Journal of Experimental Psychology: General

Beyond Analogy: Pragmatic Constraints on Metaphor Production and Comprehension

Nicholas Ichien and Keith J. Holyoak

Online First Publication, April 13, 2026. https://dx.doi.org/10.1037/xge0001919

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001919

Beyond Analogy: Pragmatic Constraints on Metaphor Production and Comprehension

Nicholas Ichien and Keith J. Holyoak
Department of Psychology, University of California, Los Angeles

The psychological relationship between metaphor and analogy has been the focus of continuing controversy. Using both production and comprehension paradigms, we investigated the processing of metaphors versus formally equivalent analogies. To precisely match the metaphor and analogy tasks, we generated materials based on Aristotelian metaphors, derived by translating proportional verbal analogies in the form *A:B::C:D* into metaphors in the form "*A* is the *C* of *B*" (omitting the *D* term). For both analogies and metaphors, we had participants rate their goodness (Experiment 1), produce expressions (Experiments 2a, 2b, and 2c), and select the best completion (Experiment 3). People assigned higher goodness scores to analogies and metaphors that were relationally valid, and the two expression types showed similar sensitivity to individual differences in both executive functions and verbal ability. But although a valid analogical basis proved *necessary* to create a successful metaphor, this was not sufficient. We found that good metaphors honor additional pragmatic constraints predictable from a set of lexical variables related to word generalness concreteness and familiarity. Good metaphors (but not analogies) describe topic concepts (denoted by *A* terms) in terms of their more general and familiar domain concepts (denoted by *B* terms) and more general vehicle concepts (denoted by *C* terms). Our findings demonstrate both formal commonalities and pragmatic differences in the processing of metaphors and analogies.

Public Significance Statement
Ever since Aristotle, philosophers, linguists, and psychologists have struggled to pin down how metaphor (a form of language) relates to analogy (a form of reasoning). We investigated the relationship between the processing of metaphors versus formally equivalent analogies. Metaphors depend on valid analogical relationships; however, in order to provide insight into its topic, a metaphor also has to obey additional pragmatic constraints related to properties of words. Metaphor involves reasoning but goes beyond it.

Keywords: metaphor, analogy, individual differences, reasoning, semantic cognition

Supplemental materials: https://doi.org/10.1037/xge0001919.supp

Over two millennia before the dawn of modern cognitive science, Aristotle proposed in his *Poetics* (c. 335 BCE) that metaphor is rooted in analogy. Metaphors and analogies typically aim to convey insight about one thing (the *topic* or *target*) in terms of another, generally more familiar concept (the *vehicle* or *source*). Metaphor and analogy appear to be intimately related (Gentner et al., 2001), yet despite decades of analysis and experimentation by philosophers, linguists, psychologists, neuroscientists, and artificial intelligence

Agnieszka Konopka served as action editor.
Nicholas Ichien https://orcid.org/0000-0002-0928-0809
Materials, data, and analysis scripts for all experiments are available on the Open Science Framework at https://osf.io/vb52z/. Experiments 1 and 2a were not preregistered. Experiment 2b's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/tkb9p. Experiment 2c's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/f56mk. Experiment 3's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/wfyvz. Data were analyzed using R 4.3.1. The ideas and data appearing in this article have not been disseminated publicly at a conference meeting, posted on a listserv, shared on a website, or otherwise. No artificial intelligence–assisted technologies were used in the creation of this article. Research procedures were approved by the University of California, Los Angeles Institutional Review Board (IRB Protocol 25-0307). Both authors declare no conflicts of interest.
reparation for this article was provided by National Science Foundation Social and Behavioral Science Postdoctoral Research Fellowship Grant 2313985 awarded to Nicholas Ichien.

Nicholas Ichien played a lead role in data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, and visualization and an equal role in conceptualization, writing–original draft, and writing–review and editing. Keith J. Holyoak played a lead role in supervision, a supporting role in methodology, and an equal role in conceptualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Nicholas Ichien, Department of Psychology, University of California, Los Angeles, 1285 Psychology Building, Box 951563, Los Angeles, CA 90095-1563, United States. Email: ichien@g.ucla.edu

researchers, precisely *how* they are related remains an unsettled issue. An analogy is a comparison based at least in part on similar *relations*; for example, a finger is related to the hand the same way a toe is related to the foot, making finger/hand analogous to toe/foot (for a general review, see Holyoak, 2025). Analogies can be verbal, as in this example, but also can be based on visual or other sensory inputs (Weinberger et al., 2022). Analogical reasoning is thus multimodal in nature, functioning as a core component of fluid intelligence (Snow et al., 1984).

In contrast, metaphors are more intrinsically linked to language. A metaphor arises when one concept is described using terms drawn from a distinctly different domain. Like an analogy, a metaphor involves a comparison. Sometimes the comparison is explicit in the syntax of the sentence (e.g., "Time is a thief," where "time" is the target and "thief" the source). However, in many cases, the comparison is more implicit (e.g., "Our attention is being fracked," where the verb "fracked" evokes a comparison between the exploitation of human attention by new media and a modern method for extracting natural gas from deep underground).

But although metaphors resemble analogical comparisons, there has been considerable debate regarding the extent to which the psychological processes involved in the production and comprehension of metaphors depend on explicit analogical reasoning (for a review, see Holyoak & Stamenković, 2018). The completion of verbal analogies can indeed cue the subsequent production of metaphors, suggesting some mutual overlap in processing (George et al., 2025). However, the precise relation between analogy and metaphor appears to depend on the particular metaphor under investigation. Some metaphorical comparisons that are highly conventional (e.g., "Life is a journey") can spin off a wide range of variations (e.g., "My job is a dead end") that seem to be understood with little apparent effort (Lakoff & Johnson, 1980). Metaphors may sometimes be processed as categorization statements (e.g., "thief" might be interpreted as an abstract category of which "time" is a member; Glucksberg & Keysar, 1990). Some metaphors may evoke a looser process of semantic integration in which features of word meanings are blended (e.g., semantic features of "fracked" might modify those of "attention"; Kintsch, 2000). It has been argued that explicit analogical reasoning is more likely to be engaged for novel metaphors (Bowdle & Gentner, 2005) or for more complex literary metaphors (Holyoak, 2019).

A salient gap in previous research on the relationship between metaphor and analogy is that the two have never been directly compared in a single study. To fill this gap, in the present study, we obtained judgments about closely matched metaphors and analogies, using evaluation, production, and comprehension paradigms. To generate matched materials, we made use of the syntactic form on which Aristotle focused in his *Poetics* (c. 335 BCE). To use one of his examples, we first create a standard proportional analogy in the form *A:B::C:D*, such as *old age* : *life* :: *evening* : *day*. The analogy is valid because the relation *A:B* (*old age* : *life*) is similar or identical to the relation *C:D* (*evening* : *day*). This analogy can be directly translated into a matched metaphor with the syntactic form "*A* is the *C* of *B*," that is, "Old age is the evening of life." The *A* term in the analogy becomes the target in the metaphor, playing a role defined in relation to *B*, while *C* becomes the source, which plays a matching role with respect to the unstated *D*. The metaphor is "valid" in the sense that it is based on the same relational structure as the corresponding proportional analogy.

The precise formal equivalence between proportional verbal analogies and Aristotelian metaphors allows us to examine both their commonalities and (perhaps more importantly) their differences. Analogies typically serve pragmatic functions such as helping to solve a difficult target problem or making a convincing argument. However, proportional analogies (standardly used in intelligence tests) are *nonfunctional* (Holyoak, 2025), in the sense that they serve no purpose beyond simply being evaluated as analogy problems. The source/target distinction does not clearly apply, and the terms can be freely rearranged as long as matching relations are maintained. For example, Aristotle's example could be transformed into *day* : *evening* :: *life* : *old age*. This version would correspond to the potential metaphor, "Day is the life of evening."

But while the transformed analogy remains valid, the corresponding metaphor does not seem very meaningful. A successful or "apt" metaphor necessarily serves the function of illuminating its target by comparing it to the source. Aptness is high when the source is perceived as providing a unique and accurate description of the target, such that salient properties of the source are attributed to the target (Al-Azary & Katz, 2021; Blasko & Connine, 1993; Chiappe et al., 2003; Jones & Estes, 2006). We hypothesized that while Aristotelian metaphors necessarily correspond to valid proportional analogies, they are also subject to additional pragmatic constraints that determine their aptness. In Aristotle's original example, "Old age is the evening of life" seems to offer insight into the nature of old age and, hence, constitutes an apt metaphor. In contrast, "Day is the life of evening" does not seem informative or interesting. To clarify these pragmatic constraints, we focus on two elements that we propose distinguish Aristotelian metaphors from analogies: the relation between *A* and *B* terms and that between *A* and *C* terms. In metaphors, these elements respectively constitute the *topic–domain* relation and the *topic–vehicle* relation. We discuss each in turn below.

Topic–Domain Relations

A basic feature of Aristotelian metaphors (i.e., "*A* is the *C* of *B*") is explicit mention not only of the topic (i.e., the *A* term) but also of its *conceptual domain* (i.e., the *B* term). For instance, in "Old age is the evening of life," "life" denotes the conceptual domain of the topic "old age." This feature contrasts with the "*A* is *C*" format (commonly used in psychological research on metaphor), in which the domains of both topic and vehicle are left implicit. The more explicit Aristotelian format highlights the conceptual asymmetry between the *A* term ("old age"), which refers to a role, and the *B* term ("life"), which refers to a broad domain.

As part of an analogy, an asymmetric semantic relation linking *A* to *B* implies a meaningful converse relation linking *B* to *A* (Lu et al., 2019): For example, just as old age is a part of life, life has old age as one of its parts. Swapping term *A* with *B* (and *C* with *D*) thus yields a valid analogy—*life* : *old age* :: *day* : *evening*. But though the corresponding metaphor, "Life is the day of old age," remains "valid," it conspicuously lacks aptness. The pragmatic violation is that the topic of an apt metaphor should be a specific role, not a general domain. It follows that if an analogy with the topic–domain relation *A:B* makes an apt metaphor, its converse *B:A* generally will not, because only the former establishes a role as the topic. This asymmetry is consistent with the view that at least some types of metaphors are interpreted as category statements (Glucksberg & Keysar, 1990). When the topic term *A* constitutes a

member of its domain *B*, this will encourage a metaphorical interpretation in which *A* is placed into an ad hoc category in the corresponding domain linked to *C* (i.e., the unstated *D* term). Reversing the *A* and *B* terms will generally block the formation of an ad hoc category.

To test this hypothesis, we exploited an asymmetry in the *generalness* of domains relative to particular roles within those domains. If an *A* term and *B* term instantiate a topic–domain relation, then, in addition to being semantically related to one another, the *B* term should denote a more general concept than does the corresponding *A* term. More concretely, between *old age* and its associated domain *life*, the latter is a more general term than is the former. In the present study, we hypothesized that in apt metaphors (but not in formally matched analogies), the generalness of *B* (i.e., topic domains) would be higher than that of *A* (i.e., topics). In addition, we considered other potential asymmetries in lexical properties of *A* and *B* terms (see below).

Topic–Vehicle Relations

A second key relational element in a metaphor is the link between its topic and the vehicle to which it is compared. In an Aristotelian metaphor, the vehicle is signaled by *C*, which typically plays a role with respect to the unstated domain *D*. Vehicle domains are critical in metaphor comprehension and are dissociable from any particular concept that instantiates them. For example, Keil (1986) showed that once elementary school children (Grades K, 2, and 4) were able to comprehend metaphors describing a topic (e.g., "the idea") in terms of a particular instance of a given vehicle domain (e.g., plant terms, as in "The idea blossomed"), they were also able to comprehend distinct metaphors that instantiated another concept drawn from the same vehicle domain (e.g., "The idea was mowed down"). Knowledge about the relation between topic and vehicle domains thus appears to be a crucial determinant of which metaphors an agent can grasp (see also Kelly & Keil, 1987). Similarly, highly familiar conceptual domains, such as body parts or other human traits, constitute particularly useful domains from which children draw source terms for analogies (Gentner, 1977; Inagaki & Hatano, 1987).

A major theoretical issue in research on metaphor concerns the relation between the topic and the vehicle, which is commonly asymmetrical. People tend to prefer canonical metaphors (e.g., "My job is a jail") over variants in which topic and vehicle terms are reversed (e.g., "A jail is my job"; Chiappe et al., 2003; Connor & Kogan, 1980; Glucksberg et al., 1997; Ichien et al., 2024). The categorization account of metaphor comprehension provides one account of this asymmetry, according to which a metaphor describes a topic as a member of an ad hoc category denoted by the vehicle term (Glucksberg & Keysar, 1990). Proponents of this view concede that metaphors are not always interpreted in this way (Glucksberg & Haught, 2006). They may sometimes be processed as comparisons (e.g., similes or analogies), gradually acquiring an asymmetric instance–category relation with conventionalized use (Bowdle & Gentner, 2005; Ortony, 1979).

However, Glucksberg and Haught (2006) showed that semantic properties of novel metaphors can prompt an inherently asymmetric category interpretation in the absence of conventionalization, particularly when vehicles include a modifier that prompts a necessarily figurative interpretation (e.g., "advertising wart" as in "A billboard is an advertising wart"). More broadly, topic–vehicle asymmetries in unconventionalized metaphors may involve the relative generalness of topics and vehicles: Terms that denote more general concepts should be better suited to serve as the vehicle/category in a metaphor. It is thus possible that relative generalness may distinguish both the *A* and *B* terms in a metaphor (*A* plays a specific role within the more general domain *B*) and also the *A* and *C* terms (the topic *A* is introduced as a specific member of a more general ad hoc category denoted by *C*). The two pairs of terms are distinguished from each other by the further constraint that *A* and *B* are closely associated semantically whereas *A* and *C* are semantically distant.

Another attempt to clarify the topic/vehicle asymmetry, conceptual metaphor theory, focuses on the relative *concreteness* of the topic and vehicle, proposing that metaphor enables reasoners to use concrete vehicles to think about abstract topics (Kövecses, 2002; Lakoff & Johnson, 1980). Consistent with this hypothesis, Katz (1989) found an overall preference for completing metaphors (e.g., "Sociology is the ____ of science") with vehicles drawn from concrete rather than abstract domains (e.g., instances of *birds* as opposed to *nations*; see also Xu, 2010). Moreover, participants scoring higher on a test of analogical reasoning were more likely to select concrete vehicles to complete metaphors with more abstract topics (Katz, 1989).

