COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12979

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12979

Testing the Reference of Biological Kind Terms

Michael Devitt, Brian C. Porter

City University of New York, Graduate Center

Received 11 September 2020; received in revised form 11 February 2021; accepted 29 March 2021

Abstract

Recent experimental work on "natural" kind terms has shown evidence of both descriptive and non-descriptive reference determination. This has led some to propose ambiguity or hybrid theories, as opposed to traditional description and causal-historical theories of reference. Many of those experiments tested theories against referential intuitions. We reject this method, urging that reference should be tested against usage, preferably by elicited production. Our tests of the usage of a biological kind term confirm that there are indeed both descriptive and causal-historical elements to the reference determination of some natural kind terms. We argue that to accommodate our results and earlier ones, we should abandon the common assumption that any one theory of reference fits all natural kind terms. Rather, it is likely that some terms are descriptive, some causal-historical, some ambiguous, and some hybrid. This substantive conclusion is accompanied by a methodological one. Our experiments, like some earlier ones, found participants contradicting both each other and themselves. We argue that these contradictions indicate a lack of linguistic competence with the term. We conclude that these experiments have been faulty, because they test terms that are novel to participants and/or use fantastical vignettes. We provide some suggestions for future research.

Keywords: Reference of natural kind terms; Experimental semantics; Linguistic usage; Ambiguity theory; Hybrid theory; Referential intuitions; Truth value judgments; Elicited production

1. Introduction

1.1. Theories of reference

A theory of reference tells us in virtue of what a term picks out its referent. In the case of the proper name "Einstein," the theory explains its reference to the famous physicist; in the case of the biological kind term "tiger," its reference to certain striped carnivorous felines; in the case of the chemical kind term "water," its reference to the clear potable liquid that

Correspondence should be sent to Brian Cross Porter, City University of New York, Graduate Center, 365 5th Ave, Room 7113, New York, NY 10016. E-mail: bporter@gradcenter.cuny.edu

we drink and swim in. This paper is about theories of reference for "natural kind terms,"[1] particularly biological kind terms, and about the methodology of testing such theories.

Prior to the 1970s, virtually all theories of reference were description theories of one sort or another. A description theory takes the reference of a term to be determined by the identifying descriptions that competent speakers associate with the term, hence typically by descriptions of observable superficial properties.

In the 1970s, Kripke (1980) and Putnam (1973, 1975) convinced many philosophers that description theories of reference were incorrect for terms like "Einstein," "tiger," and "water." Kripke and Putnam looked at several real and hypothetical cases, and argued that, intuitively, such terms refer even when speakers do not associate appropriate identifying descriptions with them. The paradigmatic examples were proper names: Kripke argued that, for example, the name "Einstein" refers to the famous physicist Einstein, despite the fact that many competent speakers —perhaps most competent speakers—simply do not know enough about Einstein to associate any description with "Einstein" that identifies him.

Kripke argued similarly about natural kind terms, including biological terms like "tiger." Even if ordinary competent speakers *did* associate a description that identifies tigers—perhaps "large four-legged feline with orange and black stripes"—it would not be in virtue of this association that "tiger" refers to tigers. For, speakers could turn out to be wrong about tigers. Kripke (1980: 120) considers a scenario in which we discover that the animals thought to be tigers are really three-legged and not striped but, due to an optical illusion, have *appeared* to be four-legged and striped. The associated description would then be false of these animals and yet, Kripke claims, "tiger" would still refer to them.

Similarly, Putnam argues that the reference of terms like "water" cannot be determined by factors internal to the speaker's mind, like her association of descriptions of water. Putnam introduces the "Twin Earth" thought experiment, in which a liquid on another planet ("Twin Earth") has all of the superficial properties of water familiar to competent users of "water," but is not made of H_2O. The standard intuition among philosophers of language has been that the Twin Earth liquid is *not* water, and so "water" cannot refer to the Twin Earth liquid *even though it fits the description* that competent speakers associate with the term.

Arguments like these led many philosophers to abandon description theories of natural kind terms in favor of causal-historical theories. On this sort of theory (Devitt & Sterelny, 1999: 88–90), the reference of a term is "fixed" in an initial dubbing and (probably) later "groundings." The people who fix the reference of the term use the term to communicate, and thereby pass on the ability to use the term with that reference. Later uses of a term thus "borrow" their reference from the previous uses of the term in a causal network that goes back to earlier fixings. Even if speakers' associated descriptions do not apply to anything at all, their term still refers to whatever kind the term is causally connected to. A biological term like "tiger" does not refer to an animal in virtue of its having the superficial properties picked out by speakers' associated descriptions but rather in virtue of its having the same deep structural properties (the same underlying "essence") as the animals that ground the network.

However, these antidescriptivist conclusions and, more importantly, the methodology that led to them have been subject to a severe challenge from a number of "experimental semanticists," a challenge to be discussed in Section 1.2. These philosophers have found evidence in their experiments of both descriptivist and causal-historical reference determination. This has led them to propose "ambiguity" and "hybrid" theories. The terminology has been somewhat inconsistent, but on our usage, the "Ambiguity Theory" takes terms to have two distinct linguistic meanings, one descriptivist and one nondescriptivist. So each token of, say, "water" or "tiger" will have its reference either determined descriptively or determined nondescriptively, depending on which meaning is in play.[2] On the "Hybrid Theory," a term has just one linguistic meaning which determines the reference of any token partly by what its associated descriptions are true of and partly causal-historically.[3] In this paper, we present experiment results testing these theories against the usage of biological kind terms. We will ultimately argue that none of these theories fits *all* natural kind terms and advocate for an eclectic approach on which each theory gets it right for *some* natural kind terms.

In addition to the above theories of reference, there is another phenomenon that may play a role in explaining the experimental results: indeterminacy.[4] In some cases, there may be no fact of the matter whether the term refers to this or that. Indeterminacy in reference has been mentioned by some, but it has not featured prominently in the discussion. Each of the above theories can allow for indeterminate reference in at least some cases. As we will see, some of the theories (especially the Hybrid Theory) will need to appeal to indeterminacy to explain our results.

1.2. The challenge of experimental semantics

This challenge is both methodological and substantive. The seminal work is by Machery, Mallon, Nichols, and Stich (2004) on proper names. They questioned the "armchair" methodology of Kripke and Putnam that relies on philosophers' own referential intuitions (RIs) to test theories of reference. Instead, Machery et al. tested theories of reference against *the folk's* referential intuitions, a method (RI) they helped to popularize. In RI tests, participants are presented with a vignette and asked who a speaker is talking about (or what a term refers to). Machery et al. found that though many folk do indeed have intuitions predicted by the causal-historical theory, many others have descriptivist intuitions: there is considerable variation, including cultural variation. Since then, there has been an explosion of experimental work testing theories of reference.[5] Much of this work has focused on proper names, but we will attend only to that on natural kind terms. These experiments have cast doubt on the idea that reference is generally to be explained just descriptively or just nondescriptively.

We start with some experiments that predate Machery et al. (2004). Braisby, Franks, and Hampton (1996) conducted experiments based on Putnam's Twin Earth thought experiment. Instead of RI, they used another popular method: truth value judgment (TVJ) tests. In such tests, participants are presented with a vignette that includes a certain term, and then asked whether a statement using that term is true or false (or whether they agree with the statement, or to what extent they agree with the statement, etc.). Braisby et al., testing "cat" and other common natural kind terms, found that only 58% of responses in one experiment, and only 76% in the other experiment, were consistent with the predictions of the causal-historical theory. Interestingly, they also found that participants were willing to contradict themselves: A significant number of participants assigned the same truth value to statements like "Tibby is a cat, though we were wrong about her being a mammal" and "Tibby is not a cat, though she is a robot controlled from Mars." Braisby et al. infer that participants must be switching between two different interpretations of the term. They, therefore, conclude that their results provide evidence against the Kripke–Putnam view of natural kind terms, and in favor of the Ambiguity Theory.

However, Jylkkä, Railo, and Haukioja (2009), using RI, ran an experiment in which they gave participants "the option to answer explicitly ambiguously": they could choose "'on the one hand yes, on the other hand no'" (p. 55). Participants overwhelmingly chose one of the *non*-ambiguous answers. This suggests that despite the interpersonal variation in the responses, participants do not seem to believe that there are two distinct "interpretations" of the natural kind term. For, if they did, they would presumably choose the "ambiguous response."

Genone and Lombrozo (2012), also using RI, present results that they take to support the Hybrid Theory of reference: "most participants use both descriptive and causal information in making reference judgments" (2012; 795).[6] This means that the Ambiguity Theory of reference is incorrect, since the Ambiguity Theory predicts that participants will use *either* descriptive *or* causal information in each case, but not both.

Nichols, Pinillos, and Mallon (2016) used both RI and TVJ tests to argue for the Ambiguity Theory. Their results suggest that contextual factors influence whether participants make descriptivist judgments or causal-historical judgments about the reference of natural kind terms. As in Braisby et al. (1996), some participants in the Nichols et al. experiments seemingly contradict themselves. Nichols et al. argue that this is best explained by the Ambiguity Theory, on which participants can switch back and forth between a clearly descriptivist interpretation of the term and a clearly causal-historical interpretation. Nichols et al. also suggest that in some cases, it will "simply be indeterminate which reference-fixing mechanism is in play" (2016: 161), and so there need not always be a fact of the matter as to which of the two interpretations is being used.

Tobia, Newman, and Knobe (2020), using TVJ, also provide evidence that participants are willing to make different judgments in different contexts. For example, participants were more likely to use superficial properties in making natural kind membership judgments in a legal context, but more likely to use deep structural properties in making membership judgments in a scientific context. They take this "dual character pattern" of judgments about kind membership to support the Ambiguity Theory of natural kind terms, on which these terms have two distinct "senses": a descriptive sense and a causal-historical sense.

