COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1552–1585

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12609

The Reference of Proper Names: Testing Usage and Intuitions

Michael Devitt, Nicolas Porot

Philosophy Program, Graduate Center, City University of New York

Received 22 July 2017; received in revised form 19 January 2018; accepted 16 February 2018

Abstract

Experiments on theories of reference have mostly tested referential intuitions. We think that experiments should rather be testing linguistic usage. Substantive Aim (I): to test classical description theories of proper names against usage by "elicited production." Our results count decisively against those theories. Methodological Aim (I): Machery, Olivola, and de Blanc (2009) claim that truth-value judgment experiments test usage. Martí (2012) disagrees. We argue that Machery et al. are right and offer some results that are consistent with that conclusion. Substantive Aim (II): Machery et al. provide evidence that the usage of a name varies, being sometimes descriptive, sometimes not. In seven out of eight tests of usage, we did not replicate this variation. Methodological Aim (II): to test the reliability of referential intuitions by comparing them with linguistic usage. Earlier studies led us to predict that we would find those intuitions unreliable, but we did not. Our results add to evidence that tests of referential intuition are susceptible to unpredictable wording effects.

Keywords: Reference; Referential intuition; Linguistic usage; Truth-value judgment; Proper name; Description theory

1. Introduction

1.1. Background and aims

It is usual to think that a term's referential relation to reality is the core of its meaning. In the case of a proper name the relation in question is to a certain object. This raises the fundamental question: *In virtue of what* does the proper name refer to that object? A theory of reference attempts to answer this question. Until the 1970s, all popular answers

Correspondence should be sent to Nicolas Porot, Philosophy Program, Graduate Center, City University of New York, 365 Fifth Ave., Room. 7113, New York, NY 10016. E-mail: nporot@gradcenter.cuny.edu

were description theories of one sort or another. Thus, according to the "classical" Frege–Russell description theory, the reference of a name is determined by a description associated with it by competent speakers; for example, in the case of "Gödel," by the description, "the person who proved the incompleteness of arithmetic"; in the case of "Jonah," by one capturing the high points of the biblical story of Jonah. Then came the revolution in the theory of reference led by Saul Kripke (1980). Kripke made seemingly devastating criticisms of a variety of description theories, including the classical one. Thus, in response to the classical theory of "Gödel," he imagined the following scenario:

> Suppose that Gödel was not in fact the author of the theorem. A man named "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 it was thereafter attributed to Gödel. ... So, [on the view in question] since the man who discovered the incompleteness of arithmetic is in fact Schmidt, we, when we talk about "Gödel," are in fact always referring to Schmidt. But it seems to me we are not. We simply are not. (1980, pp. 83–84)

And, in response to the classical theory of "Jonah," Kripke had this to say:

> Suppose that someone says that no prophet ever was swallowed by a big fish or a whale. Does it follow, on that basis, that Jonah did not exist? ... while Biblical scholars generally hold that Jonah did exist, the account not only of his being swallowed by a big fish but even going to Nineveh to preach or anything else that is said in the Biblical story is assumed to be substantially false. (1980, p. 67)

Yet, according to the description theory, the falsity of the biblical story *entails* that Jonah did not exist. Kripke completed his critique of description theories by proposing an alternative "better picture" (p. 94), according to which the reference of a name is determined by a network of causal chains stretching back to the name's introduction. Theories of this sort are called "causal-historical."

How do philosophers of language tell which theory of reference is right? Edouard Machery, Ron Mallon, Shaun Nichols, and Stephen Stich ("MMNS") looked critically at this *modus operandi* (2004). They noted that theories of reference are tested by consulting referential *intuitions* and that those intuitions are nearly always those of the philosophers themselves "in their armchairs"; for example, those of Kripke about "Gödel" and "Jonah."[1] MMNS responded by testing the intuitions of the folk, in particular those of undergraduates in Rutgers and Hong Kong about cases like Kripke's Gödel and Jonah ones. Whereas philosophers of language, we can presume, almost all share Kripke's antidescriptivist intuitions about these cases,[2] the experiments revealed considerable variation in the intuitions of the undergraduates; indeed, the intuitions of Hong Kong undergraduates leaned toward descriptivism in the Gödel cases.[3] MMNS presented this result as casting doubt on philosophers' intuition-based methodology for theorizing about reference and hence, of course, on the resulting theories of reference.

A criticism aimed at MMNS by Genoveva Martí (2009) and Michael Devitt (2011a, 2012b,c) was that MMNS had tested the wrong thing.[4] Experimentalists should not be testing theories against *anyone's* referential intuitions but rather testing them against the reality that these intuitions are about: Theories should be tested against *linguistic usage*.[5] The right response to armchair philosophy is not to pull up more armchairs for the folk.[6]

What lies behind the standard methodology of relying on intuitions? The answer suggested by the literature is that competent speakers of a language have some sort of privileged access to the referential facts about their language. Devitt has argued that this answer is mistaken (Devitt, 1996, pp. 48–54, 72–85; 2006a, pp. 95–121; 2006b, 2010, 2012a, 2015a). Intuitions about language, like intuitions in general, "are empirical theory-laden central-processor responses to phenomena, differing from many other such responses only in being fairly immediate and unreflective, based on little if any conscious reasoning" (Devitt, 2006a, p. 103; 2006b, p. 491). A speaker's competence in a language does, of course, give her ready access to the *data* of that language, the data that the intuitions *are about*, but it does not give her privileged access to the *truth* about the data. It is *not* a methodological consequence of this view that referential intuitions should have no evidential role in theorizing about reference. It *is* a consequence, however, that they should have that role only to the extent that they are likely to be reliable, only to the extent that they are *reliable indicators*. And that likelihood needs to be assessed empirically, using independent evidence about reference relations obtained from linguistic usage. But we need that independent evidence anyway. This evidence can be found in the corpus but it can be found more easily using the technique of "elicited production," taken from linguistics (Thornton, 1995, p. 140). That is the method that we favor.

Machery resists the epistemological contrast between evidence from usage and intuitions, concluding a criticism of Martí (2009, 2012, 2014) with the claim that this contrast "violates the usual linguistic methodology in an apparently ad hoc manner, and it is implausible" (Machery, 2014, p. 15). But, as Devitt has argued (2006a,b, 2010), the contrast applies as much to linguistics as to the theory of reference. The linguists' practice of using intuitive judgments about acceptability, synonymy, and so on as evidence is open to the same criticism as the philosophers' practice of using intuitive judgments about reference. However, our contrast "violates the usual linguistic methodology" much less than the usual philosophical one because linguists, unlike philosophers, do not rely *solely* on intuitions as evidence: They also test usage by looking to the corpus, elicited production, reaction times, eye tracking, and electromagnetic brain potentials.

Our view of the source of intuitions raises a question about the reliability of folk referential intuitions.[7] Past experiments have provided evidence that these intuitions in Gödel cases are unreliable. First, these intuitions have proved quite susceptible to wording effects. Although the results of MMNS have been replicated several times, it has been found that small changes in wording yield somewhat different results.[8] Second, there is worrying evidence that the experimental task may be beyond many participants. The MMNS prompt asks participants to say who "John," a character in their vignette, "is talking about" when he "uses the name "Gödel." In one experiment (Sytsma & Livengood, 2011, pp. 326–327), participants who had answered this

question were then asked how they had understood the question: Is it about who John *thinks* he is talking about or about who John is *actually* talking about? Remarkably, 44 out of 73 chose the former, providing clear evidence that they had misunderstood the question. (If we ask whether it rained at Trump's inaugural, we are not asking whether Trump, or anyone else, *thinks* it rained.) Third, if Kripke's antidescriptivism is right about names, as our present tests of usage suggest, then folk intuitions would have to be antidescriptivist to be reliable. But these past studies show that the folk intuitions are far from consistently antidescriptivist.

In sum, there are very good reasons for preferring to test theories of reference against linguistic usage rather than referential intuitions. Our general substantive aim in this paper is to do just that. But we also have a general methodological aim: to examine ways of testing theories of reference. More particularly:

Substantive Aim (I). We have conducted experiments, using the method of elicited production to test usage on Gödel and Jonah cases.[9] Our main aim in so doing was to address the substantive issue of the reference for proper names. We shall see that these results count strongly against classical description theories and hence give some indirect support to causal-historical theories of names.

Methodological Aim (I). In response to Martí's criticism that experimentalists should be testing usage, Machery, Christopher Olivola, and Molly De Blanc ("MOD") conducted what were in effect "truth-value judgment" tests. They took these to be tests of usage and to rebut Martí's criticism because their results were in sync with tests of referential intuitions (2009). But Martí doubts that these truth-value judgment tests really test usage. Indeed, the tests raise in her "pretty much the same concerns" (2012, p. 74) as did MMNS's original tests. We aim to address the status of truth-value judgment tests, arguing in Section 1.4 that they are (somewhat imperfect) tests of usage and so should not raise those same concerns in Martí. And we will present some experimental results that are consistent with that conclusion.

Substantive Aim (II). If we are right that truth-value judgment experiments really do test usage, then MOD's results, particularly the variation in usage that they revealed, are damaging to theories of reference, even more so than MMNS's results; see Sections 1.2–1.3 for discussion. So our second substantive aim was to conduct truth-value judgment tests to see if they replicated MOD's variation in usage. The tests mostly revealed no such variation. Our elicited production tests revealed none too.

Methodological Aim (II). We aim to test whether the folk's referential intuitions on Gödel and Jonah cases are reliable (in the sense of being likely to be true) by comparing them with the results of tests of elicited production and truth-value judgments. Past tests of these referential intuitions led us to predict that we would find these intuitions at odds with the usage results and hence have evidence that they are unreliable; see above. But that was not what we found in our experiments (although we still think that folk intuitions are unreliable for reasons that we will articulate).

In sum, we will be reporting and comparing the results of experiments on Gödel and Jonah cases that test usage, truth-value judgments, and referential intuitions.

The description theories of names tested directly by MMNS and by us are the classical Frege–Russell sort. Evidence against such theories is likely to count also against "cluster" description theories of the Searle–Strawson sort (Searle, 1969, pp. 162–174; Strawson, 1959, pp. 180–183, 190–194). Kripke's criticism has prompted many other sorts of description theories which are not in contention here (but which have other problems; Devitt & Sterelny, 1999, pp. 60–62). Our unqualified uses of "the description theory" and cognates should be taken to be talking only of classical theories.

