judgment per participant).¹²

5 Discussion

5.1 Summary of Results

In summary, our results are the following:

1. *Global Contextual Effects*: We found clear contextual effects across all the types of scenarios tested (knowledge, color and miscellaneous).
2. *Distinctions between Types of Scenarios*: We uncovered previously unnoticed distinctions between types of scenarios of interest to contextualists. The contextual effects

-
- | | |
|----------------------------|------------------------|
| (i) a. * False, he is not. | (ii) a. * True, he is. |
| b. False, he is. | b. True, he is not. |

¹²Notice that we reproduce Buckwalter's absence of contextual effect even while accommodating DeRose's recommendation not to separate stakes and mentioned possibilities of error. That recommendation was designed to reinforce the contextual effect and make it visible, but it did not have that effect in the first block participants encountered.

		Knowledge	Color	Miscellaneous
⊕	Low			
	High			
<i>Statistical diff.</i>		$F(1, 37) = 1.4, p = .24$	$F(1, 38) = 42, p < .001$	$F(1, 31) = 4.8, p < .05$
⊖	Low			
	High			
<i>Statistical diff.</i>		$F(1, 36) = .75, p = .39$	$F(1, 36) = 6.7, p < .05$	$F(1, 30) = 20, p < .001$
<i>Stat. interaction</i>		$F(1, 35) = .05, p = .82$	$F(1, 36) = 50, p < .001$	$F(1, 61) = 22, p < .001$

Figure 7: Mean ‘local’ results for the knowledge scenarios, the color scenarios and the miscellaneous scenarios. (We report between-subject statistics for the miscellaneous scenarios because not all participants got a scenario of each relevant condition in the first block).

were weaker for knowledge scenarios than for color and miscellaneous scenarios, and the contextual effect for knowledge scenarios (but not for color or miscellaneous scenarios) disappeared when only the first block of responses was considered, where it is not plausible that there were contrast effects.

3. *Against DeRose’s Design*: DeRose’s proposed design of context shifting experiments was found to be flawed because it predicts a null result and because it varies both the context of use and the polarity of the sentence used without the means to isolate the effects of either factor.
4. *Rule of Accommodation vs. ‘Truth Bias’ for Positive Sentences*: We were also able to investigate the role of the alleged rule of accommodation, which we did not find evidence of. Instead, we found evidence of a ‘truth bias’ favoring positive over negative sentences.

5.1.1 Global Contextual Effects. Our discovery of global contextual effects across all the scenarios tested is a response to the growing sense of skepticism (discussed in Buckwalter 2010, Schaffer and Knobe 2010 and DeRose 2011) about the intuitions reported by contextualists that have long served as the empirical foundation of debates over the merits of contextualism and competing theories.

5.1.2 Distinctions between Types of Scenarios. Why do the knowledge scenarios display a weaker contextualist effect than the color and miscellaneous scenarios? Answering that question is a topic for further research, though there is a straightforward difference between the knowledge scenarios and the color and miscellaneous scenarios that might account for the difference in strength of the contextualist effect: The knowledge scenarios are simply *longer* (and hence more complex) than the color and miscellaneous scenarios. In addition, all but one of the color scenarios and both of the miscellaneous scenarios are based on context shifting experiments written by Charles Travis. Travis is particularly good at writing short scenarios that make the contrast between contexts especially vivid

and memorable, while the knowledge scenarios tend to be bland and forgettable—it is possible that those stylistic differences play a role in the relative strength of the responses the scenarios generate.¹³ Whatever explains the differences between the strength of the global results for the knowledge scenarios and the color and miscellaneous scenarios, it is interesting that the knowledge scenarios, which have received the most attention in the experimental literature on contextualism, turn out to be the hardest scenarios for the demonstration of contextual effects.

