ELSEVIER

Cognition 59 (1996) 247–274

COGNITION

Essentialism, word use, and concepts

Nick Braisby[a], Bradley Franks[b,*], James Hampton[c]

[a]*Department of Psychology, London Guildhall University, Old Castle Street, London E1 7NT, UK*

[b]*Department of Psychology, London School of Economics, Houghton Street, London WC2A 2AE, UK*

[c]*Department of Psychology, City University, Northampton Square, London E1 7NT, UK*

Received 25 July 1994, final version accepted 29 September 1995

Abstract

The essentialist approach to word meaning has been used to undermine the fundamental assumptions of the cognitive psychology of concepts. Essentialism assumes that a word refers to a natural kind category in virtue of category members possessing essential properties. In support of this thesis, Kripke and Putnam deploy various intuitions concerning word use under circumstances in which discoveries about natural kinds are made. Although some studies employing counterfactual discoveries and related transformations appear to vindicate essentialism, we argue that the intuitions have not been investigated exhaustively. In particular, we argue that discoveries concerning the essential properties of whole categories (rather than simply of particular category members) are critical to the essentialist intuitions. The studies reported here examine such discovery contexts, and demonstrate that words and concepts are not used in accordance with essentialism. The results are, however, consistent with "representational change" views of concepts, which are broadly Fregean in their motivation. We conclude that since essentialism is not vindicated by ordinary word use, it fails to undermine the cognitive psychology of concepts.

1. Introduction

The essentialist approach to the meaning of natural kind terms has been one of the most influential theories to emerge in the borderline between the philosophy of language and the philosophy of mind.[1] The approach has

* Corresponding author. Tel.: +44 171 955 7704; fax: +44 171 955 7565; e-mail: B.Franks@lse.au.uk. Order of authors is alphabetical.

[1] "Essentialism" refers to the intuitions underlying rigid designation (e.g., Kripke, 1972).

SSDI 0010-0277(95)00698-2

profound implications not only for perennial philosophical problems relating to the nature of mind and meaning, but also for the cognitive psychology of concepts and language use. In terms of the latter, the approach has been taken to suggest that contemporary empirical approaches necessarily fail to investigate conceptual content, and so are flawed (e.g., Margolis, 1994; Rey, 1983, 1985, 1992). We suggest that unpacking the argumentation offered in favour of the essentialist approach leads to important empirical implications. Our aim is to evaluate the intuitions about word use that Kripke and Putnam deploy in support of essentialism, by putting them to an empirical test.

We first summarize the essentialist approach and its argumentation, and note the crucial role played by intuitions about ordinary language use. Potential problems with these intuitions then form the foundation for two empirical studies. Finally, we consider the implications of our results for the essentialist position.

2. The essentialist view

The essentialist view (Kripke, 1972, 1980; Putnam, 1975) holds that natural kind terms refer to categories, whose members share the same essential properties. It entails the following:

(1) Essential properties determine reference.
(2) Non-essential (or contingent) properties do not determine reference.
(3) Reference is determined independently of people's beliefs about which properties determine reference.

For example, consider the natural kind term *gold*. By 1, *gold* refers to anything possessing the essential (by hypothesis, atomic) structure of gold. In contrast, by 2, *gold* does not refer to things that only share superficial characteristics (such as colour and density). And since most people lack knowledge of gold's atomic structure, it follows that the reference of the term *gold* is independent of people's beliefs about its essence. This picture gives rise to Putnam's claim, "Cut the pie any way you like, 'meanings' just ain't in the *head*!" (Putnam, 1975b, p. 227): it is always possible that people are mistaken in what they take the essential properties to be. Such fallibility, however, does not imply fallibility of reference, since, by 1 and 2, reference is determined by essential properties, not properties mistakenly assumed to be essential. For example, though it is assumed that all cats are mammals, it is conceivable that one could discover that all cats are robots controlled from Mars. Despite *assuming* the essential properties for *cat* to include *being a mammal*, they actually include *being a robot*. According to the Kripke–Putnam view, in such a case *cat* never referred to a subset of mammals even though people may have thought it did. Rather, *cat* always

referred to the same category of objects: it is simply that these objects were once thought to be mammals but later discovered to be robots controlled from Mars.[2]

Essentialism's proposal that actual essences alone determine a word's reference has two potentially troubling consequences for the cognitive psychology of concepts. The first concerns whether concepts can be mental entities and still determine reference. Since, according to essentialism, agents may revise their beliefs concerning essential properties without this affecting reference, it follows that concepts, if considered traditionally as mental entities, do not possess reference-determining content. Alternatively, since whatever content determines reference must be non-mental, if concepts determine reference, they must be non-mental – an apparently paradoxical implication for the cognitive psychology of concepts. However, these are the lines of argument pursued by Rey (1983) and Margolis (1994).[3] The upshot appears to be that a tradition of empirical research on concepts at best requires radical reinterpretation, and at worst should be jettisoned.

The divergence between essentialism and cognitive psychology is further exemplified in their approaches toward the context-(in)sensitivity of concepts. Since essences are, by definition, invariant, essentialism critically entails that conceptual content is context-invariant. Essentialism implies that people who are ignorant of essences would not be governed by those essences in their use of words to classify referents. By contrast, people who are aware of essences, and who use a term conventionally (i.e., following the meaning of the term in the language) *will* use that term according to the essence, since the essence determines the extension (governing which descriptions of objects with the term are truthful), and thus is constitutive of the conventional meaning of the term. So all conventional uses of a term are determined by the essence. Putnam's discussion (Putnam, 1975b), in fact, goes further than this: he suggests that it is not only the conventional meaning of a term (what he refers to as the "predominant" sense for that term) that is given by the essence, but also many of the *non*-predominant or *non*-conventional senses also. This implies that any uses of a term not governed by the essence must be *non*-conventional and the only way that essentialism could allow for contextual variation in conceptual content would be by assuming that such variations are due to non-conventional uses. Although non-conventional uses of terms are widespread, it is not clear that the demarcation of the conventional from the non-conventional in terms of being governed by essence or not, would map onto any widely held

[2] Essentialism makes predictions about word use before and after a discovery about the essence, but not about the intervening period when beliefs about essences are changing. Although the question arises as to what occurs during this period, it is unclear what criteria people would employ in their use of a term when the essence is subject to such doubt.

[3] This is the externalist thesis about mental content: external, real-world facts are partly constitutive of the nature of thought (see McGinn, 1989). Whereas essentialism implies externalism, essentialism is not necessary for externalism (e.g., Burge, 1979).

linguistic/pragmatic view of that distinction (see, for example, Bach, 1987; Sperber & Wilson, 1986).

By contrast, the cognitive psychology tradition treats concepts as mental entities and assumes a close relation between concepts and reference (but for dissenting opinions, see Malt, 1991, and Barsalou, Yeh, Luka, Olseth, Mix, & Wu, 1993). For example, Murphy (1991) suggests that concepts can be equated with Fodor's (1987) notion of narrow content and, moreover that the meaning of a word is a concept. Also, recent work in cognitive psychology has been taken to demonstrate the context sensitivity of conceptual content. This interpretation is supported by shifts in the use of words in classification, both within and between individuals, and in shifts in typicality (e.g., Barsalou & Sewell, 1984; Clark, 1983; Hampton, 1988; Medin & Shoben, 1988; Roth & Shoben, 1983). A more radical context sensitivity is evidenced by apparent contradictions in individuals' judgements (Hampton, 1982; McCloskey & Glucksberg, 1978), which are hallmarks of a word having different senses (Evans, 1982; Frege, 1952/1892). Evans, for example, explicitly links different senses/conceptual contents to different patterns of assent/dissent in conventional word use. Indeed, such variation in conceptual content is a central feature of (at least) two accounts of concepts: sense generation (Braisby, 1990, 1993, 1994; Braisby, Franks, & Myers, 1992; Franks, 1991, 1995a; Franks & Braisby, 1990) and "generalised" prototype theory, in which context can alter the settings of thresholds and feature weights (Hampton, 1988, 1992). Such an approach is also clearly articulated in the theoretical work of Barsalou (1987, 1993). Henceforth, we refer to such broadly Fregean theories as "representational change" theories.[4]

In sum, the treatment of concepts within psychology is doubly problematic if essentialism is true. First, either concepts are not "in the head" yet determine reference, or are "in the head" and do not determine reference: within psychology it has typically been assumed that they are both "in the head" and that they determine reference. Second, since essentialism requires word meaning to be context-insensitive, a very different account must be given of psychological demonstrations of context-sensitivity of conceptual content.