1.2. Martí's criticism of MMNS

Martí criticized MMNS for failing to test theories of reference against linguistic usage. We agree, of course, but we are not happy with the way she presented this criticism.

Here, is how we think the criticism should be presented. MMNS tested referential intuitions when they should have tested linguistic usage, for the reasons given in the last section. Referential intuitions are *metalinguistic judgments about the referential properties of an expression or utterance*; for example, in MMNS's experiment, the judgment that John, in his use of "Gödel," is "talking about" the person who got hold of the manuscript. In contrast, linguistic usage consists largely of the subconscious processes of producing and understanding linguistic expressions. These processes are very speedy and might well be called "intuit*ive*" but, even if the ones that should count as evidence express *judgments*, those judgments are not about *reference*, hence they are not *referential* intuitions; rather they are judgments about *the nonsemantic world*. Martí provides a nice example (that partly inspired our experiment). She suggests that MMNS's Gödel story should have ended:

> One day, the fraud is exposed, and John exclaims: "Today is a sad day: we have found out that Gödel was a thief and a liar."
>
> What do you think about John's reaction? (2009, p. 47)

Now suppose that this elicited the response, "John's right: Gödel was a thief." We might say that this was an intuitive judgment about *Gödel* but it was *not* an intuitive judgment about *the reference of the name* "*Gödel*." Yet this example of the method of elicited production would provide good and direct evidence that a competent speaker understands "Gödel" as referring to the character in the vignette who stole the proof rather than to Schmidt who discovered the proof; hence it would provide good and direct evidence that that is who "Gödel" does refer to.

Martí presents her criticism of MMNS rather differently:

> MMNS test people's intuitions about *theories* of reference, not about the *use* of names. But what we think the correct theory of reference determination is, and how we use names to talk about things are two very different issues. (2009, p. 44)

There are two problems with this description. First, what MMNS test is not people's intuitions about *theories of* reference, of which the folk likely have none, but about *reference itself*: participants are asked to judge who John is "talking about" (2004, p. B6) not *in virtue of what* John was referring to that person.[10] Second, the test that Martí is urging is not a test of "*intuitions about*" the use of names but a test of *the use of names itself*; see above. MMNS's mistake is not that they tested the *wrong kind* of intuition but rather that they tested *a referential intuition rather than usage*.

In responding to Martí's criticism, MOD (Machery, Olivola, and De Blanc) clearly understand that the mistake that MMNS were accused of making was failing to test *usage*. Nonetheless, going along with Martí's incorrect idea that this mistake was a failure to test *intuitions about* usage, they introduce the term "linguistic intuition" for what was not tested (2009, p. 689). This is a strikingly infelicitous term for linguistic usage[11] and so we will use it only in scare quotes. However, MOD use the unobjectionable term "metalinguistic intuition" for what *was* tested.

1.3. MOD's response to Martí

In their response, MOD tested both the "linguistic intuitions" and metalinguistic intuitions of participants in three countries, India, Mongolia, and France on a Gödel case.[12] They used the following vignette:

> Ivy is a high school student in Hong Kong. In her astronomy class, she was taught that Tsu Ch'ung Chih was the man who first determined the precise time of the summer and winter solstices. But, like all her classmates, this is the only thing she has heard about Tsu Ch'ung Chih. Now suppose that Tsu Ch'ung Chih did not really make this discovery. He stole it from an astronomer who died soon after making the discovery. But the theft remained entirely undetected and Tsu Ch'ung Chih became famous for the discovery of the precise times of the solstices. Everybody is like Ivy in this respect; the claim that Tsu Ch'ung Chih determined the solstice times is the only thing people have heard about him. (2009, p. 690)

The test of "linguistic intuitions" went on:

> Having read the above story and accepting that it is true, when Ivy says, "Tsu Ch'ung Chih was a great astronomer," do you think that her claim is: (A) true or (B) false? (p. 690)

So this is what is often called a "truth-value judgment test" (Gordon, 1998). The test of metalinguistic intuitions went on:

> Having read the above story and accepting that it is true, when Ivy uses the name "Tsu Ch'ung Chih," who do you think she is actually talking about:

(A) the person who (unbeknownst to Ivy) really determined the solstice times?
or
(B) the person who is widely believed to have discovered the solstice times, but actually stole this discovery and claimed credit for it? (p. 690)

MOD "found the same pattern of answers in all three countries" (p. 691). That pattern was of failing

> to find differences between these two kinds of intuitions . . .[This] suggests that linguistic and metalinguistic intuitions are largely congruent. People's intuitions about what proper names refer to in counterfactual cases seem to be in sync with the way they use these proper names to make judgements about the characters described in these cases. (p. 692)

A notable feature of this congruence is that MOD found the same within-culture variation in "linguistic intuitions" as they and MMNS found in metalinguistic intuitions: "Finally, and perhaps most damning to Marti's argument, we repeatedly found a large within-culture variation in people's linguistic intuitions" (p. 693).

Indeed, if these findings about "linguistic intuitions"—that is, about linguistic usage—reflect reality then they are much more damaging to the theory of reference than MMNS's findings. According to MMNS, their test of metalinguistic intuitions casts doubt on philosophers' intuition-based methodology for theorizing about reference and hence, of course, on the resulting theories of reference. There is room for argument about MMNS's claim—see note 4—but even if it is right it leaves a way out for those who want to defend a theory of reference: change from an intuition-based to a usage-based methodology. But if MOD's claim is right then, so far as reference is concerned, linguistic usage itself varies in the community, being sometimes descriptive, sometimes not. What could we make of that?

(i) We could say that a name is systematically ambiguous in that it is associated with more than one convention, a descriptive convention (several?) and a causal-historical one. So members of the community use it sometimes according to one convention, sometimes according to another. This sort of view of a term is not appealing but it has been urged for a few names—those of authors, like "Shakespeare" (Devitt, 2011a, p. 428 n. 9; 2012b, p. 12 n. 8)—and for all natural kind terms (Nichols, Pinillos, & Mallon, 2016). In any case, although the view could explain a variation in usage from context to context, it could not explain the variation in question, a variation in the one context.
(ii) We could say that, within what seems to be the one speech community, one group of people participates only in a descriptive convention for names, another, only in a causal-historical convention. So, to that extent, the groups speak different languages. But that seems most unlikely to be a plausible explanation of MOD's results.

(iii) So it is likely that we would have to say that, though a name has just one convention, reference has nothing to do with it: the meaning and role of names is to be explained in other terms. This would have the alarming consequence that semantics has no place for theories of reference.[13]

So a lot is at stake if MOD's truth-value judgment tests really do provide evidence of the claimed variation in linguistic usage. But do the tests provide this? Martí doubts that they do. We now turn to that issue.

1.4. Martí's response to MOD; Truth-value judgments as tests of usage

A truth-value judgment, like judging Ivy's claim "false," is prima facie quite different from an elicited production, like responding to John's damning claim about Gödel with "John's right: Gödel was a thief." And this difference leaves Martí quite dissatisfied with MOD's experiment as a response to her criticism of MMNS:

> The question I proposed in Martí (2009) is meant to elicit responses in which subjects use "Gödel," it does not ask subjects to reflect on someone's use of "Gödel." Does MOD's linguistic question do that? It seems to me that it does not for. . .the questions still ask subjects to reflect on someone's practice using "Gödel," it doesn't require of them to use "Gödel." (2012, pp. 74–75)

In brief, Martí does not think that MOD's truth-value judgments test usage.

We think that she is wrong about this. We shall argue, to further our Methodological Aim (I), that whereas an elicited production is a pure test of usage, a truth-value judgment is a *somewhat imperfect* test of it. Our reason for thinking that a truth-value judgment tests usage is indicated by one of MOD's replies to the following anticipated objection:

> Because "Tsu Ch'ung Chih" is mentioned, not used, in the question for the linguistic case, one might object that this case does not really elicit a linguistic intuition and that our study is thus not relevant to the issue at hand. (2009, p. 691)

MOD's reply is that

> asking people to assess the truth-value of a sentence relates to (though is not identical with) their desire to assert it and thus to what they would say. (ibid.)

The appropriateness of this reply needs to be explored. It is important not only to MOD's disagreement with Martí but more generally to establish the evidential significance of truth-value judgment tests to the theory of reference.

Deflationists about truth have emphasized a feature of truth terms (like "true") that should be accepted by all: *a truth term has an extremely useful "expressive" role*. The

term can play this role because it yields equivalences like the classic one between "Snow is white' is true' and "Snow is white." When a truth term is attached to the quotation name of a statement it yields a statement that is equivalent to that statement: it undoes the effect of quotation marks. (Attention to this led to the name, "the disquotational theory of truth.") Indeed, when the truth term is attached to *any* device for referring to a statement it yields a similar equivalence; it is a "denominalizing" device. Thus, if Jack responds to Jill's remark, "George W. Bush was America's worst president," by saying, "That's true," what he says is equivalent to what she said. And if we were to say, "What Jill said is true," our remark would also be equivalent to Jill's. The generalization of this is the equivalence thesis: all appropriate instances of the "equivalence schema"

$$s \text{ is true iff } p$$

hold, where an appropriate instance substitutes for "*p*" the sentence referred to by what is substituted for "*s*."[14] The expressive role of "true" exploits this equivalence.

Given the expressive role of "true," a participant's asserting "true" in response to MOD's question is short for asserting something that is a translation of Ivy's "Tsu Ch'ung Chih was a great astronomer." As a competent user of the term "true," that is what the participant is asserting; or, putting this in popular proposition-speak, he is asserting the same proposition as he takes Ivy to have asserted. We should break this down into two processes: first, the participant understands Ivy's utterance, *including her use of* "*Tsu Ch'ung Chih*," a certain way, just as we understand the uses of language by others day in and day out; second, he asserts whatever he has understood Ivy to say. Let's say that the man Ivy's teachers *thought* made the discoveries about the solstices was X, and the victim astronomer who *actually* made the discoveries according to the vignette was Y. Given that the participant knows very well from the vignette that Y, not X, made the discoveries, we theorists can be certain about one important thing from the participant's assertion of "true": the participant has understood Ivy as referring to Y not X; the proposition that the participant understands and then asserts is about Y not X. This provides evidence from usage, and not from a referential intuition (unlike MMNS's paper), and so is just the sort of evidence we want.[15] And it is evidence for the description theory. So it is surprising that a third of the participants provided it!

