COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13055

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13055

Linguistic Distributional Knowledge and Sensorimotor Grounding both Contribute to Semantic Category Production

Briony Banks,[a] Cai Wingfield,[a] Louise Connell[a,b]

[a]*Department of Psychology, Lancaster University*
[b]*Department of Psychology, Maynooth University*

Received 17 December 2020; received in revised form 22 June 2021; accepted 9 September 2021

Abstract

The human conceptual system comprises simulated information of sensorimotor experience and linguistic distributional information of how words are used in language. Moreover, the linguistic shortcut hypothesis predicts that people will use computationally cheaper linguistic distributional information where it is sufficient to inform a task response. In a pre-registered category production study, we asked participants to verbally name members of concrete and abstract categories and tested whether performance could be predicted by a novel measure of sensorimotor similarity (based on an 11-dimensional representation of sensorimotor strength) and linguistic proximity (based on word co-occurrence derived from a large corpus). As predicted, both measures predicted the order and frequency of category production but, critically, linguistic proximity had an effect above and beyond sensorimotor similarity. A follow-up study using typicality ratings as an additional predictor found that typicality was often the strongest predictor of category production variables, but it did not subsume sensorimotor and linguistic effects. Finally, we created a novel, fully grounded computational model of conceptual activation during category production, which best approximated typical human performance when conceptual activation was allowed to spread indirectly between concepts, and when candidate category members came from both sensorimotor and linguistic distributional representations. Critically, model performance was indistinguishable from typical human performance. Results support the linguistic shortcut hypothesis in semantic processing and provide strong evidence that both linguistic and grounded

All three authors contributed equally to the work.

Correspondence should be sent to Briony Banks, Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster, LA1 4YF, UK. E-mail: b.banks@lancaster.ac.uk; or to Louise Connell, Department of Psychology, Maynooth University, Maynooth Co., Kildare, Ireland. E-mail: louise.connell@mu.ie

representations are inherent to the functioning of the conceptual system. All materials, data, and code are available at https://osf.io/vaq56/.

Keywords: Concepts; Category production; Semantic fluency; Sensorimotor simulation; Linguistic distributional information; Computational cognitive model

1. Introduction

The ability to carve up the world into categories—to group like with like and treat distinct entities as equivalent—is arguably the most fundamental process in cognition (Harnad, 2006). It allows us to infer similarities based on category membership and hence make predictions about objects and events in the world around us. Critically, categories lend structure to our conceptual system (i.e., semantic memory), and so a proper understanding of categories rests on understanding how concepts are represented in the conceptual system.

Current theories of the human conceptual system hold that both simulated information (i.e., partial replays of sensorimotor, affective, and other experience) and linguistic distributional information (i.e., statistical relationships of how words and phrases co-occur in language) are critical to conceptual representation and processing (Barsalou, Santos, Simmons, & Wilson, 2008; Connell, 2018; Connell & Lynott, 2014b; Louwerse & Jeuniaux, 2008; Vigliocco, Meteyard, Andrews, & Kousta, 2009). Our knowledge of concepts in long-term semantic memory (e.g., "what is a *cat*?") is largely formed through sensorimotor experience of the world (e.g., seeing, hearing, or stroking a cat; Barsalou, 1999), which provides grounding to both abstract and concrete concepts (Connell & Lynott, 2012; Connell, Lynott, & Banks, 2018). This modality-specific experience can then later be simulated during conceptual processing. For instance, neural regions active during perception in a particular modality (e.g., auditory cortex for processing sounds) are also recruited for processing words whose meaning relates to that modality (e.g., *thunder*: Bonner & Grossman, 2012; also see Hauk, Johnsrude, & Pulvermüller, 2004). Behavioral studies have further demonstrated complex interactions between the sensorimotor information that is processed during perception/action and that simulated during conceptual processing (Connell & Lynott, 2010, 2014a; Connell, Lynott, & Dreyer, 2012; Dils & Boroditsky, 2010; Zwaan & Taylor, 2006). For example, when people passively hold a ball between their hands, judgments about object size are faster for manipulable objects, compared to non-manipulable objects (Connell et al., 2012), suggesting a functional role for perceptual systems in conceptual processing of objects that is consistent with the sensorimotor experience of those objects. It is critical to note that sensorimotor information is not restricted to concrete concepts like *cat* but also underpins the representation and processing of abstract concepts like *justice* (e.g., Connell & Lynott, 2012; Lynott, Connell, Brysbaert, Brand, & Carney, 2020). Indeed, some forms of perceptual experience—particularly interoception (i.e., sensations inside the body)—are demonstrably more important to the representation of abstract concepts than concrete (Connell et al., 2018).

However, concepts are also formed through our experience with language. Linguistic distributional knowledge comprises the words and phrases that label concepts and the statistical

distributions of how they occur together, and is critical to conceptual representation and processing (Connell, 2018; Louwerse, 2011). Linguistic distributional information about "cat," for example, includes other words that appear in the same or similar contexts, such as "kitten," "mouse," and "food." The role of linguistic distributional information has been demonstrated in a wide range of semantic tasks (Goodhew, McGaw, & Kidd, 2014; Louwerse & Connell, 2011; Louwerse & Jeuniaux, 2010; for a review, see Wingfield & Connell, 2020). For instance, the frequency with which two words co-occur in the same context predicts how quickly they can be understood as a novel conceptual combination, as well as how often the conceptual combination process is likely to succeed (Connell & Lynott, 2013). Even semantic decision—that is, the classification of a given word as abstract or concrete—can be predicted by how closely the context of the target word resembles the contexts of the decision options (e.g., whether the contexts of "cat" and "concrete" are more similar than the contexts of "cat" and "abstract"; Wingfield & Connell, 2020).

Both sensorimotor and linguistic distributional information typically interact to drive conceptual processing, depending on the exact context or cognitive task (Connell, 2018). Although sensorimotor simulation can provide a more detailed, precise conceptual representation when required, linguistic distributional information may provide a quicker and more efficient way of processing concepts, compared to sensorimotor simulation (Barsalou et al., 2008; Connell & Lynott, 2014b; Louwerse, 2011). In certain circumstances, linguistic distributional information could therefore be used as a response heuristic—that is, a linguistic shortcut—when a rapid representation of a concept in the form of its linguistic label is sufficient for the task in question (Connell, 2018; also see Connell & Lynott, 2013, 2014b; Lynott & Connell, 2010). Hence, although concept labels are ultimately grounded in sensorimotor simulation, a given label does not *have* to be grounded every time it is processed (Louwerse & Connell, 2011).

1.1. Category production

A common way of testing how concepts are structured and accessed from long-term memory is with a category production task (also called semantic or verbal fluency), whereby a participant is presented with a category label such as ANIMAL,[1] and asked to name concepts belonging to that category (Battig & Montague, 1969; Cohen et al., 1957; McEvoy & Nelson, 1982; Van Overschelde, Rawson, & Dunlosky, 2004). Despite such frequent use in cognitive research, as well as in neuropsychological settings as a clinical tool (e.g., Cerhan et al., 2002), the mechanisms driving responses in category production tasks are not well understood. Traditionally, the process of listing members of a given category is assumed to reflect access to taxonomic categorical structures in semantic memory (e.g., Rosch, Simpson, & Miller, 1976; Warrington & McCarthy, 1987). There is also some evidence that the process involves using a controlled search that relies on executive function and working memory (Baddeley, Lewis, Eldridge, & Thomson, 1984; Rosen & Engle, 1997; Unsworth, Brewer, & Spillers, 2013). However, these accounts do not fully explain how or why particular concepts are accessed and selected over others.

Sensorimotor and linguistic distributional information may offer such an explanation, as they contain useful categorical information which could dynamically drive responses in category production. In terms of linguistic distributional information, there is already some evidence that the relationship between category members and category labels (e.g., between *cat* and ANIMAL) in corpus-derived linguistic space is an effective predictor of category membership (Connell & Ramscar, 2001; Riordan & Jones, 2011).[2] That is, because the words "cat" and "animal" share more linguistic contexts than do "bat" and "animal," *cat* may be named as a member of the category ANIMAL more readily than *bat*.

In terms of sensorimotor information, many theories of conceptual structure favor a process whereby categorical distinctions emerge from commonalities in the way we perceive and interact with the world around us. For instance, because our experience with cats and dogs tends to involve a haptic experience of fur, visuo-haptic experience of a four-legged body shape, and so forth, they and other similar concepts tend to group together to form the category of ANIMAL. Some of the strongest evidence for such emergent structure comes from computational models that define concepts in terms of abstracted feature sets, such as *cat* [has_fur, miaows, is_independent] (McRae, de Sa, & Seidenberg, 1997; Tyler, Moss, Durrant-Peatfield, & Levy, 2000). However, such models tend to include features that are too abstract to have a clear sensorimotor correspondence (e.g., *clock* [used_to_tell_time]). Thus, even though concepts may group together according to the similarity of their featural representations, there is currently a lack of evidence that sensorimotor experience alone (i.e., without abstracted features) is enough to accomplish it. Nonetheless, because the sensorimotor experience of *cats* tends to be quite similar to that of *animals*, or at least more so than *bats* and *animals*, we propose that is why *cat* may be named as a member of the category ANIMAL more readily than *bat*.

1.2. The current study

In the present paper, we report a two-part study that investigated the role of linguistic distributional and sensorimotor information in predicting the rank order, frequency, and time course of responses in a category production task. To do so, we used a corpus-based measure of linguistic distributional information, based on co-occurrence frequencies in a large corpus of English, to create a measure of *linguistic proximity* between the category name and member name (e.g., between ANIMAL and *cat*). We also used a novel measure of sensorimotor similarity, based on ratings of sensorimotor strength across 11 different dimensions (six perceptual modalities and five action effectors) from Lynott et al. (2020), to create a measure of *sensorimotor similarity* between a category and member concept (e.g., between ANIMAL and *cat*). We hypothesized that, when presented with a category name (e.g., ANIMAL), people would name member concepts (e.g., *dog, cat, horse*…) more often and earlier in a list when they were (a) more similar in sensorimotor experience to the category concept, and (b) more often encountered in proximity to the category name in linguistic contexts. We first report a pre-registered behavioral experiment of category production and a follow-up pre-registered examination of typicality ratings based on these category production responses and subsequently report a novel computational model to test the importance of indirect (spreading)

activation across sensorimotor and linguistic distributional representations in fitting human performance.

2. Experiment 1a: Category production

In this pre-registered experiment (pre-registration, data, analysis code, and results are available at https://osf.io/vaq56/), we predicted that both sensorimotor similarity and linguistic proximity would contribute to explicit and implicit measures of category production (i.e., how often each member concept is listed for a particular category, how early in a list each member is named, how often each member is listed first, and how quickly the first member is listed). In addition, following predictions of the linguistic shortcut hypothesis that people will make use of word-to-word distributional information when it is sufficient for the task in question, we expected linguistic proximity would dominate, uniquely predicting responses over and above sensorimotor similarity.

2.1. Method

2.1.1. Participants

Sixty-four participants recruited from Lancaster University took part for payment of £3.50. Three participants were excluded as they were non-native speakers of English (i.e., after debriefing revealed that they had misunderstood the screening criteria), and one was excluded for providing too few responses ($M < 2$ responses per category). Of the remaining 60 participants, all had English as their native language, 46 were female, the mean age was 21.72 years ($SD = 5.73$), and 52 were right-handed. The sample size was determined by sequential hypothesis testing with Bayes factors (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; see pre-registration) in JASP (version 0.14; JASP Team, 2020) using the default JZS prior with a scale parameter of $r = 1$ for the effect size prior. We stopped data collection at $N\text{min} = 60$ when our Step 3 models for two of the three item-level dependent variables cleared the specified grade of evidence ($BF_{10} > 5$). For details of variables and models, see Section 2.1.5., statistical analyses; full statistics of sequential analysis are available as Supplemental Materials at https://osf.io/vaq56/.

2.1.2. Materials

We selected 117 categories that encompassed a range of specificity, such as basic level (e.g., BIRD), superordinate (e.g., ANIMAL), and subordinate (e.g., WATER BIRD) categories and a number of ways of splitting high-level conceptual domains, including abstract/concrete (e.g., EMOTION/ANIMAL), living/non-living (e.g., BIRD/BOAT), animate/inanimate (e.g., FISH/FRUIT), artifact/natural (TOOL/FLOWER), and biological/nonbiological (e.g., VEGETABLE/FURNITURE) concepts. Most categories (both concrete and abstract) were selected from the categorization literature (e.g., Battig & Montague, 1969; Capitani, Laiacona, Mahon, & Caramazza, 2003; Larochelle, Richard, & Soulières, 2000; McEvoy & Nelson, 1982; Rosch, 1975; Uyeda & Mandler, 1980; Van Overschelde et al., 2004). We selected concrete

categories that have been frequently investigated in studies of semantic categories (e.g., ANIMAL, FRUIT, MUSICAL INSTRUMENT), as well as others from both concrete and abstract domains that have featured less frequently (e.g., BIRD OF PREY, ROOM IN A HOUSE, CRIME, EMOTION, MILITARY TITLE). However, as few category production studies have included a large number of abstract categories, we also added some novel categories that (to the authors' knowledge) have not previously been studied in a category production task. Some of these novel abstract categories were subordinate categories (e.g., VIOLENT CRIME, NEGATIVE EMOTION) or modified variants (e.g., ROYAL TITLE) of those already selected from the literature, while others were created de novo by the authors from categorical distinctions in WordNet for abstract entities (e.g., FRACTION, SOCIAL GATHERING, PERSONAL QUALITY). All categories were piloted on participants not involved in the main experiment to ensure that they were easily understood. Categories were divided into three lists of 39 categories each, counterbalanced as much as possible across the categorical distinctions described above. Categories that constituted a subset of another category (e.g., WATER BIRD, BIRD) were not included in the same stimulus list. Four categories (BREAD, CIRCUS ACT, FOOTWEAR, and CONTINENT) that were not featured in the main experiment were used as practice items.

2.1.3. Procedure

Participants triggered the start of each trial by pressing the space bar on the keyboard. They were then presented with a fixation cross for 500 ms followed by the category name presented in capital letters in the center of the screen. They were instructed to name aloud as many concepts as possible that belonged to each category, within a maximum of 60 s. The category name was displayed on the screen until participants could not name any more concepts and pressed the space bar to end the trial or until the trial timed out automatically after 60 s. The words "Press space bar when ready" were then displayed on the screen until participants triggered the next trial; timing between categories was thus self-paced, and participants could take a short break between categories if required. Participants first carried out four practice trials and were then randomly assigned to one of the three category lists. Each list was presented to 20 participants, and categories from each list were presented in a randomized order for each participant. Verbal responses were audio-recorded through a unidirectional headset microphone. Verbal responses were transcribed during the task by the experimenter (hidden from the participant's view behind a panel screen) and later verified via audio recordings. The entire experimental procedure took approximately 20 min, after which participants provided demographic information and were debriefed by the experimenter.