3. Essentialist intuitions about word use

The essentialist argument involves what has been labeled a "modal step" by Bealer (1987), because it moves from discoveries of facts about the

[4] Representational change theories, however, do not posit ambiguity for apparently unambiguous natural kind terms. Nor are the different contents for a kind term represented, say, as a list in a mental lexicon, with context selecting senses appropriately (the problematic "sense selection assumption" – Clark, 1983; Franks & Braisby, 1990).

properties of a natural kind, to a claim that those facts are *necessary*, thus strengthening the knowledge claim about the properties of the natural kind. Bealer also discusses the pivotal role of intuitions in justifying the modal step. He notes Putnam's assertion that, for example, if beliefs about the macroscopic (ex hypothesi, non-essential) properties of water are discovered to be false, then that sample will remain water. The *intuition* of the constancy of kind identity under such changes is critical to the argument for essentialism. As Kripke notes, "I think it [intuition] is a very heavy voice in favour of anything, myself. I really don't know, in a way, what more conclusive evidence one can have" (Kripke, 1980, p. 42). Rey echoes this point: "Kripke and Putnam argue for their view primarily on the basis of ordinary linguistic intuitions regarding 'natural kind terms'" (Rey, 1983, p. 252). Bealer puts it more starkly: "Without these intuitions Putnam would have no argument" (Bealer, 1987, p. 302). Four such essentialist intuitions can be discerned in the writings of Kripke and Putnam.

Intuition A: general essentialist intuition: This involves circumstances were agents are aware (or believe that they are aware) of the category's essence. In such circumstances, the intuition is that agents will use that essence as the criterion for the conventional application of the term. So classification of entities as members or non-members of the kind category will be determined according to their possession of the essence.

Intuition B: essential properties of a kind: This concerns a counterfactual circumstance that contains Intuition A as a component. This is a circumstance in which (a) the agent is aware of the essence of the kind, and (b) that essence is discovered to be radically different from the essence that the agent had previously presumed the kind to have. Such discoveries include changes from one natural kind to another (e.g., from H_2O to XYZ) and changes from a natural kind to an artefact category (e.g., from cats' being mammals to their being robots controlled from Mars). The intuition is illustrated by Putnam's scenario in which it is discovered that all cats are robots controlled from Mars. It would not follow from this discovery, he argues, that people would say *cats do not exist*, or *there are no cats*. Rather, it is merely that cats are not as they had previously been thought. As he notes: "we will still *call* them 'cats'", and, moreover, "not only will we still *call* them *cats*, they are cats" (Putnam, 1975a, p. 143). Kripke offers a similar example involving gold: discovering that gold is not really a metal would not entail that people would say *gold does not exist*, or that *there is no such thing as gold*. The appeal to a common intuition about widespread word use subsequent to such discoveries is evident: Kripke asks, "would there on this basis be an announcement in the newspapers:' 'it has turned out that there is no gold. Gold does not exist. What we took to be gold is in fact not gold?'" and answers, "it seems to me there would be no such announcement" (Kripke, 1972, p. 316).

Intuition C: non-essential properties of a kind: This concerns circum-

stances where discoveries are made about the non-essential properties of a kind. Here, as noted above, the intuition is that such a discovery has no effect on ordinary language use concerning the kind.

Intuition D: essential properties of an individual: The fourth intuition concerns circumstances where it is discovered that a particular individual, contrary to prior belief, does not possess the essence associated with a particular kind. The essentialist intuition here is that language users should accept the truth of statements to the effect that the individual is not (and has never been) a member of the kind in question.

Thus language users are presumed to share these essentialist intuitions and, once aware of the facts, use terms accordingly. However, the intuitions may be disputed from two angles. One concerns the evidential status of intuitions as a means of making the modal step: Bealer (1987) suggests that no traditional characterisation of the role and nature of intuitions is consistent with essentialism. However, regardless of whether intuitions are *in general* convincing as support for the modal step, one can also challenge the acceptability of the *particular* intuitions deployed by Kripke and Putnam – this is the line we follow.

Intuition A presupposes that, as Kripke claims, "One might discover essence empirically' (Kripke, 1972, p. 322). Moreover, he asserts that "In general, science attempts, by investigating basic structural traits, to find the nature, and thus the essence ... of the kind" (Kripke, 1980, p. 138). Canfield (1983), Shapere (1982) and Nersessian (1984) suggest that this view mis-characterises the nature of scientific practice. For example, Shapere argues, "it is not just one property or set of properties – the essential ones – that determines ... how scientists will apply terms in new situations; all the (true) properties may ... play a role, and ... the properties and behaviour of other entities ... may also play a role" (Shapere, 1982, p. 7). That is, after discovering an essence (supposing it could be discovered by science), going on actual current practice, scientists would not henceforward restrict themselves to applying the natural kind term only to entities that possessed that essence. We could only apply a kind term rigidly on the basis of the kind's essence if we had knowledge that justifies the modal step. Without such modally justified knowledge, an essentialist application of a term would amount to no more than dogmatism. In sum, "the alleged linguistic practice ... [that is defining of essentialism] ... neither would, nor should nor does take place" (Shapere, 1982, p. 4). Thus, Intuition A may not be in line with empirical observations of scientific activity; part of our goal is to investigate whether it is equally out of line with non-scientists' word use. If essentialism places too stringent a constraint on scientific use of language, one might reasonably suspect that it is also too stringent a requirement for ordinary, non-scientific use of natural kind terms (as argued by Donnellan, 1983).

Arguments against Intuition B (where agents are aware of the discovery

that essences are different from their previously assumed nature) are provided by Canfield (1983). He allows that, in a case where the essence of tigers is found to differ from that previously presumed, people *might* concur with Kripke and Putnam, that the entities with the "new" essence are still tigers, but we were in fact previously wrong about what constituted tigerhood. However, he claims that one could also say that "it turns out that there are no tigers" (Canfield, 1983, p. 115), and explicate this by saying that "The tigers of the jungle that were feared as man-eaters, that Blake wrote about, and so on, do not exist; it was all a quite incredible illusion." We might view this claim that there are no tigers as a claim about tigers qua four-legged, striped felines, say, and not a claim about the particulars so described. In a vocabulary slightly different from Canfield's: interpreting *tiger* in terms of the set of particulars (what we will refer to as a "particularist" reading), *tigers do not exist* is false; by contrast, interpreting tiger in terms of the properties associated with being a tiger, such as *four-legged*, *striped*, *feline*,... (what we will refer to as "property" reading), *tigers do not exist* is true. Hence, the statement of the Kripke–Putnam intuition (i.e., that *tigers do exist*) is true on a particularist reading, but false for the property reading relating to the "old" presumed essence. Canfield's over-all point is that *either* statement would be acceptable, depending upon the reading taken of the kind term. So a rational agent could by turns dissent from or assent to the statement *tigers do not exist*, because that agent could interpret the kind term differently (or, possibly, interpret the existence claim differently). Such context sensitivity of interpretation of a conventional use of a kind term, in knowledge of what is taken to be the essence of the kind, is clearly at odds with essentialism's dictates.