Another difference we observed between the knowledge scenarios and the color and miscellaneous scenarios is that while we reproduced the Buckwalter null result in the first ‘block’ of knowledge cases, we did find a significant result as participants got more familiar with the task (or were exposed to contrast effects), but that was *not* the case with the color and miscellaneous scenarios. In the color and miscellaneous cases, people showed the contextual effect (between ‘Low’ and ‘High’ contexts) right from the start, in the first block of results. Is perceiving a contrast between ‘High’ and ‘Low’ contexts necessary to bring about the contextualist effect in the knowledge scenarios? Or is it simply that the knowledge scenarios are relatively complex and participants need some familiarity with the task before they can make a competent judgment? Both of these explanations are live options.

5.1.3 Against DeRose’s Design. We confirmed DeRose’s prediction that speakers would find both ‘I know that p’ in the ‘Low’ context and ‘I don’t know that p’ in the ‘High’ context true. But we also showed how DeRose’s proposed design is flawed. First, when designing an experiment, one aims to isolate a particular variable to see whether it is having an effect. That is what the standard design of context shifting thought experiments does, by holding the sentence fixed, and varying the context in which it is used. DeRose varies both the context *and* the sentence, so his design does not put him in a position to identify the change in context as the factor that is influencing participants’ responses. Second, if used in a quantitative survey, DeRose’s design would predict a *null result*, namely that there is no difference in (true) responses to the *Positive–Low* and *Negative–High* cells. Third, DeRose has never asked for intuitions about the *Negative–Low* cell. We designed our experiment to include that neglected cell. Including that cell allows us to see how the polarity of the sentence interacts with the context, an effect which DeRose’s design obscures.

5.1.4 Rule of Accommodation vs. ‘Truth Bias’ for Positive Sentences. While we showed that there is evidence of a form of ‘truth bias’ for positive sentences in participants’ judgments, we did not find evidence of DeRose’s rule of accommodation. DeRose’s prediction that subjects would find both the use of a sentence and a use of the sentence’s negation

¹³There is another feature of the knowledge scenarios that is missing from the color and miscellaneous scenarios. Because knowledge is factive, the knowledge scenarios must include a statement that the fact that the knowledge ascription concerns actually obtains. So, for example, in the ‘bank’ scenario, the scenario concludes with a statement that the bank is in fact open on Saturday. This statement comes last and right after the knowledge ascription. The presence of this (partially) confirming statement might contribute to the bias in favor of positive over negative sentences in the knowledge scenarios.

true in the same context seems to be confirmed by the results for the knowledge scenarios, given that participants judged both positive and negative sentences ‘true-ish’ in the same context.¹⁴ But only the knowledge scenarios, and not the color or miscellaneous scenarios, show that pattern. That casts doubt on the existence of the principle of accommodation as DeRose understands it. As mentioned above, we are curious as to how one might go about verifying the effect of the rule of accommodation as DeRose conceives it, given that it is supposed to affect all uses of sentences.

5.2 Possible Objections

5.2.1 The Task. All methods of eliciting responses to linguistic experiments, whether they employ a binary true/false judgment task, or a Likert scale with labelled points, or the continuous true/false scale we employed, play a role in shaping the responses participants give. For example, a binary true/false judgment task demands that participants make sharp judgments, even when their responses may in fact be much more nuanced. That could obscure interesting differences between participants’ responses to scenarios. For example, judgments about the color scenarios are more clear-cut than are judgments about the knowledge scenarios, a fact which might not emerge if one were using a binary true/false judgment task. And no type of response corresponds directly to the binary, TRUE/FALSE (or 1/0) outputs of semantic theory, even those elicited by a binary true/false judgment task. Semantic theory has to be combined with theories of how participants will perform in response to particular experimental material and in response to particular kinds of tasks before predictions about actual participants’ responses are possible.

We have encountered some specific objections to the task that we asked participants to perform, which involves setting a value on a continuous scale from FALSE to TRUE, that go beyond those general issues about using experimental data as evidence for or against particular semantic and pragmatic theories. In this section, we consider those objections and offer our replies.

