Received: 21 March 2019 | Revised: 3 July 2019 | Accepted: 21 August 2019

DOI: 10.1111/mila.12278

SUBMITTED ARTICLE

WILEY

Reports from Twin Earth: Both deep structure and appearance determine the reference of natural kind terms

Jussi Haukioja[1] | Mons Nyquist[1] | Jussi Jylkkä[2]

[1]Department of Philosophy and Religious Studies, Norwegian University of Science and Technology, Trondheim, Norway

[2]Department of Psychology, Åbo Akademi University, Turku, Finland

Correspondence
Jussi Haukioja, Department of Philosophy and Religious Studies, Norwegian University of Science and Technology, 7491 Trondheim, Norway.
Email: jussi.haukioja@ntnu.no

Funding information
Abo Akademi University Endowment, the BrainTrain project; Jussi Jylkkä, personal grant

Following the influential thought experiments by Hilary Putnam and others, philosophers of language have for the most part adopted semantic externalism concerning natural kind terms. In this article, we present results from three experiments on the reference of natural kind terms. Our results confirm some standard externalist assumptions, but are in conflict with others: Ordinary speakers take both appearance and underlying nature to be central in their categorization judgments. Moreover, our results indicate that speakers' categorization judgments are gradual, and proportional to the degree of similarity between new samples and familiar, "standard" samples. These findings pose problems for traditional theories, both externalist and internalist.

KEYWORDS
causal theory of reference, experimental philosophy, psychological essentialism, semantic externalism, semantic internalism, thinking style

1 | INTRODUCTION

Mainstream externalist theories of reference for natural kind terms hold that their reference is at least partly determined by the deep structure of the samples the term is or has been applied to, and not solely by the relevant speakers' mental states. For instance, externalists typically hold that being H_2O is necessary and sufficient for anything to belong in the extension of the term "water," even when the term is used by someone ignorant of the molecular structure of water. This is one upshot of Hilary Putnam's influential Twin Earth thought experiment (Putnam, 1975). However, Twin Earth type cases can only assess whether deep structure is considered as necessary for natural category membership, not whether it is sufficient. Existing experimental studies on ordinary speakers' use of natural kind terms have not explicitly distinguished between the necessity and sufficiency questions (e.g., Malt, 1994; Braisby, Franks, & Hampton, 1996; Genone & Lombrozo, 2012; Jylkkä, Railo, & Haukioja, 2009; Nichols, Pinillos, & Mallon, 2015; Tobia, Newman, & Knobe, 2019).

In this article, we present results from three experiments that were conducted on ordinary speakers' usage of natural kind terms. We used two types of case: *Twin Earth* scenarios, where participants had to categorize samples of a natural kind that had identical appearance but a different deep structure from the standard samples, and *reverse Twin Earth* scenarios, where deep structure, but not appearance, was shared with the standard samples. In Experiments 1 and 2, we used scenarios where the underlying nature and the appearance of samples found on a new planet were either identical to, or completely different from, the standard samples. In Experiment 3, we examined whether underlying structure and appearance have to be *completely* shared with standard samples in order for a new sample to belong to the same kind, or whether *similarity* to standard samples in one or both respects would be enough. As an additional research question, we examined whether speakers' categorization judgments are associated with their thinking style (rational or experiental; Epstein, Pacini, Denes-Raj, & Heier, 1996).

1.1 | Theoretical background and previous work

Hilary Putnam's *Twin Earth* case (Putnam, 1975) is one of the most celebrated thought experiments in philosophy of language. It is generally thought to give considerable support for *semantic externalism*: the view that the meaning and extension of some linguistic expressions, in particular natural kind terms, are at least partly dependent on factors external to the individual speaker. "Meanings just ain't in the head," as Putnam (1975, p. 227) memorably put it. Meanings are not in the head, because what is in the head is not sufficient for determining what natural kind terms refer to.

In his thought experiment, Putnam asks us to imagine a planet, Twin Earth, that is very much like our Earth: We may even imagine that each of us has a duplicate on Twin Earth, sharing all our internal properties, our behavioral history, and so on. However, the liquid called "water" on Twin Earth does not consist of H_2O, but of XYZ, where "XYZ" is an abbreviation for a complex chemical formula. XYZ is "indistinguishable from water at normal temperatures and pressures" (Putnam, 1975, p. 223). Putnam then imagines that a spaceship from Earth visits Twin Earth. At first the Earthlings will, according to Putnam, assume that "water" has the same meaning on Earth and on Twin Earth. When apprised of the chemistry, however, the Earthian spaceship will report somewhat as follows: "On Twin Earth the word 'water' means '*XYZ*'." (Putnam, 1975).[1] Our word "water" simply does not apply to XYZ, but only to H_2O, even though the two liquids

[1] We think it is far more likely that the astronauts—given that they in fact use natural kind terms in the way Putnam hypothesized—would report something like the following: "On Twin Earth, there is no water, but a very similar liquid, XYZ." As far as the philosophical consequences that Putnam wants to draw from his example are concerned, this would do the job equally well.

have identical appearances and functional properties. Similarly, the Twin Earthlings' tokens of a phonetically identical word do not apply to H_2O, but only to XYZ.

Relatedly, after having presented his causal–historical theory of reference for proper names, Kripke (1980, pp. 116–139) argues that terms for natural kinds ("water," "gold," "tiger," and so on) and natural phenomena ("lightning," "heat") function semantically much like proper names do. He argues—also by thought experiment—that the descriptive properties that ordinary speakers associate with natural kind terms are at most contingently possessed by members of the corresponding kinds.

Both Putnam and Kripke conclude that the reference of natural kind terms is determined by underlying structure.[2] A natural kind term applies to all and only those things that belong to the same kind as the "standard samples," that is, most of the things we have actually applied the term to. And, if we follow Putnam in understanding kinds as individuated by underlying structure, we get the following: A natural kind term applies to all and only those things that have the same underlying structure as standard samples.

Note that this formulation takes sharing underlying structure to be both necessary and sufficient for belonging in the extension of a natural kind term. In what follows, it will be helpful to distinguish between these: For a natural kind term T, we should ask two separate questions:

> *The necessity question*: Is sharing underlying structure with standard samples *necessary* for belonging in the extension of T?
>
> *The sufficiency question*: Is sharing underlying structure with standard samples *sufficient* for belonging in the extension of T?

Kripke and Putnam answered both questions affirmatively, and many philosophers of language today would agree. However, we should note that the Twin Earth thought experiment, by itself, only speaks to the first question. If successful, the thought experiment tells us that a completely watery substance cannot be water unless it is composed of H_2O; it does not tell us whether nonwatery H_2O is water. An affirmative answer to the sufficiency question is often assumed without separate argument, but one might well accept an affirmative answer to the necessity question while answering the sufficiency question in the negative, or remaining agnostic about it (Häggqvist & Wikforss, 2015; cf. Steward, 1990, where the sufficiency question is discussed using a thought experiment resembling our reverse Twin Earth cases).

Kripke and Putnam argue, on the basis of Twin Earth and other thought experiments, for a view that combines externalism about meaning, a causal–historical theory of reference, and essentialism about natural kinds. We will here call this combination of views simply "externalism," and contrast it with "internalism," which claims that reference is determined by associated manifest properties, making the view both internalist and nonessentialist. In reality, the theoretical landscape is considerably more varied; many internalists have sought to accommodate judgments about Twin Earth and related cases by incorporating rigidification or causal

[2] Putnam and Kripke do not say very much about what it is for two things to share underlying structure. Putnam stresses that the *same liquid* (or, more generally, *same kind*) relation is a theoretical one: "[W]hether something is or is not the same liquid as *this* may take an indeterminate amount of scientific investigation to determine" (Putnam, 1975, p. 225). In general, underlying structures in these discussions have been assumed to be empirically discoverable properties that play a causal-explanatory role in the determination of the kind's manifest properties (e.g., Leslie, 2013, p. 125)—taking molecular structure as the underlying nature of water is then dependent on the empirical assumption that molecular structure does, indeed, play this role for water.

elements into descriptivism. We will return to these issues at the end of the article, but to simplify the discussion, we use the labels "internalism" and "externalism" to denote the two extremes: Internalism answers "no" to both the necessity and the sufficiency question, while externalism answers "yes" to both. Our main aim in this article is to find out whether the kinds of judgments typically taken as strong evidence for Kripke–Putnam essentialism are as widespread as externalists typically have assumed, not to come to any final conclusion on externalism and internalism more generally.

Putnam's judgments concerning the Twin Earth case have not gone completely unchallenged. Although most philosophers of language have agreed with him, according to what is sometimes called the *common concept strategy* (Crane, 1991; Mellor, 1977; Segal, 2000), "water" expresses the same concept when uttered by an Earthling and her twin: a concept that has both H_2O and XYZ in its extension. On this view, Putnam and his astronauts reporting to Earth were just wrong: XYZ is merely another kind of water.

