Knowledge, Stakes, and Mistakes

WESLEY BUCKWALTER
University of Waterloo

JONATHAN SCHAFFER
Rutgers University

Do stakes affect knowledge ascription? According to a prominent claim in recent epistemology, people are less likely to ascribe knowledge to a high stakes subject for whom the practical consequences of error are severe, than to a low stakes subject for whom the practical consequences of error are slight, even when "traditional" epistemic factors like evidence and belief are held fixed. In short:

> *Folk stakes sensitivity*: All else equal, people are less likely to ascribe knowledge to a high stakes subject than to a low stakes subject.

To illustrate, *Folk stakes sensitivity* says that all else equal, people are less likely to say that Sarah knows that her health insurance is current, if she is about to get hit by a bus and incur huge medical bills.

Folk stakes sensitivity holds intrinsic interest to anyone interested in the psychological basis of knowledge ascription. It holds further interest to epistemologists by featuring as the lead premise in an abductive argument to the radical conclusion—defended by Fantl & McGrath (2002), Hawthorne (2004), and Stanley (2005)—that knowledge itself is sensitive to stakes. It holds a different sort of interest to experimental philosophers as an empirical claim made from the armchair.

We offer an opinionated discussion of the "state of the art" on *Folk stakes sensitivity*. We draw on a first wave of empirical studies—due to Feltz & Zarpentine (2010), May *et al.* (2010), and Buckwalter (2010)—which cast doubt on the claim, and a second wave of empirical studies—due to Pinillos (2012) and Sripada & Stanley (2012)—said to vindicate it, as well as new studies of our own. We think that the balance of evidence to date best fits a very different view, namely that of:

> *Folk stakes insensitivity*: All else equal, people are equally likely to ascribe knowledge to a high stakes subject as to a low stakes subject.

In other words, we doubt that stakes affect knowledge ascription, holding all else equal.

Overview: In §1 we review the epistemological background in which *Folk stakes sensitivity* has arisen, and the empirical background in terms of the first wave studies casting doubt on it. This is to set the stage for a critical discussion of the

second wave studies said to offer vindication. In §2 we discuss Pinillos's studies. We present new results showing that Pinillos has found a stakes effect, but not an effect on knowledge ascription (we argue that his effect is on the modal element in his complex probes). In §3 we turn to Sripada & Stanley's studies. We present new results showing that Sripada & Stanley have found an effect on knowledge ascription, but not a direct stakes effect (we argue that their effect is a salience effect). In §4 we conclude overall that the balance of evidence to date best fits *Folk stakes insensitivity*.

1. Background

1.1 Epistemological Background

Folk stakes sensitivity is a behavioral level claim, concerning what people are likely to say. It begins from the claim that stakes affect knowledge ascription, in the sense that people are less likely to ascribe knowledge to a high stakes subject than to a low stakes subject. This initial claim arose in the epistemology literature in discussion of paired cases such as DeRose's (1992: 913) bank cases, as well as Cohen's (1999: 58) airport cases, and Fantl & McGrath's (2002: 67–8) train cases. Minimal pair versions of the bank cases differing solely over stakes run as follows (from Feltz & Zarpentine 2010):

> *Bank high*: Hannah and her sister Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, it is very important that they deposit their paychecks by Saturday. Hannah notes that she was at the bank two weeks before on a Saturday morning, and it was open. Hannah says to Sarah, "I know that the bank will be open tomorrow."

> *Bank low*: Hannah and her sister Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they do not have an impending bill coming due, it is not very important that they deposit their paychecks by Saturday. Hannah notes that she was at the bank two weeks before on a Saturday morning, and it was open. Hannah says to Sarah, "I know that the bank will be open tomorrow."

It was widely agreed in the epistemology literature that "our intuitions"—understood as "what ordinary speakers will count as 'knowledge'" (DeRose 2005: 72)—are that Hannah is right to claim knowledge in cases like *Bank low* but wrong to claim knowledge in cases like *Bank high*, due to a difference in the practical consequences of error for Hannah. So the received view in epistemology has become that people are less likely to ascribe knowledge to a high stakes subject than to a low stakes subject, as per:

> *Stakes effect*: People are less likely to ascribe knowledge to a high stakes subject than to a low stakes subject.[1]

 Stakes effect should be understood as a causal claim, but one that is neutral along several dimensions. It is neutral, for instance, on the shape of the effect. Perhaps the likelihood of knowledge ascription decreases linearly as the stakes rise,

or perhaps there is a threshold, or perhaps the relationship is more complicated still. It is neutral as well on the magnitude of the effect. Perhaps the likelihood of knowledge ascription differs dramatically between high stakes and low stakes cases, or perhaps the difference is smaller. We ignore these nuances. We are interested in whether there is any effect at all, regardless of shape or magnitude.

Folk stakes sensitivity starts from *Stakes effect* and adds the proviso that stakes continue to affect knowledge ascription even when all other epistemic factors are held fixed ("all else equal"). In order to clarify what is meant by this added proviso, two examples may prove helpful. First, it might be that stakes merely affect knowledge ascription by means of affecting belief ascription. For instance, it might be that stakes merely affect whether a given subject's level of credence is sufficiently high to cross a variable threshold for belief. That would count as vindicating *Stakes effect* but not *Folk stakes sensitivity*. Indeed in that case the stakes effect would reveal nothing new about knowledge (given that knowledge requires belief), but only something about belief.[2] Secondly, it might be that stakes merely affect knowledge ascription by means of affecting contextual salience. For instance, it might be that stakes merely affect whether certain alternatives become contextually relevant to a given ascriber. That would again count as vindicating *Stakes effect* but not *Folk stakes sensitivity*. In that case the stakes effect on knowledge ascriptions might support a controversial claim about knowledge (e.g., that knowledge requires the elimination of the contextually relevant alternatives), but nothing specifically tied to stakes. (This example is especially pertinent since we argue that it actually provides the best interpretation of Sripada & Stanley's (2012) studies: §3.)

We suspect that all of the authors who endorse *Stakes effect* also endorse *Folk stakes sensitivity*. For it is only with the added "all else equal" proviso that the prospect arises of stakes playing a distinctive epistemic role. Indeed, the main debate seems to take for granted the truth of *Folk stakes sensitivity*, and merely concern what its "truth" reveals. Thus Fantl & McGrath (2002), Hawthorne (2004), and Stanley (2005) use the behavioral level claim of *Folk stakes sensitivity* as the lead premise in an abductive argument for the following radical semantic and metaphysical claims about the epistemic role of stakes:

> *Pragmatism* (semantic): The truth of a knowledge ascription depends in part on what is at stake for the subject.

> *Interest relativity* (metaphysical): Knowledge depends in part on what is at stake for the subject.

Essentially, the idea is that the best explanation for why ordinary speakers behave as per *Folk stakes sensitivity* is that ordinary speakers are responding to truth value differences (as per *Pragmatism*), and the best explanation for why there are truth value differences is that there is a difference as to whether the knowledge relation obtains (as per *Interest relativity*).[3] If the effect of stakes did not persist when all other epistemic factors are held fixed, then there would be no call for recognizing a distinctive epistemic role for stakes, or anything beyond these other epistemic factors.

Likewise the main response to this abductive argument in the literature—found in various forms in Rysiew (2001), Brown (2005), Nagel (2008), Bach (2010), Gerken (*forthcoming*), and Turri (*forthcoming*), *inter alia*—has been to grant *Folk stakes sensitivity* but deny the inference to *Pragmatism* or *Interest relativity*, on grounds that *Folk stakes sensitivity* equally follows from a standard Moorean "intellectualist" semantics operating in combination with independently plausible assumptions about conversational pragmatics and/or performance errors. These authors are not merely granting *Stakes effect*, for they also seem to be saying that the stakes effect is arising for non-epistemic reasons (conversational pressures, conceptual errors), and not through a stakes effect on some other epistemic factors.[4]

Leaving open the question of exactly who endorses what, it is important to separate *Stakes effect* from the stronger claim of *Folk stakes sensitivity*, since only the latter yields the prospect of a distinctive epistemic role for stakes. Our focus is thus on whether *Folk stakes sensitivity* is true. We remain neutral on what if anything it would reveal if true (for instance, whether *Pragmatism* or *Interest relativity* would be supported). Of course if *Folk stakes insensitivity* is true instead (as we argue), then the issue is moot.

Likewise we remain neutral on the status of other arguments that might involve *Folk stakes sensitivity*. For instance, there is the prospect of a reversed argument that takes the falsity of *Folk stakes sensitivity* as its lead premise, and concludes with the falsity of *Pragmatism*, on grounds that if *Pragmatism* were true then *Folk stakes sensitivity* would be true. Indeed, if *Folk stakes sensitivity* also follows from a standard Moorean intellectualist semantics (plus some independently plausible assumptions), as some respondents to the abductive argument mentioned above seem to hold, then this reversed argument could in principle equally falsify Mooreanism. We remain neutral on which if any views would be falsified if *Folk stakes sensitivity* is false.

We are likewise not discussing other arguments concerning *Pragmatism* or *Interest relativity*. For instance, there is a second argument for *Interest relativity*—found in Fantl & McGrath (2002, 2009), Hawthorne (2004), and Hawthorne & Stanley (2008)—from the theoretical role of knowledge in practical reasoning, which is not premised on *Folk stakes sensitivity*.[5] This line of argument falls beyond the scope of our discussion (although a small connection surfaces in §2.2).

Finally, we are not focusing on analogous issues concerning the epistemic role of mistakes salient to the ascriber. For there is an analogous claim that salience affects knowledge ascription.[6] This is the claim that people are less likely to ascribe knowledge in a high salience context in which uneliminated possibilities of a mistake are relevant, than in a low salience context in which uneliminated possibilities of a mistake are irrelevant (even when "traditional" epistemic factors like evidence and belief are held fixed). This is the behavioral level claim of:

> *Folk salience sensitivity*: All else equal, people are less likely to ascribe knowledge in a high salience context than in a low salience context.

And there are analogous abductive arguments from *Folk salience sensitivity* to radical semantic and metaphysical claims, such as:

> *Contextualism* (semantic): The truth of a knowledge ascription depends in part on which mistakes are salient for the ascriber.
>
> *Contrastivism* (metaphysical): Knowledge depends in part on the value of a contrast argument for the mistakes in play.

Again the idea is that the best explanation for why ordinary speakers behave as per *Folk salience sensitivity* is because ordinary speakers are responding to truth value differences (as per *Contextualism*), and the best explanation for why there are truth value differences is that there is a difference as to whether the knowledge relation obtains (as per *Contrastivism*).[7] Issues concerning *Contextualism* and *Contrastivism* fall beyond the scope of our discussion, but the status of *Folk salience sensitivity* will surface in our discussion of Sripada & Stanley's studies (§3), since we argue that their effect is really a salience effect and not a stakes effect, and so confirms *Stakes effect* but not *Folk stakes sensitivity*. In other words, we think that their studies are best understood as (re-)confirming *Folk salience sensitivity*.

1.2 Empirical Background

Given the prominent role that *Folk stakes sensitivity* has played in recent epistemology, and its status as an empirical claim made from the armchair, it should be unsurprising that experimental philosophers have cast their eyes in its direction. And given the track record of such armchair claims, it may also be unsurprising that the empirical results have not matched what the armchair-bound had imagined. Indeed there was a first wave of empirical studies due to Feltz & Zarpentine (2010), May *et al.* (2010), and Buckwalter (2010). These studies, conducted largely contemporaneously and independently, converged in finding little to no support for *Folk stakes sensitivity*. Indeed, these studies found little to no support even for the weaker claim of *Stakes effect*.[8]

Feltz & Zarpentine began with a range of bank cases drawn verbatim from Stanley (2005). For Stanley these cases were supposed to be clear examples of how a stakes manipulation can flip people's intuitions from "knowledge" (in low stakes) to "ignorance" (in high stakes), and this intuition flip was supposed to be the primary data that his theory explains. But Feltz & Zarpentine found no statistically significant difference in responses between Stanley's low stakes and high stakes cases. Participants tended to cluster around the midpoint ("neutral") on both cases, and no effect of stakes was detected. This looks to be a far cry from the strong flip from "knowledge" to "ignorance" which DeRose, Stanley, and most other epistemologists had predicted from the armchair.

