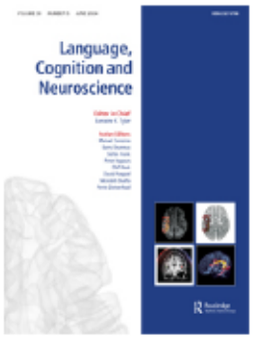

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/plcp21

Words do not just label concepts: activating superordinate categories through labels, lists, and definitions

Lilia Rissman & Gary Lupyan

To cite this article: Lilia Rissman & Gary Lupyan (2024) Words do not just label concepts: activating superordinate categories through labels, lists, and definitions, Language, Cognition and Neuroscience, 39:5, 657-676, DOI: 10.1080/23273798.2024.2350526

To link to this article: https://doi.org/10.1080/23273798.2024.2350526

LANGUAGE, COGNITION AND NEUROSCIENCE
2024, VOL. 39, NO. 5, 657–676
https://doi.org/10.1080/23273798.2024.2350526

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Words do not just label concepts: activating superordinate categories through labels, lists, and definitions

Lilia Rissman and Gary Lupyan

Department of Psychology, University of Wisconsin – Madison, Madison, WI, USA

ABSTRACT

We investigate the interface between concepts and word meanings by asking English speakers to list members of superordinate categories under one of three conditions: (1) when cued by a label (e.g. *animals*), (2) an exemplar list (e.g. *dog, cat, mouse*), or (3) a definition (e.g. "living creatures that roam the Earth"). We find that categories activated by labels lead to participants listing more category-typical responses, as quantified through typicality ratings, similarity in word embedding space, and accuracy in guessing category labels. This effect is stronger for some categories than others (e.g. stronger for *appetizers* than *animals*). These results support the view that a word is not merely a label for a concept, but rather a unique way of accessing and organizing conceptual space.

ARTICLE HISTORY
Received 17 July 2023
Accepted 23 April 2024

KEYWORDS
concepts; categories; lexical semantics; distributional semantics

1. Introduction

Consider a banana, an orange, and some grapes. If you are a typical American adult, you would have extensive knowledge about these items and you would likely think of them as belonging to the same category. If asked to name this category in English, you would likely say: *fruit*. How necessary is this label for invoking the category? Different theories of the relationship between language and thought suggest different answers to this question (Snodgrass, 1984). In some theories, words are thought to label regions of conceptual space (Murphy, 2002; Rogers & McClelland, 2004), whereas in other theories, word meanings relate to, but are not reducible to, conceptual structure (Levinson, 1997; Malt & Majid, 2013). Both perspectives have a long history in cognitive science, in part because each finds support in existing empirical evidence. We attempt to reconcile previous findings by examining the role that superordinate words like *fruit* (which pick out larger categories than the "basic-level" words *banana* and *orange*) play in activating the categories they denote. We investigate how superordinate categories are processed in different contexts, shedding light on the relationship between semantic structure and conceptual structure – whether, and to what extent, word meanings are reducible to concepts.

1.1. Words and concepts: how are they related?

People are skilled categorisers, fluidly organising entities in the world in multiple, cross-cutting ways. For example, the items *cheeseburger, pizza,* and *hot dog* could be categorised as food, or things to eat at a baseball stadium, or items that are partially coloured brown, etc. (see Ross & Murphy, 1999).[1] This general ability to form categories is not dependent on language. People can extract categories like different species of birds from co-varying clusters of features in the environment (see Malt, 1995; Rogers & McClelland, 2004 for review) and children can represent categories without knowing words for them (see Westermann & Mareschal, 2014 for review). For example, 3- to 4-month-old infants can already categorise horses vs. cats (Eimas & Quinn, 1994) and 12-month-old infants can categorise novel objects (e.g. staplers vs. teapots), which helps scaffold word learning (Pomiechowska & Gliga, 2018). These concept representations appear to be distributed across multiple regions of the brain, rather than being represented as local, atomic units (Binder & Desai, 2011; Kemmerer, 2019; Ralph et al., 2017; Tong et al., 2022).

How does language interface with this conceptual system? One widespread view is that words label concepts (Gliozzi et al., 2009; Rogers & McClelland, 2004; Sloutsky et al., 2001). For example, one Introductory Psychology textbook describes concepts in the following way: "To have even a simple thought such as 'my cousin borrowed my hoodie,' first you'd need to understand the concept of *cousin*, the concept of *borrow*, and the concept of *hoodie*" (Pomerantz, 2023, p. 210). This passage is from a section on conceptual knowledge, not language – the fact that the author uses words to

CONTACT Lilia Rissman lrrgsh@rit.edu Department of Psychology, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY, 14623, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/23273798.2024.2350526.

refer to concepts suggests that the two may not need to be distinguished. Researchers who study conceptual knowledge (also called semantic memory) often assume that labels are one of the attributes of a category (Kumar, 2021; Martin, 2016; Patterson et al., 2007; Ralph et al., 2017; Sloutsky & Fisher, 2012). As described by Martin (2016), semantic information about objects "can be accessed through multiple modalities … information about how dogs look is accessed automatically when we hear a bark, or when we read or hear the word 'dog'" (980). From this perspective, knowing the label (*dog*) that English speakers use to refer to the concept DOG is on par with knowing what dogs like to eat or what their fur feels like.

Compatible with this view, Murphy (2002) argues that words map onto concepts: "a word gets its significance by being connected to a concept or coherent structure in conceptual representation" (Murphy, 2002, pp. 388–389). From this perspective, word meanings and concepts not only align with each other, but are structurally parallel:

> there isn't one theory of concept learning … and a qualitatively different one for nouns, say. The reason is that … the same phenomena occur in nonlinguistic concepts and in word use, and so any theory of one will serve to a large degree as a theory of the other. (Murphy, 2002, p. 399)

If we imagine conceptual structure as a Christmas tree, words might be viewed as ornaments, attaching to conceptual space branches (concepts). The ornaments attach to those branches which are labeled but omit those branches which are not (see Murphy, 2002, Figure 11.4). On this view, there is only one cognitive structure to be accounted for (see also Jackendoff, 1983; Langacker, 1987; Pustejovsky, 1995).

The issue of whether a theory of concepts will serve as a theory of word meanings is distinct from the question of how different languages carve up conceptual space. Cross-linguistic diversity in semantics is widespread (Blasi et al., 2022; Evans & Levinson, 2009; Kemmerer, 2019; Levinson & Wilkins, 2006; Thompson et al., 2020). For example, Dutch uses separate verbs for cutting with a single blade (e.g. cutting bread with a knife; *snijden*) vs. cutting with two blades (e.g. cutting hair with scissors; *knippen*) whereas English uses the verb *cut* for both these types of events. This type of semantic diversity does not necessarily contradict the view that words label concepts. Given that many regions of conceptual space are not labeled, it may simply be that different languages constitute different configurations of ornaments on the conceptual tree. That is, Dutch *snijden* and *knippen* label the single blade and double blade concepts, respectively, and English *cut* labels the region including both these concepts.

Users of different languages have shared knowledge of conceptual/perceptual categories, even if these languages provide different sets of labels (Kemmerer, 2019; Malt & Majid, 2013; Rissman, Liu, et al., 2023). For example, speakers of English, Spanish, and Chinese sort pictures of household containers in similar ways despite naming them in strikingly different ways (Malt et al., 1999). At the same time, having different sets of labels may reconfigure conceptual/perceptual categories in subtle ways (Athanasopoulos & Casaponsa, 2020; Lupyan et al., 2020; Wolff & Holmes, 2011). This debate is orthogonal to the issue of whether word meanings and concepts are structurally parallel – even if conceptual structure is subtly different across users of different languages, words may still map onto concepts as labels.

Rather than being mere ornaments, however, an alternative is that a language's vocabulary provides its own system of semantic organisation, where the categories denoted by words interface with, but are not reducible to, conceptual and perceptual categories (Bierwisch & Schreuder, 1992; Enfield, 2022; Keller, 1998; Levinson, 1997; Malt, 2024). For one, words are often more abstract and schematic than the thoughts they invoke in any particular instance. For example, if I say on October 30 *my flight is tomorrow night,* my interlocutor will take me to be traveling on Halloween, but the same words said on December 31 would imply traveling on New Year's Day. *Tomorrow* is an abstract indexical, but its interpretation depends on context and is necessarily specific (see Bierwisch & Schreuder, 1992; Levinson, 1997). As described by Levinson (1997),

> this one-to-many mapping shows that SR [semantic representation] and CR [conceptual representation] cannot be isomorphic. The persuasive view that SRs are schematic, incomplete, or semantically general suggests that SR is not simply a subset of CR either, but a representational medium with a different vocabulary and syntax. (p. 19)

Along similar lines, words are thought to be semantically restricted in ways that conceptual knowledge is not. For example, many verbs lexicalise manner (e.g. *scrub, bounce*) or result/path (e.g. *clean, cross*) but not both, even though people have no difficulty in conceptualising an event with both a manner and a result (e.g. someone scrubbing a table clean) (Beavers & Koontz-Garboden, 2020; Grimshaw, 2005; Rappaport-Hovav & Levin, 2010; Talmy, 1985).

Words also appear to play a special role in activating category knowledge compared to other cues. When English-speaking adults hear an auditory cue paired

with a picture and report whether the picture matches the cue, participants are faster when the cue is a label (e.g. "cow") rather than a characteristic sound (e.g. a cow mooing) (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). This label-advantage is strongest for the most typical exemplars suggesting that what the label is doing is preferentially activating category-diagnostic features (the things that make a cow different from its semantic neighbours) (see also Lupyan, 2008; Lupyan & Spivey, 2010). These findings cast doubt on the proposal that words are simply a feature associated with a concept on par with features such as characteristic appearance, sounds, and textures.

