Behavior Research Methods
https://doi.org/10.3758/s13428-019-01312-3

Verbal analogy problem sets: An inventory of testing materials

Nicholas Ichien[1] · Hongjing Lu[1,2] · Keith J. Holyoak[1]

Abstract
Analogical reasoning is an active topic of investigation across education, artificial intelligence (AI), cognitive psychology, and related fields. In all fields of inquiry, explicit analogy problems provide useful tools for investigating the mechanisms underlying analogical reasoning. Such sets have been developed by researchers working in the fields of educational testing, AI, and cognitive psychology. However, these analogy tests have not been systematically made accessible across all the relevant fields. The present paper aims to remedy this situation by presenting a working inventory of verbal analogy problem sets, intended to capture and organize sets from diverse sources.

Keywords Analogy, Relational Reasoning, Language, Education, Artificial Intelligence

Introduction

Analogical reasoning is the ability to grasp and exploit similarities based on relations between sets of entities, rather than solely on features of the individual entities themselves (Holyoak, 2012). The general ability to think in terms of explicit relations is highly developed in humans, and indeed may constitute a discontinuity between human intelligence and non-human intelligence (Penn, Holyoak, & Povinelli, 2008). This cognitive process supports human performance in a wide range of activities (Holyoak & Thagard, 1995), including classroom learning (Richland, Zur, & Holyoak, 2007; Jee et al., 2010), engineering design (Chan & Schunn, 2015), and scientific reasoning (Dunbar, 1995; Gentner & Jeziorski, 1993). Performance on analogy tests is highly correlated with measures of fluid intelligence (Snow, Kyllonen, & Marshalek, 1984). Analogy problems also provide an important source of benchmarks for assessing the performance of models of relational processing developed by researchers in artificial intelligence (AI) (e.g., Mikolov, Chen, Corrado, & Dean, 2013; Turney, 2013).

Sets of explicit analogy problems provide useful tools for investigating the mechanisms underlying analogical reasoning. Such sets have been developed by researchers working in the fields of educational testing, AI, and cognitive psychology. However, these analogy tests have not been systematically made accessible across all the relevant fields. The present paper aims to remedy this situation by presenting a working inventory of verbal analogy problem sets, intended to capture and organize sets from diverse sources.

Analogy tests can be constructed in pictorial formats (e.g., Krawczyk et al., 2008); but here we focus on verbal problem sets, which themselves are extremely varied in form and content. First, analogies are frequently used in education as a method of training and testing verbal aptitude in children at various grade levels. Standardized tests (e.g., School and College Ability Test, or SCAT; Secondary School Admission Test, or SSAT; and the Scholastic Aptitude Test, or SAT), as well as other educational materials (e.g., englishforeveryone.org) provide useful sources of analogy problem sets tailored to specific stages of cognitive development (e.g., Turney, Littman, Bigham, & Shnayder, 2003). Second, recent advances in AI have been facilitated by a number of analogy problem sets used as benchmarks for computational models (e.g., Mikolov et al., 2013a; Mikolov, Sutskever, Chen,

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-019-01312-3) contains supplementary material, which is available to authorized users.

✉ Nicholas Ichien
ichien@ucla.edu

Hongjing Lu
hongjing@ucla.edu

Keith J. Holyoak
holyoak@lifesci.ucla.edu

[1] Department of Psychology, University of California, Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1563, USA

[2] Department of Statistics, University of California, Los Angeles, 405 Hilgard Ave., Los Angeles 90095-1563, CA, USA

Published online: 02 January 2020

Corrado, & Dean 2013; Mikolov, Yih, & Zweig 2013; Gladkova, Drozd, & Matsuoka, 2016; Turney, 2008). Third, research in cognitive psychology has produced an extensive literature on analogical reasoning in humans. Researchers in this area have produced several analogy problem sets for which data on human performance is available (e.g., Morrison et al., 2004; Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010, 2012; Kmiecik, Brisson, & Morrison, 2019). The general aim of the present inventory is to provide a centralized source of analogy problem sets drawn from multiple domains so as to facilitate further research on analogical reasoning in both humans and machines.

Analogies can be expressed in a variety of forms, and a given form can be varied to produce different types of analogy problems (see Table 1). Consider the familiar four-term *A:B::C:D* verbal analogy form, in which the meanings of two words, *A* and *B*, are related to each other in the same or a similar way as two other words, *C* and *D* (e.g., *dog:puppy :: kitten:cat*). Analogy problems in this form involve a minimal mapping process: Solving an *A:B::C:D* analogy involves grasping a single source relation (*A:B* term) and reasoning about a single target relation (*C:D* term).

Within this form, several types of analogy problems can be created. In one type, which we will call *evaluative* problems, an intact four-term sequence of words is presented (as in the above example), and the reasoner is asked to judge whether or not the sequence forms a valid analogy (e.g., Kmiecik et al., 2019).