In discussing thought experiments, externalists typically appeal to "what we would say" about various cases, where "we" is intended to cover more than just academic philosophers of language. Empirical results concerning ordinary speakers' usage of natural kind terms should, then, be potentially relevant to the evaluation of theories of meaning and reference for such terms.[3] There is a long tradition of research in psychology on *psychological essentialism*, the view that humans take hidden deep structure to at least partly determine membership in natural categories. Most of the empirical work on psychological essentialism has focused on children's categorization (typically of animals; cf. Keil, 1989; Medin & Ortony, 1989; Rips, 2001; Gelman, 2003, 2004). These studies have found that at least children represent natural kinds as having deep structure or "essence": some (hidden) property that is central for something to belong in the kind, and which causes the characteristic observable properties of members of the kind. The precise nature of the essence need not be known, but instead the concept can include a placeholder for some unknown essence (Medin & Ortony, 1989; Rips, 2001). However, a typical finding in the literature is that deep structure is not alone necessary and/or sufficient for category membership, which is also affected by appearance (e.g., Braisby et al., 1996; Braisby, Braisby, 2001, Braisby, 2004; Hampton, Estes, & Simmons, 2007). This has motivated a causal homeostasis account of concepts, where deep structure and appearance are taken to be causally linked, so that neither alone is sufficient to determine category membership (Hampton et al., 2007; Rehder & Kim, 2010). On this approach, deep and superficial properties cannot be separated, because appearance is also evidence about deep structure. In general, psychological essentialism is primarily a theory of categorization, not of semantic externalism, but it can be argued that, at least in the case of natural kind terms, externalist language use presupposes essentialist language use: Speakers cannot take an *external* deep structure to determine extension unless they consider *some* deep structure to determine extension (Jylkkä et al., 2009).

Also worth noting is Malt's (1994) study, performed on adults. Malt found that subjects' beliefs about the amount of H_2O in a liquid sample did not strongly predict whether the sample was categorized as "water." The subjects called liquids "water" although they did not believe they had H_2O as the main ingredient, or believed that they contained only a small percentage of H_2O. On the other hand, Malt found that subjects did not categorize as "water" many liquids,

[3] Precisely how experimental results are relevant to philosophy of language is, of course, a complex and controversial topic. For our purposes here, it is enough to assume that, should it be found that ordinary usage systematically conflicts with a theory of reference, that would count as evidence against the theory in question. In what follows, we will take this as uncontroversial—for more detailed discussion of these issues, see Cohnitz and Haukioja (2013).

such as juice or coffee, although they believed the liquids were high in H_2O. In all the samples called "water," by contrast, H_2O was always judged to be present, to some extent. Malt tentatively suggests that this may indicate that a liquid's containing H_2O is necessary for being called "water," though not sufficient.

Experimental work has also been carried out on the semantics of natural kind terms, directly (Braisby et al., 1996; Genone & Lombrozo, 2012; Jylkkä et al., 2009; Nichols et al., 2015; Tobia et al., 2019). The results obtained in these studies do not point in a uniform direction, but all of them cast doubt on the externalist orthodoxy: While speakers in many situations do use natural kind terms in ways that are consistent with externalism, in others they seem to apply them on the basis of observable properties, as traditional internalism would have it. However, these studies do not give us a clear and uniform understanding of when speakers rely more on deep structure, when on observable properties, and why (for discussion, see Hansen, 2015; Martí, 2015; Häggqvist & Wikforss, 2015; Cohnitz & Haukioja, 2020). Moreover, no published studies have looked at ordinary speakers' reactions to the standard kinds of Twin Earth cases, which philosophers of language typically put most weight on. This is quite surprising, given that the thought experiments have had such a central role, and that (as noted above) not all philosophers have agreed about how ordinary speakers *would* use and interpret natural kind terms in them. Furthermore, none of the previous studies have clearly distinguished between the necessity and sufficiency questions.

If it were found that ordinary speakers' use of natural kind terms displays the kind of split pattern suggested by the existing studies, with speaker usage in some situations agreeing with externalism, in others with internalism, that would put considerable pressure on both of the views in their "pure" form. However, as argued in more detail in Cohnitz and Haukioja (2020), it would be premature to conclude on that basis that externalism (or internalism) is false. We know that speakers may be disposed to make systematic errors, and it is not clear why such a pattern of use should be taken as evidence against externalism (or internalism) and not as evidence of systematic error in the speakers' use of the term, in one direction or the other. In the absence of clear criteria for when a particular application of a term is to be deemed erroneous, we simply cannot draw such direct conclusions from the data. What, then, could be evidence for taking a particular application of a term as erroneous? One plausible answer to this question is: If the speaker themselves is disposed to retract their application, either as a result of thinking more closely about the case, or as a result of learning that their usage does not line up with that of other speakers, in particular that of the relevant experts, then we are prima facie justified in taking their first application of the term as erroneous.[4] None of the studies listed above have tried to look at such self-corrective behaviour (but see Braisby, 2001, 2004, who found limited evidence of deference).

1.2 | The present study

In the study reported here, we have taken steps to address the issues mentioned earlier. Our experiment looks directly at Twin Earth style cases, using different variants to address both the necessity question *and* the sufficiency question. Unlike previous studies, we have also

[4] This view of what to count as errors is, we take it, intuitively plausible, but it can be also motivated by general dispositionalist theories of reference determination (cf. Cohnitz & Haukioja, 2013; Johnson & Nado, 2014). Of course, merely the fact that the speaker correct themselves does not guarantee that the original application was erroneous: Sometimes we correct ourselves when we in fact were right. All we are assuming here is that, *on the whole*, when our initial judgments and our later, more considered judgments diverge, the latter are to be given more weight.

supplemented the survey methodology with an elicited production task, attempting to get speakers to *use* the relevant natural kind term, rather than merely report their truth value judgments.[5] We have also given the subjects the opportunity to reconsider their initial judgments, trying to detect both self-correction and deference.

We conducted three experiments on lay speakers' usage and understanding of natural kind terms. In Experiments 1 and 2, the subjects were presented with Twin Earth and reverse Twin Earth thought experiments (cf. Section 1), where new samples of a naturally occurring substance, phenomenon, and so forth, were found, such that these were either identical in appearance but completely different in underlying structure, or vice versa, compared to samples of a familiar natural kind. The aim of Experiments 1 and 2 was to test whether lay speakers' language use and interpretation is driven solely by underlying structure, or whether appearance properties also play a role: In particular, we tested whether sharing underlying structure with standard samples is, as externalists have assumed, both necessary and sufficient for belonging in the extension of a natural kind term. The setup of Experiment 3 was similar, but the differences in appearance and underlying nature were less dramatic. The aim of this last experiment was to examine whether lay speakers' use and interpretation of natural kind terms is only sensitive to underlying structure and/or appearance that is completely shared with standard samples, or whether similarity along one of the dimensions is sufficient, and whether natural category membership is understood as graded.

In addition to probing the participants' use of natural kind terms, we investigated whether their responses were associated with their thinking style, rational or experiental (Epstein et al., 1996). This was motivated by previous research suggesting that there is substantial inter-individual variation in semantic judgments (Jylkkä et al., 2009), and that the background of an individual may be associated with their semantic judgments (Machery, Mallon, Nichols, & Stich, 2004). We hypothesized that internalist semantic judgments, where the appearance of the samples is considered as central, would be associated with an experiental thinking style, whereas externalist judgments would be associated with a rational thinking style. We assumed that relying on deep structure that is beyond appearance can be considered as rational (in contrast to experiental) thinking, because there the participant relies more on the causally efficacious underlying properties than on immediately available appearance properties. A 10-item version of the rational–experiential inventory (REI-10) was used to measure experiential versus rational thinking style (Epstein et al., 1996). All three experiments were preregistered and the anonymized data are openly accessible at http://osf.io/jdf7g/.

2 | EXPERIMENTS 1 AND 2

2.1 | Method in Experiments 1 and 2

Experiments 1 and 2 had identical setups, except for the natural kinds used. A total of five different natural kind terms were used: "water," "lightning," and "tiger" in Experiment 1; and "water," "diamond," and "gold" in Experiment 2 (i.e., the probes for "water" described below were used in both experiments). For each natural kind term, two types of scenario were

[5] Merely asking for speakers' metalinguistic judgments, or truth value judgments, gives rise to various potential sources of systematic error (cf. Martí, 2009; Cohnitz, 2015). Elicited production has recently also been used to study the reference of proper names (cf. Devitt & Porot, 2018)

prepared, one Twin Earth-like case (hereafter, "TE-case"; same appearance, different underlying nature), and one "reverse Twin Earth case" (hereafter, "reverse-TE-case"; different appearance, same underlying nature).[6] Each participant was presented with one or two TE-cases and one or two reverse-TE-cases, always a total of three cases featuring three different natural kinds. The order of the natural kinds and scenario types (TE and reverse-TE) were counterbalanced.

Each of the probes presented to the subjects had the same structure. First, the subjects were given a description of a case, in the form of a short story and an image. For "water," for example, the TE-case description read as follows[7]:

> Imagine that you are a member of a large group of astronauts who have traveled to another solar system. The group is composed of experts in various fields. You have landed on a planet that has not been previously explored. You find, to your great excitement, that the planet contains a rich plant life. The planet also contains a number of seas, lakes and rivers that contain a liquid that looks like this: [an image of a watery liquid]. The liquid evaporates from the surface of seas and lakes and falls down as rain, just like on Earth. However, when chemists in your group analyze the liquid, they find that it does not have the chemical structure H_2O, but rather a complex structure that you abbreviate as XYZ. XYZ is clear, odourless, tasteless, thirst-quenching, and supports life: In everyday circumstances it appears just as H_2O, but can easily be distinguished from it in the laboratory.

