



Cognitive Science 47 (2023) e13335

© 2023 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.13335

## Are Natural Kind Terms Ambiguous?

Jussi Haukioja,<sup>a</sup> Jeske Toorman,<sup>a</sup> Giosuè Baggio,<sup>b</sup> Jussi Jylkkä<sup>c</sup>

<sup>a</sup>*Department of Philosophy and Religious Studies, Norwegian University of Science and Technology*

<sup>b</sup>*Department of Language and Literature, Norwegian University of Science and Technology*

<sup>c</sup>*Department of Psychology, Åbo Akademi University*

Received 26 September 2022; received in revised form 21 July 2023; accepted 18 August 2023

---

### Abstract

Recent experimental studies have claimed to find evidence for the view that natural kind terms such as “water” are ambiguous: that they have two extensions, one determined by superficial properties, the other by underlying essence. In an online experiment, we presented to 600 participants scenarios describing discoveries of novel samples that differ in deep structure from samples of a familiar kind but are superficially identical, such as a water-like substance that is not composed of H<sub>2</sub>O. We used three different types of question sets to probe whether the participants considered the sample as a member of the kind or not. Our results did not confirm the predictions of the ambiguity view. They were, rather, consistent with views that take underlying essences to be the sole criterion for membership in a natural kind.

*Keywords:* Reference; Categorization; Causal theory of reference; Descriptivism; Experimental semantics; Natural kind terms; Ambiguity

---

### 1. Introduction

Putnam (1975) once asked us to imagine a faraway planet, Twin Earth, that is identical to our planet in all respects except one: there is a difference in the chemical composition of the substance called “water.” On our planet, the substance is composed of H<sub>2</sub>O. On Twin Earth, it is composed of a different chemical compound, abbreviated as XYZ. In a similar vein, Kripke (1980) once asked us to imagine an animal that looks and behaves exactly like a tiger. Despite looking and behaving like a tiger, however, this animal is of a different species, and thus has internal properties different from tigers as we know them.

---

Correspondence should be sent to Jeske Toorman, Department of Philosophy and Religious Studies, Norwegian University of Science and Technology, Trondheim 7491, Norway. E-mail: jeske.toorman@ntnu.no

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Most philosophers agree with Putnam and Kripke that entities or substances with internal properties different from the samples of a natural kind that we are familiar with ought not to be categorized as instances of that kind and do not belong in the extension of the relevant natural kind term (i.e., the term denoting that kind). If most of the liquids we call “water” are in fact composed of  $H_2O$ , what inhabitants of Twin Earth call “water” ought not to be categorized as water by us, and our term “water” does not apply to it. If tigers, as we are familiar with them, are in fact members of the species *Panthera tigris*, no animal of a different species is a tiger or is in the extension of “tiger.” This is taken to be so, in both cases, despite similarity, or even (qualitative) identity, in superficial properties.

Insofar as philosophers agree with Kripke and Putnam, their judgments have been taken to undermine what we will call “the superficial properties view.” According to the superficial properties view, natural kind terms refer to all and only whatever shares a specific set of superficial properties with familiar samples of a kind. The relevant superficial properties, in turn, are taken to be those we can easily observe, such as color, taste, smell, and at least some of the kind’s functions. The most familiar superficial properties view is classical descriptivism, which holds that the referent of a natural kind term is whatever has the superficial properties competent speakers semantically associate with that term.

The same judgments have been taken to support what we will call “the underlying essence view.” According to the underlying essence view, natural kind terms refer to all and only entities or substances which share whatever is the underlying essence of a sufficient portion of familiar samples of the kind. The underlying essence, in this case, is standardly taken to be that which causes the superficial properties of the entities with which we are familiar, and which can be discovered by the natural sciences. For instance, the essence might be *being composed of  $H_2O$*  when it comes to “water,” and it might be *having a specific kind of DNA*, or *having a specific lineage*, when it comes to “tiger.” The most familiar underlying essence view is the causal-historical theory of reference, but causal descriptivist views (e.g., Kroon, 1987) would, for example, also fall in this category.

The superficial properties view and the underlying essence view are views about language. It should be noted, however, that their relevance extends to psychological theories of concepts. Psychological theories of concepts make predictions concerning our categorization behavior. Theories of reference, on the other hand, can be taken to make predictions concerning term application. But data about term application and data about categorization often go hand in hand: in studies of categorization behavior, subjects are typically invited to respond verbally, that is, to apply terms. Term application, on the other hand, is itself typically an instance of categorization behavior. As such, evidence in favor of either one of these types of view can count in favor of the other type of view.<sup>1</sup>

Of particular interest here are *prototype theory* and *psychological essentialism*. According to prototype theory, people represent categories as prototypes, which can be described as probabilistic representations of the typical features found in members of the category, bound together by family resemblance (Rosch, 1978). On such a view, an entity is to be categorized as a member of a category if its similarity with the prototype exceeds a specific threshold value associated with the concept. The features represented in the prototype are usually taken to be observable, superficial properties in the same sense as the superficial properties that

are taken to determine the reference of natural kind terms according to the superficial properties view. It is thereby to a large extent analogous to the superficial properties view, and evidence against the superficial properties view may, therefore, in many instances also count as evidence against certain variations of prototype theory.

According to psychological essentialism, people represent natural kinds as having an “underlying reality or true nature, shared by members of a category, that one cannot observe directly but that gives an object its identity and is responsible for other similarities that category members share” (Gelman, 2003, p. 8). The most common variety of psychological essentialism is a placeholder version according to which it is not specific essence beliefs that are of relevance, such as a belief to the effect that the essence of water is H<sub>2</sub>O. The relevant essence belief, rather, is a general one: natural kinds have an essence, and the essence is just whatever gives rise to category members’ similarities (Medin & Ortony, 1989). The view is thereby almost directly analogous to the underlying essence view. Consequently, the evidence against the underlying essence view may also count against psychological essentialism.<sup>2</sup>

Recent work in experimental philosophy of language and cognitive psychology can be taken to indicate that, contrary to most philosophers, nonphilosophers categorize natural kinds by virtue of both superficial properties as well as underlying essence. It thereby threatens the underlying essence view widely accepted by philosophers, as well as the psychological essentialism widely accepted by psychologists, on two counts. First, it seems to indicate that superficial properties do, after all, determine reference and categorization. Second, it seems to indicate that it is not the case that *either* superficial properties *or* underlying essence determine reference and categorization, but rather that *both* superficial properties *and* underlying essence do. We start with a brief overview of the recent literature before turning to the study we most directly take issue with, namely that of Tobia, Newman, and Knobe (2020). Most of these studies are concerned with the question of whether subjects’ responses are in accordance with classical descriptivism or with the causal-historical account of reference determination.<sup>3</sup> As noted, the former account, at least when it concerns the versions tested, is a version of the superficial properties view, while the latter is a version of the underlying essence view. Hence, if the empirical data indicate that the reference of natural kind terms is determined by a conjunction of the facts by virtue of which these accounts take reference to be determined, the result is that facts about superficial properties as well as facts about underlying essence might determine reference and categorization.

### 1.1. Previous experimental work

A substantial body of research on psychological essentialism indicates that sharing underlying essence is people’s sole criterion for category membership, when it comes to natural kind categories. In a famous study by Keil (1989), for instance, children are told about doctors performing a “special operation” on an animal to make it look like another animal. Despite now looking like another animal, children tended to judge that the animal in question is still a member of the category to which it initially belonged, suggesting that animals’ superficial properties do not guide their categorization judgments. Similarly, Gelman and Wellman (1991) told children about an animal which had its insides removed, and thereby all likely candidates for

that animal's underlying essence, whereas from the outside, the animal still looked the same. In this case, children tended to judge that the animal in question is no longer a member of the category to which it initially belonged, suggesting that the underlying essence is relevant to their categorization judgments. However, recent work in both cognitive psychology and experimental semantics suggests that the picture may be more complicated than main-stream essentialist and causal-historical views would have it. On the one hand, results in psychology call into question whether essence is the *sole* criterion for categorization.<sup>4</sup> On the other hand, recent work in the experimental philosophy of language suggests that superficial properties may, after all, have a role to play in determining the reference of natural kind terms.

Genone and Lombrozo (2012), for instance, presented subjects with scenarios in which they varied both whether two individuals had the same beliefs about a natural kind as well as whether the natural kind first given the name of that kind in those individuals' respective community was the same. After asking if the individuals in question were thinking about the same natural kind, they found that subjects did not consistently rely on one of these two factors to judge the sameness of thought. Genone and Lombrozo conclude that "intuitions [...] support a hybrid theory of reference that includes both causal and descriptive factors" (p. 717).

Nichols, Pinillos, and Mallon (2016) described a case from the history of science in which researchers held false beliefs about a kind of mammal named "catoblepas." These beliefs, though false, were nevertheless based on reports of encounters with real animals, namely, wildebeests. Nichols et al. found that they could shift subjects' willingness to affirm that catoblepas exist by priming them with a story that either did or did not imply referential continuity. Moreover, they found that more subjects agreed with the statement "Catoblepas refers to wildebeests" than with the statement "catoblepas exist." On the assumption that subjects know that wildebeests exist, and in order to avoid attributing inconsistent responses to subjects, they propose that "natural kind terms (and plausibly names as well) are ambiguous, such that in some cases the reference is determined descriptively and in other cases non-descriptively" (p. 160).

