# When Do Words Promote Analogical Transfer?

*Ji Y. Son[1], Leonidas A. A. Doumas[2], and Robert L. Goldstone[3]*

## Abstract:

The purpose of this paper is to explore how and when verbal labels facilitate relational reasoning and transfer. We review the research and theory behind two ways words might direct attention to relational information: (1) words generically invite people to compare and thus highlight relations (the Generic Tokens [GT] hypothesis), and/or (2) words carry semantic cues to common structure (the Cues to Specific Meaning [CSM] hypothesis). Four experiments examined whether learning Signal Detection Theory (SDT) with relational words fostered better transfer than learning without relational words in easily alignable and less alignable situations (testing the GT hypothesis) as well as when the relational words matched and mismatched the semantics of the learning situation (testing the CSM hypothesis). The results of the experiments found support for the GT hypothesis because the presence of relational labels produced better transfer when two situations were alignable. Although the CSM hypothesis does not explain how words facilitate transfer, we found that mismatches between words and their labeled referents can produce a situation where words hinder relational learning.

## Keywords:

analogical reasoning, transfer, problem solving, relation learning, similarity

The authors wish to express thanks to John Hummel, David Landy, and Linda Smith for helpful suggestions on this work. This research was funded by the Department of Education, Institute of Education Sciences grant R305H050116, and the National Science Foundation REESE grant 0910218. Correspondence concerning this article should be addressed to Ji Son, Department of Psychology, California State University, Los Angeles, 5151 State University Drive, Los Angeles, California, or by email at json2@calstatela.edu.

[1]California State University, Los Angeles; [2]University of Hawaii; [3]Indiana University



http://dx.doi.org/10.7771/1932-6246.1079

Although there is much debate on the connection between language and thought (e.g., Whorf, 1956; Gumperz & Levinson, 1991; see Gentner & Goldin-Meadow, 2003 for a review), there is general agreement that words are useful for learning new concepts. For example, even when words and meanings are unknown, as is the case with very young children or with the use of novel words, linguistic labels facilitate learning (e.g., Lupyan, Rakison, & McClelland, 2007). Research with young children has shown that words facilitate category learning more than non-linguistic cues (Balaban & Waxman, 1997; Waxman & Booth, 2003; Waxman & Markow, 1995). In addition, adults show faster learning and more robust retention when novel categories are associated with linguistic labels relative to non-linguistic cues (Lupyan, 2008).

To date, a great deal of work has focused on the general phenomenon of the usefulness of words in learning situations, but comparatively little empirical work has focused on the reason for this usefulness. What makes words so useful in learning contexts?

Relational reasoning—reasoning based on the relations between objects or features of objects—is a rich domain for looking at potential cognitive benefits of words because it is a highly demanding cognitive skill and many studies have shown that words make the task of relational reasoning easier. Relational thinking plays a central role in human cognition. It underlies our ability to perceive and understand the spatial relations among an object's parts (Hummel, 2000; Hummel & Biederman, 1992; Hummel & Stankewicz, 1996), comprehend arrangements of objects in scenes (Green & Hummel, 2006; Markman & Gentner, 1993; Richland, Morrison, & Holyoak, 2006), and comprehend abstract analogies between otherwise very different situations or systems of knowledge (e.g., between the structure of the solar system and the structure of the atom; Gentner, 1983; Gick & Holyoak, 1980, 1983; Holyoak & Thagard, 1995). However, despite its centrality in human cognition, relational thinking is cognitively demanding. In contrast to simpler reasoning about object features or single objects, reasoning about relations requires more working memory and makes greater demands on attention (e.g., Halford, Wilson, & Phillips, 1998; Hummel & Holyoak, 1997).

There are at least two reasons for the greater cognitive demands of relational thinking. First, relations are properties that hold over *collections* of objects rather than single objects in isolation (Doumas, Hummel, & Sandhofer, 2008). The relation *same-shape* (*x*, *y*), for instance, is a property of any two objects with the same shape, but not of any specific *x* or *y*. Two identical shoes are the *same-shape* in exactly the same way that two triangles are the *same-shape*, although *same-shape* is not a feature of either any single shoe or any specific triangle. If one of the identical shoes were paired with a cup, the sameness relation would disappear. By contrast, an object property such as color remains a property of an object whether it is paired with a green object or another red object. Because relations are less spatially and temporally stable than the features of single objects, they are easily overshadowed by more salient object features such as a color. Even highly perceptual

relations such as spatial relations (e.g., above, in, under; Loewenstein & Gentner, 2005) are less stable than featural qualities (e.g., has a star on it).

Second, relational reasoning is cognitively demanding because representing structure is complex (Doumas & Hummel, 2005; Doumas et al., 2008; Gentner, 1983; Hummel & Holyoak, 1997, 2003). It requires representing (1) the relation and (2) the objects involved in the relation independently of one another, and (3) the bindings of these objects to particular relational roles (Doumas & Hummel, 2005). For example, representing the relation *bigger* (shoe, cup) requires representing the relation *bigger* and the two objects, the shoe and cup, independently of one another. Consequently, we understand that in the expression *bigger* (shoe, cup), the shoe is *larger* and the cup is *smaller*, and that in the expression *bigger* (cup, shoe), the same elements play the opposite roles (the cup is *larger* and the shoe is *smaller*).

The structure inherent in mental representations makes them very powerful for the purposes of reasoning (e.g., Doumas et al., 2008; Hummel & Holyoak, 2003), but this power comes at a cost. Considerable empirical evidence indicates that adults process concrete features and concrete categories faster than relational ones (Gentner & Kurtz, 2005; Kurtz & Gentner, 2001) and relational categories seem to be acquired later in development as well (Hall & Waxman, 1993; Keil & Batterman, 1984; Smith, Rattermann, & Sera, 1988).

Studies that show how relational language enables relational reasoning typically come from developmental research. These studies often teach children relational categories with or without linguistic labels and then test for generalization. For example, in a series of studies Kotovsky and Gentner (1996) investigated how labels affected four-year-old children's sensitivity to relations such as symmetry and monotonicity. In Kotovsky and Gentner's studies, children were taught triads of shapes in a symmetric (i.e., xXx) or monotonically increasing pattern (i.e., xXX). The symmetric cards were called "even" and the increasing cards were called "more-and-more." Then, children were asked to determine which of two triads was the best match to a target triad, where the best match involved relations with different dimensions (e.g., a size-based pattern of xXx matched black-white-black) or different dimension values (e.g., xXx to OoO). Children who learned the relational labels were able to make relational choices more frequently than children who did not. Kotovsky and Gentner (1996) suggest that acquiring a word for the xXx-patterned triads allowed children to notice the relational similarities among them.

Often experiments regarding words and relational reasoning are designed to demonstrate that words facilitate relational reasoning but they do not allow us to distinguish between different ways words might help. By one account, favored by Kotovsky and Gentner (1996), the word "even" cues children to compare different triads and to extract the subtle relational similarity, thus directing their attention. However, there is an alternative possibility that the labels "even" and "more-and-more" help direct attention by virtue of

their semantics. Perhaps the meanings of these labels, more than the mere act of giving common labels to situations, helps children attend to relational information over other sources of similarity. By this account, "even" suggests balance or symmetry, which allows this aspect of "xXx" to be emphasized.

The purpose of this paper is to explore how and when verbal labels facilitate relational reasoning. First, we review the research and theory behind two ways words might direct attention to relational information: (1) words invite learners to compare, highlight, and represent relations (the Generic Tokens [GT] hypothesis), and/or (2) words carry semantic cues to common structure (the Cues to Specific Meaning [CSM] hypothesis). Given these two (non-mutually exclusive) possibilities, we can make predictions about when words boost relational learning. Four experiments examine these predictions.

## Words as Generic Tokens (GTs) to Represent Difficult Concepts

We have already discussed how relations are difficult to process because they require more representational capacity and more processing resources than simple objects in isolation. The crux of the GT theory is that associating a simple symbol (i.e., a word) with a complex situation (i.e., a relation) might make it easier to access or think about the situation. Linguistic labels, and other useful symbols, are typically stable across contexts because they are relatively unchanged by idiosyncratic differences in context (e.g., tokens of the word "dog" said at different times are highly similar) and are non-iconic to their referents (e.g., the word "dog" does not particularly look like a dog). Because words enjoy the combination of being relatively context-free and non-iconic, their GT qualities allow them to stand for potentially subtle relations. When relations are tied to an object-like word, they might seem more concrete. However, it is important to note that this function of words does not necessitate that all word and language processing is inherently symbolic and propositional. In fact, there are theories about the mechanism of language processing (e.g., Elman, 1995) that suggest that language has the appearance of being symbolic and context-free even though the underlying mechanism may be dynamic, continuous, and sensitive to context in real-time (see also Clark, 1998; Dennett, 1991; Spivey, 2007).

Words as GTs may stabilize highly variable perceptual experiences—a function particularly useful in learning relational concepts. Having the same label for similar relations can implicitly induce comparison (Brown, 1958; Gentner & Namy, 2004; Namy, 2001), a powerful mechanism for structural abstraction (Dixon & Bangert, 2004; Doumas & Hummel, 2005; Doumas et al., 2008; Gentner, 2005; Gentner & Namy, 1999; Gick & Holyoak, 1983). Symbolic juxtaposition (Gentner & Medina, 1998)—applying the same word to different instances—is a natural cue to compare instances partly because of our conventional and ubiquitous practice of labeling categories.

Although symbolic juxtaposition might suggest that words are only effective when applied to multiple situations, even having one labeled instance may be effective because of our general convention of labeling concepts/categories. Some might consider that the very existence of a word implies the existence of a category/concept (Quine, 1960) and indeed cross-cultural research has suggested that concepts such as *exact numerosity* (Pica et al., 2004) or particular spatial categories (Bowerman & Choi, 2003) are used and acquired because of the arbitrary labels that stand for these ideas. Even cases of limited "language" training, such as laboratory-raised nonhuman primates, suggest that understanding numerosity (Boysen & Bernston, 1989) and relational similarity (Thompson & Oden, 1993; Thompson, Oden, & Boysen, 1997) are mediated by symbolic tokens.

Comparison may drive the discovery of relational similarity but words provide stable tokens to represent any newly discovered similarities. In other words, once acquired, words provide a new level of object-like computation over the actual relations (Clark, 1998). Support for this generic function of words comes from Richard Catrambone's research on how words seem to help novices chunk newly learned procedures into meaningful and better remembered groups (Catrambone, 1996, 1998). Also, separate words applied to subtly different objects help differentiate objects that are difficult to discriminate (Goldstone, 1994). These results suggest that words have generic properties, apart from their meanings, that may foster more efficient encoding and categorization.

## Words as Cues with Specific Meanings (CSM)

Thinking about words as generic tokens places the emphasis on the ability of words to efficiently capture complex ideas and make manipulation of these ideas easier. However, language probably derives much of its power from connections to real experiences. When known words are used, children also seem to show consistent benefits in detecting relational similarities. An experiment reported by Rattermann and Gentner (1998) showed that brief training with known words significantly increased relational responding in children compared to children who did not receive word training. In their task where toddlers could make matches by relative size similarity or object similarity, children typically made object matches. However, when objects were named with labels that children of this age spontaneously use to mark monotonic size changes (e.g., *daddy, mommy, baby*), children were able to make relational matches. However, this benefit was not found when objects were labeled with arbitrary words (*jiggy, gimli, fantan*). This result indicates that associations between words and past experiences significantly influence whether words can highlight relations. Likewise, Loewenstein and Gentner (2005) found that some sets of words promote relational responding more effectively than others. Labeling locations in a three-tiered box as {*top, middle, bottom*} promoted children's ability to use spatial information more effectively than the labels {*on, in, under*}. Both studies suggest that the

specific content of the words, or the relational framework they invoke, matters for providing cognitive benefits.