A final reason to doubt that words merely label concepts is that behaviour in a range of tasks takes on a more categorical character under the influence of category labels. For example, when people are cued with the category label "triangle", they draw more typical triangles than when they are cued with the definition of a triangle such as "three sided polygon". The two cues also produce different recognition profiles in a speeded recognition task and different profiles in an unspeeded inferential reasoning task – all pointing to the two types of cues activating different representations (Lupyan, 2017). In lower-level tasks such as visual discrimination, hearing labels like "green" changes what colours people are able to accurately discriminate, supporting the idea that colour labels induce more categorical visual representations compared to nonverbal ways of conveying the idea of greenness (Forder & Lupyan, 2019). If the categories invoked by words have clearer boundaries than the categories present in conceptual knowledge, then word meanings are not reducible to concepts. As articulated by Malt (2019), conceptual knowledge should not be assumed to be packaged in word-sized chunks:

> although "categorization" is often cited as a crucial operation for navigating the world … , it is not obvious that interpreting experiences, as opposed to communicating about them, requires the association of some element of immediate experience with a bounded chunk of stored knowledge. (p. 8)

1.2. Superordinate nouns as a window into the linguistic/conceptual interface

Do people activate superordinate concepts independently from language, such that the word *fruit* merely labels the concept FRUIT? Several studies suggest yes. For example, when people rate the similarity of two items (e.g. *watermelon, strawberry*), their ratings do not differ depending on whether they are cued with the category label (e.g. *fruit*) or not (Barsalou, 1982; Ross & Murphy, 1999). In addition, when people judge whether a word is a member of a category (e.g. *soda – beverages*), people are not faster after reading a sentence that primes the category (e.g. "The soda was poured into a glass by the waiter") (Ross & Murphy, 1999). Finally, Waxman and Gelman (1986) asked three- and four-year-old children to categorise objects using three types of cues: labels (e.g. "this puppet only likes animals"), an exemplar list (e.g. "This puppet only likes things like a dog, or a horse, or a duck, and other things like that"), or by telling them that the exemplar objects (i.e. the dog, horse, and duck) make "a really good group". Children then divided a larger set of objects into one basket that the puppet would like and a second basket that the puppet would not like. Although three-year-olds made more correct classifications given the superordinate label (*animal*, *clothing*, or *food*) than the other two cues, four-year-olds were at ceiling for all three cue types suggesting that a list of exemplars and a superordinate label activated similar categories in the minds of four-year-olds (adults were not tested). Although individuals in these tasks may be self-generating the labels even when labels are not explicitly given, this set of results is at least compatible with the idea that people can activate concepts independently from their labels.

As distinguishing adults' understandings of concepts vs. word meanings is tricky business, it is important to replicate these findings with different methods. In addition, these studies tested a small set of superordinate terms, and did not focus on analyzing differences across individual terms. If a concept is represented independently of its label, it should be possible to activate the concept through an alternate means (i.e. without using the label). By contrast, if word meanings are more schematic than concepts and impose more categorical divisions on conceptual space, then the categories activated by words should differ systematically from categories activated through other types of cues. We conducted a series of experiments to see whether it is possible to activate the same superordinate category that is activated by the conventional superordinate label, in the absence of an overt label.

We test two non-label alternatives to cueing categories: exemplar lists (e.g. *banana, orange, grapes*) and definitions (e.g. "foods that are grown from a seed and have seeds in them"). We then compare people's understanding of these categories with their understanding of superordinate nouns (e.g. *fruit*). We test exemplar lists because prior research suggests that basic-level members of a category can activate superordinate categories (Barsalou, 1982; Ross & Murphy, 1999). In addition, conveying a category

through representative members is widely attested across languages (Mauri, 2017; Mauri & Sansò, 2018). We test definitions because this is a plausible way of activating a category without using the word for it. In fact, many theorists have assumed that even if English lacks a translational equivalent for a word in another language (e.g. Dutch *knippen*), this has little consequence for communication because English speakers can just use the word's definition (e.g. "cut with two blades") (see Rissman, Liu, et al., 2023 for review).

If words cue categories that are more abstract and discrete than exemplar-cued categories, then the relationship between words and exemplars is unidirectional: *fruit* activates the typical exemplars apple, banana, and orange, but encountering an apple, a banana, and an orange does not necessarily activate the intension of *fruit*. In other words, cueing participants through the exemplar list *banana, orange, grapes* might activate a similar category as *fruit*, but not an identical one, either in terms of the scope of the category or its degree of abstraction. For example, if people tend to match exemplars, then their understanding of *fruit* may be broader than their understanding of the exemplar-based category *banana, orange, grapes* – a blackberry is a good example of *fruit* but is relatively dissimilar perceptually to each of those specific exemplars.

The definitions we use in this study (e.g. "living creatures that roam the Earth") are phrasal descriptions of categories and so are structurally similar to "ad hoc" categories such as "flowers that aren't toxic to cats" (Barsalou, 1983, 1991). Our definitions are composed of superordinate nouns that are modified to restrict the category of items under consideration such that they include items denoted by the target superordinate term and exclude items not denoted by the term. At the same time, our definitions are potentially different from ad hoc categories because the category being defined is already well-known to our participants. Although the definition "living creatures that roam the Earth" is not conventionalised, it corresponds to a category that *is* named with a conventional term (i.e. *animals*). This *in-principle* equivalence allows for a strong test of the hypothesis that exemplar lists and definitions activate the same conceptual content as a conventional superordinate term. Encountering a list of salient exemplars like *orange, banana, grape* (in the context of a task asking people to think of a category that includes these items) may lead people to activate a representation very similar to that activated by *fruit*. Alternatively, it could lead them to implicitly gloss the list with the conventional label *fruit*. In either case, we should then fail to find a difference in semantic knowledge activated by the conventional label vs. the exemplar list. In contrast, finding such differences would show that even in cases where people are very familiar with the conventional labels, the labels serve to activate different semantic knowledge than what can be activated in their absence.

2. Experiment 1

We investigated category representation by asking English speakers to list members of various categories. We varied how these categories were cued: through a label, an exemplar list, or a definition. If labels play a unique role in activating categories, we predict two ways in which the responses of people cued by labels will differ from the responses of people cued by exemplar lists or definitions. First, responses will be more central to the category activated by the label – for example, we predict that when cued by the label *fruit*, participants' responses will be more typical of the category *fruit* than when they are given an exemplar list (e.g. *orange, banana, grapes*) or a definition. Second, we predict that labels will lead to greater alignment between respondents compared to exemplar lists and definitions. This prediction follows from studies showing that labels activate more typical representations (Lupyan, 2017; Lupyan & Thompson-Schill, 2012). Because there are fewer ways to be typical than atypical, activating more typical representations should lead to greater alignment, a prediction supported by several recent studies (Rissman, Liu, et al., 2023; Suffill et al., 2022, under review).

2.1. Participants

We recruited 264 English-speaking adults on Amazon Mechanical Turk (Experiment 1A: $N_{female} = 89$, $N_{male} = 82$, $N_{other} = 3$, age range = 20–70, median age = 36; Experiment 1B: $N_{female} = 37$, $N_{male} = 52$, $N_{other} = 1$, age range = 23–71, median age = 38). An additional 38 participants were tested but were excluded for providing repetitive or non-word responses ($N = 7$) or because they viewed a duplicate set of trials as another participant ($N = 31$; see Design and procedure). Participants received $1.00 for completing the study. Different participants were tested across each of Experiments 1A-B, Experiment 2, and the Typicality rating task. Participants in all studies were located in the United States and self-identified as native speakers of English. Across all studies, 8% of participants reported being comfortable using a language other than English. Informed consent was obtained for all participants in this study. This research was approved by the University of Wisconsin – Madison Institutional Review Board, #2020-0683.

2.2. Materials

In Experiment 1A, we used the 20 superordinate terms shown in Table 1. We selected terms that have been studied in previous work on superordinate categories (see Wisniewski et al., 1996), that are familiar to English speakers, and that have sufficiently many category members such that participants could easily list six members (cf., *precipitation*, which has relatively few category members). Given our goal of addressing whether some labels play a stronger role in cueing categories than others, we tested a heterogeneous set of superordinate nouns. This set includes both natural kind and artifact category terms (e.g. *animals* vs. *weapons*) and mass and count terms (*food* vs. *games*). It also includes nouns at differing levels of abstraction (e.g. *pets*, *mammals*, and *animals*), which correspondingly have different numbers of category members, as well as goal-based categories such as *hobbies* and *games* (see Barsalou, 1983; Chrysikou, 2006). We did not have a priori hypotheses about how these different dimensions of word meaning would interact with any observed effect of cue type (labels vs. exemplars vs. definitions).

In Experiment 1B, we tested 20 definitions corresponding to each of the 20 superordinate terms in Table 1. We selected one definition for each term using the following procedure: first, we asked 15 adult English speakers on Amazon Mechanical Turk to write definitions for each of the 20 terms.[2] From this set, we selected three definitions per term that we judged to be the best approximations of the term's meaning. We then asked 16 adult English speakers on Amazon Mechanical Turk to rate on a 0–100 scale how good the definition was for the term (0 – terrible, 100 – excellent). To avoid ties, we asked participants to give different ratings to each of the three definitions. For each term, we selected the definition with the highest mean rating; these are listed in Table 1. Across terms, mean ratings ranged from 73 to 92 with an overall mean of 82.[3]

We used Turker-generated definitions with the goal of understanding how non-technical definitions that might be produced in everyday communication activate category representation. It is possible, however, that these rough-and-ready definitions are inferior to those produced by trained lexicographers. To assess this possibility, we gathered ratings of definitions from three online dictionaries: Oxford Languages, Merriam-Webster, and American Heritage. For each term in Table 1, we selected the first definition that appeared under the noun entry for that term, unless that definition corresponded to a different sense than our Turker-generated definition (e.g. the first definition of *pet* appearing in the Merriam-Webster dictionary is "a pampered and usually spoiled child"). Definitions were pluralised and technical jargon was removed (e.g. "kingdom (Animalia)" was removed from the definition of *animal*). We also removed exemplar lists from definitions. For example, the Merriam-Webster definition of vegetable is "a usually herbaceous plant (such as the cabbage, bean, or potato) grown for an edible part that is usually eaten as part of a meal" – we removed the exemplar list "(such as the cabbage, bean, or potato)". Dictionary definitions are listed in Supplementary Materials.

Fourteen English speakers on Amazon Mechanical Turk viewed four definitions of each term on a single screen (the three dictionary definitions and our Turker-generated definition) and were asked to rate each definition from 0 to 100. Participants were told to rate each definition according to how well they think it captures the meaning of the word. As in the previous rating study, participants were told that they should not assign the same value to more than one definition. The order of

Table 1. Superordinate terms and definitions tested in Experiments 1A-B.