Feltz & Zarpentine went on to test a range of other cases, including cases involving trucks crossing rickety bridges (in the low stakes case the bridge is over a three foot ditch, while in the high stakes case the bridge is over "a yawning thousand foot drop"). In these cases they had the protagonist claim knowledge, and participants were asked whether they thought that what this character said was true (for instance, in their rickety bridge case, participants were asked to evaluate the following sentence: "When John thinks to himself, 'I know that my truck will make it across the bridge,' what he thinks is true.") No statistically significant

difference was found in the evaluation of these knowledge ascriptions between low and high stakes cases. On the basis of not finding any stakes effect on knowledge ascription in any of these experiments, Feltz & Zarpentine (2010: 694) concluded: "[T]he practical facts in these situations do not qualitatively change knowledge attributions and they are not likely to be a fundamental or important feature of our ordinary knowledge attributions."

May *et al.* (2010) focused on simplified versions of DeRose's bank cases. They also had cases in which a protagonist claimed knowledge, but then asked participants to evaluate a third person knowledge ascription for this character (participants were asked to rate "Hannah knows that the bank will be open on Saturday"). Their stakes manipulations produced a very small effect, which marked a slight increase in willingness to ascribe knowledge. That is, their participants in *both* the low stakes and high stakes tended to ascribe knowledge, with just a slight weakening of that tendency with high stakes. They (2010: 273) thus drew the following conclusion on stakes:

> [I]t seems that a person's practical interests affect our confidence in attributing knowledge, but this difference does not determine whether we do credit her with knowledge, at least in these bank cases. Thus, some doubt is cast on the cogency of attempts to support such a version of [*Interest relativity*] by appealing to its capturing common-sense or ordinary linguistic practice with respect to these bank cases.

Buckwalter (2010) likewise focused on versions of DeRose's bank cases. Buckwalter initially ran vignettes without a protagonist claiming knowledge, in which participants were asked to make a knowledge ascription themselves (participants were asked "Does Bruno know that the bank will be open on Saturday?" and given the options to agree or disagree). Buckwalter found no statistically significant effect of stakes. Participants tended to ascribe knowledge in both cases (there was a rate of agreement of 83.3% in the baseline bank case and 81.7% in the high stakes variant). DeRose (2011) critiqued studies that did not include a protagonist self-ascribing knowledge as failing to speak to "standard contextualism." So Buckwalter re-did his studies using a design similar to that used by Feltz & Zarpentine, in that he had vignettes in which a protagonist claimed knowledge, and participants were asked whether they thought that what this character said was true ("On a scale of 1 to 5, circle how much you agree or disagree that Bruno's assertion, 'I know the bank will be open on Saturday' is true.") Buckwalter continued to detect no statistically significant difference due to stakes, and he (2010: 403) naturally wound up doubting the existence of any stakes effect on knowledge ascription.[9]

Putting this together, three independent studies using diverse stimuli converged in finding little to no stakes effect on knowledge ascription.[10] This is a far cry from the clear flip from "knowledge" to "ignorance" predicted from the armchair. On reviewing these findings, Schaffer & Knobe (2012: 675) worried that the current debate may be "founded on a myth." Indeed, these findings are more or less what one would expect, if *Folk stakes insensitivity* were true instead.
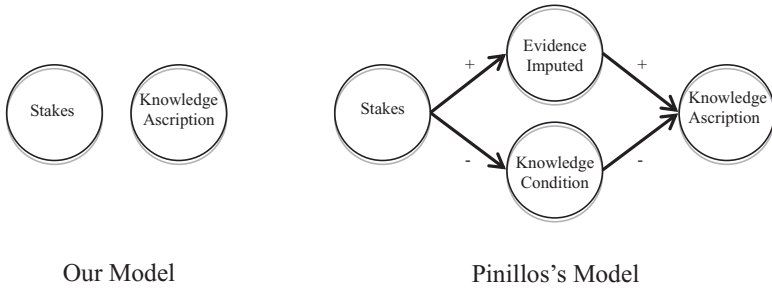
Figure 1. Rival explanatory models for first wave empirical data discussed in §1.2.

2. Second Wave Results: Pinillos

So far all the evidence seems to converge on the conclusion that stakes do not affect knowledge ascription: *Stakes effect* and *Folk stakes sensitivity* both look to be false, and *Folks stakes insensitivity* looks to hold instead. The stage is set to discuss a second wave of empirical results—due to Pinillos (2012) and Sripada & Stanley (2012)—said to reverse this verdict. We begin with Pinillos.

2.1 Pinillos's Evidence Seeking Experiments
Pinillos (2011) re-opened the discussion of *Folk stakes sensitivity*, starting from a criticism of the first wave studies. His criticism is that these studies did not ensure that participants in the low stakes and high stakes conditions were imputing a constant level of evidence to the subject of the knowledge ascription. After all, these studies used vignettes in which the subject (the protagonist) self-ascribes knowledge. And there is independent psychological evidence that people expect high stakes subjects to gather more evidence before self-ascribing knowledge (Nagel 2008). So Pinillos suggested that participants in the high stakes condition might have been cued to impute more evidence to the subject than did participants in the low stakes condition: perhaps participants in the high stakes condition read the self-ascription of knowledge as indirectly indicating sufficient evidence given the stakes, and thereby imputed some additional background evidence to the subject in the high stakes situation. In this way Pinillos pointed out that the first wave results are compatible with the truth of *Folk stakes sensitivity* plus offsetting differences in the level of evidence imputed.

We do not find Pinillos's criticism telling for two main reasons. The minor reason is that Pinillos's criticism is based on an antecedently improbable interpretation of the first wave data. While we interpret the general absence of any statistically significant difference in terms of the absence of an effect, Pinillos instead is suggesting the possibility of two component effects (one on what evidence is imputed, and one on the conditions required for knowledge), which moreover perfectly cancel out to leave a net effect of zero. These two models are depicted in figure 1.

Pinillos's suggested model is of course not impossible, but—pending positive evidence otherwise—it must be viewed as improbable.[11]

The major reason that we do not find Pinillos's criticism telling is that there is in fact empirical evidence against his model, since the first wave results have been reproduced using vignettes *without* a self-ascription of knowledge (which is what Pinillos thought was cueing differences in the level of evidence imputed). Indeed Buckwalter's initial studies were done in just this way, without a self-ascription of knowledge (§1.2). And—as we report in §2.3—we tested still another case without a protagonist self-ascription of knowledge, and once again found no effect of stakes.[12] So overall we think that the first wave results withstand Pinillos's criticism. (This matters when it comes to assessing the overall balance of evidence.)

In any case, Pinillos (2012) used his criticism of the first wave results as a springboard for a novel experimental design involving *evidence seeking*. In Pinillos's design, a protagonist is described in a given practical situation, and participants are asked how much evidence this protagonist would need in order to know. This design gets rid of any self-ascription of knowledge, thereby removing what Pinillos thought was an implicit cue for greater evidence in the high stakes case. Moreover, instead of trying to control for a constant level of evidence between the low stakes and high stakes vignettes, this design does not impute any evidence at all, but instead invites participants to say how much evidence they think will be needed for knowledge. To illustrate, Pinilllos used paired vignettes such as:

> *Typo low*: Peter, a good college student, has just finished writing a two-page paper for an English class. The paper is due tomorrow. Even though Peter is a pretty good speller, he has a dictionary with him that he can use to check and make sure there are no typos. But very little is at stake. The teacher is just asking for a rough draft and it won't matter if there are a few typos. Nonetheless Peter would like to have no typos at all.

> *Typo high*: Peter, a good college student, has just finished writing a two-page paper for an English class. The paper is due tomorrow. Even though Peter is a pretty good speller, he has a dictionary with him that he can use to check and make sure there are no typos. There is a lot at stake. The teacher is a stickler and guarantees that no one will get an A for the paper if it has a typo. He demands perfection. Peter, however, finds himself in an unusual circumstance. He needs an A for this paper to get an A in the class. And he needs an A in the class to keep his scholarship. Without the scholarship, he can't stay in school. Leaving college would be devastating for Peter and his family who have sacrificed a lot to help Peter through school. So it turns out that it is extremely important for Peter that there are no typos in this paper. And he is well aware of this.[13]

Participants, after reading one of these vignettes, were then asked:

> *Typo probe know*: How many times do you think Peter has to proofread his paper before he knows that there are no typos? ____ times.

With this new experimental design, Pinillos found a stakes effect, and indeed a strong and robust one. For instance, participants in the *Typo low* condition gave a median answer of '2', while participants in the *Typo high* condition gave a median answer of '5'. The overall effect in these two conditions is perhaps best displayed in figure 2.
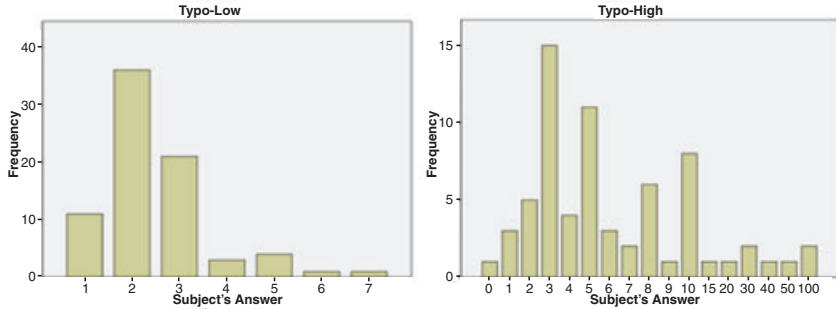
Figure 2. Histograms from Pinillos (2012) showing the distribution of participant responses in *Typo low* and *Typo high* vignettes.

Pinillos was able to reproduce these sorts of results using a range of different vignettes, and also in a within-subjects design. On this basis he (2012: 200) concluded: "[F]olk attributions of knowledge are indeed sensitive to stakes..."

One might worry that Pinillos's *Typo high* is significantly longer and more detailed than his *Typo low*. It is of course preferable to use more "minimal pairs" which minimize differences (including differences in cognitive load) between cases. That said, in the experiments we report below, we were able to minimize the differences somewhat and still reproduce Pinillos's results, by replacing his *Typo high* with:

> *Typo high short*: Peter, a good college student, has just finished writing a two-page paper for an English class. The paper is due tomorrow. Even though Peter is a pretty good speller, he has a dictionary with him that he can use to check and make sure there are no typos. There is a lot at stake. The teacher is a stickler and guarantees that no one will get an A for the paper if there is a typo. Peter needs an A on the paper to get an A for the class, and he needs an A for the class to keep his scholarship. If he loses the scholarship he will have to leave school, which would be devastating for him. So it is extremely important for Peter that there are no typos in the paper.

So we can not only confirm that Pinillos's effect exists, we can even add that its existence is not a mere artifact of the added length and detail of his high stakes variants. So far, so good.[14]

Though we ultimately disagree with Pinillos's interpretation of his effect, we applaud him for having re-opened the discussion in this way. His experimental design is original, his results are powerful (reflecting both the large number of participants he recruited, and the clear trend in the responses he received) and easily replicable, and they are integrated with the use of psychological techniques not often seen in experimental philosophy (for instance Pinillos used the Cognitive Reflection Test to address concerns as to how carefully his participants were thinking).

2.2 Pinillos's Effect is not An Effect on Knowledge Ascription

We are impressed with Pinillos's studies and believe that he has found a real effect of stakes, but doubt that his stakes effect has anything to do with *knowledge ascription*. Pinillos's studies—unlike any of the other studies under discussion—do not use

simple knowledge ascriptions, but instead use fairly complex probes which include additional elements (e.g. "Peter has to proofread his paper ___ times before he knows that there are no typos"). Our first clue that Pinillos's effect has nothing to do with knowledge ascription actually comes from one of his own studies, on which he found that his effect persists *even in probes without 'know'*.

   For these studies, Pinillos used vignettes about a protagonist, Peter, who has to correctly count the pennies in a medium sized jar. In the low stakes version Peter stands to win a pair of movie tickets, while in the high stakes version Peter stands to win a lot of money which he needs to save his mother's life. Pinillos first ran these vignettes using a probe with 'know':

   Count probe know: Peter should count the pennies in the jar at least ___ times before he comes to know the correct answer.

He continued to find a stakes effect. Nothing surprising yet. But then he re-ran the experiment using a probe without 'know,' which instead concerns action:

   Count probe act: Peter should count the pennies in the jar at least ___ times before turning in his final answer.

He continued to find a stakes effect, even when *Count probe know* was replaced by *Count probe act*. Indeed, he found that the choice between these two probes made no statistically significant difference to participants in the high stakes condition, or to participants in the low stakes condition.

