Behav Res
DOI 10.3758/s13428-017-0903-9

Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences

Paul H. Thibodeau[1] · **Les Sikos**[2,3] · **Frank H. Durgin**[2]

Abstract What makes some metaphors easier to understand than others? Several psycholinguistic dimensions have been identified as candidate answers to this question, including appeals to *familiarity* and *aptness*. One way to operationalize these dimensions is to collect ratings of them from naive participants. In this article, we question the construct validity of this approach. Do ratings of aptness actually reflect the aptness of the metaphors? Are ratings of aptness measuring something different from ratings of familiarity? With two experiments and an analysis of existing datasets, we argue that ratings of metaphoric sentences are confounded by how easily people are able to understand the sentences (*processing fluency*). In the experiments, a context manipulation was designed to affect how fluently people would process the metaphors. Experiment 1 confirmed that the manipulation affected how quickly people understood the sentences in a response time task. Experiment 2 revealed that the same manipulation influenced ratings of such dimensions as familiarity and aptness. Finally, factor analyses—on the ratings data from Experiment 2 and from several existing datasets—revealed two underlying sources of variance in sentence-level ratings of metaphors (the "big two" dimensions of metaphoric sentences): processing fluency and figurativeness. We discuss the implications of these findings for theories of figurative-language processing by emphasizing more careful treatment of subjective ratings of metaphoric sentences, and by suggesting the use of alternative methods to manipulate and measure such dimensions as familiarity and aptness.

Keywords Metaphor · Analogy · Measurement · Conventionality · Language · Processing fluency

What makes some metaphors easier to understand than others? For example, why are people faster to process the metaphor "Memory is a warehouse" than the metaphor "A fisherman is a spider?"

Several psycholinguistic dimensions have emerged as candidate answers to this question, which correspond to different psychological theories of how metaphors are processed. One possibility is that "Memory is a warehouse" is a more *familiar* or *conventional* expression,[1] and common metaphors may be easier to interpret than novel metaphors (Blank, 1988; Bowdle & Gentner, 2005; Gentner & Markman, 1997; Gentner & Wolff, 1997; Giora, 1997, 1999, 2007; Miller, 1979; Wolff & Gentner, 2011). To understand a novel metaphor like, "A fisherman is a spider," the topic (*fisherman*) and vehicle (*spider*) must be compared—in order to find properties of the vehicle (*spider*) that also describe the topic (*fishermen*; e.g., both are patient hunters). For more common metaphors, like "Memory

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0903-9) contains supplementary material, which is available to authorized users.

✉ Paul H. Thibodeau
paul.thibodeau@oberlin.edu

[1] Department of Psychology, Oberlin College, 120 W Lorain Street, Oberlin, OH 44074, USA

[2] Swarthmore College, Swarthmore, PA, USA

[3] Universität des Saarlandes, Saarbrücken, Germany

[1] A distinction is often made between two types of conventionality in the metaphor-processing literature. The term *conventionality* describes how commonly a metaphor vehicle expresses a specific figurative meaning (e.g., how often "warehouse" is used to mean "large storage space") and, thus, how quickly a metaphor vehicle can bring that figurative meaning to mind (Bowdle & Gentner, 2005). The term *familiarity* describes the commonness of a metaphorical expression like "Memory is a warehouse" (e.g., Katz, Paivio, Marschark, & Clark, 1988). In other words, *conventionality* refers to a word-level property, whereas *familiarity* refers to a sentence-level property.

Published online: 27 May 2017

is a warehouse," a figurative meaning of the expression may be stored in the mental lexicon, thereby speeding comprehension by eliminating the need to engage in a comparison process.

Another possibility is that "Memory is a warehouse" is more *apt*: That is, properties of the metaphor vehicle, *warehouse*, may describe the topic, *memory*, better than the properties of *spider* describe *fisherman* (Glucksberg & Keysar, 1990). According to this view, metaphors are processed as class inclusion statements—through categorization rather than comparison (e.g., Chiappe & Kennedy, 1999; Chiappe, Kennedy, & Chiappe, 2003; Glucksberg & Haught, 2006; Jones & Estes, 2005, 2006). People may interpret metaphors quickly when the vehicle (warehouse) is an *apt* characterization of the topic (memory), regardless of whether the metaphor is novel or familiar.

In addition, metaphorical sentences may differ in *metaphoricity*, or how figurative (nonliteral) they are; in *imagery*, or how clear a mental picture they produce; or a in a variety of other ways that could influence their comprehensibility.

How should these psycholinguistic dimensions of metaphorical sentences be operationalized? A common method is to collect subjective ratings from naive participants. For instance, several stimulus sets of metaphoric sentences have been developed with normed ratings of such dimensions as familiarity, aptness, metaphoricity, and imagery (e.g., Cardillo, Schmidt, Kranjec, & Chatterjee, 2010; Cardillo, Watson, & Chatterjee, 2016; Katz, Paivio, Marschark, & Clark, 1988; Roncero & de Almeida, 2015).

Because so much would seem to hinge on whether we are able to measure these constructs accurately, an examination of the ratings tasks used to assess them has great theoretical significance. On the one hand, subjective ratings of metaphorical sentences tend to be reliable—in the sense that participants agree that certain metaphors are more *familiar* than others, and that certain metaphors are more *apt* than others. For instance, one study found that ratings of 260 nonliterary metaphors were internally consistent along seven dimensions—with an average Cronbach α = .88 (Katz et al., 1988). Less clear, however, is what the ratings actually reflect. Can naive participants operationalize abstract qualities of metaphorical sentences? Do ratings of aptness actually reflect the aptness of the metaphors? Do ratings of aptness measure something different from ratings of familiarity?

One concern is that when people are asked to judge sentences for abstract qualities like aptness or familiarity, they may misattribute processing fluency to the dimension they are being asked to evaluate (Alter & Oppenheimer, 2009; Jacoby, Allan, Collins, & Larwill, 1988; Jacoby & Whitehouse, 1989; Kahneman, 2011). People may thus be substituting a hard question (e.g., aptness: how well does the metaphor vehicle capture important properties of the topic of the sentence?) with an easier one (e.g., how easy was the metaphor to understand?). In other words, subjective ratings of dimensions like aptness or familiarity may represent an indirect (and unintentional) measure of processing fluency, rather than aptness or familiarity per se (Thibodeau & Durgin, 2011).

This possibility would raise questions about how to interpret the results of experiments in which ratings of these dimensions are used to predict, for instance, how quickly people process metaphors (e.g., Blank, 1988; Blasko & Connine, 1993; Bowdle & Gentner, 2005; Cardillo, Watson, Schmidt, Kranjec, & Chatterjee, 2012; Chettih, Durgin, & Grodner, 2012; Chiappe & Kennedy, 1999; Chiappe et al., 2003; Citron & Goldberg, 2014; Damerall & Kellogg, 2016; Jones & Estes, 2005, 2006; Keysar, Shen, Glucksberg, & Horton, 2000; Thibodeau & Durgin, 2008). Using ratings data to predict response time data would be tantamount to demonstrating that one measure of processing fluency (e.g., ratings of familiarity or aptness) predicts another measure of processing fluency (e.g., comprehension time).

This possibility may also explain unanswered questions in the metaphor-processing literature. For example, although familiarity and aptness are, in theory, very different properties of metaphorical sentences, ratings of the two dimensions tend to be highly correlated—on the order of r = .9 (Jones & Estes, 2006; Thibodeau & Durgin, 2011). The empirically tight relationship between ratings of familiarity and aptness may stem from the confounding influence of processing fluency on ratings of each of the dimensions. That is, ratings of familiarity and aptness may be highly correlated with one another because they are both measures of processing fluency.

