2017

# Causes and predictors of thematic intrusion on human similarity judgments

Garrett R. Honke
*Binghamton University--SUNY*, ghonke1@binghamton.edu

# CAUSES AND PREDICTORS OF THEMATIC INTRUSION ON HUMAN SIMILARITY JUDGMENTS

BY

GARRETT HONKE
BA, University of Texas at Austin, 2008
MSc, Binghamton University (SUNY), 2012

Dissertation

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Cognitive and Brain Sciences
in the Graduate School of
Binghamton University
State University of New York
2017

Accepted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Cognitive and
Brain Sciences in the Graduate School of
Binghamton University
State University of New York
2017

December 30, 2017

Kenneth J. Kurtz, Chair
Department of Psychology, Binghamton University
State University of New York

Sarah Laszlo, Faculty Advisor
Department of Psychology, Binghamton University
State University of New York

Vladimir Miskovic Member
Department of Psychology, Binghamton University
State University of New York

Daniel Mirman, Outside Examiner
Department of Psychology, University of Alabama
at Birmingham

# ABSTRACT

Most theoretical accounts of psychological similarity maintain that similarity judgments are based on shared features (and shared relations among those features, e.g., the commonalities between SPATULA and LADLE). Accounts rarely include associations between targets of comparison (e.g., the association between EGG and SPATULA) as a contributor to similarity judgments. This position is taken despite the fact that people will often choose associates over things with shared features and relations in similarity judgment tasks. So-called dual-process models—where thematic integration and feature (and relation) based comparison are component processes of perceived human similarity—have been proposed to handle this apparent failure to account for human similarity judgments. The present experiments were designed to further explore the thematic association effect on similarity with the goal to test the hypothesis that confusion about similarity and association (rather than a radical theoretical redirection, e.g., the dual-process model) is the cause of the reported thematic association influence on similarity judgments. Experiment 1 introduces a novel task for collecting similarity judgments of real world concepts—the Anti-Thematic Intrusion (ATI) task—and tests alternative task instructions as a possible driver of thematic intrusion on similarity. Experiment 2 examines the effect of the isolated components of the ATI task as compared to the classic two-alternative, forced choice similarity judgment task to determine what changes from the classic task are most influential for reducing thematic intrusion. Experiment 3 was conducted to confirm that the concept sets used in Experiments 1 and 2 did not produce biased responding. Having explored task, instruction and concept-based effects, Experiment 4 investigated behavioral and electrophysiological differences among individuals to attempt to clarify how differences between individuals correspond to similarity judgment behavior. The results were not expected in that the strength of the thematic association effect on similarity was weaker than predicted; Experiments 1, 2, and 4 show that overall association-based preferences were only present in situations strongly biased toward producing that response type. It was also found that taxonomic pair matching reliably increased across the time course of the task. Changes in the properties of the task and the instructions attenuate the effect, suggesting that the intrusion of thematic relationships on similarity judgments is not an unyielding feature of the similarity judgment process (as dual-process accounts propose) but instead (at least in part) due to interpretation of the task goal and confusion about similarity and association-based relatedness. Finally, this confusion is identifiable by less differentiation in the EEG signal elicited by these competing semantic relations, where people who produce more similarity-based responding also produce more distinctive ERP waveforms for taxonomic and thematic category members.

# DEDICATION

*For Sabina and Ben*

# ACKNOWLEDGEMENTS

Above all—thanks to Ken, who I owe an immeasurable debt of gratitude for shepherding me through the last 5 years. Thanks for always keeping the door open and thanks for being so supportive of my research interests, even when they took us to unexpected places.

Thanks to Vlad and Dan for agreeing to be on the committee. Thanks to Sarah—your hospitality and enthusiastic attitude toward my continued development have been a critical source of encouragement for me during our work and time together.

Major thanks are due to past academic advisors and friends who have helped me so much along the way—Art, Raedy, Christian, Nina, Micah, Anja and the other members of the Similarity and Cognition and Cognition and Language Labs. Thanks especially to Dedre for introducing me to the topic of this dissertation.

Thank you to all the current and past members of the Learning and Representation in Cognition Lab and Brain and Machine Labs that have been a part of this project and my life over the past 5 years. Kim, JD, Sean, Dan, Matt, I've been incredibly lucky to have such a great lab group. Special thanks go out to Sean, Dan, and JD for comments on early drafts of this thesis. Thanks also to Liz, Kate, Aira, Mavi, and the BAMlab RAs for adopting me and being willing to help with the project at every turn. Thanks to Nolan for being such a great "big" brother. Thanks to Gina for everything.

Thanks to Mom, Dad, and Kyle, Sandra and Eduardo and fam. Thanks to Wes Anderson for the entertainment and the awesome color palette featured in this thesis. A significant portion of my graduate education was funded by Kenneth J. Kurtz' IES Cognition and Student Learning Grant #R305A120554. Additional support was provided by Sarah Laszlo's Brain and Machine Laboratory. Thanks are also due to the Binghamton University–State University of New York graduate funding program.

# Contents

# List of Tables

# List of Figures

# Introduction

Higher-level cognition requires that the similarity between targets of comparison can be perceived and relied on when needed. Despite this critical role, exactly what determines the perceived similarity of real-world concepts remains a matter of debate. There is no comprehensive theoretical account that can perfectly predict human similarity judgments for real-world concepts—particularly for the case of similarity judgments in the presence of thematic association, i.e., spatiotemporal contiguity among targets of comparison (Kurtz & Gentner, 2001, in preparation). To address this prediction failure, it has been suggested that theories of similarity be extended to include thematic association as a contributing factor (Chen et al., 2013; Estes, 2003; Estes, Golonka, & Jones, 2011; Jones & Love, 2007; Simmons & Estes, 2008; Wisniewski & Bassok, 1999). This paper argues that the effect of thematic association on similarity judgments is not caused by the integration of taxonomic similarity and thematic association in similarity processes and, thus, does not call for revision of theoretical accounts of similarity.

Thematic association appears to affect perceived similarity when objects or concepts have taxonomic similarity (i.e., shared features and relations)[1] and even when the only relationship between the objects is their co-occurrence in a situation, event or action (i.e. theme). The association effect on similarity has been used as evidence for the dual-process model: the proposal that human similarity judgments result from an integration of taxonomic similarity and thematic association (Chen et al., 2013; Estes, 2003; Estes et al., 2011; Simmons & Estes, 2008; Wisniewski & Bassok, 1999). Speaking frankly, this proposed theoretical extension is a step too far. Even in light of existing criticism—where the value of similarity as a theoretical construct and predictor of human behavior has been questioned (Goodman, 1972), and especially when the respect to which things should be compared is undefined (Medin, Goldstone, & Gentner, 1993; Murphy & Medin, 1985)—the idea that theoretical similarity should be redefined to account for *perceived* similarity among things that share no features or relations conflicts with what is known about the higher-order cognitive processes that require similarity as a stable construct. The initial goal of the experiments presented here was to further explore the hypothesis that the thematic association effect on similarity is primarily driven by the triad task, one of three proposed sources of the thematic similarity effect, i.e., task constraints, stimulus properties and individual

[1]Membership in the same superordinate category and similarity in function (e.g., SCISSORS and LAWNMOWER) have also been proposed as defining criteria (Chen et al., 2014; Lin & Murphy, 2001).

biases for taxonomic or thematic information (Kalénine & Bonthoux, 2008; Mirman & Graziano, 2012; Simmons & Estes, 2008; Wisniewski & Bassok, 1999). When the reported thematic similarity bias was more difficult to produce than initially expected, however, the focus was shifted to consider why thematic similarity effects were weaker than predicted and what this result means for theoretical accounts of similarity, namely the *confusability* and *dual-process* or *integration* accounts of similarity.

## 1.1 Taxonomic Similarity and Thematic Association

The apprehension of taxonomic similarity—while difficult to predict—is required for core cognitive processes. Taxonomically-similar 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; they look alike and often play the same roles in situations (e.g., ORCA, DOLPHIN) (Goldwater, Markman, & Stilwell, 2011; A. B. 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 instances of a schema (e.g., PRESSURE and TEMPERATURE fill the same role in the *flow* schema instantiated by water transfer and heat transfer, respectively). This recognition is a powerful and necessary tool for reasoning in unfamiliar domains (Kurtz, Miao, & Gentner, 2001).

Thematic associates, generally, 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 can be a substitute for MILK in some contexts but—unlike taxonomic category members—this relationship is unidirectional. Having MILK or knowing about its properties is not helpful if you need COW. In the most useful case thematic associates have a corresponding relationship and, thus, provide one piece of information about the relational structure of the theme (e.g., the causal relationship between BOWLING BALL and BOWLING PIN). We note that this definition is not universally accepted; it has been argued that objects must have corresponding roles to qualify as thematic associates (e.g., Estes et al., 2011) and there are examples of even more restrictive definitions (for review, see Mirman, Landrigan, & Britt, 2017). These more restrictive definitions fail to represent the full complexity and variability

of thematic association.

We adopt an expansive 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); 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 externally related (Lin & Murphy, 2001). When complementary roles do exist, they exhibit a large degree of variation: they can be any *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).[2]

The critical point is that thematic associates are quite varied and do not possess the level of information provided by objects that share taxonomic category membership. Restricting the definition of thematic association 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., object manipulable associates like STAPLER and PAPER, Canessa et al., 2007) would disqualify a large cross-section of valid examples of thematic association. The concepts ARCADE and BOWLING PIN might not have readily identifiable corresponding roles, but they are valid members of the bowling theme. With these definitions in hand, we turn to the hypothesis that thematic association has a place in the similarity judgment system and, therefore, theoretical accounts of similarity.

## 1.2 Theoretical Accounts of Similarity

The guiding framework for this work is that cognitive models that propose an integration of these distinct semantic relations undervalue the importance of being able to distinguish between similarity (the property that determines the coherence of a taxonomic category) and thematic association. Membership in a taxonomic category is more informative than membership in a theme. The use of similarity for recognition and retrieval of real-world concepts is a known cognitive bottleneck (Forbus, Gentner, & Law, 1995; Halford, Wilson, & Phillips, 1998), particularly when overt physical similarities are minimal (Gentner, Rattermann, & Forbus, 1993; Holyoak & Koh, 1987). Adding thematic associates to the pool of possible retrieval candidates would compound this challenge. Therefore, the evidence must be strong to warrant a revision of existing theories of similarity in this manner.

A brief overview of existing models of similarity is needed to understand how theoretical definitions of similarity could be revised to include thematic association. Shepard theorized that similarity could be represented as the distance between entities in a

[2]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).

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 alignment, where the relational structure of entities is aligned via the comparison process and structural and featural correspondences are used to judge similarity (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983; Gentner & 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). While these theories can account for a diverse range of psychological phenomena, *they all fail to predict the effect of thematic association on similarity.* These models have no mechanism to account for the co-occurrence of concepts as a driver of similarity; they are strictly concerned with features (and sometimes relations). Co-occurrence is extrinsic, it is not a feature.

It bears repeating, these theoretical accounts cannot address the supposed effect of thematic association on human similarity judgments for real-world concepts—where things that are thematically associated (e.g., DOG and BONE) are identified as more similar than things that share more featural and structural commonalities (e.g., DOG and CAT) (Bassok & Medin, 1997; Gentner & Brem, 1999; Greenfield & Scott, 1986; Lin & Murphy, 2001; Mirman & Graziano, 2012; Simmons & Estes, 2008; Skwarchuk & Clark, 1996; Smiley & Brown, 1979; Wisniewski & Bassok, 1999).

Dual-process models—where 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 through their integration into a theme. Thus, the increase in perceived similarity 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 (though from a different perspective) for a combination of taxonomic similarity and thematic association as components of one system (Sloman, 1996, 2014). In this proposal, coherence for taxonomic and thematic categories comes from the unitary *associative* system. To be clear, we accept that concepts that share taxonomic similarity *and* thematic association (e.g., FORK, SPOON) are more related (i.e., participation of *taxonomically-similar* items in a shared theme further increases perceived similarity) as compared to taxonomically-similar but thematically unrelated concepts. It is another matter, however, to revise theoretical definitions of what it means to be similar so that thematic category members that share *no taxonomic similarity* can be construed as equally similar or more similar than taxonomic category members. There is some ambiguity as to whether proponents of the dual-process integration account take the strong view outlined here or if the interaction of taxonomic and thematic information under this account is more nuanced; see Gentner and Brem (1999) for a survey of the possible variations of this hypothesis. We therefore rely on the stated proposal that judgments of thematic associates as more similar

than taxonomic category members must be accounted for by theoretical accounts of similarity as a defining feature of the dual-process or integration perspective.


Figure 1.1: 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.

## 1.3 Thematic Integration or Thematic Intrusion?

What supports the proposal to incorporate thematic association into theories of similarity? Evidence does appear to suggest that similarity judgments are biased by information that is 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 (Figure 1.1) between a pair of concepts that share features and relations (taxonomic category members) and a pair of concepts that co-occur in a theme (thematic associates). A 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 (as explained above). 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 1.1).

Structure-Mapping Theory would predict that (2) should be rated as most similar to the standard (Gentner, 1983; Gentner & Markman, 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, where

Table 1.1: Stimulus Example from Bassok and 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 |

similarity judgment, thematic relatedness judgment, and commonality and difference listing were affected by whether the targets were taxonomic or thematic category members (Wisniewski & Bassok, 1999). Wisniewski and Bassok (1999) argue that the process recruited for these tasks depends on task constraints and the similarity of the objects themselves: when entities have commonalities, their relational structure and features are compared and a process (e.g., structural alignment) is used to produce a similarity judgment; conversely, when entities have low taxonomic similarity, the integration process is invoked. When targets are integrated, the perceived similarity of the objects increases. When targets are compared, the alignment of the targets makes their differences more salient and perceived similarity decreases. In other words, it's easy to spot the differences of similar things because they are easy to compare; different things are difficult to compare so their differences aren't 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 bias in responding to forced-choice triads 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 1.2 for a survey of task instructions). Lin and Murphy (2001) investigated ten variations of the triad task, finding thematic biases with college-aged samples in a close replication of Smiley and Brown (1979) and other triad-style tasks. The study 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.

Table 1.2: 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 & Roberson, 2004 |
| This is a [CONCEPT], find another one. | Davidoff & Roberson, 2004; Gentner & Brem, 1999 |

The literature is not without reports of taxonomic response preferences. As mentioned, the fifth graders 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 a similar age cohort. 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.[3]

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 plastic bag to children during a sorting activity increased the frequency of taxonomic responding (E. M. Markman et al., 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 as accurately as possible produced a majority of taxonomic matches in the triad task.

Despite the mixed results, a widely accepted account of responding preferences in the classic task—the 2AFC triad task with instructions only to choose the most similar match—is that people are (at the least) ambiguous responders and often they are biased toward selecting thematic matches. This responding pattern is attributed to three factors: task constraints, stimulus properties, and individual biases for taxonomic or thematic information (Kalénine & Bonthoux, 2008; Mirman & Graziano,

[3]Note: Justifying similarity judgments has not reliably produced majority taxonomic responding across the work surveyed here, e.g., Greenfield and Scott (1986).

2012; Murphy, 2001; Simmons & Estes, 2008; Wisniewski & Bassok, 1999). The central interests of the present work are (1) the unresolved question of why thematic associates affect similarity in the simplest of paradigms (the forced-choice triad) and (2) the validity of the proposal to revise the theoretical definition of similarity due to this behavior. If thematic association is not accepted as an integral component of perceived human similarity, what alternative to the dual-process integration account can explain this behavior?

## 1.4 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 reject the view, however, that this effect is grounds for including thematic association as a contributing factor in theoretical models of similarity. Similarity judgments, from simple geometric shapes to complex causal systems, depend on featural and relational commonalities because key cognitive processes (e.g., induction, inference, generalization) rely on their stability to do their work. How then can the observed behavioral effects of thematic association on similarity be explained? 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). We append to this proposal the hypothesis that this confusion occurs because taxonomic and thematic categories rely on the same machinery of categorization, where the key and defining distinction between them is the source of their category coherence. Category coherence and category member similarity are not the same (Barsalou, 1983; Conaway & Kurtz, 2017; A. B. Markman & Stilwell, 2001; Murphy & Medin, 1985). There is ample evidence that categories can carry a mix of intrinsic and extrinsic information (Barr & Caplan, 1987) and different types of categories differentially rely on this information for coherence. Therefore, a more parsimonious hypothesis for the thematic integration effect is that people simply confuse category coherence with taxonomic similarity and interpret their goal in similarity judgment tasks as "find a match that seems most coherent".

Instead of using this evidence to suggest that similarity is whatever people say it is, a more conservative 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 category coherence is called for in a situation. In other words, we accept the proposal that different tasks and objects of comparison (stimuli) elicit different processes (Wisniewski & Bassok, 1999) but reject the dual-process integration account, i.e., the proposal that these processes must both be components of the similarity judgment system (Chen et al., 2013; Estes, 2003; Estes et al., 2011; Simmons & Estes, 2008). Rather than making radical changes to the theoretical definition of similarity, it would be more parsimonious to attribute this confusion to the categorization system—known for its varying reliance on diverse sources of category coherence (Barsalou, 1983; Conaway & Kurtz, 2017; A. B. Markman & Stilwell, 2001; Murphy

& Medin, 1985).

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 into 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 information about the competing semantic relations (and sources of category coherence) and the presence of an alternative type of category coherence (i.e., 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 are designed to test the predictions of the confusability and dual-process accounts by clarifying (1) task-based, stimulus-based and individual-based determinants of similarity judgments and (2) the supposed strength of the thematic response bias. Experiment 1 presents an *Anti-Thematic Intrusion* (ATI) task designed to head off two hypothesized causes of thematic responding under similarity instructions: the prioritized positioning of the standard concept and the forced-choice aspect of the task. Experiment 2 further clarifies the role of these hypothesized task-based causes of thematic responding. Experiment 3 addresses the interpretation that the pattern of results found in Experiments 1 and 2 might be attributable to the experimental stimuli (not the manipulations of task and instructions) and returns to the results of Experiments 1 and 2 to analyze the effect of similarity and association strength (as determined by pairwise ratings) on responding preferences. Finally, Experiment 4 investigates the correspondence between similarity judgments in the classic triad task and electrophysiological responses to taxonomic and thematic category members.

# Experiment 1: Anti-Thematic Intrusion Task

## 2.1 Introduction

In Experiment 1, we set out to create and test a task that eliminates the effect of thematic intrusion on similarity judgments (Figure 2.1). This Anti-Thematic Intrusion (ATI) task departs from the classic forced-choice triad task in two ways: (1) the addition of distractors removes the two-alternative forced-choice (2AFC) aspect of the task and (2) the privileged (or prioritized) position of the standard is removed so that people must choose two items from the presented concepts.

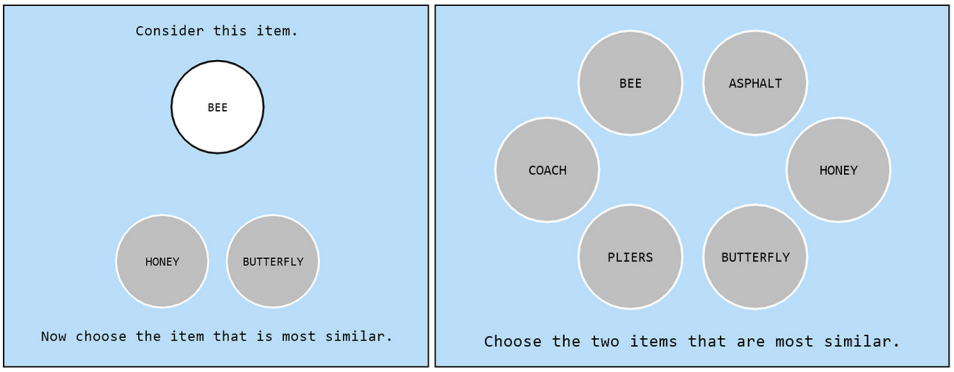


Figure 2.1: Classic Triad and Anti-Thematic Intrusion Tasks. Left: Classic 2AFC triad task with similarity instructions. Right: Depiction of the ATI task where the goal is to choose the two concepts that are most similar. No concepts are prioritized and a set of three distractor concepts are presented along with intended taxonomic and thematic matches (standard, taxonomic target, thematic target).

### 2.1.1 Task Design and Thematic Intrusion

#### Forced-choice Similarity Judgments

Why might a forced-choice decision between a taxonomic match and a thematic match cause increased thematic responding in the task? We propose that providing only a thematic and taxonomic match in a forced-choice task might implicitly suggest that both options are equally valid answers to the question of what is most similar. In

other words, the forced-choice triad task creates an implicature where people interpret the task as something more like a free-association or relatedness exercise than a task asking for judgments of taxonomic similarity. This might change the interpretation of the task goal by implicitly suggesting that the task is not to *Choose the option that is most similar* but to *Choose the option that feels most connected to you*. In fact, 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, the addition of distractors should work against the interpretation that there are two equally valid options to choose from and the task is to identify the pair that seems most *related*.

### Prioritizing the Standard

There are several ways in which the provision of a prioritized standard (i.e., a concept to be considered first before the response options are available) might increase the thematic response bias. Prioritizing the standard could increase the salience of context-dependent (CD) conceptual information at the expense of context-independent (CI) information (Barsalou, 1982). This is critical because CD information will most often be thematic or associative in nature while CI information consists of intrinsic properties (a source of taxonomic category coherence). When the standard and thematic match are considered, it is possible that CD information becomes more salient and this increases the thematic response bias.

A survey of the concept sets used in past research shows that noun phrases such as DOG and BONE (Smiley & Brown, 1979) or CHICKEN and LASAGNA (Ross & Murphy, 1999) have been included in the experimental stimuli. The co-occurrence frequency of these conventional noun phrases might bias people toward a thematic match (cf. Simmons & Estes, 2008, for a report of persistent thematic bias with frequency matched materials). Generally, thematic matches will have a higher co-occurrence frequency than taxonomic matches. This is because members of a taxonomic category are more likely to fill the same roles in a situation. While taxonomic matches will fit in the same location in a sentence frame, thematic matches will most often appear in corresponding positions and lists.

Therefore, thematic matches will more frequently conform to predictions about the next word(s) in a sequence (Kamide, Altmann, & Haywood, 2003). The increased co-occurrence frequency combined with a misinterpretation of the task objective as a free-association task could contribute to a bias to choose thematic matches. The classic triad task does not allow for presentation order counterbalancing. The standard is location invariant—switching the position of the thematic target and the standard produces an invalid taxonomic match. This is important because concept pairs might have stronger "forward" relationships than "backward" relationships (e.g., SPIDER and WEB vs. WEB and SPIDER, Jenkins & Russell, 1952; Nelson, McKinney, Gee, & Janczura, 1998). Thus, forward association strength may be part of the explanation for thematic responding in similarity judgment tasks. 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 perhaps the combination of added distractors and the removal of a prioritized standard (i.e., the ATI task) will have a different result.

### Task Instructions and Goals

Lastly, the instructions of the task are important. They exhibit considerable variability across investigations (see Table 1.2 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?

If the thematic response bias is due to a simple 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 (see below) are basically a control. They are included to confirm the baseline responding pattern under the most straightforward 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 quite 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. Lastly, the Alien condition was hypothesized to produce more taxonomic responding than the other instructions because it renders the usefulness of similarity in the task more salient. If people think their similarity judgments will be interpreted by another mature and functioning, earth-bound adult, perhaps they also assume that the receiver will understand that their response doesn't mean similar but instead "similar with respect to the fact that they occur in the same theme". The Alien instructions might make people more likely to think about why similarity (commonalities in relational structure and features) is useful—e.g., in the service of inductive reasoning—and provide responses that are most likely to support that goal for a naïve individual.

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 the semantic relations—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 at study.

### 2.1.2 Experiment 1 Design

This set of considerations produced an experiment with three conditions featuring the ATI task with distinct instructions (2.1). Due to an initial oversight, the Similar condition did not include a recurrent on-screen reminder of task instructions—this condition is included in this report because it shows the effect of not providing a reminder of the task goal. Thus, Experiment 1 features four between-subjects conditions examining the ATI task under three different sets of instructions, plus one additional condition using the Similar instructions and no consistent reminder of the task goal.

## 2.2 Method

### 2.2.1 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 (Peirce, 2007). The stimuli consisted of semantically-related concept triads adopted from previous experiments (Gentner & Brem, 1999; Hendrickson et al., 2015; Lin & Murphy, 2001; Wisniewski & Bassok, 1999) and novel triads developed for this project. In addition to the classic three-concept structure, three semantically-unrelated concepts were added to each concept triad (see Experiment 3 for norming data). The added freedom of choosing two concepts required that the taxonomic and thematic response options were 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 (all 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 was identical for all conditions. The preliminary instructions and the on-screen trial instructions varied by condition. Due to a programming error, a reminder about the task instructions intended to appear on every trial in the Similar condition (e.g., "Choose the two items that are most similar") was not presented in the experiment interface. In this case, participants read the initial instructions but were not reminded about the goal of the task for the remainder of the experiment. Including this "No Reminder" condition,

the experiment had four between-subjects conditions: three conditions with distinct instructions—Similar, Alike, and Alien (see below)—with the Similar condition featuring two sub-conditions, a sub-condition that presented instructions on every trial and one that lacked the reminder.

### 2.2.2 Procedure

Participants provided informed consent, were randomly assigned to condition, and seated at computer terminals in private testing rooms. The experiment was presented 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 instructions for the Similar condition 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.

To address the possibility that the thematic response bias observed in previous studies was due to a simple misunderstanding of the meaning of the concept *similar*, the Alike condition instructions removed any mention of the term:

> 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 condition instructions depart most from previous work. They are motivated by the idea that adults might (1) interpret the goal of the task as choosing the most related concepts overall or (2) assume the audience would understand the respect to which the thematic selection was provided and then use this judgment to determine their response. It was thought that providing a context that renders the taxonomic similarity of the concepts less mundane would increase taxonomic responding:

> 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 presentation of instructions, participants initiated the experiment and 59 randomized trials were presented. Each trial started with the presentation of a fixation cross followed by a concept array where concept placement was randomized. Participants responded by clicking the two concept buttons they judged as conforming to the task instructions (e.g., "Choose the two most similar options") and confirmed their selection by clicking the "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.

## 2.3 Results

### 2.3.1 Results Overview

Recall that the main interest in Experiment 1 was to test a novel experimental paradigm—the Anti-Thematic Intrusion task—designed to head-off several hypothesized causes of the reported thematic response bias. The frequency of each type of match is presented in Figure 2.2. The analysis was conducted in `R` (R Core Team, 2017); all data and analyses in this report are available in the supplemental materials.

Table 2.1: Experiment 1 Taxonomic Responding Pattern

| Condition | $N$ | Mean Proportion Taxonomic Responses | Taxonomic Responding Exact Binomial Test $p$ | 95% Binomial Confidence Intervals | |
|---|---|---|---|---|---|
| | | | | 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 |

### 2.3.2 General Taxonomic Responding Patterns

We follow the convention of reporting the number of participants who produced reliably biased responding and follow-up with the overall response frequency. We first note that—at the trial level—three of the four conditions produced majority taxonomic responding, where only the No Reminder condition showed the opposite pattern of majority thematic responding (see Table 2.1 for taxonomic response frequency). A two-stage binomial test procedure was used to determine the number of participants that produced 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. Trials were classified as thematic, however, when the intended thematic match was made (thematic target and standard) and

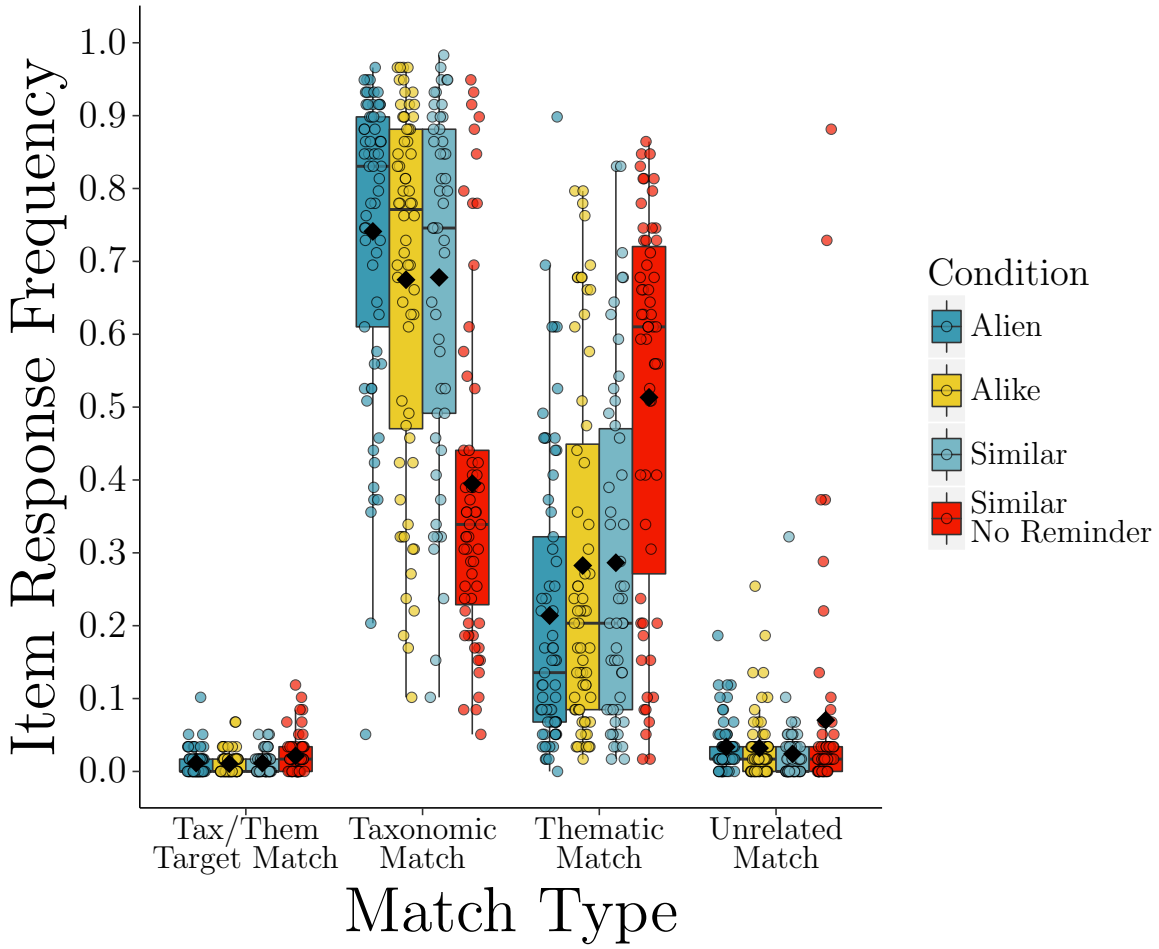


Figure 2.2: Frequency of matches by match type and condition for each participant from Experiment 1. Participants are represented by one point (positioned by condition) for each match type. Tukey's box plots show the median and interquartile range and diamonds represent the mean frequency of each type of match by condition. Taxonomic matches were made more frequently than any other match. 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% when only considering the taxonomic and thematic matches where *chance* = .5 (37–38 consistent matches out of 59 trials; $.036 < p <= .067$).

when 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 bias was calculated, these classifications were used as the DV in two-tailed binomial tests to determine if there were more (or less) people consistently responding taxonomically than what would be expected by chance. The outcome of this analysis was that every condition featured a taxonomic response bias 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. The prevalence of taxonomically biased responding found here was unexpected given past reports and it appears that consistent presentation of task instructions is critical to achieving a reliable taxonomic response bias.