   Pinillos took the persistence of the stakes effect with *Count probe act* as additional confirmation for a stakes sensitive view of knowledge, by drawing on Hawthorne & Stanley's (2008) *reason-knowledge principle* connecting knowledge to reasons for action:

   Reason-knowledge principle: Where one's choice is *p*-dependent, it is appropriate to treat the proposition that *p* as a reason for acting iff you know that *p*.

Pinillos reasoned that, if ordinary people are behaving in line with *Interest relativity* plus *Reason-knowledge principle*, there will be a stakes effect not just on knowledge ascription but on the connected matter of ascription of the actions appropriately supported by reasons. So for Pinillos, the fact that the stakes effect persists with *Count probe act* is a further point in favor of the package of *Interest relativity* plus *Reason-knowledge principle*.[15]

   We think this result is instead the first clue that Pinillos's stakes effect has nothing to do with knowledge ascription, in the sense that what is triggering the effect is something wholly distinct from the occurrence of 'know' (we will offer a speculation as to what that might be in §2.3). With this very concern in mind, Buckwalter (*forthcoming*) returned to *Typo*, replicated Pinillos's original results ($N = 100$) with 'know,' and then was able to reproduce the result using yet another probe which replaced 'know' with 'believe':

   Typo probe believe: How many times do you think Peter has to proofread his paper before he believes that there are no typos? ___ times.
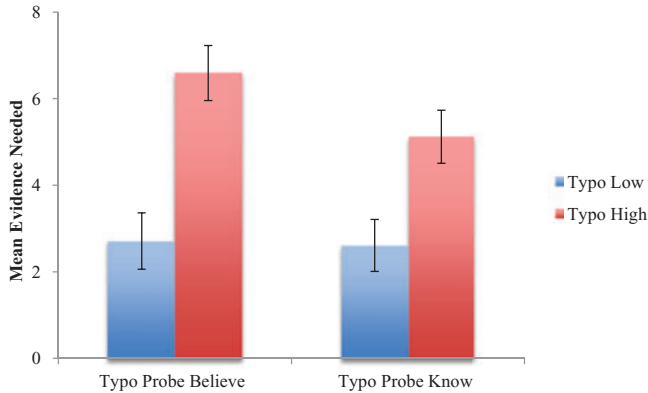
Figure 3. Mean participant responses from Buckwalter (*forthcoming*) in *Typo low* and *Typo high*, grouped by *Typo probe know* and *Typo probe believe*.

Comparing his *Typo probe know* and *Typo probe believe* results, Buckwalter found a significant effect of stakes, but no significant effect of the attitude verb ('know' versus 'believe'), nor any interaction between these two factors.[16] Buckwalter's results may be visualized in figure 3. So again it seems as if the stakes effect has nothing to do with knowledge ascription.

But Pinillos has a plausible reply on offer,[17] since he—following Williamson (2000: 47)—endorses the following *belief-knowledge principle*:

Belief-knowledge principle: One should believe that *p* iff one knows that *p*.

So for Pinillos, Buckwalter's results are to be explained as just what *Interest relativity* plus *Belief-knowledge principle* predicts. The stakes effect on what Peter knows will spill over into a connecting stakes effect on what Peter should believe, given that knowledge is the norm of belief.

Pinillos's reply is not just plausible but fruitful, insofar as it suggests a further empirical test. At this point, there are two main hypotheses open for why his stakes effect persists when the knowledge probe is replaced by either an action probe or a belief probe. On Pinillos's hypothesis, his stakes effect is an effect on knowledge ascription, and spills over to action and belief claims due to knowledge serving as the norm for each of these notions. On our hypothesis, his stakes effect is not an effect on knowledge ascription at all, and persists in the action and belief probes precisely due to its disconnection with knowledge. So it seems that a key test between these hypotheses would involve replacing the knowledge claim with some other claim not plausibly connected to knowledge as action and belief might be thought to be. On Pinillos's hypothesis one should not expect the stakes effect to persist, while on our hypothesis one should.

With this in mind we thought to replace the knowledge claim with a claim about guessing. Guessing is certainly not intimately connected with knowing: there is no plausible "*Guessing-knowledge principle*" that looks anything like the
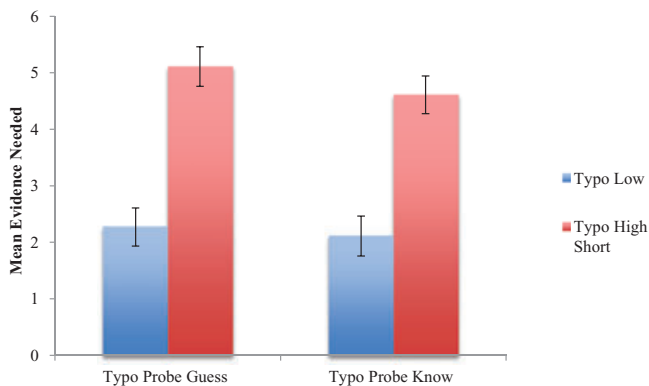
Figure 4. Mean participant responses in *Typo low* and *Typo high short*, grouped by *Typo probe know* and *Typo probe guess*.

Reason-knowledge principle or the *Belief-knowledge principle*. Knowledge is evidently not the norm of guessing. (Indeed it is standardly thought that guessing is *incompatible* with knowing: even if one manages to guess rightly and manages to believes the result, one has still come to a true belief on the basis of a guess, which still falls short of knowledge.)[18] Accordingly we re-ran Pinillos's study ($N = 186$), using his *Typo low* together with our *Typo high short*, but this time using a probe that replaced 'know' with 'guess':

> *Typo probe guess*: How many times do you think Peter has to proofread his paper before he guesses that there are no typos? _____ times.

On Pinillos's hypothesis one should not expect the stakes effect to persist with *Typo probe guess*, while on our hypothesis one should.

In accord with our prediction, we found that Pinillos's stakes effect persisted even with *Typo probe guess*. There was a statistically significant effect of stakes, but no significant effect of the attitude verb ('know' versus 'guess') was detected, nor was any interaction detected between these two factors.[19] These new results may be visualized in figure 4.

But perhaps—as Pinillos (*personal communication*) suggested—there is simply a further independent stakes effect on guessing ascription. It is true that *Pragmatism* and *Interest relativity* are silent on guessing, so do not rule out a stakes effect there as well. So we admit that this is a possible line for the friend of *Folk stakes sensitivity* to take. That said, we do not consider it to be a promising line, for two reasons. First, the posit of an independent stakes effect on guessing ascription is entirely *ad hoc*. Secondly, the posit of two independent stakes effects that in our studies display *no distinguishable difference in magnitude* (in the sense that no statistically significant effect of the attitude verb was detected) is antecedently improbable.

Still, Pinillos's reply suggests yet a further empirical test. One can look at yet another attitude that is clearly not governed by a knowledge norm, and that is also
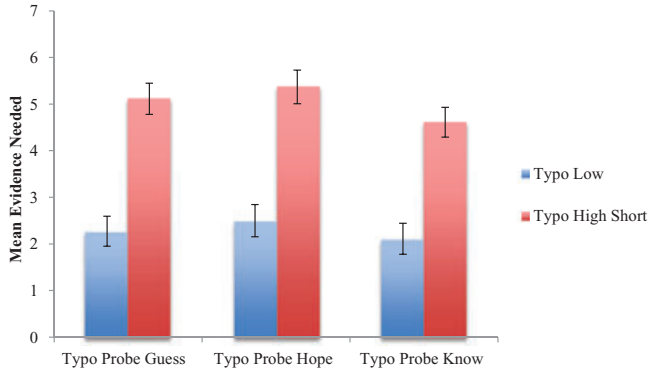
Figure 5. Mean participant responses in *Typo low* and *Typo high short*, grouped by *Typo probe know*, *Typo probe hope*, and *Typo hope guess*.

independent of guessing. So we ran a further study ($N = 80$) using *Typo low* and *Typo high short* again, but now using a probe with 'hope':

Typo probe hope: How many times do you think Peter has to proofread his paper before he hopes that there are no typos? _____ times.

We predicted that we would continue to replicate Pinillos's effect. Since there is no connection between hoping and knowing, such a continuation could not be explained by any "*Hope-knowledge principle.*" And since hoping and guessing are independent (if one hopes that *p*, one may still guess that ~*p*; indeed any combination of hoping and guessing is possible), persistence of the stakes effect would seemingly require Pinillos to posit yet a third independent stakes effect.

As we predicted, we found that Pinillos's stakes effect persisted even with *Typo probe hope*. Stakes were shown to be statistically significant, but no significant effect of the attitude verb was detected, nor was any interaction detected between these factors.[20] These new results may be cumulatively visualized in figure 5.

What we find especially relevant is not just that Pinillos's stakes effect persisted when 'know' was replaced with 'believe,' and still persisted with 'guess,' and continued to persist with 'hope,' but that the nature of the effect seemed entirely unaltered through this process. It would be utterly *ad hoc* to posit yet a third independent stakes effect (one for knowledge, one for guessing, and one for hoping), all three of which moreover display no discernible difference in magnitude in our studies.

We think that these results reveal that, whatever Pinillos's stakes effect concerns, it is not an effect on knowledge ascription (in the sense that what is triggering the effect is entirely unrelated to the occurrence of 'know'). Pinillos's stakes effects persists—with no discernible difference in magnitude—not merely when the knowledge ascription is replaced by arguably connected claims about action and belief, but even when the knowledge ascription is replaced by uncontroversially disconnected claims about other attitudes like guessing and hoping. So we conclude

that Pinillos has found a stakes effect, but not one that has anything to do with knowledge ascription.[21]

2.3 Toward an Alternative Explanation

So if Pinillos has found a stakes effect, but not one that has anything to do with knowledge ascription, then what does his effect concern? Strictly speaking this question is supererogatory for us (our burden is met by showing that his stakes effect has nothing to do with knowledge ascription), but in any case we attempt an answer.

One answer—though not the one we advocate—is that the participants are reading *Typo probe know* and all of its variants in the same shallow way, without even processing the material in the 'before' clause, as if it read:

> *Typo probe shallow*: How many times do you think Peter has to proofread his paper [*blah blah blah*]? _____ times.

That would explain why the attitude verb (if any) embedded in the 'before'-clause makes no difference, on grounds that it is not processed. But—as Pinillos (*personal communication*) noted—this option is empirically implausible. On this option one would expect more cognitively reflective participants to give different responses from less cognitively reflective participants, since more cognitively reflective participants are less likely to indulge in shallow readings like *Typo probe shallow*. But as Pinillos's original results indicated, higher scores on the Cognitive Reflection Test did not affect responses.

A second answer—and our preferred one—is that Pinillos is instead seeing a stakes effect on the modal element 'has' embedded in his complex probes. This 'has' is most naturally read as a *deontic modal* (a "deontic modal" is one interpreted normatively). For instance, *Typo probe know* is most naturally read as asking how many times Peter *should*, given his goals in life and the circumstances he finds himself in, proofread his paper before he knows that there are no typos.[22] It is entirely unsurprising to find a stakes effect on deontic modality. Everyone can agree that the practical consequences of error matter when it comes to considering what one needs to do, normatively speaking. As the stakes get higher, Peter should proofread his paper more times, period. *A fortiori* he should proofread his paper more times *before forming any attitude*. Knowledge plays no role whatsoever in the matter, which is why replacing 'know' with 'believe' or 'guess' or 'hope' makes no difference whatsoever to the data.

One empirical difference between these two answers concerns predictions for what would happen with the following probes:

> *Typo probe breathe*: How many times do you think Peter has to proofread his paper before he takes another breath of oxygen? _____ times.

> *Typo probe fire*: How many times do you think Peter has to proofread his paper before he uses the fire extinguisher to put out the fire that will otherwise quickly engulf him? _____ times.
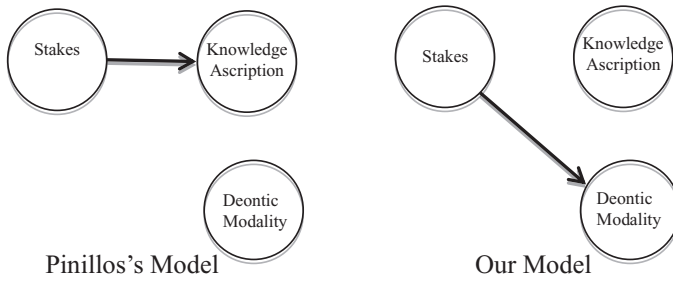
Figure 6. Rival explanatory models for the empirical data discussed in §2.2. The label "Knowledge ascription" refers to the knowledge ascribing aspect of Pinillos's complex probes (triggered by 'know'), and the label "Deontic modality" refers to the deontic modal aspect of these probes (triggered by 'has').