Table 1. Form-based taxonomy of analogy problems included in the present inventory

Analogical form	Question/Task type	Example question/Task description	Correct response
A:B::C:D	*Evaluative* (e.g., Green et al., 2010)	*answer:riddle :: jersey:number*	invalid
	One-term Generative (e.g., Green et al., 2012)	*ash:fireplace :: lint:?*	*pocket*
	Two-term Generative (e.g., Popov et al., 2017)	*book:writer :: ?:?*	*blueprint:architect*
	Multiple-Choice (e.g., Sternberg & Nigro, 1980)	*yes:no :: true:*? A. *false* B. *right*	A. *false*
	Matrix (e.g., Weinberger et al., 2016)	*helmet:head* *mitt:baseball* *amnesia:memory* *boat:anchor* *inventor:invention* *eraser:pencil* *spray:skunk* *mother:daughter* *antenna:signal* *kneepad:knee* *porkchop:pig* etc.	*mitt:baseball* *kneepad:knee*
Story / long text	*Retrieval* (e.g., Wharton et al., 1994)	Having read some target story and some cue stories	N/A
	Generative (e.g., Gentner & Toupin, 1986)	Having read some story, recreate that story using new characters	N/A
	Problem Solving (e.g., Gick & Holyoak, 1980)	Use analogous story to solve radiation problem (Duncker, 1945). Radiation problem: How can you use radiation rays to destroy a patient's tumor while avoiding destroying their healthy tissue?	Send a series of small radiation rays into the body that all converge at the tumor.
	Extended Mapping (e.g., Turney, 2008)	*bacterial mutation:slot machine ::...* 1. *bacteria:?* 2. *genes:?* 3. *mutating:?* 4. *reproducing:?* 5. *dying:?* options: *losing* *reels* *slot machines* *spinning* *winning*	1. *slot machines* 2. *reels* 3. *spinning* 4. *winning* 5. *losing*

Table 2. Content-based taxonomy of analogy problems included in the present inventory

Content distinction	Instance	*A:B::C:D* example
Relation type (especially relevant to AI problem sets)	Syntactic	*high:higher :: sad:sadder* (common relation: *adjective:comparative*)
	Semantic	*up:down :: rise:fall* (common relation: *opposite*)
Semantic distance (especially relevant to Cognitive Psychology problem sets)	Near / within-domain	*blindness:sight :: deafness:hearing* (shared domain: *human senses*)
	Far / cross-domain	*blindness:sight :: poverty: money* (*A:B* domain: *human senses*; *C:D* domain: *social resources*)
Word familiarity (especially relevant to Education problem sets)	Familiar	*happy:sad :: fat:skinny*
	Unfamiliar	*jubilant:melancholy :: corpulent:gaunt*

In a second type, which we will call *generative* problems, the reasoner is presented with incomplete analogies and is asked to complete the analogy by generating the missing terms, without being given a set of options from which to choose. This second problem type can be subdivided into two distinct subtypes: *one-term* generative problems (e.g., Green et al., 2012) and *two-term* generative problems (e.g., Popov, Hristova, & Anders, 2017). In the former, the incomplete analogy omits the *D* term only (e.g., *dog:puppy :: kitten:?*), and the reasoner is simply tasked with generating the *D* term. In the latter, the incomplete analogy omits both the *C* and *D* terms (e.g., *dog:puppy :: ?:?*), and the reasoner is tasked with searching through semantic memory for a complete exemplar of the relation instantiated by *A:B*.

In a third type, which we will call *multiple-choice* problems, the reasoner is presented with incomplete analogies, where either the *D* term is omitted or where both the *C* and *D* terms are omitted, just as with generative problems. However, the reasoner is also presented with a small number (e.g., four) of options to complete the analogy and is asked to select the best term(s) from that set of options (e.g., *dog:puppy :: kitten:?*; *a. child, b. tiger, c. cat, d. toy*; Turney et al., 2003). While we distinguish between one-term and two-term generative problems above, we do not draw a parallel distinction between one-term and two-term multiple-choice problems. Regardless of whether only the *D* term or both the *C* and *D* terms are omitted, solving multiple-choice problems simply involves selecting the missing terms from a small, highly constrained set of options.

Finally, in a fourth type, which we will call *matrix* problems, reasoners are provided with an *A:B* word pair (e.g., *helmet:head*) along with an large number (e.g., 20) of word pair options (e.g., ***mitt:baseball***, *amnesia:memory*, *boat:anchor*, *inventor:invention*, *eraser:pencil*, *spray:skunk*, *mother:daughter*, *antenna:signal*, ***kneepad:knee***), and are asked to select any options that independently form a valid analogy with the *A:B* word pair (Weinberger, Iyer, & Green, 2016; plausible responses in the parenthetical example above are bolded). Given the large number of word-pair options provided for each *A:B* word pair, matrix problems involve a larger search space for valid analogies than do the multiple-choice problems described above (typically with four options). However, the search space for matrix problems is still constrained to a predefined set of options, in contrast to the unconstrained search space that characterizes generative problems.

Verbal analogies can also be created that do not conform to the simple *A:B::C:D* format. Instead, an elaborated *source* analog (usually relatively familiar or concrete) can be related to a *target* analog (usually less familiar or more abstract). Solving analogy problems in this broader form, which we term *extended mapping* analogies, involves finding multiple relational correspondences between the two analog domains. This complex mapping process contrasts with the comparatively simple mapping process involved in solving *A:B::C:D* problems as described above. The source and target may take the form of rich narratives expressing different but analogous plots, ideas, or characters (e.g., Gentner & Toupin, 1986; Wharton, Holyoak, Downing, Lange, Wickens, & Melz., 1994) or expressing problems to be solved (e.g., Gick & Holyoak, 1980). The source and target may also be presented in a more skeletal form as two sets of constituent concepts (e.g., Turney, 2008).

In addition to variations in form, analogies can vary in the nature of their content (see Table 2). A wide range of abstract semantic relations based on word meaning can be used to generate four-term verbal analogies (Bejar, Chaffin, & Embretson, 1991). For example, the two word pairs in the analogy *up:down :: rise:fall* instantiate the relation *opposite*, whereas those in the analogy *joke:laughter :: injury:pain* instantiate the relation *cause:effect*. Syntactic or morphological relations based on word form can also be used to generate four-term verbal analogies (e.g., Mikolov et al., 2013c). For example, the two word pairs in the analogy *high:higher :: sad:sadder* instantiate the morphological relation *adjective:comparative*, and those in the analogy *make:makes :: support:supports* instantiate the relation *infinitive verb: present verb*.

