Ergo

Stakes, Scales, and Skepticism

Kathryn B. Francis*[1,2,3], Philip Beaman[2], and Nat Hansen**[1]

[1]Department of Philosophy, University of Reading, Edith Morley Building, Whiteknights Campus, Reading, RG6 6BT, United Kingdom

[2]School of Psychology and Clinical Language Sciences, University of Reading, Harry Pitt Building, Earley Gate, Reading, RG6 7BE, United Kingdom

*kathrynbfrancis@googlemail.com
**n.d.hansen@reading.ac.uk

[3]Present address: Division of Psychology, School of Social Sciences, University of Bradford, Bradford, West Yorkshire, BD7 1DP, UK

Stakes, Scales, and Skepticism*

Kathryn B. Francis, Philip Beaman, and Nat Hansen

Forthcoming in *Ergo*

Abstract

There is conflicting experimental evidence about whether the "stakes" or importance of being wrong affect judgments about whether a subject knows a proposition. To date, judgments about stakes effects on knowledge have been investigated using binary paradigms: responses to "low" stakes cases are compared with responses to "high stakes" cases. However, stakes or importance are not binary properties—they are scalar: whether a situation is "high" or "low" stakes is a matter of degree. So far, no experimental work has investigated the scalar nature of stakes effects on knowledge: do stakes effects increase as the stakes get higher? Do stakes effects only appear once a certain threshold of stakes has been crossed? Does the effect plateau at a certain point? To address these questions, we conducted experiments that probe for the scalarity of stakes effects using several experimental approaches. We found evidence of scalar stakes effects using an "evidence-seeking" experimental design, but no evidence of scalar effects using a traditional "evidence-fixed" experimental design. In addition, using the evidence-seeking design, we uncovered a large, but previously unnoticed framing effect on whether participants are *skeptical* about whether someone can know something, no matter how much evidence they have. The rate of skeptical responses and the rate at which participants were willing to attribute "lazy knowledge"—that someone can know something without having to check—were themselves subject to a stakes effect: participants were more skeptical when the stakes were higher, and more prone to attribute lazy knowledge when the stakes were lower. We argue that the novel skeptical stakes effect provides resources to respond to criticisms of the evidence-seeking approach that argue that it does not target knowledge.

(Abstract: 275 words)

* Thanks to Alexander Dinges, Julia Zakkou, audiences at the UCL Experimental Philosophy Workshop, the Buffalo Experimental Philosophy Workshop, the University of Illinois Chicago, the University of Reading philosophy department and Centre for Cognition Research, and two anonymous referees for very helpful comments on this paper. Funding from the Leverhulme Trust Research Project Grant RPG-2016-193 made this research possible. [1] See Gerken (2017, §2.5.b) for a longer summary of existing studies on practical effects on knowledge (stakes effects are one variety of practical effects—see §2 below for discussion).

1. Background: Experimental studies of knowledge and stakes

Is it easier to know something when very little is at stake in being right or wrong? If one believes that *knowledge is sensitive to stakes*, it might be the case that it is indeed easier to know trivial things. The view that knowledge is sensitive to stakes holds that whether a subject knows some proposition *p* depends on how much is at stake for that subject in being right or wrong about *p*; if knowledge is sensitive to stakes, it could be the case that the more that is at stake, the harder it is for the subject to know that *p*.

Advocates of the stakes-sensitivity of knowledge have assumed that their own intuitions that it is easier to *know that something is the case* in lower-stakes situations than in higher-stakes situations are representative of how ordinary people (that is, non-philosophers) would judge those cases. Experimental attempts to verify that assumption have been mixed, however. While May et al. (2010) found evidence of an effect of stakes on knowledge, Buckwalter (2010) and Feltz and Zarpentine (2010) did not. Subsequent studies made the empirical case for the stakes-sensitivity of knowledge look more promising: Sripada and Stanley (2012), Pinillos (2012), and Pinillos and Simpson (2014) all found evidence that participants' judgments of when an individual knew that *p* were sensitive to stakes. But a large cross-cultural study of judgments about cases involving the stakes-sensitivity of knowledge found effects only for Spanish, Japanese, and UK participants, and not for any of the other 16 sampled nationalities (Rose et al. 2017). Theoretical challenges to findings of stakes sensitivity have also been raised. Buckwalter (2014) and Buckwalter and Schaffer (2015) target what appear to be the strongest evidence in favor of the stakes-sensitivity of knowledge, namely Pinillos's "evidence-seeking" studies (Pinillos 2012; Pinillos and Simpson 2014). These arguments make the case that the appearance of stakes effects on knowledge are actually stakes effects on other features that appear in the experimental prompts, so no conclusions about knowledge *per se* follow from the experimental data.[1]

In this paper, we aim to advance our understanding of the empirical foundations of the stakes sensitivity of knowledge by looking at an aspect of the interaction of stakes and knowledge that has not received any experimental attention, namely the scalar nature of stakes. To date, stakes effects on knowledge have been investigated using binary paradigms: responses to "low" stakes cases are compared with responses to "high stakes" cases. However, stakes is not a binary property but a *scalar* property: whether a situation counts as "high" or "low" stakes comes in degrees and depends on what it is being compared with (Anderson and Hawthorne 2019, Hansen 2014). No experimental work has investigated the scalar nature of stakes effects on knowledge: Do stakes effects increase as the stakes get higher? Do stakes effects only appear once a certain threshold of stakes has been crossed? Do stakes effect plateau at a certain point?

To address these questions, we conducted experiments that probe for the scalarity of stakes effects using several experimental approaches. In our first experiment, which adopts the classic "evidence-fixed" design employed in the earlier experimental studies of stakes

[1] See Gerken (2017, §2.5.b) for a longer summary of existing studies on practical effects on knowledge (stakes effects are one variety of practical effects—see §2 below for discussion).

effects on knowledge (e.g., Sripada and Stanley, 2012), we ask participants to rate their level of agreement with claims that *S knows that P* or *S doesn't know that P*. To anticipate our results: Across several epistemic scenarios that vary the type of stakes, from personal injury to reputation, we did not find any evidence of the stakes effects on judgments about knowledge, even when comparing relatively low and relatively high points on the scale of stakes. Since this failure to find an effect is at odds with the effect reported in Sripada and Stanley (2012), we conducted a pre-registered replication of Sripada and Stanley's study. Our study did find a small effect of stakes on knowledge using the evidence-fixed design, but in a different condition than Sripada and Stanley's study**.**

The second series of experiments we conducted employed the "evidence-seeking" approach developed in Pinillos (2012). We asked participants to judge how much evidence a subject needs to collect before she counts as knowing various propositions. Results from the evidence-seeking experimental design revealed stakes effects across multiple scenarios and indicate that there is variability in the structure of stakes effects when different scales of stakes are at issue (e.g. when number of lives or money or degrees of embarrassment are at stake).

In addition, by testing both positive and negative polarity versions of the evidence-seeking prompts in an attempt to identify a threshold for ascribing knowledge, we also were able to uncover a large framing effect on participants' willingness to say that a subject in a scenario *never* can know that something is the case: Participants tended to respond to the negative prompt ("How many times can S check F and still *not* know that P?") by saying that S can *never* know that P at much higher rates than in the equivalent positive prompt ("How many times does S need to check F before she knows that P?"). Finally, we found evidence that the rate at which participants gave skeptical "never" responses in the negative frame and the rate at which participants gave "lazy knowledge" responses (*S knows without having to check*) in responses in the positive frame were *themselves* subject to a stakes effect in some of our evidence-seeking experiments. We argue that a stakes effect on skeptical "never" responses provides a new way of responding to an important theoretical criticism of the evidence-seeking approach to uncovering a stakes effect on knowledge.

2. The sensitivity of knowledge to stakes

Intellectualists about knowledge hold that "factors in virtue of which a true belief amounts to knowledge are exclusively truth-relevant, in the sense that they affect how likely it is that the belief is true" (DeRose 2009, p. 24). Those who believe that knowledge is stakes-sensitive are a type of *anti-intellectualist* about knowledge: anti-intellectualists "hold that whether a subject knows something or not depends in part on such non-truth-relevant 'practical' matters as the cost (to the subject) of being wrong" (Ibid., p. 25; see also Gerken 2017, p.34). Holding that knowledge is sensitive to stakes is only one way of being an anti-intellectualist. For example, one might hold that whether some subject knows something depends partly on whether various possibilities are salient to the subject (Hawthorne 2004; see Dinges 2017 for discussion), or on how much time is available to the subject (Shin 2014). Our focus in this paper will be on stakes-sensitivity about knowledge, the view that

whether a subject knows some proposition *p* depends on how much is at stake for that subject in being right or wrong about *p*.

Anderson and Hawthorne (forthcoming) precisify the notion of stakes at work in theories of the stakes sensitivity of knowledge (usually characterized as the stakes of being wrong about some proposition *p*) in the following way:

> A natural place to start is to articulate a measure of how much turns on *p* in performing a certain action A (where the core ideology focuses on a three-place relation between agents, actions, and propositions). Henceforth, we shall call this the '*p*-stakes of the action.'… the *p*-stakes of an action is a matter of the gap between the utility of what would happen if one performed that action and *p* were the case, and what would happen if one performed that action and *p* were not the case.[2]

To illustrate the notion of the "p-stakes" of an action, consider the following action, proposition, and contrasting pair of scenarios:

Action: Leaving the apartment without checking to see if I turned the stove off.

p: The stove is turned off.

Not-so-bad-scenario: I'm going out to check the mail, and I'll come back in five minutes, when I can check to see if the stove is off.

Bad scenario: I'm going out of town for a week and can't check to see if the stove is off until I'm back (and I live alone).

Compare the gap between the utility of leaving the apartment without checking to see if I turned the stove off given *p* versus given not *p* in the bad scenario versus the not-so-bad-scenario:

Bad scenario: if *p* is the case, my apartment does not explode because I left the gas on, while if *p* is not the case, it does.

Not-so-bad-scenario: if *p* is the case, my apartment does not explode because I left the gas on, while if *p* is not the case, it smells bad and I get a headache.

The greater differential in utility between outcomes in the bad scenario is what constitutes its "higher stakes" status. The stakes-sensitivity of knowledge can thus be understood as

[2] Anderson and Hawthorne go on to problematize the notion of "stakes sensitivity", but for our current purposes of evaluating the empirical evidence for theories that invoke the notion of stakes sensitivity, we will bracket their criticisms while exploiting their helpful precisified notion of stakes. See Armendt (2019) for additional dimensions ("odds" and "shape") along which stakes can vary beyond their size.

the claim that all else being equal, whether a subject is in the bad, high-stakes scenario or the not-so-bad, low-stakes scenario can make a difference to whether the subject knows that *p*.

3. What is a scalar stakes effect on knowledge?

Both the ordinary notion of "stakes" and the precisified notion given in the previous section are scalar, rather than binary, properties. That is, they admit of degrees of application beyond 0 and 1, and those degrees can be compared using expressions like "x is a higher-stakes scenario than y". It is possible to arrange scenarios in terms of increasing stakes. For example, the "bad" scenario discussed above is nowhere near as bad as things could get. For example, suppose that while I'm on vacation, I leave my cat at home with an automatic feeder. The utility of not checking to see if the gas is still on, given not-*p*, is that she's blown to smithereens along with my apartment. The stakes in that scenario are therefore even *higher* than the "bad" scenario. One can imagine how things could be even worse (imagine if I had two cats)—and thereby also how the stakes could be even higher. There is no obvious upper bound to the stakes scale—for any given "high stakes" scenario, there will be a scenario that will have even higher stakes. There is therefore no absolute notion of "high" stakes: whether stakes are high is a relative notion (Kennedy and McNally 2005).

Given that the notion of stakes is scalar, are stakes effects on knowledge (if there are any) also scalar? As stakes go up, how are judgments about knowledge affected? Do people attribute knowledge less and less as the stakes go up, or is there a "plateau" beyond which further increases in stakes stop affecting judgments about knowledge? How large do the differences in stakes have to be before one of the scenarios counts as "high" vs. "low" stakes?

In order to answer these questions, we designed experiments that created a variety of *stakes scales*, along which the stakes of being wrong about a particular proposition varied. To return to the example concerning whether the stove was left on, discussed in the previous section, instead of just a "bad" (high) and a "not-so-bad" (low) scenario, it could be supplemented with more scenarios that fill out the relevant stakes scale, as follows:

Action: Leaving the apartment without checking to see if I turned the stove off.

p: The stove is turned off.

Stakes 1, Not-so-bad-scenario: I'm going out to check the mail, and I'll come back in five minutes, when I can check to see if the stove is off.

Stakes 2, Bad scenario: I'm going out of town for a week and can't check to see if the stove is off until I'm back (and I live alone).

Stakes 3, Very bad scenario: I'm going out of town for a week and can't check to see if the stove is off until I'm back, and I live with a cat.

Stakes 4, Terrible scenario: I'm going out of town for a week and can't check to see if the stove is off until I'm back, and I live with a cat, a gray parrot, and I have several unpublished papers by J.L. Austin in my library, one of which contains a heretofore unknown and totally convincing response to external world skepticism.

One possibility that becomes clear when the stakes scale is expanded beyond two degrees is that the difference in stakes between any two adjacent points on the scale (between "not-so-bad" and "bad", for example) might not be big enough to trigger a stakes effect on knowledge. Previous experiments which failed to uncover a stakes effect, all of which only use two points on a stakes scale, might simply have failed to pick points on the stakes scale that were distinct enough for a stakes effect to show up.

In order to examine the scalarity of stakes effects on judgments about knowledge, we designed two experiments. The first experiment used "evidence-fixed" prompts: participants were asked, across four different points on the relevant stakes scale for six different scenarios (concerning paramedics racing to the scene of an accident, scientists checking the formula for a vaccine, mountaineers checking their climbing rope, participants on a game show thinking about an answer to a question, a moderator for a talk looking up the pronunciation of a guest speaker's name before introducing them, and a homeowner checking on her home sprinkler system in response to the threat of an arsonist), to agree or disagree with sentences attributing knowledge and denying knowledge to the protagonist of each scenario. In each scenario, the amount of evidence available to the protagonist remained fixed, while the point on the stakes scale was varied. The first experiment is discussed in §4. We also conducted a pre-registered replication of Sripada and Stanley's (2012) "evidence-fixed" study, which will be discussed in relation to our first experiment, and the details of which are presented in Appendix II.

The second experiment used the "evidence-seeking" prompts introduced in Pinillos (2012). Instead of asking participants to agree or disagree with statements that the protagonist knows or doesn't know a proposition, participants were asked to indicate how

much evidence the protagonist would have to gather in order to know that the relevant proposition is true (in the positive polarity condition), or how much evidence the protagonist could gather and still not know the relevant proposition (in the negative polarity condition). The scenarios and points on the relevant stakes scales were the same as in the "evidence-fixed" experiment. The second experiment is discussed in §5**.**

4. Experiment 1: The "Evidence-Fixed" Design

4.1 Experimental materials

Six scenarios were developed in which different types of stakes were manipulated (lives; physical injury; embarrassment; money; damage to objects of personal value). Four versions of each scenario were created in which the stakes were scaled in magnitude (see Supplementary Material for all scenarios and versions). For example, the 'vaccine' scenario involves changing the number of lives at stake should a vaccine be made incorrectly and administered to research participants:

> Stakes 1: Low
> Elaine is a medical researcher. Her task is to create a vaccine for a virus. Elaine has done this before, and she has a check list that specifies all of the steps she needs to take to make the vaccine. Elaine is following all of the steps correctly. Elaine's assistant has informed her that there is **one human research participant** who has volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participant **will give them mild cold-like symptoms**.

In the above 'low' stakes version of the scenario, one individual will experience mild symptoms. These stakes are then incrementally raised in the remaining versions of the scenario:

> Stakes 2
> …**one human research participant** who has volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participant **will kill him within days**.

> Stakes 3
> …**15 human research participants** who have volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participants **will kill them all within days**.

Stakes 4: High
…**100 human research participants** who have volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participants **will kill them all after several days of excruciating pain**.

After reading each scenario, participants were asked to respond to a knowledge prompt along a 7-point Likert-type scale (the endpoints of the Likert scale were labeled as follows: *1-Strongly disagree, 7-Strongly agree*). Whether participants received a negative or positive knowledge prompt was determined by initial condition assignment (positive polarity condition; negative polarity condition). For example, having read the vaccine scenario (above) participants in the positive polarity condition were asked to rate their level of agreement with the statement:

Elaine **knows** that she is making the vaccine correctly

Participants in the negative polarity condition were asked to rate their level of agreement with the statement:

Elaine **doesn't know** that she is making the vaccine correctly

To exclude any participants who failed to understand the experimental instructions, all participants had to respond correctly to a control question at the beginning of the experiment (Prompt 1) and a control question at the end of the experiment (Prompt 2) in order to be included in the data analysis. Participants who responded "agree" or "strongly agree" to Prompt 1 were removed and participants who responded "disagree" or "strongly disagree" to Prompt 2 were removed.[3]

Prompt 1
You have a fair coin, with heads on one side and tails on the other, that you flip into the air and catch on the back of your hand without looking at it.

