A NEW PERSPECTIVE CONCERNING EXPERIMENTS ON SEMANTIC INTUITIONS

Justin Sytsma and Jonathan Livengood[1]

Abstract: In two fascinating articles, Machery, Mallon, Nichols, and Stich [2004, forthcoming] use experimental methods to raise a specter of doubt about reliance on intuitions in developing theories of reference which are then deployed in philosophical arguments outside the philosophy of language. Machery et al. ran a cross-cultural survey asking Western and East Asian participants about a famous case from the philosophical literature on reference (Kripke's Gödel example). They interpret their results as indicating that there is significant variation in participants' intuitions about semantic reference for that case. We argue that this interpretation is mistaken. We detail a type of ambiguity found in Machery et al.'s probe but not yet noted in the response literature. We argue that this *epistemic ambiguity* could have affected their results. We do not stop there, however: Rather than rest content with a possibility claim, we ran four studies to test the impact of this ambiguity on participants' responses. We found that this accounts for much of the variation in Machery et al.'s original experiment. We conclude that in the light of our new data, their argument is no longer convincing.

Keywords: Reference; Semantic Intuitions; Experimental Philosophy; Machery, Mallon, Nichols, and Stich; Kripke; Gödel

In two provocative articles, Edouard Machery, Ron Mallon, Shaun Nichols, and Stephen Stich [2004, forthcoming] have argued that there is a problem with the standard methodology for work on reference in modern analytic philosophy. They argue that this methodology attempts to construct theories consistent with our intuitions about prominent fictional and non-fictional cases [2004: B3]. Philosophical method presupposes that intuitions for such cases are sufficiently uniform across the relevant populations of people.[2] Call this presupposition the *uniformity conjecture*. The uniformity conjecture is a testable, empirical claim (once sufficiently specified). However, philosophers have offered no empirical evidence that the conjecture is true.

---

[1] This is a preprint of an article whose final and definitive form will be published in the *Australasian Journal of Philosophy* [2011]; the *Australasian Journal of Philosophy* is available online at: http://www.tandf.co.uk/journals/. This research was supported by the Wesley C. Salmon Fund, University of Pittsburgh. The authors wish to thank Jonah Schupbach, Ron Mallon, Edouard Machery, Max Deutsch, Benny Goldberg, Peter Gildenhuys, Jenny Nado, and an anonymous reviewer for thoughtful comments on previous drafts. We would also like to thank the audiences in Boulder, Kalamazoo, New Brunswick, and Bloomington.

[2] Just what the relevant populations are and what it means for the intuitions within these populations to be sufficiently uniform will be discussed below.

Contrariwise, Machery et al. provide evidence that the uniformity conjecture is false, presenting experimental results which point to significant variation in people's intuitions about one prominent fictional case (Kripke's Gödel example). Specifically, their results suggest that there is significant cross-cultural variation between Westerners and East Asians with regard to this case, as well as significant intra-cultural variation within each group.

In this paper we challenge Machery et al.'s interpretation of their empirical results, arguing that the data do not show what they think they show and thus do not provide convincing evidence against the uniformity conjecture.[3] One could offer two broad types of challenge to Machery et al.'s interpretation: theoretical and empirical. One could provide a purely *theoretical* critique that points out possible confounds in their study. A number of authors have raised objections of this type [Ludwig 2007; Marti 2009; Deutsch 2009]. Instead of following this well-worn path, we offer an empirical critique of Machery et al.'s study. We present data from a series of new experiments that cast doubt on the reliability of the original study.

We detail two ambiguities that can be found in Machery et al.'s study. The first ambiguity centers on the epistemic perspective that is adopted by participants in deciding who is denoted by each of the answer choices given in Machery et al.'s forced-choice test question; this ambiguity has not been articulated in the existing literature. The second ambiguity centers on whether the question is interpreted as asking about the speaker's intentions or not; this ambiguity was noted by Kripke in giving the Gödel example [1972: 85fn36] and has been discussed in the response literature. While our focus is on the first type of ambiguity, we find both to be potentially problematic.

[3] As such, for the purposes of this paper and for the sake of argument, we will simply grant that standard philosophical methodology requires the uniformity conjecture. Further, we also assume with Machery et al. that it is possible for there to be variation in intuitions about the semantic reference of terms between people. See, Devitt [forthcoming] and Jackman [2009] for examples of alternative lines of critique.

Unlike previous critics, however, we do not merely point out a potential problem. Rather than simply arguing that ambiguity *might* have affected participants' responses, we ran a series of experiments to test whether it *actually does*. The results are striking and suggest that ambiguity accounts for much of the variation found in Machery et al.'s study. We conclude that Machery et al.'s instrument is broken: Responses to their Gödel probe do not reliably indicate participants' intuitions about the semantic reference of the term 'Gödel' (what we will refer to as *semantic intuitions*). As such, their results do not provide the basis for a compelling case against the uniformity conjecture.[4]

The structure of the present paper is as follows. In Section 1, we critically evaluate Machery et al.'s argument. In Section 2, we present our interpretation of their results. In Section 3, we present the results of the four experiments that we carried out in order to test our interpretation. Finally, in Section 4, we argue on the basis of these experiments that no convincing case has been made against the uniformity conjecture.

1. Semantics, Cross-cultural Style

It is plausible that one standard philosophical methodology assumes that there is uniformity in our intuitions. These shared intuitions play an evidential role in such philosophical theorizing, constraining our philosophical theories about a given phenomenon.[5] If a theory turns out to be inconsistent with our intuitions about key cases, one has reason to reject it. Machery et al. argue

[4] Note that this is not to argue that the uniformity conjecture is true, nor is it to argue that there is no variation in semantic intuitions across or within cultures. Rather, it is simply to argue that Machery et al.'s current empirical case against the uniformity conjecture fails.

[5] It should be noted that there is ongoing debate about the nature of intuitions, what role they play in philosophical theorizing, and whether the role of intuitions is that of evidence (see, for example, Williamson [2008]). For the sake of argument, we will sidestep these controversies, adopting Machery et al.'s use of 'intuitions' and accepting that such intuitions play a prominent evidential role in philosophical theorizing about reference.

that this is the case in much work on philosophical theories of reference: Our intuitions, and specifically our semantic intuitions, are supposed to serve as evidence.