1.3. Assessment

Some of the cited experiments exemplify the common practice of testing theories of reference against the folk's referential intuitions, RI. Martí (2009, 2012, 2014) and Devitt (2011, 2012a, 2012b), in responding to Machery et al. (2004), have argued against this methodology: "Relying on referential intuitions is not scientifically respectable," regardless of whether they come from the folk or philosophers (Devitt, 2015; 53).[7] Linguistic competence does not provide privileged access to the truth about reference. Rather, RIs are empirical judgments about linguistic data. Whether or not these are good evidence depends on whether or not the person intuiting is reliable about such semantic matters, which is itself an empirical issue. Apart from this objection in principle to RI, referential intuitions have been found to be unreliable in practice. They have sometimes been shown to be at odds with tests of usage.[8] They have often been shown to be susceptible to disturbing wording effects: Minor changes in the wording of the prompt can have significant impact on the intuitions participants report.[9] These criticisms of RI make us dubious of the results of Jylkkä et al. (2009), Genone and Lombrozo (2012), and some of the results of Nichols et al. (2016).

In rejecting RI, Martí and Devitt proposed an alternative methodology: Rather than testing theories of reference against *intuitions about* linguistic usage, test them directly against linguistic usage itself. And that is what we proposed to do in our experiments.

What about the value of TVJ tests, used in several of the cited experiments? Their value came into question in discussions of a paper by Machery, Olivola, and Blanc (2009). These researchers claimed to test usage in responding to Marti's (2009) criticism of Machery et al. (2004), yet their tests were in fact TVJ. Martí objected that these are *not* tests of usage; indeed, they raise in her mind "pretty much the same concerns" (2012, p. 74) as did Machery et al.'s (2004) use of RI. Devitt and Porot (2018) side with Machery et al. (2009) on this. They argue that TVJ tests exploit the disquotational property of the truth predicate: Asserting that "*p*" is true is equivalent to asserting that *p*. So, testing whether or not participants assert that a statement is true is indeed a test of usage: It *implicitly* uses the statement. However, a TVJ test "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*" (2018: 1561), rather than allowing participants to produce their own statements in a pure test of usage.

Accepting this view of TVJ tests inclines us to think that the results of the experiments by Braisby et al. (1996), Tobia et al. (2020), and most of the experiments by Nichols et al. (2016) should be taken cautiously as evidence of referential reality. But the seemingly contradictory responses in Braisby et al. (1996) and Nichols et al. (2016) raise a worry: Perhaps these experiments are not good tests of reference; perhaps, we should embrace a "Faulty Test Hypothesis." In Section 5.2, we will.

Setting that methodological worry aside, what do these experimental results show? Their striking message is that, contrary to standard opinion, the referential reality of natural kind terms is neither simply descriptivist nor simply causal-historical.

The most common attempt to accommodate this has been the Ambiguity Theory: These terms have both a descriptivist meaning and a nondescriptivist meaning. This theory gives a nice explanation for the contextual variations in usage demonstrated by Tobia et al. It also has a story, though not necessarily a convincing one, for the variations amounting to contradictions demonstrated by Braisby et al. (1996) and Nichols et al. (2016). But it has a problem. What determines which of the two meanings a given token has? Nichols et al. rightly see this as a matter of which meaning the speaker intends (2016; 161). But, they note,

> when people are using uncontested natural kind terms, it is far from clear that they intend one of the reference conventions rather than the other. In typical uncontested

> cases of natural kind terms, the causal-historical and the descriptive mechanisms are in 'harmony '. When I say, 'There is water in the Hoover dam,' what I say is true under both reference conventions for 'water'. So in this case, the lack of a determinate answer to 'which reference convention is operative' is harmless. (163-164)

It may be harmless to the Ambiguity Theory of natural kind terms if *occasionally* there is no determinate matter of fact which meaning is intended but it is not harmless if this is *normally* the case. If a term in our language is to be plausibly declared ambiguous, whether homonymous or polysemous, it should be *regularly* used one way and *regularly* used another, reflecting *two linguistic conventions*.[10] This does not seem to be the case with the folk's use of terms like "water." It seems rather that the idea of such term having two meanings has never occurred to the folk. And this appearance gets some confirmation from Jylkkä et al.'s (2009) experiment, albeit an RI test, in which only 17% of participants' selected the "ambiguous response." This is a serious problem for the Ambiguity Theory.

Hybrid theories are another attempt to accommodate the results: A natural kind term has only one meaning, which avoids the "intended meaning" problem, but that one meaning yields two factors in reference determination: Reference is partly determined descriptively and partly nondescriptively. Nichols et al. criticize this view:[11] They "do not really see" what this unified meaning could be that "would explain the shifts in uses" revealed in their experiments (162 n.27). Indeed, if a term has only one meaning, there should not be the observed variations in usage. This is a serious problem for the Hybrid Theory.[12]

However, it is important to note that the Hybrid Theory has a certain indeterminacy built into it, which can help it explain the results. Normally, each of the reference-determining factors pull in the same direction. But when they do not, as in Nichols et al. (2016), there can be no determinate matter of fact about the term's reference.

Both the Ambiguity and Hybrid theories are defended by appeal to the results of RI and TVJ tests. We are very critical of RI and urge that theories should instead be tested against *usage*. TVJs do this, but imperfectly because they prime certain usage. We thought it possible that the messy results that led to the Ambiguity and Hybrid theories were due to the flaws of RI and TVJ tests, and that a "pure" test of usage would show that the Kripke–Putnam causal-historical theory was correct after all.

The method of "elicited production" (EP) provides a "pure" test. In EP tests, participants are prompted to produce statements using a term that appears in a vignette, so that linguistic usage of that term can be directly examined. The method had been used to test theories of proper names (Devitt & Porot, 2018; Domaneschi, Vignolo, & Di Paola, 2017) but not, to our knowledge, to test theories of natural kind terms. Our plan was to use it.

1.4. Our aims

Our aim in these experiments was to use EP to test the extant theories of reference for biological kind terms. We hoped that this pure and direct test of linguistic usage would support the Kripke–Putnam causal historical theory of reference. The apparent contradictions and disagreements found in previous experiments could then be explained away as arising from flaws in RI and TVJ tests. However, this was not what we found. So, we ran some follow-up TVJ experiments. The cumulative results of these experiments led us, first, to a lot of rethinking about reference and to the conclusion that the causal-historical theory is probably not true for many natural kind terms in common folk use (5.1). The results lead us, second, to embrace the methodological Faulty Test Hypothesis for many experiments, including ours (5.2).

2. Elicited production test

2.1. Vignette

We used the following vignette, based on those used by Nichols et al. (2016):

> Researchers in the Middle Ages, who are now considered early biologists, described a distinctive kind of animal which they called "catoblepas." They claimed that these animals were like bulls, but with heads so heavy they had to keep their heads down at all times. These early biologists also thought that these animals had scales on their backs, and that their breath was poisonous to humans. We now know, of course, that there have never been any animals that meet this description. Historians have recently discovered, however, that the descriptions arose from some of those biologists observing some actual animals, and coming to mistaken views about them. The animals they observed were in fact wildebeests, which are migratory antelopes that still roam Africa today. Wildebeests are not bulls, but they do have large heads with horns. They do not keep their heads down at all times, but they often hold their heads low to the ground in order to eat grass. They do not have scales, but it is now thought that the biologists wrongly took their rough manes for scales. Furthermore, the diet of wildebeests does include many poisonous plants, and the early biologists seem to have mistakenly thought that this would make their breath poisonous.

Some differences between our vignette and the initial one used by Nichols et al. (2016) are worth emphasizing. (a) We expanded on the role of wildebeests in the origins of the false description, and on the causes of the mistake. We thus provided more of the information that causal-historical theorists think is relevant to reference. (b) We had a worry, supported by a pilot test, that participants would interpret catoblepas as "mythical" animals loosely *based on* wildebeests, rather than as the posit of a mistaken (proto-)scientific theory. So, we removed Nichols et al.'s talk of catoblepas having a "death" gaze. We also explicitly called the researchers biologists and had them observe wildebeests directly. (c) We stated that wildebeests "still roam Africa today," because some participants in the pilot tests seemed to think that wildebeests do not exist. (d) We mentioned *but did not use* "catoblepas." Devitt and Porot (2018: 1562) point out that a vignette's use of a term under investigation has biased some experiments against description theories. It is not clear that this was a problem with Nichols et al.'s vignette (which uses "catoblepas"), but we were taking no chances.

Table 1
Results of initial EP experiments

% Responses Inconsistent With Descriptivism	*N*	Independence From Chance (*p*- Value)
57%	53	.56

We used this vignette in all of our experiments.

2.2. *Methods*

We recruited 80 participants through Amazon MechanicalTurk$^{\text{TM}}$, who were compensated for their participation.[13] After answering demographic questions and a mandatory attention-check question, participants were directed to another page with our vignette.

Underneath the vignette, participants were given the following question:

> In light of this information, please say what if anything you think textbooks on the history of science should say about Catoblepas. Please give reasons but be brief.[14]

Description theories of reference predict responses indicating that catoblepas do not exist, since nothing exists that fits the associated descriptions. Antidescriptivist theories like the causal-historical theory predict responses that take Catoblepas to exist because they are wildebeests that have been misdescribed.

We coded participant responses according to instructions written before the study.[15] Responses judged to be in accordance with descriptivist predictions were coded 0; responses judged to be in accordance with antidescriptivist predictions were coded 1; responses which were not clearly in accordance with either category were coded D and discarded. We had anticipated that we would compare our coding with those of trained independent coders but, given the results below, we decided that this was unnecessary.

2.3. *Results*

None of the participants needed to be removed for failing the attention check. Thirty responses were judged 1, antidescriptivist; 23 were judged 0, descriptivist; and 27 were discarded. The results are not significantly different from chance responses (50% descriptivist) using a two-sided Fisher's exact test; see Table 1.

Most of the responses were either emphatically a descriptivist 0, emphatically an antidescriptivist 1, or an obviously irrelevant D. For example:

Emphatic 0: "Catoblepas never existed. Early biologists did not use their skills enough to produce valid research."
Emphatic 1: "They should say that Catoblepas are wildebeests and they still roam Africa."
Obvious D: "Literally anything. I've never heard them even mention them before."

