Similarity is as Similarity Does? A Critical Inquiry into the Effect of Thematic Association on Similarity

Garrett Honke and Kenneth J. Kurtz

Binghamton University (SUNY)

Word Count: 14,191

Author Note

This article is adapted from Garrett Honke's doctoral dissertation. We thank Greg Murphy, Mitchell Rabinowitz and the members of the Learning and Representation in Cognition (LARC) and Brain and Machine (BAM) Laboratories at Binghamton University for their valuable feedback.

Correspondence concerning this article should be addressed to Garrett Honke, Department of Psychology, Binghamton University (SUNY), PO Box 6000, Binghamton University, Binghamton, NY 13902-6000.

E-mail: ghonke1@binghamton.edu

Abstract

Leading theories of psychological similarity are based on the degree of match in semantic content between compared cases (i.e., shared features, low dimensional distance, alignable relations). Broader forms of semantic relatedness such as the degree of association between cases (e.g., egg and spatula) are generally not considered to contribute to similarity judgments. However, empirical work has demonstrated a behavioral tendency to choose associated pairs over proximal pairs (i.e., high semantic content overlap) in similarity judgement tasks. As a result, dual-process models have been proposed that posit thematic integration in addition to content match as component processes of similarity. The present experiments investigate the thematic association effect in similarity in order to more clearly determine whether such a theoretical redirection is warranted. An alternative viewpoint is that confusion between similarity and association is the cause of the reported thematic bias. Experiment 1 introduces a modified similarity judgement task and addresses the impact of task instructions as a potential causal factor underlying the thematic association effect on similarity. Experiment 2 specifically compares the novel similarity task to a traditional two-alternative, forced choice triad task. Experiment 3 addresses the possibility of bias in the stimulus sets used in Experiments 1 and 2. Across the experiments we find association-based responding to be much less prevalent than in previous demonstrations: the traditional finding of a thematic preference only occurred when participants were specifically asked to select based on associativity ("goes with"). Modifications to conventional methodology that minimize biasing factors clearly attenuate the effect of association on similarity. We interpret these findings as evidence that that the thematic association effect derives from intrusions on psychological similarity, not from an additional component intrinsic to psychological similarity.

Keywords: similarity, taxonomic categories, thematic categories

Similarity is as Similarity Does? A Critical Inquiry into the Effect of Thematic Intrusion on Similarity Judgment

Taxonomic similarity, while not definitively specified in psychological theory, captures the basic notion that taxonomically-matched entities are good candidates for generalization. Inferences made about members of a taxonomic category are productive (e.g., FLOUR, CORNMEAL). Members of a taxonomic category reliably share features and relations in common; they tend to look somewhat alike or to play the same roles in situations or both (e.g., ORCA, DOLPHIN) (Goldwater, Markman, & Stilwell, 2011; Markman & Stilwell, 2001). They fill the same positions in similar schemas and events (e.g., DEER, ANTELOPE). Objects with taxonomic similarity are used for the same tasks (e.g., SHOVEL, SPOON). Critically, it must be possible to recognize similarity (commonalities in relational structure and attributes) without interference from associated entities—particularly in the service of mapping relational similarities between cases (e.g., PRESSURE and TEMPERATURE fill the same role in the *flow* schema instantiated by water transfer and heat transfer, respectively).

Thematic associates do not behave in this manner. They are less useful for induction about natural kinds (Lin & Murphy, 2001; E. M. Markman, Cox, & Machida, 1981). In contrast to the rich possibilities of inference and generalization with taxonomic category members, thematic associates are (in the simplest case) only connected by their theme. A theme consists of many possible roles and role fillers; every object present in a theme might fill a different role (Kurtz & Gentner, 2001); in this way, thematic associates lack the constraints of taxonomic category members. Thematic associates cannot be relied on as good substitutes for one another. Consider the example of COW and MILK: COW might be considered a substitute for MILK in some contexts but—unlike taxonomic category members—this relationship is highly limited and unidirectional (MILK is not helpful when what you need is a COW). Thematic associates often do have a corresponding relationship (e.g., the causal relationship between BOWLING BALL and BOWLING PIN) and this can be consequential for using and organizing general knowledge. We note that it has been argued that objects must have corresponding roles to qualify as thematic associates (e.g., Estes et al., 2011). However, restricting the definition of thematic association (for review, see Mirman, Landrigan, & Britt, 2017) to things with corresponding roles (e.g., Estes et al., 2011), things with high word co-occurrence frequencies (Jackson, Hoffman, Pobric, & Lambon Ralph, 2015) or sub-types of associates (e.g., manipulable object associates like STAPLER and PAPER; Canessa et al., 2008) would disqualify a large cross-section of valid examples of thematic association.

We take a permissive view—drawing on the idea that thematic associates can be viewed in terms of categorization (see also Jones & Love, 2007; Lin & Murphy, 2001)—where thematic category coherence only requires that two things co-occur in a situation (e.g., BOWLING PIN and ARCADE); therefore, members *only* need to exhibit spatiotemporal contiguity in an existing theme to be thematic associates (Kurtz & Gentner, 2001; Mirman et al., 2017). A thematic category gains its coherence from the participation of members in a situation, event or action. Again, thematic relations can be complementary in their roles, but it is sufficient if they are only extrinsically related (Lin & Murphy, 2001). When complementary roles do exist, they exhibit a large degree of variation: they can be any type of *productive* (e.g., SNOW and AVALANCHE), *temporal* (e.g., SNOW and WINTER), *spatial* (e.g., SNOW and MOUNTAIN), *causal* (e.g., SNOW and SHOVEL), *possessive* (e.g., SNOW and TREE; cf. Jones & Love, 2007) or *functional* (e.g., SNOW and SKI) association between things (Estes et al., 2011).[1]

Traditional theoretical accounts of similarity handle the taxonomic element in a variety of ways, but they operate without any consideration of thematic association as a contributor to psychological similarity. For example, similarity could be represented as the distance between entities in a multi-dimensional feature space (Shepard, 1957; 1987)—entities are encoded as points in the feature space and the proximity of the points represents similarity. Tversky's Contrast Model is a set-theoretic approach where similarity is defined as a calculation of feature overlap between entities (Tversky, 1977; Tversky & Gati, 1978). Gentner's Structure Mapping Theory holds that similarity is derived from structural

[1] This list is not exhaustive, and these classifications are not mutually exclusive (i.e., SNOW and TREE can be construed as having the possessive association, the spatial association, or both).

alignment—where the relational structure of entities is aligned via the comparison process and the presence of matching sets of structural correspondences is the primary determinant of similarity (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983; Gentner & A. B. Markman, 1995). Theoretical accounts of similarity relying on the Bayesian perspective have also been proposed (Anderson, 1991; Tenenbaum & Griffiths, 2001) where the probability of the features of an object given a category label is used to determine degree of membership in a category (and thus similarity). We know of no proposals as to how such theories could account for the thematic association effect since they locate similarity in the match between intrinsic content, not in the co-occurrence of concepts.

Dual process models—in which similarity is derived from a combination of taxonomic similarity and thematic association—have been proposed to handle this apparent failing (Chen et al., 2013; Estes, 2003; Estes et al., 2011; Wisniewski & Bassok, 1999). The idea is that concepts that share little or no taxonomic similarity gain perceived similarity due to thematic integration. Thus, the increase in perceived similarity of the concepts is due to co-occurrence in a theme, i.e., SHIP and SAIL increase in perceived similarity because they co-occur with OCEAN, PLANK, SAILOR, etc. (Golonka & Estes, 2009). Sloman also argues (from a different perspective) for a combination of taxonomic similarity and thematic association as components of one system (Sloman, 1996; 2014).

BUTTER

JELLY **KNIFE**

Figure 1. Example of the canonical forced-choice triad task for similarity judgment. The goal of the task is to choose the alternative (bottom row) that is most similar to the standard (top row). BUTTER and JELLY are taxonomic category members and BUTTER and KNIFE are thematic associates.

Thematic Integration versus Thematic Intrusion

What supports the proposal to incorporate thematic association into theories of similarity? The existing evidence suggest that similarity judgments are influenced by information not related to taxonomic similarity. This behavior is particularly salient in empirical investigations that pit taxonomic category members against thematic associates, where the task is a match-to-sample, forced choice triad (see Figure 1) with a match featuring concepts that share features and relational structure (taxonomic category members) and a match of concepts that co-occur in a theme (thematic associates). The frequently reported result is that people choose thematic matches significantly more often than taxonomic matches. It is hypothesized that this behavior is due to thematic integration. There is evidence of thematic organization in sorting behavior as well, where children (E. M. Markman et al., 1981) and adults (Lawson, Chang, & Wills, 2017; Murphy, 2001) often favor theme-based categories in free sorting tasks. Thematic integration also appears to occur for action phrases (Rabinowitz & Mandler, 1983) and complete sentences. Apparent theme-based similarity effects in judgments of complete sentences are what initially lead to the proposal of a thematic integration-based source of perceived similarity (Bassok & Medin, 1997). Here, people were presented with sentences that exhibited varying levels of featural and relational matches (Table 1).

Table 1.
Stimulus example from Bassok & Medin, 1997

	Sentence Stimuli	Similarity to Standard
1.	The carpenter fixed the chair.	Standard
2.	The electrician fixed the radio.	Relation + Object Dependence
3.	The plumber fixed the radio.	Relation

4.	The carpenter fixed the radio.	Relation + Single Object Match
5.	The carpenter sat in the chair.	Double Object Match

Structure Mapping Theory would predict that (2) should be rated as most similar to the standard (Gentner, 1983; A. B. Markman & Gentner, 1995). This overall pattern was found, but it was also noted that (5)—the example with no relational similarity but two matching objects—was also viewed as similar. Examination of response justifications uncovered that when people viewed (5) as similar to the standard, this rating was often justified by integrating the sentences, e.g., (1) and (5) are similar because the carpenter fixed the chair and then sat down to test his repair (Bassok & Medin, 1997). Follow-up work with the three-concept triad task found a similar pattern: similarity judgments, thematic relatedness judgments, and commonality and difference listing tasks were affected by whether the targets were taxonomic or thematic category members (Wisniewski & Bassok, 1999). Wisniewski and Bassok argue that the process recruited for these tasks depends on task constraints and the similarity of the objects themselves: when objects have taxonomic similarity, features and relations are compared and a process (e.g., structural alignment) is used to produce a similarity judgment; conversely, when objects have low taxonomic similarity, the integration process is invoked. When targets are integrated, the perceived similarity of the objects increases. When comparison processes are recruited, the alignment of the targets makes their differences more salient and perceived similarity decreases. In other words, it is easy to spot the differences between similar things because they are easy to compare; different things are difficult to compare so their differences are not as easy to identify (Gentner & Gunn, 2001). The integration effect is perhaps most salient in cases where concepts that are present in a common theme (e.g., KEYBOARD and MOUSE) are chosen over more similar matches in forced-choice triad tasks (e.g., responding MOUSE to "What is most similar to KEYBOARD, TYPEWRITER or MOUSE?").

How prevalent is thematic integration-based responding in similarity judgment tasks? Smiley and Brown (1979) found that the majority of their sample exhibited a consistent responding bias (taxonomic or thematic). The youngest (preschool and first grade) and oldest (66–85 years) age cohorts produced a reliable thematic match bias in the forced-choice triad task, but fifth graders and college-aged adults did not. All age groups (3–15 years) produced a thematic response bias in a cross-sectional investigation of the triad paradigm where the stimuli were pictorial and response justifications were solicited (Greenfield & Scott, 1986). Skwarchuk and Clark (1996) found thematic response biases across three experiments and *11 conditions* where only one condition across the series produced a taxonomic response preference (see Table 2 for a survey of the task instructions). Lin and Murphy (2001) investigated ten variations of the similarity judgment task, finding thematic biases with college-aged samples in a close replication of Smiley and Brown (1979) and other triad-style tasks. The results uncovered thematic responding on 73% of trials in the direct replication of Smiley and Brown (Experiment 3), 70% thematic in a similar paradigm except with the addition of response justification (Experiment 5), 56% thematic in a conceptual replication replacing the word stimuli with pictures (Experiment 4), and a similar pattern of results in several other conditions (Lin & Murphy, 2001). Simmons and Estes (2008) also report thematic response biases in the standard triad task with similarity-based instructions. These results suggest that the presence of thematic associates should have a strong effect on similarity judgments that is easy to detect in behavioral paradigms like the forced-choice triad task, the pairwise similarity rating task, and others.

The literature is not without reports of taxonomic response preferences. As mentioned, the fifth grade students and college-aged adults sampled in Smiley and Brown (1979) produced a majority of taxonomic responses in the triad task—though the results of Greenfield and Scott (1986), Skwarchuk and Clark (1996) and Lin and Murphy (2001) report the opposite pattern with similar samples. The Lin and Murphy (2001) report also features examples of responding biased toward taxonomic matches, notably when people were asked to list similarities (Experiment 7) and differences (Experiment 8) between concepts before completing the triad task (note that justifying similarity judgments has not always produced majority taxonomic responding across the work surveyed here, e.g., Greenfield & Scott, 1986).

Considering the conflicting evidence of taxonomic and thematic responding biases, it might be better to ask why responding preferences are so flexible. Work by E. M. Markman and colleagues provides an example of *how* fluid responding preferences can be—simply providing a bag to children during a sorting activity (i.e., removing the spatial arrangement component of the task) increased the frequency of taxonomic responding (E.M. Markman, Cox & Machida, 1981). Explicit direction with examples also lowers the frequency of thematic responses. Gentner and Brem (1999) found that people who initially had a bias for thematic responding produced a majority of taxonomic matches in the triad task after a moderate amount of training and guidance. Hendrickson, Navarro and Donkin (2015) report a similar pattern of results where people directed to choose taxonomic matches and be as accurate as possible produced a majority of taxonomic matches in the triad task.

Despite the mixed results, the generally accepted view of responding preferences in the classic task (the 2AFC triad task with instructions to choose the most similar match) is that people consistently respond in one way or the other, but often select a majority of thematic responses. This responding pattern has been attributed to three factors: task constraints, stimulus properties, and individual biases for taxonomic or thematic information (Kalenine & Bonthoux, 2008; Mirman & Graziano, 2012; Murphy, 2001; Simmons & Estes, 2008; Wisniewski & Bassok, 1999). The central questions guiding the present work are how and why thematic association impacts similarity in the simplest of paradigms (the forced-choice triad)—and what should be concluded from the available evidence about psychological similarity. If thematic association is not an integral component of similarity, what else can explain the prevalence of the observed behavior?

The Confusability Account

The research above suggests that thematic associates affect similarity judgments in similarity rating and forced-choice response tasks under a variety of instructions. We question the view that this effect is grounds for including thematic association as a contributing factor in theoretical models of similarity. An alternative proposal—the confusability account—is that this behavior is the result of confusion, where thematic association intrudes on the process(es) used to derive similarity judgments (Gentner & Brem, 1999). People may get a clear sense of relatedness from the thematic associates and fail to carefully monitor themselves from treating such a match as a basis for responding to the task. Relatedly, they may simply be looking for any sense of coherence between the concepts. When they find a match that clicks, they treat it as a basis to respond to the task. One theoretical view is that attention can be flexibly focused on different dimensions or semantic relations based on their consistency with task goals (Nguyen & Murphy, 2003); this flexibility can sometimes produce confusion regarding what type of match is called for in a situation. We see value in the proposal that different tasks and objects of comparison (stimuli) elicit different processes (Wisniewski & Bassok, 1999)—without making the leap to a dual process view in which these processes must both be components of the similarity judgment system (Chen et al., 2013; Estes, 2003; Estes et al., 2011; Simmons & Estes, 2008).

The distinction between the confusability and dual-process integration accounts has been investigated by putting taxonomic and thematic relations in direct competition under time pressure. Gentner and Brem (1999) provided a definition for exactly what similarity is intended to mean to participants and then presented forced-choice triads where the task was to choose the option most similar to a standard; the options were a taxonomic match and either a thematic match or an unrelated distractor. Under a 1000 ms deadline, people produced more errors in selecting the taxonomic match. They had less trouble, however, when the distractor was unrelated to the standard, and when the deadline was increased to 2000 ms. Why does time pressure increase the thematic integration effect on similarity judgments? It is not clear how a dual-process integration account would explain an increase in the weighting of thematic information for similarity judgments at shorter timescales. Under the confusability account, however, the explanation is clear—people have not had time to resolve the competing semantic relationships and the presence of thematic association interferes with the processing of taxonomic similarity. Interestingly, the intrusion effect does not seem to work both ways. Thematic distractors appear to facilitate superordinate taxonomic categorization decisions (Lin & Murphy, 2001, Experiment 10). Even for the simpler task of object identification, co-presentation of a thematic associate facilitates picture naming while co-presentation of a taxonomic category member inhibits picture naming (de Zubicaray, Hansen, & McMahon, 2013).

The present experiments were intended to comprehensively investigate the thematic association effect: Do people truly think that concepts with a situation in common are more similar than concepts with shared features and alignable relations? We are especially concerned with the task itself. Does the standard procedure impose information processing demands that bias responding? Do participants clearly understand what they are supposed to do?

