Truth, Fallibility, and Justification:
New Studies in the Norms of Assertion[*]

Abstract: This paper advances our understanding of the norms of assertion in two ways. First, I evaluate recent studies claiming to discredit an important earlier finding which supports the hypothesis that assertion has a factive norm (i.e. assertions should express truths). In particular, I evaluate whether it was due to stimuli mentioning that a speaker's evidence was fallible. Second, I evaluate the hypothesis that assertion has a truth-insensitive standard of justification. In particular, I evaluate the claim that switching an assertion from true to false, while holding all else objectively constant, is irrelevant to attributions of justification. Two pre-registered experiments provide decisive evidence against each claim. In the first experiment, switching from mentioning to not mentioning fallibility made no difference to assertability attributions, thereby disproving the criticism concerning fallibilty. By contrast, switching an assertion from true to false decreased the rate of assertability attribution from over 90% to less than 20%, thereby replicating and vindicating the original finding supporting a factive norm. In the second experiment, switching an assertion from true to false decreased the rate of justification attribution from over 80% to 10%, thereby undermining the hypothesis that assertion's standard of justification is truth-insensitive. The second experiment also demonstrates that perspective-taking influences attributions of justification, and it provides initial evidence that the standard of justification for assertion is stricter than the standard for belief.

Introduction

Much of what you or I know about the world we get second-hand through communication. Communication would be worthless without trust. If you don't trust someone, you aren't going to rely on their word. And if others consistently refuse to trust what you tell them, then it's hard to see the point in even trying.

[*] This is the penultimate version of a paper forthcoming in *Synthese*. Please cite the final, published version if possible.

Trust is a double-edged sword. By trusting others, you stand to benefit from their talents and cooperation. Yet trust also leaves you vulnerable to manipulation and betrayal. A balance must be struck, but how? How do we establish and maintain the bonds of trust?

This question arises for members of any species that relies on communication. Researchers in the interdisciplinary field of animal communication studies have found that animals establish and maintain trust by instinctively following certain behavioral rules (Bradbury & Vehrencamp 2011).

One rule is to attend preferentially to “information constrained” signals, or signals that only individuals with particular knowledge would produce. For example, a sparrow needs to distinguish other sparrows (“conspecifics”) who are invading its territory from those who innocently occupy neighboring territory. A sparrow accomplishes this based on whether the conspecific imitates the song the sparrow just sang (“song matching”), or sings a different song that the sparrow sang previously (“repertoire matching”). A neighbor would have had time to memorize other songs that the sparrow sang, but an invader would not. This makes repertoire matching an information-constrained signal of neighborhood. Sparrows instinctively rely on it when deciding how to respond to nearby conspecifics (Beecher et al. 2000).

A wealth of observational and experimental evidence supports the hypothesis that a broadly similar rule is part of human communication (for reviews see Turri 2016b; Turri 2017b). The *knowledge-rule hypothesis* states that according to the rules of our social information-sharing practice, assertions should express knowledge. (Note: the proposal does not say that this is the *only* rule.) This is an empirical hypothesis that predicts a central tendency to link judgments about what *true* and *known* to judgments about what *should be asserted* in the behavior of competent language users. The central tendency has been detected many times across a range of contexts, including cross-culturally (Turri & Park 2018). When combined with its grounding in the broader scientific study of communication and evidence from social observation and linguistic development, the experimental findings place the knowledge-rule hypothesis on extremely strong footing and, at this point, far beyond reasonable doubt.

Part of what's at stake, therefore, in evaluating the knowledge-rule hypothesis is whether it reveals something deep and important about human information-sharing practices, or whether all the convergent evidence is instead just a massive coincidence.

In good scientific spirit, critics have recently challenged the hypothesis with new experimental evidence. One series of studies allegedly found that participants tended to judge that agents should assert false propositions (Kneer 2018). However, follow-up studies revealed a confound and when it was removed the results replicated earlier findings supporting a factive norm of assertion (i.e. one that makes truth essential to assertability, as a knowledge rule does) (Turri 2018).

Another series of studies claimed to undermine a previous finding supporting a factive norm (Reuter & Brössel 2018). The finding is that switching a proposition's truth-value from true to false causes participants to switch from mostly judging that it should be asserted to judging that it shouldn't be asserted, even while holding fixed all other objective features of a situation (Turri 2013). Researchers successfully replicated the original finding supporting a factive norm, but they also claim to have found two critical flaws in earlier experiments.