However, contrary evidence demonstrated an overall preference for metaphors, relative to similes, when vehicles were abstract (Gibb & Wales, 1990; see also Harris et al., 2006). One attempt at resolving these discrepant findings involved the relation between concreteness and semantic neighborhood density (the number of close semantic associates a term has; Al-Azary & Buchanan, 2017; Reilly & Desai, 2017). Al-Azary and Buchanan (2017) demonstrated that this variable interacted with concreteness in metaphor processing, such that topic–vehicle asymmetries only occurred in metaphors based on terms with dense (rather than sparse) semantic neighborhoods, in which case participants found metaphors with abstract topics and concrete vehicles more sensible than those with concrete topics and concrete vehicles.

But more broadly, the hypothesis that concreteness is the core lexical variable distinguishing topic from vehicle has been challenged. Winter and Srinivasan (2022) showed that asymmetries in concept pairs involving metaphor-based semantic change (e.g., topic terms denoting *bark* tend to undergo meaning change triggered by vehicle terms denoting *skin*, but not vice versa) were better predicted by relative word frequency, rather than concreteness. These findings suggest that the apparent influence of concreteness may be better characterized as an emphasis on some sort of familiarity (Noble, 1954) or accessibility (Dancygier & Sweetser, 2014; also see Littlemore et al., 2018). Rather than describing *abstract* topics in terms of *concrete* vehicles, metaphors may tend to describe *obscure* topics in terms of *familiar* vehicles. Couching the asymmetry between topics and vehicles in terms of familiarity is consistent with the typical characterization of asymmetries between source and target domains in analogies. The source is generally more familiar or better understood than the target, so that a reasoner's knowledge of the source (specifically, of functional relations among its elements) can provide a useful basis for elaboration of the target (Hesse, 1966). Skepticism about the direct relevance of concreteness to metaphor comprehension dovetails with broader concerns over whether it is a major factor influencing lexical processing, over and above established psycholinguistic variables such as context availability,

familiarity, and age of acquisition (Brysbaert et al., 2016; Connell & Lynott, 2012; Kousta et al., 2011; Löhr, 2022, 2024).

To address this controversy regarding the basis for asymmetries between topics and vehicles in metaphors (which in turn may bear on the relation between metaphor and analogy), in the present study, we related processing of both metaphors and analogies to measures of *generalness*, *concreteness*, and *familiarity* as measured by norms of *word prevalence*, an estimate of the proportion of language speakers in a population who know a given word (Brysbaert et al., 2019).

Individual Differences in Metaphor Processing

In addition to examining the relations among individual terms within metaphors and analogies, we also assessed how individual differences in cognitive abilities impact people's ability to process each kind of expression. Specifically, we assessed the impact of fluid intelligence (closely related to executive functions) and verbal ability in metaphor and analogy processing. There is evidence that both of these factors reliably predict individual variability in metaphor comprehension (Beaty & Silvia, 2013; Chiappe & Chiappe, 2007), but their relative importance varies depending on the nature of the metaphors (Stamenković et al., 2019, 2020, 2023). Using a task involving cued production of metaphors, Beaty and Silvia (2013) found that fluid intelligence was strongly related to the production of creative novel metaphors. The general picture is that verbal ability impacts processing across all metaphor types, whereas fluid intelligence (on which analogical reasoning depends heavily) plays a greater role for more complex metaphors, especially when presented without a supportive context.

To summarize, the central goal of the present study was to directly compare interpretations of proportional analogies with those of corresponding metaphors. We aimed to confirm that Aristotelian metaphors are grounded in valid analogies and to determine what additional pragmatic factors may constrain the aptness of metaphors—that is, what lexical qualities make a metaphor seem insightful? We investigated both the production and comprehension of analogies and metaphors, while also assessing the impact of individual differences in both fluid intelligence and verbal ability.

Transparency and Openness

Materials, data, and analysis scripts for all experiments are available on the Open Science Framework at https://osf.io/vb52z/. Experiments 1 and 2a were not preregistered. Experiment 2b's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/tkb9p. Experiment 2c's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/f56mk. Experiment 3's hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/wfyvz. Data were analyzed using R 4.3.1.

Experiment 1

The first goal of Experiment 1 was to assess whether metaphor quality is subject to the same *formal* constraints that determine analogy quality. Specifically, we sought to confirm that proportional metaphors derived from valid analogies (using a procedure described in Aristotle's *Poetics*, c. 335 BCE) make better metaphors than those derived from invalid analogies. The second goal of this experiment was to identify nonformal, pragmatic features that might dissociate metaphor quality from analogy quality. To do so, we elicited and analyzed explicit ratings of both metaphor and analogy goodness from human participants. Experiment 1 was not preregistered, but materials, data, and analysis code are available at https://osf.io/vb52z/.

Method

Participants

We recruited a total of 202 participants from Prolific Academic (M_{age} = 37.38, SD_{age} = 13.54; 82 women, 114 men, two nonbinary people, and four people who did not report their gender). Selection criteria included English fluency and an approval rate over 98% on Prolific. We estimated that this task would take about 10 min and paid participants $2 (a rate equivalent to $12 per hour). Research procedures were approved by UCLA's Institutional Review Board.

Materials

We derived metaphors (e.g., "A furnace is a pocket of ash") and analogy phrases (e.g., "A furnace is to ash, just as a pocket is to lint") from a subset of 20 four-term analogies introduced in Green et al. (2010). Green et al.'s original stimuli were constructed out of 40 word-pair triplets, each consisting of an *A:B* pair of source terms (e.g., *ash : fireplace*), *near C:D* target terms that are semantically associated with their source terms (e.g., *soot : chimney*), and *far* target terms that are semantically distant from their source terms (e.g., *lint : pocket*).

Early work established a general preference for metaphors involving terms from semantically distant domains (relative to semantically similar domains; Katz, 1989; Tourangeau & Sternberg, 1981, 1982). For example, "A furnace is a chimney of soot," which is derived from a near analogy, barely makes sense as an expression of any kind, let alone as an apt metaphor. Accordingly, we used only far analogies to create metaphor and analogy phrase stimuli in the present experiments. These materials are similar to those used in Katz (1989) in that they each relate instances of distinct semantic categories, using a mix of concrete categories (e.g., birds, fish, human, dwellings, and sports) and abstract categories (e.g., fuels, nations, sciences, and weather phenomena). However, whereas the equivalents of *A* and *B* terms and of *C* and *D* terms in Katz's materials exclusively instantiate an instance–category relation (i.e., *United States : nation*), the present materials involve word pairs instantiating a wide range of semantic relations (e.g., *soot : chimney* instantiates a thing–place relation, and *flock : goose* instantiates a whole–part relation), enabling us to examine a more diverse range of metaphorical expressions.

From the 40 far analogies provided by Greene et al., the present authors selected 20 "basis" analogies that seemed to allow generation of at least one reasonably apt metaphor. These basis analogies are provided in our Supplemental Material. For each of these basis analogies, we constructed analogy phrases expressing one of four formally valid variants in which the source and target pairs shared a common relation. In addition to the canonical *A:B::C:D* form (e.g., "A furnace is to ash, just as a pocket is to lint"), valid forms were *B:A::D:C* ("Ash is to a fireplace, just as lint is to a pocket"), *C:D::A:B* ("A pocket is to lint, just as a furnace is to ash"), and *D:C::B:A*

("Lint is to a pocket, just as ash is to a furnace"). We then converted each valid analogy form into a corresponding "valid" metaphor by rearranging their terms according to the formula described in Aristotle's *Poetics*—*A:B::C:D* corresponds to *A:C:B* (e.g., "Ash is lint of a fireplace"), *B:A::D:C* corresponds to *B:D:A*, *C:D::A:B* corresponds to *C:A:D*, and *D:C::B:A* corresponds to *D:B:C*. This procedure yielded 80 valid analogy phrases and 80 "valid" metaphors.

Each basis analogy was also used to construct an analogy phrase expressing a formally invalid variant (e.g., *D:B::A:C*, "A pocket is to a fireplace, just as ash is to lint") and a corresponding "invalid" metaphor (e.g., *D:A:B*, "A pocket is ash of a fireplace"), yielding 20 invalid analogy phrases and 20 matched "invalid" metaphors. Altogether, the stimuli used in Experiment 1 thus consisted of 100 analogy phrases and 100 corresponding metaphors, with five phrases of either kind derived from each basis analogy (i.e., four phrases that were formally valid variants of a given basis analogy and one phrase that was an invalid variant).

Procedure

All participants completed a task in which they were instructed to rate the quality of a series of expressions. Metaphors are typically evaluated on aptness, which reflects the extent to which a metaphor's vehicle or source term provides a unique and accurate description of the topic or target, such that salient properties of the vehicle are attributed to the latter (Al-Azary & Katz, 2021; Blasko & Connine, 1993; Chiappe et al., 2003; Jones & Estes, 2006). However, because aptness does not clearly apply to analogies, we adopted the more neutral language of "goodness" that was similarly appropriate for both metaphors and analogy phrases. On each trial, participants were presented with either an analogy phrase or a metaphor, and they were asked to rate that expression on a symmetric 7-point scale with the following labeled points: −3 = *terrible*, −2 = *very bad*, −1 = *bad*, 0 = *neither good nor bad*, 1 = *good*, 2 = *very good*, and 3 = *excellent*. Each trial explicitly instructed participants to provide their ratings according to either "how good of an analogy" or "how good of a metaphor" they were shown, depending on whether that trial displayed an analogy phrase or a metaphor. Trials eliciting ratings of analogy phrases were presented in a separate block from those eliciting ratings of metaphors, and the order of these blocks was counterbalanced across participants. Each block consisted of 10 trials: eight derived from a formally valid analogy and two derived from an invalid analogy. Trials within a block were presented in a random order.

An individual participant was shown one of 10 stimulus lists that systematically varied the particular expression that was derived from each basis analogy and whether that expression was an analogy phrase or a metaphor. In the entire task, 20 ratings in total were elicited from each participant, and each rating was of an expression derived from a distinct basis analogy. This procedure ensured that a given participant was never shown multiple expressions derived from the same basis analogy and thus never rated both an analogy phrase and its corresponding metaphor.

Results and Discussion

Analysis Software

The analyses reported below used the following R packages in R Version 4.3.1 (R Core Team, 2023). The *ordinal* package was used to fit cumulative link mixed-effects models of goodness ratings (Christensen, 2024), we used the *buildmer* packages to determine the maximal random-effect structure that would converge when fitting each model (Barr et al., 2013; Voeten, 2023), and we used the *emmeans* package to assess estimated marginal means and trends of fit models (Lenth, 2023). We report differences between estimated means as M_{diff} and estimated marginal trends as β. Statistical models adopted contrast coding for all categorical predictors (i.e., expression type and expression validity).

Goodness Ratings of Expressions Derived From Valid Versus Invalid Analogies

An average of 20 participants rated each expression (SD = 4, range = [14, 29]). Table 1 provides descriptive statistics for these ratings, and Figure 1 shows histograms of these ratings, broken down by expression type.

In order to confirm that participants rated expressions derived from valid analogies as better than expressions derived from invalid analogies, we fit a cumulative link mixed-effects model of trial-level ratings with a logit link function and a symmetric threshold (reflecting the −3 to +3 rating scale that was symmetric about 0). This model had the following two-way interaction fixed-effect term: *Expression-Type* (metaphor vs. analogy phrase) × *Expression-Validity* (valid vs. invalid); as well as the two following random slope terms: (1 + *expression-type* × *expression-validity*|*participant*) and (1 + *expression-type* + *expression-validity*|*basis-analogy*), instantiating the maximal random-effect structure for this model. A likelihood ratio test comparing this full model to a reduced model that lacked the interaction term but that was otherwise equivalent to the full model showed that omitting the interaction term increased model prediction error, ΔAIC = 15, $\chi^2(1) = 17.11$, $p < .001$. This result suggests that the impact of analogy validity on goodness ratings of expressions varied as a function of expression type. Specifically, expressions derived from valid analogies were rated as better than those derived from invalid analogies, both for metaphors ($M_{\text{diff}} = 1.85$, $SE = 0.18$, $z = 10.63$, $p < .001$) and analogy phrases ($M_{\text{diff}} = 2.64$, $SE = 0.21$, $z = 12.60$, $p < .001$). This result confirms that the distinction between expressions derived from valid versus invalid analogies impacts people's judgments of expression quality for both analogy phrases and their corresponding metaphors.

Overall, analogy phrases derived from valid analogies were rated as better than corresponding metaphors derived from valid analogies ($M_{\text{diff}} = 0.77$, $SE = 0.13$, $z = 5.74$, $p < .001$). However, there was no general difference in ratings of expressions derived from invalid analogies ($M_{\text{diff}} = 0.02$, $SE = 0.20$, $z = 0.11$, $p = .92$), suggesting that participants judged metaphors and analogy phrases derived from invalid analogies as similarly mediocre. It is possible that

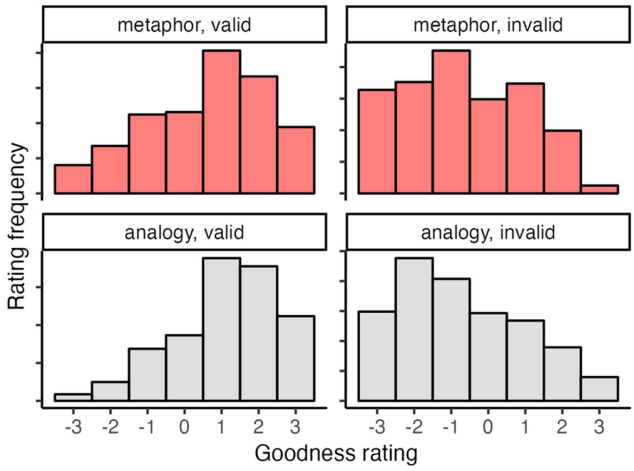
Table 1
Descriptive Statistics for Ratings of Goodness of Expressions, Broken Down by Expression Type and Relational Validity

Expression validity	Metaphor	Analogy	Overall
Valid	1.56 (0.91)	2.10 (0.80)	1.83 (0.71)
Invalid	0.34 (1.41)	0.31 (1.41)	0.32 (1.19)

Note. Cell values indicate the mean ratings, and parenthetical values indicate the standard deviations.