More broadly, another content distinction is between *near* or *within-domain* analogies and *far* or *cross-domain* analogies

(e.g., Green et al., 2010; Holyoak & Koh, 1987). For example, *blindness:sight :: deafness:hearing* is a relatively near analogy, whereas *blindness:sight :: poverty: money* is semantically more distant. Solving far analogies is generally more difficult for human reasoners than is solving near analogies. Moreover, generating solutions for far but not near analogies facilitates, and is indeed a manifestation, of creative thinking in humans. Specifically, this process increases a tendency to think in terms of the relations between entities rather than in terms of their attributes (Vendetti, Wu, & Holyoak, 2014). This distinction is not exclusive to *A:B::C:D* analogy problems. Semantic distance has also been manipulated in longer narratives, and studies show that retrieving a semantically distant narrative to solve an analogous target problem is more difficult than retrieving a semantically close narrative (Keane, 1987).

Finally, the content of an analogy can vary with respect to the general familiarity of the individual words out of which it is constructed. For example, *happy:sad :: fat:skinny* is composed of relatively familiar words, whereas *jubilant:melancholy :: corpulent:gaunt* is composed of less familiar words. The difficulty of an analogy problem can be easily manipulated by using more or less familiar words (the latter yielding more difficult problems). Notably, this difficulty does not arise directly from increased complexity of the analogical mapping process. Rather, decreasing word familiarity decreases the likelihood that a human reasoner will fully grasp the individual word meanings, a precondition for successful analogical mapping. This variation in content highlights the importance of semantic knowledge in reasoning about analogies.

Aside from content, another important dimension of variation among analogy problem sets is their sheer size. For investigating insight, or spontaneous use of analogy to solve problems, even a single problem may suffice (e.g., Gick & Holyoak, 1980). For other purposes, such as neuroimaging studies, a set of roughly 100 well-controlled problems may be necessary (e.g., Green et al., 2010). For the purpose of evaluating AI models, it may be useful to have available a set of several thousand analogy problems (e.g., Gladkova et al., 2016).

We will now provide an overview of analogy problem sets that have been developed in education, AI, and cognitive psychology. This overview is not intended to be exhaustive, but rather to give a general guide to the types of datasets made available in the present inventory.

Problem Sets from Education

Verbal analogies are a popular way to train and test children's verbal aptitude. A number of educational materials (e.g., EnglishForEveryone.org) and standardized tests (e.g., the SCAT) provide useful sources of large verbal analogy problem sets. These materials are designed for children at different grade levels, and they include problem sets that can serve as benchmarks of analogical reasoning across development. For example, the SCAT test includes three versions representing three levels of difficulty. The first version is administered to children in 2nd and 3rd grade, the second version to children in 4th and 5th grade, and the third version to children in 6th and 7th grade. Combined with other standardized tests (e.g., SSAT, and older versions of the SAT) as well as other educational material, analogy problems drawn from education sources represent levels of difficulty appropriate for kindergarten through 12th grade.

The current inventory includes problem sets compiled from testing and test preparation materials for the SSAT (Kotchian & Simmons, 2012; Enrollment Management Association, 2017a & b; Varsity Tutors, 2007-2019a & b), SCAT (Center for Talented Youth Johns Hopkins University, 2013a, b, & c), and the SAT (Turney et al., 2003), as well as additional educational materials from MindWare (2007) and EnglishForEveryone.org (n.d.; see Table 3). These sets consist solely of multiple-choice problems with either one or two terms missing from *A:B::C:D* verbal analogies. Since these problems conform to a single type, variations in difficulty across grade levels are primarily attributable to the content of the analogies. As previously mentioned, content can vary according to the relation underlying an analogy (e.g., *synonym* versus *antonym*), the semantic distance between analogs (e.g., *mad:angry :: down:sad* versus *quick:fast*), and the familiarity of the words that comprise an analogy (e.g., *mad:angry :: down:sad* versus *livid:irate :: despairing::melancholy*). Relation and word familiarity are important dimensions for analogy difficulty; analogies designed for older students tend to feature more specific semantic relations and less familiar words. For example, compare an analogy from the SAT written for high school students, *lull:trust :: cajole:compliance*, with an item from the Elementary Level SSAT written for students in third grade, *listen:music :: read:book*.

While these problem sets are not accompanied by explicit human performance data, typical human performance can be inferred from mean scores on the tests. For example, Turney et al. (2003) compiled a set of 374 analogies from unofficial SAT preparation web sites, the Educational Testing Service (ETS) website (http://www.ets.org/), a book of actual SAT questions (Claman, 2000), and from other SAT guidebooks. Turney and Littman (2005) estimated human performance on this problem set using accuracy of the average test taker (college-bound high school seniors) on the SAT Verbal section. This approach assumes that analogy problems on the SAT Verbal subtest are as difficult as the rest of the SAT Verbal section, and that the analogy problems in this collection are as difficult as the analogy problems on the official SAT.

Overall, while all of these analogy problems are multiple-choice problems in the *A:B::C:D* format, the sets provide a wide range of diversity in problem difficulty, and are helpfully organized across multiple levels of difficulty corresponding to different grade levels as well as different stages of cognitive development.