It is important that the participant in asserting "true" may be doing *no more* than what we have just described. The first step of what the participant did was understanding Ivy's utterance in a certain way, a way that takes her to be referring to Y. This sort of understanding of another's language is an ordinary exercise of our linguistic competence. We learn from psycholinguistics that it involves subconscious, subpersonal, automatic, extraordinarily fast processing, *and that is mostly all that it involves* (Gernsbacher & Kaschak, 2003; Pickering, 2003; Tannenhaus, 2003). Where understanding is difficult—for example, with multiple center embedding—it *may* be helped by "central processor," relatively slow reasoning, leading to a conscious judgment about another's utterance. But such high-level processes are a very small part of language understanding. Standardly, we process one expression as a VP, another as a c-commanded NP,

and so on, without any central processor *judgment that* the expression is a VP, c-commanded, or whatever. Hence, we have no reason to assume that such high-level judgments play *any role at all* in the participant's understanding of Ivy and hence of her assertion of "true." We have no reason to suppose that the participant *judges that* Ivy is referring to Y. The participant *may*, of course, make such a judgment, but what his assertion of "true" provides evidence of is his *understanding Ivy as* referring to Y not of his *judging that* she refers to Y. (One might be tempted to resort to a popular weasel word and say that the participant "tacitly" made a referential judgment, but this temptation is not helpful and should be resisted.)

What do we get from a participant's assertion of "false"? As with "true," the participant understands Ivy's utterance, including her use of "Tsu Ch'ung Chih," a certain way. But this time he goes on to assert the negation of whatever he has understood Ivy to say. Given that the participant knows very well that Y, not X, made the discoveries, we theorists can be certain from the participant's assertion of "false" that he has *not* understood Ivy as referring to Y. Does the participant understand Ivy as referring to X and hence provide evidence against the description theory? Probably so, but the evidence is not as clear. The problem is that the participant may have a problem understanding Ivy's "Tsu Ch'ung Chih," not assigning it to either X or Y, and that may be sufficient for her assertion of "false." There is also the worry, to be discussed in the next section, that the vignette's use rather than mention of "Tsu Ch'ung Chih" is biased against the description theory. Still, the two thirds of the participants who responded "false" do provide some evidence against the description theory.

Return now to Martí's concerns about MOD's truth-value judgment test. We have seen that MOD's question does not "still ask subjects to reflect on someone's practice using 'Gödel.' True, "it doesn't require of them to use 'Gödel'" but it does require them to say either "true," which refers to Y and hence *exemplifies a usage* of "Gödel" to refer to Y, or "false," which probably refers to X and hence probably *exemplifies a usage* of "Gödel" to refer to X. And that is why this is a test of usage. True also, the test requires a metalinguistic judgment but, given the equivalence thesis, one might say that this is only trivially metalinguistic: it is syntactically so but not semantically so. Nor, we might add, is the test, in any interesting sense, a test of intuitions about truth conditions.

The truth-value judgment test is a test of usage but, unlike elicited production, it is a somewhat imperfect one. Its imperfection lies in the fact that *it primes a certain usage: it "puts words into the mouth" of the participant*; in MOD's case the prime comes from Ivy's utterance: "Tsu Ch'ung Chih was a great astronomer." So we conclude, furthering our Methodological Aim (I), that truth-value judgment tests are "primed" tests of usage. We then predict that these tests will yield similar results to elicited production tests. The tests we shall be discussing do mostly yield such results. And those tests also yield evidence against MOD's alarming claim that there is much variation in the use of names, thus furthering Substantive Aim (II).

We turn now to our experiments, starting with our vignettes.

2. Initial experiments

2.1. The vignettes

Gödel Case Vignette: Tsu Ch'ung Chih:

> Students in astronomy classes in Hong Kong are told that a man called "Tsu Ch'ung Chih" first determined the precise time of the summer and winter solstices. This is the only thing that typical Hong Kongers ever hear about this man. Now suppose that that man did not make the discovery he is credited with. He stole it from an astronomer who died soon after making the discovery. But the theft remained entirely undetected and so the man that Hong Kongers have been told about became famous for the discovery of the precise times of the solstices.

This vignette is based on the one used by MOD, quoted earlier in Section 1.3, which is along similar lines to the one used by MMNS. However, we introduced several changes. We (a) shortened the vignette, (b) used anaphoric devices instead of the name "Tsu Ch'ung Chih" after introducing the name, (c) replaced the factive "taught" with the nonfactive "told" in the first sentence, and (d) removed the "Ivy" character from the vignette.

The most important of the changes we made was (b). That change was motivated by the fact that MMNS' and MOD's vignettes are incoherent if a (classical) descriptivist theory of names is true.[16] For example, consider MOD's vignette reproduced in Section 1.3. It contains five uses of the name "Tsu Ch'ung Chih." Now consider the question: Who do these uses refer to? There is presumably no doubt that the authors of this vignette, MOD, are fully competent with the name "Tsu Ch'ung Chih." And the referent of this name out of the mouths of the fully competent is the astronomer called "Tsu Ch'ung Chih" that Ivy's teacher was talking about, someone who did as a matter of fact make the discovery about the solstices (we presume). So *that* is who those five uses of the name in the vignette refer to. But then MOD's use of the name in the following passage disconfirms the description theory: "Now suppose that Tsu Ch'ung Chih did not really make this discovery. He stole it from an astronomer who died soon after making the discovery." If MOD's use of "Tsu Ch'ung Chih" refers to the famous astronomer in virtue of their associating with it the description, "the discoverer of the precise times of the solstices," this passage is not something that they would be disposed to say. They would not, in one and the same breath, both refer to Tsu Ch'ung Chih and suppose away the basis of that reference. To avoid this bias against the description theory, our vignette does not contain any use of the name "Tsu Ch'ung Chih."

Does our vignette's use of the indefinite description, "a man called 'Tsu Ch'ung Chih'," still bring a bias problem? We think not. It is a subtle matter deciding what to say about the reference determination of this description, hence of the anaphoric "that man" in the supposition that "that man did not make the discovery." Arguably, what we should say has the consequence that the referent must be a man called "Tsu Ch'ung

Chih." But it surely does *not* have the consequence that the referent must have discovered the solstices. So the supposition that he did not make that discovery is coherent. More important, it is hard to see how the supposition, which does not use the name "Tsu Ch'ung Chih" could be inconsistent with a (classical) description theory of that name.

Jonah Case: Ambiorix:

> In Belgium, every schoolchild is told about a man named "Ambiorix" in history classes. They are told that he led the Gauls against the Romans when the Romans invaded Gaul, and that he killed scores of legionnaires single-handedly. These tales are all that is on the records today under "Ambiorix." Recently, it's been discovered that those tales are just legends. The leader known in ancient times as "Ambiorix" never fought the Romans, and he never killed a single legionnaire. The discovery reveals, however, that he was a much admired general and an important peacetime leader. He also had a charismatic personality, attracting the attention of those around him wherever he went. In ancient Gaul, leaders were often respected for their accomplishments on the battlefield. So his friends and fellow soldiers, who admired him greatly, spun those tales about him.

This vignette is new. (MMNS used a Jonah vignette in their 2004 study, but the story told was different.)

There were three types of prompt in our study: (1) free-responses in the elicited production tasks, (2) forced-choice in the truth-value judgment tasks, and (3) forced-choice in the referential intuition tasks. We describe each in more detail below.

(1) Elicited production ("EP"): In this task, participants were given a free response prompt about the vignette. These prompts were phrased so as to invite one of two types of response. One type is predicted by the description theory, the other by the antidescriptivist causal-historical theory. For example, some of the participants who saw the "Tsu Ch'ung Chih" vignette were asked

> Having read the above story and accepting that it is true, what is your opinion of Tsu Ch'ung Chih?

Suppose some participant understands and uses "Tsu Ch'ung Chih" to refer to Y (the person who actually made the discoveries), as predicted by the description theory (but see the "New-Meaning" objection in Section 2.4). He would then be likely to have a positive opinion about Tsu Ch'ung Chih, on the basis of his supposed scientific accomplishment. Now suppose another participant understands and uses the name "Tsu Ch'ung Chih" to refer to X (the person who stole the discoveries), as predicted by the causal-historical theory. She would then be likely to have a negative opinion of Tsu Ch'ung Chih, on the basis of his supposed theft of a discovery. We treated participant responses as descriptivist if we coded them to be of the first type, antidescriptivist if we coded them to be of the second type. Our coding followed a set of instructions written in advance of the study; see Appendix S1. In

cases in which the participants' responses were judged not be clearly of either of these two types (for example, because the participant's opinion is not clearly positive or negative or is clearly based on irrelevant considerations), the response was discarded from analysis.

(2) Truth-value judgment ("TVJ"): Half of these were written so that description theories predict that participants will respond "True" ("TVJ-D"), while antidescriptivist theories, such as the causal-historical theory, predict participants will respond "False." The other half were written so that description theories predict the participants will respond "False," while antidescriptivist theories predict participants will respond "True" ("TVJ-CH"). They were coded accordingly

(3) Referential intuition ("RI"): As in previous experiments on reference (Machery et al., 2004, 2009), participants answered forced-choice questions about the reference of a name used in the vignettes. In those experiments, participants were asked whom some character in the vignette—for example, Ivy—was "talking about" when they used a name. Having eliminated that character—see (d) above—we simply asked participants who the name referred to. For example, some participants who were shown the Tsu Ch'ung Chih vignette also saw the following prompt:

> Having read the above story and accepting that it is true, who does the name "Tsu Ch'ung Chih" refer to?
>
> a The man who discovered the summer and winter solstices.
> b The man who stole the discovery and took credit for it.

The "talking about" prompts have been criticized as ambiguous (Deutsch, 2009; Ludwig, 2007). The experimenters wanted the participant's intuition about who the name conventionally refers to, its *semantic* reference. But the participant might take the prompt as asking about who the character means by the name, its *speaker* reference. So the experiment may not have yielded what was wanted. This criticism led to many further experiments aimed at testing and removing the ambiguity (Machery et al., 2015; Sytsma & Livengood, 2011; Sytsma et al., 2015). Our prompt about what a *name* refers to seems to avoid this criticism because it is hard to see how it could be construed as a question about what some unmentioned particular *speaker* is talking about with the name. However, we think that the criticism is mistaken anyway[17] and so our change in prompt was not motivated by it.