Please indicate whether you **agree** or **disagree** with the following statement about the scenario:

You know the coin landed heads
1 (strongly disagree) – 7 (strongly agree)

Prompt 2
You have a fair coin, with heads on one side and tails on the other, that you flip into the air and catch on the back of your hand without looking at it.

[3] The controls were simplified versions of the coin flipping scenario used in Horvath and Wiegman (2016), Swain et al. (2008), and Weinberg et al. (2001).

Please indicate whether you **agree** or **disagree** with the following description of the scenario:

You don't know that the coin landed heads
1 (strongly agree) – 7 (strongly disagree)[4]

4.2 Procedure

Participants in both the positive and negative polarity conditions were presented with six scenarios in four different degrees of stakes each (24 pairs of scenario + degree of stakes) in a randomized block design. Using this design ensures that all scenarios occur once in a sequence (or block) before any of them is repeated.

4.3 Participants

One hundred and twenty participants were recruited from MTurk and paid $1.75 each for participating. This research received ethical approval from [removed for blind review] and informed consent was obtained from all participants. Procedures were performed *ex post* to identify suspicious or low-quality responses in all datasets.[5] Following screening procedures, eight responses were flagged as Virtual Private Server (VPS) responses and subsequently removed. As an additional screening procedure, participants who responded incorrectly to one of the above control prompts were removed from further data analysis. A total of 15 participants were removed having failed one or more of these checks, leaving a final sample of 97 participants (44 females, 52 males, 1 non-binary gender identity) between 20 and 71 years old ($M = 39.64$ years, $SD = 12.00$ years). Participants were randomly assigned to the positive polarity condition ($N = 55$) or the negative polarity condition ($N = 42$).

4.4 Hypothesis

If knowledge is sensitive to stakes, we should observe differences in levels of agreement with the knowledge prompts in at least some of the different degrees of stakes of a particular scenario. For example, we should find some difference between the "low stakes" and the "high stakes" versions of at least one of the six scenarios we tested, in either the positive or negative polarity. Independent of the stakes sensitivity of knowledge, there should also be an effect of switching the polarity of the prompt: If participants are willing to agree with the statement that the protagonist of a story *knows that p* in a particular scenario with a particular degree of stakes, participants should be less willing to agree with

[4] This is the only reversed Likert scale of all experiments reported in this paper.

[5] Given recent concerns and emerging evidence that the integrity of MTurk-based studies has been compromised by bots or responses from individuals using Virtual Private Servers (VPS) (faking their location), screening procedures were performed by identifying identical GPS locations with unique IP addresses, determining whether IP addresses derived from an Internet Service Provider (ISP) or data center, and evaluating open-ended responses against a set of criteria (for full details regarding this procedure see Dennis, Goodson, & Pearson (August 17, 2018)).

the statement that the protagonist *doesn't know that p* in the same combination of scenario and degree of stakes.

4.5 Results

Overall and contrary to our hypothesis, there were consistent levels of agreement and disagreement across all stakes versions of all scenarios (see Figure 1). There was no main effect of stakes in any of the scenarios, meaning that we did not detect any influence of what or how much was at stake on levels of agreement with knowledge prompts (positive or negative). Our hypothesis regarding polarity was supported; across all scenarios, participants were more willing to agree that the protagonist *knows that p* and less willing to agree that the protagonist *doesn't know that p* (for a full summary of individual statistics for each scenario, see Appendix I).

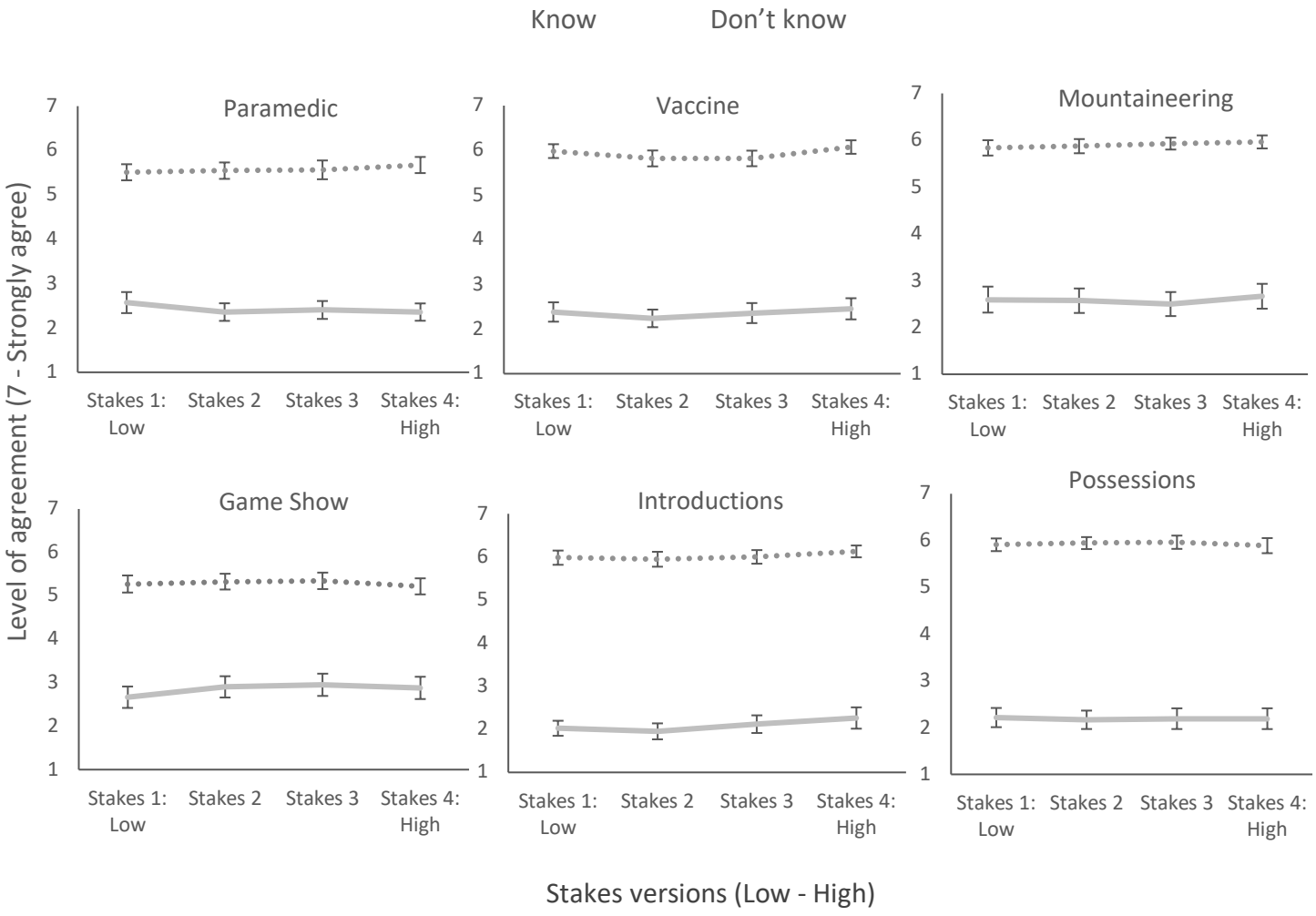


Figure 1. Levels of agreement and disagreement with knowledge prompts across all degrees of stakes for all six scenarios. Levels of agreement and disagreement are consistent across all stakes scales (note the lines are level with little slope). Levels of agreement are significantly higher for positive polarity prompts than for negative polarity prompts. Error bars represent +/- 1 Standard Error (SE).

4.6 Discussion

Using the evidence-fixed approach, we found no evidence of a stakes effect on judgments about knowledge. This finding was consistent across all six of our scenarios when varying what was at stake—whether what was at stake was people's lives, money, objects of personal value, or the degree of the protagonist's embarrassment.

While these findings are in accordance with several previous evidence-fixed studies which failed to find an effect, they conflict with findings from Sripada and Stanley (2012), who did find evidence of a stakes effect on judgments about knowledge using an evidence-fixed design. Given the conflicting findings, we carried out a pre-registered replication of Sripada and Stanley's (2012) experiment (see Appendix II for full method and analyses). Using Sripada and Stanley's between-groups experimental design, groups of participants were presented with one of six scenarios ($N = 58 – 68$ per group). Investigating three pairs of vignettes with a low and high stakes version, we asked all participants to respond to a question about the protagonist's strength of evidence (*what is the strength of S's evidence that P*), followed by a judgment about the protagonist's knowledge (level of agreement/disagreement with: *S knows that P*). In the "basic" vignettes, we replicated Sripada and Stanley's finding of an effect of stakes on quality of evidence: we found a medium-sized effect of stakes, with participants stating that the protagonist's evidence was weaker in the high stakes case (see the left-hand side of Figure 2). But we did not replicate their finding of a stakes effect on quality of evidence in the "implicit/explicit" or "ignorant" vignette pairs. In judgments about knowledge, we found a small effect of stakes with participants being less inclined to agree with the knowledge claim in the "basic" vignette pair (see the right-hand side of Figure 2). We did not find evidence of a stakes effect on judgments about evidence or knowledge in the "implicit/explicit" or "ignorant" vignette pairs; these controlled for stakes being implicitly versus explicitly described (implicit low/explicit high) and the protagonist being ignorant of the stakes involved (ignorant low/ignorant high) (see *Implicit/Explicit* and *Ignorant* bars in figures). This is the opposite pattern of stakes effects that Sripada and Stanley found; they found effects in the "implicit/explicit" and "ignorant" vignette pairs, but not in the "basic" vignette pairs.

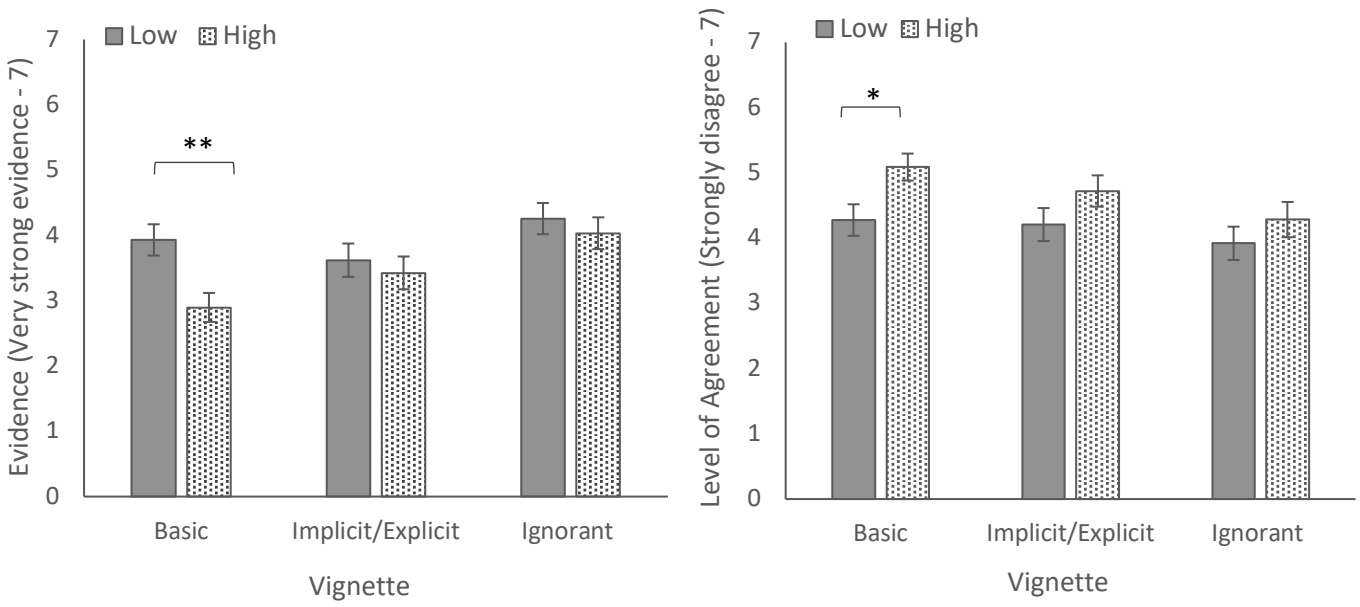


Figure 2: Strength of evidence across the three vignette pairs (*left hand panel*) and Levels of agreement and disagreement with the knowledge claim across the three vignette pairs. (*right hand panel*). Error bars represent +/- 1SE. * = small effect size. ** = medium effect size (Cohen's *d*). In the strength of evidence comparison (left panel) there was a medium effect of stakes in the basic vignette pair, ($t(124) = 3.17$, $p = .002$, $d = 0.57$) with strength of evidence higher in the low stakes scenario. The effect of stakes was not significant in the other vignettes [Implicit/Explicit: $t(117) = 0.54$, $p = .588$; Ignorant: $t(120) = 0.84$, $p = .511$]. For the levels of agreement comparison (right panel) there was a smaller effect of stakes in the basic vignette pair, ($t(124) = -.2.57$, $p = .011$, $d = 0.46$) with levels of agreement are higher in the low stakes scenario. Once again there was no significant effect of stakes in the other vignettes [Implicit/Explicit: $t(117) = -1.48$, $p = .142$; Ignorant: $t(120) = -0.98$, $p = .329$].

The results of our evidence-fixed experiment alongside the findings from our replication of Sripada and Stanley (2012) indicate that while the "evidence-fixed" experimental design is capable of uncovering stakes effects on judgments about knowledge, those effects are hard to find and, if found, are small. One additional worry about the effects found in the original Sripada and Stanley (2012) study—as well as our replication—is that the question asking participants whether they agree with the knowledge claim always follows the quality of evidence question—the order of those questions is not varied. This ordering could potentially be influencing how participants are responding to each prompt and contributing to the observed effect.[6] In our larger, more diverse "evidence-fixed" experiment, we did not ask participants about the quality of the evidence available to the protagonist before asking them whether they agreed with a claim about knowledge. That difference could account for the divergent finding in the different evidence-fixed experiments we conducted. We return to this possibility in our General Discussion (§6).

[6] This possibility is noted by Gerken (2017, p.267 n.8).

Although the results of the "evidence-fixed" design have provided only mixed results, the "evidence-seeking" design used in Pinillos (2012) and Pinillos and Simpson (2014) has consistently found evidence of stakes effects. For that reason, we therefore conducted an "evidence-seeking" version of our experiment to further probe for the scalarity of stakes effects.

5. Experiment 2: The "Evidence-Seeking" Design

5.1 Materials and Procedure

The same set of scenarios were presented to participants in the same randomized block design as in Experiment 1. As before, one group of participants received positive prompts and the other received negative prompts.

In this experiment, the original knowledge prompts from Experiment 1 were replaced with evidence-seeking prompts. For example, in response to the vaccine scenario, participants in the positive polarity condition were asked:

> How many times does Elaine need to consult her check list before she **knows** that she is making the vaccine correctly?
>
> ____ times

and participants in the negative polarity were asked:

> How many times can Elaine consult her check list and still **not know** that she is making the vaccine correctly?
>
> ____ times

As in Pinillos and Simpson (2014), after reading the positive polarity prompt, participants were asked to respond as follows:

> enter a whole number: 1, 2, 3... etc. If you think Elaine knows without having to check, write "0". If you think Elaine will never know no matter how many times she checks, write "never"

Participants were asked to respond as follows in the negative polarity condition:

> enter a whole number: 1, 2, 3... etc. If you think Elaine will never know no matter how many times she checks, write "never"[7]

[7] We did not include the option to write "0" *if you think S knows without having to check* in the negative polarity condition because the response sounded odd in response to the negative prompt.

5.2 Participants

One hundred and twenty participants were recruited from MTurk and paid $1.75 each for participating. This research received ethical approval from [removed for blind review] and informed consent was obtained from all participants. As before, screening procedures were performed *ex post* to identify suspicious or low-quality responses in all datasets. Following screening procedures, three suspicious responses were flagged and removed. Additionally, participants who responded incorrectly to one of two control prompts (these were the same coin control prompts used in Experiment 1) were removed from further data analysis. A total of 8 participants were removed having failed these checks leaving a final sample of 109 participants (54 females, 55 males) between 21 and 74 years old (M = 38.98 years, SD = 11.76 years). Participants were randomly assigned to a positive polarity condition (N = 58) or a negative polarity condition (N = 51).

5.3 Hypothesis

If knowledge is sensitive to stakes, we expect to find some effect of changing stakes on responses to a given polarity in a given scenario. For example, we should find some significant difference in responses to the "low" and "high" stakes conditions for positive or negative polarity prompts in at least one of the six scenarios we considered.