Experiment 1

We set out to create a modified version of the classic experimental task to pit taxonomic against thematic options to determine whether characteristics of the traditional set-up create an unanticipated bias favoring thematic responding. Our first concern was the two-alternative forced-choice (2AFC) triad task (see Figure 1). Participants faced with the traditional task begin by activating the meaning of the target concept and then selecting one of the two options. If the thematic option is primed to a greater extent due to the high degree of association, this could facilitate processing, direct attention, mediate concept construal, or otherwise create a bias toward thematic responding. These information processing factors would be independent of the similarity judgement in principle but not in practice—they are not what we are seeking to measure but they are showing through. Therefore, we sought to eliminate the target-first nature of the task. In addition, the triad task assesses something other than what we intend for another reason. Instead of telling us which form of relationship (taxonomic or thematic) is more important to similarity, we are finding out which is considered stronger when given a direct choice between the two. This task reflects an evaluation of each option with respect to the other. We are not looking to assess such preference judgments, but to pit competing notions of relatedness against one another. In some cases, this may amount to the same thing, but in others it may not. It would be preferable to assess the relative impact of the two kinds of relatedness against a background of unrelated options, as opposed to the concept pairs being pitted directly against one another. Finally, the presentation of the two match options could implicitly suggest that both options are equally valid answers to the similarity question, which could in turn encourage people to answer an “easier” question rather than the one they are asked (Kahneman & Frederick, 2002; Shah & Oppenheimer, 2009). Past work may have promoted this construal of the task goal with instructions that explicitly say that there is no right or wrong answer (e.g., Lin & Murphy, 2001). Therefore, we sought to eliminate the direct choice between thematic-versus-taxonomic in the task. To accomplish these goals, we turned each traditional triad (target item and two choices) into a six-item ring with no focal target item and three unrelated distractors. The task is to identify the most similar pair—this accomplishes our scientific goal while circumventing the potential biases. We note that semantic judgment tasks that require two choices (as opposed to 2AFC) are not novel. In one case, the choose-two format seems to have decreased taxonomic responding (Lin & Murphy, 2001, Experiments 2 & 6). These experiments presented a concept in a prioritized position (i.e., at the top of the triad with the number 1), so the combination of added distractors and the removal of a prioritized standard may produce a different result.

Our second major focus was how participants interpreted the task and how the specific wording and delivery of instructions could impact responding—could this be the reason for majority thematic responding? Instructions exhibit considerable variability across investigations (see Table 2 for a sample of previous task instructions). We are not the first to notice that instructions have a crucial effect on response preferences (Lin & Murphy, 2001; Simmons & Estes, 2008; Skwarchuk & Clark, 1996; see Mirman et al., 2017, for review), but the present work is novel in that all instructional variations are designed to head off possible confusion about the meaning of similarity (and the task goal). To that end, we developed instructions that address one key question: Is it possible that people are misinterpreting the goal of the task when they are asked to *choose the most similar option* to the standard? Could they simply be misunderstanding what they are being asked to do when asked to choose *similar* options?

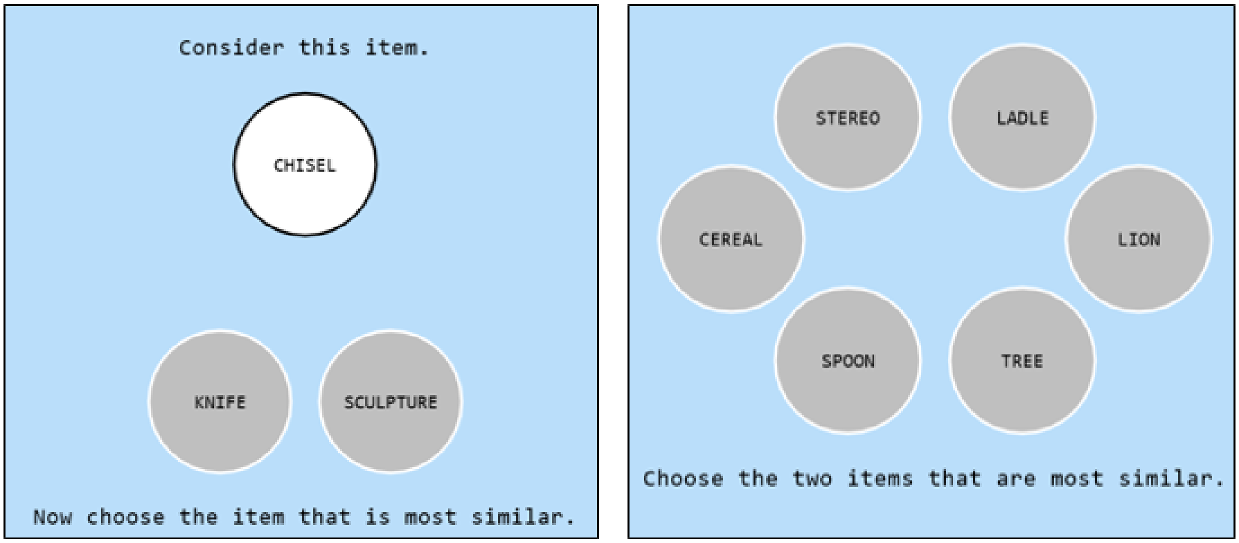
Table 2. Variation in Task Instructions

Task Instructions	Article
Choose the option that goes best with the base.	Smiley & Brown, 1979
Choose the option that is most similar to [STANDARD].	Gentner & Brem, 1999; Simmons & Estes, 2008
Pick the response option that is most like [STANDARD].	Simmons & Estes, 2008
Choose the alternative that is most related.	Skwarchuk & Clark, 1996
Choose an Alternative that is most similar and goes together	Skwarchuk & Clark, 1996
Choose the two options that can be called by the same name.	Lin & Murphy, 2001
Choose the option that goes best with [STANDARD] to form a category.	Lin & Murphy, 2001
Choose two items that best form a category.	Lin & Murphy, 2001
Find another the same as this.	Davidoff & Robinson, 2004
This is a [CONCEPT], find another one.	Davidoff & Robinson, 2004; Gentner & Brem, 1999

If the thematic response bias is due to a misinterpretation of the use of “similar” in the instructions, we should find that instructions that attempt to clarify the meaning produce more taxonomic responses. To test this hypothesis, we examined three sets of instructions—Similar, Alike and Alien. The Similar instructions (provided below) are included to establish the baseline responding pattern under the most straightforward set of instructions. The Alike instructions are subtly different; “similar” is replaced with “alike” to test the possibility that a direct misunderstanding of the term “similar” is to blame for the thematic response bias. These instructions are close to instructions used previously (Simmons & Estes, 2008) where “like” was used in place of similar, i.e., “Pick the response option that is most like the [Standard]”. We note that “like” can be interpreted quite broadly, e.g., “COW is most like MILK because they are found on farms”, and thus, “alike” should be a closer approximation to the meaning of similar. The Alien condition represents an entirely fresh approach: the task goal is to select the appropriate concept pairs in order to teach an alien about life on earth and what things are similar to another. This is hypothesized to produce more taxonomic responding than the other instructions because it reduces the impact of assumptions about concept knowledge or pragmatic factors and highlights the functional role of similarity as the basis for organizing world knowledge. It is known that people will adjust their report of basic conceptual knowledge based on what they think they are expected to say, what they deem too obvious to be expected, and/or too mundane to report (Murphy, 2002). The Alien instructions may make people more likely to think about similarity in a fundamental sense and less likely to veer toward idiosyncratic construals.

While it is important that the meaning of similar is understood, it would be too heavy-handed to explicitly identify the difference between taxonomic similarity and thematic relatedness with concrete examples. This has been done; it appears to increase the frequency of taxonomic matches (Gentner & Brem, 1999). Our question of interest is directly related to how the concepts are interpreted as similar; it would be too much to explicitly highlight the difference between taxonomic and thematic semantic relations—if this was done sufficiently clearly, then subject responses would only be parroting back what was asked for. This issue is handled in this series of experiments by omitting any concrete examples or definitions of the semantic relationships.

These considerations motivated an experimental design with three conditions featuring the pair selection task from a ring of six options under three different sets of instructions. A fourth condition addressed an additional related issue about the impact of instructions and interpretation of the task. Rather than repeating the specific instruction on screen for each trial, we explored what happens if participants are given initial baseline instructions asking for the most similar pair, but the instruction is not reiterated with the presentation of each trial. If the instruction to consider "similarity" is not reinforced, do people drift toward a default mode of responding (perhaps having little to do with similarity) that the instructions are essentially guiding individuals to monitor against?



Consider this item.

Now choose the item that is most similar.

Choose the two items that are most similar.

Figure 2. Triad versus Ring Tasks. Left: Classic 2AFC triad task with similarity instructions. Right: Depiction of the novel task where the goal is to choose the two concepts that are most similar. No concepts are prioritized and a set of three distractor concepts are presented along with intended taxonomic and thematic matches (standard, taxonomic target, thematic target).

Method

Participants and Materials

Undergraduate students from Binghamton University were recruited from the Psychology Department pool and participated for credit toward the completion of a course requirement. Participants (*N* = 238; Native English, *n* = 204) were randomly assigned to condition. The experiment was administered with Psychopy, a Python-based experiment presentation software package (Pierce, 2007). The stimuli consisted of semantically-related concept triads adopted from previous experiments (Gentner & Brem, 1999; Lin & Murphy, 2001; Hendrickson, Navarro & Donkin, 2015; Wisniewski & Bassok, 1999) and novel triads developed for this project. In addition to the classic three-item triads, three semantically-unrelated concepts were added to each concept triad (see Experiment 3 for norming data). By removing the presence of a target concept, the new method requires that the taxonomic and thematic response options are not semantically related; this consideration guided the exclusion of several concept sets from previous investigations (e.g., CHAIR, BED, CARPENTER). This process resulted in 59 concept sets presented in a random order (concept sets are provided in Appendix A).

Each trial presented the six concepts of a set (a standard, one taxonomically-related option, one thematically-related option, and three unrelated options) organized around the center of the screen as clickable buttons. The task interface was identical for each of the four conditions (right panel, Figure 2). The preliminary instructions and the on-screen trial instructions varied by condition. In the No-Reminder condition, the usual reminder about the task instructions (e.g., “Choose the two items that are most similar”) was not presented in the experiment interface, thus, in this case participants read the initial instructions but were not reminded about the goal of the task for the remainder of the experiment. The experiment included four between-subjects conditions: 3 conditions with distinct instructions—Similar, Alike, and Alien (see below)—and No-Reminder that presented the Similar instructions initially but without continual reiteration.

Procedure

The experiment was conducted as a part of an experimental session that included other unrelated studies. It started with the presentation of on-screen instructions that varied by condition. The initial instructions for the Similar and No-Reminder conditions are as follows:

> Hello! In this study, you are going to see a series of different sets of items (words). For each set, your goal is to find the two items in the set that are most similar to one another. When you've chosen the two items that are most similar, use the mouse to select the items and then press continue to confirm your selection.

The Alike instructions were matched but use the term “alike” and never mention similarity:

> Hello! In this study, you are going to see a series of different sets of items (words). For each set, your goal is to find two items in the set that are ***most alike***. When you've chosen the two items that are **most alike**,

> use the mouse to select the items and then press continue to confirm your selection.

The Alien instructions were as follows:

> Hello! In this study, you are trying to teach an alien from outer space about life on earth. Specifically, you need to teach the alien about things that we have on earth that are similar to each other. We will be showing you a series of different sets of items (words). **Can you demonstrate to your alien friend which pair of items in each set are things that are similar to one another?** When you've chosen the two items that are most similar, use the mouse to select the items and then press continue to confirm your selection.

After the instructions, participants responded to 59 trials with randomized presentation order and item positioning. Each trial started with the presentation of a fixation cross followed by a ring arrangement of six items. Participants responded by selecting a pair of items in response to task instructions (e.g., “Choose the two most similar options”) and then confirmed their selection by clicking a “CONFIRM” button. Options could be selected or deselected at will until the final choices were confirmed. All actions, final responses, and timing data were recorded.

Results

The two goals of this study were: (1) to test an alternate version of the tradition experimental paradigm modified to reduce potential bias by removing the specific evaluation of options relative to a target and removing the direct choice between thematic and taxonomic options by including distractors; and (2) to test alternate forms of instruction to investigate the impact of what participants are told to do (and whether they are continually reminded). The frequency of each type of response is presented in

Figure 3 with individual data points shown. Data and analyses for the entire project are available in the supplemental materials archived on the Open Science Framework.[2]

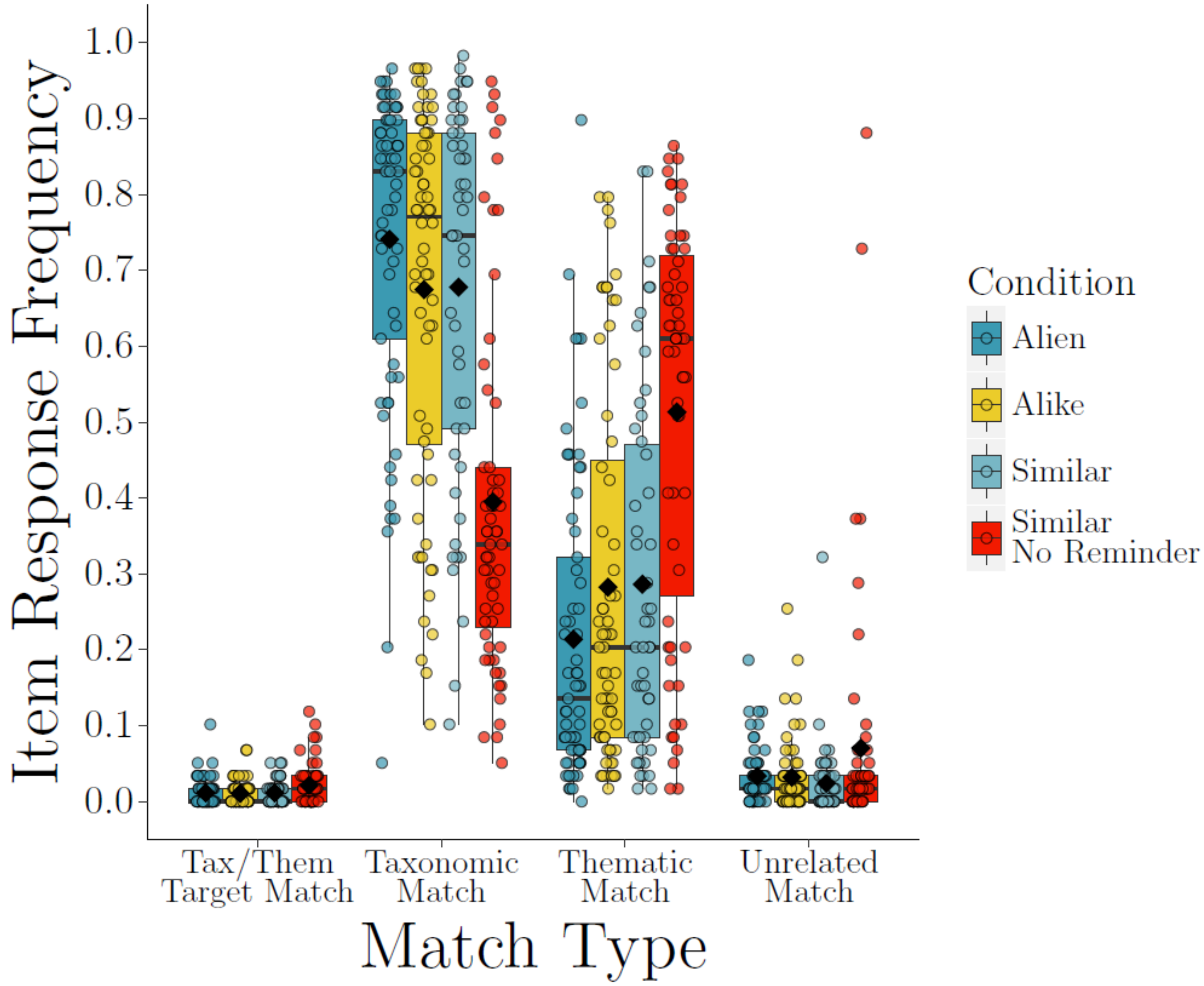


Figure 3. Frequency of matches by match type and condition for each participant from Experiment 1. Participants are represented by one point (positioned by condition) for each match type. Tax/Them target matches are the rare cases in which the participant selected the thematic and taxonomic match options, but not the standard from the embedded concept triad. Tukey's box plots show the median and interquartile range and diamonds represent the mean frequency of each type of match by condition.

[2] https://osf.io/6z9nf/

The first major pattern present in these data is that taxonomic matches were made more frequently than any other type of match using the pair selection task with six-item rings. An individual preference for taxonomic matches (based on all taxonomic and thematic responses) is statistically significant at a response frequency rate of 64.4% relative to 50% chance rate (37-38 consistent matches out of 59 trials; $.036 < p <= .067$). The strong majority of taxonomic responding contrasts sharply with the existing literature.

At the trial level, three of the four conditions produced majority taxonomic responding; only the No-Reminder Condition showed the opposite pattern of majority thematic responding (see Table 3). A two-stage binomial test procedure was used to determine the number of participants producing a significant majority of taxonomic responses compared to what would be expected by chance. First, one-sided binomial tests were used to determine if the participant made taxonomic matches more than chance. The number of trials with taxonomic matches was the DV and the null hypotheses was chance responding or more thematic responding. Only trials where the intended taxonomic match was chosen were counted as taxonomic trials. In contrast, trials were classified as thematic when the intended thematic match was made or the thematic and taxonomic targets were chosen—a more conservative classification approach following from the idea that taxonomic category members can often share thematic associates (e.g., BEER and JUICE are taxonomic category members that could both be construed as thematically associated to PARTY). Trials where unrelated distractors were chosen were excluded from the analysis so that the test would be a direct comparison of taxonomic and thematic choices (chance = .5; this exclusion had no effect on the analysis outcome). After participant response preference was calculated, these classifications were used as the DV in two-tailed binomial tests to determine if there were more (or fewer) people consistently responding taxonomically than what would be expected by chance. The outcome of this analysis was that every condition produced a taxonomic response preference except for the No-Reminder condition. The Alike and Alien conditions had reliably more taxonomic responders than would be predicted by chance; the Similar condition had the same pattern, but the result was only marginally significant. The No-Reminder condition had reliably more thematic responders than would be predicted by chance.