In light of this, it was clear that no reasonable coding would come close to differing *significantly* ($p<.05$) from a roughly 50-50 split between descriptivist and antidescriptivist responses. So, we decided against training independent coders. Nonetheless, we accepted the offer of a philosophy graduate student to code the responses. He received no training or information about the study beyond the coding instructions. He coded 46.5% of the responses that he did not discard antidescriptivist, and mostly agreed with our codings (58 out of 80 responses)[16] The only disagreements were over whether or not to discard a response. This reflects the genuine difficulty of coding some of these responses, particularly without training. Consider this response, for example:

> "THey [sic] are a good example of ancient people being correct but not knowing they are."

After much debate, we coded this 1 because if the participant did not take catoblepas to exist, then there was nothing that the ancient people could be correct about. (The student coded this D.) And consider:

> "It should be mentioned that their breath was believed to be poisonous to humans. However, this was proved to be false."

We found the precise reference of "this" unclear and coded this D. (The student coded it 1.) Finally, consider this interesting response:

> "They were actually wildebeasts [sic] so there is no such thing as Catoblepas."

This implies *both* that catoblepas are wildebeests, *and* that catoblepas do not exist. So, it is a D. (The student coded it 0.) This apparently contradictory response presaged what was to come with our TVJ tests.

2.4. Discussion

The results of the EP test were neither what we expected nor what we had hoped for. Far from showing that the Kripke–Putnam causal-historical theory is correct after all, they confirmed the main conclusions of earlier RI and TVJ tests: Reference is to be explained partly descriptively and partly causal-historically (nondescriptively).

The Ambiguity Theory is one way of doing so. It has the serious problem of lacking evidence of two regular uses (1.3), but it has a neat explanation of our EP results. It can say that the context in this test was neutral, so that it did not prompt one meaning of "catoblepas" over the other. So not surprisingly, about half of participants resolved the ambiguity one way, half, the other.[17]

The Hybrid Theory can attempt to explain the EP results along similar lines. As noted above (1.3), the Hybrid Theory builds a certain indeterminacy into reference: Where the two reference-determining factors, causal-historical and descriptive, pull in different directions, as they do with "catoblepas" in our EP test, reference is indeterminate. The Hybrid Theory can then mimic the Ambiguity Theory's explanation: About half of participants resolved the indeterminacy one way, half, the other.

However, convincing or unconvincing these explanations may be, it seems certain that neither a pure description nor a pure causal-historical theory could explain all our EP results. But there is another theory that could, a "Different Idiolects Theory": Some participants use "catoblepas" only with a descriptivist meaning, some, only with a causal-historical meaning, such that a "pure" theory is correct for each idiolect. However, this theory does not fit well with earlier results, particularly the context relativity and contradictory results found in Nichols et al. (2016) and Tobia et al. (2020). And, of course, it is a highly implausible view of what seems to be the one speech community; hence, to our knowledge, nobody has promoted it.

Finally, it is possible that there is no relevant variation in usage. Our experiment may not in fact be a good test of reference against usage. We aired a similar methodological worry in Section 1.3 about earlier experiments, leading to the Faulty Test Hypothesis. This will be discussed in Section 5.2.

We concluded that either our experiment was not a good test, or the correct theory of natural kind terms is not simply descriptivist or simply causal-historical. The results of our EP test gave us little evidence for one explanation over the other, so we decided that we needed more tests to clarify the choice. We started with a forced-choice truth-value judgment (TVJ) test.

3. Truth value judgment test 1 (TVJ1)

3.1. Methods

We recruited 80 participants for our first TVJ test and presented them with the same vignette. After answering demographic questions and a mandatory attention-check question, participants were directed to another page with our vignette. This time, participants were given one forced-choice truth value judgment question. Half of participants were given a statement for which the causal-historical theories would predict responses of "true" and descriptive theories, "false" (TVJ-CH). The other half were given a statement for which descriptive theories would predict responses of "true" and causal-historical theories, "false" (TVJ-D). This was done to counteract participants' bias toward answering "true."

TVJ-CH: On the basis of this historical report, please say whether the following statement is true or false:

> Catoblepas were real animals that were falsely described in the Middle Ages.
>
> How confident are you in your response, on a scale of 1 to 10?
> 1 = not at all confident, 10 = very confident.
> 1 2 3 4 5 6 7 8 9 10

TVJ-D: On the basis of this historical report, please say whether the following statement is true or false:

> Catoblepas did not really exist.

Table 2
Results of initial TVJ experiments

Prompt	% Responses Inconsistent With Descriptivism	Average Participant Confidence	*N*	Independence From Chance *p*-Value
TVJ-CH	89%	8.368	38	$p = .00034$*
TVJ-D	17.5%	8.325	40	$p = .00411$*

*Differs significantly ($p<.05$) from chance (50%).

Table 3
Comparisons of EP and TVJ results

Comparison	*N*	*p*
EP versus TVJ-CH	91	$p = .00092$
EP versus TVJ-D	93	$p = .013$
TVJ-D versus TVJ-CH	78	$p = .519$

> How confident are you in your response, on a scale of 1 to 10?
> 1 = not at all confident, 10 = very confident.
> 1 2 3 4 5 6 7 8 9 10

3.2. Results

Two participants were removed from the study because they failed the mandatory attention check. Of the 38 participants who were presented with the TVJ-CH question, 34 (89%) answered "true"; 4 (11%) answered "false." Of the 40 participants who were presented with the TVJ-D question, 33 (82.5%) answered "true"; 7 (17.5%) answered "false." The responses to both TVJ-D and TVJ-CH were significantly different from chance (50%). The results can be seen in Table 2.

Comparisons to the EP results can be found in Table 3.

The majority of participants reported high confidence in their answers: 92% rated their confidence at least a 7 out of 10. There was no significant correlation between how participants answered the TVJ question and how confident they claimed to be.

3.3. Discussion

These two TVJ results are strikingly different from our EP ones and from each other. Whereas our EP results were roughly 50-50, our TVJ-CH ones are overwhelmingly antidescriptivist, and our TVJ-D results are overwhelmingly descriptivist. As noted in Section 3.1, we decided to do both a CH and a D test "to counteract participants' bias towards answering 'true'." However, the effect of this bias should be relatively small,[18] and it alone cannot explain our results.

Our TVJ1 results are particularly striking when compared to those of Devitt and Porot (2018) on proper names. Although our TVJ-CH results are similarly antidescriptivist to theirs, our TVJ-D ones are not. Their TVJ-D results for the name “Tsu Ch’ung Chih” are 95% antidescriptivist, and for the name “Ambiorix,” 81%.[19] In contrast, our TVJ-D results for “Catoblepas” are only 17.5% antidescriptivist; they are 82.5% *descriptivist*. So, in their experiment unlike ours, there was no apparent bias toward agreement with the descriptivist statement.

The Ambiguity Theory seems to have a good explanation of these results. Consider any utterance containing an ambiguous expression. In a context that does not otherwise favor one interpretation of the expression over another, hearers tend to resolve the ambiguity charitably so that the utterance comes out true.[20] So the Ambiguity theory can say that this is what happened in our TVJ tests: Participants tended to interpret “catoblepas” nondescriptively in the CH test so that its statement comes out true, and descriptively in the D test so that its statement comes out true. This, together with the already-noted agreement bias, explains the results.

However, it is not all good news for the Ambiguity Theory. The high confidence of the participants in their answers, whether those answers are descriptivist or nondescriptivist, is a *prima facie* problem. If “catoblepas” has two meanings that yield opposite answers to the TVJ1 questions, then participants should surely be aware of this. So, if they choose one answer over the other *only out of charity*, it is surprising that they should have high confidence that their answer is correct. Participants’ apparent high confidence instead suggests that they are simply not aware of two distinct meanings. This is in line with the results of Jylkkä et al. (2009: 55) discussed in Section 1.2, in which participants overwhelmingly rejected the explicitly “ambiguous response.”

According to the Hybrid Theory, “catoblepas” has only one meaning. That may seem to make explaining the TVJ variations in usage hopeless from the start. But the theory can attempt to exploit the built-in indeterminacy mentioned in 1.3 and 2.4 to mimic the Ambiguity Theory’s explanation of the results: Participants tend to resolve that indeterminacy charitably by taking the reference of “catoblepas” to be determined primarily by the causal-historical factor in the CH test, and primarily by the descriptive factor in the D test. But the high confidence of the participants is as problematic for this explanation as for the ambiguity explanation.

The unpromising Different Idiolects Theory (2.4) does not fare well either: it cannot explain why one TVJ result is descriptivist and the other antidescriptivist, thus contradicting each other. If there were two idiolects split roughly 50-50 among the participants, as our EP results would suggest, both TVJ1 results should have been roughly 50-50 descriptivist and antidescriptivist. Yet, neither TVJ1 was.

Finally, as before (1.3, 2.4), we must consider the possibility that the explanation of these contradictory results is simply that our tests are not successful tests of reference. Perhaps the Faulty Test Hypothesis is true; this will be discussed in Section 5.2.

In trying to interpret these results, it occurred to us that participants who responded to the D statement may not have thought of the possibility of a CH response, and vice-versa.[21] We, therefore, decided we needed to test *intra*participant, so that participants see both the D statement and the CH statement. This led us to a second TVJ test.

4. Truth value judgment test 2

4.1. Methods

We recruited 102 participants for our intraparticipant TVJ test. Participants answered the usual demographic and mandatory attention-check questions. Participants were divided into three groups:

Group 1 (TVJ-CHthenD) consisted of 31 participants. They were presented with the following three prompts, in order: the vignette and the TVJ-CH prompt; the vignette and the TVJ-D prompt; and the following prompt on a separate page:

> Please explain your answers to the previous two questions. Give reasons, but be brief.

Group 2 (TVJ-DthenCH) also consisted of 31 participants. They were presented with the same three prompts but with the order of the first two reversed, so that TVJ-D came before TVJ-CH.

