TWO SORTS OF BIOLOGICAL KIND TERMS: THE CASES OF ‘RICE’ AND ‘RIO DE JANEIRO MYRTLE’

Forthcoming in *Philosophy and Phenomenological Research*

Michael Devitt Brian Porter

City University of New York, University of Pittsburgh

 Graduate Center

**Abstract**

Experiments have led some philosophers to conclude that the reference determination of natural kind terms is neither simply descriptive nor simply causal-historical. Various theories have been aired to account for this, including ambiguity, hybrid, and different-idiolects theories. Devitt and Porter (2021) hypothesized that some terms are covered by one theory, some another, with a place for all the proposed theories. The present paper tests hypotheses that the term ‘Rio de Janeiro Myrtle’ is simply causal-historical but the term ‘rice’ is hybrid. For, whereas the former term is of scientific but little practical interest, the latter is not: rice is a significant part of the human diet. So, we predicted there would be two factors to the reference determination of ‘rice’: a superficial-descriptive one and a deep-causal one. Our experiments confirmed these hypotheses using the methods of elicited production and truth value judgments. We take our results to support the hybrid Theory of ‘rice’ rather than the ambiguity or different-idiolects theory. We were not testing ‘myrtle’ but, surprisingly, our results implied that ‘myrtle’ was partly descriptive and so like ‘rice’ but not ‘Rio de Janeiro Myrtle’. A follow-up experiment confirmed these puzzling results. More investigation is needed.

**1. Introduction**

*1.1 Theories of Reference*

Kripke (1980) and Putnam (1973, 1975) argued that a “Description Theory” of reference is wrong for many terms, including “natural” kind terms. People are often too ignorant to supply descriptions that would identify what their terms intuitively refer to. Most of us use ‘elm’ to refer successfully to elms but could not describe those trees well enough to distinguish them from beeches. And Putnam’s Twin Earth is imagined to be exactly like Earth except that the liquid on Twin Earth that looks and functions like our water turns out not to be H2O but a different chemical, XYZ. The intuition of Putnam and many others was that XYZ is not water, though any plausibly available descriptions of water would apply to XYZ also. The argument applies equally to biological terms like ‘tiger’.

A two-step “Causal-Historical Theory” has been proposed to replace the Description Theory for these terms (Devitt and Sterelny 1999, pp. 88-90). First, the reference of a term is “fixed” in a dubbing and (probably) later “groundings”. Second, people who fixed the reference of the term use it to communicate with others, thereby passing on the ability to use the term with that reference. Later uses of a term thus “borrow” their reference from previous uses in a causal network that goes back to earlier fixings. Even if the descriptions speakers associate with a term do not apply to anything, the term still refers to whatever kind the term is causally connected to. A biological term like ‘tiger’ does not apply to an animal in virtue of its having the superficial properties picked out by speakers’ associated descriptions, but rather in virtue of having whatever properties were relevant to the initial fixing and any later groundings of the term’s causal network. Defenders of the Causal-Historical Theory have usually taken the relevant properties to be deep structural properties discovered by scientists. On this view, ‘tiger’ refers to an animal iff the animal has the same deep structural properties (the same underlying “essence”) as the animals that ground the network.[[1]](#footnote-1)

However, these anti-descriptivist developments in the theory of reference have come in for both substantive and methodological criticisms in “experimental semantics”. Our recent “Testing the Reference of Biological Kind Terms” (2021; hereafter “Testing”) is an example. That paper ends with suggestions for future experimental work on reference. The present paper takes up those suggestions.

This paper describes the results of experiments testing two different biological kind terms. We argue that these results show that, contrary to common opinion, *no one* theory of reference fits all biological kind terms. The folk’s use of a term like ‘rice’ that refers to a kind serving important “practical” interests will reflect those interests and hence be partly descriptive and only partly borrowed; so, a “Hybrid Theory” of reference is needed. We contrast ‘rice’ with ‘Rio de Janeiro Myrtle’, a term that refers to a kind of scientific interest but little or no practical interest to the folk. So, the folk’s use of it will be fully borrowed from the scientists who grounded it, and hence be covered by the Causal-Historical Theory.

In the remainder of this section, we summarize recent experimental work and then outline our conclusions in “Testing”.

*1.2 Recent Experimental Work*

Recent experiments have led some philosophers to an “Ambiguity Theory”, according to which natural kind terms have two distinct linguistic meanings: one descriptivist, and one nondescriptivist (Braisby et al. 1996; Nichols et al. 2016; Tobia et al. 2020). Each token of ‘water’ or ‘salmon’ will have its reference either determined descriptively or nondescriptively, depending on which meaning is in play in a given context. Others have urged a Hybrid Theory, according to which 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 (Genone and Lombrozo 2012).

Experimentalists have also been highly critical of the methodology of relying onphilosophers’ own referential intuitions. Machery et al. (2004) presented lay participants with a vignette and asked who a speaker is talking about in using a proper name. This method (“RI” for “referential intuition”) was then used by some experimentalists on natural kind terms (Jylkkä et al. 2009; Genone and Lombrozo 2012). Others (Braisby et al 1996; Nichols et al. 2016; Tobia et al. 2020) presented participants with a vignette including a target term, and asked whether a statement using that term is true or false, or to what extent they agree with the statement (“TVJ” for “truth value judgment”).

Genoveva Martí (2009, 2012, 2014) and Michael Devitt (2011, 2012a, 2012b) argued against RI, claiming that referential intuitions should not be used to test theories of reference at all, whether the intuitions are from the philosophers or the folk.[[2]](#footnote-2) They urged instead that theories be tested against usage. One effective way to do this is by the method of elicited production (“EP”). In EP tests, participants are prompted to produce statements about a situation described in a vignette, thus eliciting the *use* of a target term that appears in the vignette. This is a “pure” test of usage. It was not used by any of the experiments mentioned so far. Does a TVJ test usage? Contra Martí (2012), Devitt and Porot (2018) argue that TVJ is a “somewhat imperfect” test of usage, for it “puts words into the mouth of the participant” (p. 1561), rather than prompting participants to produce their own statements. We agree.

Our critical views of RI made us doubt the results of Jylkkä et al (2009), Genone and Lombrozo (2012), and some of the results of Nichols et al (2016). In contrast, our view of TVJ suggests that the results of Braisby et al (1996), Tobia et al (2020), and most of those of Nichols et al (2016) should be taken cautiously as evidence of the referential reality of natural kind terms. The striking message of these results is that, contrary to standard opinion, that reality is neither simply descriptivist nor simply causal-historical.

The most common attempt to accommodate this message has been the Ambiguity Theory: these terms have both a descriptivist and a nondescriptivist meaning. This theory has a serious problem. Participants appear able to distinguish two *potential* senses of the term, two “speaker meanings”, but this is not sufficient support for the Ambiguity Theory. Those two meanings may arise from *novel* uses of the term of the sort that all sides in the semantics-pragmatics dispute treat as a matter of “pragmatics” not “semantics”. Before declaring these meanings *linguistic*, and hence adopting the Ambiguity Theory, we need evidence of two conventions in the participants’ language. So, we need evidence that the term is *regularly* used one way and *regularly* used another.[[3]](#footnote-3) We clearly lack this evidence for the folk’s use of terms like ‘water’ and ‘salmon’.[[4]](#footnote-4)

The Hybrid Theory, another attempt to accommodate the experimental results, says that a term has only one meaning which yields two factors in reference determination: reference is partly determined descriptively and partly nondescriptively. “But if a term has only one meaning, how can the Hybrid Theory explain the observed variations in usage?” The theory must appeal to a certain indeterminacy. Normally, each of the reference-determining factors pull in the same direction. When they don’t, there is no determinate matter of fact about the term’s reference. In those circumstances, some participants resolve the indeterminacy one way, some the other.

According to the Ambiguity Theory a term like ‘salmon’ has both a descriptivist meaning and a nondescriptivist meaning. The theory predicts that when the term is used in certain contexts it is likely to have one meaning, when in others, the other. Tobia et al (2020) confirmed this prediction with an experiment on ‘salmon’. They found that participants in a TVJ test, having to judge whether an item was a salmon, were more likely to judge by superficial properties in a legal context, but more likely to judge by deep structural properties in a scientific context.

It will be important to note that, though the Hybrid Theory predicts a variation in usage when its two reference-determining factors pull in different directions, it does not predict that this variation is *determined by context*. The Ambiguity Theory, however, predicts that context will determine which of the two meanings is appropriate. So, any failure to find a variation in context counts in favor of the Hybrid over the Ambiguity Theory.

There was one striking aspect of many of these experiments (Braisby et al 1996; Nichols et al. 2016; Tobia et al. 2020) that none of these theories could explain convincingly: participants frequently contradicted themselves. Consider this one example from Braisby et al. (1996): a significant number of participants assigned the same truth value to “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” (1996). The contradictory phenomena made us wonder if there was something wrong with the experimental designs. We labeled this “the Faulty Test Hypothesis”.