Term (Experiment 1A)	Definition (Experiment 1B)	Definition mean rating (SD)
Animals	Living creatures that roam the Earth	80.9 (18.2)
Appetizers	Foods that are eaten before the main course	82.1 (14.4)
Chores	Jobs done at home to keep the home running smoothly	79.9 (16)
Clothing	Garments that we wear on our body	78.2 (16.7)
Desserts	Sweet foods that are usually eaten after a meal	81.8 (14.7)
Diseases	Ailments that can hurt or infect people or animals	78.1 (16.6)
Flowers	Things which are part of a plant and have colourful petals and stems	80.2 (18.7)
Food	Items consumed for nourishment and sustenance	87.5 (10.7)
Fruit	Foods that are grown from a seed and have seeds in them	74.1 (21.8)
Furniture	Things used in the home or workplace to sit, eat, sleep or hold items	86.1 (15.2)
Games	Activities that people participate in for fun	78.1 (19)
Hobbies	Activities or interests people do or participate in for fun/leisure	86.7 (13.5)
Mammals	Warm-blooded animals that give birth to live young and produce milk	92.1 (10)
Pets	Domesticated animals kept as helpers or companions by humans	88.6 (13)
Plants	Organisms that grow in soil and use photosynthesis	86.6 (12.9)
Tools	Objects that are used to fix or create things	77.3 (18.7)
Toys	Things that are made for children to play with	73 (17.3)
Vegetables	Foods that are parts of plants but are not fruits, nuts, or seeds	NA
Vehicles	Types of machinery used to transport humans or cargo	88.1 (10.6)
Weapons	Things that are used to harm others or for self defense	75.8 (22.6)

the definitions on the screen was randomised. Participants rated definitions for all 20 superordinate terms, which were presented in random order. An additional 27 English speakers were tested but excluded for failing one or more attention check trials. In these trials, participants viewed three dictionary definitions along with an incorrect definition of an additional superordinate (e.g. for *drugs*, the incorrect definition was "organic compounds that contain only carbon, hydrogen, and oxygen and that originate chiefly as products of photosynthesis"). To pass an attention check trial, participants needed to rate the incorrect definition lower than all three dictionary definitions.

2.3. Design and procedure

In Experiment 1A, participants viewed one of two Cue Types: Labels or Exemplars (see Figure 1). Participants in the Label condition viewed the terms in Table 1 and were asked to list six members of each category. We gave participants the examples of *colours* and *beverages* (i.e. given the category *colours*, they might list *red, blue, green, yellow, orange, pink*). Each participant listed category members for 10 superordinate labels, selected randomly from the total set of 20 and presented in a random order. In the Exemplar condition each participant was yoked to a participant from the Label condition, viewing the first three category members listed by the yoked Label condition participant. Participants in the Exemplar condition were told that these three words belonged to a category and that their task was to list three more members of the category. As in the Label condition, we gave Exemplar participants the examples of *colours* and *beverages* (i.e. "if you were given the words red, blue, and green, you might list yellow, orange, and pink".) Each Exemplar participant was yoked to a unique Label participant. Participants in both conditions were instructed to list the first category members that came to mind. Exemplar-based categories as discussed in previous studies typically include at least three members – we asked participants in the Label condition to list six members in order to align the Label and Exemplar conditions, as shown in Figure 1.

Experiment 1B added a Definition condition. The procedure for Experiment 1B was the same as for the Label condition in Experiment 1A, except that participants were given a phrase naming a category and were asked to list six members of that category. Participants were given the examples *actions which are illegal* and *things you can drink* (e.g. "if you were given the phrase 'Things you can drink,' you might list the following category members: soda, wine, beer, water, iced tea, coffee").

2.4. Data preprocessing

We standardised spelling and inflectional variants of responses to reduce inconsequential variability. For example, *action-figure* and *action figure* were replaced by *action figures*; *cleaning the room* was replaced with *cleaning room*. We retained the variant that was most common across all responses.

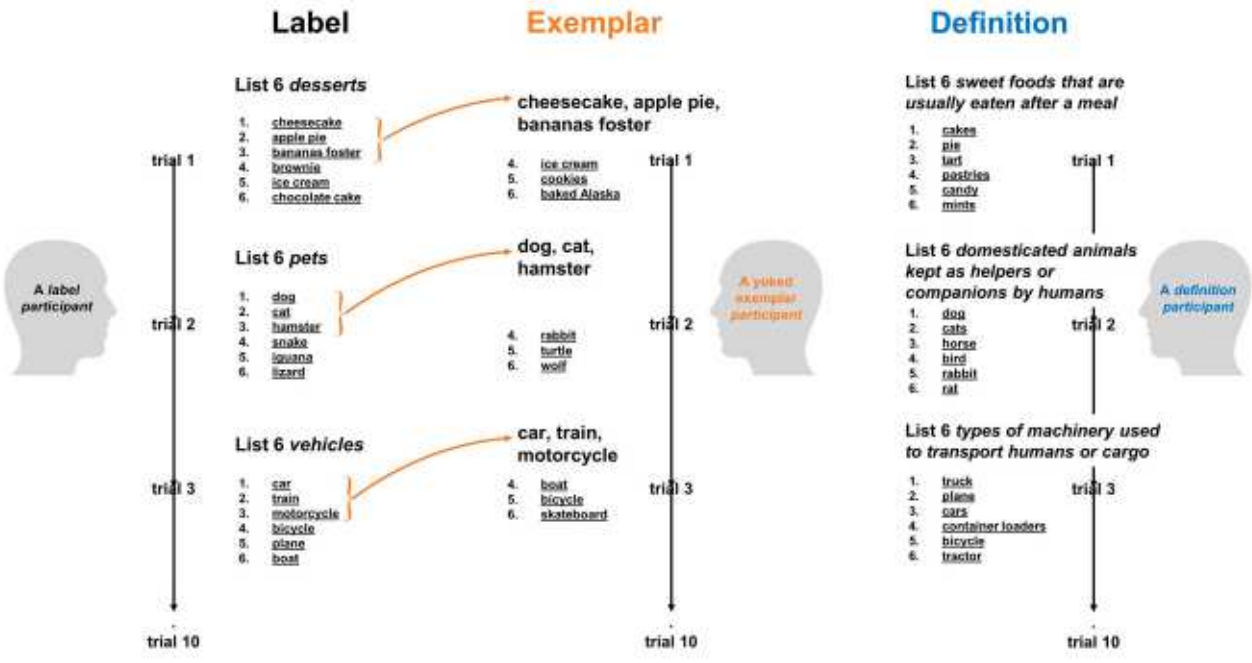

Figure 1. Schematic representation of Experiment 1 design with sample participant responses.

2.5. Typicality ratings

We collected typicality ratings for all term/response pairs produced in Experiments 1A-B (e.g. *fruit/banana*; $N = 2886$). Responses across all three Cue Types were evaluated relative to the superordinate term associated with each trial. For example, the participant in the Exemplar condition in Figure 1 produced the response *baked Alaska*, and the typicality of *baked Alaska* was evaluated relative to the term *desserts*. Similarly, for participants in the Definition condition who viewed the category definition "Sweet foods that are usually eaten after a meal", their responses were evaluated relative to the term *desserts*.

To collect typicality ratings, we recruited an additional 490 English speakers on Amazon Mechanical Turk ($N_{female} = 201$, $N_{male} = 283$, $N_{other/NA} = 6$, age range = 20–78, median age = 37). Participants were asked "How typical is the example [response] of the category [label]?" and rated each term/response pair on a scale from 1 to 8, where 1 corresponded to *Not an example*, 2 to Not typical at all, and 8 to *Very Typical*. Participants were given the example: "if the category is 'sports', you would probably rate football as more typical than lacrosse". Participants viewed one or two terms chosen at random from the total list of 20, and responses were rated in term blocks (e.g. a *plants* block followed by an *appetizers* block). The median number of ratings per term/response pair was 12 ($N_{min} = 7$). Participants received between $1.25 and $1.75 depending on the number of trials they completed (ranging from 50 to 120 trials).

2.6. Response similarity in word embedding space

Quantifying similarity between people's responses would require an enormous number of pairwise comparisons. We therefore relied on word embeddings learned by a distributional semantic model. These models are trained to predict words based on their surrounding context and are an increasingly common tool used in cognitive science to model semantic relationships (Boleda, 2020; Bolognesi, 2020; Thompson et al., 2020). The output of these models is a set of vectors that positions each word at a unique point in a high-dimensional space. We used word embeddings trained on English Wikipedia + Statmt news corpus using the *fast-text* algorithm (300 dimensions with subwords) (Bojanowski et al., 2017). Using these embeddings, we computed the cosine distance between responses. In the analyses that follow, "similarity" refers to 1 minus the cosine distance. One shortcoming of this method is that we lacked embeddings for the 13% of our responses that were multi-word phrases (e.g. *heart disease*). For these, we treated the response as the vector sum of the component words. This procedure produces reasonably good representations of compound words such as those we are dealing with here (Boleda, 2020).[4] For compound word calculations, we excluded the word *and* (e.g. for *chips and dip*, we summed the vector for *chips* and the vector for *dip*).

As additional verification, we replicated all our analyses using the subs2vec embeddings derived from movie and TV show subtitles (van Paridon & Thompson, 2020) which contains entries for many compound words. Analyses using the subtitles embeddings are included in Supplementary Materials. For almost all analyses, the Wikipedia and subtitles embeddings yielded equivalent findings; divergent results are noted in the text.

2.7. Frequency and generality by superordinate term

We anticipated that some terms would show larger Cue Type condition differences than other terms. Although there are many possible reasons for such condition-by-term interactions, we can gain initial insights by considering term frequency and generality. These two covariates are proxy measures for important properties of individual categories such as how common/culturally relevant the category is, the frequency of the term in ambient language, and the specificity of the category (although specificity is related to concreteness, the two are theoretically distinct (Bolognesi, 2020)). We quantified frequency as log-transformed word counts from the Corpus of Contemporary American English (both written and spoken language included) (Davies, 2023).

We collected ratings of the generality of each term from 34 English speakers on Amazon Mechanical Turk (see Bolognesi & Caselli, 2022; Lewis et al., 2021). Participants rated how general or specific a word's meaning is on a 1–5 scale (1 being very specific and 5 being very general). Participants were instructed that "*animal* is quite general, *dog* is more specific, and *poodle* is even more specific". Participants rated the generality of 295 superordinate terms/phrases, and we obtained 10–12 ratings for each term.

2.8. Results

2.8.1. Analytic approach

We compared participants' responses across the three Cue Type conditions (Label, Exemplar, and Definition) in three ways. First, we quantified the *centrality* of each generated response relative to the category denoted

by each superordinate label (Section 2.8.2). For example, if a participant in the Label condition listed *brownie, ice cream, chocolate cake*, and a participant in the Exemplar condition listed *ice cream, cookies, baked Alaska*, how central is each of these responses to the category *desserts*? Second, we analyzed the *diversity* of the responses produced in each condition: did participants produce a wider range of responses in one condition over the others (Section 2.8.3)? Third, we analyzed the within-trial similarity of the responses and whether this differed across conditions (Section 2.8.4).

