Conventional metaphors elicit greater real-time engagement than literal paraphrases or concrete sentences

Serena K Mon[1], Mira Nencheva[1],

Francesca M M Citron[2], Casey Lew-Williams[1], Adele E Goldberg[1*]

[1] Princeton University, Princeton, NJ 08544

[2] Lancaster University, UK

Abstract

Conventional metaphors (e.g., *a firm grasp on an idea*) are extremely common. A possible explanation for their ubiquity is that they are more engaging, evoking more focused attention, than their literal paraphrases (e.g., *a good understanding of an idea*). To evaluate whether, when, and why this may be true, we created a new database of 180 English sentences consisting of conventional metaphors, literal paraphrases, and concrete descriptions (e.g., *a firm grip on a doorknob*). Extensive norming matched differences across sentence types in complexity, plausibility, emotional valence, intensity, and familiarity of the key phrases. Then, using pupillometry to study the time course of metaphor processing, we predicted that metaphors would elicit greater event-evoked pupil dilation compared to other sentence types. Results confirmed the predicted increase beginning at the onset of the key phrase and lasting seconds beyond the end of the sentence. When metaphorical and literal sentences were compared directly in survey data, participants judged metaphorical sentences to convey "richer meaning," but not more information. We conclude that conventional metaphors are more engaging than literal paraphrases or concrete sentences in a way that is irreducible to difficulty or ease, amount of information, short-term lexical access, or downstream inferences.

Conventional metaphors are extremely common in everyday language (Lakoff & Johnson, 1980; Lakoff, 1993; Littlemore, 2019). Although the specifics differ, conventional metaphors are attested across languages (e.g., Boers, 2003; Dobrovol'skij & Piirainen, 2005; Ibarretxe-Antuñano, 2005). In English, a student beginning a thesis may be nervous about *the road ahead*. She may *hit a rough patch*, which could throw her *off track*, and she may eventually *find her way* or *hit a dead end*. These expressions treat the student as a traveler, the events as locations in space, and difficulties as obstacles along the path, thereby mapping otherwise concrete concepts onto abstract interpretations. Since there often exist literal ways of expressing quite similar meanings, a question arises as to *why* conventional metaphors are so often used. In choosing one expression over others to express a particular message, a great many factors play a role, including relative accessibility and subtle differences in content (e.g., Goldberg, 2019). Several recent studies report a different type of factor that may play a role in the selection of metaphorical expressions: they may be more engaging than literal paraphrases.

The first hint that metaphorical language may be more engaging can be traced to a meta-analysis that compared neural activation for figurative language (including novel and conventional metaphors) with literal language across 22 fMRI studies (Bohrn, Altmann, & Jacobs, 2012). Among other differences, Bohrn et al. (2012) reported greater left amygdala activation for figurative language (see also Forgács et al., 2012 for a similar finding). Notably, the amygdala is recognized to be activated by emotional, salient, and evolutionarily relevant stimuli (Costafreda, Brammer, David, & Fu, 2008; Cunningham & Brosch, 2012; Garavan, Pendergrass, Ross, Stein, & Risinger, 2001; Hamann & Mao, 2002; Seeley et al., 2007). Relatedly, the amygdala has also been implicated in "motivated attention" or the detection of input that is relevant to task goals (see Schaefer & Gray 2007 for review). In order to remain neutral about whether the increased amygdala activation reported in previous work on conventional metaphors was due to emotional or cognitive engagement or some combination, we here interpret greater amygdala activation as indicating greater attention to task-relevant stimuli, or what we describe here as greater *engagement.*

Our current focus is on conventional (familiar) metaphorical expressions because we wish to better understand why they are so common in everyday language. By directly comparing conventional metaphors to carefully matched literal controls, several fMRI studies have confirmed greater amygdala activity. Citron and Goldberg (2014) reported greater amygdala activation for conventional metaphors related to taste, e.g., *a sweet compliment*, compared to literal paraphrases that differed only by a single word, i.e., *a nice compliment*. While the taste domain may be particularly engaging (Winter, 2016), increased amygdala activity has been replicated for a range of conventional metaphors beyond those referring to taste, both in sentences and in short stories (Citron, Güsten, Michaelis, & Goldberg, 2016; Citron, Michaelis, & Goldberg, 2020). Greater amygdala activation has also been found in studies comparing idioms to non-idiomatic sentences with matched emotion-related content (Citron, Cacciari, Funcke, Hsu, & Jacobs, 2019). Idioms are highly conventional and are typically based on metaphorical mappings (Gibbs, Bogdanovich, Sykes, & Barr, 1997). For instance, the idiom *let the cat out of the bag* treats a secret as something needing to be physically contained; someone with a secret might be admonished to *keep his trap shut*, with the understanding that if he doesn't *throw away the key*, he might *spill the beans* or *the tea.* This work had matched conventional and literal sentences on explicit ratings of emotional valence and arousal to ensure that the semantic content conveyed by the figurative expressions was not itself more emotionally-charged; sentences were also matched on the basis of familiarity, length and complexity to control for possible difference in cognitive demand.

In what follows, we report the first study of conventional metaphorical processing to use pupillometry, which affords a different measure of focused attention to task-relevant stimuli, or engagement, in an effort to better understand whether, when, and why conventional metaphors appear to evoke greater focused attention in the comprehender than literal paraphrases. Although pupillometry has been used for many years (Hess & Polt, 1964; Kahneman & Beatty, 1966), it has gained in popularity over the past decade due to the availability of sensitive eye trackers that make testing easier and more reliable (Bradley, Miccoli, Escrig, & Lang, 2008; Lavin, Martin, & Jubal, 2014). Pupil responses are well-suited to exploring our questions for several reasons. First, dilation is tightly coupled to the firing of neurons in the locus coeruleus (LC), which is anatomically and functionally connected to the amygdala (Sterpenich et al., 2006), a key brain area implicated in metaphor processing as just reviewed. The LC contains neurons synthesizing norepinephrine (NE), and the LC-NE system which directly mediates pupil dilation (Alnæs et al., 2014; Aston-Jones & Cohen, 2005; Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014) and is recognized to index focused attention and task engagement (Aston-Jones & Cohen, 2005; Corbetta, Patel, & Shulman, 2008; Eckstein, Guerra-Carrillo, Miller Singley, & Bunge, 2017; Laeng, Sirois, & Gredebäck, 2012; Sirois & Brisson, 2014). That is, when illumination is held constant, pupils dilate in response to increased activation of the sympathetic nervous system evoked by focused attention to task-relevant stimuli, or engagement.

Evidence that pupil dilation is evoked by focused attention to relevant stimuli comes from both the emotion and cognitive domains. More emotionally arousing stimuli, whether positive or negative, evoke greater pupil dilation compared to emotionally neutral ones in the non-verbal visual domain (Bradley et al., 2008; Kinner et al., 2017; van Steenbergen, Band, & Hommel, 2011), and in the non-verbal auditory domain (Partala & Surakka, 2003). Greater pupil dilation has been found when participants read negatively valenced sentences which they reported elicited higher ‘emotional impact’[1], in comparison to neutral sentences (Iacozza, Costa, & Duñabeitia, 2017). The same study also found greater pupil dilation in response to emotive words in participants’ native language than in their second language, which also suggests stronger affective engagement (Iacozza et al., 2017).

Other work has linked pupil responses to engagement in cognitive tasks that do not necessarily evoke emotion. Greater pupil dilation has been associated with processing of increasingly complex sentences (Just & Carpenter, 1993), less frequent words (Kuchinke, Võ, Hofmann, & Jacobs, 2007), incongruent or conflicting stimuli in the Stroop task (Laeng, Ørbo, Holmlund, & Miozzo, 2011), words that are more challenging to imagine when asked to do so, regardless of degree of pleasantness or unpleasantness (Paivio & Simpson, 1966), sentences with a grammar-prosody incongruency (Engelhardt, Ferreira, & Patsenko, 2010), increasingly degraded speech regardless of intelligibility (Winn, Edwards, & Litovsky, 2015), increasing memory load in a dual vs. single task paradigm (Karatekin, Couperus, & Marcus, 2004), and reward-prediction errors during a decision-making task (Lavin et al., 2014). While these examples are suggestive of greater cognitive effort, greater effort is not required for increased engagement. Indeed, dilation has been found to correlate with the perceived salience of stimuli that are neither more effortful to process nor more emotional (Liao, Kidani, Yoneya, Kashino, & Furukawa, 2016). Pupil size is also recognized to increase in response to previously encountered stimuli in visual or verbal recognition tasks, even though recognizable stimuli are, if anything, easier to process than new stimuli (Bradley & Lang, 2015; Kafkas & Montaldi, 2012; Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012; Võ et al., 2008).

[1] In particular, the rating scale ranged from 1 (low, neutral impact) to 7 (high, negative impact), so this was a combination of arousal and valence.

We here interpret pupil dilation as an index of engagement or focused attention to task-relevant stimuli, thereby remaining neutral about its relationship to emotional processing or increased cognitive demands. We revisit the issue of how best to characterize the effect in Study 3.

A key advantage of using pupillometry in the current context is that it allows us to examine the time course of metaphor processing, since changes in pupil size are measurable at time scales of 100 ms or less, while it takes several seconds for changes in the blood oxygen level-dependent (BOLD) signal in fMRI work to be detected. A better understanding of the time course of any effect of greater engagement can help narrow down its potential cause. For instance, conventional metaphors may result in more inferences than literal paraphrases (Thibodeau, Hendricks, & Boroditsky, 2017); the sentence, *The soccer player fell short of scoring enough goals,* suggests that the soccer player was responsible*,* while the literal paraphrase — *The soccer player wasn't able to score enough goals* — is agnostic about apportioning blame. If conventional metaphors regularly lead to more downstream inferences, we would expect to see a difference beginning sometime after initial exposure to the metaphorical phrase (Bott & Noveck, 2004; McElree, Traxler, Pickering, Seely, & Jackendoff, 2001), particularly if no context is provided to support faster processing of subtle inferences (Gildea & Glucksberg, 1983; Ortony, 1978).

Alternatively, it is possible that the effect is due to the concrete words contained in metaphors. EEG studies have found that concrete words used literally evoke a frontal negativity roughly 200-300 ms post stimulus compared to abstract words, which has been interpreted as a neural signature of concreteness (Barber, Otten, Kousta, & Vigliocco, 2013; Welcome, Paivio, McRae, & Joanisse, 2011). The possibility that greater engagement may be due to the concrete literal meanings of the words involved presupposes that the literal meanings are activated by conventional metaphors, and there is evidence that they are. Neuroimaging studies have found that conventional textural metaphors (e.g., *a rough problem*) elicit activation of somatosensory areas (Lacey, Stilla, &

Sathian, 2012), action metaphors (e.g., *destroy an argument*) elicit activation of motor areas (Desai, Binder, Conant, Mano, & Seidenberg, 2011; Samur, Lai, Hagoort, & Willems, 2015), conventional taste metaphors elicit activation of the gustatory cortex (Citron & Goldberg, 2014), and conventional metaphors related to the sense of smell activate olfaction-related regions (Pomp et al., 2018). In fact, the frontal negative ERP component evoked by concrete words in literal expressions has recently been found to be evoked by metaphorical language as well (Lai, Howerton, & Desai, 2019).[2] Also relevant is the fact that certain words are recognized to be semantically "richer" than other words; in particular, words that tend to be described by a longer list of features or words that appear in a broader range of contexts tend to be recognized and classified faster than other words (e.g., Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008). It is therefore important to compare the same or similar words used metaphorically and in concrete descriptions to see if any effect is specific to metaphor comprehension or is instead due to the choice of words used.

In order to address the possibility that greater engagement is due to the activation of concrete conceptual domains, we compare conventional metaphors with literal uses of sensorimotor-related words, as well as with literal paraphrases of the conventional metaphors. One fMRI study of words related to the sense of smell included all three types of sentences, but this study did not find evidence of greater amygdala activation for metaphorical language compared to literal paraphrases

[2] Other ERP work on metaphorical processing has investigated the existence and timing of the N400 component, which is a measure of semantic access and integration rather than engagement (Bambini, Ghio, Moro, & Schumacher, 2013; Coulson & Van Petten, 2002; Iakimova, Passerieux, Laurent, & Hardy-Bayle, 2005; Lai & Curran, 2013; Lai, Curran, & Menn, 2009). ERP studies to date have not addressed whether conventional metaphors are more engaging.

or concrete descriptions, for reasons potentially related to the specific source domain of olfaction (Pomp et al., 2018). In particular, the amygdala, together with the piriform cortex, constitutes the primary olfactory cortex, and projects to the orbitofrontal cortex (OFC) which is considered the secondary olfactory cortex. The strong relevance of the amygdala for olfaction may have masked any additional involvement due to engagement from reading figurative expressions. Thus a comparison of conventional metaphors with both paraphrases and concrete descriptions warrants investigation.

In what follows, we report a preregistered pupillometry experiment comparing undergraduate participants' pupil dilation in response to hearing sentences containing conventional metaphors, literal paraphrases, and sentences containing words from the same concrete domains as the conventional metaphor but used literally. Greater engagement, defined as focused attention to task-relevant stimuli, is operationalized as greater pupil dilation in comparison to control sentences. If metaphors evoke greater pupil dilation than both literal paraphrases and concrete sentences, it will support the claim that sentences containing conventional metaphors are more engaging. On the other hand, if the concrete sentences evoke greater or equivalent pupil dilation as the metaphorical sentences, it will suggest that greater concreteness (more sensorimotor information) drives greater engagement.

Since changes in pupil size are detectable at fine time scales (100 ms or less), the current work also allows us to investigate *when* any differences in pupil dilation occur and how long they last. If a difference is only evident downstream, it will suggest that metaphors evoke distinct or additional inferences than other types of sentences. If a difference is detectable early and is short-lived, it will suggest the effect is related to lexical access; moreover, if a difference is evident at the key phrase in the case of both metaphorical expressions and concrete expressions, it would support the idea that increased engagement is due to a concreteness effect, rather than metaphoricity. Finally, if a difference is detectable early and is long-lasting, it will suggest that the meaning of metaphorical expressions evokes more focused attention which begins immediately and persists throughout its integration into the overall interpretation of the sentence.