We turn now to our own results in “Testing”.

*1.3 “Testing”*

In “Testing”, we started with an EP test which confirmed earlier experiments: the results were part descriptivist (43%), part nondescriptivist (causal-historical) (57%). We then ran TVJ tests that presented participants with either (a) a statement that the Causal-Historical predicts is true, the Description theory, false; or (b) a statement that the Description Theory predicts is true, the Causal-Historical, false. Significantly more than half of participants chose “true” regardless of which statement they were given: 89% and 82.5%, respectively. So the two groups contradicted each other. We then ran more TVJ experiments, in which groups again contradicted each other and, more strikingly, 50% of participants contradicted *themselves*.

None of the theories we have identified can explain these results. Nor can a further theory, the “Different Idiolects Theory”: some participants use the target term ‘catoblepas’ only with a descriptivist meaning, some, only with a causal-historical meaning, such that a “pure” theory is correct for each idiolect. The Different Idiolects Theory is not promising but we will keep it in mind in considering our results below.

We concluded that the Faulty Test Hypothesis largely applied to our experiments: they were not good tests of reference against usage (2021, sec. 5.2). So, what had gone wrong? The supposition that a test of the participants’ uses of ‘catoblepas’ provides evidence of its reference, in effect assumes two things: first, that simply on the basis of reading the vignette, participants for whom ‘catoblepas’ was a novel term will become linguistically competent with it; second, that, as a result of this competence, participants will be able to identify catoblepas well enough to give an appropriate response to the prompts about the vignette. Yet the results strongly suggest that many participants are not able to do that. We concluded that the “root problem” was the first assumption (2021, sec. 5.2).

For participants to be linguistically competent with ‘catoblepas’, the word has to have a meaning in their language. Yet this is a novel term to them and, with an important proviso, simply reading the vignette will *not* be sufficient to bring that competence about. That was why the Faulty Test Hypothesis largely applied to our experiments.

The qualification “largely” is necessitated by the important proviso. That proviso stems from the central idea of the Causal-Historical Theory, vividly demonstrated with proper names: a person can borrow the speaker’s reference with a term in a communication *thereby coming to participate in the convention for the term and becoming linguistically competent with it*. Now, suppose that ‘catoblepas’ were covered by the Causal-Historical Theory. Then just reading the vignette should make participants linguistically competent with the term: they borrow reference from researchers in the vignette. So, in our experiments, that theory predicted consistently non-descriptive results. We did not get such results. So, our experiments directly falsify the Causal-Historical Theory of ‘catoblepas’ and to that extent the Faulty Test Hypothesis does not apply.

However, the experiments were faulty in not providing any evidence that bore directly on any of the other theories of the conventional linguistic reference of ‘catoblepas’. Nonetheless, we argued that the experiments did test *speaker* reference and hence provided *indirect* evidence of the linguistic reference of biological kind terms in general. We concluded that our experiments supported the view that the reference determination of biological kind terms, taken as a group, is neither purely causal-historical nor purely descriptive (sec. 5.1). So, the Description Theory is no more true of them all than is the Causal-Historical Theory.

We also concluded (sec. 5. 3) that the Faulty Test Hypothesis that largely applies to our TVJ experiment applies equally to any others that use novel terms (Jylkkä et al. 2009; Genone and Lombrozo 2012; Nichols et al. 2016), and may also apply to experiments using ‘Twin Earth’-style thought experiments (Braisby et al. 1996; Experiments 1 and 2 by Tobia et al. 2020; Haukioja et al 2020). Where the situation raised in the test vignette is complex and fantastical, folk participants may do little more than guess at what the term means or how to use it.

Despite all these methodological worries, we argued that many of these past results provided some indirect support that reference determination is part descriptive and part not. We thought that the best evidence for this view comes from Experiment 3 of Tobia et al. (2020, pp. 198-201), which likely avoids the Faulty Test Hypothesis. That experiment was one inspiration for our present experiments.

In light of the evidence that folk natural kind terms cannot all be explained simply by the Description Theory or the Causal-Historical Theory, we proposed that “reference determination for natural kind terms is almost definitely more varied than any of the extant theories suggest” (2021, sec. 5.1). None of the five identified theories applies to all natural kind terms, yet each theory has some place in the full story of the reference determination of these terms. In particular,

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’, ‘whale’). (2021, sec. 5.1)

These speculations about the semantics of natural kind terms led to some suggestions for future tests (sec. 5.4), which guided our aims in this paper.

Before describing these aims we must mention a further complication that was emphasized to us by Andrea Bianchi. In testing the reference of natural kind terms, it is common to associate anti-descriptivism with a particular version of the Causal-Historical Theory: the view that the reference borrowed by the folk is fixed in deep structural properties. This is largely because the Description Theory and this version of the Causal-Historical Theory have historically been the two primary competing theories. However, it is of course possible for the Causal-Historical Theory’s picture of reference borrowing to be true, while the reference is fixed in practical functions rather than deep structures.[[5]](#footnote-5) Our experiments were not designed to test such a theory directly. However, as we will argue, we think our results provide evidence against it as a theory of biological terms.

*1.4 Our Aims*

We chose to test the folk use of two biological terms, one that is of high practical interest and one that is of scientific interest but of little to no practical interest. We predicted that these two terms would be covered by different theories of reference. We tested participants by EP followed by TVJ, finishing with a request for an explanation of answers.

We chose ‘rice’ for the “Practical” term. We take ‘rice’ to be a good example of a biological term that is widely used by folk because of their practical interest in its referent: rice is a significant part of the human diet. ‘Rice’ will obviously not be a novel term for participants (as ‘catoblepas’ was in “Testing”) and hence it does not lead to the Faulty Test Hypothesis. We supposed that the reference of the folk’s use of ‘rice’ was fixed largely by the folk themselves in their many acquaintances with rice in dining room and kitchen. We predicted that there would be two factors to its reference determination, a superficial-descriptive one and a deep-causal one. That prediction was confirmed.

We chose ‘Rio de Janeiro Myrtle’ for the “Scientific” term. It refers to an extinct species in the myrtle genus and is almost certainly used causal-historically by biologists– see the discussion of ‘echidna’ in “Testing” (sec. 5.1). Its referent has little or no practical interest to the folk. So, we predicted that folk would borrow their reference from the biologists and their use of the term would be covered by the Causal-Historical Theory. Given that this is the prediction we are testing, the fact that ‘Rio de Janeiro Myrtle’ will surely be a novel term for participants, does not lead to the Faulty Test Hypothesis. Just as our earlier experiments directly tested the Causal-Historical Theory of the novel term ‘catoblepas’, so too our present experiments directly test that theory of ‘Rio de Janeiro Myrtle’. But whereas that theory was found wanting for ‘catoblepas’, we predicted that it would be confirmed for ‘Rio de Janeiro Myrtle’: we predicted that the results would be consistently nondescriptive. And so they were.

In brief, we predicted sharply different results for a Practical term and a Scientific term. And that is what we found.

Where a folk term has both a descriptive and causal-historical element to its reference determination, it might be covered by the Hybrid Theory, the Different Idiolects Theory, or the Ambiguity Theory. The latter two theories require two meanings for the folk use of ‘rice’. The Different Idiolects Theory requires that these meanings be exemplified by the regular use of ‘rice’ descriptively by one group of folk and causal-historically by another group. The Ambiguity Theory requires that these meanings be exemplified in (near enough) each member of the folk regularly using ‘rice’ in each of the two ways. We did not test for these regularities but expect that they are not to be found.[[6]](#footnote-6) If that is right then the Hybrid Theory applies to the folk use of ‘rice’.

We noted that Tobia et al (2020) found that participants were more likely to use the allegedly ambiguous term ‘salmon’ descriptively in a “Practical” context - a legal one - and nondescriptively in a “Scientific” context. We tested for this contextual variation in our terms. We predicted no such difference for the Scientific term, ‘Rio de Janeiro Myrtle’, because its reference would always be borrowed from biologists. We found no difference. We thought that there might be such a difference for the term ‘rice’, a Practical term like Tobia et al’s ‘salmon’, but found none. This ‘rice’ result is inconsistent with the prediction of the Ambiguity Theory. The result is consistent, however, with the Hybrid Theory which, as noted, makes no prediction of variation in context (though it could accommodate such a variation).

