*Metaphor and Symbol*, 26: 206–226, 2011

ISSN: 1092-6488 print / 1532-7868 online
DOI: 10.1080/10926488.2011.583196

Psychology Press
Taylor & Francis Group

# Metaphor Aptness and Conventionality: A Processing Fluency Account

Paul H. Thibodeau
*Stanford University*

Frank H. Durgin
*Swarthmore College*

Conventionality and aptness are two dimensions of metaphorical sentences thought to play an important role in determining how quick and easy it is to process a metaphor. Conventionality reflects the familiarity of a metaphor whereas aptness reflects the degree to which a metaphor vehicle captures important features of a metaphor topic. In recent years it has become clear that operationalizing these two constructs is not as simple as asking naïve raters for subjective judgments. It has been found that ratings of aptness and conventionality are highly correlated, which has led some researchers to pursue alternative methods for measuring the constructs. Here, in four experiments, we explore the underlying reasons for the high correlation in ratings of aptness and conventionality, and question the construct validity of various methods for measuring the two dimensions. We find that manipulating the processing fluency of a metaphorical sentence by means of familiarization to similar senses of the metaphor ("in vivo conventionalization") influences ratings of the sentence's aptness. This misattribution may help explain why subjective ratings of aptness and conventionality are highly correlated. In addition, we find other reasons to question the construct validity of conventionality and aptness measures: for instance, we find that conventionality is context dependent and thus not attributable to a metaphor vehicle alone, and we find that ratings of aptness take more into account than they should.

Most people agree that the metaphor *"Memory is a warehouse"* is easier to understand than the metaphor *"A fisherman is a spider."* But is this because the former is more familiar than the latter or because the metaphor vehicle *warehouse* captures the important features of *memory* better than the metaphor vehicle *spider* captures the important features of *a fisherman*?

Recent work in the metaphor-processing literature has identified these two dimensions of metaphorical sentences—conventionality and aptness—as candidates for explaining variation in metaphor processing fluency, the speed and ease with which people process a metaphor (e.g., Bowdle & Gentner, 2005; Clement & Gentner, 1991; Glucksberg, 2008; Glucksberg & Haught, 2006a, 2006b; Jones & Estes, 2005, 2006). Conventionality reflects the familiarity of a metaphor whereas aptness reflects the degree to which a metaphor vehicle captures important features of a metaphor topic.

Address correspondence to Paul H. Thibodeau, Stanford University, Department of Psychology, Building 01-420, 450 Serra Mall, Stanford, CA 94305. E-mail: pthibod1@stanford.edu

It has been argued that gaining a better understanding of how conventionality and aptness impact metaphor processing fluency will yield insight into the mechanisms that underlie metaphor processing. Two alternative models of metaphor processing have been proposed. One relies on a categorization-based mechanism (Glucksberg, 2001, 2008; Glucksberg & Haught, 2006a, 2006b; Glucksberg & Keysar, 1990; Glucksberg, McGlone, & Manfredi, 1997; Honeck, Kibler, & Firment, 1987; Kennedy, 1990). On this view, a metaphor vehicle is the prototypical member of a dynamically created category and a metaphorical sentence serves to identify the metaphor topic as a member of this category. For instance, in the example above, *warehouse* would be considered the prototypical member of the category of "vast but bounded spaces for storage," and *memory* would be considered an exemplar of this category. Variance in the speed and ease with which people process a metaphor, according to this approach, is related to the ease with which this taxonomic relationship can be established. That is, if *memory* is already thought of as a vast but bounded space for storage, then it will be considered highly apt and it will be a good fit to the *warehouse* category. As a result, it will be quick and easy to process.

An alternative model proposes that the mechanisms that underlie metaphor comprehension differ depending on the familiarity of the metaphor. This is the career of metaphor hypothesis (Bowdle & Gentner, 2005). On this view, novel and conventional metaphors are processed as category assertions (as detailed above), but novel metaphors require a relatively more intensive comparison-based process like structure mapping (Clement & Gentner, 1991; Gentner, 1982, 1983; Gentner, Bowdel, Wolff, & Boronat, 2001; Gentner & Wolff, 1997; Holyoak & Thagard, 1995; Indurkhya, 1987; Kittay & Lehrer, 1981; Murphy, 1996; Verbrugge & McCarrell, 1977). The theory behind this view is that when people encounter a completely novel metaphor like *"A fisherman is a spider*," they have to map the relational structure of *spider* onto that of *a fisherman* to understand the sentence. In this case, one might liken a patient spider to an enduring fisherman and a spider's net to a fisherman's line to generate a mental image of the sentence. On the other hand, when people encounter a highly conventional metaphor like *"Memory is a warehouse"* they may be able to directly retrieve an interpretation from memory.

Recent experimental evidence suggests that both dimensions are important determinants of the processing fluency of a metaphor. Several studies have found that the more conventional a metaphor, the faster people process it (Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997) while other studies have found that the more apt a metaphor, the faster people process it (Blasko & Connine, 1993; Chiappe & Kennedy, 1999; Chiappe, Kennedy, & Chiappe, 2003; Glucksberg & McGlone, 1999; Jones & Estes, 2005, 2006).

However, in recent years, it has become clear that operationalizing the two constructs may not be as easy as asking naïve raters for subjective ratings. For instance, Jones and Estes (2006) found that subjective ratings of conventionality and aptness are highly correlated and possibly confounded. This finding was surprising because it had been thought that the two constructs were orthogonal—a metaphor need not be conventional to be apt (e.g., *"Beavers are lumberjacks"*); similarly, a metaphor need not be apt to be conventional (e.g., *"The clue is a red herring"*).

Assuming that the conventionality and aptness of a given set of metaphors is not inherently related, there are several reasons why ratings of the two constructs could become so highly correlated. Raters could misattribute a metaphor's aptness for its conventionality. If this were the case, then ratings of aptness would actually reflect the conventionality of a metaphor. Alternatively, raters could misattribute a metaphor's conventionality for its aptness, in which case ratings of

conventionality would actually reflect the aptness of a metaphor. Or, finally, raters could misattribute some other variable to ratings of both constructs. If this were the case, then ratings of aptness and conventionality would at least partially reflect some other mediating variable.

Jones and Estes (2006) take the position that ratings of conventionality actually reflect the aptness of a metaphor. They attempt to resolve this confound by utilizing an alternative method for operationalizing conventionality (which we call metaphor-vehicle conventionality in contrast to metaphor-sentence conventionality). This method involves yoking the conventionality of a metaphor to a metaphor vehicle: for instance, according to this approach, all *blueprint* metaphors (e.g., *"A syllabus is a blueprint"*; *"A game-plan is a blueprint"*) are equally conventional.

Metaphor-vehicle conventionality is measured in two steps: First, a single figurative property that is commonly elicited by the given metaphor vehicle is identified. For instance, *blueprint* metaphors often highlight some notion of "planning," so we might identify the property "provides a plan" with the metaphor vehicle *blueprint*. Second, raters judge the degree to which this property is associated with the metaphor vehicle by rating "how conventional it is to use the concept (e.g., BLUEPRINT) to represent the given property (e.g., provides a plan)" (p. 23). Jones and Estes find that this method of operationalizing conventionality yields ratings that are not correlated with ratings of aptness, thus successfully differentiating the two constructs. However, a potential drawback of this approach, as we detail in Experiments 1 and 2, is that these ratings may not reflect a notion of familiarity. Adopting the metaphor-vehicle conventionality approach commits us to the view that the metaphor *"A monkey is a blueprint"* is just as conventional as *"A syllabus is a blueprint,"* which does not seem true.