Objection: ‘By allowing for graded judgments, the continuous scale task seems to encourage participants to reinterpret “true” and “false”. It’s like asking about bachelorhood using a graded scale. You would likely find that people are happy to move well beyond the 50% mark on the scale if a character has many typical characteristics of bachelors but fails a defining criterion. (For example, a person who married to gain citizenship but does not and has never lived with his legal spouse.)’¹⁵

Reply: First, it is not clear that asking for responses to be placed on a scale from ‘false’ to ‘true’ encourages participants to *reinterpret* ‘false’ and ‘true’, because participants may be bringing a gradable understanding of ‘true’ and ‘false’ to the experiment. That is, the binary understanding of truth and falsity at work in standard semantic theories is theoretical concept; it is a further question whether ordinary speakers attach such an understanding to ‘true’ and ‘false’, or some other understanding. The ease with which

¹⁴Note however that ‘true-ish’ does not mean true, because the 50% middle may not correspond to an actual frontier between true and false.

¹⁵Thanks to an anonymous reviewer for articulating this objection.

participants operate with the continuous scale true/false task may, as the objection suggests, be evidence that they are reinterpreting ‘true’ and ‘false’, or it may be evidence that they do not understand ‘true’ and ‘false’ in binary terms to begin with. Which of those possibilities is the case is a topic for further research.

Second, the fact that we reproduced existing results in similar conditions (in the first ‘block’ of our experiment) using the continuous scale task is some evidence that we are uncovering the same phenomena as experiments that used more traditional response tasks.

Objection: ‘I can imagine many ways in which participants could reinterpret “true” and “false”. For example, they might reinterpret them in terms of whether they would agree with the statement in the sense of supporting it against any denials of it. Alternatively they might respond on the basis of whether they would be prepared to say the same thing in the same circumstances. Neither of these notions necessarily tracks the semanticist’s notions’.¹⁶

Reply: We agree that there are many possible ways that participants may be reinterpreting ‘true’ and ‘false’ as they appear on the scale. But the same observation holds as well for other types of response task. It is not obvious that participants are interpreting ‘true’ and ‘false’ as the semanticist uses those terms, even when prompted with a binary true/false judgment task. In fact, we think that the use of binary truth value judgments in response tasks can create the misleading impression that the responses of ordinary speakers do bear directly on the outputs of semantic theory.¹⁷

Objection: It has been suggested in discussion that the continuous task is mysterious or ambiguous for participants. For instance, it is not specified whether the midpoint on the red bar task represents some unidentified unacceptability of the use of the sentence or some hesitation about how to respond to the task.

Reply: We believe that this *a priori* worry about our design does not cast doubt on the actual interpretation of the results obtained. This is because we do not interpret any single datapoint in isolation, e.g., saying that an ‘80%’ answer means truth, or that ‘50%’ means uncertainty or intermediate judgment. Instead, we focused our attention and interpretation on *contrasts* that emerge between conditions whether or not there is any ambiguity involved in the intermediate part of the scale. In the worst case, even if there were such an ambiguity, it may introduce noise in the data, but it would not produce artificial contrasts that would challenge our interpretations. And there are two additional reasons for being satisfied with the present task. First, in our experience, participants reported being extremely comfortable with the continuous task for semantic and pragmatic judgments (Chemla 2009a,b, Chemla and Spector 2011, Chemla and George 2011, Chemla and Schlenker 2012). Second, in these previous studies and in the present case, if we ‘binarize’ our results by classifying participants’ answers as TRUE when their response was above the midpoint and FALSE otherwise (even excluding responses that fall within a gray zone

¹⁶Thanks again to an anonymous reviewer for raising this objection.

¹⁷A similar point has been raised about attempts to generate evidence for or against particular theories of the technical notion of *what is said* by asking experimental participants to judge ‘what is said’ by various uses of sentences. See Bach (2002) for a convincing criticism of such an attempt.

around the midpoint), we find exactly the same contrasts. In fact, the main difference between the continuous results and the 'binarized' results is that the statistical tests are less powerful when applied to the 'binarized' version.