Regarding the effect of polarity, we do not expect to find a significant difference between responses to the negative and positive polarity prompts. That's because participants should be searching for the threshold of how much evidence is required for knowledge when presented with either the positive or the negative prompts. To explain why, consider a particular example of a positive prompt in the "high" stakes version of the vaccine scenario:

Positive: *How many times does Elaine need to consult her check list before she **knows** that she is making the vaccine correctly?*

If participants respond to the positive prompt with *3 times*, they should give a similar number in response to the negative prompt in the same combination of scenario and degree of stakes:

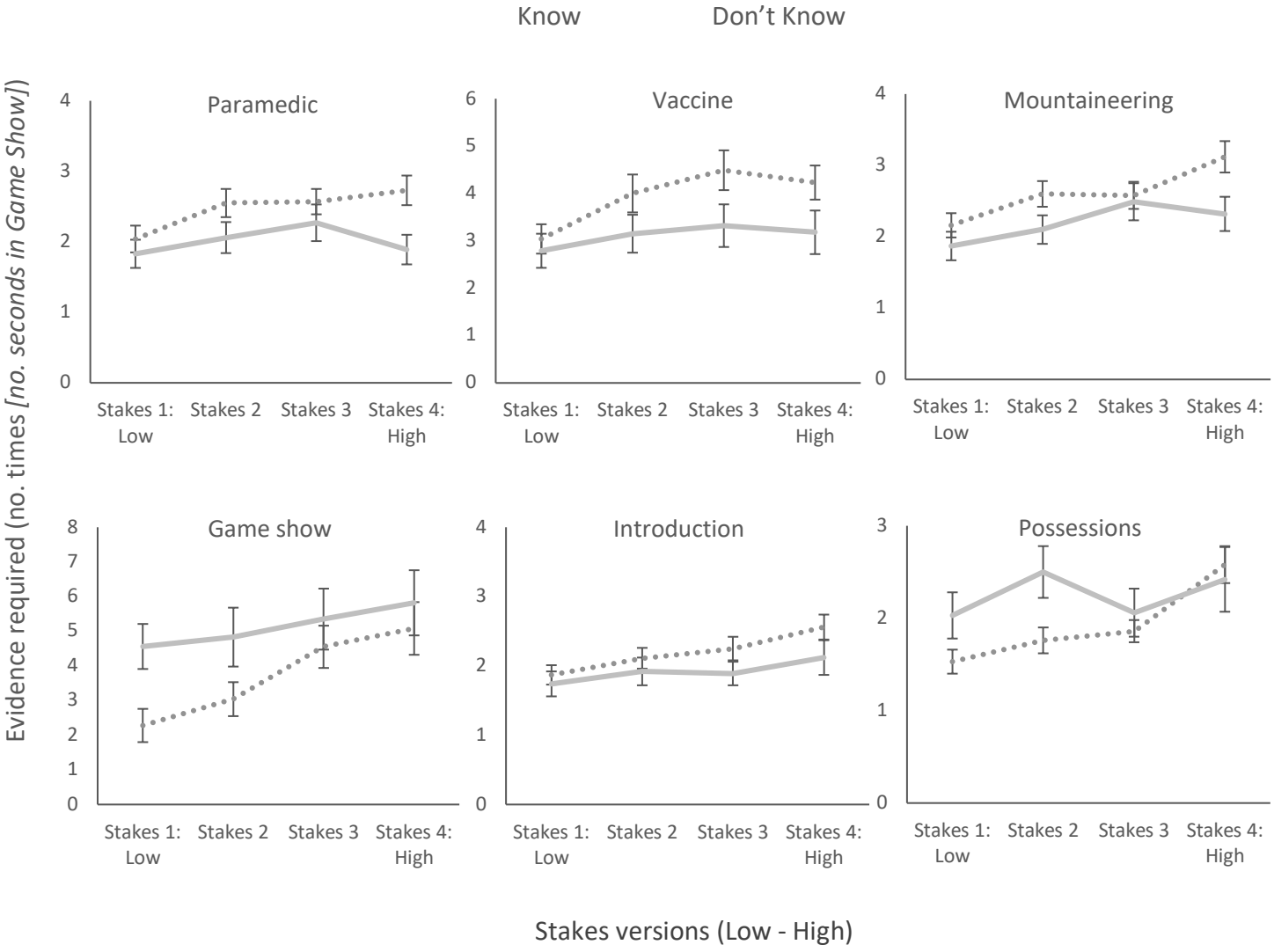
Negative: *How many times can Elaine consult her check list and still **not know** that she is making the vaccine correctly?*

It would be unexpected to find that participants would say that Elaine could consult her check list a significantly different number of times and *still not know that she's making the vaccine correctly* while saying that she needs to consult her check list only 3 times *before she knows that she's making the vaccine correctly.*[8]

[8] More precisely, we expect that responses to the negative polarity prompts should have values one lower than responses to the positive polarity prompts. If S needs to check N times before she knows that P, then the

5.4 Results

Overall and in line with our hypothesis, there were effects of changing stakes on responses across both the negative and positive polarities in all scenarios (in the possessions scenario, this change in responses was observed in the positive polarity only) (see Figure 3).[9] Across scenarios, as the stakes increased, there was a general pattern of participants stating that more evidence would be needed. In terms of the scalarity of these stakes effects, there was a clear pattern of lower stakes scenarios requiring less evidence than higher stakes scenarios (for a full summary of scenario and stakes scales analyses, see Appendix III). We did not observe any significant difference in responses to positive and negative polarity prompts (for a full summary of individual statistics for each scenario, see Appendix III).



maximum number she can check and still not know that P would be N-1. But we expected that we wouldn't be able to detect this difference, even if it does exist. Thanks to an anonymous referee for asking about this.

[9] The data summarised here follows outlier removal. For details regarding how outliers were removed and for replications of all analyses prior to outlier removal, see Appendix III.

Figure 3. The amount of evidence participants state is needed in order for the protagonist to know/or that can be had while the protagonist still doesn't know across all stakes versions of all six scenarios. Amount of evidence required increased as the stakes were raised in each scenario (note positive gradient of lines). In the possessions scenario, this effect was observed in the positive polarity only. Error bars represent +/- 1SE.

In the evidence-seeking task, we also collected two additional types of response from participants:

> enter a whole number: 1, 2, 3... etc. If you think Elaine knows without having to check, write "**0**" (*in the positive polarity condition*). If you think Elaine will never know no matter how many times she checks, write "**never**" (*in both the positive and negative polarity conditions*)

Previous papers simply discarded these "never" responses from further analysis (Pinillos 2014, p. 21) and in terms of zero responses, our previous analyses did not include values less than or equal to 0.[10]

Upon closer inspection, we noticed something unexpected about these responses: There was a much larger number of "never" answers given in response to the negative polarity prompt concerning how much evidence a protagonist could have and *still not know that p* than in response to the positive polarity prompt. "Never" answers made up 22% of the overall responses in the negative polarity group but only 2% of responses in the positive polarity group (see Figure 4). In effect, many more participants were responding to the prompt "skeptically"—that is, responding that they thought that *no matter how many times she checks, S will never know that p*.[11] We also observed a larger percentage of "0" responses to the positive (8.5%) than in the negative prompts (4%), but that isn't surprising given that participants weren't explicitly given the option to respond with "0" if the subject *knows without having to check* in the negative polarity prompts (see Figure 4). The meaning of a "0" response when given in response to a positive vs. a negative polarity prompt is therefore probably different: in response to a positive prompt such a response means S knows without having to check; in response to a negative prompt, it's not clear what a "0" response means.

[10] When the distribution of the dependent variable (in this case, *amount of evidence required*) is specified as a gamma distribution, any values that are less than or equal to 0 are not used in subsequent analysis.
[11] Pinillos and Simpson (2014, p. 40 n. 18) suggest that "never" responses "may indeed reveal a skeptical attitude toward the possibility of knowledge".

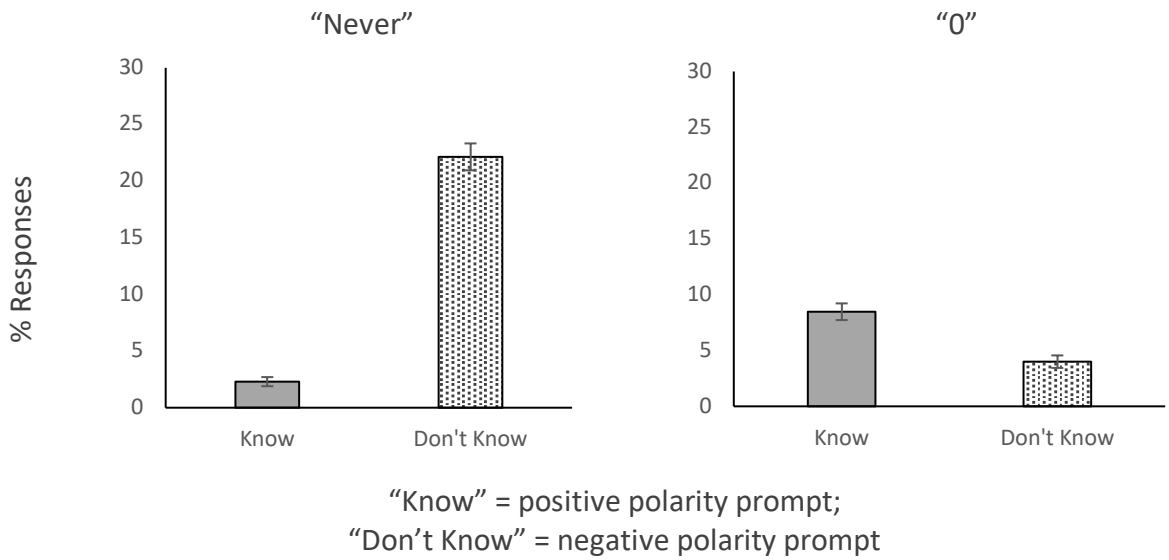


Figure 4. The percentage of "never" responses (left) and zero responses (right) given in response to positive and negative prompts (across all scenarios). The proportion of "never" responses was significantly higher in the negative polarity ($\chi^2(1) = 188.52$, $p < .001$). The proportion of "0" responses was significantly higher in the positive polarity ($\chi^2(1) = 28.51$, $p < .001$). Error bars represent +/- $1SE_p$ (converted to %).

Once we noticed these large disparities in "never" and "0" responses between the positive and negative polarity prompts, we wondered whether there might be a stakes effect on the rate at which participants gave these responses. If knowledge is sensitive to stakes, it's possible that the number of these declarations (*S will **never know** no matter how many times she checks* and *S **knows** without having to check*) will vary when the stakes vary.

In order to investigate this, we analysed the frequency of "never" and "0" responses in each stakes scale across all scenarios. We found no main effect of stakes in global analysis across all stakes scales for "never" responses (see left-hand side Figure 5).[12] But we did find an overall pattern of participants being less likely to say that *S knows P without having to check* ("0") as the stakes increased in the positive polarity (see right-hand side Figure 5).[13]

[12] Analysis GEE, poisson [loglinear]) model (count data) revealed a main effect of polarity, (Wald $X^2[1] = 34.33$, $p < .001$). There was no main effect of stakes, ($p = .522$). The interaction of polarity x stakes was not significant, (Wald $X^2[3] = 7.28$, $p = .064$).

[13] Analysis (GEE, poisson [loglinear]) revealed a main effect of stakes, (Wald $X^2[3] = 11.02$, $p = .012$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 8.65$, $p = .034$). There was no main effect of polarity, ($p = .168$). When interpreting the interaction, comparisons using sequential Bonferroni indicated a significant difference between the zero counts in the stakes 1 [*low*] scenarios and the stakes 3 scenarios ($p = .031$) and between the stakes 1 [*low*] scenarios and the stakes 4 [*high*] scenarios ($p = .031$) with a lower number in the higher stakes scenarios. This effect was present for the positive polarity only.

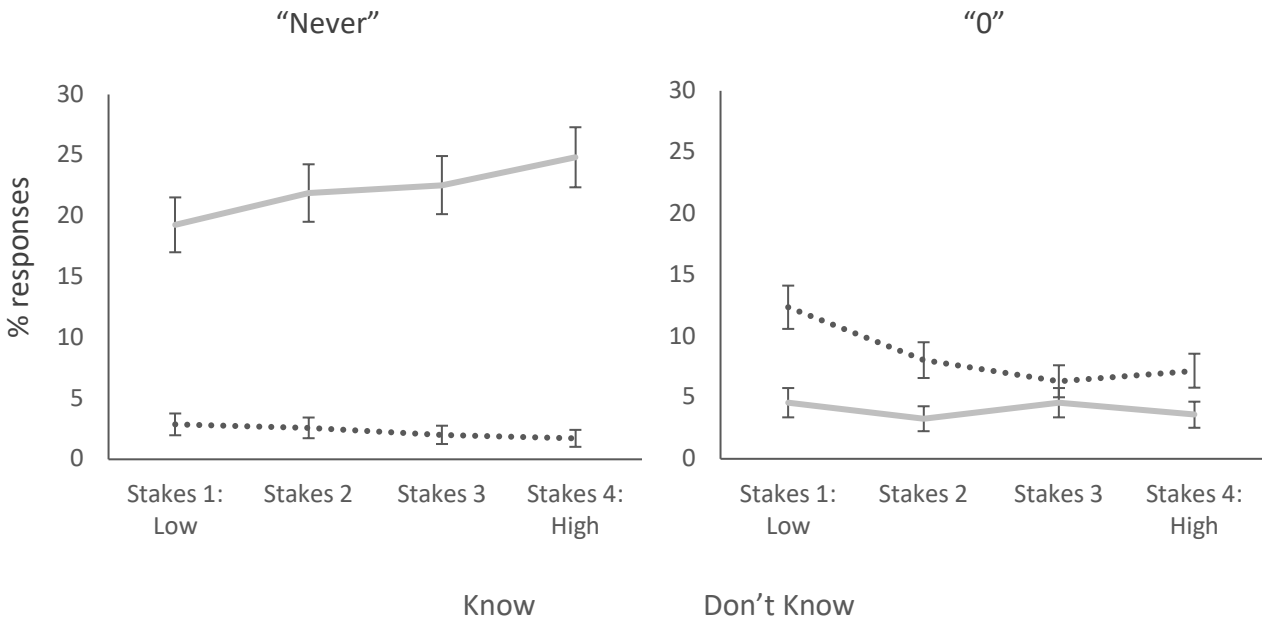


Figure 5. The percentage of "never" responses (left-hand side) and zero responses (right-hand side) given in each stakes scale (across all scenarios) in each polarity condition. The frequency of "never" answers given in response to negative prompts increased as the stakes were raised (note positive gradient of line) although this was not significant in inferential analyses. The frequency of zero answers given in response to positive prompts significantly decreased as the stakes were raised (note negative gradient of line). Error bars represent +/- $1SE_p$ (converted to %).

5.5 Discussion

Using the evidence-seeking approach, we found evidence of a stakes effect: As the stakes were raised, individuals stated that the protagonist would need to gather more evidence in order to know something or that she could gather more evidence and still not know something. This finding was consistent across the majority of our scenarios (bar the scenario involving personal possessions in which a stakes effect was found for the positive polarity only). We further found evidence that these stakes effects were scalar, not binary. In the majority of scenarios, participants stated that less evidence would be required by the protagonist in stakes 1 [*low*] cases when compared to the stakes 4 [*high*] cases. However, the details of how degrees on the scale of stakes affected knowledge judgments varied across different scenarios. For both the paramedic and vaccine scenarios (the vignettes in which lives were at stake), the stakes 1 [*low*] case was significantly different to the stakes 3 case with the amount of evidence plateauing after this point. That could be due to participants reaching a saturation point of how much the number of lives at stake affect knowledge. In the mountaineering scenario, the stakes 1 [*low*] case was significantly different to all of the subsequent stakes cases (i.e., stakes 2, stakes 3, and stakes 4 [*high*]). In the possessions scenario, the reverse pattern was found with the stakes 4 [*high*] case being significantly different to all previous stakes cases (i.e. stakes 1 [*low*], stakes 2, and stakes 3) Finally, in the game show scenario, the stakes 1 [*low*] case was significantly

different to the stakes 3 and stakes 4 [*high*] cases. But the stakes 2 case was also significantly different to the stakes 4 [*high*] case.

The variability we observed in scalar stakes effects across the different scenarios is unsurprising, given that (i) the scenarios differed in their details, including what the scale of stakes was measuring (lives, money, embarrassment, non-monetary possessions, personal injury), and (ii) we relied on our own intuitive, non-systematic sense of what would make for noticeable differences between different degrees of stakes on each scale. Future investigations of scalar stakes effects could systematically construct particular types of stakes scales in order to evaluate more precisely how the shape of particular stakes scales affect knowledge judgments. For example, according to Kahneman and Tversky's (1979) prospect theory, people experience monetary losses and gains with diminishing sensitivity: moving from $5 to $10 is experienced as a bigger gain than going from $95 to $100. And people are "loss-averse"—losing $5 hurts more than gaining $5 feels good. With these factors in mind, it would be possible to more systematically construct monetary stakes scales with the aim of making the differences between degrees on those scales more regular. That would also serve to bring the discussion of stakes understood as differences in expected utility (as on the Anderson and Hawthorne proposal discussed in §2) into contact with what we know about human decision making under risk and uncertainty. That is a potentially rich area for future experimental work on stakes effects.