On the one hand, they claim that it is "improper to ask" participants what an agent *should* say, because "should" might pertain to practical matters irrelevant to assertability. This concern, although reasonable, was ruled out by prior experiments that collected a range of judgments pertaining to practical matters and found that, even when controlling for their influence, attributions of assertability were powerfully influenced by truth-value (Turri 2017a; see also 2015a).[1] On the other hand, they claim that the scenario participants read made evidential fallibility salient in a subtly illicit way. Moreover, and more importantly, they claim to have found that simply removing mention of fallibility reversed the central tendency: now most participants judge that a false assertion should be made. They interpret this as evidence that assertion does not have a factive norm but instead has a norm of justification that is insensitive to truth.

The present research is dedicated to evaluating the remainder of this challenge, pertaining to fallibility and justification. In particular, I report a pre-registered attempt to replicate the reported finding regarding fallibility. The finding did not replicate: the results were exactly the opposite of what the critics report and align closely with the original finding supporting a factive norm. In light of the failed replication, I proceed to test the hypothesis that justification is a central norm of assertion, where this status is alleged to be independent of whether an assertion is true. The findings do not support the existence of such a norm. Instead they support the conclusion that assertion's standard of justification is deeply truth-sensitive. In the process, they also demonstrate that perspective-taking affects judgments about justification, and they provide evidence that the assertion's standard of justification is stricter than belief's.

Before proceeding, it is important to note some unfortunate details of the research I will be responding to. For their first experiment, the critics claim to test stimuli that are "almost identical" (Reuter & Brössel 2018: 8) to those used in earlier research (Turri 2013). But this is false. In the critics' version, the agent asserts the proposition, but this does not happen in the original. In

[1] Critics have pressed other objections pertaining to the terminology used to probe for assertability attributions (Kneer 2018; cf. Turri 2013: p. 281). I address this issue elsewhere in research currently in progress.

the critics' version, the agent's evidence is characterized as "malfunctioning," but this does not happen in the original. In the critics' version, the agent is explicitly characterized as ignorant ("unbeknownst"), but this does not happen in the original. The critics' version included the test statement at the end of the scenario, but this is not true of the original. The critics included three answer options, whereas the original includes only two. Finally, the critics' version contains at least one grammatical error not found in the original. The critics' method section is unclear in certain respects, so there could be other differences that aren't evident from the published record. But the ones already noted are enough to disqualify this as a legitimate replication attempt. In the first experiment reported below, I am careful to use the exact original materials (from Turri 2013) when testing whether the original finding was due to mentioning evidential fallibility.

For their second experiment, the critics claim to test whether "the justified belief account prevails" over accounts focused on truth by testing "scenarios that directly pitted truth against justification" (Reuter & Brössel 2018: 11). They observe differences across condition and claim that this "strongly suggests" justification, but not truth, is relevant to assertability. But the observation is confounded in so many ways that the results are basically uninterpretable. Differences include the following (Reuter & Brössel 2018: 12). The assertion's content switches from an affirmation to a denial. The basis for the agent's assertion changes. The description of the agent's mental state changes. The response option for affirming assertability changes. The response option for denying assertability changes. The text of the scenario is also ungrammatical due to punctuation errors. An experiment with this many conspicuous differences and no proper controls is uninformative. In the second experiment reported below, I am careful to test the potential effect of only tightly controlled differences on the attribution of justification.

General methods

The following statements are true of all studies reported here. All manipulations, measures, and exclusion criteria are reported. All participants were adult residents of the United States. I recruited and tested people using an online platform of Amazon Mechanical Turk (https://www.mturk.com), TurkPrime (Litman, Robinson, and Abberbock 2017), and Qualtrics (https://www.qualtrics.com). Participants completed a brief demographic questionnaire after testing. I used R 3.5.2 for all analyses (R Core Team 2018). All stimuli, data, and code are available through an Open Science Foundation project (https://osf.io/sk78y/). All studies were pre-registered.

Experiment 1

This experiment attempted to replicate the alleged finding that a key result supporting a factive norm of assertion was an artifact of mentioning fallibility in experimental stimuli. In the process, the experiment also constitutes a pre-registered replication attempt of the key result.

Method

I decided in advance to recruit 50 participants per condition, plus some extra as a precaution against attrition (see pre-registration).