Table 3
Taxonomic Responding Rates in Experiment 1

Condition	N	Mean Proportion Taxonomic Responses	Taxonomic Responding Exact Binomial Test p	95% Binomial Confidence Interval	
				Lower	Upper
Alien	$n = 65$	.74	$p < .001$	.63	.85
Alike	$n = 63$	.68	$p < .001$	.59	.82
Similar	$n = 50$	.68	$p = .065$	.49	.77
No Reminder	$n = 59$	.40	$p < .001$	.12	.35

To compare across conditions, generalized linear mixed-effects regression models (GLMER; Bates, Maechler, Bolker, & Walker, 2014) were built that predict taxonomic responding under different instructional manipulations. We start by describing the maximal model, which includes condition and trial as fixed effects and participant, trial and concept set (item) as random effects. As explored below, responding preferences had a considerable amount variability across the time course of the experiment. Thus, random (by-subject) intercepts and slopes were included to account for the effect of this change across the experiment (see supplementary materials for data and code). The model uncovered a pattern where all conditions with reiterated task instructions produced more taxonomic responding than No-Reminder Condition (see Figure 4), Alien: Beta Estimate = -2.161, $SE = 0.28$, Wald $Z = 7.620$, $p < .001$; Alike: Beta Estimate = 1.631, $SE = 0.29$, Wald $Z = 5.647$, $p < .001$; Similar: Beta Estimate = 1.633, $SE = 0.31$, Wald $Z = 5.284$, $p < .001$. When the No-Reminder condition is dropped from the model, the results show that the Alien condition produced more taxonomic responding than the Alike condition (Beta Estimate = 0.554, $SE = 0.28$, Wald $Z = 1.963$, $p = .0496$). The same pattern held for Alien relative to the Similar condition (Beta Estimate = 0.563, $SE = 0.30$, Wald $Z = 1.859$, $p = .063$) but only reached marginal significance. The differences between the conditions with reiterated instructions are marginally significant when the No-Reminder condition is included in the model (Alien vs. Alike, $p = .058$; Alien vs. Similar, $p = .078$). These results provide tentative support for the hypothesis that instructions designed to better measure psychological similarity (as opposed to just asking about “similarity”) attenuated the thematic association effect on similarity, though the marginal (and near-marginal) differences between conditions make it difficult to make strong conclusions about the generalizability of this effect. In sum, the Alien condition produced more taxonomic responding than the other conditions, but this difference was only marginally significant for the Alien–Alike comparison under the most conservative analysis.

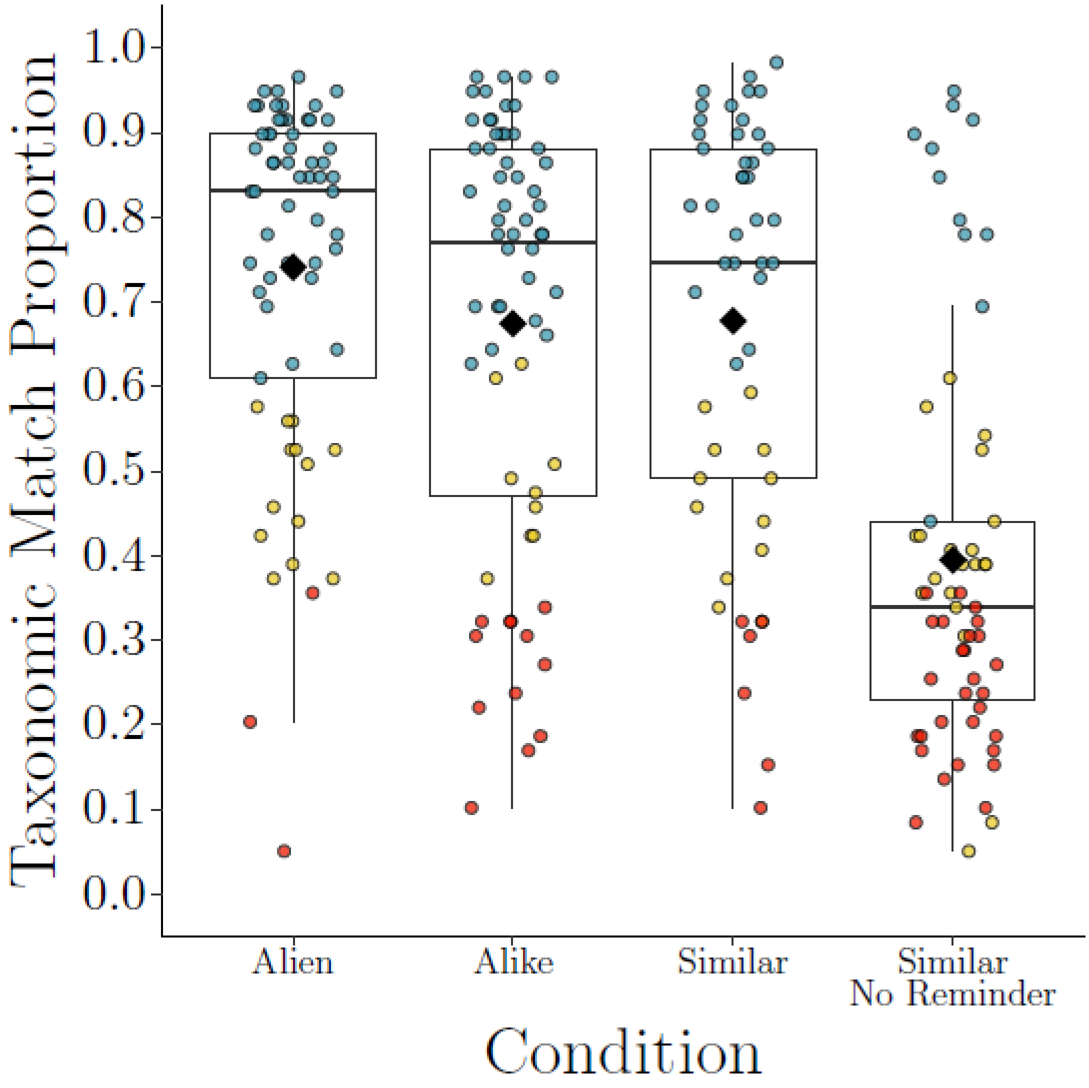


Figure 4. Proportion of taxonomic matches by condition for Experiment 1. Participants are represented as points, diamonds present the condition means, and Tukey's box plots present the median and interquartile range of mean taxonomic responding. Points are colored based on response preference classification.

Time course analysis

Trial was a significant fixed-effect predictor of taxonomic responding in both models (even when accounting for the variance of individual participant slopes and intercepts). This means that the frequency of taxonomic responding increased across the time course of the experimental session (Beta Estimate = .011, *SE* = 0.003, Wald *Z* = 4.187, *p* < .001). Analyzing the conditions in isolation produced a different pattern where trial was a reliable predictor of taxonomic responding for all conditions except the No-Reminder condition (*p* = .94) and the Alien condition (*p* = .34). (Note: the conditions that produced the most and least taxonomic responding in this exploratory analysis were the conditions that did not have reliable increases in taxonomic matching across trials.) A post-hoc explanation for the lack of a trial effect for the Alien condition could be that the instructional manipulation worked as expected. The time course curves (see Figure 5) are an unexpected and possibly quite revealing finding. The Alike and Similar groups start out showing no sign of a preference between thematic and taxonomic responding, but across the course of the task, they catch up to the strong taxonomic preference of the Alien group. The Similar/No-Reminder group shows a non-significant thematic preference at the outset and maintains this pattern across the course of the task. When the participants are not continually reminded of what they are supposed to be doing, they show no signs of spontaneously (since there is no feedback) refining their task interpretation. Notably, the Alien group showed reliably more taxonomic responding in the first 10 trials of the experiment (*p*s < .001). The pattern of taxonomic responding increasing for the least well-specified instructions conditions as the experiment progressed—only if the instruction is reiterated—raises a number of important implications (addressed below)

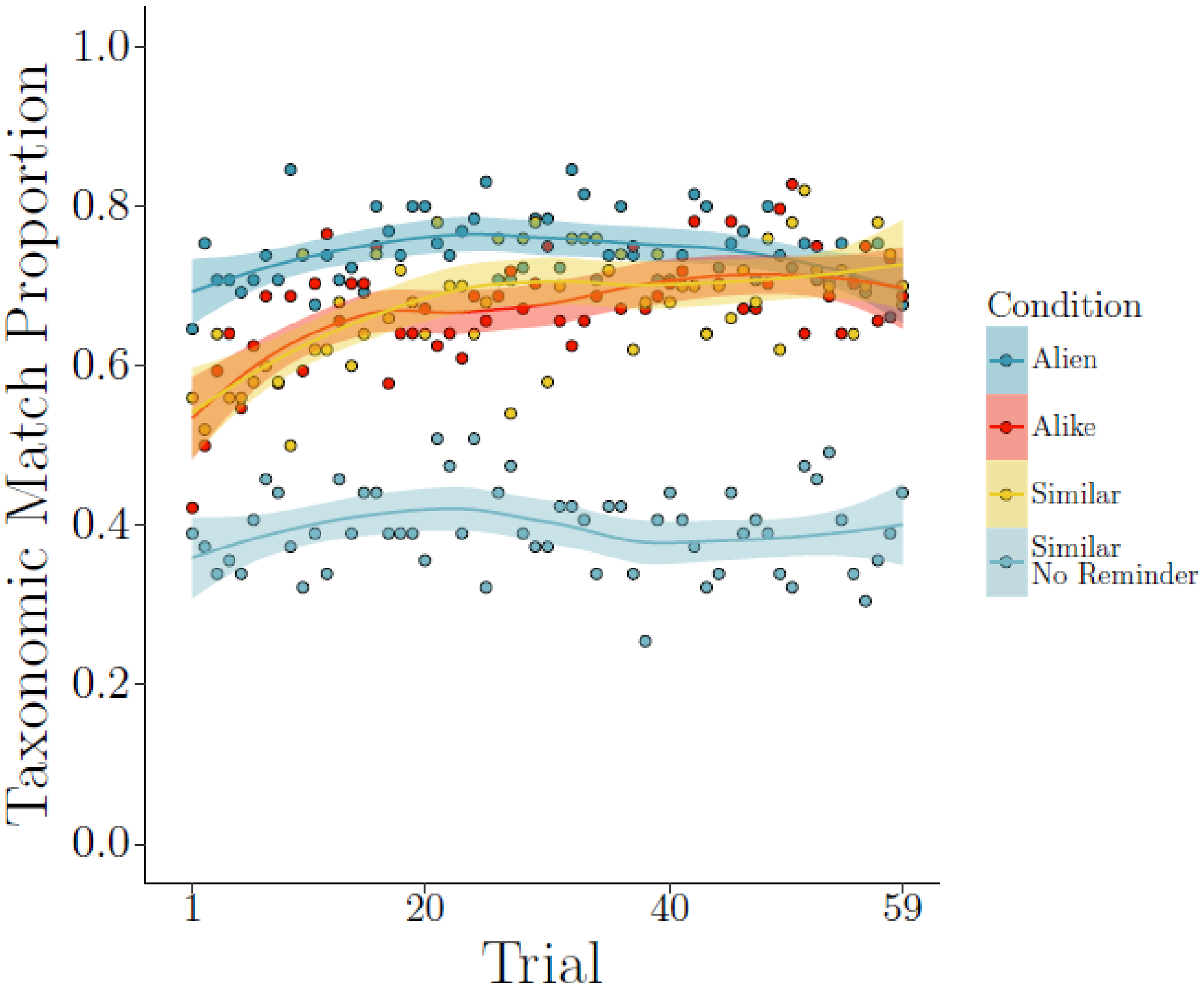


Figure 5. Taxonomic responding frequency across trials in Experiment 1. Points represent mean taxonomic responding by trial for each condition.

Response latency analysis

Previous research suggests that thematic category members are processed faster than taxonomic category members (Estes et al., 2011; Gentner & Brem, 1999; Mirman & Graziano, 2012). One issue that has been raised about the methodology of deadline-based experimental paradigms, however, is that imposing a deadline (e.g., Gentner & Brem, 1999) may fail to capture a comprehensive account of the processing time-course of these semantic relations (Hendrickson, Navarro & Donkin, 2015). We recorded trial response time in this free-choice, speed-irrelevant task (i.e., no directive to focus on speeded response was provided). Note that the cell count is quite different for each of the four possible matches (i.e., Taxonomic Target + Standard, Thematic Target + Standard, Taxonomic Target + Thematic Target, Match including Unrelated Distractor) and these frequency differences should be considered when interpreting the results (see Figure 3). Match frequency is presented in Table 4.

Table 4
Frequency of Matches and RT by Match Type in Experiment 1

Match Type	Frequency (Count)	Mean of Participant Median RT
Standard and Taxonomic Target	62.4% (8,765)	6.75 seconds
Standard and Thematic Target	32.2% (4,519)	8.05 seconds
Taxonomic Target and Thematic Target	1.4% (192)	11.25 seconds
Match including an Unrelated Distractor	4% (566)	12.54 seconds

An LMER model (featuring the maximal random effects structure: participant nested within condition) was built to predict median response time (in seconds) with match type included as the sole fixed effect. The results show that taxonomic matches were completed faster than thematic matches (Beta Estimate = -1.335, *SE* = 0.42, *t* = -3.198, *p* = .002); and thematic matches were reliably faster than matches with unrelated distractors (Beta Estimate = 4.379, *SE* = 0.48, *t* = 9.136, *p* < .001) and matches with the taxonomic and thematic targets (Beta Estimate = 3.298, *SE* = 0.54, *t* = 6.091, *p* < .001). These effects were robust to the removal of outliers (+/- 2.5 SD). By condition, we find this general trend of faster taxonomic trials for every group *except* the No-Reminder group. This exploratory analysis seems to suggest that—unlike previous work with the 2AFC triad task—concepts that share taxonomic category membership "pop out" when participants receive the pair selection task with distractors (see Figure 6). We will return to this possibility in Experiment 2 which provides a direct comparison of the two tasks.

Additionally, we find that response time patterns appear to follow taxonomic response frequency. This is a surprising correspondence that (to our knowledge) has not been explored in this domain. It seems that the response time difference between taxonomic and thematic responding tracks closely to the frequency of taxonomic matches (Alien $_{\text{Beta Est.}}$ = -2.42, *p* = .005; Alike $_{\text{Beta Est.}}$ = -1.39, *p* = .036; Similar $_{\text{Beta}}$

$_{\text{Est.}}$ = -1.78, *p* = .14; No Reminder $_{\text{Beta Est.}}$ = 0.26, *p* = .67). These results may provide a new framing for response time effects in this research area. It is possible that a heretofore unconsidered contributor to latency in processing these semantic relations is the interpretation of the task or the ambiguity of the task goal. Response time might be an effective stand-in for other measures of task ambiguity in similarity judgement tasks—a strategy that has been used in other contexts for inquiries into the comparison process (Gentner & Kurtz, 2006).