**2. First Test: Rice and Rio de Janeiro Myrtle**

*2.1 Vignettes*

Each participant was presented with one of four vignettes: a rice vignette with either a Practical context or Scientific context continuation; or a Rio de Janeiro Myrtle vignette with either a Practical context or Scientific context continuation:

**Rice Vignette**:

Rice is a seed that is an important part of the diet of millions around the world. It comes from the grass plant *Oryza sativa*. Growing *Oryza sativa* is a difficult task, which requires a great deal of water. This has led to food shortages around the world, often caused by drought. Scientists at the Maxwell Laboratory have been working to combat this problem. Rather than trying to modify the genes of *Oryza sativa*, they are developing a new plant altogether. This plant will produce seeds that have the exact same look, taste, and nutritional content as the seeds of *Oryza sativa*. But the new plant is genetically very different and is not *Oryza sativa*. Because of this genetic difference, the plants will be much easier to grow. It will require much less water, and can be grown in a wider variety of climates. Scientists estimate that this new plant can produce a lot more food than *Oryza sativa* on half as much water. The scientists believe, therefore, that this new method can solve the food shortages caused by insufficient rice production.

The two different continuations of the vignette were as follows:

*Rice Continuation in Practical Context:*

Suppose that these scientists succeed. The Maxwell Laboratory starts growing the plant and producing the new seeds. Now suppose that one of the Maxwell laboratory workers also works for a local Asian restaurant. His duties include providing grains for the restaurant. On this day, he takes along a bag of the new seeds and tells the chef about them. The chef decides to serve them in all the dishes listed on the menu as including rice.

*Rice Continuation in Scientific Context:*

Suppose that these scientists succeed. The Maxwell Laboratory starts growing the plant and producing the new seeds. Now suppose that one of the Maxwell laboratory workers also works as a teaching assistant in botany classes at the local college. His duties include providing plant samples for the students’ practical testing requirements. On this day he is asked to provide rice and decides to give the students some of the new seeds that he has brought with him from the laboratory.

**Rio de Janeiro Myrtle Vignette:**

The Rio de Janeiro Myrtle, *Campomanesia lundiana*, a species of the Myrtle genus *Campomanesia*, were thought to have gone extinct in the mid-19th century. Then recently some botanists announced that they had found Rio de Janeiro Myrtles growing in Asia, far from their old habitat in Brazil. Suppose, however, that scientists at the Maxwell Laboratory subject these newly discovered trees to genetic testing and conclude that they are very different from *Campomanesia lundiana*. The newly discovered trees are simply the result of convergent evolution: two very different evolutionary histories producing some of the same features. To the human eye, the newly discovered trees are exactly like the Rio de Janeiro Myrtles that used to grow in Brazil; the only differences are microscopic. Given their different genetic make ups and evolutionary histories, biologists do not consider these newly discovered trees to be members of the same species as the Rio de Janeiro Myrtles nor in the genus *Campomanesia*.

The two different continuations of the vignette were as follows:

*Rio de Janeiro Myrtle Continuation in Practical Context:*

Now suppose that one of the Maxwell laboratory workers also volunteers for the local park service in his spare time. His duties include planting trees and bushes along the walkways in the parks. One day he is asked to plant some myrtle trees in a new park that is being built, and he decides to plant some of the newly discovered trees that he brought home from the laboratory.

*Rio de Janeiro Myrtle Continuation in Scientific Context:*

Now suppose that one of the Maxwell laboratory workers also works as a teaching assistant in botany classes at the local college. His duties include providing plant samples for the students’ practical testing requirements. On this day he is asked to provide a myrtle tree and decides to give the students some of the newly discovered trees that he has brought with him from the laboratory.

We think that these vignettes avoid the Faulty Test Hypothesis. They are not fantastical or otherwise outside the realm of real-world possibilities. And, as noted (1.4), ‘rice’ is not a novel term for participants and, given the proviso, the novelty of ‘Rio de Janeiro’ myrtle is acceptable if it is indeed causal-historical as predicted.

The vignettes and contexts were inspired by those used in Experiment 3 in Tobia et al. (2020), but some differences are worth emphasizing:

(1) Tobia et al.’s vignettes use genetically modified fish (salmon) as examples of organisms that have the superficial properties of a biological kind but not the deep structural properties the Causal-Historical Theory usually takes to be relevant. We think this is problematic, because it is not obvious that the genetically modified fish lack the necessary “essence” to be salmon. They share a history with salmon, and genetic modification is what evolution is all about. We think it is likely to be indeterminate whether the genetically modified fish are biologically salmon in just the way that it is indeterminate whether organisms immediately ancestral to salmon are salmon. To avoid this indeterminacy, our vignette explicitly gives the novel specimens histories and genes that are totally different from the kind in question.[[7]](#footnote-7)

(2) Tobia et al. end their vignette with the following sentence:

The laboratory issues a report stating that while the fish do not belong to the same scientific category as familiar salmon, this difference is immaterial for any purpose other than scientific classification.

We were worried about the final clause of this sentence. Tobia et al. prompt participants to say whether they agree with the statement, “The fish from the laboratory are salmon” in various contexts. The worry is that the final clause comes dangerously close to telling participants directly that, for nonscientists, these fish count as salmon, particularly in a nonscientific context. So, participants are almost being told to agree with the prompt’s statement, particularly in the legal context. This quasi-instruction may be responsible for the apparent contextual effects that Tobia et al. found in their results. We told participants only that biologists do not consider the new specimens to be rice/Rio de Janeiro Myrtles, giving no guidance about what nonscientists consider the specimens to be (in any context).

(3) Tobia et al. chose a legal context for their nonscientific “Practical” one. This raises the worry that the participants view of whether the laboratory fish are salmon will be influenced by their view of whether the *law* would count them as salmon, rather than simply answering whether, in *their* view, they really are salmon. Yet what we want is an example of the participants’ usage without any such influences. We thought that our restaurant and park contexts provide good examples of that.

(4) Tobia et al. tested not only in scientific and legal contexts but also in “neutral” ones. We simplified by dropping neutral contexts.

*2.2 Methods*

We recruited 320 participants through Amazon’s MechanicalTurk, who were each compensated for their participation.[[8]](#footnote-8) After answering demographic questions and a mandatory generic attention check question, participants were asked whether they had heard of rice/myrtle trees (depending on which vignette they were going to get). Participants were then directed to another page with a vignette, context, and EP prompt for each of the four conditions.

The form of these prompts was influenced by an informal test we ran on friends and family. We gave them the Rice Practical vignette with the simple prompt, “Describe briefly, in your own words, what the chef served”. Several of those tested thought that the prompt was a trick question, or that we were asking them to judge the morality or legality of the chef’s actions. To avoid these misunderstandings, we expanded the four prompts to include explicit instructions to ignore whether the actions were appropriate, and a reassurance that this was not a trick question. The resulting prompts are illustrated by these two:

**Rice EP Prompt in Practical Context**:

Leaving aside whether this is an appropriate thing for the chef to do, please describe, in your own words, in a sentence or two, what the chef served in all those dishes. (This is not a trick question; just describe what was served.)

**Rio de Janeiro Myrtle EP Prompt in Scientific Context**:

Leaving aside whether this is an appropriate thing for the laboratory worker to do, please describe, in your own words, in a sentence or two, what the laboratory worker provided to the students. (This is not a trick question; just describe what was provided.)

The Description Theory of reference predicts responses indicating that the novel seeds (trees) are rice (Rio de Janeiro Myrtles), since they fit the superficial descriptions speakers associate with them. The Causal-Historical Theory predicts responses indicating that the novel seeds (trees) are *not* rice (Rio de Janeiro Myrtles) because they lack the deep genetic and evolutionary properties that would, according to scientists, make them rice (Rio de Janeiro Myrtles). The Different Idiolects Theory predicts a significant number of responses that the seeds are rice (Rio de Janeiro Myrtles) and a significant number that they are not. So too do the Ambiguity Theory and the Hybrid Theory.

The Description, Causal-Historical, and Different Idiolects Theory each predict that responses will not vary between the Practical and Scientific contexts for either term. In contrast, the Ambiguity Theory predicts variation in context. Thus, if ‘rice’ is ambiguous, it seems clear that the descriptivist meaning is more appropriate in the Practical context described, and the causal-historical meaning, in the Scientific context described. Finally, the Hybrid Theory does not predict a variation in context but can accommodate one. [[9]](#footnote-9)

We coded participants’ EP responses according to instructions written before the study. Minor changes had to be made to the instructions to account for the surprising fact that some participants indicated that the new trees were myrtles, but not Rio de Janeiro Myrtles. Our coding instructions wrongly assumed that participants’ views of whether the new plants were myrtles would reflect their views of whether the new plants were Rio de Janeiro Myrtles. To fix this issue, the instructions for “Rio de Janeiro Myrtle” were separated into two sets: one for ‘myrtle’, one for ‘Rio de Janeiro Myrtle’.

Responses judged to be in accordance with descriptivist predictions were coded D; responses judged to indicate uncertainty about whether or not the seed (tree) was rice (Rio de Janeiro Myrtle) were coded N (for neutral); responses judged to be in accordance with nondescriptivist (causal-historical) predictions were coded CH; responses which were not clearly in any of these three categories were coded X and discarded. We thought that we might have to compare our coding with those of trained independent coders but, given the results below, we decided that this was unnecessary.

