teorema
Vol. XXXVI/1, 2017, pp. 00-00
ISSN: 0210-1602
[BIBLID 0210-1602 (2017) 34:2; pp. 00-00]

Experimenting on Contextualism: Between-Subjects vs. Within-Subjects[1]

Adrian Andrzej Ziółkowski

RESUMEN

Según el contextualismo, la gran mayoria de las expressiones del lenguaje natural son sensibles al contexto. Verificar si esta alegación se refleja o no en las intuiciones de la gente común, suscita algunas interesantes cuestiones metodológicas como: ¿qué diseño experimental es mas apropriado para poner a prueba el contextualismo, el intra-sujetos, o en correspondiente entre-sujetos? La thesis central de este artículo es que debería preferirse el modelo entre-sujetos.

El primer experimento aspira a evaluar la diferencia entre los resultados conseguidos para las mediciones intra-sujetos (donde todos los partecipantes evaluan todos los contextos) y para las mediciones entre-sujetos (donde los encuestados que evaluan diferentes contextos son divididos en grupos diferentes). Se muestra que el modelo intra-sujetos proporciona datos que parecen respaldar el contextualismo. Sin embargo, yo presento una interpretación alternativa y invariantista de estos resultados, mostrando entonces que el modelo intra-subjetos no permite distinguir entre contextualismo e invariantismo. El segundo experimento elabora adicionalmente la cuestión de cómo percibir el contraste entre los contextos puede afectar a los juicios de los sujetos. Muestro que algunos tipos de contextos pueden provocar intuiciónes opuestas cuando se contrastan con diferentes contextos.

PALABRAS-CLAVE: *contexstualismo, invariantismo, filosofía experimental, metodología, modelo experimental, intra-sujetos, entre-sujetos.*

ABSTRACT

According to contextualism, vast majority of natural-language expressions are context-sensitive. When testing whether this claim is reflected in Folk intuitions, some interesting methodological questions were raised such as: which experimental design is more appropriate for testing contextualism – the within- or the between-subject design? The main thesis of this paper is that the between-subject design should be preferred.

The first experiment aims at assessing the difference between the results obtained for within-subjects measurements (where all participants assess all contexts) and between-subject measurements (where respondents evaluating different contexts are distinct groups). It is shown that the within-subject design provides data that seems to support contextualism. However, I present an alternative, invariantist interpretation of these results, therefore showing that the within-subject design does not allow to empirically distinguish between contextualism and invariantism. The second experiment further elaborates the issue of how perceiving the contrast between contexts can affect subjects' judgments – I show that certain kinds of contexts may elicit opposite intuitions when contrasted with different contexts.

KEYWORDS: *Contextualism, Invariantism, Experimental Philosophy, Methodology, Experimental Design, Within-Subjects, Between-Subjects.*

I. INTRODUCTION

Contextualists claim that context-sensitivity is a widespread phenomenon in natural language, i.e. they believe that the conversational context in which a certain expression is uttered is very likely to affect the meaning of that expression. Thus, according to them, a given sentence, when used in different conversational contexts, usually expresses different propositions. On the other hand, inviariantists (also known as 'minimalists') argue that the meaning of most natural language expressions remains stable across contexts. They agree that there are some obviously context-sensitive terms, like indexicals (e.g. 'I', 'here', 'tomorrow'), but they also claim that if a sentence does not contain such terms, it expresses the same proposition in each utterance, regardless of the conversational context.[1]

In recent years, the discussion between contextualists and invariantists has moved in a new direction. Philosophers started seeking evidence to support of one of these views. Evidence that had not been previously put forward by their proponents. This new direction relies on methods of philosophizing introduced by experimental philosophers, according to whom we can gain new insights and formulate new arguments based on systematic empirical studies of folk intuitions. Since the supporters of contextualism offered many diverse cases that are supposed to elicit intuitions in favor of their view, there is a rich source of thought experiments that can be adopted in experimental studies on contextualism.

Contextualist thought experiments usually focus on describing pairs of conversational contexts in which a certain sentence is uttered by some speaker. Their authors argue that the difference in context affects the truth conditions of the utterance in question; which favors their view over invariantism. Experimental philosophers are interested in checking whether this crucial claim would be in fact reflected in folk judgments regarding contextualist thought experiments or, more specifically, in folk verdicts concerning truth-conditions of sentences uttered in different contexts. Importantly, there are two different ways of adopting contextualist thought experiments in empirical studies – subjects evaluating different contexts may be in separate groups (between-subject design) or both contexts can be presented to all subjects, giving them the opportunity to see the contrast between contexts when forming their judgments (within-subject design). Recently some philosophers (Hansen, Chemla, 2013; Hansen, 2014) argued that the latter approach is more suitable for providing conclusive data regarding the contextualism-invatiantism debate.

The main aim of this paper is to investigate the difference between the two abovementioned alternative experimental designs in case of studies concerning contextualism. Contrary to H&C,[2] I argue that the between-subject design is more appropriate for experiments examining folk intuitions about the influence of conversational context on the meaning of expressions. I provide reasons to believe that, in fact, the within-subject design tends to elicit judgments more favorable to contextualism than the between-subject design. However, I also claim that this effect can be explained as a result of an influence of factors different than the conversational context. Therefore, the alleged additional support to contextualism observed in within-subject experiments should not be considered relevant to the discussed issue. In my argumentation I refer to results of my own experiments. I also show that nevertheless, even for the between-subject design, the data collected in my experiments confirm contextualists' predictions to a degree that cannot be ignored.

The first section sketches the background for my experiments. Firstly, I discuss the way of delineating between contextualism and invariantism that corresponds to the methods used in experimental philosophy to test which view receives more support from folk judgments. Secondly, I briefly summarize the results of previous studies regarding contextualism, focusing mostly on the H&C experiment, their methodological proposal and arguments they present in favor of it.

In the second section, I report the results of my first study which examines the difference in folk judgments for contextualist thought experiments tested in within- and between-subject designs.

The third section focuses on the data collected in my second study, which shows that there are conversational contexts that are evaluated differently depending on whether they are contrasted with some other contexts or assessed independently.