Wojtek Rostworowski (in correspondence) has posed an excellent question about RI. He points out that just as there is an equivalence thesis for the ordinary term "true," so too is there one for the ordinary term "refer": all appropriate instances of "*a* refers to *x* iff $n = x$" hold, where an appropriate instance substitutes for "*n*" the name referred to by what is substituted for "*a*." So why wouldn't an argument analogous to our earlier one about truth-value judgment tests (Section 1.4) show that our RI are *also* primed tests of usage, thus challenging the contrast we draw in Section 1.1? Now, the key assumption in that earlier argument was that "true" has an expressive role that exploits its equivalence thesis. So the question Rostworowski raises is: Does "refer" *also* have an expressive role that exploits *its* equivalence thesis? If it does, then when RI participants give a response that amounts, for

example, to "The name 'Tsu Ch'ung Chih' refers to the man who stole the discovery and took credit for it," they are expressing the thought, "Tsu Ch'ung Chih is the man who stole the discovery and took credit for it." Now there is no comparable equivalence thesis to be exploited for claims about who a certain speaker is "talking about" and so those claims generate no comparable challenge to our Section 1.1 contrast.[18] Thus, Rostworoski is raising the possibility that in changing to claims about what a name "refers to," we have unwittingly turned a test of intuitions into a (primed) test of usage. Our best brief response to this is that although, given its equivalence thesis, it is indeed possible that "refer," like "true," has this expressive role we rather doubt that it, unlike "true," does have it in the ordinary language of the participants. Still, the possibility should not be ignored.

2.2. Methods

2.2.1. Participants

Participants were recruited through Amazon MechanicalTurk™. A total of 256 participants responded to the study and were compensated for their participation.

2.2.2. Materials

2.2.2.1. Survey: After agreeing to a consent form, answering demographic questions, and responding to a mandatory attention-check question (Appendix A), participants were directed to another page. At the top of this page was a vignette. Half the participants were presented with the Tsu Ch'ung Chih vignette; the other half, the Ambiorix vignette (see Section 2.1). Each participant was asked just one question about a vignette. Participants presented with the Tsu Ch'ung Chi vignette were asked one of questions 1–4, below. Participants presented with the Ambiorix vignette were asked one of questions 5–8, below. Accordingly, there were eight questions in the study, one for each vignette-question pair. After responding to the question, participants could choose to continue to the next page to complete the survey.

Tsu Ch'ung Chih Questions:

1. **EP:** Having read the above story and accepting that it is true, what is your opinion of Tsu Ch'ung Chih?
2. **TVJ-D:** Having read the above story and accepting that it is true, please indicate below whether you think the following statement is true or false.

   Tsu Ch'ung Chih was a great astronomer.

   True / False

   Confidence in your response:

   (not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

3. **TVJ-CH:** Having read the above story and accepting that it is true, please indicate below whether you think the following statement is true or false.

   Tsu Chu'ung Chih was a thief and a liar.

   True / False

   Confidence in your response:

   (not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

4. **RI:** Having read the above story and accepting that it is true, who does the name "Tsu Ch'ung Chih" refer to?

   a The man who discovered the summer and winter solstices.
   b The man who stole the discovery and took credit for it.

   Confidence in your response:

   (not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

Ambiorix Questions:

5. **EP:** In light of this discovery, please say what if anything you think Belgium history books should now say about Ambiorix. Please give reasons but be brief.

6. **TVJ-D:** Taking the preceding passage to be historically accurate, please say whether the following statement is true or false:

   Ambiorix is a legendary person who did not really exist.

   True / False

   Confidence in your response:

   (not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

7. **TVJ-CH:** Taking the preceding passage to be historically accurate, please say whether the following statement is true or false:

   Ambiorix was a real historical figure about whom stories have grown.

   True / False

Confidence in your response:

(not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

8. **RI:** In light of this discovery, who has the name "Ambiorix" been referring to in history classes?

   a The man who those legends were first spun about.
   b No real person at all.

   Confidence in your response:

   (not confident at all) 0 1 2 3 4 5 6 7 8 9 (very confident)

For the forced-choice questions (TVJ and RI), the order of answer-choices was randomized from one participant to the next in order to counteract any effect of answer-choice order on participants' responses.

2.2.2.2. Coding: We coded EP responses according to instructions written in advance of the study (Appendix S1). We checked our codings against those of two pairs of independent coders: The first pair were two philosophy graduate students at the Graduate Center, CUNY; the second, two undergraduates at CUNY colleges enrolled in introductory philosophy courses and reporting no familiarity with philosophy of language. The independent coders had no information about the study beyond the coding instructions. The undergraduates were given preliminary training using data from the pilot of this study. The graduates were not trained. Responses judged in accordance with antidescriptivist predictions were coded "1," responses judged in accordance with descriptivist predictions were coded as "0"; responses which were judged not to be clearly in either of these categories (perhaps because of irrelevance) were discarded.

2.3. Results

Participants: 41 of the 256 participants were removed from the study because they failed the mandatory attention check (Appendix A).[19] In addition, 8 participant responses were discarded in the Ambiorix EP task (but none in the Tsu Ch'ung Chih EP task) because, as per our coding instructions, they could not be coded as matching either descriptivist or antidescriptivist predictions.

The results of the different questions are listed in Table 1 and the comparison of results in Table 2.[20]

We tested responses for each question against chance responses (50% descriptivist) using a two-sided Fisher's exact test (Table 1). We used this test for all of the comparisons reported in this paper, unless otherwise noted.

The results of both EP tests were significantly more antidescriptivist than chance (Table 1). The combined TVJ-D and TVJ-CH result for each vignette was not

Table 1
Results of initial experiments

Prompt	% Responses Inconsistent With Descriptivism	Average Participant Confidence (0–9)	Independence From Chance	Odds Ratio, [95% CI:Upper, Lower]
TCC: EP	95*	N/A	$N = 20$, $p = .002$	.053, [.01, .047]
TCC: TVJ-D	95*	7.00	$N = 22$, $p = .002$	.048 [.01, .042]
TCC: TVJ-CH	81*	7.45	$N = 31$, $p = .016$	.240, [.08, .074]
TCC: RI	80	8.75	$N = 20$, $p = .057$	.250, [.06, 1.02]
Ambiorix: EP	100*	N/A	$N = 14$, $p = .006$	.035, [.002, .689]
Ambiorix: TVJ-D	67	8.69	$N = 21$, $p = .096$	.500, [.13, 2]
Ambiorix: TVJ-CH	96*	8.57	$N = 26$, $p < .001$	.040, [0, .34]
Ambiorix: RI	92*	8.13	$N = 24$, $p = .003$	.091, [.02, .48]

Note. *Differs significantly from chance (50%).

Table 2
Comparisons of Results

Comparisons	*N*	*p*	Odds Ratio, [95% CI: Upper, Lower]
TCC: TVJ and EP	73	.295	.346, [.040, 3.006]
Ambiorix: TVJ and EP	61	.184	.160, [0.01 to 3.385]
Ambiorix: TVJ-D and EP	35	.027	.067, [.004, 1.279]
Ambiorix: TVJ-CH and EP	40	1	.586, [.022, 15.346]
Ambiorix: TVJ-D and TVJ-CH	47	.015	.080, [.009, .719]
TCC: RI and EP	40	.342	.211, [.0213, 2.079]
Ambiorix: RI and EP	38	.602	.310, [.144, 72.053]

significantly different from the EP result (Table 2). With the Ambiorix vignette, however, the TVJ-D result was significantly, and surprisingly, less antidescriptivist than the EP result and the TVJ-CH result and so we have included those comparisons. Finally, the RI results for each vignette were not significantly different from the EP results.

2.4. Discussion

In six out of eight conditions, participant responses were decisively in favor of antidescriptivism, thus giving indirect support to causal-historical theories of reference for proper names. None provided evidence in favor of the description theory and only two results were less than strongly at odds with that theory.

Substantive Aim (I). We have argued that usage, not intuition, was the sort of evidence that should be brought to bear in selecting among theories of reference (Section 1.1). So the most important of our results by far are those from EP (elicited production), for this method is a pure test of usage. The results of those tests on both Gödel and Jonah cases were strongly antidescriptivist: Almost all participants used names in a way inconsistent with the predictions of the classical description theory.[21] Many previous tests of referential intuitions have cast doubt on antidescriptivism for names. We think that our results

from testing usage undermine the significance of those earlier experiments for the theory of reference.[22] But we must address a possible objection.

We shall call this "the New-Meaning objection." Although our results seem clearly to favor antidescriptivist theories of reference, there is an alternative explanation of these results that a classical description theory might appeal to. We have been supposing that the reference-determining descriptions for the participants' use of the names "Tsu Ch'ung Chih" and "Ambiorix," would be the *same* as those for the populations of Hong Kongers' and of Belgium students, respectively. But, the objection runs, that may not be so.

Now, given what participants learn from a vignette, some of the descriptions that they *associate with* the name in question are certainly different from those associated by the original population. *If those new associations were reference-determining* in the participants' language, then the description theory could explain the results; for example, it could explain—and explain as well as the causal-historical theory—why participants responded, "Tsu Ch'ung Chih is a thief and a liar." Might those new associations really be reference-determining? They *might*, but we think it most unlikely that they *are*. We must be brief with our reasons.

To suppose that the new associations are reference-determining is to suppose that the participants take the story narrated in a vignette to have led to a new conventional meaning and reference for the name; to have led to a community, *including the vignette's narrator*, that uses the name according to a new convention. Now meanings can change, of course, but they don't change every time we get new information about the world. In particular, a proper name typically does not change its meaning on discovery that it is "empty." Indeed, it was a triumph of the description theory that it gave a nice account of "Zeus does not exist," an account that *presupposes* that the "Zeus" means now just what it meant before it was discovered that nothing fits the descriptions associated with it. So we think it clear that the New-Meaning objection fails in the Ambiorix case: The description theory would not take the discovery in that case to have created a new meaning of "Ambiorix" and would predict, just as we supposed, that participants will take Ambiorix not to exist. And nothing in the Tsu Ch'ung Chih vignette invites the idea that "Tsu Ch'ung Chih" has developed a new meaning. So we think that a *plausible* description theory would not suppose that there was a new meaning there either.