In this paper, we take a close look at why ratings of aptness and metaphor-conventionality are so highly related and explore whether the standard and alternative methods for measuring the constructs are valid and reliable. In Experiment 1, we replicate Jones and Estes (2006) finding that subjective ratings of aptness and conventionality are highly correlated and evaluate the various methods for operationalizing conventionality. Unlike Jones and Estes (2006), we do not argue that conventionality ratings are aptness ratings in disguise. Instead, we take the position that people's metacognitive awareness of the processing fluency of the metaphors that they rate bleeds into their judgments of both constructs. Previous work by Jacoby and others has found that people tend to misattribute a general sense of processing fluency to abstract target dimensions that they are trying to rate (Alter & Oppenheimer, 2009; Jacoby, Allan, Collins, & Larwill, 1988; Jacoby & Whitehouse, 1989; Kelley & Jacoby, 1996). That is, raters may unwittingly rate the metaphor *"A memory is a warehouse"* high in conventionality and aptness *not* because the metaphor is especially familiar to them or because *warehouse* is a particularly good description of *memory*, but, in part, because the sentence is easy to process. If this is indeed what is going on then it would be problematic to use these ratings to predict reading time (RT) data, since RTs are a direct measure of processing fluency.

## EXPERIMENT 1

In Experiment 1, we set out to evaluate two methods of operationalizing the conventionality of metaphors: metaphor-sentence conventionality and metaphor-vehicle-conventionality. Metaphor-sentence conventionality is measured by directly rating metaphorical sentences for

their familiarity whereas metaphor-vehicle conventionality involves measuring the rated association strength between a figurative property and a metaphor vehicle.

First, we look at whether these methods yield conventionality ratings that are correlated with ratings of aptness. Jones and Estes (2006) find that metaphor-sentence conventionality ratings are highly correlated with and possibly confounded by aptness ratings but that metaphor-vehicle conventionality ratings are uncorrelated with aptness ratings.

Second, we compare both types of conventionality ratings to frequency counts of the metaphorical sentences in a corpus. Since the conventionality construct is intended to measure familiarity, we should expect to find a relationship between the actual prevalence of a given metaphor and its rated conventionality. That is, for either method of conventionality ratings to have construct validity, it should reflect the actual pervasiveness of the metaphor. Indeed, we believe that pursuing a method for generating conventionality measurements from a corpus of natural language would be a worthwhile future endeavor. This would provide a more objective means of valuing metaphors along this dimension. However, the approach that we adopt here is only a first step in this direction.

## Method

*Participants.* Seventy-two Swarthmore College undergraduates participated in the experiment in partial fulfillment of a course requirement. All were native English speakers.

*Materials and design.* In the analyses below, we will compare data collected by Jones and Estes (2006) to data that we gather here. We will use three pieces of data from the Jones and Estes data set: reported metaphor-vehicle conventionality ratings, reported aptness ratings, and metaphor reading time data, which were graciously provided to us by Jones and Estes. We will gather two additional pieces of data here: metaphor-sentence conventionality ratings and corpus frequency counts.

All of the data mentioned above was collected on the same set of 128 metaphoric sentences from Jones and Estes (2006). The stimulus set was modified from a sample of 100 metaphors that had been used previously for similar experiments (e.g., Katz, Paivio, Marschark, & Clark, 1998; McGlone, 1996; McGlone & Manfredi, 2001). Jones and Estes took the original list of 100 metaphors as a high-apt usage of the 100 metaphor vehicles. They created a corresponding low-quality version for each metaphor by substituting a less appropriate topic for the original topic. For instance, *"A rooster is an alarm clock"* would be the original (reasonably apt) usage of the *alarm clock* vehicle and *"A robin is an alarm clock"* would be the low-quality substitution. After gathering aptness ratings and metaphor-vehicle conventionality ratings for each of the 200 sentences, Jones and Estes selected 64 pairs of original and modified statements to be included in the final set of 128 metaphors. This yielded 32 metaphoric sentences in each of the four quadrants of a two-by-two table: high-apt and high-conventional, high-apt and low-conventional, low-apt and high-conventional, and low-apt and low-conventional. We emphasize that the "low-apt" stimuli were created with the intention of inducing low ratings of aptness, whereas (as we will argue below) it is probably best to classify these, a-theoretically, as "modified" variants of the original metaphor sentences.

To get an approximate sense of how frequently each of these 128 sentences are used in natural discourse we used the Google search engine to generate frequency counts for each of

the sentences. While we can gain a general sense of frequency from using the Google search engine, it should be noted that there are drawbacks to using Google as a corpus (e.g., Kilgarriff & Grefenstette, 2003).

To gather metaphor-conventionality ratings, we asked each participant to directly judge the conventionality of a subset of 32 metaphoric sentences from the set of 128 (eight from each quadrant of the two-by-two table). Ratings were made on a seven-point Likert scale. The instructions read as follows:

> For this section of the survey, you will be asked to judge the conventionality (commonness) of a variety of simple metaphoric sentences. Expressions can vary in conventionality with respect to any idea that they are intended to communicate. For example, consider the following two descriptions of the mind: a conventional one, *The mind is a computer*, and an unconventional one, *The mind is a food processor*. Both of these metaphors make sense and convey a similar idea: Information can be *stored* and *processed* and *digested* and *swallowed*. Nonetheless, the first metaphor, *The mind is a computer*, is clearly more common or conventional than the second one, *The mind is a food processor*.

*Procedure.* Because the metaphor-conventionality ratings task was untimed, participants were tested in groups in a classroom. Ratings were made on paper surveys. The session lasted approximately 30 minutes.

## Results and Discussion

With regard to the metaphor-conventionality ratings, we computed a mean metaphor-conventionality score for each of the 128 sentences in the set. The distribution of the mean conventionality ratings ranged from 1.6 to 7.0. The overall mean rating was 4.27 and the standard deviation was 1.47. We then correlated these conventionality scores with aptness ratings, vehicle-conventionality ratings, reading time data, and frequency counts of the same sentences.

With regard to the Google-generated frequency counts, we found that the distribution of frequency counts ranged from 1 to 4,420,000. The mean frequency was 89,729 (the median was 2) and the standard deviation was 502,654. Because the distribution was highly skewed, we log-transformed the frequency counts before comparing them with conventionality or aptness ratings.

*Metaphor-sentence conventionality and aptness.* Our findings confirmed those of Jones and Estes (2006): metaphor-sentence conventionality ratings and aptness ratings are highly correlated, $r = .87$, $t(126) = 20.11$, $p < .001$ (see Figure 1).

This could be because, as Jones and Estes (2006) argue, ratings of metaphor-conventionality actually reflect variation in aptness. However, this could also be because when people rate metaphoric sentences for aptness or conventionality their overall sense of how easy the metaphor is to process bleeds into their judgment.