After answering the EP prompt, participants were taken to another page and given a TVJ question for each of the four conditions. Here are two:

**Rice TVJ Prompt in Practical Context**:

Which one of the following statements do you think comes closest to the truth about what the chef served?

* It was rice
* It was not rice
* It's hard to say whether or not it was rice

**Rio de Janeiro Myrtle TVJ Prompt in Scientific Context**:

Which one of the following statements do you think comes closest to the truth about what the laboratory worker provided?

* They were Rio de Janeiro Myrtles
* They were not Rio de Janeiro Myrtles
* It's hard to say whether or not they were Rio de Janeiro Myrtles

For each of these TVJ questions, the Description Theory predicts that participants will choose the affirmative answer, while the Causal-Historical Theory predicts that participants will choose the negative answer. We intended the third answer to be neutral between ambiguity and hybrid theories, indicating awareness of more than one potential use or meaning of the term, and uncertainty as to which meaning/use to apply here.

Participants were then given, on a separate page, the following prompt to explain their answers to the EP and TVJ questions:

**Explanation Prompt**:

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

Following the method of Tobia et al (2020), we then gave participants two content-based attention check questions:

**Rice Attention Questions (both contexts):**

1. Are the seeds from the new plant identical to rice in terms of all their observable properties (e.g., appearance, taste, etc.)?
	* Yes
	* No
2. Are the seeds from the new plant identical to rice in terms of their genetic structure and evolutionary history?
	* Yes
	* No

**Rio de Janeiro Myrtle Attention Questions (both contexts):**

1. Are the newly discovered trees identical to the Rio de Janeiro Myrtle in terms of all their observable properties (e.g., appearance)?
	* Yes
	* No
2. Are the newly discovered trees identical to the Rio de Janeiro Myrtle in terms of their genetic structure and evolutionary history?
	* Yes
	* No

*2.3 Results*

Given our main predictions, the key results are: (1) Participants are significantly more likely to give CH answers for ‘Rio de Janeiro Myrtle’ than for ‘rice’ (*p* = 0.0001, 0.0001); the odds of getting a CH response for ‘Rio’ is 11 times higher than the odds for ‘Rice’ in the EP; 5 times higher in the TVJ. (2) Participants are significantly more likely to give CH than D answers for ‘Rio de Janeiro Myrtle’ (*p* = 0.0256, 0.0004, 0.0002, <0.0001); the odds of getting a CH response are 10 greater than chance in the EP, 18 times greater in the TVJ. (3) In contrast, on EP conditions, CH and D answers for ‘rice’ do not differ significantly from chance (*p* = 0.3359, 0.7961). However, on TVJ conditions, answers favor CH significantly (*p* = 0.0002, 0.0353).

Some other results are notable. Participants in the TVJ conditions are significantly more likely to choose the “hard to say” answer for ‘rice’ than for ‘Rio de Janeiro’ (*p* = 0.0035); the odds of a “hard to say” response are nearly 11 times higher for ‘rice’ than for ‘Rio’. There is no significant variation across Practical and Scientific contexts (*p* = 0.8273, 0.2352). Surprisingly, the EP answers for ‘myrtle’ are significantly more descriptivist than those for ‘rice’ (*p* = 0.0265) and so, of course, much more so than those for ‘Rio de Janeiro Myrtle’ (*p* = <0.0001).

Of the 320 participants recruited, 32 were removed for failing the initial generic attention check, 90 were removed for failing the content-based attention checks,[[10]](#footnote-10) and 1 was removed for obviously making no reasonable attempt to answer both the EP and Explanation prompts.[[11]](#footnote-11) All but one of the participants in the two Rice conditions reported that they had heard of rice; 36.5% of participants in the two Rio de Janeiro Myrtle conditions reported that had heard of myrtle trees. We found no significant differences in the results for the participants who had heard of myrtle trees and those who had not.

To our surprise, we discovered that many participants clearly indicated in their EP response and in their Explanation response that they took the trees in the Rio de Janeiro Myrtle vignette to be myrtles, but not to be Rio de Janeiro Myrtles. For example, one participant gave this EP response: “He gave them the Myrtle that looks like the Brazil Myrtle. But in reality it’s the new one that has no genetic makeup of the ORIGINAL one”. We therefore introduced a separate coding for ‘myrtle’ and for ‘Rio de Janeiro Myrtle’. The results of our codings can be found in Table 1 and Figure 1 below (with ‘CH’ coding a causal-historical response; ‘D’, a descriptivist response; ‘N’, a “neutral” response; ‘X’, a discard):

**Table 1**

*EP Results*

|  |  |  |
| --- | --- | --- |
| Test |  CH D N\_\_\_\_\_  n % n % n %  | X |
| Rice Practical | 22 | 62.9 | 13 | 37.1 | 0 | 0 | 29 |
| Rice Scientific | 17 | 56.7 | 13 | 43.3 | 0 | 0 | 29 |
| Rio de Janeiro Practical | 15 | 88.2 | 2 | 11.7 | 0 | 0 | 19 |
| Rio de Janeiro Scientific | 19 | 100 | 0 | 0 | 0 | 0 | 29 |
| Myrtle Practical | 6 | 28.6 | 15 | 71.4 | 0 | 0 | 15 |
| Myrtle Scientific | 8 | 47 | 9 | 53 | 0 | 0 | 21 |

*Note.* This table reports results of our elicited production tests. For each response type (CH, D, and N) percentages are out of all *undiscarded* responses, i.e. responses not coded X.

**Figure 1**

*EP Response percentages*

The results of the TVJ questions can be found in Table 2 and Figure 2 below:

**Table 2**

*TVJ Results*

|  |  |
| --- | --- |
| Test |  CH D “hard to say”  n % n % n %  |
| Rice Practical | 49 | 76.5 | 10 | 15.6 | 5 | 7.8 |
| Rice Scientific | 35 | 59.3 | 13 | 22 | 11 | 18.6 |
| Rio de Janeiro Practical | 32 | 88.9 | 3 | 8.3 | 1 | 2.8 |
| Rio de Janeiro Scientific | 36 | 94.7 | 2 | 5.3 | 0 | 0 |

*Note.* This table reports results of our truth value judgment tests. For each response type, percentages are out of all responses.

**Figure 2**

*TVJ Response percentages*

Almost all of these results are statistically significantly independent from chance (33%), but we don’t think this fact is particularly informative. This significance is largely attributable to the fact that no participant was coded N (“neutral”) in the EP, and relatively few participants chose “hard to say”. Some comparisons, which we think are more enlightening, can be found in Table 3 below.

**Table 3**

*Comparisons*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | *N* | *p* | *Odds Ratio* | *df* |
| Comparing rates of CH responses: Rice vs. Rio de Janeiro\*\* |
| EP  | 101 | 0.0001\* | 0.09 | 1 |
| TVJ | 197 | 0.0001\* | 0.19 | 1 |
| Comparing rates of CH responses: Practical Context vs. Scientific Context\*\*\* |
| EP  | 101 | 0.8273 | 0.89 | 1 |
| TVJ  | 197 | 0.2352 | 1.56 | 1 |
| Comparing rates of “hard to say” responses: Rice vs. Rio de Janeiro Myrtle\*\* |
| TVJ  | 197 | 0.0035\* | 10.92 | 1 |
| Comparing rates of CH responses: Rio de Janeiro Myrtle vs. Myrtle\*\* |
| EP  | 74 | <0.0001\* | 3.43 | 1 |
| Comparing rates of CH responses: Rice vs. Myrtle\*\* |
| EP  | 103 | 0.0265\* | 2.57 | 1 |

*Note.* This table compares rates at which participants gave a particular response (either CH or “hard to say”) across different tests. All p-values are calculated using a two-sided Fisher’s Exact Test.

‘\*’ = significant at p<0.05

‘\*\*’ = combines results for Practical and Scientific contexts

‘\*\*\*’ = combines results for Rice and Rio de Janeiro Myrtle

The table above compares percentages *out of all responses*.[[12]](#footnote-12) But we are also interested in what percentage of participants chose the CH (causal-historical) response, or the D (descriptivist) response, *among participants* *who chose either CH or D,* ignoring the participants who chose “hard to say” (and ignoring the N coding option). For the EP, these are the same percentages given in the original table. For the TVJ, these percentages can be found in Table 4.

**Table 4**

*TVJ percentages without “hard to say”*

|  |  |  |
| --- | --- | --- |
| Test | CH | D  |
| Rice Practical | 83.1% | 16.9% |
| Rice Scientific | 72.9% | 27.1% |
| Rio de Janeiro Practical | 91.4% | 8.6% |
| Rio de Janeiro Scientific | 94.7% | 5.3% |

*Note*. This table reports percentages of CH and D responses among participants who chose either CH to D responses.