As GTs, *mommy* and *jiggy* are essentially equivalent (both are equally good symbolic tokens). However, if words are thought of as CSMs, not all words are predicted to be equally beneficial. The fact that *mommy* works well as a relational label may be the consequence of *mommy* having rich associations to experiences that suggest medium size (especially in the context of *daddy* and *baby*). However, the acquisition of relational meanings is not at all straightforward. Hall and Waxman (1993) have attempted to teach children a relational word by providing a definition. They taught children an arbitrary word, *murvil* (with the equivalent meaning as the word "passenger"), and even defined it for them (i.e. "This is a murvil because it is riding in a car"). Despite the provision of a relational word and an explicitly relational definition, children were not able to learn that *murvils* are any and all dolls that sit in cars. Instead, children interpret the label *murvil* as the name of dolls that look like the doll that was named. This suggests that it is not only difficult to learn the *murvil* category (how to generalize the label) but also to learn the explicitly provided relational concept. Because of the label's lack of rich associations to other words and experiences, there is no relational benefit from using an arbitrarily *defined* word.

There might be a continuum of words (and their meanings) from semantically empty (i.e., *jiggy, murvil*) to semantically rich and matching the referent (e.g., *daddy* to refer to something large) and some in between (e.g., semantically rich but not matching, such as using the word *daddy* to refer to something small). We focus our research (with adults) on the semantically meaningful end of the spectrum, looking at semantically meaningful words that can either match or mismatch their referents. Semantically mismatching words may be a better control for matching words since they control for the meaningfulness, but not the appropriateness, of the label. Also, it is possible that there are additional memory demands from having to learn a nonsense term like *jiggy*.

## Rationale of Experiments

The majority of the experiments reviewed above illustrate difficulties that children have with relational similarity, but even for adult learners, novel abstract relations are difficult to acquire (e.g., Goldstone & Sakamoto, 2003). This paper examines the dual role of words, as GTs and CSMs, in adult relational reasoning in order to test how linguistic labels can affect relational reasoning. Our central question concerns how and when words confer benefits in relational reasoning. Is it because labels act as GTs that are easier to manipulate and remember than entire relational systems? Or, is it because the specific semantic content of the words provides clues to a situation's underlying relational structure? We conducted four experiments to investigate how words confer benefits in relational reasoning. In each experiment, participants were presented with a tutorial, a corresponding tutorial quiz

followed by a structurally similar transfer situation and a corresponding transfer quiz. Each experiment tested two conditions: a Word condition with relational labels included in the tutorial situation, and a Control condition without those labels.

The behavior of interest was the ability of learners to utilize relational knowledge from the tutorial situation in a new transfer context. The underlying system of relations that participants learned and transferred was Signal Detection Theory (SDT). SDT is a way of understanding decision making that involves uncertainty. Typically an SDT situation involves some sort of evidence upon which a categorical decision is made, the decision itself (e.g., "yes/no," "in/out," "healthy/sick," "signal/noise"), and the actual status of the decided entity (whether it was actually signal or noise). Although the evidence is informative as to whether something is signal or noise, it is often imperfect so the decision has some uncertainty. Under these conditions, there are ways to maximize the likelihood of making hits (deciding "signal" when the signal is actually present) and minimizing false alarms (deciding "signal" when the signal is not present). A parallel expression of the same idea is to maximize correct rejections (deciding "noise" when signal is not present) and minimizing misses (deciding "noise" when the signal is actually present). SDT provides an informative framework for understanding a variety of decision-making situations under uncertainty. The relational words that we used were: *evidence, target* (signal), *distracter* (noise), *hit, miss, correct rejection,* and *false alarm*. We did not use the traditional SDT terms *signal* and *noise* because those are grounded in the historical development of SDT that is probably not intuitive to our participants.

We crossed two aspects of similarity in order to test the effects of GTs and CSMs as well as their interactions. If words are GTs that represent relations efficiently, then regardless of the semantics of the relational labels, they should provide a benefit. Especially when working together with comparison (Doumas et al., 2008; Markman & Gentner, 1993) to drive the discovery of relational similarity, the presence of GTs that can represent these extracted relations may be beneficial. More alignable (relationally comparable) SDT stories will benefit from GTs more than less alignable SDT stories. To test this prediction, Experiments 1 and 2 used tutorial and transfer situations that were more alignable and Experiments 3 and 4 contained situations that were less alignable (see the columns of Table 1). If the generic properties of words work together with useful comparisons, then alignable and thus more comparable stories should show an advantage to learning with relational words (Experiments 1 and 2).

However, if the CSM aspect of words is critical for directing attention to relations, the similarity of words' meanings to the referents in the story should also modulate relational learning. To test this prediction, Experiments 1 and 3 had greater similarity and Experiments 2 and 4 had less similarity between the relational words and the story elements they referenced (see the rows of Table 1). Given that the relational label *target* (especially in contrast to *distracter*) is a positive term, Experiments 1 and 3 paired it with a positive

**Table 1**
The overall design of the four experiments was created by manipulating whether the relational words semantically align with the tutorial (rows) and whether the tutorial story semantically aligns with the transfer story (columns). *Positive target* means that the SDT target in the story is semantically positive, such as healthy athlete or sweet melon. *Negative target* means that the referred element is negative, such as sick patient or infected melon.

| | **Stories align** | **Stories do not align** |
|---|---|---|
| **Relational words semantically overlap with tutorial elements** | Experiment 1<br>Positive target tutorial<br>Positive target transfer | Experiment 3<br>Positive target tutorial<br>Negative target transfer |
| **Relational words do not semantically overlap** | Experiment 2<br>Negative target tutorial<br>Negative target transfer | Experiment 4<br>Negative target tutorial<br>Positive target transfer |

element in the tutorial situation (*healthy athletes*) while *distracter* was paired with the corresponding negative element (*unhealthy athletes*) so that the relational labels were semantically aligned with the story elements. Even though positivity could be construed as a superficial feature, it may provide a semantic clue toward the relational structure. By contrast, in Experiments 2 and 4 the positive label *target* referenced a negative story element (*sick patient*) while the negative label *distracter* referenced a positive story element (*healthy patient*). Table 2 shows the complete set of relational labels aligned with their intended referents in the tutorial and transfer stories. If the semantic overlap between relational words and their referents during learning is important, we should see greater benefits of relational words in Experiments 1 and 3. A semantic mismatch between relational labels and their referents might also lead to a deleterious influence of relational words in Experiment 2 and 4.

We used three different measures: a learning quiz to test whether words have any impact on initial learning, a transfer quiz to test appreciation of the implicit relational similarities between tutorial and transfer stories, and an analogy quiz (matching correspondences between story contexts) to see if subjects can explicitly make connections between the simulations.

## Experiment 1

The conditions of Experiment 1 provide the best chances of producing a benefit for learning with relational words because this experiment provides both semantic alignment between tutorial and transfer elements as well as semantic overlap between the relational labels and their tutorial referents.

**Table 2**

Table 2 presents the relational labels with their story referents from all four experiments. Participants in the Word conditions were presented with a tutorial that included both the relational labels and story referents while corresponding Control tutorials only presented the story referents. There were no relational labels in any of the transfer contexts.

| **Relational Labels** (Explicitly presented in the Word condition tutorial) | **Positive Target Tutorial** (Exp. 1 & 3) | **Negative Target Tutorial** (Exp. 2 & 4) |
|---|---|---|
| *Target* | Healthy athlete | Sick patient |
| *Distracter* | Unhealthy athlete | Healthy patient |
| *Evidence* | Cell strength | Cell distortion |
| *Hit* | Healthy diagnosed "healthy" | Sick diagnosed "sick" |
| *Miss* | Healthy diagnosed "unhealthy" | Sick diagnosed "healthy" |
| *False alarm* | Unhealthy diagnosed "healthy" | Healthy diagnosed "sick" |
| *Correct rejection* | Unhealthy diagnosed "unhealthy" | Healthy diagnosed "healthy" |
| (None of these labels were presented in transfer) | **Positive Target Transfer** (Exp. 1 & 4) | **Negative Target Transfer** (Exp. 2 & 3) |
| *Target* | Sweet melon | Infected melon |
| *Distracter* | Bitter melon | Normal melon |
| *Evidence* | Melon weight | Melon weight |
| *Hit* | Sweet melon exported | Infected melon sent to analysis center |
| *Miss* | Sweet melon rejected | Infected melon sold |
| *False alarm* | Bitter melon exported | Normal melon sent to analysis center |
| *Correct rejection* | Bitter melon rejected | Normal melon sold |

## *Method*

### *Participants and Design*

Eighty-seven undergraduates from Indiana University participated in this experiment for credit. A computer program randomly assigned half of these participants to be in the Word condition ($N = 44$) and the other half were assigned to the Control condition ($N = 43$). Three additional participants who took less than 15 minutes to complete the experiment were excluded from analysis. When participants were debriefed at the end of the experiment,

they reported how much they previously knew about SDT. All of our participants did not know it at all or had heard of it but did not know what it was about.

*Materials and Procedure*

All undergraduates read through a computer-based SDT tutorial made up of pictures and explanatory text (screenshots are provided in Figure 1; full tutorials and corresponding quizzes from all four experiments are available online, http://www.calstatela.edu/centers/learnlab/sdt). The tutorial was a 47-screen self-paced slide show covering basic SDT concepts such as the difference between evidence for a decision, the decision, and the actual status of the decided entity (either signal or noise). Students were shown how a decision boundary could lead to two ways of making the right decision (*hits* and *correct rejections*) and two ways of being incorrect (*misses* and *false alarms*). This was followed by two examples where the decision boundary was moved in order to show the relationship between these categories. Additionally, participants were shown what would happen if the signal distribution shifted along the evidence continuum.

The principles of SDT were embedded in the context of a doctor trying to pick out healthy athletes to play for the university by examining blood cell strength. In the tutorial story, athletes with strong cell samples were more likely to be healthy than those with weak cell samples. Although cell strength was an imperfect indicator of health, the doctor tried to optimize his decisions based on this imperfect evidence. The Word condition differed from the Control condition in only one respect: interspersed into the tutorial were relational labels presented alongside contextual elements. Healthy athletes were labeled *targets* and the unhealthy athletes were *distracters*. Those that the doctor deemed "healthy" were labeled *"target"* with quotation marks around both the story element and the relational term indicating that this is only the doctor's decision rather than the actual status of the athlete. *Hit, miss, correct rejection,* and *false alarm* were also included in the Word condition's tutorial. Other than the addition of the labels, the tutorials for the Word and Control conditions were identical.