*Metaphor-sentence and metaphor-vehicle conventionality.* To decouple aptness and conventionality ratings, Jones and Estes (2006) relied on ratings of metaphor-vehicle conventionality rather than metaphor-sentence conventionality. These ratings have the advantage of being uncorrelated with ratings of aptness. As a result, it should not be surprising to find that this measure of metaphor-vehicle conventionality is also uncorrelated with ratings of

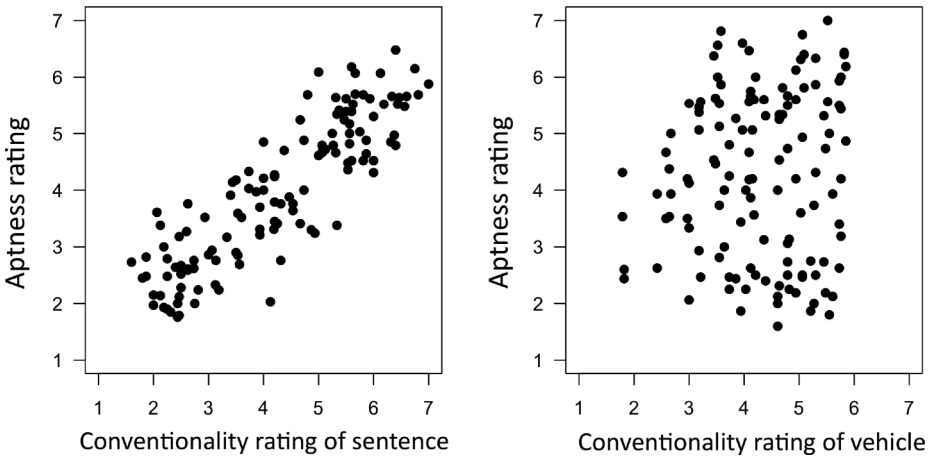


FIGURE 1 Aptness ratings (as reported by Jones and Estes, 2006) are compared with two methods for rating conventionality. On the left, raters judge the conventionality of metaphoric sentences as a whole. On the right, raters judge the conventionality of metaphor vehicles alone.

metaphor-sentence conventionality, $r = .07$ $t(126) = .74$, $p = ns$. However, this might also signal a problem with the construct of metaphor-vehicle conventionality.

*Conventionality and observed frequency.* We found a high correlation between ratings of metaphor-sentence conventionality and measured frequency, $r = .37$, $t(126) = 4.52$, $p < .001$, and a low correlation between metaphor-vehicle conventionality and measured frequency, $r = .035$, $t(126) = 4.52$, $p = ns$. This suggests that ratings of metaphor-sentence conventionality are related to the actual prevalence of the sentences used in the study but that ratings of metaphor-vehicle conventionality are not. Since metaphor-sentence conventionality and aptness ratings are highly correlated, it is not surprising that aptness ratings are also highly correlated with this measure of observed frequency, $r = .41$, $t(126) = 5.07$, $p < .001$. Although it may be that frequency of use is a result of aptness, it is equally likely that ratings of aptness are contaminated by actual familiarity (i.e., metaphor-sentence conventionality).

*Reading time and observed frequency.* Log frequency counts (actual conventionality) were highly correlated with how quickly and easily people are able to comprehend metaphors, $r = -0.249$, $t(126) = 2.89$, $p < .01$. Although it is possible that ease of processing also predicts metaphor frequency, the implication that metaphor frequency predicts processing speed provides additional evidence for the value of metaphor-sentence conventionality as a theoretical construct.

*Interim conclusions.* When analyzed at the sentence level, ratings of metaphor-sentence conventionality and of aptness are indeed highly correlated. But both measures are also highly correlated with the observed frequency of occurrence of the relevant sentences in language databases, which in turn strongly predicts sentence RTs. Given the relationship with observed frequencies in the language, the claim that raters are successful at rating metaphor aptness when asked to so, but fail to accurately judge metaphor conventionality seems poorly supported.

Conversely, although using metaphor-vehicle conventionality ratings instead of metaphor-sentence conventionality ratings successfully decouples ratings of the two dimensions, this alternative method of defining conventionality does not exhibit construct validity: This may be because the metaphoric sense of *alarm clock* that applies to *rooster* is not really the same sense that applies to *robin* (e.g., a rooster is regarded as reliable, loud, and connected to a fixed timer—sunrise—whereas none of these properties are associated with a robin.) We suggest that ratings of metaphor-vehicle conventionality do not predict the interpretability of metaphoric sentences because not all metaphoric senses that employ the same nominal metaphor vehicle are equally familiar (i.e., the specific metaphor sense that is evoked by a particular topic can vary quite a bit).

## EXPERIMENT 2

In Experiment 1 we found that observed sentence frequency was correlated with reading time. Here we test whether the direct manipulation of familiarity with a specific metaphor sense also affects comprehension time. In particular, we contrasted the effects of familiarization when applied to the same sense or a different sense of a specific metaphor vehicle.

We adapted a paradigm designed by Bowdle and Gentner (2005) called "in-vitro conventionalization." Bowdle and Gentner found that they could facilitate the processing of a metaphoric sentence by first exposing people to examples of similar metaphors (i.e., metaphors that use the same vehicle and instantiate a similar meaning as the target sentence). For instance, if the target sentence is *"Education is a lantern,"* then the priming sentences might read *"A mentor is a lantern"* and *"An encyclopedia is a lantern."* Because the prime sentences and the target sentence all highlight the power of lanterns to metaphorically light the darkness of ignorance, the target metaphor becomes easier to understand. At the same time, Bowdle and Gentner found that literal prime sentences like *"A camp light is a lantern"* or *"A torch is a lantern"* did not facilitate the processing of a related target metaphoric sentence, thereby ruling out the possibility that the facilitation effect was driven by lexical priming.

To clarify the importance of priming the specific metaphoric sense, we include here another class of prime: alternative-sense primes. Whereas same-sense metaphor primes are metaphoric sentences that use a target sentence's vehicle to instantiate a similar meaning as the target sentence, alternative-sense primes are metaphoric sentences that use a target sentence's vehicle to instantiate a meaning that is dissimilar to that of the target sentence. For example, an alternative-sense sentence for lantern could evoke the concept of a signaling device as in *"Education is a lantern"* or *"A flag is a lantern."*

If, as we argue, the conventionality of a metaphor is specific to its contextual sense, as defined by its relationship to the topic of the sentence, then same-sense metaphor primes alone should facilitate the processing of related target sentences.

### Method

*Participants.* Eighty Swarthmore College and Stanford University undergraduates participated in the experiment in partial fulfillment of a course requirement or pay. All were native English speakers. Data from nine participants was not included in the analysis because these participants either did not follow instructions or did not complete the experiment.

*Materials and design.* The experiment consisted of two phases: a priming phase and a test phase. Sentences for the test phase were taken from the appendix of Jones and Estes (2006)—one sentence for each of the 64 metaphor vehicles. For each metaphor vehicle, we randomly chose whether to select the Original or the Modified version of the sentence; however, we made sure to select 32 Original and 32 Modified sentences. We then randomly assigned 16 of the 64 sentences to be "fillers" (8 were Original and 8 were Modified versions) and the remaining 48 sentences to be "target" sentences.

Sentences for the priming phase were created by the authors. Each target sentence (e.g., *Education is a lantern*) was the basis for three types of prime sentences: *same-sense*, in which a similar metaphoric mapping was instantiated (e.g., *"A mentor is a lantern"*); *alternative-sense*, (e.g., *"A flag is a lantern"*); and *literal-sense* (e.g., *"A camp light is a lantern"*). Two sentences of each type were created for each target sentence for a total of 288 prime sentences (see the appendix for the complete stimulus set). In our analyses below, we compare RTs for target sentences (e.g., *"Education is a lantern"*) as a function of the three different types of priming conditions as well as an unprimed condition.

Four versions of the experiment were created. In each case the test phase consisted of the same 48 test sentences (plus 16 fillers) and the prime phase consisted of 24 same-sense primes (for 12 of the target sentences), 24 alternative-sense primes (for another 12 of the target sentences), and 24 literal primes (for another 12 of the target sentences). In each version there were also 12 target sentences that were not primed. Between participants, each target sentence was presented in each prime context.