On a shallow processing story (assuming that these probes would not change processing depth) one would predict the effect to persist unchanged, just because the probe is effectively unchanged, always being processed as per *Typo probe shallow*. On our preferred deontic modal story, one would predict the effect to change, because Peter should, given his goals in life and the circumstances he finds himself in, breathe oxygen and extinguish threatening fires even more urgently than he should proofread his paper. We find the prediction of our preferred deontic modal story plausible, but leave the matter to future research.[23]

Comparing Pinillos's explanation in terms of knowledge ascription to our preferred option in terms of deontic modality is comparing two different causal models of our intuitive judgments, which are depicted in figure 6. Pinillos's model is a model of *Folk stakes sensitivity*, while our model is a model of *Folk stakes insensitivity*. (Indeed our model for this data does not even posit anything confirming *Stakes effect*.) We are suggesting that Pinillos's effect is best understood via our model, and hence that it is a stakes effect on deontic modality rather than a stakes effect on knowledge ascription.

We think that our model is preferable to Pinillos's model for two main reasons. The minor reason is that our model is antecedently more probable, since it provides a *conservative* explanation of the data. Pinillos's model posits a stakes effect on knowledge ascription, which seems unlikely given the first wave results (which withstand Pinillos's criticism: §2.1). Our model instead posits a stakes effect on deontic modality, which should be entirely uncontroversial since the consequences matter for what one should do. In other words, since Pinillos's complex probes include both 'know' and a modal element, and since the modal element is clearly liable to a stakes effect, there seems to be no motivation for positing an effect on knowledge ascription to boot.

The major reason we think that our model is preferable to Pinillos's model is that it best fits the full range of empirical evidence. We have already argued (§2.1) that Pinillos's stakes effect persists (completely unaltered) when 'know' is removed but the deontic modal 'has' is retained. To this we now add the converse result, which

is that Pinillos's stakes effect disappears when 'know' is retained but the modal removed. In order to test for this, we ran another experiment ($N = 100$) based on Pinillos's *Typo* cases. Recall that Pinillos finds that participants in *Typo Low* give a median answer of '2' to the question of how many times Peter needs to proofread his paper before he knows that there are no typos, while participants in *Typo High* give a median answer of '5' (§2.1). So we set up paired high stakes and low stakes cases in which Peter has in fact proofread his paper twice. We then told participants to suppose that there were in fact no typos, and asked participants whether or not Peter knows that there are no typos.[24]

Thus we use the following paired vignettes:

Two reads low: Peter is a good college student who has just written a two-page paper for an English class. The paper is due tomorrow. Peter would like his paper to be free of typos, but very little is at stake. The teacher is just asking for a rough draft, and it won't really matter if there are a few typos. Peter is naturally a pretty good speller, plus he has a dictionary with him which he has already used to check the paper carefully, twice over. In fact there are no typos in his paper.

Two reads high: Peter is a good college student who has just written a two-page paper for an English class. The paper is due tomorrow. Peter would like his paper to be free of typos, and there is a lot at stake. The teacher is a stickler and guarantees that no one will get an A for the paper if there is a typo. Peter needs an A on the paper to get an A for the class, and he needs an A for the class to keep his scholarship. Otherwise he will have to leave school, which would be devastating for him. Peter is naturally a pretty good speller, plus he has a dictionary with him which he has already used to check the paper carefully, twice over. In fact there are no typos in his paper.

Participants received one of these two vignettes, and were then asked to answer the following question (on a seven item scale from '1' to '7' anchored with negative to positive agreement terms):

Two reads probe know: To what extent would you agree or disagree with the following claim: "Peter knows that there are no typos in his paper."

We take Pinillos's model—together with his interpretation of his original *Typo* studies—to posit a stronger tendency to agree with the knowledge claim in *Two reads probe know* after *Two reads low* than after *Two reads high*. After all—given Pinillos's interpretation of his original *Typo studies*—in *Two reads low* Peter has proofread the paper just as many times (twice) as he needs before people will tend to say that he knows that there are no typos; while in *Two reads high* Peter still needs to proofread the paper around three more times before people will tend to say that he knows. In contrast, our model posits no difference between these cases.

As our model predicts, no statistically significant difference was found between the *Two reads low* and *Two reads high* conditions.[25] Indeed we find a strong tendency to ascribe knowledge in both cases, as can be seen in figure 7.

One might worry that our *Two reads* studies diverge from Pinillos's *Typo* studies by adding in that Peter's proofreads were done "carefully" and with his dictionary.[26] Perhaps in doing so we raised both *Two reads low* and *Two reads high* well past the
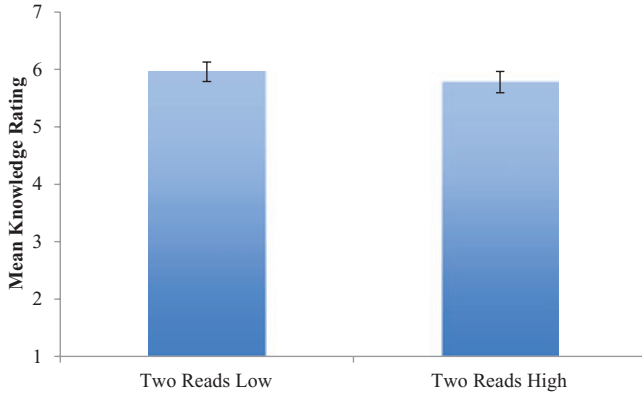
Figure 7. Mean agreement with *Two reads probe know* in *Two reads* cases by stakes. Scales ran 1–7.

threshold for knowledge, and perhaps had we not done so Pinillos's stakes effect might have persisted. So we ran a further test:

Two reads uncareful low: Peter, a good college student, has finished writing a two-page paper for an English class. The paper is due tomorrow. Even though Peter is a pretty good speller, he has a dictionary with him that he can use to check and make sure there are no typos. There is very little at stake. The teacher is just asking for a rough draft, and it won't really matter if there are a few typos. Peter has proofread his paper twice. In fact there are no typos in his paper.

Two reads uncareful high: Peter, a good college student, has finished writing a two-page paper for an English class. The paper is due tomorrow. Even though Peter is a pretty good speller, he has a dictionary with him that he can use to check and make sure there are no typos. There is a lot at stake. The teacher is a stickler and guarantees that no one will get an A for the paper if there is a typo. Peter needs an A on the paper to get an A for the class, and he needs an A for the class to keep his scholarship. If he loses the scholarship he will have to leave school, which would be devastating for him. So it is extremely important for Peter that there are no typos in the paper. Peter has proofread his paper twice. In fact there are no typos in his paper.

Participants ($N = 60$) received one of these two vignettes, and were then asked to answer two questions (on a seven item scale from '1' to '7' anchored with negative to positive agreement terms). For reasons that will emerge in §3.1, the first question we asked was:

Two reads probe evidence: What is the strength of Peter's evidence that there are no typos in his paper?

And the second question was just *Two reads probe know*. Just as we expected, no stakes effect was detected on either the evidence or the knowledge probe, as per figure 8.[27]
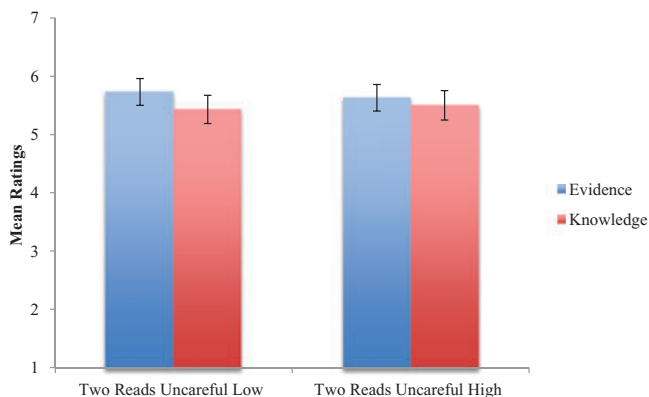
Figure 8. Mean agreement with *Two reads probe evidence* and *Two reads probe know* in *Two reads uncareful* cases by stakes. Scales ran 1–7.

Of course it remains possible that Pinillos's stakes effect did persist in these studies, but we merely failed to detect it. But—pending further evidence otherwise— such an interpretation seems unlikely. Overall, since Pinillos's stakes effect (i) persists when 'know' is removed but the modal retained, and (ii) disappears when 'know' is retained but the modal removed, by standard methods of causal reasoning we conclude that his effect is an effect on the modal element and not on 'know.'[28]

Perhaps a still better explanation of Pinillos's effect can be offered. Our main point remains that Pinillos's effect—whatever its explanation—has nothing to do with knowledge ascription. The balance of evidence to this point still best fits *Folk stakes insensitivity*. Indeed the balance of evidence to this point still does not even support *Stakes effect*.

3. Second Wave Results: Sripada & Stanley

We have just argued that Pinillos has found a stakes effect, but not one that has anything to do with knowledge ascription. It remains to discuss the other component of the second wave of studies said to vindicate *Folk stakes sensitivity*, namely the studies due to Sripada & Stanley (2012).

3.1 Sripada & Stanley's Experiments
Sripada & Stanley (2012) began from a range of criticisms of various first wave studies. In particular, they (2012: §1) expressed the following four worries:

- The probes use an embedded knowledge ascription, which may add a distracting complication.
- The probes do not control for a possible minimal reading of 'know' as only requiring true belief.

- The low stakes vignettes may fail to convey low stakes situations to the participants, since (for Gricean reasons) participants may assume that the description provided of the situation has some importance.
- The high stakes vignettes fail to make adjustments for how participants will expect high stakes subjects to behave (as per Pinillos's worry: §2.1).

The dialectical status of these worries is not ultimately clear to us, since Sripada & Stanley—to their credit—acknowledged that their own evidence does not bear out all of their worries.[29] Still we take all of these worries at face, in part because we think each can be answered.

Indeed, we think that our *Two reads* studies (§2.3) overcome all of Sripada & Stanley's worries directly. Our studies did not use an embedded knowledge ascription, thus avoiding the first worry (Buckwalter's original studies also have this feature, and avoid this worry). Our *Two reads uncareful* studies used an evidence question before the knowledge question, exactly as per Sripada & Stanley's preferred technique for overcoming the second worry. Our *Two reads* studies should quite clearly convey low stakes, and so overcome the third worry. Or at least they should convey low stakes as well as Pinillos's studies do, since they use the same techniques which Pinillos's studies use (Sripada & Stanley themselves embrace Pinillos's techniques as successful). And because our studies do not specify whether or not Peter is done gathering evidence, but instead simply ask whether he has yet achieved knowledge after two careful proofreads, the fourth worry is overcome. No adjustments are needed in our *Two reads* study for what participants will expect of high stakes subjects. Perhaps participants in the high stakes condition expect Peter to go on to do more proofreading, or to ask a friend for additional help, but that should play no role in the question of whether Peter has yet achieved knowledge after two proofreads.

By way of buttressing our results with this second worry in mind, we also re-did our original *Two reads* study using exactly Sripada & Stanley's preferred technique for overcoming the worry, which is to preface the knowledge question with a question about evidence (see details below). Accordingly, in our revised design ($N = 100$), before we asked *Two reads probe know* we first asked *Two reads probe evidence*. We yet again found no statistically significant effect of stakes on knowledge ascription.[30] (Likewise we found no statistically significant effect of stakes on strength of evidence imputed.[31]) Indeed, prefacing the knowledge question with a question about evidence did not seem to have any significant effect on responses to the knowledge question, as can be seen in the combined figure 9.

Of course it remains possible that there was a stakes effect which we merely failed to detect, but pending positive evidence this claim must be regarded as unlikely. So we take our *Two reads* studies to provide empirical confirmation for the claim that the first wave results can withstand every one of Sripada & Stanley's four worries. (This matters when it comes to assessing the overall balance of evidence.)