II. BACKGROUND

II. 1. *How does Experimental Philosophy Distinguish Between Contextualism and Invariantism?*

One can choose different criteria to draw the distinction between contextualism and invariantism (sometimes also referred to as 'minimalism'). It is possible for one account under a certain criterion to count as invariantist, while according to another criterion be classified as contextualist. As an example, let us consider Jason Stanley's (2005) Interest Relative Invariantism (IRI), whose main claim is that the truth-value of knowledge ascriptions may depend on the practical interest of the agent to whom the knowledge is being attributed. If we decide to classify views by asking whether the account in question allows sentences to express propositions *outside* of the conversational context, IRI would be in fact counted among invariantist theories, since IRI claims that each knowledge attribution expresses the same complete proposition in every context. A view that would be classified as contextualist on this criterion is, for example, the one defended by François Recanati in *Truth Conditional Pragmatics* (2010). However, if we choose a criterion pointing at how often pragmatic (contextual) factors influence the truth-value of utterances according to the view in question, IRI would have to be classified as a contextualist view, because, according to Stanley, such an influence is a widespread phenomenon in cases of knowledge attributions. An example of an account that counts as invariantism on this criterion is the view proposed by Cappelen and Lepore (2005).

Most experimental philosophy studies concerning contextualism focus on the latter method of delineating between the rival theories in question. On one hand, invariantist views are the ones according to which context-sensitivity is a relatively rare phenomenon, restricted to well-known cases such as indexicals or demonstratives. On the other hand, contextualist accounts are the ones that predict that context-sensitivity is a common trait of expressions. Therefore, experimental philosophers' interest in contextualism usually aims at establishing whether folk judgments regarding truth value of sentences vary with conversational context in which the assessed sentence is uttered. In order to achieve this goal, they ask the participants of their experiments to evaluate pairs of vignettes, each describing a different conversational context in which the same sentence is being used by a speaker (from now on I will refer to such pairs as 'scenarios'). These compared vignettes are constructed in such a way that while contextualists would predict a shift in the truth-value of the utterance in question (due to contextual influence on truth conditions), invariantists would have to claim that the truth-value (and meaning) remains constant, because the only thing that makes the vignettes different is the conversational context. For the sake of clarity, let us from now on call contexts in which contextualists predict negative judgments 'rejection contexts' and the ones in case of which they predict positive judgments – 'acceptance contexts'.

Since the participants of experiments are asked to provide judgments about the truth-value of the crucial utterances, the method of measuring their intuitions about the influence of context on meaning is indirect, based on a widely shared opinion that reference is determined by meaning. The difference of subjects' judgments concerning the truth-value of the utterance in question between contexts can be interpreted as empirical evidence for contextualism. Unfortunately, lack of such differences is not enough to provide support to invariantism – note that it is possible to link different meanings with different uses of a sentence and still reasonably claim that the truth-value of that sentence is identical in both cases. Therefore, most experimental work on the discussed issue focuses more on contextualism than on invariantism. The studies designed along the lines of the approach described above are sometimes called *context shifting experiments*.

II. 2. *Results of Previous Studies Concerning Contextualism*

Most experiments carried out so far have focused on context-dependence of knowledge ascriptions [e.g. Buckwalter (2010); May *et al.*, (2010); Feltz and Zarpentine (2010); Pinillos (2011); Sripada and Stanley (2012)], and based on famous Bank Cases first introduced to the literature by Keith DeRose (1992). DeRose constructed these cases to illustrate and argue for his version of *epistemic contextualism.* According to DeRose the meaning of the predicate '…know that…' depends on what is at stake and how salient the possibility of error is (both factors are constituents of the conversational context). However, what needs to be stressed is that not all studies mentioned above explicitly aimed at providing data on epistemic contextualism; but rather on Interest Relative Invariantism. Nevertheless, the way context shifting experiments distinguish between invariantism and contextualism classifies both DeRose and Stanley as supporters of contextualism, because their views predict a shift in truth-evaluations between certain contexts.[3]

A lot of interesting work has been done on the topic of context-dependence of knowledge attributions. I will not discuss these results in details here. Instead, I will focus on experiments seeking support for contextualism as a broader view that suggests widespread context-dependence of many other expressions, not only knowledge ascriptions. However, two things about the experiments on knowledge attributions need to be pointed out. Firstly, majority of previous studies on this issue failed to provide strong support to epistemic contextualism (or IRI) – folk judgments regarding knowledge ascriptions were either completely insensitive to stakes and salience of error, or the influence of these factors on judgments was marginal, and far from shifting the subjects' verdicts from positive to negative. Secondly, and what is most important for the issue discussed in this paper, the data collected in some experiments show that conducting the experiment in within-subject design usually increases the magnitude of the observed difference in judgments across contexts in comparison to studies utilizing the between-subject design [see for example May *et al.* (2010)].

Based on this observation, Nat Hansen and Emmanuel Chemla (2013) carried out a complex study testing many different contextualist scenarios using a within-subject design. The participants of their experiments assessed ten cases – four knowledge scenarios (testing context-sensitivity of the predicate '…know that…'), four color scenarios (testing context-sensitivity of adjectives describing colors) and two 'miscellaneous' scenarios in case of which the expected context-sensitivity cannot be clearly classified. Following DeRose's (2011) suggestions, H&C decided to present every scenario in two variants: in one variant subjects were asked to evaluate an affirmative claim, while in the other variant their task was to evaluate its negation. DeRose offered a defense strategy for contextualism by stipulating that lack of confirmation observed in previous experiments is an effect of a pragmatic phenomenon known as accommodation [Lewis, 1979]. It is a tendency of hearers to look for such an interpretation of utterances on which they turn out to be true. In some cases, this tendency may lead hearers to form non-standard interpretations of utterances – according to DeRose that might explain subjects' positive judgments in cases where an agent ascribes knowledge to herself, but contextualism predicts that the context of utterance shifts the meaning in such a way that the ascription is false. To check these stipulations, H&C decided to ask their subjects to asses *both* affirmative and negative claims in *both* kinds of contexts in order to test whether DeRose's suspicions about the role of accommodation in shaping subjects' judgments are in fact right. As a result, each scenario was presented to participants of their experiments in four variants, which means that each context of each scenario was presented twice. In total, each participant of their study had to give 44 judgments (10 target scenarios and one control scenario). Their experiment was designed for 2x2 ANOVA model with repeated measurements including two factors: context (rejection vs. acceptance)[4] and polarity (affirmative vs. negative claim).