Comparisons of results to random chance are in Table 5. In each row, the ratio of CH responses to D responses (leaving out N and “hard to say” responses) is compared to random chance (50%) using a two-tailed Fisher’s exact test.

**Table 5**

*CH vs. D: Comparisons to 50% chance*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | *N* | *p* | *Odds Ratio* | *df* |
| Rice Practical |
| EP  | 35 | 0.3359 | 1.69 | 1 |
| TVJ | 59 | 0.0002\* | 4.9 | 1 |
| Rice Scientific |
| EP  | 30 | 0.7961 | 1.31 | 1 |
| TVJ  | 48 | 0.0353\* | 2.69 | 1 |
| Rio de Janeiro Myrtle Practical |
| EP | 17 | 0.0256\* | 7.5 | 1 |
| TVJ | 35 | 0.0002\* | 10.67 | 1 |
| Rio de Janeiro Myrtle Scientific |
| EP  | 19 | 0.0004\* | undefined | 1 |
| TVJ | 38 | <0.0001\* | 18 | 1 |
| Myrtle Practical |
| EP  | 21 | 0.2082 | 0.4 | 1 |
| Myrtle Scientific |
| EP  | 17 | 1 | 0.89 | 1 |

*Note.* This table compares the ratio of CH and D responses in the EP and TVJ tests to 50% random chance. All p-values are calculated using a two-sided Fisher’s Exact Test.

‘\*’ = significant at p<0.05

No EP response was coded N. For each of the terms, the majority of responses were either emphatically a descriptivist D, emphatically a nondescriptivist (causal-historical) CH, or an obvious discard X. See, for example:

**Emphatic D:** “The chef was serving a new type of rice that is easier to grow. ”

**Emphatic CH:** “The chef served a grain that was developed by scientists to be similar to rice but was not rice”

**Obvious X:** “He served seeds of a new plant.”

To check our EP codings, we compared the D and CH codings with the TVJ answers in the four conditions. 89 out of 101 EP codings matched the TVJ answers: participants whose EP responses were coded D chose the D answer in the TVJ test; those coded CH, the CH answer. The 12 exceptions were in the Rice conditions. 5 participants who were coded D on the EP chose the CH answer on the TVJ; 7 who were coded D, chose “hard to say”. A re-examination showed that our coding of these 12 was straightforwardly in accord with the coding instructions and specimen examples. The instruction for Rice Practical is: “It is clear from a participant’s response that they think that the seeds served at the restaurant and provided by the lab worker were rice….” This requires that, for example, the responses that began, “It is a new rice seed…” and “rice that grew with much less water…”, be coded D. Similar coding instructions for Rice Scientific require that responses such as, “The TA gave the students starter plants from the genetically changed rice…” and “They provided seeds for the new variant of rice…”, be coded D. Each of these 12 exceptions gave similar evidence in the EP response that the participant believed that the seeds/plants were rice. Nonetheless, these participants show in their TVJ answers and Explanations that they either believe that the seeds/plants are not rice or that they have doubts whether they are. How is this to be explained? We take these participants to have been *speaking loosely* in their “free” EP response, much as when one refers to almond milk as “milk”. This looseness disappeared when they were put on the spot by the “forced-choice” TVJ tests and subsequent Explanations.

We concluded that an independent coder was unnecessary. Our EP codings were largely confirmed by the TVJ results. Where they were not, this was not due to any failing that an independent coder could resolve.

In some conditions, participants with discarded EP responses chose TVJ answers at rates noticeably different from those of participants with undiscarded EP responses. The most striking case was Rice Practical: only 6.9% of participants with discarded EP responses chose the D TVJ option, while 22.9% of participants with undiscarded EP responses chose the D TVJ option. We think that this is to be explained by our cautious coding together with the nature of our EP prompt. Participant were only asked what the chef served; they were not asked what the chef did *not* serve. A participant was coded D if they called the stuff that the chef served “rice”. But they were only coded CH if they explicitly said that the stuff the chef served was *not* rice, which meant answering a question they were not directly asked. Many did not do this, giving answers like “he served a rice-like seed” or “the chef sold genetically modified plants” which, given our cautious coding, were discarded as not explicitly saying whether or not the new seeds were rice. This meant that, in answering the EP prompt, participants who use “rice” descriptively were much less likely to be discarded than participants who use “rice” causal-historically. This is an unavoidable consequence of our open-ended prompt.

*2.4 Discussion*

Our aim was to test the reference of two biological terms, ‘rice’, a term of high practical interest, and ‘Rio de Janeiro Myrtle’, a term of scientific interest but of little to no practical interest. Our main predictions in section 1.4 were:

1. That these two terms are covered by different theories of reference. This was confirmed with high significance in both EP and TVJ tests across all conditions and contexts. Causal-historical, nondescriptive (CH) responses were very much more likely for ‘Rio de Janeiro Myrtle’ than for ‘rice’ (p = 0.0001 on both EP and TVJ in combined contexts).
2. That participants will borrow their reference of ‘Rio de Janeiro Myrtle’ from the biologists and so use the term causal-historically, with no descriptive element to its reference determination. And that is what we found. The term’s CH responses on the EP were 88.2% (Practical context) and 100% (Scientific context); on the TVJ 88.9% (Practical), 94.7% (Scientific). The differences between the CH and D scores (with “hard to say” eliminated) are also significant, and the odds of a CH response are over 10 times greater than chance.
3. That there are two factors to the reference determination of ‘rice’: a superficial-descriptive D-factor and a deep-causal CH-factor. Now if ‘rice’ were purely causal-historical, we would expect responses much like the decisively CH ones for ‘Rio de Janeiro Myrtle’ just discussed. Yet, as (1) reveals, that was not what we found: the “rice” responses were statistically significantly more descriptivist than the ‘Rio de Janeiro Myrtle’ ones. Indeed, in the EP test, D responses were 37.1% (Practical) and 43.3% (Scientific). The “hard to say” TVJ responses provide further evidence of two factors to the reference determination of ‘rice’. Those factors would pull in different directions, likely yielding some “hard to say” responses. In contrast, the single CH factor for ‘Rio de Janeiro Myrtle’ should yield few “hard to say” responses. So, there should be significantly more “hard to say” responses for ‘rice’ than for ‘Rio de Janeiro Myrtle’. That is what we found (p = 0.0035 and odds ratio = 10.96 in combined contexts).

Although the TVJ responses for ‘rice’ (with “hard to say” eliminated) were significantly more CH than D, the effect size (as measured by the odds ratios) was comparatively small: the odds ratio was less than 5 in both “rice” conditions, and over 10 in both “Rio de Janeiro Myrtle” conditions. There was still a notable proportion of D responses in both “rice” conditions: 16.9% (Practical) and 27% (Scientific). And our prediction was not that participants will follow the causal-historical factor and the descriptive factor at equal rates when the factors pull in different directions. The comparatively high percentages of D and “hard to say” responses are strong evidence of two factors in the reference determination of ‘rice’, even if participants are significantly more likely to resolve the tension between the two factors causal-historically than descriptively.

We take these results to be evidence against the usual “deep properties” version of the Causal-Historical Theory for a Practical term like ‘rice’. But that is not the only version of the Causal-Historical Theory; we must also consider the Bianchi complication mentioned in section 1.3. The usual version of the Causal-Historical Theory takes “groundings” to fix reference in deep structural properties. But there is another version of the theory according to which groundings fix reference in something else. In the case of ‘rice’, the proposal is likely to be that groundings fix reference in rice’s *practical role as food* rather than in the deep structural properties of *Oryza Sativa*. On the usual version, the reference fixers, who would mostly be just rice eaters not scientists, would typically be quite ignorant of the relevant deep structural properties. On the alternative Bianchi version, those fixers would likely be only a little bit ignorant of the relevant practical role.

All versions of the Causal-Historical Theory predict that participants will look for information in the vignette on what the relevant experts would pick out as rice. On the usual version, participants will look for the opinions of scientists (biologists) about whether the seeds have the deep properties of rice. On the alternative version, participants will look for the opinions, presumably, of food industry experts about whether the seeds can play the practical role of rice. But the vignettes say nothing explicitly about what a food expert picks out as rice; our experiments were not designed to test this version of the theory. The closest thing to a food expert in the vignettes is the chef in the Practical context, but the vignette does not say whether she thinks that what she serves is rice. Still, participants might take the chef as *implicitly* calling what she serves “rice”. But in that case, this version of the Causal-Historical Theory predicts that the participants will follow suit and also call the new seeds “rice”. This is not what we found: the responses were more CH than D, significantly so in the TVJ (p = 0.0002). The vignette in the Scientific context has no food expert, just the irrelevant biology experts. So, participants should be fairly stumped, yielding a 50-50 result. That was true for EP but significantly wrong on TVJ (p = 0.0353). We therefore take our results to be evidence against this version of the Causal-Historical Theory for ‘rice’.