In any case, Sripada & Stanley took these criticisms of the first wave studies to motivate their own sort of experimental design. Working backwards through the bullet points above, they recommended working with 'ignorant' cases in which
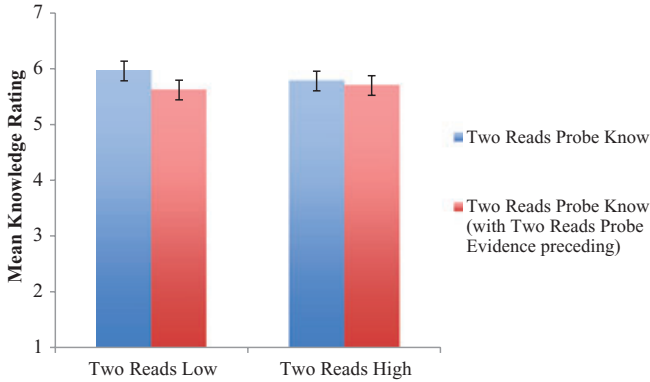
Figure 9. Mean agreement with *Two reads probe know* in *Two reads cases* with and without being preceded by *Two reads Probe Evidence*. Scales ran 1–7.

the subject is unaware of what is at stake, so that participants will not expect the stakes to impact behavior. They recommended minimizing the description of the low stakes situation, so as not to convey any major importance. They recommended first asking about the quality of the subject's evidence, so as to trigger a more demanding reading of 'know.' And they recommended just directly asking participants if the subject at issue knows.

Putting these recommendations together, Sripada & Stanley designed three new studies, obtaining their best results with the following paired vignettes, which thus serve as our focus:

Ignorant low: Hannah has a gene that causes her to experience a slightly dry mouth when she eats Mongolian pine nuts. Hannah has absolutely no idea that she has this gene, nor is there any way she could know that she has this gene. One evening, Hannah and her sister Sarah are at a brand new Mongolian restaurant that has just opened up. Hannah orders a plate of noodles. When her food is brought to the table, Hannah notices something that looks like nuts sprinkled on her noodles and wonders what it is. Sarah says, "I heard that Mongolian dishes are often served topped with Mongolian pine nuts." Hannah notes that the menu says her dish does not contain Mongolian pine nuts. Based on this, Hannah forms the belief that the noodles are not topped with Mongolian pine nuts. If it turns out that the noodles are topped with Mongolian pine nuts, then when Hannah eats the dish, her mouth will get a little dry. Since Hannah has plenty to drink with her meal, it does not matter very much whether or not the noodles are topped with Mongolian pine nuts.

Ignorant high: Hannah has a gene that makes her seriously allergic to Mongolian pine nuts. Eating only a single Mongolian pine nut will cause her to go into shock and die. Hannah has absolutely no idea that she has this gene, nor is there any way she could know that she has this gene. One evening, Hannah and her sister Sarah are at a brand new Mongolian restaurant that has just opened up. Hannah orders a plate of noodles. When her food is brought to the table, Hannah notices something that looks like nuts sprinkled on her noodles and wonders what it is. Sarah says, "I heard that Mongolian
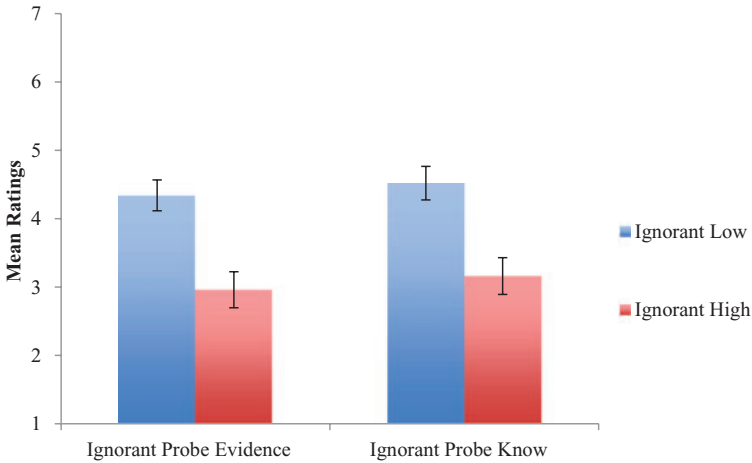
Figure 10. Mean evidence and knowledge ratings from Sripada & Stanley (2012) in *Ignorant low* and Ignorant high, grouped by probe. Scales ran 1–7.

dishes are often served topped with Mongolian pine nuts." Hannah notes that the menu says her dish does not contain Mongolian pine nuts. Based on this, Hannah forms the belief that the noodles are not topped with Mongolian pine nuts. If it turns out that the noodles are topped with Mongolian pine nuts, then when Hannah eats the dish, she will go into shock and die. Since eating even a single Mongolian pine nut will cause her to die, it matters a lot whether or not the noodles are topped with Mongolian pine nuts.

Participants ($N = 100$) were given one of these two vignettes above, and then asked both of the following questions below:

Ignorant probe evidence: What is the strength of Hannah's evidence that her noodles are not topped with Mongolian pine nuts?

Ignorant probe know: Suppose it turns out that her noodles are not topped with Mongolian pine nuts. Please rate how strongly you agree or disagree with the following sentence: "Hannah <u>knows</u> her noodles are not topped with Mongolian pine nuts."

Sripada & Stanley observed a "modest" but statistically significant effect of stakes on both evidential strength ascription and knowledge ascription,[32] as can be seen in figure 10. On the basis of this and related results, they (2012: 16) concluded: "the modest-sized effects observed in these experiments . . . are what should be expected given sensible versions of [*Interest relativity*]."

   We doubt that such modest results should be very encouraging for the friend of *Pragmatism* or *Interest relativity*. They remain a far cry from the strong flip from "knowledge" to "ignorance" which DeRose, Stanley, and most other epistemologists had predicted from the armchair, and they remain far more modest than the strong and robust effect Pinillos reports (despite Sripada & Stanley's (2012: 18) claim that "Pinillos's results mirror and complement our own.") Indeed to our eyes

Sripada & Stanley's effects looks most comparable to the very small effect reported by May *et al*. in the first wave studies (§1.2). The reader should keep in mind that the cases tested—including the bank cases and Sripada & Stanley's *Ignorant* cases— were actually put forward by advocates of *Folk stakes sensitivity* as the best cases for eliciting a stakes effect. One might be reluctant to build a radical epistemology around so modest a result. But let this pass: we have a different criticism to make.

Before issuing criticism, we first want to credit Sripada & Stanley for continuing the discussion as they have. Their experimental design is original, their results are statistically significant and replicable (§3.3), and they look quite clearly to concern knowledge ascription. (Unlike Pinillos who used complex probes with both 'know' and a modal element, Sripada & Stanley used simple knowledge ascription probes.) Indeed we think that Sripada & Stanley have successfully provided the first sustained empirical vindication of *Stakes effect*, and that their results offer real insight into knowledge ascription, albeit not exactly the one they suggest.

3.2 Sripada & Stanley's Studies Confound Stakes and Salience

We are impressed with Sripada & Stanley's data and convinced that they have found a modest but real effect on knowledge ascription (and on ascription of evidential strength as well). But we doubt that their effect is a direct stakes effect. We think that their studies confound what stakes are present for the subject with what mistakes are salient for the ascriber, and thereby do not necessarily support *Folk stakes sensitivity*. They might instead be supporting *Folk salience sensitivity* (the behavioral claim which may be used to support *Contextualism*: §1.1). In other words, we think that their studies fail to keep "all else equal" since they fail to keep fixed the salience of alternatives.

To begin with, we note that participants in *Ignorant high* can be expected to experience some anxiety for Hannah. Perhaps such anxiety is the result of direct sympathy for real and fictional characters, in which participants directly take on some concern for Hannah's well-being (see Loewenstein & Small 2007 for a discussion of the psychological determinants of sympathy). Or perhaps this anxiety is a byproduct of the kind of mental simulation which we might use in attributing mental states to others (Gordon 1986, Goldman 2006), arising from imagining ourselves in Hannah's dangerous situation in which even a nibble of Mongolian pine nut can prove lethal.

The added anxiety experienced by the participants in *Ignorant High* can then be expected to trigger a fear of error alongside concomitant worries about a wider range of possibilities. Indeed this is part of the standard story about the psychological effect of stakes. In this vein, Kruglanski & Mayseless (1987: 835; c.f. Mayseless & Kruglanski 1987; Kruglanski & Webster 1996) note that "heightened fear of invalidity was found to intensify the generation of alternative hypotheses." As they (1987: 835) explain, "[A] fear of invalidity effects an unfreezing of a currently entertained belief and facilitates the generation of numerous alternative hypotheses in order to avoid commitment to a potentially erroneous judgment."

So we think that Sripada & Stanley's studies suffer a *confound*. Sripada & Stanley's vignettes did not merely manipulate what stakes are present for the subject

(Hannah). They also manipulated which alternatives are salient for the ascriber (the experimental participants). Their *Ignorant high* condition can be expected to trigger anxiety in the participants and thereby "to intensify the generation of alternative hypotheses," presumably including the hypothesis that the menu might be mistaken.[33]

The confounding of stakes and salience is especially concerning since there is good independent evidence for a salience effect on knowledge ascription, as per *Folk salience sensitivity*. While the matter is somewhat contested, the latest results—due to Schaffer & Knobe (2012), and successfully reproduced by Buckwalter (*forthcoming*)—show a salience effect.[34] In this vein, Schaffer & Knobe (2012: 695) ran a baseline bank case alongside a bank case with high salience triggered by the addition of a character in the vignette saying: "Well, banks do change their hours sometimes. My brother Leon once got into trouble when the bank changed hours on him and closed on Saturday. How frustrating! Just imagine driving here tomorrow and finding the door locked." They found that this induced a statistically significant difference: participants in their baseline case tended to agree with the knowledge ascription (mean rating: 5.54 out of 7), while participants in their salient mistake condition tended to disagree (mean rating: 3.05 out of 7).

Putting this all together, we have two strands of independent evidence that already suffice to explain Sripada & Stanley's findings without positing a direct stakes effect that persists when all else is equal. Now add to the mix the fact that stakes effects on knowledge ascription were not found on any of the first wave studies (which withstand Sripada & Stanley's criticisms: §3.1), and it seems gratuitous to posit a direct stakes effect on knowledge ascription suddenly arising in Sripada & Stanley's *Ignorant* cases. The conservative explanation for their effect is via salience.[35]

At the very least, we think that, due to the confounding of stakes and salience, Sripada & Stanley's studies fail to establish *Folk stakes sensitivity*. They fail to show that *all else equal*, people are less likely to ascribe knowledge to a high stakes subject than to a low stakes subject, simply because they fail to leave all else equal.

3.3 *Sripada & Stanley's Effect is not a Stakes Effect but a Salience Effect*

So if Sripada & Stanley's studies confound stakes and salience, what would happen if one re-did their studies while controlling for salience? Perhaps this question is supererogatory for us (our burden only being to show that Sripada & Stanley's studies fail to establish *Folk stakes sensitivity*), but in any case it is the crucial question looming. For what is a confound, but an invitation for further research?

However, several difficulties arise. Ideally, one would want to be able to manipulate stakes and salience independently, and measure people's willingness to ascribe knowledge under all four combinations of high and low stakes with high and low salience. But given that some cases of high stakes may trigger high salience, it is not clear how (if at all) one might create any analogue of Sripada & Stanley's *Ignorant high* that is low salience. Ideally, given that we are effectively suggesting that salience might serve to *fully mediate* stakes, one would also want to be able to measure salience in the *Ignorant* cases, so as to establish that high stakes is
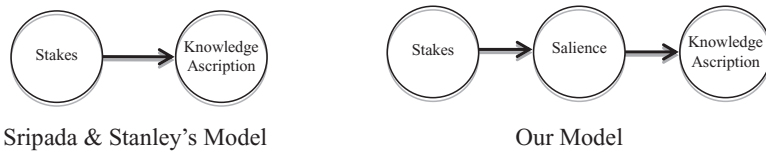
Figure 11. Rival explanatory models for empirical data discussed in §3.2.

causing high salience, that high salience is then causing low agreement with the knowledge ascription, and that salience fully predicts knowledge ascription when stakes are fixed.[36] But it is not clear how to measure salience (especially since asking people whether a scenario is salient may itself have the effect of making the scenario salient), and again it is not clear how to hold fixed high stakes while still manipulating salience.