The participants' task in the priming phase of the experiment was to rate the *metaphoricity*—the degree to which the sentences were metaphorical – of the sentences on a scale from 1 to 5. We chose this task to ensure that participants would carefully read the prime sentences.

The participants' task in the test phase of the experiment was modeled on Jones and Estes (2006). Participants were asked to read and think of an interpretation of the given sentence. Before each sentence appeared, a '******' was presented on the screen for 500 milliseconds. When it disappeared, a target sentence took its place. Participants were instructed to press the spacebar once they had an interpretation in mind. When they pressed the spacebar, they were prompted to type in their interpretation.

The experiment was implemented in Java. At the beginning of the experiment, participants were given oral and written instructions about the nature of the experiment. The oral instructions emphasized speed and accuracy as well as the logistics of the computer program. The written instructions explained the task. They read as follows:

> On the screens that follow, please read each sentence, think about what it means, and judge how metaphorical it is. Indicate a rating from 1 (not metaphorical) to 5 (very metaphorical) by typing the number at the top of the keyboard. Then press spacebar to record your answer and move on to the next sentence.

After the participants completed the prime phase, the program explained the second half of the experiment. These instructions read as follows:

> In this part of the experiment, you will again read sentences one at a time; however, this time all of the sentences will be metaphorical. Your task is to carefully read the sentence and decide what it means.

When you have an interpretation in mind, press the spacebar and then type your interpretation into the textbox that appears. After you write your interpretation, press the return (enter) key and prepare for the next sentence.

*Procedure.* Participants were tested in individual testing rooms. The session lasted approximately 30 minutes. The order of the sentences was randomized in both the study and test phase of the experiment. However, the order of the test phase was set so that the 16 filler sentences were the first 16 sentences of the test phase. They were used to acclimate the participant to the test phase of the experiment. We did not analyze data collected on the filler sentences.

## Results and Discussion

*Metaphoricity ratings.* In designing the study, we were not interested in the metaphoricity ratings of the prime sentences *per se*; however, results from this task can be analyzed to ensure that the metaphoric primes (same-sense and alternative-sense) were considered more metaphoric than the literal primes. To test this, we fit a linear mixed-effects regression model with random intercepts for subjects and items (Baayen, Davidson, & Bates, 2008). T-tests on the coefficients from the model revealed that literal primes (mean = 1.39, $SD = .79$) were rated as significantly less metaphoric than the alternative-sense primes (mean = 3.17, $SD = 1.37$; $t = 57.9$, $p < .001$) and the same-sense primes (mean = 3.49, $SD = 1.30$; $t = 48.8$, $p < .001$). It also revealed that same-sense primes were judged as slightly more metaphoric than alternative sense primes, $t = 9.02$, $p < .01$.

*Reading times.* Extreme reading times were trimmed so that response times below 50 ms (button-press errors) were removed, as were response times above 15,000 ms. Average reading times as a function of Priming Condition are shown separately for Original and for Modified metaphoric sentences in Figure 2.

We conducted the statistical analyses of the RT data with log-transformed RTs to correct for skewness in the data. The linear mixed-effects regression model found that participants were significantly faster to read target sentences when they were preceded by same-sense metaphor primes than when they were preceded by alternative-sense metaphor primes ($t = 2.29$, $p < .05$), literal primes ($t = 2.15$, $p < .05$), or no prime ($t = 2.93$, $p < .01$). Although Original metaphoric sentences were comprehended much more rapidly than Modified metaphoric sentences ($t = 4.73$, $p < .001$), this factor did not interact with prime condition. Adding prime metaphoricity to the model as a covariate did not alter the reliability of the various effects.

These findings replicate and extend those of Bowdle and Gentner (2005). On the one hand, we have confirmed that processing one metaphoric sentence can lead to speeded processing of another sentence that employs the same metaphor vehicle. This replicates the observation that the experimental manipulation of familiarity with a specific metaphor renders similar metaphors easier to understand. However, because we also tested alternative-sense primes, we have additionally shown that simply sharing a metaphor vehicle is not enough to facilitate the processing of a target sentence. The prime and target metaphors must also instantiate a similar meaning.

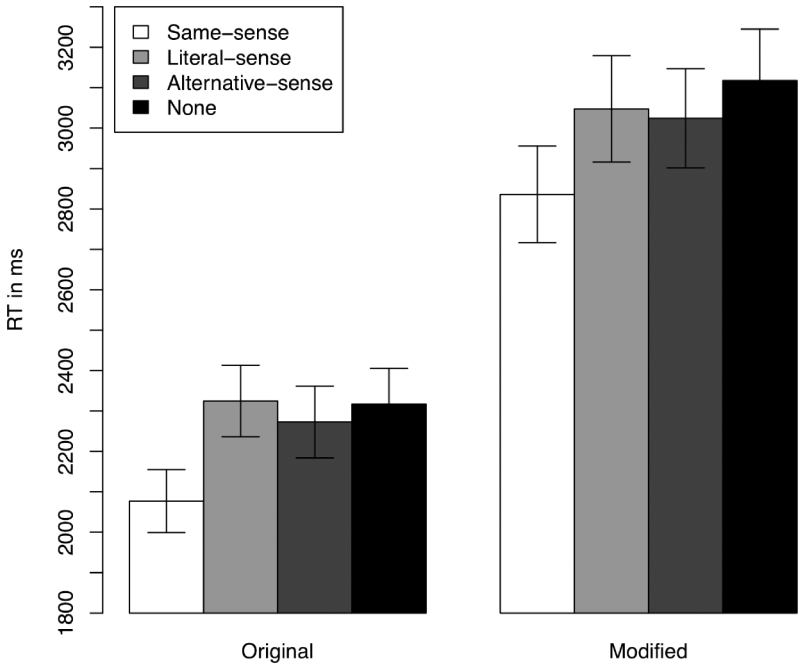


FIGURE 2 Mean RTs for target sentences grouped by prime-type.

## EXPERIMENT 3

Given that ratings of aptness were shown to be correlated with actual sentence frequency in Experiment 1, it is natural to ask whether the priming manipulation of Experiment 2 also has an effect on ratings of aptness. According to the definition of the construct, aptness reflects the degree to which the metaphor vehicle of a sentence captures important features of the metaphor topic. However, if ratings of aptness are more holistic and take into account ease of processing, then the same manipulation that affected comprehension time should affect ratings of aptness. That is, priming a metaphoric sense seems likely to influence ratings of aptness. To test this we repeated Experiment 2, but substituted ratings of aptness for the comprehension time measure.

### Method

*Participants.* Sixty-two Swarthmore College and Stanford University undergraduates participated in the experiment in partial fulfillment of a course requirement or pay. All were native English speakers.

*Materials and design.* The materials used in this experiment were identical to those in used in Experiment 2 and the design was similar to that of Experiment 2, except that the participant's task in the test phase of the experiment was to generate aptness ratings for the target sentences.

Following prior research, aptness was defined as "the extent to which the statement captures important features of the topic" (Chiappe, Kennedy, & Smykowski, 2003, p. 97) and participants were asked to indicate an aptness rating from 1 (not apt at all) to 7 (very apt).

*Procedure.* Because the task was untimed, participants were tested in groups in a computer classroom. The session lasted approximately 30 minutes. The order of the sentences was randomized in both the study and test phase of the experiment as in Experiment 2.