The present studies

To gauge the construct validity of sentence-level ratings of metaphors, we present the results of two experiments and an analysis of four existing large-scale datasets of metaphorical sentences (from Cardillo et al., 2010, 2016; Katz et al., 1988; Roncero & de Almeida, 2015). In both experiments, we manipulated the processing fluency of metaphors by changing the context in which they were presented.

In Experiment 1, we showed that the context manipulation affected how easily people processed the metaphoric sentences in a response time task. In Experiment 2, we showed that the same context manipulation affected ratings of the *comprehensibility*, *familiarity*, *aptness*, *surprisingness*, and *figurativeness* of the metaphoric sentences. Both of these findings raise concerns about the validity of familiarity and aptness ratings, in particular because of how these constructs have been defined and treated in the metaphor-processing literature (e.g., Cardillo et al., 2010, 2016, Katz et al., 1988; Roncero & de Almeida, 2015).

We then conducted a factor analysis on the ratings data from Experiment 2 to gain a better understanding of (a) the

number of linguistic dimensions that can be reliably measured by asking naive participants to rate various properties of metaphorical sentences and (b) what these ratings actually reflect. We also compared the ratings data from Experiment 2 to the comprehension time data from Experiment 1 to address these questions. The results suggest that the ratings data are almost fully explained by two underlying sources of variance—one that reflects how easily participants can interpret the metaphors (processing fluency), and another that reflects how metaphorical the sentences are (figurativeness). As we expected, the first factor, processing fluency, was highly predictive of the reading times from Experiment 1.

Finally, we investigated data from four norming studies of metaphorical sentences (Cardillo et al., 2010, 2016; Katz et al., 1988; Roncero & de Almeida, 2015), by conducting a factor analysis on each. Consistent with the analysis of the ratings data from Experiment 2, we found that most variability in the subjective ratings of metaphorical sentences in these datasets could be explained by two underlying factors, which also seem to reflect processing fluency and figurativeness. We discuss the implications of these findings for theories of figurative-language processing in the General Discussion by emphasizing more careful treatment of subjective ratings of metaphoric sentences. We also suggest alternative methods for quantifying these dimensions and investigating mechanistic theories of metaphor processing.

Experiment 1

Experiment 1 was designed to confirm that manipulating the context for metaphoric target sentences can affect how quickly people process the metaphors (Gerrig & Healy, 1983; Gibbs & Gerrig, 1989; Giora, 2003; Nayak & Gibbs, 1990; Ortony, Schallert, Reynolds, & Antos, 1978; Thibodeau & Durgin, 2008). Participants were timed reading metaphoric target sentences like "My lawyer attacked him with punches that dropped him to the mat," which figuratively compares a lawyer to a boxer. The metaphoric target sentence was always presented at the end of a brief vignette, which provided a context for interpreting the target sentence (see Table 1).

In some cases, the context instantiated a metaphoric mapping that was consistent with the target sentence (*matched* condition). In other cases, the context instantiated an inconsistent metaphoric mapping (e.g., by describing my lawyer as a shark; *mixed* condition), or no metaphoric mapping (*literal* condition). Importantly, as is shown in Table 1, the three versions of the initial vignette always described a similar situation, which was designed to elicit a similar, figurative interpretation of the target sentence.

There are at least two reasons to expect people to read the target sentence faster in a *matched* context. One possibility is that the *matched* context may establish a conceptual mapping between the source and target domains that makes the target sentence easier to process: When people are already thinking of a lawyer as a boxer, they can process a target sentence that extends this metaphor more quickly (Bowdle & Gentner, 2005; Gibbs, 2011; Giora, 2003; Thibodeau & Durgin, 2008). Another possibility is that a *matching* context primes people to read related language more quickly (i.e., lexical priming): Exposure to phrases like "up against the ropes" prepares people to process semantically related words like "punches" (McGlone, 2011). The present article is not designed to distinguish between these theoretical accounts. Instead, as we noted above, the goal of Experiment 1 was simply to show that this particular context manipulation affects how fluently people process this particular set of metaphoric target sentences. In Experiment 2, we tested whether the same context manipulation affected ratings of dimensions such as the aptness and familiarity of the same metaphoric target sentences.

Method

The data for Experiment 1 were collected as part of an electroencephalographic study of metaphor processing (Sikos, Thibodeau, Strawser, & Durgin, 2013). An article detailing the full scope of the theoretical questions it was designed to address and the results of the event-related potential (ERP) data are the topic of a forthcoming article (Sikos, Thibodeau, Strawser, Klein, & Durgin, 2013). Here we focus on the response time data as a simple measure of processing fluency. For brevity and clarity, we discuss the methodological details that are relevant to the comprehension time data here; additional details are included in the supplementary materials.

Table 1 Example stimuli

Condition	Context
Matched	Going in, I was really worried about the trial. But during the cross- examination, the main defense witness turned out to be little more than a featherweight. Once my lawyer had him up against the ropes, there was no way we were going to lose.
Mixed	Going in, I was really worried about the trial. But during the cross- examination, the main defense witness turned out to be little more than shark bait. Once my lawyer sunk his teeth into him, there was no way we were going to lose.
Literal	Going in, I was really worried about the trial. But during the cross- examination, the credibility of the main defense witness turned out to be questionable. Once my lawyer confronted him with the contradictions in his testimony, there was no way we were going to lose.

The metaphorical target sentence "My lawyer attacked him with punches that dropped him to the mat" was presented in a matched, mixed, or literal context.

Participants Ninety-six Swarthmore College students (48 males, 48 females; mean age = 19.8 years, range = 18–23) participated in the experiment in exchange for either course credit or payment.

Materials Thirty-six stimulus bases were developed, like the one shown in Table 2. For each, two different metaphoric mappings (e.g., A: lawyers are boxers vs. B: lawyers are sharks) were used to generate conceptually parallel initial vignettes that contained two or more context sentences. A third conceptually parallel vignette (C) used nonmetaphoric language to communicate the same basic situation. In addition, two metaphoric target sentences were created for each stimulus base: one for each metaphoric mapping.

The target sentences were designed to be interpretable following a description that included *matched* or *mixed* metaphors or nonmetaphoric, *literal*, language. That is, the study was not designed 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 metaphoric target sentences, so that the metaphoric target sentence would be processed more easily in the matched than in the mixed or literal contexts (Thibodeau & Durgin, 2008).

The 36 stimulus bases provided a framework for creating 72 matched, 72 mixed, and 72 literal items. Items were created for the three conditions by pairing contexts and target sentences: matched (Context A + Extension A; Context B + Extension B), mixed (Context A + Extension B; Context B + Extension A), and literal (Context C + Extension A; Context C + Extension B).

Table 2 Example base for generating stimulus items

Context
A. Going in, I was really worried about the trial. But during the cross-examination, the main defense witness turned out to be little more than a featherweight. Once my lawyer had him up against the ropes, there was no way we were going to lose.
B. Going in, I was really worried about the trial. But during the cross-examination, the main defense witness turned out to be little more than shark bait. Once my lawyer sunk his teeth into him, there was no way we were going to lose.
C. Going in, I was really worried about the trial. But during the cross-examination, the credibility of the main defense witness turned out to be questionable. Once my lawyer confronted him with the contradictions in his testimony, there was no way we were going to lose.
Extension
A. My lawyer attacked him with PUNCHES that dropped him to the mat.
B. My lawyer attacked him with JAWS that tore him apart.