### 2.3.3 Taxonomic Response Frequency and Instructions

To compare across conditions, generalized linear mixed-effects regression models (GLMER; Bates, Maechler, Bolker, & Walker, 2014) were built that predicted taxonomic responding in the ATI task 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 supplemental materials for data and code).

The model uncovered a pattern where all conditions with explicit task instructions produced more taxonomic responding than the No Reminder condition (Figure 2.3), Alien: $\hat{\beta}$ = -2.161, $SE$ = 0.28, Wald $Z$ = 7.620, $p < .001$; Alike: $\hat{\beta}$ = 1.631, $SE$ = 0.29, Wald $Z$ = 5.647, $p < .001$; Similar: $\hat{\beta}$ = 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 ($\hat{\beta}$ = 0.554, $SE$ = 0.28, Wald $Z$ = 1.963, $p$ = .0496) and the Similar condition ($\hat{\beta}$ = 0.563, $SE$ = 0.30, Wald $Z$ = 1.859, $p$ = .063), though the taxonomic responding difference between Alien and Similar only reached marginal significance. The differences between the conditions with consistent on-screen 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 with a subtle change to avoid similarity language attenuated the thematic association effect on similarity judgments, though the marginal (and near-marginal) differences between conditions make it difficult to

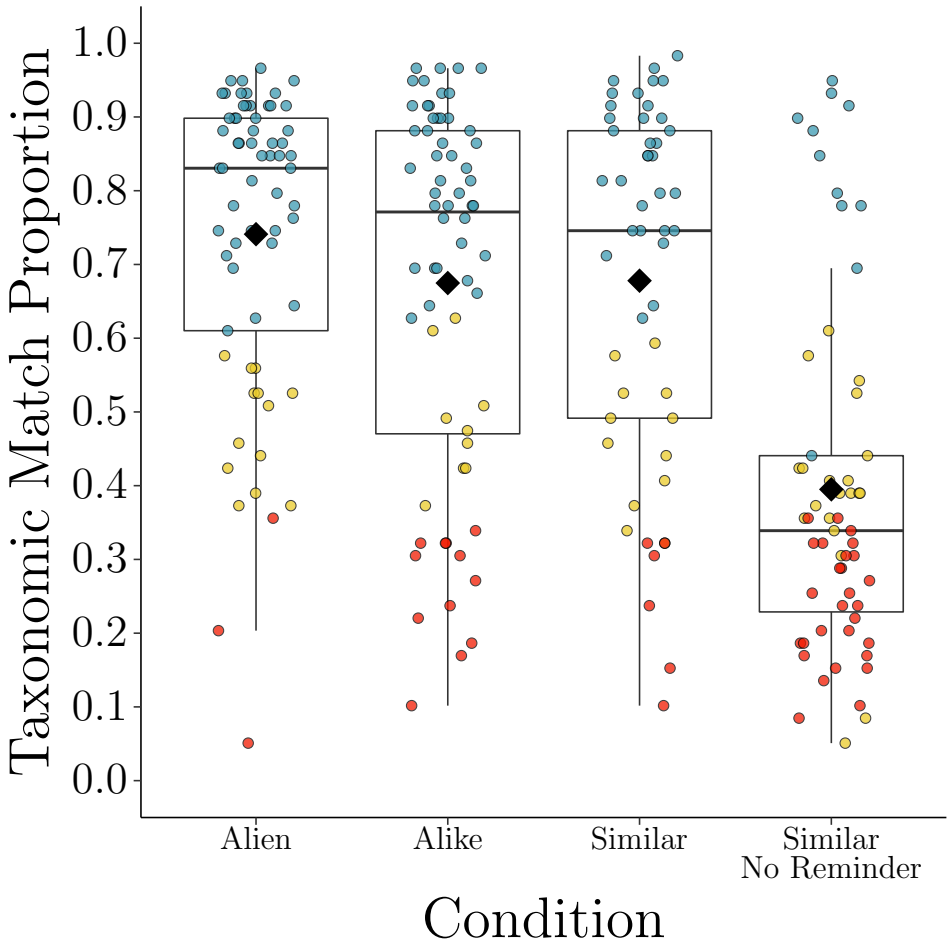


Figure 2.3: Proportion of taxonomic matches by condition for Experiment 1. Participants are represented as points, diamonds present the condition means, and Tukey's box plots present the median and interquartile range of mean taxonomic responding. Points are colored based on response bias classification. The Alien condition produced more taxonomic responding than the other conditions (this difference was marginally significant for the Alien–Alike comparison under the most conservative analysis approach).

make strong conclusions about the generalizability of this effect.

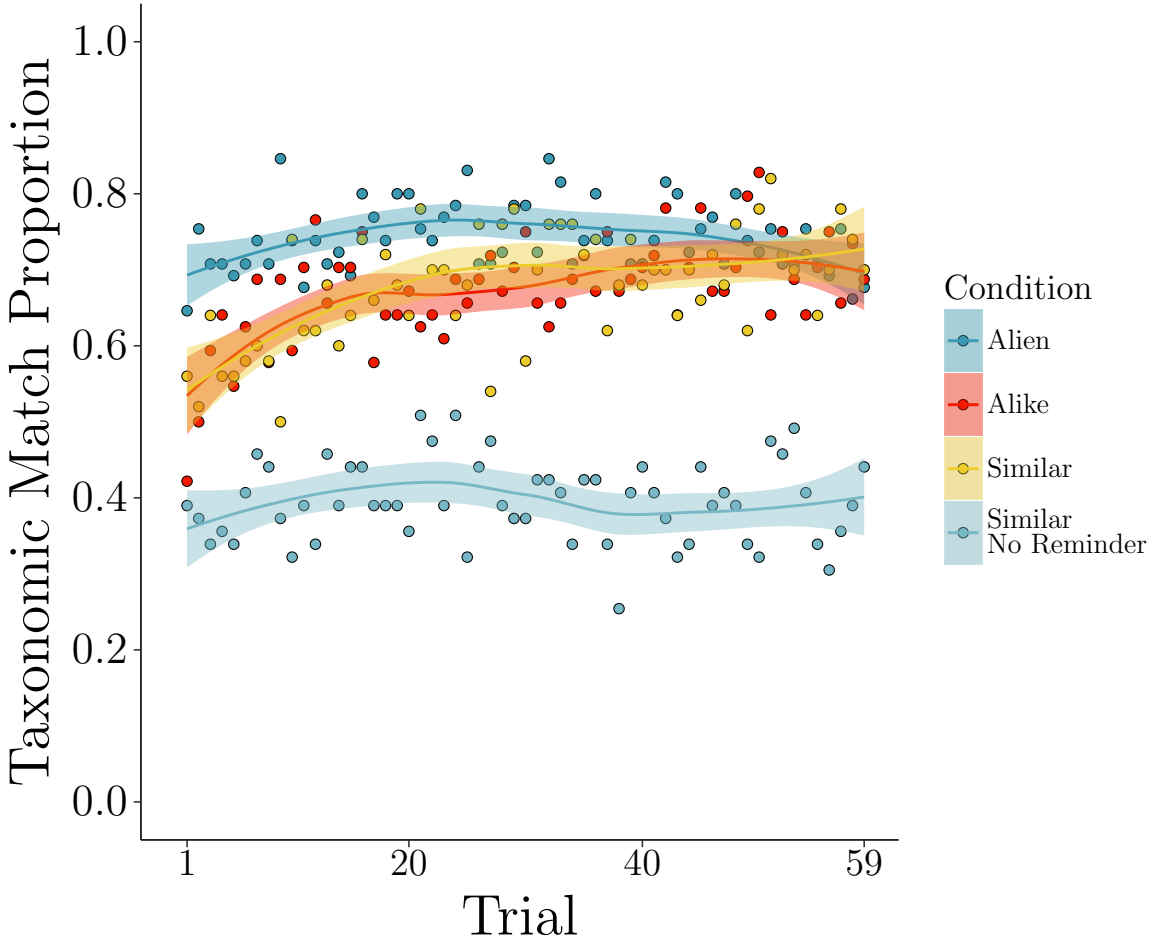


Figure 2.4: Taxonomic responding frequency across trials in Experiment 1. Points represent mean taxonomic responding by trial for each condition. Overall, taxonomic matches were more frequent as the experiment progressed.

### 2.3.4 Taxonomic Responding Across Trials

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 ($\hat{\beta} = .011$, $SE = 0.003$, Wald $Z = 4.187$, $p < .001$). Analyzing the conditions in isolation, however, 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 ($p = .34$).[1] A post-hoc

[1]Note: The conditions that produced the most and least taxonomic responding were the conditions that did not have reliable increases in taxonomic matching across trials.

explanation for the lack of a trial effect for the Alien condition could be that the instructional manipulation worked as expected; consistent with the confusion account, there was less confusion about the goal of producing taxonomic matches in this condition. The data support this idea in that there was more taxonomic responding in the Alien condition in the first 10 trials of the experiment (as compared to all other conditions with a linear model, $p$s $< .001$). This general pattern—where taxonomic responding increased as the experiment progressed—is perhaps the most interesting result of this analysis, as it is difficult to reconcile with the dual-process model (Figure 2.4). Further exploration of this issue is provided below.

### 2.3.5 Trial Response Time

Past 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 et al., 2015). Therefore, although it does not have direct bearing on the main goals of this work, we recorded trial response time in this free choice, speed-irrelevant task (i.e., no directive to focus on speeded responding was provided) and analyzed these data with LMER. First (as might be expected given the previously reported results) it should be noted that the cell count is quite different for each of the four possible matches (i.e., Taxonomic Target and Standard, Thematic Target and Standard, Taxonomic Target and Thematic Target, Match including an Unrelated Distractor) and these frequency differences should be considered when interpreting the results (see Figure 2.2). Match frequency is presented in Table 2.2.

Table 2.2: Experiment 1 Frequency of Matches and Response Time by Match Type

| Match Type | Frequency (count) | Mean of Participant Median RTs |
|---|---|---|
| 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 ($\hat{\beta}$ = -1.335, $SE$ = 0.42, $t$ = -3.198, $p$ = .002) and thematic matches were reliably faster than matches with unrelated distractors ($\hat{\beta} = 4.379, SE = 0.48, t = 9.136, p < .001$) and matches with the taxonomic and thematic targets ($\hat{\beta} = 3.298, SE = 0.54, t = 6.091, p < .001$). These effects were robust to the removal of outliers ($\pm 2.5$ $SD$). Within condition, we find

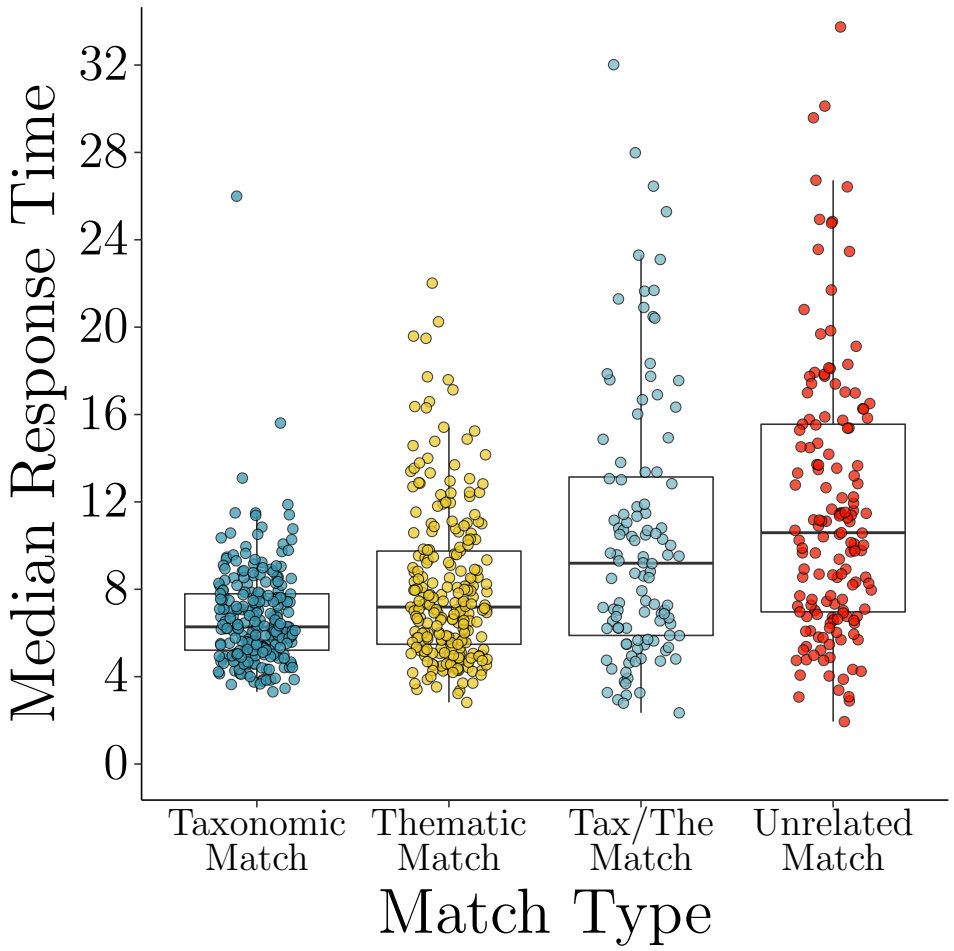


Figure 2.5: Experiment 1 median response time by match type. Median response time for each possible match type. Participant median response times are presented by points for each match type where they produced at least one match. Taxonomic matches were completed faster than all other matches.

this general trend of faster taxonomic trials for every group *except* the No Reminder group. This analysis seems to suggest that—unlike the 2AFC triad task—concepts that share taxonomic category membership "pop out" in the ATI task (see Figure 2.5). We will return to this possibility in Experiment 2 where a direct comparison of the tasks is possible.

Lastly, response time patterns appear to follow taxonomic response frequency. This is a surprising contribution that (to our knowledge) has not been explored. It seems that the response time difference between taxonomic and thematic responding tracks closely to the frequency of taxonomic matches ($Alien_{\hat{\beta}} = -2.42, p = .005$; $Alike_{\hat{\beta}} = -1.39, p = .036$; $Similar_{\hat{\beta}} = -1.78, p = .14$; $NoReminder_{\hat{\beta}} = 0.26, p = .67$). These results may provide a new framing for response time effects in this research area. It's possible that a heretofore unconsidered contributor to response time between these semantic relations is the interpretation of the task or the ambiguity of the task goal. It's possible that 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).

### 2.3.6 Summary of Results

This analysis produced several unexpected results. A majority of reliable thematic responding (and responders) was only found when the interface lacked an explicit reminder of the task goals (the No Reminder condition). We interpret this to mean that the ATI task (with different variations of similarity instructions) appears to produce more taxonomic responding than would be expected from a survey of past work in the domain. There are important similarities and differences to consider in relation to the previous work. Experiment 1 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). The studies are similar, however, in that the Alike and Similar conditions did not have reliably different taxonomic responding rates. Questions remain as to what is driving the difference between these reports. We note that the current experiment features roughly twice as many concept sets and—perhaps most interestingly—the frequency of taxonomic responding increased across the time-course of the experiment. These results coupled together suggest that a possible limitation of previous interpretations of the thematic intrusion effect is that it takes some time for people to settle in to a consistent responding pattern. Shorter experiments or those that rely on aggregation-based statistics might underestimate the prevalence of taxonomic responding. The same issue applies for previous attempts at characterizing individuals as taxonomic or thematic responders (Lin & Murphy, 2001; Simmons & Estes, 2008; Smiley & Brown, 1979).

How might the inclusion of participant and item in a mixed-effects analysis approach have affected the results reported above? A simple generalized linear model (i.e., no random effects) predicting taxonomic responding with condition and trial as fixed effects produced a different pattern of results, where the marginal differ-

ences between the Alien condition and the Alike ($\hat{\beta} = -0.322, SE = 0.05$,Wald $Z = -6.349, p < .001$) and Similar conditions ($\hat{\beta} = -0.307, SE = 0.05$,Wald $Z = -5.691, p < .001$) are reliable. So, in this case the inclusion of random effects produced more conservative results; allowing the model to calculate a random intercept for each participant accounts for variance that is attributed to the condition effect under the simple GLM. Thus, we take a conservative view of the effect of the Alien instructions on taxonomic responding—there is some evidence that the Alien instructions increased taxonomic responding, but this increase must be interpreted cautiously. This analysis is a key example of the effectiveness of individualized models of thematic intrusion over aggregation-based approaches.

Comparisons to the responding pattern found in the No Reminder condition are less ambiguous. In this condition, a factor as seemingly benign as a repeated reminder of the goal to "Choose the two most similar options" had a large effect—participants were 5 times more likely to select a taxonomic match when an instructional reminder was present (or a 5.12 log-odds ratio compared to the condition with identical instructions but no reminder). The absence of this instructional reminder completely reversed the pattern of responding observed in the other conditions. This result is difficult to reconcile with the dual-process account. We would expect the effect of thematic intrusion on similarity judgments to be more resilient if similarity judgments were produced by a composite taxonomic similarity and thematic association system.

## 2.4 Discussion

The observed increase in taxonomic responding across the experimental session, the reversal of the taxonomic response bias in the No Reminder group and the overall high frequency of taxonomic responding in the conditions with consistent presentation of instructions all lend support to the confusability account. While it is not clear why the effect of thematic integration would attenuate during the experimental session under to the dual-process model, an interpretation based on the confusability account is straightforward—people become better able to distinguish between competing semantic relations as the task proceeds. The reversal of the taxonomic responding preference in the absence of a persistent reiteration of the task goal also fits this explanation. As people get further away from the initial instructions, the goal of the task becomes less clear. Even with the reminder present, the condition that relied on *similar* with no additional clarification in instructions did not produce a reliable taxonomic bias at the participant level. This adds support to the hypothesis that interpretation of the task goal vis-a-vis *similar* is a driver of the thematic intrusion effect.

Finally, the ATI task seems to have generally increased the frequency of taxonomic responding. A survey of past research suggests that the thematic responding bias (and thus, the thematic intrusion effect on similarity) is robust, where it would not be uncommon to find thematic matches occurring most frequently. While any conclusions relying on comparison to previously published results should be interpreted cautiously, it must be acknowledged that the present results diverge from prior re-

ports at the trial *and* participant levels (Greenfield & Scott, 1986; Lin & Murphy, 2001; Simmons & Estes, 2008; Skwarchuk & Clark, 1996). Considering that this argument relies on cross-study comparison, this is perhaps the weakest conclusion in this section. Therefore, in Experiment 2 we present a systematic investigation of the components of the ATI task in relation to the classic triad paradigm.

# Experiment 2: Task Properties and Thematic Intrusion

## 3.1 Introduction

Considering the surprisingly high rates of taxonomic responding observed in Experiment 1, it is necessary to confirm that this pattern is replicable and that our materials and/or process have not confounded the results. For this reason, a Standard Thematic Triad condition featuring the classic triad paradigm with thematically-biased instructions was included in Experiment 2 to confirm that a thematic response bias could be produced under appropriate circumstances. The "goes with" version of instructions featured in previous research—*choose the item that goes best with the item above*—has been found to reliably produce thematic responding, so it was chosen as the task goal for the Standard Thematic Triad condition (Lin & Murphy, 2001; Skwarchuk & Clark, 1996). Finally, close attention was paid to the time-course of taxonomic responding across the experimental session to determine if the increase in taxonomic responding observed in Experiment 1 could be replicated in Experiment 2.

Table 3.1: Experiment 2 Conditions and Design

| Condition | Prioritized Standard | Distractors Present | Instructions |
|---|---|---|---|
| Standard Thematic Triad | YES | NO | GOES WITH |
| Standard Triad | YES | NO | SIMILAR |
| Random Triad | NO | NO | SIMILAR |
| Random Hex (ATI) | NO | YES | SIMILAR |
| Standard Hex | YES | YES | SIMILAR |

## 3.2 Method

### 3.2.1 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 3.1)—a 2 x 2 + 1 between-subjects design. The experimental materials (concept sets) were identical to those of Experiment 1.

### 3.2.2 Procedure

Participants provided informed consent and then were randomly assigned to condition and seated at computer terminals in private testing rooms. All conditions (save the Thematic Bias Condition) received the same similarity-based instructions (emphasis added to highlight the key difference):

> 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 Thematic Bias 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.

Aside from the difference in instructions between the Thematic Bias Condition and the four Similarity Conditions, each of the Similarity Conditions featured a different task and interface. The goal of these interface changes was to pin down exactly what components of the ATI Task were responsible for the observed increase in rates of taxonomic responding in Experiment 1. The interface differences are outlined in Table 3.1 and visual depictions are provided in Appendix B. For the Random Triad condition concepts were placed in random positions equidistant from the fixation point (screen center) and the other concepts. Concepts were presented in fixed locations (the apexes of the triad) for the Standard Triad and Thematic Bias Triad conditions (where the two response options were randomly placed in the left and right positions). In the Random Hex condition concepts were randomly placed in positions organized around the screen center. Concepts were presented randomly in a trapezoid for the Standard Hex condition (with the standard presented directly above). Trials were randomly ordered and presented sequentially, each following the presentation of a fixation cross.

## 3.3 Results

### 3.3.1 Results Overview

Recall that the central goal of Experiment 2 was to clarify the distinct effects of the components of the ATI Task; the experiment was designed to directly compare the effects of the presence of distractor concepts and a prioritized standard on taxonomic responding. Sub-goals were to confirm that the observed patterns of (1) an overall taxonomic response bias and (2) an increase in taxonomic responding across trials were replicable, and (3) that an overall thematic response bias could be produced with the classic triad paradigm including instructions biased toward thematic responding. We were also interested to see if the response time results from Experiment 1 (where taxonomic matches were completed faster) could be reproduced. See Figure 3.1 for the overall frequency of matches.

According to an exact binomial test analysis procedure identical to that of Experiment 1, only the Random Triad condition had enough consistent taxonomic responders to suggest a reliable preference (see Table 3.2). We note that the opposite approach—examining if less 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 Random Triad condition had enough participants exhibiting the taxonomic response bias to be reliably higher than 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 3.1).

Table 3.2: Experiment 2 Taxonomic Responding Pattern

| Condition | $N$ | Mean Proportion Taxonomic Responses | Taxonomic Responding Exact Binomial Test $p$ | 95% Binomial Confidence Intervals | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| Standard Thematic Triad | $n = 55$ | .25 | $p < .001$ | .01 | .15 |
| Standard Triad | $n = 57$ | .66 | $ns$ | .44 | .71 |
| Random Triad | $n = 57$ | .74 | $p < .001$ | .60 | .84 |
| Random Hex (ATI) | $n = 55$ | .61 | $ns$ | .42 | .70 |
| Standard Hex | $n = 62$ | .63 | $ns$ | .43 | .69 |

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 reports of reliable thematic response biases with the 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

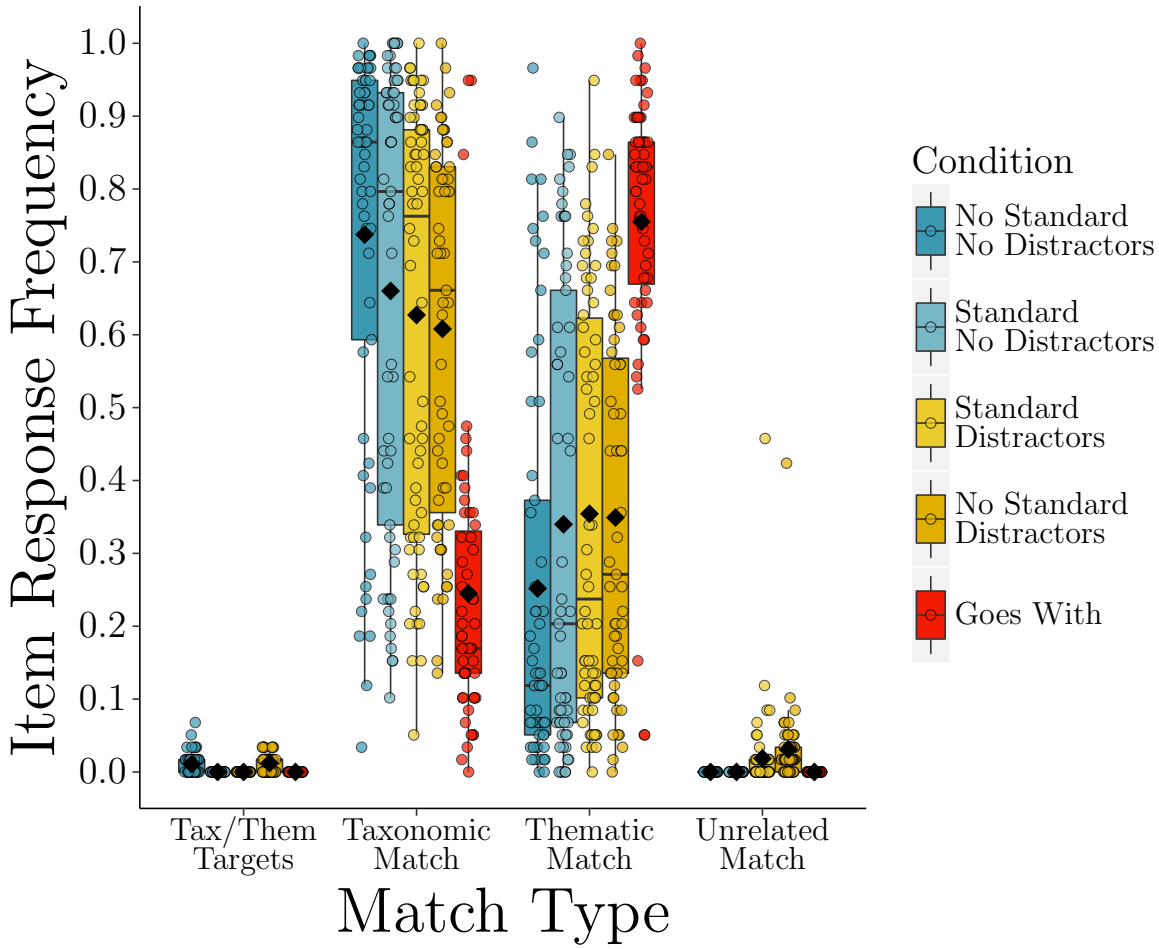


Figure 3.1: Frequency of matches by match type and condition for each participant from Experiment 2. Participants are represented by one point (positioned by condition) for each match type. Tukey's box plots show the median and interquartile range and diamonds represent the mean frequency of each type of match by condition. 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%.

the ATI task increased taxonomic responding relative to this baseline. We return to consider the implications of this result after addressing the initial analysis goals of Experiment 2.

There are two possible approaches to analyzing the effect of the ATI task on taxonomic responding: a comparison of taxonomic responding between the 5 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 thematically-biased condition removed. We begin with the latter.