Group 3 (TVJ-ChooseCHorD) consisted of 40 participants. They were presented with the vignette, and the following prompt:

> On the basis of this historical report, please say which of the following statements you think is true:
>
> Catoblepas did really exist, but were falsely described in the Middle Ages.
>
> Catoblepas did not really exist.
> How confident are you in your response, on a scale of 1 to 10?
> 1 = not at all confident, 10 = very confident.
> 1 2 3 4 5 6 7 8 9 10

This was not a forced choice question; participants were able to select both statements.

4.2. Results

In Group 1 (CHthenD), 28 participants (90.3%) answered “true” to the TVJ-CH prompt, and 3 (9.7%) answered “false.” Then, 15 participants (48.4%) answered “true” to the TVJ-D prompt, and 16 (51.6%) answered “false.” Twelve of the 31 participants (38.7%) apparently contradicted themselves: They answered “true” to both prompts; no participant answered “false” to both.

In Group 2 (DthenCH), 24 participants (77.4%) answered “true” to the TVJ-D prompt, and 7 (22.6%) answered “false.” Then, 24 participants (77.4%) answered “true” to the TVJ-CH prompt, and 7 (22.6%) answered “false.” Strikingly, 19 of the 31 participants (61.3%) apparently contradicted themselves: 18 answered “true” to both prompts; 1 answered “false” to both.

Table 4
Participant responses in follow-up TVJ experiments

a: ChthenD

		TVJ-D	
		True	False
TVJ-CH	True	12	16
	False	3	0

b: DthenCH

		TVJ-D	
		True	False
TVJ-CH	True	18	6
	False	6	1

c: ChooseCHorD

		Chose D?	
		Yes	No
Chose CH?	Yes	2	22
	No	16	0

In Group 3 (ChooseCHorD), 24 participants (60%) selected "Catoblepas did really exist, but were falsely described in the Middle Ages." Eighteen participants (45%) selected "Catoblepas did not really exist." Two participants selected both statements, indicating that they considered both statements true. Of the 38 who chose only one statement, 22 (57.9%) chose the antidescriptivist one, 16 (42.1%) the descriptivist one.

Group 3 participants reported high confidence; the mean rating was 8.1, with 85% of participants selecting a 7 or higher. Only three participants chose 5 or less; no participants chose 1 or 2. The two participants who selected both statements reported confidence levels of 7 and 8. Groups 1 and 2 were not asked to rate their confidence.

The results of the three groups can be seen in Tables 4a, 4b, 4c, and 5.

Interestingly, 50% of participants given the CHthenD and DthenCH prompts (31 out of 62) seemingly contradicted themselves: They either answered "true" to both statements, or in one case answered "false" to both statements. Here is a representative sample of explanations of apparently contradictory answers to TVJ-CHthenD:

A1. "There were indeed animals that had similarities to wildebeests. but they clearly were not wildebeests. the name given to these animals at the time were not in fact the correct name."

A2. "They gave the wildebeest a name, and that name is doesn't represent the true nature of the wildebeest."

A3. "Catobleepas [sic] were real animals but improperly described and as such were not actually a separate animal but a mistaken wildebeest."

A4. "The animals were wildebeests, but that name wasn't used. The name catobles [sic] was paired with descriptions that were incorrect, and since those attributes have never existed, the animal named that never existed."

Table 5
Results of follow-up TVJ experiments

Prompt	% Responses Inconsistent With Descriptivism	Average Participant Confidence	*N*	Independence From Chance *p*-Value
TVJ-CH (from DthenCH)	77%	N/A	31	$p = .062$
TVJ-D (from DthenCH)	23%	N/A	31	$p = .062$
TVJ-CH (from CHthenD)	90%	N/A	31	$p = .00162$*
TVJ-D (from CHthenD)	52%	N/A	31	$p = 1.0$
ChooseCHorD (including contradictions)	60%	8.1	40	$p = .5$
ChooseCHorD (excluding contradictions)	58%	8.1	38	$p = .646$

*Differs significantly ($p < .05$) from chance (50%).

A5.“The animals weren’t what researchers thought in the Middle Ages. They were wildebeests, which do exist. So the animals thought to exist did not, but were an actual species.”
A6.“They were real animals, but the name they were given were not real animals, they were wildebeests and therefor [sic] a differnt [sic] animal than what they named.”

Here is a representative sample of explanations of apparently contradictory answers to TVJ-DthenCH:

B1.“They don’t exist under that term, however, they were a real entity with a name that is currently used for such animals.”
B2.“The researchers saw wildebeests, which are real. They gave a false name to an animal that they thought existed and they got some facts wrong about this animal that doesn’t exist.”
B3.“They didn’t actually exist as animals, but the name described a real animal just a different one.”
B4.“They don’t exist because the people at the time didn’t know they were looking at wilderbeest [sic]. Also, their description of the animal was mixed with fantasy as well.”
B5.“The creature as described did not exist, but was an actual animal that was inaccurately described in history.”
B6.“The animal catoplebus [sic] did not exist. The wildebeast [sic] was mistakenly identified as this animal.”

Not all participants provided substantive explanations of their answers; one participant just wrote “no.” B3, for example, just seems confused: It is not clear how the name “catoblepas” could “describe” an animal other than catoblepas. But of the answers that provided substantive explanations of the apparent contradictions—about 18 in total—we abstracted three strains of thought: Participants think that catoblepas did not exist but nonetheless that they are real because

1. The catoblepas story arose from sightings of wildebeest, which are real; for example, A1, A5, B2, and B4;

2. Wildebeest were incorrectly named “catoblepas”; for example, A1, A2, A4, A6, B1, and B2;
3. Wildebeest were misidentified with catoblepas; for example, A3, A5, B5, and B6.

It is not immediately clear what conclusions can be drawn from these contradictory responses, or from the explanations of those responses. This will be discussed in the next section.

4.3. Discussion

4.3.1. Groups 1 and 2

Unsurprisingly, the results for the TVJ-CH prompt in CHthenD (90.3% “true”) are similarly antidescriptivist to those for it in TVJ1 (85.4% “true”); and the results for the TVJ-D prompt in DthenCH (77.4% “true”) are similarly descriptivist to those for it in TVJ1 (78.6% “true”). So just as in the earlier TVJ experiment, the CH and D answers are very different from our 50-50 EP ones and from each other. And our discussion in Section 3.3 applies to these results, too.

The striking new result is that 50% (31 of 62) of participants in groups 1 and 2 gave answers to their second question that were inconsistent with their answers to their first. None of the theories we have entertained can explain this. (1) The Ambiguity Theory explains each of the contradictory answers by claiming that the participant tends to resolve the ambiguity of “catoblepas” charitably to make the statement in an answer true. This requires *different* interpretations of “catoblepas” in the first and second answer; the participants must be alternating between the descriptivist and nondescriptivist interpretations. Yet the participants’ explanations of their answers show no sign of this: There is no indication that participants are even *aware* of any ambiguity. (2) The Hybrid Theory does not fare much better. As with the TVJ1 results (3.3), the Hybrid Theory must exploit the built-in indeterminacy to mimic the Ambiguity Theory’s explanation: the contradictory answers come from the participant’s charitable resolution of the indeterminacy. This requires that participants to switch between the different factors in determining reference rather than between the different meanings of the Ambiguity Theory. And this gives the Hybrid Theory a slight advantage. For, while none of the participants’ explanations can plausibly be read as demonstrating any awareness of two meanings, some could plausibly be read as demonstrating awareness of referential factors pulling in different directions (see, e.g., A2, A3, and B5). However, this alone is not enough to explain the contradictory results. (3) The Different Idiolect Theory can obviously not explain this variation of usage *within* a participant.

These contradictory responses, and the explanations given for them, are worrying. Could the participants really be so incompetent at telling whether catoblepas exist, using the information in the vignette? The responses cast doubt on the legitimacy of our tests, and of earlier tests of the reference of natural kind terms. The case for the Faulty Test Hypothesis mounts. We shall consider the case in Section 5.2.

4.3.2. Group 3

Just as in the EP test, the responses in ChooseCHorD show a nearly 50-50 split between descriptivist answers and antidescriptivist answers. Here, the confidence ratings pose an even larger problem for the Ambiguity and Hybrid Theories than before. Participants were presented with *both* D and CH sentences, so even charity cannot point to one unique answer. Participants would in effect be guessing at which meaning/factor was relevant, which should not result in high confidence in their answers. Only the unpromising Different Idiolects Theory—which cannot explain our other TVJ results—seems to have a good explanation of the group 3 results: Half of the participants use the term descriptively and half use the term causal-historically.

5. General discussion

5.1. Explaining the results

How can these results be explained? Each of the three theories, Ambiguity, Hybrid, and Different Idiolects, can explain some of the results. But none comes close to a complete explanation that covers the apparent contradictions. For that, a large role must be given to the methodological Faulty Test Hypothesis, to be discussed in Section 5.2. But, as we shall then see, the faultiness does not undermine the view that our results provide quite strong evidence of:

(a) Descriptivist and nondescriptivist reference determination of biological kind terms *within the community*.

This is supported by the following patterns:

1. Faced with both nondescriptivist and descriptivist options at once, participants' choices were *close to 50-50*, with only an insignificant preference for the nondescriptivist one (EP and ChooseCHorD).
2. Faced with the nondescriptivist statement without having been presented with the descriptivist statement, an extremely significant proportion of participants chose the *nondescriptivist* one (TVJ-CH+TVJ-CHthenD, $p = .000000632$).
3. Yet, faced with descriptivist statement without having been presented with the nondescriptivist statement, a highly significant proportion of participants chose the *descriptivist* one (TVJ-D+TVJ-DthenCH, $p = .00036$.)

Next, we take the results to provide quite strong evidence of:

(b) Descriptivist and nondescriptivist reference determination of biological kind terms *within individuals*.

This is supported by TVJ1, TVJ-CHthenD, and TVJ–DthenCH, including the participants' explanations. Note also that two participants in ChooseCHorD chose *both* the descriptivist and nondescriptivist statements.