Context A, B, or C was presented with Extension A or B. Target words are indicated by capitalization. The target words were not capitalized in the response time task (Exp. 1), but they were capitalized in the ratings task (Exp. 2).

In every case, the first use of the metaphor in the target sentence (extension) was nominal and sentence-medial (e.g., "My lawyer attacked him with PUNCHES/JAWS that…"), and the 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 ERPs, 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 sentence served as its own control because it appeared in all three contexts.

In addition to the target items, 72 unrelated fillers were created with a similar structure: a short multisentence context followed by a concluding target sentence. One third of the fillers contained a semantic anomaly; one third of the fillers contained a syntactic anomaly or misspelling; and one third of the fillers were well-formed.

Procedure Participants were exposed to 12 items from each experimental condition (matched, mixed, literal)—one item from each stimulus base—for a total of 36 metaphoric target sentences. These target items were randomly interspersed with the 72 unrelated filler items, so that each participant responded to 108 items total (33.3% contained a metaphor in the target sentence). The target items and fillers were blocked so that participants were not exposed to more than two target items in succession. The order of the items was randomized across participants.

For each trial, context sentences were presented in full on a single screen under the heading "Background" and stayed on the screen until the participant pressed a button to continue. Then a fixation cross appeared in the center of the screen for 1,100 ms, followed by the target sentence, which was presented one word at a time. Each word of the target sentence was presented for 400 ms, followed by a 150-ms interstimulus interval, with the exception of critical words (e.g., PUNCHES/JAWS), which were presented for 450 ms in order to facilitate analysis of the ERP effects. When the target sentence ended, a question mark appeared on the screen.

Participants were instructed to respond as quickly as possible if the story made sense (including metaphorically) by pressing a key on the keypad. They were given 3,000 ms to make this judgment; a failure to respond in this time period was interpreted as comprehension failure. Participants were told that both speed and accuracy were important.

We chose this sensibility judgment task because we were interested in how people process metaphors in naturalistic settings. Previous work had shown that the ways people are instructed to interpret metaphorical language in an experiment can influence the ways they process metaphors (Bohrn, Altmann, & Jacobs, 2012). More cognitively demanding tasks like valence or imagery judgments tend to take more time and

to engage the brain in qualitatively different ways than less cognitively demanding tasks like sensibility judgments (Yang, Edens, Simpson, & Krawczyk, 2009).

Analysis We conducted analyses on the response time data using mixed models with the lme4 and lmerTest libraries in R (Bates, Maechler, Bolker, & Walker, 2014; Kuznetsova, Brockhoff, & Christensen, 2015). We treated participants and items as random effects (i.e., allowing the model to fit random intercepts by participants and items)—with random slopes by condition computed by participant and by item (Baayen, Davidson, & Bates, 2008). The means and standard deviations presented in the text were averaged by item and then by condition. Only data from trials in which participants indicated that they understood the metaphoric target sentence were analyzed.

Results

As expected, the context manipulation affected how quickly participants read the metaphoric target sentences, $F(2, 48.2) = 9.29$, $p < .001$. As is shown in Fig. 1, participants understood the metaphoric target sentence more quickly in the matched condition ($M = 1{,}494$ ms, $SD = 179$) than in the mixed ($M = 1{,}573$ ms, $SD = 180$) or the literal ($M = 1{,}612$ ms, $SD = 238$) condition, ts > 3.7, ps < .001. There was no difference in comprehension times between the mixed and literal conditions, $t(47.9) = 1.37$, $p = .177$. These results suggest that the target sentence was processed more fluently when it was preceded by a context that included matching metaphoric language than when it was preceded by a context that included mixed metaphoric language or literal language.

Discussion

The critical finding from Experiment 1 is that people processed the metaphoric target sentences more fluently in the matched than in the mixed or literal contexts, which is consistent with prior work on the role of context in metaphor processing (Gerrig & Healy, 1983; Gibbs & Gerrig, 1989; Giora, 2003; Nayak & Gibbs, 1990; Ortony et al., 1978; Thibodeau & Durgin, 2008). We tested for a similar context effect on ratings of these same target sentences in Experiment 2.

Experiment 2

In Experiment 2, we tested whether the same manipulation of processing fluency would affect ratings of the *surprisingness*, *comprehensibility*, *familiarity*, *metaphoricity*, and *aptness* for the same metaphoric target sentences. We expected that ratings of all five dimensions would be influenced by the manipulation, which would suggest that sentence-level ratings of metaphoric sentences are confounded by processing fluency (Thibodeau & Durgin, 2011). Such a finding would raise concerns about the validity of aptness and familiarity ratings in particular, which are most relevant to cognitive psychological theories of how people process metaphors (e.g., Bowdle & Gentner, 2005; Glucksberg & Keysar, 1990). One might expect, for example, people to judge a metaphoric target sentence as less surprising and more comprehensible when it is in a matched context. But, given how metaphor aptness and familiarity have been defined in the metaphor-processing literature, ratings of these dimensions should be less susceptible to context manipulations (e.g., Roncero & de Almeida, 2015).

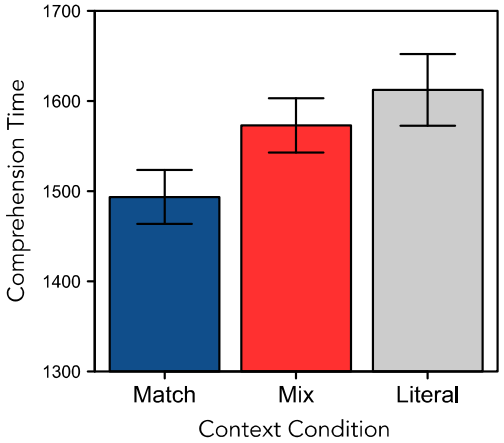

Fig. 1 Mean comprehension times for target metaphors by condition. Error bars denote standard errors of the means

Method

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 completion code into Turk, leaving data from 1,193 participants for analysis.

Survey materials and procedure The stimuli for Experiment 2 were identical to those of Experiment 1, except that there were no filler items in Experiment 2. Each participant read one item (i.e., context–target pairing) from each of the 36 stimulus bases described in the Method section of Experiment 1, such that all participants rated 12 matched, 12 mixed, and 12 literal items. The order of items was pseudorandomized so that no one saw more than two items from a given condition in sequence. Participants were asked to attend to the (uppercase) target word—the word that instantiated the metaphoric

extension in the target sentence—when making their ratings (see Table 2).

There were two versions of the questionnaire.[2] One asked participants to rate the target words along four dimensions: surprisingness, comprehensibility, familiarity, and metaphoricity. The other asked participants to rate target words along the single dimension of aptness (see the definitions of these dimensions in Table 3). Following the instructions, participants saw an example question and had an opportunity to reread the definitions of the dimensions. Then they rated two practice 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 First, we tested for an effect of the context manipulation on the five rated dimensions by fitting separate, by-item, repeated measures analyses of variance (ANOVAs). Of note, the effects of context on the ratings data were large, and a mixed-model approach to analyzing the data yielded results similar to those from the ANOVAs. For clarity and brevity, and because we have made the stimuli available for other researchers with mean ratings of each target metaphor by item, we present the repeated measures ANOVAs by items in this section of the Results.