The results support a Causal-Historical Theory of ‘Rio de Janeiro Myrtle’, and are evidence against a Causal-Historical Theory of ‘rice’. The results for ‘rice’ are compatible with the Hybrid Theory, the Different Idiolects Theory and, aside from the variation problem below, the Ambiguity Theory. As noted (1.2, 1.4), the latter two theories require the regular uses of ‘rice’ with two distinct meanings: in the case of Ambiguity, two uses by each person; in the case of Different Idiolects, one use by some people, another, by other people. Given the implausibility that there are such regularities, we take our results to support the Hybrid Theory of ‘rice’.

We found no statistically significant differences between the Practical and Scientific contexts for any of the terms tested, whether by EP or TVJ. This is as predicted by the Causal-Historical Theory. So, it is what we predicted for the Scientific term, ‘Rio de Janeiro Myrtle’; the term’s reference is always borrowed from biologists. The results are also *compatible with* the Hybrid Theory which we favored for the Practical term ‘rice’. The Hybrid Theory predicts a variation in usage when its two reference-determining factors pull in different directions, as they do in these experiments, but it does not predict that this variation is *influenced by context*.

In contrast, the lack of variation goes against the Ambiguity Theory. Thus, if ‘rice’ is ambiguous, it seems clear that the descriptivist meaning is more appropriate in the Practical context described, and the causal-historical meaning, in the Scientific context described. So, responses in the Rice Practical context should have been significantly more descriptivist than those in the Rice Scientific context. That is not what we found. Indeed, we found a greater descriptivist response in the Rice *Scientific* context than in the Rice Practical context (though not significantly greater). In sum, our results favor the Hybrid Theory over the Ambiguity Theory.

These results are strikingly different from Tobia et al’s results with ‘salmon’, results that partly led them to embrace the Ambiguity Theory. They found that participants were more likely to use the allegedly ambiguous term ‘salmon’ descriptively in a Practical context - a legal one - and nondescriptively in a Scientific context. We suspect that the explanation of this different in results is to be found in the differences between their vignette and ours, itemized in section 2.1, particularly difference (2). As noted above, Tobia et al.’s vignette comes dangerously close to explicitly telling participants that the new fish are not salmon in a Scientific context, and are salmon in other contexts. It’s possible that without such instructions, participants would not answer differently across different contexts.

To our surprise, we found another statistically significant result: in our EP codings, ‘myrtle’ is more descriptivist than ‘rice’ (*p* = 0.0265) and so, of course, much more so than ‘Rio de Janeiro Myrtle’ (*p* = <0.0001). The term ‘myrtle’ is likely not of practical interest for most participants, and the vignette states that although the new trees look like Rio de Janeiro Myrtles (and therefore myrtles), biologists do not consider them to be in the same species as Rio de Janeiro Myrtles, *nor in the genus Campomanesia*. Given that the vignette refers to *Campomanesia* as “the Myrtle genus”, this *almost* tells participants that biologists do not consider the new trees to be myrtles. We therefore supposed that ‘myrtle’ would be as non-descriptivist as ‘Rio de Janeiro Myrtle’, but this is not what we found.

We were not sure what to make of these ‘myrtle’ results. We noted earlier (2.3) that the TVJ results for ‘rice’ seemed to show that *some* apparently descriptivist EP responses arose from the loose use of the term (analogous to the use of ‘milk’ in ‘almond milk’). Could *all* the apparently descriptivist EP responses for ‘myrtle’ be similarly explained away? Since we had not asked a TVJ question for ‘myrtle’, we couldn’t answer. We therefore decided to run a follow-up test that asked such a question.

**3. Follow-up Test: Myrtle and Rio de Janeiro Myrtle**

*3.1 Vignette and Context*

Since we were only interested in answering our question about ‘myrtle’, we used just the Rio de Janeiro Myrtle vignette. We had concluded that context made no statistically significant difference in responses and so used just one continuation, that in Scientific Context.

 We made a minor change to our vignette. One potential explanation for the highly descriptivist use of ‘myrtle’ in the original test was that participants had simply not understood that biologists not only considered the new trees not to be Rio de Janeiro Myrtles but also not to be myrtles. To remedy this, we changed the last sentence of the vignette to end “nor in the Myrtle genus” rather than “nor in the genus *Campomanesia*”. We left the Continuation in Scientific Context unchanged.

*3.2 Methods*

We recruited 100 participants through Amazon’s MechanicalTurk, who were each compensated for their participation.[[13]](#footnote-13) After answering demographic questions and a mandatory generic attention check question, participants were asked whether or not they had heard of myrtle trees. Participants were then directed to another page with the vignette, context, and the same Rio de Janeiro Myrtle Scientific EP prompt used in the original experiment.

Next, on a separate page, participants were presented with two TVJ questions: the original Rio de Janeiro Myrtle Scientific prompt, and a new prompt asking about the term ‘myrtle:

**Myrtle TVJ Prompt:**

Which one of the following statements do you think comes closest to the truth about what the laboratory worker provided?

* They were myrtles
* They were not myrtles
* It's hard to say whether or not they were myrtles

The order of the two questions was randomized, but the two questions appeared together on the same page.

 Participants were then given the following prompt on a separate page:

**Explanation Prompt:**

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

The Explanation prompt was edited slightly to accommodate the fact that participants were now being asked to explain their answers to three questions.

Finally, participants were given the original Rio de Janeiro Myrtle Scientific content-based attention check questions on a separate page.

* 1. *Results*

Of the 100 participants recruited for the follow-up test, 18 were removed for failing the initial generic attention check, 44 were removed for failing the content-based attention checks,[[14]](#footnote-14) 3 were removed because they had already participated in the original test, and 6 were removed for obviously making no reasonable attempt to answer both the EP and Explanation prompts.[[15]](#footnote-15) 14 out of the remaining 29 participants reported having heard of myrtle trees. We found no significant differences in the results for the participants who had heard of myrtle trees and those who had not.

The EP and TVJ results are listed in the tables and figures below:

**Table 6**

*Follow-up EP Results*

|  |  |  |
| --- | --- | --- |
| Test |  CH D N\_ \_  n % n % n %  | X |
| Rio de Janeiro Myrtle | 13 | 100 | 0 | 0 | 0 | 0 | 16 |
| Myrtle  | 8 | 47 | 8 | 53 | 0 | 0 | 13 |

*Note.* This table reports results of our follow-up elicited production tests. For each response type (CH, D, and N) percentages are out of all *undiscarded* responses, i.e. responses not coded X.

**Table 7**

*Follow-up TVJ Results*

|  |  |
| --- | --- |
| Test |  CH D “hard to say”  n % n % n %  |
| Rio de Janeiro Myrtle | 24 | 82.8 | 2 | 6.9 | 3 | 10.3 |
| Myrtle | 19 | 65.5 | 7 | 24.1 | 3 | 10.3 |

*Note.* This table reports results of our follow-up truth value judgment tests. For each response type, percentages are out of all responses.

**Figure 3**

***Follow-up EP results***

**Figure 4**

***Follow-up TVJ results***

**Table 8**

*Follow-up Comparisons*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | *N* | *p* | *Odds Ratio* | *df* |
| Comparing rates of CH responses: Rio de Janeiro Myrtle vs. Myrtle |
| EP  | 29 | 0.0033\* | undefined | 1 |
| TVJ | 58 | 0.2299 | 2.53 | 1 |
| Comparing rates of CH responses: Rice Scientific vs. Myrtle |
| EP  | 46 | 0.7604 | 1.31 | 1 |
| TVJ  | 88 | 0.6457 | 0.77 | 1 |

*Note.* This table compares rates at which participants gave a CH response across different tests. The data for Rice Scientific is taken from the original experiment; all other data comes from the follow-up test. All p-values are calculated using a two-sided Fisher’s Exact Test.

‘\*’ = significant at p<0.05

**Table 9**

*Follow-up TVJ percentages without “hard to say”*

|  |  |  |
| --- | --- | --- |
| Test | CH | D  |
| Rio de Janeiro Myrtle | 92.3 | 7.7 |
| Myrtle | 73.1 | 26.9 |

*Note*. This table reports percentages of CH and D responses among participants who chose either CH to D responses in our follow-up TVJ test.

**Table 10**

*Follow-up CH vs. D: Comparisons to 50% chance*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | *N* | *p* | *Odds Ratio* | *df* |
| Rio de Janeiro Myrtle |
| EP  | 13 | 0.0052\* | undefined | 1 |
| TVJ | 26 | 0.0016\* | 12.0 | 1 |
| Myrtle |
| EP  | 16 | 1 | 1 | 1 |
| TVJ  | 26 | 0.1534 | 2.71 | 1 |

*Note.* This table compares the ratio of CH and D responses in our follow-up EP and TVJ tests to 50% random chance. All p-values are calculated using a two-sided Fisher’s Exact Test.