H&C designed a very interesting method aiming at randomizing the presentation of vignettes in an optimal manner which they refer to as 'block design'. Due to lack of space, I will not go into details here. However, it is worth mentioning that the block design allowed H&C to minimize the probability of two versions of one scenario appearing one after another, and extrapolate how the results would look if they used a between-subject design.

The data H&C subjected to statistical analysis included answers given by 39 respondents. They argue that the results they obtained provide strong support to contextualism. In fact, subjects were much more likely to deny the utterances in rejection contexts than in acceptance contexts with regards to color and "miscellaneous" scenarios (for the positive sentences, for the negative sentences the effect direction was, of course, opposite). For most of these scenarios the effect size was really robust. On average, people rejected the utterance in one context while accepting it in the other. However, the results were completely different when it comes to knowledge scenarios. Most comparisons showed no significant differences between judgments for opposite contexts. This result replicates the data collected in previous experiments investigating folk intuitions concerning context-sensitivity of knowledge attributions.

Contrary to DeRose's (2011) suspicions, H&C did not observe any symptoms of accommodation in answers given by the participants of their experiment. The distribution of judgments between contexts for positive and negative sentences was symmetrical. Interestingly, they noticed a different bias in folk judgments – subjects were, on average, more likely to accept positive sentences than negative ones.

Using some traits of their block design, H&C tried to extrapolate how the results might have looked if the participants of their experiment weren't able to perceive the contrast between contexts. In order to do this, they ran additional statistical analysis focusing only on answers given by subjects for the first presented block, which included variants of scenarios representing all four combinations of factors (context x polarity) for knowledge and color cases, and two combinations for miscellaneous cases. Therefore, H&C assumed that comparing the judgments given in the first block can allow us to "simulate" the between-subject design (although not fully, because the idea here rests on comparing judgments concerning different scenarios). The analysis revealed no significant difference for knowledge cases, but confirmed the existence of contextual effects in color and miscellaneous cases. However, the size of the obtained effects was somehow smaller than in case of full-blown within-subject measurements.

The H&C experiment was the first study focusing on contextualism that was not restricted to the issue of context-dependence of knowledge attributions, and the first study in which robust contextual effects were observed. Therefore, I attempted to replicate their findings for two scenarios in case of which the size of the observed effect was the most profound. Apart from the attempted replication, I also tried to provide better evaluation of the difference between the results one can obtain for the between-subject and the within-subject measurements. The details are discussed in the following section.

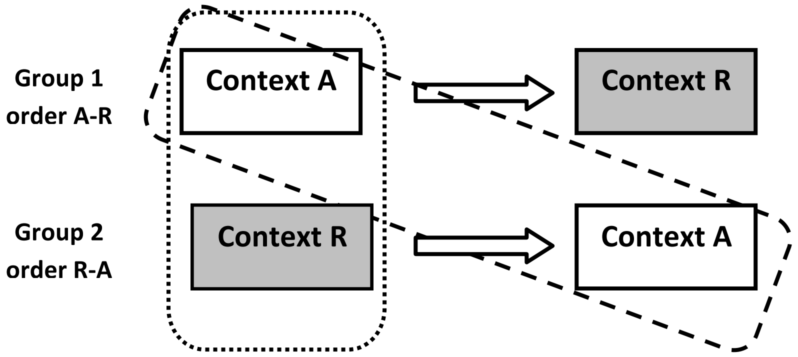
III. First study

III. 1. *Experimental Design*

My first experiment aimed at providing further data concerning two cases used in the H&C study – *Milk* and *Weight* (see Appendix) – by comparing the sizes of the contextual effect for between- and within-subject design. As mentioned above, such a procedure was in fact included in the original study, but the sample size for the between-subject comparison was considerably small, in my experiment I checked whether the picture would remain the same for a bigger sample. Since H&C did not observe any important difference between judgments for positive and negative sentences that would be relevant to the contextualism-invariantism debate (I assume that the truth-bias towards positive sentences is not relevant here), all participants of my study were asked to assess affirmative sentences.

The participants of my first experiment were randomly assigned to either the *Milk* or *Weight* scenario. Each subject gave her judgment about both contexts, but, depending on random assignment, the order of presentation started either with the acceptance context or the rejection context. Different variants of scenarios were presented and evaluated separately. After assessing one context, subjects did not have the opportunity to change their previous answers. The design of the experiment for each scenario is presented in the graph below.

Graph 1. The experimental design for each scenario and possible ways of analyzing the data it can provide.



An experimental design of this nature makes it possible to perform both between- and within-subject comparations on data collected in one experiment. Comparing first judgments given by representatives of different groups (the dotted frame in the above picture), works in an analogous way as a full-blown between-subject design. On the other hand, collapsing the answers from both groups and comparing two judgments given by each respondent works in the exact same way as a within-subject design with presentation of contexts counterbalanced for order. Moreover, and most importantly, we can also evaluate how much of the effect size is 'added' in the within-subject design by comparing the verdicts for each context depending on whether it was presented as first, or primed with the presentation of the opposite context (the dashed frame in the above picture).

The participants of my experiment gave their judgments using a 5-point scale, with the extremes defined as 'true' and false'. The crucial question in every experimental variant read as follows: *Please evaluate the* [protagonist name]'s *claim: '*[*target sentence*]*' using the scale below, where '1' means 'false' and '5' means 'true'.*

The survey was published online, on a website hosted by Experimental Philosophy Lab (KogniLab) at the University of Warsaw <www.kognilab.pl>.