Devitt and Porter (2021), using a version of the vignette from Nichols et al., asked subjects to perform an elicited production task, as well as to judge the truth-value of two statements presented simultaneously. Regarding the latter task, one of these statements was predicted to be judged as true if the causal-historical account is true, the other if classical descriptivism is true. In both tasks, about half of the subjects responded in a way taken to be in accordance with classical descriptivism and about half of them in a way taken to be in accordance with the causal-historical account. Moreover, when subjects were presented with the same two statements separately, about half of them responded in a way taken to be in accordance with one theory of reference for one statement, but contradicted themselves by responding in accordance with the other theory of reference for the other statement.

Devitt and Porter take these results, together with certain theoretical considerations, to motivate an "eclectic approach" (p. 18). The reference of natural kind terms, according to them, is determined both by appearance and by causal-historical factors. On their view, however, which of these determines reference can vary depending on the natural kind term in question as well as with particular uses of tokens of a natural kind term type.<sup>5</sup>

## 1.2. Tobia et al. (2020)

Thus, recent experimental work can be taken to suggest that natural kind terms are ambiguous between two senses, one based on superficial properties and the other on underlying essence. Continuing this line of research, Tobia et al. (2020) claim to find evidence in favor of the view that nonphilosophers' categorization behavior of natural kinds shows what the authors call a "dual character" pattern of judgment (p. 184). That is to say, according to the authors, nonphilosophers are indicated to categorize natural kinds according to two distinct sets of criteria. One set of criteria is such that natural kinds are categorized in the way they are by virtue of their superficial properties only. The other set is such that natural kinds are categorized in the way they are by virtue of underlying essence only.

Their evidence for this view stems from the following observations. Subjects were presented with scenarios in which they were told about a newly discovered entity or substance that by its superficial properties is identical to members of an antecedently familiar natural kind, but differs from them in its underlying essence. When asked to make a forced choice, 16% of the subjects on average chose a statement to the effect that the newly discovered entity or substance was a member of the familiar kind, about 29% on average chose a statement to the effect that the newly discovered entity or substance was not a member, and about 55% on average chose a statement to the effect that the newly discovered entity or substance is a member in one sense, but not a member in another sense (experiment 1a in Tobia et al.). After being presented with a standard Twin Earth-style scenario, for instance, only 46% of the subjects chose one of the following two statements:

1. The liquid from Twin Earth is water.
2. The liquid from Twin Earth is not water.

Instead, 54% chose the following statement:

- 3 There is a sense in which the liquid from Twin Earth is water, but ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid from Twin Earth is not truly water at all.

Moreover, after being presented with the same vignettes, subjects were (in their experiment 1b) asked for their level of agreement with two statements of the following form:

- 4 There is a sense in which the liquid from Twin Earth is water.
- 5 Ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid from Twin Earth is not truly water at all.

The subjects' mean rates of agreement with (4) and (5) were roughly the same, falling slightly on the side of the agreement. On a scale from 1 (disagree) to 7 (agree), the mean rating for statements like (4) was 4.73 and the mean rating for statements like (5) was 4.39, where 4 is the midpoint of the scale.

Tobia et al. take these results to indicate that philosophers and nonphilosophers do not agree in the way they categorize natural kinds. Insofar as nonphilosophers' categorization behavior is taken to have a bearing on theories of reference, they take these results

additionally to undermine a theory of reference predicting that natural kind terms refer exclusively to substances or entities sharing the relevant underlying essence. Their results, instead, are taken to support a view according to which nonphilosophers categorize natural kinds according to two distinct sets of criteria that operate separately, and a theory of reference predicting that tokens of a natural kind term type sometimes refer exclusively to entities sharing underlying essence, whereas at other times, they refer exclusively to entities sharing superficial properties. According to Tobia et al., then, their results suggest that natural kind terms are ambiguous and that natural kind concepts are dual-character concepts.<sup>6</sup>

Two features distinguish Tobia et al.'s study from that of Genone and Lombrozo, Nichols et al., and Devitt and Porter. First, whereas the latter sets of authors all test what nonphilosophers take other speakers' usage of a linguistic expression to refer to, Tobia et al. are testing subjects' own categorization behavior of the natural kinds at issue.<sup>7</sup> Subjects' own categorization behavior, arguably, constitutes more direct evidence for word use and thereby for referential relations, and certainly more direct evidence for how they would categorize a natural kind themselves.

Second, Tobia et al.'s target terms are terms with which subjects are plausibly familiar, and which have traditionally been used in the philosophy of language to undermine the superficial properties view and support the underlying essence view. Genone and Lombrozo, Nichols et al., and Devitt and Porter's target terms, by contrast, are terms with which subjects can plausibly be taken to be unfamiliar, and which have not figured centrally in earlier debates. Genone and Lombrozo use a fictional disease called "tyleritis" and a fictional mineral called "evensium" as their natural kinds of concern, both of which they came up with themselves. Nichols et al. and Devitt and Porter are concerned with testing the reference of a term of an existing kind, namely, *catoblepas* (an existing kind, in any case, if the causal-historical account is true). "*Catoblepas*" is, however, not a term used before, nor is it likely a term test subjects have heard of previously. Tobia et al., on the other hand, use "water," "gold," and "tiger," the former two used both by Kripke and Putnam, the latter used by Kripke alone, and all of which are terms familiar to almost everyone.

Whether it is of relevance that the used terms are ones with which test subjects have previous familiarity depends on the details of the theory of reference or concepts on which subjects' responses are taken to have a bearing. When it comes to using terms traditionally employed in the philosophy of language, however, the relevance is clearer: it allows for a smoother inference to the effect that nonphilosophers' categorization behavior does or does not correspond to those of philosophers. If nonphilosophers do not categorize natural kinds by virtue of underlying essence only, one might readily conclude not just that philosophers were wrong in endorsing the underlying essence view, but additionally that philosophers were wrong in taking their own category judgments to be of relevance to theorizing about reference.

We find that the experimental setup used by Tobia et al. has shortcomings, both when it comes to the vignettes, and to the statements presented to the subjects. When it comes to the vignettes used, they state that while the novel samples (e.g., XYZ<sup>8</sup>) do not belong to the same scientific category as the familiar substances, species, and so on (e.g., water), this difference is "immaterial for any purpose other than scientific classification." This could conceivably affect

the subjects' subsequent classifications in two opposite directions. First, and as also noted by Devitt and Porter (forthcoming), this can be naturally read as telling the subjects that the novel samples *should* be categorized as belonging to the familiar kind, when *not* performing scientific classification: this could make subjects more likely to classify the novel samples as category members. On the other hand, this phrase might make the subjects categorize the novel samples as not being category members because the difference in underlying essence is taken to matter *only* for scientific purposes.

When it comes to the statements presented to the subjects, we think there are two problems. First, expressions such as “there is a sense in which XYZ is water” are open to various interpretations. In order for agreement with statements like (4) and (5) to count as direct evidence for the ambiguity view, or for the view that subjects are truly classifying XYZ as water, the subjects would need to understand this expression as stating that *there is a literal sense of the word “water” on which XYZ is water*. But we cannot simply assume that this is how they understand it: the statement could also be understood nonliterally, for example, as stating that XYZ can be *treated* as water, at least by some of the subjects.

Second, the way Tobia et al. formulated the member and nonmember statements contains an obvious asymmetry. In experiment 1b, subjects were found to agree to an equal extent with statements (4) and (5) (see above). But in the nonmember statement (5), unlike in the member statement (4), subjects are invited to think about what it *ultimately, really* means to be *truly* water *at all*. The asymmetry should make us cautious in drawing very definite conclusions on the basis of the fact that the two statements received similar rates of agreement. To the extent that we do rely on comparisons between subjects' rates of agreement with member and nonmember statements, it would be more informative to compare ratings for pairs of statements that are otherwise identically formulated.

Because of the two problems, one might even claim that Tobia et al.'s results can be accommodated by the underlying essence view and views like psychological essentialism. Proponents of these theories could, namely, dig in their heels and insist that the subjects' agreement with (5) should count as evidence for *their* theory—as showing that the subjects think that, *literally speaking*, the liquid from Twin Earth is *not* water—and go on to try to explain away the subjects' agreement with (4) as “loose talk,” based on the unclarity of expressions like “there is a sense in which x is K.”<sup>9</sup> We are not recommending that the underlying essence theorists should simply rest content with this response, and ignore Tobia et al.'s results—far from it—but we do think that its availability should make us view Tobia et al.'s results as inconclusive, and that more data are needed.

### 1.3. The present study

In the present study, we used a similar setup to that used by Tobia et al., but aimed at avoiding the shortcomings described above. First, we rewrote the vignettes with the intention of not giving test subjects any information that could be interpreted as an indication as to how the new samples ought to be classified. The vignettes were also simplified, and technical vocabulary which subjects might not be familiar with (such as “elemental atom” in their “gold” vignette<sup>10</sup>) was removed.