## Results and Discussion

*Metaphoricity ratings.* As in the analysis of Experiment 2, the metaphoricity rating task was designed to ensure that participants closely read the primes. As a result, we were not particularly interested in the results of this aspect of the experiment; however, evaluating these data affords an opportunity to ensure that the metaphoric primes were considered more metaphoric than the literal primes. To test this, we fit a linear mixed-effects regression model with random intercepts for subjects and items. The model confirmed that the alternative-sense metaphor primes (mean = 3.36, $SD = 1.37$, $t = 44.38$, $p < .001$) and same-sense metaphor primes (mean = 3.69, $SD = 1.26$, $t = 52.49$, $p < .001$) were judged more metaphorical than the literal primes (mean = 1.41, $SD = .79$). Additionally, as in Experiment 2, the same-sense primes were rated as significantly more metaphoric than the mix primes ($t = 6.4$, $p < .01$).

*Aptness ratings.* To analyze the results from the aptness rating task, we fit a linear mixed-effects regression model with random intercepts for subjects and items. We found that participants rated target sentences significantly higher in aptness when they were preceded by same-sense metaphor primes (mean = 4.43, $SD = 1.95$) than when they were preceded by alternative-sense metaphor primes (mean = 4.04, $SD = 2.06$; $t = 4.92$, $p < .001$) or literal primes (mean = 4.20, $SD = 2.07$; $t = 3.00$, $p < .01$). Although same-sense priming did not reliably increase aptness ratings relative to the un-primed condition overall (mean = 4.33, $SD = 2.03$; $t = 1.51$, $p = ns$), alternative-sense primes reliably reduced aptness ratings for target sentences compared to the un-primed condition, $t = 3.41$, $p < .01$. Thus, aptness ratings for metaphoric sentences are affected by experimental manipulations, such as priming, that are irrelevant to the theoretical construct of aptness.

As can be seen in Figure 3, aptness ratings were much higher for the Original metaphor sentences than for the Modified set, but the pattern of priming effects is similar for both. However, whereas the same-sense primed metaphors showed no overall increase in aptness compared to unprimed metaphors, within the Modified metaphors the contrast between unprimed sentences and those primed with same-sense metaphors was marginally reliable, $t = 1.87$, $p = .0621$. Including prime metaphoricity in the model did not alter the statistical conclusions.

The principal findings of the present experiment were that (1) priming with the same metaphoric sense led to an increase in rated aptness relative to priming with a literal sense, whereas (2) priming an alternative sense of a metaphor vehicle led to a reduction in rated aptness relative to the unprimed condition. Whereas we had predicted that processing fluency would increase ratings of aptness, the finding that priming an alternative sense reduced ratings of aptness implicates the idea that aptness ratings may reflect failures of correspondence as well as successes.

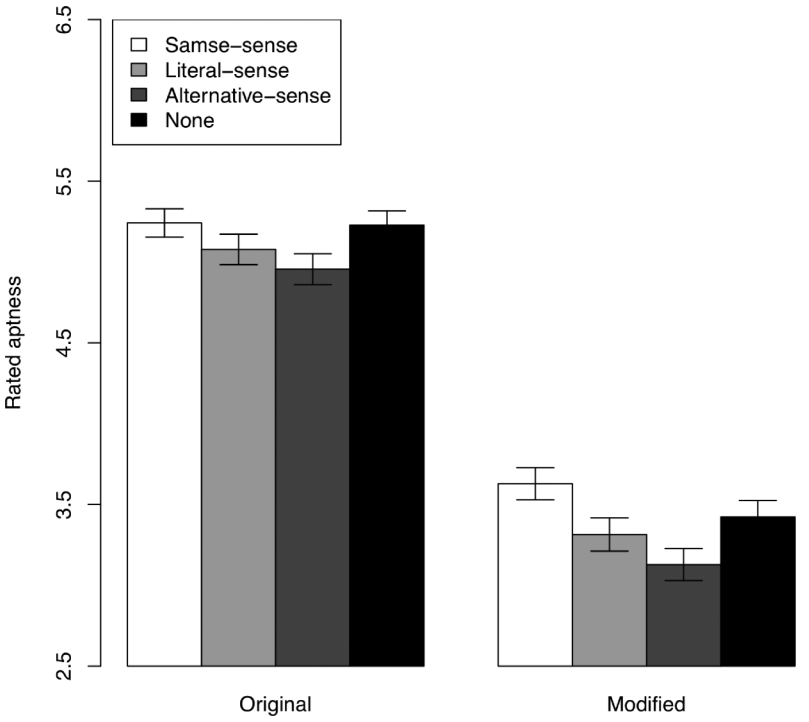


FIGURE 3 Mean aptness rating for target sentences grouped by prime-type.

## EXPERIMENT 4

Aptness nominally refers to the extent to which the metaphor vehicle captures important aspects of the topic. Thus, ratings of aptness are intended to measure the degree to which salient features of the metaphor vehicle are appropriate descriptions of the metaphor topic. Salient features of the metaphor vehicle that are irrelevant to the metaphor are outside the purview of the construct. Yet in Experiment 3 it was observed that participants reduced their aptness ratings when they had recently been exposed to an alternative sense of the target metaphor.

In this experiment, we asked raters to list salient properties of metaphor vehicles in the stimulus sentences, noting which ones were applicable to the metaphor topic and which ones were not. If ratings of aptness only reflect the presence of positive feature overlap, then a model that uses two predictors—one for positive feature counts and one for negative feature counts—should be no better than a model with only one predictor—for positive feature counts. However, if ratings of aptness reflect a more holistic consideration of the metaphorical sentence, then we might expect a model that includes a predictor for negative feature counts to be better than a model with only one predictor for positive feature counts.

## Method

*Participants.* We recruited sixty-five Swarthmore College undergraduates, all of whom were native English speakers, to contribute data in exchange for course credit.

*Materials and design.* They were presented with 128 metaphoric sentences (all taken from Jones & Estes, 2006) and instructed to list both salient properties of the vehicle that applied to the topic as well as salient properties of the vehicle that did not. The instructions read as follows:

> In the following pages, you will read short metaphoric sentences and you will be asked to list properties (if any) of the metaphor that capture important features of the subject of the sentence. Where appropriate, you should also list salient characteristics of the metaphor that seem inappropriate for the subject.
>
> For example, for the metaphor *My lawyer is a shark*, you might list *aggressive, dangerous, ruthless (cold-blooded),* and *tenacious* as properties of sharks that capture important features of the subject (lawyer). You should not list such properties as "is a vertebrate" even though both lawyers and sharks are vertebrates. For "miracle", as in "That I passed that test was a miracle." You might list *unlikely*, *surprising* and *fortunate*. There may also be cases where only one property seems to matter, such as "His stubble was sandpaper." Where "rough" seems to be the intended property. Conversely, if a metaphoric sentence seems not to have any appropriate properties, but implies an inappropriate one, such as the sentence "His spectacles were sandpaper", you would list "rough" under the second column of "inappropriate implied properties."

*Procedure.* Because the task was untimed, participants were tested in groups in a classroom. The rating task was done on paper surveys. The session lasted approximately 30 minutes.