Participants

Out of 227 participants recruited, 17 (7%) failed a comprehension question and were excluded from further analysis (pre-registered exclusion), yielding a final sample of 210. Their mean age was 37.89 years (range = 20-70, SD = 11.81), 47% (98 of 210) were female, and 96% reported native competence in English.

Materials and procedure

Participants were randomly assigned to one of four conditions in a 2 (truth-value: false, true) × 2 (inclusion: unmentioned, mention) experimental design. Participants first read a brief scenario about an agent who is asked a question. Then they rated whether the agent should make an assertion. Then they went to a new screen and answered a comprehension question from memory. The truth-value factor manipulated whether the relevant proposition was false or true. The inclusion factor manipulated whether the scenario mentioned that the agent's evidential source was imperfect. For mentioned conditions, the scenario's exact text was taken verbatim from Experiment 1 of Turri (2013). For unmentioned conditions, the text was exactly the same except for deleting one sentence, "Maria knows that the inventory is not perfect, but it is extremely accurate," which some researchers argue is problematic (see the Introduction). Here is the text of the scenario, with the sentence of interest in curly braces and the truth-value manipulation bracketed, followed by the test item and comprehension question (response options rotated randomly):

> Maria is a watch collector. She owns so many watches that she cannot keep track of them all by memory alone. So she maintains a detailed inventory of them. She keeps the inventory up to date. {Maria knows that the inventory isn't perfect, but it is extremely accurate.}
>
> Today Maria is having guests over for dinner. Soon after dinner is served, one of her guests asks, "Maria, do you have a 1990 Rolex Submariner in your watch collection?"

> Maria consults her inventory. It says that she does have a 1990 Rolex Submariner in her collection. [But this is one of those rare cases where the inventory is wrong: she does not have one /And this is just another case where the inventory is exactly right: she does have one].
>
> Should Maria tell her guest that she has a 1990 Rolex Submariner in her collection?
>
> - Yes
>
> - No
>
> Is there a 1990 Rolex Submariner in Maria's collection?
>
> - Yes
>
> - No

Coding

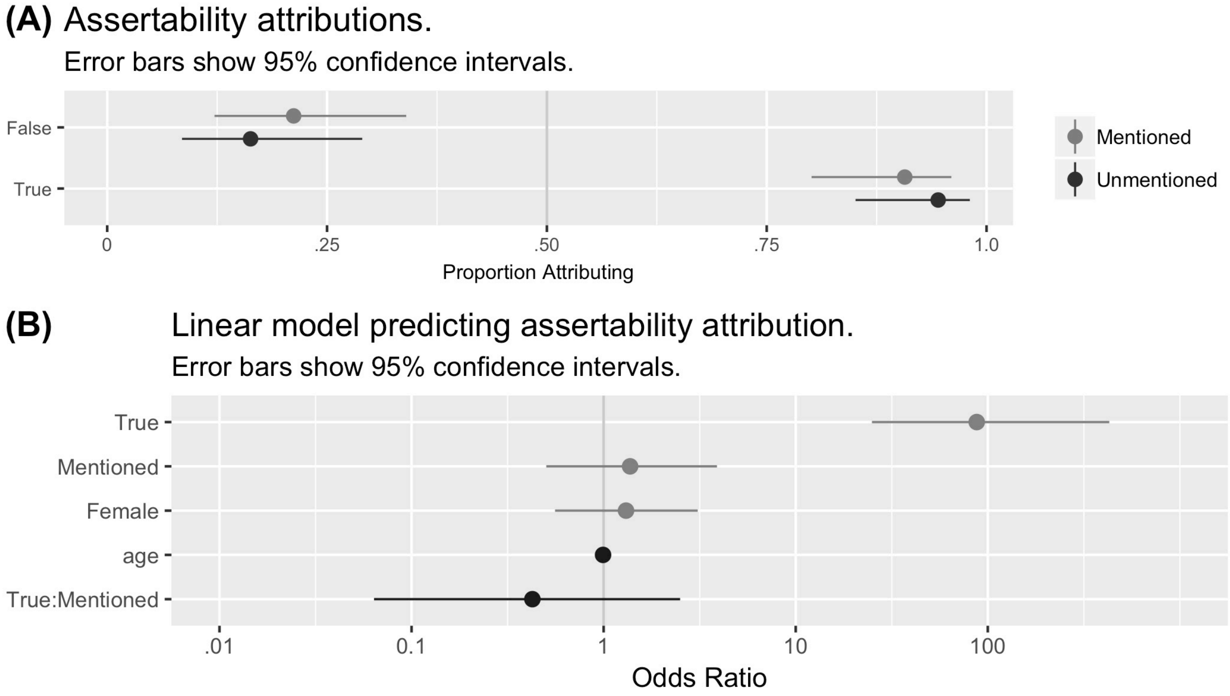
I interpreted "Yes" as an attribution of assertability (coded as 1) and "No" as a denial (coded as 0).