The reverse-TE-case for "water" was the following:

> Imagine that you are a member of a large group of astronauts who have traveled to another solar system. The group is composed of experts in various fields. You have landed on a planet that has not been previously explored, but is listed as a candidate for containing life. You find, to your disappointment, no signs of life. Chemists in your group conduct a routine analysis of substances found on the surface of the planet. To their great surprise, they find that a substance found on the planet, which they took to be a kind of mineral, has the molecular structure H_2O. For some mysterious reason, H_2O on this planet is solid, not liquid, up to about 800°C, and it is greenish. H_2O is widely found on the planet as lumps that look like this: [an image of a greenish mineral]. The chemists can think of no plausible explanation for these puzzling phenomena.

All 10 cases are summarized in Table 1.

After this, the subjects were presented with Question 1 (Q1), the elicited production task:

Question 1

[6] The main conclusion of the Twin Earth thought experiment is often taken to be that internal duplicates (like $Oscar_1$ and $Oscar_2$ in Putnam's original presentation) can refer to different kinds with their natural kind terms. However, to get to that conclusion, Putnam relies on an assumption about how our natural kind terms work: That XYZ is not in the extension of *our* term "water." Once that assumption is accepted, we can go on to imagine Twin English speakers, note that the situation is completely symmetrical, and draw the conclusion that internal duplicates can refer to different kinds with their (orthographically identical) natural kind terms. Our vignettes do not probe subjects' judgments concerning internal duplicates: Our aim is to see whether Putnam's assumption about how our natural kind terms work is correct.

[7] This is a direct translation from Norwegian.

> Right after having been informed of these discoveries, you are told to report on your findings about the [liquid (TE-case for water)/substance (reverse-TE-case for water), etc.] back to your base on Earth. Draft a report in your own words (no more than 2 lines):

Question 1 (Q1) serves a dual purpose. First, as an elicited production task, it provides direct evidence of language use. Second, having to summarize the findings ensures that the subjects have understood the scenario description before going on to answer Question 2 (presented on a separate sheet of paper, which they were instructed to read only after having answered Q1).

Question 2 (Q2) provided the subject with two clearly formulated alternatives, where one is consistent with externalism, the other with internalism.[8] The order of the alternatives was randomized. For example, Q2 for the TE-case for "water" was the following:

Question 2

> There is disagreement among your group on how to report your findings back to Earth, and you decide to take a vote. In the vote, you are given two options, (a) and (b):
>
> (a) We have found a new water-like substance on this planet, but it is not water. It consists of XYZ, not H_2O.
>
> (b) We have found a new kind of water on this planet. It consists of XYZ rather than H_2O.

Which do you think is a better description of your findings, (a) or (b)? Please answer by circling a number on the scale.

1 2 3 4 5 6 7
(a) is clearly better (b) is clearly better

In this case, (a) would represent the externalist answer and (b) the internalist answer. For the reverse-TE-case for "water," the question was as follows:

Question 2

> There is disagreement among your group on how to report your findings back to Earth, and you decide to take a vote. In the vote, you are given two options, (a) and (b):
>
> (a) We have found a strange new form of water on this planet. It is solid in temperatures under 800°C, and it is greenish.
>
> (b) We have found a strange substance on this planet. It has the molecular structure H_2O, but it is not water, but a mineral.

Which do you think is a better description of your findings, (a) or (b)? Please answer by circling a number on the scale.

1 2 3 4 5 6 7

[8] Again, we are using the labels "externalist" and "internalist" for the opposite patterns of response, and not assuming that externalist theories are necessarily unable to account for the "internalist" responses, or vice versa.

TABLE 1 The scenarios used in Experiments 1 and 2

	Water		Lightning		Tiger		Diamond		Gold	
	TE	Reverse-TE	TE	Reverse-TE	TE	Reverse-TE	TE	Reverse-TE	TE	Reverse-TE
Appearance	Watery: Clear, odourless, tasteless liquid; an image of water was used.	Greenish, solid up to 800° C; an image of a greenish mineral was used.	As lightning on earth: Sudden flashes of light accompanied by loud booming sounds; an image of lightning was used.	Narrow cloud-like structures that appear and fade away over a few seconds; accompanied by a whistling sound; a manipulated image of a cloud formation was used.	Creatures appearing just as tigers on Earth: Striped, four-legged, etc.; an image of a tiger was used.	Creatures looking quite unlike tigers on Earth; an image depicting the prehistoric mammal *megatherium* was used.	Diamond-like: Transparent, very hard, disperses light; an image of an uncut diamond was used.	Nontransparent, soft, spongy material; an image of a sea sponge was used.	Gold-like: Yellow, malleable, heavy; an image of gold nuggets was used.	Light, brittle and bluish; an image of a blue mineral was used.
Underlying nature	XYZ	H_2O	Rapid biochemical reactions in bacteria floating in the atmosphere	Electrical discharge between clouds and the surface of the planet	Genetic material not stored in DNA but in previously unknown structures; could not reproduce with Earthian tigers	DNA identical to tigers on Earth, biologists conclude that the creatures could reproduce with Earthian tigers	Not composed of carbon, but a complex compound abbreviated as “FGH.”	Composed of carbon atoms in exactly the same configuration as diamonds on Earth	Complex compound abbreviated as “LMN”	Au

(a) is clearly better (b) is clearly better

In this case, (a) is the externalist answer and (b) the internalist. The choice between two alternatives in Q2 naturally provides us with evidence of the subjects' categorization judgments. In addition, possible patterns of mismatch between single subjects' answers to Q1 and Q2 could give evidence of corrective behavior: For example, if subjects whose answer to Q1 was categorized as internalist were to be found to tend to choose the externalist answer in Q2, that could be interpreted as evidence for externalism: While such subjects' first reactions (Q1) were internalist, on further consideration, and after having been made aware of two well-formulated alternatives, they would converge toward externalism. Naturally, the pattern could arise in the opposite direction as well.

After this, the subjects were presented with a third question ("Deference"), trying to detect their disposition to defer to experts. This was the same for the TE-case and for the reverse-TE-case:

Question 3

> Suppose that you find out that all the chemists[9] in your expedition have voted unlike you, and chosen the other alternative. How confident are you, after learning of this, that the option you chose was the more accurate one?
>
> 1 2 3 4 5 6 7
>
> not at all confident absolutely confident

Finally, the subjects were presented with a fourth question ("Imaginability") that asked how easy it was for them to imagine the depicted scenario, on a scale of 1 (very difficult to imagine) to 7 (very easy to imagine).

The answers to Q1 and Q2 were coded according to whether they were in accordance with internalism or externalism. Answers to Q1, the elicited production task, were coded by two independent raters on a scale from −2 (clearly internalist) to 2 (clearly externalist). A typical clearly internalist answer would be something like "we have found a new kind of water, composed of XYZ," a clearly externalist answer would be, for example, "there is no water on this planet, but another substance, XYZ, which appears very much like water," while neutral answers typically avoided using the term "water" at all. Answers to Q2, originally on a scale from 0 to 7, were recoded onto a scale from −3 (internalist option clearly better) to 3 (externalist option clearly better), depending on which alternative the participants preferred.

The thinking style questionnaire, REI-10, was administered after the scenarios. It consists of 10 items in total: five items that load on a "faith in intuition" (FI) factor, and five items that load on a "need for cognition" (NFC) factor. Subjects who score high on FI can be considered as having an experiential thinking style, whereas subjects high on NFC arguably have a rational thinking style. All the items are depicted in Table 2. Separate sum scores for FI and NFC were used in the analysis.

[9] The relevant experts were described as chemists, meteorologists, biologists, or mineralogists, depending on the natural kind in question.

TABLE 2 Items of the REI-10 questionnaire

Need for cognition (NFC)	
1	I do not like to have to do a lot of thinking. (R)
2	I try to avoid situations that require thinking in depth about something. (R)
3	I prefer to do something that challenges my thinking abilities rather than something that requires little thought.
4	I prefer complex to simple problems.
5	Thinking hard and for a long time about something gives me little satisfaction. (R)
Faith in intuition (FI)	
6	I trust my initial feelings about people.
7	I believe in trusting my hunches.
8	My initial impressions of people are almost always right.
9	When it comes to trusting people, I can usually rely on my "gut feelings."
10	I can usually feel when a person is right or wrong even if I cannot explain how I know.

Note: All items were answered on a 5-point scale from "completely false" to "completely true." Items marked with "R" were reverse-coded.[10]

2.2 | Experiment 1 participants

The participants ($N = 116$, 58 male) were university students from the Norwegian University of Science and Technology with average age 21.6 ($SD = 4.3$), with no or very little prior exposure to philosophy of language. The group of subjects included both students in science and engineering (83 subjects), and students in humanities and the social sciences (33 subjects). Seventy-seven percent of the participants reported Norwegian as their mother tongue. Based on their answers to Q1, they all appeared to be fluent, judged by a native speaker. Their self-reported knowledge of physics was 2.65 ($SD = 1.01$) and of chemistry 2.4 ($SD = .75$, on a scale from 1 to 5 in both).

2.3 | Experiment 1 results

The inter-rater agreement in Q1 was ICC = .59, $p < .001$ (two-way random, absolute agreement). Final value was determined through consensus where there was no agreement in the initial rating.

