What Do We Learn from Rating Metaphors?

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
Department of Psychology, Oberlin College
120 W. Lorain St; Oberlin, OH 44074 USA

Les Sikos (lsikos1@swarthmore.edu)
Frank H. Durgin (fdurgin1@swarthmore.edu)
Department of Psychology, Swarthmore College
500 College St; Swarthmore, PA 19081 USA

Abstract

What makes some metaphors easier to understand than others? Theoretical accounts of metaphor processing appeal to dimensions like *conventionality* and *aptness* to explain variability in metaphor comprehensibility. In a typical experiment, one group of naive participants rates a set of metaphoric sentences along these dimensions, while another is timed reading the same sentences. Then, the ratings are used to predict response times in order to identify the most relevant linguistic dimension for metaphor comprehension. However, surprisingly high correlations between ratings of theoretically orthogonal constructs and the results of an experiment in which a context manipulation affected ratings of metaphor *conventionality* and *aptness* suggest that these measures should be treated as dependent, rather than explanatory, variables. We discuss the implications of this perspective for theories of language processing.

Keywords: Metaphor, analogy, measurement, conventionality, language

Introduction

What makes some metaphors easier to understand than others? Theoretical accounts of metaphor processing appeal to dimensions like *conventionality* (e.g., Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997) and *aptness* (e.g., Glucksberg & Haught, 2006; Jones & Estes, 2006; Chiappe, Kennedy, & Chiappe, 2003) to explain variability in metaphor comprehensibility. In this context, *conventionality* reflects the familiarity of a metaphor; *aptness* reflects the degree to which a metaphor vehicle captures important features of the topic.

In a typical experiment, one group of naive participants rates a set of metaphoric sentences along these dimensions, while another is timed reading the same sentences. Then, the ratings are used to predict response times in order to identify the most relevant linguistic dimension for metaphor comprehension (e.g., Blank, 1988; Chiappe et al., 2003; Chiappe & Kennedy, 1999; Giora, 1997; Glucksberg, McGlone, & Manfredi, 1997; Jones & Estes, 2005, 2006).

However, surprisingly high correlations between ratings of theoretically orthogonal constructs raise questions about how well naive participants can actually operationalize these linguistic dimensions (Jones & Estes, 2006; Thibodeau & Durgin, 2011). For instance, Table 1 shows correlations between ratings of *familiarity*, *naturalness*, *imageability*, and *figurativeness* from data collected by Cardillo, Schmidt, Kranjec, and Chatterjee (2010), which are all significant at the $p < .001$ level, even though theoretical accounts of metaphor processing have argued that (at least some of) these constructs are orthogonal. A principal components analysis reveals that 70.3% of the variance across the four dimensions is captured by a single variable that loads positively on *familiarity*, *naturalness*, and *imageability*, and negatively on *figurativeness*, *p*s < .001. Similar patterns are found in existing datasets that seek to norm metaphoric stimuli for experimental work (e.g., Campbell & Raney, 2015; Cardillo et al., 2010; Katz, Paivio, Marschark, & Clark, 1988; Roncero & de Almeida, 2014).

Table 1: Correlations between ratings of familiarity, naturalness, imageability, and figurativeness from data collected by (Cardillo et al., 2010).

	Natural	Image	Figurative	PC
Familiarity	.936	.509	-.598	.931
Naturalness		.571	-.579	.942
Imageability			-.361	.706
Figurativeness				-.748

The goal of the present paper is to highlight this issue (i.e., what are we actually measuring when we ask people to rate sentences for *conventionality*, *aptness*, *metaphoricity*, etc?) by showing that ratings of these dimensions change as function of the context in which metaphoric sentences are presented. In the experiment, metaphoric target sentences were situated within one of three contexts: *matched*, *mixed*, or *literal* (see Table 2). In every case, participants read a brief description of a scenario, followed by a sentence that included a target metaphor. In some cases (the *matched* and *mixed* conditions), the initial description (i.e., *Context*) included metaphoric language, which was either consistent with the target metaphor (i.e., in the *matched* condition, the target metaphor extended a previously instantiated conventional metaphor) or inconsistent with the target metaphor (i.e., in the *mixed* condition, the target metaphor came from a different metaphor family than the metaphor used in the initial description of the scenario). In the *literal* condition, the target metaphor followed a non-metaphoric description of the scenario.

Table 2: Example base for generating stimulus items. Context A, B or C was presented with Extension A or B. Target words are indicated by capitalization.

Context

A. In big cities across America, crime has become an epidemic that can't be cured. It is beginning to infect small towns as well.