Table 3. Summary of analogy problem sets taken from education materials

Problem set	Source	Grade	# problems
*Analogy Challenges	Mindware. (2007). *Analogy Challenges (Beginner Level)*.	K-2	124
SCAT Elementary Level	Center for Talented Youth Johns Hopkins University (2013a). School and college ability tests: Sample questions for 2nd and 3rd graders.	2–3	10
SCAT Middle Level	Center for Talented Youth Johns Hopkins University (2013b). School and college ability tests: Sample questions for 4th and 5th graders.	4–5	10
SCAT Upper Level	Center for Talented Youth Johns Hopkins University (2013c). School and college ability tests: Sample questions for 6th graders and all higher grades.	6–8	10
SSAT Official Guide Elementary Level – Grade 3	Enrollment Management Association (2017a). *The Official Study Guide for the Elementary Level Grade 3*.	3	7
SSAT Official Guide Elementary Level – Grade 4	Enrollment Management Association (2017b). *The Official Study Guide for the Elementary Level Grade 4*.	4	7
*SSAT for Dummies Middle Level	Kotchian, V. & Simmons, C. (2012) *SSAT and ISEE For Dummies*.	5–7	30
*SSAT for Dummies Upper Level	Kotchian, V. & Simmons, C. (2012) *SSAT and ISEE For Dummies*.	8–11	30
SSAT Varsity Tutors Elementary Level	Varsity Tutors. (2007–2019a). SSAT elementary level verbal: Analogies. https://www.varsitytutors.com/ssat_elementary_level_verbal-help/analogies.	3–4	109
SSAT Varsity Tutors Middle Level	Varsity Tutors. (2007–2019b). SSAT middle level verbal: Analogies. https://www.varsitytutors.com/ssat_middle_level_verbal-help/analogies.	5–7	100
SAT	Turney, P. D., Littman, M.L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pp. 482–489.	9–12	374
English for Everyone	EnglishForEveryone.org. Analogies Worksheets. https://www.englishforeveryone.org/Topics/Analogies.htm.	1–12	24–30

NB: All problem sets provide *multiple choice A:B::C:D* analogy problems. They are not organized according to relation

*These problem sets are not publicly available; however, their sources are available for purchase at the following links:

Mindware (2007): http://bloomingminds.com/analogy-challenges-beginner-level.html

Kotchian & Simmons (2012): https://www.dummies.com/store/product/SSAT-and-ISEE-For-Dummies.productCd-1118115554.html

Problem Sets from Artificial Intelligence

Verbal analogies have garnered a considerable amount of interest in AI research, most notably in the area of natural language processing (NLP). Developing NLP applications that can understand natural language requires a kind of dictionary that represents word meanings in some way. *Word embeddings* are a type of lexicon in which word meanings are represented as real-valued vectors in *n*-dimensional space (typically 300 dimensions, each with a continuous value). These gained popularity after Mikolov et al. (2013a) demonstrated that word embeddings produced by their Word2vec model using a recursive neural network could solve some verbal analogies. As a consequence, the Google analogy test on which Mikolov et al. (2013a) demonstrated this success became a popular benchmark that has since been used to test several other algorithms for producing word embeddings (e.g., Baroni, Dinu, & Kruszewski., 2014; Faruqui, Tsevskov, Yogatama, Dyer, & Smith., 2015; Schnabel, Labutov, Mimno, & Joachims, 2015; Zhai, Tan, & Choi., 2016). In general, verbal analogies have become a popular tool for evaluating the overall quality of word embeddings (Gladkova & Drozd, 2016; Gladkova et al., 2016).

The current inventory includes four problem sets that have been used to test NLP word embeddings: the Google analogy test (Mikolov et al., 2013a), the Google phrase analogy test (Mikolov et al., 2013b), the Microsoft Research test (MSR; Mikolov et al., 2013c), and the Bigger Analogy Test Set (BATS; Gladkova et al., 2016; see Table 4). Separately, the current inventory includes an additional dataset, the Turney set (Turney, 2008), which was developed outside of recent attempts to evaluate word embeddings, and features extended mapping problems. Because the Turney set differs in format from the four NLP sets, we will first describe the latter as a group, followed by a separate description of the former.

NLP Analogy Sets

All the NLP sets are very similar in general structure, consisting exclusively of evaluative problems within the *A:B::C:D* format. The analogy items are explicitly categorized according to their relations. For example, the MSR contains 8000 analogy problems organized into eight categories representing different relations. Across these four problem sets, the MSR includes syntactic relations only, the Google

Table 4. Summary of analogy problem sets taken from the AI literature

Problem set	Source	Relations (#)	Problems (#)	Relation types
Google analogy test	Mikolov, T., Chen, K, Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.	14	19,544	Semantic & syntactic
Google phrase analogy test	Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 3111–3119.	5	3218	Semantic
Microsoft Research (MSR)	Mikolov, T., Yih, W., & Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pp. 746–751.	8	8000	Syntactic
Bigger Analogy Test Set (BATS)	Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pp. 47–54, San Diego, California.	40	*99,200	Semantic & syntactic
Turney	Turney, P.D. (2008), A uniform approach to analogies, synonyms, antonyms, and associations, In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 905–912, Manchester, UK.	140	20	Semantic

NB: All problem sets, except for Turney, consist of *evaluative A:B::C:*D analogy problems. Turney consists of *extended mapping story / long text* analogy problems

*BATS consists of word pairs that can be combined to produce A:B::C:D evaluative analogy problems

phrase analogy problems include semantic relations only, and the Google analogy problems and the BATS feature a mix of syntactic and semantic relations. The analogies in these sets were all constructed by collecting a set of word or phrase pairs that instantiate a given relation, and then combining all word or phrase pairs from a given relation set to create four-term analogies (each consisting of two word or phrase pairs), using all possible permutations. For example, the analogies instantiating the relation *adjective:adverb* in the Google analogy test were constructed by combining 32 word pairs each instantiating that relation (e.g., *amazing:amazingly*, *apparent:apparently*) into all possible permutations, yielding a set of 992 (32 × 31) analogy problems for that relation.