The tutorial teaches some basic concepts of SDT without using the traditional normal distributions typically used in SDT classes or textbooks because of the limited time constraints of the experiment. Pilot experiments teaching students SDT with traditional normal distributions contrasted with other attempts using frequency bar graphs supported the claim that frequency information is far easier to understand than probability information (in both general cognition, Gigerenzer & Hoffrage, 1995, and pedagogy, Bakker & Gravemeijer, 2004). We speculated that the overlapping region of the traditional distributions (i.e., where the evidence could be indicative of either targets or distracters; see Figure 1) was particularly crucial for understanding SDT but also particularly confusing for students. Because we were not interested in teaching graph reasoning per se, we developed bar graphs that utilized non-overlapping spaces and color codes tailored to

represent critical concepts of SDT (see Figure 1). Non-overlapping regions of the screen (i.e., top and bottom of Figure 1c) were used to represent two different distributions (i.e., actually healthy and actually unhealthy people). Colored labels ("H" and "U") provided a perceptual indicator for the categorization the detector has made (i.e., diagnosed "healthy" versus "unhealthy"). The tutorial implemented the combination of these features because SDT requires an understanding of two distinctions: (1) which cases are in which categories and (2) what categorizations have been made by the detector.

Each case is represented by an idealized cell in a box outline. The evidence is the strength of the cell and this is indicated by how dark and how large each cell is. The cases are spatially ordered, from left to right, by increasing cell strength. The columns of cases (see Figure 1b) indicate how frequent a particular level of cell strength is. If a particular level of strength indicates a higher likelihood of predicting actual health, there are more actually healthy cases (green boxes) in the column than actually unhealthy (red boxes). If a particular level of strength indicates a lower likelihood of predicting actual health, there are more actually unhealthy cases (red boxes) in the column than actually healthy (green boxes). Columns that include both red and green boxes can be seen as analogous to the overlapping regions of traditional SDT distributions because the same level of strength could belong to either category. Although in a typical SDT distribution there is an actual physical overlap to signify that both categories can occur with the same cell strength level, in our diagrams, we illustrate this with instances from both categories stacked in the same column. The cases are then separated by category into two different bar graphs (see Figure 1c) to clearly show the actual status of these athletes. This reflects the SDT distinction between actual target and distracter distributions, as contrasted with the doctor's decision.

After reading through the tutorial, participants answered eight multiple-choice questions about the tutorial's doctor situation that could be answered correctly by applying SDT principles. Difficult quiz questions were purposefully used to ensure that participants needed to use SDT principles rather than relying on common sense. (Tutorial quiz questions have been included in the Appendix and are available online.)

Then, participants received an opportunity to transfer what they had learned to a different context. Participants read a few paragraphs (included in the Appendix) presented on three slides describing a small town that wants to export sweet melons and avoid sending out bitter melons. Sweet melons, laden with nectar, tend to be heavier, so this town decides to sort the melons by weight (even though weight is not a perfect indicator of sweetness). Heavy melons are exported and sold while light melons are rejected. However, all of the melons were subject to consumer reports that allow the town to find out which melons are actually sweet/bitter. An eight-question transfer quiz was administered. At the end of the experiment participants were told that these two stories were analogous and asked to explicitly place elements of the two stories in correspondence with each other in a six-question multiple-choice mapping quiz.

**Figure 1.** These are screenshots of the tutorial in the Word condition, showing relational labels such as "targets" and "distracters" alongside elements of the story, "healthy" and "unhealthy" people. Each patient's blood test (the evidence for the diagnosis) is represented by what is in each rectangle, the diagnosis is represented by the letter, and the actual status of the patient is represented by the color of the outline.

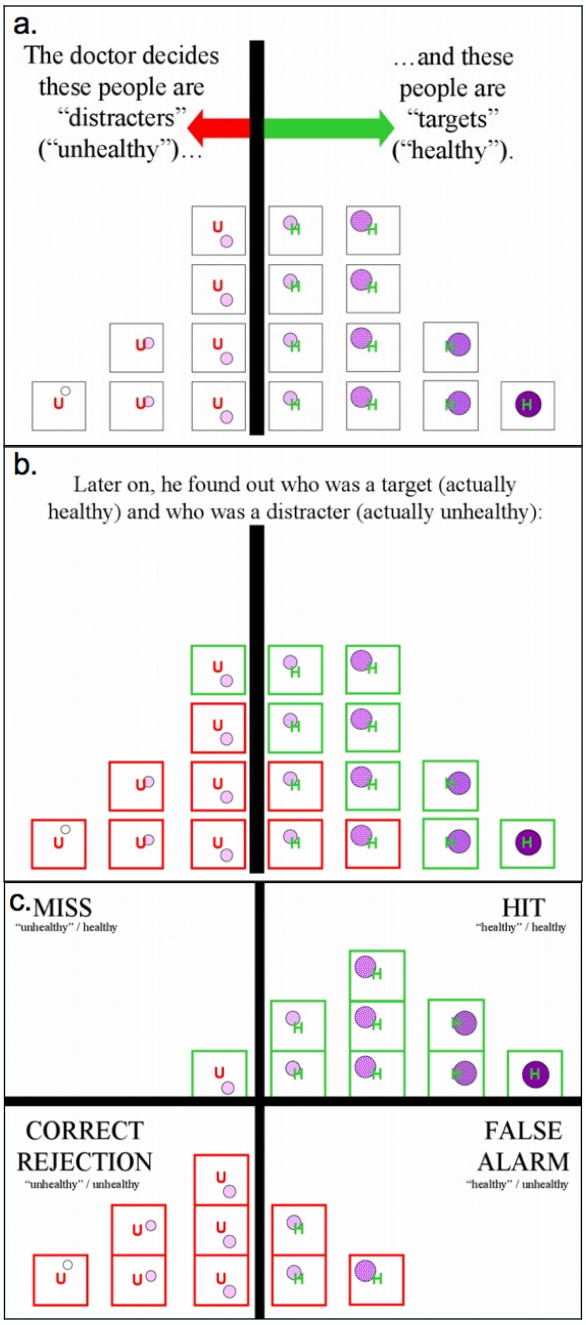
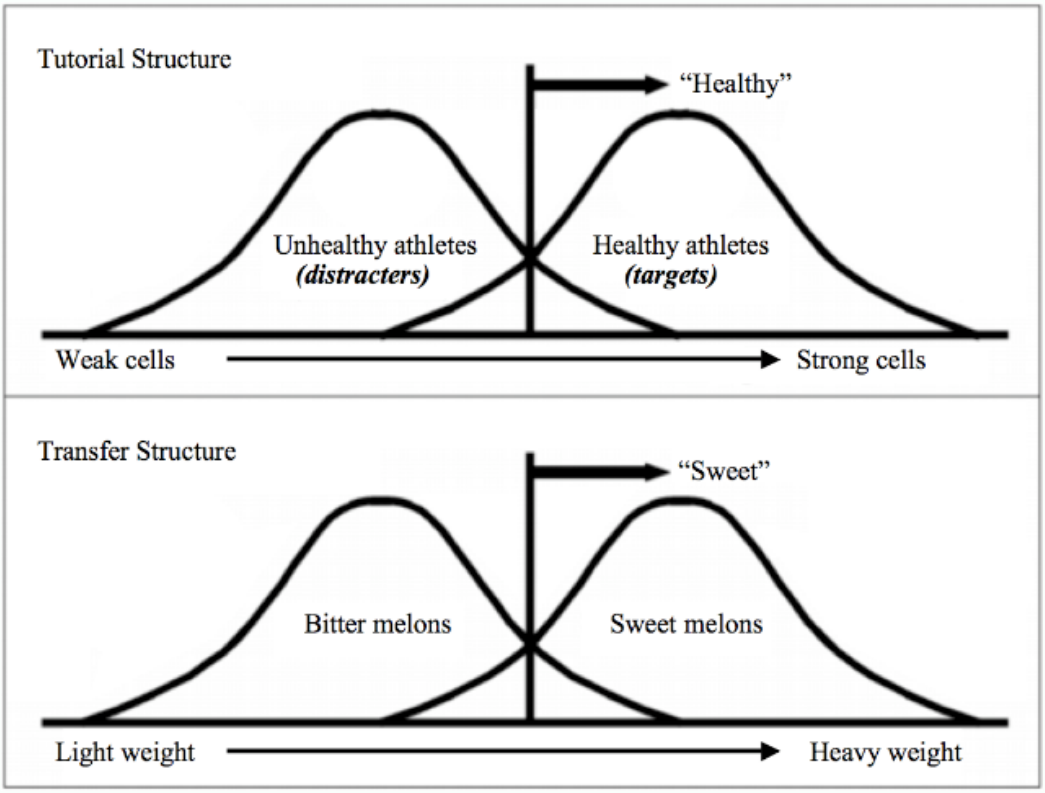


**Figure 2.** Overlapping normal curves are typically used to represent the structure of SDT. This figure shows the SDT structure of both tutorial and transfer stories used in Experiment 1.




***Results***

Because this study examines the impact of learning relational words not only in immediate learning situations but also performance on other relationally similar examples, there are two dependent measures of interest here, the scores on the tutorial quiz and the transfer quiz. The relationship between the quizzes and the experimental manipulation were analyzed with a mixed-design 2 x 2 ANOVA (quiz type x condition). There were no main effects for quiz type, $F(1, 85) = 0.04$, nor word manipulation, $F(1, 85) = 0.08$, but this analysis confirmed that there was a significant interaction, $F(1, 85) = 8.63$, $p < 0.01$ (see Table 3). Participants in the Word condition had showed an improvement (a positive difference) from transfer and tutorial scores ($M = 0.07$, $SD = 0.25$) while the Control condition participants showed a decline ($M = -0.08$, $SD = 0.23$). A paired $t$-test confirmed that this change in performance (transfer score - tutorial score) was significantly different between Word and Control conditions, $t(86) = 8.63$, $p < 0.01$.

Students who learned more from the tutorial should be predicted to do better on the transfer quiz regardless of condition. But we also wanted to know whether learning SDT with relational labels helped make SDT concepts more flexible and transferable to new contexts. An ANCOVA first revealed that the tutorial score is a significant covariate, $F(1, 84) = 32.17$, $p < 0.001$, on transfer performance. More surprisingly, this analysis also revealed that the word manipulation is still a significant factor influencing transfer performance, $F(1, 84) = 5.62$, $p < 0.05$, with words predicting better transfer performance than no words. This was a small effect size (calculated using Cohen's d, $d = -0.25$) but the experimental manipulation was also subtle. Even though the Control participants look better than those in the Word condition on the tutorial quiz (not significantly different though, $t(86) = 3.03$, $d = 0.37$), these Word participants seemed to have transferred more of what they learned.

If words direct attention to structure or provide comparison opportunities that highlight structure, one might expect the Word condition to outperform the Control condition even in the mapping quiz. The mapping results are shown in Table 4. However, we found that there were no differences in performance on the mapping quiz, $t(83) = 0.78$, with Control and Word participants both scoring well, averaging 0.74 ($SD = 0.17$) and 0.71 ($SD = 0.16$), respectively. A second look at our mapping questions indicates that the questions might have been too well constrained to show differences. A question typically asked about one element of either the tutorial or the transfer story (e.g., "heavy melon"—evidence of being a target) and presented four possible answers. Two of the answers would be justifiable SDT answers, "patient with strong cells" (evidence of being a target) and "patient with weak cells" (evidence of being a distracter) while the other two would be incorrect according to SDT structure (i.e., healthy person or sick person, targets and distracters). If a participant could narrow down his or her choice to the two SDT answers, then a loose similarity between the dimensions of heaviness and strength could lead to a correct match: "heavy melon" with "strong cells."