Unfortunately, appeals to 'our intuitions' are not generally made precise when they appear in philosophical work. With regard to theories of reference, Machery et al. claim that 'our' has two reasonable referents. It could refer to people at large or it could refer to professional philosophers. Machery et al.'s criticism of the second interpretation seems to be most charitably read as a vehement challenge [2004: B9]: They ask for a compelling justification for elevating the intuitions of philosophers over those of other groups and express their strong doubts that such a justification will be forthcoming. We will not pick up the gauntlet here and grant the first interpretation. Given that Machery et al.'s study surveyed non-philosophers, we think that granting that 'our intuitions' typically refers to people at large provides them with the strongest possible argument. We show that their argument is unconvincing even here.

The uniformity conjecture then has two parts. The first part claims that there exists widespread agreement among the intuitive judgments people make with respect to philosophical thought experiments or case studies concerning reference. Universal agreement would be an unreasonably strong requirement. With that exception, we will leave vague just how widespread the agreement needs to be (although we will comment on this in Section 4). The second part of the uniformity conjecture claims that agreement persists if one conditions on membership in theoretically interesting groups. If one found a systematic change in intuitions conditional on gender or age or race, then the uniformity conjecture would be false. Although the restriction to theoretically interesting groups is again vague, as a first-pass we mean to indicate groupings that might reasonably have been predicted to show variation in the relevant intuitions. In particular, it would be unreasonable to require that agreement persist conditional on membership in just any

group that might be constructed, since if there is any variation at all in intuitions, some way of Gerrymandering could be found on which agreement would not persist.

Machery et al.'s work presents challenges to both parts of the uniformity conjecture, though they focus on the second. They make their case with an empirical study showing that there is both intra-cultural and cross-cultural variation in responses to a key intuition pump from the literature on reference. Machery et al. presented two groups of English-speaking undergraduate students—one from Rutgers University and one from the University of Hong Kong—with probes modeled on stories in Kripke's *Naming and Necessity*. All probes were presented in English. Two probes were based on Kripke's Gödel example and two probes were based on his Jonah example. Only the Gödel-style probes produced results calling the uniformity conjecture into doubt. While we agree with Ludwig [2007] and Devitt [forthcoming] in thinking that Machery et al. were too quick to embrace the case that indicates variation across culture and to ignore the case that does not, we again confine ourselves to the more interesting case in order to strengthen Machery et al.'s argument. The two Gödel-style probes differed only in whether Western or Chinese names were used and each participant was given both probes. The Western-name probe reads as follows:

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt", whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel. When John uses the name "Gödel", is he talking about:
>
> (A) the person who really discovered the incompleteness of arithmetic? or
> (B) the person who got hold of the manuscript and claimed credit for the work?

Machery et al. scored answers as either 0 or 1 (0 indicating an (A) answer, 1 a (B) answer), then added the scores for each participant on the two probes, resulting in a scale running from 0 to 2 (where 0 corresponds with 0% (B) answers, 1 with 50% (B) answers, and 2 with 100% (B) answers). They found a mean of 1.13 for Western participants compared to a mean of 0.63 for East Asian participants. We find that this scoring procedure obscures the results. As such, we will compare results in terms of a simple percentage of (B) answers: Aggregating across the two probes, Machery et al. found that 56.5% of Western participants answered (B), compared to 31.5% of East Asian participants.

The difference between the scores for the two groups is statistically significant, and Machery et al. conclude that there is cross-cultural variation in the responses to this probe. There is also significant variation *within* each sample, as Machery et al. note [2004: B8]. As such, if the responses of participants to the probe reliably indicate their semantic intuitions, then Machery et al.'s results are strong evidence against both parts of the uniformity conjecture. In what follows, we argue that the antecedent of this conditional is false. The Gödel probe fails to reliably indicate participants' semantic intuitions because the question it asks is ambiguous.

2. Alternative Interpretation

Machery et al. are well aware that their study, on its own, is not definitive. They write that they "have no illusions that our experiment is the final empirical word on the issue" [2004: B8]. They continue:

> our experiment does not rule out various pragmatic explanations of the findings. Although we found the effect on multiple different versions of the Gödel case, the test question was very similar in all the cases. Perhaps the test question we used triggered different interpretations of the question in the two different groups. [B8]

To this it should be added that ambiguity in the question could also explain the variation found *within* each group. In fact, we think that this is the case. We note two distinct ambiguities that can be found in the test question. First, the question used in Machery et al.'s Gödel probe does not clearly indicate whether the (A) and (B) answer choices are to be read from the *narrator's epistemic perspective* (the narrator relaying information of which John is ignorant) or rather from *John's epistemic perspective* (as the speaker using the name 'Gödel'). This is potentially problematic because if participants read the choices from John's perspective, then their answers would not necessarily correspond with their semantic intuitions. Second, the test question is ambiguous with regard to speaker's reference and semantic reference. This is potentially problematic because if participants understand the question in terms of speaker's reference, then their answers would not necessarily correspond with their semantic intuitions. We discuss each type of ambiguity in turn.

2.1 Epistemic Perspective Ambiguity

Focusing on the Western-name version of Machery et al.'s Gödel probe, the test question asks who John is talking about when he uses the name 'Gödel' and participants are given a forced-choice between the descriptions labeled (A) and (B). One problem with this question is that it does not adequately specify whose epistemic perspective should be adopted in deciding who these descriptions denote, raising the possibility that different participants might associate the same description with different people from the story.

While Machery et al. expect the descriptions to be read from the narrator's perspective, the question might plausibly lead participants to instead adopt John's perspective. Since there is a knowledge asymmetry between John and the narrator, we expect them to give different

descriptions of the man Gödel. Specifically, from the narrator's point of view, 'the person who really discovered the incompleteness of arithmetic' denotes Schmidt and 'the person who got hold of the manuscript and claimed credit for the work' denotes Gödel; but, from John's perspective 'the person who really discovered the incompleteness of arithmetic' denotes Gödel.[6] Indeed, John has never heard of Schmidt![7] This asymmetry between John's knowledge and the narrator's knowledge means that it is possible for a participant to think that when John uses the name 'Gödel' he is talking about Gödel and nonetheless for that participant to legitimately answer (A) because she reads that description from John's perspective as denoting Gödel.