Figure 1

Histograms of Goodness Ratings (Experiment 1), Broken Down According to Condition (Metaphors at Top, Analogy Phrases at Bottom), and the Validity of the Analogy From Which a Given Expression Was Derived (Valid on Left, Invalid on Right)



Note. The *y*-axis reflects the frequency with which expressions were given the goodness rating specified along the *x*-axis. See the online article for the color version of this figure.

simply converting valid analogies to metaphors degrades expression quality. However, an alternative explanation is that, among expressions derivable from valid analogies, participants had stronger preferences between particular metaphors than they did between particular analogy phrases (among which they were more or less indifferent). For example, whereas participants showed a clear preference for the metaphor "A constellation is a flock of a star" ($M_{\text{rating}} = 5.32$, $SD_{\text{rating}} = 1.45$) over "A star is a goose of a constellation" ($M_{\text{rating}} = 3.39$, $SD_{\text{rating}} = 1.50$), they rated the corresponding analogy phrases "A constellation is to a star, just as a flock is to a goose" ($M_{\text{rating}} = 5.69$, $SD_{\text{rating}} = 1.08$) and "A star is to a constellation, just as a goose is to a flock" ($M_{\text{rating}} = 5.35$, $SD_{\text{rating}} = 1.43$) very similarly.

To adjudicate between these two explanations, we compared the goodness ratings for those valid metaphors for which the mean rating was highest among the four valid variants of a given basis analogy (e.g., "A constellation is a flock of a star") with those for the analogy phrases with the corresponding form (e.g., "A constellation is to a star, just as a flock is to a goose"). If participants had a ubiquitous preference for analogy phrases over metaphors, then even the best rated metaphors should be judged inferior to their corresponding analogy phrases. On the other hand, if participants' general preference for analogy phrases over metaphors is instead explained by low ratings for dispreferred metaphors, then this apparent preference should disappear when we consider only the preferred metaphors. Consistent with the latter explanation, a paired Wilcoxon signed-rank test of mean goodness ratings revealed no reliable difference between the highest rated metaphors ($M_{\text{rating}} = 5.24$, $SD_{\text{rating}} = 0.50$) and their corresponding analogy phrases ($M_{\text{rating}} = 5.27$, $SD_{\text{rating}} = 0.52$; $W = 188$, $p = .76$). Thus, although formal validity appears to have contributed to the rated goodness of analogy phrases fairly uniformly across valid variants, participants appeared to adopt further criteria when rating metaphors. In the following analyses, we examine possible candidates for these criteria. We also provide direct analyses of the lexical features that distinguish the highest rated and lowest rated metaphors in the Supplemental Material.

Goodness Ratings and Lexical Feature Norms

We used lexical feature norms to clarify the determinants of goodness ratings that go beyond the formal validity of the analogies from which expressions were derived. We report the results from two analyses. The first analysis aimed to dissociate metaphors and analogies through a constraint based on a topic–domain relation between *A* terms and *B* terms that applies in metaphors but not analogies. To do so, we were particularly interested in the relative *generalness* of *A* terms and *B* terms. In addition to assessing relative generalness between *A* and *B*, we also analyzed relative values in two further lexical features, word *familiarity* and *concreteness*.

The second analysis characterized the constraints on *A* terms and *C* terms, which respectively constitute topics and vehicles in metaphors. This analysis served to adjudicate whether topic–vehicle asymmetries in metaphors are better explained by differences in word *generalness*, *familiarity*, or *concreteness*. We specifically examined whether better rated metaphors (and possibly analogies) were those with more general, concrete, or familiar *A* terms or *C* terms. We introduced our own measure of generalness, which reflects participants' explicit judgments of how general is the concept denoted by a given word. To collect these judgments, we elicited best–worst judgments among each of the four terms constituting the basis analogies in our data set. In the context of generalness, these judgments reflected which of four terms was the most general and which was the least general (or most specific). We used this approach rather than rating scales for two reasons. First, Hollis and Westbury (2018) showed that norms derived from best–worst judgments of the semantic features age of acquisition, valence, arousal, and concreteness yield better predictive validities of lexical decision response times than those derived from rating scales. Second, this format enabled us to elicit comparisons between terms within each basis analogy. The Supplemental Material provides details of our procedure both for collecting these judgments and for computing the corresponding norms using Lipovetsky and Conklin's (2014) analytic best–worst solution.

We adopted established norms to estimate familiarity and concreteness. To measure word familiarity, we used norms of word "prevalence" computed by Brysbaert et al. (2019), which provide data from a large-scale task in which participants indicated whether they knew a given word or not. These norms reflect word knowledge as the proportion of language users who know a given word. In addition to prevalence, we also considered frequency as an additional measure of familiarity. Specifically, we attempted to use Zipf frequencies, standardized logarithmic values of word frequencies reported in the SUBTLEX-US norms, which reflect the occurrence of words as used in American television shows and films (Brysbaert & New, 2009; van Heuven et al., 2014). Although frequency norms were highly correlated with prevalence norms ($\rho = .40$), they were correlated to an even greater extent with our own generalness norms ($\rho = .62$), making models including both frequency and generalness particularly difficult to interpret. We thus decided to omit frequency. Our Supplemental Material reports comparisons between models reported below for the present experiment and for Experiments 2a,

2b, and 2c that each include generalness as a predictor, with models that replace generalness with frequency but that are otherwise equivalent. Overall, these comparisons show that inclusion of generalness norms results in better model fit than does inclusion of Zipf frequencies.

Finally, we used concreteness norms collected by Brysbaert et al. (2014), in which participants rated on a scale from 1 to 5 the extent to which a given word denoted something that they could directly experience with one of their five senses. Table 2 provides descriptive statistics and pairwise correlations between lexical variables for the individual word stimuli used in our experiments.

Relation Between *A* and *B* Terms. We now describe our first set of analyses, which sought to test a potential dissociation between metaphors and analogy phrases in the role assignments of *A* terms and *B* terms. We hypothesized that Aristotelian metaphors, but not analogy phrases, would be rated better to the extent that they include an *A* term that denotes a given topic and a *B* term that denotes that topic's conceptual domain. Since the goal of the present analysis was to test a hypothesis about the *relative* features of *A* and *B* terms, we fit a statistical model that predicted trial-level goodness ratings as a function of the *differences* between *A* term lexical features and *B* term lexical features.

Participant goodness ratings as a function of these asymmetry measures are shown in Figure 2. We used a cumulative link mixed-effects model with a logit link function and a symmetric threshold (reflecting the symmetric rating scale from −3 to +3) and a series of three *Expression-Type* (metaphor vs. analogy phrase) × *Lexical Feature* two-way interaction terms, one for each of the lexical features: *relative generalness*, *prevalence*, and *concreteness*. This model also adopted the maximal random-effect structure with (1 + *condition*|*participant*) and (1 + *condition*|*basis-analogy*) random slopes.

In order to test whether metaphors were distinct from analogies with respect to differences in lexical features of *A* and *B* terms, we fit three reduced models (i.e., one for each lexical feature) that were each identical to the full model described above other than omitting a single two-way interaction term. We used likelihood ratio tests to compare the full model to each reduced model. Omitting each of the *Expression-Type* × *Relative Generalness*, $\Delta AIC = 0.0$, $\lambda_{LR}(1) = 2.96$, $p = .09$; *Expression-Type* × *Relative Prevalence*, $\Delta AIC = 2.0$, $\lambda_{LR}(1) = 3.99$, $p = .05$; and *Expression-Type* × *Relative Concreteness*, $\Delta AIC = 2.0$, $\lambda_{LR}(1) = 0.74$, $p = .39$, interaction terms failed to significantly increase model prediction error.

Table 2
Descriptive Statistics and Pairwise Correlations Between Lexical Variables for Stimuli Used Across All Studies (Significant Correlations Bolded)

		Pairwise correlation (Spearman's ρ)		
Lexical feature	*M* (*SD*)	1	2	3
1. Generalness	<0.01[a] (0.64)	—	**.28**	−.17
2. Prevalence	2.31 (0.23)	**.28**	—	−.03
3. Concreteness	4.35 (0.75)	−.17	−.03	—

[a] Generalness norms are ultimately based on the difference in proportion of trials on which a given term was selected as the most general and least general, according to best–worst judgments. If this difference is 0, the analytic best–worst solution that we use to compute these norms will evaluate to 0. Since all trials on which a given term was judged involved the same four terms (i.e., those constituting the same basis analogy), the mean rating across terms is necessarily 0 or very close to it.

In order to further interpret these results, we examined simple trends of the extent to which featural differences between *A* and *B* terms predicted goodness ratings, separately for metaphors and analogies. Details of these trends are shown in Table 3. Each trend reflects the extent to which an increase in a featural difference between *A* terms and *B* terms increases the probability of a higher goodness rating. As predicted, relative generalness between *A* terms and *B* terms predicted goodness ratings for metaphors, such that higher ratings were given to expressions in which *B* terms were more general than *A* terms. However, we also saw the same trend for analogies, which we did not predict. Corresponding tests of simple trends show that for prevalence, a greater difference between terms predicted an increase in rated goodness of metaphors but not analogy phrases. In particular, better rated metaphors were those with *B* terms that were higher in prevalence than *A* terms, suggesting that high-quality metaphors were those with a *B* term that was broadly more familiar than the *A* term. Finally, we did not observe any simple effects of relative concreteness in predicting goodness ratings for either metaphors or analogies.

Relation Between *A* and *C* Terms. Next, we aimed to identify any asymmetries between *A* terms and *C* terms, which respectively denote the topic and vehicle of Aristotelian metaphors. We specifically examined whether these asymmetries could be explained by differences in word *generalness*, *prevalence*, or *concreteness*. As in the analysis above, we aimed to explain goodness ratings in terms of the *relative* features of *A* terms and *B* terms, so we adopted the same modeling approach in predicting trial-level goodness ratings in terms of the *differences* between lexical features (Figure 3).

We fit a cumulative link mixed-effects model to predict trial-level goodness ratings as a function of featural differences between *A* and *C* terms. Our full model had a series of three two-way *Expression-Type* (metaphor vs. analogy) × *Feature* interaction terms, each of *relative generalness*, *concreteness*, and *prevalence*, and the maximal random-effect structure, with (1 + *expression-type*|*participant*) and (1 + *expression-type*|*basis-analogy*) random slope terms. We also fit three reduced models that each omitted a different two-way interaction term but that were otherwise equivalent to the full model. Omitting each of the *Expression-Type* × *Relative Generalness*, $\Delta AIC = 1.0$, $\lambda_{LR}(1) = 1.47$, $p = .22$; *Expression-Type* × *Relative Prevalence*, $\Delta AIC = 2.0$, $\lambda_{LR}(1) = 0.22$, $p = .64$; and *Expression-Type* × *Relative Concreteness*, $\Delta AIC = 2.0$, $\lambda_{LR}(1) = 0.12$, $p = .73$, interaction terms failed to increase model prediction error.

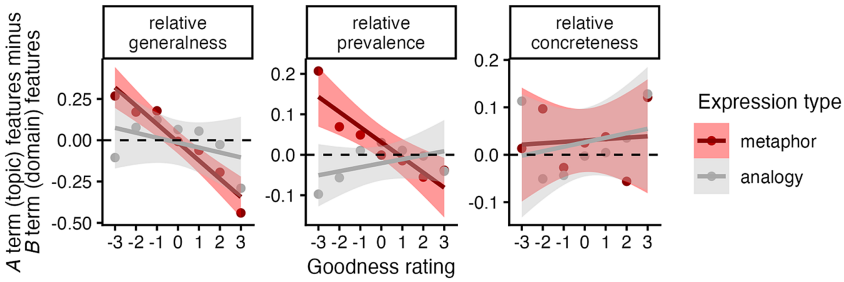
To examine these impacts further, we examined simple trends for each feature, details of which are shown in Table 4. Metaphors with vehicle terms denoting more general concepts than topic terms were rated as better, and there was no corresponding effect for analogy ratings. Broadly, these results favor the view that metaphors are distinguished from their corresponding comparison statements in virtue of instantiating ad hoc categorization (Glucksberg & Keysar, 1990). This view attributes topic–vehicle asymmetries unique to metaphor to semantic properties that contribute to the vehicle's role as a general category of the topic.

Experiment 2

The rating task in Experiment 1 revealed that the judged quality of metaphors is constrained by pragmatic factors that go beyond the

Figure 2

Goodness Ratings of Expressions Derived From Valid Analogies as a Function of the Relative Difference Between Semantic Features of A and B Terms, Which, in Aristotelian Metaphors, Correspond to the Topic and Its Conceptual Domain (Experiment 1)



Note. Values greater than 0 indicate that *A* terms are higher in some feature than *B* terms, and vice versa for values less than 0. Dashed lines indicate no difference. See the online article for the color version of this figure.

factors influencing the judged quality of analogies. In Experiment 2, we investigated whether the *production* of metaphors is also constrained by the same pragmatic factors. Experiment 2a was not preregistered. Hypotheses, methods, data collection, and analysis plan for Experiments 2b and 2c were preregistered at https://osf.io/tkb9p and https://osf.io/f56mk, respectively. Materials, data, and analysis code for all three experiments are available at https://osf.io/vb52z/.

Method

Participants

We recruited a total of 299 participants from Prolific Academic across Experiments 2a, 2b, and 2c: 99 participants in Experiment 2a (M_{age} = 32.91, SD_{age} = 10.09; 46 women, 51 men, and two nonbinary), 100 in Experiment 2b (M_{age} = 39.34, SD_{age} = 13.88; 46 women and 53 men), and 100 in Experiment 2c (M_{age} = 36.95, SD_{age} = 24.09; 45 women and 55 men). Selection criteria included English fluency and an approval rate over 98% on Prolific. We estimated that the rating task would take about 30 min and paid participants $6 (a rate equivalent to $12 per hour). Research procedures were approved by UCLA's Institutional Review Board.