‘\*’ = significant at p<0.05

As in the original experiment, no EP response was coded N. For each of the terms, the majority of responses were either emphatically D, emphatically CH, or an obvious discard X. For example, “The worker provided a myrtle tree from the newly discovered trees, not the thought-to-be-extinct trees.” was emphatically D for the ‘myrtle’ coding and emphatically CH for the ‘Rio de Janeiro Myrtle’ coding, while “The lab worker provided the students with slivers of the tree's bark.” was an obvious X for both codings.

To check our EP codings, we again compared the D and CH codings with the TVJ answers. 22 out of 29 EP codings in the two conditions (13 undiscarded “Rio de Janeiro Myrtle” codings and 16 undiscarded “myrtle” codings) matched the TVJ answers. 6 of the 7 exceptions were in the Myrtle condition. 4 of these were coded D on the EP but chose the CH answer on the TVJ; and one was coded D but chose the “hard to say” answer. The D coding for each of these 5 was obvious: 4 participants said that the laboratory worker provided “myrtle trees” or “a myrtle tree”, and one said “the newly discovered Myrtle”. Yet their TVJ answers and Explanations show that they either believe that the plants are not myrtles or that they have doubts whether they are. We take these participants to have been speaking loosely in their “free” EP response, just as we took the similar responses of 12 participants in the Rice conditions discussed in section 2.3. The only difficult case was the participant coded CH who chose D in the TVJ. Their EP response was: “The lab worker misrepresented what the samples were and brought in samples where they looked externally similar to the Brazilian Myrtles”. This participant was coded CH because, given that the lab worker was asked “to provide a myrtle tree”, his only possible misrepresentation was to provide a non-myrtle tree. Yet the participant not only chose D in the TVJ but also gave the following Explanation: “they both were myrtles, but they were different species.” It seems that the participant must have supposed that the lab worker misrepresented the plants *as Rio de Janeiro Myrtles,* even though that is not in the vignette. Finally, in the Rio condition, one response was: “The laboratory worker presented students with a fake version of the tree that he was told to bring”. This has to be coded CH and yet the participant chose “hard to say” for the TVJ; they did, however, choose the CH option for the ‘myrtle’ TVJ. The participant’s Explanation suggests that they may have misread the “Rio de Janeiro Myrtle” TVJ as asking whether or not the trees in the vignette were *located* in Rio de Janeiro: “It is hard to say where they are, because it could be anywhere, and they dont [sic] really talk about it. It is not a version of the tree, as scientist say it is not.”

There was no significant difference in choice of TVJ options between participants whose EPs were discarded and participants whose EPs were not discarded.

We again concluded that an independent coder was unnecessary. As can be seen, our EP codings that were not confirmed by the TVJ results are obvious, aside from the “difficult case” above.

*3.4 Discussion*

The EP and TVJ results for ‘Rio de Janeiro Myrtle’ are again highly nondescriptivist (CH 100% and 92.3%, respectively), reinforcing our conclusion (2) from the original test. However, our concern in this follow up was with ‘myrtle’ and whether it really differed from ‘Rio de Janeiro Myrtle’ in being quite descriptivist, as our earlier EP results suggested. Both our EP and TVJ results confirmed this suggestion:

1. ‘Rio de Janeiro Myrtle’ was significantly more CH than ‘myrtle’ in the EP (p = 0.0033), and was more CH than ‘myrtle’ in the TVJ, though not significantly so (p = 0.2299).
2. The CH responses for ‘myrtle’ were not significantly different from those for the partly descriptive ‘rice’ in our original Rice Scientific experiment (EP: p = 0.7604; TVJ: p = 0.6457). (It will be remembered that on the original EP, ‘myrtle’ came out significantly *more* descriptivist than ‘rice’.)
3. ‘Rio de Janeiro Myrtle’ was significantly more CH than D (EP: p = 0.0052; TVJ: p = 0.0016), but this was not true of ‘myrtle’ (EP: p = 1; TVJ: p = 0.1534).

In section 2.4 we wondered whether apparently descriptivist EP responses for ‘myrtle’ in the original test could be explained away as loose talk. In section 3.3, we saw that *some* of those responses in the follow-up test can indeed be thus explained away, but certainly not all. Furthermore, we now have to explain the many descriptivist TVJ responses for ‘myrtle’: 8 participants who chose the CH response for ‘Rio de Janeiro Myrtle’ chose the D response for ‘myrtle’. One such participant gave the following Explanation: “I would call them Myrtles. They are nearly the same, and I don't understand why scientists would classify them as such.” Another wrote “Being that they look exactly the same to the human eye, that leads me to believe that it is a type of Myrtle, just not a Rio de Janeiro Myrtle due to it's [sic] biology being different.” A third participant, who chose CH for ‘Rio’ and “hard to say” for ‘myrtle’, wrote “It is not the extinct myrtle, but I still thikn [sic] it would be a myrtle since the differences are so micropscoic [sic].”. These participants think that the new plants are (or may be) myrtles *because they are similar to myrtles*. Yet they do not think that they are Rio de Janeiro Myrtles, even though they are similar to Rio de Janeiro Myrtles! These participants are borrowing reference from the scientists for ‘Rio de Janeiro Myrtle’, as predicted by the Causal-Historical Theory, while simultaneously *not* borrowing the reference of ‘myrtle’ from the scientists, contrary to that theory. This demands an explanation, but it is hard to provide one.

 Here’s an attempt to explain these descriptivist TVJ results for ‘myrtle’. Participants find it so implausible that the new trees could look exactly like the Rio de Janeiro Myrtle without being closely related that they decide, contrary to what the biologists conclude in the vignette, that the new trees must in fact be close relatives (and therefore myrtles). So, the quoted Explanations indicate disbelief in the biologists rather than descriptivism! In support of this, one other participant who chose the D answer for ‘myrtle’ in the TVJ gave the following Explanation: “I don't think they were Rio myrtles because there were many differences. But, I wonder if they are still part of the myrtle family because they look so similar on the outside”. Another participant who chose the hard-to-say answer gave this Explanation: “they were probably cousins of the original Myrtle trees”. These Explanations suggest that, despite being told that biologists do not consider the trees to be in the Myrtle genus, some participants take the nearly identical superficial features to be evidence that the trees *must* be closely related, and therefore they must be myrtles.

This would be an odd position but, interestingly, it would make these responses compatible with the Causal-Historical Theory after all. According to the usual Causal-Historical Theory, the reference of ‘myrtle’ is fixed, largely by scientists, in deep structural properties that are often unknown to those scientists. It is *not* part of the theory that those experts are infallible in identifying which objects have whatever those deep properties may be. The experts should certainly be better than anyone else at doing this, but they can make mistakes. On the position under consideration, participants think that the biologists have made just such a mistake. These participants have indeed borrowed their reference of ‘myrtle’ from the biologists who fixed it. But they nonetheless doubt the reported opinion of biologists that the new plants are not among the plants that ‘myrtle’ refers to. These participants are skeptical, rather than descriptivist.

In “Testing”, we argued that many twin earth-style experiments may be too fantastical and implausible to provide reliable data (sec. 5.3). In one experiment, Haukioja et al. (2020) imagine a planet where something with the H2O structure of water “for some mysterious reason” is solid and looks like a greenish mineral. As we pointed out: “The problem lies in the mystery” (2021: Postscript). For, that mystery raises the specter of the Faulty Test Hypothesis. Some of the explanations given in our follow-up test suggest that the problem may be even worse than we suggested: something as relatively mundane as convergent evolution can lead participants to make probabilistic judgments that directly contradict the information in the vignette. Participants conclude that the biologists reported in the vignette were *wrong* to claim that the new trees are not closely related to the Rio de Janeiro Myrtle because it is just too implausible that a tree could look so similar *without* being closely related. We fear that our experiments, like many in the past, may have overestimated how much strangeness some participants are willing to stomach in a vignette.

 This attempt to explain away the apparently descriptivist results for ‘myrtle’ may be part of the story, but we very much doubt that it is all of it. Setting it aside, what can we conclude about ‘myrtle’? *Prima facie*, the results for ‘myrtle’ are very similar to the earlier ones for ‘rice’. Those led us to conclude that ‘rice’ is covered by the Hybrid Theory, but there is a good reason for not drawing this conclusion about ‘myrtle’. ‘Myrtle’ is crucially different from ‘rice’. All but one of our participants in the ‘rice’ conditions reported having heard of rice; we can safely assume that our participants are competent with ‘rice’. But in this follow-up test, less than half (14 out of 29) reported having heard of myrtle trees. For these participants, ‘myrtle’ is a novel term, just as ‘catoblepas’ was for the participants in “Testing”; see section 1.3 above. Furthermore, merely having heard of myrtles is a low bar, which does not guarantee competence with ‘myrtle’. So, most of our participants probably lacked the linguistic competence necessary for identifying myrtles. As such, our results for ‘myrtle’ are much like our earlier results for ‘catoblepas’ in “Testing”: they provide evidence against the Causal-Historical Theory, but do not adjudicate between the other alternatives.