We don't think that the objections that have been offered to the continuous scale true/false task problematize the results of our study. But we would welcome a version of our study that uses a binary truth value judgment task or a more traditional Likert scale in place of the continuous judgment task. Our view is that none of these tasks are less problematic than the continuous scale task that we employed. Only an improved understanding of how ordinary speakers access and report semantic phenomena would help us decide which task is optimal. We also want to emphasize that the response task we used is a detachable component of the design of context shifting experiments that we recommend in this paper.

5.2.2 The Scenarios. We have discussed the specifics of some of our scenarios in section 5.1.2. For instance, we mentioned that the knowledge scenarios were shorter than the color and miscellaneous scenarios, and we suggested that this superficial difference could affect participants' responses in significant ways. Here we would like to mention a specific worry about the wording used in our version of the bank scenario.¹⁸ In our version of the bank case, in the *Low* contexts, Sylvie says 'Lots of banks are closed on Saturdays', while in the *High* contexts, she says 'Banks are typically closed on Saturdays'. It is arguable that 'Lots of banks are closed on Saturdays' and 'Banks are typically closed on Saturdays' express different probabilities of banks being closed on Saturdays. If that's right, then the *Low* and *High* contexts in the bank scenario differ in more ways than just in terms of stakes and mentioned possibilities of error. That may produce a difference in response to the bank scenario that is not due to one of the controversial factors the contextualist is interested in, but due to variation in an uncontroversially epistemically relevant feature of the context. So, whereas we claim to find a global contextual effect in the knowledge scenarios, that result may be undermined by the possibility that the contextual effect is produced by variation in mentioned probabilities, which is not one of the controversial factors that contextualists are interested in.

We aimed to construct the scenarios using systematic recipes: we always started with the actual scenarios as they appear in the literature, then we tried to minimize the differences between different cells on the one hand while adding minimal variations to obtain new cells on the other hand. But it is possible that there are additional unintended differences between cells like the one discussed above. In the worst case, these differences would align with the predictions of contextualism and provide an alternative explanation for what otherwise seems to be a contextual effect. (Note that if the differences work *against* the predictions of contextualism, they actually strengthen the conclusion for a contextualist effect when it is found, so only a specific kind of difference is worrisome). We would like to encourage the reader to look at the scenarios (see appendix), and to screen for features that they think might affect the interpretation of the effects that we describe. The reader can then refer to the actual results for the corresponding scenario in Fig. 6

¹⁸This worry was raised in discussion by Andy Egan and Jeff Pretti.

(p. 19) and check to make sure that a scenario that may seem biased is not the one that is driving the overall effect. Applying this recommendation to the bank scenario, for example, it is apparent that it displays the weakest contextual effect of the four knowledge scenarios, so it is not driving the overall contextual effect on display in the knowledge scenarios.

Moreover, it is unlikely that each of our scenarios is affected by some unanticipated differences of the kind described above that would jeopardize our interpretation. Given that the contextual effect is present for all scenarios, and roughly to the same degree of strength within groups of scenarios, it is likely that this effect is due to the main manipulation, and is not due to a bunch of specific, unforeseen issues polluting *each* of the scenarios.¹⁹

And though our revisions to the scenarios used in the literature may introduce unintended differences between contexts, we also uncovered and eliminated some previously unnoticed confounding differences between contexts. For example, in DeRose's original bank scenario, the 'Low' and 'High' contexts differ not only in terms of stakes and mentioned possibilities of error, but also in terms of *where* and *how* the protagonist is credited with his evidence that the bank is open on Saturday (see p. 3 for the original 'bank' scenario). In the 'Low' context, the protagonist gives his evidence in direct speech at the end of the story ("I was just there two weeks ago on Saturday. It's open until noon"), but in the 'High' context, the protagonist's evidence is given in indirect speech near the beginning of the story ('... explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon'). Locating the statement of evidence in direct speech at the end of the story in the 'Low' context makes it more salient than placing it in indirect discourse at the beginning of the 'High' scenario, which may affect responses in the two contexts. So, overall, we do not think that whatever extraneous differences our modifications introduce (if any) are more problematic than the extraneous differences in existing scenarios, and we made significant improvements to existing scenarios by eliminating extraneous contrasts between contexts.