With regards to the "never" responses, we found a large framing effect in the different polarities: people were more likely to respond with "never" (*S can never know that P*) when presented with a negative prompt. That result is surprising. We predicted that there should be no significant difference in numerical responses to the positive and negative polarity prompts, so we assumed that there should be no significant difference between the "never" responses to the positive and negative polarity prompts as well. What could be driving this framing effect? One possibility is that participants are being encouraged by the negative prompt to consider possible situations in which someone could fail to know that *P* in spite of having a great deal of evidence that *P* is the case. When that possibility is suggested, it might appear reasonable to respond that S will never know, no matter how much evidence she has, because she can't rule out the possibility that she is in such a situation. When confronted with the positive polarity prompt, in contrast, participants are not being encouraged to consider such a situation.[14]

The fact that we found a stakes effect on "0" responses in the positive polarity group is a previously unnoticed stakes effect: people are less likely to say that *S knows P without having to check* as the stakes increase. Although we did not find a significant stakes effect on the "never" responses, there was a pattern of people stating that *S can never know that*

[14] When you ask someone how many times one can check something and still not know it, it's reasonable to consider the fact that one can check as often as one likes and still not know it if, for example, one is checking carelessly. That is one way of understanding how the negative prompt encourages a greater frequency of "never" responses, by making salient the ever-present possibility of error. In contrast, when you ask someone how often one needs to check something before one knows it, it's reasonable to consider the fact that one might not need to check at all because there are other ways of knowing it besides checking. That could explain why the positive prompt encourages a greater frequency of "0" responses. Thanks to Alexander Dinges for discussion of this issue.

P in greater numbers as the stakes were raised. As we will discuss below, in a follow-up experiment we did observe a stakes effect on "never" responses.

In order to illuminate these unexpected findings, we ran two follow-up studies that addressed two worries that we had with the existing experimental design:

1. The existing negative prompt included the phrase (in bold): "*how many times can S check and* ***still not know*** *that P*" which may have triggered the presupposition that the protagonist does not know that P, which might have contributed to the framing effect on "never" and "0" responses.
2. The option to "write '0' if you think *S knows without having to check*" was only included after positive prompts (not negative prompts) which may have contributed to the framing effect on "never" responses by giving the positive polarity group a different response anchor.

To address our first concern, we ran a follow-up experiment (*Symmetrical Experiment*) which replicated the existing paradigm but with symmetrical negative and positive prompts, removing the phrase "*still not know"* (see Appendix IV for full experimental details). And to address our second concern, we ran another follow-up experiment (*Matched Experiment*) which replicated the design of the Symmetrical Experiment but simply removed the option to "write '0' if you think *S knows without having to check"* from the positive polarity prompts (see Appendix IV for full experimental details).

The framing effect on "never" responses was preserved across both experiments when controlling for the wording of the prompt and when removing the additional prompt option in the positive polarity (see Figure 6).

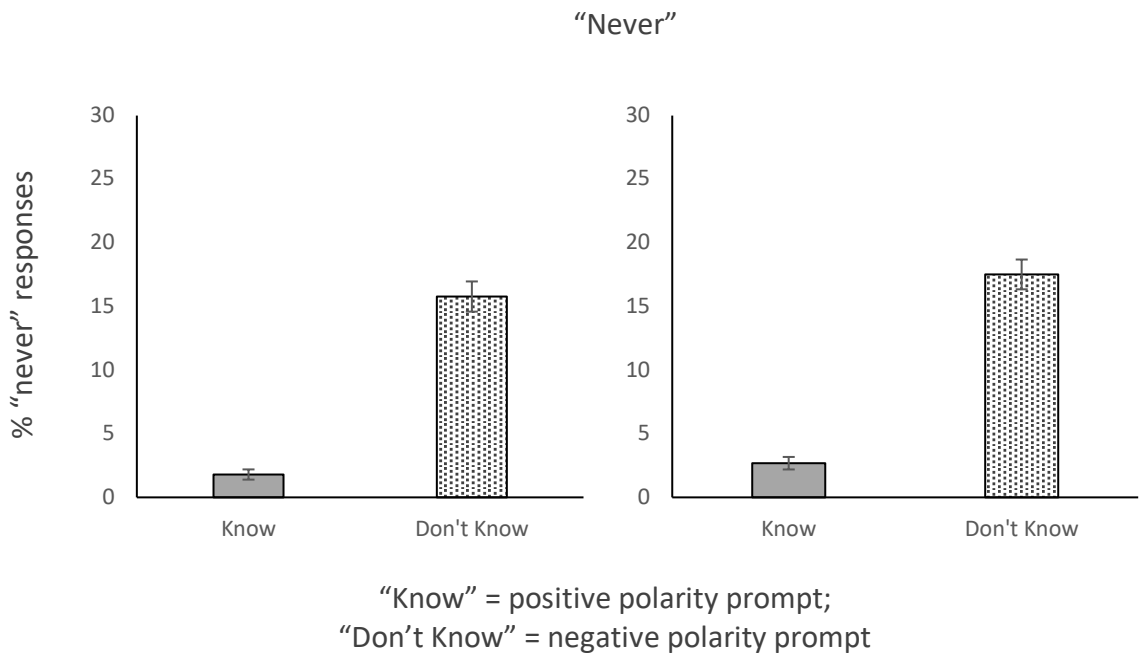


Figure 6. The percentage of "never" answers in the Symmetrical Experiment (left-hand side) and the Matched Experiment (right-hand side) given in response to positive and negative prompts (across all scenarios). The proportion of "never" responses was significantly higher in the negative polarity in the Symmetrical Experiment ($\chi^2(1) = 115.32$, $p < .001$) and in the Matched Experiment ($\chi^2(1) = 113.72$, $p < .001$). Error bars represent +/- $1SE_p$ (converted to %).

In the second, Symmetrical, experiment, we found a stakes effect on "never" responses (see figure 7).[15] (For full analyses, see Appendix IV).

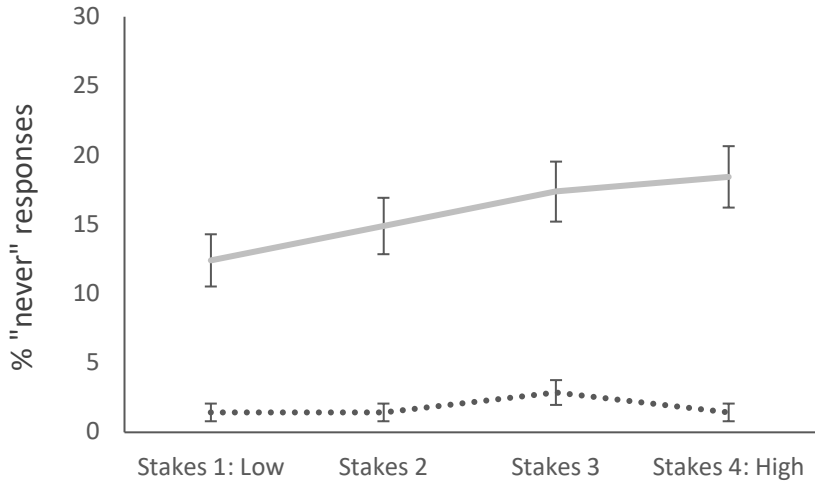


Figure 7. The percentage of "never" responses given in each stakes scale (across all scenarios) in each polarity condition of the Symmetrical Experiment. The frequency of "never" answers given in response to prompts increased as the stakes were raised from low stakes to stakes 3.

The finding of a stakes effect on "never" responses is important because it offers a response to a powerful theoretical objection that has been made to the use of evidence-seeking experiments as a way of investigating stakes effects on knowledge. Buckwalter and Schaffer (2015, p. 214) argue that the reason that changing stakes affect judgments about evidence-seeking prompts is that the evidence-seeking prompts contain a deontic modal ("need", or "has" in Pinillos 2012), and it is uncontroversial that what is at stake affects judgments about what someone needs to do; it is the effect of changing stakes on the deontic modal that is driving participants' responses, rather than any effect of stakes on *knowledge*. However, when participants respond with "never", they are responding to a secondary prompt that does not contain a deontic modal, namely:

*If you think Elaine will never know no matter how many times she checks, write **"never"**.*

[15] In order to investigate stakes effects on these responses, a GEE analysis with poisson (loglinear) model was performed on the frequency of "never" responses in the follow-up experiment using symmetrical prompts. Analysis revealed a main effect of polarity, (Wald $X^2[1] = 20.82$, $p < .001$), a main effect of stakes, (Wald $X^2[3] = 14.65$, $p = .002$), and a significant interaction of polarity x stakes, (Wald $X^2[3] = 7.88$, $p = .049$) (see Figure 7).

Since we found evidence that the rate at which participants respond with "never" is itself sensitive to stakes, that looks like evidence of a genuine stakes effect on participants' judgments about whether someone can ever know that p which cannot be explained as an effect of stakes interacting with an interpretation of a deontic modal.[16]

We offer this as a tentative response only, because we only found a stakes effect on skeptical "never" responses in one out of three of the evidence-seeking experiments we conducted. In the third and final Matched Experiment that we ran, as in the first experiment, we did not find a stakes effect on skeptical "never" responses. (See Appendix IV for detailed analyses.) The mixed findings across our three evidence-seeking experiments regarding the stakes effect on skeptical "never" responses might arise from the fact that the effect is small—this is something that will have to be addressed in future research designed specifically to investigate stakes effects on skeptical judgments.

A referee points out that our finding of a stakes effect on both positive *and* negative knowledge prompts ("S doesn't know that P") also presents a challenge to Buckwalter and Schaffer's objection to the results of evidence-seeking experiments. The negative prompt, "How many times can S [check] and not know that P?" contains a modal expression ("can"), but it's not a *deontic* modal—it's either an ability modal or a modal expressing metaphysical possibility. While stakes clearly affect what one should do (for example, in how we interpret the deontic modal "have" in the positive prompt), stakes don't obviously affect what one is able to do, or what is possible in a metaphysical sense. The stakes effect we found on responses to the negative prompts therefore isn't easily explained in the same way that Buckwalter and Schaffer explain the stakes effect on positive prompts in evidence-seeking experiments.

A stakes effect on "0" responses was found in the Symmetrical Experiment (see analyses in Appendix IV) although this time, the effect was seen across both prompts. So, participants were more likely to respond with "0" in lower stakes scenarios when presented with both positive and negative prompts.

As well as allowing us to investigate the "never" responses further, these follow-up experiments also served as opportunities to replicate the stakes effects found in the first evidence-seeking experiment. Overall, we found that the stakes effects observed in our first evidence-seeking experiment replicated across one or both of the follow-up experiments (for a full summary of individual scenario analyses see Appendix IV), with the paramedic scenario (number of lives at stake) and the game show scenario (amount of money at stake) consistently producing stakes effects across all three of the evidence-seeking experiments.

[16] A referee asks whether "never" responses could be consistent with the Buckwalter and Schaffer view, if such responses were understood as meaning that the subject should *never* perform the relevant action (checking the rope to see if it's secured in the climbing scenario, e.g.). We can't rule out this interpretation, but we find it implausible that participants who respond with "never" would say that the subject should never perform the relevant action.

6. General Discussion and Conclusions

6.1 Evidence-fixed versus evidence-seeking results

Using the classic evidence-fixed design employed in earlier experimental studies of stakes effects on knowledge, we did not find evidence of a stakes effect on judgments about knowledge across several epistemic scenarios.[17] However, in a second series of experiments which employ the evidence-seeking approach developed in Pinillos (2012), we did find evidence of a stakes effect across multiple scenarios.

Based on our findings and the results of previous studies, there are two types of competing explanations for this pattern of negative and positive results in the two types of experiment. The first type of explanation, which is favorable to the existence of genuine stakes effects on knowledge, is that there is some feature of the evidence-fixed design that obscures such a stakes effect. Pinillos (2012, p. 198) and Sripada and Stanley (2012, p. 10) hypothesize that stakes effects on knowledge exist, but they can be difficult to observe in evidence-fixed experimental designs because participants assume that protagonists in higher stakes situations will have gathered more evidence than those in lower stakes scenarios, leading to a tendency to judge that subjects know that P at *greater* rates in higher stakes situations, which would suppress any effect of higher stakes *lowering* the tendency to judge that subjects know that P. The possibility that participants may be revising their sense of how much evidence the protagonist has upwards in the high stakes case could therefore potentially explain why participants are equally likely to agree with the statement that the protagonist knows in the high stakes case as in the low stakes case.

Another factor that might be obscuring an underlying stakes effect in the evidence-fixed design is that the quality of the subject's evidence that P is so high in all conditions (from low to high) that stakes aren't having an observable effect on judgments about whether the subject knows that P, because judgments that the subject knows will already be at or near ceiling. The data represented in Figure 1 is compatible with this possibility, since agreement with "S knows that P" is consistently high across scenarios and degrees of stakes, and agreement with "S doesn't know that P" is consistently low across scenarios and degrees of stakes. It's possible that scenarios in which the subject's evidence is lower quality might leave room for the stakes effect to show up in participants' responses.[18]

A second type of explanation of the divergent patterns of results is not favorable to the existence of genuine stakes effects on knowledge. One version of this type of explanation holds that the failure to detect a stakes effect in most evidence-fixed experiments is because there is no stakes effect on knowledge, while the finding of an effect in the evidence-seeking experiments arises not from a stakes effect on *knowledge*, but a stakes effect on the deontic modal ("have") that appears in the evidence-seeking prompt (Buckwalter and Schaffer 2015).

---

[17] We did find evidence of a small effect of stakes on knowledge when replicating Sripada and Stanley's (2012) evidence-fixed experiment.

[18] Thanks to an anonymous referee for suggesting this possibility.

Our finding of a stakes effect on sceptical "never" responses in our second "Symmetrical" experiment, and our finding of a stakes effect on responses to the negative polarity prompts provide a novel response to this explanation: since neither the prompt to which the "never" responses are directly offered, nor the negative polarity prompts, contains a deontic modal, the observed stakes effect cannot be explained away as the effect of stakes on a deontic modal.[19]

Another version of the second type of explanation (which is not favorable to the existence of a stakes effect on knowledge) is proposed in Gerken (2017). Gerken explains the apparent stakes effect in Pinillos's evidence-seeking experiments as resulting from what he calls an "Epistemic Actionability-Proxy" heuristic, which leads participants to interpret the prompts in the evidence-seeking design as "concerning *how much evidence S should gather before it is reasonable to act on P*, rather than concerning the nature of knowledge" (p. 271).

While Gerken's explanation accounts for the stakes effect we found in responses to positive polarity prompts, it doesn't easily account for our finding of a stakes effect on "never" responses, and on responses to the negative polarity prompts ("How many times can S check and not know that P?"). Gerken might argue that "never" responses indicate that participants think that the subject in the scenario should never perform the relevant action (checking that the steps for creating the vaccine have been correctly carried out, e.g.), but that strikes us as implausible (see footnote 16, above). Also problematic for Gerken's explanation is the fact that we found a stakes effect in response to negative polarity prompts, which can't be interpreted as asking how much evidence S should gather before it is reasonable to act on P. He could potentially argue that the negative prompts are proxies for a question about how much evidence S could gather and yet *not* act on P, but as Nagel (2011) discusses, sentential negation (such as occurs in our negative prompts) generally triggers effortful type-2 processing, which would interfere with Gerken's heuristic-based (type-1) explanation of responses to the evidence-seeking prompts.

Like Gerken, Nagel (2008) provides an explanation of stakes effects in terms of a psychological effect that is not (directly) a stakes effect on knowledge, namely "need-for-closure". "Closure" is the arrival at a settled belief; prior to closure, one's mental state is "open" or non-committed (p. 287). "Low need-for-closure" is a state in which a subject is "strongly averse to inaccurate or premature judgment, as in [high-stakes scenarios]", while subjects in low-stakes scenarios are in a state of "neutral" need-for-closure (p. 288). Subjects who have a lower need-for-closure seek more evidence before settling on a belief, and are characterized by lower degrees of confidence in the belief even once settled. Given that subjects in high-stakes scenarios are also in a state of low need-for-closure, Nagel argues that stakes effects might be driven by the fact that it is *belief formation* that is directly sensitive to stakes, while knowledge is only sensitive to stakes indirectly (assuming belief is a component of knowledge). Nagel's competing explanation of stakes

[19] Our finding comports with the finding of a stakes effect on knowledge "retractions"—which also can't be explained away as the result of a stakes effect on a deontic modal—described in Dinges and Zakkou (ms.)

effects in terms of need-for-closure is not ruled out by the stakes effects we found in the evidence-seeking experiments.[20]

While our experiments do not rule out the possibility that stakes effects could be explained in terms of need-for-closure, Pinillos (2012, p. 202) ran an experiment in which participants were explicitly told that the subject in the scenario "forms the belief" that P before they are asked to judge how many times the subject has to check before she knows that P. Even with this modification to the evidence-seeking design, Pinillos still found a significant stakes effect on responses to the knowledge prompt, which provides reason to doubt that stakes are affecting knowledge indirectly through affecting belief formation.[21]

As discussed above, Gerken (2017) and Nagel (2008) use mechanisms drawn from cognitive psychology to explain apparent stakes effects in a way that is consistent with "intellectualist invariantism", the view that practical factors like stakes do not directly affect knowledge. Another approach to making intellectual invariantism compatible with the apparent stakes effects on knowledge is to invoke pragmatic linguistic mechanisms, like conversational implicature, to explain what participants are responding to when asked to judge whether a subject knows something.[22] To the best of our knowledge, no one has proposed a pragmatic explanation of the apparent stakes effect revealed in evidence-seeking experiments. But we did consider one possible pragmatic confound present in the first version of our evidence-seeking experiment, in the form of the presupposition trigger "still" that appeared in the negative prompts in our first evidence-seeking experiment ("How many times can S [check] and still not know that P?"). But we replicated our findings of a stakes effect and framing effect (in which there were far greater numbers of "never" responses in response to negative prompts than to positive prompts) even when "still" was removed from the negative prompts (see §5.5).