B. In big cities across America, crime has become a beast that is roaring out of control. It is beginning to prey on small towns as well.

C. In big cities across America, crime has become a problem that can't be solved. It is beginning to affect small towns as well.

Extension

A. There is no ANTIDOTE strong enough to cure it.

B. There is no CAGE strong enough to restrain it.

Experiment

Methods

Participants We recruited 1,200 people to participate in an online questionnaire through Amazon's Mechanical Turk. We restricted our sample to participants living in the United States who had a good performance record (a minimum 90% approval rating on previous tasks). We used Turk's exclusion capabilities (and internal tracking of IP addresses) to prevent people from participating in the task more than once. Upon completing the survey, participants were given a nine digit random number to paste back into the Mechanical Turk interface. Seven participants pasted an incorrect or incomplete completion code into Turk, leaving data from 1193 participants for analysis.

Stimuli The stimuli were modeled after those developed by Thibodeau and Durgin (2008, Exp. 3). One paired-item stimulus base is shown in Table 2. For each stimulus base, two different metaphoric mappings (e.g., A: CRIME IS AN INFECTIOUS DISEASE VS. B: CRIME IS A WILD ANIMAL) were used to generate conceptually parallel initial vignettes that contained two or more context sentences. A third conceptually parallel vignette (C: CRIME IS A PROBLEM) used non-metaphoric language to communicate the same basic situation.

Target sentences were designed to be interpretable following a description that included *matched* or *mixed* metaphors or non-metaphoric, *literal*, language. That is, the study was not designed to to compare across situations that involved arriving at entirely different semantic interpretations of the target sentence. However, we did expect the manipulation to modulate the processing fluency of the target metaphors. Specifically, we expected that the metaphors would be be processed more easily in the *matched* than *mixed* or *literal* contexts (Thibodeau & Durgin, 2008).

A total of 36 paired-item stimulus bases were developed. Three experimental conditions of two items each (a total of 72 items in each condition) were constructed from each stimulus base by pairing contexts and target sentences: *Matched Context* (Context A + Extension A; Context B + Extension B), *Mixed Context* (Context A + Extension B; Context B + Extension A), and *Literal Context* (Context C+ Extension A; Context C + Extension B).

All target metaphors were first instated nominally with a word that was sentence-medial (e.g., "There is no ANTIDOTE/CURE strong enough..."), and words leading up to the critical figurative noun were identical across sentences. These constraints were imposed in order to make the stimuli well-suited for investigation with methodologies such as ERP, since integration processes compete with metaphor comprehension processes when metaphor vehicles are presented at the ends of sentences (Friedman, Simson, Ritter, & Rapin, 1975; Osterhout, 1997). Note that each target word served as its own control because it appeared in all three contexts.

Survery Materials and Procedure Each participant read one item (i.e., context–target pairing) from each of the 36 stimulus bases described above, such that all participants rated 12 *matched*, 12 *mixed*, and 12 *literal* items. The order of items was pseudo-randomized such that no participant saw more than two items from a given condition in a row. Participants were asked to attend to the (uppercase) target word — the word which instantiated the metaphoric extension — when making their ratings.

In an effort to collect both contrastive and holistic ratings, we created two versions of the questionnaire. One asked participants to rate target words along four dimensions: *surprisingness*, *comprehensibility*, *conventionality*, and *metaphoricity*. The other asked participants to rate target words along the single dimension of *aptness* (see definitions of dimensions in Table 3). Following the instructions, participants saw an example question and had an opportunity to re-read the definitions of the dimensions. They then rated two filler items — a *matched* pairing and a *literal* pairing — before going on to rate the 36 experimental items.

Each item was presented, individually, above one or four 7-point Likert scales. Participants were prompted for ratings on each of the dimensions with a question. For example: "Surprisingness: how surprising was the metaphoric WORD as it is currently used?" The scale ranged from "very low" to "very high" for each dimension.

Analysis Principal components analysis (PCA) is a statistical procedure for revealing the internal structure of a dataset containing possibly correlated variables in a way that best ex-

Table 3: Definitions of dimensions as presented to participants in the instructions of the experiment.