The Ambiguity and Hybrid theories can explain both (a) and (b). The Different Idiolects Theory can explain (a) but not (b). We aim to motivate an eclectic approach to the explanation of (a) and (b), with a place for all these theories, but with particular emphasis on the Hybrid Theory. We shall start with some thoughts about reference relations and their origins.

It is important to keep the following piece of theoretical background in mind: *not all* the terms in any language could be covered by description theories. Description theories are "essentially incomplete" (Devitt & Sterelny, 1999: 60): They explain the reference of one term, say "bachelor," by appealing to the referential properties of others, say "adult unmarried male." But what then explains the reference of those other terms? Perhaps, we can use description theories to explain them, too. This process cannot, however, go on forever: there must be some expressions whose referential properties are not parasitic on those of others, else language as a whole is cut loose from the world. Description theories pass the referential buck, but the buck must stop somewhere. It stops with theories that are at least partly causal-historical, explaining reference in terms of direct relations to reality.

All terms *could* be causal-historical, but it seems a priori unlikely that they are. The evidence for (a) and (b) is evidence that they are not. However, let us start with two biological kind terms that almost certainly are: "echidna" and its "scientific" equivalent, "*Tachyglossidae*."

For reasons that will soon become apparent, we start with the *metaphysics* of the biological kind, echidna/*Tachyglossidae*. This taxon is an Australian biological family made up of three genera. What is it to be a member of the taxon? What is the nature/essence of the taxon? The consensus in philosophy of biology has been that the essence is a matter of having a certain evolutionary history (Sterelny & Griffiths, 1999: 8). There have been some dissenting voices, arguing that the essence is at least partly an underlying intrinsic, largely genetic, property.[22] So, there is controversy. But, crucially, aside from a few fringe "pheneticists," no biologist or philosopher of biology thinks that the nature/essence is a set of superficial phenotypic properties. So, we can safely assume that the nature *is not* such a set. Turn now to *semantics* and the *terms* "echidna" and "*Tachyglossidae*." It is uncontroversial that these terms, *as used by biologists*, refer to the biological taxon we have just discussed. So, *without the arguments of Kripke and Putnam (1.1), without experiments, indeed without further ado, we can confidently dismiss description theories of these uses*. For, according to description theories, the reference of these terms will be determined by commonly associated descriptions of the superficial phenotypic properties of echidnas (being spiny, anteaters, etc.). Yet we know from biology, together with the uncontroversial claim, that *this is not so*. The terms, as used by biologists, can refer to animals that lack those properties and not refer to some that have them. For an animal to be referred to by a term for a biological taxon, it must have a certain "deep" property, a certain history and/or underlying intrinsic nature, the details of which are still controversial. Of course, biology might be wrong; but no theory of reference, now or in the foreseeable future, could show that it is wrong. The evidence for semantic theories does not compare with that for biological ones.

The biologists' use of "echidna" provides a persuasive example of a term covered by a causal-historical theory that relates the term to a deep nature. Terms like this get their meaning in serving the *explanatory* needs of science: It is the deep nature of echidnas that explains their place in the causal nexus. And that is why biologists introduce a term that refers to animals in virtue of their having this (likely unknown) deep nature. This is not to say that all the terms of biology must be of this nondescriptive sort. Biologists have an explanatory interest in kinds that are not Linnaean taxa, such as *predators*. An animal's property of being a predator plays its explanatory role in virtue of a relatively superficial nature: that of preying on other animals. So, perhaps, we can assume without further ado that the biologists' reference of "predator" is determined by its association with a description of that property. Plausibly, scientists use some natural kind terms with a purely causal-historical meaning, and others with a purely descriptive meaning.

What about folk uses of biological terms? Here, we need to keep in mind the central idea of causal-historical theories: that the meaning and reference of a term can be "borrowed" via a chain of communications from those who fixed the reference of the term (1.1). Any descriptions that folk associate with such a term will have no role in reference determination; description theories of such a folk term will be false, too. And biological terms are among those that Kripke and Putnam argue are in fact borrowed from the scientists. Putnam has a vivid example: "elm" (1975: 226–7). He, like most of us folk, cannot uniquely describe elms or pick one out in a crowd of trees. His word "elm" refers to elms in virtue of his borrowing, ultimately from biologists. This strikes us as plausible, but of course we need evidence that it is so. Our experiments can be seen as a partly failed attempt to find such evidence. Indeed, these experiments have led us to (a) and (b), which imply that the causal-historical theory seriously overestimates the role that reference borrowing plays in folk uses of biological kind terms.

Where did we enthusiasts for the causal-historical theory go wrong? We seem to have overlooked that folk uses of many so-called "natural kind terms" serve largely *practical* rather than explanatory needs.[23] Here, we think a description theory is certainly possible and often plausible: those terms get their meanings from their practical roles and without reference borrowing. Consider the term "water," for example. It certainly has an explanatory role in chemistry but earlier had, and still has, a very large practical role in ordinary life. So perhaps folk, in applying the term to the familiar liquid, were and are primarily interested in identifying water by its practically significant superficial properties. They may not have much interest in identifying water by an unknown explanatorily significant underlying structure that causes the superficial properties (and much else). Perhaps, a similar story applies to the folk's use of "salmon" because of the important role that salmon play in our lives. If these stories are right, the reference of such folk terms may be largely determined by descriptions of those superficial properties; reference borrowing from scientists may play little to no role.

We can, therefore, distinguish two sorts of term introduction: a causal-historical one driven by scientific explanatory interests that ties terms to deep properties; and a descriptivist one driven occasionally by those interests (e.g., "predator") but particularly by folk practical interests that ties the term to superficial properties. Where a term has a role in science that picks out a kind of no practical interest to folk—perhaps "molybdenum" is an example—it is plausible to think that, insofar as folk use it at all, their reference will be borrowed. So, it may have just the one causal-historical meaning in the community. But where the folk use of a biological kind term serves largely practical interests, reference may well be determined descriptively without any reference borrowing from biologists.

We have contrasted the explanatory interests of scientists with the practical interests of folk. But this is misleading about folk. Folk obviously have explanatory interests as well as practical ones. Indeed, there is no sharp line between practical and explanatory interests: Explanations serve practical interests; underlying structures cause superficial properties of practical concern. Even before science, there was proto-science, yielding some terms that must be explained at least partly by causal-historical theories. Indeed, we have already noted the important background knowledge that the reference of some terms in a language must be, at least partly, determined by "direct relations to reality."

Consider now terms like "water" and "salmon" with roles in both science and ordinary life. Perhaps, we should say that such terms are *ambiguous*. Tobia et al. (2020) are led to this Ambiguity Theory by the "dual character pattern" of kind judgments revealed in their experiments on such terms: Sometimes the judgments reflect identification by "deep, causal properties," sometimes by "superficial properties" (196). Their most persuasive experiment uses a vignette about a fish created by modifying the genes of salmon:

> In the scientific context, participants were less inclined to categorize the entity as a member of the natural kind [salmon]. In the legal context, they were more inclined to categorize it as a member. (201)

The Ambiguity Theory has a serious problem, as noted in Section 1.3. For a term to be plausibly declared ambiguous, we need evidence in *regularities of usage* that it has two *conventional linguistic* meanings: Speakers should use the term regularly with one meaning and also *regularly* with the other. There is surely no sign of these regularities in the folk use of "salmon" (or "water"): no sign that on many occasions the folk mean to express a thought about "superficial-salmon" *not* "deep-causal-salmon"; and on many other occasions, the reverse. It is not sufficient support for the Ambiguity Theory that participants in an experiment appear able to distinguish two *potential* senses of a term, two "speaker meanings." Those two meanings may arise from *novel* uses of a term of the sort that all sides in the semantics-pragmatics dispute treat as a matter of "pragmatics" not "semantics." Before declaring them semantic, and hence adopting the Ambiguity Theory, we need evidence of two conventions in the participants' language.

This having been said, it seems plausible that "water" is ambiguous *for chemists*: When doing science, they regularly use "water" to refer to "deep-causal-water," but when they leave the laboratory and talk to the folk, they regularly use it to refer to "superficial-water." If this is so, then we may have found a small place for the Different Idiolects Theory: The folk may differ from the chemists in not having these two regular uses, such that "water" is ambiguous in the chemists' idiolect but unambiguous in the folks'.

Compare this now to the Hybrid Theory. On this view, the reference of a term is determined partly by causal-history, reflecting deep explanatory interests, and partly by description, reflecting practical interests. We think that this is plausible for folk terms like "water" and "salmon." Given our practical interests, it is likely that descriptions of the practically significant properties of the kind will play a prominent role in determining its reference, though it may be indeterminate which descriptions play that role. And, as just noted, the folk also have explanatory interests; even the beginning of language likely involved some proto-science. So, it is plausible to think that causal links to some "deep, causal properties" are playing a role with "water" and "salmon." If so, we have a hybrid theory that is partly descriptive and partly causal-historical.[24] Normally, each part yields the same decision about whether the term refers to certain entities, but sometimes not. Sometimes, the parts pull in different directions, as many experiments suggest. And when they do, there is no determinate matter of fact about whether the term refers to those entities. This provides an alternative explanation of the apparent "dual character" of kind judgments that the problematic Ambiguity Theory seeks to explain.

Some other possibilities need to be considered. A term that starts descriptive, meeting practical needs, may become hybrid as science impacts ordinary life. Perhaps that is the story with "water"[25] and "salmon," and it certainly seems to be with "whale" and "fish." According to current usage, whales are not fish. Yet it is fairly clear that this was not so according to the usage before the 20th century; see *Moby Dick*, for example. Dupré (1999) argues persuasively that this example of the impact of science on ordinary language came from a *mis*reading of science: The folk were "duped" (p. 465); "neither 'whale' nor 'fish' is really a scientific term" (p. 466). Consider "whale," for example. Why do we not apply this term to dolphins and porpoises? Not for any good scientific reason but because, as Dupré puts in nicely, "they aren't big enough" (465). The history of "whale" exemplifies the drive for an ordinary language that meets explanatory needs as well as practical ones, even though the drive has led us somewhat astray here.