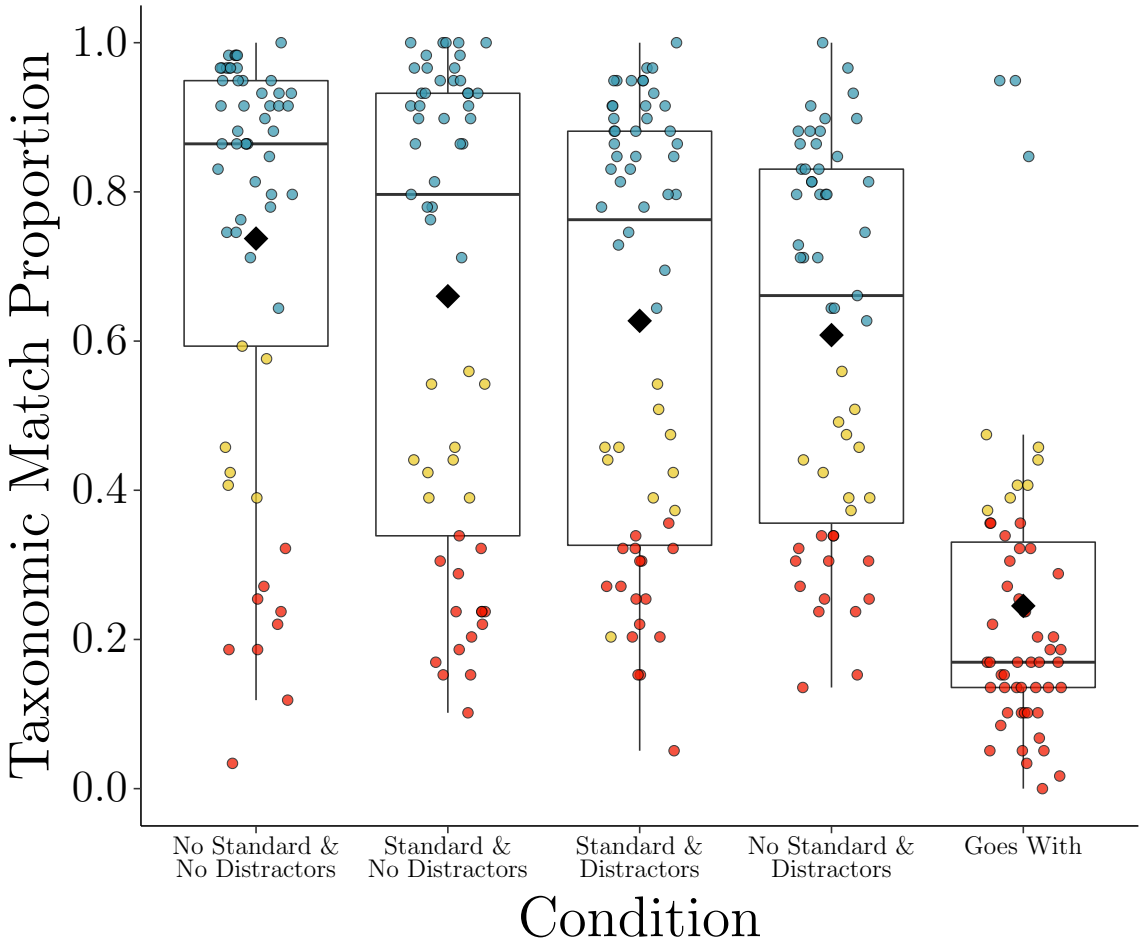


Figure 3.2: Proportion of taxonomic matches by condition for Experiment 2. Participants are represented as points, diamonds present the condition means and Tukey's box plots present the median and interquartile range of mean taxonomic responding. Points are colored based on response bias classification. The Random Triad (No Standard and No Distractors) produced the most taxonomic responding—reliably more taxonomic matches than all conditions except the Standard Triad (Standard and No Distractors) condition. Conditions with similarity-based instructions produced more taxonomic responding than the Standard Thematic Triad (Goes With Instructions) condition.

### 3.3.2 ATI Component Analysis

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 ($\hat{\beta} = -.0733, SE = 0.27$,Wald $Z = -2.705, p = .007$) and trial ($\hat{\beta} = 0.025, SE = 0.003$,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 was reliable) (see Figure 3.2).

In other words, contrary to what was hypothesized, presentation of distractors produced *less frequent* taxonomic responding and standard prioritization had no effect (see Table 3.2). Replicating the trial-effect results from Experiment 1, taxonomic responding increased in frequency as the experiment progressed. Note that the Standard Thematic Triad condition was not included in this analysis as it was designed to produce thematic responding—the addition of an instructions factor (two levels: taxonomic-biased, thematic-biased) was considered, but the resulting groups were deemed to be too unequal in terms of $n$ (224 participants with similarity instructions vs. 62 with "goes with" instructions); due to this imbalance, exploratory models including the thematically-biased condition failed to converge.

### 3.3.3 Condition Analysis

The condition-based analysis is different from the task component analysis 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 3.1). The random effects structure was the same across analyses and trial was maintained as a fixed-effect predictor. The broad pattern of results is as follows: all conditions with similarity-based instructions produced more taxonomic responding than the Standard Thematic Triad ("goes with" instructions), the Random Triad condition had a higher taxonomic response rate than every condition except the Standard Triad condition.

To restate, the Random Triad condition produced reliably more taxonomic responding than all conditions except the Standard Triad condition ($p = .29$); Random Triad vs. Random Hex, $\hat{\beta} = 1.064, SE = 0.37$, Wald $Z = 2.868, p = .004$; vs. Standard Hex, $\hat{\beta} = 0.819, SE = 0.36$, Wald $Z = 2.263, p = .024$; vs. Standard Thematic Triad, $\hat{\beta} = 3.322, SE = 0.38$, 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 ($\hat{\beta} = 2.927, SE = 0.38$, Wald $Z = 7.711, p < .001$) and marginally higher than the Random Hex condition ($\hat{\beta} = 0.669, SE = 0.37$, Wald $Z = 1.807, p = .071$). Both the Standard Hex ($\hat{\beta} = 2.503, SE = 0.37$, Wald $Z = 6.797, p < .001$) and Random Hex ($\hat{\beta} = 2.257, SE = 0.38$, Wald $Z = 5.953, p < .001$) conditions produced more taxonomic responding than the Standard Thematic Triad group. We note that the condition-based models presented here occasionally fail to

converge and require additional iterations. 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 number; the results (i.e., parameter estimates) are also consistent across a set of suggested optimizers.

To close the task-level analysis section, we present simple generalized linear models (GLM) of the effects of the key ATI task components. Note that these models are less conservative than the mixed-effects models presented above, as they do not include participant, concept, or trial level random intercepts or slopes. Given the marginal results of the Experiment 1 GLMER, however, it seems important to try to characterize the variance that the mixed effects approaches are accounting for—especially because the convergence failures above might be due to lack of statistical power adequate to model the participant and item-level variance. A GLM model was built to test the hypothesis that standard prioritization, distractor presentation and the interaction suggested by Figure 3.2 would be reliable predictors of taxonomic responding without random participant and item effects (taxonomic responding predicted by the fixed effects of trial, distractor presentation, standard prioritization and the distractor by standard interaction, excluding the Standard Thematic Triad condition). This is exactly what was found. The model produced a reliable interaction between standard prioritization and distractor presentation, where the absence of a prioritized standard and distractors (the Random Triad condition) produced the highest levels of taxonomic responding ($\hat{\beta} = 0.453, SE = 0.07$, Wald $Z = 6.19, p < .001$). As for the other fixed effects, distractor presentation produced less taxonomic responding ($\hat{\beta} = -0.15, SE = 0.05$, Wald $Z = -2.89, p < .001$) and removal of standard prioritization produced more taxonomic responding ($\hat{\beta} = 0.37, SE = 0.05$, Wald $Z = 6.92, p < .001$) overall. It appears that having a prioritized standard produces more taxonomic responding when distractors are present and less taxonomic responding when they are absent. We return to consider this interaction in the general discussion.

### 3.3.4 Time-Course Analysis

Responses were more likely to be taxonomic matches as the experiment progressed (Figure 3.3), regardless of whether the condition-based ($\hat{\beta} = 0.020, SE = 0.003$, Wald $Z = 7.733, p < .001$) or task component-based ($\hat{\beta} = 0.025, SE = 0.003$, Wald $Z = 8.027, p < .001$) analysis approach is considered (see Figure 3.3). 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 (Section 2.3.4). Replication of the trial effect from Experiment 1 strengthens the evidence for the idea that people seem to produce more taxonomic responding as they work through the randomly-ordered concept sets. It is difficult to reconcile this result with the dual-process integration account, where thematic association should affect similarity judgments equally across the experiment. Instead, something is happening that shifts the weighting of taxonomic and thematic information. At the very least,

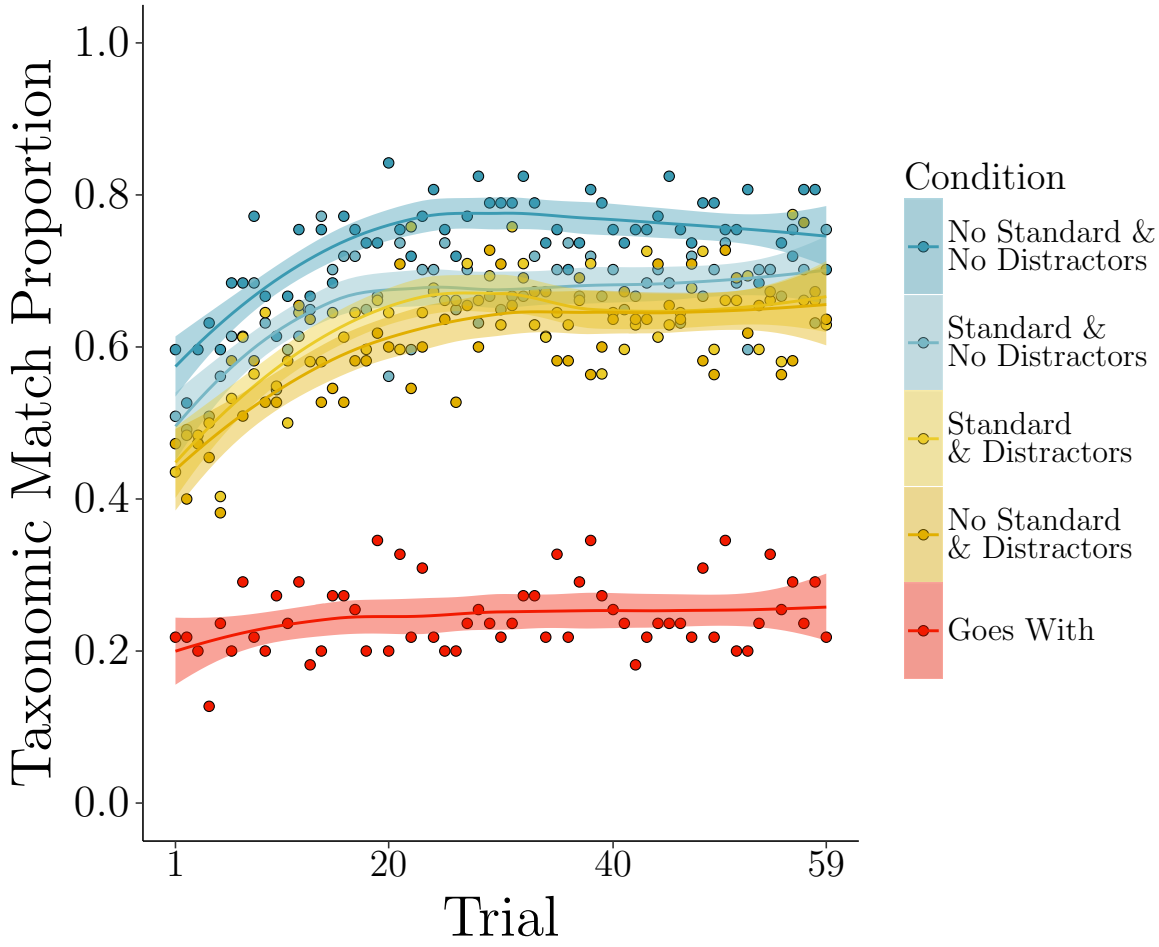


Figure 3.3: Taxonomic responding frequency across trials in Experiment 2. Points represent mean taxonomic responding by trial for each condition. Taxonomic responding increased in frequency across the experimental session for all conditions except the Standard Thematic Triad (Goes with) condition.

the variability observed here suggests that claims about response preference stability across time might need to be tempered (especially at timescales larger than the individual sessions where these effects have been observed).

Table 3.3: Experiment 2 Frequency of Matches and Response Time by Match Type

| Match Type | Frequency (count) | Mean of Participant Median RTs |
|---|---|---|
| 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 |

### 3.3.5 Trial Response Time

As in Experiment 1, trial duration was recorded and analyzed (Table 3.3). 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 ($\hat{\beta} = -0.901, SE = 0.31, t = -2.887, p = .004$) and thematic matches were faster than trials that included unrelated concepts ($\hat{\beta} = 4.496, SE = 0.57, t = 7.822, p < .001$) and the taxonomic and thematic targets ($\hat{\beta} = 1.976, SE = 0.58, t = 3.221, p = .001$) 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 condition featuring the ATI task, i.e., the Random Hex condition, no standard with distractors ($p = .44$). 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 the thematic pair, $\hat{\beta} = 0.80, SE = 0.18$, Wald $Z = 4.44, p < .001$. As in Experiment 1, these results are purely exploratory. Nevertheless, the overall pattern suggests that some conditions produce more competition between the taxonomic and thematic matches than others. The notable null results in the similarity-based instructions conditions with no standard prioritization and distractors could suggest that the task goal was less clear—these are the similarity-biased conditions that featured the least frequent taxonomic responding.

## 3.4 Discussion

While (1) the overall bias toward taxonomic matches and (2) the increase in the frequency of taxonomic matches across the time-course of the experiment were replicated, Experiment 2 has produced as many questions as answers. A surprisingly high

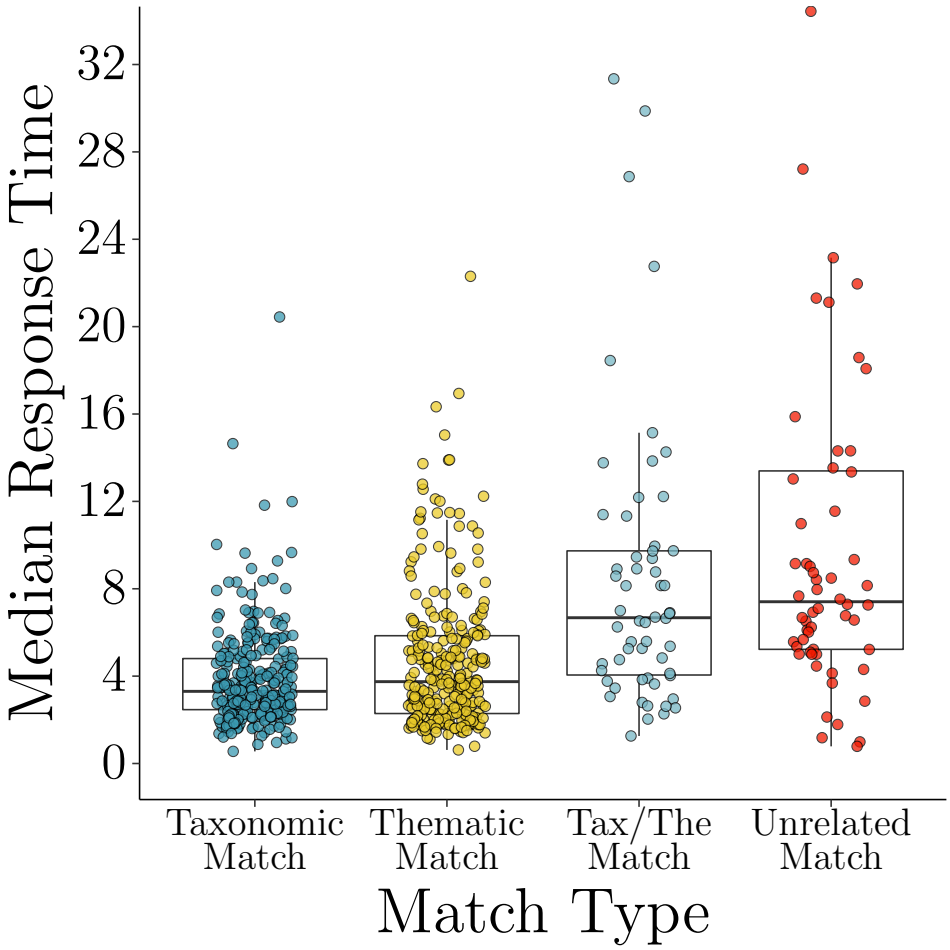


Figure 3.4: Experiment 2 median response time by match type. Median response time for each possible match type. Participant median response times are presented by points for each match type where they produced at least one match. Trials with taxonomic matches were completed faster than trials with other matches. Thematic trials were completed faster than trials with unrelated matches and matches with the taxonomic and thematic targets.

rate of taxonomic responding was found in the standard triad condition, where—based on past reports—it was expected that ambiguous or thematically-biased responding would be found. 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 Random Triad condition. The contribution of the 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 Random Triad condition (no distractors and no prioritized standard) produced the most taxonomic responding observed; 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; as is, the evidence is not strong enough to make more concrete conclusions about an interaction effect.

Moving on to the response time analysis, the results provide further evidence for the counterintuitive claim 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 uncovered 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 the key driver of response time for the different possible matches is the type of semantic relation that is consciously being searched for. Under this account, past reports of the speed advantage for thematic matches in unconstrained, non-deadline tasks are due 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) the frequency of taxonomic responding observed here appears to be higher than what has been previously reported. Recall that there are three hypothesized drivers of thematic intrusion on similarity judgments: task constraints, concept properties, and individual differences. Having explored the effect of instruction and task manipulation, we turn to address the effect of variability across concepts.

# Experiment 3: Concept Properties and Thematic Intrusion

## 4.1 Introduction

The previous experiments have uncovered an unexpected pattern; all conditions with similarity-based instructions and consistent task instructions have produced 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 often majority thematic) results presented in the past reports. At this point it must be asked—is the unexpected pattern discovered here due to some artifact in the experimental design or materials? Experiment 2 featured a direct replication of the classic triad task where many of the concept sets were adopted from previous projects. Yet, we did not find an overall thematic responding bias. Therefore—despite the inclusion of concept set variance as a random effect in our analyses—the possibility that the concept sets created for this investigation are responsible for driving this effect must be considered. This issue is the motivation for Experiment 3.

In Experiment 3, ratings were collected to determine the perceived relatedness (similarity and association) of the taxonomic and thematic relations in the materials used for this report. In a between-subjects design, people were asked to rate the similarity of pairs of concepts that share taxonomic similarity, thematic association and no relationship or asked to rate how well the same pairs of concepts *go together*. No instructions were provided to explain what was meant by "go together" or "similarity". This is an intentional omission—it would be undesirable for the behavior produced in Experiments 1 and 2 to be subject to confusability and thematic intrusion while the Experiment 3 ratings were not. Competing semantic relations (i.e., taxonomic *and* thematic pairs) from the same set were not included in the presented concept pairs for a given person; it was thought that the presentation of both pairs from one set might hint at the key distinction at study 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 semantic relationships and purpose of the study would not be recognized. Ad hoc post-task interviews were used to confirm that participants were naïve to the systematic relationships at study.

There are two goals for Experiment 3. First, it's necessary to confirm that

the concept sets used for Experiments 1 and 2 were not biased toward taxonomic responding—the simplest explanation of the observed taxonomic responding rates in Experiments 1 and 2. The ideal result for this analysis is that the intended taxonomic pairs have higher similarity ratings than association ratings and the intended thematic pairs are 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 or more related. A sub-goal for this analysis is to confirm that the unrelated distractors were indeed unrelated—rated lower on taxonomic similarity and thematic association than the semantically related matches. As for the second major goal of Experiment 3—it is possible that the rating data can provide further insight into the results of Experiments 1 and 2, particularly the effect of increased taxonomic responding across trials. For example, (while it may be unlikely) could it be that the increase across trials is due to an unfortunate failure of randomization, where 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.

## 4.2 Method

### 4.2.1 Participants and Materials

Participants ($N = 202$) were recruited, compensated and consented in an identical fashion as the previous experiments. They were randomly assigned to the Taxonomic Similarity condition or the Thematic Association condition—the difference being the question used to solicit ratings. The experiment was administered on PCs with Psychopy (Peirce, 2007). The concept sets were the same as the previous experiments. For each concept set, a taxonomic or thematic match and two unrelated concepts (reduced from the possible 3 pairwise comparisons to minimize the number of trials) 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, or 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. The taxonomic rating group was asked to consider and rate the similarity of the items. In the thematic group, people were asked to consider and rate how well the items go together (a depiction of the task interface is provided in Appendix C). No instructions were provided to guide participants about what is meant by similar or what it means to go together. The preliminary task instructions were the same across conditions:

> Hello! In this experiment you are going to be rating a series of different pairs of items (words). For each pair, your goal is to carefully consider

the items and rate them based on the question that is presented on the screen.

So, for each new pair of words you will read the words and the accompanying question and then provide your answer on the rating line.

Press any key to see an example of your task.

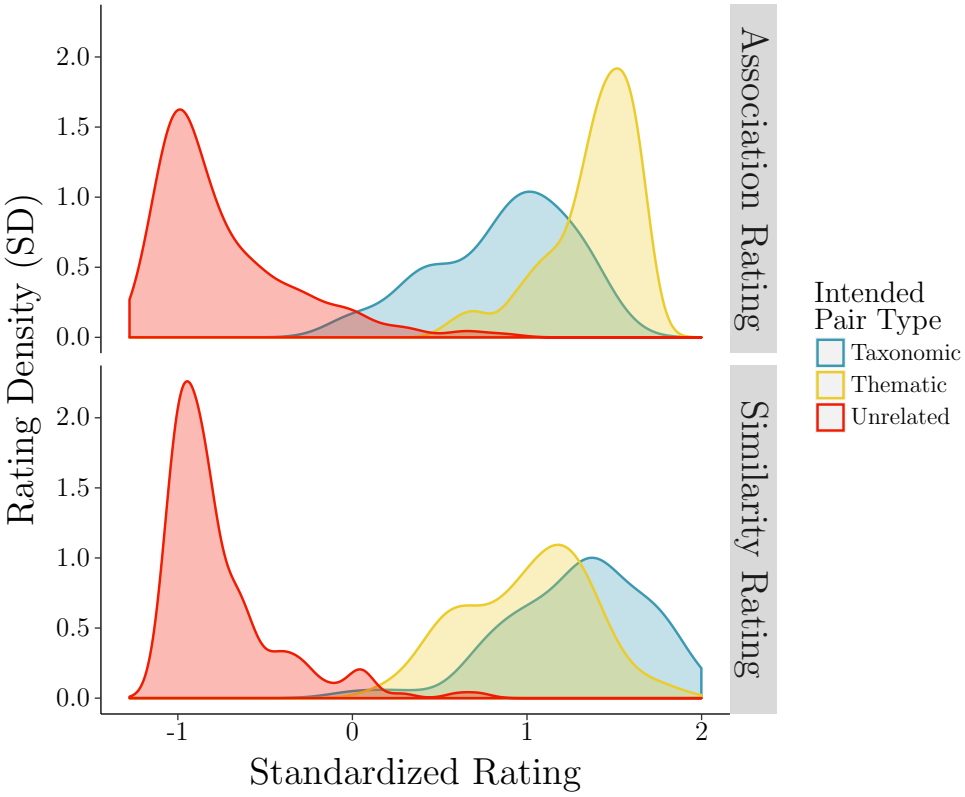


Figure 4.1: Density plot of standardized ratings for the association (top) and similarity (bottom) rating tasks. Taxonomic pairs were rated as more similar, thematic pairs were rated as more associated, and unrelated pairs were rated lowest on similarity and association. Taxonomic and thematic pairs in the same concept set were not reliably different in the magnitude of their standardized similarity and association ratings (respectively).

### 4.2.2 Procedure

Participants were seated in private testing rooms, presented with on-screen instructions and began the experiment when they were ready. Pair-wise 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 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. The ratings were collected as ratio-scale values ranging from 0 to 100.

## 4.3 Results and Discussion

There are two main analysis questions for Experiment 3: (1) 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) use the rating data to gain insight into the results of Experiments 1 and 2.

### 4.3.1 Similarity and Association Ratings

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 association rating (Table 4.1); 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 (thematic pairs rated higher on similarity than thematic relatedness: HAPPY, SMILE; FLOSS, TOOTHBRUSH; RIVER, RAPIDS; TRAILER, TRUCK; FIELD, GRASS). The aggregated results show that taxonomic pairs were rated as more similar than the thematic pairs ($\hat{\beta} = -7.489, SE = 0.55, t = -13.6, p < .001$) and unrelated pairs ($\hat{\beta} = -54.90, SE = 0.47, t = -115.70, p < .001$) 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 ($\hat{\beta} = -15.64, SE = 0.56, t = -27.88, p < .001$) and the unrelated pairs ($\hat{\beta} = -65.81, SE = 0.49, t = -135.44, p < .001$) in an LMER model predicting thematic ratings with an identical predictor structure (Figure 4.1). Perhaps most importantly, similarity scores for taxonomic pairs ($z$-scores of similarity ratings subtracted by $z$-scores of association ratings for each taxonomic pair) were not reliably different from association scores for thematic pairs ($z$-scores of association ratings subtracted by the corresponding similarity ratings for each thematic pair) from the same concept set ($M_{Difference} = 0.047$ $SD$) according to a paired $t$-test, $t(58) = 1.117, p = .27$ (see Figure 4.2).

Table 4.1: Experiment 3 Concept Ratings

| Pair Type | Similarity Rating Mean (SD) | Association Rating Mean (SD) | Similarity Rating Mean Response Time | Association 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 |

### 4.3.2 Taxonomic Responding and Concept Ratings

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 (Section 2.3.3). 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 (Section 3.3.2). Critically, 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. The interpretation that the trial effect could be explained by variation among concepts sets is not supported by these results.

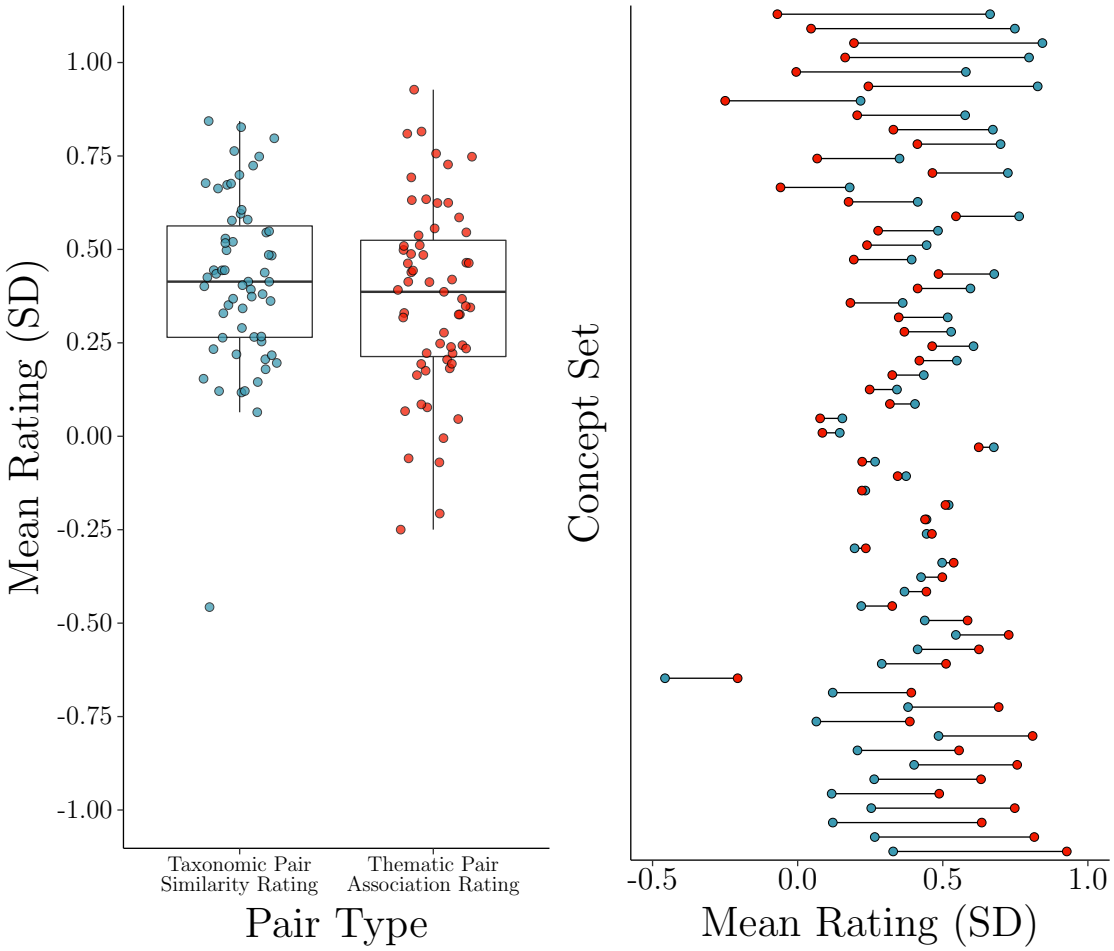


Figure 4.2: Visualization of the concept ratings overall (left) and paired with the match from the same concept set (right). The left panel depicts the mean similarity and association ratings for the taxonomic and thematic pairs, respectively. The right panel depicts the paired difference of the similarity (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).

# Experiment 4: Electrophysiological Markers of Thematic Intrusion

## 5.1 Introduction

Experimental inquiries into the task and concept-based determinants of thematic intrusion on human similarity judgment show that individual responding preferences can persevere in even the most biasing of circumstances. Similarity judgment tasks with unambiguous instructions increase the frequency of taxonomic matching. Providing a standard for comparison increases taxonomic matching when distractors are present (but it has the opposite effect when they are absent). The characteristics of the concepts themselves (as measured by association and similarity ratings) also have predictive value for similarity judgments. However, these factors alone do not perfectly predict similarity judgment behavior. Holdouts can be found in every sample; there are always people who produce the opposite responding pattern in situations that bias most of the sample to produce consistent taxonomic or thematic responding.