## Results and Discussion

For each of the 128 items, the average number of positive listed features and negative listed features was computed. On average, people listed 1.71 ($sd = 0.788$) positive features and 1.29 ($sd = .58$) negative features for each sentence. We then correlated average feature lists with reported aptness ratings and fit linear mixed-effects models to the data. As expected, positive features were positively correlated with published ratings of aptness, $r = .43$, $t(126) = 5.30$, $p < .001$; however, interestingly, negative features were also strongly correlated with published ratings of aptness, $r = -.31$, $t(126) = -3.65$, $p < .001$. Further, a model including predictors for both applicable and inapplicable feature counts explained more variance than a model that included a predictor only for applicable feature counts, $\chi^2[1] = 2282.8$, $p < .001$. In other words, aptness ratings appear to reflect both the presence of salient applicable features and the presence of salient inapplicable features.

For the 64 Original metaphor sentences used by Jones and Estes (2006), the mean numbers of positive and negative features were 1.79 and 1.27, respectively (a ratio of 1.4:1). For the 64 Modified metaphor stimuli they developed, the two means were 1.62 and 1.34, respectively (a ratio of 1.2:1). Thus, the stimulus set that was developed by Jones

and Estes (2006) to make a theoretical argument about aptness may have differed in rated aptness partly as a result of differences that are not relevant to the theoretical construct of aptness. That is, the Modified metaphor sentences had a greater number of salient metaphor features that failed to map onto the topic than did the Original sentences, $t(63) = 2.56$, $p < .05$, while having fewer salient metaphor features that successfully mapped onto the topic, $t(63) = 4.20$, $p < .001$. This observation casts reasonable doubt on the theoretical sufficiency of ratings of aptness as an explanatory variable in scientific theories of metaphor processing. Ratings of aptness are indeed sensitive to variables that predict comprehension, but it is not clear that all of those variables have much to do with the theoretical construct of aptness. Moreover, we note that rather than listing a single salient feature (as per the method of generating metaphor-vehicle conventionality ratings), the mean number of salient features listed by our participants for each metaphor sentence was 3.

We do not suggest that this feature listing method directly reflects the calculation that participants make when rating the metaphoric sentences for aptness. Rather, we use this data to make the point that participants seem to be making a more holistic judgment about the interpretability of the metaphor at hand when they are judging aptness and that they may be considering more than one aspect of the vehicle.

## GENERAL DISCUSSION

In this paper, we have found reason to question previous assumptions and methods for operationalizing conventionality and aptness. With regard to conventionality, we have argued that the construct cannot be defined for vehicles independent of topics. Just because *"A syllabus is a blueprint"* is a conventional metaphor does not mean that all metaphors that employ *blueprint* as a metaphor vehicle are equally conventional. We speculate that using corpus-based frequency counts could provide a more objective measure of metaphor familiarity in the future (although the details of such a method are left for future work).

With regard to aptness, we find that subjective ratings of the construct (as gathered by Jones & Estes, 2006, for example) are sensitive to aspects of metaphors that go well beyond the imputed target dimension of aptness. Specifically, they reflect the presence of salient inapplicable properties of the metaphor vehicle as well as variation in metaphor processing fluency (i.e., by "in vitro conventionalization"). While the processing fluency of a metaphor will likely be impacted by true aptness, it will likely also be impacted by conventionality, the salience of inapplicable features, and other contextual factors, which could include perceived ease of structure mapping. For example, Glucksberg and Haught (2006a) found that aptness ratings were increased for a novel metaphor when an adjective was added that was literally true of the topic (e.g., *"A billboard is an advertising wart"*). However the inclusion of such adjectives also delayed comprehension times by about a second, suggesting that a more extended process of dynamic category formation may have contributed to a more satisfying ultimate alignment.

Finally, we have also found that subjective ratings of aptness are correlated with corpus frequency (i.e., arguably the most direct and objective measure of conventionality), which suggests that part of the reason for the high correlation between ratings of aptness and conventionality may be that raters mistakenly attribute a metaphor's familiarity to its aptness, rather than (or possibly in addition to) the other way around.

How do these findings help us answer the question—what makes one metaphor easier to understand than another? We speculate that in order for a metaphor to be used conversationally it must be "good." If the vehicle does not have features that apply to the topic (i.e., it is not apt), then the metaphor will be uninterpretable, and is unlikely to be good. If it has some features that apply, but other salient features that do not, the metaphor may be confusing and lead to failures in communication. This would tend to make the metaphor seem less good. Therefore, we propose that aptness is normally a necessary but not a sufficient condition for a metaphor to gain traction in natural language. Specifically, there may be some quality threshold that metaphoric sentences must meet in order to become meaningful and prevalent in natural discourse. Arguments from "aptness" however, risk being tautological rather than explanatory. If aptness ratings are really measuring perceived metaphor quality, then aptness ratings are not explanatory of that quality, but only indicative of it.

Among metaphors that meet this aptness threshold, it seems to us that conventionalization certainly plays a role in determining the speed and ease of metaphor processing. And data from experiments that employ the "in vitro conventionalization" paradigm (e.g., Experiment 2 of this paper as well as Bowdle & Gentner, 2005) bear this out. Thus, our findings may well support the career of metaphor hypothesis – the notion that in comprehending novel metaphors, an interpretation emerges from comparison of the topic and vehicle. However, over time, as metaphoric mappings become familiarized, we rely less on the mapping process and more on established figurative meanings of metaphor vehicles. That the power of metaphoric communication resides in their generativity rather than merely their aptness is supported by evidence that metaphoric mappings of even conventional metaphors are alive (Thibodeau & Durgin, 2008).

## CONCLUSIONS

A summary of our chief theoretical results is shown in Table 1. In our experiments we have shown that conventionalizing (priming) an appropriately-used metaphor increases its speed of comprehension whereas priming an alternative sense of the metaphor decreases its rated aptness. More generally we have shown that ratings of aptness are sensitive to the presence of salient

TABLE 1
Central Theoretical Observations and Considerations

| |
|---|
| Conventionality Measures |
| Cannot be reliably defined for vehicles independent of topics. |
| Could be defined by more objective corpus analysis of sentences. |
| Aptness Ratings |
| Are correlated with corpus frequency (i.e., conventionality). |
| Are strongly influenced by salient features of metaphor vehicles that fail to map onto the specific topic (contrary to standard definitions of aptness). |
| Are influenced by manipulations of processing fluency (i.e., priming), and thus may not be explanatory of processing fluency (comprehension time). |
| Conventionalization and Aptness Manipulations |
| Conventionalization may normally occur only for metaphors that are (categorically) apt. |

features that do not apply to the topic. Ratings of aptness thus do not capture the theoretical explanatory construct (aptness) they are meant to measure. For these reasons rated aptness may often be a measure of processing fluency rather than a predictor of it. Thus recent claims that aptness-ratings explain metaphor processing fluency require re-evaluation.

In addition, we have argued that the conventionality of a metaphor must be considered in light of the meaning that the metaphor vehicle instantiates: Seemingly conventional metaphors can be rendered hard to interpret by ill-use. Aptness ratings can provide a heuristic evaluation of metaphor-sentence quality, but insofar as aptness ratings are actually measures of perceived processing fluency, they are irrelevant to the evaluation of "aptness" as an explanatory construct. In some cases they may even turn out to be measuring the conventionality of a specific metaphoric sense (e.g., the weak sense in which a robin might turn out to be describable as an alarm clock).

## REFERENCES


Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Blank, G. D. (1988). Metaphors in the lexicon. *Metaphor and Symbolic Activity*, *3*, 21–36.

Blasko, D.G., & Connine, C.M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 295–308.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*, 193–216.

Chiappe, D. L., & Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, *6*, 668–676.

Chiappe, D. L., Kennedy, J. M., & Chiappe, P. (2003). Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics*, *31*, 51–68.