Mean ratings to each of the questions by scenario type in Experiment 1 are summarized in Table 3.

To examine differences in responses between scenario types, linear mixed effects models (lme4 in R) were used with rating to a question as dependent variable, scenario type (TE or reverse-TE) as fixed factor (simple coded), and "subject" and "natural kind" as random factors.

Reverse-TE scenarios were answered more internalistically than TE scenarios in both Q1 (E = −.37, SE = .080, $t = -4.61$, $p < .001$) and Q2 (E = −3.49, SE = .17, $t = -20.83$, $p < .001$).

[10] Note that Item 5 was not reverse-coded in Epstein et al., 1996, but we believe this is a mistake: clearly, Item 5 should be inversely associated with a higher score on NFC.

	TE	Reverse-TE
Q1	.27 (.82)	−.11 (.68)
Q2	1.84 (1.51)	−1.68 (1.64)
Deference	4.37 (1.76)	4.36 (1.74)
Imaginability	4.99 (1.58)	4.57 (1.65)

TABLE 3 Mean ratings (SD) to the questions depending on scenario type in Experiment 1

There was no difference between scenario types in the participants' tendency to defer to experts (E = .064, SE = .12, $t = .52$, $p = .60$). Reverse-TE scenarios were judged to be less imaginable than TE scenarios (E = −.35, *SE* = .095, $t = -3.75$, $p < .001$).

As a post hoc test, we examined whether the difference in imaginability between TE and reverse-TE scenarios moderates responses to Q1 and Q2. The interaction scenario type × imaginability was not significant in the case of either question (*p*'s > .51), indicating that difference in ratings to Q1 and Q2 between scenario types was not due to difference in imaginability. In the models with imaginability as predictor, reverse-TE scenarios were not judged more internalistically in Q1 ($p = .44$), but they were in Q2 (E = −3.47, SE = .53, $t = -6.579$, $p < .001$).

2.3.1 | Differences between natural kinds

To examine if there are differences in responses to Q1, Q2, deference, or imaginability depending on natural kind, we used a model with one of the questions as dependent variable, kind and type as predictors, and subject as random factor. We report main effects and interactions from ANOVA (Type III with Satterthwaite's method) that is based on the lmer model.

In Q1, there was an interaction between kind and type (F = 3.81, $p = .023$) as well as a main effect of kind (F = 4.70, $p = .010$) and type (F = 21.39, $p < .001$; see Figure 1a). With respect to all natural kinds, reverse-TE cases were judged more internalistically, but the effect was stronger for "tiger" than the other kinds (see Figure 1a). Likewise, in the case of Q2, there was an interaction between kind and type (F = 3.22, $p = .023$) and a main effect of kind (F = 5.20, $p < .01$) and type (F = 437.39, $p < .001$). From Figure 1b one can see that the effect of type was weaker for "water" than for the other kinds, although across all kinds reverse-TE cases were rated more internalistically; additionally, water was rated more externalistically than the other kinds. In the case of deference, there was an interaction between kind and type (F = 4.51, $p = .012$), and for imaginability there was a main effect of type ($F = 13.64$, $p < .001$; see Figure 1c,d, respectively).

2.3.2 | Consistency between Q1 and Q2

To examine consistency between Q1 and Q2 across scenario types, we Z-transformed the Q1 and Q2 ratings and created a new simple coded categorical variable Question (Q1 or Q2) that specified which question the Z-coded response was from.[11] This approach enabled us to examine interactions between question and scenario type. We used a model with Z-score as

[11] The Z-transformation was conducted to make responses to Q1 and Q2 comparable, because they were on different scales. The transformed Z-variable indicates how many standard deviations a particular answer deviates from the mean. For instance, a Z-score of −1.5 means that the response is 1.5 SD lower than the average response. The transformation was done separately for Q1 and Q2.

TABLE 4 Mean ratings (*SD*) to the questions depending on scenario type in Experiment 2

	TE	Reverse-TE
Q1	.45 (.60)	.015 (.74)
Q2	1.57 (1.76)	−1.15 (1.91)
Deference	4.48 (1.71)	4.43 (1.66)
Imaginability	4.86 (1.57)	4.57 (1.65)

dependent variable; question and scenario type as fixed factors; and subject and kind as random effects. There was a significant main effect of scenario type (F = 242.46, p < .001) and an interaction between scenario type and question (F = 64.05, p < .001). In both scenario types, responses to Q1 and Q2 were in the same direction (externalistic in TE and internalistic in reverse-TE), but stronger in Q2 (see Figure 2).

2.3.3 | Associations between REI and categorization judgments

Associations between responses to Q1 and Q2, and the NFC and FI factors in REI were analyzed with linear mixed effects models with either Q1 or Q2 as dependent variable and both NFC and FI as continuous predictors. In Q1, higher FI was associated with more internalistic ratings (E = −.027, SE = .013, t = −2.04, p = .042), but in Q2, neither NFC or FI predicted the ratings (p's > .19).

2.4 | Experiment 2 participants

The participants (N = 133, 50 male) were university students from Norwegian University of Science and Technology with average age 21.11 (*SD* = 2.30). The group of subjects included both students in science and engineering, and students in humanities and the social sciences (41 in science and engineering, 92 in humanities and the social sciences). Eighty-three percent of the participants reported Norwegian as their mother tongue. Based on their answers to Q1 they all appeared to be fluent, judged by a native speaker. Their self-reported knowledge of physics was 2.11 (*SD* = 1.05) and of chemistry 2.23 (*SD* = .89; on a scale from 1 to 5 in both).

2.5 | Experiment 2 results

The inter-rater agreement in Q1 was ICC = .71, p < .001 (two-way random, absolute agreement), final value was determined through consensus. Mean ratings to each of the questions by scenario type are summarized in Table 4.

Differences in ratings to all the questions between scenario types were investigated in linear mixed effects models in the same way as in Experiment 1. Answers were more internalistic in the reverse-TE than in TE scenarios in both Q1 (E = −.42, SE = .065, t = −6.51, p < .001) and in Q2 (E = −2.74, SE = .18, t = −15.20, p < .001). There was no difference in tendency to defer to experts between scenario types (E = −.057, SE = .11, t = −.50, p = .62). Reverse-TE scenarios were judged to be less imaginable than TE scenarios (E = −.29, SE = .075, t = −3.89, p < .001).

As in Experiment 1, as a post hoc test, we examined whether the difference in imaginability could moderate responses to Q1 or Q2. The interaction between type and imaginability was not significant in either case (p's > .15), indicating that the difference in ratings to Q1 and Q2 between scenario types was not due to difference in imaginability. In the model with imaginability as predictor, reverse-TE cases were judged more internalistically in both Q1 (E = −.69, SE = .20, $t = -3.38$, $p < .001$) and Q2 (E = −2.81, SE = .56, $t = -5.00$, $p < .001$).

2.5.1 | Differences between natural kinds

Interactions between type and kind were examined in the same manner as in Experiment 1. With Q1 as dependent variable, there was an interaction between type and kind (F = 5.83, $p = .0032$), as well as a main effect of type (F = 42.60, $p < .001$) and kind (F = 3.40, $p = .035$; see Figure 3a). As to the main effects, Gold was rated more externalistically than the other kinds, and all kinds were rated more internalistically in reverse-TE cases. The interaction between kind and type is likely due to "diamond," which shows a larger difference between scenario types than the other kinds (see Figure 3a). With Q2 rating as dependent variable, there was a significant interaction between kind and type (F = 5.47, $p = .005$) as well as a main effect of kind (F = 5.12, $p = .006$) and type (F = 235.21, $p < .001$; see Figure 3b). With respect to deference, all effects were nonsignificant, except for a near-significant interaction between kind and type (F = 2.44, $p = .089$; see Figure 3c). In imaginability, there was only a main effect of type (F = 15.03, $p < .001$; see Figure 3d).

2.5.2 | Consistency between Q1 and Q2

Consistency between Q1 and Q2 and interactions between *question* and *scenario type* were examined in the same manner as in Experiment 1. There was a main effect of *type* (F = 213.05, $p < .001$), as well as an interaction between type and question (F = 23.52, $p < .001$). In both questions the judgments were in the same direction (internalistic in reverse-TE and externalistic in TE), but stronger in Q2 (see Figure 4).

2.5.3 | Associations between REI and categorization judgments

Associations between the REI factors, NFC and FI, and Q1 and Q2 were examined in the same way as in Experiment 1. There were no significant associations between the REI factors and Q1 (p's > .09) or Q2 (p's > .70).

3 | EXPERIMENT 3

In Experiments 1 and 2, the underlying natures and appearances of the new samples were either identical to, or completely different from, the standard samples. Experiment 3 examined whether judging a new sample as belonging to a kind is proportional to the sample's similarity to standard samples of the kind, or whether identity along one, or both, of the dimensions is required for a sample to be categorized as belonging to the kind.