Dimension	Definition
Surprisingness	Sometimes metaphors are used very naturally and do not seem surprising; in other cases, it can feel like they come out of nowhere. This is the dimension of 'surprisingness' that we'd like you to rate.
Comprehensibility	By this we simply mean, how easy or difficult it is to understand the statement. Some expressions are easier to understand, like "My mother is a saint," while others might be harder to understand "My mother is a paperclip."
Conventionality	Expressions can also vary in how conventional they are for conveying the idea that they are supposed to communicate. For example, consider the following two descriptions of a person running fast: a conventional or common way, 'he was running like the wind,' and a much less conventional way, 'he was running like a Porsche on a German highway.'
Metaphoricity	Finally, even metaphoric expressions may vary in how much they are figurative expressions rather than literal. For example, if I say that the sun is a ball, someone might interpret that as being literally true if they thought that 'ball' meant any object that was spherically shaped. Another person might think the expression quite metaphorical because balls are toys, and the sun is not a toy. You will be asked to judge the metaphoricity of the expressions below by indicating whether you find them to be very metaphorical or close to literal.
Aptness	Metaphors can vary in the extent to which they capture important features of the topic being described. This is the dimension of 'aptness.' For instance, it would be apt to express that someone is a fast runner by saying they are a rocket, but less apt to express that someone is a fast runner by saying they are an astronaut.

plains the common variance of the data (Dunteman, 1989). A primary goal of the present analysis was to identify and characterize the principal components (PCs) of the ratings, while checking for correlations among the various dimensions.

We test for effects of the context manipulation by fitting separate, by-item, repeated measures ANVOAs for each PC (separate models since the principal components analysis yields orthogonal components).

Results and Discussion

We found that extracting two PCs explained 92.8% of the variance across the five dimensions. The first component explained 74.1% of the variance and loaded highly on four of the dimensions: *comprehensibility*, *conventionality*, *aptness*, and negatively on *surprisingness*; this component loaded less heavily on *metaphoricity* (see Table 4). As expected, these dimensions seem to converge on the notion of processing fluency (ease) with respect to metaphor interpretation. As a result we call this component *Processing Fluency*.

The second component explained 18.6% of the variance across the five dimensions (71.8% of the variance that remained after accounting for the first component), and loaded most strongly on *metaphoricity*. As a result we call this second PC *Figurativeness*. Correlations between the individual rated dimensions and PCs are presented in Table 4. Note that ratings of *aptness* (collected separately) and *comprehensibility* (collected together with other dimensions) correlated exclusively with *Processing Fluency*, whereas *conventionality* and *surprisingness* ratings contributed to *Figurativeness* as well.

Table 4: Correlations among rated dimensions and the first two (orthogonal) principal components of the ratings data. All correlations significant at the $p < .001$ level except those noted: $^{*}p < .05$ and $^{\diamond}p > .05$.

.

	Surp	Meta	Conv	Compr	Apt
PC1: Ease	-.919	.436	.933	.969	.928
PC2: Fig	.311	.893	$-.187^{*}$	$.021^{\diamond}$	$.055^{\diamond}$
Aptness	-.821	.422	.766	.883	
Comprehens	-.848	.428	.897		
Conventional	-.902	.260			
Meta'icity	$-.142^{*}$				

Figure 1 shows mean *Processing Fluency* and *Figurativeness* as a function of *Context* (*matched*, *mixed*, or *literal*). An ANOVA by items indicated that *Processing Fluency* was reliably affected by *Context*, $F(2,142) = 133.759$, $p < .001$, $\eta^2 = 0.653$. Paired comparisons further revealed that *Processing Fluency* was higher in the *matched* context than in either the *literal*, $t(71) = 13.519$, $p < .001$, or the *mixed* context, $t(71) = 12.941$, $p < .001$, and that *Processing Fluency* did not differ between the *literal* and *mixed* contexts, $t(71) = 0.995$, $p = .323$.

These findings suggest that a supportive discourse context increased participants' metacognitive sense of processing fluency of the required metaphoric mapping. In principle, such benefits could be argued to emerge from lexical priming alone (e.g., McGlone, 2011). However, Thibodeau and Durgin (2011) found that lexical priming of the metaphor vehicle was insufficient to produce shifts in ratings of *aptness*, whereas priming the specific metaphor mapping was both necessary and sufficient. For instance, people read a target sentence like "*Education is a lantern*" faster after they had read sentences like "*A mentor is a lantern*" (i.e., when the metaphor vehicle conveyed a similar meaning) but not after they had read sentences like "*A camp light is a lantern*" (i.e., when the metaphor vehicle was used non-metaphorically) or after they had read sentences like "*A flag is a lantern*" (i.e., when the metaphor vehicle conveyed a different meaning). Therefore, a supportive context appears to prime the relationship between the relevant figurative and literal semantic domains, rather than merely the lexical item *per se*. Because the ratings concern metaphor quality (rather than response latency), they show that interpretation of the metaphor (the mapping) is indeed facilitated by the context manipulation.