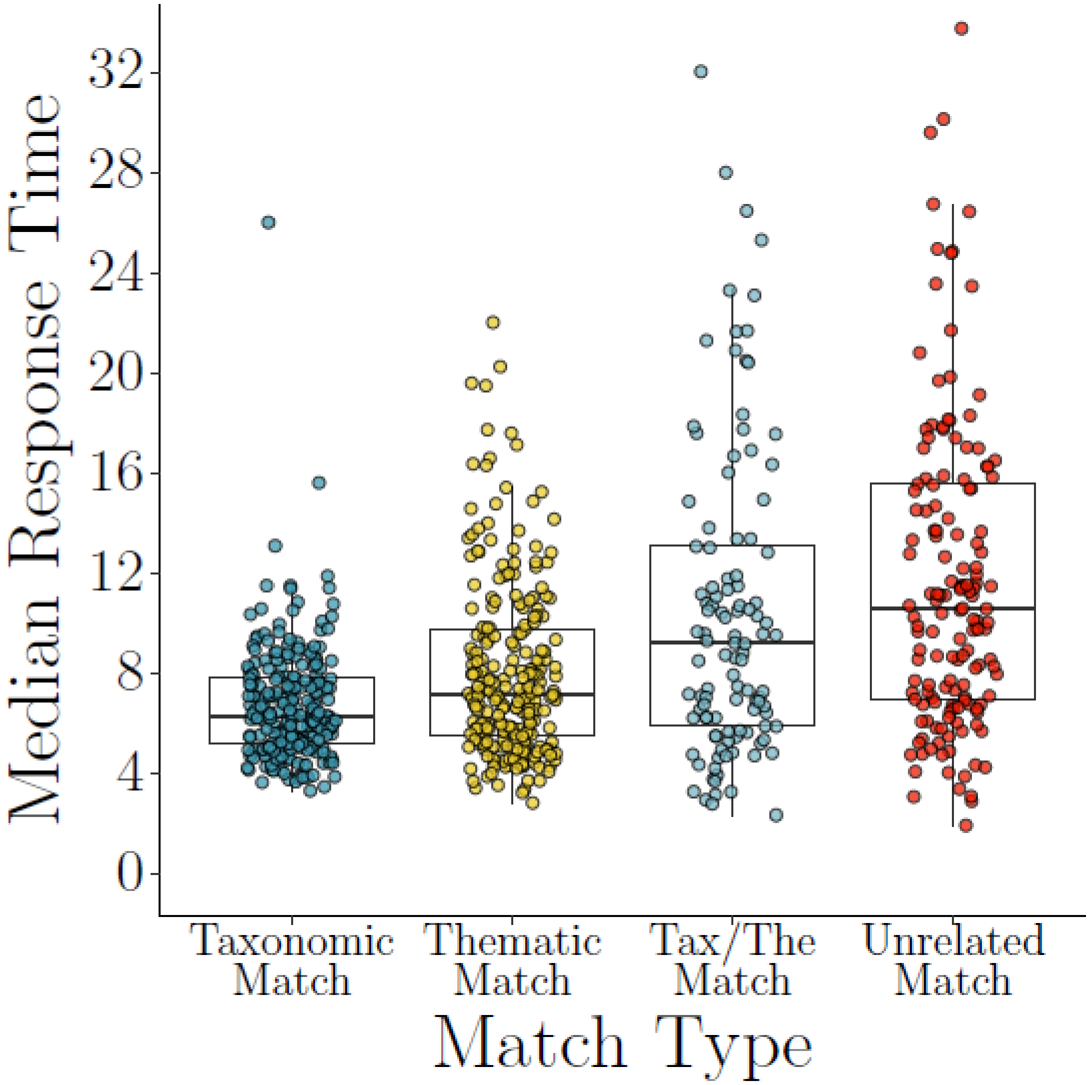


Figure 6. Median response time by match type in Experiment 1. Median response time for each possible match type. Participant median response times are presented by points for each match type where they produced at least one match.

Discussion

Experiment 1 produced several unexpected results. The traditional finding of majority thematic responding was only found when the procedure did not include continual reminders of the instructions (the No-Reminder condition). When the instructions are focused to guide participants away from broad interpretations of what might count as similarity (the Alien group), we see a strong taxonomic preference that starts at the outset of the task and persists across the course of the experiment. When the instructions are more ambiguous but reiterated on each trial, performance gravitates from initial incoherence toward a similar level of taxonomic preference. When the instructions are ambiguous and allowed to recede from memory, participants show a persistent thematic preference that would seem to reflect a poor task understanding and a failure to self-correct. It is difficult to see how a dual process view would explain these observations: why would thematic responding decrease across the task and why would thematic responding only dominate when people are allowed to follow their biases for what to look for rather than being reminded of the actual task goal?

We have produced results that conflict with those presented in Simmons and Estes (2008), where “similar to” and “like” instructions produced reliable thematic response preferences (Experiment 1a and Experiment 1b, Simmons & Estes, 2008). One possible explanation is our use of specialized instructions, but we observed a robust taxonomic preference even in the Similar group. Another possible explanation is our use of the pair selection task with distractors. Our initial experiment did not include a direct comparison of the novel task versus traditional triads, so that is a clear goal of Experiment 2.

One concern about the concept sets at study is the presence of “partonomic” relationships between thematic associates (Tversky & Hemenway, 1984), e.g., SEATBELT and CAR are thematic associates where one concept can be a part of the other. This subclass of thematic associates has competing possible interpretations and appears on 12/59 of the experimental concept sets. One could construe the relationship as having a high degree of match in semantic content; the seatbelts of a car are *literally* the car under this interpretation. As stated above, we take a broad view of thematic associates which includes concepts that share this partonomic relationship; the role of partonomic association has not been consistently controlled for in previous investigations. The presence of partonomic associates in the thematic pairings raises a potential confound in the data that could artificially increase the rate of taxonomic matching. To test this possibility, we used a simple bootstrap-based sampling approach to analyze the frequency with which partonomic and non-partonomic concept sets resulted in taxonomic matches in Experiment 1 (excluding the thematically-biased No Reminder condition). The outcome was that taxonomic matching in samples (200,000 iterations) of partonomic and non-partonomic trials (n = 5000 observations per class per sample) were reliably different less than .01% of the time—far from the threshold of 95% needed to conclude that responding differed between the sub-classes. In other words, we find no evidence that participants responded with taxonomic matches more or less frequently on partonomic vs. non-partonomic trials.

A somewhat unintentional difference from past work is that our experiment included roughly twice as many concept sets. We found an increase in the frequency of taxonomic responding across the course of the experiment (see Figure 5) such that the robust taxonomic preference did not take hold until the second half. The shorter experiments in the literature therefore seem likely to underestimate the prevalence of taxonomic responding since participants are using early trials to formulate and stabilize their interpretation of the task. The same issue applies for previous attempts at characterizing individuals as taxonomic or thematic responders (Lin & Murphy, 2001; Simmons & Estes, 2008; Smiley & Brown, 1979).

The observed increase in taxonomic responding across the course of the experiment, the reversal of the taxonomic response bias in the No-Reminder group, and the overall high frequency of taxonomic responding in the conditions with reiterated instructions collectively support the confusability account. Thematic responding declines with instructions that more sharply delineate the goal or with sufficient task experience to better delineate the goal. Finally, by comparison to previously published data (Greenfield & Scott, 1986; Lin & Murphy, 2001; Simmons & Estes, 2008; Skwarchuk & Clark, 1996), the present results suggest that switching from triads to the new task of pair selection with distractors may have contributed to the observed reversal in overall responding preferences (taxonomic preference over thematic preference). However, given that we did find a thematic preference using this task in the No-Reminder group and the general uncertainty of cross-study comparisons, this is a less compelling conclusion than the others from this experiment. Our next goal is to pursue a replication that includes a systematic investigation of the changes in task format relative to the classic triad paradigm.

Experiment 2

Considering the surprisingly high rates of taxonomic responding observed in Experiment 1 with the traditional similarity instruction, it is important to establish that this pattern is replicable and free of confounds. Additionally, the increased rate in taxonomic responding across the course of the experimental session is a novel finding with considerable implications that requires replication. We developed an experimental design using the same materials with a crossing of two factors: the inclusion of distractors (Triad vs Ring, i.e. six-item arrangements embedding the three critical concepts among three unrelated distractors) and the use of a Standard (one item presented as a base for the options to be related to) versus No-Standard (all items have equal status and the task is pair selection). All of these conditions use the standard similarity task with reiterated instruction on each trial. In addition, a fifth condition was included (see Table 4 for full design) using traditional triads with a standard, but the instructions were altered to ensure that it is possible to elicit strong thematic responding within our methodology. In this condition, participants receive reiterated instructions to choose the item that “goes with” the standard—an instruction that perhaps maximally invites participants to choose associations rather than likenesses. The “goes with” version of instructions featured in previous similarity judgment triad task research—*choose* *the item that goes best with the item above*—has been found to reliably produce thematic responding (Lin & Murphy, 2001; Skwarchuk & Clark, 1996).

Table 4
Design of Experiment 2

Condition	Prioritized Standard	Distractors Present	Instructions
Standard Thematic Triad	YES	NO	GOES WITH
Standard Triad	YES	NO	SIMILAR
No-Standard Triad	NO	NO	SIMILAR
No-Standard Ring	NO	YES	SIMILAR
Standard Ring	YES	YES	SIMILAR

Method

Participants and Materials

Undergraduate students from Binghamton University were recruited from the Psychology Department pool and participated for credit toward the completion of a course requirement. Participants (*N* = 286; Native English, *n* = 251) were randomly assigned to one of five conditions (see Table 4) that make up a 2x2 + 1 between-subjects design. The experimental materials (concept sets) were identical to those of Experiment 1.

Procedure

All of the 2x2 conditions received the same similarity-based instructions (here we show in bold the differences from the Standard Thematic Triad condition):

> Hello! In this study, you are going to see a series of different sets of items (words). For each set, your goal is to find the two items in the set that **are most similar to one another**. When you've found the two items that **are most similar**, use the mouse to select the items and then press continue to confirm your selection.

The Standard Thematic Triad condition was provided with these instructions:

> Hello! In this study, you are going to see a series of different sets of items (words). For each set, your goal is to find the two items in the set

> that **go together best**. When you've found the two items that **go together best**, use the mouse to make your selection and then press continue to confirm.

Each of the 2x2 (similarity) conditions featured a different task interface. The goal of these interface changes was to attempt to pin down what components of the task were responsible for the high rates of taxonomic responding observed in Experiment 1. The interface differences are shown in Table 4 and visual depictions are provided in Appendix B. For the No-Standard Triad condition, concepts were placed in random positions equidistant from the fixation point (screen center) and the other concepts. For the Standard Triad and Standard Thematic Triad conditions, concepts were presented in fixed locations with the standard at the top and the two response options randomly placed in the lower left and right positions. In the No-Standard Ring condition, concepts were randomly placed in positions organized around the screen center. For the Standard Ring condition, concepts were presented at random locations in a trapezoid with the standard presented directly above. Trials were randomly ordered and presented sequentially, each following the presentation of a fixation cross.

Table 5
Experiment 2 Taxonomic Response Preferences

Condition	*N*	Mean Proportion Taxonomic Responses	Exact Binomial Test *p*	95% Binomial Confidence Interval	
				Lower	Upper
Standard Thematic Triad	*n* = 55	.25	*p* < .001	.01	.15
Standard Triad	*n* = 57	.66	*ns*	.44	.71
No-Standard Triad	*n* = 57	.74	*p* < .001	.60	.84
No-Standard Ring	*n* = 55	.61	*ns*	.42	.70
Standard Ring	*n* = 62	.63	*ns*	.43	.69

Results

Recall that the central goal of this experiment was to replicate the pattern of dominant taxonomic responding from Experiment 1 and clarify the distinct effects of the two components of the novel task format: the presence of distractor concepts and the use of a prioritized standard. In addition, we sought to verify that we could generate an overall thematic preference using traditional triads and instructions biased toward thematic responding. This experiment also provides the opportunity to evaluate whether the response latency results from Experiment 1 (where taxonomic matches were completed faster) would be reproduced. The pattern of results for the overall frequency of matches can be seen in Figure 7.

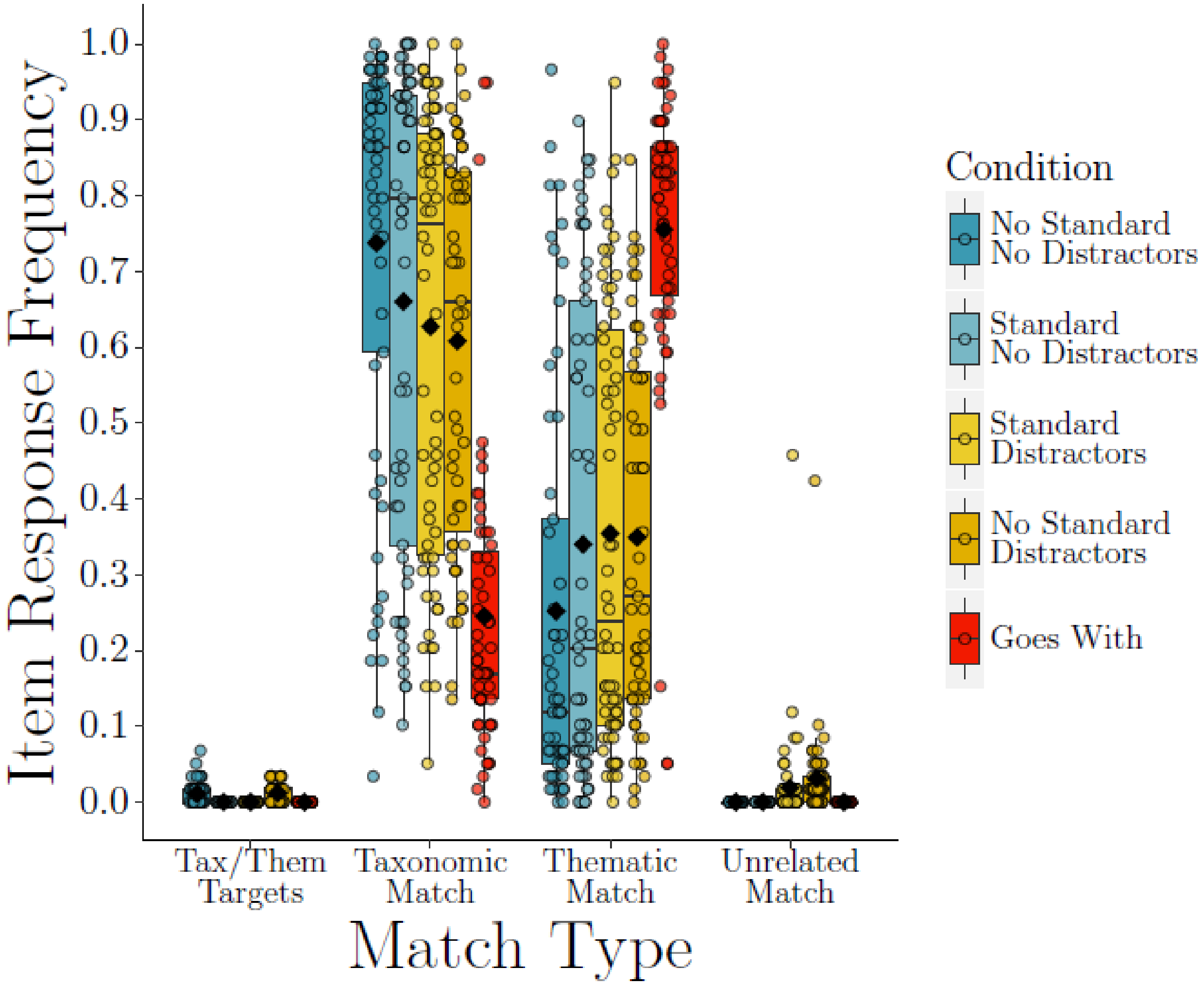


Figure 7. Frequency of matches by match type and condition for each participant from Experiment 2. Participants are represented by one point (positioned by condition) for all response types. Tukey’s box plots show the median and interquartile range and diamonds represent the mean frequency of each type of match by condition.

General Taxonomic Responding Patterns

Taxonomic matches were more frequent than any other type of match; participant response preferences are statistically significant at an item response frequency greater than 64.4%. According to a binomial analysis procedure identical to that of Experiment 1, only the No-Standard Triad condition had enough consistent taxonomic responders to suggest a reliable preference (see Table 5). We note that the opposite approach—examining if fewer thematic responders were present—found reliably fewer thematically-biased responders than would be predicted by chance in every condition with similarity instructions (*p*s < .005). In other words, while only the No-Standard Triad condition had enough participants exhibiting the taxonomic response bias to reliably surpass what would be expected by chance, all similarity conditions had fewer thematic responders than would be expected. In addition, the Standard Thematic Triad condition worked as expected; a reliable majority of participants produced a thematic response bias (*p* < .001). Despite the lack of a clear cut taxonomic responding bias in terms of the number of participants within each similarity-based condition, there was more taxonomic responding overall (Figure 7).

It was not anticipated that the classic triad paradigm would produce a taxonomic response bias (even if only when aggregated across participants). Recall that the central motivation for this work stems from repeated demonstrations of a reliable thematic response preference with exactly this task. It is therefore striking that the classic triad condition produces some of the highest rates of taxonomic responding found in this report. After all, the reason for including the classic paradigm was to compare how the components of our modified task increased taxonomic responding relative to this baseline. We return to consider the implications of this result after addressing the confirmatory analysis goals of Experiment 2.

There are two possible approaches to analyzing the impact of task characteristics on taxonomic responding: a comparison of taxonomic responding between the five distinct conditions and a factor-based approach that examines the contribution of the two task components (distractor presentation and standard prioritization) in isolation with the Standard Thematic Triad condition removed. We begin with the latter approach. GLMER models were built with the maximal random effects structure, where distractor presentation, standard prioritization and trial were included as fixed effects and participant, concept set (item) and trial were included as random effects (random slopes and intercepts were calculated by-participant for the effect of trial). The analysis uncovered reliable effects of distractor presentation (Beta Estimate = -.0733, *SE* = 0.27, Wald *Z* = -2.705, *p* = .007) and trial (Beta Estimate = 0.025, *SE* = 0.003, Wald *Z* = 8.027, *p* < .001), but standard prioritization (*p* = .84) and its interaction with distractor presentation (*p* = .23) were not reliable predictors; including the interaction produced a model where only the trial predictor remained a reliable effect. In other words, contrary to what was expected, presentation of distractors produced *less frequent* taxonomic responding and standard prioritization had no effect (see Table 5).