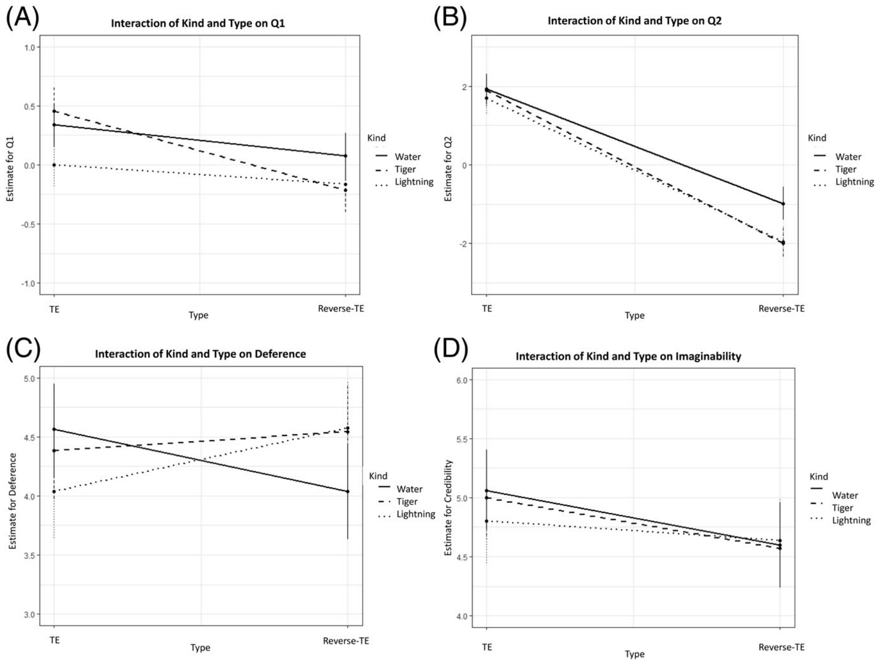

FIGURE 1 Ratings to the questions by kind and type in Experiment 1. Error bars represent 95% confidence intervals

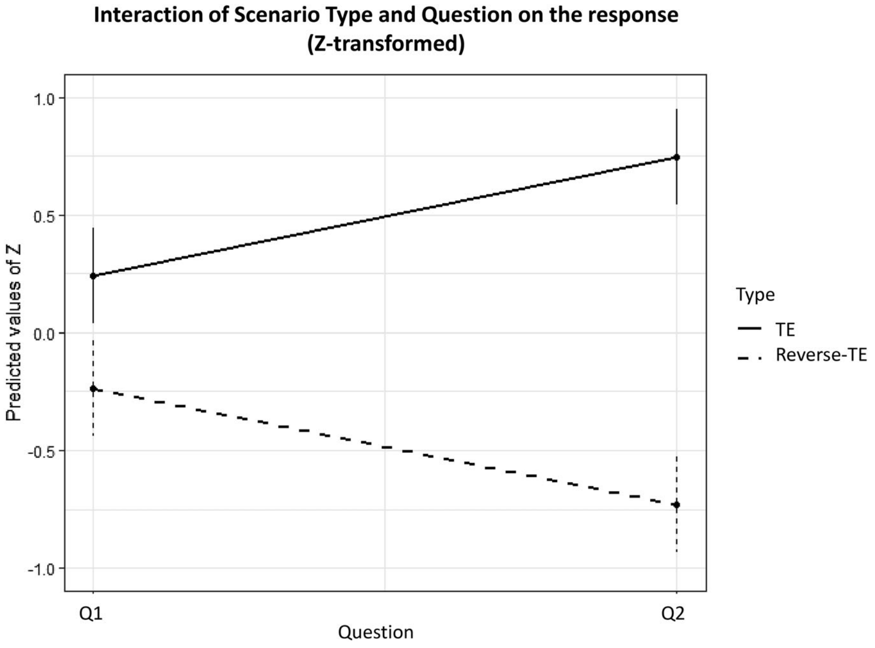

FIGURE 2 Interaction between question type (Q1 or Q2) and scenario type (TE or reverse-TE) in Experiment 1

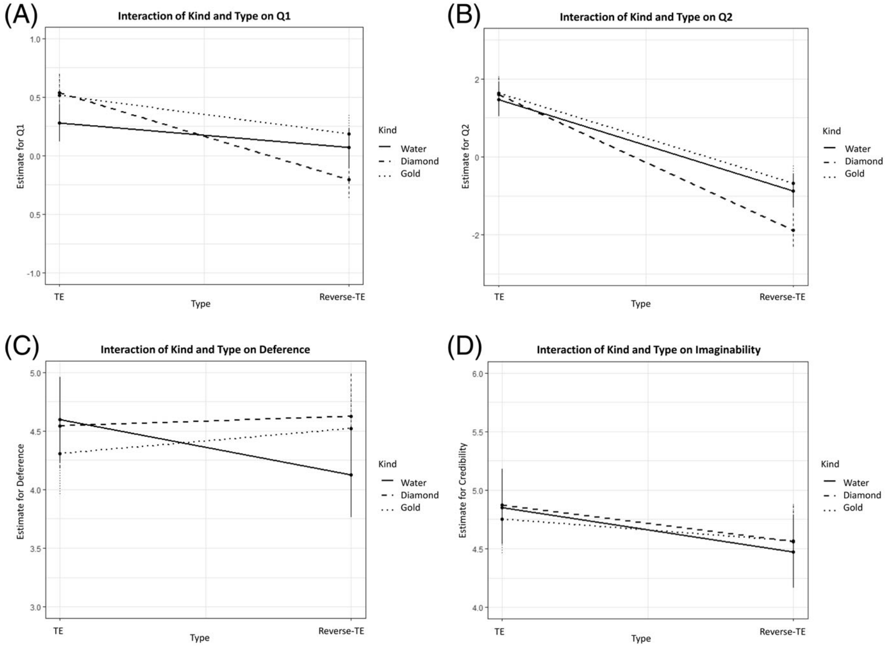

FIGURE 3 Ratings to the questions by *kind* and *type* in Experiment 2. Error bars represent 95% confidence intervals

3.1 | Method

Two different natural kind terms were used in Experiment 3: "water" and "diamond." In addition to using TE and reverse-TE cases as in Experiments 1 and 2, we also manipulated how radically the new samples deviated from the standard samples. For this purpose, "near" and "far" variations of the scenarios were formulated; in the former, the samples were not very different from the actual ones, whereas in the latter, they were markedly different, while not being as radically different as in Experiments 1 and 2. Thus, for each natural kind term, four different scenarios were prepared: TE-near, TE-far, reverse-TE-near, and reverse-TE-far. Each participant answered to one version of each natural kind (e.g., reverse-TE-near for "water" and TE-far for "diamond").

The cases had a similar structure to the ones used in Experiments 1 and 2. First, the subjects were given a description of a scenario, in the form of a short story and an image. For "water", for example, the TE-cases were as follows:

> Imagine that you are a member of a large group of astronauts who have traveled to another solar system. The group is composed of experts in various fields. You have landed on a planet that has not been previously explored. You find, to your great excitement, that the planet contains a rich plant life. The planet also contains a number of seas, lakes and rivers that contain a liquid that looks like this: [an image of a watery liquid].
>
> [near:] The liquid evaporates from the surface of seas and lakes and falls down as rain, just like on Earth. However, when chemists in your group analyze the liquid, they find that, although it consists of hydrogen and oxygen, the chemical bonds between the atoms are of a previously unknown type, and the liquid cannot be

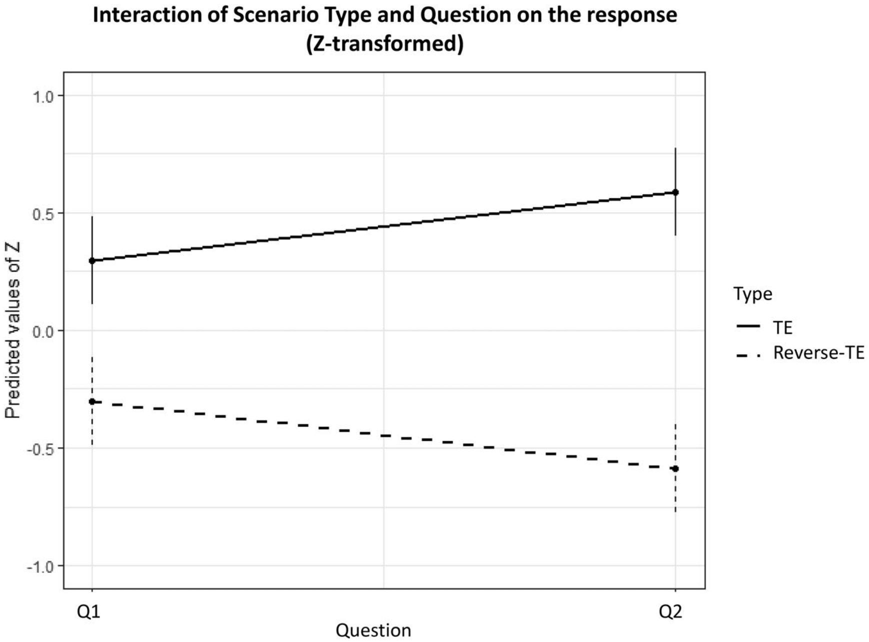

FIGURE 4 Interaction between *question type* (Q1 or Q2) and *scenario type* (TE or reverse-TE) in Experiment 2

> described as H_2O. The liquid is clear, odorless, tasteless, thirst-quenching, and supports life: In everyday circumstances it appears just as H_2O, but can easily be distinguished from it in the laboratory. The chemists have not found an explanation for why the chemical bonds are different from those found on Earth.
>
> [far:] The liquid evaporates from the surface of seas and lakes and falls down as rain, just like on Earth. However, when chemists in your group analyze the liquid, they find that its atoms do not contain oxygen, but a previously unknown element they call X. X behaves chemically very much like oxygen and bonds with hydrogen in the same way: The liquid consists of H_2X. The liquid is clear, odorless, tasteless, thirst-quenching, and supports life: In everyday circumstances H_2X appears just as H_2O, but can easily be distinguished from it in the laboratory.