Even though my study focused on the exact same scenarios as the one used in the H&C experiment, strictly speaking, it is not a replication of their study. I did not use the same experimental design (the block design), the way of measuring subjects' verdicts was also slightly different (continuous vs. 5-point scale). Most importantly, their block design prevented different variants of each scenario from being presented directly after each other (it was possible, but highly improbable), while in my experiment subjects were presented with the second variant of a scenario right after evaluating the first variant. However, even though my experiment is not an exact replication of H&C study, the methodology I adopted is not different to their experiment to such an extent that would not allow comparing my results with their data.

III. 2. PARTICIPANTS

Subjects were recruited through the Internet by an invitation sent via e-mail. They were asked to forward the invitation to other people that might be interested in taking the survey, which created a 'snowball effect' and resulted in collecting enough data. The respondents did not receive any pay for their participation in the experiment.

In total, 156 subjects filled in the survey, but only 128 answers were included in further statistical analysis, since 28 respondents either reported having a degree in philosophy (BA or higher) or admitted not being English native speakers. All statistics presented below concern the probe containing 128 responses. 62 subjects were assigned to the *Milk* scenario, whereas 66 subjects gave their opinions about the *Weight* scenario. 53.9% of respondents were female. 46.1% of respondents were male. The average age was 33.4 years (with standard deviation of 13.7 years). The youngest participant was 19 years old. The eldest participant was 77 years old. What needs to be stressed here is that the distribution was clearly skewed towards younger people – majority, 70% of subjects were 20 to 35 years old.

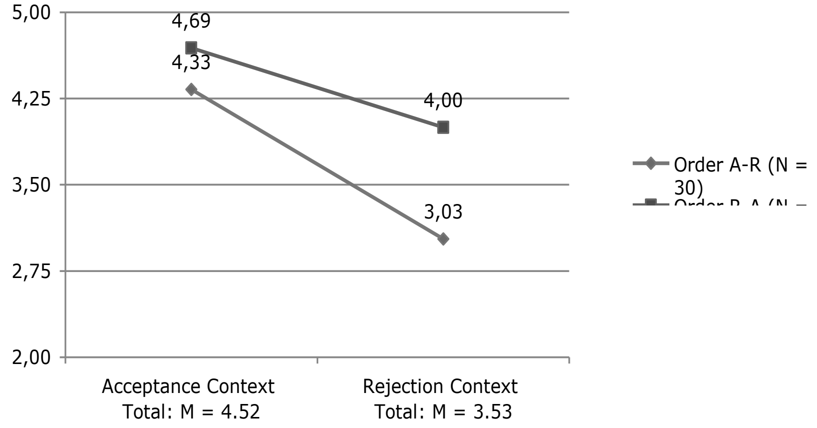
III. 3. *Results*

I analyzed the data by using a mixed ANOVA model including one within-subject factor (context) and one between-subject factor (order of context presentation), computed separately for each scenario. To test whether the contextual effect obtains for the between-subject measurements, I ran additional independent-samples t-test comparisons.

III. 3. 1. *Milk*

The *Milk* scenario analysis revealed a significant main within-subject effect of context, which means that when perceiving the contrast between contexts, subjects in fact evaluated the milk-utterance differently depending on the context of utterance – $F(1, 60) = 39.38$; $p < 0.001$; $\eta^2 = 0.396$. In accordance with contextualists' predictions, subjects were more likely to judge that the target utterance is true in the acceptance context (*M* = 4.52; *SD* = 0.95) than in the rejection context (*M* = 3.53; *SD* = 1.57). This replicates the result obtained for the *Milk* scenario in the H&C study. However, the size of the observed contextualist effect is considerably smaller in my experiment. Instead of judging that the utterance in the rejection context is false, subjects' answers for this context were rather ambivalent. Main results of the experiment for the *Milk* scenario are illustrated in the chart below.

Chart 1. Mean ratings in different contexts and orders of presentation for the *Milk* scenario.



The interaction between context and order of presentation reached the level of statistical trend – *F*(1, 60) = 3.74; *p* = 0.058; η^2 = 0.059. Thus, the order in which opposite contexts were presented influenced to some degree the judgments given by subjects. Post-hoc comparisons (based on Bonferroni correction) of the evaluations for each context between orders of presentation revealed that while priming did not affect the judgments concerning the acceptance context (no prime: *M* = 4.33; primed: *M* = 4.69), it did significantly change the answers for the rejection context (no prime: *M* = 4.0; primed: *M* = 3.03). Subjects who first got familiar with the acceptance context were much more reluctant to accept the target utterance in the rejection context than respondents who started with evaluating the rejection context.

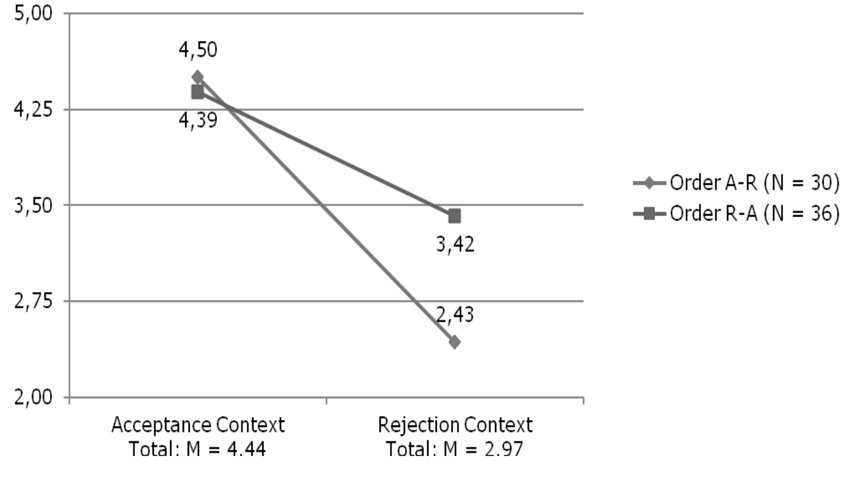
Most interestingly, however, the contextual shift in truth-evaluations obtained for related measurements did not remain in case of the between-subject comparisons. An adequate t-test comparison did not reach the threshold for statistical significance – *t*(60) = 1.12; *ns*. In their first evaluations (without priming), subjects were similarly likely to accept the target utterance in the acceptance context (*M* = 4.33; *SD* = 1.03) and in the rejection context (*M* = 4.0; *SD* = 1.30). Perceiving the contrast between opposite contexts was therefore necessary for the contextual effect to appear.