6.2 Framing effects and skepticism

We uncovered a large framing effect on participants' willingness to say that a subject in a scenario *never* can know that something is the case. These skeptical responses appeared at a much greater rate when participants responded to negative polarity prompts.

6.3. Advantages of our methodology

In order to investigate the stakes sensitivity of knowledge, we have incorporated a diverse set of scenarios that not only vary the type of things at stake but vary how much is at stake. Previous research has predominately incorporated a single pair of vignettes that involve more or less commonplace scenarios. By including a variety of stakes, from extreme cases

[20] Thanks to an anonymous referee for raising this objection.
[21] Pinillos's modified experiment does not rule out the possibility that stakes are affecting knowledge by affecting confidence, however. See Bach (2005, §V) for defense of that possibility.
[22] Brown 2006 and Rysiew 2007 offer such pragmatic explanations of patterns of judgments that might appear to lend support to epistemic contextualism (see Hansen and Chemla 2013 and Grindrod et al. 2018 for experimental evidence that such patterns exist); Dinges 2018 and Stoutenberg 2017 are recent challenges to such pragmatic explanations of patterns of contextualist judgments.

involving dozens of lives at risk in spectacular circumstances, to less extreme cases involving degrees of embarrassment, as well as vignettes that scale these stakes in magnitude, we can begin to build a finer-grained picture of the stakes sensitivity of knowledge than is possible from previous studies. Aside from being statistically more powerful, the variety of scenarios we employed also speaks to the generality of the effect – where we have consistently found (in the evidence-seeking design) stakes, framing and scalar effects across scenarios we can be more confident that such results are not unforeseen artefacts of the particular vignettes employed.

Stakes effects were elusive in our original evidence-fixed study, although a registered replication of Sripada and Stanley (2012) confirmed that such effects could be found (though the overall pattern of stakes effects on knowledge that we observed in our replication did not match those observed by Sripada and Stanley). In both the original Sripada and Stanley study and our replication, questions about the quality of the evidence available to the protagonist always preceded questioning the participants whether they agreed with a claim about knowledge, potentially contributing to the finding of an effect. In contrast, in our first experiment, where we did not find evidence of a stakes effect, we did not ask participants to assess the quality of evidence available. A direct comparison of evidence-quality question-present versus question-absent conditions is needed to clarify whether this difference contributes to determining whether stakes effects are observed in an evidence-fixed design.[23]

The contrast between the mixed results in evidence-fixed designs and the more consistent results in evidence-seeking designs serves to reinforce the notion that evidence-seeking designs are likely to be more informative experimental tools for further research on stakes effects.

Finally, because we included both negative and positive polarity prompts, we have been able to determine the role played by the polarity of a prompt is affecting judgments about knowledge. Most importantly, if we had not included this positive-negative prompt distinction, then we would not have been able to detect and interpret the large framing effects observed in our evidence-seeking experiments or uncover the stakes effects on "0" and "never" responses that we observed.

Though no single experimental investigation of a stakes effect on knowledge can definitively settle the existence of such an effect, the results of this study provide new reasons to think that such an effect exists.

[23] Sripada and Stanley (2012, p. 7) argue that the quality of evidence question is needed to focus participants' attention on the question of evidence in the knowledge question. Without it, they think, participants may be inclined to focus only on the factivity of the knowledge question.

Bibliography

Anderson, C. and Hawthorne, J. (2019). "Knowledge, Practical Adequacy, and Stakes", *Oxford Studies in Epistemology,* 6, 234-256.

Armendt, B. (2019). "Deliberation and Pragmatic Belief", in Kim, B., and McGrath, M. (eds), *Pragmatic Encroachment in Epistemology*, London: Routledge.

Bach, K. (2005). "The Emperor's New 'Knows'", in Preyer, G., and Peter, G. (eds), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, Oxford: Oxford University Press.

Brown, J. (2006). "Contextualism and Warranted Assertability Maneuvers", *Philosophical Studies,* 130(3), 407-435.

Buckwalter, W. (2010). "Knowledge isn't closed on Saturday: A study in ordinary language", *Review of Philosophy and Psychology,* 1(3), 395–406.

Buckwalter, W. (2014). "The Mystery of Stakes and Error in Ascriber Intuitions", in J. Beebe (ed), *Advances in Experimental Epistemology*, London: Bloomsbury.

Buckwalter, W., & Schaffer, J. (2015). "Knowledge, stakes, and mistakes", *Noûs*, *49*(2), 201-234.

Dennis, S. A., Goodson, B. M., & Pearson, C. *Mturk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note* (August 17, 2018). Available at SSRN: https://ssrn.com/abstract=3233954.

DeRose, K. (2009). *The Case for Contextualism*. Oxford: Oxford University Press.

Dinges, A. (2018). "Knowledge, Intuition, and Implicature". *Synthese*, 195(6), 2821-2843.

Dinges, A. (2017). "Anti-Intellectualism, Egocentrism, and Bank Case Intuitions", *Philosophical Studies*, online first, 1-17.

Dinges, A. and Zakkou, J. (ms.) "Much at Stake in Knowledge", unpublished ms.

Gerken, M. (2017). *On Folk Epistemology: How We Think and Talk about Knowledge*, Oxford: Oxford University Press.

Grindrod, J., Andow, J., and Hansen, N. (2018). "Third-Person Knowledge Ascriptions: A Crucial Experiment for Contextualism", *Mind & Language*, Early View, 1-25.

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

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

Horvath, J., and Wiegman, A. (2016). "Intuitive Expertise and Intuitions about Knowledge", *Philosophical Studies*, 173(10), 2701-2726.

Hawthorne, J. (2003). *Knowledge and Lotteries*. Oxford: Oxford University Press.

Kahneman, D. and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, 47(2), 263-291.

Kennedy, C. and McNally, L. (2005). "Scale Structure, Degree Modification, and the Semantics of Gradable Predicates", *Language*, 81(2), 345-381.

Nagel, J. (2008). "Knowledge Ascriptions and the Psychological Effects of Changing Stakes", *Australasian Journal of Philosophy*, 82(2), 279-294.

Nagel, J. (2011). "The Psychological Basis of the Harman-Vogel Paradox", *Philosopher's Imprint*, 11(5), 1-28.

Pinillos, A. (2012). "Knowledge, Experiments, and Practical Interests", in Brown, J. Gerkken, M. (eds), *New Essays on Knowledge Ascriptions*, Oxford: Oxford University Press.

Pinillos, A. and Simpson, S. (2014). "Experimental Evidence Supporting Anti-Intellectualism about Knowledge", in J. Beebe (ed), *Advances in Experimental Epistemology*, London: Bloomsbury.

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I., Cohnitz, D., Cova, F., Dranseika, V., Lagos, A. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., Hristova, E., Jraissati, Y., Kadreva, V., Karasawa, K., Kim, H., Kim, Y., Lee, M., Mauro, C., Mizumoto, M., Moruzzi, S., Olivola, C. Y., Ornelas, J., Osimani, B., Romero, C., Lopez, A. R., Sangoi, M., Sereni, A., Songhorian, S., Sousa, P., Struchiner, N., Tripodi, V., Usui, N., Vazquez del Mercado, A., Volpe, G., Vosgerichian, H. A., Zhang, X., and Zhu, J. (2017), "Nothing at Stake in Knowledge", *Noûs*, early view, 1-24.

Rysiew, P. (2007). "Speaking of Knowing", *Noûs*, 41(4), 627-662.

Shin, J. (2014). "Time Constraints and Pragmatic Encroachment on Knowledge", *Episteme,* 11(2), 157-180.

Sripada, S. and Stanley, J. (2012). "Empirical Tests of Interest-Relative Invariantism", *Episteme*, 9(1), 3-26.

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

Stoutenburg, G. (2017). “Strict Moderate Invariantism and Knowledge-Denials”, *Philosophical Studies*, 174(8), 2029-2044.

Swain, S., Alexander, J., and Weinberg, J. (2008). “The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp”, *Philosophy and Phenomenological Research*, 76(1), 138-155.

Weinberg, J.M., Nichols, S., and Stich, S. (2001). “Normativity and Epistemic Intuitions”, *Philosophical Topics*, 29(1&2), 429-460.

Appendix I: Experiment 1

1. Individual Scenario Analyses

For each individual scenario 2 x 4 mixed model Analysis of Variance (ANOVA) was performed on levels of agreement, with polarity (know, doesn't know) as the between-subjects factor and stakes scale (one [*low*], two, three, four [*high*]) as the within-subjects factor. This analysis was replicated with a generalized estimating equation (GEE) using a linear model. Note that, where these were conducted, the non-significant results of multiple comparison follow-up tests in the non-parametric analyses are likely due to the weak global stakes effect ($\eta^2 = .03$).

1.1 Paramedic Scenario

ANOVA found no main effect of stakes (p =.813) and no interaction of polarity x stakes (p =.333). There was a main effect of polarity, ($F(1, 95) = 149.03$, p <.001, $\eta^2 = .61$) with lower levels of agreement for the negative polarity prompts. GEE similarly revealed no main effect of stakes (p =.795) and no interaction between stakes x polarity (p =.500). There was a main effect of polarity, (Wald $X^2[1] = 154.42$, $p < .001$) as above.

1.2 Vaccine Scenario

ANOVA found no main effect of stakes, (p = .075) and no interaction of polarity x stakes (p =.817). There was a main effect of polarity, ($F(1, 95) = 212.46$, $p < .001$, $\eta^2 = .69$) with lower levels of agreement for the negative polarity prompts. GEE revealed a main effect of stakes, (Wald $X^2[3] = 8.58$, $p = .035$) but no interaction between stakes x polarity (p =.863). There was a main effect of polarity, (Wald $X^2[1] = 206.30$, $p < .001$) as above. Follow-up tests (sequential Bonferroni) examining the main effect of stakes were non-significant (ps >.092).

1.3 Mountaineering Scenario

ANOVA found no main of stakes (p = .650) and no interaction of polarity x stakes (p = .776). There was a main effect of polarity ($F(1, 95) = 163.29$, $p < .001$, $\eta^2 = 63$) with lower levels of agreement for the negative polarity prompts. GEE revealed a no main effect of stakes (p = .617) and no interaction between stakes x polarity (p =.789). There was a main effect of polarity, (Wald $X^2[1] = 147.06$, $p < .001$) as above.

1.4 Game Show Scenario

ANOVA revealed no main of stakes (p = .252) and no interaction of polarity x stakes (p = .513). There was a main effect of polarity ($F(1, 95) = 72.55$, $p < .001$, $\eta^2 = .43$) with lower levels of agreement for the negative polarity prompts. GEE revealed a no main effect of stakes (p = .088) and no interaction between stakes x polarity (p =.417). There was a main effect of polarity, (Wald $X^2[1] = 71.27$, $p < .001$) as above.

1.5 Introductions Scenario

ANOVA found no main of stakes (p = .055) and no interaction of polarity x stakes (p = .803). There was a main effect of polarity ($F(1, 95) = 278.95$, $p < .001$, $\eta^2 = .75$) with lower levels

of agreement for the negative polarity prompts. Follow-up tests (Bonferroni) examining the stakes effect were non-significant (*ps* >.091). GEE revealed no main effect of stakes (p = .074) and no interaction between stakes x polarity (p =.871). There was a main effect of polarity, (Wald $X^2[1] = 275.46$, $p < .001$) as above.

3.6 Possessions Scenario

ANOVA found no main of stakes (p = .983) and no interaction of polarity x stakes (p = .954). As expected, there was a main effect of polarity ($F(1, 95) = 323.81$, $p < .001$, $\eta^2 = .77$) with lower levels of agreement for the negative polarity prompts. GEE revealed no main effect of stakes (p = .995) and no interaction between stakes x polarity (p =.931). There was a main effect of polarity, (Wald $X^2[1] = 308.64$, $p < .001$) as above.

Appendix II: Registered Replication of Sripada and Stanley (2012)

1. Open Science Protocol

Adopting the experimental design of Sripada and Stanley (2012), we preregistered the experiment (background, methods, and power analysis) using the Open Science Framework repository (osf.io/sqeau). Our preregistration was submitted prior to data collection and is accessible to the public.

2. Participants

Four hundred and thirty participants (183 females, 246 males, 1 non-binary gender identity) between 20 and 67 years (M = 35.99 years, SD = 10.55 years) were recruited from MTurk and paid $0.50 for participating. This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants. As before, screening procedures were performed *ex post* to identify suspicious or low-quality responses in all datasets. Following screening procedures, 63 VPS and further suspicious responses were flagged and removed. As in the original experiment, participants were randomly assigned to one of six conditions (N = 58 – 68 per condition).

3. Materials and Procedure

As in the original study (Sripada & Stanley, 2012), we investigated three pairs of vignettes, each of which had a low and high stakes version. After reading one vignette, participants were asked to respond to two questions:

1. What is the strength of Hannah's evidence that her noodles are not topped with pine nuts?

2. Suppose it turns out that her noodles are not topped with pine nuts. Please rate how strongly you agree or disagree with the following sentence:

   "Hannah knows her noodles are not topped with pine nuts."

Participants responded to the first question (*evidence prompt*) along a 7-point Likert-type scale (*1-Very weak evidence, 7-Very strong evidence*) and to the second prompt (*knowledge prompt*) along another 7-point Likert-type scale (*1-Strongly agree, 7-Strongly disagree*). Questions were presented in the same order (as above).

Appendix III: Experiment 2

Overall, 32 "never" responses were given in the positive polarity condition and 271 "never" responses were given in the negative polarity condition. These responses were removed from main analyses and analysed separately. Responses across both polarity conditions were positively skewed across all stakes conditions and all scenarios. Two analyses were subsequently performed across the stakes versions of each scenario. Given the violations of normality and homogeneity of variance in the data a Generalised Estimating Equation (GEE) gamma (log link) analysis was conducted first with stakes (low; one; two; high) as within-subjects factor and polarity (positive; negative) as between-subjects factor. A second GEE analysis was subsequently conducted following the removal of extreme outliers. Datasets were log-transformed and outlier detection performed on the normalised distributions. Outliers were identified as those outside the range of: *median +/- 2.5 x Median Absolute Deviation (MAD)*. Analysis was then performed on the non-transformed, skewed data with outliers now removed.

1. Individual Scenario Analyses

Across all scenarios, there was a general pattern of participants stating that more evidence would be required by the protagonist in order to know something (positive polarity)/still not know something (negative polarity) as the stakes increased and this pattern remained following outlier removal.

1.1 Paramedic Scenario

Initial analysis ($N = 337$, zero values ignored) revealed no main effect of stakes, ($p = .082$) and no main effect of polarity ($p = .092$). There was a significant interaction of polarity x stakes, (Wald $X^2[3] = 11.29$, $p = .010$). Follow-up tests using sequential Bonferroni revealed a significant difference between the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario ($p = .024$) in the positive polarity only.

Having extracted extreme outliers ($N = 444$ to $N = 325$, zero values ignored), the second analysis revealed a main effect of stakes, (Wald $X^2[3] = 11.14$, $p = .011$). There was no main effect of polarity, ($p = .055$) and no significant interaction of polarity x stakes, ($p = .215$). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 3 scenario ($p = .007$). These effects were across polarity.

1.2 Vaccine Scenario

Initial GEE analysis ($N = 383$, zero values ignored) revealed a main effect of stakes, (Wald $X^2[3] = 19.91$, $p < .001$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 15.17$, $p = .002$). Follow-up tests using sequential Bonferroni found no significant differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset. There was no main effect of polarity, ($p = .058$).

Having extracted extreme outliers ($N = 436$ to $N = 355$), the second analysis likewise revealed a main effect of stakes, (Wald $X^2[3] = 9.17$, $p = .027$). There was no main effect of polarity, ($p = .065$) and no interaction of polarity x stakes, ($p = .676$). Comparisons using

sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 3 scenario (p =.032). These effects were across polarity.

1.3 Mountaineering Scenario

An initial analysis (N = 374, zero values ignored) found no main effect of stakes, (p =.190) and no interaction of polarity x stakes, (p = .533). There was a main effect of polarity, (Wald X^2[1] = 4.23, p =.040). The second analysis revealed a main effect of stakes, (Wald X^2[3] = 18.62, p <.001). There was no main effect of polarity, (p =.090) and no interaction of polarity x stakes, (p =.061). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 2 scenario (p =.011), the stakes 1 [*low*] scenario and the stakes 3 scenario (p =.003), and the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p <.001). These effects were across polarity.

1.4 Game Show Scenario

Initial analysis (N = 339, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 18.30, p <.001). There was no main effect of polarity, (p =.532) and no interaction of polarity x stakes, (p = .061). Comparisons using sequential Bonferroni found no differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset.

Having extracted extreme outliers (N = 436 to N = 297, zero values ignored), the second GEE analysis likewise revealed a main effect of stakes, (Wald X^2[3] = 13.48, p =.004). There was no main effect of polarity, (p =.281) and no interaction of polarity x stakes, (p =.502). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 3 scenario (p =.019) and the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p =.007). There was also a significant difference between the stakes 2 scenario and the stakes 4 [*high*] scenario (p =.021). These effects were across polarity.