Table 3

Simple Trends for the Extent to Which Differences in Lexical Features Between A and B Terms Predict Goodness Ratings for Metaphors (Left) and Analogy Phrases (Right) in Experiment 1

Lexical feature	Metaphor				Analogy phrase			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
rel gen.	**−0.30**	**0.06**	**5.31**	**<.001**	**−0.16**	**0.05**	**3.13**	**.002**
rel prev.	**−0.55**	**0.17**	**3.28**	**.001**	−0.08	0.17	0.44	.66
rel conc.	−0.11	0.06	1.78	.08	−0.03	0.06	0.54	.59

Note. Values in bold indicate statistically significant effects. β = trend estimated from the cumulative link mixed-effects model (described in Experiment 1, Results and Discussion); *SE* = standard error; *z* = *z* statistics; rel gen. = relative generalness; rel prev. = relative prevalence; rel conc = relative concreteness.

Materials and Procedure

Experiments 2a, 2b, and 2c used the same stimuli as in Experiment 1 in an arrangement-based expression production task, described below and illustrated in Figure 4.

Arrangement-Based Expression Production Task. In order to assess metaphor and analogy production, we had participants complete a task in which on each trial they were provided with four terms in a shuffled order and a phrase structure with open slots. These slots could provide the format for an analogy (e.g., "____ is ____, just as ____ is to ____"), or else a metaphor ("____ is ____ of ____"). The task was to arrange terms within this structure in order to create a novel expression (e.g., "a landscaper is to a lawn, just as a stylist is to hair"). The left panel of Figure 4 depicts example trials of this task. Participants could use their mouse to drag terms, one at a time, into a blank box within that trial's phrase structure. Terms could be moved until the participant was satisfied with the expression. Once each box contained a term, a button appeared, which participants could click to submit their response.

Each set of four terms constituted a distinct basis analogy from Green et al. (2010). An individual participant was given 10 sets of terms to arrange within a metaphor phrase structure in one block and 10 distinct sets of terms to arrange within an analogy phrase structure in another block. The order of these blocks was counterbalanced across participants, as were the particular basis analogies that were paired with metaphor and analogy blocks. An individual participant was never tested with both a specific analogy and its corresponding metaphor.

Experiments 2a, 2b, and 2c varied different properties of this arrangement task, as summarized in Table 2. One property is the phrase structure into which participants were asked to arrange terms, which is summarized in the first three rows of Table 5. In Experiments 2a and 2b, participants were given an analogy phrase structure that expressed the canonical four-term sequence: "____ is to ____, just as ____ is to ____." This phrase structure had one more blank to fill than the three-term metaphor phrase structure used across Experiments 2a, 2b, and 2c, "_____ is _____ of _____." To equate the number of terms in the phrase to be produced, Experiment 2c introduced an analogy phrase structure that likewise had three

Table 4

Simple Trends of the Extent to Which Differences in Lexical Features Between A and C Terms Predict Goodness Ratings of Metaphors (Left) and Analogy Phrases (Right)

Lexical feature	Metaphor				Analogy phrase			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
rel gen.	**−0.16**	**0.06**	**2.71**	**.007**	−0.06	0.06	1.03	.30
rel prev.	−0.17	0.19	0.96	.34	−0.06	0.18	0.31	.75
rel conc.	−0.10	0.06	1.61	.11	−0.07	0.06	1.17	.24

Note. Values in bold reflect statistically significant effects. β = trend estimated from the cumulative link mixed-effects model (described in Experiment 1, Results and Discussion); *SE* = standard error; *z* = *z* statistics; rel gen. = relative generalness; rel prev. = relative prevalence; rel conc = relative concreteness.

blanks to fill (i.e., "___ is to ____, just as ____ is to what?"), thereby posing an analogy-based question. This question format seemed more naturalistic than the possible declarative version (e.g., "___ is to ___, just as ___ is to something"). Note that although metaphors and analogies required participants to drag different numbers of terms in Experiments 2a and 2b (i.e., 3 vs. 4), the set of possible responses for either condition was matched at 24 across Experiments 2a, 2b, and 2c (since arranging three terms and omitting one term from a set of four is formally equivalent to arranging four terms from a set of four).

Experiments 2a, 2b, and 2c also varied in the specified criterion according to which participants were instructed to arrange terms, as summarized in the bottom three rows of Table 5. Experiment 2a used the same criterion for both metaphors and analogies, instructing participants to produce the "most meaningful expression" that they could, without explicitly mentioning either analogy or metaphor at any point during the task. In contrast, in Experiments 2b and 2c, participants were instructed to produce the "best metaphor" they could with the terms provided during metaphor trials and were instructed to produce the "best analogy" (Experiment 2b) or "best analogy question" (Experiment 2c) during analogy trials.

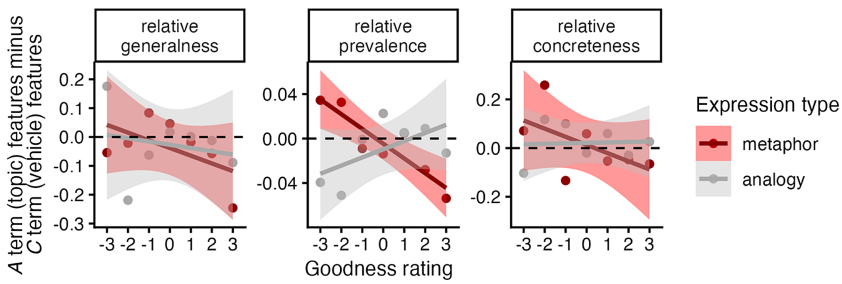
Finally, in Experiment 2a only, participants were asked to rate their own participant-generated expressions. The arrangement task in Experiment 2a instructed participants to produce expressions based on how "meaningful" they were (without explicit reference to whether an expression was a metaphor or analogy); accordingly, we asked participants to use these same criteria to evaluate the expressions they produced. After each arrangement trial in Experiment 2a, participants were shown the expression that they had just produced and were asked to rate how meaningful they found that expression on a unipolar 5-point scale, according to the following labeled points: 1 = *not at all*, 2 = *barely*, 3 = *somewhat*, 4 = *very*, and 5 = *extremely*. We adopted a unipolar scale (rather than the symmetric scale used in Experiment 1) because the dimension of "meaningfulness" does not have as clear an antonym as does "goodness" (i.e., "badness"). These ratings allowed us to assess whether participants found metaphors and analogy phrases derived from valid basis analogies to be more meaningful than those derived from invalid analogies, even when raters had generated those expressions themselves.

Individual Difference Tasks. In addition to the main task described above, participants each completed two tasks aimed at respectively measuring individual differences in domain-general reasoning ability and verbal knowledge: Raven's Advanced Progressive Matrices (RAPM) and the Mill Hill Vocabulary Test (MH; Raven, 1958). Both of these measures reliably predict individual variability in metaphor comprehension both for young adults (Beaty & Silvia, 2013; Chiappe & Chiappe, 2007) and older adults (Ichien et al., 2025), and the former is a robust predictor of individual variability in analogical reasoning (Gray & Holyoak, 2020; Ichien et al., 2022). These tasks were included in the present experiments to compare the cognitive capacities that might account for participants' ability to produce metaphors and analogies.

The present experiment used a 12-item short-form version of RAPM (Arthur et al., 1999). Each problem in this task features a 3 × 3 matrix of simple geometric shapes, in which the bottom-right cell of that matrix is missing. Participants are instructed to use the visuospatial pattern instantiated by the other cells in the matrix to guide their selection from among eight answer options of the one that best fills the missing cell to complete that pattern. The MH consists of 20 items, each of which includes a target word and six answer options. Participants are instructed to select which of the answer options is most synonymous with the target word. Participants completed the three tasks described above in a fixed

Figure 3

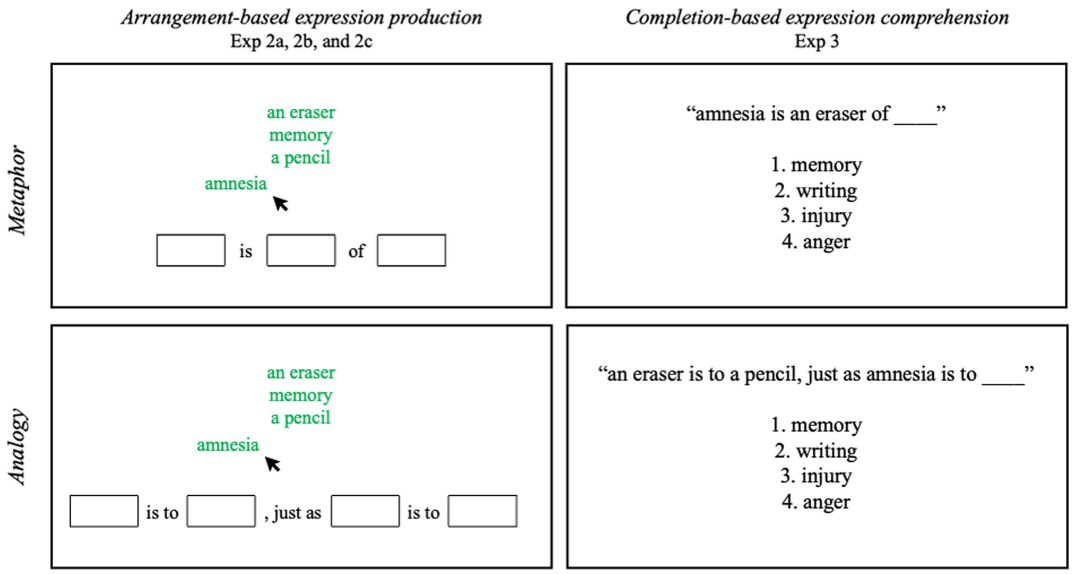
Goodness Ratings of Expressions Derived From Valid Analogies as a Function of the Difference Between Semantic Features of A and C Terms, Which, in Aristotelian Metaphors, Correspond to the Topic and Vehicle (Experiment 1)



Note. Values greater than 0 indicate that *A* terms are higher in some feature than *C* terms, and vice versa for values less than 0. Dashed lines indicate no difference. See the online article for the color version of this figure.

Figure 4

Example Trials From Production (Left) and Comprehension Tasks (Right), Respectively Used in Experiments 2a, 2b, and 2c and in Experiment 3



Note. See the online article for the color version of this figure.

order: RAPM, MH, and then the arrangement-based expression production task.

Results and Discussion

Analysis Software

The analyses reported below use the following R packages in R Version 4.3.1 (R Core Team, 2023). The *ordinal* package was used to fit cumulative link mixed-effects models of meaningfulness ratings (Christensen, 2024), the *lme4* package was used to fit logistic mixed-effects models of valid responses and of expression type (Bates et al., 2015), the *buildmer* packages were used to determine the maximal random-effect structure that would converge when fitting each model (Barr et al., 2013; Voeten, 2023), the *emmeans* package was used to assess estimated marginal means (reported as M_{diff}) and estimated marginal trends (reported as β) of fit models (Lenth, 2023), and the *psych* package was used to compute McDonald's ω_t for each individual difference measure (Revelle, 2023). Statistical models adopted contrast coding for all categorical predictors (i.e., expression type and expression validity).

Meaningfulness Ratings of Expressions Derivable From Valid Versus Invalid Analogies (Experiment 2a Only)

In order to assess whether participants rated their own expressions derived from valid analogies as more meaningful than expressions derived from invalid analogies, we fit a cumulative link mixed-effects model of trial-level meaningfulness ratings (on a unipolar 5-point scale) with a logit link function and a flexible threshold. This model had an *Expression-Type* (metaphor vs. analogy phrase) × *Expression-Validity* (valid vs. invalid) two-way interaction term, along

Table 5

Summary of Variations for Arrangement-Based Expression Production Task Across Experiments 2a, 2b, and 2c

Experiment	Metaphor	Analogy phrase
Phrase structure		
2a	"_____ is _____ of _____"	"____ is to ____, just as ____ is to ____"
2b	"_____ is _____ of _____"	"____ is to ____, just as ____ is to ____"
2c	"_____ is _____ of _____"	"____ is to ____, just as ____ is to what?"
Instructed criteria		
2a	"most meaningful expression"	"most meaningful expression"
2b	"best metaphor"	"best analogy"
2c	"best metaphor"	"best analogy question"

Note. In addition to the variations listed, Experiment 2a elicited expression ratings: After each arrangement trial, participants were shown the expression that they had just produced and were asked to rate its meaningfulness on a 5-point unipolar scale.

with two random slope terms: (1 + *expression-type* × *expression-valid*|*participant*) and (1 + *expression-type* × *expression-validity*|*basis-analogy*). A likelihood ratio test comparing this full model to a reduced model that lacked the interaction term but that was otherwise equivalent to the full model showed that omitting the interaction term did not reduce model prediction error, $\Delta AIC = 1$, $\chi^2(1) = 0.89$, $p = .34$, indicating that the impact of analogy validity on expression meaningfulness did not differ as a function of expression type. Consistent with the findings in Experiment 1, participants rated expressions derived from valid analogies as more meaningful than those derived from invalid analogies (estimated effect = 1.28, $SE = 0.17$, $z = 7.51$, $p < .001$), and they rated analogy phrases as more meaningful than metaphors overall (estimated effect = 1.05, $SE = 0.23$, $z = 4.59$, $p < .001$). Because participants in Experiment 2a only rated the expressions they actually produced, we were unable to compare ratings for the "best" metaphors as we had done in Experiment 1 (but see below for a comparable analysis of production frequencies).

Production of Expressions Derived From Valid Analogies

In all three studies included in Experiment 2, we assessed whether the expressions that participants actually produced during the arrangement task implicitly reflect a constraint to produce expressions consistent with valid analogies. Valid response rates (the rates at which participants produced expressions derivable from valid analogies) are shown in Figure 5 and in Table 6. As is evident from visual inspection, these rates were well above chance in all three studies (chance valid response rate = 4/24 or .167), both for metaphors (Experiment 2a: $V = 4{,}950$; Experiment 2b: $V = 5{,}049$; Experiment 2c: $V = 5{,}011$; all ps $< .001$) and analogy phrases (Experiment 2a: $V = 4{,}931$; Experiment 2b: $V = 5{,}028$; Experiment 2c: $V = 4{,}995$; all ps $< .001$).