2.1.4. Ethics and consent

The study received ethical approval from the Lancaster University Faculty of Science and Technology Research Ethics Committee. All participants read information detailing the purpose and expectations of the study before giving informed consent to take part. Consent included an agreement to share publicly all transcribed and alphanumeric data in anonymized form; participants could additionally opt-in to sharing publicly their original voice recordings with anonymized filenames (52 out of 60 participants consented to do so).

2.1.5. Data preparation, design, and analysis

All transcribed responses used British English spellings, as the linguistic proximity measure was based on a British English linguistic corpus (see below). Unintelligible responses were disregarded. Idiosyncratic responses (i.e., member concepts named by only one participant) were excluded from analysis (22% of the participant-level data; 3890 of 17,707 responses), resulting in an average of 6.18 member concepts per category ($SD = 3.90$) per individual participant, and a total of 2551 distinct category–member pairs. A further 322 pairs were excluded from analysis due to missing values on predictor variables (i.e., log word frequency—LgSUBTLWF, sensorimotor similarity, linguistic proximity), and one pair was excluded where the member concept comprised a repetition of the category (*relationship* for the category SOCIAL RELATIONSHIP); thus, a total of 2228 pairs were analyzed, comprising an average of 19.04 member concepts per category ($SD = 10.59$; range = [5, 64]). Response times were measured from the onset of the category name until the onset of speech to name the first member concept; disfluencies were disregarded and speech onset was calculated in Praat (Boersma & Weenink, 2018).

2.1.5.1. Linguistic proximity. To operationalize linguistic distributional information, we calculated a measure of linguistic proximity based on word co-occurrence between each category name (e.g., BIRD) and each member concept named by participants (e.g., *pigeon*). Using a corpus of 200 million words of British English television and film subtitles[3] (see van Heuven, Mandera, Keuleers, & Brysbaert, 2014), we counted 6-gram co-occurrence frequencies for each category–member word pair (e.g., how often the word BIRD appeared in the same context as *pigeon*, with zero, one, two, three, four, and five intervening words) and calculated the positive pointwise mutual information score (PPMI; Bullinaria & Levy, 2007), which reflects only co-occurrences that are more frequent than expected.[4] Linguistic proximity is therefore a measure of first-order co-occurrence, reflecting the extent to which two words co-occur more often than chance, which can capture a complex range of semantic relations (Sahlgren, 2006; Wingfield & Connell, 2020).

For multiword category names (e.g., WATER BIRD, SOCIAL GATHERING), we created a composite representation by using a multiplicative function (Mitchell & Lapata, 2010) to combine individual *n*-gram distributions before calculating co-occurrence with member concepts; function words were not included in composite representations (e.g., for the category DAY OF THE WEEK, we used DAY and WEEK only). The composite representation for the category SOCIAL GATHERING, for example, therefore comprised high PPMI scores for words that co-occurred more frequently than expected with *both* SOCIAL and GATHERING, but scores of zero for words that co-occurred frequently with only *one* of SOCIAL or GATHERING or that co-occurred less frequently than expected for both. For multiword member concepts, we first minimized distortions in our measure by removing repetitions of the category name from any responses that included it (e.g., for the category TREE, the response *oak tree* was shortened to *oak*) and by removing words that the experimenters judged to be redundant to the core meaning of the concept (e.g., for the category SOCIAL GATHERING, the response *going to the cinema* was shortened to *cinema*). For any remaining multiword member concepts, we measured how well the category name cued the individual terms in

the member concept by calculating separate co-occurrence frequencies for each word in the member name and applying an additive function (e.g., for the category SPORT and the member *ice hockey*, we summed the 6-gram counts of SPORT–*ice* and SPORT–*hockey*). Plural and singular responses for the same member concept (e.g., *cat*, *cats*) were counted as separate lexical items due to their potentially different distributional patterns. As described above, any responses that did not appear in the corpus were excluded from analysis; one member concept of *bobsledging* (sic; listed for WINTER SPORT) was replaced with the conventional form *bobsledding*. The final linguistic proximity measure for each category–member pair ranged in theory from 0 to infinity (actual range = [0.00, 78.03], $M = 5.67$, $SD = 9.70$), with higher values indicating a greater tendency for the words to appear in proximity to one another (i.e., higher frequency of co-occurrence). For instance, in the category ANIMAL, the lowest linguistic proximity score was for *ant* (0.00) and the highest was for *cheetah* (5.33).

2.1.5.2. Sensorimotor similarity. To operationalize a measure of sensorimotor similarity that was fully grounded in perceptual and action experience alone (i.e., without the use of abstracted features), we took the novel approach[5] of calculating sensorimotor similarity based on multidimensional ratings of sensorimotor strength. We used Lynott et al.'s (2020) Lancaster Sensorimotor Norms for 40,000 concepts, in which people rated the extent to which they experienced a particular concept via six perceptual modalities (auditory, gustatory, haptic, interoceptive, olfactory, visual) and by performing an action with five action effectors (foot, hand, head, mouth, torso), where each dimension was separately rated along a scale from 0 (*not at all*) to 5 (*greatly*). Each concept was therefore represented by an 11-dimensional vector, and we calculated sensorimotor similarity based on the Minkowski distance of parameter 3 between the vectors of the category concept (e.g., BIRD) and each member concept (e.g., *pigeon*). Minkowski-3 distance is similar to Euclidean distance (which corresponds to Minkowski distance of parameter 2) but places less weight on low-value dimensions; we chose to use it here because—when measured from the origin for each concept vector—Lynott et al. (2020) found that it represented the best composite measure of sensorimotor strength for predicting semantic facilitation in word recognition. To convert the Minkowski-3 distance measure to a more intuitive measure of sensorimotor similarity on a 0–1 scale, we divided each distance value x by the maximum possible distance between vectors and subtracted it from 1: $1 - (x / 11.120)$. This linear transform did not affect statistical inferencing but did make the results more interpretable.

As these sensorimotor norms were based on American English, we made a number of substitutions to the original British English transcriptions of our category production dataset before extracting the relevant sensorimotor vector. British English spellings were substituted with American English (e.g., COLOUR changed to COLOR), and terms were substituted with alternatives where dialectal differences meant that the British English term was absent from the sensorimotor norms, and we were confident that the alternative term labeled the same referent concept (e.g., *courgette* changed to *zucchini*; *paracetamol* changed to *acetaminophen*). Plural forms were substituted with the singular if the plural was absent from the sensorimotor norms (e.g., *cats* changed to *cat*). Some multiword terms were present in the sensorimotor norms (e.g., CITRUS FRUIT). Where a multiword term was not present, we shortened it by

removing repetitions and redundancies as per the linguistic proximity measure and also by removing modifiers where the experimenters judged that the sensorimotor experience would be indistinguishable (e.g., *king cobra* changed to *cobra*; *one quarter* changed to *quarter*). For both category and member concepts, we then created a composite representation using the same multiplicative function as for the linguistic proximity measure. The final sensorimotor similarity measure for each category–member pair ranged in theory from 0 to 1 (actual range = [0.48, 0.96], $M = 0.81$, $SD = 0.08$, with higher values indicating greater similarity of sensorimotor experience (i.e., closer in sensorimotor space). For example, in the category ANIMAL, the lowest sensorimotor similarity score occurred for *butterfly* (0.641) and the highest for *cat* (0.899). Linguistic proximity and sensorimotor similarity correlated only weakly, $r = .05$.

2.1.5.3. Dependent variables. Four different dependent variables were extracted for analysis. Three were explicit item-level measures that are traditionally reported in category production research (e.g., Battig & Montague, 1969): *production frequency* (i.e., the number of participants who name a particular member concept within its category); *mean rank* (i.e., the mean ordinal position of a particular member concept within its category); and *first-rank frequency* (i.e., the number of participants who name a particular member concept *first* within its category). The final dependent variable was *response time (RT)* for the first-named member concept per category and participant, which represented an implicit measure of processing effort in category production.

2.1.5.4. Statistical analyses. For each item-level dependent variable (*production frequency*, *mean rank*, and *first-rank frequency*), we carried out Bayesian linear regressions in JASP (version 0.14; JASP Team, 2020) using JZS default priors (r scale = .354) and Bayesian adaptive sampling, and their null-hypothesis significance test (NHST) counterparts (i.e., ordinary least squares linear regressions), in three hierarchical steps. Step 1 comprised a baseline model of log word frequency of the named member concept (LgSUBTLWF measure from the English Lexicon Project: Balota et al., 2007), Step 2 added sensorimotor similarity of the category–member pair, and Step 3 added linguistic proximity of the category–member pair. Although not specified in the pre-registration due to an error of omission, we tested whether Step 2 increased model fit over and above Step 1 (i.e., whether sensorimotor similarity predicts category production as hypothesized). As pre-registered, we then tested whether Step 3 increased model fit over and above Step 2 (i.e., whether linguistic proximity independently predicts category production) both via NHST of R^2-change and via model comparison using Bayes factors. Due to very high Bayes factor values, we report natural log Bayes factors throughout for clarity. We report parameter estimates from the Step 3 ordinary least squares regression model.

For *RT*, we ran linear mixed-effects models in R using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) and calculated marginal R^2 using the MuMIn package (Barton, 2017). Standardized coefficients were calculated in R by running the relevant analyses using standardized variables. We excluded 31 individual trials for which response times were $>3SD$ from the participant mean. Step 1 comprised crossed random effects of participant and

Table 1
Sample categories and member concepts produced by participants in Experiment 1a

Category	Member Concepts
ANIMAL	Cat, dog, giraffe, lion, elephant, horse, cow, rabbit, tiger, pig, sheep, fish, guinea pig, hamster, snake, zebra, monkey, mouse, rat, bird, donkey, leopard, whale, cheetah, chicken, dolphin, ferret, penguin, shark, badger, deer, duck, fox, frog, gorilla, kangaroo, lizard, squirrel, ant, antelope, bat, bear, butterfly, camel, cheetahs, ducks, eagle, elephants, flamingo, gazelle, goose, gorillas, hedgehog, jaguar, panda, pigeon, pigs, polar bear, raccoon, rhino, seagull, swan, whales, wolf
BOAT	Ferry, sailing, yacht, canoe, rowing, ship, speedboat, cruise ship, fishing, cruise, dinghy, kayak, canal, lifeboat
CRIME	Murder, burglary, robbery, fraud, theft, rape, assault, arson, shoplifting, manslaughter, stealing, vandalism, embezzlement, hate, knife, possession of drugs, violence
SUPERNATURAL BEING	Ghost, vampire, werewolf, ghosts, zombie, spirits, witches, angel, devil, poltergeist, vampires, alien, demon, demons, god, spirit, zombies
TOOL	Hammer, screwdriver, drill, saw, spanner, wrench, nails, screws, chisel, pliers, axe, bolts, crowbar, knife, shovel
VEGETABLE	Broccoli, carrot, cabbage, cauliflower, lettuce, cucumber, spinach, aubergine, onion, peas, parsnips, swede, beetroot, Brussels sprouts, carrots, courgette, green beans, parsnip, pepper, potato, sprouts, asparagus, beans, butternut squash, kale, peppers, sweet potato, tomato, turnip, leek, potatoes, pumpkin, radish, tomatoes, turnips

Note. Member concepts are listed in order of descending production frequency, excluding idiosyncratic responses, and distinguishing plurals.

item (category) and fixed effects relating to lexical properties of the member concept that can influence the time required to speak a word aloud: log word frequency (LgSUBTLWF), phonological Levenshtein distance (PLD20; Yarkoni, Balota, & Yap, 2008), and the number of syllables. Step 2 added sensorimotor similarity as a fixed effect, and Step 3 added linguistic proximity. Models were compared using NHST likelihood ratio chi-square tests and Bayes factors calculated via *BIC* (Bayesian information criterion; e.g., Wagenmakers, 2007). Descriptive statistics and zero-order correlations for all variables are provided as Supplementary Materials (Table S1).

2.2. Results and discussion

Sample categories, and the list of member concepts produced by participants, can be seen in Table 1. While some categories attracted relatively consistent responses across participants (e.g., all participants listed *milk* as a DAIRY PRODUCT, and 19 out of 20 listed it first), most categories had highly diverse sets of responses. For example, the most popular BOAT was a *ferry* (joint with *sailing boat* and *yacht*), yet only eight out of 20 participants produced it as a member concept, and only one of those did so as a first response. Similarly, while participants produced a total of 64 different member concepts for ANIMAL, the average number produced by an individual participant was only 21 (range 6–29).

Table 2
Experiment 1a hierarchical linear regressions of each category production measure on sensorimotor and linguistic predictors

Step	Model Comparison	Total R^2	ΔR^2	F	Log BF
Category Production Frequency					
1	Baseline (lexical) model versus empty model (BF_{10})	.023		51.74***	22.37
2	Sensorimotor similarity versus baseline (BF_{21})	.043	.020	46.13***	19.97
3	Linguistic proximity + sensorimotor similarity versus sensorimotor similarity only (BF_{32})	.087	.044	108.30***	49.92
Mean Rank					
1	Baseline (lexical) model versus empty model (BF_{10})	.001		1.20	−2.45
2	Sensorimotor similarity versus baseline (BF_{21})	.071	.070	167.85***	77.69
3	Linguistic proximity + sensorimotor similarity versus sensorimotor similarity only (BF_{32})	.083	.013	30.95***	12.66
First-Rank Frequency					
1	Baseline (lexical) model versus empty model (BF_{10})	.040		28.38***	11.19
2	Sensorimotor similarity versus baseline (BF_{21})	.054	.013	9.40**	2.42
3	Linguistic proximity + sensorimotor similarity versus sensorimotor similarity only (BF_{32})	.090	.036	26.63***	10.77
RT			χ^2		
1	Baseline (lexical) model versus random effects model (BF_{10})	.005		8.11*	−7.29
2	Sensorimotor similarity versus baseline (BF_{21})	.005	.000	0.00	−3.78
3	Linguistic proximity + sensorimotor similarity versus sensorimotor similarity only (BF_{32})	.012	.007	8.39**	0.41

Note. ΔR^2 = change in R^2; log BF = natural log Bayes factor, where negative values indicate evidence for the null model and positive values indicate evidence for the alternative model: log BFs ≥ 3.00 (equivalent to BFs ≥ 20) constitute strong evidence; log BFs ≥ 1.10 (equivalent to BFs ≥ 3) constitute positive evidence; and log BFs between –1.10 and 1.10 (equivalent to BFs of 0.33 and 3) constitute equivocal evidence.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

2.2.1. Category production frequency

As predicted, both sensorimotor similarity and linguistic proximity predicted the frequency of naming a particular member concept for a category. Bayes factors (see Table 2) indicated very strong evidence both for the inclusion of sensorimotor similarity in Step 2 and the inclusion of linguistic proximity in Step 3 (i.e., above and beyond sensorimotor similarity); this pattern was consistent with NHST F tests of change in R^2. In the Step 3 model, higher sensorimotor similarity led to higher *production frequency* (i.e., people listed a member concept more often when its sensorimotor profile was very similar to the category concept), unstandardized $B = 7.46$, $SE = 1.15$, standardized ß = 0.13, $t = 6.50$, $p < .001$. Likewise, higher linguistic proximity led to higher *production frequency* (i.e., people listed a member concept more often when its label co-occurred very frequently with the category label), unstandardized $B = 0.10$, $SE = 0.01$, standardized ß = 0.21, $t = 10.41$, $p < .001$. As shown by the standardized

coefficients, the effect of linguistic proximity on *production frequency* was larger than that of sensorimotor similarity.