To create a database of stimuli, we first conducted an extensive norming study that yielded 60 sentence triples (see Table 1 for examples). Each triple contains a sentence with a conventional metaphor key phrase (M); a literal paraphrase of the key phrase (L); and a sentence using the same or similar words as the key phrase used literally to describe a concrete scene (C). The sentences were matched across conditions on explicit ratings of complexity, plausibility, emotional valence and arousal, and familiarity of the key phrases. We also collected gradient measures of metaphoricity and imageability as well as semantic similarity ratings for the metaphorical and literal phrase based on human ratings, and ratings using Latent Semantic Analysis (Dumais, 2004).

Table 1: Examples of stimuli sentence triples with key phrase underlined.

	Metaphorical Sentence (M)	**Literal Sentence (L)**	**Concrete Sentence (C)**
1	The contestant's bitter comments disgusted the judges.	The contestant's derisive comments offended the judges.	The contestant's bitter drink disgusted the judges.
2	The matter was out of the editor's hands after she sent the text.	The matter was out of the editor's control after she sent the text.	The phone fell out of the editor's hands after she sent the text.
3	The chef acquired more confidence after the positive reviews.	The chef felt more self-assured after the positive reviews.	The chef acquired more customers after the positive reviews.

4	The band couldn't hide from their past.	The band couldn't avoid their past.	The band couldn't hide from the press.
5	The celebrity's story was distorted by the tabloids.	The celebrity's story was misrepresented by the tabloids.	The celebrity's voice was distorted by the special effects.

Finally, in Study 3, we report data from preregistered follow-up surveys which were conducted in order to assess whether any difference in processing may be recognized by the listener as related to emotional processing or cognitive processing. For this we asked three new groups of participants to decide which sentence from each metaphorical and literal sentence-pair conveyed “richer meaning,” “more emotion,” or “more information.”

Preregistration and Open Science

For Studies 1 and 2, norming criterion, exclusion criteria, number of participants, and main analyses were preregistered at AsPredicted.org http://aspredicted.org/blind.php?x=ae2ki4 (included in SI). The full dataset of stimuli (60 sentence triples) along with the results of norming for each sentence are publicly available at https://osf.io/5ywfn/?view_only=caa6f32c944a43418f2193d27dfea874, as are the full results and analyses: https://osf.io/dsn9w/?view_only=529c69a39b624dca8d2712847bf176e2. For Study 3, the experiment design and hypotheses related to the emotionality and informativity surveys (https://osf.io/x3da5/wiki/home/?view_only=15b7e7f996df42b2805094c18ae93ca0) and richer meaning surveys (https://osf.io/46zge/) were preregistered at Open Science Framework.

Study 1: Norming Study

Method

Participants

A total of 1,021 native English speakers, recruited through the Cloud Research platform (Litman, Robinson, & Abberbock, 2017), took part in the norming task and were paid for their time, with groups of 51-62 unique participants assigned to each survey.

Procedure

An initial set of 70 sentence triples consisted of a sentence for each of the following conditions: Metaphor (M, e.g., *The actor gave his co-star a sweet compliment.*), Literal (L, e.g., *The actor gave his co-star a kind compliment.*), and Concrete (C, e.g., *The actor gave his co-star a sweet candy.*). See Table 1 for examples. Each sentence in the M condition contained a key phrase that corresponded to the conventional metaphor (e.g., *sweet compliment*). The corresponding phrase in the L condition was intended to express the same meaning literally (e.g., *kind compliment*). The corresponding phrase in the C condition was intended to evoke the same sensory information as in the M condition (e.g., *sweet candy*).

Each participant judged one sentence from each of the 70 original hand-created triples on a single gradient scale. That is, judgments were collected separately for metaphoricity, concreteness (imageability), familiarity (subjective ratings of frequency), complexity, plausibility, emotional valence, emotional intensity (arousal). Gradient judgments of semantic similarity between two sentences of each triple were also collected (M & L; M & C). The variables of interest were metaphoricity and concreteness: we intended that the metaphorical sentences should be rated the most metaphorical, and the concrete sentences should be rated the most imageable. Imageability ratings were used as a proxy for concreteness since the latter have been found to show a more dichotomous trend (Kousta et al., 2011); we subsequently collected concreteness ratings as well, and confirm that concreteness and imageability ratings are strongly correlated for our stimuli (*r* = .83) (see also Winter et al. 2017) . We also confirmed that the metaphorical sentences and their literal paraphrases are highly similar in meaning on the basis of both human ratings and objective Latent Semantic Analysis comparisons (Dumais, 2004).

The reason to match conditions on complexity, plausibility and familiarity was to control for any differences in effort or difficulty. That is, if any condition included language that was more complex, less plausible or less familiar, we would expect that condition to require more effort. We asked participants to rate how familiar they felt the key phrases to be, rather than relying on corpus frequencies, because there is no straightforward way to identify metaphorical uses of words automatically. Subjective judgments of familiarity are known to correlate well with corpus frequency, particularly in spoken language (Tanaka-Ishii & Terada, 2011), and all of our sentences were presented auditorily.

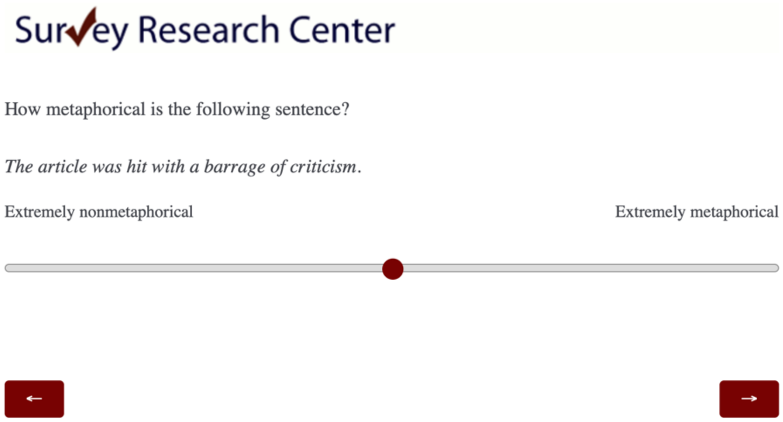
We included emotional valence and arousal because these factors are recognized to increase engagement for both metaphorical and non-metaphorical language. Since our goal is to determine whether language including conventional metaphors is more engaging because it is metaphorical, independently of whether the content expressed is explicitly emotional, we matched conditions for these factors as well (see also Citron & Goldberg, 2014; Citron et al., 2016; Citron et al., 2020).

Norming was conducted using Qualtrics with separate groups of participants recruited from Amazon Mechanical Turk through Cloud Research, a prescreening platform (Litman et al., 2017). The survey was designed so that each participant rated a set of sentences on one feature using a sliding scale (Figure 1). For all of the features, except familiarity and semantic similarity, participants rated the extent that one sentence from each triple contained the feature of interest (which sentence from each triple was counterbalanced across participants). In the case of familiarity, participants were presented with the sentences, but with the corresponding key phrase capitalized, and were asked to rate how often they had come across the capitalized phrase. For semantic similarity, participants were presented with two sentences at a time and asked to rate how similar they were in meaning.

At the beginning of the survey, participants read a definition of the feature and an example sentence exhibiting the assigned feature (Table 2). Participants then practiced rating a new sentence (or sentence-pair in the case of semantic similarity) to test that they understood the definition. Feedback for this practice sentence was given if a participant did not rate in the expected direction (e.g., incorrectly rating *The young girl was a budding programmer.* as "extremely nonmetaphorical"), but no other feedback was given during the task. Participants who rated sentences on familiarity and complexity features were not given feedback because these features were assumed to vary more subjectively.

Following the practice sentence, participants were randomly assigned to one of three lists for the main rating portion. In the initial round of norming, a total of 522 online participants were recruited and a set of 70 sentence triples was rated on each variable. Each list contained an equal number of sentences from each condition (23-24 sentences were presented for each condition). For semantic similarity, there were two possible lists, each containing half of the M&L sentence-pairs and half of the M&C sentence-pairs. For all of the features, sentence order was randomized for each participant and no feedback was provided.

Thirty-three of the sentence triples were revised and rated in a second round of norming, with a new group of 499 online participants. The final norming results were aggregated over the 37 non-revised sentence triples and the 33 revised sentence triples in order to select the final 60. A total of 9 non-native speakers also participated in the norming study but their ratings were not included in subsequent data analyses.



Survey Research Center

How metaphorical is the following sentence?

The article was hit with a barrage of criticism.

Extremely nonmetaphorical

Extremely metaphorical

Figure 1: Example of a sentence and sliding scale presented to participants assigned to rate the metaphoricity feature.

Table 2: Norming features with definition and example sentence presented during the norming survey.

Feature	Definition and Example Sentence
Metaphoricity	Words are not always used literally. For example, the following sentence is somewhat metaphorical: *The website's rules were tightened to reduce profanity.* Notice that rules can't be literally tightened or loosened. Instead, we mean that the rules were made stricter.
Imageability (Concreteness)	Some sentences describe a scene that is easy to imagine. For example, the following sentences are easy to imagine: *The maple trees in front of the house were a dazzling array of red, gold, and yellow.* *The alarm rang very loudly and the father jumped out of bed.* In comparison, the sentence below is more difficult to imagine: *The thorough considerations led to a wise decision.*

Emotional Valence	Some sentences describe positive or negative scenarios. For example, the following sentence describes a negative event: *The child lost his new toy on the subway.* In contrast, the following sentence describes a positive event: *The child won a new toy at the carnival.*
Emotional Intensity (Arousal)	Some sentences describe more emotionally intense scenarios than others. For example, the following sentences describe an intense event: *The tightrope walker slipped while practicing without a net.* In contrast, the following sentence describes an event that is not intense: *The committee's thorough decision was published in the newspaper.*
Familiarity	Some phrases are used more frequently in everyday speech than others. For example, the following sentence uses *large* to describe a room: *The host's voice echoed in the large room.* You may come across the phrase *the large room* more often than the phrase *the capacious room.*
Complexity	Some sentences are easier to read than others. For example, the following sentence may be difficult to read: *The advertising firm managed to make a prototype that displayed all of the holograms before the deadline.* In comparison, the following sentence may be easier to read: *The company managed to finish their projects before the deadline.*
Plausibility	Some sentences describe more natural or plausible events than others. For example, the following sentence describes an implausible event: *This morning the weatherman saw a pig flying on a cloud.*
Semantic Similarity	There are different ways to express the same meaning. For example, the following two sentences have a very similar meaning: *She took a grueling hike to reach the top of the mountain.* *She made a difficult hike to reach the top of the mountain.*

Although familiarity, complexity, plausibility, valence, and intensity were matched across Metaphor, Literal and Concrete conditions, to be conservative, we included them as continuous factors in the analyses anyway as described below.

Results

Descriptive statistics as boxplots from the norming data are provided in Figures 2 and 3 and the statistical comparisons of each feature across condition are provided in Table 3.

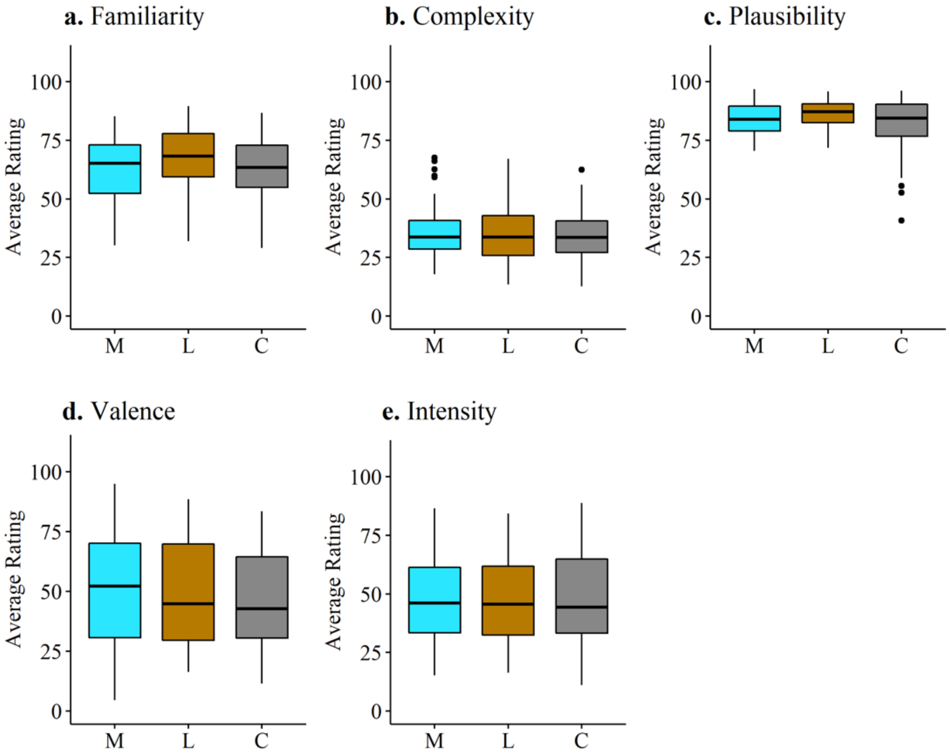


Figure 2: Boxplots for Metaphor (M), Literal Paraphrase (L) and Concrete (C) conditions on each of the 5 matched variables across 60 sentence triples (60 sentences for each condition): a) familiarity, b) complexity, c) plausibility, d) valence, and e) intensity. All features were rated on a scale from 0 (not at all complex, not at all plausible, not at all intense, etc.) to 100 (extremely complex, extremely plausible, extremely intense, etc.).