In order to further characterize responses derived from valid analogies as normative responses, we assessed how individual differences in domain-general reasoning ability and verbal knowledge predicted valid response rates. Each of these constructs has been shown to predict individual variability in both metaphor comprehension and analogical reasoning (Beaty & Silvia, 2013; Chiappe & Chiappe, 2007; Gray & Holyoak, 2020; Ichien et al., 2022, 2025). Recall that we measured reasoning ability using the RAPM and verbal knowledge using the MH. Each of these measures had adequate internal reliability across Experiments 2a, 2b, and 2c (both ω_ts = .79). Table 6 shows the pairwise correlations between valid response rates and individual difference measures.

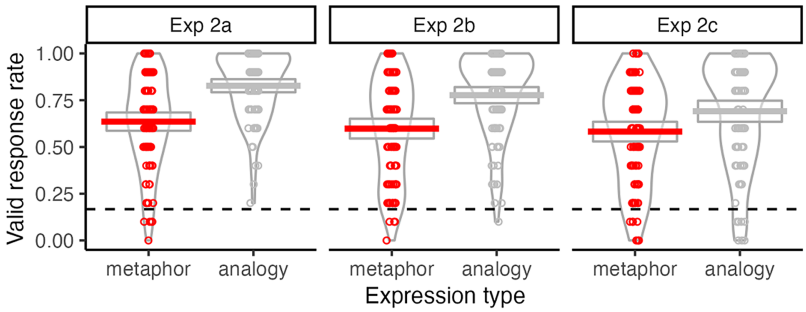
To carry out this analysis, we fit a logistic mixed-effects model to trial-level responses, coded as either valid or invalid, depending on whether or not a given response was derivable from a valid analogy. This model had three two-way interaction terms: *Study* (Experiment 2a vs. Experiment 2b vs. Experiment 2c) × *Expression-Type* (metaphor vs. analogy phrase), *RAPM-Score* × *Expression-Type*, and *MH-Score* × *Expression-Type*, respectively. This model also had the following random slope terms: (1 + *expression*|*participant*) and (1 + *expression*|*basis-analogy*).

First, we examined whether the production of valid responses relied on domain-general reasoning ability or verbal knowledge to a different extent for metaphors versus analogy phrases. For this purpose, we used likelihood ratio tests that compared the full model to two reduced models that respectively omitted the *RAPM-Score* × *Expression-Type* and *MH-Score* × *Expression-Type* interaction terms but that were otherwise identical to the full model. Removing the RAPM interaction term significantly increased model prediction error, $\Delta AIC = 4.1$, $\chi^2(1) = 4.59$, $p = .03$. Simple trends of RAPM and valid metaphor response rates ($\beta = 1.59$, $SE = 0.30$, $z = 5.23$, $p < .001$) and valid analogy phrase rates ($\beta = 2.41$, $SE = 0.33$, $z = 7.22$, $p < .001$) were both significant. However, the significant interaction term indicates that, aggregated across the three studies, participants relied more heavily on domain-general reasoning ability to produce valid analogy phrases than metaphors. On the other hand, removing the MH interaction term did not significantly increase model prediction error, $\Delta AIC = 0.5$, $\chi^2(1) = 0.96$, $p = .33$, and MH performance reliably predicted valid response rates for both expression types ($\beta = 2.94$, $SE = 0.38$, $z = 7.86$, $p < .001$). This latter result indicates that verbal knowledge was equally important in both metaphor production and analogy production.

Next, in order to assess the effect of expression type on valid response rates and determine whether it differed across studies, we used a likelihood ratio test to compare the full model described above with a reduced model that only differed in that it dropped the

Figure 5

Valid Response Rates (Rates at Which Participants Produced Metaphors [Red] and Analogy Phrases [Gray] Ultimately Derived From Valid Analogies), Broken Down by Study



Note. See the online article for the color version of this figure.

Table 6

Descriptive Statistics and Pairwise Correlations (Spearman's ρ) Between Valid Response Rates for Expression Production Tasks and Individual Difference Measures Used in Experiments 2a, 2b, and 2c (All ps < .001)

Experiment	*M* (*SD*)	Pairwise correlation (Spearman's ρ) 1	2	3	4
1. Valid metaphor					
2a	0.64 (0.25)	—	.56	.37	.34
2b	0.60 (0.27)				
2c	0.58 (0.27)				
2. Valid analogy					
2a	0.83 (0.18)	.56	—	.39	.43
2b	0.78 (0.22)				
2c	0.69 (0.29)				
3. Mill Hill					
2a	0.66 (0.17)	.37	.39	—	.20
2b	0.67 (0.18)				
2c	0.67 (0.17)				
4. Raven's progressive matrices					
2a	0.50 (0.25)	.34	.43	.20	—
2b	0.39 (0.24)				
2c	0.43 (0.26)				

Study × *Expression-Type* interaction term. Doing so did increase model prediction error, $\Delta AIC = 5$, $\chi^2(2) = 8.96$, $p = .01$, indicating that the impact of expression type on valid response rates differed by study. Though the impact of expression type on valid response rates was larger in Experiments 2a ($M_{diff} = 1.30$, $SE = 0.20$, $z = 6.64$, $p < .001$) and 2b ($M_{diff} = 1.21$, $SE = 0.19$, $z = 6.24$, $p < .001$) than in Experiment 2c ($M_{diff} = 0.71$, $SE = 0.19$, $z = 3.82$, $p = .001$), participants reliably produced valid analogy phrases at higher rates than they did valid metaphors in all three studies. These differences proved to be reliable when tested with a series of paired Wilcoxon signed-rank tests ($V = 209.5$, $V = 284$, $V = 886$; all $ps < .001$; see Figure 4). These results parallel the rating results from Experiments 1 and 2a (described above), in which participants rated analogy phrases derived from valid analogies as better and more meaningful than they did metaphors derived from valid analogies. Here, participants produced analogy phrases expressing valid analogies more often than they did metaphors that were ultimately derivable from valid analogies.

Recall, however, a follow-up analysis we performed on Experiment 1 ratings, in which the general preference for analogy phrases over metaphors was no longer present when we compared the best rated metaphors with their corresponding analogy phrases. Similarly, when we restrict our analyses in Experiment 2 to production frequencies of these best rated metaphors (selected based on goodness ratings from Experiment 1) and their corresponding analogy phrases, the overall advantage of analogies disappears. Across experiments, paired Wilcoxon tests revealed no difference between the rate at which best rated metaphors were produced (Experiment 2a: $M = 0.33$, $SD = 0.24$; Experiment 2b: $M = 0.26$, $SD = 0.18$; Experiment 2c: $M = 0.28$, $SD = 0.21$) and the extent to which the corresponding analogy phrases were produced (Experiment 2a: $M = 0.23$, $SD = 0.16$, $V = 152$, $p = .08$; Experiment 2b: $M = 0.22$, $SD = 0.12$, $V = 134$, $p = .29$; Experiment 2c: $M = 0.23$, $SD = 0.12$, $V = 123.5$, $p = .10$). Overall, there was actually a slight numerical advantage for the best rated metaphors, relative to their matched analogies. This pattern of results is consistent with the hypothesis that participants were more discriminating among valid metaphors than they were among valid analogy phrases, among which they were relatively indifferent. We test this explanation further in the next set of analyses.

Preferences Among Expressions Derivable From Valid Analogies

We hypothesized that though production of both metaphors and analogies is constrained by formal factors (as reflected in participants' preference toward valid metaphors and analogies, relative to invalid expressions), metaphors are further constrained by pragmatic factors. We therefore predicted that among expressions derived from formally valid analogies, participants would be more selective in *which* metaphors they produce. In contrast, production of analogy phrases was not expected to vary systematically within the set consistent with a valid basis analogy.

Recall that for each basis analogy, there were four unique valid responses (analogies: *A:B::C:D*, *B:A::D:C*, *C:D::A:B*, and *D:C::B:A*; metaphors: *A:C:B*, *B:D:A*, *C:A:D*, and *D:B:C*). As a measure of variability, we computed the entropy among the valid responses that participants produced. For a given basis analogy, we used the frequency of producing each valid response x within the set of four unique valid responses X to estimate its production probability, $p(x)$. We used these frequencies to compute the information entropy among valid responses for each individual basis analogy, separately for metaphors and analogy phrases (see Figure 6).

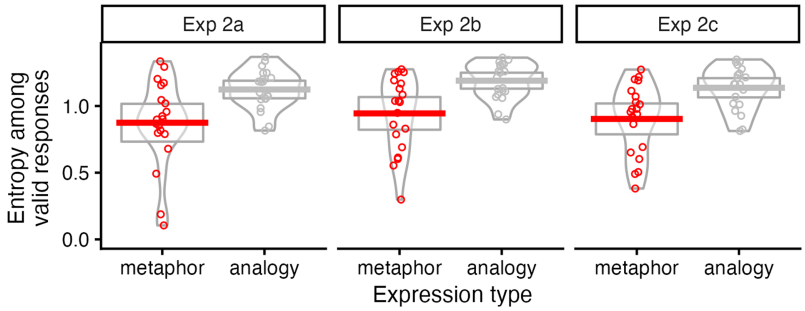
A series of paired Wilcoxon signed-rank tests showed that the entropy among the valid analogy phrases that participants generated (Experiment 2a: $M_{entropy} = 1.12$, $SD_{entropy} = 0.15$; Experiment 2b: $M_{entropy} = 1.19$, $SD_{entropy} = 0.14$; Experiment 2c: $M_{entropy} = 1.14$, $SD_{entropy} = 0.16$) was consistently higher than that for valid metaphors (Experiment 2a: $M_{entropy} = 0.87$, $SD_{entropy} = 0.32$, $V = 37$, $p = .009$; Experiment 2b: $M_{entropy} = 0.94$, $SD_{entropy} = 0.28$, $V = 23$, $p = .001$; Experiment 2c: $M_{entropy} = 0.90$, $SD_{entropy} = 0.26$, $V = 10$, $p < .001$). This pattern suggests that participants favored producing specific metaphors within the set of four valid possibilities, whereas they were relatively indifferent among the four valid possibilities for analogy phrases.

Expression Production and Lexical Feature Norms

We next sought to clarify the particular lexical features that characterize the metaphors and analogy phrases that participants freely generated. As in Experiment 1, these analyses examined the impact of the lexical features *generalness*, *prevalence*, and *concreteness*. We report two analyses of lexical features. The first analysis focused on relative features of *A* terms and *B* terms, which we hypothesize constitute a topic–domain relation in Aristotelian metaphors but not analogy phrases. We expected to see this relation manifest itself through differences in generalness between *A* and *B* terms in metaphors, but potentially also in differences between other lexical features (as was found in Experiment 1). The second analysis focused on the relation between *A* terms and *C* terms, which respectively correspond to the topic and vehicle of a metaphor. We specifically aimed to compare the impact of differences in

Figure 6

Entropy Among Valid Metaphors (Red) and Valid Analogy Phrases (Gray), Broken Down by Study



Note. See the online article for the color version of this figure.

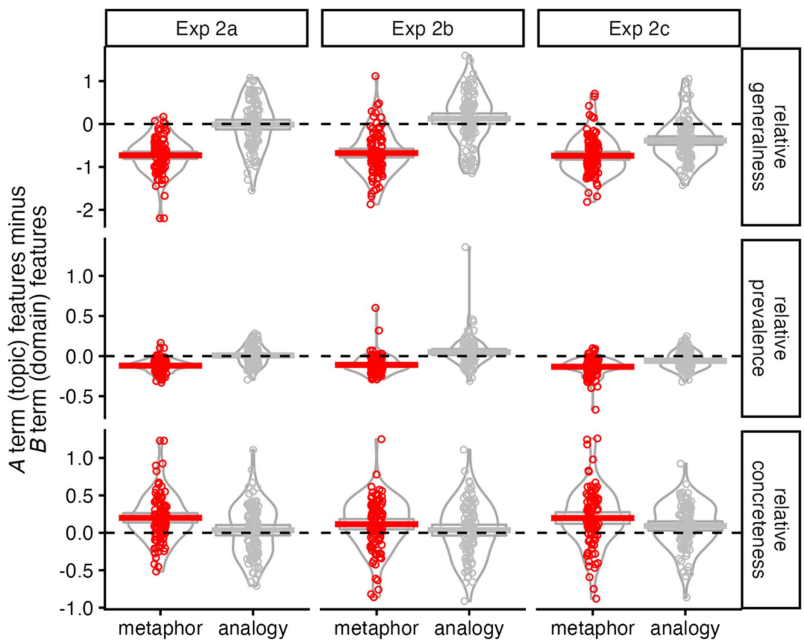
generalness with that of differences in prevalence and in concreteness on metaphor and analogy production.

Relation Between *A* and *B* Terms. In order to assess the potential constraint of a concept–domain relation in expression production, we fit a statistical model to predict whether terms used in participant-generated expressions were part of a metaphor or an analogy phrase, based on their lexical features. As in Experiment 1, the present analyses considered the extent to which *relative* features, computed as differences in lexical features of *A* terms and *B* terms (i.e., features of *A* terms minus features of *B* terms), predicted expression type. As in Experiment 1, our main hypothesis was that relative generalness would distinguish metaphors and analogies such that participant-generated metaphors would consist of more general *B* terms than *A* terms. We restricted this analysis to *A* and *B* terms in expressions derived from valid analogies. Lexical features of these terms are shown in Figure 7.

We fit a logistic mixed-effects model of expression type (*metaphor* = 1, *analogy phrase* = 0) with a set of three *Experiment* (Experiment 2a vs. Experiment 2b vs. Experiment 2c) × *Lexical Feature* two-way interaction terms, one for each lexical feature. This model also adopted a

Figure 7

Relative Lexical Features of A Terms and B Terms in Valid Metaphors (Red) and Analogy Phrases (Gray) Generated by Participants in Experiments 2a (Left Column), 2b (Middle Column), and 2c (Right Column)



Note. Values greater than 0 indicate that *A* terms are higher in some feature than *B* terms, and vice versa for values less than 0. Dashed lines reflect no difference in terms. See the online article for the color version of this figure.