1.5 Introduction Scenario

Initial analysis (N = 376, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 17.02, p =.001) and a main effect of polarity, (Wald X^2[1] = 4.06, p =.044). There was no interaction of polarity x stakes, (p = .629). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 3 scenario (p =.016) and the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p =.021). These effects were across polarity.

Having extracted extreme outliers (N = 436 to N = 364), the second GEE analysis likewise revealed a main effect of stakes, (Wald X^2[3] = 12.59, p =.006). There was no main effect of polarity, (p =.171) and no interaction of polarity x stakes, (p =.757). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p =.005). These effects were across polarity.

1.6 Possessions Scenario

Initial GEE analysis (N = 336, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 21.68, p <.001) and a significant interaction of polarity x stakes, (Wald X^2[3] = 10.19,

p =.017). There was no main effect of polarity, (p = .541). Comparisons using sequential Bonferroni found no differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset.

Having extracted extreme outliers (N = 436 to N = 322), the second GEE analysis revealed a main effect of stakes, (Wald X^2[3] = 17.04, p =.001) and a significant interaction of polarity x stakes, (Wald X^2[3] = 8.34, p =.040). There was no main effect of polarity, (p =.137). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p <.001), the stakes 2 scenario and the stakes 4 [*high*] scenario (p <.001), and the stakes 3 scenario and the stakes 4 [*high*] scenario (p =.001) in the positive polarity only.

2. Never and Zero Responses

In order to interpret the full range of responses given in the evidence-seeking experiments, separate analyses were performed on both "never" and "0" responses which could be given in response to the following prompts:

Never responses: *If you think S will never know no matter how many times she checks, write "never"*

Zero responses: *If you think S knows without having to check, write "0"*

The number of "never" responses given in response to scenarios in the negative polarity were significantly higher across all three experiments when compared to the positive polarity conditions (see the table below).

Experiment	**Know** (**"Never" count**)	**Don't Know** (**"Never" count**)	**Chi-Square**
Original Prompts	32	271	$\chi^2(1) = 188.52, p < .001$

The number of "0" responses given in response to scenarios in the positive polarity were higher in both the original prompts and symmetrical prompts experiments when compared to the negative polarity conditions (see the table below). Note that zero responses are not recorded for the *matched prompts follow-up* experiment as the response option was removed to create the *matched* design (i.e. *If you think S knows without having to check, write "0"*).

Experiment	**Know** (**"0" count**)	**Don't Know** (**"0" count**)	**Chi-Square**
Original Prompts	118	49	$\chi^2(1) = 28.51, p < .001$

Appendix IV: Follow-up evidence-seeking experiments

1. Symmetrical Experiment

1.1 Materials and Procedure

The same set of scenarios were presented to participants in the same randomized block design as in Experiment 2. For this follow-up experiment, the prompts were modified to ensure symmetry between the positive and negative polarities. For example, in response to the vaccine scenario, participants in the positive polarity condition were asked:

> What is the minimum numbers of times Elaine needs to consult her check list before she **knows** that she is making the vaccine correctly?
>
> ____ times

and participants in the negative polarity were asked:

> What is the maximum number of times Elaine can consult her check list and **not know** that she is making the vaccine correctly?
>
> ____ times

As in the first evidence-seeking experiment, after reading the positive polarity prompt, participants were asked to respond as follows:

> enter a whole number: 1, 2, 3... etc. If you think Elaine knows without having to check, write "0". If you think Elaine will never know no matter how many times she checks, write "never"

Participants were asked to respond as follows in the negative polarity condition:

> enter a whole number: 1, 2, 3... etc. If you think Elaine will never know no matter how many times she checks, write "never"

1.2 Participants

One hundred and twenty-one participants were recruited from MTurk and paid $1.75 each for participating. This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants. As before, screening procedures were performed *ex post* to identify suspicious or low-quality responses in all datasets. Following screening procedures, one suspicious response was flagged and removed. Additionally, participants who responded incorrectly to one of two control prompts (as in the first evidence-seeking experiment) were removed from further data analysis. A total of 12 participants were removed having failed these checks. Additionally, three participants were removed having already completed the first evidence-seeking experiment, leaving a final sample of 105 participants (45 females, 59 males) between 21 and 65 years old ($M = 37.09$ years, $SD = 10.67$ years). Participants were randomly assigned to a positive polarity condition ($N = 58$) or a negative polarity condition ($N = 47$).

1.3 Missing Values, Normality, and Outliers

Overall, 25 “never” responses were given in the positive polarity condition and 178 “never” responses were given in the negative polarity condition. These responses were removed from main analyses and analysed separately. As in the first evidence-seeking experiment, data were non-normal with responses in the both conditions being positively skewed. Two analyses were subsequently performed across the stakes versions of each scenario; given normality and homogeneity of variance violations in the data, a Generalised Estimating Equation (GEE) (non-parametric equivalent) with stakes (one [*low*]; two; three; four [*high*]) as within-subjects factor and polarity (positive; negative) as between-subjects factor) was initially conducted and the results of a second GEE analysis following the removal of extreme outliers are also reported.

1.4 Summary of Descriptive Statistics (prior to outlier removal and following removal)

Across all scenarios, there was a general pattern of participants stating that more evidence would be required by the protagonist in order to know something (positive polarity)/still not know something (negative polarity) as the stakes increased. This pattern of responses remained with extreme outliers removed There was a general pattern of participants stating that more evidence would be required by the protagonist in order to know something (positive polarity)/still not know something (negative polarity) as the stakes increased.

In terms of replicating the stakes effects observed in the first evidence-seeking experiment, we found a main effect of stakes in the paramedic and game show scenarios (see Figure 9). However, we did not observe a stakes effect in the vaccine, mountaineering, introductions, or possessions scenarios. Additionally; across four of the six scenarios, we did not observe significant differences in responses to positive and negative polarity prompts. An effect of polarity in the paramedic and mountaineering scenarios, in which evidence scores were higher for the positive polarity. A full breakdown of analysis by scenario follows this summary.

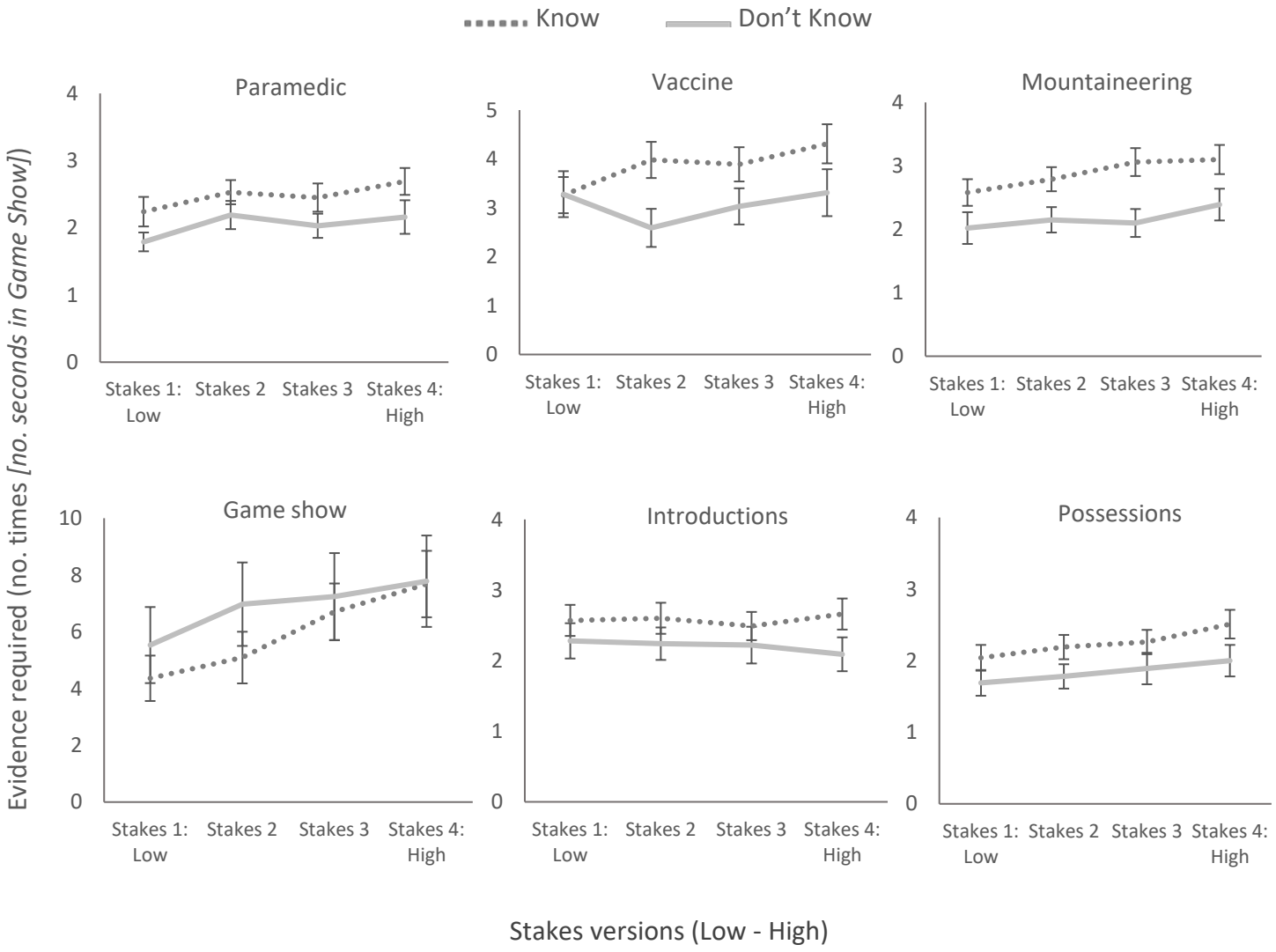


Figure 9. The amount of evidence participants state is needed in order for the protagonist to know/still not know across all stakes versions of all six scenarios. These figures show the data after removal of extreme outliers (i.e., the data used in the second analyses reported below). Amount of evidence required increased as the stakes were raised across polarities in the paramedic and game show scenarios (note positive gradient of lines) however in this experiment, there was no main effect of stakes in the vaccine, mountaineering, introduction, or possessions scenarios. Error bars represent +/- 1 SE.

1.5 Individual Scenario Analyses

1.5.1 Paramedic Scenario

Initial analysis ($N = 359$, zero values ignored) revealed a main effect of stakes, (Wald $X^2[3] = 16.12$, $p = .001$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 16.26$, $p = .001$). Follow-up tests using sequential Bonferroni revealed no significant differences

between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset. There was no main effect of polarity, (p = .096).Having extracted extreme outliers (N = 359 to N = 339, zero values ignored), further analysis confirmed a main effect of stakes, (Wald X^2[3] = 11.50, p =.009) and a main effect of polarity, was also observed (Wald X^2[1] = 4.19, p =.041). There was no significant interaction of polarity x stakes, (p =.933). Follow-up tests using sequential Bonferroni revealed no significant differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to a small global effect. Sequential Bonferroni comparisons did reveal that evidence scores were significantly higher in the positive polarity group when compared to the negative polarity group, (p =.043).

1.5.2 Vaccine Scenario

Initial GEE analysis (N = 365, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 9.65, p =.022), and main effect of polarity, (Wald X^2[1] = 14.64, p <.001) and a significant interaction of polarity x stakes, (Wald X^2[3] = 19.46, p <.001). Follow-up tests using sequential Bonferroni revealed no significant differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset. Having extracted extreme outliers (N = 365 to N = 325), further analysis found no main effect of stakes, (p =.087), no main effect of polarity, (p =.100) and no significant interaction of polarity x stakes, (p =.063).

1.5.3 Mountaineering Scenario

Initial analysis (N = 389, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 32.62, p <.001) and a main effect of polarity, (Wald X^2[1] = 4.05, p =.044). There was no interaction of polarity x stakes, (p = .083). There was a main effect of polarity. Comparisons using sequential Bonferroni found no significant differences between the polarities. In terms of differences between the stakes scenarios, sequential Bonferroni comparisons found no significant differences between stakes scenarios. These non-significant follow-up tests are likely due to large variances in the dataset. Having extracted extreme outliers (N = 389 to N = 362, zero values ignored), another GEE analysis found no main effect of stakes, (p =.112) and no interaction of polarity x stakes, (p =.514). There was a main effect of polarity, (Wald X^2[1] = 7.61, p =.006) with evidence scores higher in the positive polarity group when compared to the negative polarity, (p =.005).

1.5.4 Game Show Scenario

Initial analysis (N = 337, zero values ignored) found no main effect of stakes, (p =.377), no main effect of polarity, (p =.625) and no interaction of polarity x stakes, (p = .057). Having extracted extreme outliers (N = 337 to N = 313, zero values ignored), further GEE analysis revealed a main effect of stakes, (Wald X^2[3] = 10.24, p =.017). There was no main effect of polarity, (p =.556) and no interaction of polarity x stakes, (p =.757). Comparisons using sequential Bonferroni indicated that there was a significant difference between the stakes 1 [*low*] scenario and the stakes 3 scenario (p =.013) and the stakes 1 [*low*] scenario and the stakes 4 [*high*] scenario (p =.016). These effects were across polarity.

1.5.5 Introduction Scenario

An initial GEE analysis ($N = 375$, zero values ignored) revealed a significant interaction of polarity x stakes, (Wald $X^2[3] = 9.28$, $p = .026$). There was no main effect of polarity, ($p = .130$) and no main effect of stakes, ($p = .093$). Comparisons using sequential Bonferroni comparisons found no significant differences between stakes scenarios. After extracting extreme outliers ($N = 375$ to $N = 355$), another GEE analysis found no main effect of stakes, ($p = .961$), no main effect of polarity, ($p = .168$) and no interaction of polarity x stakes, ($p = .756$).

1.5.6 Possessions Scenario

An initial GEE analysis was performed using a poisson[1] (log link) model ($N = 381$). Analysis revealed a main effect of stakes, (Wald $X^2[3] = 92.12$, $p < .001$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 47.71$, $p < .001$). There was no main effect of polarity, ($p = .061$). Comparisons using sequential Bonferroni found no differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset. Having extracted extreme outliers ($N = 336$, zero values excluded), another GEE analysis with gamma (log link) model was performed on the evidence scores. Analysis revealed no main effect of stakes, ($p = .119$) and no interaction of polarity x stakes, ($p = .989$). There was a main effect of polarity, (Wald $X^2[1] = 4.47$, $p = .034$) with higher levels of evidence given in response to positive polarity prompts[2].

2. Matched Experiment

2.1 Materials and Procedure

The same set of scenarios were presented to participants in the same randomized block design as in Experiment 2. For this follow-up experiment, the same set of symmetrical prompts were used in both the positive and negative polarity groups. However, the additional option to write “0” following the positive polarity prompts, was removed:

> *Original*
> enter a whole number: 1, 2, 3... etc. If you think Elaine knows without having to check, write "0". If you think Elaine will never know no matter how many times she checks, write "never"
>
> *Modified*
> enter a whole number: 1, 2, 3... etc. If you think Elaine will never know no matter how many times she checks, write "never"

2.2 Participants

One hundred and twenty participants were recruited from MTurk and paid $1.75 each for participating. This research received ethical approval from the Department of Philosophy, University of Reading, UK and informed consent was obtained from all participants. As

[1] Gamma distribution was not used in this instance as the data violated assumptions for analysis (errors in computing the inverse log-link function).
[2] Results replicated using Poisson distribution (log function).

before, screening procedures were performed *ex post* to identify suspicious or low-quality responses in all datasets. Following screening procedures, seven suspicious responses were flagged and removed. Additionally, participants who responded incorrectly to one of two control prompts were removed from further data analysis. A total of nine participants were removed having failed these checks. Additionally, 15 participants were removed having already completed the first evidence-seeking experiment, leaving a final sample of 89 participants (33 females, 56 males) between 20 and 70 years old ($M = 34.71$ years, $SD = 10.84$ years). Participants were randomly assigned to a positive polarity condition ($N = 45$) or a negative polarity condition ($N = 44$).

2.3 Missing Values, Normality, and Outliers

Overall, 29 "never" responses were given in the positive polarity condition and 185 "never" responses were given in the negative polarity condition. These responses were removed from main analyses and analysed separately.

As in the first evidence-seeking experiment, data were non-normal with responses in the both conditions being positively skewed. Two analyses were subsequently performed across the stakes versions of each scenario; given normality and homogeneity of variance violations in the data, a Generalised Estimating Equation (GEE) (non-parametric) and a second GEE analysis following the removal of extreme outliers.

2.4 Summary of Descriptive Statistics (prior to outlier removal and following removal)

Across all scenarios, there was a general pattern of participants stating that more evidence would be required by the protagonist in order to know something (positive polarity)/ still not know something (negative polarity) as the stakes increased. This pattern of responses remained with extreme outliers removed. There was a general pattern of participants stating that more evidence would be required by the protagonist in order to know something (positive polarity)/still not know something (negative polarity) as the stakes increased.