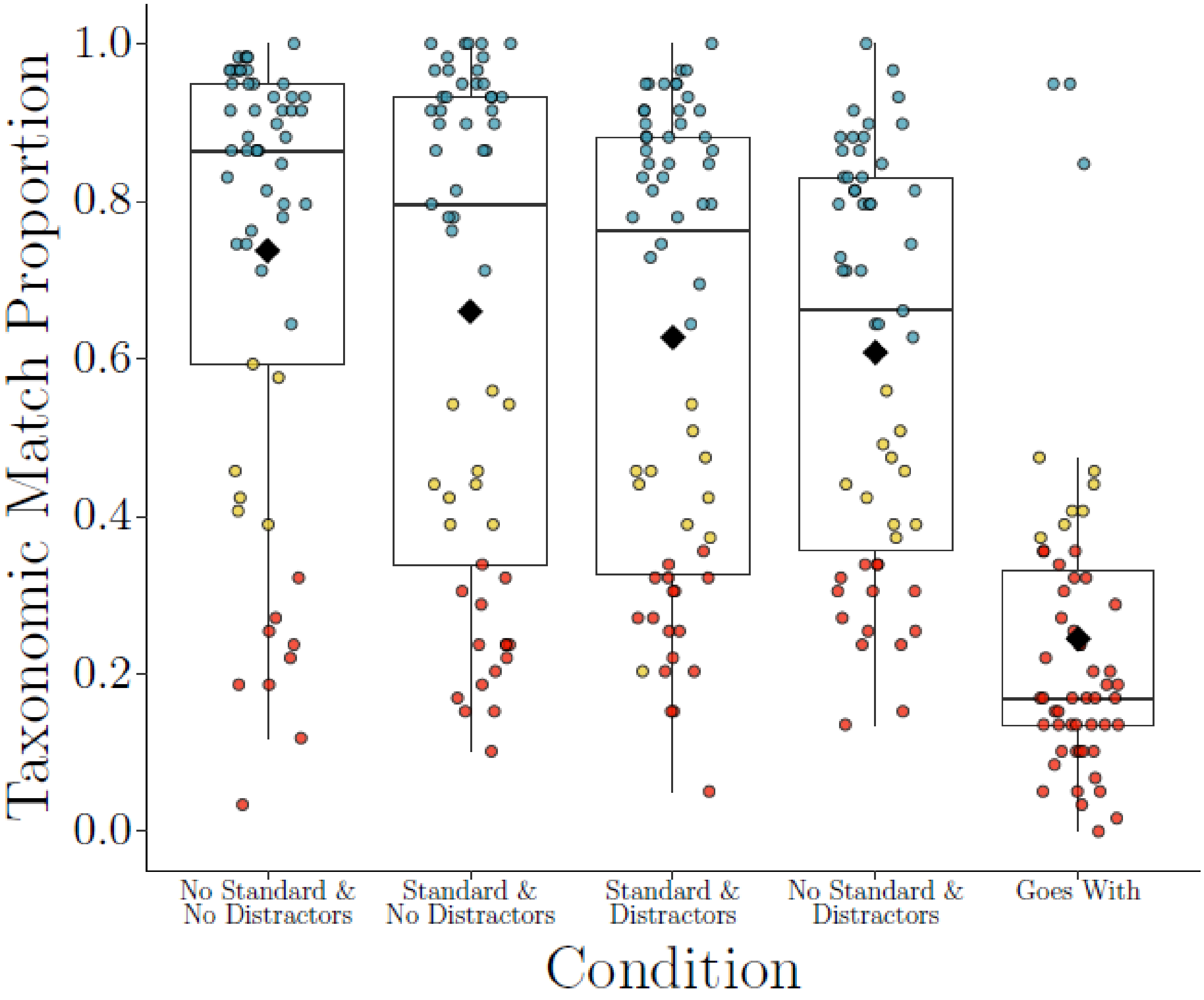


Figure 8. Proportion of taxonomic matches by condition in Experiment 2. Participants are represented as points, diamonds present the condition means and Tukey's box plots present the median and interquartile range of mean taxonomic responding.

The condition-based analysis differs in that the two-level distractor presentation and standard prioritization factors were replaced with a categorical "condition" factor with five levels: the conditions of Experiment 2 (see Table 4). The final model included condition and trial as fixed-effect predictors and the same random effects structure as the factor-based analysis above. The broad pattern of results is as follows: the No-Standard Triad group produced the most taxonomic responding—reliably more taxonomic matches than all groups except for the Standard Triad condition. Conditions with similarity-based instructions produced more taxonomic responding than the Standard Thematic Triad (Goes-With instructions) condition (see Figure 8).

To restate, the No-Standard Triad condition produced reliably more taxonomic responding than all conditions except the Standard Triad condition ($p = .29$); No-Standard Triad vs. No-Standard Ring, Beta Estimate = 1.064, $SE = 0.37$, Wald $Z = 2.868$, $p = .004$; vs. Standard Ring, Beta Estimate = 0.819, $SE = 0.36$, Wald $Z = 2.263$, $p = .024$; vs. Standard Thematic Triad, Beta Estimate = 3.322, $SE = 0.38$, Wald $Z = 8.671$, $p < .001$. The taxonomic responding rate in the Standard Triad condition was reliably higher than responding in the Standard Thematic Triad condition (Beta Estimate = 2.927, $SE = 0.38$, Wald $Z = 7.711$, $p < .001$) and marginally higher than the No-Standard Ring condition (Beta Estimate = 0.669, $SE = 0.37$, Wald $Z = 1.807$, $p = .071$). Both the Standard Ring (Beta Estimate = 2.503, $SE = 0.37$, Wald $Z = 6.797$, $p < .001$) and No-Standard Ring (Beta Estimate = 2.257, $SE = 0.38$, Wald $Z = 5.953$, $p < .001$) conditions produced more taxonomic responding than the Standard Thematic Triad group.[3]

Given the marginal results of the Experiment 1 GLMER and the convergence failures encountered above, it seems important to try to characterize the variance that these mixed effects approaches are accounting for—especially because occasional convergence failures in the analysis might be due to a lack of adequate statistical power to model the participant and item-level variance in a mixed-effects approach. Therefore, to close our inquiry into the task-based contributors to taxonomic response frequency, we present simple generalized linear models (GLM) of the effects of the presence of distractors and a prioritized standard. Note that these models are less conservative than the mixed-effects models presented above; they do not include participant or trial-level random intercepts or slopes—consequently, this also increases the power of these tests. The GLM model was built to test the hypothesis that standard prioritization, distractor presentation, and their interaction would be reliable predictors of

[3] We note that the condition-based models presented here occasionally fail to converge and require additional iterations. The favored approach for mixed-effects regression in R (and particularly the procedure for identifying convergence failure) is still under development (see supplemental materials for analysis code and results). The results presented here do stabilize when the optimization procedure includes more epochs than the default; the results are also consistent across a set of suggested optimizers.

taxonomic responding without accounting for random participant and item effects (the Standard Thematic Triad condition was excluded). This is exactly what was found. The model produced a reliable interaction between standard prioritization and distractor presentation (Beta Estimate = 0.453, *SE* = 0.07, Wald *Z* = 6.19, *p* < .001), where the task with no prioritized standard and no distractors (the No-Standard Triad condition) produced the highest levels of taxonomic responding. As for the individual fixed effects, distractor presentation produced less taxonomic responding (Beta Estimate = -0.15, *SE* = 0.05, Wald *Z* = -2.89, *p* < .001) and the removal of the standard produced more taxonomic responding (Beta Estimate = 0.37, *SE* = 0.05, Wald *Z* = 6.92, *p* < .001). In sum, these results suggest that the presence of a standard produces more taxonomic responding when distractors are present and less taxonomic responding when they are absent.

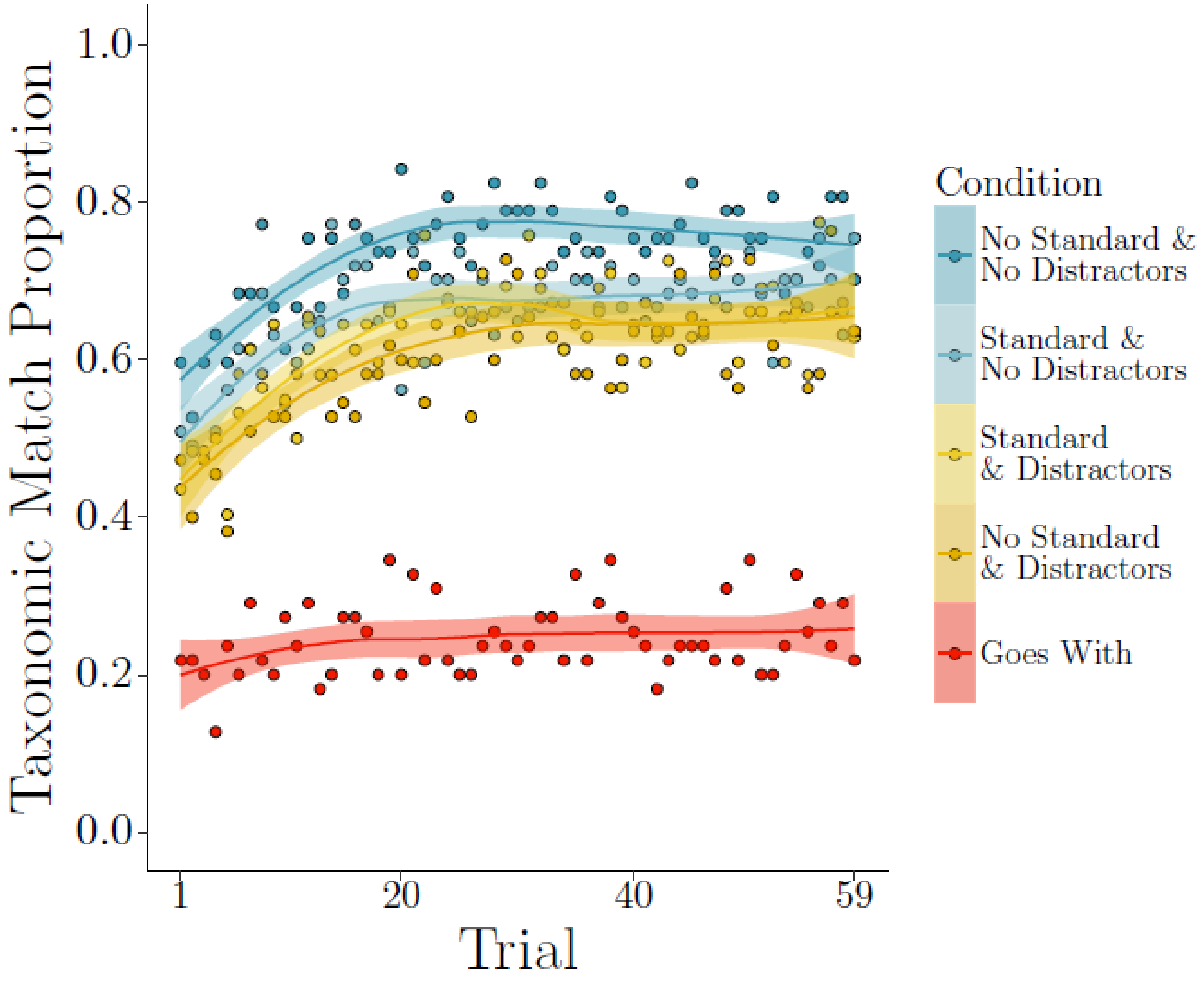


Figure 9. Taxonomic responding frequency across trials in Experiment 2. Points represent mean taxonomic responding by trial for each condition.

Time Course Analysis

Taxonomic responding increased in frequency across the experimental session for all conditions except the Standard Thematic Triad (Goes-With) condition. Responses were more likely to be taxonomic matches as the experiment progressed under the condition-based (Beta Estimate = 0.020, *SE* = 0.003, Wald *Z* = 7.733, *p* < .001) and the factor-based (Beta Estimate = 0.025, *SE* = 0.003, Wald *Z* = 8.027, *p* < .001) analysis approaches (see Figure 9). This effect holds within condition for every condition except the Standard Thematic Triad (*p* = .196) paralleling the results from the No-Reminder condition in Experiment 1.

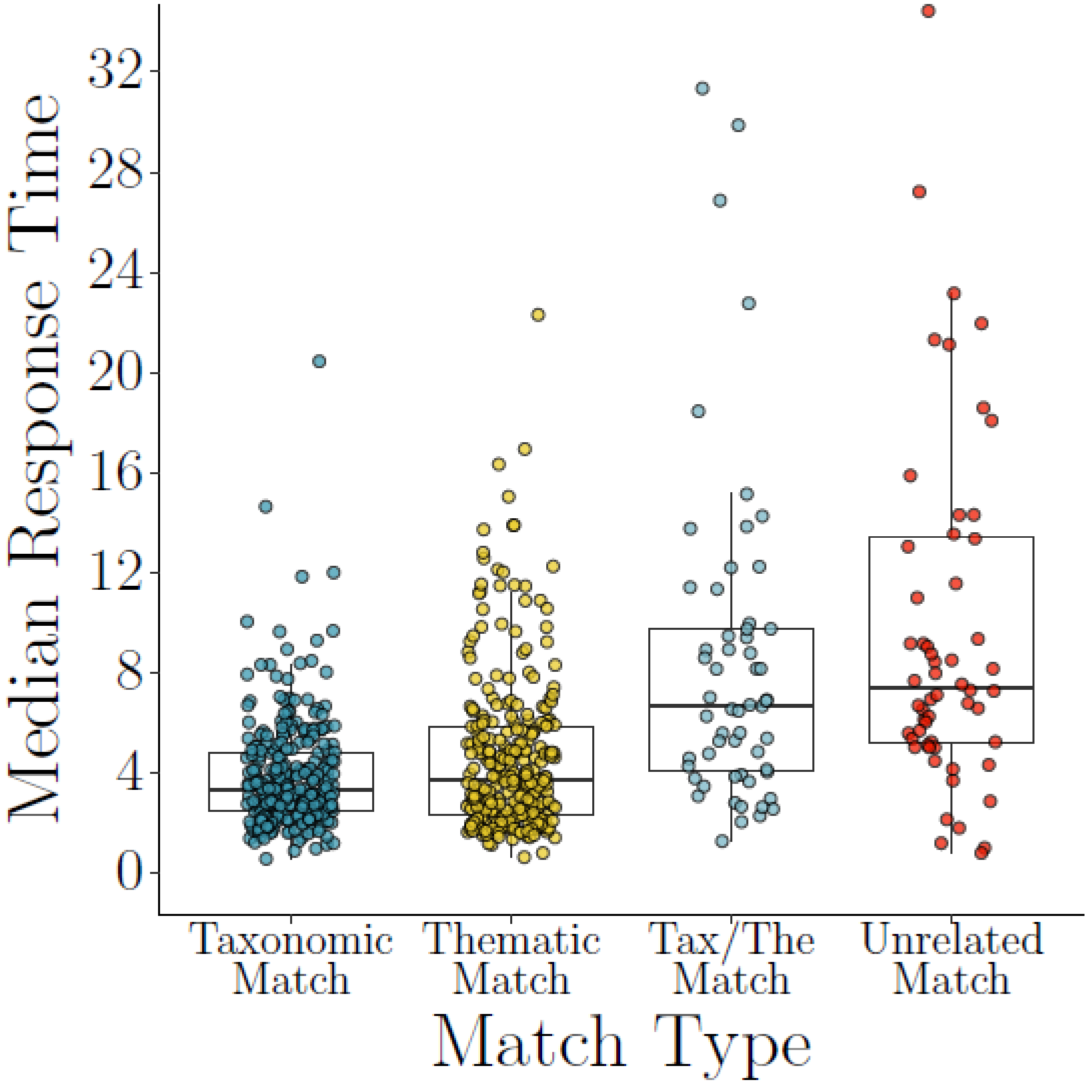


Figure 10. Experiment 2 median response time by match type. Median response time for each possible match type. Participant median response times are presented by points for each match type where they produced at least one match.

Response Latency Analysis

As in Experiment 1, trial duration was recorded and analyzed (Table 5). The response time results for Experiment 2 closely parallel those of Experiment 1, where trials that resulted in a taxonomic match were completed faster than trials with a thematic match (Beta Estimate = -0.901, *SE* = 0.31, *t* = -2.887, *p*

= .004) and thematic matches were faster than trials that included unrelated concepts (Beta Estimate = -4.496, *SE* = 0.57, *t* = -7.822, *p* < .001) and taxonomic and thematic targets (Beta Estimate = -1.976, *SE* = 0.58, *t* = -3.221, *p* = .001) as matches (see Figure 10). Recall that the latter two types of responses are quite infrequent, and thus, the cell count between different types of matches is imbalanced. Within condition, only one similarity-biased condition did not exhibit the response time effect for taxonomic matches—the No-Standard Ring condition (*p* = .44). Note that a null result was also found under the same conditions in Experiment 1. Reversing the general pattern for the similarity-biased conditions (and fitting with the idea that response time is closely associated with the most frequent target for a given task and set of instructions), the fastest type of match in the Standard Thematic Triad condition was thematic, Beta Estimate = 0.80, *SE* = 0.18, Wald *Z* = 4.44, *p* < .001.

Table 5
Experiment 2 Frequency of Matches and RT by Match Type

Match Type	Frequency (Count)	Mean of Participant Median RT
Standard and Taxonomic Target	57.9% (9,762)	3.86 seconds
Standard and Thematic Target	40.7% (6,869)	4.77 seconds
Taxonomic Target and Thematic Target	.5% (75)	8.43 seconds
Match including an Unrelated Distractor	1% (168)	11.21 seconds

Discussion

The present results replicated two key findings from Experiment 1: the overall bias toward taxonomic matches and the increase in frequency of taxonomic responding over time. These results throw new light on our initial concerns—one does not need to modify the characteristics of the task to reverse the traditional thematic preference; we see uncommitted (50/50) responding at the outset and a clear shift toward dominant taxonomic responding with experience. The traditional task (Standard Triad) shows this pattern and falls in the middle of the pack relative to the tested task variations. To be clear, these factors were proposed relative to the expectation of a traditional thematic preference. A surprisingly high rate of taxonomic responding was found in the Standard Triad condition. All similarity-based conditions produced more taxonomic responses (60% or greater)—even though the only condition with a reliable taxonomic bias at the participant level was the No-standard Triad condition. A conservative interpretation of the analysis would suggest that the contribution of factors initially hypothesized to increase thematic responding was negligible. The presence of distractors appears to have *increased* thematic responding. Forcing participants to choose two examples (removing the prioritized role of the standard) had no effect on the proportion of taxonomic matches. In terms of the descriptive pattern of results: the No-Standard Triad condition produced the most taxonomic responding and among the distractor conditions a prioritized standard produced more taxonomic responding than the absence of a prioritized standard. It is possible that more statistical power would make these descriptive patterns reliable (see the general discussion for further elaboration on this point), but as is the evidence is not strong enough to make more concrete conclusions about an interaction effect when controlling for participant and item random effects.