Second, we conducted an exploratory factor analysis on the ratings data: a principal components analysis (PCA) with normalized variables (scaled, centered), using singular value decomposition (Mardia, Kent, & Bibby, 1980). PCA is a statistical procedure for revealing the internal structure of a dataset containing possibly correlated variables in a way that best explains the common variance of the data (Dunteman, 1989). We considered the eigenvalues of the factors and the amount of variance in the ratings data that each factor explained in order to make a decision about how many factors to retain (Henson & Roberts, 2006). Then we tested for an effect of the context manipulation on the retained factors.

Third, we compared the ratings data from Experiment 2 to the response time data from Experiment 1. We used a mixed-model approach for this analysis, as we described in the Analysis section of Experiment 1 (with predictor variables included as fixed effects, with random slopes and intercepts computed by participants and by items), using the lmerTest library in R to compute F ratios for the fixed effects. We also report a comparison of nested models, for which the difference in likelihood ratios approximates a χ^2 distribution with the number of added parameters as its degrees of freedom (Baayen et al., 2008).

[2] A secondary research goal was to investigate trade-offs associated with having participants rate metaphors for such psycholinguistic dimensions as familiarity and aptness in isolation versus in parallel. Since this investigation was preliminary in nature and secondary to our primary research goal, the motivations and results pertaining to it are discussed in the supplementary material.

Results

Effects of context on ratings Table 4 and Fig. 2 show the effects of the context manipulation on ratings of the five dimensions. The analysis revealed that a matched context made the target metaphor seem less surprising, more comprehensible, more familiar, more metaphorical, and more apt. That is, in each of the models, the omnibus difference was driven by the matched condition relative to the mixed and literal conditions, ts > 5, ps < .001. Only one comparison between the mixed and literal conditions yielded a difference: The target sentences were rated as more apt in the literal context, $t(71) = 3.261$, $p = .002$.

The effect of the context manipulation was not only statistically significant, but also ecologically significant. For example, a matched context made the metaphors seem more apt, by 1.22 units as compared to a mixed context, and by 1.03 units as compared to a literal context. The values that marked the first and third quartiles of aptness ratings were 3.77 and 5.13, respectively (median = 4.52; min = 2.08; max = 6.14). Thus, the difference between a mixed and a matched context (1.22 units) was similar in magnitude to the difference between being judged in the lowest versus the highest quartile of aptness (1.36 units).

Exploratory factor analysis Correlations between the ratings of surprisingness, metaphoricity, familiarity, comprehensibility, and aptness are shown in Table 5. Pairwise comparisons revealed a significant correlation between each pair of dimensions at the $\alpha = .05$ level. Of note, ratings of metaphoricity were somewhat less correlated with the other four dimensions (range of magnitudes: $r = .14$ to .43); ratings of surprisingness, familiarity, comprehensibility, and aptness were highly correlated with one another (range of magnitudes: $r = .77$ to .90). This pattern of results suggests that ratings of the five dimensions may be driven by one or two underlying factors.

To test this possibility, we conducted a PCA on the ratings data, the results of which are shown in Table 6. The Kaiser criterion suggests retaining factors with an eigenvalue greater than 1, whereas other decision rules suggest considering the marginal amount of variance explained by each factor (Henson & Roberts, 2006). With both of these suggestions taken into account, we retained the first two principal components. Although the eigenvalue for the second component was less than 1, it was close to this cutoff (0.93), and, taken together, the first and second components explained approximately

Table 3 Definitions of dimensions, 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."
Familiarity[a]	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.

[a] This dimension was labeled "conventionality" for participants.

93% of the variance in the ratings data. Note that the eigenvalue for the third component was much smaller (0.21) and captured relatively little variance in the ratings data: 4%, as compared to 74% for the first component and 19% for the second component.

As is shown in Table 6 and Fig. 3, the first component loads highly on the inverse of the ratings of surprisingness, and positively on ratings of comprehensibility, familiarity, and aptness; it also loads onto ratings of metaphoricity to a lesser degree. We use the label *processing fluency* to describe this first component, since it seems to capture how easily people were able to understand the metaphors. The second component loads highly on ratings of metaphoricity, and, to a lesser extent, on ratings of surprisingness and on the inverse of ratings of familiarity. We refer to this dimension as *figurativeness*, to differentiate between this factor and ratings of metaphoricity.

Table 4 Effect of context on ratings of surprisingness, comprehensibility, familiarity, metaphoricity, and aptness by context

Dimension	Means (and *SD*s)			Model results	
	Matched	Mixed	Literal	F	η^2
Surprisingness	3.85 (.58)	4.47 (.72)	4.46 (.75)	111.5	.61
Comprehensibility	5.29 (.48)	4.77 (.72)	4.72 (.75)	69.2	.49
Familiarity	4.51 (.63)	4.15 (.71)	4.15 (.76)	46.9	.40
Metaphoricity	4.87 (.36)	4.65 (.39)	4.70 (.38)	32.3	.31
Aptness	5.16 (.59)	3.94 (.84)	4.13 (.86)	187.2	.73

Means (and standard deviations) are shown by condition for each dimension. The results of separate repeated measures ANOVAs are also shown ($df_1 = 2, df_2 = 142$), along with a measure of effect size (η^2). All ANOVAs are significant at the $p < .001$ level.

Effects of context on factors A repeated measures ANOVA on the first factor, Processing Fluency, revealed an effect of the context manipulation, $F(2, 142) = 133.8$, $p < .001$, $\eta^2 = .65$. Processing fluency was greater in the matched condition ($M = 1.14$, $SD = 1.41$), than in the mixed ($M = –0.62$, $SD = 1.84$) or the literal ($M = –0.52$, $SD = 1.97$) condition, ts > 12, ps < .001. There was no difference in processing fluency between the mixed and literal conditions, $t(71) = 0.99$, $p = .323$.

A repeated measures ANOVA on the second factor, Figurativeness, also revealed an effect of the context manipulation, $F(2, 142) = 7.54$, $p < .001$, $\eta^2 = .10$. Figurativeness was greater in the matched condition ($M = 0.13$, $SD = 0.93$) than in the mixed ($M = –0.13$, $SD = 1.01$) or the literal ($M = 0.00$, $SD = 0.95$) condition, ts > 2, ps < .05. In addition, figurativeness was marginally greater in the literal than in the mixed condition, $t(71) = 1.82$, $p = .073$.

The effects of the context manipulation on these "big two" factors are shown in Fig. 4. Of note, although the context manipulation affected both components, it had a much stronger influence on processing fluency (95% confidence interval [CI] for $\eta^2 = [.56, .71]$) than on figurativeness (95% CI for $\eta^2 = [.02, .19]$). Thus, the processing fluency component seems to be a clear reflection of how easily participants were able to understand the target sentences, whereas the figurativeness component may reflect a comparison between the implied meaning of the target word and its literal sense (above and beyond the effect of processing fluency). The effect of condition figurativeness suggests that presenting the target sentence in a matched context facilitates this judgment.