The reverse-TE-cases for “water” were as follows:

> Imagine that you are a member of a large group of astronauts who have traveled to another solar system. The group is composed of experts in various fields. You have landed on a planet that has not been previously explored. You find, to your great excitement, that the planet contains a rich plant life. The planet also contains a number of seas, lakes, and rivers that contain a liquid that looks like this:
>
> [near:] [an image of a transparent, reddish liquid] The liquid evaporates from the surface of the seas and lakes and falls down as rain, just like on Earth. However, the liquid is slightly reddish and has a faint, fruity smell. When chemists in your group analyze the liquid, they find that it consists of H_2O. The color and smell are

> not due to any impurities in the liquid, even pure H_2O displays these properties on this planet. The chemists cannot come up with any explanation for this puzzling phenomenon.
>
> [far:] [an image of a transparent jelly-like substance] The liquid evaporates from the surface of the seas and lakes and falls down as rain, just like on Earth. However, the liquid is jelly-like: It both flows and evaporates much slower than water on Earth. When chemists in your group analyze the liquid, they find that it consists of H_2O. The jelly-like appearance is not due to any impurities in the liquid, even pure H_2O is jelly-like on this planet. The chemists cannot come up with any explanation for this puzzling phenomenon.

After this, as in Experiments 1 and 2, the subjects were presented with Question 1 (Q1), the elicited production task:

Question 1

> Right after having been informed of these discoveries, you are told to report on your findings about the liquid back to your base on Earth. Draft a report in your own words (no more than 2 lines):

Question 2 (Q2) provided the subject with two clearly formulated alternatives, where one is consistent with externalism, the other with internalism, with the order of the alternatives randomized. For example, Q2 for the TE-far-case for “water” was the following:

Question 2

> There is disagreement among your group on how to report your findings back to Earth, and you decide to take a vote. In the vote, you are given two options, (a) and (b):
> (a) We have found a new water-like substance on this planet, but it is not water: It is not H_2O, but H_2X. X is a new element very much like oxygen.
> (b) We have found water on this planet: It is not H_2O, but H_2X. X is a new element very much like oxygen.

Which do you think is a better description of your findings, (a) or (b)? Please answer by circling a number on the scale.

1	2	3	4	5	6	7
(a) is clearly better				(b) is clearly better		

In this case, (a) would represent the externalist answer and (b) the internalist answer. For the reverse-TE-near case for “water,” the question was as follows:

Question 2

> There is disagreement among your group on how to report your findings back to Earth, and you decide to take a vote. In the vote, you are given two options, (a) and (b):

(a) We have found a new water-like substance on this planet, but it is not water: It is reddish and has a slight fruity smell.
(b) We have found water on this planet: it is reddish and has a slightly fruity smell.

Which do you think is a better description of your findings, (a) or (b)? Please answer by circling a number on the scale.

1 2 3 4 5 6 7
(a) is clearly better (b) is clearly better

In this case, (a) is the internalist answer and (b) the externalist. (The order of the options was randomized.) After this, the subjects were presented with a third question ("deference"), trying to detect their disposition to defer to experts. Finally, the subjects were presented a fourth question ("imaginability") that asked how easy it was for them to imagine the depicted scenario, on a scale of 1 (very difficult to imagine) to 7 (very easy to imagine). These last two questions were identical with those in Experiments 1 and 2.

The answers to Q1 and Q2 were coded in the same way as in Experiments 1 and 2. The REI questionnaire was not used in this study due to time constraints.

3.2 | Participants

The participants ($N = 45$, 32 male) were science and engineering students from the Norwegian University of Science and Technology, with average age 20.56 ($SD = 1.74$). Eighty-seven percent reported Norwegian as their mother tongue. Based on their answers to Q1 they all appeared to be fluent, judged by a native speaker. Their self-reported knowledge of physics was 3.18 ($SD = .72$) and of chemistry 3.00 ($SD = .85$) (on a scale from 1 to 5 in both).

3.3 | Results

Inter-rater reliability for Q1 was ICC = .69 ($p < .001$, details same as before), final ratings were reached through consensus. Mean ratings to the four questions by scenario type (TE/reverse-TE) and distance (near/far) are summarized in Table 5.

Differences in the ratings between type and distance were examined with linear mixed effects models with one of the questions at a time as dependent variable, type and distance as

TABLE 5 Mean ratings (*SD*) to the questions depending on scenario type in Experiment 3

	TE		Reverse-TE	
	Near	Far	Near	Far
Q1	.17 (.65)	.36 (.58)	.17 (.83)	.045 (1.05)
Q2	1.35 (1.92)	2.23 (1.15)	.22 (2.00)	−.64 (2.26)
Deference	4.39 (1.50)	4.73 (1.96)	4.22 (1.31)	3.91 (1.57)
Imaginability	5.04 (1.52)	5.09 (1.66)	4.65 (1.75)	4.95 (1.62)

predictors, and kind and participant as random effects. ANOVA (Type III with Satterthwaite's method) on the lmer model was used to examine main effects and interactions. With Q1 as dependent variable, none of the effects were significant (*p*'s > .30). In Q2, there was an interaction of type and distance (F = 5.63, *p* = .020), as well as a main effect of type (F = 27.60, *p* < .001; Figure 5). For deference, all effects were nonsignificant (*p*'s > .35), except for a near-significant main effect of type (F = 3.49, *p* = .068). With respect to imaginability, all effects were nonsignificant (*p*'s > .13).

We did not examine differences between natural kinds, because the sample size did not allow examining three-way interactions.

3.3.1 | Consistency between Q1 and Q2

Data were prepared for consistency analysis in the same way as in Experiments 1 and 2. We used a model with the Z-score from Q1 and Q2 as dependent variable; type, question, and distance as predictors; and subject and kind as random effects. There was a main effect of type (F = 17.98, *p* < .001): TE-scenarios were answered more externalistically than reverse-TE scenarios. Additionally, there was an interaction between type and question (F = 7.80, *p* = .0060): Answers were in the same direction to both questions, but stronger in Q2 (see Figure 6). Finally, there was an interaction between type and distance (F = 5.38, *p* = .022), which was similar as in the previous analysis (see Figure 5).

4 | DISCUSSION

In this study, we conducted three experiments to probe whether lay speakers' language use and understanding are in accordance with semantic externalism or internalism. We had three main goals: (a) to investigate the relative weight of deep structure and appearance in semantic judgments; (b) to examine how consistent speakers are in their semantic judgments; and (c) to examine whether there is gradualness in semantic judgments depending on how similar a novel sample is to standard samples of a familiar natural kind. Additionally, we examined associations between thinking styles and semantic judgments. We found evidence of ambiguity between superficial and deep features in categorization, that speakers' judgments are internally relatively consistent, and that categorization judgments may be graded depending on how radically underlying structure or appearance deviated from the actual samples. We found tentative evidence that thinking style may be associated with semantic judgments. Overall, the results are in line with previous studies, where judgments have not been clear-cut between internalism and externalism (Braisby et al., 1996; Genone & Lombrozo, 2012; Jylkkä et al., 2009; Malt, 1994; Nichols et al., 2015; Tobia et al., 2019), or between superficial properties and deep structure (e.g., Gelman, 2003, 2004; Keil, 1989; Medin & Ortony, 1989; Rips, 2001). Instead, the results can be taken to converge with a causal homeostasis view of natural kind concepts, where deep and superficial properties are causally linked and cannot be completely separated (Hampton et al., 2007; Rehder & Kim, 2010). In the present study, our aim was to more directly assess the relative weights of superficial and deep features in philosophical scenarios, and to test whether deep structure is sufficient for category membership when there is conflict in superficial properties.