6 Concluding Remarks

In one sense, the fact that we found evidence for contextual effects in all the scenarios we tested is not surprising. That is because there was widespread agreement that intuitions elicited by context shifting experiments indicated a contextual effect before the recent surveys that found no such effect. Our results, like the earlier armchair intuitions that long served as the empirical foundation of the contextualist debate, do not tip the balance in favor of contextualism over its theoretical competitors. They simply confirm that, in line with contextualist predictions, responses to target sentences are affected by changes in context.

¹⁹In statistical terms, this discussion would rely on a per item analysis (instead or on top of the per subject analysis that we offer). We did not run the per item analysis because we did not have many items and would thus find it unreliable. We could not include more items because it would have made the experiment very long, and we were more interested in testing all the relevant conditions within items and participants and across blocks. We trust however that the result of a per item analysis would go in the right direction.

In addition to responding to skeptics about the intuitions reported by contextualists, we hope that our discussion will guide others in the design of rigorous and efficient experiments. The methodological lessons of our investigation apply not only to the design of context shifting experiments in quantitative surveys, but also to the design of traditional first-personal thought experiments as well. For example, whether one is developing a quantitative survey or a thought experiment, differences between contexts other than those differences that one is trying to evaluate the effects of should be systematically eliminated. And one should be aware that the responses that one elicits are shaped in part by different kinds of tasks, and that one may uncover differences by using a more finely-grained task (like setting a value on a continuous scale) that would be obscured by the traditional absolute truth value judgment task.

Nat Hansen
Institutionen för idé- och samhällsstudier
Umeå Universitet

Emmanuel Chemla
Laboratoire de Sciences Cognitives et Psycholinguistique
École Normale Supérieure

A Experimental material

This appendix provides the details of the stories we used in our experiment.²⁰ Instead of listing each story separately (4 minimally different stories for each scenario), we present the scenarios schematically to highlight the relevant differences between cells (that is, between cells featuring ‘High’ and ‘Low’ contexts, and positive and negative sentences).²¹ The convention for reading these schematic scenarios is straightforward: non-bracketed material is constant across all cells; bracketed material is labeled according to which cells it occurs in. Of course, participants in the study encountered non-schematic versions of these scenarios, presented as a single paragraph, as in the following example of the Positive–Low cell from the color scenario presented in §A.2.3:

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**’.

The construction of these scenarios was constrained by two requirements. We wanted to be able to hold the target sentence fixed across ‘Low’ and ‘High’ contexts, and we also

²⁰For a list of the original versions of the scenarios on which these are based, see footnote 6.

²¹In the knowledge scenarios we tested, we systematically manipulated stakes and mentioned possibilities of error by changing different sentences in the stories. These changes are marked as, e.g., ‘Low (stakes)’ and ‘Low (error)’. In the other scenarios, the relevant change is also indicated beside ‘Low’ and ‘High’, e.g., ‘Low (Decorator)’ and ‘High (Botanist)’ in the painted leaves scenario (§A.2.1).

wanted to be able to hold the context fixed as much as possible while varying the polarity of the target sentence. Consequently, we often used prompting questions ('do you know p?') to make room for both positive and negative target sentences after the same text. Other adjustments between the different cells were also necessary; all differences between cells are explicit in the schematic representations below.

A.1 Knowledge cases

A.1.1 Bank

See Fig. 2, p. 14.

A.1.2 Truck

John is a passenger in a truck that is part of a convoy of trucks driving along a dirt road. His co-worker Jim is driving. They come to what looks like a rickety wooden bridge...

- { Low (stakes): ... over a three foot ditch.
- { High (stakes): ... over a yawning thousand foot drop.

They stop and John radios ahead to find out whether the other trucks in the convoy have made it safely over. He is told that all 15 trucks in the caravan made it over without a problem. John reasons that if they made it over, ...

- { Low (error): ... his truck will probably make it over as well.
- { High (error): ..., he will probably make it over as well, but he wonders whether the other trucks might have weakened the bridge enough that he won't make it across.