4. Empirical investigations of essentialist intuitions about word use

Some empirical studies appear to support essentialist intuitions about word use. Rips' (1989) investigation of the relationship between similarity and category membership judgements suggests an independence between the two, with essential properties determining membership, and non-essential properties determining similarity. Similarly, Keil (1986) found that, where an individual is discovered to lack the non-essential properties of the category, classification behaviour did not change, but where the individual is discovered to lack essential properties, it did change (but see Hampton (1995), for conflicting evidence).

Malt (1994), however, found that subjects used *water* to talk about substances that contain only a relatively small amount of H_2O (e.g., puddle water), or even substances that were not aqueous (e.g., oil). Her interpretation is that the extension of *water* is not circumscribed by the possession of an essence (H_2O): rather, it is more flexible and strongly influenced by

socio-historical factors. Such findings, however, might be argued to reflect non-conventional word use (e.g., using *water* to talk about oil), and so may not relate directly to essentialism. Moreover, in examining only actual word use rather than counterfactual scenarios, the essentialist intuitions are not directly evaluated. Also, the counterfactual scenarios used by Rips and Keil concerned discoveries and transformations of properties of individuals rather than whole kinds. But scenarios concerning discoveries about whole kinds are of crucial importance to a full test of essentialist intuitions (cf. Intuitions B and C).

The essentialist intuitions correspond to four types of counterfactual scenario, as follows (implications of essentialism for word use are indicated in parentheses):

Essential category: Presumed *Essential* properties of a category are discovered not to be the actual essential properties for that *Category* (from 3, word use should not change). For example, discovering that *all* individuals thought to be cats are, in fact, robots, should *not* change use of *cat* with respect to these individuals.

Non-essential category: *Non-essential* properties of a category are discovered not to be true of that *Category* (from entailment 2 above, word use should not change). For example, discovering that *all* cats do not purr, but that the purring is made by parasites in their fur, should *not* induce changes in the use of *cat* with respect to the category of cats.

Essential individual: *Essential* properties of a category are discovered not to be true of an *Individual* presumed to be a category member (from 1, word use should change). For example, the discovery that, though all cats are mammals, one individual once through to be a cat is really a robot, *should* change the way that it is classified as a *cat*.

Non-essential individual: *Non-essential* properties of a category are discovered not to be true of an *Individual* presumed to be a category member (from 2, word use should not change). For example, discovering that, though cats purr, an individual presumed to be a cat cannot purr, should *not* change use of *cat* with respect to that individual.

This study examines whether essentialist intuitions are vindicated under the critical Essential category and Non-essential category scenarios, scenarios that have not previously been used: Rips and Keil used Essential individual and Non-essential individual scenarios.[5] Moreover, to measure

[5] Non-Essential Individual scenarios are not considered, since these scenarios appear to have yielded results clearly consistent with essentialism (in the work of Keil and Rips).

subjects' intuitions more sensitively, we introduced different types of statement with which they could agree or disagree. Keil and Rips used single classification judgements: Keil's subjects gave binary judgements (e.g., whether a horse painted black and white was a *horse* or a *zebra*) and Rips' subjects rated times on a 10-point scale (e.g., with *bird* at one end and *insect* at the other), indicating to which category the item was most likely to belong. Judging *likely* kind membership, however, assumes that agents are not aware of the facts about essences, and so again is only indirectly related to the essentialist intuitions.

Judgements of truth and falsity were obtained for three different statement types (Existential, Qualified and Membership) that tap essentialist intuitions in different ways. Existential statements exactly reflect the wording used by Kripke and Putnam in articulating the essential intuitions, and are of the type *Xs exist*, where *X* labels a natural kind.

As noted in connection with Canfield's arguments, Existential statements can be ambiguous between property and particularist interpretations. Thus, Existentials may be judged *in accordance with* essentialism without this implying essentialist belief. For example, consider scenario Essential category and *cats do exist*. On a particularist reading of *cat* (where *cat* refers solely to the set of objects described in the scenario), then the statement is true: the set of particulars continues to exist after the discovery. However, a response of True need not reflect an essentialist belief, for essentialism would hold not just that these particulars exist but *crucially* that they continue to be members of the same natural kind. To avoid particularist interpretations, we included Qualified statements which make explicit reference to the set of objects referred to in the scenario. For example, one should not be able to judge the Qualified statement *there are no such things as cats, only robots controlled from Mars* to be false simply by taking a particularist interpretation of *cat*, since the sentence itself militates against such an interpretation.

Membership statements reflect the view that essentialist beliefs may have little to do with *existence* but whether entities are judged to be members of the same category before and after a discovery. Hence, concerning a putative feline Tibby, Membership statements were of the form, *Tibby is a cat, though we were wrong about her being a mammal*. Our claim is that essentialist belief is best measured by Membership statements, these being better indicators of essentialism than Qualified, which are better than Existentials.

Further, positive (+) and negative (−) variants of each statement type were used to examine whether subjects make apparent self-contradictions (contra essentialism, but consistent with representational change): if *cat* has different senses, then sentences such as *cats exist* and *cats do not exist* may be given the same truth evaluation; if it has one sense only, the positive and negative forms of the sentence must receive different truth evaluations.

EXPERIMENT 1

5. Method

5.1. Subjects

Subjects were 28 undergraduates from the London School of Economics and City University, London. All had English as their first language.

5.2. Design

The experiment used three counterfactual scenario types: Non-essential category, Essential individual and Essential category. Seven natural kind categories of the kind which figure prominently in the argumentation of Kripke and Putnam were used: cat, water, tiger, gold, bronze, lemon and oak. Examples of the scenarios follow.

Non-essential category: You have a female pet cat named Tibby whom you believe to be able to miaow. Indeed, she seems to miaow quite loudly. However, in spite of the fact that most people assume that cats can miaow, scientists have discovered that the noise of miaowing is actually created by small parasites which live in their fur and that, in fact, *no* cats can actually miaow. Indeed, Tibby, once thoroughly cleared of parasites, cannot miaow.
Essential individual: You have a female pet cat named Tibby who has been rather unwell of late. Although cats are known to be mammals, the vet, on examining Tibby carefully, finds that she is, in fact, a robot controlled from Mars.
Essential category: You have a female pet cat named Tibby. For many years people have assumed cats to be mammals. However, scientists have recently discovered that they are *all*, in fact, robots controlled from Mars. Upon close examination, you discover that Tibby too is a robot, just as the scientists suggest.

Examples of the statements, which subjects had to judge as true or false follow:

Existential (+): Cats do exist
Existential (−): Cats do not exist
Qualified (+): Cats do exist, and people's beliefs concerning cats have changed
Qualified (−): There are no such things as cats, only robots controlled from Mars
Membership (+): Tibby is a cat, though we were wrong about her being a mammal

Membership (−): Tibby is not a cat, though she is a robot controlled from Mars

5.3. Hypotheses

We predicted agreement with essentialism in Non-essential category and Essential individual conditions, but less agreement in the Essential category condition. We also predicted that Existential statements would show greater (apparent) agreement with essentialism than Membership statements, with Qualified statements intermediate.

5.4. Procedure

Subjects were given a scenario and a set of statements as a example. For each category, subjects were then presented with the scenario types (21 scenarios in all: 3 scenario types × 7 natural kinds). Scenarios were presented in a fixed order. The order of statements was random for each scenario. The design incorporates 126 statements (21 scenarios × 6 statement types). Each subject judged all statements as True or False under the instruction that they were to assume that the described scenario was true.