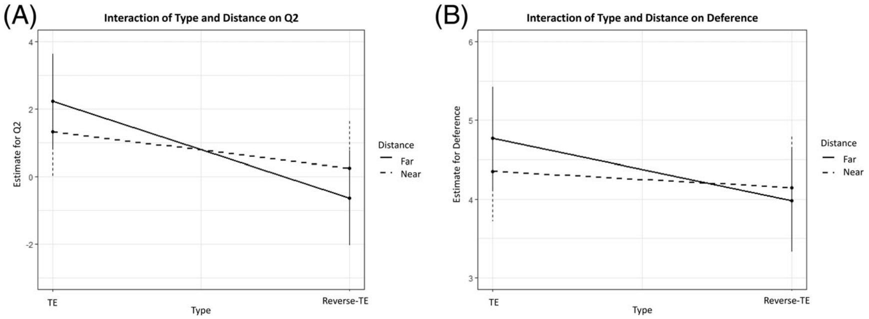

FIGURE 5 Ratings to Q2 and deference by distance and type in Experiment 3

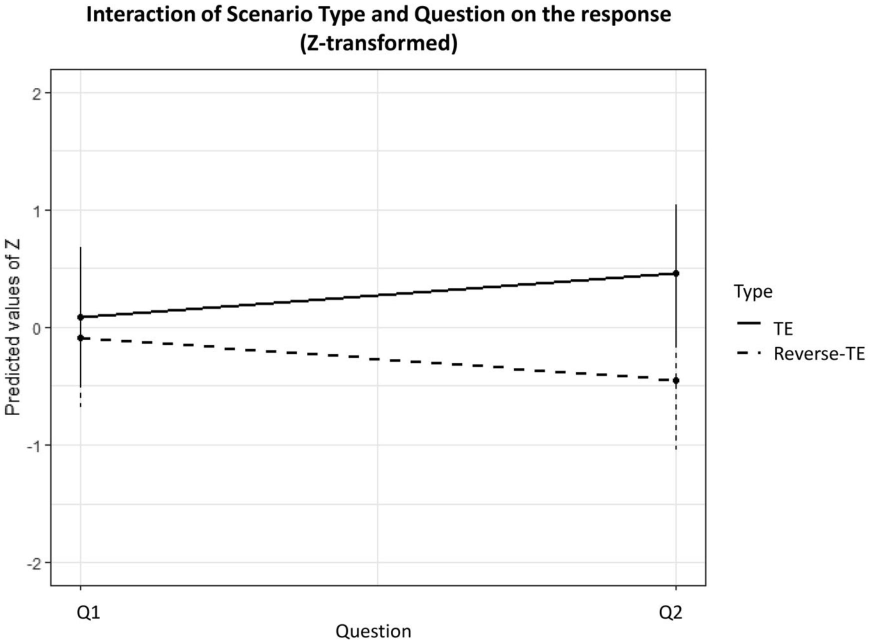

FIGURE 6 Interaction between scenario type and question in Experiment 3

The results of Experiments 1 and 2 are problematic both for traditional internalism and for mainstream externalism. When it comes to TE-cases, externalist judgments were the norm: XYZ is not water, and likewise for other natural kinds in structurally similar cases. This is, of course, welcome news to the externalist, and troublesome for the traditional internalist. However, the subjects' responses to reverse-TE-cases were directly in conflict with Kripke–Putnam essentialism: H_2O with a completely nonwatery appearance was consistently categorized as not being water, and likewise for other kinds. While there was minor variation depending on the kind used, the pattern was remarkably consistent across the full range of natural kinds. The results of Experiment 3 indicate that categorization judgments are gradual, in proportion to the degree of similarity between new samples and standard samples. This can be considered as casting doubt on a background assumption of both traditional externalism and internalism, namely that category membership is all-or-none. In what follows, we will discuss these findings and their implications in more detail.

4.1 | Experiments 1 and 2

In response to Q1, the elicited production task, the majority of subjects answered in a way that was neutral between externalism and internalism. In hindsight, this is perhaps not very surprising, given that the subjects were asked to report on the findings, based on a description of the scenario that had been carefully formulated so as to be neutral: Many subjects simply borrowed the neutral descriptions from the scenario description. Nonetheless, when subjects *did* deviate from the neutral expressions, the pattern was clear: For TE-cases, the subjects were unwilling to use the relevant natural kind term to describe the newly found samples or phenomena, indicating that they thought sharing underlying structure was necessary; for reverse-TE-cases, the subjects tended to describe the new samples as not belonging to the relevant natural kind term, indicating that they did not think sharing underlying structure was sufficient.

In response to Q2, the choice between two well-formulated alternatives, the pattern could hardly have been clearer. In response to TE-cases, the subjects consistently chose the externalist alternative as the better one, while for reverse-TE-cases, the internalist alternative was preferred. There was some variation depending on the kind term used: For "water" and "gold," the internalist alternative in reverse-TE-cases was not preferred as strongly as for "lightning," "tiger," and "diamond." Nonetheless, the effects were quite consistent: For all five kinds, TE-cases were judged in accordance with externalism, while reverse-TE-cases were not.

At the very least, this indicates that externalists have been far too hasty in implicitly assuming that an affirmative answer to the necessity question—which is here affirmed by the subjects' responses to TE cases—directly entails an affirmative answer to the sufficiency question.[12] If the data from Experiments 1 and 2 are taken at face value, a more substantial reevaluation of mainstream externalism is in order. On the basis of these results, it would be tempting to conclude that the subjects took both sharing underlying nature *and* sharing appearance as necessary and sufficient for belonging in the extension of a term. This would be too hasty, however. In Experiments 1 and 2, the underlying nature and the appearance of the new samples were either identical to or completely different from the standard samples. Hence, the results from these experiments indicate that neither a watery substance with a structure completely different from H_2O, nor a completely *non*-watery substance with structure H_2O, belong in the extension of "water," and similarly for other natural kind terms. Completely different underlying structure, and completely different appearance, are sufficient for excluding a sample from the extension of the natural kind term, but we do not yet know whether *any* difference along one of the dimensions is enough for exclusion, or whether nonidentity in one or the other respect can be tolerated, as long as the underlying structure or the appearance is to some degree *similar* to that found in the standard samples. Experiment 3 was conducted to make progress on this question and will be discussed later.

As explained earlier, we also attempted to find evidence of self-correction (through mismatches between answers to Q1 and Q2) and deference (through answers to deference). We did not find evidence of more self-correction in one direction rather than another; the only pattern that was found was that the participants' judgments tended to become stronger in Q2,

[12] This result is fully in line with Malt's tentative suggestion in her 1994 paper (Malt, 1994).

compared to Q1. This is probably due to the fact that, unlike in Q2, many of the answers to Q1 were neutral. When it comes to deference, no consistent differences between scenario types were found, indicating that the participants were equally certain about their judgments in both types of scenarios.

In Experiments 1 and 2 we also examined whether the participants' general thinking style, measured with the REI-10, would predict their use of natural kind terms. We hypothesized that internalistic judgments would be associated with experiental thinking, and externalistic judgments with rational thinking. These hypotheses received modest support: In Experiment 1, faith in intuition (hypothetically tapping on experiental thinking) significantly but weakly predicted internalist semantic judgments. Although the effect was weak, it was in line with our hypothesis, suggesting that general personality features or thinking styles could underlie semantic judgments. This raises similar questions as the study by Machery et al. (2004), which claimed to find cross-cultural differences in semantic intuitions concerning proper names: Whose intuitions should philosophers' theories of reference capture?[13]

In the present study, it is possible that we failed to discover robust associations between thinking styles and semantic intuitions because we used a very brief questionnaire of thinking styles, the validity of which could be questioned. Future studies should use more extended questionnaires, including personality measures such as the "international personality item pool" (Goldberg et al., 2006), and the full version of the rational–experiental questionnaire.

4.2 | Experiment 3

Experiment 3 was performed in order to get a more nuanced picture of the role of underlying structure and appearance in semantic categorization. The main finding was that the differences between TE and reverse-TE cases were stronger in the far-cases compared to the near-cases (as indicated by the significant interaction in Figure 5a). That is, the participants tended to categorize a K-like substance as belonging to kind K more strongly when the deep structure of the substance was similar to actual samples of K, compared to when it was more radically different. In reverse-TE cases, samples with very different appearance, but with the same deep structure as standard samples of K, were more likely to be categorized as non-K than samples with an appearance only slightly different from standard samples.

These findings present a challenge to traditional theories of reference. Both externalist and internalist theories have assumed that the extensions of our natural kind terms have sharp boundaries. The results of Experiment 3 put pressure on this assumption. Whether, and how, the results can be made consistent with an assumption of sharp boundaries will depend on one's background assumptions about metasemantics; this is a complex issue that we cannot hope to settle here. The dispositionalist view we find plausible for independent reasons (Cohnitz & Haukioja, 2013) suggests the following reasoning. The extensions of natural kind terms can have sharp boundaries if the speakers are, on closer examination, and through deference to experts, disposed to converge on one or the other pattern of response, also in the cases where samples are initially not classified as clearly belonging, or not belonging, to the extension of the term. Whether speakers *are* so disposed is clearly an empirical question. In our experiment, we tried to find evidence of such corrective and deferential behavior, but did not succeed. This could be because the subjects were not disposed to self-correct or defer, or because our

[13] For an excellent overview of the massive critical discussion that followed Machery et al. (2004), see Hansen (2015).

experimental setup was not able to detect such dispositions. At present, we simply do not know which is the case.

An alternative reaction is simply to take the data at face value, indicating that membership in the extension of a natural kind term depends on similarity with standard samples in a continuous manner, rather than being clear-cut. On this view, the extensions of natural kind terms would be directly dependent on speakers' patterns of judgment using the term, and such judgments would be seen as driven by the relevant natural kind concepts, understood as mental representations. Our findings could then be interpreted in line with a prototype or exemplar theory of concepts (see Margolis & Laurence, 1999). However, the theory would be quite different from traditional prototype theories, where the features associated with prototypes are thought of as directly empirically accessible. Our results suggest that our natural kind concepts are associated with prototypes involving both appearance properties and (placeholders of) underlying properties that are not directly accessible, and of which we may even be ignorant. On the other hand, our results could be interpreted as being consistent with a causal homeostasis account of natural kind concepts, where deep and superficial properties are taken to be causally linked (Hampton et al., 2007; Rehder & Kim, 2010). On this line of reasoning, the participants might consider a change in appearance as reflecting some unknown changes in deep structure. However, in doing so, the participants would have to ignore or not believe what we told them about the deep structure of the samples.