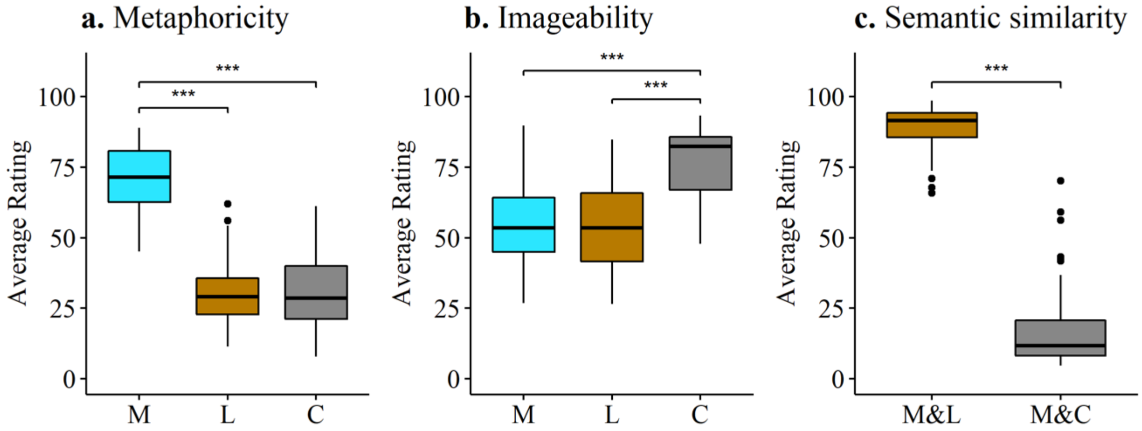


Figure 3: Boxplots on the variables intended to differ across conditions: a) metaphoricity, b) imageability/concreteness, c) semantic similarity, for each condition (Metaphor, Literal Paraphrase, and Concrete) across the 60 sentence triples (60 sentences for each condition). All features were rated on a scale from 0 (extremely nonmetaphorical, extremely difficult to imagine, not at all similar in meaning) to 100 (extremely metaphorical, extremely easy to imagine, extremely similar in meaning).

Table 3: Test statistics from the non-parametric Mann-Whitney U test comparing the norming rating distributions between pairs of conditions. As intended, conditions were matched on explicit ratings of emotional valence, intensity, familiarity, complexity, plausibility. M and L were highly similar; M was more metaphorical than L which did not differ from C; and C was easier to imagine than M which did not differ from L.

Feature	M vs. L Comparison	M vs. C Comparison	L & C Comparison
Metaphoricity	$W = 3558$, $p < .0001$ ***	$W = 3548$, $p < .0001$ ***	$W = 1813$, $p = .95$
Imageability	$W = 1815.5$, $p = .94$	$W = 433$, $p < .0001$ ***	$W = 507$, $p < .0001$ ***

Emotional Valence	$W = 1819$, $p = .92$	$W = 1958$, $p = .41$	$W = 1930.5$, $p = .50$
Emotional Intensity	$W = 1902.5$, $p = .59$	$W = 1789$, $p = .96$	$W = 1701.5$, $p = .61$
Familiarity[3]	$W = 1480$, $p = .09$	$W = 1886$, $p = .65$	$W = 2173.5$, $p = .05$
Complexity	$W = 1846.5$, $p = .81$	$W = 1891$, $p = .64$	$W = 1877.5$, $p = .69$
Plausibility	$W = 1478$, $p = .09$	$W = 1887$, $p = .65$	$W = 2144$, $p = .07$
Semantic Similarity	**M&L vs M&C Comparison** *Human judgments:* $W = 3598$, $p < 2.2\text{e-}16$ *** *Latent Semantic Analysis:* $t(119) = 5.28$, $p < .0001$ ***		

We use imageability as a proxy for concreteness although the two are not identical. In response to a reviewer's suggestion, we conducted an additional survey in which we explicitly asked a new group of participants ($N = 53$) for judgments of how 'concrete' each sentence was. In the current stimuli, concreteness ratings were strongly correlated with ratings based on imageability ($r = .83$). Our analyses include the ratings described in the preregistration (querying 'imageability.')

[3] Familiarity approached significance in comparisons between M&L (.09) and L&C (.05). However, since L&C conditions were the most divergent in terms of mean familiarity (67 vs. 62), and M fell in between (63), familiarity ratings are unlikely to be responsible for the hypothesized result, namely that L&C should pattern alike in terms of pupil dilation, while M is predicted to differ. A similar pattern is evident in judgments of plausibility. To be conservative, as described below, we included familiarity, plausibility and the other factors that matched across conditions as continuous fixed effects in the models predicting pupil dilation in Study 2 (see Figure 3 for factors). Neither familiarity nor plausibility were significant predictor of pupil dilation in any model.

Discussion

The norming study identified sentence triples that form a database of 60 sentences containing conventional metaphors, 60 literal paraphrases and 60 concrete descriptions. These sentence triples met the following criteria based on the aggregated norming:

- All conditions were matched on explicit ratings of emotional valence, emotional arousal, complexity, familiarity, plausibility.
- M & L were rated as highly semantically similar both in a human rating task and according to Latent Semantic Analysis.
- As planned, sentences in the Metaphor condition were rated as significantly more metaphorical than sentences in either of the other two conditions.
- Also as planned, sentences in the Concrete condition were rated to be more imageable than sentences in the Metaphor or Literal conditions.

The normed stimuli were used in the main pupillometry study and in three final surveys, as detailed below.

Study 2: Pupillometry

Method

Participants

Sixty-nine (38 women, 31 men*, M* = 19.71 years, *SD* = 2.09) native English speakers were brought to the lab to participate in the experiment, recruited from the Princeton Psychology Subject Pool and Princeton Paid Research Pool and compensated with either course credit or $8. No data was collected for 3 participants due to technical difficulties. We thus collected data for the preregistered target number of 66 participants, which was based on a power analysis of results from separate preliminary pilot data (*N* = 23) which is not included, using 0.80 power for a 2-tailed t-test with alpha of 0.05. Five participants with lower than 70% accuracy on an attention check were excluded based on preregistered exclusion criteria. Data from a total of 61 participants were analyzed. The protocol was approved by the Princeton IRB (#4951).

Sentence Recordings

All 180 target sentences (60 triples), 2 practice sentences, and 12 filler sentences were recorded using the software Praat and all sentence recordings were normalized to an average intensity of 60 dB. The duration of sentence recordings (without silence) ranged from 2.02 to 5.39 sec. Two seconds of silence were concatenated to the end of each sentence to enable analysis of pupil size changes in the moments following each sentence. An additional (jittered) 250 to 750 ms interstimulus interval was not analysed.

Fillers used for comprehension task (attention check)

In order to ensure that participants paid attention and did their best to interpret each sentence, we randomly interspersed 12 filler sentences, which were each immediately followed by a multiple-choice comprehension question. Each condition (M, L, & C) was assigned 4 filler sentences. Comprehension questions were non-trivial as they were designed to encourage participants to comprehend each sentence (see Table 4). While the sentences were all presented auditorily, comprehension questions were presented on a screen, with the order of answer options randomized. Text for the comprehension questions, instruction slides, and the fixation cross was adjusted to be isoluminant.

Table 4: Examples of filler sentences and comprehension questions.

Condition	**Filler**	**Comprehension Question**	**Correct Answer**	**Incorrect Choices**
C	The airplane was hit by a barrage of bullets.	What most likely happened to the airplane?	It crashed	It arrived late to the airport It flew again the next day It got lost
L	The article received a great deal of criticism.	Who is most likely receiving the criticism?	the writer	the intern the advertiser the graphic designer
M	The couple's relationship was spinning its wheels, not going anywhere.	How was the couple likely feeling?	discouraged	content protective astonished

Procedure

After providing written consent, participants were asked to sit in front of a computer monitor and EyeLink 1000 Plus eye tracker, which was calibrated for each participant. Participants were told that they would listen to 72 sentences and answer randomly interspersed comprehension questions. They were asked to keep their eyes on the fixation cross while listening to the sentences and to respond to comprehension questions using the keyboard.

Practice trials consisted of two sentences followed by one comprehension question. After that, participants listened to one of three lists of sentences. Each list contained one sentence from each triple (20 from each condition, counterbalanced across participants), randomly ordered for each participant. Each trial began with an interstimulus interval (ISI) randomly jittered between 250 and 750 ms (*M* = 500 ms) followed by a sentence presented auditorily. Filler sentences/comprehension questions occurred randomly after every 2-8 target trials (see Figure 2). Pupil size data were recorded at 500 Hz during each trial.

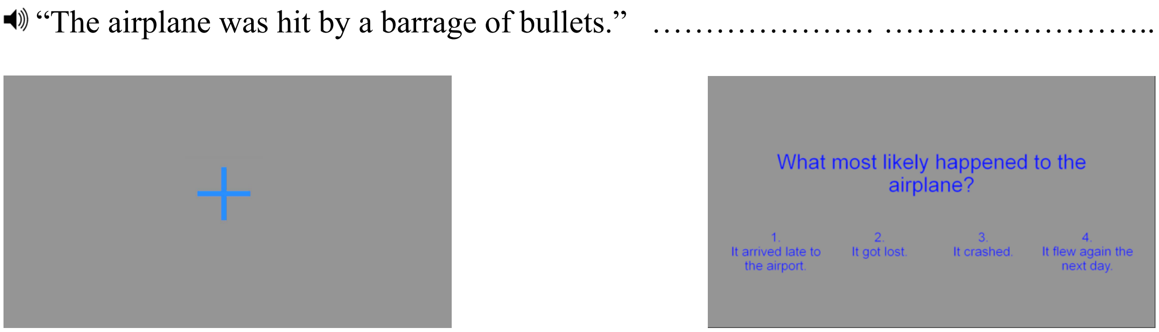

“The airplane was hit by a barrage of bullets.” ………………… ……………………..
What most likely happened to the airplane?
1. It arrived late to the airport.
2. It got lost.
3. It crashed.
4. It flew again the next day.

Figure 4: Example of filler sentence and comprehension question pair presentation.

Pupillometry Preprocessing

Blinks and other artifacts were removed following procedures from Merritt, Keegan, and Mercer (1994) and Nencheva, Piazza, and Lew-Williams (2020). A baseline for each sentence was calculated using the average pupil size during the first 100 ms of the onset of the key phrase. This baseline was chosen to account for pupil size variations due to the location of the key phrase in the sentence, which varied to some extent across different sentence triples so they could not be anticipated. Relative pupil dilation was calculated by dividing the pupil dilation data at each time point for each sentence by the corresponding baseline. Trials with missing data during the baseline or spanning more than half of the trial were excluded from subsequent analyses. Stineman interpolation was used to estimate missing data over durations shorter than 100 ms.

Pupillometry Data Analysis

The time course of average relative pupil dilation (% compared to baseline) for M, L, and C conditions is represented in Figure 5. Data files were analyzed over four intervals relative to each sentence's key phrase: the portion of the sentence immediately preceding it (sentence onset; *M* = 0.88 sec, *SD* = 0.46), the key phrase (*M* = 1.40 sec, *SD* = 0.51), the remaining portion of the sentence following it (rest of sentence; *M* = 1.00 sec, *SD* = 0.82), and the first 2 seconds of silence after the sentence (divided into three equal durations of 0.67 sec each). Relative pupil dilation was calculated by dividing pupil dilation by the average pupil size during the baseline (first 100 ms of the key phrase for each sentence), and the average pupil size was calculated for each interval. Error bars represent standard errors of the mean. An additional ISI was randomly jittered between 250 to 750 ms, which provided time for pupil size to reset after each trial, and was not included in the analysis.

Results and Analyses

We tested the effect of Metaphoricity at the key phrase (Models I and II) and over the subsequent period that spanned 2 seconds beyond each stimulus sentence (Models III and IV). Because Metaphoricity judgments lie on a continuum, we performed analyses in two ways: with either Condition as a categorical variable (Models I and III), or with Metaphoricity as a gradient variable (Models II and IV). In all models, by-participant and by-item intercepts were included as random effects. Random slopes were omitted to compare models with and without factors of interest. Even though conditions were matched on the normed values across conditions, we conservatively also included complexity, familiarity, emotional intensity, valence, and plausibility ratings as factors in testing the role of Condition, as well as in testing the gradient Metaphoricity variable. All norming ratings were standardized before being included in the models. For full results of models I-IV see Appendix A.

Figure 5 displays the relative pupil size across the duration of the trial for each condition: sentences containing metaphors (M), literal paraphrases (L), and concrete descriptions (C).

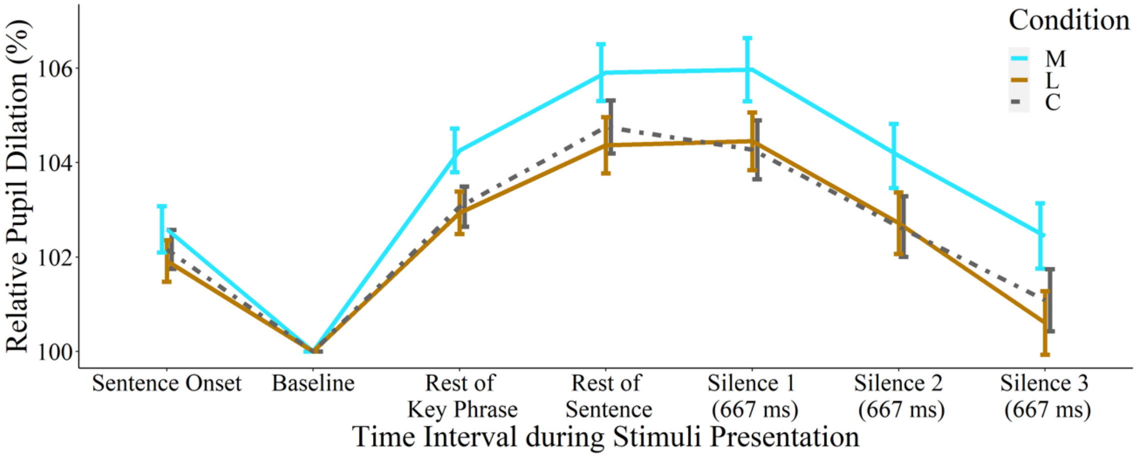


Figure 5: Time course of average relative pupil dilation for Metaphor (M), Literal (L), and Concrete (C) conditions during: sentence onset, baseline (first 100 ms of the key phrase), key phrase after baseline, the rest of sentence, and 2 seconds of silence before additional jittered ISIs. The key comparison is the degree of pupil dilation between conditions.

