Socratic Questionnaires

Nat Hansen, Kathryn B. Francis, and Hamish Greening

Abstract

When experimental participants are given the chance to reflect and revise their initial judgments in a dynamic conversational context, do their responses to philosophical scenarios differ from responses to those same scenarios presented in a traditional static survey? In three experiments comparing responses given in conversational contexts with responses to traditional static surveys, we find no consistent evidence that responses differ in these different formats. This aligns with recent findings that various manipulations of reflectiveness have no effect on participants' judgments about philosophical scenarios. Although we did not find a consistent quantitative effect of format (conversation vs. static survey), conversational experiments still provide qualitative insights into debates about how participants are understanding (or misunderstanding) the scenarios they read in experimental studies, whether they are replacing difficult questions with questions that are more easily answered, and how participants are imagining the scenarios they read in ways that differ from what is explicitly stated by experimenters. We argue that conversational experiments—"Socratic questionnaires"—help show what is going on "under the hood" of traditional survey designs in the experimental investigation of philosophical questions.

1. Background and Overview

In his 1958 criticism of ordinary language philosophy, Benson Mates introduced the idea of a "Socratic questionnaire". Instead of the standard static approach to investigating meaning used by ordinary language philosophers, in which speakers of a language are asked to use an expression in various imagined situations, or judge whether a use of an expression in an imagined situation is something "we would say", Mates raised the possibility of a dynamic approach to probing meaning, by asking speakers

> prodding questions aimed at drawing the subject's attention to borderline cases, counterexamples, and various awkward consequences of his first and relatively off-hand answers. (Mates, 1958, p. 169)

Our thanks to Josh Alexander, James Andow, Alex Davies, Alexander Dinges, Sarah Fisher, Michael Hannon, Chauncey Maher, Eddy Nahmias, Julia Zakkou, audiences at the University of Nottingham, the European X-Phi Conference, the University of Iceland and the IRP group at the University of Reading for very helpful comments, Bob Bishop and Sara Vilar-Lluch for independent coding, Kimberley Tang for running the chats in experiment 3, and Wesley Buckwalter and Nick Byrd for their work as referees. Nat Hansen gratefully acknowledges support from the University of Reading's UROP program and the Alexander von Humboldt Foundation.

Responses to a Socratic questionnaire would reveal whether, when given time to reflect and consider relevant arguments, a speaker would stick with or revise their initial "off-hand" responses to philosophical questions. Mates argues that responses to a Socratic questionnaire should be considered evidence of how speakers use expressions alongside evidence from the standard, static approach.

Even though ordinary language philosophy is widely considered to have been killed off sometime in the 1960s or 1970s by the rise of systematic semantic theory and Gricean pragmatics, one of its legacies is the static approach to the investigation of meaning in experimental philosophy of language (Hansen 2014). In contemporary experimental investigations of linguistic meaning, the standard approach is to elicit one-off judgments about the use of expressions in various contexts—typically truth value judgments, or judgments of acceptability, or judgments about where on a Likert scale to locate the degree of one's agreement with some statement about a use of language. Mates's idea of a Socratic questionnaire has remained an unexplored option among those investigating the way people ordinarily use language and respond to philosophical thought experiments.

Some more recent critics of standard, survey-based approaches in experimental philosophy have raised worries about the static approach to the investigation of meaning that are related to Mates's mid-century challenge to ordinary language philosophy. To take one example, Cullen (2010) finds evidence that experimental participants sometimes are responding to different questions than the questions that experimenters intend to be asking. For example, from the open feedback given at the end of a survey-based experiment on Gettier scenarios, Cullen finds that while many participants are responding in ways that align with philosophers' prediction that the subject in a Gettier scenario doesn't know that something is the case, they are doing so for the *wrong reasons*:

> In my survey of students at Melbourne University, a number of subjects made their guesses about the purpose of the research explicit in the open feedback question. One example is [a student who]… guessed that the experiment was about "accepting that you can know anything". The experiment's conversational context raised her standard for answering "really knows" to truly Cartesian heights: she did produce the philosophers' response, viz., that the agent only believes that p, *but for an entirely different reason to philosophers'.* This was a very common theme in the students' feedback....Many who answered "only believes" claimed to have done so because, to quote another student's response, "nobody can ever truly KNOW anything". (Cullen 2010, n. 6)

Along similar lines, Turri (2013) found that the standard way that Gettier scenarios are presented to participants can obscure their significance. In a series of experiments using a more explicit, three-part structure of presentation in which different parts of Gettier scenarios are presented on

different screens and participants are instructed to keep track of the truth of the target proposition at each stage, Turri found that participants' judgments aligned with standard philosophical predictions about knowledge judgments more closely than in standard presentational formats.

And in the neighboring field of experimental linguistics, Schütze (2020) has recently argued that collecting responses to static surveys

> is useful only to the extent that you are confident you know what [participants] are basing those ratings on. If you are not very confident, you should find out, and often the best way is to ask them. (Schütze 2020, p. 190)

In this paper, we compare judgments made in traditional static surveys with judgments made at various points in Socratic questionnaires in which participants are asked, in different phases of the conversation, to (a) make judgments about a series of scenarios (the same as those presented in the traditional survey format), (b) explain why they gave the responses they did, and (c) respond to the fact that other participants give conflicting responses to the scenarios. By comparing the results from traditional surveys and the Socratic questionnaire, we can evaluate whether a more naturalistic conversational setting, in which participants can ask for clarification, discuss, hedge, and revise their initial responses, has an effect on judgments about philosophical thought experiments.[1]

We conducted three experiments to evaluate the effect of "format" (Socratic questionnaire vs. traditional static survey) on participants responses to a variety of philosophical scenarios. In our first, exploratory, experiment, we found some differences in how format affects participants' responses to different types of philosophical scenarios. In one type of scenario (a "trolley"-type scenario set in a nuclear power plant that is on the verge of meltdown in which participants are asked whether they would sacrifice one person to save many) we found no effect of the shift from a traditional survey format to a Socratic questionnaire. For "Travis"-type scenarios in which participants make truth value judgments about the color of walls that are made of white plaster but are painted brown, in our first experiment we found evidence that responses in the Socratic questionnaire differed from responses in a traditional static survey, but we did not find that effect in our second or third experiments. For "evidence-seeking"-type scenarios in which participants are asked to evaluate how long someone needs to think about an answer before she knows it when

[1] There are a few recent examples of philosophers discussing the possibility of conducting conversation-based experiments: Nadelhoffer and Nahmias (2007, n. 27) report that "Some graduate students at Florida State University recently ran a pilot study that involved presenting participants with various cases about intentional action and allowing them to discuss and debate the cases among themselves. At the end of the study they took further surveys to examine how the students' views changed (or did not change). And while the results were inconclusive—owing primarily to some problems with the design of the studies—their strategy is certainly one that could prove useful in the future". Hannon (2018) discusses the possibility of using conversational experiments to evaluate participants' reflective judgments about the plausibility of skeptical arguments. Andow (2016) surveys the potential uses of various qualitative methods in experimental philosophy, including the use of structured interviews.

the stakes are low, and again when the stakes are high, we found evidence in our first two experiments that responses participants give in the Socratic questionnaire differ from responses in a traditional static survey, but when we controlled for how participants were recruited (offering equal pay and identical descriptions of task and time requirements) in our third experiment, we did not find evidence of an effect of format on responses to the evidence-seeking scenarios. An aggregate data analysis of responses to the Travis-type and evidence-seeking type scenarios in all three experiments does not reveal any effect of format on participants' responses.

Although we did not find consistent quantitative evidence of an effect of conversational format on participants' responses, we present qualitative evidence, drawn from the details of participants' explanations and defenses of their judgments, that the way participants understand some scenarios and the questions they are being asked differs from the understanding that philosophers have assumed they have.

These qualitative findings indicate that dynamic, conversational experiments should be added to the philosopher's experimental repertoire, even if their employment ends up raising more questions about how we should investigate philosophical questions than they answer.

2. Experiment #1: Comparing a Socratic Questionnaire with a Traditional Survey

We designed an experiment to test the hypothesis that responses to scenarios presented in a Socratic questionnaire would differ from responses to the same scenarios presented in a traditional survey format.

2.1 Experimental Materials

We selected three pairs of scenarios from recent experimental literature in different areas of philosophy for participants to respond to.

Color: This scenario is a version of a "Travis case" (Travis 1989, Hansen and Chemla 2013, Grindrod et al. 2019), which asks participants to make truth value judgments about a statement, "The walls in our apartment are brown". In the RUG version of the color scenario, the conversational context involves finding a rug to match the walls of the speaker's apartment. In the GAS version of the color scenario, the conversational context involves an investigation into whether the walls of the apartment are made of brown or white plaster because brown plaster has been found to emit a poison gas.

Nuclear: This scenario is a version of a "trolley case" (Foot 1967, Thomson 1985, Christensen et al., 2014). Participants are asked to imagine that they work in a nuclear power plant that is about to suffer a meltdown, threatening the lives of everyone in a nearby town. In the INSERT scenario (an "impersonal case"), participants are told that they can stop the meltdown by inserting material manually, which will kill one employee but will save the thousands of inhabitants of the town. In

the PUSH scenario (a "personal" case), participants are told that the only way they can stop the meltdown and save the inhabitants of the town is by pushing a foreman into the cooling circuits, killing him. Participants were asked whether they would insert material in the INSERT scenario, and whether they would push the foreman in the PUSH scenario.

Game Show: This is an "evidence-seeking" design, inspired by Pinillos (2012) and used in experiments in Francis et al. (2019) to evaluate the stakes-sensitivity of knowledge judgments. In the LOW stakes Game Show scenario, the subject of the scenario is asked a trivia question about the capital of Tanzania when only $1 is at stake. In the HIGH stakes Game Show scenario, the subject stands to win or lose $1,000,000 depending on her answer. Participants were asked the same question in both the low and high stakes scenarios: How long does the subject need to think about her answer before she knows that the capital of Tanzania is Dodoma?

We chose these particular scenarios because we have done previous experimental work with them and were interested in whether the effects of context that they were originally designed to detect would replicate in the Socratic questionnaires format, and because we thought they were different enough from each other in terms of the relevant factors contributing to contextual effects (conversational "point" or question under discussion in the "Travis"-style scenarios, personal vs. impersonal considerations in the "trolley"-type scenarios, and stakes of being wrong in the "evidence-seeking" scenarios) that they had a chance of revealing variability in how conversational format affects participants' responses.

These scenarios and prompts appeared unchanged in both the Socratic questionnaire and the traditional survey.[2]

2.2 Design of the Socratic Questionnaire

The conversational component of the experiment was implemented using ChatPlat, an online tool for placing participants into interactive chats and archiving the text of the ensuing conversations (see **Figure 1** for the appearance of the ChatPlat interface).[3] One participant at a time entered into a chat with the Research Assistant (RA), who used a pre-written script to guide the conversation.[4]

[2] The full scenarios and prompts are available here: https://osf.io/9mnha/?view_only=a3cf063dcbf64f0aabf08ee9b9eed04c

[3] For an example of a study in experimental psychology that uses ChatPlat, see Huang et al. (2017).

[4] The complete script is available here: https://osf.io/4b9yk/?view_only=0860ffe03c4845eeb09c03f5497d6935

(19:50:35) System: >> All chat participants have arrived. You may now chat!
(19:50:35) System: >> User 1 has Connected
(19:50:49) User 1: Hello?
(19:51:17) Admin: Hi! Thanks for participating in this study. Feel free to ask questions in the chat at any point.
(19:51:26) Admin: You will be given six short scenarios to read, and I'll be asking for your reactions to them. After you respond to the six scenarios, I'll ask you to explain why you responded in the way that you did, and then I'll ask some further questions about what you thought about the scenarios. Finally, I will ask you some basic demographic information.
(19:51:30) Admin: Does that sound okay?
(19:51:38) User 1: 😃 😃
(19:52:05) Admin: Okay Great. Now just some formalities to begin with.

Figure 1: ChatPlat interface ("Admin" is the RA, "User 1" is the participant)

The scenarios were presented randomly in counterbalanced blocks to minimize the number of times two versions of the same scenario type (Gameshow, Nuclear, Color) appeared side-by-side (see **Figure 2**).