Note that Machery et al.'s goal in giving the Gödel probe is to ascertain people's intuitions about the semantic reference of the name 'Gödel'; as such, the issue is neither the specific descriptions given nor the person a given participant thinks they best apply to. Rather, the intuition that Machery et al.'s probe is supposed to test is whether untutored people follow Kripke in thinking that the name 'Gödel' as used by John refers to Gödel, even though John associates a description with the name that best fits somebody else (Schmidt). If a reasonable proportion of participants answered (A), not because they held different semantic intuitions than Kripke, but because they adopted John's epistemic perspective instead of the narrator's in deciding who the descriptions denote, then Machery et al.'s results would not be relevant to the uniformity conjecture.

[6] As Edouard Machery has pointed out [personal communication], the occurrence of the adverb 'really' in the (A) description emphasizes the narrator's perspective, not John's, since it indicates a contrast between the real discoverer and an implied imposter. This point is well taken. We agree that 'really' serves to emphasize the narrator's perspective, but we find that it is insufficient emphasis. We suggest that naïve participants are strongly disposed to read the (A) description from John's perspective, taking it to denote Gödel. In fact, in pilot testing of the three probes discussed in Section 3.1, we removed the word 'really'. We found a strong preference for (A) on these probes. For the original probe we found 9.1% (B) answers (N=11), for the John's perspective probe we found 0% (B) answers (N=12), and for the narrator's perspective probe we found 15.4% (B) answers (N=13). In contrast, for Machery et al.'s original wording of the Gödel probe question, we found 41.2% (B) answers (N=17).

[7] Throughout we will use Schmidt (without quotes) to talk about the man in the story who (unbeknownst to John) really discovered the incompleteness of arithmetic and Gödel (without quotes) to talk about the man in the story who is widely believed to have discovered the incompleteness of arithmetic, but actually got hold of the manuscript and claimed credit for the work.

2.2 Speaker's Reference Ambiguity

The epistemic ambiguity noted above does not depend on how the participant assesses John's intentions in using the name 'Gödel'. In fact, the participant need not even think about John's intentions. If participants were to focus on who John intended to be talking about in answering the test question, however, this would also be problematic for Machery et al.'s use of the Gödel probe. We hold that this is a second ambiguity that might have affected the results.

Thus, one plausible reading of the test question in Machery et al.'s Gödel probe is as asking who John *intends* to be talking about in using the name 'Gödel'. Read in this way, to answer the question a participant would first need to imagine what John's intentions might be. This is likely to depend on the type of 'Gödel' statement that the participant considers. For some such statements, the most charitable reading would be to treat John as intending to talk about Schmidt ('Gödel is a mathematical genius'); but, for others the most charitable reading would be to treat John as intending to talk about Gödel ('Gödel has probably won many awards for the incompleteness proof'). For a statement of the first sort, the *speaker's reference* of the name 'Gödel' diverges from its *semantic reference*. Assuming that 'Gödel' actually refers to Gödel (semantic reference), John might nonetheless be taken to intend to be talking about Schmidt (speaker's reference). A participant might therefore answer (A) despite sharing Kripke's intuitions about the semantic reference of the name 'Gödel'; she does so because she thinks that John intends to be talking about Schmidt.

Unlike the epistemic ambiguity detailed in Section 2.1, the speaker's reference ambiguity in the Gödel example has been previously noted. In fact, Kripke suggests this ambiguity in giving the Gödel example in *Naming and Necessity* [1972: 85fn36] and brings up the Gödel example in articulating the distinction between speaker's reference and semantic reference

[1977: 261]. And not surprisingly, the ambiguity has been brought up in responses to Machery et al.'s empirical work, notably by Kirk Ludwig [2007] and Max Deutsch [2009]. The idea is straightforward. Some participants might have reasoned, 'John intends to be talking about the person who discovered the incompleteness of arithmetic and hence intends to be talking about Schmidt'. If such participants also read description (A) from the narrator's epistemic perspective as denoting Schmidt, then they should answer (A), not because they hold intuitions about the semantic reference of 'Gödel' that differ from Kripke's but because they are answering with regard to speaker's reference. In this way, while it is possible for the speaker's reference ambiguity to affect participants' responses, for it to do so they must first navigate the epistemic ambiguity noted above. As such, we will focus on the epistemic ambiguity in the studies that follow; nonetheless, our response to Machery et al. does not depend on there being only one type of ambiguity at work. Further, as our goal is simply to test whether Machery et al.'s results reflect ambiguity in the test question, we will not attempt to distinguish between these two types of ambiguity experimentally. It is therefore possible that in attempting to clarify the epistemic ambiguity in the studies discussed in Section 3, we also clarified the speaker's reference ambiguity.

3. Four Studies

In the previous section we argued that epistemic ambiguity *might have* impacted Machery et al.'s results. Fortunately, we can control for this ambiguity and experimentally test its *actual* impact. This is what we did in the studies described below. We found that the epistemic ambiguity does indeed have a significant impact on how participants responded to Machery et al.'s probe.

3.1 Study 1: Original, John's Perspective, and Narrator's Perspective

To test the impact of the epistemic ambiguity on participants' responses to Machery et al.'s Gödel probe, we developed two variations on their original question—one emphasizing John's perspective and one emphasizing the narrator's perspective. We hypothesized that: (1) clarifying the question to encourage taking John's perspective would result in a *lower* percentage of (B) answers than found for the original question; and, (2) clarifying the question to encourage taking the narrator's perspective would result in a *higher* percentage of (B) answers than found for the original question. The vignette was the same for each probe, using the same text as that given by Machery et al. on the Western-name version of the Gödel probe (see Section 1); only the test questions differed. The three variations read as follows:

> **Original:** When John uses the name "Gödel," is he talking about: (A) the person who really discovered the incompleteness of arithmetic? Or, (B) the person who got hold of the manuscript and claimed credit for the work?
>
> **John's Perspective:** When John uses the name "Gödel," does John think he is talking about: (A) the person who the story says really discovered the incompleteness of arithmetic? Or, (B) the person who the story says got hold of the manuscript and claimed credit for the work?
>
> **Narrator's Perspective:** When John uses the name "Gödel," is he actually talking about: (A) the person who the story says really discovered the incompleteness of arithmetic? Or, (B) the person who the story says got hold of the manuscript and claimed credit for the work?