In addition to testing whether Cue Type influenced participants' responses, we analyzed how responses differed depending on when in the trial they were produced (first response, second response, etc.). Participants in the Label and Definition conditions produced six responses, corresponding to Response Numbers 1 through 6. Participants in the Exemplar condition only produced three responses, but we label these as having Response Numbers 4, 5, and 6, to maintain comparability across the three Cue Type conditions (see Figure 1).

We fit mixed effects linear regression models using the *lme4* package for R (Bates et al., 2014; R Core Team, 2022). We evaluated whether a predictor makes a significant contribution to a model using likelihood ratio testing. The independent variable Response Number was taken to be a continuous variable. Unless noted otherwise, models included random intercepts for Participants and Terms as well as Term by Cue Type random slopes.[5] Continuous variables were centered and scaled. We used the *lmerTest* package (Kuznetsova et al., 2017) and Satterthwaite approximation to compute *p*-values for fixed effects (see Luke, 2017). 95% confidence intervals were calculated using *summarySEwithin* from the *Rmisc* package. Stimuli materials, data files, analysis scripts, and Supplementary Materials are available at: https://osf.io/y7g2r.

2.8.2. Semantic centrality of responses

We analyzed the relationship between the responses and category structures denoted by the superordinate terms in two ways: first, we analyzed the typicality ratings described in Section 2.5 (e.g. "how typical is the example 'cheesecake' of the category 'desserts'?"). Second, we used the word embeddings described in Section 2.6 as a measure of each response's similarity to the centroid of each superordinate term in word embedding space (defined in Supplementary Materials). Measures of similarity-to-centroid were found to have a moderate positive correlation with human ratings of typicality: $r(2736) = .57$, $p < .001$. The results using typicality and similarity-to-centroid were highly similar; the latter are reported in Supplementary Materials.

In the analyses below, we compared Response Numbers 1 through 6 between the Label and Definition conditions but only Response Numbers 4 through 6 between the Label and Exemplar conditions.

The mean typicality of each response relative to the superordinate term is shown in Figure 2. We fit one model to the Label/Exemplar data and another model to the Label/Definition data, as Response Numbers 1–3 were absent from the Exemplar condition. We used sum contrast coding for the Cue Type variable: coefficient values reflect the difference between the Label condition and the overall mean.

In the Label vs. Exemplar model, responses were more typical in the Label condition ($\beta = .080$, $CI_{95} = [.021, .14]$, $t = -2.66$, $p < .05$). Responses decreased in their typicality as Response Number increased ($\beta = -0.064$, $CI_{95} = [-0.088, -040]$, $t = -5.25$, $p < .001$), and there was an interaction between Cue Type and Response Number such that the difference between the Label and Exemplar conditions was greater for higher Response Numbers ($\beta = .032$, $CI_{95} = [.008, .056]$, $t = 2.62, p < .01$). This result shows that when participants just saw exemplar lists such as *cheesecake, apple pie, bananas foster*, they produced responses that were judged as being less typical members of the superordinate. We will refer to this effect as the Label Advantage.

In the Label vs. Definition model, responses were again more typical in the Label condition ($\beta = .22$, $CI_{95} = [.12, .32]$, $t = 4.38$, $p < .001$) and responses decreased in their typicality as Response Number increased ($\beta = -0.19$, $CI_{95} = [-0.20, -0.17]$, $t = -22.22$, $p < .001$). Adding the interaction between Cue Type and Response Number did not significantly increase the likelihood of the model. This result demonstrates a Label Advantage for definitions: participants produced more typical category members when cued with the label rather than with a definition of the term.

As described in Section 2.5, participants in the typicality rating task could select that a response was not just atypical but was not even a member of the category, e.g. all participants judged that a *cedar* is not an example of the category *flowers*. We asked to what extent the Label Advantage for typicality ratings was driven by Not-example responses, as opposed to atypical but within category responses (e.g. *trout* as a type of *appetizer*). We modeled the typicality data including the mean rate of Not-example responses for each response/term pair as a regressor in the model (e.g. 1.0 for *cedar/flowers*, .5 for *shovel/weapons* and 0 for *trout/appetizers*). When Not-example ratings were controlled

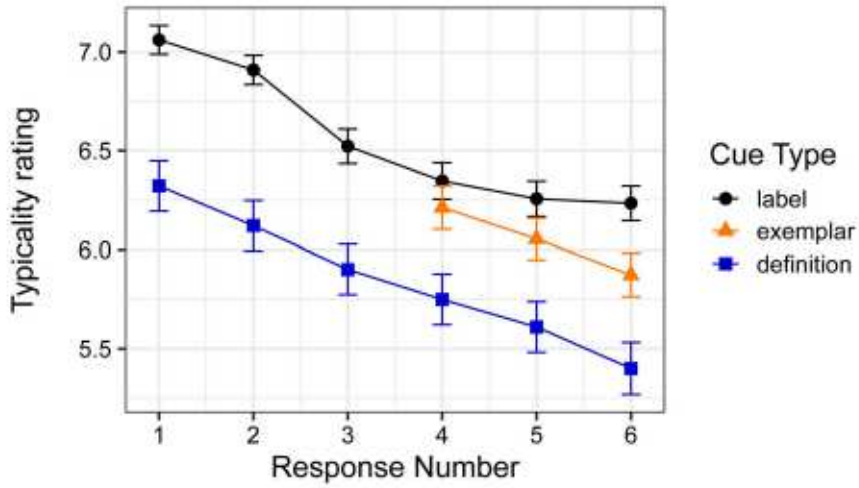

Figure 2. Mean typicality ratings for each response/term pair, grouped by Cue Type and Response Number (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Typicality was judged on a scale from 1 (Not an example) to 8 (Very Typical). Error bars show 95% confidence intervals of the mean.

for, responses in the Label condition were not significantly more typical ($p > .1$). This shows that responses were more typical in the Label condition than the Exemplar condition because Exemplar participants were more likely to produce responses that were arguably not instances of the superordinate category. By contrast, responses in the Label condition were still significantly more typical than responses in the Definition condition controlling for likelihood of Not-example ratings ($\beta = .058$, $CI_{95} = [.029, .086]$, $t = 3.97$, $p < .01$).

We next analyzed the size of the Label Advantage for each individual term in each condition. As shown in Figure 3A, the Label Advantage was positive for most terms. The advantage of labels over exemplars was small for some terms (e.g. *fruit, animals, chores*, and *tools*) but large for others (e.g. *toys, mammals, hobbies*, and *appetizers*). We modeled whether the frequency and generality of the superordinate terms predicted whether responses in the Label condition were more typical than responses in the Exemplar condition. We observed a significant interaction between Cue Type and frequency, such that the Label Advantage was greater for less frequent terms ($\beta = -.057$, $CI_{95} = [-.098, -.016]$, $t = 2.70$, $p < .05$). In a separate model, we observed a significant interaction between Cue Type and generality, such that the Label Advantage was greater for more specific terms ($\beta = -.055$, $CI_{95} = [-.096, -0.14]$, $t = 2.61$, $p < .05$). Rather than manifesting equally across all superordinate categories, these results show that the advantage of labels over exemplars depended on properties of the individual categories. We did not fit a model with both frequency and generality as predictors, as these measures are moderately correlated ($r(18) = .68$, $p < .001$).

Turning to definitions, the size of the Label Advantage for individual terms reflects how well our definitions approximate the meaning of the superordinate term. For example, the Label Advantage for *pets* was close to zero, suggesting that "domesticated animals kept as helpers or companions by humans" activates a similar category representation as *pets*. By contrast, the Label Advantage for *fruit* was relatively large, meaning that "foods that are grown from a seed and have seeds in them" does a relatively poor job of activating the category *fruit*. In models of the Label/Definition data, we found that frequency of the superordinate terms interacted with Cue Type such that more frequent superordinate terms showed a greater Label Advantage ($\beta = .11$, $CI_{95} = [.034, .20]$, $t = 2.79$, $p < .05$). The generality of the term did not interact with Cue Type ($p > .1$).

As described in Section 2.2, we tested rough-and-ready definitions of the terms, rather than dictionary definitions produced by trained experts. It may therefore be the case that we observe a Label Advantage over definitions simply because we chose imprecise definitions for the terms. To test this possibility, we compared the ratings of our Turker-generated definitions to ratings of dictionary definitions. For each term, we identified which dictionary definition received the highest rating and then computed the difference between this best definition and our Turker definition. On a scale from 0 to 100, the mean difference across terms was 7.9, and difference scores ranged from –16.2 (*furniture*; our definition was better) to 26.8 (*diseases*; our definition was worse). In a model of response typicality with both the difference scores and Cue Type as predictors, the interaction between difference scores and Cue Type

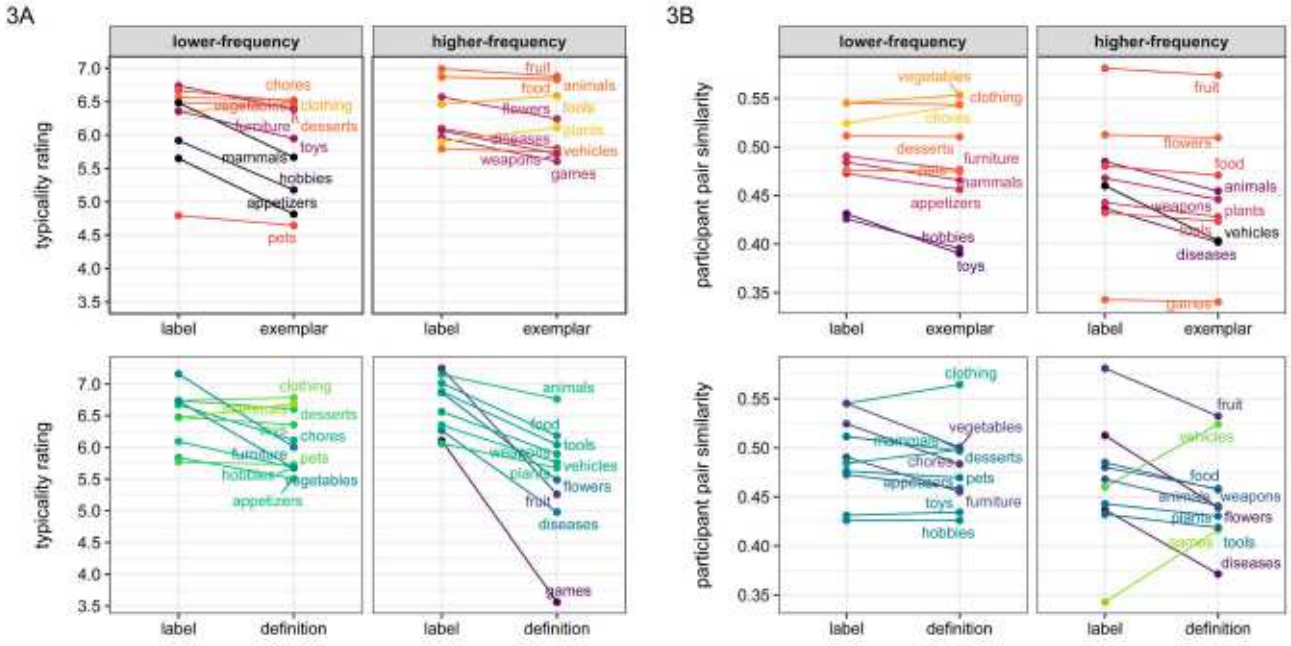

Figure 3. A. Mean typicality of responses, grouped by Cue Type and superordinate term (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Terms are grouped by lower vs. higher frequency. Darker colours indicate a stronger Label Advantage for that term. B. Similarity between participants in word embedding space, grouped by Cue Type and superordinate term (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Terms are grouped by lower vs. higher frequency. Darker colours indicate a stronger Label Advantage for that term.

was non-significant ($p > .1$). This suggests that although the Label Advantage over definitions was greater for some terms than others, this is likely not an artifact of the definitions that we chose. In other words, we observed a Label Advantage not because we tested a worse definition than some existing superior alternative, but because for many terms, adequate definitions are genuinely difficult to construct.