In terms of replicating the stakes effects observed in the first evidence-seeking experiment, we found a main effect of stakes across all scenarios excluding the introduction scenario (see Figure 10). As in the first evidence-seeking experiment, the main effect of stakes in the possessions scenario was observed in the positive polarity only and unique to this experiment, so was the main effect of stakes in the game show scenario. Across four of the six scenarios, we did not observe significant differences in responses to positive and negative polarity prompts. However, in this experiment, we did observe an effect of polarity in the vaccine and possessions scenarios; evidence scores were higher for the positive polarity in the vaccine case but lower for the positive polarity in the possessions case. A full breakdown of analysis by scenario follows this summary. As previously, an initial GEE analysis was performed using a gamma (log link) model with stakes (low; one; two; high) as within-subjects factor and polarity (positive; negative) as between-subjects factor and then repeated once extreme outliers had been identified and removed.

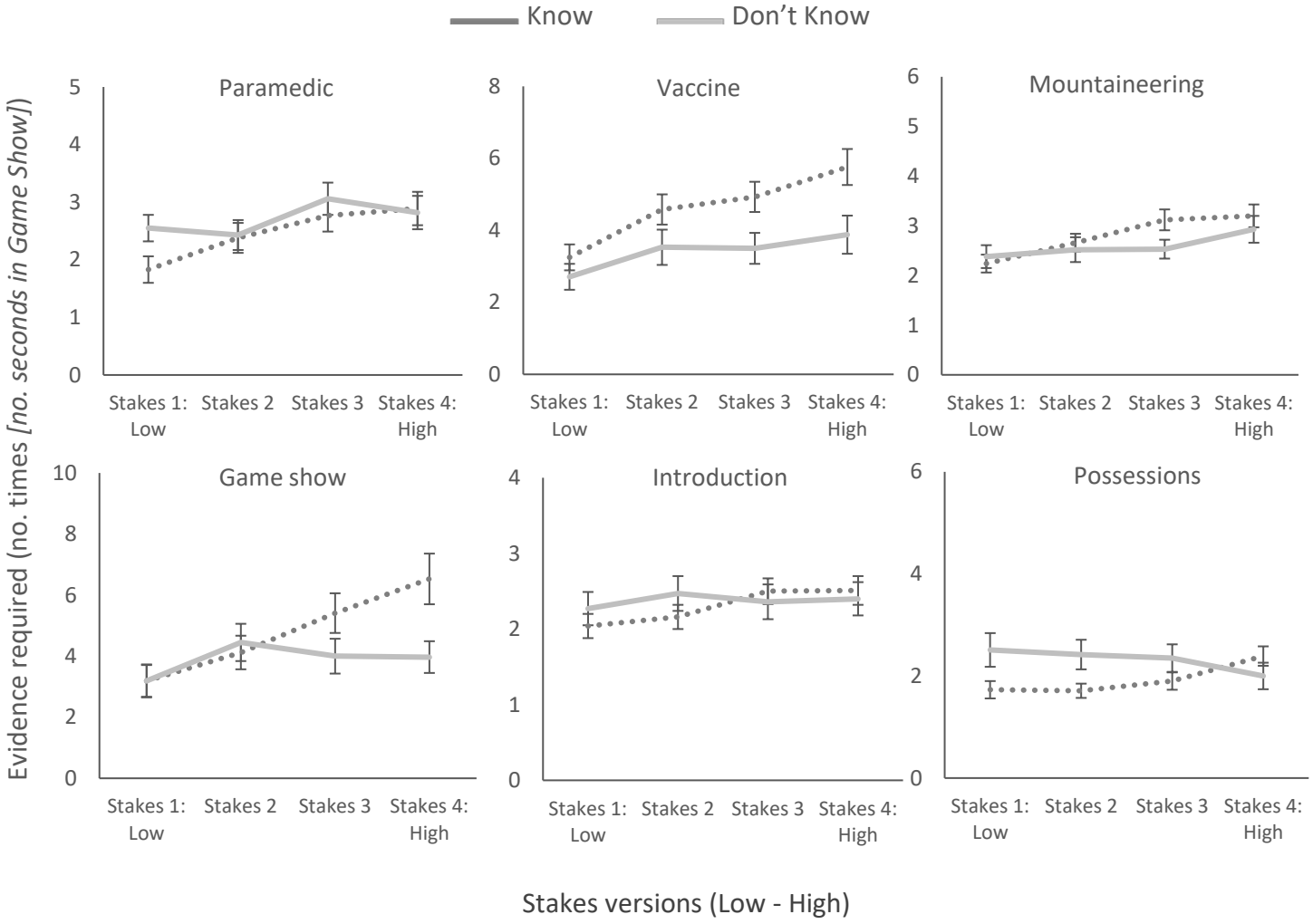


Figure 10. The amount of evidence participants state is needed in order for the protagonist to know/still not know across all stakes versions of all six scenarios. Amount of evidence required increased as the stakes were raised across all scenarios (note positive gradient of lines) except the introduction scenario. In both the possessions and game show scenario, the main effect was observed in the positive polarity only. Error bars represent +/- 1 SE.

2.5 Individual Scenario Analyses

2.5.1 Paramedic Scenario

Iinitial GEE ($N = 316$, zero values ignored) revealed a main effect of stakes, (Wald $X^2[3] = 11.25$, $p = .010$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 16.10$, $p = .001$). There was no main effect of polarity, ($p = .312$). Follow-up tests using sequential Bonferroni a significant difference between evidence scores between the stakes 1 [*low*] paramedic scenario and the stakes 2 scenarios ($p = .034$), the stakes 3 scenario ($p < .001$), and the stakes 4 [*high*] scenario ($p < .001$). There was also a significant difference in levels of evidence between the stakes 2 scenario and the stakes 4 [*high*] scenario ($p = .012$). These effects were present for the positive polarity only.

Having extracted extreme outliers ($N = 316$ to $N = 285$, zero values ignored), another GEE analysis confirmed a main effect of stakes, (Wald $X^2[3] = 27.70$, $p < .001$). There was no significant interaction of polarity x stakes, ($p = .076$) and no main effect of polarity, ($p = .269$). Follow-up tests using sequential Bonferroni revealed a significant difference in evidence scores between the stakes 1 [*low*] paramedic scenario and the stakes two ($p < .001$) and stakes 4 [*high*]scenario ($p = .001$). There was also a significant difference between the stakes 2 scenario and stakes 3 scenario ($p = .005$).

2.5.2 Vaccine Scenario

Initial GEE analysis ($N = 318$, zero values ignored revealed a main effect of stakes, (Wald $X^2[3] = 21.54$, $p < .001$). There was no main effect of polarity, ($p = .604$) and no interaction of polarity x stakes, ($p = .137$). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] vaccine scenario and the stakes 2 scenario ($p = .020$), the stakes 3 scenario ($p = .031$), and the stakes 4 [*high*] scenario ($p = .033$). These effects were across polarity.

Having extracted extreme outliers ($N = 318$ to $N = 294$), another GEE analysis confirmed a main effect of stakes, (Wald $X^2[3] = 28.19$, $p < .001$) and a main effect of polarity, (Wald $X^2[1] = 5.33$, $p = .021$) with higher evidence scores in the positive polarity. There was no interaction of polarity x stakes, ($p = .624$). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*]vaccine scenario and the stakes 2 scenario ($p = .002$), the stakes 3 scenario ($p < .001$), and the stakes 4 [*high*] scenario ($p < .001$). These effects were across polarity.

2.5.3 Mountaineering Scenario

An initial GEE analysis revealed a main effect of stakes, (Wald $X^2[3] = 23.91$, $p < .001$). There was no main effect of polarity, ($p = .625$) and no interaction of polarity x stakes, ($p = .159$). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] mountaineering scenario and the stakes 4 [*high*] scenario ($p < .001$), between the stakes 2 scenario and the stakes 4 [*high*] scenario ($p < .001$), and between the stakes 3 scenario and the stakes 4 [*high*] scenario ($p = .002$). These effects were present across polarity.

Having extracted extreme outliers ($N = 305$ to $N = 284$, zero values ignored), another GEE analysis confirmed a main effect of stakes, (Wald $X^2[3] = 16.88$, $p = .001$). There was no main effect of polarity, ($p = .395$) and no interaction of polarity x stakes, ($p = .225$). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] mountaineering scenario and the stakes 3 scenario ($p = .013$) and the stakes 4 [*high*] scenario ($p < .001$). There was also a significant difference between the stakes 2 scenario and the stakes 4 [*high*] scenario ($p = .025$).

2.5.4 Game Show Scenario

Initial analysis ($N = 330$, zero values ignored) revealed a main effect of stakes, (Wald $X^2[3] = 38.37$, $p < .001$) and a significant interaction of polarity x stakes, (Wald $X^2[3] = 9.97$, $p = .019$). There was no main effect of polarity, ($p = .084$). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] game show scenario and the stakes 3 scenario ($p = .002$) and the stakes 4 [*high*] scenario (p

<.001). There was also a significant difference between the stakes 2 scenario and the stakes 3 scenario (p =.001) and stakes 4 [*high*] scenario (p <.001). There was also a significant difference between the stakes 3 scenario and the stakes 4 [*high*] scenario (p =.001). These effects were present in the positive polarity only.

Having extracted extreme outliers (N = 330 to N = 275, zero values ignored), further analysis with gamma (log link) confirmed a main effect of stakes, (Wald X^2[3] = 13.58, p =.004) and a significant interaction of polarity x stakes, (Wald X^2[3] = 8.58, p =.035). There was no main effect of polarity (p =.205). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] game show scenario and the stakes 3 scenario (p =.025) and the stakes 4 [*high*] scenario (p =.001). There was also a significant difference between the stakes 2 scenario and the stakes 4 [*high*] scenario (p =.049). These effects were present in the positive polarity only.

2.5.5 Introduction Scenario

Initial analysis (N = 331, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 9.23, p =.026). There was no main effect of polarity, (p =.295) and no interaction of polarity x stakes, (p = .326). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] introduction scenario and the stakes 4 [*high*] scenario (p =.041). This effect was present across polarity

Having extracted extreme outliers (N = 331 to N = 312), further analysis found no main effect of stakes, (p =.155), no main effect of polarity, (p =.709) and no interaction of polarity x stakes, (p =.266).

2.5.6 Possessions Scenario

Initial GEE analysis (N = 297, zero values ignored) revealed a main effect of stakes, (Wald X^2[3] = 9.03, p =.029) and a main effect of polarity, (Wald X^2[1] = 4.14, p =.042) with higher levels of evidence in the negative polarity. There was no interaction of polarity x stakes, (p = .248). Comparisons using sequential Bonferroni found no differences between stakes scenarios in either polarity. These non-significant follow-up tests are likely due to large variances in the dataset.

Having extracted extreme outliers (N = 297 to 285), further analysis found no main effect of stakes, (p =.876) and no main effect of polarity, (p =.099). There was a significant interaction of polarity x stakes, (Wald X^2[3] = 14.57, p =.002). Follow-up tests using sequential Bonferroni revealed significant differences in evidence scores between the stakes 1 [*low*] possessions scenario and the stakes 4 [*high*] scenario (p =.039). This effect was present for the positive polarity only.

3. Never and Zero Responses

In order to interpret the full range of responses given in the evidence-seeking experiments, separate analyses were performed on both “never” and “0” responses which could be given in response to the following prompts:

Never responses: *If you think S will never know no matter how many times she checks, write "never"* (in the positive polarity condition and in the negative polarity condition)

Zero responses: *If you think S knows without having to check, write "0"* (in the positive polarity condition)

The number of "never" responses given in response to scenarios in the negative polarity were significantly higher across all three experiments when compared to the positive polarity conditions (see Table below).

Experiment	**Know** **("Never" count)**	**Don't Know** **("Never" count)**	**Chi-Square**
Symmetrical Prompts	25	178	$\chi^2(1) = 115.32$, $p < .001$
Matched Prompts	29	185	$\chi^2(1) = 113.72$, $p < .001$

The number of "0" responses given in response to scenarios in the positive polarity were higher in both the original prompts and symmetrical prompts experiments when compared to the negative polarity conditions (see Table below). Again, as in the first evidence-seeking experiment, this isn't surprising given that participants weren't explicitly given the option to respond with "0" if the subject *knows without having to check* in the negative polarity prompts. The meaning of a "0" response when given in response to a positive vs. a negative polarity prompt is therefore probably different. Note that zero responses are not recorded for the *matched prompts follow-up* experiment as the response option was removed to create the *matched* design (i.e. *If you think S knows without having to check, write "0"*).

Experiment	**Know** **("0" count)**	**Don't Know** **("0" count)**	**Chi-Square**
Symmetrical Prompts	84	48	$\chi^2(1) = 9.82$, $p = .002$

In order to investigate stakes effects on these responses, a GEE analysis with poisson (loglinear) model was performed on the frequency of "never" responses in the follow-up experiment using symmetrical prompts. Analysis revealed a main effect of polarity, (Wald $X^2[1] = 20.82$, $p < .001$), a main effect of stakes, (Wald $X^2[3] = 14.65$, $p = .002$), and a significant interaction of polarity x stakes, (Wald $X^2[3] = 7.88$, $p = .049$).

When interpreting the interaction, comparisons using sequential Bonferroni did not find any significant differences. Given the findings of these follow-up tests, we are unable to interpret this interaction. When interpreting the main effect of stakes, comparisons using sequential Bonferroni indicated a significant difference in the frequency of "never" responses between the stakes 1 [*low*] scenarios and the stakes 3 scenarios ($p = .016$). This effect was present across polarities (see Figure 11).

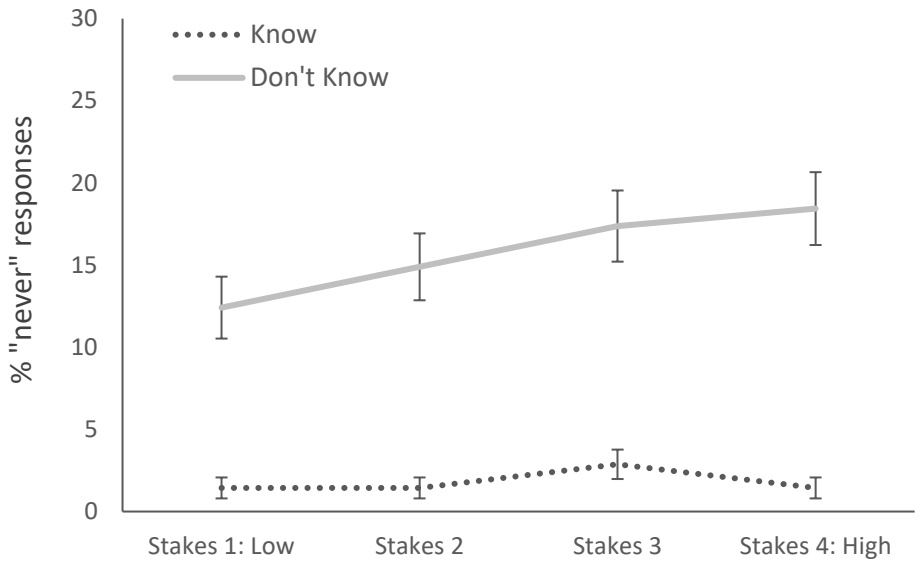


Figure 11. The percentage of "never" responses given in each stakes scale (across all scenarios) in each polarity condition of the Symmetrical Experiment. The frequency of "never" answers given in response to prompts increased as the stakes were raised from low stakes to stakes 3. *Note*: this figure is *figure 7* in the manuscript.

In order to investigate stakes effects on zero responses in the Symmetrical Experiment, a GEE analysis with poisson (loglinear) model was also performed on the frequency of zero responses. Analysis revealed a main effect of stakes, (Wald $X^2[3] = 8.73$, $p = .033$). There was no main effect of polarity, ($p = .599$) and no interaction of polarity x stakes, ($p = .330$).

When interpreting the main effect of stakes, comparisons using sequential Bonferroni indicated a significant difference between the zero counts in the stakes 1 [*low*] scenarios and the stakes 3 scenarios ($p = .045$). The difference between the stakes 2 scenarios and the stakes 3 scenarios fell short of significance ($p = .050$). This effect was present across polarities (see Figure 12).

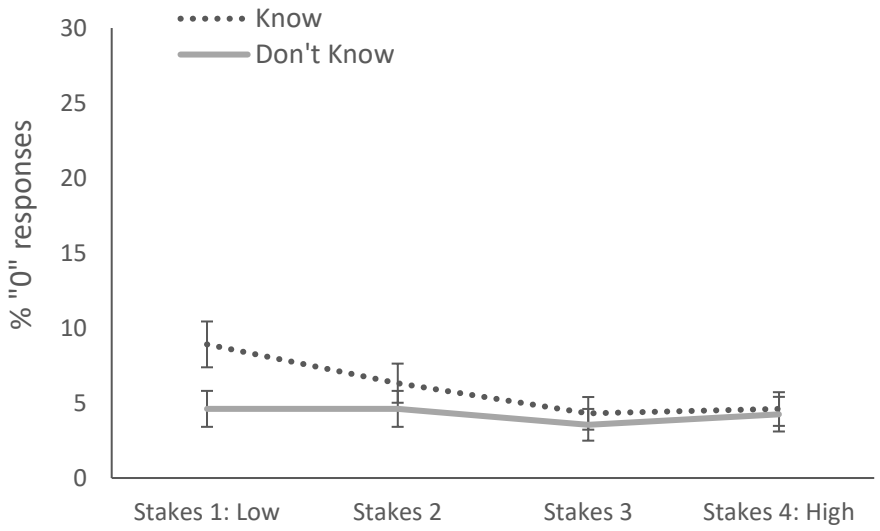


Figure 12. The percentage of zero responses given in each stakes scale (across all scenarios) in each polarity condition of the Symmetrical Experiment. The frequency of zero answers given in response to prompts decreased as the stakes were raised from low stakes to stakes 3.