The study was run in a classroom setting on undergraduates at the University of Pittsburgh. We used a between subjects design with each participant assigned to just one of the three conditions. While the descriptions given for the (A) and (B) answers were counter-balanced for order, we will report answers as (A) or (B) in terms of the ordering presented above. After answering, participants were asked to explain their answers and to fill in a brief biographical questionnaire. As our goal was to test the intuitions of 'Western' participants, 10 participants were excluded

because English was not their native language.[8] The remaining 189 participants were 41.3% female, with an average age of 21.0, and ranging in age from 18 to 44.

We found that on the original probe, 39.4% of participants answered (B); on the John's perspective probe, 22.0% of participants answered (B); and on the narrator's perspective probe, 57.4% of participants answered (B). The results are shown graphically in Figure 1. As predicted, in comparison to the original probe, the percentage of (B) answers was *significantly lower* for the John's perspective probe (with a joint p-value of 0.044) and *significantly higher* on the narrator's perspective probe (with a joint p-value of 0.034).[9]

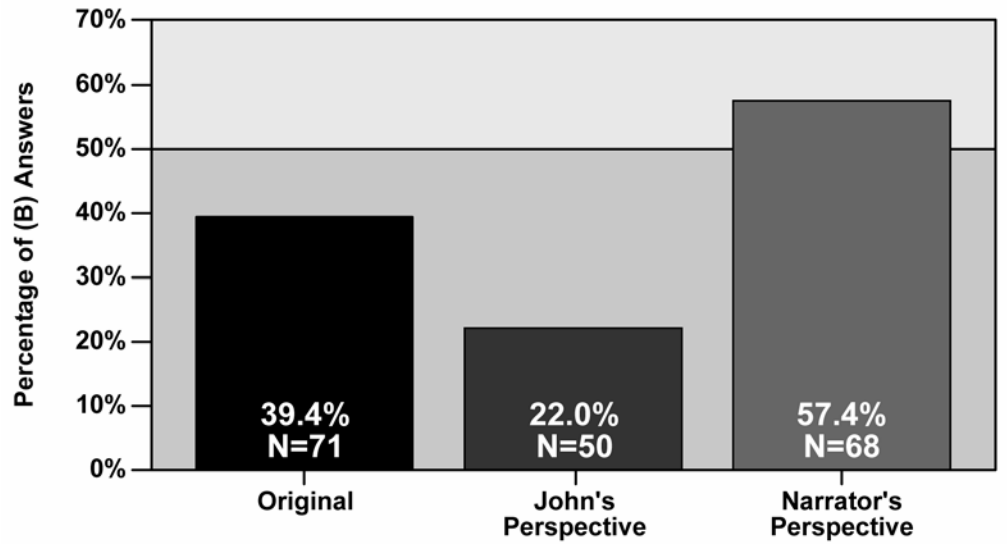

Figure 1: Study 1, Results.

[8] In keeping with Machery et al.'s work, we also asked participants about their ethnic background, whether they were born in the United States, how many of their parents/grandparents were born in the United States, and so on. While a more or less restrictive criterion for Western status could be applied, here, we were unsure how to draw the dividing lines within a population with diverse ethnic backgrounds and immigration histories. While the category of native English-speakers is clearly not equivalent with Western, it has the advantage of being a simple criterion that includes American students of African and Middle Eastern descent, for example, while minimizing concerns about language fluency. It is worth noting that if we instead used the criterion of simply excluding participants of East Asian ethnic background, the results are virtually identical: For the original probe we get 39.7% (instead of 39.4%); for the John's perspective probe we get 21.6% (instead of 22.0%); and, for the narrator's perspective probe there is no difference at all.

[9] We computed z-scores to determine whether the original probe differed from the John's perspective or narrator's perspective probes. We jointly tested the hypotheses that the John's perspective probe is identical to the original probe and that the original probe is identical to the narrator's perspective probe against the alternatives that the John's perspective probe is less than the original probe and that the original probe is less than the narrator's perspective probe. We determined joint p-values for these tests using a Bonferroni correction.

Surprisingly, we failed to replicate Machery et al.'s findings for the original probe. While they found that 56.5% of the 31 Western undergraduates they surveyed answered (B), only 39.4% of the 71 Western undergraduates that we surveyed answered (B). Our results are actually closer to the results Machery et al. reported for the 41 East Asian undergraduates they surveyed (31.5% of whom answered (B)). In fact, there is no statistically significant difference between our sample of Westerners and Machery et al.'s sample of East Asians (a z-test comparing the two proportions yields a p-value of 0.380). We discuss this further in Section 4.

While the results of our first study support our hypothesis, we nonetheless found a fairly high degree of variation on the narrator's perspective probe (with 42.6% of participants answering (A) and 57.4% answering (B)). It might be argued that this variation is itself sufficient to call the uniformity conjecture into doubt. Again, the test question for this probe was designed to emphasize the narrator's perspective and it is therefore reasonable to conclude that the (A) answers indicate that those participants have descriptivist semantic intuitions about the Gödel case. However, in reviewing the explanations given by those participants who answered (A) on the narrator's perspective probe, we found that only 8 out of 29 (27.6%) expressed what we found to be straight-forwardly descriptivist intuitions. Most participants indicated that they were specifically considering John's perspective, giving an explanation that was compatible with having causal-historical semantic intuitions about the case. Of course, it is possible that experimenter bias was affecting our assessment of the participants' explanations. To control for this, we employed two graduate students in the Department of History and Philosophy of Science at the University of Pittsburgh to independently code the explanations.

Each of the two coders was familiar with Kripke's *Naming and Necessity*, including the distinction between the descriptivist view and the causal-historical view. They were given a copy

of the narrator's perspective probe and a transcript of the explanations given for (A) answers. They were instructed to code the explanations (1) as indicating only descriptivist intuitions, (2) as focusing only on John's perspective, (3) as including indications of both descriptivist intuitions and a focus on John's perspective, (4) as simply being unclear, or (5) as other. There was 75.9% agreement pair-wise between the two codings. The first coder found that six of the explanations expressed descriptivist intuitions with an additional four explanations being coded as both; the second coder found that nine of the explanations expressed descriptivist intuitions with an additional four explanations being coded as both. This suggests that a significant proportion of the variation found for the narrator's perspective probe reflects residual ambiguity, spurring us to develop a second version of the probe that further emphasizes the narrator's perspective.