6. Results

Responses were coded 1 for True, 0 for False. There were two main aspects to our analysis: the first concerned the over-all response patterns and their level of agreement with essentialism; the second concerned the number of response patterns which contain apparent self-contradictions. Table 1 contains the over-all response pattern, and that predicted by essentialism. Table 2 contains the responses recoded to indicate agreement with essentialism. All analysis was performed on these recoded scores, so significant

Table 1
Over-all response patterns in Experiment 1 and the predictions made by essentialism (E)

	Scenario type					
	Non-essential category		Essential individual		Essential category	
Statement	E	Actual	E	Actual	E	Actual
Existential (+)	7	6.96	7	6.79	7	6.43
Existential (−)	0	0.21	0	0.32	0	0.79
Qualified (+)	7	6.43	0	2.75	7	5.82
Qualified (−)	0	0.36	0	0.61	0	2.14
Membership (+)	7	6.36	0	3.43	7	6.07
Membership (−)	0	0.54	7	5.32	0	2.63

Table 2
Response (shown in Table 1) recoded to indicate agreement with essentialism (7 = max. agreement, 0 = no agreement)

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential (+)	6.96	6.79	6.43
Existential (−)	6.79	6.68	6.21
Qualified (+)	6.43	4.25	5.82
Qualified (−)	6.64	6.39	4.86
Membership (+)	6.36	3.57	6.07
Membership (−)	6.46	5.32	4.37

differences indicate significant differences in terms of agreement with essentialism.

There was a marginal effect of natural kind (Friedman, $\chi^2 = 11.80$, $df = 6$, $p = .07$). This did not interact with other factors and was absent in Experiment 2 (see below), so no interpretation is offered. As can be seen from Tables 1 and 2, essentialism appears vindicated only under certain combinations of scenario and statement type. In particular, essentialism fares poorly in Essential category and Essential individual scenario types under Membership statements. There was a strong effect of scenario type (Friedman $\chi^2 = 26.96$, $df = 2$; $p < .0001$): multiple comparisons indicated that the Non-essential category scenario type showed greater agreement with essentialism than the other two scenario types ($p < .025$), these not differing from each other (Essential category mean rank = 1.56, Essential individual = 1.63, Non-essential category = 2.81). There was also an effect of statement type (Friedman $\chi^2 = 56.34$, $df = 5$; $p < .0000$): multiple comparisons reveal an over-all difference between Existential, and Qualified and Membership statements in their degree of agreement with essentialism ($p < .025$): Existential statements showed greatest agreement, Membership showed least, and Qualified were intermediate (Membership mean rank = 2.40, Qualified = 3.14, Existential = 4.95). In general, then, our predictions concerning agreement with essentialism were supported.

The other strand to our analysis concerns the occurrence of apparent self-contradictions as indicated in Table 3. A contradiction was said to occur when a subject gave either False responses or True responses to both the positive and negative forms of a statement, within the same scenario (e.g., responding False to *Water does exist* and *to Water does not exist* within a single scenario). In such cases, subjects would be making a logical contradiction, unless the kind term had different senses for the two forms of the statement. In the following analysis we omitted Qualified statements in Essential individual scenario types since only in this combination could subjects respond False or True to both forms of the statement type, without

Table 3
Number and proportion of apparent contradictions obtained in Experiment 1

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential	2% (4)	2% (4)	4% (7)
Qualified	12% (24)	N/A	33% (64)
Membership	13% (25)	34% (66)	31% (60)

using different senses for the kind term. Since this cell contains the largest number of contradictions, our analysis errs in favour of essentialism.

Table 3 shows that most contradictions were obtained for Membership statements under Essential individual and Essential category scenario types and for Qualified statements under Essential category scenario types. In these conditions, around one-third of responses were apparent self-contradictions. An analysis of variance was performed on the number of self-contradictions under the types of scenario and statement. This showed a main effect of scenario type, $F(2, 54) = 132.37$; $p < .0001$, with Non-essential category scenarios showing fewer contradictions than Essential individual scenarios, which showed fewer contradictions than Essential category scenarios (mean number of contradictions, in that order, were 0.64, 1.32 and 1.61). There was also a main effect of statement type, $F(2, 54) = 112.28$; $p < .0001$, with Existential statements showing fewer contradictions than Qualified statements, which showed fewer than Membership statements (mean number of contradictions, in that order, were 0.21, 1.58 and 1.86). Additionally, there was an interaction between scenario and statement type, $F(3, 93)$ (adj.) $= 181.28$; $p < .0001$: as noted, there was a general trend over the statement types (Existential < Qualified < Membership statements), but this was less marked in scenarios of type Non-essential category than in the other two scenarios. Further, the general trend over the scenario types (Non-essential category < Essential individual < Essential category) was less marked for Existential statements than for the other statement types. This confirms that only certain combinations of scenario and statement type result in substantial levels of contradiction.

7. Discussion

The results were generally in line with our predictions. There are main effects of both scenario type and statement type on degree of agreement with essentialism. We predicted that the Essential category scenario type would show less agreement with essentialism that the other scenario types: the other scenario types had, according to previous studies, already lent support to essentialism. The results were only slightly different from the

prediction. The Essential category and Essential individual conditions showed significantly less agreement with essentialism than did the Non-essential individual condition. However, essentialism holds that there is no significant difference between scenario types. Again, contra essentialism, but as we predicted, Qualified statements elicited less agreement with essentialism than Existentials, while Membership statements elicited less agreement again.

The analysis of apparent contradictions also illustrates some of the above effects. Again, we found that the proportion of contradictory responses was lower with Existential than with either Qualified or Membership statements. Also, the proportion was lower with the Non-essential category scenarios than with the other two scenarios. Not only were no contradictions predicted by essentialism, but the observed proportion of contradictions underestimates inconsistency with essentialism for the following reason. There are four possible patterns of response for a pair of Existential statements: The pattern predicted by essentialism (Existential (+) = True, and Existential (−) = False), one pattern of apparent contradiction (Existential (+) = True, and Existential (−) = True), the other pattern of apparent contradiction (Existential (+) = False, and Existential (−) = False) and the opposite pattern to that predicted by essentialism (Existential (+) = False, and Existential (−) = True). Since Table 3 only indicates the cases of apparent contradiction, it *does not* indicate the total amount of disagreement with essentialism. Table 4 gives the percentage of responses which conform to essentialism, adjusted according to the above pattern.

Tables 3 and 4 show that essentialism fares poorly for Membership statements under Essential individual and Essential category scenarios. Agreement with essentialism, in these conditions, ranges from 46% to 58%. With qualified statements, agreement ranges from 58% to 60% (whereas for Existential statements agreement ranges from 88% to 95%). Although the lowest of these is above the chance level (25% – the predicted response being one out of four possible response patterns), approximately 40–54% of responses are incompatible with essentialism. Table 3 shows that with Membership statements approximately one-third of responses are cases of apparent contradiction. These findings also contradict those of Keil (1986)

Table 4
Percentage and number (in parentheses) of response patterns in Experiment 1 that conform to the predictions of essentialism

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential	97% (190)	95% (186)	88% (173)
Qualified	87% (171)	58% (114)	60% (118)
Membership	85% (167)	46% (90)	58% (114)

and Rips (1989) which appeared to support essentialism under Essential individual scenarios. Under the same scenarios, but critically using Membership statements, we find 54% of responses disagreeing with essentialism. While all these findings contradict essentialism, they are consistent with representational change views of concepts.

However, we also found a marginal effect of natural kind. Since there is no reason to suppose that natural kinds vary in conforming with essentialism, we suspect that this effect was an artefact of the experimental design. Analyses omitting the first natural kind (in all cases, *cat*) yielded the same pattern of results, except that the effect of natural kind was eliminated. Experiment 2 corrected this design flaw, aimed to replicate the results of Experiment 1, and sought to obtain further information about subjects' difficulty in carrying out the task, and their awareness of their strategies for classification.

EXPERIMENT 2

8. Method

8.1. Subjects

Subjects were 37 undergraduates from the London School of Economics and City University, London. All had English as their first language.

8.2. Design

The basic design of Experiment 2 was repeated under the same hypotheses.

8.3. Procedure

The procedure for Experiment 2 was identical to that of Experiment 1, except that the order of all scenarios was varied randomly across subjects. In this way, we hoped to ascertain whether the difference between natural kinds in Experiment 1 was due to order of presentation of natural kinds, or to some substantive difference between *cat* and the other kinds. Further, two practice items was given to subjects in order to minimise further the possibility of subjects finding the task difficult in the initial items. Subjects were then presented with 21 scenarios in all (3 scenario types × 7 natural kind categories). Scenarios were presented in random order, and order of statements was random for each scenario.