- { Low Positive: So, **he says to Jim,**
- { Low Negative: Still, **he says to Jim,**
- { High Positive: Still, **he says to Jim,**
- { High Negative: **He says to Jim,**

- { Positive: **'I know that our truck will make it across the bridge.'**
- { Negative: **'I don't know that our truck will make it across the bridge.** We should find another way across.' Jim doesn't listen to John and simply drives ahead.

Their truck does make it across the bridge.

A.1.3 Train

Sid and Johnny are at Back Bay Station in Boston preparing to take the commuter rail to Providence.

- { Low (stakes): They can take their time getting there—they're going to see friends. It will be a relaxing vacation.
- { High (stakes): They absolutely need to be in Providence, the sooner the better. Their career depends on it.

As the train rolls into the station, Sid asks Johnny, 'Do you know if this train is the express, or does it make all those little stops in Foxboro, Attleboro, etc.?'

- { Low (stakes): It doesn't matter much to the two of them whether the train is the express or not, though they'd mildly prefer that it was, since then they'd get to Providence sooner.
- { High (stakes): 'If it does, we'll miss our meeting in Providence'.

Johnny tells Sid that he remembers that the person who sold him the ticket said it was the express.

- | | | |
|---|---------------|--|
| { | Low (error): | Positive: Then he says to Sid,
Negative: But Johnny says, |
| | High (error): | But then Sid says, 'Maybe the ticket-seller misunderstood your question. Maybe you misunderstood the answer. I don't want to be wrong about this.' Johnny says, |
| { | Positive: | 'I know that it is the express.' |
| | Negative: | 'I don't know that it is the express. I'd better go and make sure.' |

It turns out that the train is the express.

A.1.4 Spelling

John,²² a good college student, has just finished writing a two-page paper for an English class. The paper is due tomorrow.

- | | | |
|---|----------------|---|
| { | Low (stakes): | The teacher is just asking for a rough draft and it won't matter if there are a few spelling mistakes. Nonetheless Peter would like to have no spelling mistakes at all. |
| | High (stakes): | There is a lot at stake. The teacher is a stickler and guarantees that no one will get an A for the paper if it has a spelling mistake. He demands perfection. John, however, finds himself in an unusual circumstance. He needs an A for this paper to get an A in the class. And he needs an A in the class to keep his scholarship. Without the scholarship, he can't stay in school. Leaving college would be devastating for John and his family who have sacrificed a lot to help John through school. So it turns out that it is extremely important for John that there are no spelling mistakes in this paper. And he is well aware of this. |

Even though John is a pretty good speller, he has a dictionary with him and he has checked the paper once to make sure it doesn't have any mistakes.

- | | | |
|---|---------------|---|
| { | Low (error): | ∅ |
| | High (error): | Before he hands it in, John's roommate reminds him that he might have missed some mistakes when he checked the paper. |

John²³ thinks to himself,

- | | | |
|---|-----------|---|
| { | Negative: | 'I won't hand this paper in yet. I don't know that everything on it is spelled correctly'. |
| | Positive: | 'I will hand this paper in. I know that everything on it is spelled correctly'. |

It turns out that everything in John's paper is spelled correctly.

A.2 Color cases

A.2.1 Leaves

Pia has a Japanese maple tree in her backyard that has russet (reddish brown) leaves. She paints the leaves of the tree green.

- | | | |
|---|------------------|--|
| { | Low (Decorator): | 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. |
| | High(Chemistry): | 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. |

²²Following the original Pinillos scenario, the name of the student was 'John' for the stories in the High cells (negative and positive), and 'Peter' for the stories in the Low cells. We reproduce them both here with 'John'.

²³We had 'He' instead of the proper name in the *Low-Positive* cell.