More work on these issues is needed. Our Experiment 3 only employed two natural kind terms, and some of the tentative patterns may be due to the way our scenarios were formulated. The choice of the properties to be varied in the near and far cases was here due to the experimenters' own pretheoretical conceptions of which appearance properties are central: that, for example, low viscosity is more central to being watery than color or smell. A more elaborate experimental setup would systematically vary different kinds of properties to find out which properties in fact are more central to a given natural kind term than others.

4.3 | General discussion

The results indicate that both standard mainstream externalism and traditional descriptivism (as exemplified by the common concept strategy) are mistaken, when it comes to the reference of natural kind terms. Mainstream externalism overemphasizes underlying structure and ignores appearance; traditional descriptivism overemphasizes appearance and ignores underlying structure. The results *could* potentially be accommodated both by causal *and* descriptivist theories of reference, but in both cases, substantial adjustments to the theory are needed.

It is widely accepted that a causal theory of reference for natural kind terms will, in any case, need to include some descriptive elements to deal with the *qua* problem (Devitt & Sterelny, 1987)—our results show that such theories should give descriptive elements a more prominent role in reference determination than previously assumed, to deal with the subjects' responses to reverse-TE-cases. The descriptive component in such theories should be thought to do more than merely select the appropriate *type* of kind named: The observable properties associated with the kind term also have to be, to some extent, satisfied by samples, in order for them to belong in the extension of the term. While this may be seen as a major concession to descriptivism, the resulting theory will nonetheless remain strongly externalist: Causal chains still have a central role to play in reference determination, and the standard externalist judgments concerning TE cases remain valid.

Likewise, a descriptivist theory of reference would need to incorporate elements of the causal–historical theory in the descriptive content of natural kind terms, to deal with the subjects' responses to TE-cases. The most straightforward way to do this would be to adopt causal descriptivism and/or rigidified descriptivism (à la Lewis, 1997); such a theory would then need to include observable properties as part of the descriptive content of natural kind terms, just as suggested earlier for the causal–historical theory. Another internalist alternative would be to adopt a cluster theory, assuming that the cluster of properties or descriptions associated with a natural kind term can gradually evolve to include microstructural descriptions, along the lines of Häggqvist and Wikforss (2015).

The graduality of judgments found in Experiment 3 presents an additional complication to traditional theories, both internalist and externalist. As noted earlier, whether this graduality can be made consistent with the assumption that natural kind terms have sharp boundaries depends on one's theoretical background assumptions concerning how reference is determined in general, and remains an open question. If it cannot, an even more radical reevaluation of theories of reference is needed. However, the results of Experiment 3 should be replicated in a larger sample, and with more natural kind terms, before the finding can be considered robust.

It is always possible for the proponent of a theory to try to explain away conflicting data as evidence of systematic error. However, given how clear the data reported above are, in particular when it comes to Q2 in Experiments 1 and 2, trying to dismiss either the internalist or the externalist answers as systematically erroneous strikes us as quite desperate, at least in the absence of a systematic theory of error that would justify such a move.

A general limitation of studies which utilize far-fetched scenarios with a discrepancy between superficial and deep properties is that such cases are often nomologically impossible. In reality, superficial properties are not detached from underlying deep structure, but instead causally determined by them. Accordingly, appearance is typically evidence about deep structure—in the actual world when we observe a liquid that appears to be water, the odds are that it truly is water. Such contingent facts can be reflected in the structure of our concepts and categorization behavior. There is evidence that natural kind concepts are *causal homeostasis concepts*, where different features are causally connected (Hampton et al., 2007; Rehder & Kim, 2010). In contrast, in philosophical thought experiments like the Twin Earth case the focus is on logical or conceptual possibility, which might make imagining such cases difficult, especially for nonphilosophers. Accordingly, we cannot completely rule out the possibility that the participants' apparently nonexternalistic responses reflect their *difficulty of imagining* such cases, or their *uncertainty* concerning whether a sample counts as water, instead of the semantics of their concepts. However, in all of our experiments the participants judged the scenarios to be relatively easy to imagine (on average, 5 on a scale from 1 [very difficult to imagine] to 7 [very easy to imagine]). If appearance and deep structure were closely tied on the conceptual level, as suggested by the causal homeostasis approach, imagining reverse Twin Earth cases should be very difficult.

5 | CONCLUSION

In this article, we have presented three experiments on ordinary speakers' use and understanding of natural kind terms. Our results suggest that both mainstream externalist and traditional internalist theories of reference are mistaken, or at the very least in need of substantial revision. Experiments 1 and 2 revealed that speakers take both appearance and underlying nature into account when categorizing new samples as either belonging, or not belonging, in the extension of natural kind terms. Experiment 3 suggests that speakers' judgments are gradual, and proportional to the degree of similarity the new samples have with respect to standard samples. Explaining and accommodating these results calls for further empirical and theoretical work.[14]

ORCID

Mons Nyquist https://orcid.org/0000-0001-5083-2233

REFERENCES

Braisby, N. (2001). Deference in categorization: Evidence for essentialism? In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Braisby, N. (2004). Deference and essentialism in the categorization of chemical kinds. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Braisby, N., Franks, B. & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, *59*(3), 247–274.

Cohnitz, D. (2015). The metaphilosophy of language. In J. Haukioja (Ed.), *Advances in experimental philosophy of language* (pp. 85–108). London: Bloomsbury Academic.

Cohnitz, D. & Haukioja, J. (2013). Meta-externalism vs. meta-internalism in the study of reference. *Australasian Journal of Philosophy*, *91*, 475–500.

Cohnitz, D. & Haukioja, J. (2020). Variation in natural kind concepts. In Å. Wikforss & T. Marques (Eds.), *Shifting concepts*. Oxford: Oxford University Press.

Crane, T. (1991). All the difference in the world. *The Philosophical Quarterly*, *41*, 1–25.

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

Devitt, M. & Sterelny, K. (1987). *Language and reality*. Cambridge, MA: MIT Press.

Epstein, S., Pacini, R., Denes-Raj, V. & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390–405. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/8765488

Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford: Oxford University Press.

Gelman, S. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*, 404–409.

Genone, J. & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, *25*, 717–742.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/J.JRP.2005.08.007

Häggqvist, S. & Wikforss, Å. (2015). Experimental semantics: The case of natural kind terms. In J. Haukioja (Ed.), *Advances in experimental philosophy of language* (pp. 109–138). London: Bloomsbury Academic.

Hampton, J. A., Estes, Z. & Simmons, S. (2007). Metamorphosis: Essence, appearance and behaviour in the categorization of natural kinds. *Memory & Cognition*, *35*, 1785–1800.

Hansen, N. (2015). Experimental philosophy of language. *Oxford handbooks online*. https://doi.org/10.1093/oxfordhb/9780199935314.013.53

Johnson, M. & Nado, J. E. (2014). Moderate intuitionism: A metasemantic account. In A. R. Booth & D. Rowbottom (Eds.), *Intuitions* (pp. 68–90). Oxford: Oxford University Press.

[14] This work was presented at the Society for Exact Philosophy and European Society for Philosophy and Psychology conferences in 2018—we want to thank the audiences at these events for their feedback. We are especially grateful to Nat Hansen, Sören Häggqvist, and Åsa Wikforss, as well as two anonymous referees for the journal, for valuable comments on earlier drafts of the paper. Jussi Jylkkä's work was supported by the Åbo Akademi University Endowment (the BrainTrain project and a personal grant.)

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

Keil, F. C. (1989). *Concepts, kinds and conceptual development*. Cambridge, MA: MIT Press.

Kripke, S. (1980). *Naming and necessity*. Oxford: Blackwell.

Leslie, S.-J. (2013). Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology*, *4*, 108–165.

Lewis, D. (1997). Naming the colours. *Australasian Journal of Philosophy*, *75*, 325–342.

Machery, E., Mallon, R., Nichols, S. & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1–B12. https://doi.org/10.1016/j.cognition.2003.10.003

Malt, B. (1994). Water is not H_2O. *Cognitive Psychology*, *27*, 41–70.

Margolis, E. & Laurence, S. (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.

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

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 Academic.

Medin, D. L. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Mellor, H. (1977). Natural kinds. *British Journal for the Philosophy of Science*, *28*, 299–312.

Nichols, S., Pinillos, A. & Mallon, R. (2015). Ambiguous reference. *Mind*, *125*, 145–175.

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

Rehder, B. & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(5), 1171–1206.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin*, *127*, 827–852.

Segal, G. (2000). *A slim book about narrow content*. Cambridge, MA: MIT Press.

Steward, H. (1990). Identity statements and the necessary a posteriori. *Journal of Philosophy*, *87*, 385–398.

Tobia, K. P., Newman, G. E. & Knobe, J. (2019). Water is and is not H_2O. *Mind & Language.* https://doi.org/10.1111/mila.12234

How to cite this article: Haukioja J, Nyquist M, Jylkkä J. Reports from Twin Earth: Both deep structure and appearance determine the reference of natural kind terms. *Mind & Language*. 2020;1–27. https://doi.org/10.1111/mila.12278