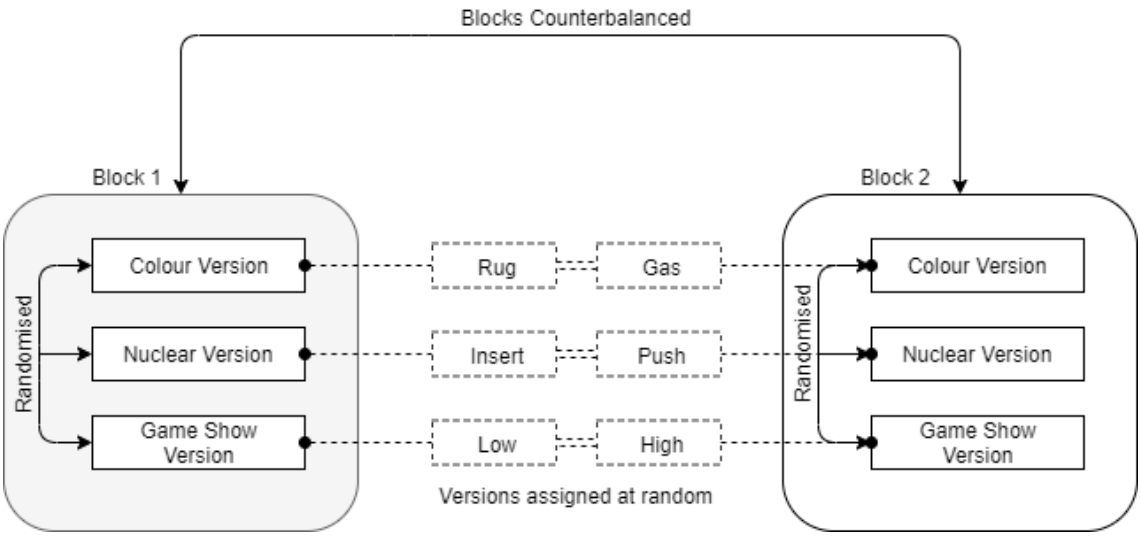


Figure 2: Diagram showing the experimental design used in both the Socratic and traditional questionnaire. Blocks contain each type of scenario (Colour; Moral; Stakes) presented randomly. Scenario versions (Rug versus Gas; Insert versus Push; Low versus High) are then randomly assigned to each block. Block order is counterbalanced.

The resulting 40 chat transcripts from experiment #1 run to 98,340 words and are available in their entirety online: https://osf.io/ys96n/?view_only=3a574b031d3f4e38a00e9af41ffd990f

The script, loosely inspired by the structure of the conversational experiments in Trouche et al. (2014), is divided into five phases:

I. *Introduction*: Participants were asked for their consent to participate and given instructions for how to respond.

II. *Initial responses to scenarios*: Participants were given the six scenarios in random order and asked for their responses.

III. *Requests for justification*: The RA noted how the participants responded in their initial responses and offered pre-formulated requests for justification depending on how they responded. For example, for participants who gave different responses in the two Game Show scenarios, the RA asked: "I notice that you gave different responses to each of the game show scenarios. Would you say a little bit about how you decided on those responses?" If participants gave the same answer to the scenarios, the RA asked: "I notice that you gave the same responses to both the game show scenarios. Would you say a little bit about how you decided on those responses?"[5]

IV. *Pushback*: After participants gave justifications for how they responded to the scenarios, the RA offered some gentle pushback against those justifications. The pushback consisted in noting that some participants gave different responses to each of the scenarios, using a standard format. So, for example if participants gave different responses to the two Game Show scenarios and justified those responses, the RA pushed back as follows:

"So you have explained why you gave different responses to the two game show scenarios. Some people think that changing the amount of money involved in the scenario does not change how long someone must consider an answer before they know it. How would you respond to this point of view?"

And if participants gave the same answer to the two Game Show scenarios and justified that response, the RA would push back as follows:

"So you have explained why you gave the same responses to the two game show scenarios. Some people think that changing the amount of money involved in the scenario changes how long someone must consider an answer before they know it. How would you respond to this point of view?"[6]

[5] One participant (Experiment 1, Chat #10) responded so slowly to the scenarios that they did not reach the justification or pushback stages of the Socratic questionnaire.
[6] One participant (Experiment 1, Chat #11) responded too slowly to reach the pushback stage of the conversation in the allotted time.

V. *Demographic information and conclusion*: The RA asked participants for demographic information (age, gender, philosophical training), thanked them for their participation, and directed them to a url that confirmed their completion of the conversation.

The design of the Socratic questionnaire makes it possible to track participants' judgments across different phases of the conversation. For example, consider the flow of Chat #16 in **Table 1:**

Scenario	**Initial Response Phase**	**Justification Phase**	**Pushback Phase**
Color, Rug	T		
Color, Gas	F	Admits error and revises (to T/T)	Confirms revision
Game Show, Low	15		
Game Show, High	30	Personal experience, no deontic modal ("will")[7]	Hedge (depends on individual differences)
Nuclear, Insert	Y		
Nuclear, Push	Y	Better consequences	Better consequences

Table 1: Schematic Representation of a Socratic Questionnaire

The participant in Chat #16 offers his initial responses, and then revises his judgment about the Gas version of the Color scenario when asked to justify his responses. He confirms that revision in the final, "pushback" stage of the conversation. He hedges, rather than revises, his responses to the Game Show scenarios in response to pushback (the hedge is his observation that different people will react to the scenario differently), and he sticks with his initial justification of his responses to the Nuclear scenarios throughout the conversation. The division of the Socratic questionnaire into phases makes it possible to evaluate how the evolving conversation affects participants' responses.

2.3 Coding Responses to the Socratic Questionnaire

Initial responses to each case were recorded. The conversations were then independently coded by the authors Hansen and Francis, taking note of where and when participants revised their answers, and noting other interesting responses that will be discussed below, in §3. We used an "inductive" approach to coding the conversations, which means that we didn't begin with a ready-made set of

[7] We discuss the significance of the presence or absence of deontic modals in justifications of the Game Show scenario in §3.3.

categories but extracted relevant categories from the text of the conversations (Chenail 2008). Hansen compiled the independent codes into a single uniform list, and then Hansen and Francis each independently coded 20 of the conversations.[8] For the purposes of the quantitative analysis discussed in §2.7 below, the most relevant codes that emerged from reading the conversations concern *revisions* to participants' initial responses and instances of participants clearly *misunderstanding* either the scenario or prompt or indicating that they were responding to a version of the scenario that they had *changed* from what was explicitly stated.

Revisions

Here is an example of a revision that happens in the justification phase of conversation #39, in which a participant revises his response to the High Stakes version of the game show scenario from 0 to 60 seconds (all text is verbatim from the chat):

(11:05:25) Admin: I notice that you gave the same responses to both the game show scenarios. Would you say a little bit about how you decided on those responses?
(11:07:36) User 1: For the scenario where Emma plays it's almost like she has nothing to lose, so she doesn't need to spend time to think
(11:08:15) User 1: For Tracy the risk was high and i feel like i would have given a different answer after i read the Emma scenario
(11:08:32) Admin: Okay what would that be?
(11:09:35) User 1: probably around 60 seconds, enough to think of some more possible capitals but not enough to worry about her first thought, which turned out to be correct as well

The participant explains his revised response in the pushback phase of the conversation, saying that when he gave his initial response he misunderstood what he was being asked:

(11:22:24) Admin: So you have explained why you gave the same responses to the two game show scenarios. Some people think that changing the amount of money involved in the scenario changes how long someone must consider an answer before they know it. How would you respond to this point of view?
(11:23:21) User 1: I share this way of thinking as well, but i think i did not completely understand at first
(11:23:41) Admin: Okay tell me a little more about that. What didn't you understand?
(11:24:54) User 1: At first i thought i should answer how much time she has on the game show, but we had no information about that
(11:25:26) User 1: I now find it really pointless that my answer was 0 seconds, at least for Tracy
(11:25:57) Admin: Why's that?
(11:27:07) User 1: Considering how unsure she was for the answer I believe no human would instantly give the answer at risk of losing 1 million dollars

[8] The complete set of codes the authors used to characterize the conversations can be found here, with the uniform codes highlighted in color: https://osf.io/9vksg/?view_only=013c3b45900143fe9fd97de838401937

Revisions occurred in 7 of the 40 Socratic questionnaires we conducted in experiment #1.[9]

Imagining additional material in the scenarios

Some philosophers have wondered to what extent participants are responding to experimental scenarios not only in terms of the material that is explicitly encoded in those scenarios and prompts, but also in terms of idiosyncratic material that participants imagine being there (Boyd and Nagel 2014, Dinges and Zakkou 2019, Horvath 2015, Sosa 2009). Austin (1956) makes this observation in relation to the cases used by ordinary language philosophers:

> When we come down to cases, it transpires in the very great majority that what we had thought was our wanting to say different things of and in the same situation was really not so—we had simply imagined the situation slightly differently: which is all too easy to do, because of course no situation (and we are dealing with imagined situations) is ever "completely" described. (Austin 1956, pp. 9–10)

Another advantage of the Socratic questionnaire is that it gives us a window onto when and what additional material is being added by participants.

One conspicuous example of imagining a scenario differently than it is written occurs in Chat #4, where the participant explains her decision to kill one to save many in both versions of the nuclear scenario in terms of imagining that the person being sacrificed in each scenario is "a bad person", which "made it easier to sacrifice him":

(16:33:16) Admin: So you have explained why you would push the foreman to his death in one nuclear scenario and why you would also release the liquid nitrogen into a chamber (killing a worker) in the other nuclear scenario. Some people think that pushing the foreman is not something you should do, even in such an extreme case. How would you respond to this point of view?
(16:34:24) User 1: I chose to assume that the number of people saved was correct, hence I decided that it was a justified sacrifice (also I imagined the foreman as a bad person)
(16:34:36) User 1: this made it easier to sacrifice him.

But by far the most frequent scenario type that participants imagine differently than they are explicitly written are the color scenarios. The most common additional material that participants added (7 out of 40 participants) is the fact that the subject of the scenario (Hugo) *doesn't know* that the walls of his apartment are made of white plaster. That addition seems to make it easier for participants to hear what he says, that the walls are brown, as true, even when the conversational context concerns the color of the plaster. Here is an example, from Chat #17, of that kind of addition in a participant's justification for their initial responses:

(14:02:41) Admin: Let's look at your answers to the color scenarios.

[9] The coded conversations can be found at this link, with revisions highlighted in red: https://osf.io/c2w3g/?view_only=bb798b4edb2542fcbf92608856d57494

(14:02:52) Admin: You said that when Hugo says “The walls are brown” in both scenarios, what he says is true. Would you explain how you decided on those responses?
(14:04:58) User 1: On both scenarios, Hugo responded with what he knew about the walls. Although the walls are made of white plaster, he doesn't know that and claiming that the walls are brown is a fact to him and he doesn't know any better.

Nowhere in either color scenario does it specify whether or not Hugo knows that the walls are made of white plaster. Other times participants add to the scenario that Hugo is lying to the building superintendent (Chat #6), that Hugo has misunderstood the superintendent’s question (Chat #12), or that Hugo doesn’t know why the superintendent is asking about the walls (Chats #24 and #25). 14 of the 40 Socratic questionnaires in experiment #1 feature participants imagining changes to the color scenarios.[10]

Failing to understand the scenarios or prompts

Imagining changes to the scenarios as presented sits on a continuum with clear cases in which participants simply *misunderstand* what the scenario is describing. The participant in Chat #7, for example, justifies his judgment that Hugo’s statement in the Rug version of the Color scenario is true on the grounds that he agrees with something that Hugo doesn’t explicitly say, but rather implicates, namely that brown and orange don’t match:

(19:51:35) Admin: You said that when Hugo says “The walls are brown” in both scenarios, what he says is true. Would you explain how you decided on those responses?
(19:52:16) User 1: In the first scenario - with the gas - Wall was painted as brown so it was brown for me, white was the plaster, not wall itself.
(19:52:47) Admin: Okay.
(19:53:11) User 1: In the second scenario - With the rug - Hugo said that wall is brown and he doesn't like orange rug. I decided that Hugo is right, orange doesn't very much brown for me.
(19:53:32) User 1: match* sorry

That indicates that this participant hasn’t understood the prompt, which asks for a judgment about whether the claim “The walls in our apartment are brown” is true or false, not whether the participant agrees with the implication that an orange rug won’t match brown walls. Six Socratic questionnaires feature misunderstandings of the color scenarios or prompts.[11]

We will discuss another other interesting qualitative discovery that emerged from the coding of the Socratic questionnaires in §7.