Maybe the reverse can happen: A term that starts causal-historical, meeting scientific needs, may refer to something of such practical interest that it becomes hybrid. Perhaps that is the story with "dinosaur." Though the term originated in paleontology to refer to a biological kind, it is now commonly used by folk, including children, who have little if any explanatory interest in the kind.

Any of these meaning changes would introduce further indeterminacy: in the middle of the change, there would be no fact of the matter whether the term has the old meaning or the new one (Devitt, 1981: 191–195) And, of course, speakers are likely to vary as to whether and when they make the change. Finally, a term may be hybrid and yet members of the speech community may differ in the weight they give to each of the two reference-determining factors: The more scientifically minded may give more weight to the causal-historical factor, the more practically minded, to the descriptive. This would yield differences in meaning in the community that may mostly go undetected.

In short, reference determination for natural kind terms is almost definitely more varied than any of the extant theories suggest. The causal-historical theory, the descriptive theory, the Ambiguity Theory, and the Hybrid Theory each claim that all (or nearly all) natural kind terms have the same sort of reference determination. But this is almost certainly false. On the approach that we are advocating, there is room for each of these theories, and even the Different Idiolect Theory, to correctly describe *some* terms. Scientists certainly use some terms causal-historically ("echidna," "molybdenum," "elm"); they likely use some terms descriptively ("predator"). To the extent that the folk borrow references from scientists, their uses of these terms will also be causal-historical/descriptive, and so the causal-historical theory and the descriptive theory will get it right for some terms. But our results seem to show that the folk do not borrow reference from scientists as much as the causal-historical theory predicts. Their uses of natural kind terms may often reflect the folk's practical interests, either instead of or in addition to the scientists' explanatory interests. So, there are likely to be two factors to the reference determination of these terms, a superficial-descriptive one and a deep-causal one ("water," "salmon," and "whale"). Where these two sets of factors feature separately in two distinct patterns of usage, the Ambiguity Theory gets it right. Should this dual usage apply to scientists but not the folk, the Different Idiolect Theory becomes part of the story. But as we have argued, the Ambiguity Theory is implausible in most cases. Where there are not two distinct patterns of regular usage, the Hybrid Theory will give the best explanation.

The interesting question, then, is not which one of these theories covers all natural kind terms but rather which terms are covered by one theory, which another. Unfortunately, much of the extant empirical evidence, laboring under a false assumption, has focused on answering the former question instead of the latter. This brings us to the Faulty Test Hypothesis.

5.2. *The faulty test hypothesis*

Our aim was to test theories of reference of biological kind terms, particularly the description theory, against usage. Clearly, any good test of the usage of a term has to be on people linguistically competent with the term. So, our experiments assume that the participants, having read the vignette, are linguistically competent with "catoblepas." Furthermore, the experiments assume that the situation described in the vignette would prompt a linguistically competent participant to say or imply "Catoblepas do not exist" or "Catoblepas exist," depending on whether the description theory of "catoblepas" is or is not true: *that is the predicted usage that tests the theory*. We thus assume that, having read the vignette, the participant is competent enough at identifying catoblepas to know whether, in the situation described in the vignette, there are no catoblepas or catoblepas are wildebeests. *Indeed, without some such assumption, usage could reveal nothing about the reference of a term.*

We are not alone, of course, in making these assumptions to test the reference of natural kind terms. We shall consider the significance of this in Section 5.3.

The EP test provided the first sign that participants may lack the required identification competence and that the methodology of our tests was flawed. The 50-50 result (2.3) suggests that the participants were no better than random guessing. And among the discarded EP responses were some that had both a descriptivist aspect (0) and also a nondescriptivist aspect (1); we gave one example (2.3). These were small signs of what was to come. For, second, consider patterns (2) and (3), described in Section 5.1 and found in our TVJ1 and TVJ2 tests (3.2, 4.2). Two groups of participants overwhelmingly agreed to the nondescriptivist statement when presented with it alone; yet two other groups overwhelmingly agreed to the descriptivist statement when presented with it alone. One group on each side was tested for confidence in TVJ1: it was high in both groups, which is not what either the Ambiguity or Hybrid theories should predict (3.2). This apparently extreme example of the Dunning–Kruger effect seems telling evidence of incompetence. But, most telling of all, groups 1 and 2 in our TVJ 2 test seem to show that this suggestion underestimates the incompetence. For here, where each participant is presented with first one statement and then the contradictory one, 50% of the participants accepted both, and in their explanations gave no indication of understanding that this was inconsistent (4.2). Finally, group 3 in our TVJ 2 was presented with a choice of the descriptivist and nondescriptivist statements at once. With high degrees of confidence, two participants chose both, thus contradicting themselves, and the rest split 50-50, much as in the EP test. All in all, this strongly suggests that the participants are not competent enough at identifying catoblepas for their use of "catoblepas" to yield persuasive evidence about its reference.

Why are participants so apparently incompetent at identifying the referent? The root problem, we suggest, comes with the initial assumption of the participants' *linguistic* competence, a competence that is supposed to enable identification of the referent. For a participant to be linguistically competent with "catoblepas," the word has to *have* a meaning in her language. Yet consider a participant's acquaintance with the word. The vignette *mentions* the word in saying that early biologists "described a distinctive kind of animal which they called 'catoblepas'" (2.1). The word is then *used* in each of the various prompts. These are the first and only experiences that the participant has of "catoblepas." So, at that moment, "catoblepas" has no regular use and no conventional meaning in her language. So, with an important proviso to be discussed, in understanding the prompt, the participant cannot assign a meaning to the word by participating in the convention for the word (as we normally do in understanding). And her later (explicit or implicit) use of "catoblepas" in response to the prompt cannot exemplify its linguistic meaning. For "catoblepas" *has* no meaning in her language. So, our experiments could not test the way that the linguistic meaning of the biological kind term "catoblepas" determines reference. That is the case for the Faulty Test Hypothesis.

Although (proviso aside) a participant reading the prompt cannot assign to "catoblepas" the meaning it has in her language, she can and will assign what she takes the experimenter to *mean by* the word: She will assign a "*speaker* meaning" to it. And when in her response she uses the word (explicitly or implicitly), that is what she will *mean by* it. What meaning is that? Well, if our speculations about meaning in Section 5.1 are anywhere near correct, she has some choices: She could treat "catoblepas" as a causal-historical term like the biologists' use of "echidna" and "elm"; or a descriptive term like perhaps "predator" and "bachelor"; or a hybrid term like perhaps "water," "salmon," and "whale." Given the choices, the following aspects of our results are not surprising: that some participants assign a descriptivist speaker meaning, some, a causal historical one (EP); that participants tend to assign a meaning that makes the statement in a prompt true (TVJ1 and 2); perhaps even, that a participant switches from one meaning to the other, hence apparently contradicting herself (TVJ2). In brief, it is not surprising that participants are rather lost in responding to the prompts. And indeed, the participants' explanations of contradictory answers (TVJ2) are probably best understood as evidence that participants are lost.

So, by taking a participant to be choosing from different sorts of speaker meanings, we can go a long way toward explaining the results. But our claim in Section 5.1 was not about *speaker* meanings but about the *conventional linguistic* meaning of biological kind terms. We claimed that our results "provide quite strong evidence" about such meanings; in particular, evidence that the reference determination of these terms within the community and within individuals has both descriptivist and nondescriptivist elements; see (a) and (b). How could there be any such evidence given that, proviso aside, "catoblepas" has no conventional meaning in the participants' language? Certainly, in those circumstances, the results provide no *direct* evidence about the meanings of biological kind terms. That is the case for the Flawed Test Hypothesis. Still, the experiments provide *indirect* evidence of the meanings of biological kind terms. For we can predict that a participant, in choosing which speaker meaning to assign to "catoblepas," will assign a meaning *of a sort that other biological kind terms have conventionally*. Since the participants assigned both descriptivist and nondescriptivist speaker meanings, our results do indeed provide evidence that there are both descriptivist and nondescriptivist elements to the reference determination of folk biological kind terms.

We must address the important proviso. We noted that, up to the moment of experiencing the prompt, "catoblepas" has no conventional meaning in a participant's language. So, we claimed, the participant cannot assign a meaning to the prompt's use of "catoblepas" by participating in the convention for the word. But the central idea of the causal-historical view is that she *can*: a person can borrow the speaker's reference with a term in a communication like this and *thereby come to participate in the convention* for the term. This is most vividly demonstrated with proper names: We all came to participate in the convention of using "Aristotle" to refer to a famous ancient philosopher simply by borrowing our reference from someone who was linguistically competent with the name. So, *if* "catoblepas" is a term that is covered by the causal-historical theory and thus can be borrowed, and *if* a participant does borrow it, then her use of it is an exercise of her linguistic competence and she should be competent at identifying whether or not wildebeests are catoblepas, given the vignette information. But these are two big "*if*"s! Concerning the first "*if*," who knows whether those "early biologists" introduced "catoblepas" the way later ones did "echidna"? The vignette is consistent with their introducing the term the way later ones likely did "predator." Concerning the second "*if*," even where a term's reference *can* be borrowed, it *may not* be. Borrowing is not compulsory; there are other things she might do, perhaps assigning a descriptivist meaning, perhaps a hybrid one. If either "*if*" is not realized, our experiments are not testing a linguistic meaning of "catoblepas." The Faulty Test Hypothesis stands. However, the proviso yields a respect in which our tests were only semi-faulty. Had the results in the tests been consistently antidescriptivist, that would have been evidence that both "*if*"*s are* realized and hence that biological kind terms are causal-historical. But, of course, the results were far from consistently antidescriptivist. So, they provide direct evidence *against* the causal-historical theory.[26]

The Faulty Test Hypothesis applies to our experiments: The experiments test speaker reference and provide only indirect evidence of linguistic reference (proviso aside). Still, we claim, they do support (a) and (b): Reference determination of biological kind terms have both descriptivist and nondescriptivist elements.