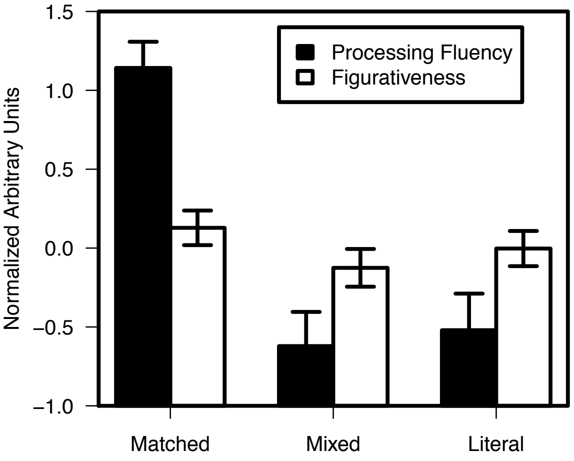

Figure 1: The mean values (across 72 items) of the first and second principal components of the rating data as a function of Context. Standard error bars are shown.

The *Figurativeness* principal component was also reliably affected by *Context*, $F(2,142) = 7.536$, $p < .001$, $\eta^2 = 0.096$. Paired comparisons indicated that *Figurativeness* was judged higher in the *matched* context than in the *mixed* context, $t(71) = 3.838$, $p < .001$. There was some evidence that *Figurativeness* estimates in the *literal* context were lower than those in the *matched* context, $t(71) = 2.099$, $p = .039$, and higher than those in the *mixed* context, $t(71) = 1.818$, $p = .073$.

However, as is evident in Figure 1, the effects of *Context* on *Figurativeness* (effect size for *matched* vs. *mixed* = 0.45) were small compared to the effects of the *Context* on enhancing *Processing Fluency* (effect size for *matched* vs. *mixed* = 1.52). Thus the *Figurativeness* component of participants' ratings may reflect a comparison between the implied meaning of the target word and its literal sense (above and beyond the effect of *Processing Fluency*, *per se*). This comparison may nonetheless be facilitated if the intended metaphoric sense of the target word is made clearer by a *matched* context.

It is important to distinguish between the measure of *Figurativeness* yielded from the principal components analysis and direct ratings of *metaphoricity*. Participants' ratings of *metaphoricity* were affected more strongly by the context manipulation, $F(2,142) = 32.483$, $p < .001$, $\eta^2 = 0.314$, compared to the principal component ($\eta^2 = 0.096$).[1] This is because ratings of *metaphoricity* were correlated with ratings of *aptness*, *surprisingness*, *comprehensibility*, and *conventionality*, and thus loaded onto *Processing Fluency*, $r(71) = .436$, $p < .001$.

In the measure of *Figurativeness* computed from the principal components analysis, on the other hand, variance in ratings of *metaphoricity* that was consistent with the other four dimensions was partialed into the measure of *Processing Fluency*, leaving the *Figurativeness* principal component to account for variance that was mostly unique to an aspect of rated *metaphoricity* (and to a lesser extent rated *surprisingness* and *conventionality*).

General Discussion

In this experiment we showed that asking people to rate metaphors for five distinct qualities produced two orthogonal principal components. The first, which we labeled *Processing Fluency*, seems to represent the ease with which a metaphor can be interpreted (another appropriate label for this dimension might be: *Ease of Interpretation*). The second, which we called *Figurativeness*, seems to represent the extent to which a word's meaning is perceived as clearly figurative rather than literal, once *Processing Fluency* is taken into account. Manipulating the context for a target metaphor had a large impact on *Processing Fluency*, but only a slight impact on perceived *Figurativeness*. That is, extended metaphors were judged by raters to be similarly metaphoric (or not) regardless of context. However, these metaphors were processed more fluently when they were presented in the context of a consistent metaphoric mapping.

One important implication of these findings relates to the explanatory power of off-line ratings of metaphor processing for theories of metaphor comprehension. One goal of language researchers has been to identify a linguistic dimension that explains why some metaphors are easier to process than others. *Aptness* has been highlighted by researchers who argue that metaphors are processed as class inclusions statements (e.g., Glucksberg & Haught, 2006; Jones & Estes, 2006; Chiappe et al., 2003), rather than through a comparison mechanism (e.g., Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997). Although it very well may be the case that such a dimension can be measured and that it can explain, *a priori*, variability in metaphor processing fluency, our results suggest that operationalizing the construct by gathering ratings from

[1] Although pairwise comparisons reveal a similar pattern of results for the direct ratings and the principal component: significant differences between the *matched* and both *mixed*, $t(71) = 7.712$, $p < .001$, and *literal* contexts, $t(71) = 5.961$, $p < .001$; a trending but non-significant difference between the *mixed* and *literal* contexts, $t(71) = 1.711$, $p = .091$.

naive participants is fundamentally flawed.