Moving on to the response time analysis, the results provide further evidence for the unexpected pattern that trials with taxonomic matches were completed faster than trials with thematic matches. Interestingly, when the conditions of Experiment 2 were analyzed in isolation the results showed that the Standard Thematic condition produced the opposite pattern from the aggregated results across conditions—thematic matches were completed faster than taxonomic matches. Perhaps then a key determinant of response time is the type of semantic relation that is consciously being searched for. Under this account, we'd expect that past reports of a speed advantage for thematic matches in unconstrained, non-deadline tasks are related to task-based biases. It is possible to take this pattern as converging evidence of ambiguity in the task when "Similarity" is featured in the instructions. This question deserves further investigation.

The results of Experiment 2 support two main conclusions: (1) task-based properties have consequences for the effect of thematic association on similarity judgments and (2) taxonomic responding is the dominant response preference for the similarity task and thematic responding is the dominant response for the Goes-With task. It does not require clarifying the similarity instructions (as in Experiment 1) or modifying the characteristics of the trial format to reverse the accepted result— a simple and effective way to elicit responding aligned to longstanding theoretical definitions of similarity is to provide a longer opportunity to figure out how best to approach a novel and (purposefully) underspecified task.

Experiment 3

Experiments 1 and 2 uncovered an unexpected pattern of reliably more taxonomic responding (with the caveat that this pattern was strongest at the trial level, but not always at the participant level). These results are in direct contrast to the mixed, but generally majority-thematic results of past reports. At this point it should be asked—could the unexpected pattern be due to some artifact in our experimental set-up? Experiment 2 featured a direct replication of the classic triad task with many of the concept sets adopted from previous projects; performance in the first half of the task was only mildly inconsistent with past results. The dramatic time course finding was that at about the point where most studies end, participants are just beginning to show a stable 2:1 taxonomic preference. To our knowledge, there is nothing non-representative about the population we tested. To be cautious—and despite the inclusion of concept set variance as a random effect in our analyses—we next investigate the possibility that the concept sets created for this investigation are responsible for driving the effects.

To accomplish this, we collected ratings to determine the perceived connection strength of the taxonomic and thematic relations in the materials used for this report. In a between-subjects design, people were asked to rate pairs of concepts (targets with the corresponding taxonomic choice, thematic choice or unrelated distractor) for either their degree of similarity or how well the concepts *go together*. No instructions were provided to explain what was meant by “go together” or “similarity.” This was intentional—to maintain an alignment with the circumstances of the experiments reported above. Competing semantic relations (i.e., taxonomic *and* thematic pairs) from the same set were not included in the presented concept pairs for a given participant; it was thought that the presentation of both pairs from one set might hint at the key distinction and skew ratings (see Simmons & Estes, 2008 for an example of this effect). Trials were a randomly-ordered mix of taxonomic, thematic and unrelated concept pairs, so it was hoped that the key semantic relationships would not be recognized. Ad hoc post-task interviews were used to confirm that participants were naïve to the systematic relationships at study.

Our goal was to verify that the concept sets used for Experiments 1 and 2 were not biased toward taxonomic responding—a potential explanation of the observed taxonomic responding rates in the experiments. A straightforward expectation is that taxonomic pairs should have higher similarity ratings than association ratings and thematic pairs should be rated higher as associates (i.e., things that go together better) and lower in terms of similarity. Critically, it is important that the similarity and association ratings of the taxonomic and thematic pairs (respectively) are not radically different within each concept set. If the similarity of taxonomic pairs is rated higher than the association strength of corresponding thematic pairs (within a concept set), it could be argued that the taxonomic matches are stronger. The present design also should confirm that the unrelated distractors were indeed unrelated—rated lower on taxonomic similarity and thematic association than the related matches. Lastly, it is also possible that the rating data can help clarify the interpretation of the results of Experiments 1 and 2—such as the effect of increased taxonomic responding across trials. For example, (while it may be unlikely) could it be that the increase across trials was due to a catastrophic failure of randomization in which the concept sets that have stronger taxonomic pairs were frequently shuffled to the back of the trial order? GLMER models from the previous experiments can be re-analyzed with the taxonomic and thematic ratings included to test for this possibility.

Method

Participants and Materials

Participants (*N* = 202) were recruited, compensated and consented in an identical fashion as the previous experiments. Participants were randomly assigned to the Taxonomic Similarity condition or the Thematic Association condition—the difference being the question used to solicit ratings. The concept sets were those used in the previous experiments. For each concept set, a taxonomic or thematic match and two unrelated concepts (from the three possible) were randomly selected to be included in the trial list. This way the thematic and taxonomic pairs from a set were never presented in the same session. This procedure produced two pairs of concepts (one related, one unrelated) from each concept set for 118 rating trials in total. On each trial, the task interface included a pair of concepts, a rating scale, and condition specific instructions for the rating task (a depiction of the task interface is provided in Appendix C).

Procedure

Pairwise rating trials consisting of two concepts from a concept set were presented in random order. Participants were presented with a pair of concepts and asked to provide ratings on a (ratio-scale) rating line (0 to 100 with tick marks in 10-point increments) with the option to provide their rating at any point along the line. The rating scale was anchored with NOT AT ALL and VERY SIMILAR for the taxonomic rating condition, and NOT AT ALL and VERY WELL (for the question of how well the items go together) for the thematic rating condition.

Results and Discussion

There are two main issues to address: (1) do we confirm that the taxonomic pairs are rated as more similar, the thematic pairs are rated as more associated and the unrelated pairs are rated lower on both? and (2) do the ratings offer any insight into the results of Experiments 1 and 2? Normalized descriptive statistics show that—with one exception (the taxonomic pair: HAPPY, SAD)—no set of taxonomic pairs had a mean similarity rating lower than its corresponding rating on the thematic scale (see Table 6; complete rating data is provided in Appendix D). Similarly, the majority of the thematic pairs (91.5%) were rated higher on thematic relatedness than their corresponding similarity rating (the thematic pairs rated higher on similarity than thematic relatedness were: HAPPY, SMILE; FLOSS, TOOTHBRUSH; RAPIDS, RIVER; TRAILER, TRUCK; FIELD, GRASS). The aggregated results show that taxonomic pairs were rated as more similar than the thematic pairs (Beta Estimate = -7.489, *SE* = 0.55, *t* =

-13.6, $p < .001$) and unrelated pairs (Beta Estimate = -54.90, *SE* = 0.47, $t = -115.70$, $p < .001$) according to an LMER model built to predict similarity ratings with pair type (taxonomic, thematic, unrelated) as a fixed-effect predictor and participant as a random predictor. The thematic pairs were rated higher as concepts that go together when compared to the taxonomic pairs (Beta Estimate = -15.64, *SE* = 0.56, $t = -27.88$, $p < .001$) and the unrelated pairs (Beta Estimate = -65.81, *SE* = 0.49, $t = -135.44$, $p < .001$) in an LMER model predicting thematic ratings with an identical predictor structure (Figure 11). Perhaps most importantly, the standardized mean similarity ratings of taxonomic pairs (*z*-scores of similarity ratings subtracted by *z*-scores of thematic relatedness ratings) were not reliably different from the standardized mean association rating of thematic pairs (*z*-scores of thematic relatedness ratings subtracted by the corresponding similarity ratings) from the same set ($M_{\text{Difference}}$ = 0.047 SD) according to a paired *t*-test, $t(58) = 1.117$, $p = .27$.

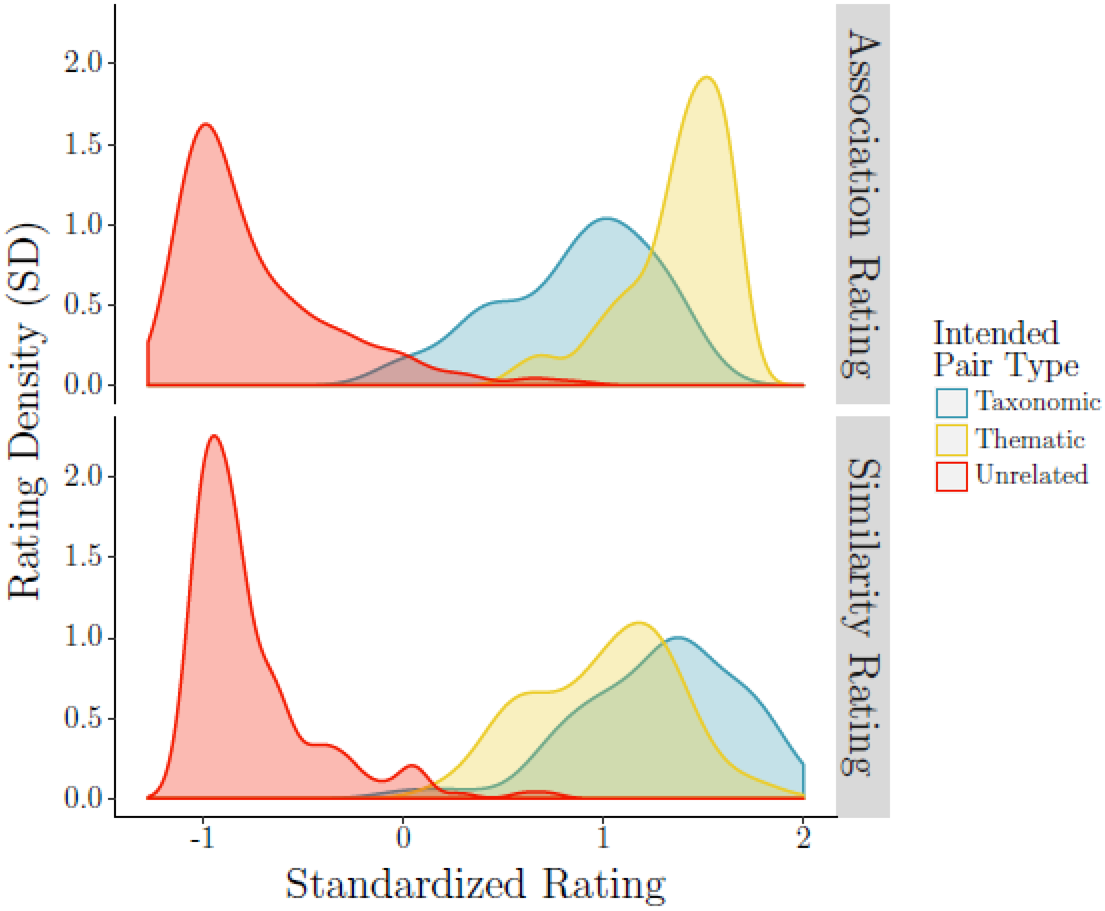


Figure 11. Density plot of standardized ratings for the similarity and association-based rating tasks.

To summarize, no evidence was found to suggest that the experimental materials are biased to produce a certain type of responding: taxonomic pairs were rated as more similar, thematic pairs were rated as more associated, unrelated pairs were rated low on both dimensions, and the strengths of similarity and association ratings within each concept set were not reliably different. This analysis suggests that the results of Experiments 1 and 2 cannot be attributed to materials that bias responding toward a particular type of match.

Table 6
Experiment 3 Concept Ratings

Pair Type	Similarity Rating Mean (SD)	Thematic Rating Mean (SD)	Similarity Rating Mean Response Time	Thematic Rating Mean Response Time
Taxonomic	68.02 (1.29)	71.08 (0.87)	4.00 seconds	3.93 seconds
Thematic	60.47 (1.00)	86.68 (1.37)	4.07 seconds	3.54 seconds
Unrelated	13.09 (-0.76)	20.90 (-0.75)	3.94 seconds	4.15 seconds

We now return to the results of Experiments 1 and 2 with the benefit of the Experiment 3 ratings. GLMER models were constructed with identical predictor structures to those presented in the previous experiments except that the random intercept term for concept set was replaced with a rating difference score for taxonomic similarity and thematic association based on the properties of the concept ratings of each set. The difference score was computed by taking the similarity score for the taxonomic pair of each concept set (standardized similarity rating – standardized association rating) and subtracting the association score of the corresponding thematic pair (standardized association rating – standardized similarity rating). The re-analysis of Experiment 1 including this difference score did not uncover any effects that diverged from those presented above. Similarly, including the difference score in a re-analysis of the factor-based (distractor and standard prioritization factors) approach from Experiment 2 did not produce a difference in reliable predictors as compared to the initial analysis. Of particular note, the trial effect (where taxonomic responding increased across trials) remained significant even when relative strength of similarity and association was accounted for in the model.

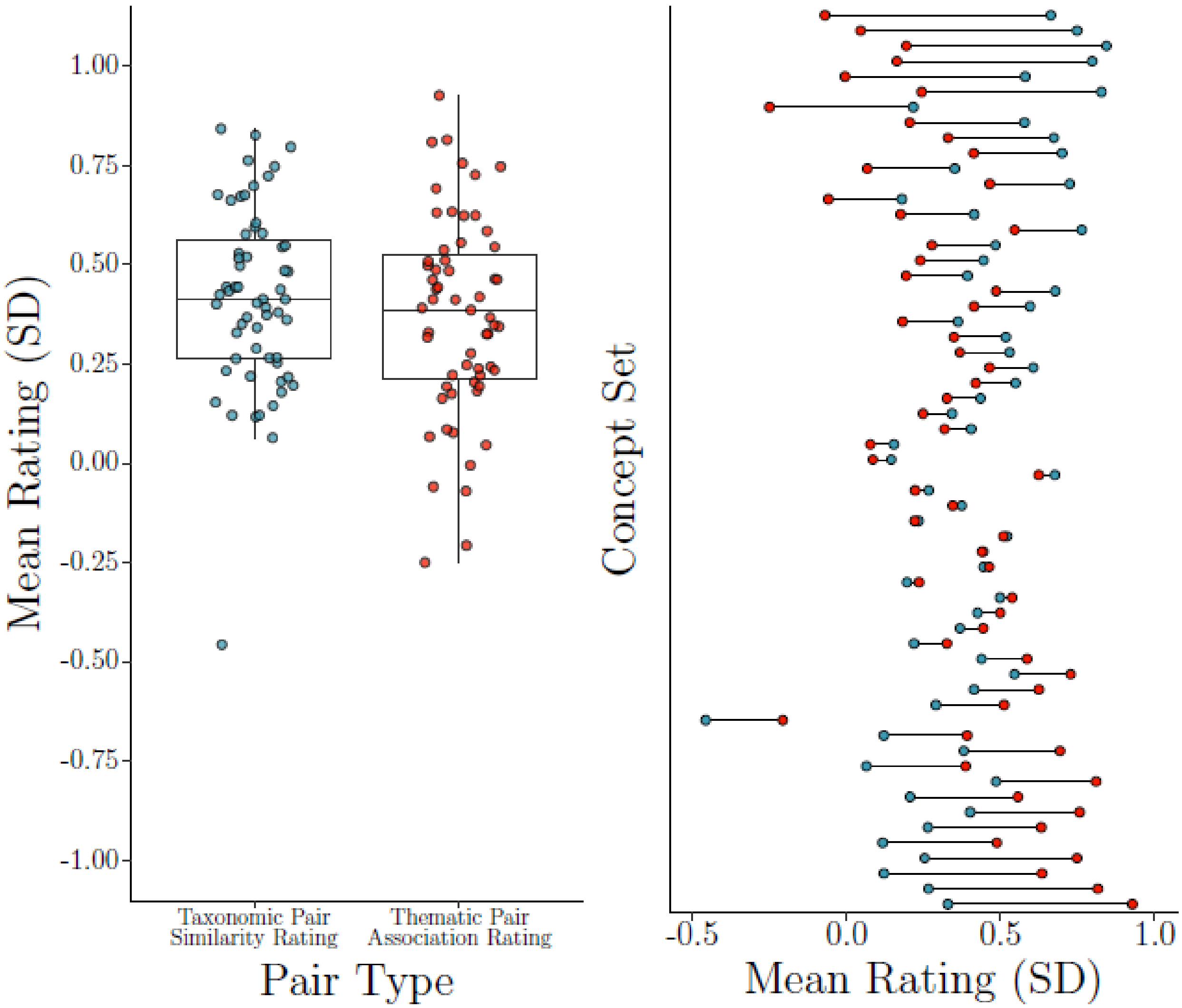


Figure 12. Visualization of the concept ratings overall (left) and paired with the match from the same concept set (right). The left panel depicts the mean similarity and association ratings for the taxonomic and thematic pairs, respectively. The right panel depicts the paired difference of the similarity (blue) and association (red) ratings within each concept set. Mean similarity and association ratings were produced by subtracting the type-consistent rating by the type inconsistent rating (i.e., taxonomic ratings are a

calculation of standardized similarity ratings subtracted by standardized thematic ratings for each taxonomic pair).