5.3. *Past experiments*

What about past experiments? Our experiments are not unique in being open to the Faulty Test Hypothesis. Nichols et al. (2016), Jylkkä et al. (2009), and Genone and Lombrozo (2012) each test novel terms ("catoblepas," "zircaum," and "tyleritis," respectively) that are introduced to participants in a vignette. These terms do not have a linguistic meaning in their participants' languages, and so their experiments are open to the same methodological objection as ours. The contradictory responses found by Nichols et al. and Jylkkä et al. add to the case that the Faulty Test Hypothesis applies to their experiments.

We earlier expressed doubts about the experiments of Jylkkä et al. and Genone and Lombrozo, and one experiment of Nichols et al., because they were RI. Clearly, the Faulty Test Hypothesis adds to these doubts. The other experiments of Nichols et al. were TVJ and so, despite the Faulty Test Hypothesis, we should assess them as we just did ours: They provide some support for (a) and (b).

We suspect that the Faulty Test Hypothesis may also apply to experiments using "Twin Earth"-style thought experiments. As Stich points out, "nonphilosophers often find such cases so outlandish that they have no clear intuitions about them" (1983: 62). If the situation raised in the test vignette is too fantastical and esoteric, folk participants may get confused and do little more than guess at what the term means or how to use it. Indeed, evidence suggests that participants tend to endorse test sentences that they do not understand (Crain & Thornton, 1998: 213). This is a problem for the TVJ test by Braisby et al. (1996), and the first two experiments by Tobia et al. (2020). Tobia et al. acknowledge the problem: "it might be thought that [Twin Earth thought experiments] are overly philosophical or esoteric" (198). The apparently contradictory responses that Braisby et al. (1996) found support the idea that the Faulty Test Hypothesis applies to those experiments. Despite this methodological worry, we think that some credence should be given to these results of Braisby et al. and Tobia et al.: They support the view that there are both descriptive and causal-historical factors in the reference determination of natural kind terms.

However, the best evidence for this view comes from Tobia et al.'s third experiment. This used more realistic scenarios involving genetic mutations rather than fantastical "other worlds." And this experiment tests natural kind terms that are already in the participants' language: "gold," "salmon," and so on. So, we think that this third experiment likely avoids the Faulty Test Hypothesis and provides good evidence that folk usage of the terms tested varies in certain contexts.

Putting these experiments together with ours, and despite the Faulty Test Hypothesis, we think that there is strong evidence that folk natural kind terms cannot all be explained simply by a description theory or a causal-historical one.

However, contrary to what authors claim, these results do not provide evidence that significantly favors either the Ambiguity or Hybrid Theory. We take the Hybrid to be generally more plausible, but further evidence is needed to distinguish the predictions of the two theories. This brings us to future experiments.

5.4. Future tests

Much of the experimental work on natural kind terms—and much of the philosophical discussion of the same—has presupposed that one of the theories of reference must be true of *all* natural kind terms: they are all either descriptive, causal-historical, hybrid, or ambiguous. We suspect that this is why introducing novel natural kind terms in a vignette seemed like a viable source of evidence: If every natural kind term functions descriptively, for example, then novel natural kind terms must necessarily function descriptively.[27] But, as we argued (5.1), this is *a priori* implausible. Some natural kind terms are clearly causal-historical (like "*Tachyglossidae*"), and some are likely descriptive (like "predator"). Without that presupposition, participants' use of a single novel term with no linguistic meaning in their language provides no direct evidence for the reference of natural kind terms in general. So, the apparent variation in usage found in all of these experiments—both within the community and within individuals—may simply be evidence of the fact that none of the aforementioned theories accurately describe *all* natural kind terms.

In the future, we need experiments that do not presuppose that one theory fits all and that avoid the methodological Faulty Test Hypothesis. So, we need experiments that:

(a) test terms that are already in the participants' language;
(b) avoid overly complex or fantastical scenarios; and
(c) do not presuppose reference is determined in the same way for *all* natural kind terms.

Condition (a) ensures that the tested term has a determinate linguistic meaning to be tested, while (b) avoids confusing or overtaxing the participants' identification competence, ensuring reliable identifications. We have just argued for the importance of (c). So, experimental semanticists should be testing a wide variety of natural kind terms. We need to test whether the folk borrow reference from scientists for terms like "aluminum" and "elm," where a causal-historical theory is most plausible. We need to test terms like "water" and "salmon," where the Hybrid and Ambiguity Theories are most plausible, to see which if either of those theories apply. And we need to test terms like "predator," where a pure description theory is most plausible.

Many past tests of usage—including our own—have been methodologically faulty. But this should not return us to the discredited testing of intuitions. Tests of linguistic usage––preferably direct tests via EP—ought to be our primary source of evidence for and against semantic theories. Following the above conditions and testing a variety of terms should ensure that future tests of usage actually test usage, thus providing reliable evidence for the semantics of natural kind terms.

6. Conclusions

Previous experiments were thought to provide evidence of both descriptivist and nondescriptivist (causal-historical) reference determination in natural kind terms. This has led some researchers to propose the Ambiguity Theory, according to which terms have two distinct linguistic meanings, one descriptivist and one not. But this theory faces a serious problem: For it to be plausible, there must be regular uses exemplifying each meaning. Others have proposed the Hybrid Theory, according to which a term has just one linguistic meaning which determines reference partly by description and partly by a causal-historical link. This avoids the Ambiguity Theory's problem but at the cost of introducing an inherent indeterminacy into the theory of reference.

The methods used in previous experiments tested reference against RI or TVJs. We think that testing theories against RIs is wrong in principle and has proved unreliable in practice with proper names. Theories should be tested against usage, preferably by the method of EP rather than TVJ. And that was what we did first in testing the reference of the biological kind term, "catoblepas." Our expectation was that this would support the causal-historical theory, but that was not what we found. This led us to a series of TVJ tests.

Our experiments, like some earlier ones, found participants contradicting both each other and themselves. We drew some methodological conclusions. We argued that many experiments, including our own, were faulty in that they assume that participants are linguistically competent with the tested terms. Where the term is novel to the participants, like our and Nichols et al.'s "catoblepas," this may well not be so. And our results suggest that it is not so with "catoblepas." So, experiments that were intended to test *linguistic* reference may only be directly testing *speaker* reference. We argued that this Faulty Test Hypothesis may also be true of some earlier experiments because they were too complex and fantastical for participants. Despite these methodological issues, we argued that our and earlier results are evidence for two substantive conclusions.

The first substantive conclusion is that there are indeed *both* descriptive and causal-historical elements to the reference determination of biological kind terms. This led us to distinguish two sorts of term introduction: a causal-historical one driven by scientific explanatory interests that ties terms to deep properties; and a descriptivist one driven particularly by folk practical interests that ties terms to superficial properties. Our second substantive conclusion is that, given these varying interests and the experimental results, the common assumption that any one theory of reference fits all natural kind terms is most likely false. Rather, it is likely that some terms are descriptive, some causal-historical, some ambiguous, and some hybrid. We argued that when both the Hybrid Theory and the Ambiguity Theory are in contention, the Hybrid Theory should usually be preferred because of the Ambiguity Theory's commitment to two distinct patterns of usage.

What about the future? Future experiments should focus on testing usage directly, using prompts that are neither overly complex nor overly fantastical, and which test terms that are already part of the participants' language. And the question to be addressed is not which theory of reference governs all natural kind terms, but which natural kind terms are governed by each theory of reference.

7. Postscript

After our paper was submitted, our attention was drawn to a very interesting recent paper by Haukioja, Nyquist, and Jylkkä (2020). They used EP and TVJ on Twin Earth (TE) but also ingenious "reverse Twin Earth scenarios, where deep structure, but not appearance, was shared with the standard samples" (p. 2). They experimented on five natural kind terms, including the biological "tiger."

(a) Their TE experiments decisively confirmed the Kripke–Putnam prediction that the referents of a term must share a "deep structure." This contrasts with our experiments, particularly EP and TVJ2-ChooseCHorD, which failed to confirm the Kripke–Putnam prediction that the referents need not share the superficial properties picked out by the descriptions speakers associate with the term. (The contrast rests on the fact that if associated superficial differences do not matter to reference, then deep, or nonassociated superficial, differences must.)

(b) What explains these different results? In thinking about this, we should note that, in one respect, it should have been *easier* to get antidescriptivist results from our experiments than from theirs! For, our antidescriptivist hypothesis is not committed to the referents having to share any particular sort of property, just to their *not* having to share the associated superficial properties. In contrast, their antidescriptivist hypothesis is committed to the strong Kripke–Putnam thesis that referents must share deep microstructural properties. Do their experiments fall prey to the Faulty Test Hypothesis? Not on the grounds that ours did, because they used familiar, not novel, terms (we think this may explain why many of our participants contradicted themselves, but none of Haukioja et al.'s did). But they may fall prey, like other TE experiments, on the grounds of being too fantastical.

(c) In experiment 3, Haukioja et al. found that "categorization judgments are gradual, in proportion to the degree of similarity between new samples and standard samples" (p. 21). They argue that this casts doubt on the traditional assumption "that category membership is all-or-none" (p. 22).

This result may provide evidence for the Hybrid Theory. Neither pure causal-historical theories, pure descriptivist theories, nor the Ambiguity Theory has a good explanation for how kind membership might come in degrees. But the Hybrid theory, which claims it is sometimes indeterminate whether reference is primarily determined by descriptive factors or causal-historical factors, may be able to offer a better explanation for categorization judgments coming in degrees.

(d) Their reverse-TE experiments led them to conclude that not only deep properties but also "observable properties associated with the kind term…have to be, to some extent, satisfied by samples, in order for them to belong in the extension of the term" (p. 24). We think that the experiments do not support this conclusion. Consider the reverse-TE scenario for "water":

> a substance found on the planet…has the molecular structure H_2O. For some mysterious reason, H_2O on this planet is solid, not liquid, up to about 800°C, and it is greenish. H_2O is widely found on the planet as lumps that look like this:[an image of a greenish mineral]. (p. 7).