Chiappe, D. L., Kennedy, J. M., & Smykowski, T. (2003). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, *18*, 85–105.

Clement, C., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, *15*, 89–132.

Gentner, D. (1982). Are scientific analogies metaphors? In D. Miall (Ed.), *Metaphor: Problems and perspectives* (pp. 106–132). Brighton, UK: Harvester Press.

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

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 199–253). Cambridge, MA: MIT Press.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language*, *37*, 331–355.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, *8*, 183–206.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. New York, NY: Oxford University Press.

Glucksberg, S. (2008). How metaphor creates categories – quickly! In R. Gibbs (Ed.), *Metaphor and thought* (pp. 67–83). Cambridge, UK: Cambridge University Press.

Glucksberg, S., & McGlone, M. S. (1999). When love is not a journey: What metaphors mean. *Journal of Pragmatics*, *31*, 1541–1558.

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language*, *36*, 50–67.

Glucksberg, S., & Haught, C. (2006a). Can Florida become like the next Florida? When metaphoric comparisons fail. *Psychological Science*, *17*, 935–938.

Glucksberg, S., & Haught, C. (2006b). On the relation between metaphor and simile: When comparison fails. *Mind and Language*, *21*, 360–378.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*, 3–18.

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

Honeck, R. P., Kibler, C. T., & Firment, M. J. (1987). Figurative language and psychological views of categorization: Two ships in the night? In R. E. Haskell (Ed.), *Cognition and symbolic structures: The psychology of metaphoric transformation* (pp. 103–120). Norwood, NJ: Ablex.

Indurkhya, B. (1987). Approximate semantic transference: A computational theory of metaphor and analogy. *Cognitive Science*, *11*, 445–480.

Jacoby, L. L., & Whitehouse, L. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*, 126–135.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 484–495.

Jones, L. L., & Estes, Z. (2005). Metaphor comprehension as attributive categorization. *Journal of Memory and Language*, *53*, 110–124.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *36*, 50–67.

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol*, *3*, 191–214.

Kelley, C. M., & Jacoby, L. L. (2000). Recollection and familiarity: Process-dissociation. In E. E. Tulving, E. Fergus, & I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 215–228). New York, NY: Oxford University Press.

Kennedy, J. M. (1990). Metaphor–Its intellectual basis. *Metaphor and Symbolic Activity*, *5*, 115–123.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*, 333–347.

Kittay, E. F., & Lehrer, A. (1981). Semantic fields and the structure of metaphor. *Studies in Language*, *5*, 31–63.

McGlone, M. S. (1996). Conceptual metaphors and figurative language interpretation: Food for thought? *Journal of Memory and Language*, *35*, 544–565.

McGlone, M. S., & Manfredi, D. (2001). Topic–vehicle interaction in metaphor comprehension. *Memory & Cognition*, *29*, 1209–1219.

Murphy, G. L. (1996). On metaphoric representation. *Cognition*, *60*, 173–204.

Thibodeau, P., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language*, *58*, 521–540.

Verbrugge, R. R., & McCarrell, N. S. (1977). Metaphoric comprehension: Studies in reminding and resembling. *Cognitive Psychology*, *9*, 494–533.

## APPENDIX: STIMULUS LIST

TABLE A1
Same-sense Stimulus List

| | *Target* | *Same-sense1* | *Same-sense2* |
|---|---|---|---|
| 1 | A business is a living organism. | A country fair is a living organism. | A super computer is a living organism. |
| 2 | A fisherman is a spider. | A sharpshooter is a spider. | An assassin is a spider. |
| 3 | Beavers are lumberjacks. | A termite is a lumberjack. | A woodpecker is a lumberjack. |
| 4 | Some bladders are barrels. | Some mugs are barrels. | Some stomachs are barrels. |
| 5 | Insults are razors. | Rumors are razors. | Prejudice is a razor. |
| 6 | Cocaine is a time bomb. | Speeding is a time bomb. | Cheating is a time bomb. |
| 7 | Education is a lantern. | A documentary is a lantern. | A mentor is a lantern. |
| 8 | Her ex-husband is a gem. | A great job is a gem. | A sibling is a gem. |
| 9 | My computer skills course is a joke. | My cooking ability is a joke. | My golf stroke is a joke. |
| 10 | His college class is a zoo. | The airport is a zoo. | The mall is a zoo. |
| 11 | The driveway is an ice rink. | The marble floor is an ice rink. | The dance floor is an ice rink. |
| 12 | Having summers off was a bear. | Flying back from China was a bear. | Doing homework is a bear. |
| 13 | A lie is a dagger. | Cold water is a dagger. | Breaking up is a dagger. |
| 14 | The good news was an earthquake. | Victory was an earthquake. | Graduation was an earthquake. |
| 15 | A zoo is a museum. | A library is a museum. | A botanical garden is a museum. |
| 16 | That criminal's pathway is a portrait. | That math theorem is a portrait. | A graph is a portrait. |
| 17 | Some teachers are encyclopedias. | Some game show contestants are encyclopedias. | Some grandparents are encyclopedias. |
| 18 | My boyfriend is a peach. | My uncle is a peach. | My kitten is a peach. |
| 19 | Jalapeno peppers are fire. | Taco sauce is fire. | Curry is fire. |
| 20 | Hostility is a veil. | Love is a veil. | An envelope is a veil. |
| 21 | Music can be medicine. | Poetry can be medicine. | Exercise can be medicine. |
| 22 | That football player is a rail. | That power lifter is a rail. | That boxer is a rail. |
| 23 | Alcohol is a crutch. | Drugs are a crutch. | Plagiarism is a crutch. |
| 24 | Control is fertilizer | Love is fertilizer. | Discipline is fertilizer. |
| 25 | A tree is an umbrella. | A roof is an umbrella. | A helmet is an umbrella. |
| 26 | Some snores are sirens. | Some whistles are sirens. | Some applause is a siren. |
| 27 | That receptionist is a breath of fresh air. | That candidate is a breath of fresh air. | The glass of water is a breath of fresh air. |
| 28 | An opponent is an anchor. | A friend is an anchor. | A goal is an anchor. |
| 29 | His marriage was a short leash. | Daily chores are a short leash. | A two year old is a short leash. |
| 30 | The basketball player was thunder. | That racehorse was thunder. | That punch was thunder. |
| 31 | The Great Plains are a board. | The Sahara Desert is a board. | Calm seas are a board. |
| 32 | Some tears are magnets. | Some lights are magnets. | Some screams are magnets. |
| 33 | My young cousin is a shrimp. | My Chihuahua is a shrimp. | My hamster is a shrimp. |
| 34 | Grandparents can be donkeys. | Bureaucrats can be donkeys. | Lobbyists can be donkeys. |
| 35 | Some fashion models are twigs. | Some greyhounds are twigs. | Some marathoners are twigs. |
| 36 | That professor is a duck. | That clown is a duck. | That comedian is a duck. |
| 37 | Some dogs are princesses. | Some cheerleaders are princesses. | Some professional tennis players are princesses. |

(*Continued*)