2.4 Design of the Traditional Survey

[10] In the overview of the coded conversations (linked to in footnote 9), changes to scenarios are highlighted in green.
[11] In the overview of the coded conversations (linked to in footnote 9), misunderstandings are highlighted in blue.

The second part of experiment #1 involved a traditional survey, in which a different set of participants were given the same six scenarios as in the Socratic questionnaires and prompted to enter their responses into a text box. Scenarios were randomized using the same approach as in the Socratic questionnaire (see **Figure 2**), and the same demographic information was collected at the end of the survey.

2.5 Participants

Socratic questionnaires

40 participants (17 females, 22 males, 1 undisclosed) between 18 and 54 years old ($M = 26.69$, $SD = 8.47$ years) were recruited from Prolific, with the requirement that they be fluent speakers of English, and paid $9 each for a chat lasting an average of 48 minutes and 24 seconds (an average hourly rate of $11.16).

Because of the exploratory nature of experiment #1, we didn't know what kind or size of effects we would expect to find when comparing the Socratic questionnaire with the traditional survey. As such, we did not run a power analysis to determine a sample size. We chose to recruit 40 participants for both the Socratic questionnaire and the traditional survey based on the fact that the study of color and knowledge scenarios in Hansen and Chemla (2013) found contextual effects with that many participants, and because 40 participants was a manageable number of conversations to conduct given time and funding constraints with the RA.

Traditional survey

41 participants (11 females, 30 males) between 18 and 65 years old ($M = 26.93$, $SD = 10.48$ years) were recruited on Prolific, with the requirement that they be fluent speakers of English and that they not have participated in the Socratic questionnaire study.[12] Participants were then directed to the Qualtrics survey. Participants were paid $2 on completion of the survey, which took approximately nine minutes to complete (average awards came to $14.61/hour). Responses to each case were recorded in the same way as the initial responses in the Socratic questionnaire.

This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants, in both the Socratic questionnaire and traditional survey formats.

2.6 Data Analysis Outline

[12] We aimed to recruit 40 participants for the traditional static survey but a sampling error led to recruiting 41.

I Initial responses: Initial responses to each scenario (Gameshow, Nuclear, Color) were compared across cases (e.g., Low Stakes, High Stakes) and across formats (Socratic vs Traditional survey).

II Revised responses: The same analyses were repeated but using any revised responses recorded in the Socratic format during the justification and pushback phases of the conversation.

III: Misunderstanding or changing scenario: The same analyses were repeated but with any responses that showed a misunderstanding of the scenario or any responses in which the content of scenarios was changed removed.

2.7 Results

In this section, we present the results of experiment #1, broken down by scenario type.

Color

Overall, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald ($\chi^2(1) = 18.39$, $p < .001$, 95% Wald CI [-3.37, -1.13], $V = .34$) with the proportion of responses saying that Hugo's claim ("The walls in our apartment are brown") is *true* significantly higher in the Rug case. There was no difference in responses across formats and no interaction of format x case (*ps* >. 230). These results remain the same when factoring in revised responses (2 identified) (see **Figure 3A**) but when we remove examples of changed or misunderstood scenarios, these results change (**see Figure 4**).[13]

As before, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald ($\chi^2(1) = 11.74$, $p = .001$) and no difference between the formats ($p = .325$). However, there was now a significant interaction of format x case, (Wald ($\chi^2(1) = 3.84$, $p = .050$, 95% Wald CI [0.00, 3.27], $V = .17$) with only the Traditional survey format showing the context effect (the proportion of *true* responses was significantly higher in the Rug case) ($p < .001$). The context effect was no longer present in the Socratic survey format ($p = .770$) (see **Figure 4**).

Nuclear

Overall, we found a personal/impersonal effect in the moral scenarios (Insert case; Push case) across both formats (Socratic; Traditional), (Wald ($\chi^2(1) = 12.86$, $p < .001$, 95% Wald CI [0.19, 2.03], $V = .28$) with the proportion of responses indicating that one person should be sacrificed to save many significantly higher in the Insert case (impersonal) case. There was no difference in responses across formats and no interaction of format x case (*ps* >. 698). These results remain the same when factoring in revised responses (2 identified) (see **Figure 3B**). The above analysis

[13] We identified 17 participants in the Socratic survey format who either misunderstood at least one scenario (11 instances total) or changed at least one scenario (e.g., added additional features) (13 instances total). Sometimes a single participant changed and misunderstood a scenario at different points in the conversation.

remains the same when factoring in revised responses and when removing responses where participants added materials or misunderstood a scenario **(see Figure 3B**).

Game Show

Overall, we found a stakes effect in the game show scenarios (Low Stakes case; High Stakes case) across both formats (Socratic; Traditional), (Wald ($\chi^2(1) = 72.00$, p <.001, 95% Wald CI [-1.47, -.72], $V = .69$)[14] with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was no difference in responses across formats and no interaction of format x case (ps > .056).

However, when revised responses to the Socratic questionnaire are taken into account, the analysis changes:[15]

The stakes effect remains across both formats (Socratic; Traditional), (Wald ($\chi^2(1) = 73.65$, p <.001, 95% Wald CI [0.16, 1.62], $V = .70$)[3] with individuals stating that more seconds are required to know the correct answer in the High Stakes case (p <.001). As before, there is no interaction of format x case (p = .074). However after taking into account revisions, there is a difference in responses between formats (Socratic; Traditional), (Wald ($\chi^2(1) = 4.18$, p =.041, 95% Wald CI [-1.47, -.72], $V = .17$) with the number of seconds being higher in the Socratic format across both cases (see **Figure 3C**).

[14] Using Generalised Estimating Equation (poisson distribution) and including zero values. Note that results remain the same when a gamma distribution is used with zero values excluded.

[15]Raters identified 8 revised responses (4 in the Game Show cases). For non-binary responses, as in the Game Show scenario, revised responses were handled as follows: a) a revision stating that there was *no epistemic change* between cases meant that the number of seconds given in the high stakes case was made equal to that of the low stakes case (1 revision), and b) a revision stating that *more time would be needed in the higher stakes case* meant that the number of seconds given in the high stakes case was amended to the format/case average (3 revisions).

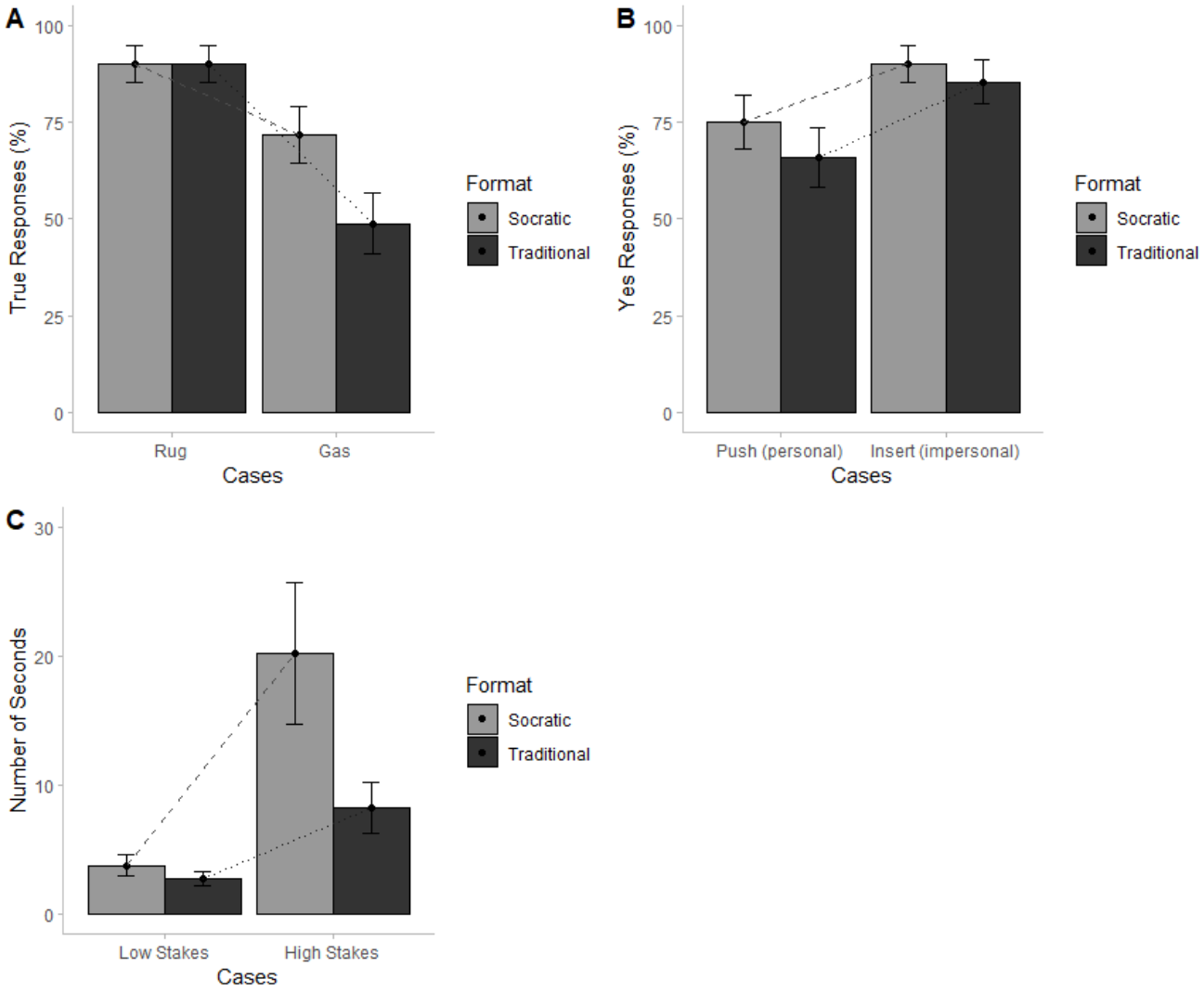


Figure 3: Responses to the different scenarios taking revised responses into account. A) Responses to the Color scenarios, B) Responses to the Nuclear scenario, C) Responses to the Game Show scenario. Differences in responses to varieties of scenarios (context effects; personal/impersonal effects; stakes effects) are found in all scenarios across formats. In the Game Show scenario, there is also a difference in responses between the formats. Error bars represent +/- 1 SE.

Given that participants could also respond to the Game Show scenarios by stating that a) the protagonist can never know no matter how many times they check (never) or by stating that the protagonist knows without having to check (zero), we also analysed these responses. There was no difference in the number of zero responses across formats ($p = .537$) and this remained the same when factoring in revised responses. However, there were a higher number of "never" answers given in response to the Traditional survey versus the Socratic questionnaire ($\chi^2(1) = 5.44$, $p = .020$, 95% CI [1.21, 3.01], $\varphi = .77$).