Data analyses focused on examining whether conventional metaphors elicited greater real-time engagement (operationalized as greater pupil dilation) compared to literal paraphrases and concrete sentences. Linear mixed models were used to test for an effect of condition on relative pupil dilation both during the key phrase (e.g., *sweet compliment / kind compliment / sweet candy*) and across the full trial, using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015).

Model I: Effect of Condition (M, L vs C), a categorical variable, at the key phrase

As predicted, the key phrase evoked more dilation in the Metaphor condition than the Concrete condition, which served as the reference condition, (β_M = 1.44, *SE=.63*, $p = 0.023$). On the other hand, the L condition was indistinguishable from the C condition (β_L = 0.23, *SE* = .65, p = 0.725). A comparison of models with and without the categorical variable condition as a fixed factor also show a significant advantage of including condition: $\chi2(2) = 6.01$, $p < 0.05$.

Model II: Effect of Metaphoricity, as continuous variable, at the key phrase

A different way to investigate the same factor of interest, Metaphoricity, is to test its role as a continuous factor instead of as a categorical difference between conditions. To do this, as preregistered, we included the standardized ratings collected in the norming study on the degree of metaphoricity for individual items. All other matched variables were also included as fixed factors. By-item and by-participant random effects were again included. The fixed effect of Metaphoricity was significant (β_M = 0.58, *SE* = .26, $p = 0.028$), indicating that pupil dilation at the key phrase is positively correlated with Metaphoricity.

In two additional analyses (Models III and IV), we investigate pupil dilation after the key phrase until the end of the trial, which includes 2 seconds of silence, before the ISI consisting of an additional 250-750 ms. of jittered silence. These additional models then include no overlap in data from Models I and II. Testing the subsequent period is critical to determine whether the effect of metaphoricity is long-lasting.

Model III: Effect of Condition after key phrase to beyond end of sentence

In order to determine whether pupil dilation remained greater in the M condition in comparison to the L and C conditions after the key phrase, we treated time as a random factor for the 4 time points following the key phrase (rest of sentence, and three silence intervals of 667 ms each). Because the data is continuous, we correct for multiple comparisons, requiring $p < 0.0125$ for an effect to be significant. By-participant and by-item random intercepts were included. We again conservatively included all of the matched variables in the analysis. The resulting model shows a significant increase in dilation for the M condition in comparison to the reference (Concrete) condition, with Bonferroni correction applied to the p value ($\beta_M = 1.43$, SE = 0.453, $p < 0.0015$). The L condition pattern was indistinguishable from the C condition ($\beta_L = 0.32$, $SE = 0.472$, $p = 0.505$). The model with the categorical condition variable is a significantly better fit than the model without it: ($\chi2(2) = 14.72$, $p = .003$).

Model IV: Effect of Metaphoricity, continuous variable, after key phrase to beyond end of sentence

A final model examined the factor of interest, Metaphoricity as a gradient factor (as in Model II), after the key phrase for the rest of the trial (as in Model III). As in other models, all of the matched variables were included as additional fixed factors, along with random intercepts for participants and items. The resulting model shows a significant increase in dilation as metaphoricity increases, with Bonferroni correction applied (requiring $p < .0125$) ($\beta_M = .50$, $SE = 0.194$, $p = 0.0097$).

Exploratory analyses of other fixed factors

In order to ascertain whether imageability as a (negative) influence was responsible for the difference in dilation rather than metaphoricity, we considered models in which the gradient imageability factor replaced Condition (in Model I) or gradient metaphoricity (in Model II) and asked whether imageability improved model fit. Model comparisons confirm it did not ($\chi2(1)$ = .1604, *p* = .688). We similarly checked whether imageability improved model fit in the models that considered dilation after the key phrase, by again substituting imageability for Condition (Model III) or gradient metaphoricity (Model IV). Model comparison again confirms that imageability is not responsible for the difference in pupil dilation ($\chi2(1)$ = 1.67, *p* = .196).

Recall that the factors matched across conditions were included in all analyses, even when testing the categorical variable, Condition (Models I and III), for which they were matched across conditions. As metaphoricity was the preregistered factor of interest (as Condition or gradient factor), any analysis of the additional five factors--complexity, familiarity, intensity, valence, or plausibility--is exploratory and requires correction for multiple comparisons in all models. Intensity was the only normed factor aside from metaphoricity to approach significance in more than a single model: specifically in Model I: *β* = 0.79, SE = .299, *p* = 0.008 and Model II: *β* = 0.78, *SE* = .299, *p* = 0.009; intensity did not correlate strongly with metaphoricity (as condition or gradient factor) in either model Model I (-.05) or Model II (-.08). Therefore metaphoricity and intensity appear to be independent influences on pupil dilation at the key phrase. Intensity did not show a significant effect in either Model III or IV after correction (see Appendix for full models). To summarize, the only factor to even approach a significant effect on dilation in all four models is metaphoricity.

Post hoc norming and analyses of the intensity of the key phrases in isolation

Recall that we preregistered and included norming data of emotional intensity of full sentence stimuli, as has been done in relevant prior work (e.g., Citron et al. 2016; Citron, Lee, &

Michaelis 2020; Müller, Nagels & Kauschke, 2021). At the suggestion of a reviewer, we additionally performed post hoc norming of the key phrases *in isolation,* since intensity was a significant factor in analyses of dilation immediately at the key phrase. For the Intensity at the Key Phrase (IKP) norming, we aimed to collect judgments from 60 participants as was done in Study 1 (N = 51-62 for each norming task). Fifty-nine people completed the survey. Analysis revealed that judgments of IKP correlated well with judgments that had been collected for intensity of the overall sentence (Pearson's r = .71). Perhaps for that reason, adding intensity at the key phrase did not improve the fit of any model Model I_{IKP}: ($\chi2(1) = .667$, $p = .41$); Model II_{IKP}: $\chi2(1) = .484$, $p = .49$); Model III_{IKP}: $\chi2(1) = .128$, p = .72); Model IV_{IKP}: ($\chi2(1) = .144$, $p = .70$). For two additional sets of full models that include IKP as a fixed factor (with and without the original intensity), see Supplementary Information. Key effects are largely replicated in these additional exploratory analyses, lending further support to the claim that metaphoricity leads to an increase in pupil dilation.

Discussion

In the first application of pupillometry to investigate metaphor processing, we find that participants' pupils dilate more when they passively listen to conventional metaphors than when they listen to carefully matched literal paraphrases or concrete descriptions. Since greater pupil dilation is evoked by increased focused attention or task engagement, the present results are consistent with previous fMRI research that had likewise argued that metaphors are more engaging than literal paraphrases on the basis of greater amygdala activation (Citron & Goldberg, 2014; Citron et al., 2016; Citron et al., 2020; Citron et al., 2019). It is unlikely that the metaphorical sentences were more difficult or effortful, since, as in previous work, the current stimuli were matched for familiarity, complexity, and plausibility, and none of these factors influenced pupil dilation in the current experiment.

By using pupillometry, we were able to determine that the greater engagement occurs as soon as the metaphorical phrase is heard and persists well beyond the end of the sentence. The immediacy of the effect undermines the idea that conventional metaphors are more engaging due to delayed inferences. The fact that the effect remains beyond the sentence argues that it is not simply due to a difference in lexical access. By comparing responses evoked by conventional metaphorical sentences to those evoked by literal paraphrases (which share the same general meaning) and by concrete descriptions (which share the same sensory information), we have further demonstrated that the greater engagement is due to metaphoricity rather than the general content or concreteness of the stimuli.

Exploratory analyses considered possible effects of the normed factors on pupil dilation in response to the current stimuli, though they had been matched across conditions (recall Study 1). Intensity (of the full sentence) was the only factor to show a reliable influence on dilation, and it only did so at the key phrase. At the suggestion of a reviewer, we performed additional

post hoc norming using the same method as in Study 1, but aimed to determine how intense the key phrases were judged to be in isolation. The norming confirmed that intensity at the key phrase correlated strongly with judgments of intensity of the overall sentence; also, including intensity at the key phrase as a fixed factor did not approach significance in any model. Finally, two sets of analyses (8 full models in total) are provided in the Supplementary Information which include the ad hoc factor of intensity at the key phrase (in addition to, or instead of, intensity at the sentence level). These additional exploratory models are consistent with the claim that metaphoricity predicts increased pupil dilation.

To summarize, the pupillometry study shows that sentences containing conventional metaphors evoke pupil dilation in comparison to both literal paraphrases and concrete descriptions, and the effect is not attributable to familiarity, complexity, valence, plausibility or intensity. We take the immediate and sustained effect of metaphoricity compared to concrete and literal sentences to confirm that conventional metaphors are more engaging, not because they use more imageable words, and not because they evoke more downstream inferences. Instead, we interpret the greater engagement to imply that conventional metaphors are directly associated with meaning that evokes increased attention immediately and throughout the interpretation of the entire sentence. In an effort to characterize whether or how people perceive the greater engagement, we performed a series of surveys in Study 3.

Study 3: Comparing metaphorical and literal sentences directly

Following previous work on stimuli-evoked pupil dilation, we interpret the heightened dilation established in Study 2 as indexing greater focused attention to task-relevant stimuli or greater engagement (Aston-Jones & Cohen, 2005; Corbetta et al., 2008; Eckstein et al., 2017; Laeng et al., 2012; Sirois & Brisson, 2014). This characterization is intended to be neutral with regard to a possible link to emotional or cognitive processing. Citron and Goldberg (2014) had suggested that metaphorical sentences are more *emotionally* engaging, as greater amygdala activity in that study was interpreted as a signature of greater emotion processing. Yet amygdala activity, like pupil dilation, is sensitive to focused attention that is attributable to emotion processing *or* to cognitive processing (Schaefer & Gray, 2007). And while we cannot attribute greater pupil dilation to greater difficulty or to an increase in delayed inferences in the current study, it remains possible that metaphors conventionally evoke more information (Thibodeau & Boroditsky, 2011; Thibodeau et al., 2017).

Therefore, in a final set of surveys, we aim to clarify more specifically what subjective quality of conventional metaphors may result in greater engagement. In particular, we asked three new groups of participants to compare Metaphorical and Literal pairs of sentences and determine which member of each pair conveyed more information to them, evoked more emotion in them, or conveyed "richer meaning" to them, respectively. The last description is motivated by Colston (2015)'s characterization of metaphors as providing "enhanced semantic meaning" (p. 73), or a means to "enrich the meaning being expressed" (p. 19). We take "richer meaning," like greater engagement, to apply to emotional or informational content without disentangling a potential distinction.

Because we aimed to compare speakers' intuitions about subtle differences in meaning or evoked emotion in the comprehender, in this study we asked participants to compare M and L sentences of each pair to one another, as direct comparisons of stimuli that are closely aligned tend to make any differences more salient (e.g., Gentner & Markman, 1994). Recall the norming study had already confirmed that these pairs were judged to be highly similar in meaning to each other and distinct from the C sentences on the basis of human judgments and objective Latent Semantic Analysis (Dumais, 2004).

Method

Participants

A total of 358 new participants from AMT via Cloud Research were recruited and paid for their participation.

Procedure

Participants were randomly assigned to one of the 3 surveys (*N* = 118, for emotionality, and *N* = 120 for each of the other two surveys). Surveys asked participants to decide which member of a M-L pair of sentences conveyed more information, evoked more emotion, or conveyed richer meaning, respectively, to them, the reader. Two or three practice trials were provided with feedback (see Table 5 for instructions and practice trials). Each survey consisted of a 2-alternative forced choice (2AFC) task in which participants compared a random subset of the 20 metaphorical and literal sentence-pairs used in the pupillometry task. Which subset of the full 60 M-L pairs was included varied randomly across participants. The order of presentation of M and L was randomized on each trial for each participant.

At the end of each survey, we asked whether participants noticed that one sentence of each pair contained a metaphor in order to determine whether explicit awareness of metaphors might lead to strategic responses.

Table 5: Instructions and practice trials for 2AFC surveys in Study 3.

Survey	**Instructions and practice trials** [with feedback: correct response boldfaced below]
Richer meaning	Please decide which of the two sentences being compared seems richer in meaning to you. **I glimpsed the sailboat on the waves**. I saw the boat on the water. I'll help you. **I got you.** **She's over the hill.** She's older.

More informative	Please decide which of the two sentences being compared seems more informative to you. **Sam got lost in the woods.** Sam walked in the woods. Keisha did well on the exam. **Keisha aced the exam.**
More emotional	Please decide which sentence is more emotional, meaning which one seems to elicit more of an emotional response in you, the reader. **Listening to the news tortured him.** Listening to the news hurt him. It was a beautiful image. **It was a stunning image.**

Results

The percentage of participants who selected the metaphorical sentence for each item in each survey is shown in Figure 6.