Second, we presented subjects with three different question sets: one identical to Tobia et al.'s (i.e., (4) and (5) above), one where the asymmetry in the first one was reversed, and one with statements including neither leading vocabulary such as “ultimately” and “really,” nor the expression “there is a sense in which x is K.” In the last question set, the subjects were simply asked for their level of agreement with statements to the effect that the novel sample is, or is not, a category member. (These correspond to “Tobia et al. questions,” “Inverted Tobia et al. questions,” and “Neutral questions” below.) This experimental setup entails clearly different predictions for the underlying essence and for the ambiguity view, as explained below.

We focused on Tobia et al.'s experiment 1b, rather than 1a, because it allowed us to reverse the asymmetry in (4) and (5) in a natural way. Moreover, we focused on what Tobia et al. call the “neutral” context, and ignored the legal and scientific contexts which Tobia et al. study in their experiments 2 and 3, because we were primarily interested in the question of whether natural kinds are ambiguous in the first place: if they are, this should be detectable in that context.

## 2. Method

Preregistration for the current study can be found at <https://osf.io/shx69/>. The experiment was conducted as an online survey. We aimed to recruit roughly 200 participants in each of three groups, on the crowdsourcing site Prolific. We targeted UK residents who were native English speakers. Based on the policy of the NTNU, no ethical approval was needed since the study did not involve the collection of sensitive information and was fully anonymous.

### 2.1. The probes

We employed four different vignettes featuring natural kind terms. Three were adapted from Tobia et al. (“gold,” “tiger,” and “water”), and one was adapted from Devitt and Porter (forthcoming) (“rice”). The first three retained the basic structure from Tobia et al., but were simplified, and partly rewritten to remove technical vocabulary, and the reference to the scientific classification mentioned above. The vignette featuring “rice” was adapted from Devitt and Porter (forthcoming), with minor modifications.<sup>11</sup> All of these had a similar structure, familiar from Twin Earth cases: novel samples were discovered or developed, sharing all superficial properties with members of a familiar kind, but not sharing their underlying essence. For example, we used the following vignette for “water”:

Suppose that in a few years, humans are able to travel to other galaxies. While exploring, they land on a planet that looks exactly like Earth in virtually all respects. It is populated by plants and animals that look exactly like the familiar plants and animals on Earth. Its landscapes and ecosystems look and function exactly like those on Earth. They dub this planet “Twin Earth”. The astronauts remove their helmets and find that they can breathe freely. They drink from the lakes and rivers and find that their contents look and taste just like water. They quench their thirst on what they collect from the lakes



and rivers while they explore the planet. When they perform a chemical analysis of this liquid, they find out that it does not consist of  $H_2O$ , but another chemical compound, not found on Earth. This is surprising, because scientists long ago discovered that all the samples of water on Earth are composed of the compound  $H_2O$ . The liquid in Twin Earth's lakes and rivers behaves exactly like  $H_2O$  in everyday circumstances, but can easily be distinguished from  $H_2O$  in the laboratory.

In addition, all subjects received three vignettes which functioned as foils; two of these featured novel samples that were only remotely similar in appearance to members of familiar kinds ("silver," "salt"), while the third featured a functional kind term ("bottle opener"). The order of the vignettes was counterbalanced. All vignettes, including the foils, are reproduced in the Appendix.

## 2.2. *The question sets*

The participants were randomly assigned to three groups, each receiving one of three different question sets. Thus, each question set was given to a separate group of participants, allowing us to examine how the question formulation affects the response. The between-subjects design was used to avoid repeating similar questions for the same participants and to avoid priming effects (if a subject first answers one type of question, this is likely to affect their response in the following questions). For "water," the questions were the following:

**QS1: "Tobia et al." question set:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the liquid from Twin Earth is water.
- (1b) Ultimately, if you think about what it really means to be water, you'd have to say there's a sense in which the liquid from Twin Earth is not truly water at all.

**QS2: "Inverted Tobia et al." question set:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the liquid from Twin Earth is not water.
- (2b) Ultimately, if you think about what it really means to be water, you'd have to say there's a sense in which the liquid from Twin Earth is truly water.

**QS3: "Neutral" question set:** How strongly do you agree with the following claims:

- (3a) The liquid from Twin Earth is water.
- (3b) The liquid from Twin Earth is not water.

Subjects reported their level of agreement on a 7-point Likert scale with 1 labeled as "disagree" and 7 as "agree." The Tobia et al. question set is identical to the one used in Tobia et al.'s experiment 1b. Our study thereby enables a conceptual replication of the Tobia et al. study through comparing the responses within the "Tobia et al. question set." The Inverted Tobia et al. question set is a negation-inverted version of the Tobia et al. question set, and was used to examine whether Tobia et al.'s original formulation of the question biased the responses. Finally, the Neutral question set is symmetrical with respect to the two

alternatives: here, the subjects are invited to make a straightforward classification judgment, without any leading vocabulary in either direction.

### 2.3. Hypotheses

The underlying essence view and the ambiguity view entail clearly different predictions about how subjects would respond to the different question sets.<sup>12</sup> The ambiguity view predicts that subjects would agree equally strongly with both (1a) and (1b), as subjects in Tobia et al.'s study did. The ambiguity view predicts that subjects would agree to the same extent with both (2a) and (2b), as well: if “water” is ambiguous, and there is a sense in which XYZ is water, alongside a sense in which it is not, and both are senses in which something *really* is or is not water, responses to (2a) and (2b) should be similar to the responses to (1a) and (1b). When it comes to the Neutral question set, our initial assumption was that, if the ambiguity view is correct, subjects would agree with (3a) and (3b) to a roughly equal extent, since XYZ is water on one but not the other of the two senses of “water” (but see our discussion in Section 4).

The underlying essence view, on the other hand, predicts that the subjects' agreement with nonmember statements should be significantly higher than with otherwise identically formulated member statements. That is, their level of agreement with (2a) should be significantly higher than (1a), and their level of agreement with (1b) should be significantly higher than (2b). Regarding the Neutral question set, the underlying essence view predicts that subjects are significantly more likely to agree with the nonmember statement (3b), than they are with the member statement (3a).

Our primary focus is on the comparisons between ratings, rather than their absolute values, for reasons we have already touched on above (1.2). Especially when it comes to the Tobia et al. and Inverted Tobia et al. question sets, absolute ratings are of limited evidential value, because the expression “there is a sense in which x is (not) K” can be interpreted in multiple different ways by the subjects, and only some of the ways are such that the responses would provide evidence of the literal meaning(s) subjects attach to natural kind terms. Comparisons between ratings can, nonetheless, provide evidence for and against the views under consideration. Even though the expression “there is a sense in which x is (not) K” leaves room for different interpretations, if subjects consistently agree more readily with statements that do *not* count the new samples as being in the extension of a natural kind term than with ones that do, that would at least count against an ambiguity view that gives equal weight to both senses.

### 2.4. Analytical approach

We used Bayes factors (BF) for each analysis, because they can yield evidence not only for the alternative hypothesis (e.g., that the formulation of the questions has an effect) but also for the null hypothesis (e.g., that the formulation of the questions has no effect).  $BF_{10}$  indicates the likelihood of the observed data if the alternative hypothesis holds, in proportion to its likelihood if the null hypothesis is true.  $BF_{10}$  is mathematically defined as follows:

$$BF_{10} = P(D|H_1) / P(D|H_0),$$

where  $P$  is likelihood,  $D$  is data, and  $H_1$  and  $H_0$  are the alternative and null hypotheses, respectively. For example, if  $BF_{10} = 3$ , the data is three times more likely if the alternative hypothesis is true. Conversely, the inverted BF, that is,  $BF_{01}$ , indicates the likelihood of the data if the null hypothesis is true compared to if the alternative hypothesis is true. The  $BF_{10}$  is interpreted as follows:  $> 100$  Extreme evidence for  $H_1$ ; 30–100 Very strong evidence for  $H_1$ ; 10–30 Strong evidence for  $H_1$ ; 3–10 Moderate evidence for  $H_1$ ; 1–3 Anecdotal evidence for  $H_1$ ; 1 No evidence for  $H_1$ . The inverted BF is interpreted in the same way, but the evidence is for  $H_0$  (Jeffreys, 1961).

### 3. Results

The total sample consisted of  $N = 605$  participants (392 female, 200 male, 10 other, and 3 not disclosing this information). Their average age was 35.13 ( $SD = 12.58$ ), and their level of education was as follows: 39% had a bachelor's degree, 26% had a higher secondary education, 13% had a master's degree, 11% had an upper vocational education, 8% had a basic vocational education, and 2% had a PhD; the remaining categories were  $< 1\%$ . The number of participants in the three groups receiving the three question sets were 190, 210, and 205, respectively; the differences are due to random fluctuations as the participants were randomly assigned to the groups. The three groups were equal with respect to education (Bayesian cross-tabulation  $BF_{01} = 95$ ), gender (Bayesian cross-tabulation  $BF_{01} = 43$ ), and age (Bayesian ANOVA  $BF_{01} = 39$ ).