(1|*basis-analogy*) random-intercept term. We omitted a random-effect term for *participant* because it resulted in a singular fit in our model, indicating that the term is redundant with the other terms in the model and predicts little to no unique variance in fit data.

To test whether relative lexical features of *A* and *B* terms in metaphors and analogy phrases varied across experiments, we fit three reduced models that omitted a different *Experiment* × *Lexical Feature* interaction term. We conducted a series of likelihood ratio tests that compared each of these reduced models to the full model. The relation between lexical features and expression type did change across experiments for generalness, $\Delta AIC = 16.4$, $\chi^2(2) = 20.45$, $p < .001$, but not for prevalence, $\Delta AIC = 3.8$, $\chi^2(2) = 0.20$, $p = .90$, or concreteness, $\Delta AIC = 0.0$, $\chi^2(2) = 4.04$, $p = .13$.

Next, we examined simple trends of each lexical feature within each experiment to fully interpret this pattern of results. Table 7 displays these effects, which each reflect how much an increase in 1 standard deviation of a given relative lexical feature for a given word increased the probability that that word is used in a metaphor rather than an analogy phrase. As predicted, relative generalness between *A* terms and *B* terms distinguished between metaphors and analogy phrases across experiments, such that metaphors were more likely to consist of *B* terms that were more general than *A* terms. The two-way interaction mentioned above between experiment and generalness can be attributed to the reduced effect of generalness in Experiment 2c, which reflects the impact of controlling for the number of terms across expression types. Relative prevalence also distinguished between metaphors and analogy phrases across experiments, as did relative concreteness, but only inconsistently across experiments. Overall, these results support the existence of a topic–domain relation between *A* terms and *B* terms in Aristotelian metaphors that distinguish them from their corresponding analogies.

Relation Between *A* and *C* Terms. As in Experiment 1, we ran a parallel analysis to that discussed in the previous section, now comparing lexical features of *A* terms and *C* terms. These terms denote the topic and vehicle in metaphors. Relative lexical features of these terms are shown in Figure 8. We were particularly interested in the extent to which any difference between the lexical features of these terms in metaphors could be traced to either word generalness, prevalence, or concreteness.

We adopted the same modeling approach as previously, now predicting whether or not a given term was used in a metaphor or an analogy phrase, based on the relative features of its *A* and *C* terms (i.e., *A* term features *minus C* term features). We restricted the analysis to valid metaphors and analogies. We fit a logistic mixed-effects model predicting whether a given term was used in a metaphor or an analogy phrase (*metaphor* = 1, *analogy phrase* = 0). This model included three *Experiment* × *Lexical Feature* two-way interaction terms, one for each of *generalness*, *prevalence*, and *concreteness*. As with the full model in the above analysis of *A:B* terms, we included a (1|*basis-analogy*) random-intercept term but not a random-effect term for *participant*, as inclusion of the latter resulted in a singular fit in our model.

We used likelihood ratio tests to compare this full model to each of three reduced models, each omitting a different two-way interaction term. None of these terms contributed significantly to model fit—*generalness*: $\Delta AIC = 15.9$, $\chi^2(2) = 0.55$, $p = .76$; *prevalence*: $\Delta AIC = 2.4$, $\chi^2(2) = 1.59$, $p = .45$; *concreteness*: $\Delta AIC = 3.2$, $\chi^2(2) = 0.74$, $p = .69$.

Finally, we used simple trends to assess the relation between lexical features of *A* and *C* terms for metaphors and analogy phrases within each experiment. Statistics for these trends are reported in Table 8. Word generalness was crucial for distinguishing asymmetries between *A* and *C* terms in metaphors as compared to analogies. Specifically, participant-generated metaphors described topics in terms of more general vehicles. In contrast, differences in prevalence and concreteness between *A* and *C* terms did not discriminate metaphors from analogies. These results converge with the simple trends observed in Experiment 1 for goodness ratings. On balance, word generalness, rather than prevalence or concreteness, appears to better explain topic–vehicle asymmetry both in rated metaphor goodness and in metaphor production.

Table 7
Simple Trends of the Extent to Which Relative Lexical Features of Words Predict Whether an Expression Was a Metaphor or an Analogy Phrase, for Each of the Three Studies in Experiment 2

Lexical feature	Experiment	β	*SE*	*z*	*p*
rel gen.	2a	**−0.54**	**0.07**	**8.21**	**<.001**
	2b	**−0.68**	**0.07**	**9.86**	**<.001**
	2c	**−0.25**	**0.07**	**3.72**	**<.001**
rel prev.	2a	**−0.68**	**0.22**	**3.06**	**.002**
	2b	**−0.74**	**0.23**	**3.18**	**.002**
	2c	**−0.59**	**0.23**	**2.54**	**.01**
rel conc.	2a	−0.06	0.07	0.79	.43
	2b	**−0.23**	**0.08**	**3.01**	**.003**
	2c	−0.03	0.08	0.32	.75

Note. These simple trends reflect how much an increase in the difference between lexical features of *A* terms and *B* terms would increase the probability that both terms are part of a metaphor rather than an analogy phrase. Values in bold reflect statistically significant effects. rel gen. = relative generalness; rel prev. = relative prevalence; rel conc. = relative concreteness; β = trend of a given lexical feature in comparing pairs of terms serving a given role within its expression, estimated from the logistic mixed-effects model (described in Experiment 2, Results and Discussion); *SE* = standard error; *z* = *z* statistics.

Experiment 3

Whereas Experiments 2a, 2b, and 2c examined the *production* of metaphors and analogy phrases, Experiment 3 aimed to assess their *comprehension* using a completion task that allowed comparison of the difficulty posed by processing each kind of expression. In Experiment 3, we tested participants' ability to complete expressions with a single term omitted in order to directly compare the difficulty of processing metaphors and analogy phrases. This task has been used to assess metaphor comprehension in previous work (Stamenković et al., 2019). We hypothesized that processing of metaphors would be more difficult than processing corresponding analogies due to the need to consider additional pragmatic constraints when processing metaphors. Hypotheses, methods, data collection, and analysis plan were preregistered at https://osf.io/wfyvz, and materials, data, and analysis code are available at https://osf.io/vb52z/.

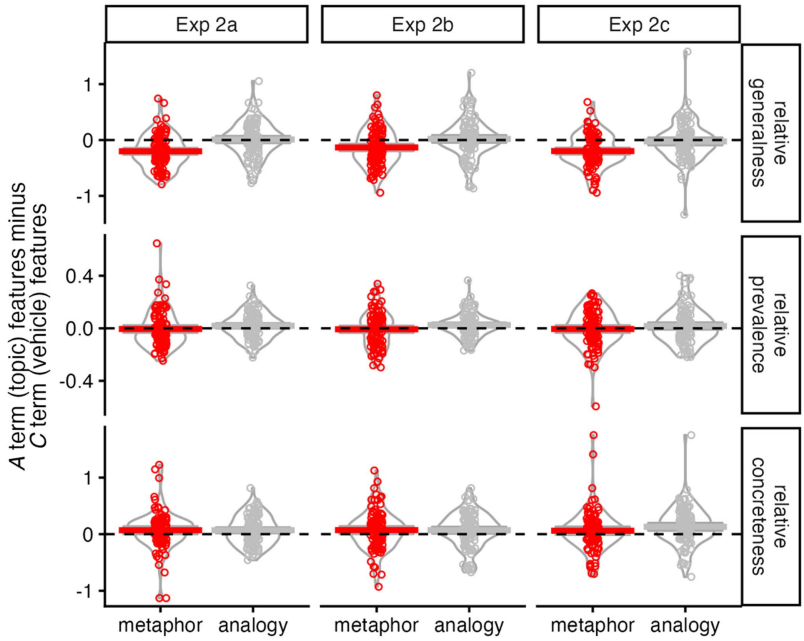
Method

Participants

We recruited a total of 300 participants from Prolific Academic for Experiment 3 ($M_{age} = 37.16$, $SD_{age} = 11.67$; 125 women, 170

Figure 8

Relative Lexical Features of A Terms (Left) and C Terms (Right) in Valid Metaphors (Red) and Analogy Phrases (Gray) Generated by Participants in Experiments 2a (Left Column), 2b (Middle Column), and 2c (Right Column)



Note. See the online article for the color version of this figure.

men, four nonbinary, and one gender not reported). This sample size (which exceeds the preregistered plan for 150) was adopted in order to match the total number of participants across Experiments 2a, 2b, and 2c, thereby enabling a more direct comparison between expression production (assessed previously in Experiment 2) and expression comprehension. Selection criteria included English fluency and an approval rate over 98% on Prolific. We estimated that the rating task would take about 20 min and paid participants $4 (a rate equivalent to $12 per hour). Research procedures were approved by UCLA's Institutional Review Board.

Table 8

Simple Trends of the Extent to Which Relative Lexical Features of Words Predict Whether an Expression Was a Metaphor or an Analogy Phrase

Lexical feature	Experiment	β	*SE*	*z*	*p*
rel gen.	2a	**−0.34**	**0.07**	**4.70**	**<.001**
	2b	**−0.27**	**0.07**	**3.63**	**<.001**
	2c	**−0.34**	**0.08**	**4.25**	**<.001**
rel prev.	2a	−0.03	0.19	0.14	.89
	2b	−0.16	0.20	0.80	.42
	2c	0.18	0.19	0.96	.34
rel conc.	2a	−0.01	0.07	0.20	.84
	2b	−0.07	0.07	0.90	.37
	2c	0.10	0.08	1.37	.17

Note. These simple trends reflect how much an increase in the difference between lexical features of *A* and *C* terms would increase the probability that both terms were part of a metaphor rather than an analogy phrase. Values in bold reflect statistically significant effects. rel gen. = relative generalness; rel prev. = relative prevalence; rel conc. = relative concreteness; β = trend of a given lexical feature in comparing pairs of terms serving a given role within its expression, estimated from the logistic mixed-effects model (described in Experiment 2, Results and Discussion); *SE* = standard error; *z* = *z* statistics.

Materials and Procedure

The stimulus set used in Experiment 3 was derived from the materials used in the previous experiments. Recall that a given basis analogy had four valid arrangements (*A:B::C:D*, *B:A::D:C*, *C:D::A:B*, *D:C::B:A*), each of which could be used to construct a different metaphor (i.e., *A:C:B*, *B:D:A*, *C:A:D*, and *D:B:C*). For each basis analogy, the particular metaphor used in Experiment 3 was the one that received the highest mean goodness rating in Experiment 1 (e.g., "A polygraph is a thermometer of honesty"). The matching analogy phrase was in the form in which its final term matched that of the corresponding metaphor (e.g., "A thermometer is to temperature, just as a polygraph is to honesty").

Completion-Based Expression Comprehension Task. All participants completed a multiple-choice expression completion task, used in past work to assess metaphor comprehension abilities (Stamenković et al., 2019). These multiple-choice problems were generated from the same basis analogies used in Experiments 1 and 2, with each basis analogy being used to generate one metaphor problem and one analogy problem. On each problem of this task, participants were presented with either a metaphor or an analogy phrase with the last term omitted (e.g., "a polygraph is a thermometer of ____" or "a thermometer is to temperature, just as a polygraph is to ____"), and they were instructed to select a term that best completed the expression from among four options: the correct

response (e.g., "honesty"), a *C* term lure that was semantically associated with the *C* term of the analogy phrase and the first term of the metaphor (e.g., "technology," which is associated with "polygraph"), an *A* term lure that was semantically associated with the *A* term of the analogy phrase and the second term of the metaphor (e.g., "weather," which is associated with "temperature"), and a nonlure option that was semantically distant from all terms in either phrase (e.g., "cotton"). All 20 metaphor problems and all 20 analogy problems used in this task can be found in the Supplemental Material.

Note that each pair of metaphor and analogy problems generated from the same basis analogy was phrased so that their final, omitted term, along with their answer options, was identical. An individual participant was only presented with one of these problems, and the particular problems that were presented in their metaphor and analogy form were counterbalanced across participants. Participants completed two blocks of this task, each consisting of 10 problems. One block featured analogy problems only, and the other block metaphor problems only. The order of these blocks was counterbalanced across participants.

Individual Difference Tasks. As in Experiments 2a, 2b, and 2c, participants also completed the RAPM and MH (Raven, 1958). Participants completed the three tasks described above in a fixed order: RAPM, MH, and then the multiple-choice expression completion task.

Results and Discussion

Analysis Software

The analyses reported below used the following R packages in R Version 4.3.1 (R Core Team, 2023). The *lme4* package was used to fit logistic mixed-effects models of response accuracy (Bates et al., 2015), the *buildmer* packages were used to determine the maximal random-effect structure that would converge when fitting each model (Barr et al., 2013; Voeten, 2023), and the *emmeans* package was used to assess estimated marginal means (reported as M_{diff}) and estimated marginal trends (reported as β) of fit models (Lenth, 2023).

In the following analysis, we assess comprehension accuracy along with individual differences in domain-general reasoning ability as measured by the RAPM and verbal knowledge as measured by the MH. As in Experiments 2a–2c, both measures had adequate internal reliability (RAPM ω_t = .76 and MH ω_t = .79). Table 9 shows the pairwise correlations between accurate responding on the multiple-choice task and RAPM and MH performance. We fit a logistic mixed-effects model of trial-level accuracy with two two-way interaction terms respectively for *RAPM-Score* × *Expression-Type* (metaphor vs. analogy phrase) and *MH-Score* × *Expression-Type*, as well as two random slope terms (1 + *expression-type*|*participant*) and (1 + *expression-type*|*basis-analogy*). To test whether domain-general reasoning ability and verbal knowledge make differential contributions to metaphor and analogy comprehension, we used likelihood ratio tests that compared this full model to two reduced models that respectively omitted the *RAPM-Score* × *Expression-Type* and *MH-Score* × *Expression-Type* interaction terms but that were otherwise identical to the full model. Neither removing the RAPM interaction term nor the MH interaction term significantly increased model prediction error—RAPM: ΔAIC = 1.9, $\chi^2(1) = 0.03$, $p = .87$; MH: ΔAIC = 0.2, $\chi^2(1) = 2.18$, $p = .14$—and each reliably predicted comprehension rates across expression types (RAPM: β = 1.47, SE = 0.25, $z = 5.82$, $p < .001$; MH: β = 2.73, $SE = 0.33$, $z = 8.30$, $p < .001$). These results highlight the importance of both domain-general reasoning ability and verbal knowledge in both metaphor and analogy comprehension.