III. 4. 2. *Weight*

Similarly with the case of the *Milk* scenario, a significant main within-subject effect of context was also obtained for the *Weight* scenario – *F*(1, 64) = 72.81; $p < 0.001$; $\eta^2 = 0.532$. When subjects were able to see the difference between contexts, they were much more likely to accept the target utterance in the acceptance context (*M* = 4.44; *SD* = 0.90) than in the rejection context (*M* = 2.97; *SD* = 1.47). It is worth noting that the size of this effect was bigger than the observed effect in the *Milk* scenario (see the η^2 coefficient). Nevertheless, the judgments for the rejection context were not clearly negative, as predicted by contextualism, but rather ambivalent.

Key results of the experiment for the *Weight* scenario are summarized in the graph presented below.

Chart 2. Mean ratings in different contexts and orders of presentation for the *Weight* scenario.



Apart from the main effect of context, a significant interaction of context and order of presentation was observed as well – $F(1, 64) = 9.44$; $p = 0.003$; $\eta^2 = 0.129$. This means that subjects' judgments concerning different context depended on the order of presentation. In analogy to the *Milk* scenario, the order of presentation had no influence on truth-evaluations made for the acceptance context (no prime: $M = 4.50$; primed: $M = 4.39$), but it made an important difference in case of the rejection context (no prime: $M = 3.42$; primed: $M = 2.43$). This last difference turned out to be significant according to an appropriate post-hoc Bonferroni test.

Contrary to the pattern of results seen for the *Milk* scenario, the contextual shift in truth-evaluations with regard to the *Weight* scenario was observed for both within- and between-subject measurements. It turned out that when giving their first judgment, subjects who evaluated the acceptance context were significantly more happy to accept the target utterance ($M = 4.50$; $SD = 0.90$) than the ones who evaluated the rejection context ($M = 3.42$; $SD = 1.34$) – $t(64) = 3.78$; $p < 0.001$. It is worth noting, though, that even in this case the size of the contextualist effect observed for the between-subject design was considerably smaller than the analogous effect obtained for the within-subject design.[5]

III. 5. *Discussion*

At first glance, it may seem like the data collected in my first experiment confirms the results obtained by H&C; therefore, providing further empirical support to contextualism. As we have seen, at least for the within-subject design, both tested scenarios – *Milk* and *Weight* – elicited a contextual shift in truth-evaluation. The strength of this effect was not as robust as contextualists would like it to be, but, nevertheless, the observed tendency seems to be much more in accordance with contextualism than invariantism.

After further consideration however, the results of my experiment confirm that, as one might have expected basing on previous observations, the judgments elicited by context shifting experiments do in fact depend on whether subjects have the opportunity to directly see the contrast between contexts. The divergence of judgments given for opposite contexts was considerably bigger for the within-subject design than for the between-subject design in case of both the *Milk* and *Weight* scenarios. When it comes to the former scenario a significant difference between rejection and acceptance contexts was observed *only* for related measurements, so seeing the contrast between contexts was necessary for the contextual shift to appear. It looks as if, at least in some situations, our conclusions concerning empirical support to contextualism may depend on the method we choose to measure folk intuitions.

Moreover, it is worth considering the way in which the order of presentation of contexts influenced the judgments given by the subjects. The comparably strong within-subject contextual effects obtained for both *Milk* and *Weight* scenarios were mostly 'produced' by the judgments given by subjects who first got familiar with the acceptance context and then evaluated the rejection context. The truth-evaluations given for the acceptance context were insensitive to priming with the opposite context. In case of the rejection contexts subjects were much more reluctant to accept the target utterance when their judgment was primed with the presentation of the acceptance context. It seems that these results reveal an interesting bias towards giving positive judgments in the first evaluated case among participants. It is much likely for the subjects to give a negative verdict in context shifting experiments if the rejection context is contrasted with the acceptance context, but initially they rather tend to give positive judgments.

What conclusions should we draw from the observed differences between rival experimental designs and influence of order of presentation on intuitions elicited by context shifting experiments? Should we follow Nat Hansen (2014) and claim that the within-subject design generates better data for the purpose of assessing contextualism empirically, because, as he argues, it helps subjects in understanding how meaning is subject to context of utterance? On the contrary, I will question the claim that the additional divergence between judgments for different contexts observed in within-subject experiments are to be taken as a support to contextualism. Therefore, I will argue that when seeking evidence in support of contextualism, we should rather focus on the results of between-subject experiments.

The first reason to oppose the abovementioned claim of Nat Hansen is the fact that the most prominent proponents of contextualism [e.g. Recanati (2010); DeRose, (1992)] do not mention the influence of *contrast* between contexts on meaning among the main theses of contextualism. In fact, as Hansen (2014) points out, when contextualists introduce thought experiments in support of their view, they usually present different contexts of utterance by directly contrasting them. But, regardless of this, the core claim of contextualism is that meaning is subject to the context of utterance *per se*, not to the contrast between different possible contexts of utterance. The latter could be the crucial factor for some contrastivist view instead. This is a theoretical reason to have doubts about Hansen's (2014) main methodological suggestion.

The second argument against the use of within-subject design in experiments concerning contextualism is that the difference between rival experimental designs observed in my study should not be explained in terms of context of utterance, but rather some other factors.

Even if we agree with the supporters of contextualism that the additional divergence in judgments observed in within-subject studies is in fact a sign of contextual influence on meaning, the reason for preferring the between-subject design would remain. When we are interested in finding empirical evidence in favor of contextualism, we should be investigating whether the context of *utterance* affects the meaning of the statement in question. However, what the within-subject design "adds" to the results obtained in case of the between-subject design should be rather explained in terms of the context of *evaluation*, if to be seen as a contextual effect at all. Seeing the contrast between compared contexts is not a factor operating at the level of context of utterance, but the context of evaluation (the context of the experimental procedure). So, the effects observed in experiments based on related measurements may be jointly produced by two different contextual phenomena, which should not be confused.