In contrast to how these ratings are typically used (i.e., as operationalizations of independent, predictor, variables), our work suggests that these dimensions may be more appropriately considered as an indirect measure of processing fluency (Thibodeau & Durgin, 2011). When people are asked to rate sentences for abstract qualities like *conventionality* and *aptness*, they misattribute how easily they processed the sentence for the dimension they are being asked to rate (Alter & Oppenheimer, 2009; Jacoby & Whitehouse, 1989; Jacoby, Allan, Collins, & Larwill, 1988; Kahneman, 2011).

This is problematic because processing fluency is supposedly what we are trying to explain by gathering ratings of the linguistic dimensions in the first place. In other words, using subjective ratings of *conventionality* and *aptness* to predict how quickly people process metaphoric sentences entails showing that an "off-line" measure of processing fluency is related to an "on-line" measure of processing fluency. As a result, it is unclear what we can learn about the mechanisms that support language processing from such experiments.

To address this issue, we suggest designing experiments that actively manipulate the relevant linguistic dimensions or operationalizing the constructs with more objective methods like corpus analysis. For instance, one way to manipulate the familiarity of a metaphor "on-line" is to repeatedly expose people to similar metaphoric expressions (e.g., "A figure skater is a butterfly"; "A ballerina is a butterfly"; Bowdle & Gentner, 2005; Thibodeau & Durgin, 2011; Chettih, Durgin, & Grodner, 2012). One way to measure a construct like *aptness* "off-line" is to use corpus-based metrics like Latent Semantic Analysis (Kintsch, 2000; Kintsch & Bowles, 2002).

We also see value in the use of statistical procedures like PCA for researchers interested in testing questions about how metaphors are processed. On this approach, ratings of dimensions like *conventionality* and *aptness* can be viewed as converging on a notion of *Processing Fluency*. Just as personality instruments and attitudinal surveys often include complementary items that facilitate reliable measurement of personality and attitudinal constructs, asking people to rate various linguistic dimensions of metaphors can elicit more reliable estimates of *Processing Fluency*. The results of the current experiment (as well as analyses of existing data sets like those of Cardillo et al., 2010) suggest that people can reliably differentiate their sense of a sentence's metaphoricity (figurativeness) from a dimension like *Processing Fluency* – particularly when a procedure like PCA is used to make these dimensions orthogonal (i.e., to partial variance in ratings of metaphoricity that seem to be affected by *Processing Fluency* into a measure of *Processing Fluency*). On this approach, one can more confidently use subjective ratings in response time and imaging studies to test mechanistic questions about figurative language processing.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review*.

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

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*(1), 193.

Campbell, S. J., & Raney, G. E. (2015). A 25-year replication of katz et al.'s (1988) metaphor norms. *Behavior research methods*, 1–11.

Cardillo, E. R., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior research methods*, *42*(3), 651–664.

Chettih, S., Durgin, F. H., & Grodner, D. J. (2012). Mixing metaphors in the cerebral hemispheres: What happens when careers collide? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 295.

Chiappe, D. L., & Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, *6*(4), 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*(1), 51–68.

Dunteman, G. H. (1989). *Principal components analysis* (No. 69). Sage.

Friedman, D., Simson, R., Ritter, W., & Rapin, I. (1975). Cortical evoked potentials elicited by real speech words and human sounds. *Electroencephalography and clinical Neurophysiology*, *38*(1), 13–19.

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

Glucksberg, S., & Haught, C. (2006). Can florida become like the next florida? when metaphoric comparisons fail. *Psychological Science*, *17*(11), 935–938.

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of memory and language*, *36*(1), 50–67.

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*, *14*(2), 240.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*(2), 126.

Jones, L. L., & Estes, Z. (2005). Metaphor comprehension as attributive categorization. *Journal of Memory and Language*, *53*(1), 110–124.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

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*(4), 191–214.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*(2), 257–266.

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, *17*(4), 249–262.

McGlone, M. S. (2011). Hyperbole, homunculi, and hindsight bias: An alternative evaluation of conceptual metaphor theory. *Discourse Processes*, *48*(8), 563–574.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and language*, *59*(3), 494–522.

Roncero, C., & de Almeida, R. G. (2014). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior research methods*, 1–13.

Thibodeau, P. H., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language*, *58*(2), 521–540.

Thibodeau, P. H., & Durgin, F. H. (2011). Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, *26*(3), 206–226.