Despite these difficulties, it remains possible to design a test that speaks to the issue. While it may be hard to fix low salience in a high stakes case, it should be relatively easy to fix high salience in a low stakes case, just by discussing the possibility of error in a concrete and vivid way. If there is a real stakes effect on knowledge ascription (as Sripada & Stanley think), then the effect should persist *even holding fixed high salience*. But if stakes is merely operating indirectly as a trigger for salience (as we think), then the stakes effect should be *screened off* when salience is fixed. In other words, we are trying to test between the two causal models of intuitive judgments about knowledge depicted in figure 11.

Sripada & Stanley's model is a model of *Folk stakes sensitivity*, while our model is a model of *Folk stakes insensitivity* since we say that *all else equal* people are equally likely to ascribe knowledge to a high stakes subject and a low stakes subject (where "all else equal" includes equal salience of mistakes to the ascriber). Seen this way, the telling question is *whether salience screens off stakes*.

With this in mind, we designed two different sorts of studies featuring two different ways to manipulate stakes while holding salience fixedly high. Our prediction was that Sripada & Stanley's effect would disappear. Our first studies in this vein ($N = 120$) used vignettes describing *both* the slight allergy to pine nuts (leading to a slightly dry mouth) and the severe allergy to pine nuts (which can kill), and merely manipulating which of these allergies Hannah is said to have.[37] So we used:

Two allergies low: Most people can eat Mongolian pine nuts without any problem at all. But a small percentage of people have a gene that makes them just slightly allergic to Mongolian pine nuts. If someone with this gene eats Mongolian pine nuts, they will experience a slightly dry mouth. And a small percentage of other people have a different gene that makes them seriously allergic to Mongolian pine nuts. If someone with this other gene eats even a single Mongolian pine nut, they will go into shock and may well die. Hannah does not have the gene that makes people seriously allergic to Mongolian pine nuts, but she happens to have the gene which makes people just slightly allergic to Mongolian pine nuts. Hannah—who has never even heard of Mongolian pine nuts before—has absolutely no idea that she has this gene, nor is there any way she could know that she has this gene. One evening, Hannah and her sister Sarah are at a brand

new Mongolian restaurant that has just opened up. Hannah orders a plate of noodles. When her food is brought to the table, Hannah notices something that looks like nuts sprinkled on her noodles and wonders what it is. Sarah says, "I heard that Mongolian dishes are often served with Mongolian pine nuts." Hannah notes that the menu says her dish does not contain Mongolian pine nuts. If it turns out that the noodles are topped with Mongolian pine nuts, then when Hannah eats the dish, her mouth will get a little dry (she will not go into shock and die like someone who has the other kind of gene). Since Hannah has plenty to drink with her meal, it will not turn out to make much of a difference to her whether or not the noodles are topped with Mongolian pine nuts.

Two allergies high: Most people can eat Mongolian pine nuts without any problem at all. But a small percentage of people have a gene that makes them just slightly allergic to Mongolian pine nuts. If someone with this gene eats Mongolian pine nuts, they will experience a slightly dry mouth. And a small percentage of other people have a different gene that makes them seriously allergic to Mongolian pine nuts. If someone with this other gene eats even a single Mongolian pine nut, they will go into shock and may well die. Hannah does not have the gene that makes people just slightly allergic to Mongolian pine nuts, but she happens to have the gene which makes people seriously allergic to Mongolian pine nuts. Hannah—who has never even heard of Mongolian pine nuts before—has absolutely no idea that she has this gene, nor is there any way she could know that she has this gene. One evening, Hannah and her sister Sarah are at a brand new Mongolian restaurant that has just opened up. Hannah orders a plate of noodles. When her food is brought to the table, Hannah notices something that looks like nuts sprinkled on her noodles and wonders what it is. Sarah says, "I heard that Mongolian dishes are often served with Mongolian pine nuts." Hannah notes that the menu says her dish does not contain Mongolian pine nuts. If it turns out that the noodles are topped with Mongolian pine nuts, then when Hannah eats the dish, she will go into shock and die (she will not merely get a dry mouth like someone who has the other kind of gene). So it turns out to make a huge difference to her whether or not the noodles are topped with Mongolian pine nuts.

We then asked three questions, the first of which was a comprehension check to ensure that people understood what Hannah had at stake, and were not confused as to which type of allergy she has:

Two allergies probe stakes: What sort of allergy does Hannah actually have to Mongolian pine nuts?

(The options were "No allergy," "Slight allergy," and "Severe allergy").[38] Following Sripada & Stanley's preferred methodology (§3.1), the second and third questions were just their *Ignorant probe evidence* and *Ignorant probe know* questions:

Ignorant probe evidence: What is the strength of Hannah's evidence that her noodles are not topped with Mongolian pine nuts?

Ignorant probe know: Suppose it turns out that her noodles are not topped with Mongolian pine nuts. Please rate how strongly you agree or disagree with the following sentence: "Hannah <u>knows</u> her noodles are not topped with Mongolian pine nuts."

We take Sripada & Stanley's model to posit a stronger tendency to agree with the knowledge claim in *Two allergies low* than in *Two allergies high*, presumably
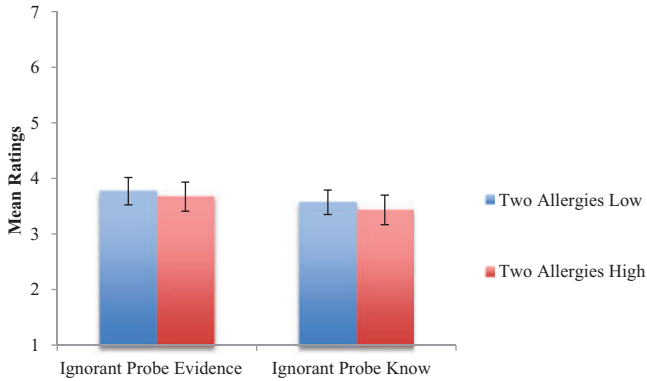
Figure 12. Mean evidence and knowledge ratings in *Two allergy* cases, grouped by probe. Scales ran 1–7.

a difference of the same "modest" strength as they observe in *Ignorant low* versus *Ignorant high*. After all, we have left the stakes difference completely intact. Our model—together with the natural assumption that the salience of a mistake to the participants is fixedly high between *Two allergies low* and *Two allergies high* (since both types of allergy are described in a concrete and vivid way)—posits no difference between these cases.

As we predicted, no statistically significant difference was found between the *Two allergies low* and *Two allergies high* conditions.[39] These results are represented in figure 12. So here is a first indication that salience screens off stakes. Sripada & Stanley's stakes effect disappears when salience is held fixedly high, as seen in our *Two allergies* vignettes.

Of course any one experimental design is open to criticism, so in order to buttress our idea that salience screens off stakes, we ran a second sort of study ($N = 240$) to see if we could generate the same screening off results by different means. To this end we incorporated the exact Sripada & Stanley vignettes in an experimental design that also included the Schaffer & Knobe (2012) technique for manipulating salience by mentioning the possibility of error in a concrete and vivid manner. This resulted in four independent conditions. Roughly half the participants were given one of Sripada & Stanley's original *Ignorant low* or *Ignorant high* stakes cases (verbatim). The other half were given variants of *Ignorant low* or *Ignorant high* that just tacked on two sentences to the end, where the final sentence ("Just imagine...") uses the Schaffer & Knobe technique for fixing high salience:

Ignorant low and salient: [*Ignorant low* plus:] Hannah is lucky that her allergy is so minor: some other people with a slightly different gene can go into severe allergic shock and die if they eat Mongolian pine nuts. Just imagine how horrible it would be if you or someone you loved had such a severe allergy without knowing it, and the restaurant menu was wrong!

Ignorant high and salient: [*Ignorant high* plus:] Hannah is unlucky that her allergy is so serious: some other people with a slightly different gene will merely experience a
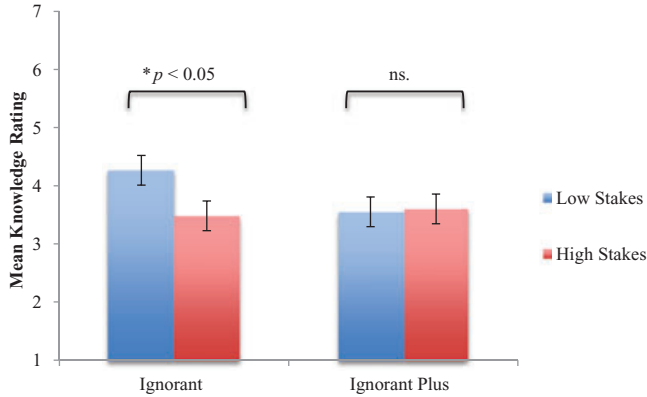
Figure 13. Mean knowledge ratings in low and high stakes *Ignorant* and *Ignorant and salient* [*Ignorant plus*] cases. Scales ran 1–7.

slightly dry mouth if they eat Mongolian pine nuts. Just imagine how horrible it would be if you or someone you loved had such a severe allergy without knowing it, and the restaurant menu was wrong!

We then asked participants the same three questions as in the previous *Two allergies* study above: the comprehension check of *Two allergies probe stakes*, followed by Sripada & Stanley's *Ignorant probe evidence* and their *Ignorant probe know*.[40]

We take Sripada & Stanley's model to posit a stronger tendency to agree with the knowledge claim in *Ignorant low and salient* than in *Ignorant high and salient*, presumably a difference of the same "modest" strength as they observed in *Ignorant low* versus *Ignorant high*. After all, the stakes difference is intact. Our model together with the natural assumption that the salience of a mistake to the participants is fixed between *Ignorant low and salient* and *Ignorant high and salient* (since the possibility of a mistake on the restaurant menu is made concrete and vivid), predicts no difference between these cases.

First, we were able to replicate Sripada & Stanley's effect when comparing results in *Ignorant low* and *Ignorant high*, albeit in an even more modest way: our means were a bit closer than theirs, and our results closer to the conventional threshold for statistical significance.[41] But we did find a statistically significant difference still.

Secondly and more crucially, in line with our prediction, the differences in knowledge ascription between high stakes and low stakes vignettes vanish when one controls for salience as in *Ignorant low and salient* and *Ignorant high and salient*.[42] These results are represented in figure 13.

So we think that our model is preferable to Sripada & Stanley's model for two main reasons. The minor reason is that our model is antecedently more probable, since it provides a *conservative* explanation of the data. Sripada & Stanley's model posits a direct stakes effect on knowledge ascription, which seems unlikely given the first wave results (which withstand Sripada & Stanley's criticism: §3.1). Our model

instead posits a salience effect on knowledge ascription, for which there is good independent evidence. The major reason we think that our model is preferable to Sripada & Stanley's model is that it best fits the full range of empirical evidence. Their effect disappears once one controls for salience, as befits a model in which salience screens off stakes from knowledge ascription.[43]

One might worry that our model, whatever its comparative merits vis-à-vis Sripada & Stanley's model might be, still remains a poor fit to the first wave results (and to our discussion of Pinillos's results). Doesn't our current model predict that the high stakes cases used in the first wave should have triggered participant anxiety, a concomitant increase in the salience of mistakes, and thus a real effect on knowledge ascription? Relatedly one might worry that our model predicts a stronger stakes effect in the Sripada & Stanley studies than the modest effect that was found, since—at least according to Schaffer & Knobe—the effect of salience is fairly strong. We think that both worries misunderstand the complicated psychological nature of salience, which comes in degrees, and which is sensitive to a wide range of positive and negative factors.[44] Our model is committed to the following empirical claims:

- The first wave studies did not notably increase the salience of mistakes.
- Sripada & Stanley's studies minorly increased the salience of mistakes.
- Schaffer & Knobe's studies majorly increased the salience of mistakes.

These claims are empirically open (indeed we find them independently plausible). Though of course they are open to refutation. In that sense, our model is "sticking its neck out" on the psychological determinants of salience.[45]

We thus think that Sripada & Stanley offer real insight into how knowledge ascription works, and we have been moved by their studies to offer a model that allows for *Stakes effect*, albeit in a mitigated and indirect way. But the underlying insight is one that provides further confirmation for *Folk salience sensitivity* (buttressing the results of Schaffer & Knobe 2012, and helping secure the lead premise in an inference to the best explanation for *Contextualism*), and not one that provides any confirmation for *Folk stakes sensitivity*.