2.2.2. Mean rank

As predicted, both sensorimotor similarity and linguistic proximity predicted the mean ordinal rank of a named member concept for its category. Again, evidence was very strong for both Steps 2 and 3 (see Table 2); this pattern was consistent with NHST *F* tests. In the Step 3 model, higher sensorimotor similarity led to lower *mean rank* (i.e., people listed a member concept earlier when its sensorimotor profile was very similar to the category concept), unstandardized $B = –11.44$, $SE = 0.89$, standardized ß $= –0.26$, $t = –12.79$, $p < .001$. Higher linguistic proximity also led to lower *mean rank* (i.e., people listed a member concept earlier when its label co-occurred very frequently with the category label), unstandardized $B = –0.04$, $SE = 0.01$, standardized ß $= –0.12$, $t = –5.56$, $p < .001$. Although linguistic proximity had an effect over and above that of sensorimotor similarity, standardized coefficients indicate that sensorimotor similarity had the larger effect on *mean rank*.

2.2.3. First-rank frequency

Again as predicted, both sensorimotor similarity and linguistic proximity predicted the frequency of naming a member concept for a category in first ordinal position, with very strong evidence for both Steps 2 and 3 (see Table 2); this pattern was consistent with NHST *F* tests. In the Step 3 model, higher sensorimotor similarity led to higher *first-rank frequency* (i.e., people listed a member concept in first place more often when its sensorimotor profile was very similar to the category concept), unstandardized $B = 4.70$, $SE = 1.51$, standardized ß $= 0.12$, $t = 3.12$, $p = .002$. Likewise, higher linguistic proximity led to higher *first-rank frequency* (i.e., people listed a member concept in first place more often when its label co-occurred very frequently with the category label), unstandardized $B = 0.06$, $SE = 0.01$, standardized ß $= 0.20$, $t = 5.16$, $p < .001$. Standardized coefficients indicate that linguistic proximity had a larger effect on *first-rank frequency* than did sensorimotor similarity.

2.2.4. Response times

Results for *RT* were mixed. Against our predictions, and unlike our findings for the explicit dependent variables above, sensorimotor similarity did not predict RTs. Model comparisons via Bayes factors showed strong evidence for Step 1 (i.e., the baseline model) over Step 2 with sensorimotor similarity, consistent with the non-significant likelihood ratio test (see Table 2). However, results for Step 3 were inconsistent: Bayesian model comparison found only equivocal evidence in favor of linguistic proximity in Step 3, compared to Step 2, whereas the likelihood ratio test showed that Step 3 was significantly better than Step 2. In the Step 3 model, sensorimotor similarity had no effect on *RT*, unstandardized $B = 0.06$, $SE = 0.61$, standardized ß $= 0.00$, $t = 0.11$, $p = .917$, but higher linguistic proximity led to faster *RT* (i.e., people were faster to list their first category member when its label co-occurred very frequently with the category label), unstandardized $B = –0.01$, $SE = 0.01$, standardized ß $= -0.09$, $t = –2.90$, $p = .004$). We interpret this finding cautiously whereby—as predicted—linguistic proximity predicted *RT*, but the effect is relatively small and with equivocal evidence.

2.2.5. Summary

Results of our confirmatory analyses strongly supported our hypothesis that both sensorimotor similarity and linguistic proximity would independently contribute to explicit measures of category production (i.e., how often each member concept is listed for a particular category, and how early each member concept is listed for a particular category). Of the two variables, linguistic distributional information dominated *production frequency* and *first-rank frequency*, but sensorimotor similarity had the larger effect on *mean rank*.[6] However, results were equivocal for the implicit measure of category production (i.e., RT to name the first member of a category), where sensorimotor similarity had no effect but linguistic proximity had a relatively small effect. Overall, member concepts that were very similar in sensorimotor experience to their category concept were more likely to be produced early and often in a category production task. Likewise, member concepts whose label frequently appeared in proximity to their category label were also more likely to be produced early and often—and to some extent, a little faster—in category production.

Previous work by Taler, Johns, and Jones (2020) found that the density of the linguistic distributional neighborhood around a member concept predicted its production frequency in the category ANIMAL, which could be related to the likelihood of the category concept successfully activating that member concept (i.e., as activation spreads from the category concept, member concepts with many neighbors in proximity may benefit from many paths of activation in a way that member concepts with few close neighbors cannot). In an exploratory analysis,[7] we examined the contribution of linguistic distributional density alongside our existing predictors of linguistic proximity and sensorimotor similarity. The best model for all three explicit measures of category production (i.e., *production frequency, mean rank, first-rank frequency*) included all three predictors, although the contribution of linguistic distributional density was relatively modest, compared to that of linguistic proximity and sensorimotor similarity (see Supplemental Materials for full data and results). Nonetheless, this exploratory analysis suggests that the density of linguistic distributional neighbors around a member concept may help to boost the activation coming from the category concept, though it is the category-to-member relationship itself (i.e., linguistic proximity and sensorimotor similarity) that remains the primary influence on category production.

3. Experiment 1b: Typicality in category production

Experiment 1a demonstrated that semantic category production relies on sensorimotor and linguistic distributional information. However, producing members of a category may also be influenced by the perceived typicality of semantic concepts (e.g., Hampton & Gardiner, 1983; Mervis, Catlin, & Rosch, 1976)—that is, how good an example a particular concept is of its category. To some extent, these measures are theoretically circular: *dog* might be the first ANIMAL that comes to mind because it is the best example of an ANIMAL; and *dog* might be the best example of an ANIMAL because it is the first example that comes to mind. Nonetheless, since linguistic distributional information has previously been shown to

correlate with typicality ratings (Connell & Ramscar, 2001), it was possible that our linguistic–sensorimotor measures in Experiment 1a could outperform typicality in predicting explicit and implicit measures of category production.

In a second experiment (pre-registration, data, analysis code, and results are available at https://osf.io/vaq56/), we therefore collected typicality ratings for all category production responses in Experiment 1a and tested whether category production performance relies on sensorimotor and linguistic distributional information to a greater extent than on typicality. We predicted that typicality ratings would predict category production responses from Experiment 1a, such that more typical members of a category would be named more frequently and earlier than atypical members. However, we also predicted that including word frequency in our baseline model would reduce the magnitude of the typicality effect (as word frequency is correlated with typicality: e.g., Moreno-Martínez, Montoro, & Rodríguez-Rojo, 2014; Navarrete, Arcara, Mondini, & Penolazzi, 2019; Schröder, Gemballa, Ruppin, & Wartenburger, 2012). Further, while typicality correlates strongly with category production measures for natural categories (e.g., ANIMAL, FRUIT, TOOL), it tends to be much weaker when a wider range of category types are examined (e.g., Casey, 1992). Thus, we hypothesized that category production would be better predicted by linguistic and sensorimotor information combined (Experiment 1a) than by typicality (the present experiment).

3.1. Method

3.1.1. Participants

One hundred and forty-one native speakers of English (88 female, mean age = 31.23 years (SD = 10.34), 111 right-handed) took part in this study via the research crowdsourcing tool Prolific and received £1.75 for participation. In order to recruit a sample with similar linguistic experience to Experiment 1a (i.e., British English as opposed to other dialects), we restricted recruitment to native speakers of English who were U.K. nationals using Prolific's screening criteria. Fourteen participants' submissions were rejected because their ratings did not pass our quality control checks (i.e., they were not paid, their data was not included in the main analysis, and they were not permitted to participate further in the study; see Data Preparation and Analysis for details). New participants were recruited via Prolific until we reached $N = 12$ for all stimulus lists; 127 participants' data were included in the final analysis, as participants were able to rate multiple lists if they wished.

3.1.2. Materials

We used all category–member word pairs from the final analysis for Study 1a as stimuli for typicality rating. Most member concepts were presented verbatim (e.g., singular and plural forms of the same word were treated as separate items), but in some cases, we added the category name if the member name on its own would be ambiguous as a referent (e.g., *straw hat* was used instead of *straw* for the category HAT).

The full item set for typicality rating comprised 2228 category–member stimuli from Experiment 1a, plus a further six stimuli included due to an initial error in data preparation of that experiment and an additional 46 stimuli (with identical format) that were rated for use in

a separate study (these latter items did not form part of the present experiment and will not be discussed further), leading to a total of 2280 items. Category–member pairs were pseudo-randomly divided into 20 stimulus lists with a number of constraints. Each category was distributed across lists as equally as possible (given the varying number of members listed for each category), and member concepts that appeared with more than one category (e.g., *eagle* was listed for both categories BIRD and BIRD OF PREY) were allocated to separate lists. In addition, production frequency (from Study 1a) and word frequency (LgSUBTLWF; Balota et al., 2007) were counterbalanced across lists (mean *production frequency* per list = 5.73 ($SD = 4.65$) ranging from 5.24 to 6.05, with no significant difference between lists, $F(19) = 0.28$, $p = .999$; mean LgSUBTLWF per list = 2.59 ($SD = 0.87$) ranging from 2.48 to 2.74, with no significant difference between lists, $F(19) = 0.98$, $p = .485$).

We selected 80 category–member pairs from our stimulus set to enable quality control checks in online data collection. These control items were selected based on their typicality ratings in previous studies (Armstrong, Gleitman, & Gleitman, 1983; Rosch, 1975; Uyeda & Mandler, 1980), where half had high typicality ratings (i.e., <1.6 on a scale of 1–7, 1 being *high typicality* and 7 being *low typicality*; mean "high" typicality = 1.34, $SD = 0.19$) and half had low typicality ratings (i.e., >3.7, mean "low" typicality rating = 4.39, $SD = 0.54$). Each control category–member pair appeared in two different stimulus lists so that, overall, each stimulus list contained four high-typicality and four low-typicality control items. Two further category–member pairs (selected from previous studies but *not* featured in our stimuli), one high typicality (TOY: *doll*) and one low typicality (VEHICLE: *surfboard*), were chosen to act as scale calibrators and were presented at the start of each stimulus list; these items were not included in the reported analyses. Each stimulus list therefore comprised 120 items (category–member word pairs), including the two calibrator and eight control items.

3.1.3. Procedure

Each stimulus list was presented in a randomized order in an online questionnaire via Qualtrics. Participants were instructed to rate each category member based on how good an example of the category they thought it was, on a scale from 1 to 5, with 1 being a "*very poor*" example, and 5 being a "*very good*" example. Participants were asked to base these ratings on their own judgments and not to worry about the responses that other people might give. For each item, the category name was presented in capital letters in a text box in the center of the screen; underneath was the framing question "How good an example of this category is/are a/an X(s)?" and the rating scale (e.g., for the category ANIMAL, "How good an example of this category is a CAT?"). Participants indicated their rating by selecting the appropriate radio button and only one response per item was allowed. They could also indicate if they did not know the meaning of the category or the category member; no ratings were recorded for such trials. The entire rating procedure took approximately 15 min. At the end of the stimulus list, participants provided demographic information and read a study debrief.

3.1.4. Ethics and consent

The study received ethical approval from the Lancaster University Faculty of Science and Technology Research Ethics Committee. All participants read online information detailing the

purpose and expectations of the study, including the caveat that data had to pass quality checks before payment would be awarded, before giving informed consent to take part. Consent included an agreement to share publicly all alphanumeric data in anonymized form.

3.1.5. Data preparation, design, and analysis

To check the quality of the online data we had collected, each participant's ratings for the control items were correlated with ratings gained from previous studies. If the Pearson's correlation coefficient was $r < .30$, *and* the variance of that participant's data was close to zero, then the participant was excluded for failing to adequately attend to and/or understand the task, and their submission was rejected. Fourteen participants were excluded on this basis (see Section 3.1.1).

3.1.5.1. Typicality ratings. We calculated the mean typicality rating per category–member pair to act as the critical predictor in the analysis below.

3.1.5.2. Dependent variables. From Experiment 1a, we took *production frequency*, *mean rank*, *first-rank frequency*, and *RT*.

3.1.5.3. Statistical analyses. As in Experiment 1a, and as per the pre-registration, we carried out Bayesian linear hierarchical regressions in JASP (version 0.14; JASP Team, 2020) for each dependent variable using JZS default priors (r scale = .354) and Bayesian adaptive sampling, and their NHST counterparts (i.e., ordinary least squares linear regressions), in two hierarchical steps. Step 1 comprised a baseline model of log word frequency of the named member concept (LgSUBTLWF: Balota et al., 2007), and Step 2 added mean typicality rating of the category–member pair. We tested whether Step 2 increased model fit over and above Step 1 (i.e., whether typicality rating predicts category production) both via NHST of R^2-change and via model comparison using Bayes factors. In a non-nested model comparison, we then compared the Step 2 typicality model against a sensorimotor–linguistic model containing word frequency, sensorimotor similarity, and linguistic proximity (i.e., the Step 3 model from Experiment 1a analyses) using Bayes factors.

For *RT*, we ran linear mixed-effects models in R using the lmerTest package (Kuznetsova et al., 2017) and calculated marginal R^2 using the MuMIn package (Barton, 2017). Standardized coefficients were calculated in R by running the relevant analyses using standardized variables. As in Experiment 1a, Step 1 comprised crossed random effects of participant and item (category) and fixed effects of log word frequency (LgSUBTLWF; Balota et al., 2007), PLD20 (Yarkoni et al., 2008), and the number of syllables. Step 2 added typicality rating as a fixed effect and was compared to Step 1 using NHST likelihood ratio chi-square tests and Bayes factors calculated via BIC. As above, the Step 2 typicality model was then compared to a sensorimotor–linguistic model (i.e., the Step 3 model from Experiment 1a analyses), using Bayes factors.

Descriptive statistics and zero-order correlations for all variables are provided as Supplementary Materials (Table S1).