### *Discussion*

The results of Experiment 1 suggest that even when the Control participants exhibited learning on the tutorial quiz, they did not transfer their learning to the new situation as well as those in the Word condition. There were no differences between conditions on the tutorial quiz, but those in the Word condition may have performed better if quiz questions also included the relational words. The exclusion of these relational words from the tutorial quiz may have limited their performance on the assessment. However, even with this disadvantage, the Word participants were able to show better transfer of their learning to the second quiz. The interaction between quiz (initial or transfer) and condition (Word or Control) underscores the importance of separating variables that affect immediate learning versus those that make knowledge readily transferable (Bransford & Schwartz, 1999; Goldstone & Sakamoto, 2003; Chi, Feltovich, & Glaser, 1981). Relational words may

be generally difficult for learners to acquire (Keil & Batterman, 1984; Gentner, 1975; Hall & Waxman, 1993), but it seems that their real benefit shows up later on. The interaction between learning relational words and the quizzes could be interpreted as evidence for words making relations more salient when seen again in a new context, thereby allowing the transfer situation to seem more similar to the tutorial situation.

**Table 3.** Tutorial and transfer quiz results from all four experiments are shown here. Positive target contexts (i.e., healthy athletes, sweet melons) have been colored green and negative target contexts (i.e., sick patients, fungus-infected melons) have been colored red.

| **Relational Words** | **Tutorial** | **Transfer** |
|---|---|---|
| Experiment 1 | | |
| *target* | Healthy athlete | Sweet melon |
| *distracter* | Unhealthy athlete | Bitter melon |
| Control | 0.55 (*SD* = .25) | 0.47 (*SD* = .27) |
| Word | 0.46 (*SD* = .23) | 0.53 (*SD* = .24) |
| Experiment 2 | | |
| *target* | Sick patient | Fungus-infected melon |
| *distracter* | Healthy patient | Normal melon |
| Control | .56 (*SD* = .22) | .37 (*SD* = .26) |
| Word | .52 (*SD* = .22) | .43 (*SD* = .19) |
| Experiment 3 | | |
| *target* | Healthy athlete | Fungus-infected melon |
| *distracter* | Unhealthy athlete | Normal melon |
| Control | .50 (*SD* = .18) | .48 (*SD* = .25) |
| Word | .52 (*SD* = .23) | .55 (*SD* = .28) |
| Experiment 4 | | |
| target | Sick patient | Sweet melon |
| distracter | Healthy patient | Bitter melon |
| Control | .57 (*SD* = .23) | .47 (*SD* = .22) |
| Word | .51 (*SD* = .22) | .36 (*SD* = .23) |

**Table 4.** Mapping results from all four experiments are shown here. In Experiments 1 and 2, the tutorial and transfer stories were designed to align semantically. However, in Experiments 3 and 4, the two stories preserved the structural alignment without the same semantic overlap. Thus, the latter experiments had two different mapping possibilities, favoring structure or semantic alignment.

| Relational Words | Experimental Example | Control condition | Word condition |
|---|---|---|---|
| Experiment 1 | | | |
| Target = Target mapping (Structure + Semantic) | healthy person = sweet melon | .74 (*SD* = .19) | .70 (*SD* = .16) |
| Target = Distracter mapping | healthy person = bitter melon | .10 (*SD* = .10) | .09 (*SD* = .14) |
| Experiment 2 | | | |
| Target = Target mapping (Structure + Semantic) | sick person = infected melon | .71 (*SD* = .21) | .72 (*SD* = .25) |
| Target = Distracter mapping | sick person = normal melon | .09 (*SD* = .10) | .07 (*SD* = .12) |
| Experiment 3 | | | |
| Target = Target mapping (Structure) | healthy person = infected melon | .24 (*SD* = .17) | .22 (*SD* = .10) |
| Target = Distracter mapping (Semantic) | healthy person = normal melon | .66 (*SD* = .17) | .64 (*SD* = .20) |
| Experiment 4 | | | |
| Target = Target mapping (Structure) | sick person = sweet melon | .20 (*SD* = .22) | .31 (*SD* = .22) |
| Target = Distracter mapping (Semantic) | sick person = bitter melon | .61 (*SD* = .31) | .37 (*SD* = .25) |

## Experiment 2

The advantage of learning labels seen in Experiment 1 may have been due to one of two factors: (1) because labels provided a generic cue to compare the doctor and melon story or (2) because the content of the labels (i.e., target—a positive label) were consistent with the corresponding elements of the tutorial (i.e., healthy athlete—a positive story element). If labels function as a comparison cue, then it is critical that the doctor and melon stories are alignable. In Experiment 1, they were alignable in that the doctor was looking for positive targets (e.g., healthy athletes) and the melon farmers were looking for

positive targets (e.g., sweet melons). In Experiment 2, the stories were alignable because the doctor looks for negative targets (e.g., sick patients) and the melon farmers also look for negative targets (e.g., infected melons). Preserving the alignability in this way forfeited the consistency between the label (i.e., target—positive label) and the tutorial element (i.e., sick patient—negative story element).

If the Word condition in this experiment promotes transfer like Experiment 1, this would provide support for the hypothesis that labels help by promoting comparison between stories, allowing participants to see common relational structure. However, if the Word condition does not promote transfer, this provides further support for the hypothesis that it is the semantic overlap between relational labels and their contextual objects that determines whether words facilitate transfer.

### *Method*

*Participants*

Seventy-five undergraduates (34 in the Control condition, 41 in the Word condition) from Indiana University participated in this experiment for credit. Seven additional participants were excluded from analysis because they took less than 15 minutes to read through the tutorial. All participants reported that they had not previously learned SDT.

*Materials*

Tutorials similar to those used in Experiment 1, with and without relational words, were used in this experiment. The main changes were to the tutorial and transfer story contexts to convert them into situations where the detector searches for a negative target. In the new tutorial story the doctor is trying to diagnose leukemia patients by examining blood samples. People with distorted cell samples are more likely to have leukemia than those with pure cell samples. Although cell distortion is an imperfect indicator of leukemia, the doctor must try to optimize his decisions. The new melon transfer story is semantically aligned to this tutorial story. The melon farming town is now trying to detect fungus-infected melons in order to send them to an analysis center. Heavier melons tend to be infected because they are carrying spores, but melon weight is not a perfect indicator of fungus. Reports from the analysis center as well as consumers allow the town to find out which melons are actually infected/normal. The alignment between the two stories is demonstrated in Table 2.

The relational labels are the same as those in Experiment 1, only applied to the elements of the new story. The positive label, *targets*, refers to actually sick people and the negative label, *distracters,* refers to actually healthy people. Those that the doctor has diagnosed are marked as "sick" and, in the Word condition, they are accompanied by the label "*target*." Those that have been diagnosed as "healthy" are labeled "*distracters*." The departure from Experiment 1 removes the consistency between relational words and the

story elements that may have aided Word participants. Note that the relational words preserve the structure of SDT and are correctly applied to the doctor context. We will sometimes refer to Experiment 2's tutorial as a negative target tutorial to draw attention to the reduced semantic overlap between the positive relational word "target" and the negative targets in the story (Table 2 shows the mappings between the relational words and story elements more fully). Similarly, Experiment 2's transfer context is a negative target transfer situation.

Once again, other than the addition of relational words, the tutorials for the Word and Control conditions were made up of the same pictures and explanatory text. Note that unlike Experiment 1, there is no mention of rejection in the transfer situation here. The relational role of the *correct rejection* in the new melon scenario is filled by normal melons that get sold to consumers instead of getting sent to the fungus analysis center. Any transfer that might be found from the relational word tutorial cannot be explained by an explicit connection between the words and the transfer context.

Other materials included an eight-question multiple-choice tutorial quiz, a transfer quiz, and a six-question mapping quiz. The wording of the quizzes was changed to reflect the new stories. One of the transfer questions was also changed from Experiment 1 (Question #1).

*Procedure*

The procedure was the same as before. First participants were presented with a tutorial involving patients, then a quiz based on the tutorial, then a transfer situation based on melons, and finally a transfer quiz. At the very end of the experiment, there was a mapping quiz between the leukemia-detecting doctor and the fungus-detecting town.

*Results and Discussion*

If words foster relational transfer by capturing the alignment across story contexts, then more similar situations should show the benefits of labeling. By this account, despite the dissimilarity between the labels and the tutorial context, the relational labels could serve as a representation of the similarity between stories to encourage transfer. A mixed-design 2 x 2 ANOVA (quiz type x condition) showed that there was a main effect of quiz type, $F(1, 73) = 34.80$, $p < 0.001$, and a significant interaction, $F(1, 73) = 4.27$, $p < 0.05$. There was no main effect of Word condition, $F(1, 73) = 0.19$. These results are shown in Table 3.

The interaction is consistent with the pattern found in Experiment 1 because even though participants in the two conditions seem to have performed similarly in the initial tutorial context, $t(74) = 0.29$, $d = 0.18$, those who also learned relational words were better able to transfer their learning to a new situation. An ANCOVA showed that the Word condition significantly outperformed the Control condition in transfer, $F(1, 72) = 4.08$, $p < 0.05$, $d = -0.70$, and the tutorial quiz score is also a significant covariate, $F(1, 72) = 41.28$,

$p < 0.001$. An initially mismatching set of relational words, when supported by alignable similarities, aids in transfer of previously learned material to a new situation. Especially because the difference between conditions arises in transfer, we can speculate that the Word participants may have used the relational words to capture the abstract similarities that arose between the two stories. The relational words in Experiment 2 may have provided cognitively easy handles for difficult relational concepts.

Unlike Experiment 1, there was a main effect of quiz type where performance on the tutorial quiz was generally better than transfer ($d = 0.61$). A paired *t*-test showed that across both conditions, there was a significant decline ($M = 0.13$, $SD = 0.20$) between tutorial and transfer quizzes, $t(74) = 5.61$, $p < 0.001$. It may be that the negative target transfer quiz was harder or that learning from the negative target tutorial was not as transferable.

In the mapping quiz, participants had to match the analogous elements, such as sick patient (target) to infected melon (also target), and correct matches were both structurally and semantically aligned. In this way, the difficulty of Experiment 2's mapping quiz was similar to that of Experiment 1. The resulting patterns of results were also similar. There was no difference in mapping scores between the Control and Word conditions, $t(74) = 0.12$ (means and standard deviations shown in Table 4).

The similarity between the results of Experiments 1 and 2 suggests relational words allow students to exploit alignable similarities between the tutorial and transfer contexts. The labels used in Experiment 1 had little in common with their corresponding tutorial elements, but the labels in Experiment 2 had even less in common with their corresponding tutorial elements. Even despite the introduction of dissimilarity, words still had a beneficial effect for transfer. However, this set of results does not rule out the possibility that the inclusion of relational words could help learners to think about the scenarios more relationally in general. If the semantics of *target* and *distracter* can guide learning of SDT, even without alignable stories, we should see benefits of tutorials with relational words.

## Experiment 3

So far we have shown that there is a beneficial effect of relational words when there are alignable similarities, which have been implicated in creating structural representations (Markman & Gentner, 2000). Experiment 3 examines whether the meanings of the relational words *alone* (even without alignable stories) could also encourage learning relational concepts. To illustrate this point, recall the Rattermann and Gentner results (1998), where *daddy-mommy-baby* were helpful lexical terms but *jiggy-fantan-gimli* were not helpful for children learning about monotonic decrease (large-medium-small). Presumably the meaning of *daddy* helps children focus on the large size. Under this explanation, part of the success of Experiment 1's Word condition may have been due to the semantic overlap

between *target*, a positively valenced relational word, and healthy athlete, a positively valenced story element.