Table 9
Descriptive Statistics for Response Rates on Multiple-Choice Completion Task and Individual Difference Measures Used in Experiment 3

Response type	*M* (*SD*)	Pairwise correlation (Spearman's ρ) 1	2	3	4
1. Metaphor					
Correct	0.68 (0.20)	—	.47	.35	.35
C lure	0.22 (0.14)				
A lure	0.08 (0.10)				
Nonlure	0.02 (0.06)				
2. Analogy					
Correct	0.76 (0.20)	.47	—	.38	.31
C lure	0.19 (0.15)				
A lure	0.04 (0.08)				
Nonlure	0.01 (0.05)				
3. Mill Hill	0.66 (0.17)	.35	.38	—	.23
4. Raven's progressive matrices	0.40 (0.22)	.35	.31	.23	—

Note. Correlations between completion task performance and individual difference measures are determined by correct response rates (all *p*s < .001).

Multiple-Choice Response Rates and Accuracies

As shown in Table 9, participants were overall very successful in selecting the correct option on these completion problems (M_{corr} = 0.72, SD_{corr} = 0.17). Among the incorrect answer options, the *C* term lure was most distracting ($M_{C\ term}$ = 0.20, $SD_{C\ term}$ = 0.12), followed by the *A* term lure ($M_{A\ term}$ = 0.06, $SD_{A\ term}$ = 0.07) and then nonlures ($M_{nonlure}$ = 0.02, $SD_{nonlure}$ = 0.05). Table 9 provides the descriptive statistics for multiple-choice response rates, broken down by expression type.

This model also revealed a main effect of expression type (M_{diff} = 0.58, SE = 0.14, z = 4.13, $p < .001$), indicating that completing metaphors (M_{corr} = 0.68, SD_{corr} = 0.20) was less accurate than completing analogy phrases (M_{corr} = 0.76, SD_{corr} = 0.20), as predicted. This difference was corroborated with a paired Wilcoxon signed-rank test (V = 7,671.5, $p < .001$). Thus, although the best metaphors and their matched analogies did not differ in rated goodness (Experiment 1) or production frequency (Experiment 2), it proved more difficult to complete the *C* term for the best metaphor than for its corresponding analogy in Experiment 3. This difference may be due to the structure of the multiple-choice questions used here, in which the analogy form explicitly stated the *D* term whereas the metaphor form did not.

General Discussion

Commonalities Between Metaphors and Analogies

Using both production and comprehension paradigms, we investigated the relationship between the processing of metaphors versus formally equivalent analogies. In order to precisely match

the metaphor and analogy tasks, we generated materials based on Aristotelian metaphors. These can be derived by translating proportional verbal analogies in the form *A:B::C:D* into metaphors in the form "*A* is the *C* of *B*" (omitting the *D* term). Each basis analogy yields four "valid" variants, defined by the factorial combination of which pair appears first (creating the variant *C:D::A:B*) and the order of the items within each pair (creating *B:A::D:C* and *D:C::B:A*). All four variants form valid analogies in which the semantic relation between the first pair is the same or highly similar to the relation between the second pair. Each analogy variant can in turn be used to generate a matching "valid" metaphor. Using a uniform set of stimuli, we had participants rate the goodness of analogies and metaphors (Experiments 1 and 2a), produce analogies and metaphors by placing randomized words into appropriate slots (Experiments 2a, 2b, 2c), and select the best completion for analogies and metaphors (Experiment 3).

Results of the rating tasks revealed that people assign higher goodness scores to both analogies and metaphors that are valid. The preference for valid forms was found both when all forms were assessed by independent raters (Experiment 1) and when participants rated the specific form they had personally produced (Experiment 2a). The latter finding reveals that people have enough metacognitive awareness to downgrade their own productions when these fail the validity test. The validity preference also manifested itself as an advantage in expression production: Participants produced valid expressions much more often than they did invalid ones (Experiments 2a, 2b, 2c). The present study thus confirms the central claim made by Aristotle (c. 335 BCE): A necessary condition for a successful metaphor in the form "*A* is the *C* of *B*" is that it must satisfy the basic requirement for a valid analogy in the form *A:B::C:D*, namely, high semantic similarity between the relation linking *A:B* and that linking *C:D*.

In addition to this formal connection, metaphors and verbal analogies showed similar sensitivity to individual differences in both executive functions (assessed using the RAPM test) and verbal ability (assessed using the MH). Across Experiments 2a, 2b, 2c, and 3, participants who obtained higher scores on each of these tests achieved greater accuracy on all tasks requiring processing of either metaphors or analogies. Performance on the two tests was correlated, but each contributed unique variance in predicting accuracy on both metaphor and analogy tasks. The findings are broadly consistent with previous research (Beaty & Silvia, 2013; Chiappe & Chiappe, 2007; Gray & Holyoak, 2020; Ichien et al., 2022). One hint of a processing difference between analogies and metaphors emerged: When data from the three variants of production tasks (Experiments 2a, 2b, 2c) were aggregated, the measure of executive function was a stronger predictor of accuracy for analogy than for metaphor tasks. This finding suggests that in production tasks, the analogy format may elicit a greater contribution from the reasoning system than does the metaphor format. Nonetheless, successful processing in either format depends on both reasoning and verbal ability.

Pragmatic Constraints Unique to Metaphor

But although a valid analogical basis is *necessary* to create a successful metaphor, it is not a sufficient condition. An analogy in the proportional format *A:B::C:D* can be *nonfunctional*, with no purpose other than "solving an analogy problem" (Holyoak, 2025). In contrast, a good metaphor needs to "make a point" by offering some insight into its topic. For example, the topic of "amnesia is an eraser of memory" is "amnesia"—the injury to memory is likened to the physical action of an eraser removing vulnerable writing done in pencil (thereby destroying the meaning of the written message, just as amnesia destroys knowledge of one's own life). This metaphor, which intuitively seems apt, is formally equivalent to the analogy *amnesia* : *memory* :: *eraser* : *pencil.*

However, none of the other three valid variants of this analogy support a good metaphor. "An eraser is the amnesia of a pencil" is at best odd; "Memory is the pencil of amnesia" and "A pencil is the memory of an eraser" are virtually meaningless. Analogical validity alone does not suffice to make a good metaphor. Rather, good metaphors honor additional pragmatic constraints. In the present study, we identified consistent quantitative differences in patterns of performance between analogies and metaphors that illuminate these pragmatic constraints that impact metaphors. In Experiments 2a, 2b, and 2c, we consistently found that people produce the four valid variants of a proportional analogy about equally often (resulting in higher entropy for production frequencies), whereas they favor producing specific variants as metaphors (resulting in lower entropy). It is worth noting that in Experiment 2c, the number of terms in the phrase to be produced was equated (three terms provided) for both metaphors and analogies; this procedure yielded the same basic pattern as found in Experiments 2a and 2b, in which metaphor problems had three terms and analogy problems had four.

Overall, our findings indicate that metaphor production is impacted by additional factors beyond formal validity of the corresponding analogy. Moreover, the "best" metaphor form, and the form most likely to be produced, proved to be predictable from patterns of asymmetries defined over lexical variables related to the generalness and familiarity of words.

Relation Between Topics and Domains

Good Aristotelian metaphors exhibit strong asymmetries in word *generalness*: In particular, *B* denotes a more general concept than does *A*. These asymmetries are consequences of the fact that in an apt Aristotelian metaphor, the topic *A* (e.g., "amnesia") typically plays a role defined with respect to a broader domain described by *B* (e.g., "memory"). Thus, even though the semantic relation linking *A* to *B* implies a meaningful converse relation linking *B* to *A* (Lu et al., 2019), the relation is an asymmetrical one. Although reversing the *A* and *B* terms (and *C* and *D* terms) maintains a valid analogy form, such reversals do not yield a corresponding apt metaphor.

An asymmetry in generalness (more general *B* relative to *A*) predicted higher goodness of metaphors (Experiment 1), though analogies showed a similar trend. In all the production experiments (2a, 2b, 2c), the magnitude of the asymmetry was reliably greater for metaphors than analogies. Differences in prevalence of *A* and *B* terms also predicted both goodness ratings and production frequencies. In general, people preferred metaphors with topics (*A* terms) that were less familiar than domain terms (*B*).

The observed relation between domains and particular topics may be related to a similar asymmetry proposed to explain how people interpret noun–noun conceptual combinations (e.g., "whale boat"). In general, interpreting such phrases involves identification of a head noun, which denotes the overall category of the compound

(e.g., "boat"), and a modifier, which provides further elaboration of the head (e.g., "whale"). The meaning of the compound can then be assigned either by finding a relation that integrates the two (e.g., *used-for*) or (less often) by attributing properties of the modifier to the head (e.g., *large*; e.g., Gagné, 2000). It has been suggested that heads and modifiers activate role concepts in semantic memory (e.g., *objective* and *tool*), which agents use to constrain the relation that they ultimately adopt to interpret the compound (e.g., *used-for*, such that a "whale boat" is a boat used to observe or hunt whales; Maguire et al., 2010; Mather et al., 2014). A similar cognitive process has been proposed as a possible mechanism by which nominal metaphors (e.g., "My lawyer is a shark") can be interpreted (e.g., "shark lawyer"; Estes & Glucksberg, 2000).

The proposed process of role activation emphasizes the lexical access of head and modifier role categories in a way that is consistent with our emphasis on the generalness of domain terms in verbal analogies (expressed as *B* and *D* terms), which will likewise tend to be activated by their corresponding topic concepts. For example, basic-level categories (e.g., *fruit*) readily prime their superordinates (e.g., *food*) and do so more strongly than they do their exemplars (e.g., *apple*; Kareev, 1982). But despite this parallel, we caution against interpreting the comprehension of Aristotelian metaphors directly in terms of this role activation hypothesis used to explain conceptual combination. Beyond obvious formal differences between our metaphors (e.g., "Amnesia is an eraser of memory") and noun–noun compounds, recall that the verbal analogies on which metaphors are based are semantically remote (e.g., *amnesia* : *memory* :: *eraser* : *pencil*). The relation between domains (e.g., between *memory* and *pencil*) that would correspond to that between head and modifier role categories is therefore less obviously discernible. It seems unlikely that reasoners would apply some explicit relation holding between these domain terms to interpret Aristotelian metaphors. Nonetheless, something akin to role activation may serve an indirect function, in which activation of the unstated domain (e.g., *pencil*), along with the stated domain (e.g., *memory*) and vehicle (e.g., *eraser*), jointly constrains the properties that are attributed to the topic, as proposed in alternative models of conceptual combination (Estes & Glucksberg, 2000; Gagné, 2000; Wisniewski, 1997). Future work is needed to clarify the relation between metaphor and conceptual combination.

Relation Between Topics and Vehicles

Pragmatic factors related to lexical features also constrained the distinction between topics (*A* term) and vehicles (*C* term) in metaphors. Metaphors were rated higher in goodness in Experiment 1 and produced more frequently in Experiments 2a, 2b, and 2c, when the topic, denoted by the *A* term (e.g., "amnesia"), was a more specific term than the vehicle, denoted by the *C* term (e.g., "eraser"). In other words, a good metaphor aims to explain a specific concept in terms of one that is more general, not the reverse. Our findings were neutral with respect to debates over whether word concreteness or familiarity distinguishes between metaphor topics and vehicles (Dancygier & Sweetser, 2014; Katz, 1989; Lakoff & Johnson, 1980). We instead provide evidence that Aristotelian metaphors are interpreted as modified categorization statements, in which terms better suited to serve as an ad hoc category occupy the vehicle role (Glucksberg & Keysar, 1990). We found evidence that word generalness (but not relative concreteness or familiarity) for *A* and *C* terms distinguished goodness ratings of metaphors from those of analogies: Metaphors with topics less general than their vehicles were judged to be better. In Experiments 2a–2c, we also found that relative generalness between *A* and *C* terms, but not concreteness or prevalence (our measure of familiarity), distinguished between participant-generated metaphors versus analogies. Across studies, expressions with *C* terms higher in generalness than *A* terms were more likely to be metaphors than analogies. Overall, our results converge with other work showing that semantic features of constituent terms can drive whether novel metaphors are interpreted as asymmetric category statements rather than symmetric comparison statements (Glucksberg & Haught, 2006).

Relationship Between Metaphor and Analogy

The present study is, to the best of our knowledge, the first to directly compare processing of metaphors and formally equivalent analogies. A good metaphor serves to illuminate its topic through an apt comparison that conveys some mix of new understanding and a suggested emotional attitude. Typically, the insight offered by a metaphor depends in part on similar relations between concepts drawn from different domains—a requirement that implies existence of a valid analogy. Our results confirm that for the type of metaphor first discussed by Aristotle (c. 335 BCE)—the form "*A* is the *C* of *B*"—a valid analogical relationship is indeed a necessary condition for an apt metaphor.

Nonetheless, metaphor depends on more than analogy. The basic pragmatic function of metaphor—offering a new insight or point of view that illuminates the topic—depends on special properties of the concepts involved, notably differences in generalness. Apt Aristotelian metaphors are those that correspond to valid analogies *and also* satisfy specific pragmatic constraints. People prefer relationally valid metaphors that describe relatively specific terms (*A* term) using a familiar and general domain (*B* term) and general vehicle (*C* term). Although we did not test similes in the present study, intuition suggests that similes are sensitive to much the same pragmatic factors as metaphors. For example, "Old age is like the evening of life" seems sensible, whereas "Life is like the day of old age" does not. Although similes clearly imply a comparison, they also appear to adhere to the additional constraints that govern metaphor.