However, I believe that another plausible approach to the results of my experiment should be formulated in invariantist, rather than contextualist terms. The participants of within-subject experiments become familiar with pairs of descriptions of very similar situations that upon direct comparison clearly differ in some respects. The crucial parts – the utterance and the question concerning its truth value – on the other hand, remain the same. If we consider some well-known theories of pragmatic phenomena, such as Grice's (1975) account of conversational maxims, we should expect that such an experimental setting would increase the tendency of subjects to differentiate their judgments. However, this tendency will not be due to the context of evaluated utterance, but rather to what subjects might believe is implicated by the crucial question addressed by the experimenter. The Gricean account of conversational maxims was initially proposed for sentences in indicative mood, but it could be easily generalized to questions. The Maxim of Quantity for questions could, for example, prevent speakers from asking the same question more than once if the speaker has no reason to expect to gain more information as a result. Note, however, that being asked the exact same question twice (but in slightly different circumstances) is exactly what happens to participants of within-subject context shifting experiments. Due to standard conversational practices, subjects might have assumed that different answers are *expected* by the experimenters, otherwise they would not ask the question twice. Therefore, the within-subject design might have simply encouraged the respondents to look for such a reading of the crucial question, on which the differentiation of their answers would be justified. But in such cases there is no guarantee that all subjects in fact expressed their verdicts concerning truth-conditions of the target utterance, instead of, for example, their warranted assertability.

The above suggestions are of course, purely speculative and, at least at this point, I am not in possession of any independent empirical evidence in support of my claim. However, the above considerations show that it's possible to, at least in principle, give an explanation of data seemingly supporting contextualism in invariantist terms. Adopting the same strategy to discard data supporting contextualism collected using between-subject design would be much more difficult. In other words – my argumentation shows that context-shifting experiments that use within-subject design do not allow to empirically distinguish between contextualism and invariantism.

In order to further support my doubts about using within-subject design in experiments concerning contextualism I decided to run a follow-up experiment. My secondary experiment aimed at showing that in some cases contrasting scenarios can not only strengthen the effects observed when there's no contrast at play, but also completely change the evaluation of a context in comparison to the condition with no contrast at all.

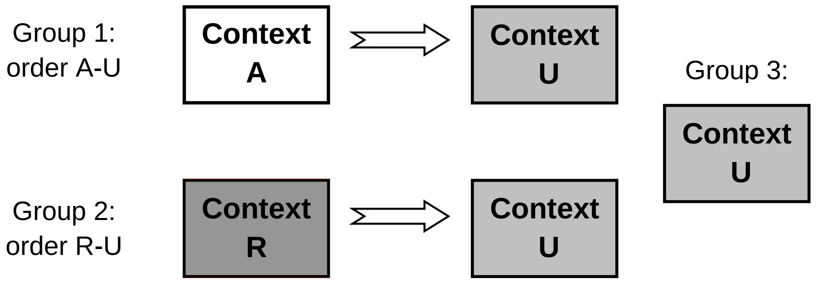
IV. Second Study

IV. 1. *Experimental Design*

While in my first study the within-subject design contrasted acceptance and rejection contexts with each other, for the purpose of the second experiment I created a third kind of context and decided to contrast it with the two others. The contexts of this type, which we may call *uncertainty contexts*, were intentionally designed to elicit ambivalent intuitions – it is really far from clear what is the truth-value of the target utterance in this case (for more details see Appendix).

The experiment aimed at comparing subjects' verdicts concerning the uncertainty context in three different conditions: 1) when it was contrasted (primed) with the acceptance context; 2) when it was contrasted (primed) with the rejection context; 3) when it was not contrasted with any context whatsoever (no priming). Each subject was randomly assigned to one of these conditions, so the comparison was based on a full-blown between-subject design, fitting a simple one-factor ANOVA model. Note that such an experimental procedure additionally allowed evaluating the size of the contextual effect between acceptance and rejection contexts (for independent measurements). The design of the experiment for each tested scenario is illustrated in the graph below.

Graph 2. The experimental design for each scenario in the second study.



All other methodologicalaspects of the second study were similar to the first experiment I discussed above. Even though the second study tested two different scenarios (triplets of contexts), I present in detail only one of them, since the results obtained for the other one did not confirm the expectations.

IV. 2. *Participants*

Subjects were recruited from internet users registered as ‘workers’ on the Amazon Mechanical Turk website <www.mturk.com>. Each participant was paid $0.3 for taking the survey.

In total, 143 respondents filled in the survey[6]. 12 of them, however, reported having a degree in philosophy (BA or higher), admitted not being native English speakers or failed to answer comprehension questions correctly. Further statistics concern answers provided by 131 participants.

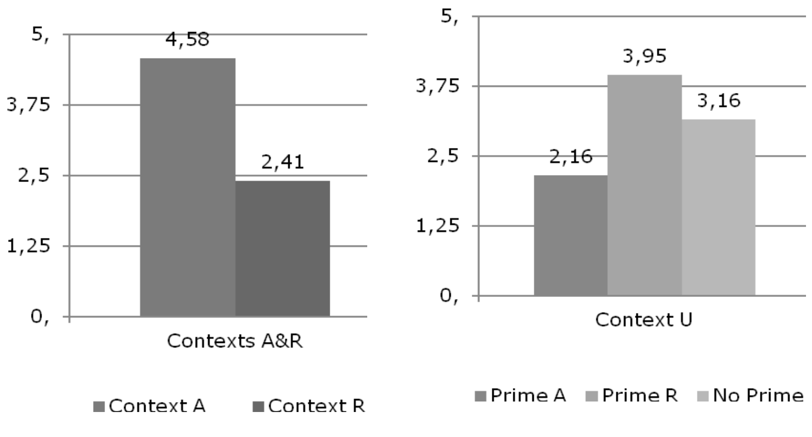
71% of subjects were male, 29% were female. The age of participants ranged from 19 to 57, with the average equalling 30.18 (and standard deviation equal 7.42). The majority of subjects, 65.4%, were 31 years old or younger.