Predicting comprehension time The statistical model described in Experiment 1—in which condition was used to predict comprehension time—was augmented to include processing fluency and figurativeness as predictor variables. Adding processing fluency and figurativeness to the model significantly improved the fit, $\chi^2(15) = 38.14$, $p < .001$. Comprehension times were related to processing fluency,

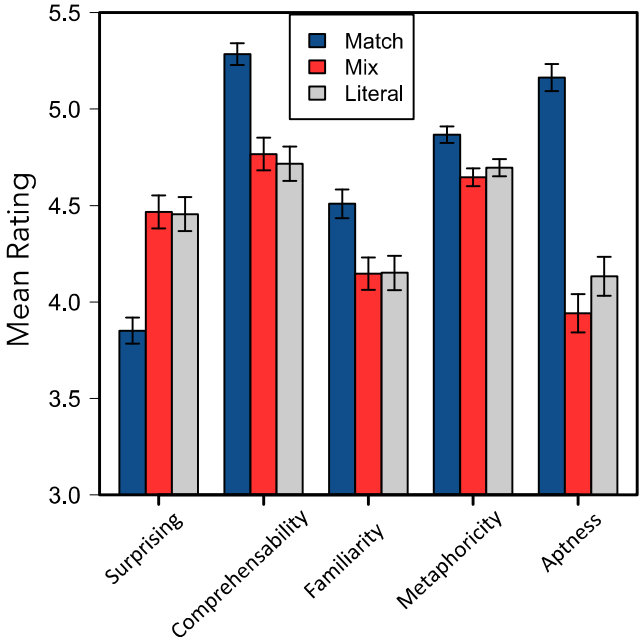

Fig. 2 Mean ratings of surprisingness, comprehensibility, familiarity, metaphoricity, and aptness by condition. Error bars denote standard errors of the means

$F(1, 43.2) = 16.12$, $p < .001$, but not to figurativeness, $F(1, 32.7) = 0.53$, $p = .470$.

Interestingly, adding these measures to the model affected the reliability of condition (match vs. mixed vs. literal) as a predictor of response times: Without processing fluency and figurativeness in the model, condition was a reliable predictor of response times, $F(2, 48.2) = 9.29$, $p < .001$; with processing fluency and figurativeness in the model, condition was no longer a reliable predictor of response times, $F(2, 46.5) = 1.48$, $p = .238$ (see Table 7). In other words, the continuous measure of processing fluency accounted for modulations of comprehension time caused by the context manipulation, as well as for variability in comprehension times within each condition (and any potential effect of figurativeness; see Fig. 5).

A second model tested for interactions between condition (mixed, matched, literal) and processing fluency and between condition and figurativeness, neither of which was significant,

Table 5 Correlations between rated dimensions

Dimension	1	2	3	4	5
1. Surprisingness		–.85***	–.14*	–.90***	–.82***
2. Comprehensibility			.43***	.90***	.88***
3. Metaphoricity				.26***	.42***
4. Familiarity					.77***
5. Aptness					

Asterisks indicate statistical significance at the * $p < .05$ and *** $p < .001$ levels.

Table 6 Results of the principal component analysis, with factor loadings, eigenvalues, and percentages of variance explained for each principal component

Dimension	PC1	PC2	PC3	PC4	PC5
Surprisingness	–.48	.32	–.03	.70	.43
Comprehensibility	.50	.02	.05	.67	–.54
Familiarity	.48	–.19	.60	.08	.60
Metaphoricity	.23	.92	.18	–.24	–.04
Aptness	.48	.06	–.77	.02	.41
Eigenvalue	3.71	0.93	0.21	0.10	0.05
Variance explained	74%	19%	4%	2%	1%

$\chi^2(6) = 5.19$, $p = .520$. As is shown in Fig. 5, the relationships between processing fluency and comprehension time were similar in the matched, mixed, and literal conditions.

Discussion

In Experiment 2, participants rated metaphoric target sentences as less surprising, more comprehensible, more familiar, more metaphorical, and more apt in a context that included matched metaphoric language, as opposed to mixed metaphoric language or literal language. For example, presenting a metaphor in a matched, rather than a mixed or a literal, context increased aptness by more than a full unit on the scale—enough to shift the metaphor from one of the least apt (bottom third) to one of the most apt (top third) items. Ratings of the five dimensions were highly correlated with one another, and an exploratory factor analysis suggested that variability in the ratings data reflected two underlying sources of variance. We interpreted the first as a measure of *processing fluency*, or how easily participants were able to understand the meaning of the sentence, and the second as a measure of *figurativeness*, or the extent to which the meaning of the sentence departed from a literal interpretation. Processing fluency was a reliable predictor of the comprehension time data from Experiment 1; figurativeness was not.

On the one hand, it is unremarkable that the context manipulation affected ratings of surprisingness and comprehensibility. The stimuli were designed so that a matched context, as compared to a literal or mixed context, would facilitate processing of the target sentence (i.e., make it less surprising and more comprehensible). On the other hand, the effect of context on ratings of familiarity and aptness raises questions about the ability of naive participants to accurately operationalize these abstract qualities of metaphors, suggesting that they are confounded by processing fluency.

An alternative possibility is that the context manipulation affected the familiarity and aptness of the target sentences in a way that would be predicted by theoretical accounts of familiarity and aptness. In other words, perhaps a matching context

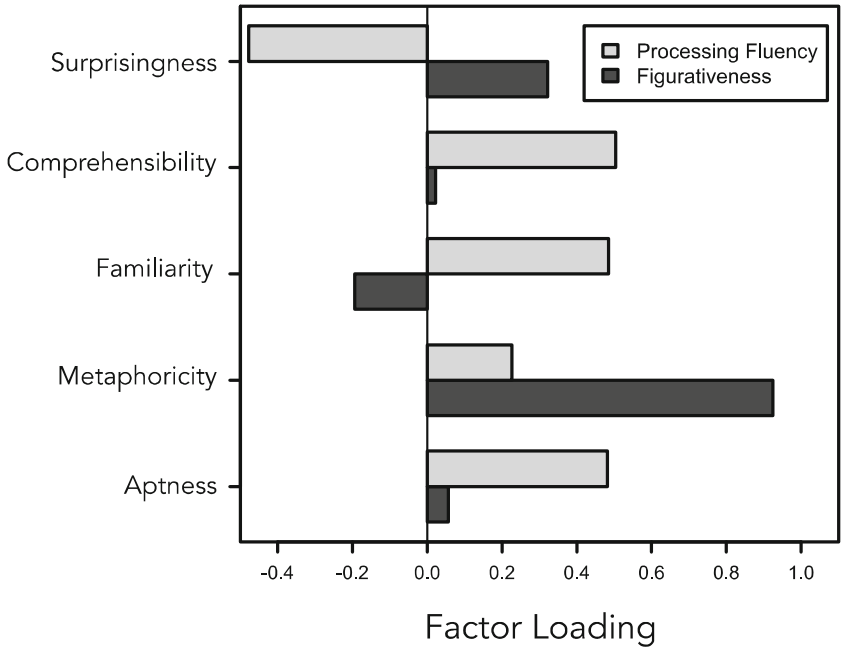

Fig. 3 Factor loadings of our rated dimensions on the two principal components: processing fluency and figurativeness

actually makes the target sentences more familiar and more apt. This interpretation would represent a severe limitation to the explanatory value of subjective ratings data, suggesting that familiarity and aptness can only be measured in the exact local context in which they appear. In our view, a more parsimonious interpretation appeals to the confounding influence of processing fluency.

The notion that the primary source of variance in the ratings data reflects processing fluency represents an important departure from how ratings data are often treated in the metaphor-processing literature (e.g., Damerall & Kellogg, 2016; Jones & Estes, 2006; Katz et al., 1988). For example, in the first large-scale norming study of metaphoric sentences, Katz et al. found similar relationships between ratings of such dimensions as aptness and familiarity. However, they did not question the construct validity of the ratings, instead noting that it was possible to identify subsets of metaphoric sentences in which ratings of familiarity were independent of the ratings of aptness.