Data analysis and predictions

The principal research questions were whether mentioning a source's fallibility would affect assertability attributions, and whether the previously observed effect of truth-value would replicate. To answer these questions, I analyzed attributions using a generalized linear model with truth-value (false, true), inclusion (unmentioned, mentioned), and participant age and sex as predictors. I followed this up with proportion tests on attribution rates in the relevant conditions. I predicted that mentioning fallibility would not affect attributions but that truth-value would, with above chance attribution rates in true conditions and below chance rates in false conditions.

Results

The linear model revealed that assertability attribution was significantly affected only by truth-value (see Figure 1). The switch from false to true increased the odds of an attribution by nearly a factor of 100. Binomial tests revealed that attribution rate in false conditions (18.8%) was significantly below chance rates, whereas it was significantly above chance rates in true conditions (92.7%) (see Table 1).



Experiment 1. (A) Rates of assertability attribution across four conditions (between-subjects) that varied truth-value (false, true) and whether a source's fallibity was unmentioned or mentioned. (B) Visualization of generalized linear model predicting attribution, showing odds ratios. Error bars show 95% confidence intervals.

Experiment 1. Proportions and binomial tests for assertability attributions across truth-value conditions.

Truth Value	n	k	proportion	95 CI low	95 CI high	test value	p	h
False	101	19	.188	.124	.275	.5	<.001	-0.674
True	109	101	.927	.862	.962	.5	<.001	1.022

Discussion

This experiment tested the criticism that an important finding on assertability supporting a factive norm was due to mentioning fallibility in the stimuli (Reuter & Brössel 2018). The results disprove the criticism. Assertability attributions were unaffected by mentioning fallibility. Replicating the original finding, truth-value powerfully affected attributions, which were below 20% when the proposition was false and over 90% when the proposition was true.

Experiment 2

In addition to falsely claiming to have identified methodological problems with earlier studies, critics also claim that assertion is associated with a truth-insensitive standard of justification determined by the speaker's evidence, rather than truth. The present experiment directly investigates whether assertion is associated with a truth-insensitive standard of justification by manipulating truth-value and probing for judgments about justification. This experiment also breaks new ground in two other ways: by testing the role of perspective-taking in attributions of justification, and by directly comparing attributions of justification for believing and asserting a proposition. The role of perspective-taking has been studied for judgments about what *should* be asserted

(Turri 2018, 2016a), but it has not previously been tested for judgments about whether an assertion is justified.

Method

I decided in advance to recruit 50 participants per condition, plus some extra as a precaution against attrition (see pre-registration).

Participants

Out of 424 participants recruited, 25 (6%) failed a comprehension question and were excluded from further analysis (pre-registered exclusion), yielding a final sample of 399. Their mean age was 38.26 years (range = 19-81, SD = 12.79), 48% (193 of 399) were female, and 94% reported native competence in English.

Materials and procedure

Participants were randomly assigned to one of eight conditions in a 2 (truth-value: false, true) × 2 (option: plain, contrast) × 2 (focus: think, assert) experimental design. Participants first read the same basic scenario used in the mention conditions of Experiment 1. Then they rated a justification attribution. Then they went to a new screen and answered a comprehension question from memory (same as in Experiment 1). The truth-value factor manipulated whether the relevant proposition was false or true. The option factor manipulated which response options participants used to rate justification attributions. The plain options allowed participants to select whether the agent's evidence "does" or "does not" justify her. The contrast option allowed participants to select whether the agent's evidence "actually does" or "only seems to" justify her.

The focus factor manipulated whether the justification attribution pertained to what the agent is justified in *thinking* or *saying*.