The Google sets and MSR are more similar to each other than either is to BATS. First, the former tests include different numbers of analogies instantiating various relations. Overall, the number of word pairs per relation varies between 20 and 70 for the former two problem sets. BATS, on the other hand, standardizes the number of word pairs per relation at 50. Second, the Google sets and MSR exclusively consist of near analogies based on word pairs that involve both relational and more direct semantic similarity. BATS, on the other hand, includes both near and far analogies.

Turney

The Turney problem set, derived from psychological and linguistic studies, was created to evaluate a computational model of analogical reasoning, the Latent Relation Mapping Engine (LRME; Turney, 2008). This set includes 20 extended mapping problems. Ten problems involve scientific analogies taken from Holyoak and Thagard (1995), and the remaining ten problems feature common metaphors taken from Lakoff and Johnson (1980). Each analogy consists of a source domain and a structurally similar target domain between which a human reasoner can map a rich set of relational correspondences. The extended mapping problems in the Turney set reduce each of these analogical domains to two sets of 5–9 key terms. For example, in one problem, *solar system* is the source and *atom* is the target. This analogy includes seven key terms: *solar system*, *sun*, *planet*, *mass*, *attracts*, *revolves*, and *gravity* constitute the *solar system* set; whereas *atom*, *nucleus*, *electron*, *charge*, *attracts*, *revolves*, and *electromagnetism* constitute the *atom* set.

Turney (2008) had 22 participants complete these extended mapping problems. For each problem they were presented with the list of terms from the source and were asked to use the items from the target to "construct an analogical mapping; that is, a one-to-one mapping" between the two sets. The data reported are the percentage of participants whose responses matched the intended mapping for each problem.

Problem Sets from Cognitive Psychology

The current inventory (see Table 5) includes fifteen problem sets created for basic research in cognitive psychology: *Green-*

eval (Green et al., 2010), *Kmiecik* (Kmiecik et al., 2019), *Green-gen* (Green et al., 2012), *Popov* (Popov et al., 2017), *Jurgens* (Jurgens, Mohammad, Turney, & Holyoak, 2012), *Sternberg* (Sternberg & Nigro, 1980; Morrison et al., 2004), the UCLA Verbal Analogy Test (*VAT*; Lu, Wu, & Holyoak, 2019b), *Weinberger* (Weinberger et al., 2016), *Wharton* (Wharton et al., 1994), *Clement* (Clement & Gentner, 1991), *Rattermann* (Gentner, Rattermann, & Forbus, 1993), *Gentner* (Gentner & Toupin, 1986), *Gick* (Gick & Holyoak, 1980), *Gick 2* (Gick & Holyoak, 1983), and *Keane* (Keane, 1987). Overall, these problems sets are highly diverse. Some include analogies in the *A:B::C:D* format, consisting of either *evaluative* (e.g., Green et al., 2010), *one-term generative* (e.g., Green et al., 2012), *two-term generative* (e.g., Popov et al., 2017), or *multiple-choice* problems (e.g., Sternberg, 1980). Others involve more elaborate sources and targets expressed as analogous plots, characters, or ideas to be used in *retrieval* (e.g., Wharton et al., 1994), *generation* (e.g., Gentner, & Toupin, 1986), and *problem-solving* tasks (e.g., Gick & Holyoak, 1980). Because several of these problem sets were developed as experimental stimuli, many are associated with data on human performance. These data mainly take the form of reaction times, accuracy rates, and explicit judgments of difficulty. Because these analogy sets are so numerous, we describe a subset and the procedures for gathering corresponding data. Each problem set described below is meant to exemplify a different formal type of analogy problem (see Table 1).

A:B::C:D Evaluative Problems (Green-eval)

The Green-eval problem set was created to examine the neural bases of analogical reasoning (Green et al., 2010, 2012). This set contains 120 *A:B::C:D* evaluative problems. The set is organized into 40 triplets of analogy problems sharing a common *A:B* term, while varying the *C:D* term. Each triplet forms either a valid within-domain analogy, a valid cross-domain analogy, or an invalid analogy. In half of the triplets, the invalid analogy is constructed out of within-domain word pairs; in the other half, the invalid analogy is constructed out of cross-domain word pairs. Both the within-cross domain classification and the valid-invalid classification were established at a level of > 90% agreement among 84 human raters. Raters responded to the following prompt to make the within-cross domain classification: "Are the items in the left word pair taken from the same semantic domain as the item in the right word pair? That is, do the two-word pairs involve similar kinds of things or different kinds of things?"

Separately, the semantic distance between word pairs in all 120 analogy problems was estimated using latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). The LSA application (http://lsa.colorado.edu) calculates the similarity between the contextual-usage meanings of words as measured by the cosine of the included angle between vectors assigned to those words within a very high-dimensional semantic space derived from a large corpus of English text. Vectors for word pairs were obtained by adding together the vectors for the individual words within word pairs, and the semantic distance between word pairs was calculated by pairwise comparisons between the *A:B* word pair and the *C:D* word pairs used to form each problem in the analogy triplet (i.e., the within- and cross-domain analogous word pairs and the non-analogous foil). The mean semantic distance between word pairs was .55 for within-domain analogies and .91 for cross-domain analogies.

Green et al. (2010) had 84 native English-speaking undergraduates evaluate all 120 analogies, using a seven-point scale to respond to the following prompt: "How difficult is it to identify the analogical connection?" In addition, 23 participants performed an experimental task in which they were sequentially presented with each four-word analogy from the 120-analogy set, and asked to indicate whether the presented word pairs constituted a true analogy (i.e., analogous word pairs) or a false analogy (i.e., non-analogous word pairs). The human data accompanying the Green-eval problem set consists of the accuracy rates and reaction times from this task.