Experiment 3 used the tutorial from Experiment 1 but did not use the well-aligned transfer story of Experiment 1 to take away the influence of easily comparable situations. Instead, the positive target tutorial was followed by the negative target transfer situation (infected melon story from Experiment 2). If relational words require the support of alignable similarities to be beneficial, we should see no benefit in the Word condition. However, if the initial semantic overlap of the relational words to the tutorial context is also effective, we should see benefits in transfer even without the analogous elements of the two scenarios being semantically aligned.

Additionally, this experiment may shed light on why Experiment 2's transfer scores were overall lower than the tutorial scores. If the negative target transfer quiz (infected melon) is simply more difficult than the positive target transfer quiz (sweet melon), then Experiment 3 should show a similar pattern of decrease in transfer performance. However, if seeking positive targets in the tutorial (healthy athletes) is a better tutorial than seeking negative targets (sick patients), then Experiment 3 should be more similar to Experiment 1 and show no overall decreases in transfer.

### *Method*

*Participants*

Sixty-one undergraduates (32 in the Control condition, 29 in the Word condition) from Indiana University participated in this experiment for credit. Eight additional participants were excluded from analysis because they took less than 15 minutes to read through the tutorial. One other participant was excluded because of previous SDT knowledge. All other participants reported not knowing SDT.

*Materials*

The tutorials, with and without relational words, used here were the same as Experiment 1. The transfer materials were the ones used in Experiment 2. This design, positive target tutorial with negative target transfer, is shown in Table 2. Importantly, there is a match in the semantics between the relational words and the tutorial elements (i.e., *target* and healthy athlete are both positive; *distracter* and unhealthy athlete are both negative) but a misalignment between the tutorial and transfer elements (i.e., healthy athlete and infected melon; unhealthy athlete and normal melon).

Because the tutorial and transfer stories were not closely aligned, the grading of the mapping quiz in this experiment was different than in the previous experiments. For a mapping question such as "What in the melon story is most analogous to the healthy athlete in the doctor scenario?" students could pick the other target-like element (infected melon). However, if we assume that targets can legally map to distracters (because in

some contexts, which element is signal and which is noise is an arbitrary decision), mapping healthy athlete to normal melon still preserves the structure of SDT. In fact, in an impoverished setting like a multiple-choice quiz, this is more appealing because of the semantic similarity.

Given the experimenters' familiarity with SDT, the tutorial and transfer situations were designed with the healthy athlete-infected melon match (the target-target match) in mind. One reflection of that intention is found in the spatial organization of the tutorial and transfer figures, with cell strength and melon weight increasing from left to right (for a schematic illustration see Procedure). Because of this spatial alignment, the targets were both on the right side (the higher end of the evidence dimension) and the distracters were on the left side (the lower end). We will call this mapping the structural answer. The mapping quiz was graded in three ways: a structural score (mapping healthy athletes to infected melons), a semantic score (mapping healthy athletes to normal melons), and a total mapping score (both answers counted as correct).

*Procedure*

The procedure was the same as before. The tutorial and the tutorial quiz were followed by the transfer situation and the transfer quiz. A mapping quiz was administered at the end of the experiment.

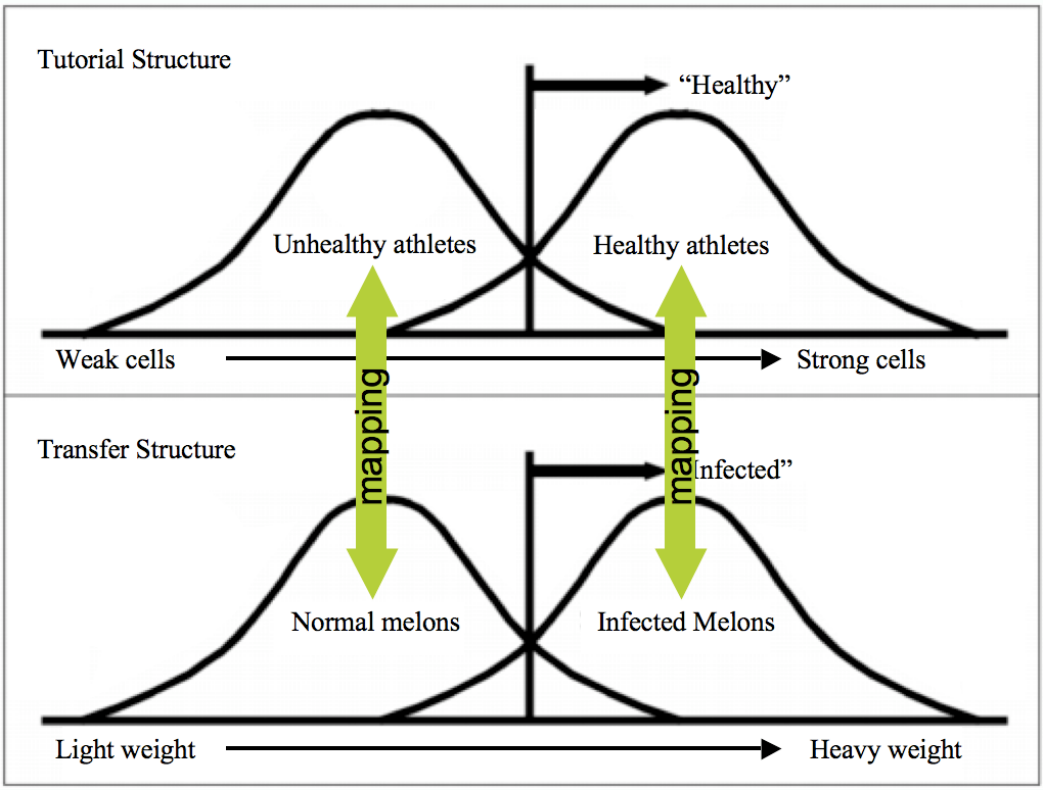
*Results and Discussion*

A mixed-design 2 x 2 ANOVA (quiz type x condition) showed important similarities and differences from previous experiments. First, there was no main effect of quiz type, $F(1, 59) = 0.019$, like Experiment 1 and unlike Experiment 2. This suggests that there is nothing inherently difficult about the negative target transfer situation in which the farmers look for infected melons. However, transfer seems less difficult overall with positive target tutorials (Experiments 1 and 3). The ANOVA also revealed no significant effect of condition, $F(1, 59) = 0.654$, nor any interaction, $F(1, 59) = 0.411$. Although these are null effects, these results are important to consider in the context of the other two experiments. The results are presented in Table 4 for ease of comparison.

Comparing the means between the Control and Word conditions, there seems to be a trend toward better transfer performance with relational words than without them. However, this is not borne out statistically, $t(60) = 0.86$. This suggests that relational words did not consistently provide advantages when the two stories were not alignable.

Mapping results showed that all participants preferred making semantic matches, 0.65 ($SD = 0.19$), over structural ones, 0.23 ($SD = .14$), $F(1, 59) = 114.14$, $p < 0.001$. Word and Control conditions showed no difference in the number of semantic mappings made, $t(60) = 0.18$, nor in the number of structural matches made, $t(60) = 0.14$. The semantically similar elements (i.e., an infected melon and an unhealthy athlete) were more influential

**Figure 3.** The structure of the stories used in Experiment 3 are shown with examples of the counterintuitive structural mapping (green arrows).




than the structural aspects of the stories (targets corresponding to targets). The sparse multiple-choice format of the mapping quiz may have biased participants toward forming local mappings.

## Experiment 4

So far we have learned that labels facilitate relational reasoning best in the context of stories with semantically matching elements, and matching the meanings of the relational words to the training scenario alone does not result in the same benefits. However, from these results we do not know what the effect of relational words are when there is no overlapping meaning between the word and the contextual element *and* no alignment between stories. We will use the Rattermann and Gentner (1998) study to illustrate a plausible, but untested, conjecture: relational responding may have suffered if they had labeled the small object *daddy* and the large object was called *baby*. Even though semantic associations alone do not significantly improve quiz performance, without them, performance might actually suffer. To test this, participants learned relational words in the context of a negative target tutorial but transferred to a positive target situation.

If there is no effect of learning relational words in this experiment, it provides further support for the hypothesis that labels interact with comparison of well-aligned stories. Also, if there is a decline between tutorial and transfer quiz performance, like in Experiment 2, this suggests that the negative target tutorial is not as effective for transfer as the positive target tutorial (Experiments 1 and 3).

### Method

*Participants*

Sixty-five undergraduates from Indiana University participated in this experiment for credit. They were randomly assigned to the Word ($N = 33$) or Control ($N = 32$) condition. Five additional participants who took less than 15 minutes to complete the experiment were excluded from analysis. At the end of the experiment, all participants reported that they had not previously learned SDT.

*Materials and Procedure*

The tutorials, with and without relational words, used here were the same as Experiment 2. The transfer materials were the ones used in Experiment 1. This design, negative target tutorial with positive target transfer, is shown in Table 2. There is less overlap between the semantics of the relational words and the tutorial elements (i.e., *target* is positive but sick patient is negative; *distracter* is negative but healthy patient is positive). Also, there is no alignment between the tutorial and transfer elements (i.e., sick patient and sweet melon play the same role, as do healthy patient and bitter melon).

Because of this lack of alignment, the mapping quiz was scored like Experiment 3. For a mapping question that inquired about sick patients, participants could pick the transfer's target (sweet melon) or the distracter (bitter melon). Although matching sick patients to bitter melons is probably more appealing because of the semantic similarity, the sick patient-sweet melon match is the structural mapping by both being the target-target match and being the spatial match. The mapping quiz was graded in three ways: a structural score (mapping sick patients to sweet melons, see Figure 3), a semantic score (mapping healthy patients to sweet melons), and a total mapping score (both answers counted as correct).

### Results and Discussion

Table 3 shows the results of the tutorial and transfer quiz broken up by condition. A mixed-design 2 x 2 ANOVA (quiz type x condition) showed no reliable interaction, $F(1, 63) = 1.70$, but showed a significant main effect of quiz, $F(1, 63) = 26.41$, $p < 0.001$, $d = 0.59$. Similar to Experiment 2, the other experiment that used a negative target tutorial, participants had significantly higher scores on the tutorial quiz, 0.54 ($SD = 0.23$), than the transfer quiz,

0.41 ($SD = 0.23$). Experiment 4 was the only experiment where the ANOVA showed even a marginal effect of condition, $F(1, 63) = 3.38, p < 0.08$.

Although words typically show benefits for fostering appreciation of relational structure in the literature, our results show a trend in the opposite direction in which the Control condition has generally better quiz scores than the Word condition (see Table 3). However, quiz-specific analysis revealed that this advantage is primarily driven by differences on the transfer quiz, $t(64) = 6.09, p < 0.05, d = 0.49$, and there was no significant difference in tutorial quiz performance, $t(64) = 0.98, d = 0.27$. Control participants showed significantly better transfer performance than those trained with relational words. An ANCOVA further revealed that even though the tutorial scores were found to be a significant covariate, $F(1, 62) = 27.11, p < 0.001$, condition was still a significant influence on transfer performance, $F(1, 62) = 4.07, p < 0.05$. In this case, learning with relational words actually was disadvantageous rather than being neutral (Experiment 3) or beneficial (Experiments 1 and 2). These results underscore the importance of alignment between scenario elements as removing alignment also removes the benefit of learning relational words. These results go further to suggest that there are hazards of teaching relational words when there is little semantic overlap between the relational word and the learning context.