While the experiments presented here suggest that the thematic response bias is not as prevalent as previously reported, the goal to eliminate thematic intrusion on similarity judgment through task manipulation might not be possible. Instead, it might be more fruitful to try to predict when thematic intrusion will occur and who will be most susceptible to its effects. If Experiments 1–3 have shown how task and concept-based properties can predict thematic intrusion, a critical component that has not been addressed is individual variation in preference or ability to identify and distinguish between taxonomic and thematic category members for the purposes of judging similarity, as previously explored by Mirman and Graziano (2012), Murphy (2001), Simmons and Estes (2008), and others. The goal of Experiment 4 is to further clarify the role of this variation in similarity judgments by looking at online processing of these semantic relations under completely unbiased conditions and connecting this processing to behavioral response patterns from the classic forced-choice, taxonomic–thematic conflict triad task. Is thematic matching in the triad task the result of confusion about the difference between taxonomic and thematic category members (e.g., Gentner & Brem, 1999)? Or, is this behavior a result of a system that integrates thematic and taxonomic information to produce similarity judgments (e.g., Bassok & Medin, 1997; Chen et al., 2013; Simmons & Estes, 2008? We hypothesize that an examination of the processing of these semantic relations in a completely unbiased

situation can help to tease apart these competing hypotheses.

In this study, we collected event-related potentials (ERPs) elicited by the passive observation of semantically related and unrelated wordforms and analyzed them in relation to overt similarity judgments of the same concepts in the classic 2AFC triad task. The idea was to examine the processing of taxonomic and thematic category members away from the effects of task design and instructions and then investigate how performance in the triad task—a task shown to produce both taxonomically and thematically-biased responding—is related to the unbiased ERP waveforms. No previous work has attempted to link ERP waveforms and similarity judgments while maintaining a purely unbiased EEG recording procedure with no intervening behavioral tasks. No previous work has looked at the relationship between ERPs and overt similarity judgments for the purpose of characterizing divergent, individualized activation and decision patterns—patterns that may be useful in differentiating between the confusability and dual-process accounts. These theoretical and methodological changes increase the likelihood that heretofore undetected ERP differences between taxonomic and thematic category members can be discovered and used to support or refute existing theoretical accounts of thematic intrusion on human similarity judgment.

### 5.1.1 Characterizing ERPs Elicited by Taxonomic and Thematic Relations.

ERP research in this domain has generally fallen short of the goal of discovering differences between ERP waveforms elicited by taxonomic and thematic category members. Thus, ERP research has failed to address the key issue that has been raised by this and other behavioral investigations: What causes people to choose more or less taxonomically-similar concepts in similarity judgment tasks? Are these behavioral differences consistent across different tasks or are they an artifact of the "match-to-sample" tasks used to collect similarity judgments? Are there neural (i.e., electrophysiological) differences that predict behavioral response biases? Can these differences be used to provide a window into why—outside of the influence of concepts, task instructions, and design—people produce different responding patterns in similarity judgment tasks?

While existing work has inadequately addressed these questions, success has certainly been found in clarifying the general ERPology (i.e., the character and form of ERPs elicited by certain stimuli, see Luck, 2014, pg. 5) of the processing of these semantic relations, particularly in relation to semantically unrelated concepts. In one such study, Chen et al. (2013) recorded ERPs while people performed a similarity or difference judgment task for a sequence of taxonomic and thematic category pairs. The analysis uncovered a reliable difference in the amplitude of the P600 component elicited by taxonomic and thematic category members—a larger (more positive) P600 for taxonomic pairs.[1] The authors argue that this P600 difference is evidence of "less

[1]It is interesting to note in relation to the behavioral data presented in Experiments 1–3, Chen et al. (2013) report no differences in similarity ratings, difference ratings, or reaction time between

syntactic flow" in the processing of taxonomic relations (Chen et al., 2013). Another study from Chen and colleagues (Chen et al., 2014) collected ERPs in a sequential concept priming experiment administered in conjunction with a lexical decision task. Study participants viewed taxonomic and thematic category pairs while indicating if the stimuli were words or non-words with a button press. The study uncovered a reduced frontal negativity effect for productive thematic associations (e.g., BEE and HONEY) as compared to hierarchical relations (taxonomic category members) and other subcategories of semantic relations not relevant for this work. In other words, more facilitative priming (evidenced by reduced negative frontal activation in the 400–550 ms time window) was found for thematic associates as compared to taxonomic category members. Note the apparent exploratory nature of these reports (particularly the spatial specificity of the conclusion, and the differences in analysis approach and results as compared to the study above).

Work by Wamain, Pluciennicka, and Kalénine (2015) had more success in uncovering ERP differences between these semantic relations at time points where semantic effects would be expected. The authors found ERP waveform differences between pictorial depictions of thematic associates and two specific sub-types of taxonomic category members (taxonomic category members that share a specific function or a general function, e.g., SAW–AXE vs. SAW–KNIFE) at short inter-stimulus intervals (66 ms). The task was to observe visual depictions of semantically related concepts and vocally name the pairs after EEG collection was finished for the trial. One difficulty in interpreting this finding is that it's possible that the ISI in this condition was too short for the semantic processing of the first stimulus in the pair to finish. Waveforms from the second stimulus presentation for each semantic pair (presented 366 ms after the first stimulus) are not distinguishable from the waveforms of the first stimulus in the pair. Thus, it is difficult to say whether or not these differences are due to priming or late processing of the first concept in the 400-600 ms time window.

Maguire and colleagues (Maguire, Brier, & Ferree, 2010) also contribute to this effort with an ERP and ERSP (event-related spectral perturbation) based design paired with a passive listening task. The authors found a distinction in the distribution of the power of certain frequencies across the scalp: more alpha power was found over the parietal areas of the brain for taxonomic category members and more theta power was found over the right frontal areas of the brain for thematic category members. The authors suggest that this increase in parietal alpha power is due to the fact that it requires additional attentional resources to process taxonomic category members—a conclusion that dovetails with the idea that (1) processing taxonomic similarity requires an effortful comparison process (e.g., Kurtz et al., 2001), (2) processing taxonomic similarity is more difficult than processing thematic association (Sachs, Weis, Krings, Huber, & Kircher, 2008) and (3) less-educated (Denney, 1974; Sharp et al., 1979; cf. Mirman & Graziano, 2012) and less "intelligent" people (Simmons & Estes, 2008) experience more thematic intrusion on similarity judgments.

taxonomic and thematic category members. Given that these differences have been reliable in other work, this may suggest a limitation of the generalizablility of this research.

### 5.1.2 Theoretical and Methodological Advances in the Present Work

A frequent goal of the studies in this area (including those reviewed above) has been to investigate possible taxonomic–thematic differences in the N400 component. The ideal N400 effect for these inquires would be a difference in *facilitative priming* between taxonomic, thematic, and unrelated word pairs as evidenced by diverging ERP waveforms roughly 300–400 ms post stimulus exposure (Kutas & Federmeier, 2011). This facilitative priming would be considered a measure of semantic connection strength. Couching this phenomenon in terms of the dual-process model, Chen et al. (2013) argue that similar levels of facilitative priming (minding the issue of endorsing $H_0$) are evidence of the integrated nature of association and comparison processes for similarity judgments. The N400 time window is not a bad place to look; it makes sense that a general behavioral preference for perceiving more similarity in a certain type of semantic match might be the result of stronger priming for that match (eliciting increased positivity in the N400 component for the match as well). Unfortunately, we know of no successes in this effort and several failures to find distinctive N400 patterns between taxonomic and thematic category members in healthy adults (Chen et al., 2013, 2014; Hagoort, Brown, & Swaab, 1996; Khateb et al., 2003; Maguire et al., 2010)—notable exceptions being the work of Wamain et al. (2015) and Hagoort et al. (1996), though for the latter, the difference was only found in comparison to right hemisphere damaged patients. There is strong evidence that the processing of taxonomic and thematic category members occurs in different systems or networks (Schwartz et al., 2011), so why do ERP approaches fail to detect differences? Or stated differently, given the apparent difficulty in finding differences between taxonomic and thematic category processing, why continue to use the ERP framework to study these semantic relations?

Methodological and theoretical adjustment could address several of the issues raised here. A common design feature of past investigations has been that the entire sample was treated the same (often analyzed using factorial analyses, e.g., ANOVA). In other words, ERPs elicited from taxonomic and thematic category members were analyzed in the aggregate. Could it be that averaging over the entire sample hides important differences in the processing of these semantic relations? Along those lines, in some cases we have shown (Experiments 1 and 2) that behavioral data analyzed with a factorial approach at the group level is anti-conservative. Whether the results are obscured by aggregation or the outcomes are anti-conservative, it seems like a good idea to use an individualized experimental design to study this individual-driven phenomenon.

#### Individualized ERPs for Individual-based Similarity Responding Patterns

Our hypothesis is that analyses that average across participants obscure important differences—people who exhibit strong taxonomic or thematic response biases might work against the calculation of a mean amplitude ERP outcome variable. Consider that the most likely manifestation of behavioral biases—if they could be detected

through ERPs—would be more facilitative priming (i.e., increased N400 positivity) for a specific type of semantic relationship. In this scenario, averaging across a sample of people who have reliable but opposite biases would cancel out differences—thematic responders would show increased facilitative priming for thematic category members, taxonomic responders would show increased facilitative priming for taxonomic category members, and these differences would not be preserved in a measure of average ERP amplitude.

Conversely, consider the hypothesis that people who are more susceptible to thematic intrusion produce less distinct ERP differences between these semantic relations—these people are included in aggregation-based approaches as well. Additionally, stimuli that have been well-normed should be expected to elicit different N400 activation patterns in an adequately-powered experiment simply by virtue of being different classes of semantic relations. In this case, differences found in the aggregate wouldn't be saying anything more than *different things produce different waveforms*. For these reasons, the present work focuses more closely on individual differences by classifying participants based on their similarity judgment behavior and then using this classification to look at ERP differences across groups.

### Effects of Intervening Tasks on ERPs and Other Methodological Concerns

There are several methodological adjustments (relative to the work surveyed here) that can increase the likelihood that real differences in the ERPs will be detected. First, previous studies have often included intervening tasks directly or indirectly related to the question(s) at study during EEG recording (e.g., lexical decision, similarity judgments, difference judgments, button pressing). Intervening tasks have large effects on the EEG signal (Luck, 2014), particularly those that require a physical response or covert decision. In addition to the biasing effects of task properties (see Experiments 1 and 2), the signal elicited by these responses cannot be distinguished from the underlying processes at study. The result is ERP data confounded by the intervening task. Similar to Maguire et al. (2010), the present design features passive EEG collection with no explicit task instructions or behavioral task related to the processing of the semantic relations at study. Instead, participants are asked to identify pseudowords when they appear in the stimulus stream. Thus, measures of semantic relation processing do not include response potentials (trials with responses are removed from analysis) and the task is simply to respond if the letter string is not recognized as a word. This effectively eliminates the risk of signal contamination from the evoked response potential while ensuring that focus is maintained on the stimulus stream.

Additionally, concepts will be presented with long enough ISIs (3–3.5 seconds) that ERP waveforms can be reliably attributed to the most recently presented stimulus and its semantic relationship with the preceding concept (i.e., removed from the processing of the preceding concept itself). Results will be presented and analyzed without averaging across electrode sites, as this type of averaging carries the risk of obscuring real effects and producing anomalous patterns (Thigpen, Kappenman, & Keil, 2017). Lastly, confirmatory data analysis will be restricted to the a priori hy-

potheses presented below—hypotheses that only relate to ERP amplitude differences in the established time window for semantic effects (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980).

#### Breadth of Taxonomic and Thematic Category Members

The types of thematic and taxonomic relations used in previous investigations have been too restrictive to make class-wide conclusions. This is not a problem for the particular studies we have outlined here, i.e., it is reasonable to investigate specific types of taxonomic (e.g., function specific taxonomic categories, Wamain et al., 2015) or thematic (e.g., productive relations, Chen et al., 2014) categories if the research interest is in those specific sub-types. It is a different matter, however, to extend the results of these investigations to taxonomic or thematic categories in general. Therefore, in this work we adopt an expansive definition where thematic category members only require temporal contiguity in an established situation and taxonomic category members are entities of the same *kind*, i.e., entities that share membership in a category of natural kinds or artifacts that can be characterized by a common set of shared features and relational structure (Kurtz & Gentner, 2001; Lin & Murphy, 2001; Mirman et al., 2017).

### 5.1.3 The Current Study

A key goal of this research is to find evidence that supports or refutes the two competing theoretical explanations of the effect of thematic intrusion on similarity judgments—the confusability and dual-process accounts. The broad hypothesis here is that facilitative priming differences between distinct semantic relationships are difficult to detect when collapsing across an entire sample. Instead, what if different behavioral patterns are due to different levels of facilitative priming for semantic relations? Or, what if similarity judgment differences are due to difficulty distinguishing between semantic types at the individual level, i.e., less distinctive EEG activation patterns between types of semantic relations? Looking for answers for these questions by averaging across an entire sample would fail if elicited ERP waveforms have a direct correspondence with similarity judgment behavior. The present study uses a novel experimental design to match concept similarity judgments with ERP waveforms elicited during the passive processing of those same taxonomic and thematic category members. This procedure has the potential to uncover presently unknown properties of taxonomic and thematic processing and how these properties might relate to confusability about the distinction between similarity and association or the integration of these distinct sources of semantic relatedness and category coherence for similarity judgment.

## Toward Characterizing Individual Differences in Taxonomic and Thematic Thinking

The general approach of linking similarity judgments to measures of individual differences such as education (Denney, 1974; Sharp et al., 1979), the Need for Cognition (NFC) scale (Cacioppo & Petty, 1982; Simmons & Estes, 2008), and online processing (Mirman & Graziano, 2012) has had success in uncovering differences between people with different profiles of similarity judgment behavior. Mirman and Graziano (2012) used the visual world eye-tracking paradigm to investigate processing time-course and competition between taxonomic and thematic category members. They found that more competition between taxonomic and thematic category members predicted taxonomic responding in the the triad task. Given this link between on-line concept processing and triad task behavior, a set of measures that assess language and reading ability were included in the current experiment. Not only are these measures (exposure to print, verbal fluency, and vocabulary) effective controls for general education and language exposure variance, but they are also important for similarity judgment behavior itself.

*Role of Reading Experience and Language Exposure.* The recognition of authors and magazines has been shown to predict orthographic knowledge and experience even when controlling for other measures of general aptitude (e.g, SAT scores) and domain knowledge (West & Stanovich, 1991). Vocabulary knowledge has a direct relationship with semantic priming. In children, words that are less well-known elicit stronger thematic priming than taxonomic priming. The opposite pattern is found for words that children can define and use correctly in a sentence (Ince & Christman, 2002). The relationship between verbal fluency and semantic relation processing is less clear. On one hand, the categories in our assessment (particularly fruits and animals) are superordinate taxonomic categories, so ease of recall of category members could be a measure of taxonomic processing ability. On the other hand, many people are successful in the task by using a free association clustering strategy (Jenkins & Russell, 1952)—like using a biome-based organization, for example, when naming living things (e.g., using the savanna biome to produce lion, elephant, antelope, rhino, zebra, etc.) or a color scheme organization to list colors (e.g., ruby, sapphire, topaz). However verbal fluency relates to the processing of taxonomic and thematic relations, the measure is predicted to help account for variance in the design that would otherwise be attributed to random error or taxonomic responding in the triad task.

*Individual Differences and Similarity Judgments.* Sharp et al. (1979) showed that education is related to taxonomic responding. Simmons and Estes (2008) found that triad task responding patterns related to NFC scores, where lower scorers produced more thematic matches. Mirman and Graziano (2012) did not find demographic differences (i.e., education, age) to be predictive of triad responding behavior. At the least, we hypothesize that including these specific reading and language exposure assessments will allow us to disentangle the contribution of these factors and similarity judgment behavior in the analysis of ERPs elicited from taxonomic and thematic category members. The outcome of these assessments was analyzed in relation to similarity judgment behavior in addition to being included in the analysis of the ERP

data.

*Choosing an Appropriate Task for Collecting Similarity Judgments.* The similarity-based task instructions found to be most ambiguous in Experiment 1 were deliberately chosen for the similarity judgment phase of Experiment 4. It is convenient for comparison to past work that these instructions coupled with the classic triad task are also the most frequently used way to assess people's similarity judgment behavior. They were also desirable because they produce a varied spread of the possible response biases (as seen in Experiment 2). It would be counterproductive to use a task like the Random Triad (no standard and no distractors) or an instructions set like the *Alien* or *Goes With* instructions because these conditions produce responding that is heavily biased toward a particular response type (see Sections 3.3.2, 2.3.3 and 3.3.3, respectively). The idea is to use a task that has the least biasing conditions to maximize the diversity of response patterns found and sample roughly equal groups of participants for the ERP comparison. Consistent with our hypothesis and data that suggests that the classic triad task is interpreted most ambiguously, it is the best task for the purposes of collecting similarity judgments in Experiment 4.

### Teasing Apart the Predictions of the dual-process and Confusability Accounts

Mixed results and methodological issues currently limit understanding of taxonomic and thematic category member processing and the ERP waveforms they elicit—particularly for the goal of teasing apart the predictions of the confusability and dual-process explanations of thematic intrusion on similarity judgments. Are there general differences in the N400 components elicited by taxonomic and thematic category members? This has been the focal question of past research. Here we change the focus to how differences in similarity judgment behavior might correspond to differences in electrophysiology at the individual level.

What do the confusability and dual-process accounts predict about ERP waveforms elicited by taxonomic and thematic pairs and corresponding similarity judgments? Chen et al. (2013) suggest that the dual-process model finds support from evidence that N400s elicited by taxonomic and thematic pairs are not reliably different. Under this view, a similarity judgment process that integrates taxonomic and thematic information should produce similar ERP waveforms—particularly for the semantically sensitive N400 component. Again, the failure to find N400 effects between taxonomic and thematic category members is presented as support for the dual-process account (Chen et al., 2013). In the present research, a failure to find ERP differences in the key semantic time window between individuals who exhibit different response patterns would also support this argument. According to this account, semantically related pairs are experienced, integration and comparison processes are engaged, their outputs are integrated, and this procedure produces a general similarity judgment that is not qualitatively different across the experience of different types of semantic relations.

In contrast, the confusability account suggests two possible alternative hypotheses. The first is that more similarity in the ERP signals between semantic relations makes

it harder to differentiate between similarity and association-based category coherence; this makes it harder to produce similarity-based responding that is unaffected by thematic intrusion. People with less differentiated ERPs might be more ambiguous with respect to their responding preference (not reliably choosing taxonomic or thematic matches consistently). In contrast, people with more differentiated ERPs might be better able to distinguish between the competing semantic relations and thus be less subject to the effect of thematic intrusion. This possibility directly relates to Gentner and Brem's argument that the similarity process is derailed when people have difficulty distinguishing between the mental output of similarity and association-based processing (Gentner & Brem, 1999).

On the other hand, a combination of differentiation and facilitative priming could be a marker of perceived similarity and responding behavior. If people are simply substituting the answer to a hard question (e.g., What commonalities—features, roles and relations—do these entities share?) with the answer to an easier question (e.g., What feels more related? What shows up together most often? What word occurs next most frequently?), it should be expected that reliable matching of a particular type in the triad task will correspond to more facilitative priming for the favored semantic relation. Recall the argument that it is more difficult to make taxonomic similarity judgments as compared to thematic association judgments (Maguire et al., 2010). For thematic matching, all that is needed is confirmation that the entities show up together. For taxonomic matching, the effortful engagement of the comparison process is needed.

To sum, the dual-process account predicts that similarity judgments are derived from a combination of comparison and thematic integration. A failure to find differences in the amplitude of N400s elicited by taxonomic and thematic pairs supports the idea that similarity judgments are the result of an integrative dual-process mechanism. The confusability account makes no prediction about the overall pattern of differences between taxonomic and thematic category members. Rather, it suggests that some people are more susceptible to thematic intrusion than others. Evidence for this susceptibility would be a correspondence between less differentiated ERP waveforms and ambiguous or thematically-biased similarity judgments. The two sub-hypotheses—differentiation vs. differentiation and facilitation—suggest different sources of confusability, where the cause of confusion is (1) the inability to easily differentiate between association and similarity or (2) the ability to differentiate and increased fluidity of processing for a particular class of semantic relation.

## 5.2 Method

### 5.2.1 Participants

Undergraduate students ($N = 61$) from Binghamton University were recruited from the Psychology Department pool ($n = 53$) or the university community ($n = 8$) and participated for credit toward the completion of a course requirement or $30.00 cash compensation, respectively (36 female; $\text{Age}_{\bar{X}} = 19.0$, $\text{Age}_{Range} = 17$–$23$). Three par-

ticipants were dropped due to experimenter error during the EEG collection phase. Three participants were missing data from part of the procedure; the demographics survey, the demographics survey and verbal fluency assessment, and exposure to print assessment, respectively. Where needed, these missing values were imputed with the `mi` package (Su, Gelman, Hill, & Yajima, 2011) in `R` (R Core Team, 2017). In the analysis below, this resulted in a total of 58 participants: 56 participants with complete data and two participants with imputed values for the assessments mentioned above. The study was approved by the Internal Review Board of Binghamton University. Participants identified themselves as right-handed, monolingual English speakers with little-to-no early life exposure to any other language, normal or corrected-to-normal vision and no history of psychiatric or neurological disorders. Participants who reported recent alcohol, prescription or recreational drug use that could affect their performance were asked to reschedule the experiment.

### 5.2.2 Materials

#### Reading and Language Exposure Assessment

Three measures of reading and language exposure were collected prior to the EEG recording phase of the experiment. *Exposure to print* was assessed with a 160 item questionnaire consisting of real and fake authors and magazine titles following from the work of Stanovich and West (1989). The task was to indicate which items in the questionnaire were real while avoiding the fake items. $d'$ values were calculated for each participant to account for individual differences in orthographic processing skill. Verbal fluency was assessed with a category member naming task where the goal was to name as many examples of a given category (fruit, colors, animals) in 60 seconds. The third assessment was a vocabulary test consisting of 30 items drawn from the Verbal Reasoning section of the Graduate Records Examination (GRE) test. The concepts used in the experiment are well below the reading level of a college-aged sample, but nevertheless it is hypothesized that this measure will help to account for the differences among participants in vocabulary ability.

#### Concept Set Generation and Presentation Order

Concept sets ($N = 100$) were created that consisted of a standard, a taxonomic match, a thematic match, and two unrelated concepts. Ideally, the same concept sets would have been used across Experiments 1–4. The concept sets from Experiments 1–3, however, included some concepts more than once (i.e., standards and taxonomic and thematic targets were often used as unrelated distractors in other concept sets). Concept repetition would have produced a confound in the EEG data due to N400 repetition effects (Deacon, Dynowska, Ritter, & Grose-Fifer, 2004; Laszlo & Federmeier, 2011; Rugg & Nagy, 1989). Therefore, new concept sets were developed using the concept sets from Experiments 1–3 as a starting point. These concept sets were normed as follows. Similarity and association ratings (collected in the same manner as Experiment 3), mean concreteness ratings (Brysbaert, Warriner, & Kuperman,

2014), and age of acquisition data (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) were visualized and examined for outliers. The 20 worst outliers in terms of concreteness, age of acquisition, and difference in similarity and association ratings (i.e., relatedness strength) were removed. This exclusion process resulted in 80 concept sets (see Tables 5.1 and 5.2 for aggregated concept set properties, comprehensive data provided in Appendix F).

Pseudowords generated from the orthographic and lexical characteristics of the experimental stimuli (i.e., frequency, length, orthographic neighborhood size, and constrained bigram frequency) were paired with concept sets in an iterative procedure that minimized the cost (difference) between the properties of the possible pseudoword matches (string length, orthographic neighborhood size, and bigram frequency) and the mean of those same properties in the real-word concept sets across 10,000 iterations of possible pseudoword–concept set combinations (pseudowords and lexical and orthographic statistics were generated from MCWord, Medler & Binder, 2005). The purpose of this process was to make sure that the pseudowords were as word-like and similar to their paired concept set as possible. Closely matching pseudowords were expected to increase the difficulty of the pseudoword identification task and thus increase attention to the word stream in the EEG recording phase.

During the EEG recording phase of the experiment, four categories of concept pairs were presented with Psychtoolbox (Brainard & Vision, 1997) in a continuous stream of letter strings. Each letter string could be preceded by a member of the same taxonomic category, a member of the same thematic category, an unrelated concept, or a pseudoword (see Figure 5.1). Four counter-balanced presentation orders were produced that followed three considerations: randomization of concept presentation within each set, randomization of concept set presentation across the EEG phase, and randomization of presentation of the taxonomic category member or thematic category member within each set. Again, the latter consideration was required because the standard could not be presented twice in the course of EEG recording due to the possible confound of N400 repetition effects for words and non-words (Laszlo & Federmeier, 2011; Rugg & Nagy, 1989).

Two randomized presentation orders were produced to satisfy the first and second considerations, where concept set order, concept order within set and taxonomic or thematic pair selection was randomly determined. Two additional orders were produced by replacing the randomly selected taxonomic or thematic matches with their alternatives from the same set to satisfy the third consideration; this process produced two sets of two randomly ordered presentation orders, four orders in total. Randomly placing the concept sets into a single stream of words and pseudowords carried the risk that unintended relationships might be produced between adjacent words. This issue was resolved within concept sets by soliciting similarity and association ratings from a separate sample of participants (as in Experiment 3, results below). Between-set correspondences were handled differently. Each counter-balanced presentation order was examined independently by a team of research assistants to confirm that concepts at the boundaries between concept sets did not have incidental taxonomic or thematic relationships. When relationships were identified (independent of how weak they were perceived to be) the presentation order was altered to break up these

incidental pairs.

**Similarity Judgment Triad Task**

In the final phase of the experiment, the semantically-related pairs from the EEG phase (the standard, taxonomic match and thematic match from each set) were presented as forced choice triads with Psychopy (Peirce, 2007). The task was identical to the standard similarity-based triad task from Experiment 2, with the only task difference being that the new Experiment 4 concept sets were used. On each trial, a standard was presented first in a prioritized position followed by a taxonomic category member and a thematic category member (randomly placed at the left and right apexes of the triad below the standard). On-screen instructions directed participants to: *Consider this item* [the standard] *Now choose the item that is most similar.* A depiction of the task is provided in Appendix B in the top left quadrant of Figure B.1. Final responses, response time and all other behavior was recorded.

### 5.2.3 EEG Recording and Processing

EEG data collection closely modeled the procedures used for previous research in the Brain and Machine Laboratory (e.g., Laszlo & Sacchi, 2015; Sacchi & Laszlo, 2016). An elastic EasyCap with 26 geodesically arranged[2], passive amplification, ring-sintered Ag/AgCl electrodes (inter-electrode impedances maintained below 2 kΩ, see Laszlo, Ruiz-Blondet, Khalifian, Chu, & Jin, 2014) was used to record the EEG signal. Two electrodes on the outer canthi of the left and right eyes and one electrode on the suborbital ridge of the left eye were used to record the electrooculogram (EOG) and monitor blinks. The EEG and EOG were referenced to the the left mastoid online; offline the EEG and EOG were re-referenced to the average of the left and right mastoids, the vertical and horizontal EOGs were re-referenced as a singular bipolar channel. The signal was recorded with a Brain Vision Brain Amp DC amplifier (low pass filtered at 250 Hz, high pass filtered with a 10 s time constant, sampled at 500 Hz with an A/D resolution of 16 bits).

A two-stage, offline artifact rejection procedure was applied to each participant's data. First, EEG data for each participant was filtered with a high-pass filter (0.05 Hz), ICA components were computed and components corresponding to blinks were visually identified and removed. Second, the EEG record was visually inspected with a participant-individualized amplitude threshold to identify and remove artifacts less well-identified by ICA (e.g., blocking, drift, horizontal eye movements, etc.). Exclusion criteria were as follows. Participants were candidates for exclusion from the analysis if less than 60% of all trials or less than 60% of a particular concept pair type were retained after the artifact rejection procedure (no participants met these criteria). An average of 89% of trials were retained per concept pair type (minimum number of trials retained across concept pair types for a single participant: 70%). The

[2]Geodesic placement refers to the equidistant positioning of electrodes on an approximately spherical surface—this arrangement differs from the 10–20 system that does not feature equidistant placement.

EEG record was binned into concept pair specific ERPs time-locked to stimulus onset with a 100 ms pre-stimulus baseline and a 998 ms post stimulus recording period. A band-pass filter of 0.1–20 Hz was applied to the ERPs for presentation purposes only (e.g., Figures 5.7 and 5.8).