IV. 3. *Results*

In order to check whether subjects reacted to the uncertainty context differently depending on the experimental condition, I subjected the data to a simple one-way ANOVA analysis. It turned out that priming influenced subjects' judgments – $F(2, 128) = 26.43$; $p < 0.001$. Post-hoc tests (based on Bonferroni correction) revealed significant differences ($p < 0.05$) between all pairs of conditions. When there was no prime involved, subjects gave ambivalent answers with an average equal to 3.16 and considerable variance ($SD = 1.24$). Subjects were less likely to accept the target utterance if the presentation of uncertainty context was primed with evaluation of the acceptance context ($M = 2.16$; $SD = 1.13$) and, accordingly, more likely to accept it if they first assessed the rejection context ($M = 3.95$; $SD = 1.08$). This means that the context of uncertainty elicited different reactions depending on whether it was primed, and how it was primed.[7]

The experiment also provided data allowing for a between-subject comparison of acceptance and rejection contexts. In this respect a significant and considerably robust contextualist effect was observed. Subjects tend to accept the claim 'There is gasoline in the garage' much less willingly in the rejection context ($M = 2.41$; $SD = 1.37$) than in the acceptance context ($M = 4.58$; $SD = 0.73$) – $t(85) = 9.19$; $p < 0.001$. The results are summarized in the graph below.

Chart 3. Mean ratings in different contexts and orders of presentation in the second experiment.



V. CONCLUSIONS

As the results of my second experiment show, in some cases contrasting one context with different vignettes can make people react to this context in a significantly different way. This observation is extremely important when discussing the choice of proper experimental design for experimenting on contextualism. The first experiment proved that non-philosophers' verdicts regarding different rejection contexts were similarly subject to priming. This means priming caused a rise in differences between judgments for acceptance and rejection contexts. Here, the influence of contrasting contexts is along the lines of contextualist predictions. Some philosophers argue that this procedure should be used in experiments on contextualism because it helps collect more reliable data. I already presented some arguments against that claim, but when it comes to the role contrasting contexts play in the case of my second experiment, the picture becomes even more complicated. Which judgments concerning the uncertainty context given by the participants of my second study are more reliable – the ones given by subjects who first assessed the acceptance context, or the ones who first evaluated the rejection context? It seems that there is no clear answer to this question. Maybe the right answer is that neither of them are fully reliable, and if we want to draw conclusions about folk intuitions concerning the truth-value of sentences in different conversational contexts, we should rather focus on judgments that weren't influenced by any specific contrast.

Unfortunately, this interesting pattern of results for uncertainty contexts was observed only in case of one of the scenarios tested in my experiments. Nevertheless, the result discussed above proves at least that it is *possible* to obtain such a result if we adopt a within-subject design of an experiment on this issue[8].

One more thing needs to be said – it is not the case that the results of my experiments undermine the *main* claim made by H&C. Their key conclusion that contextualism receives quite a lot of support from folk intuitions can be seen in the data collected by my experiments. In the cases of *Weight* and *Puddle of Gasoline* scenarios significant and robust between-subject contextualist effects were observed. The only claim that I tried to challenge here is the one according to which within-subject context shifting experiments is a proper method for testing contextualism within the framework of experimental philosophy. I think that the arguments presented in this paper should make us skeptical about this idea.

APPENDIX – VIGNETTES PRESENTED TO PARTICIPANTS

Milk – Acceptance Context

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 armchair, book and beverage in hand. The refrigerator is devoid of milk except for a puddle of milk at the bottom of it. Odile opens the refrigerator, looks in, closes it and says to Hugo, 'There is milk in the refrigerator'.

Please evaluate Odile's utterance 'There is milk in the refrigerator' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Milk – Rejection Context

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. Odile says to Hugo, 'There is milk in the refrigerator'. Please evaluate Odile's utterance 'There is milk in the refrigerator' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Weight – Acceptance Context

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. While wearing his heavy winter clothes, Hugo wants to announce the progress of his diet, and he says 'I weigh 80 kilograms'. Please evaluate Hugo's utterance 'I weigh 80 kilograms' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Weight – Rejection Context

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. 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, 'I weigh 80 kilograms'. Please evaluate Hugo's utterance 'I weigh 80 kilograms' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Puddle of Gasoline – Acceptance Context

Mary and Gregory's car ran out of fuel. Before Mary left, Gregory refilled the tank with the rest of the gasoline that they had in a canister. By accident, he spilled some of the gasoline onto the garage floor, leaving a small puddle. Except for the puddle, there is no more gasoline in the garage. Gregory's son, in exchange for borrowing his parents' car, has been working to clean the garage. "It's finished!" he says. Gregory takes a look into the garage to check on his son's work. The puddle of gasoline is still on the floor. "Hey! There is gasoline in the garage!" Gregory yells. Please evaluate Gregory's statement 'There is gasoline in the garage' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Puddle of Gasoline – Uncertainty Context

Mary and Gregory's car ran out of fuel. Before Mary left, Gregory refilled the tank with the rest of the gasoline that they had in a canister. By accident, he spilled some of the gasoline onto the garage floor, leaving a small puddle. Except for the puddle, there is no more gasoline in the garage. Gregory hears a doorbell. It's his neighbor, Ben, wearing a kitchen apron. 'Pal, I need your help. I need to light the fire in the barbecue. Do you have some gasoline?' – He asks. 'Yes. There is gasoline in the garage.' – Gregory replies. Please evaluate Gregory's statement 'There is gasoline in the garage' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Puddle of Gasoline – Rejection Context

Mary and Gregory's car ran out of fuel. Before Mary left, Gregory refilled the tank with the rest of the gasoline that they had in a canister. By accident, he spilled some of the gasoline onto the garage floor, leaving a small puddle. Except for the puddle, there is no more gasoline in the garage. A motorcycle stops in front of Gregory's home. The owner of the motorcycle approaches Gregory and says, "Pardon me but I am running low on gasoline. May I borrow a small amount to reach the nearest gas station?" "Yes. There is gasoline in the garage." Gregory replies. Please evaluate Gregory's statement 'There is gasoline in the garage' using the scale below, where '1' means 'definitely false' and '5' - 'definitely true'.