2.8.3. How diverse were the responses across different Cue Types?

Participants cued by conventional superordinate labels produced more typical/central responses. We next analyzed whether responses differed in their heterogeneity as a function of Cue Type. For example, when participants are listing *desserts*, do their responses constitute a narrower category than when participants are prompted by the exemplars *cheesecake, apple pie, bananas foster*? Although responses in the Exemplar and Definition conditions were less central overall than responses in the Label condition, these responses are not necessarily more heterogeneous. We analyzed response diversity in two ways: by computing Simpson's diversity index *D* for the responses for each term and Cue Type, and by calculating how similar participants' responses were to each other in word embedding space. For these analyses, we only included Responses 4 through 6 from the Label and Definition conditions, to allow comparison of response diversity across the three Cue Types.

2.8.3.1. Simpson's diversity D.

Simpson's *D* takes into account the distribution of response types within a category as well as the overall size of the category (see Majid et al., 2018; Rissman et al., 2022; Zettersten & Lupyan, 2020 on computation of this index). *D*-values range from 0 to 1, with 0 corresponding to complete heterogeneity (all responses are different) and 1 to complete homogeneity (all responses are the same). We computed the *D*-value for each term for each Cue Type. The mean *D*-value across the 20 terms was .03 for each of the three Cue Types. *D*-values were not significantly different between the Label and Exemplar conditions or between the Label and Definition conditions (p's $> .1$). This suggests that although responses were less central to the category in the Exemplar and Definition conditions, the responses were not themselves more heterogeneous in these conditions.

2.8.3.2. How similar were participants to each other?.

As a second test of whether responses were more diverse for some Cue Types than others, we analyzed how much individual participants aligned with each other across the three Cue Types. We observed in Section 2.8.2 that responses were less central to the category in the Exemplar and Definition conditions.

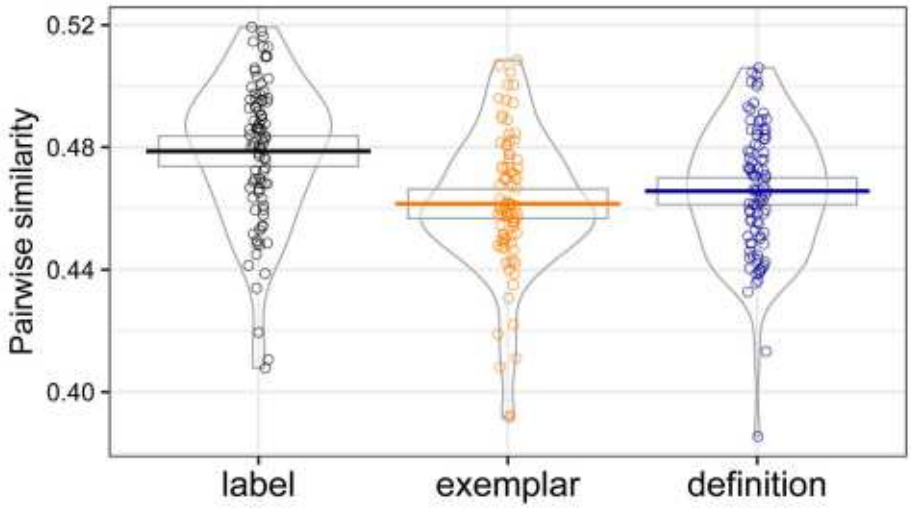

Figure 4. Mean similarity in word embedding space between each pair of participants for each Cue Type. Each point is a participant (indicating the mean similarity between that participant and every other participant).

It does not necessarily follow, however, that individuals diverged from each other more in these conditions – it could be that participants all produced the same kinds of atypical responses (e.g. *trout* as a type of *appetizer*). To test this, we calculated the similarity between the responses for each pair of participants in word embedding space. For example, if one Label participant produced *potato, leek, onion* and another Label participant produced *kale, okra, lettuce* as their fourth, fifth, and sixth responses in a *vegetables* trial, we computed the similarity between each pair of responses across the two participants (e.g. between *potato* and *kale* but not between *potato* and *leek*). We then calculated the mean similarity between each pair of participants for each Cue Type (Figure 4). For this analysis and the analyses in the subsequent section, we used treatment coding for the Cue Type variable, with the Label condition as the reference level.

Pairwise similarity was lower in the Exemplar condition than in the Label condition ($\beta = -.22$, $CI_{95} = [-.34, -.094]$, $t = -3.45$, $p < .01$), but did not differ between the Label and Definition conditions ($\beta = -.21$, $CI_{95} = [-.45, .041]$, $t = -1.64$, $p = .12$). This analysis suggests that when participants viewed an exemplar list, they interpreted the category structure of the list in divergent ways, leading them to align less with one another than when viewing a superordinate term. By contrast, viewing a definition did not lead to lower alignment. This non-result should be interpreted with caution, however, as when we calculated pairwise similarity using the subtitles embeddings (see Section 2.6), there was significantly less alignment in the Definition condition than the Label condition ($\beta = -.26$, $CI_{95} = [-.45, .066]$, $t = -2.64$, $p < .05$).

Figure 3B shows the Label Advantage for individual terms. The advantage of labels over exemplars was positive for most terms, with the largest Label Advantage for *toys* and *vehicles* and smallest for *chores* and *vegetables* (for which the Label Advantage was negative). In models of the Label/Exemplar data, neither the frequency and nor the generality of the terms interacted with Cue Type (p's > .1). There was a negative Label Advantage for many of the terms in the Definition condition (e.g. *vehicles, games, clothing*), indicating that the definitions for those terms led to less diversity and more cross-subject alignment than the terms themselves did.

2.8.4. Similarity of responses within trials

In an exploratory analysis, we used the word embeddings to calculate the mean semantic similarity between responses 4–6 and responses 1–3 in each trial. For example, if a Label participant produced *broccoli, zucchini, carrot* as their first three responses and *potato, leek, onion* as their second three responses, we calculated the mean similarity over all pairs in the Cartesian product of these two sets (i.e. *broccoli/potato, broccoli/leek*, etc.). Similarity on this measure was significantly higher in the Exemplar condition than in the Label condition ($\beta = .087$, $CI_{95} = [.012, .16]$, $t = 2.28$, $p < .05$) but did not differ between the Label and Definition conditions ($p > .1$). This suggests that participants in the Exemplar condition were more strongly tethered to the particular responses 1–3 than participants in the Label condition were.

2.9. Discussion

When people are asked to list members of a category, it matters whether the category is cued with a

superordinate label vs. an exemplar list vs. a definition. Given a superordinate label rather than an exemplar list, participants' responses were more central to the category activated by the label, as measured through typicality ratings as well as similarity of computed word embeddings – a Label Advantage. Participants' responses were also more central when cued with a label when compared with being cued by a definition. Although participants were more aligned with one another given label cues than exemplar cues, we did not find evidence that participants were more aligned with each other in their interpretation of a label over a definition. Importantly, the Label Advantage was stronger for some terms than for others, as shown in Figure 3. This variation across terms was not random: we found a greater Label Advantage for less frequent terms and more specific terms.

In Experiment 2, we investigated the robustness of the Label Advantage by testing whether the items generated in response to various cues (label, exemplars, definition) helped people infer the superordinate categories. For example, given the list *ice cream, cookies, baked Alaska* (produced originally in the Exemplar condition), how successfully can participants reconstruct that this category is *desserts*, and does participants' success vary as a function of Cue Type? This experiment is essentially Experiment 1 in reverse: rather than being given labels and producing category members, participants are given category members and asked to produce a label. The relationship between labels and category members is not necessarily symmetrical – *appetizers* can strongly activate *soup* without *soup* strongly activating *appetizers* (Rosch et al., 1976). If category exemplars can nonetheless reliably lead people back to the intended superordinate term, this may indicate that the exemplars activate a concept that is independent of the label. The crucial question is whether this occurs more often when the exemplars were produced by people in the Label condition than the Exemplar and Definition conditions from Experiment 1.

3. Experiment 2

3.1. Participants

We recruited 245 English-speaking participants on Amazon Mechanical Turk ($N_{female} = 114$, $N_{male} = 130$, $N_{other} = 1$; age $M = 38$; range = 21–69). An additional eight participants were tested but excluded for not identifying as a native English speaker ($N = 4$), for listing one of the individual exemplars rather than providing a category label as their response ($N = 3$), or for providing a thematically associated word rather than a category label (e.g. guessing *farmer* for the exemplar list *edamame, quinoa, buckwheat*; $N = 1$). Participants received $1.00. Informed consent was obtained for all participants.

3.2. Design and materials

Participants guessed category labels for three-item lists produced in the Label, Exemplar, and Definition conditions in Experiment 1 (responses 4–6 in a given trial; see Figure 5). Given the large number of responses produced in Experiment 1, we selected a sample of lists for this label guessing task. For each list in Experiment 1, we calculated the mean similarity of each of the three exemplars to the target superordinate term in word embedding space. We classified the lists for each term into tertiles of low, medium, and high similarity and randomly sampled five lists from each tertile for each term in each Cue Type condition (900 lists sampled in total). This sample was randomly divided such that each participant viewed lists from a single Cue Type and provided guesses for one list for each of the 20 terms.