In the Matched Experiment, A GEE analysis with poisson (loglinear) model was performed on the frequency of "never" responses, revealing a main effect of polarity, (Wald $X^2[1] = 18.66$, $p < .001$), but no main effect of stakes, ($p = .477$), and no interaction of polarity x stakes, ($p = .202$) (see Figure 13).[3]

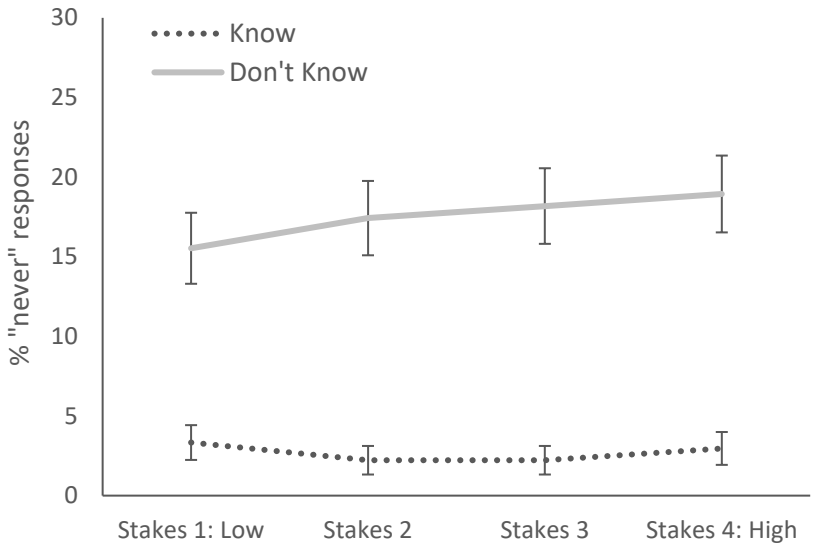


Figure 13. The percentage of "never" responses given in each stakes scale (across all scenarios) in each polarity condition of the Symmetrical Experiment. Here we find no main effect of stakes on "never" responses to the negative prompts.

[3] Note that no GEE analysis was performed for zero responses in this experiment as the response option was removed to create the *matched* design (i.e. *If you think S knows without having to check, write "0"*).

Supplementary Material

Experimental Materials

Paramedic: *lives*

ONE (LOW)
Megan, a paramedic, has just been called to an accident involving a collision between two cars. Megan is driving an ambulance carrying her team of paramedics to the scene of the accident. Megan is familiar with the surrounding area, she has GPS on her phone that she can check if necessary, and she is traveling on the right route to get to the accident. Over the radio, Megan is told that there is one person at the scene of the accident with a broken arm, which is not life-threatening. If Megan makes a wrong turn on the way to the accident, the injured person will be inconvenienced, but nothing terrible will happen.

TWO
Heather, a paramedic, has just been called to an accident involving a collision between two cars. Heather is driving an ambulance carrying her team of paramedics to the scene of the accident. Heather is familiar with the surrounding area, she has GPS on her phone that she can check if necessary, and she is traveling on the right route to get to the accident. Over the radio, Heather is told that there is one person at the scene of the accident who is bleeding badly. If Heather makes a wrong turn on the way to the accident, she will be too late and the injured person will bleed to death.

THREE
Carolyn, a paramedic, has just been called to an accident involving a collision between a bus and a truck. Carolyn is driving an ambulance carrying her team of paramedics to the scene of the accident. Carolyn is familiar with the surrounding area, she has GPS on her phone that she can check if necessary, and she is traveling on the right route to get to the accident. Over the radio, Carolyn is told that the bus flipped over and there are 15 critically injured people pinned underneath the bus. If Carolyn makes a wrong turn on the way to the accident, she will be too late and the injured people will die.

FOUR (HIGH)
Ottoline, a paramedic, has just been called to an accident involving a collision between several vehicles, including buses and trucks. Ottoline is driving an ambulance carrying her team of paramedics to the scene of the accident. Ottoline is familiar with the surrounding area, she has GPS on her phone that she can check if necessary, and she is traveling on the right route to get to the accident. Over the radio, Ottoline is told that one of the trucks was carrying flammable fuel and several vehicles, including a school bus carrying 50 children, are on fire. If Ottoline makes a wrong turn on the way to the accident, she will be too late and the children will die.

Prompts:
+ How many times does *S* need to check her GPS before she knows that she will make it to the accident without taking a wrong turn?
- How many times can *S* check her GPS and still not know that she will make it to the accident without taking a wrong turn?

Vaccine: *lives*

ONE (LOW)

Elaine is a medical researcher. Her task is to create a vaccine for a virus. Elaine has done this before, and she has a check list that specifies all of the steps she needs to take to make the vaccine. Elaine is following all of the steps correctly. Elaine's assistant has informed her that there is one human research participant who has volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participant will give them mild cold-like symptoms.

TWO

Alison is a medical researcher. Her task is to create a vaccine for a virus. Alison has done this before, and she has a check list that specifies all of the steps she needs to take to make the vaccine. Alison is following all of the steps correctly. Alison's assistant has informed her that there is one human research participant who has volunteered to trial the vaccine before it is distributed more widely. If Alison does not follow the steps correctly, it will produce a deadly combination that when administered to the research participant will kill him within days.

THREE

Georgina is a medical researcher. Her task is to create a vaccine for a virus. Georgina has done this before, and she has a check list that specifies all of the steps she needs to take to make the vaccine. Georgina is following all of the steps correctly. Georgina's assistant has informed her that there are 15 human research participants who have volunteered to trial the vaccine before it is distributed more widely. If Georgina does not follow the steps correctly, it will produce a deadly combination that when administered to the 15 research participants will kill them all within days.

FOUR (HIGH)

Julie is a medical researcher. Her task is to create a vaccine for a virus. Julie has done this before, and she has a check list that specifies all of the steps she needs to take to make the vaccine. Julie is following all of the steps correctly. Julie's assistant has informed her that there are 100 human research participants who have volunteered to trial the vaccine before it is distributed more widely. If Julie does not follow the steps correctly, it will produce a deadly combination that when administered to the 100 research participants, will kill them all after several days of excruciating pain.

Prompts:

+ How many times does *S* need to consult her check list before she knows that she is making the vaccine correctly?

- How many times can *S* consult her check list and still not know that she is making the vaccine correctly?

Mountaineering: *personal injury*

ONE (LOW)
Josephine is leading a mountain climbing expedition in the Alps with a novice climber. Josephine has tied a rope securely to the other climber, in order to protect him from falls as they move up the mountain together. Visibility is reducing, making the climb increasingly dangerous, because it is becoming harder to see the edge of the mountain trail that they are following. The drop on either side of the trail edge is around 5 feet. If not tied together securely, a slip from this height would result in minor injuries (a minor fracture, for example) to the climber who slips, but if the rope is tied securely, no one will be injured from a slip.

TWO
Kristin is leading a mountain climbing expedition in the Alps with a novice climber. Kristin has tied a rope securely to the other climber, in order to protect him from falls as they move up the mountain together. Visibility is reducing, making the climb increasingly dangerous, because it is becoming harder to see the edge of the mountain trail that they are following. The drop on either side of the trail edge is around 15 feet. If not tied together securely, a slip from this height would result in moderate injuries (a broken arm or leg and a concussion) to the climber who slips, but if the rope is tied securely, no one will be injured from a slip.

THREE
Teresa is leading a mountain climbing expedition in the Alps with a novice climber. Teresa has tied a rope securely to the other climber, in order to protect him from falls as they move up the mountain together. Visibility is reducing, making the climb increasingly dangerous, because it is becoming harder to see the edge of the mountain trail that they are following. The drop on either side of the trail edge is around 50 feet. If not tied together securely, a slip from this height would result in major injuries (a broken spine or a broken neck) to the climber who slips, but if the rope is tied securely, no one will be injured from a slip.

FOUR (HIGH)
Laura is leading a mountain climbing expedition in the Alps with a novice climber. Laura has tied a rope securely to the other climber, in order to protect him from falls as they move up the mountain together. Visibility is reducing, making the climb increasingly dangerous, because it is becoming harder to see the edge of the mountain trail that they are following. The drop on either side of the trail edge is around 1,000 feet. A fall from this height would be fatal to the climber who slips, but if the rope is tied securely, no one will be injured from a slip.

Prompts:
+ How many times does *S* need to inspect the rope before she knows that it is tied securely?

- How many times can *S* inspect the rope and still not know that it is tied securely?

Game show: *finance*

ONE (LOW)
Emma is taking part in a game show that involves answering general knowledge trivia questions. The game show host has asked Emma, "What is the capital of Tanzania?". Emma has recently read a list of the most obscure world capitals and the city "Dodoma" pops into her head. In fact, Emma is right: the capital of Tanzania is Dodoma. As this is the first round of the game show, only $1 is at stake: answering this question correctly will result in Emma winning $1, and answering incorrectly will result in her losing $1.

TWO
Debra is taking part in a game show that involves answering general knowledge trivia questions. The game show host has asked Debra, "What is the capital of Tanzania?". Debra has recently read a list of the most obscure world capitals and the city "Dodoma" pops into her head. In fact, Debra is right: the capital of Tanzania is Dodoma. As this is the second round of the game show, $100 is at stake: answering this question correctly will result in Debra winning $100, and answering incorrectly will result in her losing $100.

THREE
Lisa is taking part in a game show that involves answering general knowledge trivia questions. The game show host has asked Lisa, "What is the capital of Tanzania?". Lisa has recently read a list of the most obscure world capitals and the city "Dodoma" pops into her head. In fact, Lisa is right: the capital of Tanzania is Dodoma. As this is the third round of the game show, $10,000 is at stake: answering this question correctly will result in Lisa winning $10,000 and answering incorrectly will result in her losing $10,000.

FOUR (HIGH)
Tracy is taking part in a game show that involves answering general knowledge trivia questions. The game show host has asked Tracy, "What is the capital of Tanzania?". Tracy has recently read a list of the most obscure world capitals and the city "Dodoma" pops into her head. In fact, Tracy is right: the capital of Tanzania is Dodoma. As this is the final round of the game show, $1,000,000 is at stake: answering this question correctly will result in Tracy winning $1,000,000 and answering incorrectly will result in her losing $1,000,000.

Prompts:
+ How many minutes does *S* need to spend considering her answer before she knows that the capital of Tanzania is Dodoma?

- How many minutes can *S* spend considering her answer and still not know that the capital of Tanzania is Dodoma?

Introduction: *reputation*

ONE (LOW)
Siena teaches at a university and has been asked to introduce a guest speaker to her colleagues over lunch. There are 5 colleagues present at lunch. Siena wrote down the speaker's name—"Dr. Woodbridge"— in her notebook earlier in the day. But if Siena introduces the guest speaker by the wrong name, she will feel slightly embarrassed in front of her colleagues.

TWO
Jane teaches at a university and has been asked to introduce a guest speaker to her colleagues during a seminar. There are 20 colleagues present at the seminar. Jane wrote down the speaker's name—"Dr. Woodbridge"— in her notebook earlier in the day. If Jane introduces the guest speaker by the wrong name, she will feel embarrassed in front of her colleagues and it will reflect badly on her professional capabilities.

THREE
Agnes teaches at a university and has been asked to introduce a guest speaker to her colleagues and members of the public during a public lecture. There are 200 people present at the public lecture. Agnes wrote down the speaker's name—"Dr. Woodbridge"— in her notebook earlier in the day. If Agnes introduces the guest speaker by the wrong name, she will feel very embarrassed in front of the audience and it will reflect very badly on her professional capabilities.

FOUR (HIGH)
Nicole teaches at a university and has been asked to introduce a guest speaker on national television as part of a live interview. The interview will be viewed live by thousands of people. Nicole wrote down the speaker's name—"Dr. Woodbridge"— in her notebook earlier in the day. If Nicole introduces the guest speaker by the wrong name, she will feel very embarrassed in front of a live television audience and it will reflect very badly on her professional capabilities and on her university's reputation.

Prompts:
+ How many minutes does *S* need to check her notebook before she knows that the guest speakers name is "Dr. Woodbridge"?

- How many minutes can *S* check her notebook and still not know that the guest speakers name is "Dr. Woodbridge"?

Arson: *personal value*

ONE (LOW)
Natalie is living in an area where there have been a series of fires set by arsonists recently. Only a functioning sprinkler system can stop a fire set by an arsonist. A week ago, Natalie checked that the sprinklers were working in her storage room, which contains her garbage and recycling. If the sprinklers do not work, everything in the room is at risk from arson. But the sprinklers in the room are fully functioning.

TWO
Winnie is living in an area where there have been a series of fires set by arsonists recently. Only a functioning sprinkler system can stop a fire set by an arsonist. A week ago, Winnie checked that the sprinklers were working in her living room, which contains Winnie's laptop and hard drive containing all her family photos. If the sprinklers do not work, everything in the room, including her laptop and hard drive, is at risk from arson. But the sprinklers in the room are fully functioning.

THREE
Becky is living in an area where there have been a series of fires set by arsonists recently. Only a functioning sprinkler system can stop a fire set by an arsonist. A week ago, Becky checked that the sprinklers were working in her spare bedroom, where the family dog sleeps. If the sprinklers do not work, everything in the room, including the family dog, is at risk from arson. But the sprinklers in the room are fully functioning.

FOUR (HIGH)
Kylie is living in an area where there have been a series of fires set by arsonists recently. Only a functioning sprinkler system can stop a fire set by an arsonist. A week ago, Kylie checked that the sprinklers were working in her nursery room, where her baby sleeps. If the sprinklers do not work, everything in the room, including her baby, is at risk from arson. But the sprinklers in the room are fully functioning.

Prompts:
+ How many times does *S* need to check the sprinklers before she knows that they are working in the *X* room?

- How many times can *S* check the sprinklers and still not know that they are working in the *X* room?

Prompt variations

Symmetrical prompts (for evidence-seeking):

In a second experiment, we use these symmetrical prompts (to remove any presuppositions that might be triggered by "and still not know" in the negative prompts in the first experiment).

Paramedic

+ What is the minimum number of times *S* needs to check her GPS before she knows that she will make it to the accident without taking a wrong turn?

- What is the maximum number of times *S* can check her GPS and not know that she will make it to the accident without taking a wrong turn?

Vaccine

+ What is the minimum number of times *S* needs to consult her check list before she knows that she is making the vaccine correctly?

- What is the maximum number of times *S* can consult her check list and not know that she is making the vaccine correctly?

Mountaineering

+ What is the minimum number of times *S* needs to inspect the rope before she knows that it is tied securely?

- What is the maximum number of times *S* can inspect the rope and not know that it is tied securely?

Game show

+ What is the minimum number of minutes *S* needs to spend considering her answer before she knows that the capital of Tanzania is Dodoma?

- What is the maximum number of minutes *S* can spend considering her answer and not know that the capital of Tanzania is Dodoma?

Introduction

+ What is the minimum number of times *S* needs to check the sprinklers before she knows that they are working in the *X* room?

- What is the maximum number of times *S* can check the sprinklers and not know that they are working in the *X* room?

Arson

+ What is the minimum number of times *S* needs to check the sprinklers before she knows that they are working in the *X* room?

- What is the maximum number of times *S* can check the sprinklers and not know that they are working in the *X* room?

Note: in all evidence-seeking experiments we include the additional instructions – *if you think S* knows without having to check, write "0". If you think *S* can never know no matter how many times she checks, write "never".

Prompts (for evidence-fixed experiments):

For experiment one, we used the traditional approach of asking participants the extent to which they agree or disagree with knowledge claims:

To what extent do you agree or disagree with the following claim:

Paramedic

+Subject x [specific to scenario] knows that she will make it to the accident without taking a wrong turn.

- Subject x [specific to scenario] doesn't know that she will make it to the accident without taking a wrong turn.

Vaccine

+Subject x [specific to scenario] knows that she is making the vaccine correctly.

- Subject x [specific to scenario] doesn't know that she is making the vaccine correctly.

Mountaineering

+Subject x [specific to scenario] knows that the rope is tied securely.

- Subject x [specific to scenario] doesn't know that the rope is tied securely.

Game show

+Subject x [specific to scenario] knows that the capital of Tanzania is Dodoma.

- Subject x [specific to scenario] doesn't know that the capital of Tanzania is Dodoma.

Introduction

+Subject x [specific to scenario] knows that the guest speaker's name is "Dr Woodbridge".

- Subject x [specific to scenario] doesn't know that the guest speaker's name is "Dr Woodbridge".

Arson

+Subject x [specific to scenario] knows that the sprinklers are working in the x room [specific to scenario].

- Subject x [specific to scenario] doesn't know that the sprinklers are working in the x room [specific to scenario].