> Maria's evidence _____ her in [thinking / saying] that she has a 1990 Rolex Submariner.
>
> - does not justify / does justify (plain options)
>
> - only seems to justify / actually does justify (contrast options)

Coding

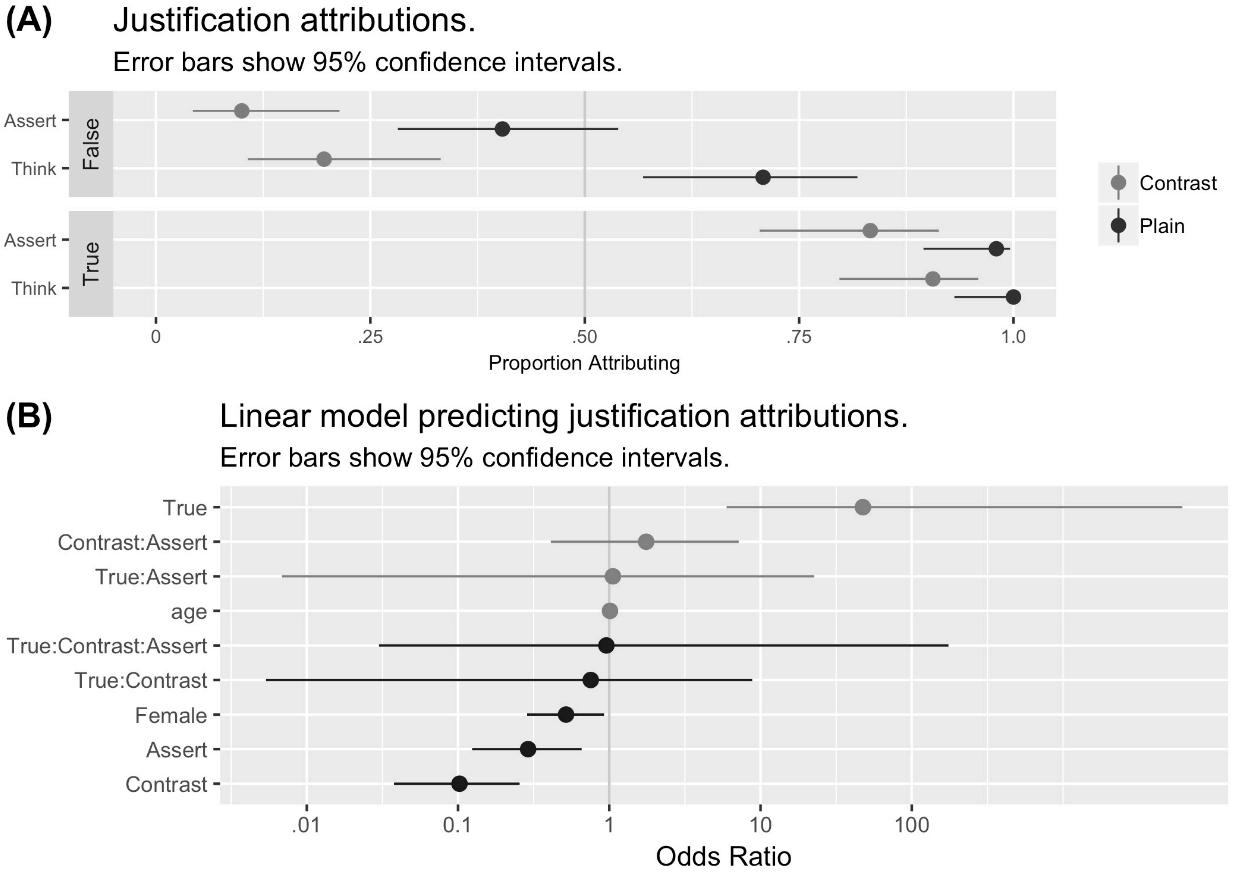
I interpreted "does justify" and "actually does justify" as an attribution of justification (coded as 1) and "does not justify" and "only seems to justify" as denials (coded as 0).

Data analysis and predictions

The principal research question was whether the three independent variables would affect justification attribution. To answer this question, I analyzed attributions using a generalized linear model with truth-value (false, true), option (plain, contrast), focus (think, assert), and participant age and sex as predictors. I followed this up with proportion tests on attribution rates in the relevant conditions. I predicted main effects of truth-value (higher in true conditions) and option (higher in plain conditions) and an interaction between truth-value and option (truth-value mattering more in contrast conditions).