- { Positive: **'Yes, the leaves on my tree are green,' Pia says.**
 { Negative: **'No, the leaves on my tree aren't green,' Pia says.**

A.2.2 Kettle

- { Low (Camping): Max fills his shiny new aluminum kettle with the makings of a stew, and sets it over the campfire. An hour later, he informs Sam that he has done this. 'That was pretty stupid', Sam replies, and rushes out to the fire.
 { High(Shopping): Everard 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. Everard says, 'Look. There's a nice kettle'. Clothilde looks closer and sees that the kettle is covered in soot.
 { Low Positive: He returns holding a soot-blackened pot and says, **'Look. The kettle is black.'**
 { Negative: He returns holding the soot-blackened kettle and says, **'Look. The kettle isn't black.'**
 { High Positive: **'Yes, the kettle is black', she says.**
 { Negative: **'No, the kettle isn't black', she says.**

A.2.3 Walls

Hugo and Odile have a new apartment. The walls of their apartment are painted beige, but are made of white plaster.

- { Low (Rug): 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. ...'
 { High(Gas): 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 poison 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,
 { Positive: **'The walls in our apartment are beige.'**
 { Negative: **'The walls in our apartment aren't beige.'**

A.2.4 Apples

{ Low (Flesh): Anne and her son are sorting through a barrel of assorted apples to find those that have been afflicted with a horrible fungal disease.
 { High(Skin): Anne and her son are investigating a horrible fungal disease that afflicts apples. This fungus grows out from the core and stains the flesh of the apple red.

- { Low (Flesh): Anne's son 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.
 { High(Skin): 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. Anne's son cuts open another Granny Smith apple that has the fungal disease.
 { Positive: Anne asks, 'Is that one red?' and **he says 'Yes, this one is red'.**
 { Negative: Anne asks, 'Is that one red?' and **he says 'No, this one isn't red'.**

A.3 Miscellaneous cases

A.3.1 Milk

- Low (Cleaning): Hugo has been given the task of cleaning the refrigerator. He has just changed out of his house-cleaning garb, and is settling with satisfaction into his arm-chair, book and beverage in hand.
- High(Coffee): Hugo is seated at the breakfast table, reading the paper. He prefers his coffee with milk. From time to time he looks dejectedly (but meaningfully) at his cup of black coffee, which he is idly stirring with a spoon.

The refrigerator is devoid of milk except for a puddle of milk at the bottom of it.

- Low (Cleaning): Odile opens the refrigerator, looks in, closes it and **says to Hugo,**
- High(Coffee): **Odile says to Hugo,**
- Positive: **'There is milk in the refrigerator'.**
- Negative: **'There isn't milk in the refrigerator'.**

A.3.2 Weight

80 kilograms is Hugo's recommended weight. One morning, after months of dieting, he steps on the scale and it reads 80 kilograms. Later in the day, heavily dressed in winter clothes but without having eaten anything, he is such that if he stepped on a scale, it would register 84 kilograms.

- Low (Diet): While wearing his heavy winter clothes, Hugo wants to announce the progress of his diet, and **he says**
- High(Bridge): Hugo is out exploring the countryside while wearing his heavy winter clothes. He comes to a trestle bridge across a deep ravine. A sign says that the bridge is quite delicate and can bear only 80 kilograms or less. **Hugo says to himself,**
- Positive: **'I weigh 80 kilograms'.**
- Negative: **'I don't weigh 80 kilograms'.**

A.4 Control Cases

Bill and Jane are at a huge speed dating party. Both Jane and Bill are very shy.

- Low (tall): Bill is 7 feet tall, but no one seems to notice him.
- High(short): Bill is 5 feet tall, but no one seems to notice him.

Jane is a bit lonely and bored, but suddenly she faces Bill. She looks at him for a moment and suddenly says

- Positive (tall): **'You are quite tall!'**
- Negative(short): **'You are quite short!'**

References

- Bach, K. (2002). Seemingly semantic intuitions. In J. K. Campbell, M. O'Rourke, and D. Shier (Eds.), *Meaning and Truth*, pp. 21–33. New York: Seven Bridges Press.
- Bezuidenhout, A. (2002). Truth conditional pragmatics. *Noûs* 36(16), 105–134.
- Buckwalter, W. (2010). Knowledge isn't closed on Saturdays. *Review of Philosophy and Psychology* 1(3), 395–406.
- Buckwalter, W. (2011). Further experimental work on the bank cases. Unpublished ms.
- Cappelen, H. and E. Lepore (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