Additional information was also collected. One question asked subjects to rate how difficult they found the task, on a 7-point scale (since "such

counterintuitive situations are notoriously hard to judge": Atran, 1989, p. 12); if they found it difficult, this might explain inconsistencies in their responses. Subjects also answered three new questions: two (Possibility questions) concerned the acceptability of making contradictory truth evaluations of statements. They were asked whether it was possible to be both consistent and give contradictory responses to the positive and negative forms of Existential and Membership statements. For example, in the Existential case, subjects were asked:

"Consider the following pair of sentences:

Cats do exist. Cats do not exist.

Do you think that it is possible for someone to be both consistent in their responses and to say 'True' to the first statement and 'True' to the second (or 'False' to the first and 'False' to the second)?"

A final, Consistency question, asked whether subjects tried to make their truth evaluations consistent within scenarios: responses were Always, Sometimes, and Never.

9. Results

As before, responses were coded 1 for True, 0 for False. Analysis of Experiment 2 had four different aspects: degree of agreement with essentialism and proportion of apparent contradictions obtained (as in Experiment 1), and the interaction of the apparent contradictions with the Consistency and Possibility questions. Table 5 contains responses recoded to indicate agreement with essentialism. All analysis was performed on these recoded scores, so different scores indicate differences in agreement with essentialism.

As expected, there was no effect of natural kind (Friedman, $\chi^2 = 4.70$,

Table 5
Responses obtained in Experiment 2 recoded so as to indicate agreement with essentialism (cf. Table 2)

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential (+)	6.89	6.41	6.19
Existential (−)	6.92	6.30	6.30
Qualified (+)	6.35	3.76	5.84
Qualified (−)	6.78	6.00	5.70
Membership (+)	6.57	3.43	6.30
Membership (−)	6.51	4.49	5.38

$df = 6$, $p = .58$). All other effects were as predicted. There was an effect of scenario type (Friedman $\chi^2 = 33.95$, $df = 2$; $p < .0001$): multiple comparisons showed the following ordering in terms of agreement with essentialism: Non-essential category >Essential category > Essential individual ($p < .025$: Essential category mean rank = 2.05, Essential individual = 1.30, Non-essential category = 2.65). There was also an effect of statement type (Friedman $\chi^2 = 73.29$, $df = 5$; $p < .0001$); multiple comparisons reveal a difference between Existential, and Qualified and Membership statements in their agreement with essentialism ($p < .025$): the ordering was, Existential > Qualified > Membership (Membership mean rank = 2.42, Qualified = 3.27, Existential = 4.81). Experiment 2, then, replicated the findings concerning agreement with essentialism found in experiment 1.

The second aspect of our analysis concerns contradictions (see Table 6): there were fewer, but in the same pattern as Experiment 1. Most occurred for Membership statements under Essential individual and Essential category scenarios and for Qualified statements under Essential category scenarios. An analysis of variance on the number of contradictions revealed a main effect of scenario type, $F(2, 72) = 227.80$; $p < .0001$, with the ordering as before (Non-essential category < Essential individual < Essential category, with means, in that order, 0.51, 0.82 and 0.91). There was also a main effect of statement type, $F(2, 72) = 183.15$; $p < .0001$, with Membership producing most contradictions, Existential producing least, and Qualified being intermediate (means, in that order, 1.07, 0.23 and 1.04). There was also an interaction between scenario and statement type, $F(2, 73)$ (adj.) = 258.31; $p < .0001$, with only certain combinations resulting in substantial levels of contradiction.

The third aspect of our analysis concerned the Possibility questions. Concerning whether one can make contradictory truth evaluations to Membership statements, 10 subjects said this was possible, 27 said it was not. An analysis of variance examining the effect of response to this question on the number of contradictions observed showed no main effect of response to this question. Further, there were no interactions between response to this question and scenario or statement type.

Concerning the possibility of giving contradictory truth evaluations to Existentials, 8 subjects said this was possible, 29 said it was not. In

Table 6
Number of apparent contradictions obtained in Experiment 2 by scenario and statement type

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential	1% (3)	4% (10)	7% (17)
Qualified	12% (30)	N/A	19% (49)
Membership	9% (24)	20% (51)	15% (40)

conjunction with the corresponding finding for Membership statements, this suggests that most subjects did not think it was possible to be consistent and give apparently contradictory truth evaluations. An analysis of variance examining the effect of response to this question on the number of contradictions observed yielded the same pattern of results as for the Membership possibility question. Thus the over-all *pattern* of contradictions did not depend on responses given to either of the Possibility questions. Finally, *t*-tests indicated that the *total* number of contradictions made was not affected by type of response made to either of these Possibility questions.

The next aspect of our analysis concerns responses to the Consistency question: 17 subjects said Always, 5 said Never, 15 said Sometimes. Thus most subjects attempted to be consistent in their responding, for at least some of the time. An analysis of variance examined whether the number of observed contradictions varied according to response to this question. There was no main effect of response to this question on the number of contradictions, nor were there any interactions between response to this question and scenario or statement type. In addition, an analysis of variance revealed that the total number of contradictions made was not affected by response to the Consistency question.

The final aspect of our analysis concerned the Difficulty question. An over-all mean rating indicating high difficulty could be taken as vitiating our results. However, the mean rating was 4.3 on a 7-point scale (i.e., only slightly on the difficult side).

10. Discussion

Experiment 2 replicated Experiment 1. Further, the earlier unexpected effect of natural kind was eliminated. Table 6 (like Table 3) underestimates the total number of disagreement with essentialism which are shown in Table 7.

The pattern of agreements with essentialism mirrors that obtained in Experiment 1 (see Table 4). It again fares poorly with Membership statements under Essential individual and Essential category scenarios, with

Table 7
Proportion of response patterns in Experiment 2 that conform to the predictions of essentialism

	Scenario type		
Statement	Non-essential category	Essential individual	Essential category
Existential	98% (254)	89% (230)	87% (225)
Qualified	88% (228)	46% (120)	73% (189)
Membership	89% (230)	47% (121)	76% (196)

agreement with essentialism ranging from 47% to 76% (compared with 46–58% in Experiment 1). For Qualified statements, agreement ranges from 46% to 73% (compared with 58–60% in Experiment 1). Existential statements show agreement ranging from 87% to 98% (compared with 88–95% in Experiment 1). Although, for Membership and Qualified statements the proportion of agreements with essentialism exceeds the chance level (25% – calculated as before), approximately 24–54% of responses are incompatible with essentialism. These findings also replicate our earlier refutation of Keil (1986) and Rips (1989). Overall, our results contradict essentialism, but concur with representational change views of concepts.

The Possibility questions showed that, for Existential and Membership statements, most subjects believed that apparently contradictory responses reflect real inconsistencies. However, this belief had no effect on the number of contradictions actually observed. The Consistency question showed that most subjects tried to be consistent some of the time. Given that most subjects believe that contradictory responses indicate inconsistency and that they tried to be consistent, these findings indicate a surprising dissociation between the use of a concept in classification, and an individual's awareness of how they use concepts.

GENERAL DISCUSSION

Two aspects of our findings cast doubt on the essentialist intuitions used to make the modal step. First, essentialism predicts that ordinary word use mirrors these intuitions and should reflect the pattern shown in Table 1. Our subjects simply did not share these intuitions. Second, the contradictions made by subjects suggest that the context insensitivity predicted by essentialist intuitions is not upheld. However, several essentialist counter-arguments might be mounted and, in this section we present arguments against these. After suggesting a representational change account of our findings, we conclude with some suggestions for further investigations of the ramifications of essentialism.