Still, what about a theory that did suppose this, whether plausible or not? Our experiment was not designed to test such a description theory but it did provide some evidence against it. If the participants took a new meaning to have been created, then they should have shown sensitivity to the fact that this would make the name ambiguous; that "Tsu Ch'ung Chih" has its old meaning for Hong Kongers and its new meaning for the narrator's community. This would bring *two* Tsu Ch'ung Chihs into play. So that description theory would predict EP responses that were not just baldly negative like "Tsu Ch'ung Chih is a thief and a liar," but ones that also included something like "but the person formerly known as Tsu Ch'ung Chih made a great discovery." We did not get such responses.

Nonetheless, we decided to conduct a follow-up EP experiment aimed at a description theory that does suppose that there was a new meaning in the Tsu Ch'ung Chih case, and hence aimed at the New-Meaning objection to all experiments on that case (Section 3.1).

Methodological Aim (I). We argued in Section 1.4 that TVJ (truth-value judgment) tests are tests of usage, although imperfect ones because they prime the participants. Hence, we predicted that TVJ results would resemble the EP results. That is what we mostly found: There was no significant difference between the results of three of our four TVJ tests and those of the EP tests. So these three results are consistent with the view that TVJ are tests of usage. We shall discuss the exception, TVJ-D for Ambiorix, below.

Substantive Aim (II). Using TVJ tests, MOD claimed to show that there is much the same variation in usage as has been found in referential intuitions. If MOD were right about this variation in usage, that would be very damaging to theories of reference (Section 1.3). Three of our four TVJ tests failed to replicate MOD's result. Participants were well above chance in reporting "True" where descriptivist theories predict they will respond "False," and in responding "False" where those theories predict they will respond "True." Our two EP tests did not replicate MOD's result either. So, provided the New-Meaning objection fails, we have five tests of usage yielding antidescriptivist results and providing no evidence of variation in usage.

We must now consider the exception. In one of the Ambiorix TVJ tasks (TVJ-D), responses were not significantly different from chance. This is particularly surprising since we did not find the same result in the other Ambiorix TVJ task (TVJ-CH): 96% of TVJ-CH responses were antidescriptivist but only 66% of TVJ-D were. Why the difference? The result is unlikely to be due to some bias toward responding "True" or "False" in the TVJ task, or to a framing effect, since we do not see a similar difference between responses to the TVJ-D and TVJ-CH questions for the Tsu Ch'ung Chih vignette. We suggest instead that this difference is due to the wording of the prompt. Specifically, one word, "legendary," seems to have introduced confusion. A comparison with the pilot experiment for this study suggests that this word-choice is at the root of the finding. Our TVJ-D prompt reads:

> Taking the preceding passage to be historically accurate, please say whether the following statement is true or false:
>
> Ambiorix is a legendary person who did not really exist.

The pilot TVJ-D prompt ended with a different sentence:

> Ambiorix is a fictional person who did not really exist.

This change seems small yet the results of the pilot study, unlike this study, were highly significantly more antidescriptivist than chance (86%, $N = 29$, $p < .001$). This suggests such a wording effect may have been responsible for the difference we found.

There are at least two ways the use of the word "legendary" may have been responsible for the result we saw. First, participants may have read "legendary" in its praise-conferring sense (as it is most naturally read in "John Wooden was a legendary basketball coach"). Such uses of "legendary" suggest that the thing described as legendary exists. Yet the sentence continues on with the dependent clause "who does not really exist."

Perhaps participants resolved this apparent conflict by agreeing with the independent rather than the dependent clause. Alternatively, participants might have been confused because the vignette *states* that traditional stories including the name "Ambiorix" are in fact legends. This may have led some who disagreed with the "who does not exist" clause to mark "True" anyway.

Therefore, we decided to conduct a follow-up experiment without the troublesome "legendary" to further our Methodological Aim (I) and Substantive Aim (II) (Section 3.1).

Methodological Aim (II). These results also weigh in on the evidential status of folk referential intuitions in theorizing about reference. In the past, tests of judgments on Gödel and Jonah cases have shown that the folk vary in their referential intuitions, being roughly split between descriptivist and antidescriptivist judgments. Thus, testing Westerners on the Gödel case, MMNS originally found only 58% had antidescriptivist intuitions; much the same result was found in several repeats of this experiment (but Sytsma & Livengood, 2011, got only 39.4% anti-descriptivist). It should be noted, however, that changes in wording in some experiments have reduced the variation: Sytsma and Livengood (2011) achieved 73.8% antidescriptivist (in "Clarified Narrator's Perspective"), Machery et al. (2015), 73.9% (in "Award Winner Gödel Case"). So our usage results imply that folk referential intuitions are mostly unreliable; alternatively, that the standard method of *testing* referential intuitions, namely with vignette-based, forced-choice surveys, does not reliably track folk intuitions. These earlier results[23] also led us to predict that our RI test would yield a variation in the referential intuitions thus adding to evidence of their unreliability. Thus, we were surprised to find little variation: The results were firmly antidescriptivist (Tsu Ch'ung Chi, 80%; Ambiorix, 92%) and in accord with the results from our usage tests (EP and TVJ); we did not replicate past variations.

We have remarked earlier (Section 1.1) on the evidence of wording effects in previous studies on the expression of referential intuitions. Our results supply further evidence. Consider our Tsu Ch'ung Chi experiment in particular. The vignette and prompt did differ in several ways from earlier Gödel experiments, as noted in Section 2.1. Still, the change that strikes us as most important—(b) avoiding uses of the name in favor of anaphoric devices—removed a clear bias *against* the description theory. So this change should have made descriptivist answers *more* likely![24]

The susceptibility of intuition testing to wording effects is even more striking when we compare the results of our study to those of the pilot experiment we ran on the Tsu Ch'ung Chi case. In that pilot, unlike in this study, we replicated the results of many previous studies: We found nearly split results on the intuitions (only 53% antidescriptivist). Yet, when we moved from the pilot to this study, we made just two small changes in the vignette and prompt. We expanded the pilot's short prompt,

The name "Tsu Ch'ung Chih" refers to
to

> Having read the above story and accepting that it is true, who does the name "Tsu Ch'ung Chih" refer to?

(Both prompts were followed by the same forced choice answer choices.) This change seems almost trivial. The change in the vignette seems more significant. We replaced the sentence

> Students in astronomy classes in Hong Kong are told about a man called "Tsu Ch'ung Chih" who first determined the precise time of the summer and winter solstices

in our pilot experiment with the sentence

> Students in astronomy classes in Hong Kong are told that a man called "Tsu Ch'ung Chih" first determined the precise time of the summer and winter solstices

in this study. We made this change to remove a possible incoherence in the pilot's vignette. It is arguably a consequence of what we should say about the referent of the long indefinite description, "a man called 'Tsu Ch'ung Chih' who first determined the precise time of the summer and winter solstices," that the referent must be a man who made that discovery. If so, we should say the same about the referent of the anaphoric "that man" in the supposition that "that man did not make the discovery" (Section 2.1). But then the supposition would be incoherent! Our change removed this possible incoherence. This strikes us as a clear improvement, but it is strange indeed that this removal would push the results against a (classical) description theory of "Tsu Ch'ung Chih."

A final thought. In Section 2.1, we raised the possibility that claims about what a name "refers to" have an expressive role and so our RI tests are actually primed tests of usage. We did not embrace the view that these tests *are* primed tests of usage, but suppose that they are. Then it would *not* be surprising that the results of these tests would be similar to those of our unprimed tests of usage. And it would *not* be surprising that they differed from those of earlier intuition tests of what a speaker is "talking about," and the like, for those locutions do not have an expressive role. One might even hope that the wording effects that plague the latter tests might spare tests like our RI, tests of intuitions about what a name refers to. But that hope is rather quickly dashed by the just-noted results of our pilot experiment.

All in all, our study adds to the case against using referential intuitions, obtained in vignette-based, forced-choice experiments, as evidence for theories of reference. First, our usage results show that the intuitions in earlier experiments are mostly unreliable. Second, we have added to the evidence that these intuitions are susceptible to puzzling wording effects which alone casts doubt on their reliability. Finally, there is the high level of misunderstanding among participants demonstrated by Sytsma and Livengood (2011) and noted earlier (Section 1.1).

To see if our failure to replicate the variation in previous tests of referential intuitions was anomolous, we decided to rerun the Tsu Ch'ung Chih RI test in a followup (Section 3.1).

So we had decided on three follow-up experiments.

3. Follow-up experiments

3.1. Introduction

First, we wanted a Tsu Ch'ung Chih experiment that, more clearly than the initial ones, escaped the New-Meaning objection. Thus, we ran another variant of the EP task. This time, we asked participants to give their opinions of Tsu Ch'ung Chih *as if they were speaking to a Hong Konger*. So participants are likely to use the name "Tsu Ch'ung Chih" as Hong Kongers do even if the participants take the name to have a new meaning as well. Our main aim was to present a disconfirmation of the description theory that was not open to the New-Meaning objection. But we were also interested in showing that the objection is mistaken. If the objection were good, then one would expect responses in the follow-up experiment to differ from those in the initial experiment. Specifically, one would expect participants to be much more positive in their opinions of Tsu Ch'ung Chih. Based on our interpretation of the results of the initial study, we predicted that there would be no such difference.

Second, we re-ran the Ambiorix TVJ-D task, with a slight modification: We removed the apparently confusing expression "is a legendary person" from the prompt, so that it read simply, "Ambiorix did not really exist."

Third, we ran more participants in the Tsu Ch'ung Chih RI task to confirm the unusual result we found in the initial experiment.

3.2. Methods

3.2.1. Participants, Materials, and Procedure

Tsu Ch'ung Chih EP: 27 participants responded to this question. Eight responses were discarded because, per our coding instructions, they could not be coded as cohering with descriptivist or antidescriptivist predictions.

The materials and procedure were as reported in the initial study, with the following exceptions. In this experiment, the prompt read:

> Having read the above story and accepting that it is true, what would you say to a Hong Konger about Tsu Ch'ung Chih? (Please write as if you are speaking to a Hong Konger).