Results

In one of the conditions (true-plain-think), 100% of participants attributed justification, resulting in complete separation when fitting the linear model. To address this, I fit the model using a penalized likelihood method (Firth 1993; Heinze 2006) (see Figure 2). Even with the penalized bias correction, the standard errors on some of the coefficient estimates remained high. There were main effects of truth-value, option, and focus. No interaction reached significance, including truth-value by option. The switch from false to true increased the odds of an attribution by nearly a factor of 50. The switch from plain to contrast options decreased the odds of an attribution by a factor of 10. The switch from evaluating thinking to asserting decreased the odds of an attribution by a factor of 3.5. Follow-up binomial tests (see Table 2) showed that justification attribution was significantly above chance in all four true conditions. In false conditions, attribution was significantly above chance for thinking evaluated with plain options, numerically below chance for asserting when evaluated with plain options, and significantly below chance for thinking and asserting when evaluated with contrast options.



Experiment 2. (A) Rates of justification attribution across eight conditions (between-subjects) that varied truth-value (false, true), answer options (plain, contrast), and the focus of evaluation

(think, assert). (B) Visualization of generalized linear model predicting attribution, showing odds ratios. Error bars show 95% confidence intervals.

Experiment 2. Proportions and binomial tests for justification attributions across eight conditions.

Tvalue	Option	Focus	n	k	prop	95 CI low	95CI high	test value	p	h
False	Plain	Think	48	34	.708	.568	.818	.5	.006	0.430
False	Plain	Assert	52	21	.404	.282	.539	.5	.212	-0.194
False	Contrast	Think	46	9	.196	.107	.332	.5	< .001	-0.654
False	Contrast	Assert	50	5	.100	.043	.214	.5	< .001	-0.927
True	Plain	Think	52	52	1.000	.931	1.000	.5	< .001	1.571
True	Plain	Assert	50	49	.980	.895	.996	.5	< .001	1.287
True	Contrast	Think	53	48	.906	.797	.959	.5	< .001	0.946
True	Contrast	Assert	48	40	.833	.704	.913	.5	< .001	0.730

Binomial tests revealed that overall attribution rate in false conditions (35.2%) was significantly below chance rates, whereas it was significantly above chance rates in true conditions (93.1%). Overall attribution rate in plain conditions (77.2%) was significantly above chance rates, but it did not differ from chance rates in contrast conditions (51.8%)

Experiment 2. Proportions and binomial tests for justification attributions across truth-value conditions (false, true).

Tvalue	n	k	prop	95 CI low	95 CI high	test value	p	h
False	196	69	.352	.289	.421	.5	<.001	-0.300
True	203	189	.931	.888	.958	.5	<.001	1.039

Experiment 2. Proportions and binomial tests for justification attributions across answer option conditions (plain, contrast).

Option	n	k	prop	95 CI low	95 CI high	test value	p	h
Plain	202	156	.772	.710	.825	.5	<.001	0.576
Contrast	197	102	.518	.448	.587	.5	.669	0.036

Discussion

This experiment tested whether justification attributions are affected by truth-value, answer options encoding an appearance/reality distinction, or shifting the focus from thinking to asserting. All three factors had an effect. For both thinking and asserting, an attribution was significantly more likely when the target proposition was true and when the answer answer options did not encode an appearance/reality distinction. Attribution was also more likely for thinking than for asserting. Overall, the attribution rate was 35% when the proposition was false and over 90% when the proposition was true. With respect to assertion specifically, justification attribution was as low as 10% when the proposition was false and over 80% when it was true. These results undermine the hypothesis that assertion is associated with a truth-insensitive standard of justification. They replicate previous findings on the truth-sensitivity of evaluations of belief (Turri 2015b). They also challenge any view hypothesizing a common standard of justification for belief and assertion. In particular, the present findings suggest that assertion's standard is more demanding than belief's.

General Discussion

Two experiments advanced our understanding of assertability, in two ways. On the one hand, researchers recently claimed to show that an important result supporting a factive norm of assertion was due to the stimuli mentioning evidential fallibility. The results from a pre-registered replication attempt disprove their claim (Experiment 1). Comparing the original stimuli to a closely matched control condition that did not mention fallibility, I found no evidence that mentioning fallibility affected assertability attributions. By contrast, replicating the original important result from the literature, I found that truth-value had an enormous effect: switching a proposition from false to true, while holding all else objectively constant, boosted the odds of an attribution by a factor of 100. In absolute terms, the rate of attribution rose from under 20% to over 90%.