Perhaps a still better explanation of Sripada & Stanley's effect can be offered. Whatever the ultimate fate of our positive model may be, our point remains that Sripada & Stanley's effect—whatever its explanation—disappears when salience is held fixed. This already challenges the idea that there is a direct effect of stakes on knowledge ascription, since direct effects do not allow for screening off. The balance of evidence to this point still best fits *Folk stakes insensitivity*. Holding all else equal (including salience), there is no persisting stakes effect on knowledge ascription.

4. Conclusions

Putting all of this together, we conclude that the balance of evidence to date best fits *Folk stakes insensitivity*: all else equal, stakes do not affect knowledge ascription. Our conclusion is based on the following three main points:

- The first wave studies—due to Feltz & Zarpentine (2010), May *et al*. (2010), and Buckwalter (2010), and extended by our *Two Reads* studies—withstand Pinillos's and Sripada & Stanley's criticisms.
- Pinillos's (2012) studies reveal a stakes effect, but not an effect on knowledge ascription: his effect persists when 'know' is removed but the modal retained, and his effect disappears when 'know' is retained but the modal removed.
- Sripada & Stanley's (2012) studies reveal an effect on knowledge ascription, but not a direct stakes effect: their effect disappears when salience is held fixed.

In short, the first wave studies provide strong standing evidence for *Folk stakes insensitivity*, and the second wave studies fail to provide evidence otherwise.

Still, the second wave studies have a lesson to teach us, which is that *Stakes effect* does look to hold (albeit only in special cases, and in a minor and mitigated way). This is more than one might have thought after the first wave studies, but less than *Folk stakes sensitivity* requires. Indeed we have argued that the underlying lesson supports *Folk salience sensitivity* instead.

Our conclusion stands in stark contrast with the received view in epistemology, which has it that *Folk stakes sensitivity* holds. We think that this armchair claim does not withstand empirical scrutiny. Despite the best attempts of the friends of *Folk stakes sensitivity* to formulate cases that might elicit their alleged stakes effect, and despite repeated and diverse attempts at empirical detection, barely any stakes effect on knowledge ascription has been successfully detected, and no stakes effect that persists all else equal has been successfully detected at all. We think this is telling. Of course we cannot foreclose the possibility that the next series of empirical results will reverse this verdict. But pending such a dramatic development, we must conclude that—given the best evidence to date—*Folk stakes sensitivity* is false.

Though we have couched our conclusion in terms of the received epistemological view on stakes not withstanding empirical scrutiny, there are at least two respects in which our conclusion fits well with other received views in epistemology. The more obvious such respect would be that our conclusion undermines a major line of argument for the radical conclusion of *Interest relativity*, and hence befits the received "intellectualist" picture. A less obvious respect is that the bank cases and their ilk were originally offered as supporting *Contextualism* (DeRose 1992). It is only when the friends of *Interest relativity*—in particular Fantl & McGrath (2002), Hawthorne (2004: 156–7), and Stanley (2005: 2–3)—entered the discussion that it emerged that the bank cases might be accounted for via *Folk stakes sensitivity* without supporting *Contextualism*. We are in effect saying that this account fails, and that the bank case intuitions are best understood via *Folk salience sensitivity*, which does render them fit to support *Contextualism*. In this sense we think that the epistemologists had things right from the start.[46]

Notes

[1] In this vein, DeRose (1992: 914–5) suggests that "the requirements for making a knowledge attribution true go up as the stakes go up," and Cohen (1999: 59) says that "if someone's life were at

stake, we might not even be willing to ascribe knowledge [that the flight stops in Chicago] on the basis of the testimony of the airline agent. We might insist on checking with the pilot." Likewise Rysiew (2001: 477) takes such cases to illustrate "a manifest flexibility in our willingness to attribute knowledge," and Nagel (2008: 279) asks "why our natural inclinations to ascribe knowledge become more stringent or lax when subjects are described as, for example, having pressing practical interests in the proposition believed."

[2] Relatedly, Weatherson suggests that stakes might affect justified belief only through belief, and (2005: 442) so concludes: "[W]e can account for the intuitions behind the pragmatic encroachment view while keeping a concept of probabilistic epistemic justification that is free of pragmatic considerations."

[3] Thus Stanley (2005: 5) motivates his defense of *Pragmatism* plus *Interest relativity* based in large part by presenting paired stakes cases and reporting "the intuitive reactions we have," explaining (2005: 13) that his "central interest is to evaluate accounts that make as much sense of these intuitions as possible."

[4] For instance, Nagel (2008: 280) offers "an intellectualist invariantist account that accommodates rather than dismisses our intuitive responses" by arguing "that our stringency with Hi and laxity with Lo is best explained by our natural recognition of differences in the truth-conducive features of high- and low-stakes belief formation."

[5] Thus Hawthorne (2004: 186) speaks of *Interest relativity* as offering "the best hope yet for respecting the intuitive links between knowledge, assertion, and practical reasoning." Overall he suggests (very tentatively) that *Interest relativity* might provide the best fit for a range of further constraints on knowledge including constraints concerning closure, epistemic possibility, and objective chance.

[6] DeRose (1992: 915) is careful to distinguish what is at stake for the subject, and what mistakes are salient for the ascriber, as distinct factors, either (or both) of which might drive bank case intuitions. (Though he wrongly classifies both as "contextual factors" when—as emerged especially through the work of Hawthorne (2004) and Stanley (2005)—stakes is *not* a contextual factor: what is at stake for a given subject in a given situation is an objective matter which cannot vary across contexts of use.) Salience is what corresponds to Lewis's (1996: 559) *Rule of Attention*: "No matter how far-fetched a certain possibility may be, no matter how properly we might have ignored it in some other context, if in *this* context we are not in fact ignoring it but attending to it, then for us now it is a relevant alternative."

[7] Indeed DeRose (2005: 72) goes so far as to say that "the best evidence" for *Contextualism* comes from "what ordinary speakers will count as 'knowledge'." Schaffer (2008: 237) employs a similar style of argument for *Contrastivism*: "[T]he most direct explanation of why knowledge ascriptions are contrast-sensitive is that the knowledge relation has a contrast slot."

[8] For some further reviews of the first wave research, see Buckwalter 2012 (279–81) and Schaffer & Knobe 2012 (§2).

[9] Buckwalter (2010) only includes the re-designed studies using a protagonist who self-ascribes knowledge. The original Buckwalter studies without the self-ascription of knowledge were discussed on the blog *Experimental Philosophy*: http://experimentalphilosophy.typepad.com/experimental_philosophy/2008/07/knowledge-on-sa.html.

It is unfortunate that many of the first wave studies tried to test both *Folks stakes sensitivity* and *Folk salience sensitivity* together, since these hypotheses are best tested via different experimental designs (c.f. Sripada & Stanley 2012: 6). In testing *Folk salience sensitivity* it may be useful to include a protagonist self-ascription of knowledge, but in testing *Folk stakes sensitivity* this is an unneeded and potentially complicating distraction.

[10] In a related vein, Phelan (*forthcoming*), in studies conducted jointly with Neta, found little to no effect of stakes on assessments of "how confident one ought to be," which he takes as proxy for assessments of evidential support.

[11] In the terminology of causal modeling, Pinillos is positing a distribution that violates the methodological precept of *faithfulness* (c.f. Pearl 2000: ch. 2). His model "overfits" the data. That said, we would emphasize that faithfulness is at most a defeasible starting point, and that the empirical evidence that attributors generally expect high stakes subjects to gather more evidence provides some support for positing the upper-left arrow in Pinillos's model. Still, we think that in order to overcome the methodological preference for faithful models, Pinillos would need to provide *specific evidence* that participants

in the first wave studies are actually reading the vignettes in this way. Indeed Sripada & Stanley (2012: 16)—who ask both about strength of evidence and about knowledge—note that their own data does not support the upper-left arrow for related cases, since they do not find a predicted difference in assessments of evidential strength between cases where the subject is aware of the stakes, and cases where the subject is ignorant of the stakes. We thank Chandra Sripada for discussion on this point.

[12] In cases without a protagonist self-ascription of knowledge, participants might still expect the high stakes protagonist to gather more evidence. But they will not receive reassurance from the protagonist that he or she has lived up to these expectations. (By Pinillos's lights, the self-ascription of knowledge was functioning to reassure participants that the protagonist has lived up to expectations). So without a self-ascription of knowledge, participants should be more inclined to treat the high stakes protagonist as having behaved recklessly. Pinillos should thus predict a residual stakes effect in these cases. We thank Joshua Knobe for helping us clarify this matter.

[13] Pinillos uses 'John' to name the protagonist in *Typo high*. We have switched the name to 'Peter' for uniformity.

[14] We report these results in Figures 4 and 5: §2.2.

[15] At this point the arguments for *Interest relativity* from *Folk stakes sensitivity* and from the theoretical role of knowledge in practical reasoning connect (as alluded to in §1.1). Pinillos is appealing to the theoretical role of knowledge in practical reasoning to try to reconcile these results with his case for *Folk stakes sensitivity*.

[16] From Buckwalter (*forthcoming*): *Typo low* belief ($M = 2.71$, $SD = 1.27$), *Typo low* knowledge ($M = 2.61$, $SD = 0.89$), *Typo high* belief ($M = 6.59$, $SD = 5.05$), *Typo high* knowledge ($M = 5.12$, $SD = 3.42$). A 2 (stakes) x 2 (attitude verb) between-subjects analysis of variance was conducted to evaluate the effect of verb type and stakes on participant free responses regarding evidence. A significant main effect was obtained for stakes, $F (1, 86) = 23.1$, $p < 0.01$. However, no main effect was found for verb type, $F (1,86) = 1.40$ $p = 0.24$, and no interaction between these two factors was detected, $F (1, 86) = 1.05$, $p = 0.31$.

[17] As posted on *Experimental Philosophy* (Pinillos's reply begins the comments thread): http://experimentalphilosophy.typepad.com/experimental_philosophy/2010/07/stakes-and-evidenceseeking.html

[18] Perhaps it is still possible to come to a true belief that *p* on the basis of *both* good reasoning and a simultaneous guess, and to still know that *p* due to the presence of the good route. All that is relevant to our argument is that there is no plausible knowledge norm governing what one might guess.

[19] *Typo low* guess ($M = 2.27$, $SD = 1.09$), *Typo low* knowledge ($M = 2.11$, $SD = 1.00$), *Typo high short* guess ($M = 5.11$, $SD = 3.50$), *Typo high short* knowledge ($M = 4.61$, $SD = 2.76$). A 2 (stakes) x 2 (verb type) between-subjects analysis of variance reveals a main effect for the factor of stakes, $F (1, 182) = 60.55$, $p < 0.001$. However a main effect was not found for verb type, $F (1,182) = 0.93$, $p = 0.34$, and no interaction effect was detected between these two factors, $F (1,182) = 0.25$, $p = 0.62$.

[20] *Typo low* hope ($M = 2.50$, $SD = 1.04$), *Typo high short* hope ($M = 5.37$, $SD = 2.61$). Combined from above, a 2 (stakes) x 3 (verb type) between-subjects analysis of variance reveals a main effect for the factor of stakes, $F (1, 260) = 99.46$, $p < 0.001$. However a main effect was not found for verb type $F (2,260) = 1.45$, $p = 0.24$, and no interaction effect was detected between these two factors, $F (2,260) = 0.191$, $p = 0.83$. We also reconfirmed these findings using the Wilcoxon-Mann-Whitney test, which revealed significant stakes effects for all predicates, but no differences between individual verbs within either *Typo low* or within *Typo high short*.

[21] Pinillos (2012: 211) includes a footnote mentioning our results with the 'belief' and 'guess' replacements, and—after mentioning the issue of whether there are knowledge norms—says: "[I]n a further set of studies, I replaced the evidence-seeking questions with standard queries about the level of agreement with simple knowledge statements 'X knows that P.' I found evidence of stakes sensitivity here as well. This is evidence that the evidence-seeking probes are tracking knowledge contrary to what the Schaffer-Buckwalter findings might suggest." But he provides no further information about these studies, and so we are unable to engage with his claim. We do not know what vignettes he ran, what probes he used, or what data he found. We do not know if he controlled for salience (§3). In any case, we present studies of our own that (in accord with the first wave studies) did not detect a stakes effect on knowledge ascriptions, in §2.3.

[22] More precisely, Pinillos's "_____ times" means something like "Peter has to proofread his paper _____ times before he knows that there are no typos." The modal element is most naturally understood as a deontic modal (indicating some norm or standard that ought to be met). So "Peter has to proofread his paper _____ times before he knows that there are no typos" is most naturally understood as something like "Given the norms in play, Peter has to proofread his paper _____ times before he knows that there are no typos." We thank Zoltán Gendler Szabó for discussion on this matter.