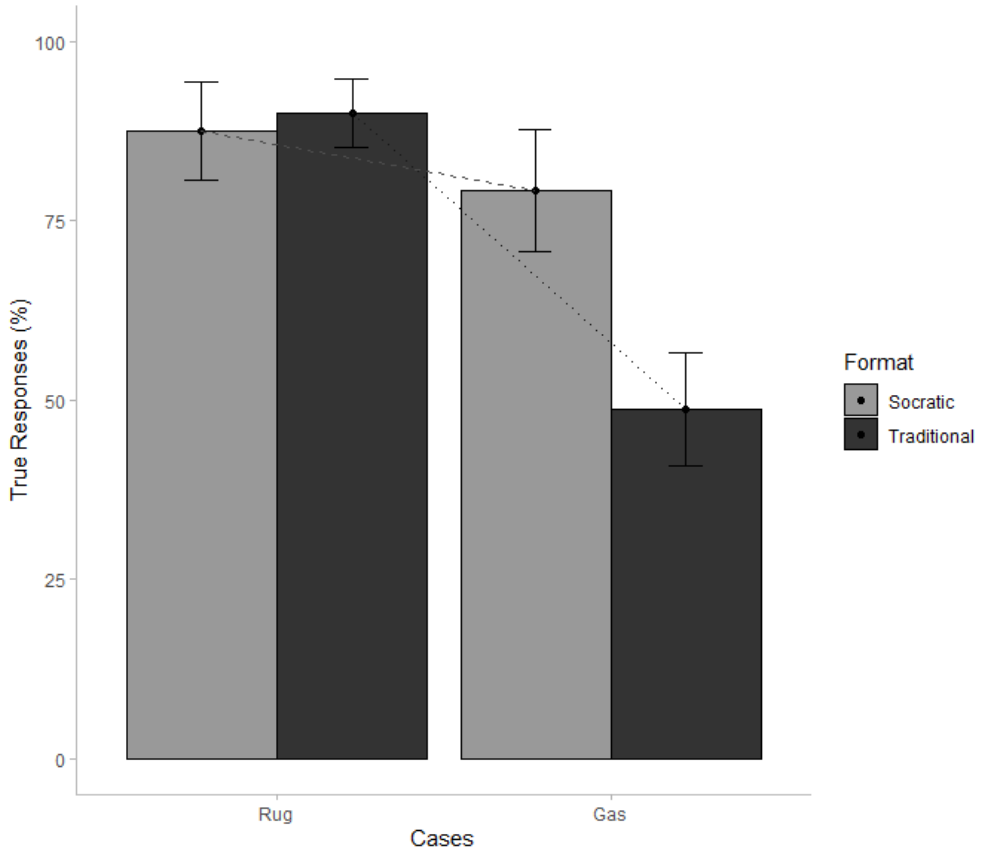


Figure 4*:* Responses to the Color scenarios when participants who misunderstand or change the scenario are removed from the Socratic format. The context effect is present in the Traditional survey format only with a greater percentage of true responses given in response to the Rug case. Error bars represent +/- 1 SE.

2.8 Discussion

Replications of previous findings (before taking into account revisions, changes, and misunderstandings)

We found a main effect of scenario in all three scenario types, meaning that the contextual manipulation in the color scenarios (RUG and GAS), the impersonal/personal manipulation in the nuclear scenarios (INSERT and PUSH), and the stakes manipulation in the game show scenarios (LOW and HIGH) all had a significant effect on participants' responses. That replicates previous findings of contextual effects on color judgments (Hansen and Chemla 2013, Grindrod et al. 2019), of impersonal/personal effects on responses to trolley cases (e.g., Greene et al., 2001; 2004), and of stakes effects on judgments about knowledge using "evidence-seeking" questions (Pinillos 2012, Pinillos and Simpson 2014, Francis et al. 2019).

A Conversational Effect in the Game Show Scenario and Color Scenario When Taking into Account Revisions, Changes, and Misunderstandings

In the game show scenarios, we found a main effect of format, meaning that responses differed significantly depending on whether they occurred in the Socratic questionnaire or in the traditional survey, but only when revised responses were taken into consideration. When considering participants' initial responses, in contrast, we found no difference between the Socratic questionnaire and the traditional survey in any of the scenario types we examined. A similar pattern was observed in the color scenarios, where we found no effect of format on the initial responses, but once responses that were based on misunderstandings of the scenario or changes being made to the scenario were removed, we observed an effect of format on the color scenarios. We did not find any effect of format on responses to the nuclear scenarios even taking into consideration revisions, changes, and misunderstandings.

Effect of Format on the Game Show Scenarios

Part C of **Figure 3** shows how the stakes effect in the game show scenarios is significantly different in the Socratic questionnaire format than in the traditional survey, once revisions are taken into consideration: The average number of seconds participants say are needed before the protagonist knows that the capital of Tanzania is Dodoma in both the HIGH and LOW stakes version of the scenario is significantly greater in the Socratic questionnaire than in the traditional survey. In addition to that difference, there are also significantly fewer "never" responses to the game show scenarios in the Socratic questionnaire than in the traditional survey.

Effect of Format on the Color Scenarios

Figure 4 shows the effect of conversation on the color scenarios once misunderstandings and responses that change the scenario are removed. The effect of context on truth value judgments is present only in the traditional survey format—it disappears in the Socratic questionnaire. The fact that we find no significant difference in truth value judgments in the RUG and GAS versions of the color scenarios once we remove responses that are misunderstandings or make changes to the scenario might be taken to lend some support to those philosophers who have worried that some apparent effects found in experimental philosophy surveys shouldn't be trusted because they incorporate performance errors. However, it is important to note that with misunderstandings and responses that change the scenario removed, the sample size included in the analysis is reduced (N = 24) which would inevitably reduce statistical power. Any interpretation of this result should therefore be tentative. And as we will discuss below, we did not find this effect in our experiments #2 and #3.

The Reflectiveness Defense

Kneer et al. (2021) examine whether various manipulations of "reflectiveness" have an effect on participants' judgments about a variety of philosophical thought experiments. Their aim is to evaluate an argument against the use of surveys in experimental philosophy they call the "reflectiveness defense" of armchair philosophy, which claims that the judgments elicited by surveys are not sufficiently reflective to give us any insight into philosophical questions: unlike highly reflective philosophers, survey respondents are "shooting from the hip" and responding to philosophical thought experiments without "the necessary deliberative care".[18] In experiments designed to prompt greater reflectiveness, which (a) forced participants to delay their responses, (b) offered financial incentives for correct responses, (c) asked participants to give reasons to support their responses, and (d) primed participants for analytical thinking by having them complete the Cognitive Reflection Test, they did not find any difference between responses to a variety of philosophical thought experiments and responses to surveys without reflectiveness manipulations.

Kneer et al. concede that the manipulations of reflectiveness that they investigate concern a "thin characterization of reflective judgment". They allow that there are "thicker" ways of characterizing reflective judgment, like the "Dialogue" conception discussed by Kauppinen in his criticism of experimental philosophy:

> [T]here is no way for a philosopher to ascertain how people would respond [reflectively] without (...) entering into dialogue with them, varying examples, teasing out implications, presenting alternative interpretations to choose from to separate the semantic and the pragmatic, and so on. I will call this approach the Dialogue Model of the epistemology of folk concepts. (Kauppinen 2007, p. 109)

In one respect, our experiment #1 provides additional support for Kneer et al.'s failure to find an effect of "thin" reflectiveness on responses to philosophical scenarios. Our Socratic questionnaires incorporated two different factors that align with the reflectiveness manipulations in Kneer et al: First, participants were informed during the Introduction phase of the Socratic questionnaire that they would be participating in a conversational study that would involve explaining their responses. That is similar to the "reasons" condition in Kneer et al, which presented "a screen which instructed participants that they would have to provide detailed explanations of their answers" (Kneer et al., 2021, §3). Second, participants were instructed to tell the RA when they had finished reading a scenario. Only when participants indicated that they were finished did the RA give them the relevant prompt. That has the effect of slowing down response times, in a way similar to the "forced delay" manipulation in Kneer et al.[19] Even with those two factors in place, we did not find an effect of format between the traditional survey responses and responses to the

[18] For discussions of the role of reflectiveness in experimental philosophy, see de Bruin (2021), Kauppinen (2007), Liao (2008), Ludwig (2007), Hannon (2018), Horvath (2010), and Nado (2015).

[19] The average time it took participants to respond to the six versions of the scenarios in the Socratic questionnaire (excluding instructions and justification and pushback phases) was just over 15 minutes, which is six minutes longer than it took for participants to complete the whole survey, including instructions.

Socratic questionnaires *when looking only at initial responses*--those responses participants gave before being asked to explain and respond to a mild challenge to their responses.

The support for Kneer et al.'s findings provded by the initial responses to Socratic questionnaires in our experiment #1 might be taken to be complicated by the fact that once we take into consideration *revisions*, *changes*, and *misunderstandings* of the scenarios, we did find differences between responses to the scenarios presented in the Socratic questionnaires and in the traditional survey. That might be understood to lend some support to Kaupinnen's (2007) "Dialogue" version of the "thick reflectiveness" challenge to experimental philosophy. But, as we will discuss below, we didn't find these differences in experiments #2 and #3 or in our aggregate data analysis of responses to all three experiments. Overall our results therefore align with Kneer et al's findings and don't lend support to the "Dialogue" version of the "thick reflectiveness" challenge.

3. Experiment #2: A More Rigorous Investigation of the Effect of Socratic Questionnaires vs. Traditional Surveys

Our experiment #1 was exploratory: we didn't know before running the experiment what kinds of responses participants would give in the course of a Socratic questionnaire, and we had to make an informed guess (not based on a power analysis) about how many participants we would need to detect an effect of format. Our findings in experiment #1 were suggestive: they seem to point towards effects of conversational format in the Color and Game Show scenarios. We ran a second experiment with the aim to evaluate whether our findings from the more exploratory experiment #1 would replicate under more rigorous conditions. In experiment #2, the number of participants we recruited was determined by a power analysis based on the results from the Game Show scenarios in experiment #1, and we knew ahead of time what kinds of responses in the Socratic questionnaires we were interested in counting, namely revisions, misunderstandings, and changes to the scenarios. And in experiment #2, we employed an independent coder to check that our classifications of these responses were not idiosyncratic. Since we did not find any effect of conversational format on the nuclear scenarios in our first experiment, we dropped those scenarios from experiment #2 to focus on the likeliest scenarios for which an effect of format would be found, namely the game show and color scenarios. Other than those changes, we used the same experimental materials and overall design, comparing responses to the scenarios given in a Socratic questionnaire with responses to the scenarios given in a traditional static survey.

3.1 Participants

Socratic questionnaires

In experiment #2, 47 participants[20] (13 Females, 32 Males, 2 undisclosed, M_{age} = 25.18, SD_{age} = 8.71) were recruited from Prolific, with the requirement that they be fluent speakers of English and not have participated in our previous studies, and paid $8 each for a chat lasting an average of 34 minutes (an average hourly rate of $14.11). The resulting 47 chat transcripts from experiment #2 run to 86,229 words and are available in their entirety online: https://osf.io/8cyxe/?view_only=59d1ac26d32a421bb5b38cb2cadd6048

Traditional survey

48 participants (22 Females, 24 Males, 2 identifying as "other", M_{age} = 25.75, SD_{age} = 6.54) were recruited on Prolific, with the requirement that they be fluent speakers of English and that they not have participated in our previous studies.[21] Participants were then directed to the Qualtrics survey. Participants were paid $1.50 on completion of the survey, which took an average of 6 minutes to complete (an average hourly rate of $15/hour). Responses to each case were recorded in the same way as the initial responses in the Socratic questionnaire.

This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants, in both the Socratic questionnaire and traditional survey formats.

3.2 Results

Inter-rater reliability

We trained an independent coder (who was blind to the hypotheses we were testing) on practice items from experiment #1 and both the independent coder and a member of the research team coded all of the scripts from ChatPlat independently.[22] To determine interrater reliability, we calculated percentage agreement scores and calculated Cohen's Kappa for each set of codes at both justification and pushback stages. Across all codes and stages, average agreement between raters was 88.11% with moderate to perfect agreement in coding of deontic modals, hedges, and revisions ($\varkappa$s = .55 – 1.00) and fair agreement in codings of misunderstandings ($\varkappa$s > .29).[23]

---

[20] Using G*Power, we determined that a total sample size of N = 94 (N = 47 in each group) would be sufficient to detect a medium-sized effect with a power of .80 (alpha = .05). Supporting this, using a simulation-based power analysis (Lakens & Caldwell, 2019) based on a 2b*2w design, we determined that N = 47 in each cell would be sufficient to detect a stakes effect (power = 100%) and context effect (power = 78%) in experiment #2 (2000 simulations performed using means from experiment #1 and alpha criterion of .05).