In this case, we checked our coding only against those of the two undergraduate coders used in the initial test.

Ambiorix TVJ-D: 33 participants responded to this question. One was removed for failing the attention check.

The methods and procedure were as reported in the initial study, with the following exception. In this experiment, the prompt read:

Taking the preceding passage to be historically accurate, please say whether the following statement is true or false:

Ambiorix did not really exist.

Tsu Ch'ung Chih RI: 30 participants responded to this question. Three were removed for failing the attention check.

The materials and procedure were as reported in the initial study.

3.3. Results

Participants gave antidescriptivist-predicted responses in all three tests. The results are reported in Tables 3 and 4.[25]

The results of the follow-up Tsu Ch'ung Chih EP were not significantly less antidescriptivist than those of the initial test: follow-up, 89%; initial, 95%. The results of the follow-up Ambiorix TVJ-D were not significantly different from those of the initial Ambiorix EP. The results of the follow-up Ambiorix TVJ-D were 15% more anti-descriptivist than those of the initial Ambiorix TVJ-D, but this difference was not significant. The results of the follow-up Tsu Ch'ung Chih RI experiment were not significantly different from those of the initial experiment: follow-up, 90%, initial, 80%. The combination of these two RI tests, of course, yields a result that is significantly more antidescriptivist than chance ($N = 50$, $p < .001$, odds ratio = .163, 95% CI [.062, .043]).

Table 3
Results of follow-up experiments

Prompt	% Responses Inconsistent With Descriptivism	Average Participant Confidence (0–9)	Independence From Chance	Odds Ratio, [95% CI]
TCC: EP	89*	N/A	$N = 19$, $p = .014$	.118, [.02, .65]
Ambiorix: TVJ-D	81*	8.15	$N = 32$, $p = .017$	.231, [.07, .71]
TCC: RI	90*	7.93	$N = 30$, $p = .002$	.111, [.03, .45]

Note. *Results significantly different from chance (50%).

Table 4
Comparisons of results

Comparison	*N*	*p*	Odds Ratio, [95% CI: Upper, Lower]
TCC: Follow-up EP and Initial EP	39	.514	.447, [.037, 5.386]
Ambiorix: Follow-up TVJ-D and Initial EP	46	.157	.141, [.007, 2.678]
Ambiorix: Follow-up TVJ-D and Initial TVJ-D	53	.329	.462, [.13, 1.642]
TCC: Initial RI and Follow-up RI	50	.416	.444, [.088, 2.245]

3.4. Discussion

Substantive Aim (I). The New-Meaning objection to our initial antidescriptivist results in the Tsu Ch'ung Chih case was that, because of the information in the vignette, the participants' use of the name "Tsu Ch'ung Chih" may not be the same as that of the population of Hong Kongers. We did not find this plausible but decided to address the objection anyway in a follow-up EP experiment. We think that this experiment was clearly not open to the objection because we asked participants to give their opinions of Tsu Ch'ung So participants would be likely to use the name "Tsu Ch'ung Chih" with its old meaning, as it is understood by Hong Kongers, *even if* the participants took the name to now have a new meaning as well. Responses were decisively antidescriptivist and hardly differed at all from those in the initial experiment. Therefore, we take our two Tsu Ch'ung Chih EP experiments to provide powerful evidence against the classical description theory of proper names. And we take the follow-up experiment to provide evidence that the New-Meaning objection is a mistaken objection to all the Tsu Ch'ung Chih experiments. If the objection were good, then one would expect responses in the follow-up experiment to be significantly more positive in their opinions of Tsu Ch'ung Chih than those in the initial experiment. They were not. We think we have accomplished our aim.

Methodological Aim (I). In our view, TVJ tests are tests of usage, although imperfect ones because they prime the participants. So TVJ results should resemble the EP results, which straightforwardly reflect usage. That was what we found in three of our four initial TVJ tests. Still, there was the troublesome exception, the initial Ambiorix TVJ-D test. A comparison of it with our pilot suggested that the word "legendary" in its prompt may have confused participants. Therefore, we dropped that word and reran the experiment with a prompt that read simply, "Ambiorix did not really exist." The results, like those of the three other initial TVJ tests, did not differ significantly from EP results[26] and are consistent with the view that TVJ are tests of usage.

Substantive Aim (II). MOD offered evidential support from a TVJ experiment on participants from India, Mongolia, and France for the alarming conclusion that there is as much variation in usage as has been found in referential intuitions. Three of our initial TVJ tests, our two initial EP tests, and our follow-up EP test, all on Americans, failed to replicate MOD's findings. But there was the exception: the initial Ambiorix TVJ-D test. Our follow-up experiment with the changed prompt (no "legendary") also failed to replicate MOD's findings: It yielded a significantly antidescriptivist result with little variation. What should we conclude about MOD's findings? It is possible, of course, that the use of proper names among Americans differs from that among Indians, Mongolians, and the French. This needs to be tested but it strikes us as very unlikely. Thus, we conclude that either MOD's findings are anomalous, or else the specific prompt and vignette used in their study were insensitive to a real pattern in the usage of proper names. The latter is compatible with a susceptibility to wording effects for TVJ tests, like the one we observed for the Ambiorix TVJ-D question in this study.

Methodological Aim (II). The additional participants in our rerun of the Tsu Ch'ung Chih RI test cement our surprising failure to replicate most previous results, including those in our

own pilot study, in tests of folk intuitions about the reference of names. It seems that the folk, in responding to our vignette and prompt, unlike to most previous ones, are reliable intuiters about Gödel cases. It is quite unclear why this should be so, particularly given that the changes from our pilot to this study are so small (Section 2.4). We pointed out in Section 1.1 that referential intuitions should be used to test theories of reference "only to the extent that they are *reliable*." Our results suggest that folk intuitions about Gödel cases at least, or the methods typically used to test them, are reliable only in special circumstances that are hard to fathom. This counts against the use of these methods to test theories of reference.

4. Conclusions

Previous experiments on theories of reference have mostly tested folk referential intuitions and have shown great variation in those intuitions. We follow Martí (2009) and Devitt (2011a, 2012b,c) in thinking that experiments should focus on testing folk usage not folk intuitions about usage (Section 1.1). Our Substantive Aim (I) was to test usage for proper names on Gödel and Jonah cases using the method of elicited production. Our experiments, including one responding to "the New-Meaning objection," were decisively antidescriptivist. These results, in combination with the ones discussed below, provide very powerful evidence against classical description theories. Hence, they give indirect support to causal-historical theories. And they undermine the significance of previous tests of referential intuitions.

MOD (2009) conducted a truth-value judgment experiment in response to Martí's criticism (2009) that MMNS (2004) should have tested usage not intuitions (Section 1.2). MOD claimed to have rebutted that criticism by showing that there is much the same variation in usage as there is in referential intuitions. Martí (2012) doubted that MOD's experiment really did test usage, seeing it rather as just another test of metalinguistic intuitions. Our Methodological Aim (I) was to investigate this issue. We argued that the truth-value judgment experiment is indeed a test of usage, albeit a somewhat imperfect one because it "primes" a usage (Section 1.4). If this is right, truth-value judgment tests should yield much the same results as elicited production tests. And that is what we mostly found with our initial truth-value judgment tests: There was no significant difference between the results of three of those four tests and our elicited production tests. When we ran a follow-up test to the one exception, dropping the apparently confusing "legendary" from its prompt, we again found no significant difference from our elicited production tests. We stand by our argument that truth-value judgment experiments test usage.

This conclusion raised another issue that led to our Substantive Aim (II). If MOD's truth-value judgment experiment did indeed test usage then it provided evidence that the usage of a name varied, being sometimes descriptive, sometimes not. That would be a damaging discovery for the theory of reference (Section 1.3). So our aim was to see if our truth-value judgment tests replicated the variation revealed by MOD's test. In three of the initial four tests, and in the follow-up of the exception, they did not: They provided no evidence of variation. Our three EP experiments provided no evidence either. We wonder, of course, why MOD's experiment did show variation. We suspect a wording effect.

Aside from that one exception, our truth-value judgment tests were significantly antidescriptivist and hence add to the evidence against descriptivist theories of reference for names.

Past tests of folk referential intuitions in Gödel and Jonah cases have revealed wording effects and participant misunderstanding. And those studies have mostly shown folk intuitions to be split, sometimes descriptivist, sometimes antidescriptivist. Assuming that antidescriptivism is right, which is what our tests of usage support, those earlier results provide further evidence that the folk are unreliable in their referential intuitions about these cases. And our own theory of intuitions inclined us to that view too. Methodological Aim (II) was a further test of this reliability. Our experiments yielded strongly antidescriptivist intuitions and hence no evidence that the folk *are* unreliable. We were surprised by this and so included a re-run of the Gödel experiment in our follow-ups. The rerun yielded the same result. This is particularly surprising given that our own pilot study, differing only in two apparently minor ways, yielded the usual unreliability. This adds to the case that the results of referential intuition experiments using vignettes are susceptible to unpredictable wording effects, casting doubt on these experiments as effective ways to test theories of reference.

Future studies in this area should develop the methodological findings of this study. (1) We need further experiments using different vignettes to discover whether elicited production tests of theories of reference are also susceptible to wording effects. However, we predict that they will be found to be much less so. For we suspect that the words that have caused the variation in intuition and truth-value judgment tests have come from the experimenters' prompts rather than the vignettes, whereas the words that matter in an elicited production test come from the participants themselves. (2) The usage of proper names should be tested in other cultures than the American one we tested. This is particularly important given the cross-cultural differences revealed by many tests of referential intuitions. We predict that the usage of proper names will not differ across cultures. (3) There should be further tests of intuitions about what a name "refers to" to see whether these are indeed reliable and not prone to wording effects. Ideally, these tests would be done in comparison with tests of intuitions about what a speaker is "talking about," using the same vignette. (4) Promisingly, the method of testing usage could be applied across other areas in the theory of reference, for example, to natural and artifactual kind terms.

Acknowledgments

Versions of this paper have been delivered in several places, starting with the Jean Nicod Institute in Paris in February 2016 and the workshop, "Experimental Semantics and the Testing of Usage," at the University of Warsaw in April 2016. We are indebted to the helpful comments it received on these occasions. Also to comments from Andrew Latham, Bianca Cepollaro, Brent Strickland, Daniel Cohnitz, David Chalmers, Edouard Machery, Eric Mandelbaum, Genoveva Martí, Georges Rey, Jesse Prinz, Jussi Haukioja, Natalia Pietrulewicz, Stephen Stich, Steven Gross, and Wojtek Rostworowski. This

research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank the Spanish MINECO project FFI2015-70707P for its support.