On the other hand, researchers also claim to have found strong evidence that assertion is associated with a truth-insensitive standard of justification, which is unaffected by objective truth-value. The studies which allegedly provide evidence for this conclusion were so multiply confounded as to be uninterpretable. In order to gain better evidence, I conducted a pre-registered study of justification attributions that closely controlled truth-value and another factor, the formulation of answer options, which has been shown to affect evaluative social judgments by triggering or inhibiting perspective-taking. The results provide strong evidence against the existence of a truth-insensitive norm of justification: switching a proposition from false to true, while holding all else objectively constant, boosted the odds of an attribution by a factor of 50 (Experiment 2). Additionally, the results also further demonstrate the importance of using response options that inhibit perspective-taking when probing for justification attributions. Finally, the results also provided interesting initial evidence against the hypothesis that assertion and belief have a common standard of justification. More specifically, it appears that assertion is associated with a more stringent standard.

In closing I would like to note that I deliberately refrained from a detailed examination of some errors in recent critical work on assertability, including misrepresentations of my own views, in order to focus on substantive questions that can help advance our understanding of the underlying issues. To the extent that I did comment on errors, it was to motivate or justify decisions made about experimental stimuli and design that might otherwise raise questions in the reader's mind. Moving forward it will be helpful for researchers to avoid certain tendencies that make unproductive contributions more likely. I'll mention three specifically. First, it is wise to avoid an "either/or" approach, a principal mark of which is to assume that results suggesting the existence of one norm are automatically evidence against the existence of another. Second, existing evidence does not support the assumption that a norm of assertion will impose an exceptionless perfect requirement. Instead the evidence suggests that, like social rules generally, norms of assertion tolerate exceptions. This means that many philosophers' weapon of choice, the counterexample, is ill-suited to advancing understanding. It also means that failing to detect an effect in one study does not erase evidence of the effect found in many other studies. Third, inspiration can legitimately come from many quarters, including previous philosophical debates. But new research on the topic isn't beholden to the unsupported assumptions, speculative objections, and epicyclical refinements of those debates. Rather than searching for the shortest rhetorical path to resurrecting dialectical stalemates of old, a better approach is to begin by honestly reviewing the imperfect but considerable and growing body of evidence relevant to the topic.

Acknowledgments — For helpful feedback and discussion, I thank Peter Brössel, Kevin Reuter, YeounJun Park, Haider Riaz, Angelo Turri, and Sarah Turri. This research was supported by the Canada Research Chairs Program and the Social Sciences and Humanities Research Council of Canada.

References

Beecher, Michael D, S Elizabeth Campbell, John M Burt, Christopher E Hill, and J Cully Nordby. 2000. "Song-type matching between neighbouring song sparrows." *Animal Behaviour* 59 (1): 21–27.

Bradbury, J. W., & Vehrencamp, S. L. (2011). Principles of animal communication (2nd ed.). Sunderland, Mass: Sinauer Associates.

Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80 (1): 27–38.

Heinze, Georg. 2006. "A comparative investigation of methods for logistic regression with separated or nearly separated data." *Statistics in Medicine* 25 (24): 4216–26.

Kneer, Markus. 2018. "The norm of assertion: empirical data." *Cognition* 177: 165–71.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock. 2017. "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences." *Behavioral Research Methods* 49 (2): 1–10.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reuter, Kevin, and Peter Brössel. 2018. "No knowledge required." *Episteme*, 1–19, DOI: 10.1017/epi.2018.10

Turri, John. 2013. "The test of truth: An experimental investigation of the norm of assertion." *Cognition* 129 (2): 279–91.

———. 2015a. "Evidence of factive norms of belief and decision." *Synthese* 192 (12): 4009–30.

———. 2015b. "The radicalism of truth-insensitive epistemology: truth's profound effect on the evaluation of belief." *Philosophy and Phenomenological Research* 93 (2): 348–67.

———. 2016a. "Knowledge and assertion in 'Gettier' cases." *Philosophical Psychology* 29 (5): 759–75.

———. 2016b. *Knowledge and the norm of assertion: an essay in philosophical science*. Cambridge: Open Book Publishers.

———. 2017a. “The distinctive ‘should’ of assertability.” *Philosophical Psychology* 30 (4): 481–89.

———. 2017b. “Experimental work on the norms of assertion.” *Philosophy Compass* 12 (7): e12425.

———. 2018. “Revisiting norms of assertion.” *Cognition* 177 (March): 8–11.

Turri, John, and YeounJun David Park. 2018. “Knowledge and assertion in Korean.” *Cognitive Science* 42 (6): 2060–80.