A:B::C:D One-Term Generative Problems (Green-gen)

While the Green analogy set contains evaluative analogy problems, it and any set of valid *A:B::C:D* analogies can be easily converted to generative analogy problems by dropping the *D* term from each problem. This was done by Green et al. (2012), resulting in the Green-gen set. Vendetti et al. (2014) also used the Green-gen set. In the latter study, a group of 54 English-speaking UCLA undergraduates generated solutions to the 80 valid analogies from the Green-gen set (40 within-domain and 40 cross-domain). Vendetti et al. (2014) showed that generating solutions to the cross-domain problems induced a relational mindset in participants (i.e., increased participants' propensity to think in terms of relations between entities rather than constitutive features of those entities). The human data accompanying this problem set consists of the accuracy rates from this task reported by Vendetti et al. (2014).

A:B::C:D Two-Term Generative Problems (Popov)

The Popov dataset (Popov et al., 2017) was created for use in a task that aimed to induce relational luring based on the unintentional and long-term priming of semantic relations. The set has also been used as a training set for a computational model of relation learning (Lu, Liu, Ichien, Yuille, & Holyoak,

Table 5. Summary of analogy problem sets taken from the cognitive psychology literature

Problem set	Source	Relations	Problems (#)	Analogical form	Question type
Green-eval	Green, A. E., Kraemer, D. J. M., Fugelsang, J., Gray, J. R., & Dunbar, K. (2010). Connecting Long Distance: Semantic Distance in Analogical Reasoning Modulates Frontopolar Cortex Activity. *Cerebral Cortex, 10*, 70–76	Indefinite	80	A:B::C:D	Evaluative
Kmiecik	Kmiecik, M. J., Brisson, R. J., & Morrison, R. G. (2019). The time course of semantic and relational processing during verbal analogical reasoning. *Brain and Cognition*, *129*, 25–34	5	720	A:B::C:D	Evaluative
Green–gen	Green, A. E., Kraemer, D. J. M., Fugelsang, J., Gray, J. R., & Dunbar, K. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *38*(2), 264–272	Indefinite	80	A:B::C:D	One-term generative
Popov	Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722–745	58	950	A:B::C:D	Two-term generative
Jurgens	Jurgens, D. A., Mohammad S. M., Turney P. D., & Holyoak K. J. (2012) SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 356–364	79	3218	A:B::C:D	Two-term generative
Sternberg	Sternberg, R. J. & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, *51*, 27–38.; Morrison, R.G., Krawczyk, D., Holyoak, K.J., Hummel, J.E., Chow, T., Miller, B., & Knowlton, B.J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*, 260–271	5	197	A:B::C:D	Multiple choice
VAT	Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*. *116*(10). 4176–4181	4	80	A:B::C:D	Multiple choice
Weinberger	Weinberger, A., Iyer, H., & Green, A. E. (2016). Conscious augmentation of creative state enhances "real" creativity in open-ended analogical reasoning. *PLoS ONE 11*, e0150773	Indefinite	10	A:B::C:D	Matrix
Wharton	Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., & Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, *26*, 64–101	Indefinite	14	Story / long text	Retrieval
Clement	Clement, C., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, *15*, 89–132	Indefinite	4	Story / long text	Retrieval
Rattermann	Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrieval from inferential soundness. *Cognitive Psychology*, *25*, 524–575	Indefinite	18	Story / long text	Retrieval
Gentner**	Gentner, D, & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277–300	Indefinite	54	Story / long text	Generative
Gick*	Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355	Indefinite	1	Story / long text	Problem solving
Gick2*	Gick, M. L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38	Indefinite	2	Story / long text	Problem solving
Keane	Keane, M. (1987). On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology*, *39*(1), 29–41	Indefinite	1	Story / long text	Problem solving

*Additional variations of ray problem analogs presented here can be found in:

Catrambone and Holyoak (1989)

Holyoak and Koh (1987)

**This dataset includes nine sets of characters, but only one of these character sets is accompanied by a story set

NB: All problem sets involve semantic relations

2019a; Lu, Wu, & Holyoak, 2019b). It contains 8–31 word pairs (e.g., *car:driver*, *jacket:cold*, *beer:mug*) instantiating each of 58 specific relations (e.g., *vehicle-driven-by*, *protects-against*, *container-of*), which were in turn categorized under one of ten relation types. Each word pair constitutes a separate two-term generative analogy problem, in which a reasoner might be presented with a word pair as an incomplete *A:B::C:D* analogy (e.g., *winter:summer :: ?:?*) and asked to generate an analogous word pair (e.g., *war:peace*). Word pairs grouped under the same specific relation offer a pool of valid responses to each analogy problem. In addition to providing a set of *two-term generative* analogy problems, word pairs in

this dataset may also be combined to create a set of *evaluative* analogy problems as in the NLP datasets (e.g., Mikolov et al., 2013a).

This dataset was constructed with Bulgarian words and using Bulgarian-speaking participants. Seventy-nine participants were presented with two-word pairs and were asked to generate three-word pairs exemplifying the same relation in descending order of the degree to which each word pair exemplified that relation. For example, a participant might be presented with Bulgarian translations of *pipe:water* and *artery:blood,* and then generate Bulgarian translations of *cable:current*, *drain:rain*, *system:medicine*, where *cable:current* is considered to best exemplify the same relation as *pipe:water* and *artery:blood* (something like *flows-through*), followed by *drain:rain* and then *system:machine*. The Popov dataset provides indirect estimates of relation typicality, based on the frequency with which participants generated a particular word pair for each relation, weighted according to the number of times that that word pair was generated first (weighted most heavily), second, or third (weighted least heavily).