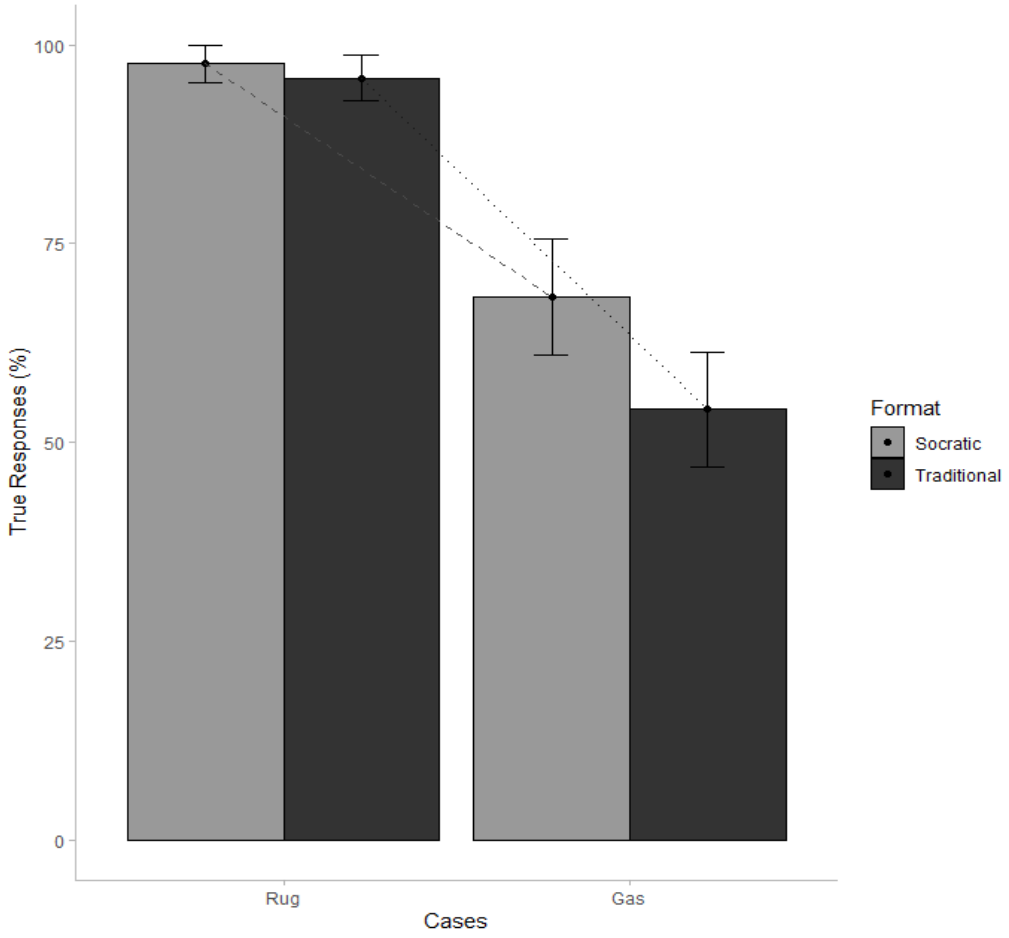
[21] We aimed to recruit 47 participants for the traditional static survey but a sampling error led to recruiting 48.

[22] The instructions that we gave to the independent coder can be found here: https://osf.io/k598r/?view_only=055c42b5add24e889f454b1266c5b4af

[23] Note that in all experiments, sets of codes where >90% of cases fell into a single nominal category, kappa was not calculated as this is problematic in skewed data. However, McNemar's tests confirmed no statistically significant differences between raters. The spreadsheet with the conversations coded by the independent coder can be found here: https://osf.io/c4hs2/?view_only=f6f090a3c9f94a9881f7e15938db56f4; the spreadsheet with the conversations

Color

Overall, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 21.91, p <.001, 95% Wald CI [-4.49, -1.45], V = .34) with the proportion of responses saying that Hugo's claim ("The walls in our apartment are brown") is *true* significantly higher in the Rug case. There was no difference in responses across formats and no interaction of format x case (ps >. 290). These results do not change even when revised responses are factored in (4 revisions were coded), and when examples of changed or misunderstood scenarios are excluded (5 examples were coded). As before, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 22.17, p <.001, 95% Wald CI [-4.49, -1.45], V = .36) and no difference between the formats (p =.394) and no interaction (p = .986) (**see Figure 5)**.



coded by the experimenters can be found here: https://osf.io/97xjq/?view_only=06f14fd2ce704f4481a92cf63e17e84e;
The full interrater reliability analysis can be found here: https://osf.io/27wpq/?view_only=f341553e119c4481849c713290cb1f44

Figure 5: Responses to the Color scenarios when participants who misunderstand or change the scenario are removed from the Socratic format in experiment #2. Unlike in experiment #1, we did not find an effect of format for the Color scenarios when responses that involved changes or misunderstandings were removed. Error bars represent +/- 1 SE.

Game Show

Overall, we found a stakes effect in the game show scenarios (Low Stakes case; High Stakes case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 59.43, p <.001, 95% Wald CI [-1.86, -0.86], *V* = .67) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was a statistically significant main effect of format (Socratic; Traditional), (Wald (χ2(1) = 8.30, p =.004, 95% Wald CI [0.02, 1.36], *V* = .25) with individuals stating that more seconds are needed in the Socratic format than in the traditional format across both cases.

When revised responses to the Socratic questionnaire are taken into account, the analysis remains essentially the same (only one example of a revision to the Game Show scenarios was coded in experiment #2): We found a stakes effect in the Game Show scenarios (Low stakes case; High Stakes case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 59.81, p <.001, 95% Wald CI [-1.86, -0.86], *V* = .67) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was a statistically significant main effect of format (Socratic; Traditional), (Wald (χ2(1) = 8.33, p =.004, 95% Wald CI [0.02, 1.35], *V* = .25) with individuals stating that more seconds are needed in the Socratic format than in the traditional format across both cases **(See Figure 6.)**.

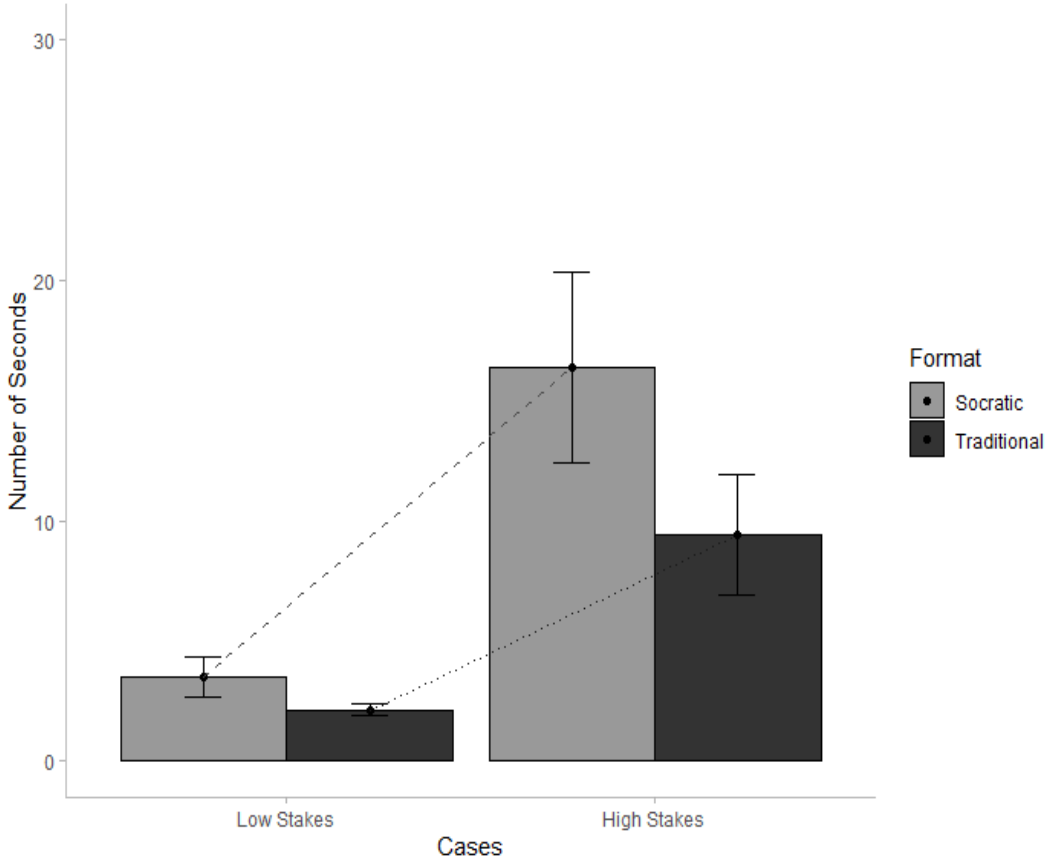


Figure 6: Differences in responses to the Game Show scenarios in experiment #2 (taking revision into consideration). There is an effect of stakes in both formats (Socratic questionnaire and Traditional survey), and there is an effect of format: participants say that more seconds are needed before the subject knows the answer in the Socratic questionnaire format than in the traditional survey. Error bars represent +/- 1 SE.

3.3 Discussion

Main findings: Failure to replicate effect of conversational format on the color scenarios; evidence of an effect of conversational format on Game Show scenarios

Even after taking revisions, misunderstandings, and changes to the color scenarios into consideration, we did not find an effect of conversational format on responses to the color scenarios in experiment #2. But we did find evidence of an effect of conversational format on responses to the Game Show scenarios in experiment #2, though we found it in a different place than we found such an effect in experiment #1: In experiment #1, the effect of conversational format appeared only once we took revisions into consideration, while in experiment #2, the effect of conversational format was present already in participants' initial responses to the Game Show scenarios.

4. Experiment #3: A New Quasi-Dynamic Survey and Controls for Features of Participant Recruitment and Attention

Given the mixed findings from our first two experiments, we conducted a third experiment with two primary aims. First, we wanted to see to what extent the effects of format that we found on responses to the Game Show scenarios in experiments #1 and #2 would replicate. Since those effects of format arose at different points in the Socratic questionnaires, we hoped a third experiment would reveal whether later revisions in participants' responses were essential to the difference, or whether the difference was already present from the very beginning of the Socratic questionnaire.

Second, we wanted to control for two potential confounds in our findings of effects of format in experiments #1 and #2. First, the Socratic questionnaires in experiments #1 and #2 differ from their counterpart static surveys in having both space for participants to explain their responses and space to respond to light pushback to their explanations. As we discuss in §8, below, Socratic questionnaires quickly reveal if participants aren't paying attention or are misunderstanding what they are being asked to do. But in our traditional surveys in experiments #1 and #2, we didn't include any attention checks. It might therefore be the case that the differences we observed between responses to the traditional survey format and to the Socratic questionnaires could be due to differences in attention. In order to address this worry in experiment #3, we designed a new, quasi-dynamic version of the questionnaire that mimicked the structure of the initial response and justification stages of our Socratic questionnaires, which tailored requests for justification depending on participants' initial responses. The quasi-dynamic questionnaire allows us to assess whether participants are paying attention and whether they are understanding the prompts they are responding to. And, as we will discuss below, we also ran an unmodified static survey identical to the survey we conducted in experiment #2 so we could see whether the quasi-dynamic survey produced different responses than the static survey.

The second potential confound we wanted to control for concerned participant recruitment. In our first two experiments, participants who were recruited to the Socratic questionnaire condition were paid several times more than participants in the traditional survey format condition, and when recruited they were told roughly how long the task would take (about 48 minutes in Experiment #1 and 34 minutes in Experiment #2, in contrast with 9 minutes and 6 minutes for the traditional survey in each experiment, respectively). Those differences might attract different types of participants or give them different attitudes about the study before it begins, and those differences might account for the differences we were attributing to differences in format in experiments #1 and #2. To control for this confound, we recruited participants for all three experimental conditions (Socratic questionnaire, quasi-dynamic survey, and static survey) using the same generic task

description, which said that participants might be assigned to one of three different experimental conditions with different durations, and paid participants in all three conditions the same amount.[24]

We used the same scenarios that we used in experiment #2 (Color and Game Show) in all three formats (Socratic questionnaire, quasi-dynamic survey, and traditional survey) in experiment #3.