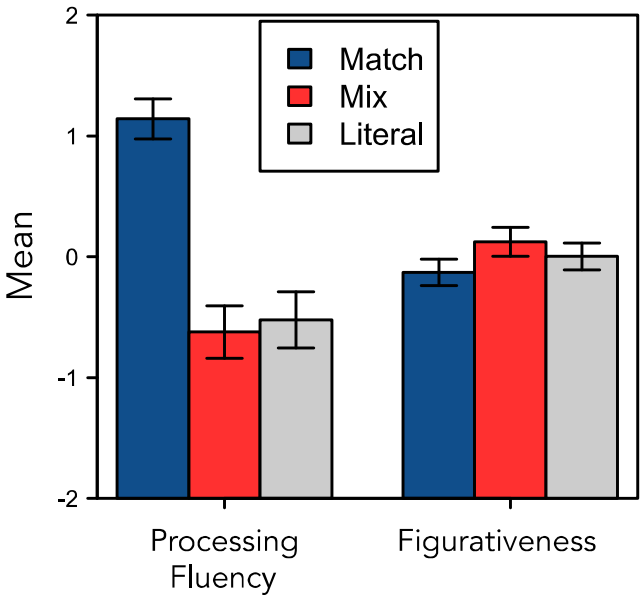

Fig. 4 Effects of our context manipulation on two factors. Error bars denote standard errors of the means

In our view, it is misleading to use ratings of familiarity as a measure of familiarity and ratings of aptness as a measure of aptness, even if it is possible to identify clusters of metaphoric sentences that differ in rated familiarity and aptness. This is because identifying subsets of metaphoric sentences that differ in familiarity but not aptness (or vice versa) may introduce artifacts that limit the representativeness of the stimulus set (see the supplementary materials for an example).

Finally, although we have shown that processing fluency confounds ratings of metaphoric sentences that are made in context, it is unclear whether processing fluency confounds ratings of isolated (decontextualized) metaphors. One possibility is that the context manipulation used in Experiment 2 was so strong that it impeded participants' ability to assess the familiarity and aptness of the target sentences. Ratings of decontextualized metaphors may be less influenced by processing fluency.

The following section investigates this question. We conducted factor analyses on four stimulus sets of single-sentence metaphors to test whether they would reveal a structure similar to what we found in Experiment 2. We expected, consistent with the results of Experiment 2, to find a "big two" factor

Table 7 Results of a mixed-effect linear model that included condition (match, mix, literal), processing fluency, and figurativeness as predictors of comprehension times

Predictor	β	*SE*	*p*
Intercept	.11	.08	.183
Condition: Match	–.12	.08	.127
Condition: Mix	–.09	.06	.124
Processing fluency	–.15	.04	<.001
Figurativeness	–.02	.03	.470

structure, with one source of variance reflecting how easily people were able to process the sentences, or processing fluency, and another reflecting the figurativeness of the sentences.

Analyses of existing datasets

Several stimulus sets of metaphorical (and nonmetaphorical) sentences have been developed to test theories of metaphor processing. Here we consider four: from Cardillo et al. (2010; Cardillo et al., 2016), Katz et al. (1988), and Roncero and de Almeida (2015). In each, naive participants rated metaphorical sentences along a set of target dimensions. The stimulus sets and normed ratings have informed important work on metaphor processing in the years since they were published (e.g., Bowdle & Gentner, 2005; Cardillo et al., 2012; Citron & Goldberg, 2014; De Grauwe, Swain, Holcomb, Ditman, & Kuperberg, 2010; Diaz, Barrett, & Hogstrom, 2011; Diaz & Hogstrom, 2011; Gentner & Wolff, 1997; Kacinik & Chiarello, 2007; Kuiken, Chudleigh, & Racher, 2010; Schmidt & Seger, 2009; Thibodeau & Durgin, 2011; Xu, 2010).

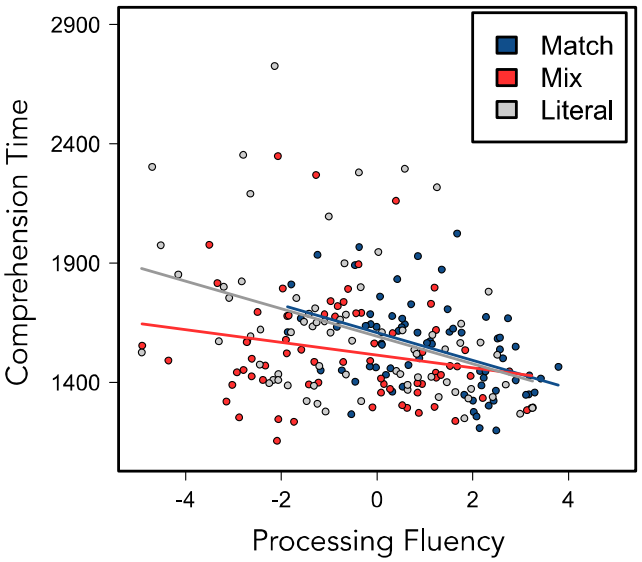

Fig. 5 Relationship between processing fluency computed from the ratings data (Exp. 2) and from comprehension times (Exp. 1), by condition

Table 8 shows the dimensions that were measured in each of the four stimulus sets. For brevity, we focus on sentence-level ratings (rather than, e.g., ratings of the imagery associated with the target or base term alone) for nonliterary metaphors (rather than literary metaphors or similes). Our goal in this section is to quantify the number of dimensions that are measured reliably in these datasets and to identify what the dimensions reflect. To facilitate comparison between the stimulus sets, we have modified some of the labels that were used in the original work. We discuss the details of each ratings task, including the labels and instructions used in the original work, in the supplementary material.

Results

A PCA was conducted separately for each of the four stimulus sets. Consistent with the results of Experiment 2, we retained the first two factors from each analysis. The factor loadings are illustrated in Fig. 6. As is shown in Table 8, a majority of the variance in the rated dimensions (between 78% and 89%) could be explained by two underlying factors. The first principal component tended to capture variability associated with the familiarity, aptness, imagery, ease of interpretation, comprehensibility, and semantic relatedness of the metaphors—variables that we interpret as reflecting how easily a metaphor's meaning can be interpreted—whereas the second factor consistently captured variability associated with the figurativeness of the metaphorical sentences.

Discussion

Our goal in analyzing existing datasets was to show that subjective ratings of psychological dimensions of metaphorical sentences tend to be highly correlated—and seem to reflect two sources of underlying variance. This is true whether the metaphoric sentences are presented in context (Exp. 2) or out of context (as in these stimulus sets). The first of the "big two" sources of variance represents an indirect measure of processing fluency (Alter & Oppenheimer, 2009; Jacoby, Allan, Collins, & Larwill, 1988; Jacoby & Whitehouse, 1989; Kahneman, 2011). A secondary, reliably distinguishable, source of variance reflects the figurativeness of the sentences.

General discussion

The study of personality and social psychology reveals that variability in many traits used to describe people can be explained in terms of a relatively small number of underlying factors (Digman & Inouye, 1986; Fiske, 1949; McCrae & Costa, 1999), and that situations are often the best predictors

Table 8 Numbers of metaphorical sentences, linguistic dimensions rated at the sentence level, and variance explained by our first two principal components in four published norm sets of nonliterary metaphors

Dataset *N* Metaphors	Dimensions rated at sentence level	Variance explained
Cardillo et al. (2010) *N* = 280	Familiarity, aptness, imageability, metaphoricity	.865
Cardillo et al. (2016) *N* = 120	Familiarity, aptness, imageability, ease of interpretation, metaphoricity	.817
Katz et al. (1988) *N* = 260	Familiarity, aptness, comprehensibility, ease of interpretation, imagery, semantic relatedness, metaphoricity	.890
Roncero and de Almeida (2015) *N* = 84	Familiarity, conventionality, aptness, metaphoricity	.781

of behavior (Asch, 1951; Milgram, 1963). Here we have made an analogous observation about metaphoric sentences: Most of the variability in subjective ratings of metaphors can be explained in terms of two underlying factors, and context has a strong influence on metaphor processing (Gerrig & Healy, 1983; Gibbs & Gerrig, 1989; Giora, 2003; Nayak & Gibbs, 1990; Ortony et al., 1978; Thibodeau & Durgin, 2008). We have referred to these dimensions as the "big two" dimensions of metaphoric sentences.