TABLE A1
(Continued)

| | *Target* | *Same-sense1* | *Same-sense2* |
|---|---|---|---|
| 38 | My rat's fur is silk. | A baby's bottom is silk. | A bird's feathers are silk. |
| 39 | Books are treasure chests. | A good doctor is a treasure chest. | A big brother is a treasure chest. |
| 40 | Many teams are jails. | Some families are jails. | Some high schools are jails. |
| 41 | Time is money. | Work is money. | Sleep is money. |
| 42 | The nearest star is a ball. | An orange is a ball. | A globe is a ball. |
| 43 | Ideas can be diamonds. | Paintings can be diamonds. | Houses can be diamonds. |
| 44 | Sadness is a volcano. | Anger is a volcano. | Joy is a volcano. |
| 45 | The senator is a fossil. | A barber is a fossil. | A judge is a fossil. |
| 46 | My grandfather's legs are steel. | A rugby player's neck is steel. | A bouncer's arm is steel. |
| 47 | That bedroom is a dump. | That dorm room is a dump. | That minivan is a dump. |
| 48 | Intelligence is a warehouse. | An external hard drive is a warehouse. | A garage is a warehouse. |

TABLE A2
Literal-sense and Alternative-sense Stimulus List

| | *Literal-sense1* | *Literal-sense2* | *Alternative-sense1* | *Alternative-sense2* |
|---|---|---|---|---|
| 1 | A mouse is a living organism. | A tree is a living organism. | Friendship is a living organism. | Romance is a living organism. |
| 2 | A black widow is a spider. | A tarantula is a spider. | A basket weaver is a spider. | A seamstress is a spider. |
| 3 | Paul Bunyan is a lumberjack. | A logger is a lumberjack. | A soldier is a lumberjack. | A gladiator is a lumberjack. |
| 4 | Some casks are barrels. | Some wooden containers are barrels. | Santa Claus is a barrel. | Some sumo wrestlers are barrels. |
| 5 | Shavers are razors. | Scalpels are razors. | Minds are razors. | Memories are razors. |
| 6 | Land mines are time bombs. | Grenades are time bombs. | Blackmail is a time bomb. | Propaganda is a time bomb. |
| 7 | A camp light is a lantern. | A torch is a lantern. | A flag is a lantern. | A uniform is a lantern. |
| 8 | A sapphire is a gem. | A ruby is a gem. | A lake surface is a gem. | A full moon is a gem. |
| 9 | A knock-knock riddle is a joke. | A prank is a joke. | Writing on his face while he is asleep is a joke. | A Whoopi cushion is a joke. |
| 10 | The place to see lions is a zoo. | A great field trip destination is the zoo. | An art gallery is a zoo. | A jewelry show is a zoo. |
| 11 | A hockey arena is an ice rink. | A frozen pond is an ice rink. | The north pole is an ice rink. | A walk in freezer is an ice rink. |
| 12 | The grizzly animal was a bear. | The large predator was a bear. | This case of books is a bear. | This solid oak desk is a bear. |
| 13 | A short knife is a dagger. | A small bladed weapon is a dagger. | A pistol is a dagger. | A fist is a dagger. |
| 14 | Shaking the house was an earthquake. | The shifting ground was an earthquake. | My parents' divorce was an earthquake. | The stock market crash was an earthquake. |
| 15 | The Smithsonian Institute is a museum. | An art gallery is a museum. | My grandmother's jewelry box is a museum. | A time capsule is a museum. |
| 16 | A picture of one's self is a portrait. | A president's picture is a portrait. | A skyline is a portrait. | A sunset is a portrait. |
| 17 | Wikipedia is an encyclopedia. | This journal set is an encyclopedia. | Some phone books are encyclopedias. | Some textbooks are encyclopedias. |
| 18 | This fruit is a peach. | This candy flavor is peach. | His cheek is a peach. | This blanket is a peach. |
| 19 | The blue flame is fire. | The light in the distance is fire. | Coffee is fire. | Fresh pizza is fire. |
| 20 | A black cloth is a veil. | A mask is a veil. | A tissue is a veil. | An undershirt is a veil. |
| 21 | Tylenol is medicine. | Marijuana can be medicine. | Cherry candy can be medicine. | Chalky milkshakes can be medicine. |
| 22 | Trains travel on a rail. | This strip of metal is a rail. | That swimsuit model is a rail. | That runner is a rail. |
| 23 | A cane is a crutch. | A wooden brace is a crutch. | A column is a crutch. | A wide bookshelf is a crutch. |
| 24 | Manure is fertilizer. | Nitrogen is fertilizer. | Broccoli is fertilizer. | Sleep is fertilizer. |
| 25 | A waterproof tarp is an umbrella. | A parasol is an umbrella. | Plastic is an umbrella. | Gore-tex is an umbrella. |
| 26 | Some horns are sirens. | Some alarms are sirens. | The announcement is a siren. | The news release is a siren. |
| 27 | A deep sigh is a breath of fresh air. | A yawn is a breath of fresh air. | Graduation is a breath of fresh air. | The election is a breath of fresh air. |
| 28 | A large iron weight is an anchor. | A sinker is an anchor. | A broken leg is an anchor. | Debt is an anchor. |

(*Continued*)

TABLE A2
(Continued)

| | *Literal-sense1* | *Literal-sense2* | *Alternative-sense1* | *Alternative-sense2* |
|---|---|---|---|---|
| 29 | A dog collar is a short leash. | A restraining rope is a short leash. | A ponytail is a short leash. | A sweatshirt hood is a short leash. |
| 30 | That loud noise was thunder. | The sound of lightening is thunder. | A crowd's cheer is thunder. | A lion's roar is thunder. |
| 31 | A wooden plank is a board. | A table is a board. | A dead body is a board. | Arthritic joints are a board. |
| 32 | Some refrigerator decorations are magnets. | This iron strip is a magnet. | Some tape is a magnet. | Some nails are magnets. |
| 33 | A small sea creature is a shrimp. | A small shellfish is a shrimp. | A swimmer is a shrimp. | A diver is a shrimp. |
| 34 | One farm animal is a donkey. | One pack animal is a donkey. | Cashiers can be donkeys. | Butchers can be donkeys. |
| 35 | A small tree branch is a twig. | A tiny piece of wood is a twig. | Some stale biscuits are twigs. | Some eggshells are twigs. |
| 36 | That quacking animal is a duck. | That mallard is a duck. | Floating trash is a duck. | Tugboats are ducks. |
| 37 | The daughter of the queen is a princess. | Some fairy tale heroines are princesses. | Some ice skaters are princesses. | Some ballet dancers are princesses. |
| 38 | A soft delicate fabric is silk. | A Chinese fabric is silk. | A computer screen is silk. | A balk head is silk. |
| 39 | Pirate booty is a treasure chest. | A gold coin container is a treasure chest. | Safe deposit box | Vaults are treasure chests. |
| 40 | A detention center is a jail. | A penitentiary is a jail. | A $2000 fine is jail. | A spanking is jail. |
| 41 | A dime is money. | A euro is money. | A class president is money. | Tom Brady is money. |
| 42 | A sphere is a ball. | This dog toy is a ball. | This party is a ball. | Playing cards is a ball. |
| 43 | Engagement rings are diamonds. | Expensive jewels are diamonds. | Shin guards are diamonds. | A cast is a diamond. |
| 44 | Mt. Vesuvius is a volcano. | Pompeii is a volcano. | A chimney is a volcano. | The Marlboro man is a volcano. |
| 45 | Petrified wood is a fossil. | A dinosaur bone is a fossil. | A stale Twinkie is a fossil. | Beef jerky is a fossil. |
| 46 | Silverware is steel. | A girder is steel. | Some facial expressions are steel. | A meditating monk is steel. |
| 47 | A landfill is a dump. | A junkyard is a dump. | A toilet is a dump. | A trashcan is a dump. |
| 48 | A post office is a warehouse. | A factory is a warehouse. | That gym is a warehouse. | That office building is a warehouse. |