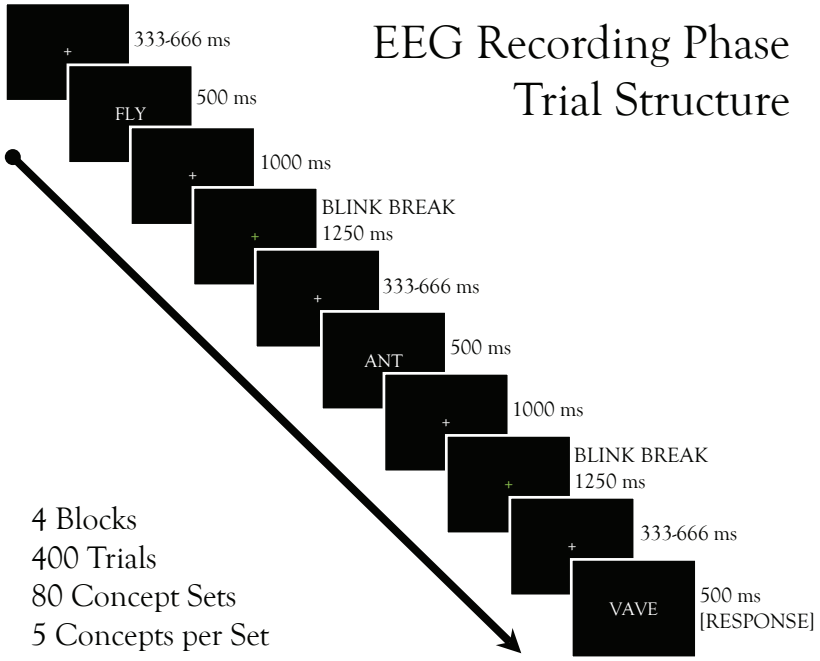


Figure 5.1: Visual depiction of the trial structure for the EEG recording phase of Experiment 4. The task goal was to observe a continuous stream of concepts and respond by pressing a button when a pseudoword appeared in the stream.

### 5.2.4 Procedure

Participants entered the lab and were provided with a verbal description of the complete experimental procedure. After attaining informed consent, the demographic survey and reading and language exposure assessments were administered and participants were fitted with the EEG cap. EEG recording occurred in a sound attenuated booth[3]. Stimuli were presented at a distance of 75 cm on 24 inch computer monitors displaying at a resolution of 1920 x 1080. Demonstrations of the EEG record and the task were provided before the start of EEG collection to (1) illustrate the importance of reducing eye and body movement during EEG collection and (1) orient participants to the pseudoword identification task. Participants were instructed to maintain control of their eye and body movements and press a button as fast as possible when the image presented on the screen contained a string of letters that was not a word. This Go/NoGo task was used to confirm that participants attended to the presented

[3]A subset of the sample ($n = 17$) completed the experiment in private testing rooms (not sound attenuated booths) due to lab construction.

stimuli. The task was designed to be unrelated to the semantic relationships of interest to avoid the introduction of evoked response potentials into the EEG data of the critical trials (semantically-related and unrelated real word pairs). Concepts were presented in a continuous stream broken into four blocks that followed one of four randomly generated and assigned counter-balanced presentation orders. Breaks were provided in between blocks (after approximately 100 trials); the task resumed when participants indicated that they were ready to start the next block.

Each trial started with a 333–666 ms fixation cross presentation that was randomly jittered to avoid anticipatory processing. Stimuli (images of letter strings) were presented for 500 ms followed by a 1000 ms post-stimulus fixation cross and a 1250 ms blink break. The next trial began immediately after the blink break terminated.

After the EEG recording was complete, the EEG cap was removed and participants were allowed as much time as needed to clean up before the triad similarity judgment task was started. The triad task was administered on computer and self-paced in an identical procedure to that of Experiments 1 and 2.

### 5.2.5 Statistical Methods

The analyses were conducted with linear mixed-effects regression (LMER: Bates et al., 2014; Kuznetsova, Brockhoff, & Christensen, 2015) models built in `R` (R Core Team, 2017) to predict ERP amplitude with semantic pair type, word properties, concept association and similarity ratings, participant reading and language experience, similarity judgment behavior, and random effects for participant, time window and concept. Critically, the use of LMER does not require the aggregation of data across participants like factorial analysis approaches; this makes it particularly valuable for the analysis of individual differences. Mean amplitude was examined with 20 ms averaged time points constrained a priori to the time window where the N400 (300–400 ms) component is most likely to be found (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). Consistent with prior research, unaveraged EEG data collected at central, parietal and occipital electrode sites (MiCe, MiPa, LDPa, RDPa, LMOc, RMOc, LLOc, RLOc, MiOc) were used to capture the broadly distributed N400 effect. A minimal ("parsimonious") random effects structure was used due to the overall size and complexity of the models—this procedure is not subject to the maxim (and general critique) to *keep it maximal*, as specifying the maximal random effects structure was not expected to significantly affect parameter estimation in this situation (See Stites & Laszlo, 2015).

The central goal of the analysis was to identify amplitude differences in ERPs that can be linked to differences in similarity judgment behavior, but the set of additional measures that were collected also have an important relationship to these behavioral patterns. Therefore, in addition to including word-based statistics (word length, frequency, orthographic neighborhood size, and constrained bigram frequency), individual differences in reading and language ability (exposure to print, verbal fluency and GRE vocabulary assessments) and concept similarity and association ratings in the modeling of the ERP waveforms, it was also important to characterize how these variables affect behavioral response patterns in the similarity judgment task. Thus,

the similarity judgment data will also be analyzed in relation to these variables.

## 5.3 Results

Recall that participants completed a series of reading and language assessments and then viewed a stream of images of letter strings where temporally adjacent strings could be taxonomic category members, thematic category members, unrelated concepts, or concept and pseudoword pairs. The session finished with a similarity judgment triad task. The results will be presented in four sections: (1) concept rating data, (2) reading and language exposure assessments, (3) behavioral task outcomes, and (4) general ERP results and behavioral–electrophysiological correspondences. A summary of each of these sections is provided in Section 5.4 to aid the reader.

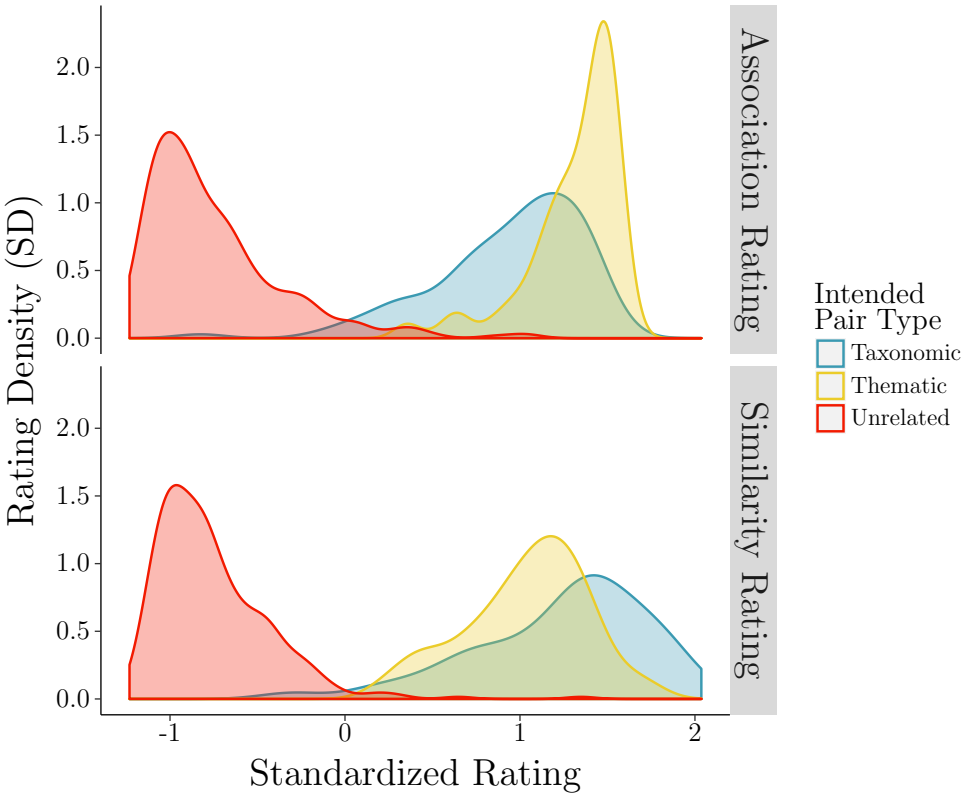


Figure 5.2: Density plot of standardized ratings for the association (top) and similarity (bottom) rating tasks. Taxonomic pairs were rated as more similar, thematic pairs were rated as more associated, and unrelated pairs were rated lowest on similarity and association. Taxonomic and thematic pairs in the same concept set were not reliably different in the magnitude of their standardized similarity and association ratings (respectively).

### 5.3.1 Concept Norming

The concept sets used for Experiment 4 were rated in an identical procedure to what was presented in Experiment 3. A separate set of participants ($N = 259$, association question condition: $n = 132$) were recruited from the Binghamton University Psychology Department Pool to collect the rating data. As in Experiment 3, the concept ratings were analyzed to confirm that taxonomic pairs were rated highest on the similarity question, thematic pairs were rated highest on the association question, and that the standardized strength of the similarity relationship for a given concept set was not reliably different from the standardized strength of the thematic relationship (Figure 5.2). Descriptive statistics are provided in Table 5.1.

#### Similarity and Association Strength

Mixed-effect LMER models were built to analyze the unadjusted association and similarity ratings. The similarity rating model (pair type as a fixed effect categorical predictor and participant as a random categorical predictor) uncovered reliably higher similarity ratings for the taxonomic pairs as compared to the thematic pairs ($\hat{\beta} = 5.488, SE = 0.55, t = 9.991, p < .001$) and the unrelated pairs ($\hat{\beta} = 55.837, SE = 0.48, t = 1117.06, p < .001$). Similarly, the association rating model showed that the thematic pairs were rated as more associated than the taxonomic pairs ($\hat{\beta} = 12.741, SE = 0.54, t = 23.43, p < .001$) and the unrelated pairs ($\hat{\beta} = 69.37, SE = 0.47, t = 147.41, p < .001$). Lastly, similarity and association scores within concept sets (calculated in the same manner as the Experiment 3 scores, see Section 4.3.2) were analyzed with a paired $t$-test to determine if semantic relationship strength within sets was biased toward taxonomic or thematic pairs. The $t$-test did not produce a reliable difference ($M_{Difference} = 0.03\ SD$) between the similarity scores of the taxonomic pairs and the association scores of the thematic pairs, $t(79) = 0.96, p = .34$. Thus, we cannot conclude that the concepts sets had more associated or more similar pairs (see Figure 5.3). The complete similarity and association rating data is provided in Appendix F, Table F.1.

Table 5.1: Experiment 4 Concept Ratings

| Pair Type | Similarity Rating Mean (SD) | Association Rating Mean (SD) | Similarity Rating Mean Response Time | Association Rating Mean Response Time |
|---|---|---|---|---|
| Taxonomic | 70.52 (1.24) | 75.26 (0.94) | 4.34 seconds | 4.08 seconds |
| Thematic | 65.05 (1.03) | 88.01 (1.31) | 4.32 seconds | 3.72 seconds |
| Unrelated | 14.68 (-0.75) | 18.59 (-0.75) | 4.38 seconds | 4.47 seconds |

#### Lexical and Orthographic Properties

The lexical and orthographic properties of the taxonomic and thematic targets in each concept set were also analyzed to determine if there were any systematic differences between the semantic relations. Paired $t$-tests confirm no differences in word length ($M_{Difference} = -0.06, t(79) = -0.22, p = .82$), word frequency ($M_{Difference} =$

11.75, $t(79) = 0.41, p = .68$), average frequency (per million) of orthographic neighbors ($M_{Difference} = 61.6, t(79) = 1.12, p = .27$) and average frequency of the constrained bigrams for the wordforms ($M_{Difference} = 26.72, t(79) = 0.11, p = .91$). Lexical and orthographic statistics are provided in Appendix F, Table F.2. Orthographic statistics were drawn from the MCWord database (Medler & Binder, 2005) and the word frequency data came from the Shaoul and Westbury (2006) USENET corpus.

Table 5.2: Experiment 4 Concept Properties

| Pair Type | Word Length | Word Frequency | Orthographic Neighborhood | Bigram Frequency | Similarity/Association Difference Score |
|---|---|---|---|---|---|
| Taxonomic | 5.66 | 52.33 | 90.16 | 1160.26 | 0.35 |
| Thematic | 5.73 | 40.18 | 28.55 | 1133.55 | 0.31 |
| Set Mean | 5.79 | 43.94 | 56.72 | 1148.00 | |

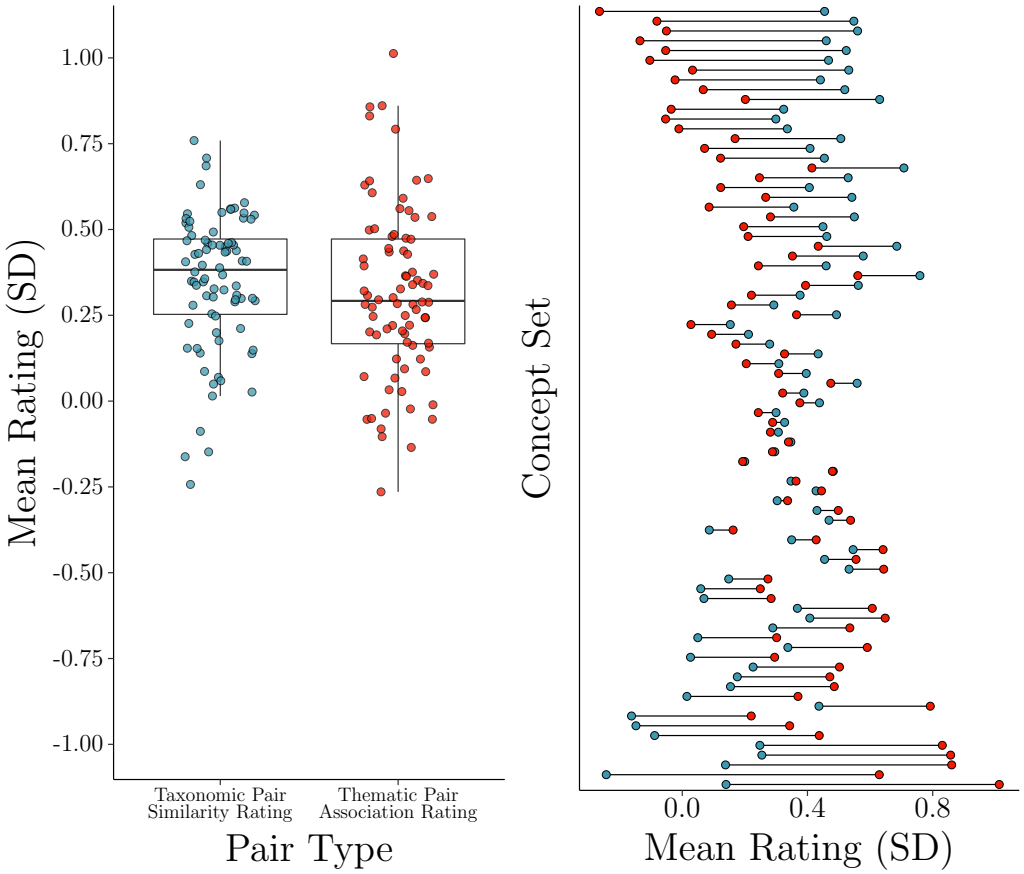


Figure 5.3: Visualization of the concept ratings overall (left) and paired with the match from the same concept set (right). The left panel depicts the mean similarity and association ratings for the taxonomic and thematic pairs, respectively. The right panel depicts the similarity (blue) and association (red) ratings paired for each concept set.

Table 5.3: Experiment 4 Behavioral Descriptives

| Responding Bias | Taxonomic Responding Mean (Med.) | Exposure to Print *d'* Mean (Med.) | Verbal Fluency Mean (Med.) | GRE Vocabulary Mean Accuracy (Med.) | Pseudoword Identification Accuracy (Med.) |
|---|---|---|---|---|---|
| Taxonomic | .88 (.89) | 1.57 (1.67) | 17.56 (18) | .58 (.55) | .93 (76) |
| Ambiguous | .48 (.47) | 1.23 (1.27) | 18.19 (18) | .53 (.55) | .88 (73) |
| Thematic | .31 (.33) | 1.27 (1.22) | 19.53 (19.67) | .44 (.45) | .89 (73) |
| Mean Total | .56 (.56) | 1.36 (1.39) | 18.43 (18.56) | .52 (.52) | .90 (74) |

### 5.3.2 Reading and Language Exposure Assessment

The reading and language exposure assessment data are presented in Figure 5.4. Recall that exposure to print was measured with *d'*, where higher values indicate more success in identifying real magazines and authors while rejecting fake magazines and authors. The verbal fluency task was to name as many members of a category as possible in 60 seconds. This produced a verbal fluency score calculated by averaging the number of distinct fruits, animals, and colors that were named in the time allotted. The GRE vocabulary assessment was a 30 item fill-in-the-blank task that was scored as a proportion correct. As mentioned above, data for one participant's verbal fluency task and one participant's exposure to print task were missing. These values were imputed in `R` with the `mi` package (Su et al., 2011).[4] The median values from 8000 hypothetical value estimations (80 trials × 100 hypothetical datasets) replaced the missing data points. The results of the reading and language exposure assessments are presented in Table 5.3. All of the measures were normally distributed according to Shapiro–Wilk tests.

### 5.3.3 Triad Similarity Judgment Task

#### Similarity Judgments in the Triad Task

The taxonomic pair was selected 56.7% of the time (mean range by participant: 12.5%–98.75%)—a lower frequency of taxonomic responses than what is needed to conclude that there was a reliable taxonomic bias at the participant level (and the lowest frequency found across Experiments 1–4). Binomial tests were conducted to classify each participant as taxonomic, thematic or ambiguous in their responding. The process resulted in 22 taxonomic biased responders, 22 thematic biased responders, and 14 ambiguous responders. When these bias frequency data are analyzed in a binomial exact test, the result is that people produce a taxonomic (or thematic) bias less frequently than would be expected by chance ($p = .087$), though this test

[4]Parameters for the missing values were estimated at the trial level with data from the triad task and the reading and language exposure assessments (i.e., participant, trial number, concept set, trial response, response time, mean verbal fluency, exposure to print *d'* and GRE vocabulary accuracy). The ERP data were excluded from the imputation procedure due to extreme processing requirements.

was only marginally significant.

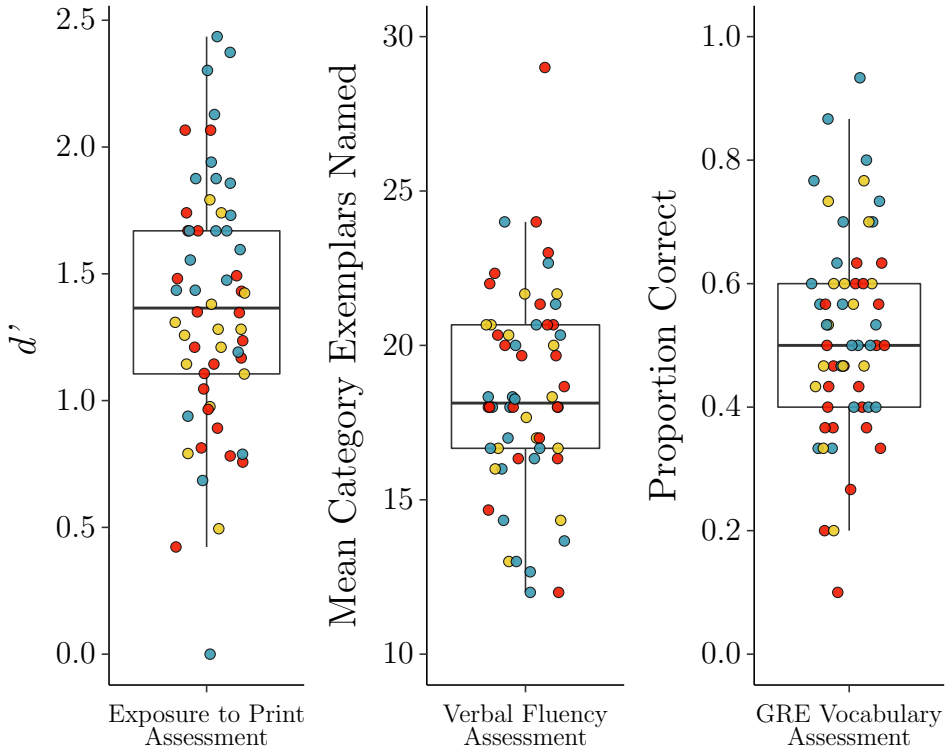


Figure 5.4: Boxplots and individual data for each of the reading and language exposure tasks. Blue, red, and yellow points present people with taxonomic, thematic or ambiguous responding preferences, respectively. The data were normally distributed with no obvious outliers. Exposure to Print and GRE Vocabulary were positively related to taxonomic responding and Verbal Fluency was negatively related to taxonomic responding.

### Response Time in the Triad Task

Overall, taxonomic matches were completed faster than thematic matches ($\hat{\beta} = 0.256, SD = 0.10, t = 2.465, p = .018$) but this effect is not found when outliers are removed ($\pm 2.5$ *SD*; $p = .11$). Consistent with the proposal that faster responding is found for the semantic relationship that is preferred or sought out, people with a taxonomic bias were faster on trials where the taxonomic pair was chosen, $\hat{\beta} = -0.92, SE = 0.20, t = -4.651, p < .001$, and people with a thematic responding bias or ambiguous response preference were faster on thematic trials, $\hat{\beta} = -0.14, SE = 0.05, t = -3.032, p = .006$ and $\hat{\beta} = -0.16, SE = 0.07, t = -2.426, p = .031$, respectively. This response bias timing effect was resilient to outlier exclusion ($\pm 2.5$ *SD*s).

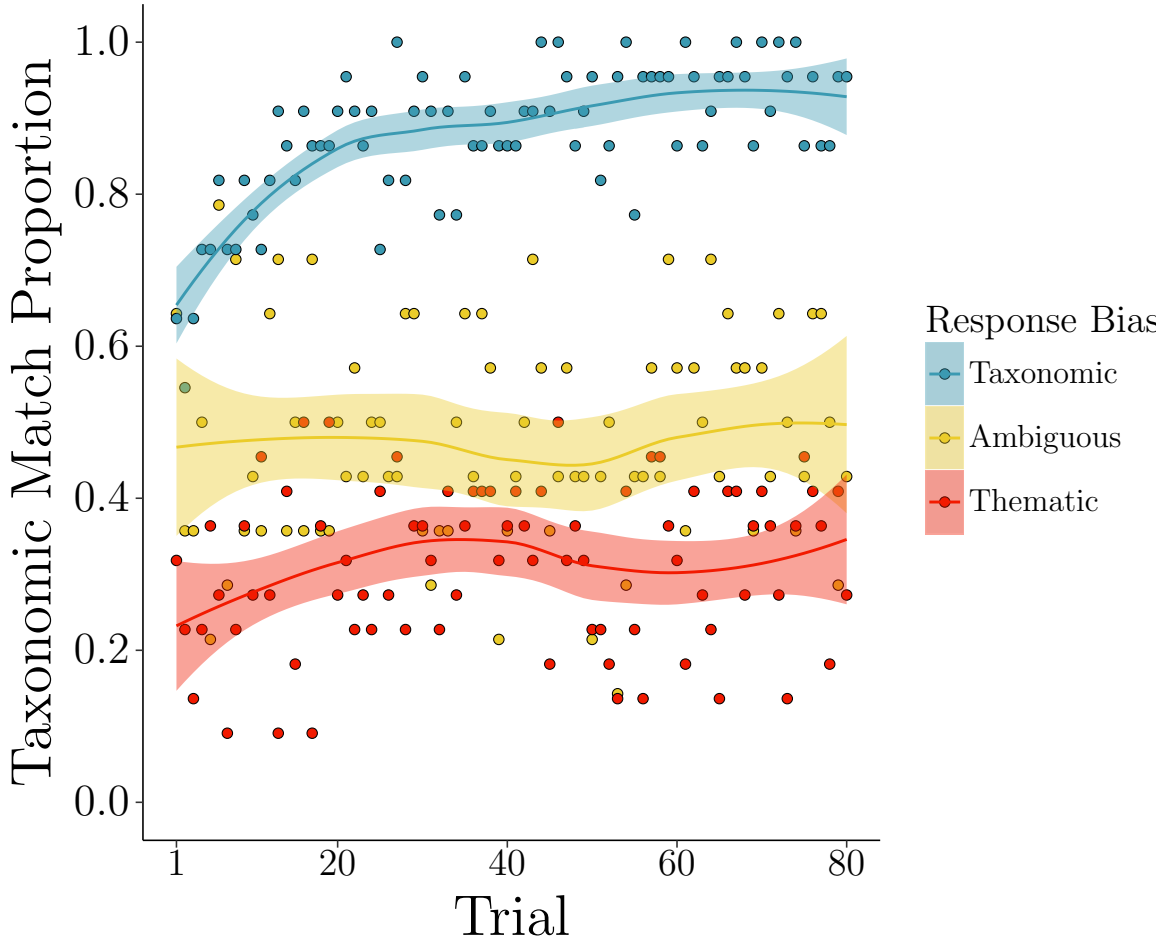


Figure 5.5: Taxonomic responding frequency across trials in Experiment 4. Points represent mean taxonomic responding by trial for response bias type. Taxonomic matching reliably increased as the experiment progressed.

### Similarity Judgments and Reading and Language Exposure

*General Relationship between Similarity Judgments and Reading and Language Exposure.* While they had clear importance for the ERP measurement goals of the study, it was less clear how these measures might relate to similarity judgment behavior. A series of regression models were built to examine this relationship. A simple GLM built to predict taxonomic responding at the trial level including trial and all three reading and language exposure measures uncovered reliable effects of all predictors ($p$s $< .001$).

A different pattern emerges when the data are analyzed with mixed effects (taking participant and concept set into account). A GLMER model built to predict trial-level taxonomic responding with fixed effect predictors for each of the reading and language exposure assessments and trial and random effects (random intercepts for participant and concept set and random slopes for trial) produced a reliable effect of trial, $\hat{\beta} = -0.16, SE = 0.07, t = -2.426, p = .031$; no other reliable effects were found and allowing the terms to interact did not change this overall pattern.[5] This is now the third instance in this report where an effect of trial has been found where people increased their taxonomic responding across the time-course of the experimental session (see Sections 2.3.4 and 3.3.4).

No reliable differences were uncovered for exposure to print, verbal fluency or GRE vocabulary when they were analyzed in isolation ($\text{ETP}_{\text{Wald } Z} = 1.378$, $\text{ETP}_p = 0.168$; $\text{VF}_{\text{Wald } Z}$ = -1.235, $\text{VF}_p = .22$). GRE vocabulary accuracy did approach significance as a predictor of ERP amplitude ($\hat{\beta} = 2.531, SE = 1.30$, Wald $Z = 1.945, p = .052$).[6] See Table 5.3 for descriptive statistics.

The simplest explanation for the conflicting results between the simple and mixed-effects models is that people differ in similar ways in terms of responding preferences and reading and language exposure. When the random intercept term for participant is included, this similarity is accounted for and adding predictors for the specific measures does not address significantly more variance. It is not safe to conclude that the simple GLM produced a spurious relationship between these variables, but the current results are not strong enough to make conclusions about how the predictors relate to similarity judgments. The *patchy* or bimodal distribution of mean taxonomic responding (Figure 5.6) could also be playing a role in the failure to find reliable effects with the mixed-effects approach.

*Individual-based Relationship between Similarity Judgments and Reading and Language Exposure.* Since the overall relationship between the survey measures and taxonomic responding frequency is not clear, it might be more informative to look at this relationship with the inclusion of the response bias classification of each participant. In line with the main hypothesis of this paper (ERP differences are detectable between participants but not in the aggregate), it is possible that differences in the survey mea-

[5]Model specification: `taxonomic.selection ~ response.bias × etp.data × mean.vf × vocab.acc + trial + (1 + trial|pid) + (1|concept.set)`

[6]The measure-isolated models only differed from the comprehensive model in that a single predictor was included from the reading and language exposure assessments (as opposed to all three measures).

sures are also obscured when response bias is not accounted for. This is what was found with the caveat that the comprehensive model including response bias, trial and the survey measures often failed to converge. A fairly safe conclusion, however, is that response bias and the survey measures interact. When the model did converge (i.e., after many additional iterations and the use of the `Nelder–Mead` optimizer), this interaction was consistently reliable for the difference between taxonomic and ambiguous responding groups (e.g., $\hat{\beta} = 6.30, SE = 2.49$, Wald $Z = 2.534, p = .011$). Unfortunately, the parameter estimates for the interaction between the taxonomic and thematic responding groups were quite volatile across model initializations, $p$s = .002–.4.

We also conducted the taxonomic responding analysis within each response bias group (as opposed to using response bias as a predictor). Again, these analyses were plagued with convergence failures. Nevertheless, an interesting pattern emerged that is worthy of mentioning even under this caveat. It was found that the survey measures and their interaction predicted taxonomic responding for the taxonomic ($p$s $< .001$) and ambiguous ($p$s = .005 – .028) bias groups. No survey measure, however, was found to be reliable for the thematic bias group. If it needs to be restated, any interpretation of the results of these models should be made with extreme caution. We take this as evidence that—similar to the convergence failures in Experiment 2—the regression models suffer from overdispersion in the outcome variable, i.e., variability in trial-level responding that is not being sufficiently addressed by the predictors of these models.

#### Pseudoword Identification

The sole purpose of the pseudoword task was to confirm that participants were paying close attention to the word stream during EEG collection, but it is possible that the ability to detect pseudowords is related to taxonomic—thematic processing (as was the case with the reading and langauge exposure assessments). Overall, participants did quite well in identifying pseudowords ($M = 72.2$; 90%). The correct identification of pesudowords was a reliable predictor of taxonomic responding, $\hat{\beta} = 0.08, SE = 0.03$, Wald $Z = 2.522, p = .012$. The analysis featured pseudoword identification and trial number as fixed effects, participant as a random intercept, trial as its random participant-level slope, and concept set as a random intercept. The effect was not reliable when the pseudoword accuracy predictor was included as a fixed-effect predictor in the mixed-effects model that featured the reading and language exposure surveys (See Footnote 5 for the model specification save the fixed-effect pseudoword accuracy predictor).