In Experiment 1, we showed that a context manipulation affected how fluently people processed metaphors. In Experiment 2, we showed that the same context manipulation affected ratings of the comprehensibility, familiarity, aptness, surprisingness, and metaphoricity of the target metaphors. A factor analysis of these ratings revealed that most of the variance (roughly 90%) could be explained by two underlying dimensions: *processing fluency*, or how easily participants were able to interpret the meaning of the given sentence, and

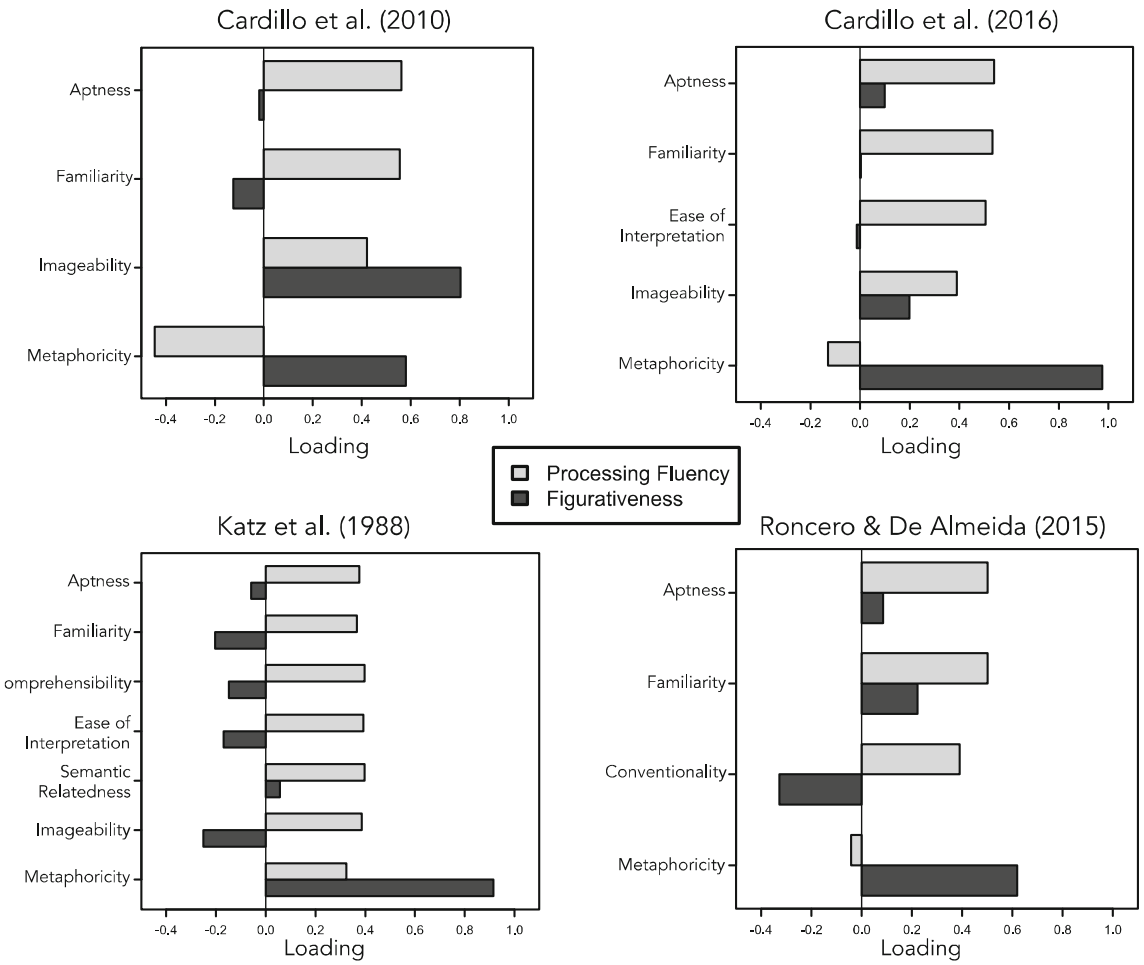

Fig. 6 Factor loadings for sentence-level ratings of four stimulus sets of metaphoric sentences

figurativeness, or the nonliteralness of the interpretation. Processing fluency varied considerably as a function of context; figurativeness was fairly stable, in comparison.

Then we analyzed four large datasets of metaphoric sentences, in which groups of naive participants subjectively rated metaphors along a variety of psychological dimensions such as familiarity, aptness, and metaphoricity (Cardillo et al., 2010, 2016; Katz et al., 1988; Roncero & de Almeida, 2015). Exploratory factor analyses on each dataset revealed that most of the variability in the rated dimensions (roughly 80%) could be explained by our two underlying dimensions: processing fluency and figurativeness.

One important implication of these findings relates to the explanatory power of sentence-level ratings of metaphors 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 inclusion statements (e.g., Chiappe et al., 2003; Glucksberg & Haught, 2006; Jones & Estes, 2006). *Familiarity* has been the focus of researchers who argue that people process metaphors by comparison (e.g., Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997). Although it very well may be the case that these dimensions can be measured and can explain, a priori, variability in metaphor processing fluency, our results suggest that operationalizing the constructs by gathering ratings from naive participants is problematic.

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 indirect measures of *processing fluency* (Thibodeau & Durgin, 2011). When people are asked to rate sentences for abstract qualities like familiarity and aptness, they mistakenly report how easily they processed the sentence (Alter & Oppenheimer, 2009; Jacoby, Allan, Collins, & Larwill, 1988; Jacoby & Whitehouse, 1989; Kahneman, 2011). As a result, more careful treatment of subjective ratings data will be necessary in psycholinguistic research on metaphor processing.

Despite these concerns, subjective ratings data can still be useful for researchers interested in testing psycholinguistic theories of metaphor processing. First, there is value in using subjective ratings data as a dependent variable, as a measure of processing fluency. Experiment 2 showed that ratings of metaphors were sensitive to a context manipulation in much the same way as a more controlled response time task (Exp. 1). Thus, ratings tasks may complement response time and neural-imaging tasks as a tool for quantifying variability in metaphor-processing fluency. One advantage of using a ratings task for this purpose is that ratings data often require fewer resources to collect than response time or imaging data. Second, statistical procedures like principal components analysis can be used to partial out the primary source of variance (processing fluency), leaving researchers with a more reliable measure of figurativeness. Such a measure can, for instance, be used to test mechanistic questions about the role of figurativeness in cognitive and neural theories of language processing.

Finally, it is important to note that not all work on metaphor processing relies on subjective ratings to quantify such dimensions as familiarity and aptness. For example, familiarity can be manipulated by exposing people to a series of related metaphors. People are faster to read a target sentence like "Education is a lantern" after hearing related metaphors like "An encyclopedia is a lantern" and "A mentor is a lantern" (Bowdle & Gentner, 2005). This effect is not simply due to lexical priming: Exposing people to literal sentences that include the same term as the metaphor vehicle (lantern), as in "A camp light is a lantern," or metaphorical sentences that include the same metaphor vehicle but that instantiate a different meaning, such as "A flag is a lantern," do not facilitate comprehension of the target sentence "Education is lantern" (Thibodeau & Durgin, 2011). These studies show that familiarity, operationalized through an experimental manipulation, plays a role in metaphor processing.