3.2 Study 2: Clarified Narrator's Perspective

As in the first experiment, we used Machery et al.'s Western-name vignette. Again, only the test question was changed. In this variation, participants were asked:

> **Clarified Narrator's Perspective:** Having read the above story and accepting that it is true, when John uses the name "Gödel," would you take him to actually be talking about: (A) the person who (unbeknownst to John) really discovered the incompleteness of arithmetic? Or, (B) the person who is widely believed to have discovered the incompleteness of arithmetic, but actually got hold of the manuscript and claimed credit for the work?

Unlike the first experiment, this study was run online and was open to philosophers as well as non-philosophers. Each participant was given the clarified narrator's perspective probe. Again, the descriptions given for (A) and (B) were counter-balanced for order. Participants were given the same biographical questionnaire, but were not asked to explain their answer. In keeping with the first experiment, 49 participants were excluded because English was not their native

language. An additional 10 were excluded because they did not indicate their native language or because they had previously taken the survey. Of the remaining 142 participants, 58 were classified as philosophers (21.1% female; average age, 28.2; age range, 20-64) and 84 as non-philosophers (51.9% female; average age, 33.1; age range, 18-61).[10]

We found that 73.8% of non-philosophers answered (B), compared to 75.9% of philosophers. The difference was not statistically significant (a z-test comparing the two proportions yields a p-value of 0.851). To determine whether the particular probe seen by a given participant is a statistically significant predictor of the answer given by that participant, we carried out a logistic regression using the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \varepsilon$$

where π is the probability that an arbitrary participant answers (B) on the probe and the I_i are dummy variables indicating which probe the participant saw.[11] A likelihood ratio test rejects the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ with a p-value less than 0.0001 (chi-square test statistic of 40.96 with three degrees of freedom).[12]

[10] Participants were classified as philosophers if they were a professor of philosophy, had completed (or were in the process of completing) a graduate degree in philosophy, or had completed (or were in the process of completing) an undergraduate major in philosophy.

[11] For example, the variable I_1 equals 1 if the participant saw the John's perspective probe, and it equals 0 otherwise. If a participant saw the original probe, then all of the I_i variables are equal to zero. The response variable being measured was the participant's choice of either the (A) description or the (B) description. Thus, the response variable is binary and can be thought of as equal to 1 or to 0. We treated (B) as equal to one. Logistic regression allows one to infer the probability that a binary response variable is equal to one given the values of the covariates, the I_i. If one wants to know whether seeing one probe rather than another makes a difference to the answer given by a participant, one tests whether the coefficients on the dummy variables are significantly different from zero. For more detailed discussion of logistic regression, see Chapter 14 of Kutner *et al*. [2005].

[12] Point estimates for the coefficients were $b_0 = -0.4290$, $b_1 = -0.8367$, $b_2 = 0.7253$, and $b_3 = 1.4651$ with 95% confidence intervals (-0.9153, 0.0412), (-1.6889, -0.0355), (0.0541, 1.4101), and (0.7963, 2.1610), respectively. One might object that from the point of view of Machery et al., the John's Perspective probe asks a different question than the other three probes, so the test described is unfair. However, using a more generous model that does not include the John's Perspective probe, namely $\text{logit}(\pi) = \beta_0 + \beta_2 I_2 + \beta_3 I_3 + \varepsilon$, the hypothesis that $\beta_2 = \beta_3 = 0$ is still rejected with a p-value less than 0.0001 (this time with a test statistic equal to 18.99 on two degrees of freedom).

3.3 Study 3: Within-subjects Design

We followed up on our first two experiments using a within-subjects design. Each participant was given all four of the Gödel probes discussed above, with two filler probes in between each Gödel probe. Unlike either of our first two studies, this study was administered face-to-face but not in a classroom setting, with participants solicited at two Pittsburgh-area cafes. Again, the descriptions given for (A) and (B) on each of the four probes were counter-balanced for order; participants were not asked to explain their answers but were asked to fill out a short biographical questionnaire. Two participants were excluded because English was not their native language (a third was excluded because he did not select answers to the Gödel probes). The average responses of the remaining 35 participants were 42.9% for the original probe, 31.4% for the John's perspective probe, 57.1% for the narrator's perspective probe, and 74.3% for the clarified narrator's perspective probe (51.4% female; average age, 41.5; age range, 21-69). The results are shown next to the responses from Studies 1 and 2 in Figure 2.

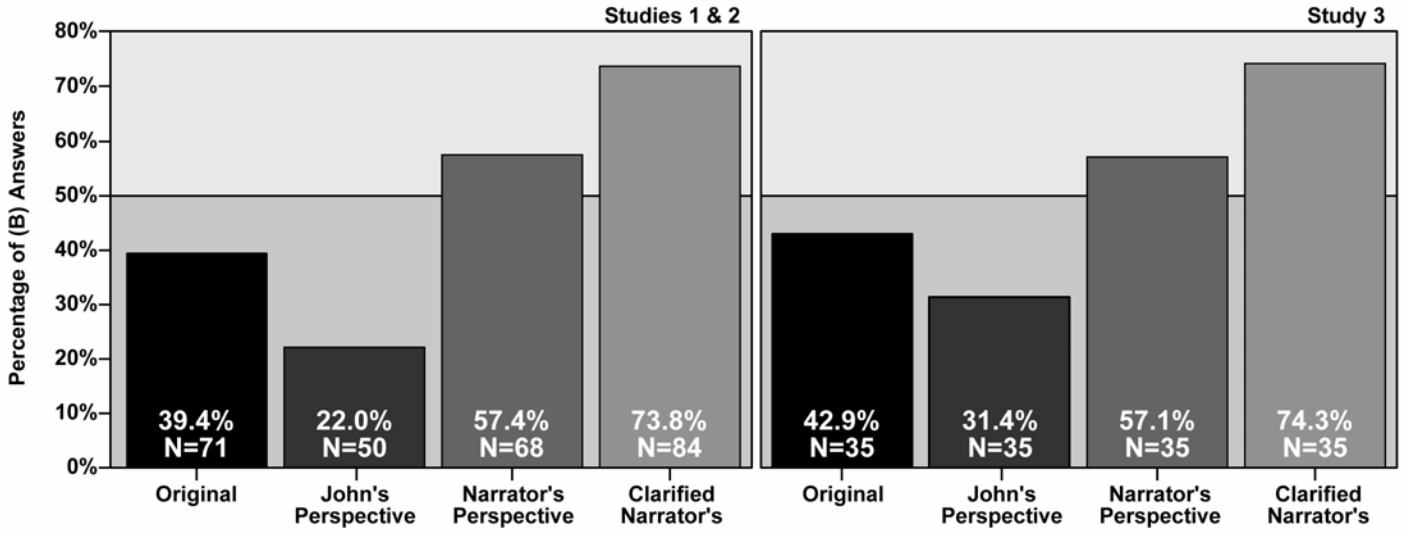

Figure 2: Results, Studies 1 and 2 (combined) on the left; Study 3 on the right.