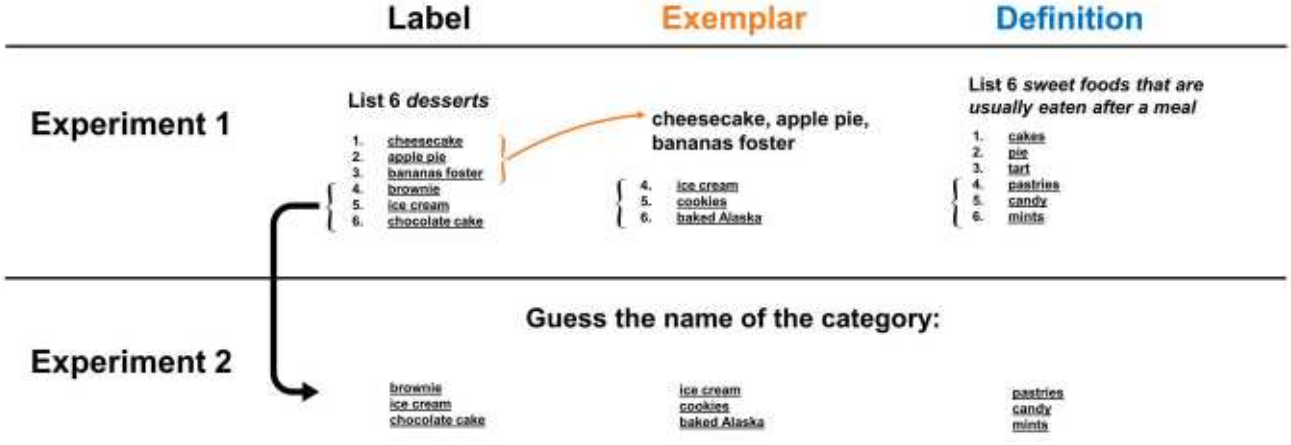

Figure 5. Schematic representation of the relationship between the designs of Experiments 1 and 2.

3.3. Procedure & data preprocessing

Participants were instructed that they would be given three words or phrases that were members of a category and they had to guess the name of the category. For example, if given "soda, wine, beer", they might guess "beverages". They were asked to provide a single word as a category label (e.g. not "primary colors"). Each participant completed 20 trials in random order.

We corrected the responses for spelling errors. Morphological variants of the superordinate terms (e.g. *fruits, appetizer*) were counted as correct. Modified variants of the superordinate terms (e.g. *farm animals*) were counted as incorrect.

3.4. Results

We analyzed the accuracy of participants' guesses using mixed effects logistic regression and used treatment coding for the Cue Type variable, with the Label condition as the reference level. Participants were less accurate at guessing the original superordinate term in the Exemplar and Definition conditions than in the Label condition ($M_{label} = 67.2\%$ vs. $M_{exemplar} = 59.6\%$ vs. $M_{definition} = 46.5\%$; Exemplar: $b = -0.43$, $CI_{95} = [-0.70, -0.15]$, $z = -3.058$, $p < .01$; Definition: $b = -1.12$, $CI_{95} = [-1.57, -0.66]$, $z = -4.77$, $p < .001$). Figure 6 shows guessing accuracy as a function of Cue Type for each superordinate term. Adding frequency to this model, we found that accuracy was higher for more frequent terms ($b = .73$, $CI_{95} = [.27, 1.18]$, $z = 3.14$, $p < .01$), and that frequency interacted with Cue Type: the advantage of labels over both exemplars and definitions was greater for more frequent terms (Exemplar: $b = -0.23$, $CI_{95} = [-0.45, -0.02]$, $z = -2.14$, $p < .05$; Definition: $b = -.59$, $CI_{95} = [-.94, -0.24]$, $z = -3.28$, $p < .01$). There was no effect of generality on the advantage of labels over exemplars, but the advantage of labels over definitions was marginally greater for more general terms ($b = -.39$, $CI_{95} = [-.79, 0.005]$, $z = -1.94$, $p = .053$).

Connecting the results of Experiments 1 and 2, we asked whether the average typicality and centroid similarity of the words in the lists predicted guessing accuracy. We found that lists composed of more typical words were easier to guess ($b = 1.22$, $CI_{95} = [1.11, 1.33]$, $z = 21.71$, $p < .001$). Adding typicality to the model eliminated the advantage of labels over exemplars. That is, guessing accuracy was worse in the Exemplar condition because these lists were less typical on average. Adding typicality to the model did not eliminate the Label Advantage over the Definition condition. We found the same pattern of results even when excluding lists where one or more words were judged to be not in the category of the superordinate. We also found that lists with higher centroid similarity were easier to guess ($b = .47$, $CI_{95} = [.38, .55]$, $z = 10.74$, $p < .001$). Adding centroid similarity to the model did not eliminate either the advantage of labels over exemplars or labels over definitions.

3.5. Discussion

Participants who viewed lists produced in the Experiment 1 Label condition were more accurate at guessing the original term than participants who viewed lists produced in the Exemplar or Definition conditions. This finding supports the results of Experiment 1: participants who are not cued with labels produce category responses reflecting different regions of conceptual space than the labels themselves. We found considerable variation across terms regarding both participants' overall accuracy in guessing the term as well as the size of the Label Advantage. As noted earlier, a superordinate may activate a category member without the category member activating the superordinate. It is therefore striking that for *fruit, clothing*, and *tools*, accuracy was high in both the Label and Exemplar conditions.

4. General discussion

The common rhetorical practice of referring to concepts using (English) words (e.g. "the concept FRUIT") reflects the perspective that words are *labels for* categories. From this perspective, conceptual space is organised in terms of categories and words refer to these categories (although not all structured regions of conceptual space are labeled). An alternate view is that words access and organise conceptual space in a unique way. For example, words are more abstract than individual thoughts and words activate more categorical representations than non-linguistic cues (e.g. a dog barking). For superordinate nouns, previous literature is compatible with either perspective. In this paper, we activated categories through superordinate labels, and contrasted these with the categories activated by two types of non-label cues: exemplar lists and definitions. Averaging across all 20 superordinate terms, we found that exemplar lists and definitions activated different categories than the labels that were originally used to generate them. Specifically, when participants in Experiment 1 were given exemplars or definitions, their responses were less central to the category of the superordinate term than when participants were given the terms themselves. Participants also diverged more from each other when given exemplar lists. Finally, participants in

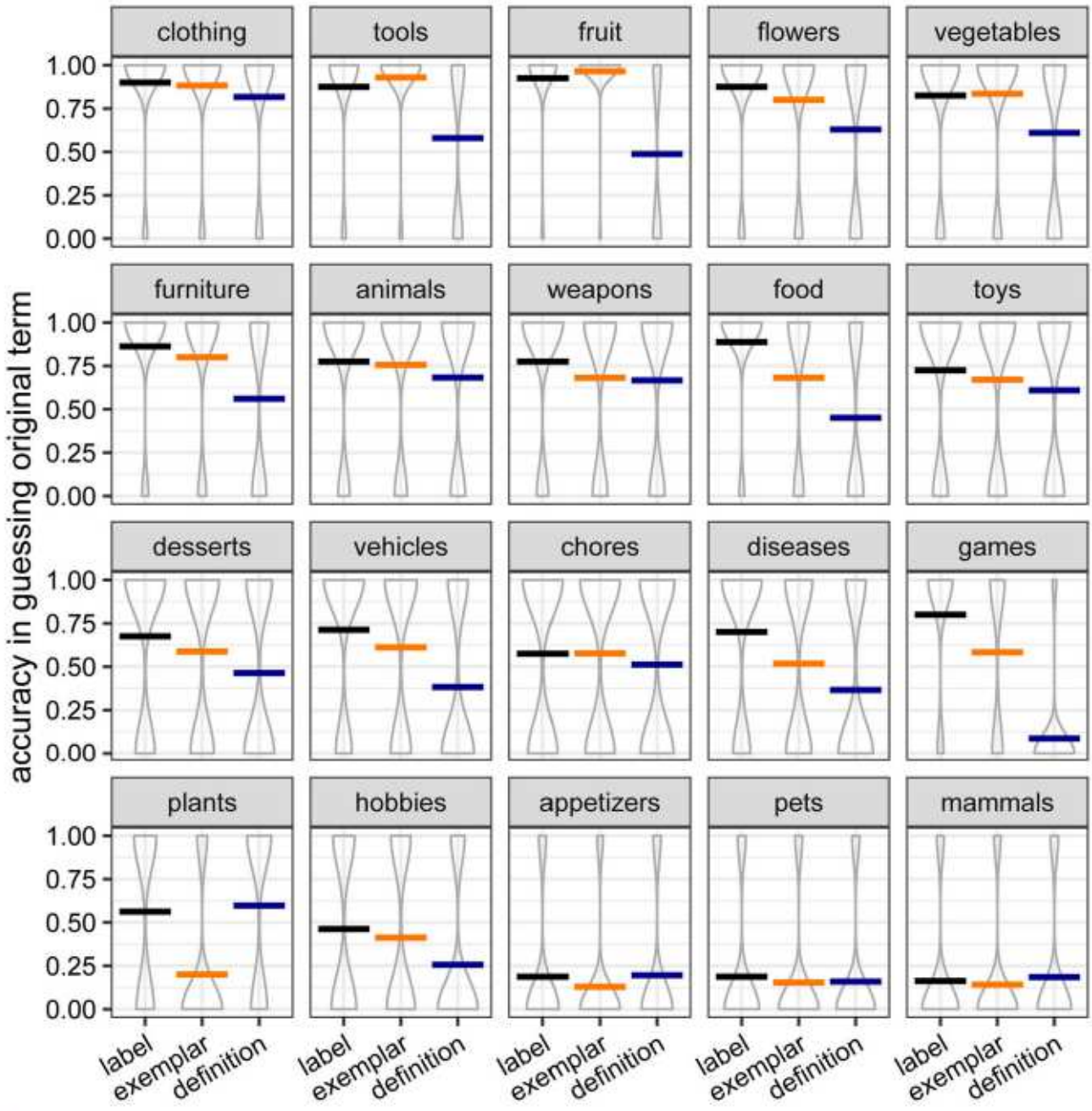

Figure 6. Guessing accuracy across each Cue Type condition for each superordinate term. The term panels are ordered by overall accuracy per term.