### 5.3.4 Electrophysiological Responses to Taxonomic and Thematic Category Members

Ideally, a comprehensive model of the ERP data (i.e., amplitude across time bins for the target channels) would be constructed that included all behavioral data and stimulus characteristics that have been collected and presented in this report, i.e.,

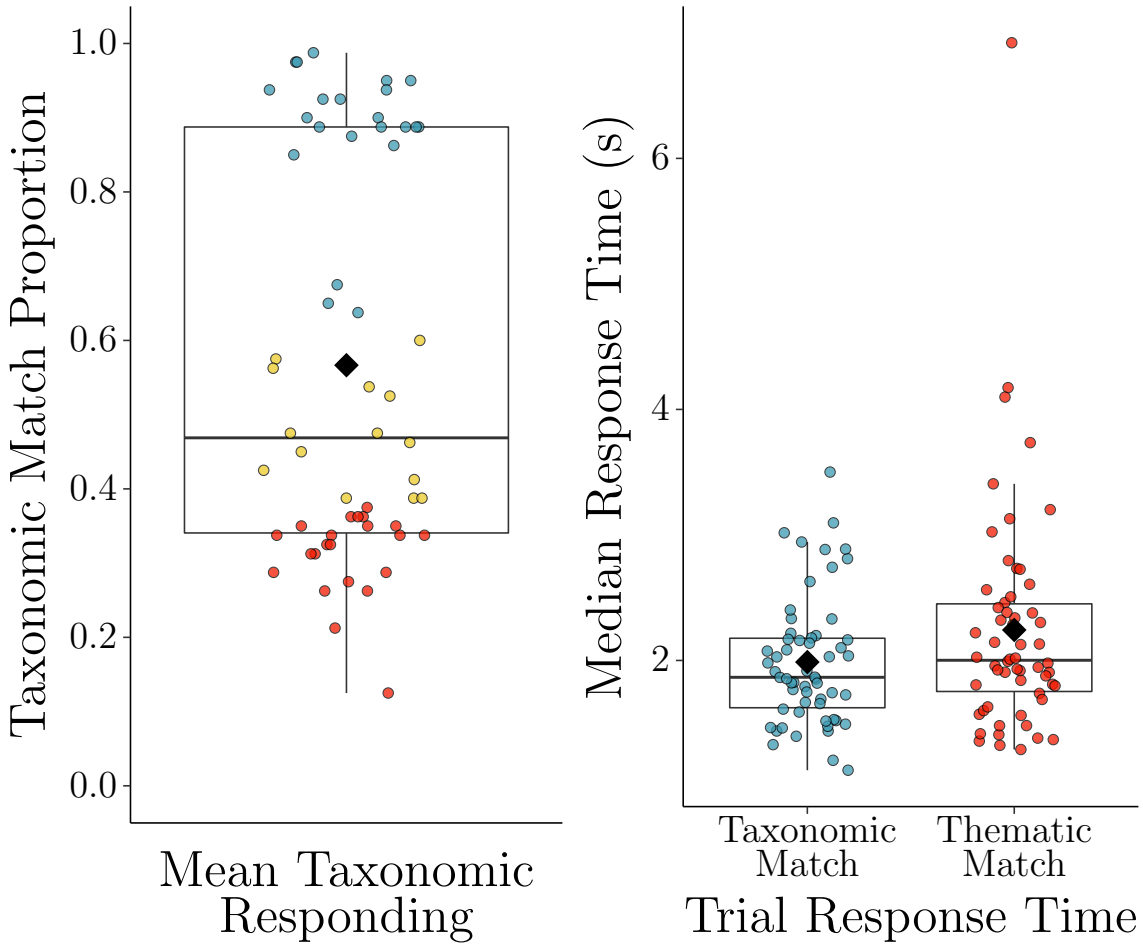


Figure 5.6: Boxplots present mean taxonomic responding (left panel) and median response time for taxonomic and thematic matches (right panel) from the triad task. Individual points present participant means and medians. Diamonds present overall means. More taxonomic responding was found overall but there was no participant-level response bias majority. Trials with a taxonomic match were generally completed faster than thematic trials but the reliability of this effect turns on 2 near-outliers.

similarity judgments, reading and language exposure outcomes, and lexical and orthographic properties of the materials. Building and presenting a model with this level of complexity is prohibitive due to technical demands, difficulty of interpretation and increased false positive rate (Luck & Gaspelin, 2017). Consistent with the presentation of results thus far, the ERP analysis is divided to present specific aspects of the problem with models that address subsets of the possible predictors. First, a general analysis of the ERPs is presented that includes no similarity judgment data. The idea here is to start with a model similar to what has been used in past research to attempt to detect differences in ERP amplitude between taxonomic and thematic pairs across an entire sample. Next, a model of similarity judgment behavior, reading and language skill and orthographic and lexical variables is presented to determine if these factors predict unique variance in N400 amplitude. Finally, the simplest possible models of the relationship between similarity judgment behavior and N400 amplitude are presented.

### General Properties of ERPs Elicited by Taxonomic and Thematic Category Members

We start with a comprehensive model of the ERPs without the effects of similarity behavior—an analysis approach similar to what has previously failed to detect differences in N400 amplitude from waveforms elicited by taxonomic and thematic category members. An LMER model was built to predict average ERP amplitude at central–posterior electrode sites from lexical and orthographic characteristics, similarity and association rating difference scores (see Section 4.3.2 for details) and—most importantly—semantic pair type.[7] The model uncovered reliable effects for semantic pair type but similarity ratings, word frequency, word length, orthographic neighborhood and bigram frequency were not reliable predictors.

In the aggregate, taxonomic category members elicited ERP waveforms with more positive N400s than thematic category members ($\hat{\beta} = .129, SE = 0.006, t = 2.09$), and unrelated concepts ($\hat{\beta} = 0.819, SE = 0.007, t = 11.71$) when accounting for other sources of stimulus-based variance (see Figure 5.7). Thematic category members elicited more positive N400s than unrelated category members ($\hat{\beta} = 0.69, SE = 0.07, t = 9.85$). To our knowledge, this is the only reported instance of N400 component differences elicited by broadly-defined taxonomic and thematic category members in healthy adults.[8]

[7]The model predicted ERP amplitude from un-averaged, trial-level data at MiCe, MiPa, LDPa, RDPa, LMOc, RMOc, LLOc, RLOc, and MiOc with the following model specification:

```
N400 amplitude ~ similarity.rating + frequency + length + orthographic.neighborhood
+ bigram.frequency + pair.type + (1 + time.start|participant) + (1|word.stimulus)
```

[8]N400 amplitude differences were reported in Wamain et al. (2015), but the stimuli were more restrictive, ERP amplitude was averaged across electrode sites, and the ISI between stimulus pairs was much shorter than the present investigation (<400 ms vs. 3.5 s). Hagoort et al. (1996) have also reported N400 differences, but these differences were only found in a comparison between healthy adults and right-hemisphere damaged adults.

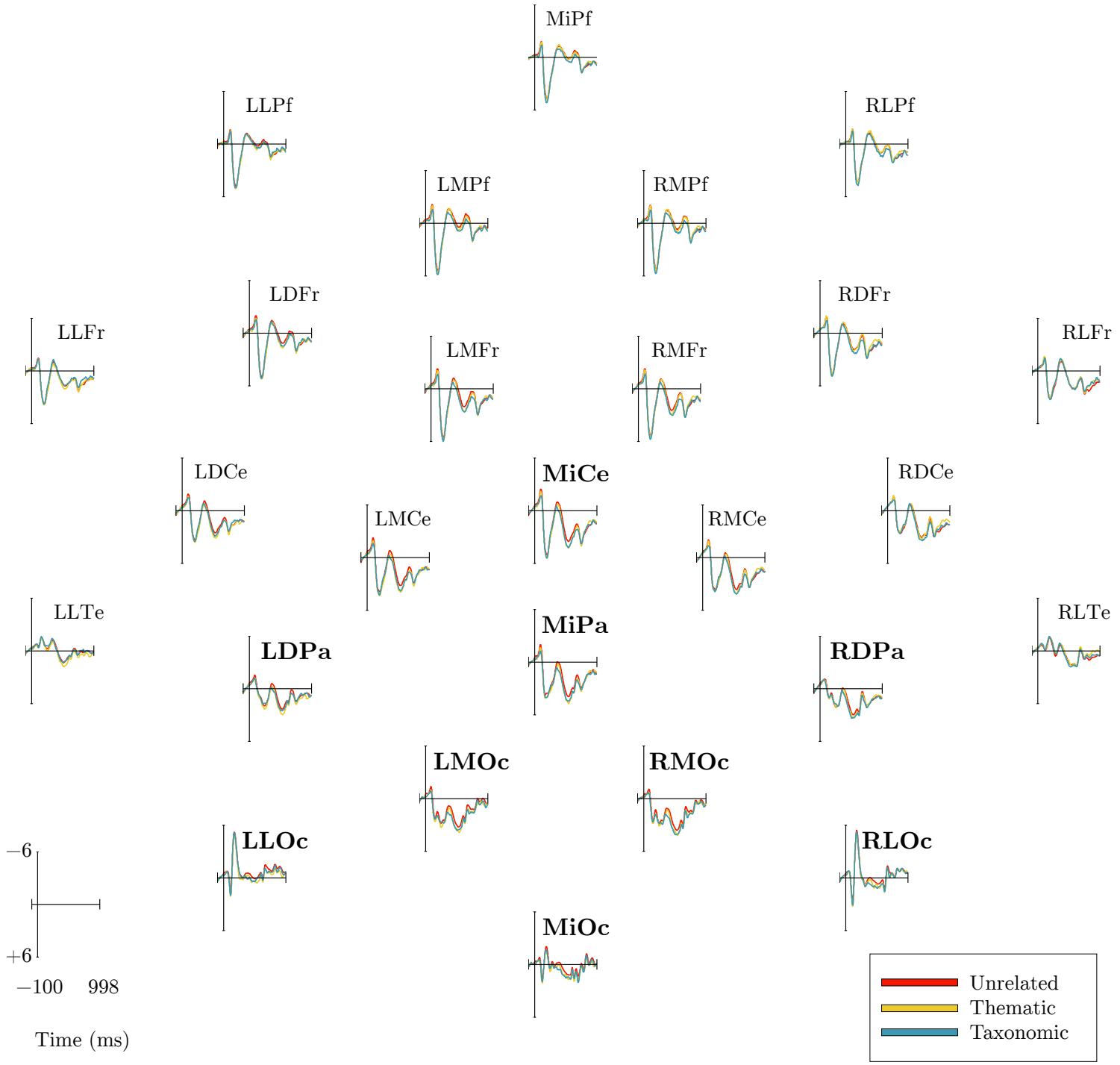


Figure 5.7: Grand averaged ERP waveforms elicited in response to taxonomic, thematic and unrelated word pairs (pseudoword trials excluded). Unrelated, thematic and taxonomic pairs are presented in red, yellow and blue, respectively. Electrode sites included in the analyses are presented in bold. An average reference was used for off-line re-referencing. The data are presented baselined and filtered with bandpass filtering at 0.1–20 Hz.

**Similarity Judgments, Reading and Language Exposure and ERP Waveforms**

As mentioned above, it is difficult to specify a single model that can comprehensively assess the contributions of the predictors in this design; a comprehensive analysis would include a series of models referenced to different combinations of the categorical predictors, including a large number of predictor terms in each. Therefore, we started by constructing a model that included all of the predictor terms necessary to address a question not answerable with fewer terms: Do the key variables of interest—similarity judgment behavior, reading and language exposure measures, and semantic pair type—interact to predict mean N400 amplitude when also accounting for the variance of task engagement (pseudoword identification accuracy) and concept properties (similarity ratings, length, frequency, bigram frequency, orthographic neighborhood size).[9] If so, further investigation of the direction of these effects would be warranted. In other words, a reliable interaction between these variables would help to validate the use of less sophisticated models without the concern that reading and language exposure (for example) can explain the effect; reliable interactions in this general model[10] would provide evidence against the interpretation that N400 amplitude differences are not directly related to similarity judgment behavior in the triad task.

The baseline reference levels for the analysis were taxonomic pairs for the semantic pair type variable and taxonomic responding bias for the response bias variable. A reliable interaction (exposure to print *d'* × verbal fluency mean × vocabulary assessment accuracy × response bias × semantic pair type) was found for each pair type by response bias group combination. The variables interacted to reliably predict amplitude differences between the taxonomic and thematic bias group for taxonomic pairs vs. thematic pairs ($\hat{\beta} = 3.21, SE = 0.87, t = 3.70$) and unrelated pairs ($\hat{\beta} = -2.89, SE = 0.68, t = -4.23$) and between the taxonomic and ambiguous bias group for taxonomic pairs vs. thematic pairs ($\hat{\beta} = 12.83, SE = 1.12, t = 11.47$) and unrelated pairs ($\hat{\beta} = 10.68, SE = 0.88, t = 12.16$). The categorical reference level for semantic pair type was set to unrelated pairs to examine the effect of the interaction for unrelated and thematic pairs between the taxonomic and ambiguous bias groups. The interaction was found to be a reliable predictor of N400 amplitude, $\hat{\beta} = 2.15, SE = 0.89, t = 2.42$.

To address the remaining comparisons, the categorical reference levels for the model were set to thematic pairs and thematic response bias and the model was recalculated. The interaction was reliable for thematic pairs vs. taxonomic pairs ($\hat{\beta} = -9.62, SE = 1.26, t = -7.62$) and unrelated pairs ($\hat{\beta} = 3.95, SE = 1.00, t = 3.94$). To analyze the final interaction effect for unrelated and taxonomic pairs between the

[9] The random effects structure and target electrode sites were identical to the previous model.

[10] The model structure was specified as:

```
N400 amplitude ~ similarity.rating + length + frequency + orthographic.neighborhood
+ bigram.frequency + pseudoword.accuracy + pair.type × response.bias
× exposure.to.print × verbal.fluency × vocabulary.accuracy + (1 +
time.start|participant) + (1|word.stimulus)
```

ambiguous and thematic bias groups, the categorical semantic pair type reference level was set to unrelated pairs and the analysis was repeated. This interaction was also reliable, $\hat{\beta} = -13.57, SE = 0.99, t = -13.63$.

The interaction of similarity judgment behavior, reading and language ability assessments and semantic pair type was found to reliably predict N400 amplitude differences for every response bias–semantic pair type comparison. Similarity ratings, word length, word frequency, orthographic neighborhood, bigram frequency and pseudoword identification were not reliable predictors in the model.

### Closer Examination of ERPs and Similarity Judgment Behavior

The models above suggest that similarity judgments and reading and language ability interact to predict differences in N400 amplitude across semantically related and unrelated concept pairs. A critical question that remains unresolved is how *exactly* these variables affect ERP amplitude. Models were built that held the categorical semantic pair type and response bias variables constant to determine (1) how semantic pairs differed in ERP amplitude within response bias groups and (2) how response bias groups differed in ERP amplitude for each semantic pair.[11] First, models built for each response bias group are presented to examine the differences between semantic pair types. Second, models built to examine differences across the response bias groups for each semantic pair type are presented. A depiction of these effects is presented in Figure 5.8.

*Semantic Pair Differences within Response Bias Groups.* The mean amplitude of ERPs elicited by semantically related and unrelated pairs in the 300–400 ms time window was analyzed within each response bias group (taxonomic, thematic and ambiguous) with LMER.[12] The goal of this analysis was to determine how the elicited waveforms of semantic pair types differed for people who produced ambiguous responding, majority taxonomic responding and majority thematic responding. The results showed that people who made more taxonomic matches in the triad task also produced N400s that were different for taxonomic and thematic pairs ($\hat{\beta} = 0.16, SE = 0.06, t = 2.67$), taxonomic and unrelated pairs ($\hat{\beta} = 0.58, SE = 0.07, t = 7.77$) and thematic and unrelated pairs ($\hat{\beta} = 0.74, SE = 0.07, t = 9.96$). People who produced more thematic matches in the triad task produced different N400s for thematic and unrelated pairs ($\hat{\beta} = 0.44, SE = 0.07, t = 6.55$) and taxonomic and unrelated pairs ($\hat{\beta} = 0.49, SE = 0.07, t = 7.40$) but no difference between taxonomic and thematic pairs ($t = 0.90$). Lastly, people who did not produce a reliable match preference (ambiguous responders) in the triad task did not produce reliable differences between any of the semantic pair types. In other words, people who produced the most taxonomic responding in the triad task also produced more differentiable ERP waveforms for taxonomic and thematic pairs. An example of the facilitative priming effect from data collected at RLOc is presented in Figure 5.9 and Table 5.4.

[11] The random effects structure and target electrode sites were identical to the previous models.

[12] Simple semantic pair model (for each response bias group):

```
amplitude ~ pair.type + (1 + time.start|participant) + (1|word.stimulus)
```

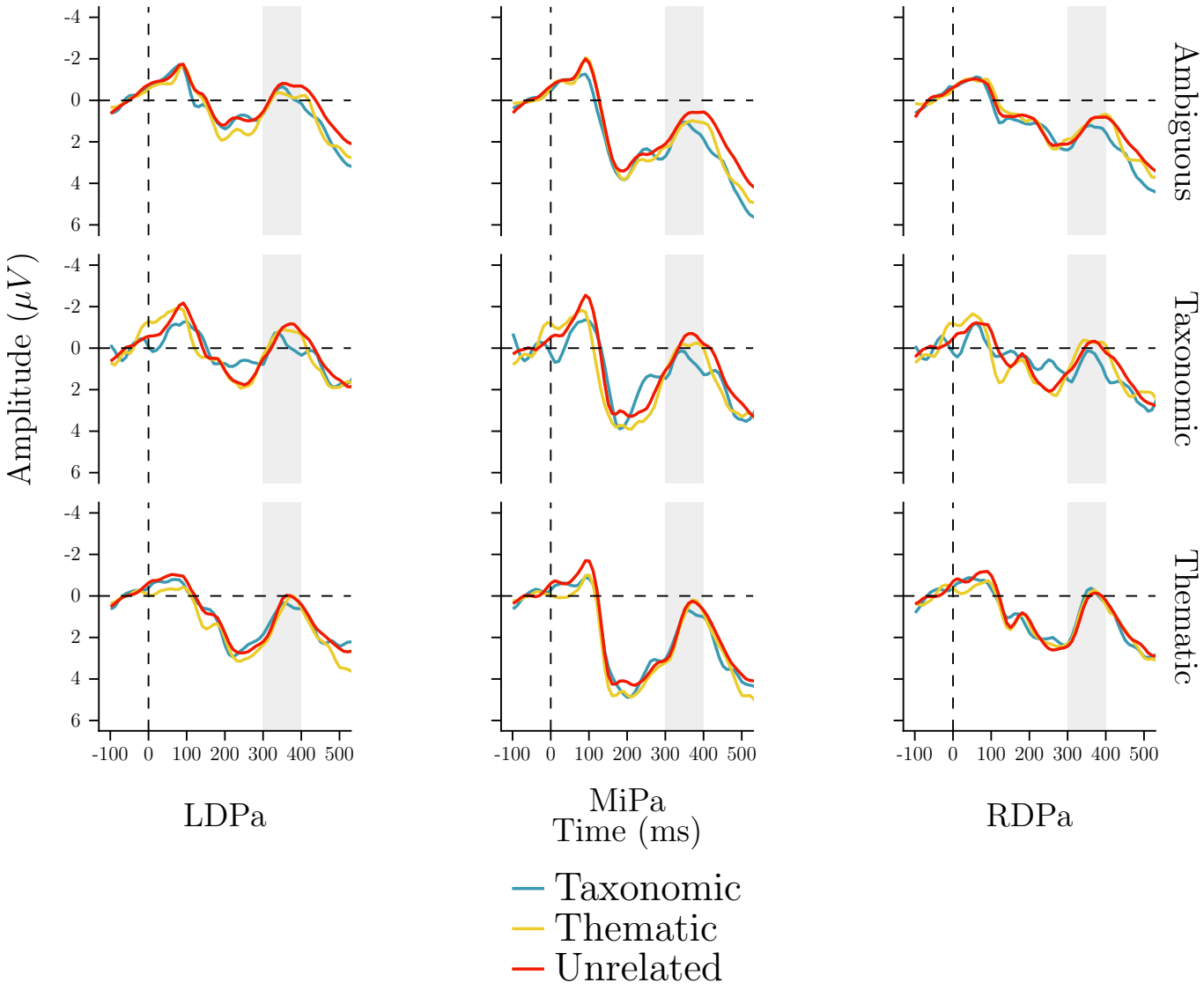


Figure 5.8: ERPs elicited from taxonomic, thematic, and unrelated word pairs. Horizontally-aligned panels present response bias groups. Vertically-aligned panels present data from LDPa, MiPa and RDPa. N400s elicited by taxonomic and thematic pairs were reliably different for the taxonomic bias group only, i.e., the group that produced reliably more taxonomic responding in the similarity judgment task was the only group to produce reliably different N400s for taxonomic and thematic pairs.

*Response Bias Differences within Semantic Pairs.* Similar to the previous analysis, LMER models were built that held one component constant (semantic pair type) to examine possible differences between the other (response bias group).[13] No reliable differences were found across response bias groups, i.e., no response bias group produced more positive or negative N400s for any semantic pair type.

Table 5.4: Experiment 4 Facilitative Priming Profiles from RLOc

| Response Bias | Facilitative Priming Profile | |
|---|---|---|
| | Taxonomic | Thematic |
| Taxonomic | 16 | 6 |
| Thematic | 13 | 9 |
| Ambiguous | 8 | 6 |

*N400 Amplitude Predicted by Semantic Pairs and Taxonomic Responding.* One possible issue with the analysis above is that the cutoff for being classified as having a particular bias ($\alpha$) is an arbitrary criterion—it turns on the difference between 49 ($p = .056$) and 50 ($p = .033$) consistent responses in an 80 trial experiment. Recall that response biases were calculated with binomial exact tests that compared the number of consistent matches to what would be expected by chance under the null-hypothesis significance testing (NHST) framework. The problem with this approach is a general issue in NHST—setting $\alpha = .05$ is an arbitrary cutoff and even the framework's originators disagreed about the importance of the cutoff as it relates to the dichotomous significance decision (Fisher, 1925; Neyman & Pearson, 1928).

Motivated by these concerns, a final set of models was constructed where mean amplitude for the N400 component was predicted by the interaction of semantic pair type and the proportion of taxonomic responses produced in the triad task (with random effects structures and electrode sites identical to the models above). The models uncovered a reliable interaction between taxonomic match proportion and semantic pair type where more taxonomic responding predicted more positive N400s for the comparison of taxonomic pairs to thematic pairs ($\hat{\beta} = 0.59, SE = 0.11, t = 5.65$) and taxonomic pairs to unrelated pairs ($\hat{\beta} = 0.50, SE = 0.08, t = 5.94$), but not thematic and unrelated pairs ($t = 1.16$). In other words, the proportion of taxonomic matches was a reliable predictor of amplitude differences between the taxonomic category members and thematic and unrelated pairs.

## 5.4 Discussion

The results of the ERP analyses show that taxonomic and thematic category members produce reliably different N400s when aggregating across the entire sample. Similar-

[13] Simple response bias group model (for each semantic pair type):
`amplitude ~ response.bias + (1 + time.start|participant) + (1|word.stimulus)`

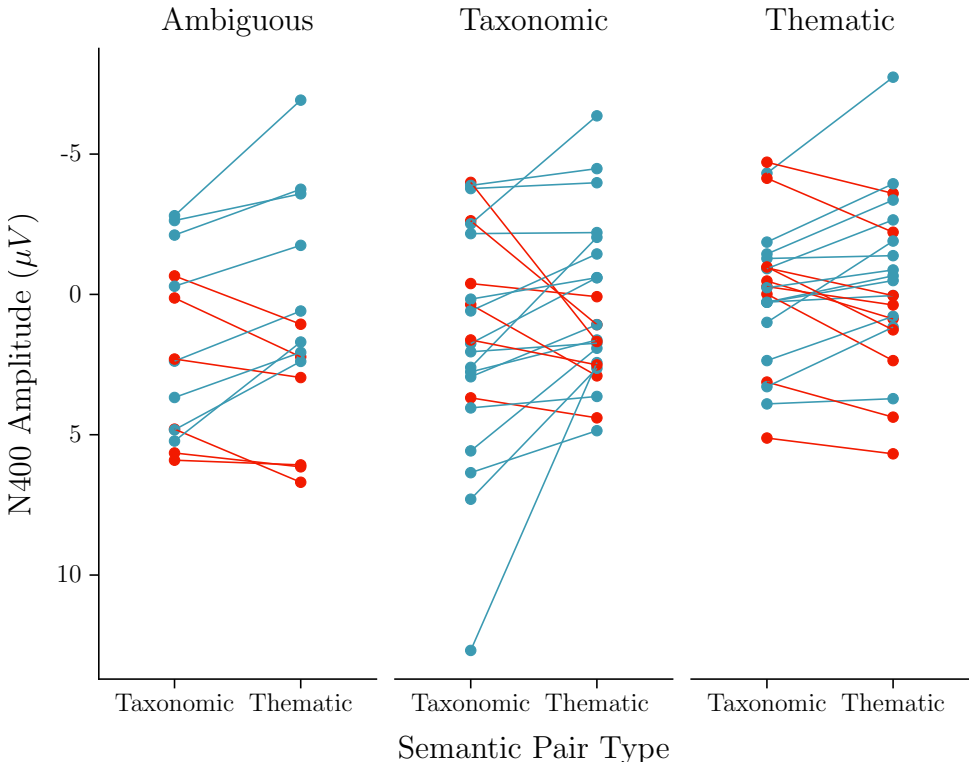


Figure 5.9: Figure depicts N400 amplitude elicited by taxonomic and thematic pairs at RLOc. Points and lines are colored to show the amplitude differences between the semantic pairs for each participant. Blue points and lines indicate that the subject exhibited a numerical pattern consistent with more facilitative priming for taxonomic category members. Red points and lines indicate the opposite. Negative is plotted up for consistency with the ERP plots. Counts of each cell are provided in Table 5.4

ity judgment behavior predicts N400 amplitude differences at the individual level. Reading and language ability (as measured by the exposure to print, verbal fluency and vocabulary assessments) is related to similarity judgment behavior but all variables predict unique N400 variance—similarity judgment behavior remains a reliable predictor and interacts with reading and language ability to predict N400 amplitude differences for taxonomic, thematic, and unrelated word pairs.

### 5.4.1 Behavioral Measures

#### Triad Task Responding

*Taxonomic Responding.* Taxonomic responding in the present study was numerically lower (56.7%) than any sample in this report where the goal was to choose the most similar match (61%–74%, see Experiments 1 & 2). Note that this is still not enough to qualify as a thematic bias in the aggregate or at the participant level. While the results across these experiments have shown that the aggregated mean is a less useful indicator of responding in this research area, the possible causes of this lower rate of taxonomic responding must be considered. The two concept sets used in this series were assessed in an identical analysis procedure and the results support the same outcome—no evidence suggests that the items are more biased toward the selection of a particular semantic relation. The taxonomic pairs were rated as more similar and the thematic pairs were rated as more associated. The standardized strength of the similarity of taxonomic pairs and the association of thematic pairs was not reliably different. The lexical and orthographic features of the concept sets were not different. Therefore—while the concept sets contained different words and word pairs—there is currently no evidence to suggest that they differ on qualities hypothesized to affect similarity judgments and processing.

An alternative explanation is that people were fatigued after the 2–3 hour procedure. This is different from the conditions of Experiments 1 and 2 where people completed a series of unrelated studies over the course of an hour-long session with no EEG setup.

Both of these explanations, however, do not work with the idea that people have stable similarity responding preferences regardless of task. Evidence for this stability comes from the fact that responding preferences did correspond to more distinct ERP waveforms across the tested semantic pairs. One final possibility is that the lower rate of taxonomic responding is a consequence of variation when sampling the phenomenon. This explanation is supported by the fact that more consistent thematic responders were found in this experiment than any other sample with similarity-based task instructions. While the current data cannot tease apart these possible explanations, it does speak to the volatility in sampling similarity judgment behavior in the triad task.

*Responding Time-Course.* Taxonomic responding increased across the time-course of the experiment. Closer inspection of the pattern shows that the effect was driven by individuals with a taxonomic response bias. Responding patterns become more stable after approximately 20–30 trials. This suggests that experiments that feature fewer

trials will produce outcomes that are adversely affected by this early stabilization in terms of parameter estimation and inference.

*Triad Task Response Time.* People who had a taxonomic responding bias completed the trials where they made a taxonomic match fastest. Conversely, people who had a thematic (or ambiguous) responding bias completed trials with thematic matches fastest (Figure 5.6). This outcome replicates the general pattern found in Experiments 1 and 2 where response time patterns were not homogeneous across experimental manipulations. Like the similarity judgments themselves, these results suggest that response time has an individualized component that should not be overlooked when making conclusions about the processing time of similarity judgments.