11. Essentialist counter-arguments

Essentialist counter-arguments may come in two guises. One would suggest theoretical difficulties in our presentation of essentialism, undermining the motivation for our investigation. The other might concur with our presentation but seek to provide an essentialist interpretation of the findings. However, a preliminary point to note is that no amount of psychological evidence will impinge upon the truth of essentialism: essentialism is an ontological doctrine, and so cannot be challenged by evidence

concerning subjects' *beliefs* about ontology reflected in word use. However, our evidence *can* cast doubt on the sharedness of the essentialist intuitions, and hence on the argumentation for the modal step.

Nonetheless, counter-arguments might claim that our findings are simply irrelevant to the essentialist case. First, it might be argued that essentialism does not imply that someone who knows the essence of a kind will inevitably employ that essence as a criterion for classification (denying Intuition A): essentialism would only require that people be "sensitive" to an essentialist as well as a non-essentialist option. Such sensitivity would then demonstrate that the Fregean relation between intension and extension (i.e., the interdefinability of intension and extension implies that grasping a term's intension equals knowing its extension) does not hold in general. It could then be claimed that our data reveal this sensitivity via the responses consistent with essentialism. Our reply has several components. The first emphasises the consistency of this sensitivity with representational change and Fregean views of concepts. These views do not imply infallibility in the relationship between a graspable sense and reference (or intension and extension): a term may include conventional content that an agent does not grasp – that is beyond the agent's ken. However, non-graspable content is not necessarily essentialist content. Hence, Fregean views *allow* that people may be sensitive to the difference between graspable content and content beyond their grasp, while the latter is not assumed to be grounded in essences.[6]

Our second reply is a denial: that is, our findings are simply not consistent with essentialism, which requires that *all* conventional uses of natural kind terms, not just some of them, are determined by essences: metaphysical essentialism is *not* a matter of degree. Moreover, it is not clear that the arguments of Kripke and Putnam are consistent with the claim that essences do not provide *the* criterion for conventional uses of kind terms where the essence is known: if the essence is not employed, the use *must* be non-conventional. However, since *any* theory of word meaning or concepts would credit people with the ability to use terms both conventionally and non-conventionally, there is nothing intrinsically essentialist about being sensitive to both options for word use: where Fregean and essentialist approaches comes apart is in their *characterisation* of conventional meanings/concepts.

A second counter-argument to our position might claim (pace Bealer) that intuitions are only critical to essentialism insofar as they accurately estimate

[6] The counter-argument takes Fregean views to accept the claim that grasping intensions entails infallibility in determining extensions. But, Fregean views claim only that a term's sense determines its reference. The argument thus treats senses as equivalent to intensions (e.g., Carnap, 1947; Montague, 1974), and implies that, since intensions determine extensions, so *senses* determine extensions. However, Fregean views do not require this idealisation of senses (indeed, such idealisations give rise to other problems; Franks, 1995b): for example, anti-realist approaches (e.g., Dummett, 1978, 1992; Tennant, 1987; Wright, 1992; Putnam, 1990) regard senses as only *defeasibly* determining extensions.

how people *would* use words if the counterfactual scenarios were *actual*, and that such intuitions are in fact inaccurate. In effect, the counter-argument claims that while word use in actual circumstances is relevant to the essentialist argumentation, *intuitions about* word use in counterfactual circumstances are not. However, this claim cannot be assessed. First, the claim that *intuitions* about word use in counterfactual circumstances would not reflect word use were those circumstances actual, cannot be decided in advance of such circumstances becoming actual: only then will word use be open to scrutiny. Hence, the claim itself can only be based on (meta-) intuitions about the accuracy of subjects' intuitions about word use. But, if intuitions in general *are* ill-founded, what could support such meta-intuitions? Since intuitions are the only form of evidence about word use under counterfactual circumstances, they are *central* to the essentialist argument. In the absence of an a priori argument in favour of essentialism, without such intuitions essentialism would represent anchorless metaphysics.

A third counter-argument to our position might claim that even conventional uses of words cannot be used to individuate conceptual content. However, since most experimental methods employ word use to indirectly elucidate the nature of conceptual content, this claim implies that we have little tangible evidence of the content of concepts in the first place. Hence, *no* experimental results concerning word use, not just those reported here, would bear on the issue of conceptual representation. Though possible, the claim compromises the great majority of experimental research into concepts. Further, severing the links between word use and concepts raises questions about the status of essentialist intuitions about word use, since word use would then permit *no* reliable inference to the nature of conceptual content.

In sum, none of these counter-arguments successfully undermines our investigation. However, perhaps essentialism could appeal to a pragmatic interpretation of our findings. Since such interpretations are mainly offered only where word use is non-conventional (e.g., using *ham sandwich* to refer to a restaurant customer; Nunberg, 1978), this appeal would hold that our subjects' word use was non-conventional and so not indicative of conceptual content. Since only conventional uses of natural kind terms are taken to be governed by essences, essentialism could claim that non-essentialist responses signal non-conventional uses, while essentialist responses signal conventional uses. However, such a pragmatic explanation of our results is untenable since the stimuli we employed directly reflect those that Kripke and Putnam take as constitutive of *conventional* word use.

12. Representational change and non-essentialist concepts

Our claim is that essentialism does not adequately predict people's intuitions about the use of natural kind terms under the critical counterfactual scenarios. We suggest that more promising accounts may be forthcoming

from the Fregean tradition, and in particular from representational change theories. An account of our findings must explain both the way in which conceptual *representations* vary according to context, and the way in which *classification* behaviour thereby varies (Braisby, 1990, 1993; Franks, 1992). The two aspects are closely intertwined: the principal explanation for classification behaviour is in terms of the underlying representations, and the contents of representations are inferred from classification behaviour. On a representational change view, concepts are viewed as elements of knowledge whose content may be represented in terms of "complex, structured descriptions" (Cohen & Murphy, 1984, p. 33), which support word use in classification. Such theories are broadly Fregean in inspiration in that concepts can have different contents in different contexts. At least some of these contents do not pertain to even a presumed essence of the category, *even in the case of conventional uses*. Accordingly, these theories are entirely consistent with our findings that word use varies according to scenario and statement type. Specifically, a sentence (e.g., *X is a cat*) may rationally be considered to be both true and false of a single entity (*X*) at one and the same time: this is precisely Frege's criterion for individuating senses. Thus, our subjects' apparent contradictions may be interpreted as indicating multiple senses of natural kind terms. Representational change theories may explain our results by positing one content of a concept pertaining to the (presumed) essence of the kind, and others pertaining to non-essential properties associated with the kind. For instance, upon discovering that *all* cats are really robots, subjects' judgements of *Tibby is a cat* as True and *Tibby is not a cat* also as True can be explained if the sense of *cat* in the first sentence pertains to the actual essence (robot), while the sense of *cat* in the second pertains to the previously presumed essence (feline). Despite a general commitment to multiple senses, different representational change theories offer different explanations of conceptual representations and classification. Here we briefly sketch two such theories: one in terms of prototype theory, the other in terms of sense generation.

Prototype theory (Hampton, 1988, 1992) would offer the following general style of explanation for the findings. The pattern of contradiction within subjects, and difference in views across subjects, indicates that the question *are there such things as cats*? has no clear-cut answer in the imagined scenario. The dilemma presented to the subject is one of a sudden conceptual change, too severe to be handled simply by adjustment of attribute weights, and in response to this change subjects may adopt one of two strategies. First, the term *cat* could be linked to the class of particulars to which it was previously linked (i.e., a particularist reading) such that the content of the concept *cat* has to radically change – *cats exist, but we were wrong about their being mammals*. Second, employing a property reading, the content of *cat* as *feline*, *mammal* . . . , could be preserved, thus requiring a new prototype concept with which to represent the class of objects that were previously (and wrongly) thought to be cats. In this case, (feline) cats

may or may not exist (although if scientists have checked all possible candidates on Earth, without finding any, their existence may be highly unlikely). However Tibby is not a cat by this second strategy, and one might well infer that there are no cats. By either strategy, two prototype concepts end up being represented in memory – one for the robots and one for the felines (note that existence of any exemplar is not a prerequisite for representing a concept as a prototype). The two concepts would have similar appearance attributes, but different attributes for internal parts, origin, etc. As a result, the two concepts would allow different entities to be classified as a cat or as a non-cat, on the basis of either exceeding or failing to exceed a threshold of similarity to the different prototypes. Taking the property reading, then, would mean that no objects would exceed threshold (since it would fail to match on a sufficient number of attributes), whilst taking the particularist reading would mean that all of the things we have always called cats would exceed the threshold. Such differential classifications of the particulars would then support the subjects' different truth evaluations of statements like *cats do not exist*. The question of which of these two prototypes deserves the label of *cat* is then undecided, and will depend on the whim of the subject.