Subsequently, a native Bulgarian speaker and fluent English speaker translated all word pairs into English. The dataset was cleaned, excluding word pairs no longer bearing any relation similarity to seeds when word pairs were translated into English, and combining word pairs that were distinguishable in Bulgarian but that had the same English translation.

A:B::C:D Multiple Choice Problems (Sternberg)

The Sternberg problem set was created to examine the development of analogical reasoning performance and strategy from mid-childhood to adulthood (Sternberg & Nigro, 1980). It contains 180 multiple-choice *A:B::C:D* analogy problems for which reasoners are asked to select the correct *D* term from two options. These problems are distributed evenly across five different relation categories: *synonymy*, *antonymy*, *category membership*, *linear ordering*, and *functional*.

Morrison et al. (2004) used a subset of the Sternberg analogy problems to examine the role of semantic inhibition in analogical reasoning. They collected human judgments of the semantic association between *C:D* (correct) word pairs and between *C:D'* (incorrect) word pairs. One-hundred fifty undergraduates were presented one pair at a time (e.g., *give:party*) and were asked to use a five-point scale to rate "how associated" the words in each pair were. From these judgments, a semantic facilitation index (SFI) was calculated for each analogy problem by subtracting the *z*-score of the semantic association rating of the *C:D'* word pair from that of the *C:D* word pair. The SFI value represents how much the semantic association between the *C:D* word pair relative to that of between *C:D'* word pair favors choosing the correct response for an analogy problem (e.g., a positive SFI implies the terms in the correct analogical option are more highly associated than are those in the foil). An additional group of 54 undergraduates were asked to solve these analogies, and their accuracy and correct response reaction times, along with SFI values, were used to select 24 analogy problems for use in a study examining analogical reasoning in frontotemporal lobar degeneration (FTLD).

A:B::C:D Matrix Problems (Weinberger)

The Weinberger problem set was created to examine how explicit instruction to think creatively influences analogical reasoning (Weinberger et al., 2016). Specifically, the Weinberger problem set was developed as a measure of open-ended analogical reasoning. It contains two matrices, each featuring five unique *A:B* stem pairs (e.g., *helmet:head*) and 20 unique completion pairs. For each stem pair, reasoners are instructed to select completion pairs to make as many valid analogies as they can, selecting only completion pairs for which they can describe how they are analogous to the *A:B* stem pair. Reasoners are allowed to select the same completion pair for multiple stem pairs in a given matrix. This option ensures that reasoners do not eliminate completion pairs from consideration for reasons other than failing to find a valid analogy with the relevant stem pair. Importantly, this problem set involves *open-ended* problems for which there is not necessarily an optimal response.

The construction of word pair matrices was informed by expert opinion, empirical testing, and computational modeling. Stem and completion pairs were drawn from stimuli used in previous studies that obtained high rates of accuracy, and the validity of analogies between stem and completion pairs was further assessed by domain experts. The two matrices of five stem pairs and 20 completion pairs were matched for the creativity of possible solutions. Specifically, they were matched on the mean semantic distance between stem pairs and completion pairs as measured using LSA (Landauer & Dumais, 1997; Landauer et al., 1998), described under the *Green-eval* problem set.

The resulting set of stem and completion pairs have a number of features that makes this problem set particularly useful for studying creativity as manifested in analogical reasoning. For example, correct completion pairs represent at least three levels of semantic distance (e.g., *kneepad:knee* – near; *mitt:baseball* – middling; *atmosphere:earth* – far) from the corresponding stem pair (e.g., *helmet:head*). This feature enables making a distinction between correct responses that are uncreative or boring from those that are creative or interesting. Further, the presence of several completion pairs that do not form a valid analogy with a given stem pair creates a distinction between responses that reflect meaningful creativity (i.e.,

correct responses that are semantically distant from the corresponding stem pair) and those reflecting meaningless and unconstrained divergence (i.e., incorrect responses, especially those that are semantically distant from the corresponding stem pair).

Story / Long-Text Retrieval Tasks (Wharton)

The Wharton problem set was created to examine the role of analogical similarity in accessing episodes in human memory (Wharton et al., 1994), and to evaluate a computational model of analogical retrieval, Analog Retrieval by Constraint Satisfaction (ARCS; Thagard, Holyoak, Nelson, & Gochfeld, 1990). It includes 14 sets of four stories, three of which were derived from materials in Seifert, McKoon, Abelson, and Ratcliff (1986) and five of which were derived from materials in the Rattermann set of 18 story sets (Gentner, Rattermann, & Forbus, 1993), which is also included in the inventory. Each set of four stories features the same basic events (e.g., receiving a call about a job, going shopping for items related to a job), and includes two stories each involving one of two unique story plots that were created by rearranging the sequence of propositions expressing these events. These unique story plots each correspond to a story theme (e.g., *counting your chickens before they're hatched*, versus *finding desperately needed employment*) dependent on more abstract causal relations. Pairs of stories sharing the same theme constitute analogous stories, and pairs of stories with different themes constitute disanalogous stories. These four-story sets are divided into two subsets, each containing one disanalogous story pair (see Table 6). Across all 14 story sets, two supersets were produced by compiling one disanalogous pair from each story set into one superset and compiling the other disanalogous pairs into the other superset. After reading the stories, 28 undergraduates responded to a questionnaire in which they used a six-point scale to indicate "how similar are the scenes being described". Analogous story pairs were rated as more similar than disanalogous story pairs, and disanalogous story pairs were rated as more similar than story pairs in different sets.