General Discussion

This project set out to investigate the reported pattern in human similarity judgments where thematic associates are identified as more similar than concepts that share taxonomic similarity. It was hypothesized that the classic triad task artificially inflates the rates of thematic responding due to ambiguity in instructions and other characteristics of the task. Instructions designed to clarify the interpretation of similarity (without explicitly providing examples or telling participants what to do) served to increase taxonomic responding. The role of design characteristics of the triad task (e.g., prioritization of the standard, forced-choice judgment) is less clear. Rendering this point somewhat moot, the overall response pattern found in this series of experiments undermined the very premise (based on the existing literature) of a thematic response bias. This reversal was found using the most traditional version of the task: selecting the most similar option in a forced-choice triad. Given these findings, the main contribution of this work is not how components of the task affect responding—our interpretation is that they did, but more work is needed to clarify their role. Rather, the surprising result of this investigation is that thematic responding appears to have been markedly overestimated in previous assessments.

Time-course of Taxonomic Responding

One of the most important discoveries of this project is the distinctive time course pattern toward increased taxonomic responding. People start out in the task unsure about how to respond and eventually settle in to a consistent responding pattern. Most frequently, this shift results in a response preference for taxonomic matches. This pattern is a challenge for the dual-process integration view. If thematic intrusion is an unavoidable consequence of producing a psychological similarity judgment—a feature, not a bug—why does this intrusion lose its effectiveness across the time course of a series of similarity judgments?

On the other hand, this pattern of responding fits with the central argument of the confusability account. When the task begins, there is little or no expectation about what type(s) of concepts will appear, how they will be connected, or how to treat the notion of similarity. Our experiments reveal that when similarity is more cleanly articulated in the instructions in terms of explaining to an alien which things on earth are similar to each other, performance is strongly taxonomic-based from the early trials on. In addition, we find that when instructions to choose the most similar match are stated initially, but not continually reiterated on each trial, participants drift toward a task interpretation that is best characterized as failing to monitor against the temptation of thematic intrusions and sliding toward (perhaps subconsciously) interpreting the task as more like the "Goes-With" instruction than the similarity instruction. It is true that this 'microcosmic taxonomic shift' has a ceiling—there is still a low level of thematic responding that occurs—and this is an interesting topic for further investigation.

Task Properties Impact Taxonomic Responding

The evidence presented here suggests that task manipulations have consequences for the variable effect of thematic intrusion on similarity judgments, as previously shown for different task manipulations (Gentner & Brem, 1999; Lin & Murphy, 2001; Mirman & Graziano, 2012; Murphy, 2001; Simmons & Estes, 2008). Experiment 1 showed that variations of similarity-based instructions can produce different rates of taxonomic responding, where instructions that sought to highlight the importance of taxonomic information for a naïve individual (the Alien condition) produced the highest level of taxonomic responding observed in Experiment 1. Conversely, the removal of a consistent reminder about the goal of the task produced the highest level of thematic responding in that experiment. This is additional evidence that an important determinant of responding preference is related to the on-line interpretation of task goals and instructions (Lin & Murphy, 2001; Nguyen & Murphy, 2003; Skwarchuk & Clark, 1996). More importantly, it is further support for the idea that thematic responding can (at least in part) be attributed to intrusion and confusion during similarity judgment processing. Without the support and clarification (differentially) provided by the instructions, thematic associates are chosen as more similar than taxonomic category members. The use and interpretation of the concept “similar” does seem to play an important role in these experiments. Where response patterns have been attributed to underlying individual differences such as cognitive ability or processing style (Simmons & Estes, 2008), a caveat may need to be added that these patterns are also affected by the interpretation of the task goal.

The effects of the task modifications and its individual components are less clear. Our initial interpretation of Experiment 1 was that the removal of a standard and addition of distractors were linked to the high rates of taxonomic responding. The results of Experiment 2, however, cast doubt on this interpretation. Removing the prioritized standard from the classic task produced the highest level of taxonomic responding, but a conservative interpretation of the analysis suggests that this was not reliably different from the classic 2AFC triad task. Taxonomic matches in the classic triad task (a conceptual, if not identical replication of previous inquiries) were more frequent than thematic matches. The addition of distractors produced fewer taxonomic matches when there was no prioritized standard, but the opposite was found—at least descriptively if not inferentially—when the standard was provided above the concept array in a prioritized position. With more power it is possible that a reliable interaction would have been observed with a mixed-effects approach, but with the present data it was only found when participant-based variance was not included in the model. The two-way interaction—where no standard with no distractors produced the most taxonomic responding and a prioritized standard with distractors produced more taxonomic responding than without distractors—does fit nicely with research on the role of working memory capacity during processing of taxonomic and thematic relations. The overall decrease in taxonomic responding associated with distractor presentation consistently found in Experiment 2 might be attributable to these working memory effects; the co-presentation of distractors has been shown to affect picture naming differentially for taxonomic and thematic category members (de Zubicaray et al., 2013; Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Rose & Rahman, 2016) and serial presentation appears to limit the effect of thematic intrusion (Rey & Berger, 2001).

The Role of Individual Differences

A pressing issue to resolve is the role of individual differences for thematic intrusion (see also Lin & Murphy, 2001; Simmons & Estes, 2008), as this component—more so than task constraints or stimulus properties—has the highest potential to affect everyday thinking. Even in the conditions that produced responding heavily biased toward taxonomic matches, there were still people who maintained their preference for thematic matches. Likewise, there were people in the Standard Thematic Triad condition that had a taxonomic responding preference. Ongoing work aims to shed more light on this subsample of "holdouts" and how their interpretation of the task and cognitive processing might differ from someone who is presumably more affected by the task constraints. Another interesting follow-up question might be to try to classify the various and competing interpretations of the task goals, instructions and definitions of similarity. Do people who match more thematic associates explicitly define similarity as a sum of taxonomic and thematic information? There is some evidence that this is the case—rating participation in a scenario as important for similarity does correspond to a thematic response preference in past work (Simmons & Estes, 2008)—but the question about whether this judgment is due to confusion remains unresolved. In relation to the present work, does this definition changes over the course of the task? Is it stable across longer periods of time? Answering these questions would go a long way to help to support or refute the confusability and dual process integration accounts.

Work at the individual differences level is important because taxonomic and thematic responding patterns might underlie more general properties of cognitive processing. While the individual differences data available today are largely correlational, it is possible that response biases have deep underlying consequences for cognition. Preferences for thematic responding are associated with low scores on the Need for Cognition (NFC) scale (Simmons & Estes, 2008). Young children, elderly adults and those with temporal lobe damage are more likely to show thematic responding preferences (Schwartz et al., 2011; Smiley & Brown, 1975; 1979). Cross-cultural differences in the prevalence of thematic responding have also been found (Ji, Zhang & Nisbett, 2004), though the reliability of these findings has been questioned (Saalbach & Imai, 2007). Formal education (or lack thereof) and occupational pressures have also been suggested as drivers of thematic responding (Denney, 1974; Sharp, Cole, & Lave, 1979)—though this evidence is more characteristic of an early view of the taxonomic response bias as the result of mature and normative cognitive functioning. Rabinowitz and Mandler (1983) explain this position well—it is the view that a taxonomic classification preference is the result of mature semantic knowledge structure, the "endpoint" of conceptual development, the typical or ideal adult functioning pattern. As might be apparent given the difficulties that exist to determine reliable response preferences, consistency in this work has been hard to come by; the link between education and taxonomic responding has also recently failed to replicate (Mirman & Graziano, 2012). While clearly more work needs to be done to clarify the relationship between response patterns and the broader cognitive implications, it is of the utmost importance that the source of the thematic response bias itself is better understood and reliably predicted before attempts are made to link it to other behavioral or demographic data.

What Made These Experiments Different?

The simplest reason why the results presented here diverge from past research is that much of that work features fewer concept sets (stimuli) than the present experiments (cf. Hendrickson et al., 2015; Skwarchuk & Clark, 1996). It is not uncommon to see as few as 20 concept sets in these investigations. We also note that the number of participants used in this investigation is larger than the average sample size in this domain, where 20-30 participants per condition is typical. This is simply too few items and participants to adequately measure the phenomenon, especially for the outcome measure of reliable response bias frequency. The semantic relationships of real-world concepts are messy—we can attempt to control for the relative strength of the taxonomic and thematic relations in these investigations (e.g., Experiment 3), but this process is necessarily imperfect and every person arrives in the lab with a distinct semantic experience of the world (Mirman & Graziano, 2012; Simmons & Estes, 2008). Regardless of the quality of concept norming, it seems that we must ultimately rely on concept pairs that vary in how well they capture the qualities of these semantic relations without confound. The best defense against this problem is to maximize the number of data-points available in terms of concept sets and sample size.

The analyses presented here have shown that people settle in to a responding pattern, but it takes some time (see Figures 5 and 9). Reliable increases in the taxonomic responding rate across trials in similarity-based tasks were consistently found. This is a clear problem for past research. In a 30 trial experiment analyzed with aggregation-based statistics, the outcome measure is averaged over all trials—the majority of which are completed as the responding preference is stabilizing. Aggregation-based statistics underestimate the strength and direction of these responding preferences.

This issue also raises another difference—the advantages of trial-level, mixed effects analysis. The analyses here include random effects of concept set and participant wherever possible. Where models did not include these random effect terms in this work, they were often found to be anti-conservative. While it should be standard practice to include experimental stimuli and participants as a source of variance when an experiment features crossed random factors (Judd, Westfall, & Kenny, 2012), the present situation of crossed items and participants necessitates this analysis approach—at the very least to attempt to address the possibility that the effects are driven by individual participants or concept sets. The reliable condition differences found without mixed-effects show the need to account for this variance, where cleaner interpretations would have been possible were it not for the inclusion of subject and concept set variance. The difficulty of this approach is also on display with the model convergence failures that occurred. The participant-level differences found here are difficult to adequately model. Our hypothesis is that this difficulty is caused by over-dispersion in the binomial outcome measure. In Experiments 1 and 2, only a few participants were reliably biased toward thematic responding, so there were far fewer trials to analyze relative to the conditions with more frequent taxonomic responding. This suggests two things: where differences emerge between the simple and mixed-effects approaches, caution should be taken in interpreting the results; differences between conditions found without random effects analyses should not, however, be completely discounted. For the present investigation, we have tried to present these ambiguous results in as much detail as possible so that readers can make their own conclusions. For future work, one possible solution to guard against the disappearance of reliable differences with random effects is to design studies with more power to adequately sample these effects.

Conclusion

In the end, we set out to solve a problem (why do people think cows are more like milk than horses?) and found that the problem was harder to reproduce than expected. We only found an overall thematic preference when people were left unguided to treat the triads however they liked and when they were specifically asked to select for associativity. We believe this reversal can be explained in part by advances in methodology (avoiding aggregation-based statistics, using mixed-effects analysis techniques, running more participants, and careful selection and verification of materials), but the most compelling element is that increasing the number of trials reveals a majority convergence toward taxonomic responding.

The experiments presented here suggest that thematic intrusion is largely controllable with experience, so theoretical accounts of similarity that propose thematic association as an inseparable component process of the similarity judgment system must confront this issue to remain viable. The data presented here fit better with an alternative account: thematic association is not a component of the similarity judgment process, it intrudes on the similarity judgment process. Thematic association is confusable with, but ultimately, distinguishable from taxonomic similarity. Psychological similarity must be relied on in the service of learning and reasoning—inference, generalization, and categorization. The inclusion of thematic association in theories of psychological similarity unnecessarily clouds a challenging, but critical issue in psychological science.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429

Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory & cognition*, *15*(5), 397-418.

Barsalou, L. W. (1982). Context independent and context dependent information in concepts. *Memory & cognition*, 10, 82–93.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, *11*(3), 211–227.

Bassok, M., & Medin, D. L. (1997). Birds of a feather flock together: Similarity judgments with semantically rich stimuli. *Journal of Memory and Language*, *36*(3), 311–336.

Canessa, N., Borgo, F., Cappa, S. F., Perani, D., Falini, A., et al. (2008). The different neural correlates of action and functional knowledge in semantic memory: an fMRI study. *Cerebral Cortex, 18,* 740 – 751.

Chen, Q., Ye, C., Liang, X., Cao, B., Lei, Y., & Li, H. (2014). Automatic processing of taxonomic and thematic relations in semantic priming—Differentiation by early N400 and late frontal negativity. *Neuropsychologia*, *64,* 54-62.Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312–1323.

Davidoff, J., & Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes, 19* (1), 137–174.

Denney, N. W. (1974). Evidence for developmental changes in categorization criteria. *Human Development*, *17,* 41–53.

de Zubicaray, G. I., Hansen, S., & McMahon, K. L. (2013). Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General*, *142*(1), 131.

Estes, Z. (2003). A tale of two similarities: Comparison and integration in conceptual combination. *Cognitive Science*, *27*(6), 911–921.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic Thinking: The Apprehension and Consequences of Thematic Relations. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 249.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, *41*(1), 1–63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive science*, *19*(2), 141–205.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science, 7*(2), 155–170.

Gentner, D., & Brem, S. K. (1999). Is snow really like a shovel? distinguishing similarity from thematic relatedness. In *Proceedings of the twenty-first annual meeting of the Cognitive Science Society* (pp. 179–184).

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory & Cognition*, *29*(4), 565-577.

Gentner, D., & Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*(4), 609-642.

Gentner, D., & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (Ed.), *Similarity in language, thought and perception* (pp. 111–147). Brussels: BREPOLS.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive psychology*, *25*(4), 524-575.

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*(3), 359–376.

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1454.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and projects* (pp. 437–447). Indianapolis: Bobbs-Merrill.

Greenfield, D. B., & Scott, M. S. (1986). Young children's preference for complementary pairs: Evidence against a shift to a taxonomic preference. *Developmental Psychology, 22,* 19–21.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(06), 803–831.

Hendrickson, A. T., Navarro, D. J., and Donkin, C. (2015). Quantifying the time course of similarity. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 908–913). Austin, TX: Cognitive Science Society.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & cognition*, *15*(4), 332-340.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: experimental and computational studies. *Cognition*, 100, 464-482.

Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral Cortex*, *25*(11), 4319–4333.

Ji, L. J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of personality and social psychology*, *87*(1), 57.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, *55*(3), 196-231.

Kurtz, K. J., & Gentner, D. (2001). Kinds of kinds: Sources of category coherence. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, 522–527.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417–446.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, 130(1), 3–26.

Markman, E. M., Cox, B., & Machida, S. (1981). The standard object-sorting task as a measure of conceptual organization. *Developmental Psychology*, *17*(1), 115.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review, 100*(2), 254–278.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of experimental psychology: General*, *141*(4), 601.

Mirman, D., Landrigan, J. F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological bulletin*, *143*(5), 499.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289–316.

Nguyen, S. P., & Murphy, G. L. (2003). An Apple is More Than Just a Fruit: Cross-Classification in Children's Concepts. *Child development*, *74*(6), 1783–1806.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1), 8–13.

Rabinowitz, M., & Mandler, J. M. (1983). Organization and information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 430–439.

Rose, S. B., & Rahman, R. A. (2016). Cumulative semantic interference for associative relations in language production. *Cognition*, *152*, 20-31.

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

Rey, E., & Berger, C. (2001). Four- and five-year-old children's categorization: Sensitivity to constraints on word meaning and influence of stimulus presentation in a forced-choice paradigm. *Cahiers de Psychologie Cognitive, 20,* 63–85.

Saalbach, H., & Imai, M. (2007). Scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General*, *136*(3), 485–501.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., Mirman, D. & Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences*, *108*(20), 8520–8524.

Shah, A. K., & Oppenheimer, D. M. (2009). The path of least resistance: Using easy to access information. *Current directions in Psychological Science*, 18, 232–236.

Sharp, D., Cole, M., & Lave, C. (1979). Education and cognitive development: The evidence from experimental research. *Monographs of the Society for Research in Child Development*, *44*, 1–92.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition*, *108*(3), 781–795.

Skwarchuk, S., & Clark, J.M. (1996). Choosing category or complementary relations: Prior tendencies modulate instructional effects. *Canadian Journal of Experimental Psychology, 50,* 356–370.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3.

Sloman, S. A. (2014). Two systems of reasoning: An update. In Sherman, J., Gawronski, B., & Trope, Y. (Eds.). *Dual process theories of the social mind.* Guilford Press.

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age. *Journal of Experimental Child Psychology, 28*(2), 249–257.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and categorization*, *1*(1978), 79–98.