Each probe was examined separately. Following the preregistration, paired and independent samples Bayesian  $t$ -tests were used. The results of the within-group analyses (question types a vs. b in the question sets “Tobia et al.,” “Inverted Tobia et al.,” and “Neutral”) are summarized in Table 1 and Fig. 1. There was decisive evidence ( $BF_{10} > 100$ ) that the participants agreed more with the nonmember statements than with the member statements, with the exception of the probe “rice” in the Tobia et al. question set. Next, we examined the between-group differences in the inverted versus noninverted questions (1a vs. 2a and 1b vs. 2b), summarized in Table 2 and Fig. 2. There was decisive evidence ( $BF_{10} > 100$ ) that the participants agreed more with the nonmember statements than with the member statements in all probes, with the exception of “rice.” All the foils were answered as expected, that is, substances differing from silver and salt both in underlying essence and in superficial properties were categorized as not silver and not salt, while devices that can be used to open bottles, but which radically differed from familiar bottle openers in their construction, were categorized as bottle openers (see Tables 1 and 2). All experimental data can be accessed at the following webpage: [https://osf.io/kvtgpp/?view\\_only=e17808acf6d54a62b3b4133de30e6c9d](https://osf.io/kvtgpp/?view_only=e17808acf6d54a62b3b4133de30e6c9d).

#### 3.1. Post hoc tests on absolute ratings

Our analytical approach could be objected to on the grounds that it solely focuses on differences between the ratings, and ignores the absolute values. There could, however, be substantial agreement with both the member (“x is K”) and nonmember (“x is not K”) statements

Table 1

Differences between the ratings of the different types of statements (a vs. b) within the three different groups (“Tobia et al.,” “Inverted Tobia et al.,” and “Neutral”)

Question set “Tobia et al.”					
	1a. “There is a sense in which x is K”		1b. “Ultimately, there is a sense in which x is not K”		BF10
	M	SD	M	SD	
Gold	3.91	1.82	5.35	1.57	> 100
Tiger	3.64	1.84	5.15	1.76	> 100
Water	4.20	1.87	5.18	1.85	> 100
Rice	4.72	1.73	4.55	1.81	1/9.63
Salt*	2.01	1.63	6.17	1.46	> 100
Silver*	2.74	1.74	5.85	1.54	> 100
Opener*	6.29	1.23	2.36	1.65	> 100
Question set “Inverted Tobia et al.”					
	2a. “There is a sense in which x is not K”		2b. “Ultimately, there is a sense in which x is K”		BF10
	M	SD	M	SD	
Gold	5.46	1.42	3.74	1.80	> 100
Tiger	5.56	1.57	3.52	1.89	> 100
Water	5.59	1.57	3.63	1.86	> 100
Rice	5.09	1.61	4.24	1.81	> 100
Salt*	6.30	1.34	2.20	1.74	> 100
Silver*	6.15	1.23	2.60	1.63	> 100
Opener*	2.31	1.80	5.97	1.56	> 100
Question set “Neutral”					
	3a. “x is K”		3b. “x is not K”		BF10
	M	SD	M	SD	
Gold	2.47	1.83	5.64	1.77	> 100
Tiger	2.29	1.66	5.82	1.61	> 100
Water	2.24	1.80	5.84	1.75	> 100
Rice	3.45	2.22	4.65	2.21	> 100
Salt*	1.40	1.11	6.57	1.19	> 100
Silver*	1.73	1.41	6.30	1.35	> 100
Opener*	6.56	1.07	1.52	1.20	> 100

*Note.* The items marked with an asterisk were foils that were used to break the pattern. Paired samples Bayesian *t*-tests were used.

even when one statement is rated higher than the other.<sup>13</sup> To examine this possibility, we used one-sample Bayesian *t*-tests to examine whether the ratings to the member and nonmember statements (of the type “x is K” or “x is not K,” respectively) significantly differ from the midpoint of the scale (i.e., the value 4 on the 7-point Likert scale). Statistically significant

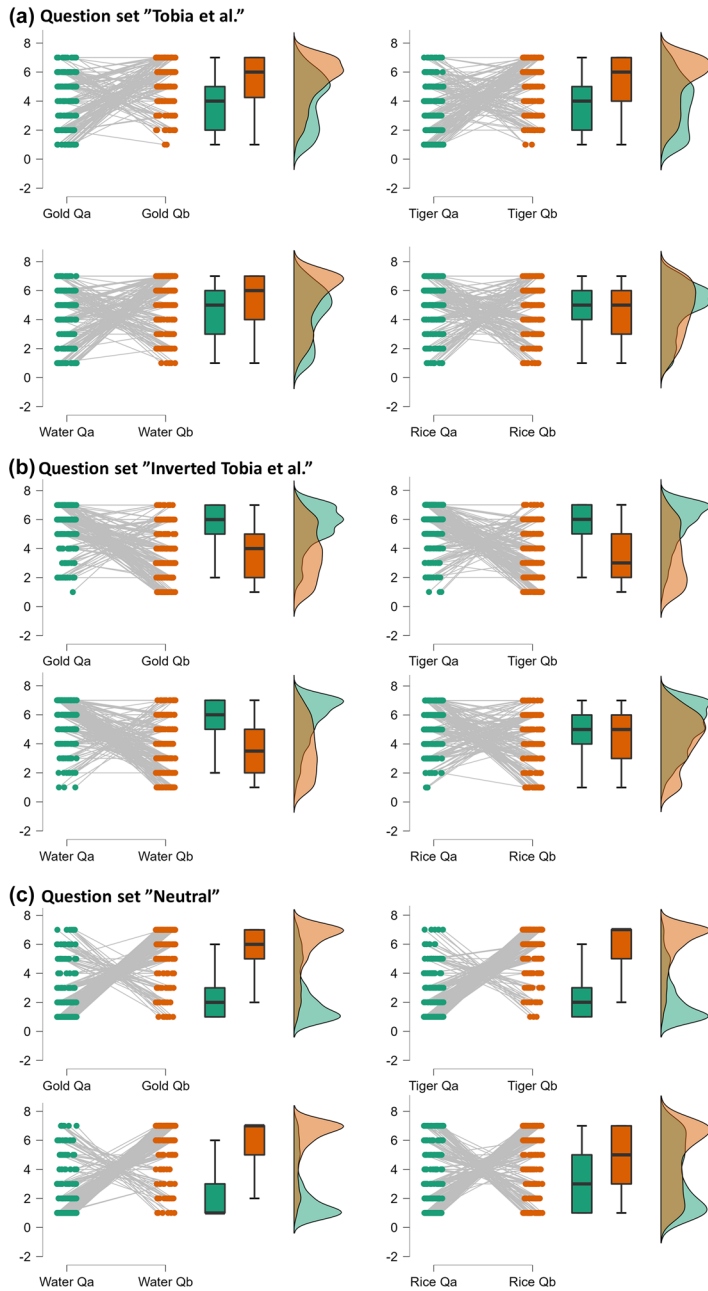


Fig. 1. Within-group differences in each Question set. In "Tobia et al." (Panel A), Qa is "There's a sense in which x is K" and Qb is "Ultimately, there is a sense in which x is not K." In "Inverted Tobia et al." (Panel B), Qa is "There's a sense in which x is not K" and Qb is "Ultimately, there is a sense in which x is K." In Question set "Neutral" (Panel C), Qa is "x is K" and Qb is "x is not K."

Table 2

Differences between the inverted versus noninverted questions (a vs. b) between the groups “Tobia et al.” versus “Inverted Tobia et al.”

Question set “Tobia et al.” versus “Inverted Tobia et al.,” question a					
	1a. “There is a sense in which x is K”		2a. “There is a sense in which x is not K”		BF10
	M	SD	M	SD	
Gold	3.91	1.82	5.46	1.42	> 100
Tiger	3.64	1.84	5.56	1.57	> 100
Water	4.20	1.87	5.59	1.57	> 100
Rice	4.72	1.73	5.09	1.61	1.17
Salt*	2.01	1.63	6.30	1.34	> 100
Silver*	2.74	1.74	6.15	1.23	> 100
Opener*	6.29	1.23	2.31	1.80	> 100

Question set “Tobia et al.” versus “Inverted Tobia et al.,” question b					
	1b. “Ultimately, there is a sense in which x is not K”		2b. “Ultimately, there is a sense in which x is K”		BF10
	M	SD	M	SD	
Gold	5.35	1.57	3.74	1.80	> 100
Tiger	5.15	1.76	3.52	1.89	> 100
Water	5.18	1.85	3.63	1.86	> 100
Rice	4.55	1.81	4.24	1.81	0.45
Salt*	6.17	1.46	2.20	1.74	> 100
Silver*	5.85	1.54	2.60	1.63	> 100
Opener*	2.36	1.65	5.97	1.56	> 100

*Note.* The items marked with an asterisk were foils that were used to break the pattern. Independent samples Bayesian *t*-tests were used.

positive or negative deviation from the midpoint would provide evidence for agreement or disagreement with the statement, in absolute terms. The analyses are summarized in Table 3. Overall, there was evidence of absolute agreement with the member statement only in the case of “rice” in the Tobia et al. question set. In contrast, in the Inverted Tobia et al. question set, there was evidence of disagreement with the member statements in the case of “tiger” and “water” ( $BF_{10}$ ’s > 4), and in the Neutral question set, there was consistent evidence of disagreement with all the member statements ( $BF_{10}$ ’s > 30). In stark contrast to the member statements, there was consistent and decisive evidence of agreement with the nonmember statements across all the question sets ( $BF_{10}$ ’s > 100).