Table 3
Experiment 1b hierarchical linear regressions of category production measures on typicality ratings, and non-nested model comparison of typicality with Experiment 1a sensorimotor–linguistic predictors

Step	Model Comparison	Total R^2	ΔR^2	F	Log BF
Category Production Frequency					
1	Baseline (lexical) model versus empty model (BF_{10})	.023	–	51.74***	22.37
2	Typicality rating versus baseline (BF_{21})	.199	.176	489.98***	217.56
Exp1a	Sensorimotor + linguistic (Exp 1a) versus baseline ($BF_{Exp1a\text{-}1}$)	.087	.064	78.33***	69.89
Exp1a	Sensorimotor + linguistic (Exp 1a) versus typicality rating ($BF_{Exp\ 1a\text{-}2}$)	–	–	–	–147.67
Mean Rank					
1	Baseline (lexical) model versus empty model (BF_{10})	.001	–	1.20	–2.45
2	Typicality rating versus baseline (BF_{21})	.051	.051	118.88***	54.84
Exp1a	Sensorimotor + linguistic (Exp 1a) versus baseline ($BF_{Exp1a\text{-}1}$)	.083	.083	100.53***	90.35
Exp1a	Sensorimotor + linguistic (Exp 1a) versus typicality rating ($Bu_{ffex\ 1a\text{-}2}$)	–	–		35.51
First-Rank Frequency					
1	Baseline (lexical) model versus empty model (BF_{10})	.040	–	28.38***	11.19
2	Typicality rating versus baseline (BF_{21})	.129	.088	68.31***	29.76
Exp1a	Sensorimotor + linguistic (Exp 1a) versus baseline ($BF_{Exp1a\text{-}1}$)	.090	.049	18.20***	13.18
Exp1a	Sensorimotor + linguistic (Exp 1a) versus typicality rating ($BF_{Exp1a\text{-}2}$)	–	–	–	–16.58
RT			χ^2		
1	Baseline (lexical) model versus random effects model (BF_{10})	.005	–	8.11*	-7.29
2	Typicality rating versus baseline (BF_{21})	.018	.013	30.27***	11.35
Exp1a	Sensorimotor + linguistic (Exp 1a) versus baseline ($BF_{Exp1a\text{-}1}$)	.012	.007	8.39*	–3.38
Exp1a	Sensorimotor + linguistic (Exp 1a) versus typicality rating ($BF_{Exp1a\text{-}2}$)	–	–	–	–14.72

Note. ΔR^2 = change in R^2; log BF = natural log Bayes factor, where negative values indicate evidence for the null model and positive values indicate evidence for the alternative model: log BFs $\geq$ 3.00 (equivalent to BFs $\geq$ 20) constitute strong evidence; log BFs $\geq$ 1.10 (equivalent to BFs $\geq$ 3) constitute positive evidence; and log BFs between –1.10 and 1.10 (equivalent to BFs of 0.33 and 3) constitute equivocal evidence.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

3.2. Results and discussion

3.2.1. Category production frequency

Typicality ratings predicted the frequency of naming a particular member concept for a category. Bayes factors and NHST F tests (see Table 3) in the Step 2 model indicated strong evidence that the more typical a concept was of its category, the more frequently it was

produced as a member of that category (unstandardized $B = 2.95$, $SE = 0.13$, standardized ß = 0.42, $t = 22.14$, $p < .001$). Contrary to our predictions, however, Bayes factors indicated stronger evidence for the typicality model, compared to the sensorimotor–linguistic model; that is, typicality ratings predicted the frequency of naming a particular member concept better than sensorimotor and linguistic distributional information.

3.2.2. Mean rank

Typicality ratings predicted the mean ordinal rank of a named member concept for its category. Bayes factors and NHST *F* tests in the Step 2 model (see Table 3) indicated strong evidence that the more typical a concept was of its category, the higher its *mean rank* within that category; that is, more typical concepts were named earlier than less typical concepts (unstandardized $B = -1.23$, $SE = 0.11$, standardized ß = -0.23, $t = -10.90$, $p < .001$). Moreover, a large Bayes factor indicated very strong evidence for a sensorimotor–linguistic model, compared to the typicality model. As predicted, sensorimotor and linguistic distributional information did better than typicality ratings in predicting how early people named a member concept.

3.2.3. First-rank frequency

Typicality ratings also predicted the frequency of naming a member concept for a category in the first ordinal position. Bayes factors and NHST *F* tests in the Step 2 model (see Table 3) indicated strong evidence that, for a given category, more typical concepts were named more frequently as a first response (unstandardized $B = 1.89$, $SE = 0.23$, standardized ß = 0.30, $t = 8.27$, $p < .001$). Contrary to our expectations, however, Bayes factors indicated stronger evidence for the typicality model, compared to the sensorimotor–linguistic model, meaning that typicality ratings could predict how often a member concept was named first in a category better than sensorimotor and linguistic distributional information.

3.2.4. Response times

Finally, typicality ratings predicted *RT*s for first-named category members, whereby more typical category members were named faster. Bayes factors and likelihood ratio tests in the Step 2 model (see Table 3) indicated strong evidence that typicality ratings predicted *RT* (unstandardized $B = -0.49$, $SE = 0.09$, standardized ß = -0.12, $t = -5.58$, $p < .001$). Contrary to our predictions, however, Bayes factors indicated stronger evidence for the typicality model, compared to the sensorimotor–linguistic model. In other words, the time taken to name the first member of a category was better predicted by typicality ratings than by sensorimotor and linguistic distributional information.

3.2.5. Exploratory analyses

Since typicality ratings unexpectedly did better than sensorimotor and linguistic distributional information in predicting many category production variables in non-nested model comparisons, it raised the question of whether sensorimotor similarity and linguistic proximity are subsumed by, or are independent of, the construct of typicality. That is, if sensorimotor–linguistic information merely reflects goodness-of-membership within a

category, then it is effectively subsumed by typicality ratings, meaning that sensorimotor similarity and linguistic proximity would predict no extra variance in category production when typicality has already been taken into account. On the other hand, if sensorimotor and linguistic distributional information are critical to category production in a way that goes beyond goodness-of-membership, then sensorimotor similarity and linguistic proximity will predict category production even when typicality has been taken into account.

To test these possibilities, we carried out exploratory analyses to determine the contribution of sensorimotor similarity and linguistic proximity to all dependent variables over and above typicality ratings. We added two additional steps to the hierarchical regressions conducted for confirmatory analyses: Step 3 added sensorimotor similarity of the category–member pair, and Step 4 added linguistic proximity. We tested whether Step 3 increased model fit over and above Step 2 and whether Step 4 increased model fit over Step 3, both via NHST of R^2-change for the category production measures and likelihood ratio for RTs and via model comparison using Bayes factors for all measures. Typicality ratings correlated relatively weakly with both sensorimotor similarity ($r = .162$) and linguistic proximity ($r = .146$). Results for all four category production-dependent measures are in Table 4.

For category *production frequency*, both the sensorimotor and linguistic variables contributed independently; however, standardized coefficients in the Step 4 model suggest these effects (sensorimotor similarity unstandardized $B = 4.03$, $SE = 1.07$, standardized ß $= 0.07$, $t = 3.77$, $p < .001$; linguistic proximity unstandardized $B = 0.08$, $SE = 0.01$, standardized ß $= 0.16$, $t = 8.21$, $p < .001$) were smaller than the effect of typicality rating (unstandardized $B = 2.70$, $SE = 0.13$, standardized ß $= 0.39$, $t = 20.14$, $p < .001$).

Both variables also contributed to *mean rank*, but this time the effect of sensorimotor similarity was larger than that of typicality rating (Step 4: sensorimotor similarity unstandardized $B = –10.24$, $SE = 0.89$, standardized ß $= –0.23$, $t = −11.48$, $p < .001$; typicality unstandardized $B = –0.95$, $SE = 0.11$, standardized ß $= –0.17$, $t = -8.48$, $p < .001$), whereas the effect of linguistic proximity was smaller than both (unstandardized $B = –0.03$, $SE = 0.01$, standardized ß $= –0.09$, $t = -4.34$, $p < .001$).

For *first-rank frequency*, model comparisons indicated that only linguistic proximity contributed above and beyond typicality. In Step 4, linguistic proximity (unstandardized $B = 0.05$, $SE = 0.01$, standardized ß $= 0.15$, $t = 4.11$, $p < .001$) had a smaller effect than typicality rating (unstandardized $B = 1.66$, $SE = 0.23$, standardized ß $= 0.26$, $t = 7.13$, $p < .001$), but sensorimotor similarity had a negligible effect (unstandardized $B = 2.71$, $SE = 1.48$, standardized ß $= 0.07$, $t = 1.83$, $p = .068$).

Finally, for *RT*, model comparisons did not favor including either sensorimotor similarity or linguistic proximity as predictors. Coefficients in the Step 4 model showed that neither linguistic proximity (unstandardized $B = –0.01$, $SE = 0.01$, standardized ß $= –0.06$, $t = –1.96$, $p = .051$) nor sensorimotor similarity (unstandardized $B = 0.93$, $SE = 0.62$, standardized ß $= 0.04$, $t = 1.52$, $p = .129$) had an effect on *RT* over and above typicality, whereas the effect of typicality remained robust (unstandardized $B = –0.50$, $SE = 0.09$, standardized ß $= –0.12$, $t = –5.41$, $p < .001$).

Overall, evidence from our exploratory analyses indicated that sensorimotor similarity and linguistic proximity independently predicted explicit measures of category production above

Table 4
Experiment 1b exploratory analysis of sensorimotor and linguistic distributional effects on category production measures above and beyond typicality

Step	Model Comparison	Total R^2	ΔR^2	F	Log(BF)
Category Production Frequency					
3	Sensorimotor similarity + typicality rating versus typicality rating only (BF_{32})	.204	.005	14.94***	4.52
4	Linguistic proximity + sensorimotor similarity + typicality rating versus sensorimotor similarity + typicality rating (BF_{43})	.228	.023	67.47***	30.32
Mean Rank					
3	Sensorimotor similarity + typicality rating versus typicality rating only (BF_{32})	.105	.053	132.57***	61.35
4	Linguistic proximity + sensorimotor similarity + typicality rating versus sensorimotor similarity + typicality rating (BF_{43})	.112	.008	18.80***	6.77
First-Rank Frequency					
3	Sensorimotor similarity + typicality rating versus typicality rating only (BF_{32})	.132	.004	2.84	–0.75
4	Linguistic proximity + sensorimotor similarity + typicality rating versus sensorimotor similarity + typicality rating (BF_{43})	.154	.021	16.90***	6.16
RT			χ^2		
3	Sensorimotor similarity + typicality rating versus typicality rating only (BF_{32})	.020	.002	2.48	–2.54
4	Linguistic proximity + sensorimotor similarity + typicality rating versus sensorimotor similarity + typicality rating (BF_{43})	.024	.004	3.83	–1.87

Note. Variables were entered as additional hierarchical steps following the confirmatory regression model of typicality ratings reported in Table 3, Step 2. ΔR^2 = change in R^2; log BF = natural log Bayes factor, where negative values indicate evidence for the null model and positive values indicate evidence for the alternative model: log BFs ≥ 3.00 (equivalent to BFs ≥ 20) constitute strong evidence; log BFs ≥ 1.10 (equivalent to BFs ≥ 3) constitute positive evidence; and log BFs between –1.10 and 1.10 (equivalent to BFs of 0.33 and 3) constitute equivocal evidence.

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

and beyond typicality ratings but not implicit measures (*RT*). When typicality had already been taken into account, linguistic proximity still predicted all three explicit measures of category production, and sensorimotor similarity predicted two (*production frequency* and *mean rank*).

3.2.6. Summary

Results of confirmatory analyses provided mixed support for our hypotheses. As predicted, typicality was a useful predictor of category production performance that was nonetheless outperformed by a combination of sensorimotor similarity and linguistic proximity in predicting *mean rank*. However—against our predictions—typicality was the better predictor (i.e.,

outperforming linguistic-sensorimotor measures) of *production frequency*, *first-rank frequency*, and *RT*. In exploratory analyses, we then found that sensorimotor and linguistic information predicted additional variance in all three explicit category production measures (*production frequency, mean rank*, and *first-rank frequency* but not *RT*) when typicality had already been taken into account, which demonstrated that linguistic distributional and sensorimotor information cannot be subsumed by typicality ratings, and instead are critical to category production in a way that goes beyond goodness-of-membership in a category.

One could view this pattern of typicality and linguistic-sensorimotor effects in terms of explaining response popularity versus order of activation, respectively. That is, typicality appeared to be most useful in explaining the popularity of a member concept within its category (i.e., *production frequency* and *first-rank frequency*, as well as latency naming this first-ranked concept), which is consistent with the idea that the "best" examples of categories come to mind more often (Mervis et al., 1976). By contrast, linguistic and sensorimotor measures were most useful in explaining the ordinal position in which a member concept is named within its category (i.e., *mean rank*), which suggests that the closer member concepts are to their category concept in terms of sensorimotor and linguistic distributional experience, the sooner they are activated. Overall, results of Experiment 1b suggest that while sensorimotor and linguistic distributional information overlap a little with typicality, they reflect a distinctly different phenomenon to that of goodness-of-membership within a category (also see Heyman & Heyman, 2019, for related findings regarding linguistic distributional information) that relates to the order of conceptual activation during category production.

4. Computational model of category production

The results of our behavioral experiments strongly supported the idea that both linguistic distributional and sensorimotor information independently contribute to category production. When presented with a category such as ANIMAL, people named a member concept more often, and earlier in rank order, when its label frequently appeared in proximity to the category label in linguistic contexts (e.g., ANIMAL and *cat* frequently co-occur within a few words of each other) *and* when its referent was more similar in sensorimotor experience to the category concept (e.g., ANIMAL and *cat* share similar sensorimotor profiles).

Nonetheless, our linguistic and sensorimotor predictors in Experiments 1a and b were limited in one critical way: They were based only on *direct* relationships between category and member concepts (e.g., ANIMAL → *cat*). While direct relationships between concepts are undoubtedly important, linguistic-simulation theories of the conceptual system have always held that activation also spreads *indirectly* between words and between sensorimotor representations of concepts (Barsalou et al., 2008; Connell, 2018; Connell & Lynott, 2014b; Louwerse, 2011). That is, it may be the case that people list *mouse* as a kind of animal not because of a strong ANIMAL → *mouse* relationship (in linguistic and/or sensorimotor terms) but rather because it is activated indirectly via one or more intermediate concepts, such as ANIMAL → *cat* → *mouse* (also see Hills, Jones, & Todd, 2012; McKoon & Ratcliff, 1992). We, therefore, wished to test whether linguistic distributional and sensorimotor information

would better fit human performance in category production if indirect relationships were taken into account. Furthermore, since neither linguistic distributional nor sensorimotor information dominated category production when considering direct category–member relationships (e.g., linguistic proximity was the stronger predictor of *production frequency*, whereas sensorimotor similarity was the stronger predictor of *mean rank*), we wished to examine whether the same would occur when considering indirect relationships or whether linguistic distributional information would primarily drive member–concept responses as originally hypothesized in Experiment 1a.