When a logistic regression is fitted to the Study 3 data as was done for Studies 1 and 2 (using the same model with the same interpretation), a likelihood ratio test rejects the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ with a p-value of 0.0019 (chi-square test statistic of 14.88 with three degrees of freedom).[13] In other words, we again see clear evidence that the original Gödel probe is ambiguous. To test this further, in our final study we asked participants how they understood the original question.

3.4 Study 4: Which question?

As in the first experiment, this study was administered in a classroom setting at the University of Pittsburgh. Participants were first given Machery et al.'s original Gödel probe. On a second page, they were then asked to indicate which of the following two restatements best corresponded with their understanding of the test question:

> (1) When John uses the name "Gödel," does *John think* he is talking about: (A) or (B).
> Or, (2) When John uses the name "Gödel," is he *actually* talking about: (A) or (B).

The descriptions given for (A) and (B), as well as the restatements given for (1) and (2), were counter-balanced for order. Four participants were excluded because English was not their native language; three more were excluded because they did not clearly answer the second question. The remaining 73 participants were 42.5% female, with an average age of 21.6, and ranging in age from 18 to 43.

We found that 29 of the 73 participants answered (B) on the original Gödel probe (39.7%). Of those, only 11 answered (1) on the follow-up question (37.9%), the clear majority indicating that they understood the original question in line with the John's perspective version

[13] Point estimates for the betas were $b_0 = -0.2877$, $b_1 = -0.4925$, $b_2 = 0.5754$, and $b_3 = 1.3486$ with 95% confidence intervals (-0.9742, 0.3773), (-1.4890, 0.4801), (-0.3651, 1.5373), and (0.3618, 2.3977), respectively. Again, the more generous model rejects the hypothesis that $\beta_2 = \beta_3 = 0$ a p-value of 0.0262 (this time with a test statistic equal to 7.29 on two degrees of freedom).

of the question used in Study 1. Of the 44 participants who answered (A) on the original question, however, 33 answered (1) on the follow-up question (75.0%). The distributions are shown graphically in Figure 3.

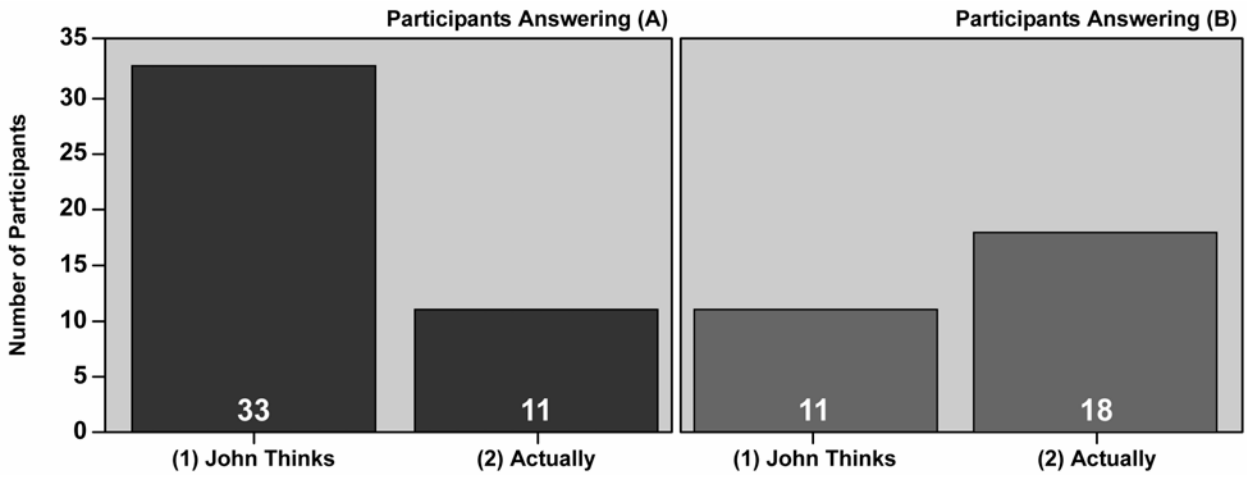

Figure 3: Study 4, results.

We expected responses on the two probes to be associated. To test this, we carried out a logistic regression using the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 X + \varepsilon$$

where π is the probability that a participant answers (A) on the Original Gödel probe and X is equal to one if the participant answered (2) on the follow-up probe (indicating the participant thought the question was asking who John was *actually* talking about) or zero if the participant answered (1) on that probe (indicating that the participant thought the question was asking who John *thought* he was talking about). A likelihood ratio test rejects the hypothesis that $\beta_1 = 0$ with a p-value of 0.0015 (chi-square test statistic of 10.11 with one degree of freedom). Responses on the two probes are associated and the conditional probability that a participant answered (A) on the original Gödel probe given that the participant answered (1) on the follow-up probe is 0.75. Our point estimates for β_0 and β_1 (1.0986 and -1.5911, respectively) fit very well with our

expectations. This is additional evidence that Machery et al.'s original Gödel probe is ambiguous in a way that undermines their results. We doubt that Machery et al. could explain the results of these four studies without giving up the claim that their original probe successfully tracks semantic intuitions (at least among Westerners).

4. Defending the Uniformity Conjecture

Machery et al.'s results provide *prima facie* evidence that semantic intuitions vary both across and within cultures. Their results cast doubt on the uniformity conjecture and thus on the standard philosophical methodology for developing theories of reference. However, the results of our studies discussed in Section 3 call this conclusion into question. We have presented strong evidence that Machery et al.'s original Gödel probe is ambiguous and that this ambiguity affects the answers given by Western participants. This raises significant doubt that Machery et al.'s results reflect participants' semantic intuitions. Since assessing semantic intuitions is the purpose of their Gödel probe, we conclude that their instrument is broken. As such, Machery et al.'s data is best interpreted as irrelevant to assessing the uniformity conjecture.