Experiment 2 were less accurate at guessing the superordinate term when their clue was a list of words produced by people in the exemplar or definition conditions. These results cast doubt on the view that words merely label concepts. We suggested earlier that words labeling concepts is akin to ornaments decorating a Christmas tree – there is only one type of mental organisation, and this is conceptual structure. If this view of the mind were correct, we would expect that exemplar lists and definitions would be more effective than they were in activating the concepts labeled by words. After all, we know that those regions of conceptual space are likely to be prominent for English speakers, precisely because they are labeled.

Although the superordinate terms in our study do not appear to be functioning only as labels, it is worth noting that the Label Advantage was stronger for some terms than for others – variation that was partially predictable from word frequency and generality. This finding suggests that the relationship between the lexicon and conceptual/perceptual structure is different for different words. In the sections that follow, we discuss

why exemplars and definitions do not in general activate the same categories as superordinate terms, we consider the status of categories for which there was a negligible Label Advantage, and we suggest future directions for studies on the relationship between words and concepts.

4.1. Labels vs. exemplars

Exemplar lists (e.g. *truck, doll, blocks*) did not consistently activate the same category as the label that was used to generate the exemplars (e.g. *toys*). This conflicts with research summarised in Section 1.1 – for example, that pairwise similarity ratings (e.g. *watermelon, strawberry*) are not affected by the presence of a superordinate label (e.g. *fruit*) (Barsalou, 1982; Ross & Murphy, 1999). A straightforward interpretation of the earlier findings is that similarity ratings are not influenced by labels because the exemplars themselves activate the concept that the label refers to. We found, however, that typical exemplars were often insufficient to activate label intensions, despite the high familiarity our participants had with the superordinates that we tested.

An immediate observation is that exemplar lists are often more ambiguous than conventional labels (e.g. *dog, cat, human* could be a group of mammals or a group of animals). As noted in Section 1, people can categorise a set of items in multiple, cross-cutting ways. For example, although the list *dog, cat, human* was produced by someone given the label *mammals*, the individual who received these exemplars produced *fish, horse, bird* as their responses 4–6. Generalising beyond this example, we found a relatively large Label Advantage for *mammals*. This result is not necessarily incompatible with the words-as-labels perspective, if we assume that speakers in Experiment 1 know the concept MAMMALS and the concept ANIMALS, and the exemplars *dog, cat, human* activate both.

We do not think, however, that this type of ambiguity is the only mechanism explaining the advantage of labels over exemplars. For one, not all terms have such clear lexical competitors as *mammals* does in *animals*. *Toys*, for example, showed a relatively large Label Advantage, but it is not obvious which term would be activated by the list *truck, doll, blocks* if not *toys*.[6] Second, the responses 4–6 of individuals in the Exemplar condition were more similar to responses 1–3 than were the responses 4–6 of individuals in the Label condition (see Section 2.8.4). This suggests that individuals in the Exemplar condition were not simply inducing more general categories than individuals in the Label condition were. Consider, for example, a person in the Label condition who was given the cue *appetizers* and produced *breadsticks, mozzarella sticks, soup* as responses 1–3, and a person in the Exemplar condition who produced *spaghetti, salad, sandwich* as a continuation of this category. The second set of responses is thematically related to the first (casual Italian-American restaurant fare), although not all of the second person's responses are appetizers. This individual does not appear to be simply activating a more general, lexicalised category such as *food*. Rather, for at least some Exemplar trials, participants appear to be spontaneously generating ad hoc categories that may or may not be encoded linguistically.

Given the ambiguity of exemplar lists, the earlier findings of Barsalou (1982) and Ross and Murphy (1999) that pairs of basic level items appeared to activate the same categories as superordinate terms may strike readers as surprising. The variation that we observed across superordinate terms may help to explain these earlier results. Across Experiments 1 and 2, we found little evidence for a Label Advantage for *fruit, vegetables, clothing, tools*, or *animals*. This suggests, for example, that if people read the list *banana, orange, grapes*, they may in fact activate a category with the same intension as *fruit*. This is even more notable because the words *food* and *produce* are available as alternate lexicalizations of this category, but participants do not appear to activate those alternatives. The most atypical fruits produced by people in the Exemplar condition were *honeydew*, *kumquat*, and *coconut* – people never produced responses such as *avocado* or *carrot*, let alone *bread* or *eggs*. The Label and Exemplar conditions impose different task demands: in the latter, participants have to identify relationships between the exemplars, form a category, and then list three more members. The absence of a Label Advantage for some terms suggests that even with these task demands, participants constructed similar categories as denoted by the labels themselves.

Importantly, the variation that we observed across individual terms was not random noise. In the analysis of typicality ratings in Experiment 1, we found that the Label Advantage was greater for less frequent terms and for more specific terms. The interaction with frequency suggests that when a label is more common in language usage, participants may be more likely to activate it when viewing a list of exemplars, reducing the Label Advantage. Alternatively, frequency in language may be a proxy for the prominence of the category in the physical and cultural environment (e.g. we more often encounter images of different fruits grouped together than images of different appetizers grouped together). Participants may be more likely to activate culturally prominent categories when viewing

exemplars, reducing the Label Advantage. These explanations are not mutually exclusive. The interaction with generality may reflect degree of ambiguity of the exemplar lists – that for more specific categories, there are more alternative terms that could be used to construct the category (e.g. *mammals → animals; appetizers → food*). In the label guessing experiment (Experiment 2), we found that the Label Advantage was greater for *more* frequent terms, the reverse effect from Experiment 1. This may be the result of a floor effect, as guessing accuracy was lower for less frequent terms. For *appetizers*, *mammals*, and *pets* (all lower frequency terms), accuracy was less than 20% across all three Cue Types. The fact that accuracy was low even in the Label condition may have made it difficult to observe a Label Advantage. Understanding what factors predict the Label Advantage is an important avenue for future research, as we discuss in more detail in Section 4.3.

4.2. Labels vs. definitions

In addition to exemplar lists, we investigated whether definitions can be used to activate the same categories as the corresponding superordinate terms. Definitions are arguably not ambiguous in the same way as exemplar lists. Multiple theorists have also argued that if a language happens to not express a category through a single word, other linguistic resources will allow language users to communicate the category. Definitions are assumed to play a key role in fueling this expressive power – if someone doesn't know what *appetizers* are, they can be told that these are "foods that are eaten before the main course". At the same time, for students of linguistics and cognitive science, one of the earliest lessons is the difficulty of defining words (see Elbourne, 2011). Despite the theoretical importance of this issue, few empirical studies have directly compared the categories activated by superordinate nouns vs. definitions. We did so in our study and found in general that term definitions activated different categories than the terms themselves.

As in the Exemplar condition, we found considerable variation across terms. For *toys*, *clothing*, *pets*, and *mammals*, the Label Advantage was minimal – the definitions we chose appeared to activate similar categories as the terms. However, definitions of many terms, especially *fruit*, *games*, *furniture*, and *vegetables*, were much less effective than the simple conventional labels. Our definitions were similar in quality to dictionary definitions and including definition quality (relative difference between Turker-produced vs. dictionary definitions) as a covariate did not decrease the Label Advantage. This result suggests that many meanings are difficult to capture in a succinct definition – even dictionary definitions may fail to adequately express a word's meaning. More frequent terms showed a greater Label Advantage, suggesting that more frequent terms are more difficult to define. The size of the Label Advantage did not vary with the generality/specificity of the term.

Demonstrating a semantic difference between labels and seemingly reasonable definitions has cultural as well as theoretical relevance. For example, in discussions of how to adapt gendered language to be inclusive of transgender and gender-nonconforming individuals, some have suggested substituting definitions for labels (e.g. "child identified female at birth" instead of *girl*) (Cordoba, 2022). Little research has been conducted on how people understand definitions as similar or different from labels, and our study takes a step towards addressing this gap.

4.3. The relationship between words and concepts revisited

If words are merely labels for concepts, then it should be possible in principle to activate these concepts through means other than the word. We attempted to activate superordinate concepts using two types of non-label cues: exemplar lists and definitions. For many terms, we found that neither of these cues was a sufficient substitute. This is not an in-principle restriction – for terms such as fruit, vegetables, and clothing, exemplar lists do in fact appear to be sufficient to activate a category with the same extension as the word. For this reason, the Label Advantage that we observed for many other terms (e.g. *appetizers, hobbies, games*) suggests that those particular words play an important role when people are processing superordinate categories. In the absence of conventional labels, the region of conceptual space corresponding to *appetizers* does not appear to be as prominent as the region of conceptual space corresponding to *fruit*. Compared to categories cued through exemplar lists, the word *appetizers* increases semantic alignment across participants to a greater degree than the word *fruit* does. Thus some words are more influential than others in guiding how participants identify relevant categories.

We tested 20 superordinate terms in this study, an improvement over many previous studies. Nonetheless, 20 terms are insufficient to gain a precise understanding of what factors enhance or diminish the Label Advantage, or how multiple factors interact with each other. We found that frequency and generality of the terms predicted some of our dependent measures, but a variety of other factors may also play a role, including:

degree of perceptual similarity of category members, degree of functional similarity of category members, and the age at which children learn the categories/terms. For those dependent measures where frequency and generality did not interact with Cue Type, it is unclear whether these covariates are truly not relevant, or whether we lacked the power to detect an effect. Future studies may successfully address these questions. Our study shows the importance of sampling a diverse range of terms when designing studies on concepts and categories, as not all categories are alike.

If a researcher can establish that a word functions as a label, then the phrase "the concept WORD", might well be appropriate. This is a high bar to clear, as even for those terms where we did not observe a large Label Advantage (in the Exemplar condition: *fruit, vegetables, clothing, animals*; in the Definition condition: *clothing, mammals, toys, pets*), the terms could still be influencing people's performance in the task. That is, if an exemplar list or definition activates an ambiguous or unclear category, then participants might draw on their word knowledge to help them formulate the category. To be sure, exemplar lists and definitions are only two ways of activating superordinate concepts, and different methods might reveal stronger evidence that words function as labels.

While different languages lexicalise perceptual/conceptual space in a variety of ways, this variability is not random. Instead, certain dimensions are reflected relatively consistently across languages (Kemmerer, 2019; Kemp et al., 2018; Majid et al., 2008; Malt et al., 2008; Rissman, Horton, et al., 2023; Xu et al., 2020; Youn et al., 2016). For example, in descriptions of tool use events, child homesigners and adult speakers of English, Spanish, and Chinese make the tool linguistically prominent for similar types of events (Rissman, Horton, et al., 2023). This suggests shared ways of categorising the world which are likely to be reflected in language. Few systematic studies have been conducted on variation in superordinate noun meaning across languages. An important question is whether those categories for which we observed a minimal Label Advantage are also more commonly lexicalised across the world's languages (i.e. they might reflect more perceptually prominent or culturally common ways of organising the world).