*Reading and Language Exposure.* The modeling difficulties outlined above require careful interpretation of the relationship between reading and language ability (as evidenced by the reading and language exposure assessments) and responding in the triad task. Analysis with a simple GLM model suggested that all of the survey measures were predictive of taxonomic responding. A mixed-effects approach produced results that were quite different—only a (marginal) effect of vocabulary assessment accuracy remained, where higher vocabulary assessment accuracy predicted more taxonomic responding ($p = .052$). Attempts to model the predictive value of the interaction between the language measures and responding preferences often failed, suggesting volatility of the estimated parameters and possible overdispersion in the trial-level responding data. However, these errors proved somewhat informative in that the interaction between reading and language ability and similarity judgments was reliable for people with ambiguous and taxonomic responding patterns but not for people with a thematic response bias.

*Pseudoword Identification.* Accurate pseudoword identification was predictive of taxonomic responding. Higher accuracy on the identification task predicted taxonomic responses at the trial level. One interpretation of this relationship is that people who were more motivated to perform well were more likely to do well on the pseudoword task and think critically about the similarity judgment task. Gentner and Brem (1999) suggest that taxonomic responding involves a level of *introspective awareness of cognitive states*. In this case, however, it's possible that pseudoword identification accuracy is measuring motivation and thus *willingness* to engage in introspection of cognitive states, i.e., a willingness to seek out the difference between the competing semantic relationships and use it as the basis for similarity decisions in the task. This hypothesis also speaks to a possible role of the reading and language exposure assessments—when pseudoword identification is included with these measures it is no longer a reliable predictor of taxonomic responding. This suggests that variance predicted by the survey measures also accounts for variance predicted by pseudoword identification accuracy.

### 5.4.2 Characterization of ERPs Elicited by Taxonomic and Thematic Category Members.

Taxonomic and thematic category members were found to elicit ERP waveforms with reliable N400 amplitude differences. Overall, N400s elicited by taxonomic category pairs were more positive than N400s elicited by thematic category pairs. These results conflict with the null effects reported by Chen et al. (2013). We return to speculate on the causes of this difference in the General Discussion. ERPs elicited by unrelated words were found to be different from taxonomic and thematic pairs; as expected, semantically-related pairs produced more positive N400s.

### 5.4.3 ERPs and Similarity Judgments.

The analysis uncovered a number of reliable correspondences between similarity judgments and ERPs elicited from the same taxonomic and thematic pairs. At the highest level, a series of reliable interactions were found between the three reading and language exposure measures, triad task responding biases, and semantic pairs. The interaction of these variables suggests that they predict unique variance in the mean amplitude of the N400 component. In other words, similarity judgments remain predictive of N400 amplitude even when accounting for effects of concept variance and reading and language ability.

Looking closer at the specific relationship between similarity judgments and mean N400 amplitude, we found that people who produced particular response biases differed in systematic ways. The taxonomic bias group produced reliably different N400s for taxonomic and thematic pairs. This difference was not found in the thematic and ambiguous responding bias groups. The effect was also found when the response bias group variable was replaced with proportion of taxonomic responses—more taxonomic responding predicted more positive N400s for taxonomic pairs relative to thematic and unrelated pairs. The results of this analysis suggest that people who show differences in their processing of taxonomic and thematic pairs are less likely to be subject to confusability and more likely to produce matches based on taxonomic similarity in the triad task.

### 5.4.4 Conclusion

In Experiment 4, we set out to test two hypothesis: that (1) the failure to detect differences between ERPs elicited by taxonomic and thematic category members was caused by processing differences between individuals and (2) an unbiased reading task could be used to clarify the competing claims of the confusability and dual process accounts of thematic intrusion.

The results provide support for the first hypothesis in that different patterns of N400 amplitude were discovered between participant response bias groupings (and taxonomic match frequencies) as determined by a distinct similarity judgment task. Contrary to this hyptohesis and prior research, however, general differences in N400s elicited by taxonomic and thematic pairs were also found. This suggests that part of

the problem in past studies could have been statistical power. In the present design more participants were recruited in an attempt to sample adequately-sized groups with different similarity judgment behavior; the size of each response bias group was comparable to the size of entire samples in studies in this research area. Sample size is likely more important for ERP studies on taxonomic and thematic categories because stimulus creation cannot be automated and the result is smaller stimulus sets than ERP investigations in other areas. Regardless of the general pattern, a novel conclusion of this work is that ERPs elicited by unbiased passive reading of taxonomic and thematic category members have a direct correspondence with similarity judgments of those same concepts in the classic forced-choice triad task.

People who produced more taxonomic matches in the triad task also produced distinct patterns of N400s between taxonomic and thematic category members (Figures 5.8 and 5.9). People who produced mostly thematic matches did not show this pattern; N400s elicited by thematic and taxonomic pairs in this group only differed from unrelated pairs. Lastly, ambiguous responders—those who did not consistently match taxonomic or thematic pairs in the triad task—showed no reliable N400 differentiation between the semantic pair types and unrelated pairs.

Returning to the existing hypotheses on the cause of thematic intrusion on human similarity judgments, the evidence suggests that the dual process model is not an adequate explanation of the thematic intrusion effect. More taxonomic responding co-occurs with more distinct ERP patterns and taxonomic and thematic category members produce reliably different N400s in the aggregate, results that contrast with the outcome and argument made in Chen et al. (2013) where a failure to find N400 differences was taken as evidence for an integrated association and similarity processing system.

The confusability account remains viable as an explanation for thematic intrusion. Again, a higher rate of taxonomic similarity-based responding corresponds to distinct N400s elicited by taxonomic and thematic pairs. The ERP data seem to support the differential sensitivity prediction of the confusability account—ambiguous responders produced no differentiation and thematically-biased responders only produced differentiation between semantically-related and unrelated pairs.

These results suggest that electrophysiological patterns elicited by the passive processing of semantically related and unrelated concept pairs are a reliable predictor of similarity judgment behavior. More reading and language skill (higher exposure to print *d'* and vocabulary assessment accuracy) predicts taxonomic matching and N400 amplitude, but individual similarity judgment behavior still explains variance in the ERP data. The strongest possible conclusion is that the failure to produce more taxonomic similarity-based matches in the triad task is attributable to less-distinctive processing of taxonomic and thematic category members, as evidenced by less distinctive N400s in the ambiguous and thematically-biased response groups. At the individual level, ERPs that don't differentiate between taxonomic and thematic category members are evidence of more difficulty in perceiving differences between taxonomic and thematic matches when making similarity judgments. Future work will follow up on the *differentiation* hypothesis to examine the flexibility of the effect under different task goals and the stability of these patterns over extended periods of

time.

# General Discussion and Conclusion

This project set out to investigate a 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 matching due to ambiguity in instructions and other characteristics of the task. The response pattern found in this series of experiments was surprising given past reports of overall thematic response biases with similar samples under similar conditions. The study was initially motivated by the goal to examine and reduce thematic intrusion on the similarity judgment process. This goal was somewhat sidelined when it was found that the most frequent responses were taxonomic matches in all conditions that were not biased against this result. This general pattern was even found under the conditions of the classic forced-choice triad task. Given these findings, the story is not that components of the Anti-Thematic Intrusion task affected thematic responding—they did, but more work is needed to clarify this effect—or that the classic triad task reliably produces a majority of thematic responses (it doesn't). Rather, the most surprising result of this investigation is that the prevalence of the thematic response bias seems to have been over-estimated in previous reports. This work was followed-up with an ERP experiment featuring an unbiased passive reading task that showed that when people produced majority thematic responding it corresponded to less distinctive ERPs as compared to taxonomically-biased responders.

We started with a concern that dual-process accounts under-emphasize the importance of distinguishing between taxonomic and thematic category members and found even less evidence supporting these accounts than previously thought. Simply stated, people were not affected by thematic intrusion enough to show an overall thematic responding bias in conditions where they were asked to judge similarity.[1] In fact, taxonomic responding reliably increased as the experiments progressed.

## 6.1 Confusability or Dual-Process Integration?

One of the most surprising discoveries of this project was the time-course of taxonomic responding. People start out in the task unsure about how to respond and eventually settle in to a consistent responding pattern (see the supplemental materials for an illustrative animation of this effect). Most frequently, this shift results in a response

[1] The only case where this did occur was when people were not repeatedly told to look for the most similar match.

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 to no expectation about what type(s) of concepts will appear or how they will be connected. After a handful of trials, people in the classic forced-choice version of the task might have learned that the two response options will always share a distinct semantic relation with the standard. Thus, it might be more clear to them that the task is to decide which of these two competing semantic relations is most similar. Participants in the Random (no prioritized standard) conditions can figure this out too, but it might be less obvious because there are frequently two thematic matches in the set (the intended thematic match and a weaker match between the intended thematic and taxonomic targets). People manage to resolve the conflict, however, and most frequently this resolution comes in the form of a taxonomic response. We suggest that—even though thematic intrusion appears to continue throughout the task—the accompanying confusion has been resolved and this leads to the higher-than-expected taxonomic responding reported here.

The results of Experiment 4 also support this conclusion. Only the taxonomic bias group showed the time-course effect. This group's ERPs were also more distinctive than the thematic bias group. So across the time-course of the experiment, people who are more sensitive to differences between taxonomic and thematic category members focus in on this distinction and choose a consistent response type. People who do not produce distinctive ERPs do not notice the differences and their responding is not different from the beginning to the end of the similarity judgment task.

## 6.2 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 Anti-Thematic Intrusion task and its individual components are less clear. Our initial interpretation of Experiment 1 was that the ATI task was driving the observed increase in rates of taxonomic responding. In the task, the prioritized role of the standard was scrapped, distractors were added to remove the direct competition between the taxonomic and thematic match, and the outcome was an overall taxonomic response preference. 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 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).

## 6.3 The Role of Individual Differences

Perhaps the most pressing issue to resolve is the role of individual differences for thematic intrusion (see also Lin & Murphy, 2001; Mirman & Graziano, 2012; 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. The results of Experiment 4 can shed some 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. We found that people who produced a taxonomic bias in the triad task also showed reliably different facilitative priming for taxonomic and thematic pairs while the thematic and ambiguous responders did not. As mentioned in the Experiment 4 discussion, this evidence dovetails with results from Mirman and Graziano (2012) that taxonomic responders are more distracted by competing semantic relations in the visual world paradigm task.

These results do not speak to the stability of these effects across time. An interesting follow-up question might be to try to classify the various and competing interpretations of the task goals, instructions, definitions of similarity, and elicited ERPs across a series of sessions. Do people who choose 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 of whether this judgment is due to confusion remains unresolved. Does this definition change over the course of the task? Are ERP differences a result of a temporary mindset or a result of stable differences in cognitive processing? How stable are these patterns across longer periods of time? Answering these questions would go a long way to help to differentiate between the predictions of competing theories of psychological similarity.

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 correlative, it is possible that response biases have deep underlying consequences for cognition. The present research shows that people who are better at identifying real authors and magazines produce more taxonomic responding. People who do better on a vocabulary assessment produce more taxonomic responding. In contrast, preferences for thematic responding are associated with lower 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, 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).[2]

Formal education (or lack thereof) and occupational pressures have also been suggested as drivers of thematic responding (Denney, 1974; Sharp et al., 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

[2]It is possible that these cultural effects explain the differences between the behavioral data in Experiments 1–4 and the aggregated ERP results of Experiment 4 and Chen et al. (2013), where the task was administered in Chinese.

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). Experiment 4 adds new information to this area by suggesting that exposure to print, verbal fluency, and vocabulary reliably predict taxonomic responding and ERP differences. 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.

## 6.4 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 and the norming work in Experiment 4, see Section 5.3.1) 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 (Figures 4.2 and 5.3). 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 after a non-negligible amount of the experiment has been completed (Figures 2.4 and 3.3). 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, the outcome measure is averaged over all trials, the majority of which are completed as the responding preference is stabilizing. Aggregation based statistics will underestimate the strength and direction of 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 stan-

dard 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 (Experiments 1 and 2) and individual-based (Experiment 4 reading and language exposure measures) differences found without mixed-effects show the need to account for this variance, where cleaner interpretations would have been possible[3] 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 presented here. The participant-level differences found here are difficult to adequately model. Our hypothesis is that this difficulty is caused by overdispersion 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. It was fortunate that roughly equal groups of taxonomic and thematic responders were sampled in Experiment 4, allowing for evenly-sized comparisons. We take all this to mean 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.

## 6.5 Conclusion

In the end, the Anti-Thematic Intrusion task could be perceived as a solution in need of a problem. We set out with the goal of investigating a frequently-reported pattern in real-world concept similarity judgments, where people showed a thematic response bias in the 2AFC conflict triad task. This turned out to be difficult to find except in situations that were biased toward thematic responding. The existence of "holdouts" in the most biasing of circumstances suggests that it is more than the task that causes this behavior. The electrophysiological evidence addresses this issue directly in that processing differences were found in a task that has no explicit instructions or biasing task constraints.

We believe the higher frequency of taxonomic responding found here can in part be explained by advances in methodology: (1) avoiding aggregation-based statistics, (2) using mixed-effects analysis techniques, (3) including more trials, (4) more participants, and (5) materials rated as appropriately taxonomically or thematically-related

[3]e.g., definitive differences in taxonomic responding between the instructional manipulations in Experiment 1; definitive differences in taxonomic responding between task components in Experiment 2; reliability of the reading and language exposure measures for predicting response preferences.

all contribute to this more nuanced view. Individual response patterns and electrophysiological patterns help to clarify how the role of individual differences interacts with these factors. Giving the standard a special status, co-presenting distractors, and clarifying instructions all affect the frequency of taxonomic responding. People that are more sensitive to taxonomic and thematic pairs (according to ERPs elicited by these semantic relations) produce more taxonomic responding. The experiments presented here suggest that thematic intrusion is controllable with experience. Theoretical accounts of similarity that propose that thematic association is 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. The ability to distinguish between these competing semantic relationships varies across people. Nevertheless, psychological similarity is required in the service of inductive reasoning—inference, generalization, and taxonomic categorization. Proposals that include thematic association in this system must be based on strong evidence and specify exactly how these competing semantic relationships are weighted for the purposes of similarity judgments. Task constraints, goals, concept properties and individual differences in the processing of taxonomic and thematic category members are important contributors to the thematic intrusion effect on similarity. Further work that identifies how these factors interact in the production of similarity judgments will be critical for the goals of advancing theoretical accounts of similarity and successfully predicting human similarity judgment behavior.

# Appendix A: Experiments 1–3 Concept Sets

Table A.1: Experiments 1–3 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

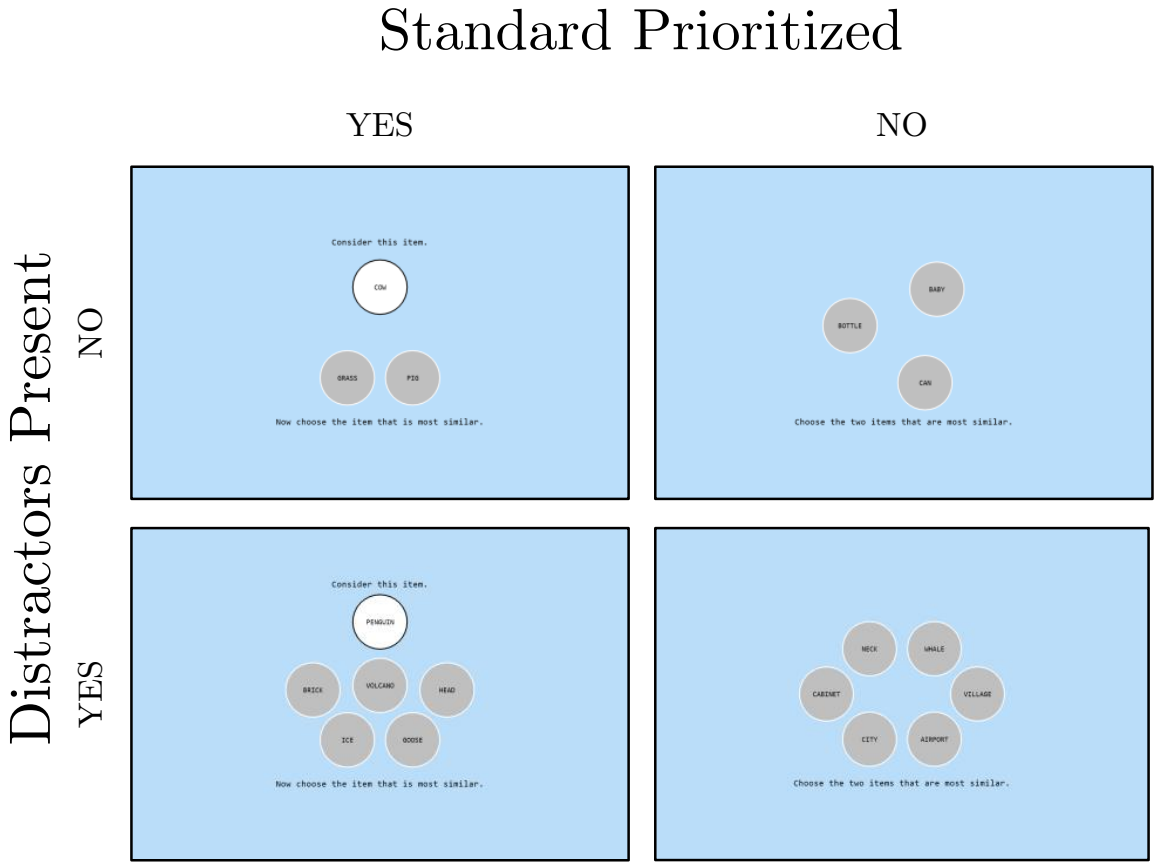


Figure B.1: Figure presents the four spatial configurations of the similarity judgment task in Experiment 2. Not pictured is the Standard Thematic Triad condition that featured the *Goes With* instructions and the classic triad task configuration (top left quadrant).

# Appendix C: Experiment 3 Task Depiction

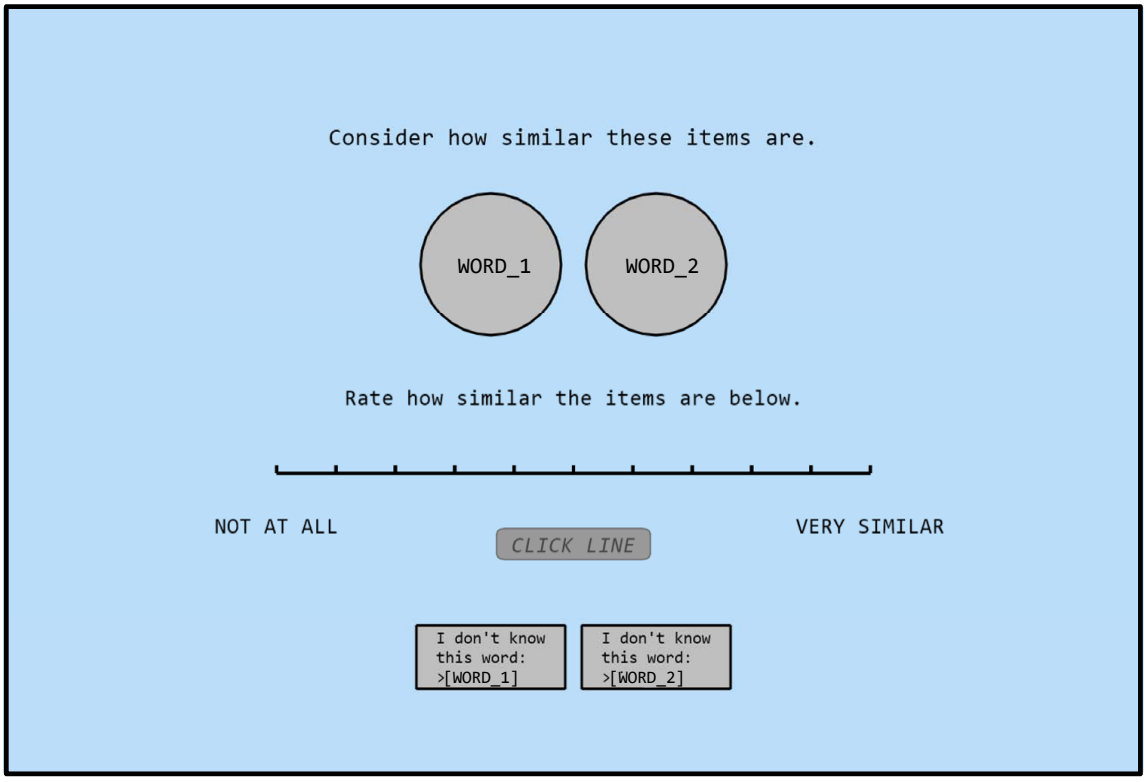


Figure C.1: Figure presents a depiction of the similarity rating task from Experiment 3. Participants were allowed to choose any point on the rating line to provide their rating. Association rating task not pictured.

# Appendix D: Experiments 1–3 Concept Properties

Table D.1: 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 |

# Appendix E: Experiment 4 Concept Sets

Table E.1: Experiment 4 Concept Sets

| Index | Standard | Taxonomic | Thematic | Unrelated | Unrelated | Pseudoword |
|---|---|---|---|---|---|---|
| 1 | CIGARETTES | ALCOHOL | LUNGS | CARPET | OUTLET | LURDUGE |
| 2 | WAITRESS | STEWARDESS | RESTAURANT | CALCIUM | SWAN | CHATAGHT |
| 3 | BEE | BUTTERFLY | HONEY | PLIERS | RECORD | INVOMBLY |
| 4 | TOOTHBRUSH | COMB | FLOSS | APE | GLASSES | RELEFUT |
| 5 | CUP | BOWL | TEA | BARBER | PHONE | SURNGE |
| 6 | SKI | SNOWBOARD | BOAT | FLOOR | STOMACH | WHICE |
| 7 | DOG | CAT | BONE | HOOD | POND | YOMECHED |
| 8 | RECEPTIONIST | HOSTESS | TELEPHONE | HAND | PARK | PAIT |
| 9 | RABBI | PASTOR | TEMPLE | DRIVEWAY | UNDERWEAR | SETIVITE |
| 10 | CABLE | CORD | TELEVISION | POT | ROCK | COSTEDED |
| 11 | GOAT | BUFFALO | FARM | CHALK | SKY | PINIER |
| 12 | FIELD | COURT | FLOWER | SCHOOL | TOAD | BEANERED |
| 13 | MINT | LOLLIPOP | BREATH | FELONY | STALLION | INYWERED |
| 14 | COOKIE | PIE | CHOCOLATE | FUR | WAVE | COLOUST |
| 15 | HORNET | WASP | STINGER | PADLOCK | RICE | BURTH |
| 16 | LAWNMOWER | SCISSORS | YARD | AUNT | BOMB | LEPELF |
| 17 | VINEYARD | ORCHARD | WINE | BEAD | DRIVER | ABQUE |
| 18 | PANDA | RACOON | BAMBOO | LAW | WHIP | NUEENG |
| 19 | BEER | JUICE | PARTY | CARRIAGE | SHOP | LOYWED |
| 20 | SPOON | LADLE | SOUP | LION | STEREO | REIEMBLY |
| 21 | HORSE | PIG | GRASS | HOTEL | MUTANT | SUEPANED |
| 22 | CAMEL | ANTELOPE | DESERT | COFFIN | ENGINE | EATENDLY |
| 23 | BLANKET | COMFORTER | PILLOW | CUCUMBER | TAR | MOUNCTE |
| 24 | TURKEY | CHICKEN | STUFFING | LETTER | SQUARE | TOMSTED |
| 25 | SHOTGUN | PISTOL | SHELL | ARK | BELT | RERANING |
| 26 | PACKAGE | CRATE | DELIVERY | CHILD | TROUT | INTH |
| 27 | SHAMPOO | BLEACH | SHOWER | CIRCLE | PIGEON | REATOWER |
| 28 | TOE | FINGER | SANDAL | MARBLE | SPIKE | HARN |
| 29 | TRUCK | BUS | TRAILER | CACTUS | CLUB | AMILES |
| 30 | BICYCLE | CAR | HELMET | BASEMENT | SKIN | NOSTE |
| 31 | BOOTS | HEELS | SHOELACE | BALCONY | BRAIN | REARAROD |
| 32 | SAXOPHONE | HARP | JAZZ | HAIR | SODA | FOMPERED |
| 33 | OYSTER | SCALLOP | PEARL | BACTERIA | LEATHER | COSSENG |
| 34 | CRIB | BED | BABY | FERRY | PATIO | LEIGS |
| 35 | POLICE | FIREMAN | HANDCUFFS | CRAB | LAUNDRY | INYOPT |
| 36 | RABBIT | SQUIRREL | CARROT | BARBELL | MOTEL | TREARDE |
| 37 | MILK | LEMONADE | COW | GUITAR | WINDOW | REEROT |
| 38 | BOTTLE | CAN | INFANT | BERRY | CLOCK | YEVER |
| 39 | BIRD | BAT | NEST | CRIMINAL | PLAYGROUND | SHUR |
| 40 | ROCKET | MISSILE | ASTRONAUT | CHEESE | SINK | GERMAL |
| 41 | SHIP | CANOE | SAILOR | GLAND | UMBRELLA | STUTABLY |
| 42 | PLATE | TRAY | NAPKIN | ANKLE | CHAUFFEUR | COOWENUL |
| 43 | CROWN | HAT | KING | NOSE | SHOVEL | LERSE |
| 44 | HURRICANE | BLIZZARD | FLOOD | BADGE | FOSSIL | GAEAID |
| 45 | LOCKER | CLOSET | JERSEY | PAINT | SPY | WAGHT |
| 46 | HEARSE | LIMOUSINE | GRAVEYARD | EYE | KITCHEN | SOLVY |
| 47 | NEEDLE | PIN | THREAD | HYDRANT | WRIST | LELICT |
| 48 | CELEBRITY | PLUMBER | FILM | FORTRESS | NECTAR | WARAENE |
| 49 | MONKEY | BEAR | BANANA | HAMMER | TOOTH | PRILY |
| 50 | OVEN | MICROWAVE | PAN | CONVICT | SCREEN | WOOUT |
| 51 | SKYSCRAPER | TOWER | ELEVATOR | HEART | HITCHHIKER | RUTISES |
| 52 | SURGEON | BUTCHER | KIDNEY | DYNAMITE | GALAXY | ISKERT |
| 53 | CHISEL | KNIFE | SCULPTURE | HATCH | MIRROR | MEDERAN |
| 54 | SHOE | GLOVE | FOOT | TIGER | WALL | SUNICED |
| 55 | FOOTBALL | BASEBALL | QUARTERBACK | NECKLACE | PLANT | SWILUARY |
| 56 | ENVELOPE | PARCEL | STAMP | MUSCLE | YOGURT | FREANDE |
| 57 | JELLY | MARMALADE | JAR | BOOK | NAIL | ACHITIED |
| 58 | SALT | PEPPER | SEA | KNUCKLE | SAW | BERFFER |
| 59 | CASKET | BOX | GRAVE | JEWEL | STREET | HARY |
| 60 | FLY | ANT | WINGS | CEREAL | CONCRETE | VAVE |
| 61 | DOOR | GATE | KNOB | FLAG | LIQUID | VINS |
| 62 | PENGUIN | GOOSE | ICE | BRICK | HEAD | COMORVED |
| 63 | CAKE | DONUT | CANDLE | ACTRESS | BROCHURE | COREWAL |
| 64 | OWL | HAWK | MOON | CIRCUIT | DIARY | CHOURN |
| 65 | HOSE | TUBE | WATER | MOTHER | RODEO | FOVIND |
| 66 | SWEATER | HOODIE | MITTENS | BATHROOM | CHALKBOARD | MARMIGLY |
| 67 | SEDAN | BIKE | SEATBELT | COTTON | SHRIMP | FEEPPER |
| 68 | PENCIL | PEN | ERASER | FLUTE | SHEEP | HALY |
| 69 | BACKPACK | SUITCASE | NOTEBOOK | BUTTER | PAINTING | BROURD |
| 70 | SEAGULL | DUCK | PIER | BEDROOM | POWDER | SHERT |
| 71 | VENOM | POISON | SNAKE | GRAFFITI | RASPBERRY | TURICAFT |
| 72 | TORTILLA | BREAD | BEANS | COLD | WIRE | BREATED |
| 73 | COMPUTER | TABLET | MOUSE | ATHLETE | COUCH | CEEY |
| 74 | CHAIR | SOFA | LEGS | ANCHOVY | BALL | AGATENG |
| 75 | BISCUITS | TOAST | GRAVY | DANCE | SNAIL | RENCTRY |
| 76 | FLOUR | CORNMEAL | DOUGH | BUTTON | SMOG | BEVERSS |
| 77 | SHIRT | BLOUSE | COLLAR | BRIDGE | POOL | QUMES |
| 78 | PATHWAY | SIDEWALK | GRAVEL | BABYSITTER | TYPEWRITER | SOOBRARE |
| 79 | SNOW | RAIN | SLED | CEMETERY | NOVEL | KITSSES |
| 80 | CITY | VILLAGE | AIRPORT | NECK | WHALE | SQUGED |

*Note: Unrelated words in Experiment 4 were only presented in the EEG recording phase of the procedure.*