4.1 Design of the quasi-dynamic survey

The quasi-dynamic survey was designed to mimic the structure of the Socratic questionnaire as closely as possible up through the justification phase. After participants gave their initial responses to the scenarios, they received prompts to explain their responses in an open-response format that depended on their initial responses. So, for example, if participants gave different responses to the Low Stakes and High Stakes versions of the Game Show scenario, they saw this prompt at the justification stage of the survey:

I'm interested in your responses to the two gameshow scenarios whereby Tracy and Emma are responding to the question about the capital of Tanzania.

I noticed that you gave different responses to each of the gameshow scenarios. Would you say a little bit about how you decided on those responses?

If participants gave the same response to each of the Game Show scenarios, they would see a different prompt. The same type of quasi-dynamic open-ended response prompts were also used for the Color scenarios.

4.2 Participants

Socratic Questionnaires

In experiment #3, 50 participants[25] (19 Females, 23 Males, 3 Non-Binary, 5 undisclosed, M_{age} = 23.19, SD_{age} = 4.38) were recruited from Prolific, with the requirement that they be fluent speakers of English and not have participated in our previous studies, and paid $6 each for a chat lasting an average of 35 minutes (an average hourly rate of $10.29). The resulting 50 chat transcripts from experiment #3 run to 96,083 words and are available in their entirety online: https://osf.io/9mdx5/?view_only=022627aecdbf4bbc8101c51ae8847713

Quasi-Dynamic Survey

[24] The generic text we used to recruit participants for Experiment #3 is available here: https://osf.io/34wtf/?view_only=dbfeeeb0aa93420aaa9adcb837ed68a7
Thanks to the two referees for this paper and the editors for pointing out these potential confounds.

[25] Using a simulation-based power analysis (Lakens & Caldwell, 2019) based on a 3b*2w design, we determined that N = 50 in each cell would be sufficient to detect a stakes effect (power = 100%) and format effect (power = 100%) in experiment #3 (2000 simulations performed using means from experiment #1 and alpha criterion of .05).

50 participants (32 Females, 15 Males, 3 identifying as "other", M_{age} = 27.60, SD_{age} = 7.46) were recruited on Prolific, with the requirement that they be fluent speakers of English and that they not have participated in our previous studies. Participants were then directed to the Qualtrics survey. Participants were paid $6 on completion of the survey, which took an average of 15.37 minutes to complete (an average hourly rate of $23.42). Responses to each case were recorded in the same way as the initial responses in the Socratic questionnaire.

Traditional Static Survey

51 participants (29 Females, 22 Males, M_{age} = 24.29, SD_{age} = 5.23) were recruited on Prolific, with the requirement that they be fluent speakers of English and that they not have participated in our previous studies.[26] Participants were then directed to the Qualtrics survey. Participants were paid $6 on completion of the survey, which took an average of 7.45 minutes to complete (an average hourly rate of $48.32). Responses to each case were recorded in the same way as the initial responses in the Socratic questionnaire.

This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants, in both the Socratic questionnaire and both survey formats.

4.3 Results

Inter-rater reliability

We trained an independent coder (who was blind to the hypotheses that we were testing) who coded 10% of the scripts from the Socratic Questionnaire and 10% of the open-ended responses to the Quasi-Dynamic Survey.[27] To determine interrater reliability, we calculated percentage agreement scores between the independent coder and research teams members who coded all scripts and open-ended responses. Cohen's Kappa was calculated for each set of codes at both justification and pushback stages in the Socratic Questionnaires and in the justification stage in the Quasi-Dynamic Survey. In the Socratic group, across all codes and stages, average agreement between raters was 88.89% with moderate to perfect agreement in codings of deontic modals, hedges, and revisions ($\varkappa$s = .74 – 1.00) and fair agreement in codings of misunderstandings ($\varkappa$s > .29)[28]. In the Quasi-Dynamic group, in the justification stage, average agreement between raters

[26] We aimed to recruit 50 participants for the traditional static survey but an additional participant was recruited through sampling error.

[27] The instructions that we gave to the independent coder were the same as we used in Experiment #2, and can be found here: https://osf.io/k598r/?view_only=055c42b5add24e889f454b1266c5b4af

[28] The spreadsheet with the conversations coded by the independent coder can be found here: https://osf.io/7ap8s/?view_only=fa2192b9626547099c75474627591b8b; the spreadsheet with the conversations coded by the experimenters can be found here: https://osf.io/z84ye/?view_only=2c1dadee9186425fa8e24bd4d0e5a678;

was 85%[29] with moderate to perfect agreement in codings of deontic modals, hedges, and revisions ($\varkappa$s = .73 – 1.00) and fair to moderate agreement in codings of misunderstandings ($\varkappa$s > .37).[30]

Color

Overall, we found a context effect in the color scenarios (Rug case; Gas case) across all formats (Socratic; Quasi-Dynamic; Traditional), (Wald (χ2(1) = 33.53, p <.001, 95% Wald CI [-3.66, -1.20], V = .24) with the proportion of responses saying that Hugo's claim ("The walls in our apartment are brown") is *true* significantly higher in the Rug case. There was no difference in responses across formats and no interaction of format x case (*ps* >. 532). These results do not change even when revised responses are factored in (9 revisions were made). As before, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 37.51, p <.001, 95% Wald CI [-3.66, -1.20, V = .25]) and no difference between the formats (p =.655) and no interaction (p = .546) (**see Figure 7)**. This analysis does not change when changes and misunderstandings are removed from the data (27 data points were removed).

The full interrater reliability analysis can be found here:
https://osf.io/c8qu6/?view_only=15b4c3942fd14d26b1494cbec351d6c8

[29] For misunderstandings and/or changes to the scenarios, inter-rater agreement was 55.56%. However, the McNemar Tests comparing raters were not statistically significant (p = .62).

[30] The spreadsheet with the open-ended responses coded by the independent coder can be found here: https://osf.io/unk27/?view_only=fe034a3a0e3f4028865c5e5557eb8e5f; the spreadsheet with the open-ended responses coded by Hansen can be found here:
https://osf.io/2ntm4/?view_only=aa03e6e01ff547c6b65bdc9b6292c5e8
The full interrater reliability analysis can be found here:
https://osf.io/5xc36/?view_only=81abd342a6ab4e3fa7b798c4ff2ccb5f

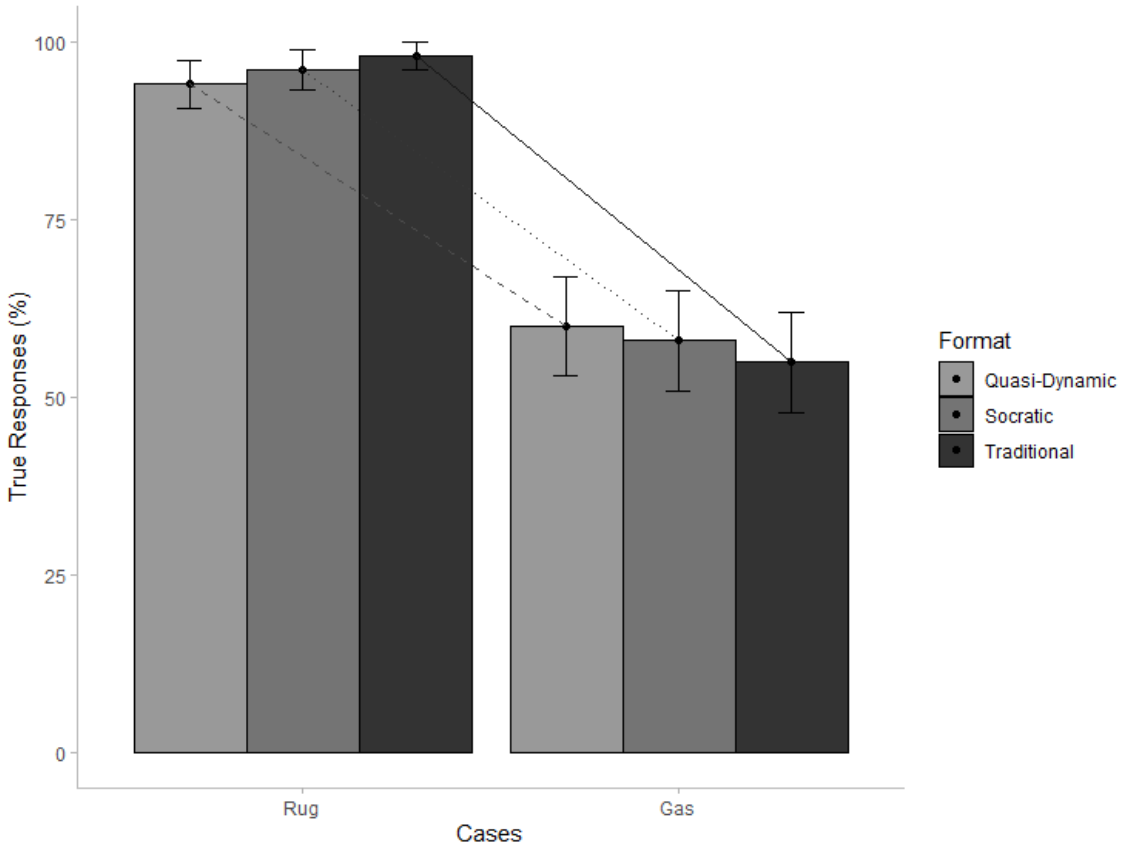


Figure 7: Responses to the Color scenarios in the quasi-dynamic, Socratic, and standard formats when revised responses are included in experiment #3. We find a context effect across all three formats. Error bars represent +/-1 SE.

Game Show

Overall, we found a stakes effect in the game show scenarios (Low Stakes case; High Stakes case) across all formats (Socratic; Quasi-Dynamic; Traditional), (Wald (χ2(1) = 114.00, p <.001, 95% Wald CI [-1.56, -0.82], V = .51) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was not a statistically significant main effect of format and no statistically significant interaction between case and format (*ps* >.13).

When revised responses to the Socratic questionnaire are taken into account (8 revisions made), the analysis remains the same. We found a stakes effect in the Game Show scenarios (Low stakes case; High Stakes case) across all formats (Socratic; Quasi-Dynamic; Traditional), (Wald (χ2(1) = 122.34, p <.001, 95% Wald CI [-1.63, -0.89], V = .52) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was not a statistically significant main effect of format and no statistically significant interaction

between case and format (*ps* >.14). (See Figure 8.). This analysis does not change when changes and misunderstandings are removed from the data (8 data points were removed).

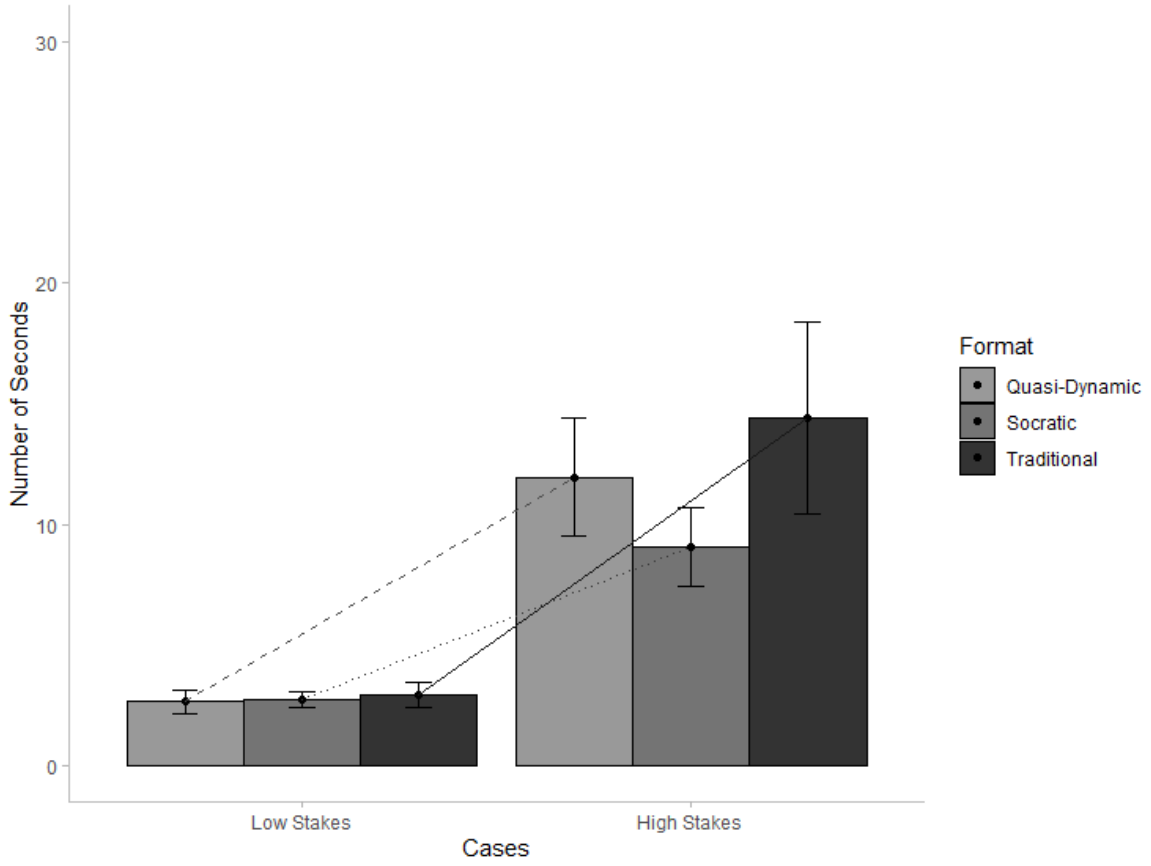


Figure 8: Differences in responses to the Game Show scenarios in experiment #3 (taking revision into consideration). There is an effect of stakes in all formats (Socratic questionnaire, Quasi-Dynamic survey, and Traditional survey)**.** Error bars represent +/- 1 SE.