### 3.2. Alternative analysis

It can be argued that, instead of the preregistered analysis that relied on *t*-tests, a multilevel model is a more appropriate method to analyze the data, given that the latter type of model enables analysis across all the probes, groups, and question types. Thus, we conducted a

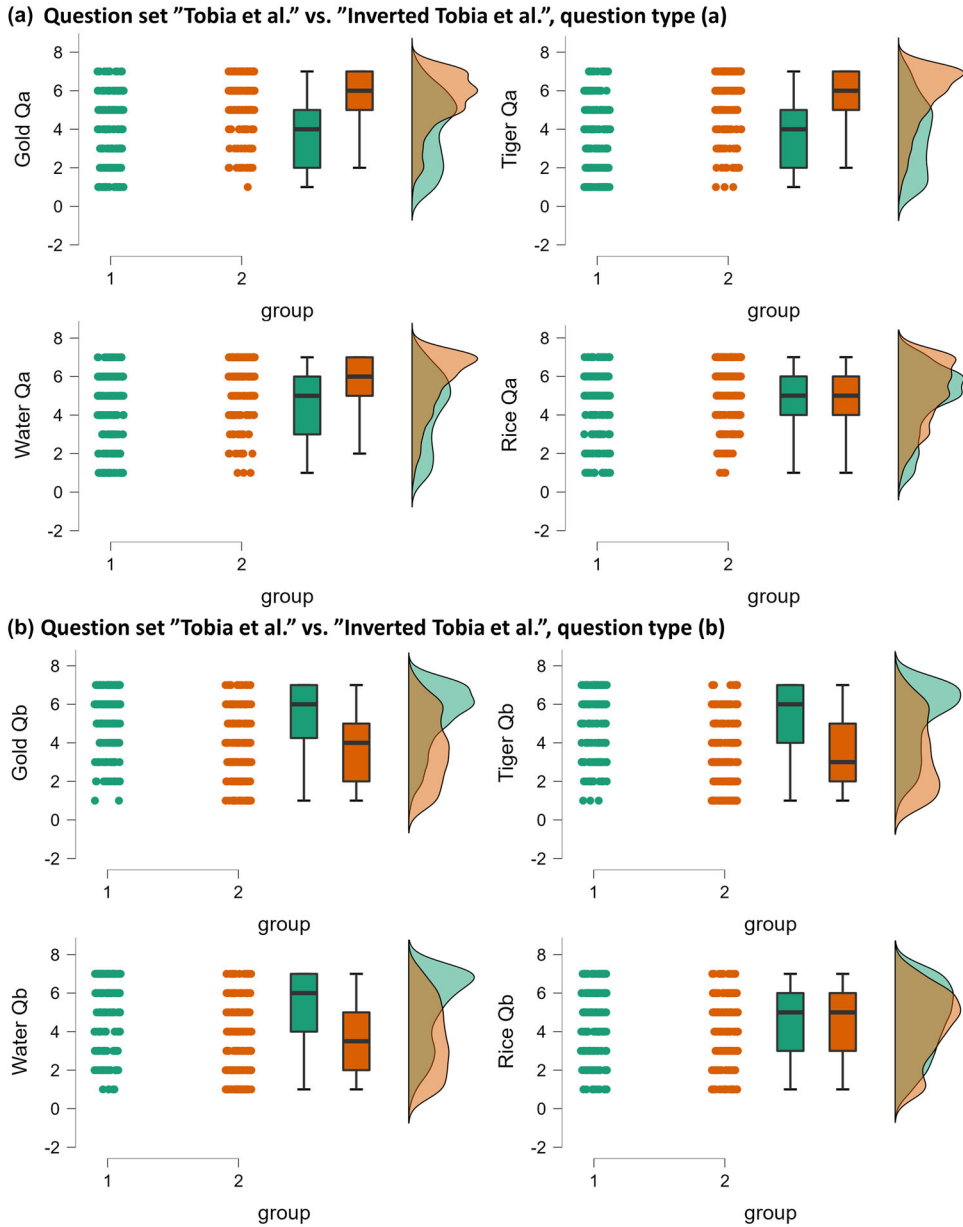


Fig. 2. Between-group (“Tobia et al.” vs. “Inverted Tobia et al.”) differences in the inverted versus noninverted questions. In Panel A, the question in Group 1 is “There’s a sense in which x is K” and in Group 2, “There’s a sense in which x is not K.” In Panel B, the question in Group 1 “Ultimately, there is a sense in which x is not K” and in Group 2, “Ultimately, there is a sense in which x is K.”

Table 3

One-sample *t*-tests on whether the ratings to the nonmember statements in each differ from the scale midpoint (i.e., value 4 on the 7-point scale)

Member statements				Nonmember statements		
QS1a				QS1b		
	M	SD	BF <sub>10</sub>	M	SD	BF <sub>10</sub>
Gold	3.91	1.82	0.10	5.35	1.57	> 100
Tiger	3.64	1.84	2.87	5.15	1.76	> 100
Water	4.20	1.87	0.22	5.18	1.85	> 100
Rice Qa	4.72	1.73	> 100	4.55	1.81	> 100
QS2b				QS2a		
Gold	3.74	1.80	0.62	5.46	1.42	> 100
Tiger	3.52	1.89	46.99	5.56	1.57	> 100
Water	3.63	1.86	4.53	5.59	1.57	> 100
Rice	4.24	1.81	0.50	5.09	1.61	> 100
QS3a				QS3b		
Gold	2.47	1.83	> 100	5.64	1.77	> 100
Tiger	2.29	1.66	> 100	5.82	1.61	> 100
Water	2.24	1.80	> 100	5.84	1.75	> 100
Rice	3.45	2.22	30.63	4.65	2.21	> 100

multilevel analysis using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). This post hoc analysis was done using frequentist statistics instead of BF for simplicity's sake. The model included Group (Tobia et al., Inverted Tobia et al., and Neutral) and the statement type (member and nonmember) as fixed factors which were allowed to interact. We tried including both Participant and Probe as random effects, but only the latter was included due to Participant resulting in singular fit. Foils were excluded from the analysis. The results are presented in Table 4 and Fig. 3.

The first two nonbolded rows in Table 4 indicate that agreement with the nonmember statements was higher in the Inverted Tobia et al. group and in the Neutral group than in the Tobia et al. group (i.e., baseline). The third row shows that the member statements received lower ratings than the nonmember statements in the Tobia et al. group. The interactions indicate that the question formulation (i.e., Group) was differentially associated with the ratings to the member and nonmember statements, which was expected given that the two statements can be a priori considered as symmetrical or mutually exclusive. Under the baseline Inverted Tobia et al., the first row indicates that the rating to the nonmember statement did not differ between the Inverted Tobia et al. and Neutral groups. However, in the Inverted Tobia et al. group, the member statements were rated lower than the nonmember statements. The interaction indicates that the member and nonmember statements were differentially related to the groups Inverted Tobia et al. and Neutral. Finally, under the baseline Neutral, we see that in this group the member statements received lower ratings than the nonmember statements.



Table 4  
Estimates from the multilevel analysis

Baseline: Tobia et al., Nonmember	<i>E</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Inverted Tobia et al.	0.37	0.09	4.02	6.02e-05***
Neutral	0.43	0.09	4.68	2.97e-06***
Member	-0.95	0.09	-10.14	< 2e-16***
Inverted Tobia et al. × Member	-0.69	0.13	-5.39	7.36e-08***
Neutral × Member	-1.93	0.13	-14.88	< 2e-16***
Baseline: Inverted Tobia et al., Nonmember				
Neutral	0.06	0.09	0.70	0.48
Member	-1.64	0.09	-18.48	< 2e-16***
Neutral × Member	-1.23	0.13	-9.77	< 2e-16***
Baseline: Neutral, Nonmember				
Member	-2.87	0.09	-31.99	< 2e-16***

Note. In all analyses, degrees of freedom *df* = 4831.

Abbreviations: *E*, estimate; *SE*, standard error.

\*\*\**p* < 0.001.

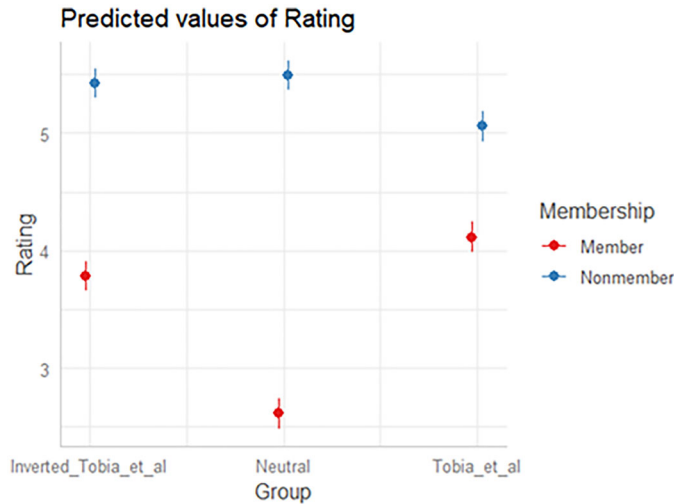


Fig. 3. Estimated means across conditions (Group and Membership) from the multilevel analysis.