To this end, we developed a computational model of conceptual processing during category production, based on linguistic distributional and sensorimotor information, and examined its fit to human data from our behavioral experiments. The model comprised two separate components, linguistic and sensorimotor, in a snapshot of the conceptual system (i.e., a time slice at the point of commencing an individual category production trial). That is, although we do not attempt to model how the nature of conceptual representations changes dynamically with recent experience, task goals, and available cognitive resources, we do not assume that conceptual knowledge is invariant or static (Connell & Lynott, 2014b). The linguistic component of the model was designed to approximate a snapshot of linguistic distributional knowledge (Wingfield & Connell, 2020) and took the form of a variant of a spreading-activation network of nodes and edges, where activation of a word node would spread out along connected edges and then activate the nodes of distributional neighbors (i.e., words that occurred most often in the same linguistic contexts). The sensorimotor component was designed to approximate a simplified snapshot of a sensorimotor simulation system (Connell & Lynott, 2014b; Connell et al., 2018) and allowed activation to spread uniformly from a concept point to a limited distance in an 11-dimensional sensorimotor space and activate any nearby concepts (i.e., with high sensorimotor similarity) within that distance. In both components, any activated word/concept could further propagate activation to new neighboring words/concepts.

We created three different versions of the model for comparison: a *control* version where linguistic and sensorimotor components allowed only direct activations and their outputs were evaluated separately; a *separate* version where linguistic and sensorimotor components allowed indirect activations, but again their outputs were evaluated separately; and a *combined* version where linguistic and sensorimotor components with indirect activations ran in parallel, and their outputs were collated and evaluated together as a single model.

4.1. Method

4.1.1. Model architecture and operation

The model was designed to approximate a full-size conceptual system of an educated, adult native speaker of English, comprising linguistic distributional knowledge for 40,000 words (based on a subtitles corpus of contemporary British English) that was grounded in sensorimotor experience from the Lynott et al. (2020) sensorimotor norms' 40,000 concepts. We first describe the architecture of the *separate model*, which comprised separate linguistic and sensorimotor components that allowed indirect activations and had their outputs evaluated separately, before outlining how the *control* and *combined* models differed. Fig. 1 shows the

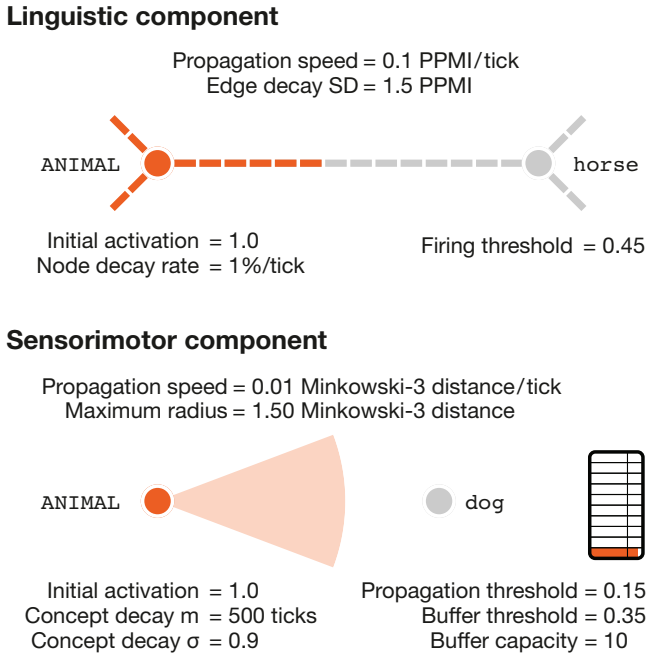

Fig 1. Overview of parameters involved in each component of the computational model. The linguistic component spreads activation in a graph from an initial category node along edges to new nodes, where both node and edge activation decays with each tick of the model clock, and new nodes fire when their accumulated activation clears a firing threshold. The sensorimotor component spreads activation from an initial concept point to other concepts points via an expanding multidimensional sphere up to a maximum radius, where concept activation decays with each tick of the model clock, and new concept points propagate their own spheres when their accumulated activation clears a propagation threshold; a fixed-capacity buffer limits concurrent candidate member concepts. Note that units of distance in linguistic and sensorimotor components are not directly comparable.

key parameters in model operation, while Fig. 2 illustrates a sample run of the combined model for the category ANIMAL.

4.1.1.1. Linguistic component. We implemented a linguistic model similar to a spreading-activation model of semantic memory (e.g., Collins & Loftus, 1975), where activation propagates on a graph of nodes and edges. As a model of linguistic distributional knowledge, the nodes corresponded to words, and edges to distributional relationships between those words, and activation spread from node to node along edges.

Our graph was built from the 40,000 most common words in the subtitles corpus used to create the linguistic proximity measures in Experiments 1a and b, with each word represented by a node, providing 92.6% coverage of the words found in categories and responses. Each edge had a length derived from the linguistic proximity score between its two endpoint words; that is, PPMI 6-gram values were converted to lengths by a linear rescaling, which inverted the global maximum and non-zero minimum, so the highest PPMI (i.e., the highest frequency of co-occurrence) resulted in the shortest edges. Activation propagated incrementally along

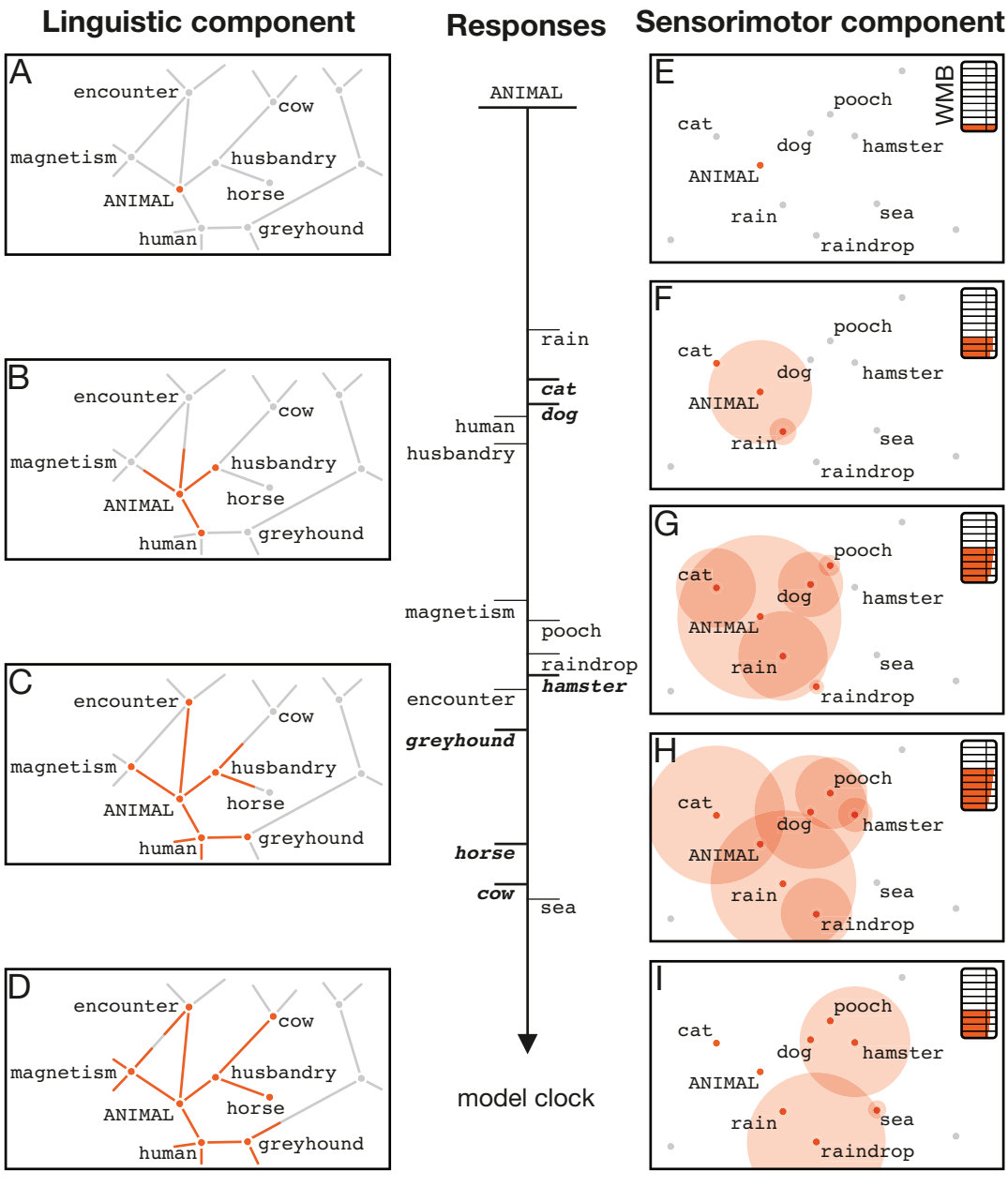

Fig 2. Operation of the combined model, illustrating indirect activation in a simplified run of the sample category ANIMAL. Activation for ANIMAL is seeded simultaneously in the linguistic component (panel (a)) and sensorimotor component (panel (e); WMB = working memory buffer), which then runs in parallel. In the linguistic component, activation spreads through the edges of the graph to activate neighboring words (b), which in turn spread activation to their neighbors (c, d). In the sensorimotor component, activation spreads through an expanding sphere to activate neighboring concept points (f), which in turn spread their own spheres of activation to further neighbors (g–i). Activation continues to spread through both components until the model is halted.

edges according to a model clock at a fixed propagation speed of 0.1 PPMI-distance per tick, whereby the shortest edge took 56 increments (*ticks*) to traverse and the longest took 243. Thus, activation would reach closer words (that co-occurred more frequently) faster than further ones (that co-occurred less frequently but still more often than chance). If two words never co-occurred in the corpus, or otherwise had a PPMI of 0 because they co-occurred no more than chance, there was no edge between them. The final graph had 26,334,191 edges.

The category production task was initialized by seeding an activation value of 1.0 at the node representing the category label (e.g., ANIMAL) and then allowing activation to propagate throughout the network (see Fig. 1). Unlike the classic spreading activation model of Collins and Loftus (1975), the sum of activation spreading out from a node was not divided between the incident edges; instead, a concept emitted its full activation value into all incident edges. Activation of word nodes decayed exponentially over time (node decay rate = 1% per tick), and activation propagating along edges decayed with a Gaussian curve (edge decay *SD* = 15.0; cf. Schweickert & Boruff, 1986) until reaching a floor value of 0.05 at which point propagation stopped. Incoming activation accumulated in a node, causing that node to fire (i.e., propagate its current activation outward) if a firing threshold (0.45) was reached. Once a node fired, it could not fire again until its activation had decayed below the firing threshold and incoming activation caused it to reach the threshold again. A node had to reach its firing threshold to be considered as a candidate category member. In this way, it could be the case that the node for a particular member concept (e.g., *horse*) received insufficient direct activation from the original category seed node (e.g., ANIMAL) to reach its firing threshold and required indirect activation (e.g., via *husbandry*) to reach the threshold and count as an activated category member. The output from the linguistic component comprised the list of activated category members, in the order they were first activated, along with the time to first activation (tick number of the model clock) for each member.

4.1.1.2. Sensorimotor component. We implemented a novel sensorimotor model based on the 11-dimensional space of the Lynott et al. (2020) Lancaster sensorimotor norms. As a model of sensorimotor conceptual knowledge, each concept corresponded to a point in an 11-dimensional sensorimotor space and activation spread within this space so that concepts with similar sensorimotor profiles could activate one other.

Our sensorimotor space was built from the 39,707 items in the Lynott et al. (2020) sensorimotor norms used to create the sensorimotor similarity measures in Experiments 1a and b. Each concept was represented as a point, or vector, within this sensorimotor space, where coordinates for a particular concept equated to the mean strength rating in each of the 11 sensorimotor dimensions. Distances within the sensorimotor space were computed as Minkowski-3 distance (see Experiment 1a) so that concepts with similar sensorimotor profiles were located close together and dissimilar concepts were far apart. The most similar concepts in sensorimotor space were 0.13 units apart in distance (e.g., *antifeminism* and *unjust*), whereas the most dissimilar concepts were 8.13 units apart (e.g., *overhear* and *string cheese*). Activation propagated through this sensorimotor space in expanding "spheres" from each activated concept point, where the radius of an activation sphere increased at a fixed

propagation speed of 0.01 Minkowski-3 distance per tick of the model clock until reaching a maximum radius (1.50), whereupon it dissipated.

The category production task was initialized by seeding an activation value of 1.0 at the concept point representing the category concept (e.g., ANIMAL) and then allowing a sphere of activation to expand throughout sensorimotor space (see Fig. 1). Other concept points that were met by the expanding sphere received its activation, attenuated by the prevalence of each concept's label (i.e., how well-known is a particular word among native speakers; Brysbaert, Mandera, McCormick, & Keuleers, 2019); in this way, we used the prevalence of a label to approximate how well-practiced a person might be at simulating its referent concept so that seldom-simulated concepts received weaker activation than oft-simulated concepts. Activation of concept points decayed according to a lognormal curve (median = 500 ticks, $\sigma = 0.9$). Incoming activation accumulated in a given concept point and caused that concept to produce its own sphere of activation (i.e., propagate its current activation outward) if it reached a propagation threshold (0.15). The continuous structure of sensorimotor space meant that any set of coordinates constituted a profile of sensorimotor experience that could theoretically underpin a conceptual representation, but only those coordinates defined as concept points (i.e., that had a lexical label and formed part of an adult conceptual system in the Lynott et al. (2020) norms) could propagate activation outward. Once a concept point produced its own sphere of activation, it could not do so again until its activation had decayed below the propagation threshold, and incoming activation caused it to re-reach the threshold. In addition, to prevent runaway activation in dense regions of sensorimotor space, activation arriving at new concept points was attenuated linearly by the number of concept points already above the propagation threshold (e.g., if 10 concepts were already above the threshold, then attenuation was slight and made little difference to the amount of activation a given concept point received; but if the number of concepts above threshold reached a limit of 3000 or more,[8] then incoming activation was attenuated to zero, meaning a given concept received no new activation).