There were no apparent differences in the number of "never" responses in the Game Show scenarios across formats (5 "nevers" in both the Socratic Questionnaire and Quasi-Dynamic Survey and 0 "nevers" in the Traditional Questionnaire). However, inferential analyses could not be completed given that no "never" responses were given in the Traditional Survey format.

5. Aggregate Data Analysis

Given the similarity in experimental designs in experiments #1, #2, and #3, we also performed an aggregate data analysis on the combined responses to the Game Show scenarios and Color scenarios taking data from Socratic and Traditional groups only.

Color

Overall, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 69.32, p <.001, 95% Wald CI [-3.64, -2.01], *V* = .35) with the proportion of responses saying that Hugo's claim ("The walls in our apartment are brown") is *true* significantly higher in the Rug case. There was no difference in responses across formats and no interaction of format x case (ps >. 168). These results do not change even when revised responses are included. As before, we found a context effect in the color scenarios (Rug case; Gas case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 80.84, p <.001, 95% Wald CI [-3.64, -2.01], *V* = .38) and no difference between the formats (p =.489) and no interaction (p = .421).

Game Show

Overall, we found a stakes effect in the game show scenarios (Low Stakes case; High Stakes case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 165.39, p <.001, 95% Wald CI [-1.55, -0.97], *V* = .65) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was not a statistically significant main effect of format and no statistically significant interaction between case and format (*ps* >.148). These results do not change even when revised responses are included. As before, we found a stakes effect in the game show scenarios (Low Stakes case; High Stakes case) across both formats (Socratic; Traditional), (Wald (χ2(1) = 171.78, p <.001, 95% Wald CI [-1.55, -0.97], *V* = .65) with individuals stating that more seconds are needed before the subject knows the correct answer in the High Stakes case. There was not a statistically significant main effect of format and no statistically significant interaction between case and format (*ps* >.136).

6. General Discussion

Our tantalizing initial findings of effects of format (Socratic questionnaire vs. traditional survey) on the Color and Game Show scenarios in experiment #1 were not borne out in our results in experiments #2 and #3. Unlike in experiment #1, we didn't find a conversational effect on responses to the Color scenarios in experiment #2 or #3. While we did find conversational effects on the Game Show scenarios in both experiments #1 and #2, the effects appeared at different stages in the two experiments: In experiment #1, it appeared once people had the chance to reflect on their responses and respond to an objection to their initial answers, while in experiment #2, the effect of conversational format was already present in participants' initial responses to the scenarios.[31] But once we controlled for confounds in format and participant recruitment in

[31] The effect of format in experiment #1 for initial responses to the Game Show scenarios was not statistically significant (p =.056) but this may be a result of experiment #1 being underpowered. This is supported by the fact that including only a few revised responses in the analysis of the Game Show scenarios in experiment #1 resulted in

Experiment #3, the various effects of format that appeared experiments #1 and #2 did not replicate. And in an aggregate data analysis of our three experiments, we did not find an effect of format.

The effects of extended conversation are therefore more elusive than we expected them to be when designing this study. There are a couple of potential explanations for this. Those who remain optimistic about the effect that conversation can have on affecting participants' responses might point out that the argumentative pushback in the Socratic questionnaires was designed to be very mild, consisting only in noting the fact that other people sometimes disagree with the answers that participants gave. If the pushback to participants' initial responses included more substantial arguments we might observe more frequent revisions of participants' initial responses. (Since we're still optimistic about finding an effect of conversation on responses, we are currently conducting experiments using Socratic questionnaires that have more substantial pushback stage, using arguments based on explanations that participants gave in defense of their responses to Socratic questionnaires in experiment #1.) For those who are more pessimistic about conversational effects, our failure to find an effect may be due to the fact that participants' initial responses are relatively stable and aren't easily affected by relatively minor situational changes of the kind that can be brought to bear in relatively short online conversations.[32]

7. Qualitative Discoveries in Socratic Questionnaires

While the comparison of Socratic questionnaires with traditional static surveys did not uncover any quantitative effects of format, we think there are valuable qualitative insights that come from reading the transcripts of the 137 Socratic questionnaires we conducted. One particular qualitative discovery that we will focus on illuminates a debate in the stakes-sensitivity literature in epistemology.

Deontic modals in justifications of responses to the game show scenario

One disputed feature of "evidence-seeking" prompts, like that used in the game show scenario, concerns the role played by the deontic modal that appears in them. In the game show scenario, participants are asked:

a statistically significant effect of format and given that in experiment #2 (which had greater power) we found a statistically significant effect of format for initial responses.

[32] See Knobe (2021) for a review of experimental evidence that supports the claim that "philosophical intuitions are surprisingly stable across both demographic groups and situations". A recent conversation-based example of stability is Santoro and Brockman (2022), who ran a pair of experiments evaluating whether short online video conversations between partisans on either side of the American political spectrum reduced "affective polarization", and found that while non-political conversations about each participant's "perfect day" reduced affective polarization, the effect "decayed within 3 months". In conversations about political topics that prompted disagreement, the experimenters did not find a reduction in affective polarization. Santoro and Brockman summarize their findings as follows: "our findings…suggest that such conversations are not enough alone to result in the durable changes in attitudes toward outpartisans and toward democracy that many hope for" (p. 11).

> How many seconds does Tracy need to spend considering her answer before she knows that the capital of Tanzania is Dodoma?

Philosophers who have used evidence-seeking questions like this have interpreted the fact that participants give different answers in low stakes and high stakes scenarios as supporting the idea that people treat knowledge as sensitive to stakes (Pinillos 2012, Pinillos and Simpson 2014, Francis et al. 2019). But critics have argued that differing responses in high and low stakes scenarios may be due to the effect of stakes not on knowledge, but on the deontic modal ("need" in the game show scenario), since it is uncontroversial that what is at stake can affect judgments about what someone needs to do (Buckwalter and Schaffer 2015).

When we look at the way participants give reasons for their responses to the game show scenario in the "justification" phase of the Socratic questionnaire, a third possibility emerges, different from both previous interpretations of what's going on in response to evidence-seeking prompts. Consider, for example, this justification given in Chat #8 in experiment #1:

(20:51:55) Admin: I notice that you gave different responses to each of the game show scenarios. Would you say a little bit about how you decided on those responses?
(20:53:05) User 1: For the first one, I decided that since she already knew the answer and the stakes were low, she wouldn't think too much about it and just answer. In the second time, a lot more money was at stake, and I think that even the people that know the answer would struggle to act quick and would take a lot longer, because of the amount of money that is at stake at that moment.

In his justification of why he gave different responses in the high and low stakes game show scenarios, this participant doesn't use a deontic modal; he says what he thinks the subject in the scenario *would* and *wouldn't* do. And later in the same conversation, in the "pushback" phase, the participant makes it clear that he is not responding to the presence of the deontic modal—he explains his response in terms of what people "tend to do". So the participant in Chat #8 doesn't seem to be responding to the presence of the deontic modal, or to the request for what needs to happen before someone *knows* something; instead, he's responding to the effect of stakes on his assessment of the non-deontic modal auxiliary "would", which figures in a prediction about what happens in certain conditions.

This is a common way of understanding the prompt: 19 out of 40 participants in experiment #1, 18 out of 47 participants in experiment #2, and 29 out of 50 participants in experiment #3 use either "would" or "will" or what people "tend to do" or say what it is "natural" to do in their justifications of their initial responses, and don't use deontic modals or "knows" in their justifications at all. That suggests that this is a case of "attribute substitution", in which participants replace a question that is difficult to answer with an easier question (Kahneman and Frederick 2002). It has been argued that questions about whether someone knows something as posed in experimental epistemology are conversationally odd (Baz 2012, Hansen 2020). If that's right, it

might be the case that participants are substituting a less odd question about how someone *would* act when different amounts of money are at stake.

While the use of a non-deontic modal auxiliary ("would", "will") is a common occurrence in the justifications of the initial responses of the game show scenario, it does not occur in every justification. Some participants do use deontic modals in their justifications ("requires", "needs", "shouldn't", "has to be"), some participants mix both deontic and non-deontic modals, and some responses don't use any type of modal in their response. The variety of explanations for their responses given by participants should make us suspicious of any uniform explanation of what is driving responses to "evidence-seeking" prompts—if participants' own justifications are any clue to what is really going on, it is likely that a variety of factors across different participants are responsible. A full understanding of what's going on in apparent stakes effects elicited by "evidence-seeking" prompts will need to disentangle these various factors.[33]

Revealing how participants are responding to different flavors of modals in their responses to the Game Show scenarios is an example of one major advantage of the Socratic questionnaire method: it gives us a glimpse, however partial and blurry, of what's going on "under the hood" when participants respond to philosophical scenarios.

8. Further Methodological Considerations: Identifying Insufficient Effort Responding and Some Limitations of Socratic Questionnaires

Both psychologists and experimental philosophers are currently facing new challenges associated with collecting empirical data online. In 2018, researchers noticed odd bot-like responses and responses from individuals using server farms, in data collected using Amazon's Mechanical Turk platform (Ahler, Roush, and Sood, 2018). These respondents are often guilty of insufficient effort responding (IER), the reduced effort used by some participants to respond to online experiments, which may include careless, inattentive, or random responding (Huang et al., 2012). Both experimental psychologists and philosophers currently adopt a number of approaches to either reduce IER or remove IER using ex ante and ex post techniques (Pölzler, 2021). For example, many researchers incorporate attention checks or comprehension checks in online experiments to improve data quality (e.g., Abbey and Meloy, 2017). However, adopting these techniques can alter responses as participants begin to look out for these "tricks" (Hauser and Schwarz, 2015). One of the most efficient and simple ways of detecting IER online is to ask participants to respond to an

[33] Thompson (2022) analyzes conversations that result from participants a talk-aloud approach to explaining their judgments about scenarios intended to probe whether people endorse the principle that ought implies can. Thompson argues that quantitative analyses of standard surveys that investigate ought implies can "misrepresent participants' rich and complex judgments", and he finds evidence that "some participants introduced good reasons for interpreting the survey in ways that might not have been intended", which aligns with our findings about deontic and non-deontic modals in the Game Show scenarios.

open-ended question (Pölzler, 2021) so that unrelated, incoherent, or nonsensical answers can be flagged and subsequently removed (e.g., Dennis, Goodson, and Pearson 2018; Francis, 2019). Socratic questionnaires—in effect, 30-45 minute Turing tests—are arguably the ultimate IER check. Engaging participants in a real-time conversation allows researchers to assess the level of attention being given to the study, whether participants comprehend the questions that they are being asked, whether they meet the study criteria in terms of language ability, and finally, whether they are in fact human.