#### 4. Discussion

Our aim in the present study was to replicate the study by Tobia et al. which found evidence that natural kind terms are ambiguous between two senses, a superficial properties sense and an underlying essence sense. Moreover, we examined how the formulation of the questions affects the results. In contrast to the findings of Tobia et al., the results were mainly

consistent with the underlying essence view and views like psychological essentialism. A possible exception is “rice,” which we will discuss below.

The first thing to notice is that the responses to the Tobia et al. questions do not mirror those of Tobia et al.’s experiment 1b: subjects agreed more with the nonmember statement (1b) than with the member statement (1a), except for the case of “rice.” Apparently, the rather minor changes we made to the vignettes were enough to make the subjects respond in a way that is more in line with the underlying essence view. We believe that our formulations of the vignettes are less biased than the ones used by Tobia et al., for the reasons given in the Introduction. If we are right, the results for this question set undermine the ambiguity view, as agreement with both the member statement and the nonmember statement was a central reason for Tobia et al. to adopt the view in the first place.

Our other findings point in the same direction, but more strongly. The interplay between the Tobia et al. questions and the Inverted Tobia et al. questions was exactly as predicted by the underlying essence view: subjects were significantly more in agreement with the nonmember statements than with the otherwise identically formulated member statements. For example, they were more inclined to accept “There is a sense in which the liquid from Twin Earth is not water” (2a) than “There is a sense in which the liquid from Twin Earth is water” (1a), and less inclined to accept “Ultimately, if you think about what it really means to be water, you’d have to say there’s a sense in which the liquid from Twin Earth is truly water” (2b) than “Ultimately, if you think about what it really means to be water, you’d have to say there’s a sense in which the liquid from Twin Earth is not truly water at all” (1b). The results for “gold” and “tiger” were very similar.<sup>14</sup>

The answers to the neutral question set also give quite clear support to the underlying essence view. These neutral statements did not invite the participants to think about what it *really*, *ultimately*, means to be an instance of the kinds mentioned, but instead simply asked for their agreement with an unqualified statement. The participants disagreed quite clearly with “x is K” type claims (e.g., that XYZ is water), and agreed with “x is not K” type claims (e.g., that XYZ is not water), as predicted by the underlying essence view.

The categorization behavior here tested is relevant not only to views about the reference of natural kind terms, but also to psychological theories of concepts. The impact of the results, therefore, carries over to the latter views. The fact that subjects tended to agree more with the member statement than the nonmember statement across all comparisons is more in line with views like psychological essentialism than with views like prototype theory (assuming the prototypical features are superficial properties). Further, the fact that we failed to replicate Tobia et al.’s results undermines a central reason Tobia et al. had for thinking that natural kind concepts are dual character concepts.

The fact that we failed to replicate Tobia et al.’s results with the first question set was a surprise to us. We were expecting the two statements in this question set to receive similar levels of agreement. We were also expecting the “ultimately” formulations to have a strong effect such that the member statement with this formulation in the Inverted Tobia et al. question set (2b) would receive significantly lower levels of agreement than the member statement without it in the Tobia et al. question set, and the nonmember statement would receive significantly higher levels of agreement than the member statement, when both contain this formulation.

In fact, the “ultimately” formulations turned out to have only a minor effect, across all terms. The comparisons we were focusing on came out as predicted by the underlying essence view (as described in Section 2.3), but this was not due to the effect of the “ultimately” formulations, but rather the fact that the member statements received lower levels of agreement than nonmember statements, not just in the failed replication of Tobia et al., but consistently across the different formulations.

We conclude that the results are quite strong: all four comparisons point in the same direction. However, it is worth noting that the subjects’ responses to the member statements (1a: “There is a sense in which  $x$  is  $K$ ”) and (2b: “Ultimately, there is a sense in which  $x$  is  $K$ ”) were, on average, quite close to 4, indicating a mixture of agreement, disagreement, and answers in the middle of the scale (as shown in Fig. 1). This could also be seen as supporting a hybrid view such as that proposed by Devitt and Porter (forthcoming), on which there is no ambiguity, but reference is determined by both superficial properties and underlying structure at the same time, and is indeterminate when the two factors pull in opposite directions, such that there is no fact of the matter as to whether, for example, XYZ is water. Our main target here was the ambiguity view, not the hybrid view, but it is worth noting that the responses to the nonmember statements (1b) and (2a), and especially to the neutral question set, do not indicate indeterminacy: the participants categorized novel samples as *not* belonging to the familiar kind (as shown by our post hoc analyses). Thus, the results are also evidence against the hybrid view (though not for “rice”). Moreover, as noted above, the underlying essence view can accommodate some degree of agreement with member statements like (1a): it is not unnatural to suppose, for example, that some subjects understand “there is a sense in which XYZ is water” as stating that XYZ *can be treated as* water (although it ultimately is not).

As noted earlier, our initial assumption was that the ambiguity view predicts roughly similar amounts of agreement with the member and nonmember statements of the neutral question set (3a and 3b). However, the ambiguity view might try to accommodate our results for these, by claiming that our vignettes made the underlying essence senses salient. If so, one might indeed expect subjects to respond in the way they did. However, our vignettes were formulated in a way that resembles the way such cases are typically presented in the philosophy of language, and it has been at least tacitly assumed that such formulations simply describe the facts that are relevant for performing the categorization, without biasing participants’ responses in one direction or another. The ambiguity theorist might protest on the grounds that the ambiguity view has traditionally not been thought of as a contender, and the possibility of the formulations making one sense more salient has, therefore, never been explicitly addressed.<sup>15</sup> At the very least, especially given the clear results we got for the neutral question set, the burden of proof is on the ambiguity theorist to show that vignettes like ours in fact do make the underlying essence sense considerably more salient than the superficial properties sense (assuming that such separate senses exist for natural kind terms, in the first place). Our subjects’ lack of agreement with the member statements in both the Tobia et al. question set and the Inverted Tobia et al. question set (1a and 2b) and the interplay between the Tobia et al. question set and the Inverted Tobia et al. question set, present in any case independent problems for the ambiguity view.

Of the four terms tested, “rice” stands out with respect to the Tobia et al. and the Inverted Tobia et al. question sets.<sup>16</sup> For the Tobia et al. questions, the subjects agreed with both the member statement and the nonmember statement, very much as subjects in Tobia et al.’s study did for “water,” “gold,” and “tiger.” There was no clear interplay between the Tobia et al. and Inverted Tobia et al. questions for “rice,” as there was for the other terms: subjects responded in the way predicted by the ambiguity view. Indeed, Fig. 2 shows that not just the mean ratings, but also the distributions of the answers were extremely similar in both comparisons for “rice” (1a vs. 2a, and 1b vs. 2b). Our results are thus to this extent compatible with an ambiguity view for “rice” (assuming the results for the neutral question set can be explained away by assuming the context makes the underlying essence sense more salient). There is, moreover, something to be said for taking the term to be ambiguous. “Rice,” as used in the vignette (as well as much of ordinary use) does not denote a species, but rather the *seeds* produced by members of the species.<sup>17</sup> As Devitt and Porter (forthcoming) note, the seeds (and their superficial properties) have great practical significance for us, and it would not be surprising if speakers thought of rice not only as the product of a specific kind of plant, but also as a (functional) kind defined by its use in cooking (thus supporting either an ambiguity view, or the kind of hybrid view proposed by Devitt and Porter).<sup>18</sup> But for “water,” “gold,” and “tiger”—the standard examples of natural kind terms during the past few decades—our results give no support for the ambiguity view, as compared to the underlying essence view.

#### 4.1. Discussion of the post hoc analyses

Our preregistered analytical approach relied on comparisons of the ratings between the member statements and nonmember statements. However, it can be objected that this ignores agreement and disagreement with the statements, in absolute terms. We agree that the absolute values cannot be completely ignored. Even if the comparisons between the rates of agreement were to conform with the predictions of the underlying essence view, if the subjects were at the same time found to clearly agree with *both* the member statements and the nonmember statements (and the difference was merely due to them agreeing even more strongly with the nonmember statements), the results would be more difficult for the underlying essence view to accommodate, and more consistent with an ambiguity view (which could claim that the underlying essence sense is more salient, in the relevant context). Thus, post hoc analyses were performed to examine whether the ratings to the member and nonmember statements differ from the scale midpoint, with the assumption that ratings that are significantly above or below the midpoint reflect agreement or disagreement with the statement, respectively. There was evidence of agreement with a member statement only in the case of “rice,” while there was consistent and decisive evidence of agreement with the nonmember statements across all the question sets. This presents strong evidence against the ambiguity view and for the underlying essence view, thus corroborating the original analyses.