While our data is specific to Westerners, having found that Machery et al.'s instrument is broken our results undermine each version of the argument against the uniformity conjecture that Machery et al. suggest. In this section, we argue that our results (1) undermine Machery et al.'s claim to have shown significant variation in semantic intuitions *within* Westerners, (2) undermine their claim to have shown significant variation in semantic intuitions *cross-culturally* between Westerners and East Asians, and (3) undermine their claim to have shown significant variation in semantic intuitions *within* East Asians.

4.1 Variation within Westerners

Our findings are most directly at odds with the claim that there is significant variation in semantic intuitions within Westerners. Western participants find Machery et al.'s original test question to be ambiguous. Specifically, alterations to the question that should be irrelevant to participants' semantic intuitions significantly alter their responses: Clarifying the question to emphasize the narrator's perspective produced a dramatic increase in (B) answers. This indicates that responses to the original test question do not reliably measure semantic intuitions. In other words, the instrument is broken and Machery et al.'s results do not constitute compelling evidence of significant variation in semantic intuitions within Westerners.

That Machery et al.'s instrument fails to reliably indicate participants' semantic intuitions is what we set out to show and we think we have been successful. Nonetheless, it might be argued that our clarified instrument does not have this problem and that it still shows some variation in responses for Westerners. Specifically, on the clarified narrator's perspective probe—the one most likely to elicit participants' semantic intuitions—we found in one study (Section 3.2) that 73.8% of Western participants answered (B); in a second study (Section 3.3) that number went up to 74.3%. It could be argued that this variation is still sufficient to threaten the uniformity conjecture. While one could make this argument, we do not find it compelling.

To begin, note that the numbers we found for non-philosophers on the clarified narrator's probe are very close to the percentage of (B) answers given by philosophers on the same probe (75.9%; see Section 3.2). We feel that this sets a reasonable bound on what counts as 'sufficient uniformity' for the uniformity conjecture. After all, as Machery et al. write, 'it has turned out that almost all philosophers share the intuitions elicited by Kripke's fictional cases, including most of

his opponents' [2004: B3]. So, either this sentiment should be re-evaluated, or the folk should be understood to have uniform semantic intuitions in this case as well.

Setting aside the comparison to philosophers, we can still ask: Is the agreement we found for Westerners really sufficient to avoid a renewed attack on the uniformity conjecture? We are uncomfortable setting a firm bound on consensus for these kinds of study. Several reasons present themselves. First, we do not know if there is a ceiling to our ability to clarify Machery et al.'s original probe. We thought our clarification for our first experiment would be sufficient. We were wrong. Fortunately, we were able to better clarify the question. It might be possible to do better still (possibly by more directly controlling for the speaker's reference ambiguity discussed in Section 2.2). Second, studies of this kind are rather noisy affairs and we should expect some variation, regardless of how clear the question or how obvious the answer. This is especially true given the difficulty of the Gödel example, not to mention its fanciful nature. Third, we simply do not know how much agreement is enough or even whether the same level of consensus should be required for each philosophical question we wish to scrutinize. If the result of our coding of participants' explanations in Section 3.1 is accurate, we would expect roughly 10%–20% of Westerners to have descriptivist intuitions about the Gödel case, and this is confirmed by our fourth study. Is this level of agreement enough to satisfy the uniformity conjecture? In our opinion, 80%–90% agreement amongst Westerners about the Gödel case would be enough to establish a clear consensus. In such an event, it would seem reasonable to flag this consensus by talking about 'our intuitions'.

The more important issue, however, is not whether 80%–90% agreement is sufficient for the uniformity conjecture, but whether the existence of a dissenting minority (whatever its size) fuels an argument against using the majority intuition as evidence. We do not feel that this level

of variation, on its own, casts significant doubt on the reliability of the intuitions of the majority of Westerners (including the majority of Western philosophers). Making that case requires a further argument that Machery et al. do not provide.

4.2 Variation between Westerners and East Asians

Machery et al.'s principal argument against the uniformity conjecture is not based on *intra*-cultural variation but on showing *cross*-cultural variation in responses to the Gödel probe. Since the studies we discussed in Section 3 were all conducted on Westerners, it may be less clear how they bear on Machery et al.'s primary argument against the uniformity conjecture. How could data on just Westerners undermine a finding of cross-cultural variation between Westerners and East Asians? Our data undermine Machery et al.'s finding of cross-cultural variation in three ways.

First, in showing that the responses of Westerners to the original Gödel probe do not reliably indicate their semantic intuitions, we make knowing how Westerners responded to the Gödel probe irrelevant to comparing their semantic intuitions to those of East Asians. In showing that Machery et al.'s instrument is broken for Westerners, we have undermined the basis for comparison to East Asians.

Second, our failure to replicate Machery et al.'s results for Westerners—more precisely, the fact that our numbers for Westerners are so similar to their numbers for East Asians—casts doubt on whether there is cross-cultural variation for Gödel probe *responses*, let alone for semantic intuitions. Responses from Westerners we sampled were not significantly different from the responses of East Asians Machery et al. sampled—39.4% (B) answers in our study, compared to 31.5% (B) answers in their sample of East Asians. Given our failure to replicate

their result, it is unclear what numbers we should use for the cross-cultural comparison. If we assume that Machery et al.'s result for East Asians is accurate, and compare it with our result for Westerners, then there is little reason to believe that there is any significant difference in semantic intuitions about this case between Westerners and East Asians. Further, our result for Westerners was closely replicated in our third and fourth study (42.9% and 39.7% of these participants answering (B) respectively). Hence, it is at best unclear whether Machery et al.'s data pose a serious cross-cultural challenge to the uniformity conjecture.

Third, in showing that the Gödel probe does not reliably indicate the semantic intuitions of Westerners, we have indirectly cast doubt on its ability to track the semantic intuitions of East Asians. If an instrument fails to work in one context where it was supposed to work, we should doubt (or suspend belief) that it works in other contexts (at least until we have confirming evidence that it does work). Of course, it is *possible* that the epistemic ambiguity is only an issue for Westerners; but, this seems implausible and attempting to challenge the uniformity conjecture in this way without first providing new experimental evidence that favors this interpretation would be hopelessly ad hoc. Having seen how simple clarifications of the test question change the responses of Western participants, it would be naïve to rely on the original probe to disclose participants' semantic intuitions, regardless of their cultural background.