Finally, in order to identify concepts that were highly activated enough to be considered as candidate category members, we implemented a simplified *working memory buffer* with a fixed capacity of 10 concepts (i.e., the estimated capacity of working memory for familiar object concepts: Dymarska, Connell, & Banks, 2020). Concepts entered the buffer if their activation reached the buffer threshold (0.35) and left the buffer if their activation decayed below the buffer threshold. Newly activated concepts displaced those with lower activation; in the case of ties, concepts were displaced according to a hierarchy of recency of entry to buffer, recency of activation emission from concept point, and finally load order of concepts into the model (alphabetical; an implementational convenience). A concept had to enter the working memory buffer to be considered as a candidate category member. Hence, it could be the case that a particular member concept point (e.g., *hamster*) did not receive any direct activation from the original category seed node (e.g., ANIMAL) before its sphere of activation dissipated, or received insufficient activation due to having a lesser-known label, and hence could not enter the buffer. Such concepts would require indirect activation (e.g., via *dog*) to clear the buffer threshold and count as an activated category member. The output from the sensorimotor component comprised the list of activated category members, in the order they

were first activated, along with the time to first activation (tick number of the model clock) for each member.

4.1.1.3. Model versions: control, separate, combined. The above sections describe the architecture and processes of the *separate model*, where the linguistic and sensorimotor components both allowed indirect activations but ran independently and produced separate output lists of activated category members.

We created the *control model* by blocking indirect activations in the separate model, whereby newly activated nodes (in the linguistic component) or concept points (in the sensorimotor component) could no longer propagate activation outward under any circumstances. That is, activation spread only from the initial seed node/concept point without repropagating, meaning that only direct activation from the category concept could cause another concept to be considered as a potential category member (i.e., one-hop activation). The output of the control model was two lists, one for each component, comprising the set of activated category members and their time to the first activation. Comparing the control and separate models allowed us to establish the utility of indirect activation in category production.

We then created the *combined model* by syncing the linguistic and sensorimotor components from the separate model so that they still allowed indirect activations, but this time ran in parallel and produced a single, combined output list of activated category members. In order to sync the components, we co-registered the speed of propagation in each component so that over all categories, the same number of clock-ticks were required to activate the first three members of a category. That is, by ensuring that both components were equally fast to activate their first category members, we put them on an equal footing before allowing them to run in parallel. In the combined model, therefore, seeding a particular category (e.g., ANIMAL) equated to simultaneously activating the category node in the linguistic component and the category concept point in the sensorimotor component, then allowing activation to spread across both components: Fig. 2 illustrates the process. The output from the combined model was a single list of activated category members, in the order they were first activated (regardless of the component responsible), along with the time to first activation (tick number of the model clock). Comparing the combined and separate models allowed us to establish whether one or both types of information—linguistic distributional or sensorimotor—was important to category production.

4.1.2. Materials

We used all 2551 distinct category–member pairs produced by participants in Experiment 1a to test the models, which comprised the 2228 pairs that were analyzed in Experiments 1a and b plus a further 323 pairs that were excluded from analysis in those experiments because they were missing values on one or more predictor variables; idiosyncratic responses were excluded. The linguistic component of the model operated on the category and member terms used to calculate linguistic proximity in Experiment 1, and the sensorimotor component operated on the category and member terms used to calculate sensorimotor similarity. Multiword items were handled differently depending on whether they were category or member concepts. For multiword category names (e.g., WATER BIRD), the initial activation level

of 1.0 was divided and seeded simultaneously for each constituent word (e.g., WATER and BIRD were both given activation of 0.5; function words excluded) in each component, with the exception of a few multiword terms that already constituted a single entry in the sensorimotor component and hence could be seeded as a single concept point (e.g., CITRUS FRUIT). For multiword member concepts (e.g., *horse riding* as a type of SPORT), we considered the concept to be activated at the point when one of the constituent words was first activated (e.g., *horse* or *riding*; function words again excluded). Coverage of the item set was 92.8% in the linguistic component, 94.4% in the sensorimotor component, and 97.0% in both components combined).

4.1.3. Evaluation of model performance

To evaluate model performance, we compared the member concept produced by participants for each category to the list of responses the model produced for each category. As noted in Experiment 1 (Section 2.2), participants tended to produce very diverse responses for most categories, whereby it was possible for two individual participants to produce entirely non-overlapping lists of member concepts for the same category. Such patterns of behavior are to be expected when past experiences, and therefore conceptual knowledge, vary enormously from person to person (Connell & Lynott, 2014b). Therefore, rather than evaluating the model based on how well it approximated *average* human performance—a measure on which many individual participants would score poorly—we opted to evaluate how well it fitted within the *bounds* of typical human performance (i.e., $M \pm 1\ SD$) as recommended by Wingfield and Connell (2020) for evaluation of cognitive models. In other words, given that the model approximates the conceptual knowledge of an adult native speaker of English, we evaluated whether the model performed about as well as an individual human.

4.1.3.1. Hit rate measures. We first established typical human performance in category production by transforming two variables from Experiment 1 that covered the entire dataset—*production frequency* and *mean rank*, and then calculating a *participant hit rate* measure for each variable. For *production frequency* (i.e., how popular was each member concept per category?), we rank-transformed the frequency of each response within its category so that rank = 1 was assigned to the response that had the highest production frequency per category (i.e., the most popular member concept for that category), rank = 2 was assigned to the response with the second-highest production frequency, and so on. As some categories had very large numbers of members that produced a long tail of ranks in production frequency that could skew evaluation, we limited the number of ranks we would consider to $M + 2SD$ of the most frequently produced member concepts, leading to the maximum *ranked production frequency* = 43, and excluding 59 category–member pairs. For *mean rank* (i.e., how early in the list of responses per category was each member concept named?), we transformed the mean rank of each response within its category by rounding down to an integer rank so that for each category, *rounded mean rank* = 1 was assigned to all responses whose mean rank was less than 2.0, *rounded mean rank* = 2 was assigned to all responses with mean rank greater than or equal to 2.0 but less than 3.0, and so on. The maximum *rounded mean rank* = 22. Finally, for each value of *ranked production frequency* and *rounded mean rank*, we computed

the *hit rate* for each individual participant (i.e., the proportion of the categories in which that participant produced the group response):

$$\text{participant}_i \text{ hit rate } (k) \frac{\text{\# participant}_i'\text{s responses with rank} - k}{\text{\# categories seen by participant}_i} \tag{1}$$

In this way, hit $\text{rate}_{\text{PF}}(1 \ldots 43)$ essentially represented how often a given participant produced the most frequent (second-most frequent, third-most frequent, etc.) response across categories, and hit $\text{rate}_{\text{MR}}(1 \ldots 22)$ represented how often a given participant produced the earliest (second-earliest, third-earliest, etc.) response across categories. Note that, due to tied ranks, it was possible for participants to score hit rate_{MR} with values greater than 1.0. Calculating the mean and standard deviation of these hit rates (i.e., hit $\text{rate}_{\text{PF}}(k)$ or hit $\text{rate}_{\text{MR}}(k)$) then allowed us to see how often participants, overall, tended to name member concepts at a given rank k. For example, on average, participants named 75.3% ($SD = 12.1\%$) of the top-most popular member concepts in each category (i.e., hit $\text{rate}_{\text{PF}}(1)$), but only 38.6% ($SD = 12.7\%$) of the fifth-most popular members (i.e., hit $\text{rate}_{\text{PF}}(6)$). Similarly, participants named 52.7% ($SD = 18.9\%$) of the member concepts with the earliest ordinal rank within its category (i.e., hit $\text{rate}_{\text{MR}}(1)$, reflecting the fact that categories with diverse responses often had no concepts with mean rank < 2), but 94.4% ($SD = 18.2\%$) of the second-earliest concept (i.e., hit $\text{rate}_{\text{MR}}(2)$).

Next, we established how often model performance fell within these bounds of typical human performance. We calculated *model hit rate* using a similar formula to the above (Eq. 1), reflecting the proportion of categories for which the model produced the human response at a particular rank k of *ranked production frequency* and *rounded mean rank*:

$$\text{model hit rate } (k) = \frac{\text{\# model responses with rank} = k}{\text{\# categories}}. \tag{2}$$

We could therefore directly compare the hit rates at a given rank (hit $\text{rate}_{\text{PF}}(k)$ and hit $\text{rate}_{\text{MR}}(k)$) of the model to those of participants. As with participants, tied ranks meant that it was possible for the model to score hit rate_{MR} with values greater than 1.0.

Finally, to summarize model performance in a single statistic per dependent variable, we computed the percentage of model hit rates that fell within 1 *SD* of the participant mean hit rate (i.e., how often did the model perform about as well as a typical participant?). In order to ensure that model hit rates of 0 did not artificially boost the model's performance (i.e., even when participants did not score well and also had hit rates close to 0), we computed summary percentages using only ranks before the human *SD* region started to include 0. The top ranks used for evaluation were 1–28 for *ranked production frequency* and 1–13 for *rounded mean rank*. Excluding the trailing ranks meant excluding category–member pairs that were produced only occasionally by participants and reducing the total number of category–member pairs by 268 for *ranked production frequency* (2283 items remaining) and by 119 for *rounded mean rank* (2432 items remaining). Summary statistics computed from the whole width of the

graph are included in the Supplementary Materials Table S2, with supporting data and results.

4.1.3.2. Choosing the optimal model. Just as participants were given limited time to produce their responses, we halted the model's operation after a fixed duration rather than let it run indefinitely. We selected a stopping point to maximize the summary hit-rate percentage for the combined model, where stopping too early led the model to activate too few category members (i.e., *model hit rate* below *participant hit rate M* – 1 *SD*), and stopping too late led the model to activate unrealistically large numbers of category members (i.e., the *model hit rate* above *participant hit rate M* + 1 *SD*). The optimal stopping points for hit rate_{PF} and hit rate_{MR} summary measures were close, but the peak of performance for both measures did not coincide. We chose to terminate the model at the best compromise point, which maximized performance for *ranked production frequency* while still achieving near-maximum performance for *rounded mean rank* (a time of 305 ticks on the model clock; see the Supplemental Materials Fig. S1 for the plot of performance on each measure across model termination times). This stopping point was used for all three model versions, control, separate, and combined, although we noted that all direct activations had been completed in the control model before this time was reached.

4.1.3.3. Design and analysis. We first tested whether allowing indirect activation improved model performance by comparing hit rates per rank in the separate model versus the control model (per component and measure), using directional signed-ranks tests in both NHST and Bayesian form (default Cauchy prior distribution with $r = .707$; JASP Team, 2020; van Doorn, Ly, Marsman, & Wagenmakers, 2020). Likewise, we tested whether combining linguistic and sensorimotor information improved model performance by comparing in turn hit rates per rank in the combined model versus separate linguistic and sensorimotor components.

In addition, in the combined model, we tested whether the linguistic component dominated model performance by counting whether each activated member concept first originated in the linguistic or sensorimotor component and using directional binomial tests in both NHST and Bayesian form (beta prior $a = b = 1$; JASP Team, 2020) to examine whether the proportion of concepts that originated in the linguistic component exceeded 50%.

4.2. Results and discussion

4.2.1. Hit rate performance per model version

Fig. 3 shows the model hit rates and summary hit-rate percentage for *ranked production frequency* and *rounded mean rank*, respectively, across each model version (control, separate, combined). Overall, a clear improvement in model performance can be seen as models progress between versions; Table 5 shows the results.

The control models' performance was below the level of participants'. With only direct category → member activations allowed, model performance on *ranked production frequency* was 53.6% for the linguistic component (i.e., model hit rates fell within 1 *SD* of the human mean a little more than half the time) and a much worse 3.6% for the sensorimotor component.

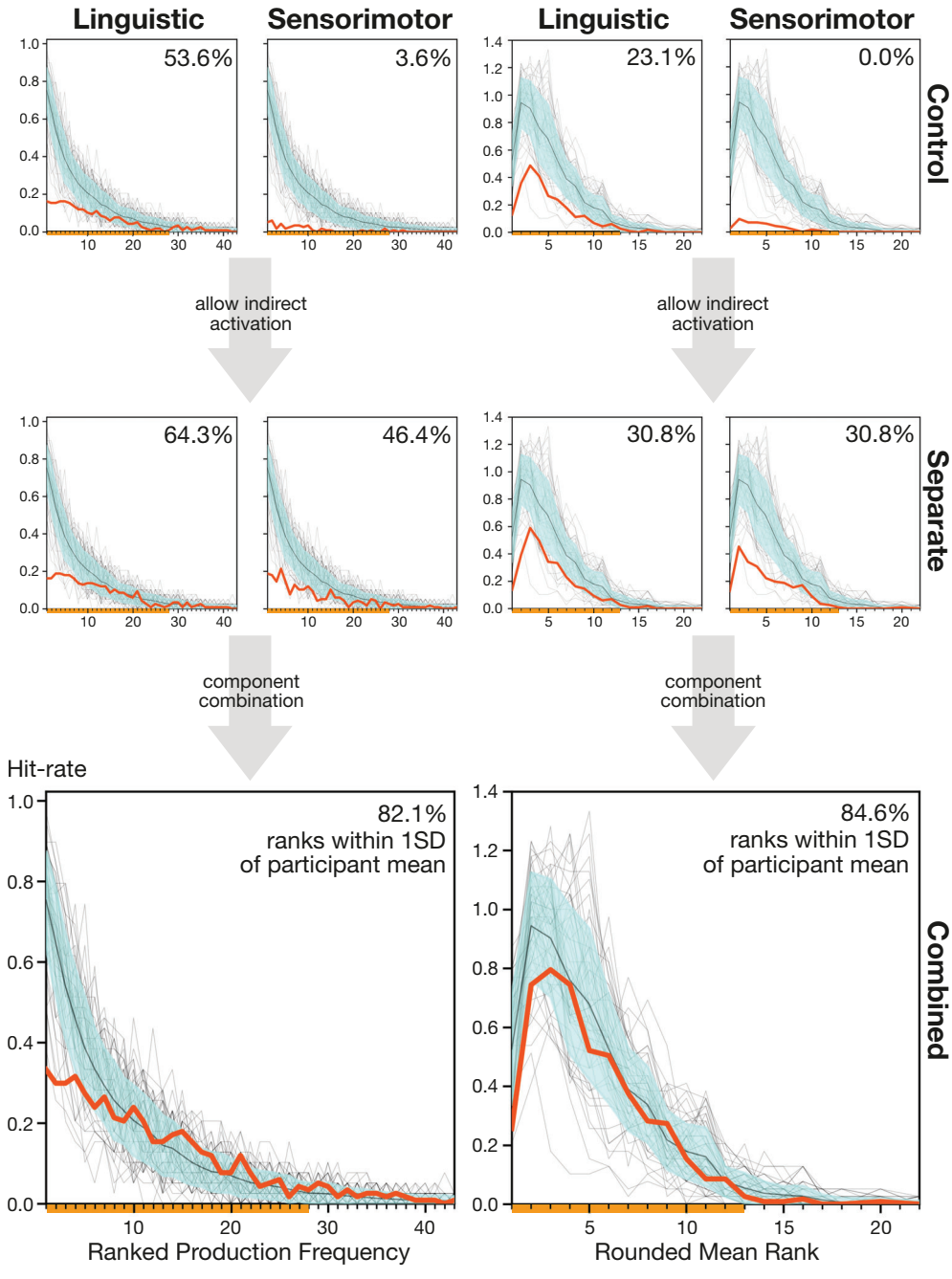

Fig 3. Hit rates for *ranked production frequency* (left) and *rounded mean rank* (right), showing the model performance (thick red line), compared to human performance (fine gray line per individual participant, fine black line for participant mean, blue shaded region representing 1 *SD* around the participant mean). The summary percentage of model performance per panel refers to the proportion of model hit rates that fell within 1 *SD* of the participant mean, calculated for high ranks (orange bar on x-axis) before the human *SD* range spanned zero.