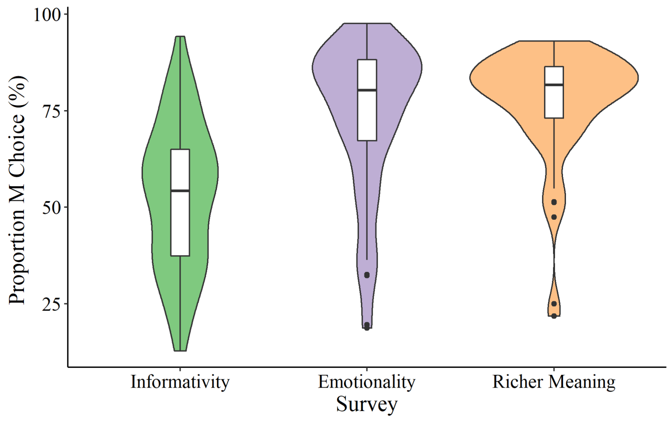


Figure 6: Distribution of percentages of participants who selected the metaphorical sentence rather than its literal paraphrase for the 60 M-L pairs. Separate surveys asked which sentence was more informative (Informativity), conveyed more emotion (Emotionality), or conveyed richer meaning (Richer Meaning) to them, the reader. Chance = 50%.

For each survey, we determined whether the 2AFC responses were distinct from chance using a generalized linear binomial model with subject and item-pair intercepts as random effects. Participants judged that metaphorical sentences conveyed richer meaning and evoked more emotion at above-chance rates: Richer Meaning (*M* = 82%, *CI* = [78-86%]); Emotion (*M* = 79%, *CI* = [73-84%]). Responses on the two surveys correlated with one another; $r^2 = 0.59$. On the other hand, participants did not choose M responses at above-chance rates when asked which sentence conveyed more information (*M* = 54%, *CI* = [47-60%]). A model that included both Richer Meaning and Emotion as predictors of M responses, and random intercepts for stimuli pair and participant, found that Richer Meaning was the stronger predictor ($\beta = 0.28$, *SD* = 0.13, $p = 0.03$). Responses from participants who reported explicit awareness of the metaphors differed little from those who did not. The mean number of M responses out of 20 in each survey, when comparing participants who said they were vs. were not aware of the metaphors, were as follows: Informativity-aware: 10.62, vs. not aware: 10.71; Emotionality-aware: 14.95, vs. not aware: 14.62; Richer Meaning-aware: 15.73, vs. not aware: 14.39.

In order to determine whether M responses were more specifically predicted by the gradient measure of metaphoricity collected from the norming ratings in Study 1, we calculated an Increase in Metaphoricity score for each M-L sentence-pair by subtracting the mean metaphoricity score of the L sentence from the mean metaphoricity score of the corresponding M sentence. We then correlated Increase in Metaphoricity scores for sentence-pairs with each survey's proportion of participants who selected the M response for each pair. Results showed that Increase in Metaphoricity scores were significantly correlated with M choices in the Richer Meaning survey (r = .29; p = 0.03), but not with Emotion choices (r = .21; p = 0.11) nor with Informativity choices (r = -.12; p = 0.36).

Discussion

We conducted three surveys in an attempt to better characterize whether and how speakers experience the differences observed in Experiment 2 between sentences containing conventional metaphors and their literal paraphrases. We hypothesized that the distinction might be based on a perception of metaphorical sentences as conveying more information, or evoking more emotion, or a quality which we chose to label "richer meaning" following Colston (2015).

Participants showed no bias toward selecting sentences with conventional metaphors when asked which sentence conveyed more information. A separate group showed some tendency to choose metaphors when asked which sentence conveyed more emotion, but gradient scores of degree of Metaphoricity from the norming study did not correlate with the proportion of participants who selected the metaphorical sentences. The third survey, which asked participants to choose which sentence conveyed "richer meaning," showed the strongest response bias toward metaphor choice, and the proportion of participants who chose metaphorical sentences correlated significantly with the norming group's ratings of degree of metaphoricity of the sentences. That is, the categorical variable (M sentence type) and the gradient measure of metaphoricity both significantly predicted the likelihood that participants would judge conventional metaphors as conveying richer meaning.

We consider “richer meaning,” like “greater engagement” (and “focused attention”), to be neutral with regard to emotional or cognitive processing. Therefore, survey results underscore our decision to remain neutral regarding this potential distinction.

General Discussion and Conclusion

Conventional metaphorical expressions are woven into the fabric of our everyday discourse. In fact it can be challenging to talk about abstract topics for more than a sentence or two without employing them (Ortony, 1975). Prior work has suggested that conventional metaphors are more engaging than literal paraphrases, on the basis of increased neural activity in comparison to literal paraphrases in the amygdala, a brain structure associated with heightened emotional arousal or focused attention to relevant stimuli. In particular, increased amygdala activation was found during the processing of sentences and stories containing conventional metaphors (Citron & Goldberg, 2014; Citron et al., 2016; Citron et al., 2020; see also Forgács et al., 2012, using compound words), and in a meta-analysis of metaphor processing (Bohrn et al., 2012).

The current study takes advantage of the fact that stimulus-evoked pupil dilation is an implicit and time-sensitive index of focused attention or task engagement, as reviewed in the Introduction (e.g., Liao et al., 2016; Preuschoff, ’t Hart, & Einhäuser 2011; Schaefer & Gray, 2007). We asked participants to listen to sentences, as their pupil dilations were recorded, with comprehension questions following filler trials to ensure participants were interpreting the sentences. One third of the sentences contained conventional metaphors, and the rest included a combination of literal paraphrases and concrete descriptions, which served as controls.

Specifically, we created a database of 60 sentence triples, each including a) a sentence containing a conventional metaphor, b) a literal paraphrase, and c) a concrete description. The sentences were normed on judgments of complexity, plausibility, familiarity, valence and intensity, and we included these gradient factors in all analyses. As intended, metaphorical and literal sentences were judged to be highly similar in meaning according to both human judgments and Latent Semantic Analysis (Dumais, 2004); the concrete descriptions were recognized to be more imageable (and more concrete) than metaphorical or literal sentences. Finally, metaphorical sentences were judged to be more metaphorical overall, while varying in their perceived degree of metaphoricity.

The current work confirms heightened engagement when participants witness sentences containing conventional metaphors compared to literal paraphrases that convey nearly the same meaning, using a wholly different method and stimuli than prior fMRI work. Moreover, the results demonstrate that the increased engagement is not due to greater imageability nor the inclusion of concrete words in metaphors. This is important because the only prior study to investigate imageability or concreteness as a possible source of greater engagement had not found significantly greater engagement even in metaphorical sentences (Pomp et al., 2018). Further work is required, but we suspect that lack of increased amygdala activity in that study was due to the fact that the sentences included words related to smell, and olfaction may independently evoke amygdala activity. The current results show increased pupil dilation when listeners comprehended sentences containing conventional metaphors in comparison to literal paraphrases (which conveyed similar meanings) or concrete descriptions (which share similar words).

Heightened pupil dilation in response to conventional metaphors was found, regardless of whether metaphoricity was treated as a categorical variable (as condition) or as a continuous variable; and regardless of whether pupil dilation was considered only at the key phrase, or from after the key phrase until two seconds beyond the end of the sentence. The four models support the same conclusion: conventional metaphors evoke greater pupil dilation in comparison to literal paraphrases or concrete descriptions. Notably, the concrete and paraphrase conditions showed dilation responses that were indistinguishable from one another, despite including different words and conveying very different meanings.

In exploratory analyses, the only normed factor aside from Metaphoricity to show a significant effect on dilation was Intensity, and this factor did not survive corrections in analyses of the extended period following the key phrase. None of the matched variables, including Intensity, correlated strongly with Metaphoricity in our stimuli. Thus analyses demonstrate the predicted specific boost in pupil dilation in response to conventional metaphors in comparison to literal paraphrases or concrete descriptions. Listeners implicitly find sentences that contain conventional metaphors to be intrinsically more engaging.

The current work contributes to an understanding of the time course of the increased engagement evoked by conventional metaphors in comparison to paraphrases or concrete descriptions. Increased dilation is measurable almost immediately, unlike responses measured by fMRI analysis, and it can also be long-lasting, as it measures physiological arousal and cognitive engagement, unlike the short-lived responses evident in analyses of ERP components. Current results show heightened dilation as soon as the metaphorical phrase is heard, in comparison to concrete or literal sentences, undermining the possibility that it is caused by downstream inferences. The dilation is sustained across the entire trial, additionally undermining the possibility that the effect is caused by lexical access of the words in the key phrase or some other immediate but short-lived process. Instead, the time course data suggest that conventional metaphors are more engaging as soon as they are recognized and remain more engaging over the course of their integration into the meaning of the entire sentence.

In an effort to determine if listeners perceived metaphorical sentences to be more emotionally engaging or more informative than literal paraphrases, we conducted a final set of surveys with three new groups of participants. Separate groups were asked to decide whether each metaphorical sentence or its literal paraphrase expressed more emotion, more information, or a third description which we take to be neutral between emotion and cognition, namely that metaphors evoke “richer meaning” (Colston, 2015).

The best predictor of choosing the metaphorical sentences over literal paraphrases came from the survey that asked which sentence conveyed richer meaning. The likelihood that participants would decide that a sentence conveyed richer meaning correlated with the gradient degree of metaphoricity, as well. Evidence that the conventional metaphors were more emotionally engaging (Citron & Goldberg, 2014) in the current study was inconclusive. Participants were more likely to choose metaphors than literal sentences when asked which sentence evoked more emotion. However, the gradient measure of Metaphoricity, collected in the norming task, did not correlate with the proportion of participants who selected metaphorical over literal sentences as conveying more emotion.

It might be tempting to interpret “richer meaning” as implying that metaphors convey more information. However, participants did not show any preference for the metaphorical sentences over literal paraphrases when explicitly asked in the final survey which conveyed more information. Current results also undermine the possibility that the greater engagement evoked by metaphorical sentences was due to their being more difficult to process, since all conditions were matched on familiarity, complexity and plausibility, and none of these factors showed a significant effect on pupil dilation in our stimuli.

The lack of evidence suggesting that conventional metaphors were perceived to convey more information than paraphrases, despite the fact that they evoked an increase in pupil dilation compared to controls, appears to contrast with a recent claim in a review of pupillometry work that links pupil dilation to the amount of information conveyed by a stimulus (Zénon, 2019). Zénon states that “changes in pupil-linked arousal all depend on… the update of brain internal models”

(2019, p. 1). The claim is supported, for example, by results of a gambling task reported by Preuschoff et al. (2011), in which participants received two cards from a fresh deck of 10 cards, labeled 1-10, on each trial. After seeing the first card, participants had to guess whether the number on the second card would be higher or lower. Notice that if the first card is low or high, it provides more useful information than if it is in the middle range. For instance, if the first card is 2, it provides a strong indication that the second card will be higher, whereas if the first card is 5, the second card is just about equally likely to be higher or lower. As predicted, both low and high numbers evoked greater dilation than numbers closer to the middle (Preuschoff et al., 2011; Zénon, 2019). Note that the sense in which both high and low numbers provided "more information" than those in the middle depended on the task, and Preuschoff et al. (2011, p. 1) themselves characterize the reported increase in pupil dilation as indexing heightened "task engagement" which is consistent with the current interpretation of pupil dilation, namely that it indexes degree of engagement during comprehension. We cannot resolve whether the engagement is best interpreted as due to emotional processing or cognitive processing, perhaps because the distinction is not germane.

The current pupil dilation results nonetheless allow us to triangulate the special sauce that conventional metaphors provide during sentence comprehension. Evidence supports the claim that sentences containing conventional metaphors are more engaging than literal paraphrases or concrete descriptions, even when other relevant variables including familiarity, emotional valence and intensity, complexity, and plausibility are taken into account. We conclude that conventional metaphors are more engaging – convey richer meaning – as soon as they are recognized and as they are integrated into the overall interpretation of the sentence. The engagement is irreducible to concreteness, difficulty or ease, amount of information, short-term lexical access, or downstream inferences.

Acknowledgements

We are grateful to Robin Gomila for statistical advice, and to three anyonymous reviewers from *JML* and associate editor Penny Pexman for helpful feedback and suggestions on two earlier versions of this paper. We also appreciate thoughtful feedback and discussion from audiences at UC Merced, Cornell, the 19th EdukCircle International Convention on Psychology, University of Michigan, and University of Alberta.

References

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, *14*(4), 1–20. https://doi.org/10.1167/14.4.1

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bambini, V., Ghio, M., Moro, A., & Schumacher, P. B. (2013). Differentiating among pragmatic uses of words through timed sensicality judgments. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00938

Beck, S. D. (2020). *Native and non-native idiom processing: Same difference* (Doctoral dissertation). Retrieved from Universität Tübingen.