While Socratic questionnaires are a powerful tool for digging deeper into participants' responses and tracking them over the course of a conversation, they also have disadvantages in comparison with standard survey methods. Socratic questionnaires are labor-intensive and expensive to conduct, in contrast with traditional surveys which can quickly and affordably collect large numbers of responses (Socratic questionnaires are like one-on-one tutorials in contrast with a large lecture). Building open-ended responses into traditional surveys (as we did in our quasi-dynamic survey in experiment #3, and as other researchers regularly do) provides a less labor-intensive way of identifying IER. But sometimes it can be difficult to tell whether participants are misunderstanding the scenario or prompts from the open-ended responses they give to the quasi-dynamic survey, because the quasi-dynamic survey lacks a key feature of the Socratic questionnaires: the ability to push back against participants' explanations. So, for example, participant #11, responding to the quasi-dynamic survey, responds to the request to say more about why he responded "true" to both of the Color scenarios by saying only: "Because it was written in both texts". That response is hard to interpret—it probably means that it was written in both scenarios that the walls were in fact brown (a reasonably common misunderstanding), which if true would be a good reason for thinking that Hugo's statements were both true. But that's just one possible reconstruction of the participant's response. In a Socratic questionnaire, we could ask a simple follow up to clarify what this response means and whether it's evidence that the participant is misunderstanding the scenario.

Another trade-off involved in using the Socratic questionnaire method is the potential introduction of unintended variables that could be influencing participants' responses, like the social pressure involved in having a live conversation with an experimenter (as opposed to simply filling out a survey), different forms of experimenter bias where the participant may try to work out what response the experimenter they are talking to wants them to give, or what the purpose of the experiment is.[34] And in an unexpected inversion of worries about IER, some participants openly wonder whether they were chatting with a bot rather than a human being![35] Socratic questionnaires

[34] See chat #11 in Experiment #2 for an example of a participant who directly asks what the purpose of the experiment is.

[35] Here is one example, from chat #2 in Experiment #3, where where a participant asks whether they are talking to a human or a computer:
(01:49:15 PM) Other Participant: Are you a real person?
(01:49:36 PM) Me: Yes, but I know that's hard to prove—feel free to ask me questions that a computer couldn't answer.

therefore give up some experimental control in favor of a more natural communicative exchange and the ability to push back on participants' explanations to pin down whether they're really understanding the scenarios and prompts they've been asked to read and respond to.

9. Conclusion

Conversational experiments have been used by researchers in fields adjacent to philosophy: in the psychology of reasoning (Trouche et al. 2014), in linguistics (Schober and Clark 1989, Schütze 2020), and sociology (Vaisey 2009). But they have seen limited use in philosophy, which is surprising given the central role that conversation plays in contemporary philosophy's own self-understanding as a descendent of Socratic dialogue.[36] While the design of the Socratic questionnaires we used in this study offer only a small step in the direction of a full Socratic elenchus, they show how even light pushback can start to reveal how participants are understanding the questions they are being asked to respond to—sometimes in ways that conflict with what philosophers have assumed.

Conversation is dynamic: we are continually updating our beliefs and adjusting, negotiating, and repairing our claims in collaboration with our interlocutors. As Euthyphro complains in his discussion with Socrates about the nature of piety, his statements "won't stay where we put them" because of Socrates's questioning:

> SOCRATES: What is holiness, and what is unholiness?
>
> EUTHYPHRO: But, Socrates, I do not know how to say what I mean. For whatever statement we advance, somehow or other it moves about and won't stay where we put it. (Plato, *Euthyphro*, 11B-C)

But it might turn out that even at the end of a demanding Socratic conversation, the person being questioned gives the same answer to a question that they started out with. For example,

---

(01:49:52 PM) Other Participant: You seem suspicious
(01:49:59 PM) Me: hahaha, uh oh
(01:50:10 PM) Me: I am a person, FYI
(01:50:30 PM) Other Participant: Im going to look up a question a computer can not actually answer to
(01:50:47 PM) Me: Ok, I'll wait! Then we can start the study.
(01:51:15 PM) Other Participant: Paul tried to call George on the phone, but he wasn't successful.
(01:51:18 PM) Other Participant: who was not successful?
(01:51:24 PM) Me: Paul
(01:51:32 PM) Other Participant: alright we can begin
(01:51:38 PM) Other Participant: though you had 50%
(01:51:44 PM) Me: Ok, great. Glad I passed the test. *sunglasses*

[36] For exceptions, see Fisher et al. (2017) and Thompson (2022).

Thrasymachus "surrenders" to Socrates's questioning about the nature of justice in Book I of the *Republic*, but he hasn't really changed his mind (Plato, *Republic*, 357b).

Adding Socratic questionnaires to our experimental repertoire gives us the ability to keep track of how and when our statements stay put or refuse to stay put as we reflect on them, defend them against objections, and either hang on to them or give them up in the face of challenges.

Bibliography

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53, 63-70.

Ahler, D. J., Roush, C. E., & Sood, G. (2018). The micro-task market for "Lemons": Collecting data on Amazon's Mechanical Turk. *Working Paper.* Epub ahead of print.

Andow, J. (2016). Qualitative tools and experimental philosophy. *Philosophical Psychology*, *29*(8), 1128–1141.

Austin, J.L. (1956). A Plea for Excuses. *Proceedings of the Aristotelian Society,* 57, 1-30.

Baz, A. (2012). *When Words Are Called For: A Defense of Ordinary Language Philosophy*. Harvard University Press.

Boyd, K., & Nagel, J. The Reliability of Epistemic Intuitions. In Machery, E., and O'Neil, E. (Eds.), *Current Controversies in Experimental Philosophy* (pp. 128-145). Routledge.

de Bruin, B. (2021). Saving the Armchair by Experiment: What Works in Economics Doesn't Work in Philosophy. *Philosophical Studies*, 178, 2483–2508.

Buckwalter, W., & Schaffer, J. (2015). Knowledge, Stakes and Mistakes. *Noûs*, *49*(2), 201–234.

Chenail, R. (2008). Categorization. In Given, L.M. (Ed.), *The SAGE Encyclopedia of Qualitative Research Methods* (pp. 72-73). SAGE.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology*, *5*.

Cullen, S. (2010). Survey-Driven Romanticism. *Review of Philosophy and Psychology*, *1*(2), 275–296.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). Mturk workers' use of low-cost "virtual private servers" to circumvent screening methods: A research note. Working Paper: University of Kentucky. doi:10.2139/ssrn.3233954

DeRose, K. (2011). Contextualism, Contrastivism, and X-Phi Surveys. *Philosophical Studies*, *156*(1), 81–110.

Dinges, A., & Zakkou, J. (2020). Much at Stake in Knowledge. *Mind & Language,* 1-21.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2017). The Influence of Social Interaction on Intuitions of Objectivity and Subjectivity. *Cognitive Science*, *41*(4), 1119–1134.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, *5*, 5–15.

Francis, K., Beaman, P., & Hansen, Nat. (2019). Stakes, Scales, and Skepticism. *Ergo*, *6*(16), 427–487.

Francis, K. (2019, August 27). Online experiments: Virtual Private Servers (VPS) and suspicious response filtering. https://doi.org/10.17605/OSF.IO/2UXK9

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*(5537), 2105–2108.

Grindrod, J., Andow, J., & Hansen, N. (2019). Third-Person Knowledge Ascriptions: A Crucial Experiment for Contextualism. *Mind & Language*, *34*(2), 158–182.

Hannon, M. (2018). Intuitions, reflective judgments, and experimental philosophy. *Synthese*, *195*(9), 4147–4168.

Hansen, N. (2014). Contemporary Ordinary Language Philosophy. *Philosophy Compass*, *9*(8), 556–569.

Hansen, N. (2020). "Nobody Would Really Talk That Way!": The Critical Project in Contemporary Ordinary Language Philosophy. *Synthese*, *197*(6), 2433–2464.

Hansen, N., & Chemla, E. (2013). Experimenting on Contextualism. *Mind & Language*, *28*(3), 286–321.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, 5(2).

Horvath, J. (2010). How (not) to react to experimental philosophy. *Philosophical Psychology*, *23*(4), 447-480.

Horvath, J. (2015). Thought Experiments and Experimental Philosophy. In C. Daly (Ed.), *The Palgrave Handbook of Philosophical Methods* (pp. 386–418). Palgrave Macmillan UK.

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, *113*(3), 430–452.

Hyland, K. (1998). *Hedging in Scientific Research Articles.* John Benjamins Publishing Company.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49–81). Cambridge University Press.

Kauppinen, A. (2007). The Rise and Fall of Experimental Philosophy. *Philosophical Explorations*, *10*(2), 95–118.

Kneer, M., Colaço, D., Alexander, J., & Machery, E. (2021). On Second Thought: Reflections on the Reflection Defense. *Oxford Studies in Experimental Philosophy, 5*, 257–296.

Knobe, J. (2021). Philosophical Intuitions Are Surprisingly Stable across both Demographic Groups and Situations. *Filozofia Nauki*, 21(2), 11–76.

Lakens, D., & Caldwell, A. R. (2019). Simulation-based power-analysis for factorial ANOVA designs. https://doi.org/10.31234/osf.io/baxsf

Lakoff, G. (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, *2*(4), 458–508.

Liao, S. M. (2008). A defense of intuitions. *Philosophical Studies, 140*(2), 247-262.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*(1), 128-159.

Mates, B. (1958). On the Verification of Statements about Ordinary Language. *Inquiry*, *1*(1), 161–171.

Nadelhoffer, T., & Nahmias, E. (2007). The Past and Future of Experimental Philosophy. *Philosophical Explorations*, *10*(2), 123–149.

Nado, J. (2015). Intuition, philosophical theorizing, and the threat of skepticism. In E. Fischer and J. Collins (Eds.), *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*, (pp. 204–221), Routledge.

Pinillos, A. (2012). Knowledge, Experiments, and Practical Interests. In J. Brown & M. Gerken (Eds.), *New Essays on Knowledge Ascriptions* (pp. 192–219). Oxford University Press.

Pinillos, A., & Simpson, S. (2014). Experimental Evidence Supporting Anti-Intellectualism about Knowledge. In J. R. Beebe (Ed.), *Advances in Experimental Epistemology* (pp. 9–43). Bloomsbury.

Plato, *Euthyphro. Apology. Crito. Phaedo*. Edited and translated by Harold North Fowler, William Preddy. Loeb Classical Library. Cambridge, MA: Harvard University Press, 1914.

Plato, *Republic.* Edited by G.R.F Ferrari, translated by Tom Griffith. Cambridge University Press, 2000.

Pölzler, T. (2021). Insufficient Effort Responding in Experimental Philosophy. *Oxford Studies in Experimental Philosophy* 4, 214–246.

Santoro, E., & Brockman D.E. (2022). The Promise and Pitfalls of Cross-Partisan Conversations for Reducing Affective Polarization: Evidence from Randomized Experiments. *Science Advances, 8*(eabn5515), 1–17.

Schober, M. F., & Clark, H. H. (1989). Understanding by Addressees and Overhearers. *Cognitive Psychology*, *21*, 211–232.

Schütze, C. T. (2020). Acceptability Judgments Cannot Be Taken at Face Value. In S. Schindler, A. Drozdzowicz, and K. Brøcker (Eds.), *Linguistic Intuitions* (pp. 189–214). Oxford University Press.

Schwarz, N. (1996). *Cognition and Communication. Judgmental Biases, Research Methods, and the Logic of Conversation.* Erlbaum.

Sosa, E. (2009). A Defense of the Use of Intuitions in Philosophy. In *Stich and his Critics* (pp. 101–112). John Wiley & Sons, Ltd.

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, 3(6), 375-387.

Thompson, K. (2022). Qualitative Methods Show that Surveys Misrepresent 'Ought Implies Can' Judgments. *Philosophical Psychology*, 2–29.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Travis, C. (1989). *The Uses of Sense: Wittgenstein's Philosophy of Language*. Oxford University Press.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, More Than Confidence, Explain the Good Performance of Reasoning Groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Turri, J. (2013). A Conspicuous Art: Putting Gettier to the Test. *Philosophers' Imprint*, 13(10), 1–16.

Vaisey, S. (2009). Motivation and Justification: A Dual-Process Model of Culture in Action. *American Journal of Sociology*, *114*(6), 1675–1715.