Sense generation (Braisby, 1990, 1993; Franks, 1995a; Franks & Braisby, 1990), in contrast to prototype theory, would explain our findings by assuming that subjects are employing two or more binary, non-fuzzy contents, which arise from different classification perspectives. In judging *there are no such things as cats* as true or false, what counts as a *cat* depends on the sense of *cat*, which is determined by the perspective adopted. One perspective then defines the property reading (pertaining to the old, presumed essence), whilst the other defines the particularist reading (pertaining to the new, actual essence). However, while diferring on non-observable essence properties, the perspectives share content concerning diagnostic or observable properties, since these are identical regardless of essence. The truth of *cats exist* can then be evaluated from either perspective: from a property perspective, the statement is false because the particulars do not possess the old, presumed essence properties; from a particularist perspective, the statement is true because they do possess the new, actual essence properties. The judgements are not contradictory because they employ the same linguistic expression with different content. This reconciles the fact that subjects made contradictory truth evaluations while believing that they were responding in a consistent way. Sense generation also holds that classification judgements are clear-cut and binary, rather than fuzzy or indeterminate. Different senses arise because different perspectives support different derivations from stable lexical conceptual content. Assuming the lexical concept for *cat* reflects the old, presumed essence as well as diagnostic attributes (i.e., the property reading), then the derivation of a property sense would require few changes to the lexical concept. By contrast, a particularist sense would involve a denial or defeat

of the old presumed essence, so that the sense inherits information about the new, actual essence. The inheritance mechanism may involve accessing some relating or implicitly attached concept that provides the additional content. In this case, the additional concept could be one for *robot*, a concept that is available from contextually provided information. Over time, this particularist or *robot* sense may be used frequently so that the *robot* content becomes lexicalised, and would not need to be generated anew for each use of the term. Since each sense is determined by different perspectives, the pragmatic acceptability of each is relative to different audiences and circumstances. For example, using the property reading of *cat* to talk to a cat expert, or the particularist reading to talk to someone unacquainted with the discovery, courts miscommunication. Thus, although senses are determined by perspectives, and so the sense a term has is a function of context, the communicative factors briefly outlined suggest important constraints on the process.

These representational change views, in positing senses that pertain to actual and presumed essences, may seem to suggest a way of retaining essentialism. However, to reiterate, essentialism holds that concepts have *only one* conventional content, pertaining to the *actual* essence of a category. Thus, representational change explanations seriously undermine essentialism. Although people may talk about things possessing essences, and we may allow that one of the conventional contents for a concept might reflect such a belief, this content would capture only a severely weakened notion of an "essence" – being explicitly non-metaphysical in nature, it is not an "essence" that essentialism would recognise.

This interpretation of belief in an essence may be an appropriate way of viewing the notion as used in the psychological essentialism approach to concepts. This holds that, "people *do* believe that things have essences", and they "behave as though they believed it" (Medin & Ortony, 1989, p. 183). We suggest that our evidence indicates that people do not, in fact, believe that things have essences, if essences are interpreted according to the model provided by Kripke and Putnam (even though people may sometimes behave as if they did). Alternatively, the notion that people believe in essences may be weakened, such that those beliefs are of essences, but that people fail to follow through the ramifications of essentialism in a coherent or complete manner. For example, it may be that when faced with actual discoveries, people would use kind terms consistent with essentialism, but that they cannot make appropriate inferences and intuitions on the basis of counterfactual scenarios. This would allow that people may believe in essences without always behaving accordingly. However, differentiating this position from one in which people do not believe in essences at all, may be a difficult empirical task. Nonetheless, motivation for supposing that people do believe in essences comes from Rips' (1989) results (Medin & Ortony, 1989) and, though these appear to support essentialism, our failure to replicate them together with our other

findings suggest otherwise. In sum, it seems that psychological essentialism may be construed as the view that conventional use of terms (and concepts) are *not* governed by the kind of essence we have discussed, nor as governed by consistent essentialist beliefs. Rather, it may be viewed as the claim that *one* of the many shades of a term's conventional content is governed by a belief about something that may be *called* an essence. Again, this interpretation seriously undermines the connection with philosophical essentialism.

13. Other approaches to empirical investigations of essentialism

We have adopted a particular approach to investigating essentialism, directly addressing implications concerning intuitions about language use under counterfactual circumstances. However, additional investigations might attempt to discern how far people *do* employ some weakened notion of an essence (as noted above), and just *how* close the notion is to true essentialism. For instance, belief in essentialism could be taken as a matter of degree, such that different degrees of essentialist belief could be indicated by the extent to which people adhere to the implications of essentialism in their use of natural kind terms. Two further possibilities are currently under investigation.

The first focuses on the notion of the "division of linguistic labour", argued by Putnam to reflect how grasp of meaning varies within a linguistic community. In brief, this suggests that different members or subsets of a community have different knowledge about the properties of natural kinds: often, only scientists are thought to have knowledge about the essence. Putnam also argues that members of the community *believe* that a division of linguistic labour holds: lay people may believe only scientists are privy to essences. The view then implies that lay people should defer to scientists in their use of natural kind terms. So, if a scientist discovers a kind's essence to be different from that which was previously presumed, then ordinary language users should follow the scientist's use of the term (such scenarios were central to our investigation). Related questions then concern whether deference varies for different domains (e.g., familiar versus unfamiliar, natural kind versus observational kind), or for different experts (e.g., natural scientist versus social scientist).

The second focuses on the possibility that deference, and apparent adherence to essentialism, may vary as a function of the radicalism of the discovery concerning essential properties. That is, if the presumed essence of a bird is discovered to be the essence of another bird (e.g., what we once thought of as sparrows are in fact robins), the adherence to essentialism may appear greater than if the discovered essence was mammalian. These investigations employ folk taxonomies of living things (cf. Atran, 1989) and examine whether allegiance to essentialism shifts as discoveries about essences stay within a genus (e.g., from one bird to another bird), shift to

another animate genus (e.g., bird to mammal), a shift to an inanimate genus (e.g., bird to fruit), shift to artifacts, and so on. Though essentialism may contend that the more the shifts have commonsense plausibility, the more essentialist will be subjects' intuitions, Shapere (1982) argues that in more scientifically plausible counterfactual discoveries, intuitions are likely to be *less* essentialist.

CONCLUSIONS

Our findings show that natural kind terms are not employed in an essentialist manner. Rather, they are used in ways that are sensitive to context and reveal patterns of apparent self-contradiction. We have suggested that the appropriate explanation of these findings is that the conventional content and use of natural kind terms varies systematically with context, as predicted by representational change theories, and not that conventional content and use are invariably associated with essences, as predicted by essentialism. Finally, our findings cast doubt upon the critique advanced by Rey and Margolis: since they claim that the cognitive psychology of concepts is undermined by essentialism, and essentialism depends upon a particular set of intuitions about word use, the lack of corroboration for those intuitions undermines their critique. Essentialism may not be as essential to a theory of concepts as has been supposed.