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boers, F. (2003). Applied linguistics perspectives on cross-cultural variation in conceptual metaphor. *Metaphor and Symbol*, *18*(4), 231–238. https://doi.org/10.1207/S15327868MS1804_1

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*(11), 2669–2683. https://doi.org/10.1016/j.neuropsychologia.2012.07.021

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. https://doi.org/10.1016/j.jml.2004.05.006

Bradley, M. M., & Lang, P. J. (2015). Memory, emotion, and pupil diameter: Repetition of natural scenes. *Psychophysiology*, *52*(9), 1186–1193. https://doi.org/10.1111/psyp.12442

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Citron, F.M.M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A.M. (2016). When emotions are expressed figuratively: Psycholinguistic and affective norms of

619 idioms for German (PANIG). *Behavior Research Methods, 48,* 91-111. doi: 10.3758/s13428-015-0581-4

Citron, F. M. M., & Goldberg, A. E. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, *26*(11), 2585–2595. https://doi.org/10.1162/jocn_a_00654

Citron, F.M.M., Lee, M., & Michaelis, N. (2020). Affective and psycholinguistic norms for German conceptual metaphors (COMETA). *Behavior Research Methods, 52,*1056-1072. doi: 10.3758/s13428-019-01300-7

Citron, F. M. M., Michaelis, N., & Goldberg, A. E. (2020). Metaphorical language processing and amygdala activation in L1 and L2. *Neuropsychologia*, *140*. https://doi.org/10.1016/j.neuropsychologia.2020.107381

Citron, F. M. M., Güsten, J., Michaelis, N., & Goldberg, A. E. (2016). Conventional metaphors in longer passages evoke affective brain response. *NeuroImage*, *139*, 218–230. https://doi.org/10.1016/j.neuroimage.2016.06.020

Citron, F. M., Cacciari, C., Funcke, J. M., Hsu, C.-T., & Jacobs, A. M. (2019). Idiomatic expressions evoke stronger emotional responses in the brain than literal sentences. *Neuropsychologia*, *131*, 233–248. https://doi.org/10.1016/j.neuropsychologia.2019.05.020

Colston, H. L. (2015). *Using figurative language*. Cambridge, UK: Cambridge University Press.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, *58*(3), 306–324. https://doi.org/10.1016/j.neuron.2008.04.017

Costafreda, S. G., Brammer, M. J., David, A. S., & Fu, C. H. Y. (2008). Predictors of amygdala activation during the processing of emotional stimuli: A meta-analysis of 385 PET and fMRI studies. *Brain Research Reviews*, *58*(1), 57–70. https://doi.org/10.1016/j.brainresrev.2007.10.012

Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition*, *30*(6), 958–968. https://doi.org/10.3758/BF03195780

Cunningham, W. A., & Brosch, T. (2012). Motivational salience: Amygdala tuning from traits, needs, values, and goals. *Current Directions in Psychological Science*, *21*(1), 54–59. https://doi.org/10.1177/0963721411430832

Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., & Seidenberg, M. S. (2011). The neural career of sensory-motor metaphors. *Journal of Cognitive Neuroscience*, *23*(9), 2376–2386. https://doi.org/10.1162/jocn.2010.21596

Dobrovol'skij, D., & Piirainen, E. (2005). *Figurative language: Cross-cultural and cross-linguistic perspectives*. Leiden, The Netherlands: Brill.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230. https://doi.org/10.1002/aris.1440380105

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology*, *63*(4), 639–645. https://doi.org/10.1080/17470210903469864

Forgács, B., Bohm, I.C., Baudewig, J., Hofmann, M.J., Csaba, P., & Jacobs, A.M. (2012). Neural correlates of combinatorial semantic processing of literal and figurative noun noun compound words. *NeuroImage, 63,* 1432-1442. https://doi.org/10.1016/j.neuroimage.2012.07.029

Garavan, H., Pendergrass, J. C., Ross, T. J., Stein, E. A., & Risinger, R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. *NeuroReport*, *12*(12), 2779–2783. https://doi.org/10.1097/00001756-200108280-00036

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152–158. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x

Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R., & Barr, D. J. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language*, *37*(2), 141–154. https://doi.org/10.1006/jmla.1996.2506

Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 577–590. https://doi.org/10.1016/S0022-5371(83)90355-9

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions.* Princeton, NJ: Princeton University Press.

Gross, M. P., & Dobbins, I. G. (2021). Pupil dilation during memory encoding reflects time pressure rather than depth of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(2), 264–281. https://doi.org/10.1037/xlm0000818

Hamann, S., & Mao, H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *NeuroReport*, *13*(1), 15–19. https://doi.org/10.1097/00001756-200201210-00008

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Iacozza, S., Costa, A., & Duñabeitia, J. A. (2017). What do your eyes reveal about your foreign language? Reading emotional sentences in a native and foreign language. *PLoS ONE*, *12*(10), 1–10. https://doi.org/10.1371/journal.pone.0186027

Iakimova, G., Passerieux, C., Laurent, J.-P., & Hardy-Bayle, M.-C. (2005). ERPs of metaphoric, literal, and incongruous semantic processing in schizophrenia. *Psychophysiology*, *42*(4), 380–390. https://doi.org/10.1111/j.1469-8986.2005.00303.x

Ibarretxe-Antuñano, I. (2005). Limitations for cross-linguistic metonymies and metaphors. In J. L. Otal, I. Navarro i Ferrando, & B. Bellés Fortuño (Eds.), *Cognitive and discourse approaches to metaphor and metonymy* (pp. 187-200). Castelló de la Plana, Spain: Publicacions de la Universitat Jaume I.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *47*(2), 310–339. https://doi.org/10.1037/h0078820

Kafkas, A., & Montaldi, D. (2012). Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. *Neuropsychologia*, *50*(13), 3080–3093. https://doi.org/10.1016/j.neuropsychologia.2012.08.001

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kang, O., & Wheatley, T. (2017). Pupil dilation patterns spontaneously synchronize across individuals during shared attention. *Journal of Experimental Psychology: General*, *146*(4), 569–576. https://doi.org/10.1037/xge0000271

Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology*, *41*(2), 175–185. https://doi.org/10.1111/j.1469-8986.2004.00147.x

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169–1180. https://doi.org/10.3758/BF03195800

Kinner, V. L., Kuchinke, L., Dierolf, A. M., Merz, C. J., Otto, T., & Wolf, O. T. (2017). What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes. *Psychophysiology*, *54*(4), 508–518. https://doi.org/10.1111/psyp.12816

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. International Journal of *Psychophysiology*, *65*(2), 132–140. https://doi.org/10.1016/j.ijpsycho.2007.04.004

Lacey, S., Stilla, R., & Sathian, K. (2012). Metaphorically feeling: Comprehending textual metaphors activates somatosensory cortex. *Brain and Language*, *120*(3), 416–421. https://doi.org/10.1016/j.bandl.2011.12.016

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on Psychological* Science, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing*, *12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z

Lai, V. T., & Curran, T. (2013). ERP evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors. *Brain and Language*, *127*(3), 484–496. https://doi.org/10.1016/j.bandl.2013.09.010

Lai, V. T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain Research*, *1284*, 145–155. https://doi.org/10.1016/j.brainres.2009.05.088

Lai, V. T., Howerton, O., & Desai, R. H. (2019). Concrete processing of action metaphors: Evidence from ERP. *Brain Research*, *1714*, 202–209. https://doi.org/10.1016/j.brainres.2019.03.005

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202-251). Cambridge, UK: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lavin, C., Martin, R. S., & Jubal, E. R. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience*, *7*(218), 1–8. https://doi.org/10.3389/fnbeh.2013.00218

Liao, H. I., Kidani, S., Yoneya, M., Kashino, M., & Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychon. Bull. Rev., 23*, 412–425. https://doi.org/10.%203758/s13423-015-0898-0

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Littlemore, J. (2019). *Metaphors in the Mind*. Cambridge, UK: Cambridge University Press.

McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., & Jackendoff, R. (2001). Reading time evidence for enriched composition. *Cognition*, *78*(1), B17–B25. https://doi.org/10.1016/S0010-0277(00)00113-X

Merritt, S. L., Keegan, A. P., & Mercer, P. W. (1994). Artifact management in pupillometry. *Nursing Research*, *43*(1), 56–59. https://doi.org/10.1097/00006199-199401000-00012

Müller, N., Nagels, A., & Kauschke, C. (2021). Metaphorical expressions originating from the human senses: Psycholinguistic and affective norms for German metaphors for internal state terms (MIST database). *Behavior Research Methods.* https://doi.org/10.3758/s13428-021-01639-w

Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, *35*(8), 4140–4154. https://doi.org/10.1002/hbm.22466

Nencheva, M. L., Piazza, E. A., & Lew-Williams, C. (2020). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Developmental Science*, e12997. https://doi.org/10.1111/desc.12997

Oliva, M., & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-23265-x

Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational Theory*, *25*(1): 45–53. https://doi.org/10.1111/j.1741-5446.1975.tb00666.x

Ortony, A. (1978). Remembering, understanding, and representation. *Cognitive Science*, *2*(1), 53–69. https://doi.org/10.1016/S0364-0213(78)80061-5

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Paivio, A., & Simpson, H. M. (1966). The effect of word abstractness and pleasantness on pupil size during an imagery task. *Psychonomic Science*, *5*(2), 55–56. https://doi.org/10.3758/BF03328277

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1-2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167. https://doi.org/10.3758/PBR.15.1.161

Pomp, J., Bestgen, A. K., Schulze, P., Müller, C. J., Citron, F. M. M., Suchan, B., & Kuchinke, L. (2018). Lexical olfaction recruits olfactory orbitofrontal cortex in metaphorical and literal contexts. *Brain and Language*, *179*, 11–21. https://doi.org/10.1016/j.bandl.2018.02.001

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*. https://doi.org/10.3389/fnins.2011.00115

Reuterskiöld, C., & Van Lancker Sidtis, D. (2013). Retention of idioms following one-time exposure. *Child Language Teaching and Therapy*, *29*(2), 219–231. https://doi.org/10.1177/0265659012456859

Samur, D., Lai, V. T., Hagoort, P., & Willems, R. M. (2015). Emotional context modulates embodied metaphor comprehension. *Neuropsychologia*, *78*, 108–114. https://doi.org/10.1016/j.neuropsychologia.2015.10.003

Schaefer, A., & Gray, J. R. (2007). A role for the human amygdala in higher cognition. *Reviews in the Neurosciences*, *18*(5), 355–382. https://doi.org/10.1515/revneuro.2007.18.5.355

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience, 27*(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., ... & Maquet, P. (2006). The locus ceruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience*, *26*(28), 7416–7423. https://doi.org/10.1523/JNEUROSCI.1001-06.2006

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96–1160. https://doi.org/10.1111/j.1467-9582.2010.01176.x

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PloS ONE*, *6*(2), e16782. https://doi.org/10.1371/journal.pone.0016782

Thibodeau, P. H., Hendricks, R. K., & Boroditsky, L. (2017). How linguistic metaphor scaffolds reasoning. *Trends in Cognitive Sciences*, *21*(11), 852–863. https://doi.org/10.1016/j.tics.2017.07.001

van Steenbergen, H., Band, G. P. H., & Hommel, B. (2011). Threat but not arousal narrows attention: Evidence from pupil dilation and saccade control. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00281

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

Welcome, S. E., Paivio, A., McRae, K., & Joanisse, M. F. (2011). An electrophysiological study of task demands on concreteness effects: Evidence for dual coding theory. *Experimental Brain Research*, *212*(3), 347–358. https://doi.org/10.1007/s00221-011-2734-8

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

Winter, B. (2016). Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience*, *31*(8), 975–988. https://doi.org/10.1080/23273798.2016.1193619

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic?: Iconicity in English sensory words. *Interaction Studies*, *18*(3), 443-464.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Zénon, A. (2019). Eye pupil signals information gain. *Proceedings of the Royal Society B*, *286*(1911). https://doi.org/10.1098/rspb.2019.1593

Appendix A: Full results of models

Model I: Effect of Condition (M, L vs C), a categorical variable, at the key phrase (pre-registered)

pupilSize_atKeyPhrase ~ Condition + familiarity + valence+ intensity + complexity + plausibility + (1 | participant) + (1 | sentence)

Fixed effects				
	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	103.25	0.52	198.95	< .0001 ***
conditionL	0.23	0.65	0.35	.725
conditionM	**1.44**	**0.63**	**2.27**	**.023 ***
familiarity	-0.39	0.29	-1.34	.180
valence	0.55	0.30	1.86	.063
intensity	**0.79**	**0.30**	**2.64**	**.008 ****
complexity	-0.53	0.28	-1.91	.056
plausibility	-0.004	0.27	-0.02	.987

Random effects		
	Variance	SD
participant (Intercept)	4.10	2.02
sentence (Intercept)	4.677e-15	6.839e-08
Residual	2.08	10.44

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnst*	*cmplxt*
conditionL	-.62						

conditionM	-.62	.51					
familiarity	.08	-.15	-.05				
valence	.04	-.02	-.07	-.06			
intensity	.02	.01	-.05	-.01	.49		
complexity	.07	-.11	-.07	.37	-.05	-.01	
plausibility	.10	-.14	-.09	-.24	-.01	.14	-.06

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25905.9	25972.6	-12942.0	25883.9	3150

Model II: Effect of Metaphoricity, a gradient measure, at the key phrase (pre-registered)

pupilSize_atKeyPhrase ~ Metaphoricity + familiarity + valence + intensity + complexity + plausibility + (1 | participant) + (1 | sentence)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	103.81	0.37	284.28	< .0001 ***
metaphoricity	**0.58**	**0.26**	**2.21**	**.028 ***
familiarity	-0.32	0.29	-1.13	.260
valence	0.51	0.30	1.71	.087
intensity	**0.78**	**0.30**	**2.61**	**.009 ****
complexity	**-0.56**	**0.28**	**-2.02**	**.043 ***
plausibility	0.02	0.27	0.08	.936

Random effects

	Variance	SD
participant (Intercept)	4.07	2.02
sentence (Intercept)	0.00	0.00

Residual	208.01	14.42

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnst*	*cmplxt*
metaphorcty	.00					
familiarity	-.001	.12				
valence	.00	-.14	-.08			
intensity	.001	-.08	-.01	.49		
complexity	.00	-.08	.35	-.04	-.003	
plausibility	.001	.00	-.26	-.01	.14	-.08

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25905.1	25965.7	-12942.5	25885.1	3151

Model III: Effect of Condition (M, L vs C), a categorical variable during the rest of sentence after the key phrase, 3 x 667 ms. of silence

pupilSize_acrossTrial ~ Condition + familiarity + valence + intensity + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	103.11	0.93	110.71	< .0001
conditionL	0.32	0.47	0.67	.50469
conditionM	**1.44**	**0.45**	**3.18**	**.00149 ****
familiarity	-0.19	0.26	-0.73	.46332
valence	0.64	0.33	1.93	.05572

intensity	0.59	0.28	2.10	.03618
complexity	-0.32	0.29	-1.09	.27622
plausibility	-0.54	0.25	-2.15	.03207

Random effects

	Variance	SD
participant (Intercept)	7.82	2.80
sentence (Intercept)	4.82	2.20
Timepoint (Intercept)	2.22	1.49
Residual	412.63	20.31

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnst*	*cmplxt*
conditionL	-.25						
conditionM	-.25	.51					
familiarity	.04	-.19	-.06				
valence	.02	-.02	-.09	-.09			
intensity	.006	.02	-.05	-.002	.38		
complexity	.04	-.13	-.09	.43	-.08	.02	
plausibility	.05	-.19	-.12	-.22	-.09	.09	-.07

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
111002.6	111091.8	-55489.3	110978.6	12493

Model IV: Effect of Metaphoricity, a gradient measure, during the rest of sentence after phrase, 3 x 667 ms. silence)

pupilSize_acrossTrial ~ Metaphoricity + familiarity + valence + intensity + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	103.70	0.89	115.99	< .0001
metaphoricity	**0.50**	**0.19**	**2.59**	**.00965 ****
familiarity	-0.11	0.26	-0.41	.68229
valence	0.62	0.34	1.84	.06778
intensity	0.58	0.28	2.07	.03902
complexity	-0.32	0.29	-1.10	.27377
plausibility	-0.50	0.25	-2.04	.04161

Random effects

	Variance	SD
participant (Intercept)	7.79	2.79
sentence (Intercept)	4.99	2.23
timepoint (Intercept)	2.22	1.49
Residual	412.74	20.32

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnst*	*cmplxt*
metaphorcty	.00					
familiarity	.00	.13				
valence	.001	-.15	-.11			
intensity	.001	-.09	-.01	.38		
complexity	.000	-.08	.40	-.07	.02	
plausibility	-.001	-.02	-.26	-.09	.09	-.09

AIC	BIC	logLik	deviance	df.resid
111005.2	111087.0	-55491.6	110983.2	12494

Supplemental Information (online)

Preregistration http://aspredicted.org/blind.php?x=ae2ki4

Pupillometry Experiment Plan

Study goal: to compare cognitive/emotional engagement and memorability of metaphorical sentences with that of literal and concrete sentences.