[23] We thank Ram Neta for discussion on this point.

[24] We did not have Peter self-ascribe knowledge, so avoiding Pinillos's complaint against the first wave studies (this is the study we alluded to in §2.1). Indeed we left open whether or not Peter was still planning on proofreading his paper further times, as perhaps might be expected in the high stakes case.

[25] *Two reads low* ($M = 5.96$, $SD = 1.20$), *Two reads high* ($M = 5.78$, $SD = 1.30$). An independent samples t-test is unable to detect significant differences between these two groups $t(98) = 0.721$, $p = 0.472$.

[26] We owe this concern to Dustin Locke.

[27] *Two reads uncareful low evidence* ($M = 5.73$, $SD = 1.26$), knowledge ($M = 5.43$, $SD = 1.33$); *Two reads uncareful high evidence* ($M = 5.63$, $SD = 1.25$), knowledge ($M = 5.50$, $SD = 1.38$). An independent samples t-test is unable to detect significant differences for stakes in evidence $t(58) = 0.309$, $p = 0.76$ or knowledge $t(58) = -0.19$, $p = 0.85$ probes.

[28] In Mill's (1950) classical terminology, the method of causal reasoning in use is the joint method of agreement and difference.

[29] In particular, Sripada & Stanley (2012: 16) acknowledge that their fourth worry—the one closest to Pinillos's worry, which they label "the suppressor hypothesis"—is not borne out in their data.

[30] *Two reads low* (with evidence preface) knowledge ($M = 5.62$, $SD = 1.11$), *Two reads high* (with evidence preface) knowledge ($M = 5.70$, $SD = 1.37$). Pooling the results from all two read cases above, a 2 (stakes) x 2 (evidence preface) between-subjects analysis was conducted. No effects were detected for stakes, $F(1, 196) = 0.08$, $p = 0.78$, evidence preface, $F(1,196) = 1.42$, $p = 0.24$, nor interaction, $F(1, 196) = 0.54$, $p = 0.46$. Indeed all scores are significantly above the midpoint ($p < 0.01$), signaling knowledge ascription in all cases.

[31] *Two reads low* (with evidence preface) evidence ($M = 5.86$, $SD = 1.07$), *Two reads high* (with evidence preface) evidence ($M = 5.88$, $SD = 1.24$). An independent samples t-test is unable to detect significant differences between these two groups $t(98) = -0.90$, $p = 0.93$.

[32] Evidential strength was measured on a seven item scale with '1' anchored to 'very weak evidence' and '7' anchored to 'very strong evidence'. Knowledge ascription in Sripada & Stanley was also measured using a seven item scale with '1' anchored to 'strongly agree' and '7' anchored to 'strongly disagree'. Scores to this latter question were then reverse coded to correspond to the evidence scale, such that higher scores represented greater agreement.

[33] Point of clarification: We are not saying that high stakes cases *always* trigger high salience, only that the specific details of *Ignorant high* managed to do so. We discuss this further in §3.3.

[34] The matter of *Folk salience sensitivity* is contested insofar as the first wave studies of May et al (2010) and Buckwalter (2010) failed to find any effect of salience. Schaffer & Knobe (2012: 694–5) diagnosed a flaw in how these studies addressed salience, and were able to find a salience effect in corrected vignettes (their results were replicated in Buckwalter *forthcoming*). Schaffer & Knobe also found a range of uniform effects associated with a range of other ways of shifting the salient contrast (such as using 'rather than'-clauses), which provides further independent support for *Folk salience sensitivity*. See Buckwalter 2012 (§2) for further discussion. (Indeed we—in collaboration with David Rose—have since found additional support for *Folk salience sensitivity*, which we hope to document in future work.)

[35] Schaffer & Knobe—writing prior to Pinillos and to Sripada & Stanley—discarded the prospect of a stakes effect on the basis of the first wave studies. But they (2012: 704, fn. 14) noted that their preferred contrastive model could account for any apparent stakes effects "if coupled with the further speculative claim that high stakes tend to trigger consideration of a wider range of scenarios." The previous discussion has been to show how Schaffer & Knobe's "speculative claim" is in fact empirically plausible.

[36] This is the Baron & Kenny (1986) test for mediation. We thank David Rose for guidance here.

[37] We used the effect size found in Sripada & Stanley's *Ignorant* cases ($d = 0.72$) as an *a priori* measure to calculate the sample size of the current experiments. (We thank Chandra Sripada and Jason Stanley for sharing the information we needed to calculate this.) Power analysis indicates that a total sample size of 64 people is needed to detect their effect with 80% power using a *t* test between means with an alpha of 0.05 (two-tailed). However since we are predicting a null result, and wish to err on the side of caution, we *doubled* the sample size required. This effectively allows us to uphold the predicted result at the level above for increasingly smaller effect sizes, up to and including what is usually considered "moderate" effect sizes ($d > 0.52$).

[38] Two participants were removed and replaced for failing the comprehension question in *Two allergies probe stakes*. We used the same scales as Sripada & Stanley (but without their reverse-coding procedure).

[39] *Two allergy low* knowledge ($M = 3.57$, $SD = 2.04$); *Two allergy high* knowledge ($M = 3.43$, $SD = 2.06$); *Two allergy low* evidence ($M = 3.77$, $SD = 1.91$); *Two allergy high* evidence ($M = 3.67$, $SD = 1.71$). An independent samples t-test was not able to detect a significant effect for stakes on knowledge $t(118) = 0.36$, $p = 0.72$, nor for evidence ratings, $t(118) = 0.30$, $p = 0.76$.

[40] Three participants were removed and replaced for failing the comprehension check.

[41] *Ignorant low* knowledge ($M = 4.27$, $SD = 2.05$); *Ignorant high* knowledge ($M = 3.48$, $SD = 1.86$); *Ignorant low* evidence ($M = 4.57$, $SD = 1.80$); *Ignorant high* evidence ($M = 3.55$, $SD = 1.70$). An independent samples t-test was used to detect a significant effect for stakes on knowledge $t(118) = 2.19$, $p < 0.05$, and for evidence ratings, $t(118) = 3.19$, $p < 0.01$.

[42] *Ignorant low and salient* knowledge ($M = 3.55$, $SD = 1.99$); *Ignorant high and salient* knowledge ($M = 3.60$, $SD = 2.03$); *Ignorant low and salient* evidence ($M = 3.70$, $SD = 1.87$); *Ignorant high and salient* evidence ($M = 3.82$, $SD = 1.90$). We were not able to detect significant differences between *Ignorant low and salient* and *Ignorant high and salient* knowledge, $t(118) = -0.136$, $p = 0.892$, nor between *Ignorant low and salient* and *Ignorant high* knowledge, $t(118) = -0.190$, $p = 0.85$.

[43] Point of clarification: we are not saying that these results uniquely secure our model. As explained at the start of §3.3, there are several difficulties arising in trying to fully rule out a stakes effect (e.g. by establishing full mediation). We are saying that the results here, insofar as they match the predictions of our model but not the predictions of Sripada & Stanley's model, provide relative confirmation for our model being preferable to theirs.

[44] Indeed part of the moral of the Schaffer & Knobe critique of the first wave studies on salience is that these studies supposed that merely mentioning a possibility thereby rendered it salient, when the effect of mentioning on salience is complicated and partly depends on whether the possibility is mentioned in a concrete and vivid way.

[45] Though our model is non-committal on Pinillos's studies, save for the claim that the salience of mistakes was not notably increased between *Two reads low* and *Two reads high*.

References

Baron, Reuben & David Kenny 1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51: 1173–1182.

Brown Jessica 2005. Comparing Contextualism and Invariantism on the Correctness of the Contextualist Intuitions. *Grazer Philosophische Studien* 69: 71–100.

Beebe, James & Wesley Buckwalter 2010. The Epistemic Side-Effect Effect. *Mind and Language* 25: 474–498.

Buckwalter, Wesley 2010. Knowledge isn't Closed on Saturday: A Study in Ordinary Language. *Review of Philosophy and Psychology* 1: 395–406.

_____ 2012. Non-Traditional Factors in Judgments about Knowledge. *Philosophy Compass* 7: 278–289.

_____ *forthcoming*. The Mystery of Stakes and Error in Ascriber Intuitions. *Advances in Experimental Epistemology*, ed. Beebe. Continuum Press.

Cohen, Stewart 1999. Contextualism, Skepticism, and the Structure of Reasons. *Philosophical Perspectives* 13: 57–89.

DeRose, Keith 1992. Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research* 52: 913–929.

_____ 2005. The Ordinary Language Basis for Contextualism and the New Invariantism. *Philosophical Quarterly* 55: 172–198.

_____ 2011. Contextualism, Contrastivism, and X-Phi Surveys. *Philosophical Studies* 156: 81–110.

Fantl, Jeremy & Matthew McGrath 2002. Evidence, Pragmatics, and Justification. *The Philosophical Review* 111: 67–94.

_____ 2009. *Knowledge in an Uncertain World*. Oxford University Press.

Feltz, Adam & Chris Zarpentine 2010. Do You Know More When it Matters Less? *Philosophical Psychology* 23: 683–706.

Gerken, Mikkel *forthcoming*. Epistemic Focal Bias. *Australasian Journal of Philosophy*.

Goldman, Alvin 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gordon, Robert 1986. Folk Psychology as Simulation. *Mind and Language* 1: 158–171.

Hawthorne, John 2004. *Knowledge and Lotteries*. Clarendon Press.

Hawthorne, John & Jason Stanley 2008. Knowledge and Action. *Journal of Philosophy* 105: 571–590.

Kruglanski, Arie & Ofra Mayseless 1987. Motivational Effects in the Social Comparison of Opinions. *Journal of Personality and Social Psychology* 53: 834–842.

Kruglanski, Arie & Donna Webster 1996. Motivated Closing of the Mind: "Seizing" and "Freezing." *Psychological Review* 103: 263–283.

Lewis, David 1996. Elusive Knowledge. *Australasian Journal of Philosophy* 74: 549–567.

Loewenstein, George & Deborah Small 2007. The Scarecrow and the Tin Man: The Vicissitudes of Human Sympathy and Caring. *Review of General Psychology* 11: 112–126.

May, Joshua, Walter Sinnott-Armstrong, Jay G. Hull & Aaron Zimmerman 2010. Practical Interests, Relevant Alternatives, and Knowledge Attributions: An Empirical Study. *Review of Philosophy and Psychology* 1: 265–273.

Mayseless, Ofra & Arie Kruglanski 1987. What Makes You So Sure? Effects of Epistemic Motivations on Judgment Confidence. *Organizational Behavior and Human Decision Processes* 39: 162–183.

Mill, J. S. 1950. *A System of Logic*. Macmillan Press.

Nagel, Jennifer 2008. Knowledge Ascriptions and the Psychological Consequences of Changing Stakes. *Australasian Journal of Philosophy* 86: 279–294.

Pearl, Judea 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Phelan, Mark *forthcoming*. Evidence that Stakes don't Matter for Evidence. *Philosophical Psychology*.

Pinillos, N. Ángel 2011. Some Recent Work in Experimental Epistemology. *Philosophy Compass* 6: 675–688.

_____ 2012. Knowledge, Experiments and Practical Interests. *Knowledge Ascriptions*, eds. Brown & Gerken: 192–221. Oxford University Press.

Rysiew, Patrick 2001. The Context-Sensitivity of Knowledge Attributions. *Noûs* 35: 477–514.

Schaffer, Jonathan 2008. The Contrast-Sensitivity of Knowledge Ascriptions. *Social Epistemology* 22: 235–245.

Schaffer, Jonathan & Joshua Knobe 2012. Contrastive Knowledge Surveyed. *Noûs* 46: 675–708.

Sripada, Chandra & Jason Stanley 2012. Empirical Tests of Interest-Relative Invariantism. *Episteme* 9: 3–26.

Stanley, Jason 2005. *Knowledge and Practical Interests*. Oxford University Press.

Turri, John *forthcoming*. Linguistic Intuitions in Context. *Intuitions*, eds. Booth & Rowbottom. Oxford University Press.

Weatherson, Brian 2005. Can We Do without Pragmatic Encroachment? *Philosophical Perspectives* 19: 417–443.

Williamson, Timothy 2000. *Knowledge and its Limits*. Oxford University Press.