#### 4.2. Discussion of the multilevel analysis

The multilevel analysis corroborated the preregistered analysis, indicating that the nonmember statements received higher ratings across all the groups (i.e., question types) than the member statements. Moreover, this analysis sheds further light on how the question

formulations affected the ratings. The difference between the member and nonmember ratings was lowest in the Tobia et al. group, followed by the Inverted Tobia et al. group, and largest in the Neutral group. This indicates that the way Tobia et al. formulated their question may have an artificially exaggerated agreement with the member statements. However, even in the Tobia et al. group, the nonmember statements received significantly higher ratings, suggesting that underlying essence is the sole criterion for category membership.




## 5. Conclusion

The present results suggest that natural kind terms are *not* ambiguous, at least when it comes to terms that have figured centrally in the debates about the semantics of natural kind terms during the last decades, such as “water,” “gold,” and “tiger.” Moreover, our results indirectly support the view that underlying essences are treated as critical for membership in natural kind categories, at least when it comes to these categories. Somewhat surprisingly, we found that the results of Tobia et al. (2020), which were taken to support an ambiguity view, were not replicated when the vignettes had been rewritten with the aim to make them clearer and remove the possibly leading formulations. Moreover, by presenting subjects with negation-inverted versions of the vignettes, we found that the subjects were clearly more inclined to disagree with the member statements and agree with the nonmember statements, when such statements were formulated in otherwise identical ways. These results are consistent with the predictions of the underlying essence view, and provide a challenge to the ambiguity view. At the very least, we claim that the burden of proof now lies on the proponents of the ambiguity view, to provide further empirical support for their view. If such evidence can not be provided, we have reason to think that the majority of philosophers of language have then not been mistaken in accepting an underlying essence view concerning terms such as “water,” “gold,” and “tiger.” Our results do, however, leave open the possibility that some terms (such as “rice”) are ambiguous in the sense proposed by a number of recent studies, but further empirical work is needed to resolve this issue.

## Acknowledgments

We are very grateful to Daniel Cohnitz, Nicolò D’Agruma, Michael Devitt, Joshua Knobe, Brian Porter, and an anonymous referee for this journal, for extremely valuable feedback on earlier versions of the paper. The material was also presented in seminars in Barcelona and Wellington, and we wish to thank the audiences for the helpful discussion. Jussi Jylkkä’s work was financially supported by the Kone Foundation (grant number 202105363).

## Open Research Badges

   This article has earned Open Data, Open Materials, and pre-registered badges. Data and materials are available at [https://osf.io/ktvgp/?view\\_only=e17808acf6d54a62b3b4133de30e6c9d](https://osf.io/ktvgp/?view_only=e17808acf6d54a62b3b4133de30e6c9d) and pre-registered are available at <https://osf.io/shx69>.

## Notes

- 1 The connections between theories of reference for natural kind terms on the one hand, and psychological theories of concepts on the other hand, are explored in more detail in Jylkkä (2008).
- 2 For a more comprehensive review of psychological theories of concepts and categorization, see Laurence and Margolis (1999).
- 3 Exceptions are Braisby, Franks, and Hampton (1996) and Jylkkä, Railo, and Haukioja (2009), both of which aim to test what the former call “the essentialist approach to word meaning”: the view that the reference of natural kind terms is determined by the underlying essence of natural kinds, independently of speakers’ beliefs about what that underlying essence is. Both studies found that speakers sometimes use natural kind terms in accordance with the superficial properties view, although Jylkkä et al. took their results, on the whole, to support what would here be called an underlying essence view.
- 4 See, for example, Hampton, Storms, Simmons, and Heussen (2009), and more recently Rose, Jaramillo, Nichols, and Horne (2022) and Machery et al. (2023). Further, Rose and Nichols (2019) argue that people’s judgments concerning category membership are driven by neither underlying (scientific) essence nor outward appearance, but rather by facts about an entity’s Aristotelian *telos*. Our primary focus in this paper is on the question of whether natural kind terms are ambiguous, and this overview will mostly focus on the work in experimental semantics that is most directly relevant to this question.
- 5 Devitt and Porter (forthcoming) take this view to be supported by two further experiments, which according to them suggest that a hybrid account is true of a biological kind term that has what they call a “practical interest,” whereas the causal-historical account is true of a biological kind term that lacks a practical interest.
- 6 Both Nichols et al. and Tobia et al. take the relevant kind of ambiguity to be an instance of polysemy: natural kind terms have different but closely related senses and, therefore, different extensions.
- 7 See also Martí (2009; 2015). In contrast to Genone and Lombrozo, Nichols et al. and Devitt and Porter do not literally ask who or what someone else is talking about, or whether other individuals are talking or thinking about the same thing. Nevertheless, descriptivism predicts reference by virtue of associated descriptive properties and the causal-historical account predicts reference by virtue of an individual standing at the end of a communicative chain. It is, however, the descriptive properties associated by other speakers, and the communicative chains ending in someone else’s use of words that are intended to guide subjects’ responses in their experiments.
- 8 Neither Tobia et al.’s nor our study use “XYZ” in the vignettes (but rather “the liquid from Twin Earth”). We use the familiar abbreviation in our discussion, for ease of exposition.
- 9 Tobia et al. (p. 201–202, 204) acknowledge that a reaction of this kind is possible.
- 10 In order to understand the crucial features of their “gold” vignette, subjects would need to know that samples consisting of atoms with 79 protons are said to consist of elemental atoms.

- 11 Devitt and Porter's vignette stated that "[rice] comes from the grass plant *Oryza sativa*," which might be interpreted as stating a necessary condition for being rice; we replaced this with "All familiar varieties of rice are harvested from the grass plant *Oryza sativa*." Their vignette also stated that the scientists *are developing* the new plant, and later on, the subjects are asked to suppose that they succeed: ours was in perfect tense, in order to be consistent with the other vignettes.
- 12 The superficial properties view would entail yet different predictions. However, as both the results obtained by Tobia et al. and by us are clearly inconsistent with this view, its predictions will not be discussed here.
- 13 We are grateful to Joshua Knobe for pointing this out.
- 14 The subjects' agreement with nonmember statements, and disagreement with member statements, was stronger for the foils featuring the natural kind terms "salt" and "silver," than for "water," "gold," and "tiger." There are multiple possible explanations for this small difference. For example, some of the subjects could be reasoning diagnostically in the nonfoils, and take the sameness of superficial properties as evidence of sameness in underlying essence. Moreover, as far as the Tobia et al. and Inverted Tobia et al. question sets are concerned, if a significant proportion of the subjects interpreted the expression "there is a sense in which x is K" as, roughly, "x can be treated as K," it is to be expected that they would be more likely to categorize the novel samples as nonmembers in the foils, and/or be more confident in their judgments, because of the difference in superficial properties.
- 15 It should be noted that, due to the nature of the cases that have to be imagined, it will be simply impossible to describe them without mentioning facts about the underlying properties, and thereby using at least some scientific terminology, thereby potentially making the underlying essence sense more salient.
- 16 With respect to the neutral question set, the subjects tended to agree with (3b) ("The seeds produced by the new plant are not rice") and disagree with (3a) ("The seeds produced by the new plant are rice"); the difference between the two was much less marked than for "water," "gold," and "tiger," but nonetheless decisive ( $BF_{10}=119.226$ ).
- 17 A potential further complication concerns the vignette: unlike in the other cases, the superficial properties (in this case of the two seeds) do not match completely, as it is explicitly stated that the plants growing from the new seeds differ from *O. sativa* in that they are "much easier to grow, require much less water, and can be grown in a wider variety of climates." Moreover, some subjects might know that there are, in fact, other species of rice besides *O. sativa*.
- 18 Of course, water is hardly *less* practically significant for us than rice is, so this significance cannot be the *only* reason for "rice" being ambiguous (if it indeed is ambiguous).

## References

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, 59(3), 247–274.

- Devitt, M., & Porter, B. C. (2021). Testing the reference of biological kind terms. *Cognitive Science*, 45, e12979.
- Devitt, M., & Porter, B. C. (forthcoming). Two sorts of biological kind terms: The cases of ‘Rice’ and ‘Rio de Janeiro Myrtle’. *Philosophy and Phenomenological Research*.
- Gelman, S., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition*, 38, 213–244.
- Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.
- Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, 25(5), 717–742.
- Hampton, J. A., Storms, G., Simmons, C. L., & Heussen, D. (2009). Feature integration in natural language concepts. *Memory & Cognition*, 37, 1150–1163.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
- Jylkkä, J. (2008). Theories of natural kind term reference and empirical psychology. *Philosophical Studies*, 139(2), 153–169.
- Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers’ language use. *Philosophical Psychology*, 22(1), 37–60.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Kripke, S. A. (1980). *Naming and necessity*. Oxford: Basil Blackwell.
- Kroon, F. W. (1987). Causal descriptivism. *Australasian Journal of Philosophy*, 65(1), 1–17.
- Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 3–81). Cambridge, MA: MIT Press.
- Machery, E., Olivola, C. Y., Cheon, H., Kurniawan, I. T., Mauro, C., Struchiner, N., & Susianto, H. (2023). Is identity essentialism a fundamental feature of human cognition? *Cognitive Science*, 47(5), 1–29.
- Martí, G. (2009). Against semantic multi-culturalism. *Analysis*, 69(1), 42–48.
- Martí, G. (2015). General terms, hybrid theories and ambiguity: A discussion of some experimental results. In J. Haukioja (Ed.), *Advances in experimental philosophy of language* (pp. 157–172). London: Bloomsbury.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge: Cambridge University Press.
- Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind*, 125(497), 145–175.
- Putnam, H. (1975). The meaning of ‘meaning’. In H. Putnam (Ed.), *Philosophical papers* (pp. 215–271). Cambridge: Cambridge University Press.
- Rosch, E. H. (1978). Principles of categorization. In E. H. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.
- Rose, D., & Nichols, S. (2019). Teleological essentialism. *Cognitive Science*, 43(4), 1–19.
- Rose, D., Jaramillo, S., Nichols, S., & Horne, Z. (2022). Teleological essentialism across development. *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 1841–1847).
- Tobia, K., Newman, G. E., & Knobe, J. (2020). Water is and is not H<sub>2</sub>O. *Mind & Language*, 35(2), 183–208.