We created a set of stimuli and then selected 60 sentence triples based on norming data. Each triple contains one 1) sentence with a conventional metaphor (M); 2) a literal paraphrase (L); and 3) a sentence using the same phrase as the metaphorical sentence to describe a concrete sense (C).

The sentences in each sentence triple, <M,L,C> were matched on the following features: emotional valence, intensity, complexity, familiarity, and plausibility. As intended, judgments of metaphoricity were higher for M (M>L,C); imageability was higher for C, and semantic similarity was higher (and high) between M and L.

After 2 preliminary sentences, in a within-subjects design, participants will listen to 60 randomly ordered sentences (1 from each triple, 20 from each condition) + 12 filler sentences. Filler sentences will be randomly interspersed after every 2-8 sentences and will be immediately followed by a corresponding multiple choice comprehension question (12 questions total). A secondary verbatim recognition memory task will be administered after the listening task.

Exclusion criteria: Participants who have lower than 70% accuracy on the comprehension questions (4 or more incorrect answers) will be excluded (chance being 25%).

Participants: 66 undergraduate students; the number was preregistered, determined on the basis of preliminary pilot data that is excluded, using 0.80 power for a 2-tailed t-test with alpha of 0.05.

DVs: 1) pupil change over the course of the target phrase relative to baseline (Baseline = average pupil size during the first 100ms of the key phrase) (and Mean per item)

2) pupillary synchrony across participants for the duration of the target phrase

Analysis 1: We will compare linear mixed effects models to test the effect of Condition (M vs. L vs. C) and a continuous score of Metaphoricity on each dependent variable (DV).

Binned by condition:

(a1) Largest convergent model

[DV~ Condition + control variables + (1 + Condition| subject) + (1+ Condition| item)], where:

- subject is participant *pair* in the case of synchrony.
- Control variables: emotional valence, intensity, complexity, familiarity, and plausibility.

(a2) Smaller model

[DV~ Condition + control variables + (1 | subject) + (1|item)]

(a3) Null model

[DV~ 1 + control variables + (1 | subject) + (1|item)]

Metaphoricity as a continuous variable:

(a1) a large model [DV~ Metaphoricity + control variables + (1 + Metaphoricity| subject)]

(a2) a smaller model [DV~ Metaphoricity + control variables + (1 | subject)]

(a3) a null model [DV~ 1 + control variables + (1 | subject)]

The effect of M,L,C or Metaphoricity on the dependent variables will be tested by comparing models (a1) and (a3) using the anova function in R, with (a1) preferred if it converges and is significantly better than model (a2) at alpha = 0.05, or by comparing models (a2) and (a3) otherwise.

Analyses

Following the preregistration, we initially tested a maximal model with random slopes included for condition for subjects and items. This model did not converge, so as preregistered, we eliminated random slopes, retaining random intercepts for subjects and items. The elimination of slopes allowed us to compare models with and without condition as intended and reported in the main text. The preregistration had erroneously specified that we exclude random intercepts for sentence-pairs when we compare sentences on the gradient measure of Metaphoricity (Models II and IV). These models also demonstrate a significant effect of metaphoricity, so we include the fuller models in the body of the paper, since including more complete random effect structure is preferable.

(Post hoc) norming of the key phrases in isolation

We preregistered and collected ratings of emotional Intensity in response to full sentence stimuli, following relevant precedent. This was motivated by the fact that participants in the pupillometry study never saw key phrases in isolation, but only saw them in the context of full sentences. Yet as mentioned in the main text, we later collected post hoc norming data on judgments of emotional intensity at the key phrase in isolation (hereafter, IKP). As in Study 1, we aimed to collect judgments from 60 participants using the Cloud Research platform for Amazon Mechanical Turk. Fifty-nine participants completed the survey. Consistent with the inclusion of intensity of the full sentences rather than IKP is that the correlation of sentence intensity and IKP was high (Pearsons r = .71). Also as in Study 1, we used the non-parametric Mann-Whitney U test to compare the rating distributions between pairs of conditions. This analysis reveals that L&M are in fact significantly different (W = 1236.5, p = 0.003), although neither C&L nor C&M are (C&L: W = 2073.5, p = 0.152; C&M: W = 1555.5, p = 0.200).

The fact that M&L differed on the IKP norming was unintended, and raises the possibility that metaphoricity was confounded with IKP. However, the fact that C&M did not differ from one another undermines the likelihood that IKP rather than metaphoricity was responsible for increases in pupil dilation, since M consistently resulted in greater pupil dilation than C. In addition, the correlation between metaphoricity and IKP was not strong (Pearson's r = .24).

Nonetheless, to rule out the possibility that IKP undermines the results reported in the main text, we performed additional analyses in two ways, creating two new sets of models Models $\mathrm{I}_{+\mathrm{IKP}}$ through $\mathrm{IV}_{+\mathrm{IKP}}$ and Models $\mathrm{I}_{\mathrm{IKP}}$ - $\mathrm{IV}_{\mathrm{IKP},}$ as described below.

Models $\mathbf{I}_{+\mathbf{IKP}}$ - $\mathbf{IV}_{+\mathbf{IKP}}$

In a first set of post-hoc analyses, we included ratings of IKP as a fixed factor in each of the Models I-IV (creating Models $\mathrm{I}_{+\mathrm{IKP}}$ through $\mathrm{IV}_{+\mathrm{IKP}}$). Although as already noted, IKP did not improve any of the models' fit compared to models without IKP, we examined whether metaphoricity/condition remained a significant influence on pupil dilation with IKP included in the models. In Model $\mathrm{I}_{+\mathrm{IKP}}$, as in Model I (without IKP), the key phrase evoked significantly more dilation in the Metaphor condition than the reference condition ($\beta_M = 1.33$, $SE = 0.65$, $p = .040$), while the Literal condition was indistinguishable from the reference condition ($\beta_L = 0.25$, $SE = 0.65$, $p = .704$). In Model $\mathrm{II}_{+\mathrm{IKP}}$, as metaphoricity increased, we now find only a marginal increase in dilation at the key phrase ($\beta_M = 0.52$, $SE = 0.28$, $p = .065$). In Model $\mathrm{III}_{+\mathrm{IKP}}$, as in Model III, the Metaphor condition evoked significantly more dilation during the period extending until 2 seconds after the key phrase in comparison to the Concrete condition, with Bonferroni correction applied [requiring $p < .0125$] ($\beta_M = 1.40$, $SE = 0.47$, $p = .003$), while Literal was not different than the C condition ($\beta_L = 0.33$, $SE = 0.47$, $p = .483$). Finally, in Model IV+$_{\mathrm{IKP}}$, as metaphoricity increases, dilation showed a marginal increase in the period extending 2 seconds beyond the sentence with Bonferroni correction applied ($\beta_M = 0.47$, $SE = 0.21$, $p = .0255$). To summarize, when we add IKP to the models, Model $\mathrm{I}_{+\mathrm{IKP}}$ and Model $\mathrm{III}_{+\mathrm{IKP}}$ show a significant effect of condition on dilation, while models that test the gradient influence of metaphoricity reveal a weaker but still marginal effect (Model $\mathrm{II}_{+\mathrm{IKP}}$ and $\mathrm{IV}_{+\mathrm{IKP}}$). The full models are below.

Model I+$_{\text{IKP}}$: Effect of Condition (M, L vs C), a categorical variable, at the key phrase

pupilSize_atKeyPhrase ~ Condition + familiarity + valence + intensity + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	103.28	0.52	198.56	< .0001 ***
conditionL	0.25	0.65	0.38	.704
conditionM	1.33	0.65	2.06	**.040 ***
familiarity	-0.31	0.31	-1.01	.315
valence	0.56	0.30	1.89	.059
intensity	0.57	0.40	1.43	.154
intensityKeyPhrase	0.33	0.40	0.82	.414
complexity	-0.50	0.28	-1.76	.078
plausibility	-0.01	0.27	-0.05	.957

Random effects

	Variance	SD
participant (Intercept)	4.09	2.02
sentence (Intercept)	0.00	0.00
Residual	207.88	14.42

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnst*	*intnKP*	*cmplxt*
conditionL	-.61							
conditionM	-.62	.49						
familiarity	.10	-.13	-.11					

valence	.04	-.01	-.08	-.05				
intensity	-.04	-.02	.10	-.22	.34			
intnstyKyPh	.07	.04	-.20	.32	.04	-.67		
complexity	.08	-.10	-.10	.40	-.04	-.11	.15	
plausibility	.09	-.15	-.08	-.24	-.01	.14	-.05	-.07

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25907.3	25980.0	-12941.6	25883.3	3149

Model II+$_{\text{IKP}}$: Effect of Metaphoricity, a gradient measure, at the key phrase

pupilSize_atKeyPhrase ~ Metaphoricity + familiarity + valence + intensity + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	103.81	0.37	284.30	< .0001 ***
metaphoricity	0.52	0.28	1.85	**.065**
familiarity	-0.26	0.30	-0.86	.393
valence	0.53	0.30	1.76	.079
intensity	0.59	0.40	1.47	.142
intensityKeyPhrase	0.28	0.41	0.70	.487
complexity	-0.52	0.28	-1.86	**.063**
plausibility	0.01	0.27	0.05	.963

Random effects

	Variance	SD
participant (Intercept)	4.06	2.02
sentence (Intercept)	0.00	0.00

Residual	207.98	14.42

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnst*	*intnKP*	*cmplxt*
metaphorcty	.00						
familiarity	.00	.00					
valence	.00	-.16	-.05				
intensity	.00	.17	-.22	.32			
intnstyKyPh	.001	-.33	.31	.07	-.67		
complexity	.001	-.14	.38	-.02	-.13	.19	
plausibility	.001	.02	-.27	-.02	.14	-.05	-.09

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25906.6	25973.3	-12942.3	25884.6	3150

Model III+$_{\text{IKP}}$: Effect of Condition (M, L vs C), a categorical variable during the rest of sentence after the key phrase, 3 x 667 ms. of silence

pupilSize_acrossTrial ~ Condition + familiarity + valence + intensity + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	103.12	0.93	110.74	< .0001
conditionL	0.33	0.47	0.70	.48343
conditionM	1.40	0.47	3.00	**.00275**

familiarity	-0.17	0.27	-0.64	.52564
valence	0.64	0.33	1.94	.05454
intensity	0.51	0.36	1.39	.16653
intensityKeyPhrase	0.14	0.38	0.36	.72023
complexity	-0.31	0.29	-1.06	.29100
plausibility	-0.55	0.25	-2.17	**.03011**

Random effects

	Variance	SD
participant (Intercept)	7.81	2.79
sentence (Intercept)	4.79	2.19
Timepoint (Intercept)	2.22	1.49
Residual	412.63	20.31

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnst*	*intnKP*	*cmplxt*
conditionL	-.25							
conditionM	-.25	.47						
familiarity	.05	-.16	-.11					
valence	.02	-.02	-.10	-.08				
intensity	-.01	-.05	.12	-.15	.28			
intnstyKyPh	.03	.10	-.24	.23	.02	-.64		
complexity	.04	-.12	-.11	.44	-.08	-.05	.10	
plausibility	.05	-.20	-.09	-.24	-.09	.14	-.11	-.08

AIC	BIC	logLik	deviance	df.resid
111004.5	111101.1	-55489.2	110978.5	12492

Model IV+$_{IKP}$: Effect of Metaphoricity, a gradient measure, during the rest of sentence after phrase, 3 x 667 ms. silence):

pupilSize_acrossTrial ~ Metaphoricity + familiarity + valence + intensity + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	103.70	0.89	116.05	< .0001
metaphoricity	0.47	0.21	2.23	.02549
familiarity	-0.09	0.27	-0.32	.74870
valence	0.63	0.34	1.86	.06486
intensity	0.49	0.37	1.34	.17923
intensityKeyPhrase	0.15	0.39	0.38	.70380
complexity	-0.31	0.29	-1.04	.29871
plausibility	-0.51	0.25	-2.07	.03915

Random effects

	Variance	SD
participant (Intercept)	7.78	2.79
sentence (Intercept)	4.95	2.23
timepoint (Intercept)	2.22	1.49
Residual	412.74	20.32