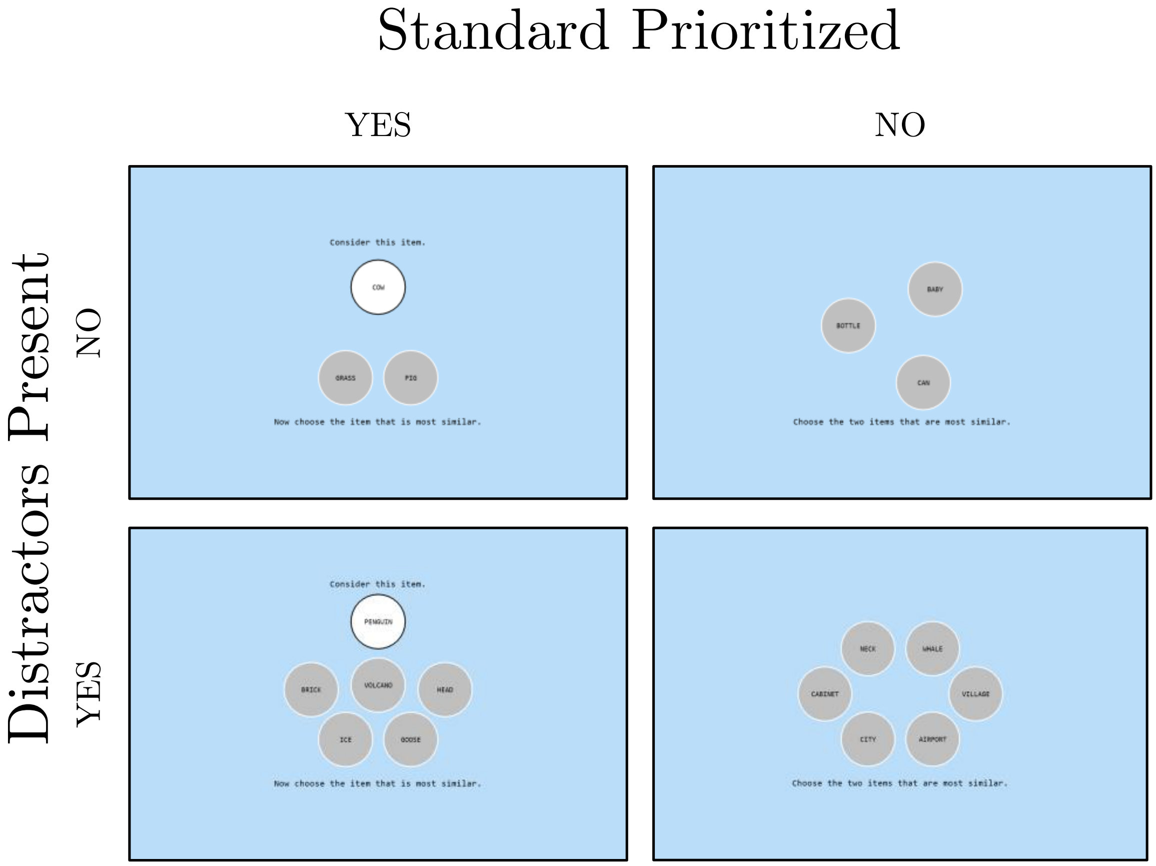
Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of experimental psychology: General*, *113*, 169–193.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology*, *39*(3), 208–238.

Appendix A: Concept Sets

Index	Standard	Taxonomic	Thematic	Unrelated	Unrelated	Unrelated
1	SPOON	LADLE	CEREAL	LION	TREE	STEREO
2	ROCKET	MISSILE	ASTRONAUT	BUG	CHEESE	WATER
3	GARLIC	ONION	VAMPIRE	HOUSE	FOOT	CODE
4	MILK	LEMONADE	COW	GUITAR	LEAF	WINDOW
5	SHIP	CANOE	SAILOR	UMBRELLA	BANANA	CHAIR
6	CAR	BIKE	SEATBELT	SHRIMP	COTTON	BISCUIT
7	CHAIR	SOFA	LEGS	BREAD	BALL	KEYBOARD
8	PANTS	DRESS	POCKET	ICE	TEETH	DOG
9	CUP	BOWL	TEA	LAMP	PHONE	TRUCK
10	BIRD	BAT	NEST	BONE	RAIN	BRACKET
11	COW	PIG	GRASS	CHISEL	PARCEL	HOTEL
12	CROWN	HAT	KING	SHOVEL	NOSE	TENT
13	SAXOPHONE	HARP	JAZZ	SODA	HAIR	PILOT
14	WAITRESS	STEWARDESS	RESTAURANT	SWAN	BEACH	CALCIUM
15	TOOTHBRUSH	COMB	FLOSS	CAKE	CUP	GLASSES
16	TRUCK	BUS	TRAILER	CLIMATE	CACTUS	CLUB
17	BICYCLE	CAR	HELMET	FISH	BEER	BANK
18	SURGEON	BUTCHER	KIDNEY	PENGUIN	MOVIE	HOUSE
19	CHISEL	KNIFE	SCULPTURE	HAMSTER	BOTTLE	MIRROR
20	FLY	ANT	WINGS	CEREAL	BUSINESS	CONCRETE
21	CRIB	BED	BABY	FERRY	BOWL	PATIO
22	SHOE	GLOVE	FOOT	WALL	CARD	TIGER
23	CIGARETTES	ALCOHOL	LUNGS	OUTLET	SOCK	CARPET
24	MONKEY	BEAR	BANANA	AIRPLANE	HAMMER	PLUG
25	FOOTBALL	BASEBALL	QUARTERBACK	CLOUD	PLANT	NECKLACE
26	SPIDER	BEE	WEB	PEPPER	SHED	TOILET
27	RABBI	PASTOR	TEMPLE	DRIVEWAY	GLOVES	APPLE
28	HAPPY	SAD	SMILE	ROOF	SEED	KEY
29	TORTILLA	BAGEL	BEANS	COLD	KNOB	SALESMAN
30	RECEPTIONIST	HOSTESS	TELEPHONE	PARK	HAND	STRING
31	CAKE	GELATO	BAKER	BROCHURE	LAKE	SON
32	COOKIE	BISCUIT	CHOCOLATE	PAGE	WAVE	FUR
33	NEEDLE	PIN	THREAD	WAX	HYDRANT	WRIST
34	DOG	CAT	BONE	POND	HOOD	QUEEN
35	BEE	BUTTERFLY	HONEY	ASPHALT	COACH	PLIERS
36	CAPTAIN	PILOT	SHIP	EAR	BENCH	FREEZER
37	PANDA	RACOON	BAMBOO	WHIP	FENDER	LAW
38	CAMEL	ANTELOPE	DESERT	CORK	ENGINE	PAMPHLET
39	COW	BUFFALO	FARM	POTATO	LIZARD	CHALK
40	RIVER	LAKE	RAPIDS	GLASS	BUDGET	FEATHER
41	COCONUT	PINEAPPLE	BEACH	CYMBAL	SOCIETY	ROD
42	BEER	JUICE	PARTY	SHOP	SNOW	WOUND
43	ROBBERY	TREASON	BANK	STEW	TUB	SHORE
44	PENCIL	PEN	ERASER	FLUTE	MINT	SHEEP
45	CROUTONS	BAGEL	SALAD	METAL	SHARK	SPOT
46	SILVER	GOLD	BULLET	STAIRS	BALLOON	LIBRARY
47	BISCUITS	TOAST	GRAVY	SNAIL	PELICAN	DANCE
48	SNOW	RAIN	SLED	CEMETARY	WORK	NOVEL
49	CITY	VILLAGE	AIRPORT	WHALE	NECK	CABINET
50	OVEN	MICROWAVE	PAN	SCREEN	BASKETBALL	BOOT
51	FIELD	COURT	GRASS	GAS	TOAD	SCHOOL
52	PENGUIN	GOOSE	ICE	VOLCANO	HEAD	BRICK
53	BOTTLE	CAN	BABY	CLOCK	BERRY	BELL
54	COMPUTER	PHONE	MOUSE	EMPLOYEE	COUCH	SALON
55	SHAMPOO	BLEACH	SHOWER	TEAM	SAUCE	CIRCLE
56	PACKAGE	CRATE	DELIVERY	TROUT	CHILD	BILL
57	SUBMARINE	AIRPLANE	OCEAN	SHEET	CROW	DOCTOR
58	LAWNMOWER	SCISSORS	GRASS	BOMB	AUNT	INTERNET
59	POLICE	FIREMAN	HANDCUFFS	CARAVAN	CRAB	LAUNDRY

Appendix B: Experiment 2 Task Depiction



Consider this item.

Now choose the item that is most similar.

Choose the two items that are most similar.

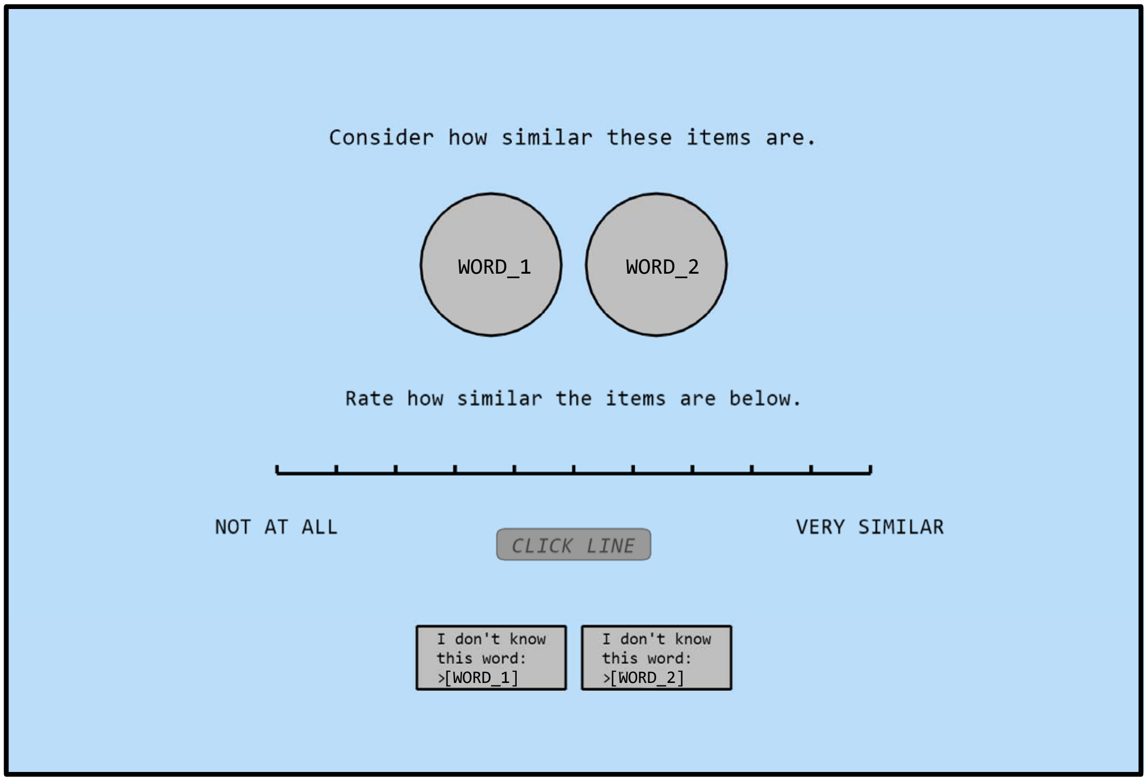
Consider this item.

Now choose the item that is most similar.

Choose the two items that are most similar.

Figure presents the four spatial configurations of the similarity judgment task in Experiment 2. Not pictured is the Standard Thematic Triad condition that featured the *Goes With* instructions and the classic triad task configuration (top left quadrant).

Appendix C: Experiment 3 Task Depiction



Consider how similar these items are.

Rate how similar the items are below.

WORD_1
WORD_2
NOT AT ALL
CLICK LINE
VERY SIMILAR
I don't know this word: >[WORD_1]
I don't know this word: >[WORD_2]

Figure presents a depiction of the similarity rating task from Experiment 3. Participants were allowed to choose any point on the rating line to provide their rating. Association rating task not pictured.

Appendix D: Concept Properties

Experiment 3 Similarity and Association Ratings

Index	Standard	Taxonomic	Thematic	Taxonomic Rating	Thematic Rating	Tax.–Unr. Rating	The.–Unr. Rating	Tax.–The. Rating Difference
1	SPOON	LADLE	CEREAL	0.545	0.727	−0.912	-0.888	-0.182
2	ROCKET	MISSILE	ASTRONAUT	0.414	0.175	−0.434	-0.224	0.238
3	GARLIC	ONION	VAMPIRE	0.121	0.392	−0.676	-0.627	-0.271
4	MILK	LEMONADE	COW	0.673	0.330	−0.840	-0.734	0.343
5	SHIP	CANOE	SAILOR	0.444	0.463	−0.882	-0.868	-0.018
6	CAR	BIKE	SEATBELT	0.435	0.326	−0.621	-0.618	0.109
7	CHAIR	SOFA	LEGS	0.362	0.182	−0.984	-0.981	0.180
8	PANTS	DRESS	POCKET	0.351	0.067	−0.521	-0.203	0.284
9	CUP	BOWL	TEA	0.254	0.748	−0.594	-0.617	-0.495
10	BIRD	BAT	NEST	0.724	0.464	−0.835	-0.873	0.260
11	COW	PIG	GRASS	0.266	0.815	−0.711	-0.576	-0.550
12	CROWN	HAT	KING	0.827	0.244	−0.826	-0.860	0.583
13	SAXOPHONE	HARP	JAZZ	0.196	0.235	−1.000	-0.942	-0.039
14	WAITRESS	STEWARDESS	RESTAURANT	0.699	0.413	−0.656	-0.745	0.287
15	TOOTHBRUSH	COMB	FLOSS	0.179	-0.059	−0.218	-0.049	0.238
16	TRUCK	BUS	TRAILER	0.663	-0.070	−0.562	-0.479	0.733
17	BICYCLE	CAR	HELMET	0.413	0.625	−0.791	-0.658	-0.211
18	SURGEON	BUTCHER	KIDNEY	0.677	0.485	−0.735	-0.474	0.192
19	CHISEL	KNIFE	SCULPTURE	0.577	0.205	−0.783	-0.802	0.372
20	FLY	ANT	WINGS	0.843	0.194	−0.868	-0.827	0.649
21	CRIB	BED	BABY	0.329	0.927	−0.952	-0.792	-0.598
22	SHOE	GLOVE	FOOT	0.393	0.193	−0.862	-0.989	0.199
23	CIGARETTES	ALCOHOL	LUNGS	0.117	0.488	−0.421	-0.492	-0.371
24	MONKEY	BEAR	BANANA	0.290	0.511	−0.765	-0.812	-0.221
25	FOOTBALL	BASEBALL	QUARTERBACK	0.267	0.222	−0.810	-0.795	0.044
26	SPIDER	BEE	WEB	0.404	0.318	−0.859	-1.002	0.086
27	RABBI	PASTOR	TEMPLE	0.154	0.077	−1.002	-0.956	0.077
28	HAPPY	SAD	SMILE	-0.457	-0.207	−0.742	-0.654	-0.250
29	TORTILLA	BAGEL	BEANS	0.444	0.439	−0.699	-0.572	0.005
30	RECEPTIONIST	HOSTESS	TELEPHONE	0.438	0.586	−0.735	-0.623	-0.148
31	CAKE	DONUT	CANDLE	0.401	0.756	−0.868	-0.800	-0.355
32	COOKIE	BISCUIT	CHOCOLATE	0.748	0.046	−0.955	-1.018	0.702
33	NEEDLE	PIN	THREAD	0.233	0.222	−0.896	-1.046	0.012
34	DOG	CAT	BONE	0.206	0.556	−0.963	-0.965	-0.350
35	BEE	BUTTERFLY	HONEY	0.425	0.498	−0.869	-0.821	-0.073
36	CAPTAIN	PILOT	SHIP	0.529	0.368	−0.974	-1.005	0.161
37	PANDA	RACOON	BAMBOO	0.380	0.693	−0.287	-0.467	-0.312
38	CAMEL	ANTELOPE	DESERT	0.520	0.509	−0.889	-0.884	0.011
39	COW	BUFFALO	FARM	0.498	0.538	−0.755	-0.669	-0.040
40	RIVER	LAKE	RAPIDS	0.580	-0.005	−0.922	-0.842	0.585
41	COCONUT	ORANGE	BEACH	0.548	0.419	−0.655	-0.707	0.129
42	BEER	JUICE	PARTY	0.797	0.164	−0.921	-0.831	0.634
43	ROBBERY	TREASON	BANK	0.121	0.634	−0.710	-0.831	-0.513
44	PENCIL	PEN	ERASER	0.595	0.413	−0.956	-1.071	0.182
45	CROUTONS	BAGEL	SALAD	0.484	0.277	−0.863	-0.850	0.206
46	SILVER	GOLD	BULLET	0.219	0.326	−0.751	-0.612	-0.107
47	BISCUITS	TOAST	GRAVY	0.374	0.345	−0.585	-0.778	0.029
48	SNOW	RAIN	SLED	0.606	0.463	−0.405	-0.390	0.142
49	CITY	VILLAGE	AIRPORT	0.342	0.248	−0.951	-1.055	0.094
50	OVEN	MICROWAVE	PAN	0.444	0.239	−0.697	-0.780	0.205
51	FIELD	COURT	GRASS	0.217	-0.250	−0.885	-0.899	0.467
52	PENGUIN	GOOSE	ICE	0.676	0.624	−0.792	-0.900	0.052
53	BOTTLE	CAN	BABY	0.485	0.810	−0.437	-0.360	-0.324
54	COMPUTER	TABLET	MOUSE	0.517	0.348	−0.712	-0.638	0.169
55	SHAMPOO	BLEACH	SHOWER	0.064	0.387	−0.713	-0.624	-0.322
56	PACKAGE	CRATE	DELIVERY	0.145	0.085	−0.617	-0.792	0.060
57	SUBMARINE	AIRPLANE	OCEAN	0.368	0.443	−0.754	-0.703	-0.075
58	LAWNMOWER	SCISSORS	GRASS	0.763	0.546	−0.946	-0.953	0.218
59	POLICE	FIREMAN	HANDCUFFS	0.264	0.632	−0.948	-1.030	-0.368