The harmful effects of learning relational words, particularly for transfer, might lead one to expect participants in the Word condition to have low mapping quiz scores as well. Total mapping scores, both structural and semantic answers together, did differ, with the Control condition scoring significantly better ($M = 0.87, SD = 0.16$) than the Word condition ($M = 0.79, SD = 0.18$), $t(64) = 4.49, p < 0.05$. When the composite score was broken down into structural and semantic mappings and analyzed with a mixed-design 2 x 2 (mapping type x condition) ANOVA, there was a main effect of mapping type, with participants generally making more semantic matches, 0.49 ($SD = 0.30$), than structural ones, 0.26 ($SD = 0.22$), $F(1, 63) = 15.34, p < 0.001$. This result supports the notion that superficial semantic similarity strongly influences mapping (Gentner & Toupin, 1986; Ross, 1989), but we should note that these explicit mappings were made after the transfer quiz, so they may or may not have been used during transfer (Ross, 1987).

There was also a main effect of condition, $F(1, 63) = 10.52, p < 0.01$, as well as a significant interaction between these variables, $F(1, 63) = 8.68, p < 0.01$. These results are shown in Table 4. Control participants made significantly more semantic mappings than the Word condition, $t(64) = 11.75, p < 0.01$. The poorly transferring Word condition showed significantly more structural choices than the Control condition, matching the explicit target in the tutorial context (sick patient) with the implicit target in the transfer context (sweet melons), $t(64) = 4.22, p < 0.05$. This is surprising since the structural choice is both counterintuitive and a more relationally sophisticated choice. As counterintuitive as the structural mapping of the two contexts may be to all of our novice participants, this alignment might be obviously seen as the "right answer" if the words introduced in the

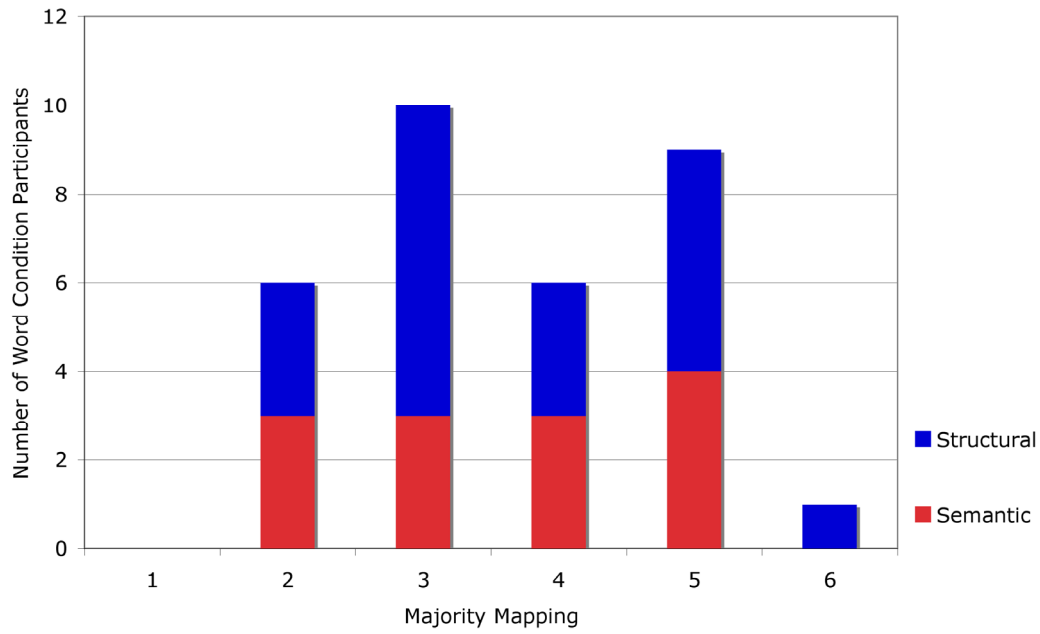
experimental condition were used to explicitly connect scenario elements. An example of an explicit connection would be if participants were asked, "Which is the target of the patient story? Which is the target of the melon story?" The Word condition participants were perhaps more likely to make the implicit connection between the targets of the two scenarios.

Even so, participants in the Word condition still made more semantic than structural mappings, suggesting one of two possibilities: (1) half of the participants in the Word condition made structural alignments and the other half made semantic ones; or (2) the individual participants in the Word condition do not have a consistently aligned view of the analogy and flip-flopped between structural and semantic mappings. To examine these two possibilities, we categorized students by how many structural and semantic mappings they made out of a possible six. Then for each participant, we registered the mapping type in which the participant showed the majority of correct answers (i.e., if a participant made two structural and four semantic mappings, we tallied the four semantic mappings) and created Figure 4 out of these majority mapping scores. If the participants in the Word condition showed high majority mapping scores (5 or 6 out of 6), this supports the first possibility, that participants are split into consistently structural and consistently semantic mappers. If participants in the Word condition tend to make only some semantic and some structural mappings (2-4 out of 6), this supports the second possibility, that each participant only makes a few of each type of mapping. Figure 4 shows that only 10 (out of 32) Word participants were consistently structural or semantic. More than half of the Word condition (22 out of 32) had a majority mapping score between 2 and 4 mappings. It seems that participants in the Word condition are influenced by both semantic and structural construals, and these construals yield a hodgepodge of inconsistent mappings. Words may cause some pull toward a structural perspective but cannot completely overcome the attractive semantic mapping. There is enough uncertainty to prevent Word participants from settling on one coherent perspective. This instability may have contributed to their poor performance on the transfer task.

## General Discussion

Taken together, these four experiments reveal a system of effects that connects to important themes of research in language and analogical reasoning. We have explored how learning and applying deep principles (such as SDT) are sensitive to interactions between similarity and language. When relational words about SDT structure were introduced with two readily alignable stories, participants in the Word condition showed better transfer than Control participants (Experiments 1 and 2). This benefit of relational words was shown even when the valence-based semantics of the relational words did not match the semantics of the elements of either tutorial or transfer contexts (Experiment 2). When

**Figure 4.** The Word participants in Experiment 4 are shown with their majority mapping score, which is the greater of their correct semantic or structural mapping scores. Note that most of these participants tend to make two to four correct mappings as opposed to five to six mappings.




the corresponding elements of the tutorial and transfer stories did not semantically align, there was either no effect (Experiment 3) or a slight disadvantage (Experiment 4) of learning relational words. Less semantic overlap between the relational labels and the learning context is more harmful than better semantic overlap, as revealed by the difference in results between Experiments 3 and 4.

The notion that words help us interpret a situation immediately before us may be generally accepted, but our experiments show that words actually continue to influence learners even in new situations presented without those relational words. Our experiments did not find any significant differences between conditions on tutorial performance; the main differences were found during transfer. Furthermore, this influence actually relies on the alignability of new transfer situations to old ones. When story similarities support structural alignments, then learning relational words fosters transfer. This suggests that words foster relational transfer best when they function as GTs—convenient, easily manipulated representations—that stand for relational structure.

Without alignment between tutorial and transfer stories, words may have either no influence (Experiment 3) or even a negative influence (Experiment 4). Note that the transfer

disadvantage in Experiment 4 was not simply due to poorer performance on the tutorial itself. This suggests that words function as CSMs as well since relational words with less effective meanings result in poorer transfer than relational words that meaningfully point to contextual elements. This finding fits with research that indicates that although words generally foster relational reasoning, not all words are equal in that ability (Rattermann & Gentner, 1998; Son et al., under review). This also supports indications that the mere presence of a relational word does not always mean that a relational category or mapping will be formed (Hall & Waxman, 1993; Keil & Batterman, 1984).

Altogether, these results further suggest that the two functions of words (as GTs and CSMs) may be considered additive. If they are additive, we have an explanation for why Experiment 1's Word condition benefited even more than the Word condition in Experiment 2 and Experiment 3. In Experiment 1, the words were effective CSMs that matched the story elements but also acted as GTs that took advantage of alignable situations to promote transfer. Experiment 2 only had words that functioned as GTs and Experiment 3 only had words that functioned as CSMs. All words, novel and meaningful, function as GTs but only some words are CSMs. So perhaps the meaningful aspect of relational labels like *daddy/mommy/baby* (Rattermann & Gentner, 1998) and *target/distracter* builds upon the way that novel labels such as *grecious/leebish* (Lupyan, 2008) function.

Beyond our manipulation of relational words, positive target tutorials (used in Experiments 1 and 3) seemed to transfer more readily than negative target tutorials (used in Experiments 2 and 4). Evidence for this comes from the decreases in quiz score from tutorial to transfer that was found in Experiments 2 and 4, even though they used different transfer tests (2 used a negative target transfer and 4 used a positive target transfer). The story about a doctor detecting leukemia may be a poorer tutorial situation than a doctor looking for healthy athletes for several reasons. Perhaps SDT inherently has something more in common with detecting positive things, because in many situations we look for things that we desire. Another speculative reason for the disadvantage of the leukemia story may be because of participants' general familiarity with that kind of situation. It may have been difficult to reconceptualize a familiar situation as an example of SDT rather than a more novel medical example. Research in other domains has found that learning with concrete, familiar situations can hinder transfer (Goldstone & Sakamoto, 2003; Kaminski, Sloutsky, & Heckler, 2008). We have explored the impact of different levels of contextualization, personalization, and familiarity in learning SDT in another line of experiments (Son & Goldstone, 2009). This work has confirmed that more familiar, personally relevant scenarios produce less robust transfer than more distant, generic scenarios.

Similarity plays a major role in analogical mapping and usually mappings that can be made on the basis of object similarity are the least effortful (Gentner & Toupin, 1986). In Experiments 1 and 2, the correct mappings were both superficially and relationally similar, but in Experiments 3 and 4 semantic and structural information did not foster the

same alignment. When the mappings are in conflict, regardless of condition, participants generally preferred semantic mappings to structural ones. When there are answer options with a high degree of superficial similarity to the target (i.e., *unhealthy* athletes corresponding to *infected* melons; Experiment 3), semantic mappings are virtually unavoidable, with or without relational language. A lesser degree of similarity (i.e., *sick* patients and *bitter* melons; Experiment 4) may have allowed mappings to be more affected by relational words.

Typically models of relational reasoning assume that mapping comes before transfer. If this were the case, then better mappings should always be accompanied by better transfer. In our experiments, structural mappings could be considered the most reflective of SDT. However, when we compare Experiments 3 and 4, the participants who made the most structural mappings, the Word condition in Experiment 4, also had one of the worst transfer scores. This suggests several possibilities. Perhaps when a system of mappings is inconsistent or incomplete, transfer suffers. Another possibility is that the ecology of a mapping task is different from transfer (the mechanisms may differ as well; see Leech, Marschal, & Cooper, 2008). Particularly in our paradigm, mapping questions were presented as separate comparisons between individual elements. This seems to reflect a very different environmental context than our transfer task, where participants must consider a full situation and the contingencies within that system. Understanding the structure within one context may be a different problem than connecting elements across two contexts. The former case requires a more global understanding of the structure, but only in one context, but the latter may be more affected by more local relations and features between contexts. However, it is important to note that mappings were probed at the end of the experiment so the participants may or may not have used these correspondences for the transfer test.