The problem lies in the mystery. On the Kripke–Putnam essentialist view, any stuff that "water" refers to must have an underlying essence which, according to scientists, has an H_2O structure. Key point: *that essence, whatever precisely it may be, along with the environment, causes all the observable properties of water*. For, that is what essences *do*! But then if the essence of the H_2O-ish stuff on reverse-TE was really that of water, how could it cause properties of being like a greenish mineral, properties that are so strikingly different from the familiar properties of water? From the Kripke–Putnam perspective, this striking difference in the causal properties of the essences of the H_2O-ish stuff on the two planets is good evidence that these essences are different, despite sharing an H_2O structure: There is more to the deep essence of water than simply having an H_2O structure (as indeed science seems to be showing anyway). So, the reverse-TE stuff is not water *because it lacks the deep essence of water, not because it lacks the right observable properties*. Kripke–Putnam can explain the reverse-TE results.

(e) There is a way forward for reverse-TE experiments: keep the suggestion that the deep properties of entities on reverse-TE are the same but remove the mystery by giving a *plausible environmental* explanation of the strikingly different observable properties. This plausibility may be hard to achieve for chemical kind terms like "water," but it should be easy for biological kind terms like "tiger" because the impact of the environment on the development of living things is very apparent. Then, if these environmentally induced differences were shown to make a difference to reference, that would count against Kripke–Putnam.

Acknowledgments

We would like to thank Karen Brøcker, John Dupré, Steven Gross, Jussi Haukioja, Michael Hillas, Frank Jackson, Ron Mallon, Shaun Nichols, N. Ángel Pinillos, Nicolas Porot, Francesco Pupa, Stephen P. Schwartz, Kevin Tobia, Jonathan Weinberg, and Tomasz Zyglewicz for their helpful comments and discussion.

Notes

[1] "Natural kind terms" may sometimes be used in a technical sense to pick out the explanatory terms of the natural sciences. Our usage is looser, covering folk terms like "tiger" and "water" with no presupposition that these terms are synonymous with scientific terms.

[2] Kitcher (1978: 535; 1993: 73) provides an example of this sort of theory.

[3] See Devitt and Sterelny (1999: 96–101) for an exploration of the possibilities for different sorts of hybrid theories.

[4] See Devitt (1974: 200–03, 1981: 138–52, 159–60, 193–5) for examples of the role of indeterminacy.

[5] For overviews of recent experimental work in the philosophy of language, see Hansen (2015) and Haukioja (2015).

[6] Genone and Lombrozo take Evans's (1973) theory of names to raise the possibility of a Hybrid Theory like theirs, on which the reference of a token is determined partly causal-historically and partly by what its associated descriptions are *true of*; see our definition (1.1). But it should be noted that the latter descriptive part has no place in Evans's view, according to which reference is to the *dominant source of the information expressed by* those descriptions. His view was really a causal theory of reference *fixing*, to which he later (1982) added reference *borrowing*, as Genone and Lombrozo note (2012: 738, n. 5). At that point, Evans's theory is a causal-historical one like Devitt's (1974, 1981). It was never a hybrid theory.

7. Although see Wikforss (2017) for a defense of the evidential value of folk intuitions.
8. Domaneschi, Vignolo, and Di Paola (2017); Devitt and Porot (2018).
9. See, for example, Sytsma and Livengood (2011), Sytsma, Livengood, Sato, and Oguchi (2015), Machery, Sytsma, and Deutsch (2015), and Devitt and Porot (2018).
10. Note that the requirement is that there be (at least) two regular uses, not that the two uses be equally common. Semanticists typically have stronger requirements for ambiguity, reflected in the popularity of "Modified Occam's Razor" (Grice, 1989: 47); for discussion, see Devitt, 2013: 297–300.
11. They also criticize Evans's view, which they wrongly take to be a hybrid (2016: 164–5); see note 6.
12. There is another way to respond to apparent variations in reference determination, a way that has not been urged in discussions of experiments on natural kind terms but has in those on proper names. It is the view that theories of reference have no place in semantics; meaning is not to be explained in terms of reference; see Devitt and Porot (2018: 1579, n. 13) for more information. We shall not consider this view.
13. The same is true of participants in all our experiments. All participants were self-identified native English speakers.
14. Sytsma and Livengood (2011), Sytsma et al. (2015), and Domaneschi and Vignolo (2019) argue that response variance in semantic experiments can be explained by some participants answering the prompt from a narrator's "perspective," while others answer from the "perspective" of a character in the vignette. To avoid these issues, we asked what *textbooks* should say. Our expectation is that this will prevent any "from a character's perspective" readings and elicit responses that reflect what the participants think is literally true of catoblepas.
15. See Supplementary Material.
16. We measured intercoder reliability with a pairwise comparison (Cohen's kappa); agreement was good ($\kappa = 0.588$).
17. Not counting the roughly 1/3 of responses that were discarded.
18. See Krosnick (1999: 552–553). This "acquiescence bias" is very real, but the available evidence suggests "an average acquiescence effect of about 10%" (ibid).
19. The 81% were in a "follow-up" experiment.
20. Musolino and Lidz (2006), Musolino, Crain, and Thornton (2000), and Crain and Thornton (1998: 52–53, 84, 103–104).
21. Nichols, Pinillos, and Mallon (2016) note that they attempted to educate participants about this prior to their experiment, in order to avoid the same issue.
22. See particularly, Walsh (2006); Devitt (2008, 2018).
23. John Dupré is entitled to say "I told you so." For, using lovely examples like "prickly pear," "lily," and "tree"; he emphasized long ago that many biological kind terms of "ordinary language" do not correspond "to *any* recognized biological taxon" (1981: 73). As he says, the functions these perform for the folk are different from the functions of scientific terms; folk terms are for kinds that are "economically or sociologically important," "furry and empathetic," "very noticeable" (80), and so on.

24. Barbara Malt's (1994) interesting experiments on "water" support this sort of view. She overlooks the indeterminacy we are positing and so is led to a somewhat different view.
25. Malt's (1994) experiments give powerful support to the view that the reference of the folk's "water" is now largely, though not entirely, linked to H_2O. But obviously it was not always. Still it might have been linked to a then-unknown deep structure from the start.
26. In contrast, the results in Devitt and Porot's (2018) experiments testing the usage of proper names were consistently antidescriptivist. So, the results do support a causal-historical theory, as claimed. Still, the Faulty Test Hypothesis otherwise applies there, too.
27. We made this assumption, at least implicitly, when designing our experiment.

References

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, *59*, 247–274.

Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Devitt, M. (1974). Singular terms. *Journal of Philosophy*, *71*(7), 183–205.

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

Devitt, M. (2008). Resurrecting biological essentialism. *Philosophy of Science*, *75*, 344–382.

Devitt, M. (2011). Experimental semantics. *Philosophy and Phenomenological Research*, *82*(2), 418–435. Retrieved from https://doi.org/10.1111/j.1933-1592.2010.00413.x

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

Devitt, M. (2012b). Semantic epistemology: Response to Machery. *Theoria*, *27*(2), 229–233.Retrieved from https://doi.org/10.1387/theoria.6225

Devitt, M. (2013). Three methodological flaws of linguistic pragmatism. In C. Penco & F. Domaneschi (Eds.), *What is said and what is not: The semantics/pragmatics interface* (pp. 285–300). Stanford, CA: CSLI.

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

Devitt, M. (2018). Historical biological essentialism. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *71*, 1–7. https://doi.org/10.1016/j.shpsc.2018.05.004

Devitt, M., & Porot, N. (2018). The reference of proper names: Testing usage and intuitions. *Cognitive Science*, *42*(5), 1552–1585.

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

Domaneschi F., Vignolo M. (2020). Reference and the ambiguity of truth-value judgments. *Mind & Language*, *35*(4), 440–455. https://doi.org/10.1111/mila.12254.

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.

Dupré, J. (1981). Natural kinds and biological taxa. *Philosophical Review*, *90*, 66–90.

Dupré, J. (1999). Are whales fish? In D. Medin & S. Atran (Eds.), *Folkbiology* (pp. 461–476). Cambridge, MA: MIT Press.

Evans, G. (1973). The causal theory of names. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *47*, 187–208.

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

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.

Grice, H. P. (1989). *Studies in the ways of words*. Cambridge, MA: Harvard University Press.

Hansen, N. (2015). *Experimental philosophy of language*. Oxford Handbooks Online.

Haukioja, J. (2015). *Advances in experimental philosophy of language* (2nd edn, 1st edn 1987). Oxford: Blackwell.

Haukioja Jussi, Nyquist Mons, Jylkkä Jussi (2020). Reports from Twin Earth: Both deep structure and appearance determine the reference of natural kind terms. *Mind & Language*, https://doi.org/10.1111/mila.12278.

Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology*, *22*, 37–60.

Kitcher, P. (1978). Theories, theorists and theoretical change. *Philosophical Review*, *87*(4), 519–547.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford: Oxford University Press.

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

Krosnick, J. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

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

Malt, B. (1994). Water is not H2O. *Cognitive Psychology*, *27*, 41–70.

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.

Musolino, J., Crain, S., & Thornton, R. (2000). Navigating negative quantificational space. *Linguistics*, *38*(1), 1–32.

Musolino, J., & Lidz, J. (2006). Why children aren't universally successful with quantification. *Linguistics*, *44*(4), 817–852.

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

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, *70*(19), 699–711.

Putnam, H. (1975). The meaning of 'meaning'. In *Mind, language and reality: Philosophical papers* (Vol. *2*, pp. 215–271). Cambridge: Cambridge University Press.

Sterelny, K., & Griffiths, P. (1999). *Sex and death*. Chicago: University of Chicago Press.

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief*. Cambridge, MA: MIT Press.

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 crosscultural 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.

Tobia, K., Newman, G., & Knobe, J. (2020). Water is and is not H2O. *Mind & Language*, *35*(2), 183–208. https://doi.org/10.1111/mila.12234

Walsh, D. (2006). Evolutionary essentialismn. *British Journal for the Philosophy of Science*, *57*, 425–448.

Wikforss, Å. (2017). Semantic intuitions and the theory of reference. *Teorema*, *36*(3), 95–116. https://doi.org/10.2307/26384624.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information