Notes

1. This view of the philosophical method is challenged by Max Deutsch (2009) and Herman Cappelen (2012). See Devitt (2015b) for a response. Surprisingly, it is also challenged by Genoveva Martí. Like us, she thinks that theories of reference should be tested against usage but, unlike us, she thinks that this is what semanticists do. She claims that the theories of Kripke and Donnellan "are based on the observation of ordinary usage" (2012, pp. 75–76). In a certain respect, we think that this is right. But the question is: What do these observations yield that is counted as *evidence* for the theories? We think, as do MMNS, that the answer is: intuitions about what a term refers to. That is, observing an act of referring, semanticists judge intuitively what the term refers to and use this judgment as evidence for the theory of reference. This is not testing a theory against usage (Devitt, 2015a, pp. 39–43).
2. This presumption is confirmed in an experiment by Machery (2012a, pp. 46–50).
3. Influenced by the work of psychologists Richard Nisbett and colleagues (e.g., Nisbett, Peng, Choi, & Norenzayan, 2001), MMNS predicted that students in Hong Kong ("East Asians") would be more likely to have descriptivist referential intuitions than the students in Rutgers ("Westerners"). And that is what they found. Yet the prediction is puzzling (Devitt, 2012b; Martí, 2012; Ostertag, 2013). In any case, we shall not be testing cross-cultural variations.
4. These papers include other criticisms of MMNS as do Ludwig (2007), Jackman (2009), Ichikawa, Maitra, and Weatherson (2012), Sytsma and Livengood (2011), Genone and Lombrozo (2012), and Ostertag (2013). Machery, Mallon, Nichols, and Stich (2013) is a reply to Devitt (2011a) and Ichikawa et al. (2012). Devitt (2012b) is a response.
5. Jennifer Nado and Michael Johnson agree (2016, p. 142).
6. We owe this nice remark to Martí, who thinks she heard or read it somewhere but cannot recall the circumstances.
7. It also suggests that, insofar as we rely on referential intuitions rather than usage as evidence, we should prefer those of philosophers because philosophers have the better background beliefs and training; they are more expert. This yielded an example of what has become known as "the Expertise Defense" against the findings of MMNS. The Expertise Defense has led to a lively exchange of opinion: Weinberg, Gonnerman, Buckner, and Alexander (2010), Machery and Stich (2012), Machery et al. (2013), Machery (2012a), Devitt (2012b,c), and Machery (2012b). See also the following exchange arising out of the analogous claim that we should prefer the grammatical intuitions of linguists over those of the folk: Devitt (2006a, pp. 108–111; 2006b, pp. 497–500), Culbertson and Gross (2009), Devitt (2010), and Gross and Culbertson (2011).

8. See "Clarified Narrator's Perspective" in Sytsma and Livengood (2011); "reverse-translation probes" in Sytsma, Livengood, Sato, and Oguchi (2015); "Award Winner Gödel Case" and "Clarified Award Winner Gödel Case" in Machery, Sytsma, and Deutsch (2015).
9. The only other use of elicited production to test theories of reference that we know of is in an important recent paper, Domaneschi, Vignolo, and Di Paola (2017). That test also provides powerful evidence against classical descriptivism.
10. See Devitt (2011a, p. 432, n. 14). Nado and Johnson take Martí's criticism to be that MMNS "don't control for which theory of reference (if any) the subjects explicitly hold." To control for this you need "to instruct the subjects to give you a judgment on the case without deducing it from" such a theory (2016, p. 148). We can see no need for this instruction because it seems most unlikely that the folk have the theory (though they do of course have referential intuitions). Nado and Johnson continue: "Martí seemed to think that one could obviate the need for such an instruction by getting subjects to use the name, but we don't see why this should be" (*ibid*). We wonder why they don't. If one had a theory of reference that might well affect one's referential intuitions (though note that almost all philosophers, whatever their theories, agreed with Kripke's intuitions—see Devitt (2012b, pp. 21–22), but it seems most unlikely to affect one's very use of a name.
11. So too is "first-level intuition," used by Cohnitz and Haukioja (2015).
12. Machery and Stich report (2012, p. 502) that Machery and Olivola later tested participants from the United States in the same way with much the same results.
13. Cohnitz and Haukioja draw attention to this consequence of the alleged variation (2015, p. 640). MMNS's final discussion (2004, B8–B9), seems to raise doubts that theories of reference do have a place in semantics. Stich had raised such doubts before (1996, pp. 37–51; 2009, p. 199) and Machery and Stich did afterward (2012, section 1.4); cf Devitt (1996, 2009). Those, like Paul Horwich who are deflationist about truth (1990) and urge a use-theory of meaning (1998, 2005), also have no place for a theory of reference in semantics; c.f. Devitt (2002, 2011b).
14. This is rough; for example, we need to guard against the semantic paradoxes and allow for indexicals.
15. It provides this evidence only on the assumption that the participant is competent in the language Ivy is speaking, namely English. (Thanks to Genoveva Martí.)
16. This comment is like that of Devitt (2012b, pp. 27–28) on MMNS's "Gödel" vignette (2004) and of Devitt (2015a, pp. 47–49) on Genone and Lombrozo's "tyleritis" vignette (2012).
17. The criticism raises a subtle theoretical issue which we cannot explore here. But, *very* briefly, on our Gricean picture (Devitt, 2015c), speakers *normally* refer with a name to what it semantically refers to. Still, a speaker can use a name (like any other word) abnormally to refer to what it does not semantically refer to, so that its speaker referent is not its semantic referent. This may happen, for example, because of mistaken identity as in Kripke's famous raking-the-leaves

case (1979, p. 14); or because of metaphor/irony. But, *in the absence of any contextual information indicating that a speaker means to refer by a name to something other than its semantic referent, an audience should take the speaker to mean the semantic referent*. There is no such contextual information in the MMNS and MOD experiments. (For a related point, see Machery & Stich, 2012, p. 506.) So, even if a participant takes the prompt to be asking about the speaker referent, her response should still yield her intuition about the semantic referent because she should be taking the speaker referent to *be* the semantic referent.

18. Cohnitz and Haukioja (2015, pp. 631–632) make a related point to Rostworowski's but take it apply not only to "refer" but also to "talking about." But, when MMNS's participants give a response that amounts, for example, to "When John uses the name 'Gödel' he is talking about the person who got hold of the manuscript and claimed credit for the work," they are clearly *not* expressing the thought, "Gödel is the person who got hold of the manuscript and claimed credit for the work."
19. We looked at the results of attention-check-failers to gauge, informally, whether the attention-check was effective. The results suggest that it worked. The rate of EP responses coded as "discard" was higher among attention-check-failers (3/11) than among attention-check passers (8/42), and in one question, the Ambiorix RI question, attention check passers were nearly unanimously antidescriptivist in their responses (2 descriptivist, 22 antidescriptivist), while attention-check-failers appeared to be at chance (3 descriptivist, 3 antidescriptivist).
20. We analyzed the confidence ratings; see Appendix B. We also checked our EP codings against those of our four independent coders. Their antidescriptivist percentages on the TCC test were as folllows: graduates, 95% and 94%; undergraduates, 94% and 100%. Those on the Ambiorix test were as follows: graduates, 100% and 100%; undergraduates, 100% and 93%. There was some disagreement among coders over 18 responses. In 17 of these, the only disagreement was over whether to discard the response. We measured intercoder reliability by doing pairwise comparisons (Cohen's kappa) of our codings to each of the independent coders'. We included disagreements about discards in the comparison. Agreement was moderate for the TCC results (average $\kappa = .405$) and substantial in codings of the Ambiorix results (average $\kappa = .68$).
21. And not just with the classical theory, contrary to what Machery claims:

> Use is sufficient to falsify some simple descriptivist theories about the reference of proper names, but not more complex descriptivist theories that appeal, for example, to deference to experts. (2014, p. 13)

Machery's talk of "deference" alludes to a description theory of the sort introduced by Peter Strawson in a footnote (1959, p. 182n). The idea is that a person *A*, in acquiring a name *N* from *B* in communication, can "borrow the reference"

of *N* from *B* by associating with *N* a description of the following sort, "the person that *B* was referring to by *N*." This sort of borrowing must terminate with an "expert" lender who can identify the referent without borrowing. But in our vignettes any such terminating "expert" would be misinformed in the crucial respect, just like the borrowers: she would not associate descriptions that picked out what our tests of usage show to be the referents: the thief X in the Gödel case, the real historic figure in the Jonah case.

22. Since the results of the pilot of this experiment are contrasted with the main study elsewhere in this discussion, it is worth mentioning that the results in that pilot study were quite similar to those we found here (virtually identical, in the Gödel case).
23. Along with our reasons (Devitt, 2011a, 2012b) for doubting that the folk would have reliable intuitions about the complicated cases considered in those experiments.
24. Though, it should be noted that change (c), replacing "taught" by "told," removed a potential bias in favor of descriptivist theories.
25. In the EP task, the antidescriptivist percentages for the two independent coders were 79% and 93%. We again measured intercoder reliability with pairwise comparisons (Cohen's kappa) of our coding to those of each of our independent coders. Agreement was good in both comparisons (κ = .588 and .565). As in the initial experiments, in most cases where there was disagreement between coders about a response (six of nine cases), the disagreements were about whether to discard the response or not.
26. They did not differ significantly from the initial TVJ-D results either. That they did not is a bit surprising. It may in part reflect the small sample size (N = 21) in the initial TVJ-D test, due to our removing participants who failed the attention check from analysis.
27. This is a version of an attention check question from Brent Strickland, who credits it to Julian De Freitas.

References

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford, UK: Oxford University Press.

Cohnitz, D., & Haukioja, J. (2015). Intuitions in philosophical semantics. *Erkenntnis*, *80*(3), 617–641. https://doi.org/10.1037/arc0000014.

Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science*, *60*(4), 721–736. https://doi.org/10.1093/bjps/axp032.

Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind and Language*, *24*(4), 445–466. https://doi.org/10.1111/j.1468-0017.2009.01370.x.