In addition, with the development of large-scale language corpora and text analysis software, it may be possible to quantify dimensions like familiarity and aptness without relying on ratings from naive participants. Corpus analysis can be used to gauge how frequently metaphors appear in public discourse (e.g., Steen et al., 2010). Tools like latent semantic analysis (LSA; Landauer & Dumais, 1997) may be useful in developing a more objective metric of aptness. LSA conceptualizes semantic knowledge, and cross-domain similarity, in terms of geometric space, which can be used to measure a quality of metaphors that is consistent with early theoretical work on aptness (Kintsch, 2000; Kintsch & Bowles, 2002; Tourangeau & Sternberg, 1981).

In conclusion, two experiments and an analysis of existing stimulus sets raise questions about the construct validity of sentence-level ratings of metaphoric sentences. When naive participants are asked to judge metaphor aptness and familiarity, their ratings are contaminated by processing fluency. Subjective ratings of these dimensions, as a result, have limited value when they are treated as explanatory variables to test psychological theories about figurative-language processing. These data, however, do contain valuable information—about how easily people can understand the sentences and about how metaphorical the sentences are—which can be reliably extracted using such data reduction techniques as exploratory factor analyses. We have called these factors the "big two" dimensions of metaphorical sentences. In addition, this work highlights an opportunity for researchers to develop new, more objective methods for operationalizing constructs like familiarity and aptness.

References

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

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. S. Guetzkow (Ed.), *Groups, leadership, and men: Research in human relations* (pp. 222–236). Pittsburgh, PA: Carnegie Press.

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. doi:10.1016/j.jml.2007.12.005

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version, 1.0-7). Retrieved from cran.r-project.org/package=lme4

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

Bohrn, I. C., Altmann, U., & Jacobs, A. M. (2012). Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia, 50,* 2669–2683.

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

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,* 651–664. doi:10.3758/BRM.42.3.651

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *NeuroImage, 59,* 3212–3221.

Cardillo, E. R., Watson, C., & Chatterjee, A. (2016). Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods, 49,* 471–483. doi:10.3758/s13428-016-0717-1

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,* 295–311. doi:10.1037/a0025862

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.

Citron, F. M. M., & Goldberg, A. E. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience, 26,* 2585–2595. doi:10.1162/jocn_a_00654

Damerall, A. W., & Kellogg, R. T. (2016). Familiarity and aptness in metaphor comprehension. *American Journal of Psychology, 129,* 49–64.

De Grauwe, S., Swain, A., Holcomb, P., Ditman, T., & Kuperberg, G. (2010). Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia, 48,* 1965–1984.

Diaz, M. T., & Hogstrom, L. J. (2011). The influence of context on hemispheric recruitment during metaphor processing. *Journal of Cognitive Neuroscience, 23,* 3586–3597.

Diaz, M. T., Barrett, K. T., & Hogstrom, L. J. (2011). The influence of sentence novelty and figurativeness on brain activity. *Neuropsychologia, 49,* 320–330.

Digman, J. M., & Inouye, J. (1986). Further specification of the five robust factors of personality. *Journal of Personality and Social Psychology, 50,* 116–123.

Dunteman, G. H. (1989). *Principal components analysis (No. 69)*. Newbury Park, CA: Sage.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44,* 329–344.

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,* 13–19.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45–56. doi:10.1037/0003-066X.52.1.45

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

Gerrig, R., & Healy, A. (1983). Dual processes in metaphor understanding: Comprehension and appreciation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 667–675. doi:10.1037/0278-7393.9.4.667

Gibbs, R. W. (2011). Evaluating conceptual metaphor theory. *Discourse Processes, 48,* 529–562.

Gibbs, R. W., Jr., & Gerrig, R. J. (1989). How context makes metaphor comprehension seem "special". *Metaphor and Symbol, 4,* 145–158.

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

Giora, R. (1999). On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics, 31,* 919–929.

Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford, UK: Oxford University Press.

Giora, R. (2007). Is metaphor special? *Brain and Language, 100,* 111–114.

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

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review, 97,* 3–18. doi:10.1037/0033-295X.97.1.3

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological Measurement, 66,* 393–416.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126–135. doi:10.1037/0096-3445.118.2.126

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,* 240–247. doi:10.1037/0278-7393.14.2.240

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, 55,* 18–32. doi:10.1016/j.jml.2006.02.004

Kacinik, N. A., & Chiarello, C. (2007). Understanding metaphors: Is the right hemisphere uniquely involved? *Brain and Language, 100,* 188–207.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: 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,* 191–214. doi:10.1207/s15327868ms0304_1

Keysar, B., Shen, Y., Glucksberg, S., & Horton, W. S. (2000). Conventional language: How metaphorical is it? *Journal of Memory and Language, 43,* 576–593.

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

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

Kuiken, D., Chudleigh, M., & Racher, D. (2010). Bilateral eye movements, attentional flexibility and metaphor comprehension: The substrate of REM dreaming? *Dreaming, 20,* 227–247.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effect models (R package version 2-0). Retrieved from http://cran.r-project.org/package=lmerTest

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1980). *Multivariate analysis*. New York, NY: Academic Press.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York, NY: Guilford Press.

McGlone, M. S. (2011). Hyperbole, homunculi, and hindsight bias: An alternative evaluation of Conceptual Metaphor Theory. *Discourse Processes, 48,* 563–574.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Miller, G. A. (1979). Images and models, similes and metaphors. In A. Ortony (Ed.), *Metaphor and thought* (1st ed., pp. 202–250). Cambridge, UK: Cambridge University Press.

Nayak, N. P., & Gibbs, R. W. (1990). Conceptual knowledge in the interpretation of idioms. *Journal of Experimental Psychology: General, 119,* 315–330.

Ortony, A., Schallert, D., Reynolds, R., & Antos, S. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior, 17,* 465–477.

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

Roncero, C., & de Almeida, R. G. (2015). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods, 47,* 800–812. doi:10.3758/s13428-014-0502-y

Schmidt, G., & Seger, C. (2009). Neural correlates of metaphor processing: The roles of figurativeness, familiarity and difficulty. *Brain and Cognition, 71,* 375–386.

Sikos, L., Thibodeau, P., Strawser, C., & Durgin, H. (2013). *Advantages of extending versus mixing metaphors: An ERP study*. Article presented at the CUNY Conference on Human Sentence Processing, Columbia, SC.

Sikos, L., Thibodeau, P.H., Strawser, C., Klein, B.J., & Durgin, F.H. (2013). *Advantages of extending vs. mixing metaphors: An ERP study*. Poster presented at the CUNY Conference on Human Sentence Processing. Columbia, SC.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU* (Vol. 14). Philadelphia, PA: Benjamins.

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,* 521–540.

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

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive Psychology, 13,* 27–55.

Wolff, P., & Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cognitive Science, 35,* 1456–1488.

Xu, X. (2010). Interpreting metaphorical statements. *Journal of Pragmatics, 42,* 1622–1636.

Yang, F. G., Edens, J., Simpson, C., & Krawczyk, D. C. (2009). Differences in task demands influence the hemispheric lateralization and neural correlates of metaphor. *Brain and Language, 111,* 114–124.