**4. Conclusion**

Past experiments led philosophers to conclude that the reference determination of natural kind terms is neither simply descriptive, as traditionally thought, nor simply causal-historical, as Kripke-Putnam claimed. Some philosophers urged an Ambiguity Theory and some, a Hybrid Theory. A Different Idiolects Theory is also a possibility. Our recent experiments led us to hypothesize that philosophers had been mistaken in thinking that one theory fits all: some natural kind terms are covered by one theory, some another, with a place for all the proposed theories. In particular, we thought that although the Causal-Historical Theory is true for some terms it is not true for all. In some cases, reference determination will 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 some terms: a superficial-descriptive one and a deep-causal one. We argued that this would usually be best accommodated by the Hybrid Theory.

Our present paper tested these hypotheses using the biological kind term ‘Rio de Janeiro Myrtle’ as an example of the former “scientific” sort, and ‘rice’ as an example of the latter “practical” sort. We tested whether participants applied the term ‘Rio de Janeiro Myrtle’ to samples that were superficially like Rio de Janeiro Myrtles but with a quite different genetic make up and history; similarly, for ‘rice’. We predicted that the two terms would be covered by different theories. This was confirmed with high significance in both elicited production and truth value judgement tests. We predicted that participants would borrow their reference of ‘Rio de Janeiro Myrtle’ from the biologists and so use the term causal-historically, with no descriptive element to its reference determination. That is what we found. And we predicted that there would be two factors to the reference determination of ‘rice’, a superficial-descriptive factor and a deep-causal factor. Again, that is what we found. We took these results to support the Hybrid Theory of ‘rice’ rather than the Different Idiolects Theory and the Ambiguity Theory, given the implausibility that there are the regularities in distinct usages that the latter theories require. These are our most significant findings, as they show that much of the debate regarding theories of reference has rested on a false assumption: that any one theory of reference is correct for the whole class of natural kind terms.

Tobia et al (2020) embraced the Ambiguity Theory partly because they found that participants were more likely to use the term ‘salmon’ descriptively in a legal context and nondescriptively in a scientific one. We tested for a similar variation in usage between “practical” and “scientific” contexts, but found no significant variation. So, our results were inconsistent with the Ambiguity Theory of ‘rice’ (but not with the Hybrid Theory).

To our surprise, we discovered that many participants in our EP test who took the sample trees not to be Rio de Janeiro Myrtles nonetheless took the trees to be myrtles. Myrtles are almost certainly of no practical interest to the folk, and so had we made any prediction about ‘myrtle’ it would have been that it is simply causal-historical like ‘Rio de Janeiro Myrtle’. Yet our results showed ‘myrtle’ to be even more descriptivist than the “practical” ‘rice’. This led us to a follow-up test in which participants, after responding to the EP prompt, were given a TVJ test that explicitly asked them whether the samples were myrtles. The results implied that the reference determination of ‘myrtle’ is indeed partly descriptivist. Perhaps some of these results can be explained away in the ways we suggest, but surely not all. It seems that some natural kind terms of no practical interest are not covered by the Causal-Historical Theory. More investigation is needed.[[16]](#footnote-16)

REFERENCES

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

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. https://doi.org/ppr201182222

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

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

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. (2021). *Overlooking conventions: the trouble with linguistic pragmatism*. Cham, Switzerland: Springer Nature.

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

Devitt, M. & Porter, B. (2021). Testing the reference of biological kind terms. *Cognitive Science*, 45, e12979. DOI: 10.1111/cogs.1297

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

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

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

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

Haukioja, J., Nyquist, M., & Jussi, J. (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.

Johnson, M. & Deutsch, M. (2021). Fake gnus! (Or: there is no experimental evidence for the lazy person’s approach to philosophy). *Inquiry*, DOI: 10.1080/0020174X.2021.1977995.

Jylkka, J., Railo, H. & Haukioja, J. (2009). Psychological essentialism and semantic externalism: evidence for externalism in lay speakers’ language use. *Philosophical Psychology*, 22, 37-60.

Kornblith, H. (1980). Referring to artifacts. *Philosophical Review*, 89, 109-114.

Kripke, S. (1977). Speaker’s reference and semantic reference. *Midwest Studies in Philosophy, II*, 255-276.

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

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.

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.

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.

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. and Newman, G. and 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 essentialism. *British Journal for the Philosophy of Science*, 57, 425-448.

1. It should be noted that the consensus in the philosophy of biology rejects this sort of intrinsic essentialism. For reactions, see Walsh 2006, Devitt 2008. [↑](#footnote-ref-1)
2. Apart from their 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: Domaneschi et al (2017); Devitt and Porot (2018). 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: Sytsma and Livengood (2011); Sytsma et al. (2015); Machery et al. (2015); Devitt and Porot (2018). [↑](#footnote-ref-2)
3. On this issue of evidence for meanings, see Devitt (2021). [↑](#footnote-ref-3)
4. Michael Johnson and Max Deutsch (2021) have a rather different criticism of the Ambiguity Theory, particularly of Nichols et al.’s case for it. Johnson and Deutsch think that accepting this theory exemplifies Kripke’s “lazy person’s approach to philosophy” (Kripke 1977, p. 268). This approach is, in effect, that of failing to be guided by Grice’s Modified Occam’s Razor, “Senses are not to by multiplied beyond necessity” (Grice 1989, p. 47). We think that this Razor, as commonly construed, is mistaken (Devitt 2021, pp. 143-159). [↑](#footnote-ref-4)
5. Kornblith’s (1980) proposal for “artifactual” kind terms is an example of such a causal-historical theory. For a brief exploration of this and other possible theories, see Devitt and Sterelny 1999, pp. 96-100. [↑](#footnote-ref-5)
6. Note that this is a view of the *folk’s* use of ‘rice’ not the *biologists’* (who we were not testing). Perhaps the biologists have both a practical and scientific use of ‘rice’; see the discussion in “Testing” of the use of ‘water’ by chemists (sec. 5.1). [↑](#footnote-ref-6)
7. We saw very little to be gained from having vignettes with novel specimens differing in histories but not genes, or genes but not histories. [↑](#footnote-ref-7)
8. We planned to recruit 320 MechanicalTurk “masters”, but were only able to recruit 145. We 0made up the balance with 175 non-masters. [↑](#footnote-ref-8)
9. Tobia et al. (2020), for example, found variation in responses between legal and scientific contexts, and take this to be evidence for the Ambiguity Theory. “Testing” argues that this may be taken as evidence in favor of either the Ambiguity *or* Hybrid theory, and as evidence *against* the Description and Causal-Historical Theory. [↑](#footnote-ref-9)
10. We lost 28% of our participants to the content-based checks. The importance of these checks is vividly demonstrated by the following. Across all four conditions, participants who failed the content attention checks had significantly more descriptivist TVJ responses than those who passed. This can be attributed almost entirely to the vast majority of those failures having failed the second content question: they wrongly reported that the new specimens *were* identical to rice/Rio de Janeiro Myrtles in their genetic structure and evolutionary history. Unsurprisingly, over 80% of these participants reported that the new specimens were rice/Rio de Janeiro Myrtles. Without the content-based attention checks, these responses would have erroneously been counted as descriptivist, when in fact the participants simply misunderstood the relevant information. [↑](#footnote-ref-10)
11. MechanicalTurk non-masters were significantly more likely than masters to fail attention checks. But among participants who passed the attention checks, there were no statistically significant differences between masters and non-masters. [↑](#footnote-ref-11)
12. In the case of the EP prompt, the percentages are taken out of all responses *coded CH, D or N.* [↑](#footnote-ref-12)
13. None of these were masters. [↑](#footnote-ref-13)
14. This time we lost 40% of our participants to the content-based checks; c.f. note 10. As before, participants who failed these checks had significantly more descriptivist responses to the TVJ prompts. This is again attributable to the majority of these participants failing the second check and reporting that the new trees *were* identical to myrtles/Rio de Janeiro Myrtles in their genetic structure and evolutionary history, contrary to what the vignette clearly states. [↑](#footnote-ref-14)
15. For example, one participant copied and pasted a part of the vignette for both their EP and Explanation. Another simply wrote “Maxwell” and “good” for their EP and Explanation, respectively. [↑](#footnote-ref-15)
16. For helpful comments and discussion, we’d like to thank Andrea Bianchi, Gregory Currie, Jussi Haukioja, Shaun Nichols, and Tomasz Zyglewicz. Some of the results presented here were presented at the University of Parma in October 2021. We thank the audience there for their helpful discussion and feedback. Portions of this paper were discussed at the Experimental Semantics Gathering in Hudson, NY in June 2022. We thank the attendees for their input. [↑](#footnote-ref-16)