## Language and Abstraction

As flexible and useful as a formalized understanding of SDT might be, conveying these schemas in highly stripped down forms, such as equations and sparse graphs, may lead to representations that are too stripped down to foster learning, much less transfer, in novices. In the experiments described here, the participants Control conditions always learned the abstractions *completely* embedded in the doctor context whereas those in the Word conditions also had exposure to decontextualizing descriptors. In general, the Control participants may have relied more heavily on the tutorial context, perhaps resulting in scaffolded performance on the tutorial quiz. However, the payoff (and the detriment in Experiment 4) for the extra work of learning decontextualized relational words was seen rather late, in transfer performance. In fact, because the alignability of the transfer contexts played such a large role in whether words were effective or not, the processes of comparison may be fundamental to fostering abstract understandings.

These findings shed light on the first of two ways (as GTs and as CSMs) in which relational words could exert their influence. Decontextualized relational words seem to interact with commonalities between two alignable contexts. This suggests that words function like GTs that represent abstractions more concretely such that they can be utilized more effectively in transfer. When contexts do not have transparent structural alignments, the availability of GTs does not facilitate transfer. Any benefit of relational words on transfer between well-aligned situations would fit with other research on words, comparison, and abstraction of relations (Kotovsky & Gentner, 1996; Gentner & Rattermann, 1991). This influence of language adds to analogy research that implicates contextual similarities in application of abstractions (Barnett & Ceci, 2002; Bassok, 1998; Ross, 1987, 1989).

Because structure is often learned *in situ* with no immediate need for decontextualization, the GT aspect of words may facilitate transfer by encouraging decontextualization and abstraction. If the very act of labeling serves as an invitation to form abstract, relational concepts (as suggested by Gentner & Namy, 1999; Gentner & Rattermann, 1991), then we may have seen all labels fostering better relational responding. Instead, the beneficial effects of labeling were mediated by how easy it was to actually form relational concepts, assuming that alignment makes it easier to compare and abstract relations. Although previous explanations of this "invitation to abstract" suggests that comparison occurs when two items share the same label, our experiments show that the presence of labels in even one entity plays an important role in relational extraction. This begs the question, how do words—that we have learned *in the past*—impact our ability to notice difficult relations in a scenario?

Our GT hypothesis suggests that the presence of relational words might encourage students to represent the relations and downplay contextual features. This differentiation between relations and concrete features would allow relational similarities to be selectively attended and also further encapsulated by the words. For example, the features that are in common between healthy athlete and *targets* could be used to redefine *target* in a manner consistent with SDT. Thus, learners would be primed with a notion of target that could be useful for understanding sweet melons in the transfer situation. If the word *target* is applied to a sick person, the meaning of the word might be adjusted to reflect a target that is bad and must be spotted and weeded out. In this case, learners would be primed with a notion of target useful for looking for infected melons.

Also, the relations represented by *target*, by being anchored by a word, can be more concretely understood, exerting a larger influence on how scenarios are interpreted. The label *target* may come to mean lots of things, including "goal (SDT signal)" and "good thing" (in the case of two overlapping good things in Experiment 1, and perhaps "thing that are important to detect" in Experiment 2), and even though the label was used with only the former meaning in mind, the latter is also acquired. Local features such as "good thing" may be important in aligning the stories, which in turn helps participants become sensi-

tive to deeper SDT structure. These local, and often superficial, similarities are critical in theories of relational reasoning, such as the Structure-Mapping theory (Gentner, 1983), as well as computational models (e.g., MAC/FAC with SME, Forbus, Gentner, & Law, 1995; LISA, Hummel & Holyoak, 1997; DORA, Doumas & Hummel, 2005). When relational and featural similarities both support a particular alignment (SME) or binding (LISA and DORA), there is a greater likelihood for success in relational responding in these models.

This theory of words as priming future relations is best captured by this metaphor: language is a filter for our perceptual experiences. This perspective would suggest that words simplify our experiences (Clark, 1997) by helping us ignore and/or highlight certain aspects. In learning new words, we may also acquire tools (Clark, 1997; Gentner, 2003; Vygotsky, 1962) that aid our capability to selectively attend. And because language has such a primary role in communication, language may also reshape the perceptions of others as well as our own experiences. In this way, the jargon used in a particular community (i.e. scientific, cultural, geographical) goes beyond communicating about ideas.

If we accept that language can act as a filter, this should also help us understand how it facilitates the processes of abstraction. One way of defining abstraction is to define it as a process of simplification through stripping away irrelevant information and retaining only critical information. Having language with words that label ideas might help us reduce the complex world, allowing us to abstract key information. This process of simplification and reduction may be at the foundation of what makes human reasoning so flexible and sophisticated. Whether our reasoning abilities are augmented through words and labels or some other cognitive tool, these boosts may be at the heart of what makes us smart.

## Appendix

Tutorial Quiz (Experiment 1)

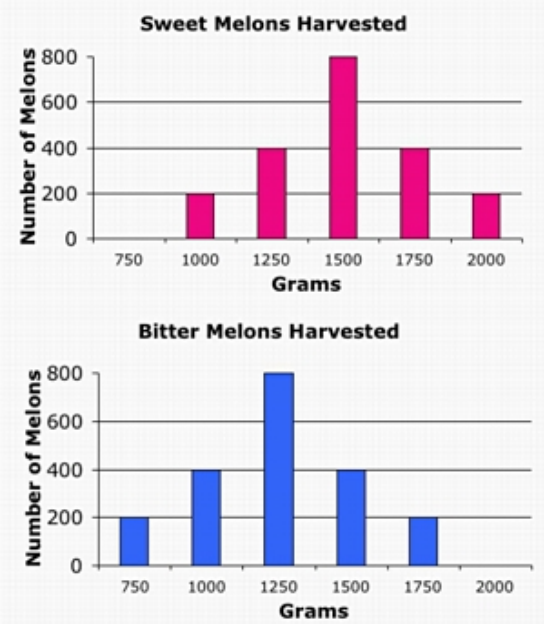
1. The numbers of actually healthy and unhealthy people are the same two months in a row. However, in the second month, the doctor is diagnosing more patients as healthy when they are actually healthy and more people as healthy when they are actually unhealthy. What must have changed in the second month?
   **a. The doctor must be diagnosing people with weaker cells as healthy.**
   b. The doctor must be diagnosing people with stronger cells as healthy.
   c. The doctor must be diagnosing more people who are actually healthy as unhealthy.
   d. The doctor must have become better at diagnosing healthy people.
2. For a particular kind of cell, the doctor knows from his experience this month that there is a 50% chance that this level of cell strength indicates anemia. What does this mean?

a. 50% of people with anemia have this level of cell strength.
b. 50% of all the patients the doctor has seen this month have anemia.
c. The doctor has seen equal numbers of people with anemia and people with strong cells this month.
**d. The doctor has seen equal numbers of healthy people with this level of strength and unhealthy people with this level of strength.**

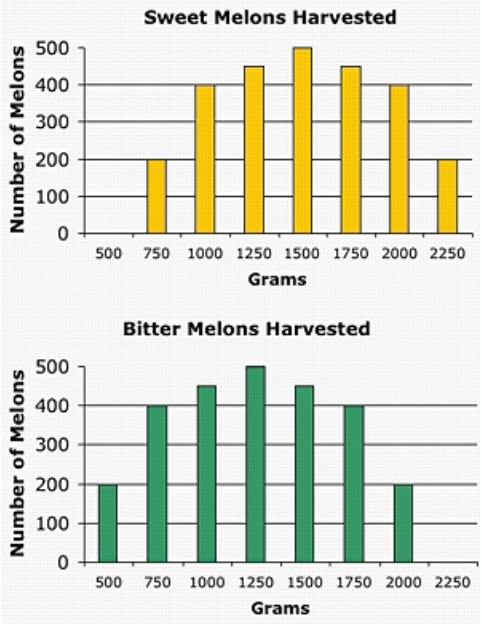
3. The doctor is looking into a new blood test for finding weakened cells. How can he find out whether this new test is better than the old one?
   a. The doctor changes his decision boundary and diagnoses more healthy people as healthy.
   b. The doctor changes his decision boundary and diagnoses more strong-celled samples as healthy.
   **c. The doctor does not change his decision boundary and diagnoses more healthy people as healthy.**
   d. The doctor does not change his decision boundary and diagnoses more weak samples as unhealthy.
4. If the doctor moves his decision boundary all the way to include even extremely strong as evidence for anemia, it means:
   a. he is generally more accurate because he is able to make less errors.
   b. he never mistakenly diagnoses healthy people as unhealthy.
   **c. he always diagnoses people as unhealthy when they are actually healthy.**
   d. he always diagnoses people as healthy when they are actually healthy.
5. This month, each sick person's cells get weaker while healthy people's cells do not get better or worse. The doctor does not know this information. If the doctor does not change his decision boundary, how does this change in the population help him?
   a. he increases the number of actually healthy people he diagnoses as healthy.
   **b. he decreases the number of actually sick people he diagnoses as healthy.**
   c. he increases the number of actually healthy people he diagnoses as sick.
   d. sick people become more common so he gets more experience diagnosing them.
6. Which of the following decision strategies will ensure that the doctor maximizes the number of actually healthy people he diagnoses as healthy?
   **a. diagnose everyone as healthy.**
   b. look more carefully at the cell distortion levels before his diagnosis.
   c. examine the previous month's ratio of healthy patients to unhealthy patients before his diagnosis.
   d. examine the previous month's ratio of patients with strong cells to patients with weak cells before his diagnosis.

7. Which is most likely to lead to inaccuracy in the doctor's diagnoses?
   a. Unhealthy people develop extremely weak cells.
   **b. Unhealthy people and healthy people have similar cell strength levels.**
   c. The people diagnosed as healthy all have similar distortion levels.
   d. Weak cells are more common among unhealthy people.
8. Very strong cells are often enriched with protein bundles. Knowing this, the doctor's accuracy can:
   a. improve at detecting who is actually healthy and unhealthy.
   b. improve at detecting who is actually unhealthy.
   c. improve at detecting who is actually healthy.
   **d. not improve based on this information.**

Transfer Quiz (Experiments 1 and 2)



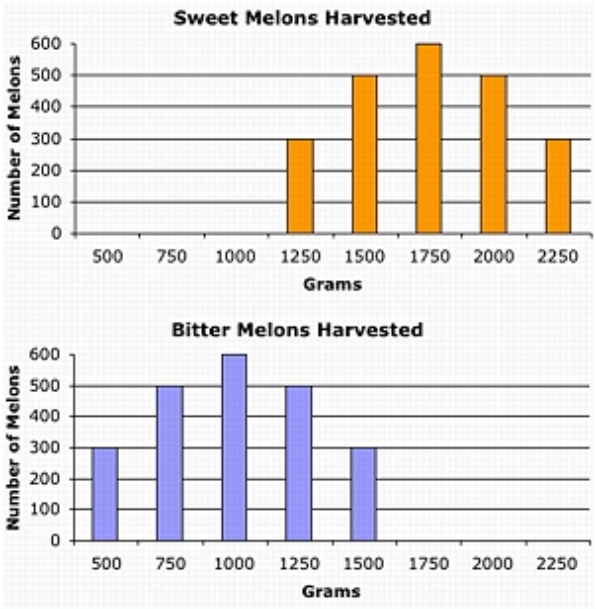

1. (Use graph above.) Approximately what percentage of all 1000 gram melons (1 kg) are sweet?
   a. 10%
   b. 25%
   **c. 33%**
   d. 50%
   e. 66%




2. (Use graph above.) There was a very bitter shipment of melons last year so the townspeople wanted to be extremely careful this year. They set a 1750 gram minimum weight but they do not know which melons are sweet or which melons are bitter. How many melons that weighed 1500 grams were rejected?
   a. 300
   b. 450
   c. 500
   d. 750
   **e. 950**
3. (Use provided graph—question 2's graph) With the minimum weight for the pluma melon set at 1750 grams, how many of bitter pluma melons are rejected?
   a. 400
   b. 950
   c. 1550
   **d. 2000**
   e. 2600
4. Some of the people in Chanterais debate over using a high-tech digital scale in place of their old-fashioned analog scale. What would be evidence that the high-tech scale is a better diagnostic?
   a. Chanterais changes their required weight and exports more sweet fruit.
   b. Chanterais changes their required weight and rejects more bitter fruit.
   **c. Chanterais does not change their required weight and rejects more bitter fruit.**

d. Chanterais does not change their required weight and rejects more lightweight fruit.