Table 5
Comparison of different versions of computational model in category production performance

Comparison	*W*	Log BF
Ranked Production Frequency		
Linguistic separate model versus linguistic control	325.0***	9.77
Sensorimotor separate model versus sensorimotor control	378.0***	8.76
Combined model versus linguistic separate model	378.0***	12.49
Combined model versus sensorimotor separate model	378.0***	10.01
Rounded Mean Rank		
Linguistic separate model versus linguistic control	78.0**	5.12
Sensorimotor separate model versus sensorimotor control	91.0***	6.18
Combined model versus linguistic separate model	91.0***	6.47
Combined model versus sensorimotor separate model	91.0***	6.63

Note. Comparisons test whether indirect activations improved performance (separate vs. control) and whether a combination of both linguistic and sensorimotor components was better than one component alone (combined vs. separate). W = Wilcoxon signed-rank test statistic; log BF = natural log Bayes factor, where negative values indicate evidence for the null model and positive values indicate evidence for the alternative model: log BFs $\geq$ 3.00 (equivalent to BFs $>$ 20) constitute strong evidence; log BFs $\geq$ 1.10 (equivalent to BFs $>$ 3) constitute positive evidence; and log BFs between –1.10 and 1.10 (equivalent to BFs of 0.33 and 3) constitute equivocal evidence.
$^{**}p < .01$, $^{***}p < .001$.

Control model performance on *rounded mean ranks* was worse at 23.1% for the linguistic component and 0.0% for the sensorimotor component. That is, although the control model produced some relevant member concepts for some categories, it activated too few category members overall to come anywhere near the level of human performance, particularly for the most important (i.e., top-ranked) member concepts.

The separate models, which allowed indirect activation to spread between concepts, performed markedly better. Model performance in the separate linguistic component fell within typical human bounds 64.3% of the time for *ranked production frequency*, which was much better than the control version in both Bayesian and NHST terms. In the sensorimotor component, performance was 46.4%, again much better than the control version. Results were similar for *rounded mean rank*, where the separate linguistic component fell within human bounds 30.8% of the time and performed more strongly than the corresponding linguistic control model in both Bayesian and NHST analyses. The separate sensorimotor component also achieved a performance of 30.8%, which was again much better than the corresponding sensorimotor control. These results show that allowing activation to spread indirectly between concepts, in both linguistic distributional knowledge and sensorimotor conceptual knowledge, critically improves how well the model can approximate human performance in category production.

Finally, the combined model, which synced the linguistic and sensorimotor components to produce a single combined list of activated category members, performed best of all. For *ranked production frequency*, the model achieved 82.1% of its hit rates within typical human performance, which strongly outperformed the separate models of both the linguistic component and the sensorimotor component in Bayesian and NHST analyses. For

rounded mean rank, model performance was 84.6%,[9] again strongly outperforming the separate linguistic component and sensorimotor component. These results indicate that both forms of information—linguistic and sensorimotor—are essential to the model's ability to capture human performance in category production.

4.2.2. Dominant component

In order to examine whether linguistic or sensorimotor information was primarily responsible for the models' success in category production, we counted whether each activated member concept in the combined model's output list originated in the linguistic or sensorimotor component at the time it was first activated. Of the 575 category–member pairs activated by the optimal stopping point,[10] more member concepts originated in the linguistic component ($N = 351$: 61.0%) than in the sensorimotor component ($N = 224$: 39.0%), $p < .001$, $\log BF_{10} = 11.86$. This result suggests that—given the excellent fit of the model to human performance—linguistic distributional information may be slightly more dominant than sensorimotor information in conceptual processing during category production. Finding that most member concepts reached threshold activation in the linguistic component before the sensorimotor component is consistent with the linguistic shortcut hypothesis (Connell, 2018; Connell & Lynott, 2014b) and related theories of the conceptual system, which hold that linguistic distributional information reaches peak activation before sensorimotor simulation (Barsalou et al., 2008; Louwerse, 2011).

4.2.3. Model versus individual human performance

Since our aim was to examine whether the combined linguistic–sensorimotor model performed about as well as an individual human in approximating typical human performance in category production, we compared their performance in a number of ways. In terms of overall category production, the model activated an average of 4.91 member concepts per category by the optimal stopping point (i.e., 575 category–member pairs divided by 117 categories), which is well within the $M \pm SD$ range reported for individual participants in Experiment 1a ($M = 6.18$, $SD = 3.90$).

In terms of hit-rate performance, it is useful to compare the summary hit rates of the model with that of participants (i.e., how often did *participants'* individual hit rates fall within 1 *SD* of the group mean?). For *ranked production frequency*, individual participant hit rates fell in typical (i.e., $M \pm 1$ *SD*) bounds between 4% and 89% of the time (average 67%), compared to 82.1% for the combined model. For *rounded mean rank*, it was between 15% and 100% for participants (average 68%), compared to 84.6% for the combined model. The wide range of summary hit rates for humans reflects the diversity of responses we saw in Experiment 1a, whereby some participants' category production behavior was relatively eccentric while others' was relatively conventional. Model behavior, in that sense, was more conventional in its performance than many participants. Overall, for both *ranked production frequency* and *rounded mean rank*, the model was at least as good as an individual participant at approximating typical human behavior in category production.

5. General discussion

The present work provides strong evidence that linguistic distributional knowledge and sensorimotor grounding both contribute to the generation of member concepts during a category production task. In Experiment 1a, we found that both measures of sensorimotor similarity and linguistic proximity independently predicted three direct measures of category production (*production frequency*, *mean rank*, and *first-rank frequency*), whereas linguistic proximity alone predicted the implicit measure of processing effort in category production (*RT*; although the effect was weak). Similarly, we found in Experiment 1b that sensorimotor similarity and linguistic proximity outperformed typicality ratings as predictors of *mean rank*. Although this pattern did not emerge for the other measures of category production, sensorimotor similarity still uniquely contributed to *production frequency* and *mean rank* over and above typicality ratings, while linguistic proximity contributed to *production frequency*, *mean rank*, and *first-rank frequency* over and above typicality ratings. Finally, using a novel computational model that incorporated both linguistic distributional and sensorimotor information, we found that the model best approximated typical human performance in *production frequency* and *mean rank* when conceptual activation was allowed to spread indirectly between words and between sensorimotor representations of concepts and when candidate category members came from both sensorimotor and linguistic distributional representations. That is, both forms of information—linguistic distributional and sensorimotor—were critical to how well the model could approximate human performance in category production, although linguistic information was responsible for activating the majority of category members. When both forms of information were included, the model performed at least as well as a typical human.

These findings support linguistic–simulation theories of the conceptual system (Barsalou et al., 2008; Connell & Lynott, 2014b; Louwerse, 2011) as the basis for conceptual processing during categorization. As predicted, the more similar a member concept's sensorimotor profile is to that of its category concept (e.g., how much the sensorimotor experience of *cat* overlaps with the experience of ANIMAL), and the more often a member concept's label co-occurs with its category label (e.g., how often the word "cat" shares a context with the word "ANIMAL"), the more often people name it, and the earlier in rank order the concept is named. Critically, we used a novel, fully grounded measure of sensorimotor experience, based on modality-specific ratings of perceptual strength and effector-specific ratings of action strength. Thus, we ensured that category and member concepts were compared on the basis of a pure sensorimotor profile rather than on the basis of a relatively restricted set of features that are limited to those attributes that can be easily verbalized (e.g., feature production norms) and/or that include abstracted information with unclear grounding (e.g., taxonomic and encyclopedic features such as *cat*: [baby is a kitten, is domesticated, has four legs]). Our findings thus provide some of the first evidence that concepts functionally group together according to the similarity of their sensorimotor representations and do not necessarily require abstracted features to accomplish it (cf. McRae et al., 1997; Tyler et al., 2000). Future work will examine in more detail how the emergent structure of categorical distinctions relies on grounded sensorimotor experience.

More specifically, our findings support the linguistic shortcut hypothesis (Connell, 2018) as the process by which people arrive at the most frequently named and first-named members of a category. In Experiment 1a, linguistic proximity was a more important predictor of *production frequency* and *first-rank frequency* than was sensorimotor similarity, and in our computational model, the linguistic component was responsible for first activating more member concepts than the sensorimotor component. These results suggest that participants are able to rely on linguistic distributional information as a computationally cheaper response heuristic (i.e., a shortcut) to activate concepts in place of full sensorimotor simulation. That is, when retrieving and selecting a suitable category member from long-term memory, representation of a concept in the form of its linguistic label may be sufficient for the task. Since linguistic distributional information tends to reach peak activation (i.e., sufficient conceptual activation to inform a response) before sensorimotor simulation (Barsalou et al., 2008; Connell & Lynott, 2014b; Louwerse, 2011), it means that responses that originate from label-to-label activation tend to occur earlier than those that originate from simulation-to-simulation activation. Further support for the linguistic shortcut hypothesis, albeit weaker, came from the analysis of *RT*s for first-named category members. As predicted, the more often the label of a member concept co-occurs with its category label (e.g., how often the word "cat" shares a context with the word "ANIMAL"), the more likely and faster people are to name that concept first as a category member; however, the effect was relatively small and the evidence equivocal in Bayesian terms. The lack of sensorimotor similarity effects on the time course of category production was unexpected, as we had initially predicted that sensorimotor similarity would still play a role even if linguistic distributional information was dominant, and had found such sensorimotor effects in related work (Van Hoef et al., 2020). It is possible that our method of eliciting verbal responses led to rather noisier RT data than if, for example, we had required participants to press a key as soon as they thought of a category member. For instance, even our lexical predictors (word frequency, number of syllables, and phonological Levenshtein distance) performed poorly in predicting *RT* variance (<1% variance explained), whereas random effects of participant and category accounted for around 36% of the variance in *RT*s. Nonetheless, this pattern of findings is consistent with the use of linguistic distributional information as a rapid linguistic shortcut when selecting category members, and future work should examine the time course of category production in more detail.

Indeed, the only good predictor of *RT* was typicality (Experiment 1b), where we replicated previous findings that typicality ratings predict category production (Hampton & Gardiner, 1983; Mervis et al., 1976); that is, the more typical a concept is of its category, the more frequently, the more quickly, and the earlier it was named. Typicality ratings have traditionally been assumed to reflect the inherently graded nature of semantic categories as part of feature-based theories of category membership (e.g., Osherson & Smith, 1981; Rosch, 1975; Rosch et al., 1976). However, the measure is somewhat problematic in that it is unclear precisely what typicality judgments are based on; for example, prototypicality effects can be observed in well-defined categories where membership relies on a single binary feature (e.g., EVEN NUMBER; Armstrong et al., 1983) and in ad hoc categories that have no stored structure in long term memory (e.g., WAYS TO MAKE FRIENDS; Barsalou, 1983), both of which are inconsistent with the notion of prototypicality and graded, feature-based category

structure. It is thus unclear what typicality actually represents in a specific, operationalized sense that goes beyond an intuitive appeal to goodness-of-membership, which means—despite its predictive ability—it is also unclear precisely *how* typicality contributes to the process of category production. Nevertheless, defining what typicality represents is beyond the scope of the current paper. As our sensorimotor and linguistic measures were only weakly correlated with typicality rating and accounted for category production responses beyond this measure, we can be confident that whatever typicality represents, it reflects different cognitive processes to sensorimotor similarity and linguistic proximity in the category production task.

Given that category production is an unconstrained, free-response task that produces highly variable data amongst participants, it is notable that our measures succeeded in predicting variance in all dependent measures. When asked to list as many members of a given category as possible in 60 s, our participants produced highly divergent responses that ranged from commonplace to downright eccentric, even when excluding idiosyncratic responses that were produced by only one participant. For example, in the categories of BOAT or INFECTIOUS DISEASE, it often happened that two individual participants produced entirely non-overlapping lists of member concepts. Systematic errors were also common as previously found in category production (e.g., Battig & Montague, 1969): For instance, our participants produced both *chocolate* and *eggs* as members of the category DAIRY PRODUCT and produced both *kiwi* and *pineapple* as members of the category CITRUS FRUIT. With such variability in the data, it is perhaps unsurprising that the fixed-effect sizes in our behavioral experiments were relatively modest in terms of explained variance. Overall, the lexical and critical predictors in Experiment 1a accounted for only 8% (*mean rank*) to 9% (*production* and *first-rank frequency*) of the variance for explicit measures of category production, and even with the addition of typicality ratings in Experiment 1b, the explained variance ranged from 11% (*mean rank*) to 23% (*production frequency*). One possible reason is that variables such as typicality tend to correlate with production frequency more weakly for ad hoc categories than for common taxonomic categories (Barsalou, 1983), and a large number of our categories were perhaps closer to goal-derived, ad hoc categories (e.g., LIVING ROOM FURNITURE, GARDENING TOOL, POSITIVE PERSONAL QUALITY) than to traditional taxonomic categories (e.g., TOOL, FRUIT, EMOTION). Regardless, our computational model very successfully replicated the human responses from Experiment 1a, including the errors (e.g., the model also activated *kiwi* and *pineapple* for CITRUS FRUIT). In other words, even though human behavior in category production tasks is somewhat variable, our model still successfully approximates it.

Our computational model of conceptual activation during category production is novel in a number of ways. Like some related models of conceptual processing in other domains (e.g., Andrews, Vigliocco, & Vinson, 2009; Johns & Jones, 2012), it implements a two-component conceptual system: one to reflect linguistic distributional knowledge and the other to reflect the sensorimotor simulation system. However, unlike such earlier work, which implemented the simulation system via discrete abstracted features that were only partly perceptual, our model is fully grounded in that each word label (e.g., *cat*) is linked to a multidimensional profile of the strength of sensorimotor experience across a range of perceptual modalities and action effectors. In addition, unlike the above and other computational models of category

production (e.g., Hills et al., 2012; Taler et al., 2020), we did not restrict the search space of the model to a few hundred concepts or a single category at a time. Instead, our model comprises some 40,000 concepts in both linguistic distributional and sensorimotor form, which approximates a full-size conceptual system for an educated, adult native speaker of English (Lynott et al., 2020). Even the modeling of linguistic distributional information was based on a corpus size of 200 million words, which (unlike linguistic distributional models trained on several billion words) approximates the lower bound of lifetime language experience in an adult speaker of English (Wingfield & Connell, 2020). With such implementational choices, we believe our model to be one of the most cognitively plausible computational cognitive models to date of the human conceptual system.