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnst*	*intnKP*	*cmplxt*
metaphorcty	-.001						
familiarity	.00	.03					
valence	.001	-.16	-.10				
intensity	.00	.19	-.15	.26			
intnstyKyPh	.001	-.39	.22	.05	-.64		
complexity	.00	-.13	.42	-.07	-.07	.14	
plausibility	-.001	.02	-.27	-.10	.13	-.09	-.10

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
111007.1	111096.3	-55491.5	110983.1	12493

Models $\mathrm{I}_{\mathrm{IKP}}$ -- $\mathrm{IV}_{\mathrm{IKP}}$

We also explored a different post hoc analysis in which we artificially matched M, L, and C conditions on IKP. Specifically, we removed 11 triples for which M was rated higher than L on IKP from dilation data. We then imputed values for those cells using the mice package in R, which allows the distribution of the rest of the data to be respected while offering the same power as the original analyses. Since IKP correlates with Intensity (at the sentence level), and since IKP did not contribute to any models when simply added (Models $\mathrm{I}_{+\mathrm{IKP}}$ through $\mathrm{IV}_{+\mathrm{IKP}}$), we substituted IKP for the preregistered sentence intensity factor and evaluated the new models (Models $\mathrm{I}_{\mathrm{IKP}}$ through $\mathrm{IV}_{\mathrm{IKP}}$). With one minor exception, the key findings are all replicated, showing a significant effect of metaphoricity/condition in all models. In particular, Model $\mathrm{I}_{\mathrm{IKP}}$, as in Model I, shows that the key phrase evoked more dilation in the Metaphor condition than the reference C(oncrete) condition ($\beta_M = 2.35$, $SE = 0.65$, $p = .0003$). A comparison of models with and without condition as fixed effect also show a significant advantage of including it: $\chi^2(2)$ = 13.28, p = .001. The one difference is that now the Literal condition also differs from the C condition in Model I$_{IKP}$ (β_L = 1.45, SE = 0.66, p = .029). In Model II$_{IKP}$, as metaphoricity increased, we find a significant increase in dilation at the key phrase (β_M = 0.978, SE = 0.28, p = .0006). In Model III$_{IKP}$, as predicted, the Metaphor condition evoked more dilation during the period extending until 2 seconds after the key phrase in comparison to the Concrete condition, with Bonferroni correction applied [requiring p < .0125] (β_M =1.98, SE = 0.45, p < .0001), while Literal was not significantly different than the C condition after correction requiring p < .0125 (β_L = 0.99, SE = 0.46, p = 0.031). Finally, in Model IV$_{IKP}$, as metaphoricity increases, dilation again increases with Bonferroni correction applied (β_M = 0.78, SE = 0.20, p = .0001). These additional analyses confirm that metaphoricity leads to an increase in pupil dilation, with IKP (artificially) matched and included in all models. These full models are provided below.

Model I$_{IKP}$: Effect of Condition (M, L vs C), a categorical variable, at the key phrase

pupilSize_atKeyPhrase ~ Condition + familiarity + valence + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence)

	Fixed effects			
	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	102.83	0.56	183.54	< .0001
conditionL	**1.45**	0.66	2.18	**.029**
conditionM	2.35	0.65	3.62	**.0003**
familiarity	-0.16	0.34	-0.48	.634
valence	0.04	0.35	0.13	.900

intensityKeyPhrase	0.69	0.35	1.99	**.048**
complexity	-0.38	0.34	-1.14	.257
plausibility	-0.19	0.31	-0.60	.547

Random effects

	Variance	SD
participant (Intercept)	3.73	1.93
sentence (Intercept)	2.71	1.65
Residual	209.79	14.48

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnKP*	*cmplxt*
conditionL	-.58						
conditionM	-.58	.49					
familiarity	.10	-.15	-.09				
valence	.05	-.007	-.13	-.002			
intnstyKyPh	.06	.06	-.20	.21	.33		
complexity	.08	-.11	-.10	.40	-.03	.10	
plausibility	.10	-.17	-.10	-.22	-.09	.02	-.06

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25962.2	26028.8	-12970.1	25940.2	3150

Model II$_{\mathrm{IKP}}$: Effect of Metaphoricity, a gradient measure, at the key phrase

pupilSize_atKeyPhrase ~ Metaphoricity + familiarity + valence + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|)
(Intercept)	104.09	0.42	248.10	< .0001
metaphoricity	0.98	0.28	3.43	**.0006**
familiarity	0.04	0.34	0.12	.908
valence	-0.08	0.36	-0.22	.829
intensityKeyPhrase	0.52	0.36	1.44	.152
complexity	-0.38	0.34	-1.12	.263
plausibility	-0.07	0.31	-0.24	.811

Random effects

	Variance	SD
participant (Intercept)	3.59	1.90
sentence (Intercept)	2.99	1.73
Residual	209.81	14.49

Number of obs: 3161; participant, 61; sentence, 60

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnKP*	*cmplxt*
metaphorcty	-.001					
familiarity	.00	.05				
valence	.001	-.22	-.02			
intnstyKyPh	.002	-.33	.20	.36		
complexity	.00	-.12	.39	-.01	.13	
plausibility	-.001	-.009	-.25	-.10	.03	-.08

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
25961.8	26022.4	-12970.9	25941.8	3151

Model III$_{IKP}$: Effect of Condition (M, L vs C), a categorical variable during the rest of sentence after the key phrase, 3 x 667 ms. of silence

pupilSize_acrossTrial ~ Condition + familiarity + valence + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	101.94	0.88	115.46	< .0001
conditionL	0.99	0.46	2.15	.03140
conditionM	1.98	0.45	4.42	**< .0001**
familiarity	0.07	0.28	0.25	.80572
valence	0.14	0.40	0.35	.72543
intensityKeyPhrase	0.25	0.32	0.78	.43713
complexity	0.46	0.33	1.39	.16481
plausibility	-0.96	0.26	-3.65	**.00027**

Random effects

	Variance	SD
participant (Intercept)	6.08	2.47
sentence (Intercept)	13.55	3.68
Timepoint (Intercept)	1.43	1.20
Residual	377.88	19.44

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*cndtnL*	*cndtnM*	*fmlrty*	*valenc*	*intnKP*	*cmplxt*
conditionL	-.25						
conditionM	-.25	.47					

familiarity	.05	-.18	-.10				
valence	.03	-.01	-.14	-.10			
intnstyKyPh	.02	.13	-.23	.15	.18		
complexity	.04	-.14	-.11	.47	-.12	.08	
plausibility	.05	-.21	-.10	-.21	-.18	-.07	-.07

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
109947.1	110036.3	-54961.5	109923.1	12493

Model IV$_{IKP}$: Effect of Metaphoricity, a gradient measure, during the rest of sentence after phrase, 3 x 667 ms. silence):

pupilSize_acrossTrial ~ Metaphoricity + familiarity + valence + intensityKeyPhrase + complexity + plausibility + (1 | participant) + (1 | sentence) + (1 | timepoint)

Fixed effects

	Estimate	SE	*t*	Pr(>|t|), Bonferroni corrected significance level: ***p* < .0125**
(Intercept)	102.93	0.84	121.85	< .0001
metaphoricity	**0.78**	0.20	3.86	**.0001**
familiarity	0.26	0.28	0.94	.34710
valence	0.06	0.40	0.14	.88542
intensityKeyPhrase	0.08	0.33	0.25	.80200
complexity	0.48	0.33	1.44	.15146
plausibility	**-0.83**	0.26	-3.25	**.00118**

Random effects

	Variance	SD
participant (Intercept)	6.02	2.45

sentence (Intercept)	13.61	3.69
timepoint (Intercept)	1.43	1.20
Residual	378.02	19.44

Number of obs: 12505; participant, 61; sentence, 60; timepoint, 4

Correlation of fixed effects

	(Intr)	*mtphrc*	*fmlrty*	*valenc*	*intnKP*	*cmplxt*
metaphorcty	-.001					
familiarity	.00	.07				
valence	.001	-.21	-.12			
intnstyKyPh	.002	-.39	.15	.21		
complexity	.00	-.12	.44	-.11	.12	
plausibility	-.001	.007	-.26	-.18	-.04	-.10

AIC	**BIC**	**logLik**	**deviance**	**df.resid**
109949.6	110031.4	-54963.8	109927.6	12494

Additional preregistered analyses

As preregistered, we performed two other analyses that yielded null results, as detailed below. First, we analyzed the degree of pupil *synchrony* across participants, which is a newer measure interpreted to indicate greater salience by measuring the extent to which dilation changes are the same for different participants (e.g., Kang & Wheatley, 2017; Nencheva, Piazza, & Lew-Williams, 2020). Kang and Wheatley (2017) had found that participants' pupil dilations were most synchronized during portions of a longer narrative judged as more emotionally salient by a separate group of raters. Adopting this perspective, we hypothesized that in addition to greater dilation, metaphorical sentences would elicit greater synchronization between participants due to greater emotional engagement beginning at the key phrase. A second analysis was also preregistered, on the basis of having participants perform a verbatim recognition memory task at the end of the main pupillometry study. The memory task was motivated by the fact that emotionally engaging stimuli tend to be recognized more accurately (Kensinger & Corkin, 2003), which led us to hypothesize that the metaphorical sentences would be better remembered.

Verbatim recognition task

A verbatim recognition memory task was created on the Qualtrics platform and conducted on a computer after the completion of the pupillometry experiment. The same participants were told they would read a series of 60 sentences and be asked whether or not they had heard the exact sentence during the previous part of the experiment (see Table S1). After a practice trial with feedback, participants responded to a list of 60 sentences based on their original exposure list. The 60 sentences (20 from each condition) included 30 old sentences and 30 new sentences (with one word replaced from original). Stimuli were presented in random order and no feedback was provided. Participants were debriefed at the end of the experiment.

Table S1: 3 Examples of old and new sentence triples with original and modified words underlined.

Sentence Type	**Metaphorical Sentence (M)**	**Literal Sentence (L)**	**Concrete Sentence (C)**
Original	The actor gave his co-star a sweet compliment.	The actor gave his co-star a kind compliment.	The actor gave his co-star a sweet candy.

Modified	The performer gave his co-star a sweet compliment.	The performer gave his co-star a kind compliment.	The performer gave his co-star a sweet candy.
Original	The musicians felt in tune with each other during the discussion.	The musicians were in agreement with each other during the discussion.	The musicians were in tune with the piano during the concert.
Modified	The musicians felt in tune with each other during the panel.	The musicians were in agreement with each other during the panel.	The musicians were in tune with the piano during the show.
Original	The business managed to stay afloat during the recession.	The business managed to stay open during the recession.	The fishing boat managed to stay afloat during the heavy storms.
Modified	The business managed to stay afloat amid the recession.	The business managed to stay open amid the recession.	The fishing boat managed to stay afloat amid the heavy storms.

Signal detection theory or d-prime (d’) was used to measure the ability to correctly distinguish old and new sentences. For each participant and each of the three conditions, a d’ value was calculated. One-sample t-tests confirmed d’ values were significantly above chance (d’ = 0) in each condition (1 sided *t*-test for M sentences: $t(60) = 6.42$, $p < 0.001$; L sentences: $t(60) = 6.61$, $p < 0.001$; C sentences: $t(60) = 6.56$, $p < 0.001$).

A mixed model was constructed to determine whether there was an effect of condition and average pupil dilation during the sentence on subsequent d’ values. For each condition, average pupil dilation during the sentence, standardized, was included as a fixed effect. Pupil dilation over the entire sentence, and not the key phrase, was included because the modified phrase occurred outside of the key phrase. By-participant and by-list random intercepts were also included. By-list random intercepts were included to account for the variation between the three lists, one of which was assigned to each participant. A model comparison was performed between this full model and an identical model but with condition removed. No significant difference was found between models ($\chi2(2) = 0.17$, $p = 0.92$). In addition, a model comparison performed between this full model and an identical model but with average pupil dilation removed likewise found no significant difference between models ($\chi2(1) = 0.17$, $p = 0.68$). These results indicate that condition and pupil dilation during stimuli presentation were not significant predictors of subsequent verbatim memory. To summarize, the verbatim memory task yielded null results: participants showed non-trivial but equivalent verbatim memory for metaphorical, literal and concrete sentences.

Although we had predicted that the metaphorical sentences would be better remembered if they were more engaging, we found no evidence of more accurate verbatim recognition for metaphors. One suggested explanation might be that the concrete sentences introduced interference with the metaphorical sentences since similar literal meanings were used in both sentence types in each triple. However, this cannot explain the null finding because each participant only saw one sentence from each triple. Prior evidence for the expected effect had included novel figurative language (Reuterskiöld & Van Lancker Sidtis, 2013) rather than the conventional metaphors in the current stimuli. And conventional metaphors may not show enhanced recognition when frequency and other factors are controlled for (Beck, 2020). The lack of a correlation with pupil dilation is likewise consistent with evidence that greater pupil dilation at encoding does not predict better recognition memory (Gross & Dobbins, 2021).

Pupil synchrony

Pupil synchrony was calculated as the average pairwise correlation between participants' pupil response time series for the duration of the key phrase for each trial. Correlations were very small and no differences were found between conditions in any analysis. In particular, for each participant pair, a correlation value was calculated for sentences in each of the three sentence conditions using the R stats library. All models included by-participant and by-item intercepts and the normed factors familiarity, valence, intensity, complexity, and plausibility as fixed effects. The linear mixed models with and without condition as an additional fixed effect were no different ($\chi 2(2) = 2.03$, $p = 0.362$), and none of the fixed factors revealed a significant effect (all $p > .35$). A second mixed model comparison that considered Metaphoricity and Imageability as gradient factors also failed to find significant differences between models that included or excluded these factors (Metaphoricity: $\chi 2(1) = 0.49$, $p = 0.485$; Imageability: $\chi 2(1) = 0.83$, $p = 0.361$), or any hint of an effect of any other fixed factor (all $p > .35$). The lack of coupled response across participants may be due to the fact that our stimuli contained unrelated sentences, compared to longer and personal narratives used in Kang and Wheatley (2017).