Acknowledgements

We would like to thank Scott Atran, Larry Barsalou, Nick Chater, Robin Cooper, Mark Crimmins, Danièle Dubois, Greg Murphy and Jean-Pierre Thibaut, and two anonymous reviewers for helpful comments and discussions. Nick Braisby was supported by a British Academy research fellowship during the early stages of this research.

References

Atran, S. (1989). Basic conceptual domains. *Mind & Language, 4*, 7–16.

Bach, K. (1987). *Thought and Reference*. Oxford: Clarendon Press.

Barsalou, L.W. (1987). The instability of graded structure: implications for nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorisation*. Cambridge, UK: Cambridge University Press.

Barsalou, L.W. (1993). Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols. In A.C. Collins, S.E. Gathercole, M.A. Conway, & P.E.M. Morris (Eds.), *Theories of memory*. Hillsdale, NJ: Erlbaum.

Barsalou, L.W., & Sewell, D.R. (1984). *Constructing representations of categories from different points of view*. Emory Cognition Project Report No. 2. Emory University, Atlanta, GA.

Barsalou, L.W., Yeh, W., Luka, B.J., Olseth, K.L., Mix, K.S., & Wu, L-L. (1993). Concepts and meaning, In K. Beals, G. Cooke, D. Kathman, K.E. McCullough, S. Kita, & D. Testen (Eds.), *Chicago Linguistic Society, 29: Papers from the Parasession on Conceptual Representations*. Chicago: Chicago Linguistic Society.

Bealer, G. (1987). The philosophical limits of scientific essentialism. In J. Tomberlin (Ed.), *Philosophical perspectives, 1: Metaphysics*. Atascadero, California: Ridgeview Publishing Co.

Braisby, N.R. (1990). Situating word meaning. In R. Cooper, K. Mukai, & J. Perry (Eds.), *Situation theory and its applications, I*. CSLI: Stanford.

Braisby, N. (1993). Stable concepts and context-sensitive classification. *Irish Journal of Psychology, 14*(3), 426–441.

Braisby, N.R. (1994). The perspective nature of complex concepts. In M. Keane, P. Cunningham, M. Brady & R. Byrne (Eds.), *AI and cognitive science '94*. Dublin: Dublin University Press.

Braisby, N.R., Franks, B., & Myers, T.F. (1992). Partiality and coherence in concept combination. In J. Ezquerro & J.M. Larrazabal (Eds.), *Cognition, semantics and philosophy*. Dordrecht: Kluwer.

Burge, T. (1979). Individualism and the mental. In P. French, T. Uehling & H. Wettstein (Eds.) *Midwest studies in philosophy* (Vol. IV) (Studies in Metaphysics). Minneapolis: University of Minnesota Press.

Canfield, J.V. (1983). Discovering essence. In C. Ginet & S. Shoemaker (Eds.), *Knowledge and mind*. Oxford: Oxford University Press.

Carnap. R. (1974). *Meaning and necessity*. Chicago: Chicago University Press.

Clark, H.H. (1983). Making sense of nonce sense. In G.B.F. d'Arcais & R.J. Jarvella (Eds.), *The process of language understanding*. Chichester: Wiley.

Cohen, B., & Murphy. G.L. (1984). Models of concepts. *Cognitive Science, 8*, 27–58.

Donnellan, K.A. (1983). Kripke and Putnam on natural kind terms. In C. Ginet & S. Shoemaker (Eds.), *Knowledge and mind*. Oxford: Oxford University Press.

Dummett, M.A.E. (1978). *Truth and other enigmas*. London: Duckworth.

Dummett, M.A.E. (1992). *Frege and other philosophers*. Oxford: Blackwell.

Evans, G. (1982). *The varieties of reference*. Oxford: Basil Blackwell.

Fodor, J.A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Franks, B. (1991). Sense generation and concept combination. In B. Franks (Ed.), *Word meaning and concepts*. ESPRIT BRA Deliverable R2.4. Centre for Cognitive Science, University of Edinburgh.

Franks, B. (1992). Folk-psychology and the ascription of concepts. *Philosophical Psychology, 5*(4), 369–390.

Franks, B. (1995a). Sense generation: a "quasi-classical" approach to concepts and concept combination. *Cognitive Science, 19*(4), 441–505.

Franks, B. (1995b). On explanation in the cognitive sciences: competence, idealisations, and the failure of the classical cascade. *British Journal for the Philosophy of Science, 45*, 475–502.

Franks, B., & Braisby, N.R. (1990). Sense generation or how to make a mental lexicon flexible. In *Proceedings of the 12th annual conference of the cognitive science society*. Cambridge, MA: MIT, July 1990.

Frege, G. (1952/1892). On sense and reference. In P. Geach & M. Black (Eds.), *Translations from the philosophical writings of Gottlob Frege*. Oxford: Blackwell.

Hampton, J.A. (1982). A demonstration of intransitivity in natural concepts. *Cognition, 12*, 151–164.

Hampton, J.A. (1988). Overextension of concept conjunctions: evidence for a unitary view of membership and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14* 12–32.

Hampton, J.A. (1992). Prototype models of concept representation. In I. Van Mechelen, J.A. Hampton, R.S. Michalski, & P. Theuns (Eds.) *Categories and concepts: Theoretical views and inductive data analysis*. London: Academic Press.

Hampton, J.A. (1995). Testing prototype theory of concepts. *Journal of Memory and Language, 34.*

Keil, F. (1986). Conceptual development and category structure. In U. Neisser (ed.), *Concepts and conceptual development.* Cambridge: Cambridge University Press.

Kripke, S.A. (1972). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Semantics of natural languages.* Dordrecht: Reidel.

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

Malt. B.C. (1991). Word meaning and word use. In P.J. Schwanenflugel (Ed.), *The psychology of word meanings.* Hillsdale, NJ: Erlbaum.

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

Margolis, E. (1994). A reassessment of the shift from the classical theory of concepts to prototype theory. *Cognition, 51,* 73–89.

McCloskey, M., & Glucksberg, S. (1978). Natural categories: well-defined or fuzzy sets? *Memory & Cognition, 6,* 462–472.

McGinn, C. (1989). *Mental content.* Oxford: Blackwell.

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

Medin, D.L. and Shoben, E.J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*(2), 158–190.

Montague, R. (1974). *Formal philosophy.* New Haven: Yale University Press.

Murphy. G.L. (1991). Meaning and concepts. In P.J. Schwanenflugel (Ed.), *The psychology of word meanings.* Hillsdale, NJ: Erlbaum.

Nersessian, N. (1984). *Faraday to Einstein: Constructing meaning in scientific theories.* Dordrecht: Nijhoff.

Nunberg, G.D. (1978). *The pragmatics of reference.* PhD Thesis. Reproduced by the Indiana University Linguistics Club, Indiana.

Putnam, H. (1975a). Is semantics possible? In *Mind, language and reality. Vol. 2: Philosophical papers.* Cambridge: Cambridge University Press.

Putnam, H. (1975b). The meaning of "meaning". In *Mind, language and reality. Vol. 2: Philosophical papers.* Cambridge: Cambridge University Press.

Putnam, H. (1990). *Realism with a human face.* Cambridge, MA.: Harvard University Press.

Rey. G. (1983). Concepts and stereotypes. *Cognition, 15,* 237–262.

Rey, G. (1985). Concepts and conceptions. *Cognition, 19,* 297–303.

Rey, G. (1992). Semantic externalism and conceptual competence. *Proceedings of the Aristotelian Society, LXXXXII,* 315–333.

Rips, L. (1989). Similarity, typicality and categorisation. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.

Roth, E.M. & Shoben, E.J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346–378.

Shapere, D. (1982). Reason, reference, and the quest for knowledge. *Philosophy of Science, 49,* 1–23.

Sperber, D. and Wilson, D. (1986). *Relevance.* Oxford: Basil Blackwell.

Tennant, N. (1987). *Anti-realism and logic.* Oxford: Basil Blackwell.

Wright, C. (1992). *Truth and objectivity.* Cambridge, MA.: Harvard University Press.