- Chemla, E. (2009a). Presuppositions of quantified sentences: experimental data. *Natural Language Semantics* 17(4), 299–340.
- Chemla, E. (2009b). Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics* 2(2), 1–33.
- Chemla, E. and B. R. George (2011). Let's agree which. Ms. LSCP, UCLA.
- Chemla, E. and P. Schlenker (2012). Incremental vs. symmetric accounts of presupposition projection: an experimental approach. *Natural Language Semantics*.
- Chemla, E. and B. Spector (2011). Experimental evidence for embedded implicatures. *Journal of Semantics* 28(3), 359–400.
- DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research* LII(4), 913–929.
- DeRose, K. (2009). *The Case for Contextualism*. Oxford: Oxford University Press.
- DeRose, K. (2011). Contextualism, contrastivism, and x-phi surveys. *Philosophical Studies* 156(1), 81–110.
- Fantl, J. and M. McGrath (2002). Evidence, pragmatics, and justification. *The Philosophical Review* 111(1), 67–94.
- Feltz, A. and C. Zarpentine (2010). Do you know more when it matters less? *Philosophical Psychology* 23(5), 683–706.
- Hansen, N. (2011). Color adjectives and radical contextualism. *Linguistics and Philosophy* 34(3), 201–221.
- Hansen, N. (2012). On an alleged truth/falsity asymmetry in context shifting experiments. forthcoming in *Philosophical Quarterly*.
- Kennedy, C. and L. McNally (2005). Scale structure and the semantic typology of gradable predicates. *Language* 81(2), 345–381.
- Kennedy, C. and L. McNally (2010). Color, context, and compositionality. *Synthese* 174(1), 79–98.
- Neta, R. and M. Phelan (ms). Evidence that stakes don't matter for evidence. Unpublished ms.
- Pinillos, N. (forthcoming). Knowledge, experiments and practical interests. forthcoming in *New Essays On Knowledge Ascriptions* (Eds. Jessica Brown and Mikkel Gerken) Oxford University Press.
- Predelli, S. (2005). Painted leaves, context, and semantic analysis. *Linguistics and Philosophy* 28(3), 351–374.
- Preyer, G. and G. Peter (Eds.) (2005). *Contextualism in Philosophy: Knowledge, Meaning and Truth*. Oxford: Oxford University Press.
- Rothschild, D. and G. Segal (2009). Indexical predicates. *Mind & Language* 24(4), 467–493.
- Sainsbury, R. (2001). Two ways to smoke a cigarette. *Ratio* XIV(4), 386–406.
- Schaffer, J. (2004). From contextualism to contrastivism. *Philosophical Studies* 119(1-2), 73–103.
- Schaffer, J. and J. Knobe (2010). Contrastive knowledge surveyed. *Noûs*, 1–34.
- Sprouse, J. (2011, Mar). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behav Res Methods* 43(1), 155–67.
- Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press.
- Szabó, Z. G. (2000). *Problems of Compositionality*. Studies in Philosophy. New York: Garland Publishing, Inc.
- Szabó, Z. G. (2001). Adjectives in context. In I. Kenesei and R. M. Harnish (Eds.), *Perspectives on Semantics, Pragmatics, and Discourse*, pp. 119–146. Amsterdam: John Benjamins Publishing Company.

- Travis, C. (1985a). On what is strictly speaking true. *Canadian Journal of Philosophy* 15(2), 187–229.
- Travis, C. (1985b). Vagueness, observation, and sorites. *Mind* 94(375), 345–366.
- Travis, C. (1989). *The Uses of Sense: Wittgenstein's Philosophy of Language*. Oxford: Oxford University Press.
- Travis, C. (1994). On constraints of generality. *Proceedings of the Aristotelian Society* 94, 165–188.
- Travis, C. (1997). Pragmatics. In B. Hale and C. Wright (Eds.), *A Companion to the Philosophy of Language*, pp. 87–107. Oxford: Blackwell.
- Wason, P. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology* 52(2), 133–142.