One important implication of our model architecture is that its goal was to capture conceptual activation during category production and not to model the entire cognitive process of category production. That is, given that the model is based on a full-size conceptual system of 40,000 concepts, it inevitably activates both relevant category members and irrelevant neighboring concepts while spreading activation from the category concept outward. For example, the category ANIMAL activates relevant *cat* and irrelevant *raindrop* in the sensorimotor component. In humans, one might assume that the nature of the category production task requires some form of top-down filter mechanism, whereby each activated concept is evaluated as a candidate category member and those candidates like *cat* that clear a threshold of evidence are named aloud (or, alternatively, those candidates like *raindrop* that fail to clear the threshold are suppressed). It was not a goal of the model to capture such a top-down process, and so we evaluated the model using the (relevant) member concepts produced by humans for each category. However, another possibility is that a top-down filter mechanism may not be necessary if the linguistic and sensorimotor components interacted with each other during spreading activation so that concepts activated in both components (e.g., sensorimotor concept point *cat* and linguistic word node "cat") would mutually boost each other above the level of concepts activated only in one component (e.g., sensorimotor concept point *raindrop*). That is, the selection of relevant over irrelevant candidate concepts may emerge spontaneously as a property of linguistic and sensorimotor components interacting and mutually reinforcing one another in cycles of spreading activation. Such inter-component interaction is a core part of how linguistic-simulation theories assume conceptual processing works, where linguistic information can activate simulated information, which in turn can activate further linguistic information, and so on (e.g., Barsalou et al., 2008; Connell & Lynott, 2014b). The computational implementation of such complex interactions is non-trivial, but we plan to investigate it in future research.

6. Conclusion

In summary, the present paper demonstrates that both linguistic distributional knowledge and sensorimotor grounding influence how particular concepts are selected over others during category production (a.k.a. semantic fluency). Through behavioral experimental and computational modeling work, we show that both sensorimotor information (e.g., how alike in sensorimotor experience is the member concept *cat* to the category concept ANIMAL?)

and linguistic distributional information (e.g., how often does the member concept label *cat* appear in the same contexts as the category concept label ANIMAL?) independently contribute to category production measures, supporting the linguistic shortcut hypothesis that label-to-label associations can often suffice in conceptual processing instead of relying wholly on more intensive sensorimotor simulation. Moreover, our novel, fully grounded computational model of conceptual activation during category production—whose performance was indistinguishable from that of human participants—shows that spreading activation indirectly between word labels, and between sensorimotor concept representations, greatly improves the predictive ability of sensorimotor and linguistic distributional information, and suggests that linguistic distributional information may play a dominant role in generating candidate member concepts during category production. While future work should investigate the time course of category production in more detail, our findings support recent theories, which include both simulated and linguistic distributional information as inherent components of semantic processing, providing strong evidence that the conceptual system is both grounded and linguistic in nature.

Acknowledgment

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 682848) to LC. Thanks to Alex Jironkin for helpful advice regarding model implementation.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data are Materials are available at https://osf.io/vaq56/.

Notes

1 For clarity throughout this paper, we follow the convention of reporting category names in uppercase (e.g., ANIMAL) and category members in lowercase italics (e.g., *cat*, *dog, aardvark*).

2 Hills et al. (2012) optimal foraging model is also based on a corpus-derived linguistic distributional space, but it focuses on member-to-member relationships (e.g., *dog* → *cat*) rather than the category-to-member (e.g., ANIMAL → *cat*) relationships we describe here. Similarly, Taler et al. (2020) examined the role of neighborhood density in a corpus-derived linguistic distributional space, but this measure concerns the member concept alone (e.g., neighbourhood density of *cat*) rather than the category-to-member (e.g., ANIMAL → *cat*) relationships we describe here. Nonetheless, both of these models support the general utility of linguistic distributional information in modeling category production.

3 Use of a corpus of approximately 200 million words is in line with recommendations regarding plausible estimates of lifetime language exposure in adults (Wingfield & Connell, 2020; also see Brysbaert, Stevens, Mandera, & Keuleers, 2016). The subtitles corpus covered 99.5% of words found in category and member concepts, with 98.1% occurring at least 10 times.

4 As recommended by Wingfield and Connell (2020), we based our choice of distributional window size (five words around target word = 6-gram) and corpus (subtitles) on the medium-to-high conceptual complexity of the task and used an empirical approach to select the most appropriate linguistic distributional model. Our selected PPMI *n*-gram measure outperformed a range of alternative measures derived from count and predict vector models, which is consistent with previous research showing that, when trained on adequately large corpora, *n*-gram measures can successfully capture semantic effects previously assumed to require more complex distributional models (Louwerse, 2011).

5 The present experiment was developed in parallel with a separate investigation using a closely related measure of sensorimotor similarity (Van Hoef et al., 2020); since both studies used this new measure at the same time, we feel both reports can legitimately describe its use as novel.

6 As an exploratory analysis suggested by an anonymous reviewer, we conducted the analyses for Study 1a (for the three category production measures) separately for abstract and concrete categories. Both abstract and concrete categories showed identical patterns of effects to the overall analysis, where sensorimotor similarity and linguistic proximity independently contributed to the three measures of category production, but neither predictor consistently dominated either type of category. That is, we did not observe the pattern that abstract categories relied predominantly on linguistic information, whereas concrete categories rely predominantly on sensorimotor information (e.g., Crutch & Warrington, 2005; Dove, Barca, Tummolini, & Borghi, 2020; Pecher & Zeelenberg, 2018; Vigliocco et al., 2009). These analyses and results are available as Supplementary Materials at https://osf.io/vaq56/.

7 We thank an anonymous reviewer for this suggestion.

8 The limit of 3000 concepts was not entirely arbitrary but rather reflected a very rough approximation of the number of concepts in long-term memory that may retain trace activation in the context of an ongoing task, such as when maintaining the plot of a novel in a situation model (e.g., Zwaan & Madden, 2005) or when asked to remember large sets of pictured objects (e.g., Brady, Konkle, Alvarez, & Oliva, 2008).

9 Note that because the optimal stopping point for the model was not at the maximum performance for *rounded mean rank* (see section "Choosing the Optimal Model"), it meant the model was technically capable of better performance on this measure. Peak performance for *rounded mean rank* occurred nine ticks later (314 ticks on the model clock) at 92.3%.

10 A further 861 category–member pairs were activated after the optimal stopping point, bringing the total to 1436 member concepts; as noted earlier (see Section 4.1.3, Choosing the Optimal Model), this level of model performance exceeded typical human bounds and is not examined further.

REFERENCES

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498. https://doi.org/10.1037/a0016261

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308. https://doi.org/10.1016/0010-0277(83)90012-4

Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General*, *113*(4), 518–540. https://doi.org/10.1037/0096-3445.113.4.518

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525x99002149

Barsalou, L. W., Santos, A., Simmons, K. W., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg & A. C. Graesser (Eds.), *Symbols, embodiment and meaning* (pp. 245–283). New York: Oxford University Press.

Barton, K. (2017). *Package 'MuMIn'*. Retrieved from https://CRAN.R-project.org/package=MuMIn

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, *80*(3, Pt.2), 1–46. https://doi.org/10.1037/h0027577

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer*. Retrieved from http://www.praat.org/

Bonner, M. F., & Grossman, M. (2012). Gray matter density of auditory association cortex relates to knowledge of sound concepts in primary progressive aphasia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(23), 7986–7991. https://doi.org/10.1523/JNEUROSCI.6241-11.2012

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526. https://doi.org/10.3758/BF03193020

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology*, *20*(3), 213–261. https://doi.org/10.1080/02643290244000266

Casey, P. J. (1992). A reexamination of the roles of typicality and category dominance in verifying category membership. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 823–834. https://doi.org/10.1037/0278-7393.18.4.823

Cerhan, J. H., Ivnik, R. J., Smith, G. E., Tangalos, E. C., Petersen, R. C., & Boeve, B. F. (2002). Diagnostic utility of letter fluency, category fluency, and fluency difference scores in Alzheimer's disease. *The Clinical Neuropsychologist*, *16*(1), 35–42. https://doi.org/10.1076/clin.16.1.35.8326

Cohen, B. H., Bousfield, W. A., & Whitmarsh, G. A. (1957). Cultural norms for verbal items in 43 categories. Technical Report No. 22, University of Connecticut, Contract Nonr. 631(00), Office of Naval Research.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Connell, L. (2018). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, *3798*(May), 1–11. https://doi.org/10.1080/23273798.2018.1471512

Connell, L., & Lynott, D. (2010). Look but don't touch: Tactile disadvantage in processing modality-specific words. *Cognition*, *115*, 1–9. https://doi.org/10.1016/j.cognition.2009.10.005

Connell, L., & Lynott, D. (2012). When does perception facilitate or interfere with conceptual processing? The effect of attentional modulation. *Frontiers in Psychology*, *3*, 474. https://doi.org/10.3389/fpsyg.2012.00474

Connell, L., & Lynott, D. (2013). Flexible and fast: Linguistic shortcut affects both shallow and deep conceptual processing. *Psychonomic Bulletin & Review*, *20*(3), 542–550. https://doi.org/10.3758/s13423-012-0368-x

Connell, L., & Lynott, D. (2014a). I see/hear what you mean: Semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General*, *143*, 527–533. https://doi.org/10.1037/a0034626

Connell, L., & Lynott, D. (2014b). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*(3), 390–406. https://doi.org/10.1111/tops.12097

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170143. https://doi.org/10.1098/rstb.2017.0143

Connell, L., Lynott, D., & Dreyer, F. (2012). A functional role for modality-specific perceptual systems in conceptual representations. *PLoS ONE*, *7*(3), e33321. https://doi.org/10.1371/journal.pone.0033321

Connell, L., & Ramscar, M. (2001). Using distributional measures to model typicality in categorization. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, Scotland (pp. 226–231).

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain: A Journal of Neurology*, *128*(Pt 3), 615–627. https://doi.org/10.1093/brain/awh349

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16396–16400. https://doi.org/10.1073/pnas.1009438107

Dove, G., Barca, L., Tummolini, L., & Borghi, A. M. (2020). Words have a weight: Language as a source of inner grounding and flexibility in abstract concepts. *Psychological Research.* https://doi.org/10.1007/s00426-020-01438-6

Dymarska, A., Connell, L., & Banks, B. (2020). Working memory for object concepts relies on linguistic labels. *Department of Psychology, Lancaster University*. Manuscript in preparation.

Goodhew, S. C., McGaw, B., & Kidd, E. (2014). Why is the sunny side always up? Explaining the spatial mapping of concepts by language use. *Psychonomic Bulletin & Review*, *21*(5), 1287–1293. https://doi.org/10.3758/s13423-014-0593-6

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, *74*(4), 491–516. https://doi.org/10.1111/j.2044-8295.1983.tb01882.x

Harnad, S. (2006). To cognize is to categorize: Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 20–42). Amsterdam: Elsevier.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

Heyman, T., & Heyman, G. (2019). Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology (2006)*, *72*(8), 2084–2109. https://doi.org/10.1177/1747021819830949

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

JASP Team. (2020). *JASP* (0.14.1) [Computer software].

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Kuznetsova, A., Brockhoff, B., & Christensen, R. H. B. (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

Larochelle, S., Richard, S., & Soulières, I. (2000). What some effects might not be: The time to verify membership in "well-defined" categories. *The Quarterly Journal of Experimental Psychology Section A*, *53*(4), 929–961. https://doi.org/10.1080/713755940

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*(2), 381–398. https://doi.org/10.1111/j.1551-6709.2010.01157.x

Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega (Ed.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 309–326). New York: Oxford University Press. https://doi.org/10.1093/acprof:oSo/9780199217274.003.0015

Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002

Lynott, D., & Connell, L. (2010). Embodied conceptual combination. *Frontiers in Psychology*, *1*, 212. https://doi.org/10.3389/fpsyg.2010.00212

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

McEvoy, C. L., & Nelson, D. L. (1982). Category name and instance norms for 106 categories of various sizes. *The American Journal of Psychology*, *95*(4), 581–634. https://doi.org/10.2307/1422189

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(6), 1155–1172. https://doi.org/10.1037/0278-7393.18.6.1155

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*(3), 283–284. https://doi.org/10.3758/BF03337190

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x

Moreno-Martínez, F. J., Montoro, P. R., & Rodríguez-Rojo, I. C. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods*, *46*(4), 1088–1097. https://doi.org/10.3758/s13428-013-0435-x

Navarrete, E., Arcara, G., Mondini, S., & Penolazzi, B. (2019). Italian norms and naming latencies for 357 high quality color images. *PLoS ONE*, *14*(2), e0209524. https://doi.org/10.1371/journal.pone.0209524

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. https://doi.org/10.1016/0010-0277(81)90013-5

Pecher, D., & Zeelenberg, R. (2018). Boundaries to grounding abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170132. https://doi.org/10.1098/rstb.2017.0132

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491–502. https://doi.org/10.1037/0096-1523.2.4.491

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, *126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* [Doctoral thesis, Stockholm University]. Digitalia Vetenskapliga Arkivet DiVA. Retrieved from http://urn.kb.se/resolve?urn=urn:nBn:sE:sU:dIva-1037

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*(2), 380–394. https://doi.org/10.3758/s13428-011-0164-y

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 419–425. https://doi.org/10.1037/0278-7393.12.3.419

Taler, V., Johns, B. T., & Jones, M. N. (2020). A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the canadian longitudinal study on aging. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *75*(9), e221–e230. https://doi.org/10.1093/geronb/gbz003

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231. https://doi.org/10.1006/brln.2000.2353

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, *12*(6), 587–595. https://doi.org/10.3758/BF03201848

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. -J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Van Hoef, R., Connell, L., & Lynott, D. (2020). The effects of sensorimotor and linguistic distance on the basic-level advantage. *Department of Psychology, Lancaster University*. Manuscript in preparation.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(02), 219–247. https://doi.org/10.1515/LANGCOG.2009.011

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, *110*(5), 1273–1296. https://doi.org/10.1093/brain/110.5.1273

Wingfield, C., & Connell, L. (2020). Understanding the role of linguistic distributional knowledge in cognition. *PsyArXiv.* https://doi.org/10.31234/osf.io/hpm4z

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 224–245). Cambridge: Cambridge University Press.

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*(1), 1–11. https://doi.org/10.1037/0096-3445.135.1.1