# Appendix F: Experiment 4 Concept Properties

Table F.1: Experiment 4 Similarity and Association Ratings

| Index | Standard | Taxonomic | Thematic | Unrelated | Unrelated | Taxonomic Rating | Thematic Rating | Tax.–Unr. Rating | The.–Unr. Rating | Tax.–The. Rating Difference |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CIGARETTES | ALCOHOL | LUNGS | CARPET | OUTLET | −0.09 | 0.44 | -0.38 | -0.86 | -0.53 |
| 2 | WAITRESS | STEWARDESS | RESTAURANT | CALCIUM | SWAN | 0.58 | 0.35 | -1.04 | -1.05 | 0.23 |
| 3 | BEE | BUTTERFLY | HONEY | PLIERS | RECORD | 0.35 | 0.43 | -0.95 | -0.92 | -0.08 |
| 4 | TOOTHBRUSH | COMB | FLOSS | APE | GLASSES | 0.3 | -0.05 | -1.13 | -1.16 | 0.35 |
| 5 | CUP | BOWL | TEA | BARBER | PHONE | 0.29 | 0.54 | -0.76 | -0.61 | -0.25 |
| 6 | SKI | SNOWBOARD | BOAT | FLOOR | STOMACH | 0.21 | 0.09 | -0.88 | -1.01 | 0.12 |
| 7 | DOG | CAT | BONE | HOOD | POND | 0.14 | 1.01 | -0.94 | -0.9 | -0.87 |
| 8 | RECEPTIONIST | HOSTESS | TELEPHONE | HAND | PARK | 0.69 | 0.43 | -0.56 | -0.73 | 0.25 |
| 9 | RABBI | PASTOR | TEMPLE | DRIVEWAY | UNDERWEAR | 0.46 | -0.14 | -1.18 | -1.12 | 0.6 |
| 10 | CABLE | CORD | TELEVISION | POT | ROCK | 0.56 | -0.05 | -0.79 | -0.86 | 0.61 |
| 11 | GOAT | BUFFALO | FARM | CHALK | SKY | 0.14 | 0.86 | -0.99 | -1.02 | -0.72 |
| 12 | FIELD | COURT | FLOWER | SCHOOL | TOAD | 0.15 | 0.03 | -0.92 | -1 | 0.13 |
| 13 | MINT | LOLLIPOP | BREATH | FELONY | STALLION | 0.71 | 0.41 | -0.78 | -0.92 | 0.29 |
| 14 | COOKIE | PIE | CHOCOLATE | FUR | WAVE | 0.09 | 0.16 | -1 | -1.12 | -0.08 |
| 15 | HORNET | WASP | STINGER | PADLOCK | RICE | 0.52 | 0.07 | -0.94 | -1.04 | 0.45 |
| 16 | LAWNMOWER | SCISSORS | YARD | AUNT | BOMB | 0.43 | 0.44 | -1.02 | -1.23 | -0.02 |
| 17 | VINEYARD | ORCHARD | WINE | BEAD | DRIVER | 0.29 | 0.16 | -0.92 | -1.05 | 0.14 |
| 18 | PANDA | RACOON | BAMBOO | LAW | WHIP | 0.55 | 0.28 | -0.46 | -0.33 | 0.27 |
| 19 | BEER | JUICE | PARTY | CARRIAGE | SHOP | 0.48 | 0.48 | -0.32 | -0.39 | 0 |
| 20 | SPOON | LADLE | SOUP | LION | STEREO | 0.49 | 0.37 | -1.13 | -0.9 | 0.13 |
| 21 | HORSE | PIG | GRASS | HOTEL | MUTANT | 0.25 | 0.86 | -1.16 | -1.13 | -0.6 |
| 22 | CAMEL | ANTELOPE | DESERT | COFFIN | ENGINE | −0.24 | 0.63 | -1.03 | -1.08 | -0.87 |
| 23 | BLANKET | COMFORTER | PILLOW | CUCUMBER | TAR | 0.56 | 0.39 | -1.02 | -1.16 | 0.17 |
| 24 | TURKEY | CHICKEN | STUFFING | LETTER | SQUARE | 0.07 | 0.28 | -0.37 | -0.4 | -0.21 |
| 25 | SHOTGUN | PISTOL | SHELL | ARK | BELT | 0.45 | 0.2 | -0.97 | -1.12 | 0.25 |
| 26 | PACKAGE | CRATE | DELIVERY | CHILD | TROUT | 0.03 | 0.3 | -0.77 | -0.77 | -0.27 |
| 27 | SHAMPOO | BLEACH | SHOWER | CIRCLE | PIGEON | 0.06 | 0.25 | -0.96 | -0.9 | -0.19 |
| 28 | TOE | FINGER | SANDAL | MARBLE | SPIKE | 0.35 | 0.34 | -0.99 | -0.92 | 0.01 |
| 29 | TRUCK | BUS | TRAILER | CACTUS | CLUB | 0.44 | -0.02 | -0.77 | -0.94 | 0.46 |
| 30 | BICYCLE | CAR | HELMET | BASEMENT | SKIN | 0.37 | 0.61 | -1.02 | -1.13 | -0.24 |
| 31 | BOOTS | HEELS | SHOELACE | BALCONY | BRAIN | 0.3 | 0.24 | -0.98 | -0.9 | 0.06 |
| 32 | SAXOPHONE | HARP | JAZZ | HAIR | SODA | 0.28 | 0.17 | -1.16 | -1.22 | 0.11 |
| 33 | OYSTER | SCALLOP | PEARL | BACTERIA | LEATHER | 0.15 | 0.27 | -0.8 | -0.85 | -0.13 |
| 34 | CRIB | BED | BABY | FERRY | PATIO | 0.56 | 0.47 | -0.76 | -0.7 | 0.08 |
| 35 | POLICE | FIREMAN | HANDCUFFS | CRAB | LAUNDRY | 0.31 | 0.28 | -0.96 | -1.12 | 0.03 |
| 36 | RABBIT | SQUIRREL | CARROT | BARBELL | MOTEL | 0.55 | 0.64 | -0.63 | -0.68 | -0.1 |
| 37 | MILK | LEMONADE | COW | GUITAR | WINDOW | 0.4 | 0.31 | -0.75 | -0.87 | 0.09 |
| 38 | BOTTLE | CAN | INFANT | BERRY | CLOCK | 0.53 | 0.64 | -1.04 | -1.15 | -0.11 |
| 39 | BIRD | BAT | NEST | CRIMINAL | PLAYGROUND | 0.44 | 0.38 | -0.78 | -0.83 | 0.06 |
| 40 | ROCKET | MISSILE | ASTRONAUT | CHEESE | SINK | 0.35 | 0.36 | -0.86 | -0.6 | -0.02 |
| 41 | SHIP | CANOE | SAILOR | GLAND | UMBRELLA | 0.34 | 0.59 | -1.08 | -1.17 | -0.25 |
| 42 | PLATE | TRAY | NAPKIN | ANKLE | CHAUFFEUR | 0.39 | 0.32 | -0.99 | -0.96 | 0.07 |
| 43 | CROWN | HAT | KING | NOSE | SHOVEL | 0.63 | 0.2 | -1.14 | -0.9 | 0.43 |
| 44 | HURRICANE | BLIZZARD | FLOOD | BADGE | FOSSIL | 0.47 | -0.1 | -0.84 | -0.84 | 0.57 |
| 45 | LOCKER | CLOSET | JERSEY | PAINT | SPY | 0.76 | 0.56 | -0.86 | -0.69 | 0.2 |
| 46 | HEARSE | LIMOUSINE | GRAVEYARD | EYE | KITCHEN | 0.34 | -0.01 | -0.8 | -0.68 | 0.35 |
| 47 | NEEDLE | PIN | THREAD | HYDRANT | WRIST | 0.45 | 0.12 | -1.12 | -0.98 | 0.33 |
| 48 | CELEBRITY | PLUMBER | FILM | FORTRESS | NECTAR | 0.54 | 0.27 | -0.73 | -0.84 | 0.28 |
| 49 | MONKEY | BEAR | BANANA | HAMMER | TOOTH | 0.15 | 0.49 | -0.79 | -1 | -0.33 |
| 50 | OVEN | MICROWAVE | PAN | CONVICT | SCREEN | 0.31 | 0.2 | -0.18 | -0.83 | 0.1 |
| 51 | SKYSCRAPER | TOWER | ELEVATOR | HEART | HITCHHIKER | 0.51 | 0.17 | -0.6 | -0.48 | 0.34 |
| 52 | SURGEON | BUTCHER | KIDNEY | DYNAMITE | GALAXY | 0.3 | 0.29 | -0.29 | -0.6 | 0.01 |
| 53 | CHISEL | KNIFE | SCULPTURE | HATCH | MIRROR | 0.05 | 0.3 | -1.06 | -0.88 | -0.25 |
| 54 | SHOE | GLOVE | FOOT | TIGER | WALL | 0.18 | 0.47 | -1.08 | -0.93 | -0.3 |
| 55 | FOOTBALL | BASEBALL | QUARTERBACK | NECKLACE | PLANT | 0.36 | 0.09 | -0.95 | -1.02 | 0.27 |
| 56 | ENVELOPE | PARCEL | STAMP | MUSCLE | YOGURT | −0.16 | 0.22 | -0.54 | -0.47 | -0.38 |
| 57 | JELLY | MARMALADE | JAR | BOOK | NAIL | 0.45 | 0.55 | -1.01 | -0.68 | -0.1 |
| 58 | SALT | PEPPER | SEA | KNUCKLE | SAW | −0.15 | 0.34 | -0.83 | -0.7 | -0.49 |
| 59 | CASKET | BOX | GRAVE | JEWEL | STREET | 0.45 | -0.26 | -0.78 | -0.92 | 0.72 |
| 60 | FLY | ANT | WINGS | CEREAL | CONCRETE | 0.53 | 0.25 | -1.11 | -1.12 | 0.28 |
| 61 | DOOR | GATE | KNOB | FLAG | LIQUID | 0.41 | 0.07 | -1.12 | -1.16 | 0.34 |
| 62 | PENGUIN | GOOSE | ICE | BRICK | HEAD | 0.44 | 0.79 | -0.67 | -0.9 | -0.36 |
| 63 | CAKE | DONUT | CANDLE | ACTRESS | BROCHURE | 0.41 | 0.65 | -0.77 | -0.61 | -0.24 |
| 64 | OWL | HAWK | MOON | CIRCUIT | DIARY | 0.33 | 0.29 | -0.95 | -0.77 | 0.04 |
| 65 | HOSE | TUBE | WATER | MOTHER | RODEO | 0.41 | 0.12 | -1 | -0.99 | 0.28 |
| 66 | SWEATER | HOODIE | MITTENS | BATHROOM | CHALKBOARD | 0.53 | 0.03 | -0.95 | -1 | 0.5 |
| 67 | SEDAN | BIKE | SEATBELT | COTTON | SHRIMP | 0.2 | 0.19 | -0.9 | -1 | 0.01 |
| 68 | PENCIL | PEN | ERASER | FLUTE | SHEEP | 0.38 | 0.22 | -1.07 | -1.04 | 0.16 |
| 69 | BACKPACK | SUITCASE | NOTEBOOK | BUTTER | PAINTING | 0.01 | 0.37 | -1.09 | -1.09 | -0.35 |
| 70 | SEAGULL | DUCK | PIER | BEDROOM | POWDER | 0.25 | 0.83 | -0.64 | -0.44 | -0.58 |
| 71 | VENOM | POISON | SNAKE | GRAFFITI | RASPBERRY | 0.43 | 0.33 | -1.04 | -1.1 | 0.11 |
| 72 | TORTILLA | BREAD | BEANS | COLD | WIRE | 0.43 | 0.5 | -0.75 | -0.96 | -0.07 |
| 73 | COMPUTER | TABLET | MOUSE | ATHLETE | COUCH | 0.23 | 0.5 | -0.58 | -0.66 | -0.28 |
| 74 | CHAIR | SOFA | LEGS | ANCHOVY | BALL | 0.47 | 0.54 | -1.16 | -1.18 | -0.07 |
| 75 | BISCUITS | TOAST | GRAVY | DANCE | SNAIL | 0.3 | 0.34 | -1.09 | -1.17 | -0.03 |
| 76 | FLOUR | CORNMEAL | DOUGH | BUTTON | SMOG | 0.32 | -0.04 | -1.1 | -1.13 | 0.36 |
| 77 | SHIRT | BLOUSE | COLLAR | BRIDGE | POOL | 0.46 | 0.21 | -0.52 | -0.62 | 0.25 |
| 78 | PATHWAY | SIDEWALK | GRAVEL | BABYSITTER | TYPEWRITER | 0.52 | -0.05 | -0.98 | -1.1 | 0.58 |
| 79 | SNOW | RAIN | SLED | CEMETERY | NOVEL | 0.46 | 0.24 | -0.64 | -0.85 | 0.22 |
| 80 | CITY | VILLAGE | AIRPORT | NECK | WHALE | 0.55 | -0.08 | -0.87 | -0.82 | 0.63 |

Table F.2: Experiment 4 Lexical and Orthographic Properties of Taxonomic and Thematic Targets

| Index | Standard | Length | | Frequency | | Neighborhood | | Bigram | |
|---|---|---|---|---|---|---|---|---|---|
| | | Tax. | Them. | Tax. | Them. | Tax. | Them. | Tax. | Them. |
| 1 | CIGARETTES | 7 | 5 | 18.7 | 15.3 | 0 | 1.1 | 229.3 | 219.2 |
| 2 | WAITRESS | 10 | 10 | 3.8 | 33.1 | 0 | 0 | 99.6 | 449.6 |
| 3 | BEE | 9 | 5 | 5.2 | 20.8 | 0 | 64.6 | 499.7 | 820.7 |
| 4 | TOOTHBRUSH | 4 | 5 | 5.7 | 1.2 | 150.2 | 2.4 | 1677.7 | 1114.3 |
| 5 | CUP | 4 | 3 | 30.7 | 89.5 | 3.7 | 45.6 | 1110.3 | 294.8 |
| 6 | SKI | 9 | 4 | NA | 55.6 | 0 | 15.3 | 129.5 | 5651.2 |
| 7 | DOG | 3 | 4 | 43.3 | 28.2 | 132.3 | 54.8 | 1462.1 | 1749.3 |
| 8 | RECEPTIONIST | 7 | 9 | 9.6 | 102.9 | 2.6 | 0.1 | 927.1 | 350.3 |
| 9 | RABBI | 6 | 6 | 3.6 | 24.5 | 0 | 0 | 714 | 1499.3 |
| 10 | CABLE | 4 | 10 | 8.2 | 104 | 58.2 | 0 | 2397.3 | 819.5 |
| 11 | GOAT | 7 | 4 | 7.3 | 69.4 | 0 | 68.2 | 137.8 | 1391.9 |
| 12 | FIELD | 5 | 6 | 128.1 | 28 | 80.8 | 5.6 | 3111.1 | 1716.4 |
| 13 | MINT | 8 | 6 | 0.4 | 57.9 | 0 | 5.9 | 253.5 | 572.3 |
| 14 | COOKIE | 3 | 9 | 12.9 | 13.4 | 21.1 | 0 | 138.4 | 253.6 |
| 15 | HORNET | 4 | 7 | 2.5 | 0.4 | 7.5 | 0.7 | 1275 | 1215.6 |
| 16 | LAWNMOWER | 8 | 4 | 4.5 | 37.6 | 0 | 49.1 | 262.3 | 1014.5 |
| 17 | VINEYARD | 7 | 4 | 5.5 | 75.6 | 0 | 50.1 | 315.5 | 4733.2 |
| 18 | PANDA | 6 | 6 | NA | 6.2 | 0 | 0 | 884.3 | 348.4 |
| 19 | BEER | 5 | 5 | 21.5 | 373.5 | 3.8 | 15.3 | 1854.3 | 1250.5 |
| 20 | SPOON | 5 | 4 | 1.2 | 20.6 | 0 | 16.5 | 800.5 | 1713.4 |
| 21 | HORSE | 3 | 5 | 18.7 | 87 | 26.1 | 25.1 | 211.7 | 1201.9 |
| 22 | CAMEL | 8 | 6 | 4 | 40.5 | 0 | 0 | 362.1 | 828.3 |
| 23 | BLANKET | 9 | 6 | 1.7 | 14.5 | 5.7 | 2.1 | 877.3 | 593.2 |
| 24 | TURKEY | 7 | 8 | 31.1 | 4.2 | 1.2 | 0.9 | 613.8 | 1769.5 |
| 25 | SHOTGUN | 6 | 5 | 15.1 | 29.7 | 1.6 | 66.5 | 586.2 | 3226 |
| 26 | PACKAGE | 5 | 8 | 2.8 | 15.2 | 1.4 | 2.1 | 1072.2 | 602.4 |
| 27 | SHAMPOO | 6 | 6 | 1.9 | 18.1 | 3.4 | 58.9 | 396.9 | 2287.2 |
| 28 | TOE | 6 | 6 | 51.8 | 1.1 | 2.9 | 0.4 | 2021.8 | 898.6 |
| 29 | TRUCK | 3 | 7 | 65.1 | 3.2 | 597.9 | 2.8 | 2755.8 | 1242.9 |
| 30 | BICYCLE | 3 | 6 | 274.9 | 9.5 | 168 | 0.1 | 1786.5 | 666.7 |
| 31 | BOOTS | 5 | 8 | 19 | 0.4 | 8.1 | 0 | 765.8 | 337.7 |
| 32 | SAXOPHONE | 4 | 4 | 2.5 | 6.7 | 40.4 | 0 | 2758.5 | 80.3 |
| 33 | OYSTER | 7 | 5 | 1 | 5.4 | 0 | 3.9 | 241.3 | 1699.9 |
| 34 | CRIB | 3 | 4 | 254.4 | 191.2 | 42.7 | 1.2 | 484.3 | 811.3 |
| 35 | POLICE | 7 | 9 | 0.7 | 2.3 | 4 | 0 | 424.2 | 111.8 |
| 36 | RABBIT | 8 | 6 | 3.7 | 2.6 | 0 | 2.6 | 417.5 | 779.4 |
| 37 | MILK | 8 | 3 | 3 | 23.3 | 0 | 128.6 | 234.2 | 1566.7 |
| 38 | BOTTLE | 3 | 6 | 1954.3 | 21.4 | 95.6 | 0 | 2766 | 500.3 |
| 39 | BIRD | 3 | 4 | 10.5 | 13.6 | 280.3 | 94.9 | 502.9 | 2975.5 |
| 40 | ROCKET | 7 | 9 | 27.3 | 1 | 0.3 | 0 | 791.3 | 120.1 |
| 41 | SHIP | 5 | 6 | 3.9 | 5.9 | 4.3 | 1.4 | 432.4 | 464.7 |
| 42 | PLATE | 4 | 6 | 21 | 4.9 | 6 | 0 | 658.3 | 258.4 |
| 43 | CROWN | 3 | 4 | 54.5 | 91.7 | 409.1 | 59.7 | 4629.3 | 1483.4 |
| 44 | HURRICANE | 8 | 5 | 2.6 | 15.6 | 0 | 158.3 | 158.7 | 806.4 |
| 45 | LOCKER | 6 | 6 | 10.5 | 13 | 55.9 | 0 | 796.4 | 607.8 |
| 46 | HEARSE | 9 | 9 | 2.7 | 4 | 0 | 0 | 714.8 | 192.8 |
| 47 | NEEDLE | 3 | 6 | 13.6 | 11.2 | 13.6 | 64.3 | 111.8 | 866.3 |
| 48 | CELEBRITY | 7 | 4 | 2.1 | 76.5 | 1.3 | 47.8 | 788.3 | 1526.5 |
| 49 | MONKEY | 4 | 6 | 63.8 | 4.3 | 73.6 | 0 | 2940.5 | 481.2 |
| 50 | OVEN | 9 | 3 | 2.1 | 26.7 | 0 | 156.4 | 170.2 | 1730.2 |
| 51 | SKYSCRAPER | 5 | 8 | 49 | 8.9 | 41.1 | 0 | 2327.6 | 251.4 |
| 52 | SURGEON | 7 | 6 | 5.6 | 4.9 | 0.1 | 0 | 1415.6 | 433 |
| 53 | CHISEL | 5 | 9 | 38.8 | 22 | 0 | 0 | 266.2 | 208.8 |
| 54 | SHOE | 5 | 4 | 4.9 | 101.1 | 6.8 | 30.4 | 793.6 | 1986.8 |
| 55 | FOOTBALL | 8 | 11 | 6.5 | NA | 0 | 0 | 174.6 | 30.3 |
| 56 | ENVELOPE | 6 | 5 | 8.4 | 13.8 | 0 | 2.7 | 756 | 1207.1 |
| 57 | JELLY | 9 | 3 | 2.6 | 11.8 | 0 | 85.3 | 188.9 | 651.5 |
| 58 | SALT | 6 | 3 | 7 | 166 | 0.7 | 152.1 | 1990.7 | 1001.7 |
| 59 | CASKET | 3 | 5 | 78.8 | 31.2 | 23.7 | 8.6 | 232.8 | 1091.1 |
| 60 | FLY | 3 | 5 | 4 | 29.6 | 4303.6 | 6.5 | 14878.9 | 464.6 |
| 61 | DOOR | 4 | 4 | 50.9 | 3.7 | 61.9 | 381.9 | 768 | 957.7 |
| 62 | PENGUIN | 5 | 3 | 6.2 | 54.4 | 11.9 | 4.1 | 1854.2 | 61.6 |
| 63 | CAKE | 5 | 6 | NA | 8 | 0 | 21.7 | 943.2 | 1682.9 |
| 64 | OWL | 4 | 4 | 4.2 | 54.8 | 1 | 31.7 | 2092.1 | 3092.8 |
| 65 | HOSE | 4 | 5 | 15.2 | 447.9 | 5.2 | 55.2 | 125.1 | 3313.5 |
| 66 | SWEATER | 6 | 7 | NA | 0.8 | 0.3 | 3.3 | 286.5 | 1086.7 |
| 67 | SEDAN | 4 | 8 | 8.3 | NA | 177.4 | 0 | 1975.9 | 350 |
| 68 | PENCIL | 3 | 6 | 19.8 | 0.3 | 64 | 0.8 | 702.9 | 1848.7 |
| 69 | BACKPACK | 8 | 8 | 13 | 7.7 | 0 | 0 | 363 | 213.4 |
| 70 | SEAGULL | 4 | 4 | 9.9 | 5.8 | 9.9 | 3.1 | 1469.1 | 915.8 |
| 71 | VENOM | 6 | 5 | 12.6 | 15.1 | 66.6 | 6.8 | 838.4 | 147.8 |
| 72 | TORTILLA | 5 | 5 | 77 | 18.3 | 30.2 | 28.2 | 1289.3 | 1902.1 |
| 73 | COMPUTER | 6 | 5 | 2.9 | 8.4 | 16.2 | 71.3 | 871.7 | 3653 |
| 74 | CHAIR | 4 | 4 | 21.4 | 117.7 | 32 | 32.9 | 989 | 610.6 |
| 75 | BISCUITS | 5 | 5 | 15.4 | 3.9 | 20.7 | 31.2 | 1086.3 | 863.3 |
| 76 | FLOUR | 8 | 5 | NA | 10.9 | 0 | 15.2 | 678.5 | 2330.5 |
| 77 | SHIRT | 6 | 6 | 8.9 | 19.1 | 0 | 12.5 | 1079.5 | 900 |
| 78 | PATHWAY | 8 | 6 | 6.2 | 11 | 0 | 14.4 | 101.8 | 587.3 |
| 79 | SNOW | 4 | 4 | 74.2 | 0.8 | 35.2 | 11 | 1825 | 554.8 |
| 80 | CITY | 7 | 7 | 140 | 53.8 | 0.4 | 0 | 706.4 | 389.5 |

# References


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

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 information in concepts. *Memory and 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.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version*, *1*(7), 1–23.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116-131.

Canessa, N., Borgo, F., Cappa, S. F., Perani, D., Falini, A., Buccino, G., ... Shallice, T. (2007). The different neural correlates of action and functional knowledge in semantic memory: an fmri study. *Cerebral Cortex*, *18*(4), 740–751.

Chen, Q., Li, P., Xi, L., Li, F., Lei, Y., & Li, H. (2013). How do taxonomic versus thematic relations impact similarity and difference judgments? An ERP study. *International Journal of Psychophysiology*, *90*(2), 135–142.

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.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, *41*(1), 60–74.

Denney, N. W. (1974). Evidence for developmental changes in categorization criteria for children and adults. *Human Development*, *17*(1), 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), 131–142.

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). 8 thematic thinking: The apprehension and consequences of thematic relations. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 249–294.

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

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

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 M. Hahn & S. C. Stones (Eds.), *Proceedings of the 21$^{st}$ annual meeting of the cognitive science society* (pp. 179–184). Mahwah, NJ: Erlbaum.

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* (pp. 111–148). Brussels, Belgium: 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), 1454–1464.

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

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*(1), 19–21.

Hagoort, P., Brown, C. M., & Swaab, T. Y. (1996). Lexical–semantic event-related potential effects in patients with left hemisphere lesions and aphasia, and patients with right hemisphere lesions without aphasia. *Brain*, *119*(2), 627–649.

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*(6), 803–831.

Hendrickson, A., Navarro, D. J., & Donkin, C. (2015). Quantifying the time course of similarity. In D. C. Noelle et al. (Eds.), *Proceedings of the 37$^{th}$ 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 se-

mantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*(3), 464–482.

Ince, E., & Christman, S. D. (2002). Semantic representations of word meanings by the cerebral hemispheres. *Brain and Language*, *80*(3), 393–420.

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.

Jenkins, J. J., & Russell, W. A. (1952). Associative clustering during recall. *The Journal of Abnormal and Social Psychology*, *47*(4), 818–821.

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), 57–65.

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*(1), 54–69.

Kalénine, S., & Bonthoux, F. (2008). Object manipulability affects childrens and adults conceptual processing. *Psychonomic bulletin & review*, *15*(3), 667–672.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133–156.

Khateb, A., Michel, C. M., Pegna, A. J., O'Dochartaigh, S. D., Landis, T., & Annoni, J.-M. (2003). Processing of semantic categorical and associative relations: An ERP mapping study. *International journal of psychophysiology*, *49*(1), 41–55.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Kurtz, K. J., & Gentner, D. (2001). Kinds of kinds: Sources of category coherence. In J. Moore & S. Keith (Eds.), *Proceedings of the 23$^{rd}$ annual conference of the cognitive science society* (p. 522-527). Mahwah, NJ: Erlbaum.

Kurtz, K. J., & Gentner, D. (in preparation). Kinds of kinds: Sources of category coherence.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, *11*(2), 99–116.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package lmertest. *R package version*, *2*(0).

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176–186.

Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F., & Jin, Z. (2014). A direct comparison of active and passive amplification electrodes in the same amplifier system. *Journal of neuroscience methods*, *235*, 298–307.

Laszlo, S., & Sacchi, E. (2015). Individual differences in involvement of the visual object recognition system during visual word recognition. *Brain and language*, *145*, 42–52.

Lawson, R., Chang, F., & Wills, A. J. (2017). Free classification of large sets of everyday objects is more thematic than taxonomic. *Acta psychologica*, *172*, 26–40.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of experimental psychology: General*, *130*(1), 3–28.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any erp experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157.

Maguire, M. J., Brier, M. R., & Ferree, T. C. (2010). Eeg theta and alpha responses reveal qualitative differences in processing taxonomic versus thematic semantic relationships. *Brain and language*, *114*(1), 16–25.

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

Markman, E. M., Cox, B., & Machida, S. (1981). The standard object-sorting task as a measure of conceptual organization. *Developmental Psychology*, *17*(1), 115–117.

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

Medler, D., & Binder, J. (2005). Mcword: An on-line orthographic database of the english language. http://www.neuro.mcw.edu/mcword/.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of experimental psychology: General*, *141*(4), 601–609.

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological bulletin*, *143*(5), 499–520.

Murphy, G. L. (2001). Causes of taxonomic sorting by adults: A test of the thematic-to-taxonomic shift. *Psychonomic Bulletin & Review*, *8*(4), 834–839.

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

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological review*, *105*(2), 299–324.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175–240.

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). Psychopypsychophysics software in python. *Journal of neuroscience methods*, *162*(1), 8–13.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

Rey, E., & Berger, C. (2001). Four-and five-year-old children's categorization: Sen-

sitivity to constraints on word meaning and influence of stimulus presentation in a forced-choice paradigm. *Cahiers de psychologie cognitive*, *20*(1-2), 63–85.

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.

Rugg, M. D., & Nagy, M. E. (1989). Event-related potentials and recognition memory for words. *Electroencephalography and clinical neurophysiology*, *72*(5), 395–406.

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.

Sacchi, E., & Laszlo, S. (2016). An event-related potential study of the relationship between n170 lateralization and phonological awareness in developing readers. *Neuropsychologia*, *91*, 415–425.

Sachs, O., Weis, S., Krings, T., Huber, W., & Kircher, T. (2008). Categorical and thematic knowledge representation in the brain: Neural correlates of taxonomic and thematic conceptual relations. *Neuropsychologia*, *46*(2), 409–418.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., ... 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.

Shaoul, C., & Westbury, C. (2006). Usenet orthographic frequencies for 111,627 english words (2005–2006). *Edmonton, AB: University of Alberta (downloaded from http://www. psych. ualberta. ca/˜ westburylab/downloads/wlfreq. download. html)*.

Sharp, D., Cole, M., Lave, C., Ginsburg, H. P., Brown, A. L., & French, L. A. (1979). Education and cognitive development: The evidence from experimental research. *Monographs of the society for research in child development*, 1–112.

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.-L., & Clark, J. M. (1996). Choosing category or complementary relations: Prior tendencies modulate instructional effects. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*(4), 356–370.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3–22.

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

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402–433.

Stites, M. C., & Laszlo, S. (2015). How do random effects structures impact LMER outcomes in an ERP study? In *Psychophysiology* (Vol. 52, pp. S116–S116).

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, *45*(2), 1–31.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138.

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.

Wamain, Y., Pluciennicka, E., & Kalénine, S. (2015). A saw is first identified as an object used on wood: ERP evidence for temporal differences between thematic and functional similarity relations. *Neuropsychologia*, *71*, 28–37.