Devitt, M. (1996). *Coming to our senses: A naturalistic program for semantic localism*. Cambridge, UK: Cambridge University Press.

Devitt, M. (2002). Meaning and use. *Philosophy and Phenomenological Research*, *65*(1), 106–121. https://doi.org/10.1111/j.1933-1592.2002.tb00186.x.

Devitt, M. (2006a). *Ignorance of language*. Oxford, UK: Clarendon Press.

Devitt, M. (2006b). Intuitions in linguistics. *British Journal for the Philosophy of Science*, *57*(3), 481–513. https://doi.org/10.1093/bjps/axl017.

Devitt, M. (2009). On determining what there isn't. In D. Murphy & M. A. Bishop (Eds.), *Stich and his critics* (pp. 46–61). Oxford, UK: Wiley-Blackwell.

Devitt, M. (2010). Linguistic intuitions revisited. *British Journal for the Philosophy of Science*, *61*(4), 833–865. https://doi.org/10.1093/bjps/axq018.

Devitt, M. (2011a). Experimental semantics. *Philosophy and Phenomenological Research*, *82*(2), 418–435. https://doi.org/ppr201182222

Devitt, M. (2011b). Deference and the use theory. *ProtoSociology*, *27*, 196–211. https://doi.org/10.5840/protosociology20112711.

Devitt, M. (2012a). The role of intuitions. In G. Russell & D. G. Fara (Eds.), *Routledge companion to the philosophy of language* (pp. 554–565). New York: Routledge.

Devitt, M. (2012b). Whither experimental semantics? *Theoria*, *27*(1), 5–36.

Devitt, M. (2012c). Semantic epistemology: Response to Machery. *Theoria*, *27*(2), 229–233. https://doi.org/theoria2012271110.1387/theoria.6225

Devitt, M. (2015a). Relying on intuitions: Where Cappelen and Deutsch go wrong. *Inquiry*, *58*(7–8), 669–699. https://doi.org/10.1080/0020174X.2015.1084824.

Devitt, M. (2015b). Testing theories of reference. In J. Haukioja (Ed.), *Advances in experimental philosophy of language* (pp. 31–63). London: Bloomsbury Academic.

Devitt, M. (2015c). Should proper names still seem so problematic? In A. Bianchi (Ed.), *On reference* (pp. 108–143). Oxford, UK: Oxford University Press.

Devitt, M., & Sterelny, K. (1999). *Language and reality: An introduction to the philosophy of language*, 2nd edn (1st edn 1987). Oxford, UK: Blackwell Publishers.

Domaneschi, F., Vignolo, M., & Di Paola, S. (2017). Testing the causal theory of reference. *Cognition*, *161*, 1–9. https://doi.org/10.1016/j.cognition.2016.12.014.

Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, *25*(5), 717–742. https://doi.org/10.1080/09515089.2011.627538.

Gernsbacher, M., & Kaschak, M. (2003). Language comprehension. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (vol. 2, pp. 723–725). London: Nature Publishing Group.

Gordon, P. (1998). The truth-value judgment task. In J. De Villiers, C. McKee, & H. Smith Cairns (Eds.), *Methods for assessing children's syntax* (pp. 211–231). Cambridge, MA: MIT Press.

Gross, S., & Culbertson, J. (2011). Revisited linguistic intuitions. *British Journal for the Philosophy of Science*, *62*(3), 639–656. https://doi.org/0.1093/bjps/axr009

Horwich, P. (1990). *Truth*, 2nd edn, 1998. Oxford, UK: Clarendon Press.

Horwich, P. (1998). *Meaning*. Oxford, UK: Clarendon Press.

Horwich, P. (2005). *Reflections on meaning*. Oxford, UK: Clarendon Press.

Ichikawa, J., Maitra, I., & Weatherson, B. (2012). In defense of a Kripkean dogma. *Philosophy and Phenomenological Research*, *85*(1), 56–68. https://doi.org/10.1111/j.1933-1592.2010.00478.x.

Jackman, H. (2009). Semantic intuitions, conceptual analysis, and cross-cultural variation. *Philosophical Studies*, *146*(2), 159–177. https://doi.org/10.1007/s11098-008-9249-6.

Kripke, S. A. (1979). Speaker's reference and semantic reference. In P. A. French, T. E. Uehling Jr, & H. K. Wettstein (Eds.), *Contemporary perspectives in the philosophy of language* (pp. 6–27). Minneapolis, MN: University of Minnesota Press.

Kripke, S. A. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Ludwig, K. (2007). The epistemology of thought experiments: First-person approach vs. third-person approach. *Midwest Studies in Philosophy*, *31*(1), 128–159. https://doi.org/10.1111/j.1475-4975.2007.00160.x.

Machery, E. (2012a). Expertise and intuitions about reference. *Theoria*, *27*(1), 37–54. https://doi.org/theoria 20122712

Machery, E. (2012b). Semantic epistemology: A brief response to Devitt. *Theoria*, *27*(2), 223–227. https://doi.org/10.1387/theoria.6223.

Machery, E. (2014). What is the significance of the demographic variation in semantic intuitions? In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 3–16). New York: Routledge.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), 1–12. https://doi.org/10.1016/j.cognition.2003.10.003.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2013). If folk intuitions vary, then what? *Philosophy and Phenomenological Research*, *86*(3), 618–635. https://doi.org/10.1111/j.1933-1592.2011.00555.x.

Machery, E., Olivola, C. Y., & de Blanc, M. (2009). Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis*, *69*(4), 689–694. https://doi.org/10.1093/analys/anp095.

Machery, E., & Stich, S. P. (2012). The role of experiments. In G. Russell & D. G. Fara (Eds.), *Routledge companion to the philosophy of language* (pp. 495–512). New York: Routledge.

Machery, E., Sytsma, J., & Deutsch, M. (2015). Speaker's reference and cross-cultural semantics. In A. Bianchi (Ed.), *On reference* (pp. 62–76). Oxford, UK: Oxford University Press.

Martí, G. (2009). Against semantic multi-culturalism. *Analysis*, *69*(1), 42–48. https://doi.org/10.1093/analys/ann007.

Martí, G. (2012). Empirical data and the theory of reference. In W. P. Kabasenche, M. O'Rourke, & M. H. Slater (Eds.), *Reference and referring: Topics in contemporary philosophy* (pp. 62–76). Cambridge, MA: MIT Press.

Martí, G. (2014). Reference and experimental semantics. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 17–26). New York: Routledge.

Nado, J., & Johnson, M. (2016). Intuitions and the theory of reference. In J. Nado (Ed.), *Advances in experimental philosophy and philosophical methodology* (pp. 125–154). London: Bloomsbury Academic.

Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind*, *125*(497), 145–175. https://doi.org/10.1093/mind/fzv196.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs analytic cognition. *Psychological Review*, *108*(2), 291–310. https://doi.org/10.1037/0033-295X.108.2.291.

Ostertag, G. (2013). The "Gödel" effect. *Philosophical Studies*, *166*(1), 65–82. https://doi.org/10.1007/s11098-012-0021-6.

Pickering, M. J. (2003). Parsing. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (vol. 3, pp. 462–465). London: Nature Publishing Group.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, UK: Cambridge University Press.

Stich, S. P. (1996). *Deconstructing the mind*. New York: Oxford University Press.

Stich, S. P. (2009). Reply to Devitt and Jackson. In D. Murphy & M. A. Bishop (Eds.), *Stich and his critics* (pp. 190–202). Oxford, UK: Wiley-Blackwell.

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.

Sytsma, J., & Livengood, J. (2011). A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy*, *89*(2), 315–332. https://doi.org/10.1080/00048401003639832.

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the land of the rising sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology*, *6*(2), 213–230. https://doi.org/10.1007/s13164-014-0206-3.

Tannenhaus, M. T. (2003). Sentence processing. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (vol. 3, pp. 1142–1148). London: Nature Publishing Group.

Thornton, R. (1995). Referentiality and wh-movement in child English: Juvenile *D-Link*uency. *Language Acquisition*, *4*(2), 139–175. https://doi.org/10.1080/10489223.1995.9671662.

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, *23*(3), 331–355. https://doi.org/10.1080/09515089.2010.490944.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Coding instructions: Elicited usage tests.

Appendix A: Attention check question

Participants were given the following prompt before beginning the survey[27]:

In order to facilitate our research, we are interested in knowing certain facts about you. Specifically, we are interested in whether you take the time to read directions; if not, then the data we collect based on your responses will be invalid. So, in order to demonstrate that you have read the instructions, please ignore the next question (i.e., don't answer it), and simply write "I have read the instructions" in the box labeled "Any comments or questions?" Thank you very much. Have you attended university?

Yes
No
Any comments or questions?

Appendix B: Confidence ratings

We wanted to make sure that the high average confidence ratings we observed in the forced choice tasks were really a sign that participants had high confidence in their responses. Therefore, we looked at the number of participants who reported confidence in the mid-to-high end of the scale (5–9), and the number who reported confidence in the low-to-mid end of the scale (0–4). In each condition, more participants reported mid-to-high confidence than low-to-mid confidence. This outcome differed significantly from an even split of mid-to-high and low-to-mid responses (50% mid-to-high responses):

Condition	Low (0–4)	High (5–9)	*N*	*X*	*p*
TCC TVJD (great astronomer)	5	17	22	6.545	0.011
TCC TVJCH (liar)	4	27	31	17.065	<.001
TCC Intuition	1	19	20	16.2	<.001
Ambiorix TVJ-CH (historical figure)	1	25	26	22.154	<.001
Ambiorix TVJ-D (legendary person)	2	19	21	13.762	<.001
Ambiorix Intuition	3	21	24	13.5	<.001
Collapsed Across Conditions	16	128	144	87.111	<.001

Might this result have been skewed by a large number of "middle range" responses? To check, we removed "4" and "5" responses, comparing only responses in the "low" (0–3) and "high" (6–9) ranges. Since in three of the questions, zero participants responded in "low" confidence range, we could not perform a chi-square goodness-of-fit test for each question. We instead collapsed across conditions. There were 115 "high" confidence responses, and just 9 "low" confidence responses. As before, this was significantly different from a chance distribution of 50% "high" confidence responses ($N = 144$, $X = 87.111$, $p < .001$).