The present study found that production and comprehension of both metaphors and matched analogies are impacted by individual differences in both executive functions and verbal ability. Although typically correlated, these cognitive abilities are separable. In production tasks (Experiment 2), the measure of executive functioning (RAPM) had a significantly greater impact for analogy than for metaphor. Other evidence indicates that executive functions and verbal ability are dissociable in regard to processing metaphors and analogies. Individuals with autism spectrum disorder, when compared to matched samples of typically developing people, generally perform less well on tests of metaphor comprehension. In contrast, studies of analogical reasoning (using nonverbal tests) have found that autism spectrum disorder groups perform as well as (and sometimes better than) typically developing groups matched in age and overall cognitive ability (Morsanyi et al., 2019). Studies of normal cognitive aging reveal an opposite dissociation between analogy and metaphor. Healthy older adults perform less well than younger adults on tests of analogical reasoning (Viskontas et al.,

2004), but as well or better on tests of metaphor comprehension (relying more heavily on verbal knowledge; Ichien et al., 2025). These lines of research support the conclusion that in different populations, analogy and metaphor are linked to distinct cognitive processes.

Constraints on Generality of Conclusions

Note that all experiments recruited human participants through Prolific Academic. We restricted our sample to English speakers with a 98% approval rate. Accordingly, the data on which we base our conclusions, though demographically representative of the population of Prolific workers across the English-speaking world, are restricted to highly motivated online users. Moreover, given that our study was exclusively run in English, conclusions about metaphor and analogy processing are accordingly limited to processing of these expressions in the English language.

The conclusions of the present study are also necessarily limited by the type of metaphor on which we focused. We chose to investigate Aristotelian metaphors because these allowed us to precisely match the formal structures of metaphors and analogies. There is reason to suspect that when the broader pool of metaphors is considered, additional pragmatic and linguistic constraints on apt metaphors come into play. In particular, some previous work has focused on differences between literary metaphors (especially those that originated in poetry, e.g., "A waterfall is a wild, unbridled horse;" Holyoak, 2019; Lakoff & Turner, 1989) and nonliterary metaphors (Katz et al., 1988). The stimuli used in the present study were nonliterary (derived from proportional verbal analogies used in previous psychological research; Green et al., 2010, 2012). Although the differences are often subtle, machine-learning algorithms are able to distinguish literary from nonliterary metaphors with high accuracy. Jacobs and Kinder (2018) found that qualities distinguishing literary metaphors rated high in goodness include high surprisal (a statistical measure of the unexpectedness of words), relative dissimilarity of source and target concepts, the combination of concrete words with relatively complex grammar and high lexical diversity, and extra difficulty in comprehending the metaphorical meaning. These properties collectively suggest that good literary metaphors are high in cognitive complexity, which may be more likely to elicit analogical reasoning (and hence place greater demands on fluid intelligence). This prediction has been supported by studies of individual differences in metaphor comprehension (Stamenković et al., 2019).

At the same time, other types of metaphors may be *less* closely linked to analogical reasoning. As noted earlier, some metaphors may evoke a looser process of semantic integration in which features of word meanings are blended (Kintsch, 2000). In particular, comprehension of predicate metaphors in which the metaphor takes the form of a verb transferred from a different domain (e.g., "The flowers *purred* in the sunlight") depends primarily on individual differences in verbal ability, with little contribution from differences in executive functioning (Stamenković et al., 2019). Future research on the relationship between processing of metaphors and analogies may usefully employ functional neuroimaging approaches (e.g., Blank et al., 2014) that in other tasks have dissociated brain networks underlying language and reasoning.

References

Al-Azary, H., & Buchanan, L. (2017). Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory & Cognition*, *45*(2), 296–307. https://doi.org/10.3758/s13421-016-0650-7

Al-Azary, H., & Katz, A. N. (2021). Do metaphorical sharks bite? Simulation and abstraction in metaphor processing. *Memory & Cognition*, *49*(3), 557–570. https://doi.org/10.3758/s13421-020-01109-2

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*(4), 354–361. https://doi.org/10.1177/073428299901700405

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Beaty, R. E., & Silvia, P. J. (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & Cognition*, *41*(2), 255–267. https://doi.org/10.3758/s13421-012-0258-5

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 295–308. https://doi.org/10.1037//0278-7393.19.2.295

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Chiappe, D., Kennedy, J. M., & Smykowski, T. (2003). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, *18*(2), 85–105. https://doi.org/10.1207/S15327868MS1802_2

Chiappe, D. L., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, *56*(2), 172–188. https://doi.org/10.1016/j.jml.2006.11.006

Christensen, R. H. B. (2024). *ordinal: Regression models for ordinal data* (Version 2023.12-4.1) [Computer software]. https://cran.r-project.org/web/packages/ordinal/index.html

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452–465. https://doi.org/10.1016/j.cognition.2012.07.010

Connor, K., & Kogan, N. (1980). Topic-vehicle relations in metaphor: The issue of asymmetry. In R. P. Honeck & R. R. Hoffman (Eds.), *Cognition*

and figurative language. Routledge. https://doi.org/10.4324/9780429432866-12

Dancygier, B., & Sweetser, E. (2014). *Figurative language*. Cambridge University Press.

Estes, Z., & Glucksberg, S. (2000). Interactive property attribution in concept combination. *Memory & Cognition*, *28*, 28–34. https://doi.org/10.3758/BF03211572

Gagné, C. L. (2000). Relation-based combinations versus property-based combinations: A test of the CARIN theory and the dual-process theory of conceptual combination. *Journal of Memory and Language*, *42*(3), 365–389. https://doi.org/10.1006/jmla.1999.2683

Gentner, D. (1977). Children's performance on a spatial analogies task. *Child Development*, *48*(3), 1034–1039. https://doi.org/10.2307/1128356

Gentner, D., Bowdle, B. F., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 199–253). MIT Press. https://doi.org/10.7551/mitpress/1251.001.0001

George, T., Christofalos, A. L., & Pambuccian, F. S. (2025). Generating distant analogies increases metaphor production. *Psychonomic Bulletin & Review*, *32*, 1402–1410. https://doi.org/10.3758/s13423-024-02628-8

Gibb, H., & Wales, R. (1990). Metaphor or simile: Psychological determinants of the differential use of each sentence form. *Metaphor and Symbolic Activity*, *5*(4), 199–213. https://doi.org/10.1207/s15327868ms0504_1

Glucksberg, S., & Haught, C. (2006). Can Florida become like the next Florida? When metaphoric comparisons fail. *Psychological Science*, *17*(11), 935–938. https://doi.org/10.1111/j.1467-9280.2006.01807.x

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*(1), 3–18. https://doi.org/10.1037/0033-295X.97.1.3

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language*, *36*(1), 50–67. https://doi.org/10.1006/jmla.1996.2479

Gray, M. E., & Holyoak, K. J. (2020). Individual differences in relational reasoning. *Memory & Cognition*, *48*(1), 96–110. https://doi.org/10.3758/s13421-019-00964-y

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264–272. https://doi.org/10.1037/a0025764

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70–76. https://doi.org/10.1093/cercor/bhp081

Harris, R. J., Friel, B. M., & Mickelson, N. R. (2006). Attribution of discourse goals for using concrete- and abstract-tenor metaphors and similes with or without discourse context. *Journal of Pragmatics*, *38*(6), 863–879. https://doi.org/10.1016/j.pragma.2005.06.010

Hesse, M. B. (1966). *Models and analogies in science*. University of Notre Dame Press.

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, *50*(1), 115–133. https://doi.org/10.3758/s13428-017-1009-0

Holyoak, K. J. (2019). *The spider's thread: Metaphor in mind, brain, and poetry*. MIT Press. https://doi.org/10.7551/mitpress/11119.001.0001

Holyoak, K. J. (2025). *The human edge: Analogy and the roots of creative intelligence*. MIT Press. https://doi.org/10.7551/mitpress/15232.001.0001

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Ichien, N., Lu, H., & Holyoak, K. J. (2022). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 108–121. https://doi.org/10.1037/xlm0001010

Ichien, N., Stamenković, D., & Holyoak, K. J. (2024). Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, *39*(4), 296–309. https://doi.org/10.1080/10926488.2024.2380348

Ichien, N., Stamenković, D., Whatley, M. C., Castel, A. D., & Holyoak, K. J. (2025). Advancing with age: Older adults excel in comprehension of novel metaphors. *Psychology and Aging*, *40*(1), 6–16. https://doi.org/10.1037/pag0000836

Inagaki, K., & Hatano, G. (1987). Young children's spontaneous personification as analogy. *Child Development*, *58*(4), 1013–1020. https://doi.org/10.2307/1130542

Jacobs, A. M., & Kinder, A. (2018). What makes a metaphor literary? Answers from two computational studies. *Metaphor and Symbol*, *33*(2), 85–100. https://doi.org/10.1080/10926488.2018.1434943

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32. https://doi.org/10.1016/j.jml.2006.02.004

Kareev, Y. (1982). A priming study of developmental changes in the associative strength of class relations. *Child Development*, *53*(4), 1038–1045. https://doi.org/10.2307/1129145

Katz, A. N. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language*, *28*(4), 486–499. https://doi.org/10.1016/0749-596X(89)90023-5

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbolic Activity*, *3*(4), 191–214. https://doi.org/10.1207/s15327868ms0304_1

Keil, F. C. (1986). Conceptual domains and the acquisition of metaphor. *Cognitive Development*, *1*(1), 73–96. https://doi.org/10.1016/S0885-2014(86)80024-7

Kelly, M. H., & Keil, F. C. (1987). Metaphor comprehension and knowledge of semantic domains. *Metaphor and Symbolic Activity*, *2*(1), 33–51. https://doi.org/10.1207/s15327868ms0201_3

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*(2), 257–266. https://doi.org/10.3758/BF03212981

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

Kövecses, Z. (2002). *Metaphor: A practical introduction*. Oxford University Press. https://doi.org/10.1093/oso/9780195145113.001.0001

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226470993.001.0001

Lakoff, G., & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226470986.001.0001

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.1.35.2) [Computer software]. https://CRAN.R-project.org/package=emmeans

Lipovetsky, S., & Conklin, M. (2014). Best-worst scaling in analytical closed-form solution. *Journal of Choice Modelling*, *10*, 60–68. https://doi.org/10.1016/j.jocm.2014.02.001

Littlemore, J., Sobrino, P. P., Houghton, D., Shi, J., & Winter, B. (2018). What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation. *Metaphor and Symbol*, *33*(2), 101–122. https://doi.org/10.1080/10926488.2018.1434944

Löhr, G. (2022). What are abstract concepts? On lexical ambiguity and concreteness ratings. *Review of Philosophy and Psychology*, *13*(3), 549–566. https://doi.org/10.1007/s13164-021-00542-9

Löhr, G. (2024). Does the mind care about whether a word is abstract or concrete? Why concreteness is probably not a natural kind. *Mind & Language*, *39*(5), 627–646. https://doi.org/10.1111/mila.12473

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Maguire, P., Maguire, R., & Cater, A. W. S. (2010). The influence of interactional semantic patterns on the interpretation of noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 288–297. https://doi.org/10.1037/a0018687

Mather, E., Jones, L. L., & Estes, Z. (2014). Priming by relational integration in perceptual identification and Stroop colour naming. *Journal of Memory and Language*, *71*(1), 57–70. https://doi.org/10.1016/j.jml.2013.10.004

Morsanyi, K., Stamenković, D., & Holyoak, K. J. (2019). Analogical reasoning in autism: A systematic review and meta-analysis. In K. Morsanyi & R. M. J. Byrne (Eds.), *Thinking, reasoning, and decision making in autism* (pp. 59–87). Routledge. https://doi.org/10.4324/9781351060912-4

Noble, C. E. (1954). The familiarity–frequency relationship. *Journal of Experimental Psychology*, *47*(1), 13–16. https://doi.org/10.1037/h0060025

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*(3), 161–180. https://doi.org/10.1037/0033-295X.86.3.161

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raven, J. C. (1958). *Guide to using the Mill Hill Vocabulary Scale with the Progressive Matrices Scales* (p. 64). H. K. Lewis & Co.

Reilly, M., & Desai, R. H. (2017). Effects of semantic neighborhood density in abstract and concrete words. *Cognition*, *169*, 46–53. https://doi.org/10.1016/j.cognition.2017.08.004

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research* (Version 2.3.9). https://cran.r-project.org/web/packages/psych/index.html

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Lawrence Erlbaum Associates.

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108–118. https://doi.org/10.1016/j.jml.2018.12.003

Stamenković, D., Ichien, N., & Holyoak, K. J. (2020). Individual differences in comprehension of contextualized metaphors. *Metaphor and Symbol*, *35*(4), 285–301. https://doi.org/10.1080/10926488.2020.1821203

Stamenković, D., Milenković, K., Ichien, N., & Holyoak, K. J. (2023). An individual-differences approach to poetic metaphor: Impact of aptness and familiarity. *Metaphor and Symbol*, *38*(2), 149–161. https://doi.org/10.1080/10926488.2021.2006046

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive Psychology*, *13*(1), 27–55. https://doi.org/10.1016/0010-0285(81)90003-7

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition*, *11*(3), 203–244. https://doi.org/10.1016/0010-0277(82)90016-6

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, *19*(4), 581–591. https://doi.org/10.1037/0882-7974.19.4.581

Voeten, C. C. (2023). *buildmer: Stepwise elimination and term reordering for mixed-effects regression* (Version 2.11) [Computer software]. https://CRAN.R-project.org/package=buildmer

Weinberger, A. B., Gallagher, N. M., Colaizzi, G., Liu, N., Parrott, N., Fearon, E., Shaikh, N., & Green, A. E. (2022). Analogical mapping across sensory modalities and evidence for a general analogy factor. *Cognition*, *223*, Article 105029. https://doi.org/10.1016/j.cognition.2022.105029

Winter, B., & Srinivasan, M. (2022). Why is semantic change asymmetric? The role of concreteness and word frequency and metaphor and metonymy. *Metaphor and Symbol*, *37*(1), 39–54. https://doi.org/10.1080/10926488.2021.1945419

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, *4*, 167–183. https://doi.org/10.3758/BF03209392

Xu, X. (2010). Interpreting metaphorical statements. *Journal of Pragmatics*, *42*(6), 1622–1636. https://doi.org/10.1016/j.pragma.2009.11.005

Received July 4, 2025
Revision received January 30, 2026
Accepted February 5, 2026 ■