Despite these objections, one might argue that our results suggest that there is *some* cross-cultural variation and that it would be ad hoc to explain this difference by appealing to the epistemic ambiguity. Such an appeal would not be ad hoc. Before seeing any data, there was just as much reason to predict that East Asians would be more likely than Westerners to answer (A) on the basis of the epistemic ambiguity as there was for Machery et al. to predict a difference on the basis of different semantic intuitions. Interestingly, this claim follows from the very same

body of empirical work that Machery et al. point to in framing their prediction that Westerners and East Asians would differ in their intuitions about the Gödel probe. Machery et al. based their prediction on recent work in cultural psychology indicating that there are some systematic cognitive differences between East Asians and Westerners. In particular, they call on the work of Nisbett and colleagues [2001] indicating a range of cultural differences, which they collect under the heading of 'holistic vs. analytic thought'. Nisbett et al. define holistic thought as 'involving an orientation to the context or field as a whole, including attention to relationships between a focal object and the field, and a preference for explaining and predicting events on the basis of such relationships' [293]. Holistic thought is supposed to be characteristic of Easterners. They define analytic thought as 'involving detachment of the object from its context, a tendency to focus on attributes of the object to assign it to categories, and a preference for using rules about the categories to explain and predict the object's behavior' [293]. Analytic thought is supposed to be characteristic of Westerners. As an instance of the difference between East and West, Nisbett et al. point to differences in their social attitudes. They write, 'China and other East Asian societies remain collectivist and oriented toward the group, whereas America and other European-influenced societies are more individualist in orientation' [295].

Thus, the more holistic way to read Machery et al.'s Gödel probe is in terms of the beliefs that would be ascribed to John by his interlocutors. The more analytic way to read the probe is in terms of the beliefs that we the readers have as informed by an omniscient narrator. The analytic approach, which is supposed to be predominant among Westerners, would read the question as being concerned with this more detached perspective. So far, the data do not support the claim that East Asians are more likely than Westerners to answer (A) on the original Gödel probe;

however, if further studies show that they are, appeal to the epistemic ambiguity to explain this difference cannot be dismissed as ad hoc.

4.3 Variation within East Asians

Machery et al. also found significant variation among East Asians and this could fuel a third challenge to the uniformity conjecture. For the reasons discussed in Section 4.1, however, we are not sure that the result of 31.5% (B) answers in their sample is sufficient to make a strong case against the uniformity conjecture. Regardless, as argued in Section 4.2, our studies support the conclusion that the original Gödel probe is not a reliable instrument for assessing semantic intuitions. Moreover, as we articulated in Section 4.2, it would be ad hoc to argue that East Asians would not change their responses were the test question appropriately clarified. This casts doubt on the finding of significant variation in semantic intuitions among East Asians.

We conclude that in the light of the studies reported in Section 3, Machery et al.'s original study fails to undermine the uniformity conjecture.

5. Conclusion

In considering Machery et al.'s empirical work on variation in semantic intuitions, we were rather suspicious of their result. We found that the question they asked participants was unclear with regards to whose epistemic perspective it was supposed to be answered from. Since the Gödel story is designed to generate a divergence between the descriptions that the speaker (John) and the narrator of the story associate with the name 'Gödel', we felt that such an ambiguity could have contributed to the large cross-cultural and intra-cultural variation that Machery et al. found. This is important because it is that variation that drives their argument against the

philosophical practice of using intuitions about such cases as evidence in developing theories of reference.

We sought to test our suspicions experimentally. By testing participants' responses to a series of clarifications to the original probe, we were able to see what impact epistemic ambiguity had on their responses. The impact was significant. What we found is that when we emphasize John's perspective, Western participants are significantly more likely to answer (A) and when we emphasize the narrator's perspective they are significantly more likely to answer (B). Upon further clarification of the narrator's perspective probe, we found that roughly three out of four Western participants (whether philosophers or non-philosophers) answered (B). We hold that this is insufficient variation to directly fuel a compelling argument against the practice of using our intuitions as evidence. Our results cast doubt on the reliability of Machery et al.'s probe for getting at participants' semantic intuitions *regardless of their cultural background* and they cast doubt on Machery et al.'s finding of significant cross-cultural variation between Westerners and East Asians.

REFERENCES

Deutsch, Max 2009. Experimental Philosophy and the Theory of Reference, *Mind & Language* 24/4: 445–466.

Devitt, Michael forthcoming. Experimental Semantics, *Philosophy and Phenomenological Research*.

Devitt, Michael and Kim Sterelny 1999. *Language and Reality: An Introduction to the Philosophy of Language, Second Edition*, Oxford: Blackwell.

Jackman, Henry 2009. Semantic Intuitions, Conceptual Analysis, and Cross-Cultural Variation, *Philosophical Studies* 146/2: 159–177.

Kripke, Saul 1972. *Naming and Necessity*, Oxford: Blackwell.

Kripke, Saul 1977. Speaker's Reference and Semantic Reference, in *Midwest Studies in Philosophy vol. II: Studies in the Philosophy of Language*, ed. P. A. French, T. E. Uehling, Jr., and H. K. Wettstein, Morris, MN: University of Minnesota: 255–276.

Kutner, M., C. Nachtsheim, J. Neter, and W. Li 2005. *Applied Linear Statistical Models, Fifth Edition*, New York: McGraw Hill.

Ludwig, Kirk 2007. The Epistemology of Thought Experiments: First Person versus Third Person Approaches, *Midwest Studies in Philosophy* 31/1: 128–159.

Machery, E., R. Mallon, S. Nichols, and S. Stich 2004. Semantics, Cross-cultural Style, *Cognition* 92/3: B1–B12.

Mallon, R., E. Machery, S. Nichols, and S. Stich forthcoming. Against Arguments from Reference, *Philosophy & Phenomenological Research*.

Marti, Genoveva 2009. Against Semantic Multiculturalism, *Analysis* 69/1: 42–48.

Nisbett, R., K. Peng, I. Choi, and A. Norenzayan 2001. Culture and Systems of Thought: Holistic Versus Analytic Cognition, *Psychological Review* 108/2: 291–310.

Williamson, Timothy 2008. *Philosophy of Philosophy*, Oxford: Blackwell.