## Appendix

### *The vignettes and question sets*

#### *Gold*

The Maxwell Mining Company discovers how to mine for metals on asteroids. It recovers a large amount of metal that tests show is identical in all observable properties to the paradigm sample of gold stored as reference M17 in the Paris Department of Precious Metals. These properties include appearance, weight, conductivity, melting point, and other markers that distinguish gold from other metals. When they perform a chemical analysis of the asteroid’s metal, however, they find out that it does not consist of atoms with 79 protons. This is surprising, because scientists long ago discovered that all the samples of gold on Earth are



composed of atoms with 79 protons. The metal from the asteroid is entirely composed of a compound not found on Earth. This compound, and the element consisting of atoms with 79 protons, have the exact same appearance and behave in exactly the same way in all everyday circumstances, but they can easily be distinguished from each other through proper tests in the laboratory.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the metal from the asteroid is gold.
- (1b) Ultimately, if you think about what it really means to be gold, you'd have to say there's a sense in which the metal from the asteroid is not truly gold at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the metal from the asteroid is not gold.
- (2b) Ultimately, if you think about what it really means to be gold, you'd have to say there's a sense in which the metal from the asteroid is truly gold.

**QS3:** How strongly do you agree with the following claims:

- (3a) The metal from the asteroid is gold.
- (3b) The metal from the asteroid is not gold.

### *Salt (foil)*

Salt is widely used as food seasoning all around the world. It is processed from salt mines, and by evaporation from sea water, and is composed of sodium chloride (*NaCl*). Salt has, however, many known negative health effects, including increased blood pressure, and increased risk of cardiovascular diseases. Scientists and nutritionists at the Kripnam Laboratory have been trying to develop a new food seasoning with a salty taste, but without adverse health effects. So far, they have not been successful, but as an accidental side product of their work, they have developed and launched a new product that tastes a lot like sugar, but does not have any of the negative health effects of either salt or sugar.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the new product is salt.
- (1b) Ultimately, if you think about what it really means to be salt, you'd have to say there's a sense in which the new product is not truly salt at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the new product is not salt.
- (2b) Ultimately, if you think about what it really means to be salt, you'd have to say there's a sense in which the new product is truly salt.

**QS3:** How strongly do you agree with the following claims:

- (3a) The new product is salt.
- (3b) The new product is not salt.

### Tiger

Conservationists in the mountains of Asia have discovered an isolated population of animals, never before encountered by humans. These animals have feline characteristics and striped orange and black fur. These 600 pound carnivores are indistinguishable from familiar tigers in their behavior and appearance. Studies on their genetics show, however, that they do not belong to any of the known subspecies of *Panthera tigris*, the species to which all previously recognized tigers belong. In fact, genetic comparisons show that the new population is more closely related to lions (members of *Panthera leo*) than to the known members of *Panthera tigris*. Scientists issue a report stating that convergent evolution has led this isolated population to be indistinguishable from familiar tigers.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the animals the conservationists found are tigers.
- (1b) Ultimately, if you think about what it really means to be a tiger, you'd have to say there's a sense in which the animals the conservationists found are not truly tigers at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the animals the conservationists found are not tigers.
- (2b) Ultimately, if you think about what it really means to be a tiger, you'd have to say there's a sense in which the animals the conservationists found are truly tigers.

**QS3:** How strongly do you agree with the following claims:

- (3a) The animals the conservationists found are tigers.
- (3b) The animals the conservationists found are not tigers.

### Silver (foil)

Silver is a soft, white, lustrous metal found in the Earth's crust that is widely used in jewelry and various industries. Scientists long ago discovered that all the samples of silver on Earth are composed of atoms with 47 protons. Scientists named this element Ag. The Maxwell Mining Company has been looking for silver in new areas. Recently, they discovered, in Slovenia, large amounts of a metal that resembles Ag, but is noticeably darker and heavier. When they perform a chemical analysis of the metal, they find out that it does not consist of atoms with 47 protons, but is entirely composed of a compound not previously found on Earth. This compound has never before been found on Earth, and the company issues a report stating that the metal from Slovenia can be used instead of Ag, for a limited number of purposes.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the metal from Slovenia is silver.
- (1b) Ultimately, if you think about what it really means to be silver, you'd have to say there's a sense in which the metal from Slovenia is not truly silver at all.

**QS2** How strongly do you agree with the following claims:

- (2a) There's a sense in which the metal from Slovenia is not silver.
- (2b) Ultimately, if you think about what it really means to be silver, you'd have to say there's a sense in which the metal from Slovenia is truly silver.

**QS3:** How strongly do you agree with the following claims:

- (3a) The metal from Slovenia is silver.
- (3b) The metal from Slovenia is not silver.

### *Water*

Suppose that in a few years, humans are able to travel to other galaxies. While exploring, they land on a planet that looks exactly like Earth in virtually all respects. It is populated by plants and animals that look exactly like the familiar plants and animals on Earth. Its landscapes and ecosystems look and function exactly like those on Earth. They dub this planet "Twin Earth." The astronauts remove their helmets and find that they can breathe freely. They drink from the lakes and rivers and find that their contents look and taste just like water. They quench their thirst on what they collect from the lakes and rivers while they explore the planet. When they perform a chemical analysis of this liquid, they find out that it does not consist of  $H_2O$ , but another chemical compound, not found on Earth. This is surprising, because scientists long ago discovered that all the samples of water on Earth are composed of the compound  $H_2O$ . The liquid in Twin Earth's lakes and rivers behaves exactly like  $H_2O$  in everyday circumstances, but can easily be distinguished from  $H_2O$  in the laboratory.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the liquid from Twin Earth is water.
- (1b) Ultimately, if you think about what it really means to be water, you'd have to say there's a sense in which the liquid from Twin Earth is not truly water at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the liquid from Twin Earth is not water.
- (2b) Ultimately, if you think about what it really means to be water, you'd have to say there's a sense in which the liquid from Twin Earth is truly water.

**QS3:** How strongly do you agree with the following claims:

- (3a) The liquid from Twin Earth is water.
- (3b) The liquid from Twin Earth is not water.

### *Bottle opener (foil)*

Bottle openers typically work by pivoting the edge of a bottle cap upward, so that the cap is removed from the rim of the bottle. The Putke Company specializes in expensive novelty items for wealthy customers. For many years, they have received requests for new kinds of bottle openers, but they have struggled to develop reliable but novel mechanisms. Recently, however, their engineers made a breakthrough, and developed a device which removes bottle caps without physical contact, using strong magnetic fields created by an electric current. This

new product has now been launched and made available for purchase. The company's sales department expects it to become a hit among novelty-seeking clients.

**QS1:** How strongly do you agree with the following claims:

- (1a) There's a sense in which the new product is a bottle opener.
- (1b) Ultimately, if you think about what it really means to be a bottle opener, you'd have to say there's a sense in which the new product is not truly a bottle opener at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the new product is not a bottle opener.
- (2b) Ultimately, if you think about what it really means to be a bottle opener, you'd have to say there's a sense in which the new product is truly a bottle opener.

**QS3:** How strongly do you agree with the following claims:

- (3a) The new product is a bottle opener.
- (3b) The new product is not a bottle opener.

### Rice

Rice is a seed that is an important part of the diet of millions around the world. All familiar varieties of rice are harvested 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 Wellmax Laboratory have been working to combat this problem. Rather than trying to modify the genes of *Oryza sativa*, they have developed a new plant altogether. This plant produces seeds that have the exact same look, taste, and nutritional content as the seeds of *Oryza sativa*. But the new plant is genetically different and is not *Oryza sativa*. Because of this genetic difference, the plants are much easier to grow, 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 substantially help combat food shortages.

How strongly do you agree with the following claims:

- (1a) There's a sense in which the seeds produced by the new plant are rice.
- (1b) Ultimately, if you think about what it really means to be rice, you'd have to say there's a sense in which the seeds produced by the new plant are not truly rice at all.

**QS2:** How strongly do you agree with the following claims:

- (2a) There's a sense in which the seeds produced by the new plant are not rice.
- (2b) Ultimately, if you think about what it really means to be rice, you'd have to say there's a sense in which the seeds produced by the new plant are truly rice.

**QS3:** How strongly do you agree with the following claims:

- (3a) The seeds produced by the new plant are rice.
- (3b) The seeds produced by the new plant are not rice.