In light of our findings and the extensive literature on variation in cross-linguistic semantics reviewed earlier, researchers would do well to avoid referring to concepts using words (as suggested by Malt, 2019), unless they cite empirical evidence that the word functions as a label for an independently represented concept. Using words to communicate about concepts may of course be a practical necessity – how else would we describe the categories of horses and cats that 3-month-olds construct? Nonetheless, in our lectures and writing, researchers should more explicitly signal whether concepts are being labeled for the sake of communication, or whether specific evidence is being presented that word meanings are reducible to conceptual structure. In the absence of such evidence, we suggest that distinguishing concepts from lexical semantics should be the default practice.

Notes

1. Although the convention of writing requires us to describe these categories using language, we do not assume that people require language to represent any of these categories.
2. Participants were instructed to not consult dictionary definitions.
3. The three definitions for *vegetables* all received relatively low mean ratings (less than 70). We therefore constructed a definition for *vegetables* that we judged to be more appropriate.
4. An alternative solution is to use embeddings from contextual models such as BERT and its variants. However, this necessitates additional assumptions about what context to use when generating the embedding of the target word(s).
5. Adding Participant by Response Number random slopes prevented the models from converging.
6. The individual in the Exemplar condition who received *truck, doll, blocks* as their cue produced the list *books, action figures, animals* as their responses 4–6. With respect to *toys*, these responses had typicality ratings of 2.2, 7.6, and 2.8, respectively.

Acknowledgments

This research was supported by NSF PAC 2020969 awarded to G. Lupyan. Thank you to Kevin Mui for JavaScript help. Thank you to Cognitive Science Society reviewers and to members of the Lupyan Lab for their feedback on this work. Thank you to all study participants.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by NSF PAC awarded to G. Lupyan, Division of Behavioral and Cognitive Sciences [grant number: 2020969].

ORCID

Lilia Rissman http://orcid.org/0000-0002-3796-2719
Gary Lupyan http://orcid.org/0000-0001-8441-7433

References

Athanasopoulos, P., & Casaponsa, A. (2020). The Whorfian brain: Neuroscientific approaches to linguistic relativity. *Cognitive Neuropsychology*, *37*(5–6), 393–412. https://doi.org/10.1080/02643294.2020.1769050

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*(1), 82–93. https://doi.org/10.3758/BF03197629

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227. https://doi.org/10.3758/BF03196968

Barsalou, L. W. (1991). Deriving categories to achieve goals. *Psychology of Learning and Motivation*, *27*, 1–64. https://doi.org/10.1016/S0079-7421(08)60120-6

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using S4 classes*. R package version 1.1-7. http://CRANR-project.org/package=lme4.

Beavers, J., & Koontz-Garboden, A. (2020). *The roots of verbal meaning*. Oxford University Press.

Bierwisch, M., & Schreuder, R. (1992). From concepts to lexical items. *Cognition*, *42*(1–3), 23–60. https://doi.org/10.1016/0010-0277(92)90039-K

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*(1), 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Bolognesi, M. (2020). *Where words get their meaning: Cognitive processing and distributional modelling of word meaning in first and second language* (Vol. 23). John Benjamins Publishing Company.

Bolognesi, M., & Caselli, T. (2022). Specificity ratings for Italian data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01974-6

Chrysikou, E. G. (2006). When shoes become hammers: Goal-derived categorization training enhances problem-solving performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 935. https://doi.org/10.1037/0278-7393.32.4.935

Cordoba, S. (2022). *Non-binary gender identities: The language of becoming*. Routledge, Taylor & Francis.

Davies, M. (2023). *The Corpus of Contemporary American English: 450 million words, 1990-present*. http://corpus.byu.edu/coca/.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, *65*(3), 903–917. https://doi.org/10.2307/1131427

Elbourne, P. (2011). *Meaning: A slim guide to semantics*. Oxford University Press.

Enfield, N. (2022). *Language vs. Reality: Why language is good for lawyers and bad for scientists*. MIT Press.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, *148*(7), 1105–1123. https://doi.org/10.1037/xge0000560

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738. https://doi.org/10.1111/j.1551-6709.2009.01026.x

Grimshaw, J. B. (2005). *Words and structure: Center for the study of language and information*. Stanford University.

Jackendoff, R. (1983). *Semantics and cognition* (Vol. 8). MIT Press.

Keller, R. (1998). *A theory of linguistic signs*. Oxford University Press.

Kemmerer, D. (2019). *Concepts in the brain: The view from cross-linguistic diversity*. Oxford University Press.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109–128. https://doi.org/10.1146/annurev-linguistics-011817-045406

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford University Press.

Levinson, S. C. (1997). From outer to inner space: Linguistic categories and non-linguistic thinking. In J. Nuyts, & E. Pederson (Eds.), *Language and conceptualization* (pp. 13–45). Cambridge University Press.

Levinson, S. C., & Wilkins, D. P. (2006). *Grammars of space: Explorations in cognitive diversity* (Vol. 6). Cambridge University Press.

Lewis, M., Colunga, E., & Lupyan, G. (2021). Superordinate word knowledge predicts longitudinal vocabulary growth. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 321–326.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*(2), 566–577. https://doi.org/10.1016/j.cognition.2008.03.009

Lupyan, G. (2017). The paradox of the universal triangle: Concepts, language, and prototypes. *Quarterly Journal of Experimental Psychology*, *70*(3), 389–412. https://doi.org/10.1080/17470218.2015.1130730

Lupyan, G., Abdel Rahman, R., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, *24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005

Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception, & Psychophysics*, *72*(8), 2236–2253. https://doi.org/10.3758/BF03196698

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170. https://doi.org/10.1037/a0024904

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*(2), 235–250. https://doi.org/10.1016/j.cognition.2008.08.009

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, , B., de Sousa, H., Cansler, B. L., Shayan, S., de Vos, C., Senft, G., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., Brown, P., … Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376. https://doi.org/10.1073/pnas.1720419115

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85–148. https://doi.org/10.1006/cogp.1995.1013

Malt, B. C. (2019). Words, thoughts, and brains. *Cognitive Neuropsychology*, 1–13. https://doi.org/10.1080/02643294.2019.1599335

Malt, B. C. (2024). Representing the world in language and thought. *Topics in Cognitive Science*, *16*, 6–24.

Malt, B. C., Gennari, S., Imai, M., Ameel, E., Tsuda, N., & Majid, A. (2008). Talking about walking: Biomechanics and the language of locomotion. *Psychological Science*, *19*(3), 232–240. https://doi.org/10.1111/j.1467-9280.2008.02074.x

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *WIRES Cognitive Science*, *4*(6), 583–597. https://doi.org/10.1002/wcs.1251

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262. https://doi.org/10.1006/jmla.1998.2593

Martin, A. (2016). Grapes—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, *23*(4), 979–990. https://doi.org/10.3758/s13423-015-0842-3

Mauri, C. (2017). Building and interpreting ad hoc categories: A linguistic analysis. In J. Blochowiak, C. Grisot, S. Durrleman, & C. Laenzlinger (Eds.), *Formal models in the study of language: Applications in interdisciplinary contexts* (pp. 297–326). Springer International Publishing.

Mauri, C., & Sansò, A. (2018). Linguistic strategies for ad hoc categorization: Theoretical assessment and cross-linguistic variation. *Folia Linguistica*, *52*(39), 1–35. https://doi.org/10.1515/flih-2018-0001

Murphy, G. L. (2002). *The big book of concepts*. MIT Press.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. https://doi.org/10.1038/nrn2277

Pomerantz, A. M. (2023). *My psychology* (3rd ed.). Worth Publishers.

Pomiechowska, B., & Gliga, T. (2018). Lexical acquisition through category matching: 12-month-old infants associate words to visual categories. *Psychological Science*, *30*(2), 288–299. https://doi.org/10.1177/0956797618817506

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Rappaport-Hovav, M., & Levin, B. (2010). Reflections on manner/result complementarity. In E. Doron & I. Sichel (Eds.), *Syntax, lexical semantics, and event structure* (pp. 21–38). Oxford University Press.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rissman, L., Horton, L., & Goldin-Meadow, S. (2023). Universal constraints on linguistic event categories: A cross-cultural study of child homesign. *Psychological Science*, *34*(3), 09567976221140328. https://doi.org/10.1177/09567976221140328

Rissman, L., Liu, Q., & Lupyan, G. (2023). Gaps in the lexicon restrict communication. *Open Mind*, 1–23.

Rissman, L., van Putten, S., & Majid, A. (2022). Evidence for a shared instrument prototype from English, Dutch and German. *Cognitive Science*, *46*(5), e13140. https://doi.org/10.1111/cogs.13140

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495–553. https://doi.org/10.1006/cogp.1998.0712

Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, *111*(1), 65–86. https://doi.org/10.1016/j.jecp.2011.07.007

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695–1709. https://doi.org/10.1111/1467-8624.00373

Snodgrass, J. (1984). Concepts and their surface representations. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 3–22. https://doi.org/10.1016/S0022-5371(84)90479-1

Suffill, E., van Paridon, J., & Lupyan, G. (2022). Verbal labels increase conceptual alignment. *Proceedings of the Joint Conference on Language Evolution*, 691–698.

Suffill, E., van Paridon, J., & Lupyan, G. (under review). *Mind melds: Verbal labels induce greater representational alignment*.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description*, *3*, 57–149.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029–1038. https://doi.org/10.1038/s41562-020-0924-8

Tong, J., Binder, J. R., Humphries, C. J., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *The Journal*

of Neuroscience, *42*(37), 7121. https://doi.org/10.1523/JNEUROSCI.1243-21.2022

van Paridon, J., & Thompson, B. (2020). Subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, *53*(2), 629–655. https://doi.org/10.3758/s13428-020-01406-3

Waxman, S. R., & Gelman, R. (1986). Preschoolers' use of superordinate relations in classification and language. *Cognitive Development*, *1*(2), 139–156. https://doi.org/10.1016/S0885-2014(86)80016-8

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369* (1634), 20120391. https://doi.org/10.1098/rstb.2012.0391

Wisniewski, E. J., Imai, M., & Casey, L. (1996). On the equivalence of superordinate concepts. *Cognition*, *60*(3), 269–298. https://doi.org/10.1016/0010-0277(96)00707-X

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *WIRES Cognitive Science*, *2*(3), 253–265. https://doi.org/10.1002/wcs.104

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280. https://doi.org/10.1016/j.cognition.2020.104280

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., & Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771. https://doi.org/10.1073/pnas.1520752113

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135. https://doi.org/10.1016/j.cognition.2019.104135