5. For 1750 gram melons, Chanterais knows from last month that there is a 25% chance that these melons are sweet. What does this mean?
   a. 25% of sweet melons will weigh 1750 grams.
   b. 25% of the melons will be sweet.
   c. 75% of the melons will be bitter.
   **d. 25% of 1750 gram melons will be sweet.**
6. If Chanterais lowers their minimum weight, which of the following would happen?
   a. They will export more sweet fruit and less bitter fruit.
   b. They will never export sweet fruit.
   c. They will export less sweet fruit and more bitter fruit.
   **d. They will export more sweet fruit.**
7. In a particular year, there is plenty of rainfall and all the melons get about 250 grams heavier. The prior year Chanterais exported melons that weighed 1500 grams or more. If they do not change their policy:
   a. Chanterais will only accept more heavy melons that are sweet.
   b. Chanterais will only reject more light melons that are sweet.
   **c. Chanterais will accept more melons that are sweet.**
   d. Chanterais will reject more melons that are lightweight.




8. How does this graph support the idea that melon weight is a good predictor of sweet melons?

a. There are fewer heavy melons that are bitter than are sweet.
b. There are fewer light melons that are bitter than are sweet.
**c. There are fewer light melons than heavy melons.**
d. There are more sweet melons than bitter melons.
e. There are more heavy melons than light melons.

Analogy questions for Experiment 1

1. A patient diagnosed as sick but is actually healthy is like what?
   a. A bitter melon that is rejected.
   b. A bitter melon that is exported. (target-to-distracter)
   **c. A sweet melon that is rejected. (target-to-target)**
   d. A sweet melon that is exported.
2. What in the doctor story is most analogous to a heavy melon?
   **a. A patient with strong cells. (target-to-target)**
   b. A patient with weak cells. (target-to-distracter)
   c. A patient who is sick.
   d. A patient who is healthy.
3. A melon that is sweet but was rejected is analogous to:
   a. A sick patient who had been diagnosed as sick.
   b. A sick patient who had been diagnosed as healthy. (target-to-distracter)
   c. A healthy patient who had been diagnosed as healthy.
   **d. A healthy patient who had been diagnosed as sick. (target-to-target)**
4. What in the melon export story is most analogous to the sick patient in the doctor scenario?
   a. A sweet melon. (target-to-distracter)
   **b. A bitter melon. (target-to-target)**
   c. An exported melon.
   d. A rejected melon.
5. The patient with anemia who has been diagnosed as sick is most like:
   a. A melon that is rejected and sweet.
   **b. A melon that is rejected and bitter. (target-to-target)**
   c. A melon that is exported and bitter.
   d. A melon that is exported and sweet. (target-to-distracter)
6. An exported melon is like:
   a. A patient who has been given weak cell results.
   b. A patient who has been given strong cell results.
   c. A patient who has been given a sick diagnosis. (target-to-distracter)
   **d. A patient who has been given a healthy diagnosis. (target-to-target)**

Analogy questions for Experiment 2

1. A patient diagnosed as sick but is actually healthy is like what?
   a. A bitter melon that is rejected.
   **b. A bitter melon that is accepted. (target-to-target)**
   c. A sweet melon that is rejected. (target-to-distracter)
   d. A sweet melon that is accepted.
2. What in the doctor story is most analogous to a heavy melon?
   **a. A patient with distorted cells. (target-to-target)**
   b. A patient with pure cells. (target-to-distracter)
   c. A patient who is sick.
   d. A patient who is healthy.
3. A melon that is sweet but was rejected is analogous to:
   a. A sick patient who had been diagnosed as sick.
   **b. A sick patient who had been diagnosed as healthy. (target-to-target)**
   c. A healthy patient who had been diagnosed as healthy.
   d. A healthy patient who had been diagnosed as sick. (target-to-distracter)
4. What in the melon export story is most analogous to the sick patient in the doctor scenario?
   **a. A sweet melon. (target-to-target)**
   b. A bitter melon. (target-to-distracter)
   c. An exported melon.
   d. A rejected melon.
5. The patient with leukemia who has been diagnosed as sick is most like:
   a. A melon that is rejected and sweet.
   b. A melon that is rejected and bitter. (target-to-distracter)
   c. A melon that is exported and bitter.
   **d. A melon that is exported and sweet. (target-to-target)**
6. An exported melon is like:
   a. A patient who has been given low distortion test results.
   b. A patient who has been given high distortion test results.
   **c. A patient who has been given a sick diagnosis. (target-to-target)**
   d. A patient who has been given a healthy diagnosis. (target-to-distracter)

## References

Bakke, A., & Gravemeijer, K. (2005). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking.* Dordrecht: Kluwer Academic Publishers.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64,* 3–26.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637.

Bassok, M. (1998). Using content to interpret structure: Effects on analogical transfer. *Current Directions in Psychological Science, 5,* 54–58.

Bowerman, M., & Choi, S. (2003). Space under construction: Language-specific spatial categorization in first language acquisition. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition.* Cambridge, MA: MIT Press.

Boysen, S. T., & Bernston, G. G. (1989). Numerical competence in a chimpanzee (Pan troglodytes). *Journal of Comparative Psychology, 103,* 23–31.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education, 24,* 61–100.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65,* 14–21.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22(4),* 1020–1031.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, *127(4),* 355–376.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes.* Cambridge: Cambridge University Press.

Dennett, D. (1991). *Consciousness explained.* New York: Little, Brown & Co.

Dixon, J. A., & Bangert, A. S. (2004). On the spontaneous discovery of a mathematical relation during problem solving. *Cognitive Science, 28,* 433–449.

Doumas, L. A. A., & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society* (pp. 606–611). Mahwah, NJ: Lawrence Erlbaum Associates.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115(1),* 1–43. doi:10.1037/0033-295X.115.1.1

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. Van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition.* Cambridge, MA: MIT Press.

Forbus, K. D., Ferguson, R. W., & Gentner, D. (1994). Incremental structure-mapping. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 313–318). Atlanta, GA: Lawrence Erlbaum Associates.

Forbus, K. D., Gentner, D., & Law, K. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19(2),* 141–205. doi:10.1016/0364-0213(95)90016-0

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. In D. A. Norman, D. E. Rumelhart, & The LNR Research Group (Eds.), *Explorations in cognition* (pp. 211–246). San Francisco: Freeman.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe & D. H. Rakison (Eds.), *Building object categories in developmental time* (pp. 245–275). Hillsdale, NJ: Erlbaum.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and cognition*. Cambridge, MA: MIT Press.

Gentner, D., & Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175). Washington, DC: APA.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65*, 263–297.

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development, 14,* 487–513.

Gentner, D., & Namy, L. L. (2004). The role of comparison in children's early word learning. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 533–568). Cambridge, MA: MIT Press.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London: Cambridge University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10,* 277–300.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102(4),* 684–704.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178–200.

Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46,* 414–466.

Green, C. B., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1107–1119.

Gumperz, J. J., & Levinson, S. C. (1991). Rethinking linguistic relativity. *Current Anthropology, 32,* 613–623.

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*, 803–831.

Hall, D. G., & Waxman, S. R. (1993). Assumptions about word meaning: Individuation and basic-level kinds. *Child Development, 64,* 1550–1570.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps.* Cambridge, MA: MIT Press.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Mahwah, NJ: Erlbaum.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*, 480–517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Hummel, J. E., & Stankiewicz, B. J. (1996). Categorical relations in shape perception. *Spatial Vision, 10*, 201–236.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science, 320,* 454–455.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23,* 221–236.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*, 2797–2822.

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

Leech, R., Mareschal, D., & Cooper, R. (in press). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology, 50,* 315–353.

Lupyan, G. (2006). Labels facilitate learning of novel categories. In A. Cangelosi, A. D. M. Smith, & K. R. Smith (Eds.), *The evolution of language: Proceedings of the 6th International Conference.* Singapore: World Scientific.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18*, 1077–1083.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32,* 517–535.

Markman, A. B., & Gentner, D. (2000). Structure mapping in the comparison process. *The American Journal of Psychology, 113,* 501–538.

Namy, L. L. (2001). What's in a name when it isn't a word? 17-month-olds' mapping of nonverbal symbols to object categories. *Infancy, 2,* 73–86.

Pica, P., Lemer, C., Izard, W., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigenous group. *Science, 306,* 499–503.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rattermann, M. J., & Gentner, D. (1998). The effect of language on similarity: The use of relational labels improves young children's performance in a mapping task. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. Sofia: New Bulgarian University.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94(3),* 249–273.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 456–468.

Smith, L. B., Rattermann, M. J., & Sera, M. (1988). "Higher" and "lower": Comparative and categorical interpretations by children. *Cognitive Development, 3,* 341–357.

Son, J. Y., & Goldstone, R. L. (2009). Contextualization in Perspective. *Cognition and Instruction, 27*, 51–89. doi:10.1080/07370000802584539

Son, J. Y., Smith, L. B., Goldstone, R. L., & Leslie, M. (under review). The importance of being interpreted: Words and children's relational reasoning.

Spivey, M. J. (2007). *The continuity of mind*. New York: Oxford University Press.

Thompson, R. K., & Oden, D. L. (1993). "Language training" and its role in the expression of tacit propositional knowledge in chimpanzees (Pan troglodytes). In H. L. Roitblat, L. M. Herman, & P. E. Nachtigall (Eds.), *Language and communication: Comparative perspectives* (pp. 365–384). Hillsdale, NJ: Erlbaum.

Thompson, R. K., Oden, D. L., & Boysen, S. T. (1997). Language-naïve chimpanzees (Pantroglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes, 23,* 31–43.

Vygotsky, L. S. (1986). *Thought and language* (*1962, trans.).* Cambridge, MA: MIT Press.

Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science, 6*, 130–137.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology, 29*, 257–302.

Whorf, B. L. (1956). *Language, thought and reality* (ed. J. B. Carroll). Cambridge, MA: MIT Press.