Philosophy Institut
University of Warsaw
Nowy Świat 69, 00-046 Warsaw, Poland
E-mail: adrian.a.ziolkowski@gmail.com

ACKNOWLEGMENTS

This research project was financed by the Polish National Science Center basing on the decision DEC/2012/05/N/HS1/03303.

NOTES

[1] This is a rough, simplified characterization for introductory purposes. For a more precise delineation, see the beginning of section II.1. below.

[2] From now on, I will use this abbreviation to refer to Hansen, Chemla (2013).

[3] Of course, it is possible to empirically distinguish between epistemic contextualism (as proposed by DeRose) and IRI (as proposed by Stanley). What matters for the former is the conversational context of the attributor, while in case of IRI the crucial factor is the practical interest of the subject to whom knowledge is ascribed. When the agent attributes knowledge to herself, both theories will give similar predictions. However, if the attributor and the agent to whom knowledge is ascribed are different persons, the predictions will diverge. Stanley (2005) presents variants of Bank Cases that illustrate this difference – *Ignorant High Stakes, Attributor Low Stakes-Subject High Stakes*, and *Attributor High Stakes-Subject Low Stakes*. Since context-dependence of knowledge attributions is not my main interest here, I will not discuss this issue in details, though.

[4] H&C used different names to call these contexts. In analogy to Bank Cases, where acceptance contexts are low-stakes, low-error salience cases and rejection contexts are high-stakes, high-error salience cases, they call these two categories 'low' and 'high' context. I find these terms a bit confusing when it comes to color and miscellaneous scenarios, in case of which stakes and error salience plays no role, so I introduced more universal labels for the opposing contexts.

[5] Interestingly, these results were replicated for Polish speakers (in the study I used Polish translations of the scenarios). There were significant priming effects for both *Milk* (N = 75) and *Weight* (N = 78) scenarios. Similarly, as in the study described above, there were significant differences in subjects' judgments between contexts for within-subject measurements in both tested cases. However, when it comes to between-subject measurements, the contextual effect was present only for *Weight*, but not for *Milk*, which is also in accordance with the data obtained from English speakers. It suggests that intuitions concerning context-sensitivity might be stable across (at least some) ethnic languages. More cross-linguistic research on that topic would be interesting.

[6] In fact the sample size was bigger, but here I provide detailed statistics for only one scenario which yielded results confirming my expectations.

[7] The data presented here should be taken skeptically, as my recent attempt to replicate these findings was not successful. During the publication process, motivated by my initial results, I was trying to find more scenarios in case of which a similar pattern of priming for certainty contexts would occur. These attempts failed, which, in turn, led me to an attempted replication of the

second study presented here. The repeated study with a slightly bigger sample size (N=165) did not find a similar priming effect. As far as uncertainty contexts are concerned, the only observed difference was that subjects were slightly more likely to judge the target sentence true when there was no prime compared to those, who first evaluated the rejection context (which is an *opposite* effect to that observed in the initial study). The only effect that was replicated was a strong, between-subjects contextualist effect in evaluations regarding acceptance and rejection contexts. Thus, it seems that the divergence in judgments regarding uncertainty contexts observed in the initial study was not due to priming, but it was rather a pure effect of uncertainty. Most probably, non-philosophers do not have strong and certain intuitions about these kind of cases, which results in instability of their judgments.

[8] In the light of the abovementioned non-replication (see footnote 8.), it does not seem that Experiment 2. provides any genuine support to the main thesis of this paper. However, that does not spoil the central argument presented here, as it rests mostly on the results of Experiment 1. The second study was only supposed to strengthen the argumentation, so even if we discard the data collected in Experiment 2., the argument remains.

REFERENCES

BUCKWALTER W. (2010), 'Knowledge Isn't Closed on Saturdays', *Review of Philosophy and Psychology*, 1, pp. 395-406.

CAPPELEN H. and LEPORE E. (2005), *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*, Oxford, Blackwell.

DEROSE K. (1992), 'Contextualism and Knowledge Attributions', *Philosophy and Phenomenological Research*, 52 (4), pp. 913-929.

— (2011), 'Contextualism, Contrastivism, and X-Phi surveys', *Philosophical Studies*, 156 (1), pp. 81–110.

FELTZ A. and ZARPENTINE C. (2010), 'Do You Know More When It Matters Less?', *Philosophical Psychology*, 23 (5), pp. 683-706.

GRICE H. P. (1975), 'Logic and Conversation', in P. Cole, J. Morgan (ed.), *Syntax and Semantics*, vol.3, New York, Academic Press.

HANSEN, N. (2014), 'Contrasting Cases', in James Beebe (ed.), *Advances in Experimental Epistemology*, Bloomsbury, pp. 71-95.

HANSEN N. and CHEMLA E. (2013), 'Experimenting on Contextualism', *Mind and Language*, 28 (3), pp. 286-321.

KING J. C. and STANLEY J. (2005), 'Semantics, Pragmatics, and The Role of Semantic Content', in Z. Szabo (eds.), *Semantics vs. Pragmatics*, Oxford, Oxford University Press.

LEWIS D. (1979), 'Scorekeeping in a Language Game', in R. Baüerle, U. Egli, A. von Stechow (eds.), *Semantics from a Different Point of View*, Berlin, Springer.

MAY J., SINNOTT-ARMSTRONG W., HULL J. G. and ZIMMERMAN A. (2010), 'Practical Interests, Relevant Alternatives, and Knowledge Attributions: An Empirical Study', *Review of Philosophy and Psychology*, 1, pp. 265-273.

RECANATI, F. (2010), *Truth-Conditional Pragmatics*, Oxford, Oxford University Press.

STANLEY J. (2005), *Knowledge and Practical Interests*, Oxford, Oxford University Press.