Wharton et al. (1994) gave UCLA undergraduates 1 min to read each story from one superset (target stories) and to rate each on a six-point scale for imagability (i.e., how easy the story was able to visualize mentally) in order to ensure semantic processing. Next, participants completed a 5-min logical reasoning experiment as a distractor task. Finally, participants were given an open amount of time to read each story from the other superset (cue stories), and write down what they were reminded from the target stories while reading each cue story. The human data associated with the Wharton problem set consist of participants' responses to this final task. Data reported are overall probabilities of being reminded of the target by the cue in each condition.

Story / Long-Text Generative Tasks (Gentner)

The Gentner problem set was created to examine the development of systematicity (i.e., sensitivity to parallels based on more complex relations) in analogical reasoning (Gentner & Toupin, 1986). The set comprises nine source stories, each of which can be adapted into either of two versions: systematic or non-systematic. Each source story follows a standard structure, within which slight adaptations determine whether the version is systematic or non-systematic. The standard structure includes three sections: an introductory section that introduces the characters, an event sequence describing some outcome, and a moral. Systematic versions include introductory sections that describe the protagonist in terms of some habit or relational trait (e.g., "There was once a very jealous cat"), whereas non-systematic versions feature introductory sections that describe the protagonist in terms of some relationally neutral trait (e.g., "There was once a very strong cat"). In addition, non-systematic versions do not include a moral section (see Table 7).

Each story version contains three character roles that can be filled in various ways to create stories to fit one of three different mapping conditions. In addition to the three character roles, each source story has three test characters (e.g., *seal*, *penguin*, and *dog*), which participants are asked to use in their main task of generating a target story analogous to the source story. Test

Table 6. Example of a story set used in Wharton problem set (adapted from Seifert et al., 1986). This example is from Wharton et al. (1994). Theme 1 can be expressed roughly as *counting your chickens before they hatch*, while Theme 2 can be expressed roughly as *finding desperately needed employment*

Set A:

Theme 1: Ernie was really encouraged about his interview for a security guard at the new factory in town. He thought he was saved. Ernie went to the shopping mall and hunted around for a dark blue security guard uniform, and bought several. The next day he received a phone call from the factory personnel director about the security guard position. Ernie was dismayed that he had wasted money. He didn't have a job.

Theme 2: Carl wasn't working at the time. He was very concerned because he had very little left in his bank account. Several days later he had lunch with the president about becoming a broker. Carl thought he had impressed people when he gave his resume to the investment firm. Carl went to the department store and tried on some suits, and got a few. He felt that he was moving up again.

Set B:

Theme 1: Ronnie thought she had it made because she thought she had done well in the audition for a keyboard player. Ronnie went to the music store, played some electric organs, and then purchased one. Later she got a message from the guitar player about playing keyboards. She wasn't in a band. Ronnie was depressed that she had run up her credit card.

Theme 2: Cindy was upset she that she had blown her savings. She wasn't employed. Cindy was really happy about her tryout as dancer for a new musical. That night she met the director about the dancer position. Cindy got over to some stores, searched for, and bought some leotards. She believed her troubles were over.

characters vary in their direct similarity, and their similarity to characters in the source story determine each mapping condition. In the high-transparency condition, source story roles are filled by characters that are physically similar to test characters filling the same role (e.g., *seal – walrus*, *penguin – seagull*, *dog – cat*). In a medium-transparency condition, source story roles are filled by characters that are not physically similar to any of the test characters (e.g., *seal – lion*, *penguin – giraffe*, *dog – camel*). In the low-transparency condition, source story roles are filled by characters that are physically similar to test characters filling different roles (e.g., *seal – cat*, *penguin – walrus*, *dog – seagull*). This low-transparency condition creates what is referred to as a "cross-mapping", such that mappings based on shared attributes (e.g., physical similarities between characters; *seal – walrus*, *penguin – seagull*, and *dog – cat*) differ from those based on shared relations (e.g., role-based similarities between characters shown in the low-transparency example above). This dissociation between attribute and relational similarity is more commonly manipulated in sets of visual analogy problems (e.g., Tohill & Holyoak, 2000), but its presence here is worth mentioning as it offers a useful way to examine the respective contributions of attribute and relation similarity in guiding inference. Since each source story can produce either a systematic or a non-systematic version, given three mapping conditions, each basic source story can produce six different specific source stories. In total, the Gentner set can produce 54 story problems.

Each analogy problem requires that the participant reproduce each story's plot using three designated test characters. After reading a given story, reasoners infer the roles of the test characters by reading that story's introduction with the story characters replaced by test characters. The particular roles that test characters fill in these introductions are determined by the mapping condition. For example, in the high-transparency condition, given the introduction, "There was once a very jealous cat who was friends with a walrus. The cat often said to the walrus, 'Don't ever play with anyone else but me.'", the test phase would begin with the following: "There was once a very jealous dog who was friends with a seal. The dog often said to the seal, 'Don't ever play with anyone else but me.'".

Gentner and Toupin (1986) tested child participants, a younger group (4–6 years old) and an older group (8–10 years old). Participants were asked to reproduce the stories using toys representing each character. Performance was assessed using six propositions representing major events and the outcome for each story. Systematic stories had a seventh proposition expressing the story's moral. A proposition was scored as correct if a participant expressed it either verbally or nonverbally with the correct characters. A proposition was scored as incorrect if a participant either omitted it from their reproduction or expressed it incorrectly by using the wrong characters. The group data are reported as mean accuracy rates for each age group and for the six story types.

Table 7. Sample story in systematic and non-systematic versions (systematic version includes indented material) from Gentner problem set. From Gentner and Toupin (1986)

Setting [a]: There was once a very jealous cat who was friends with a walrus. The cat often said to the walrus, "Don't ever play with anyone else but me."

One day the cat went away on a trip and the walrus had no one to play with. But then a seagull came to visit the walrus. He brought a wagon along and said, "Would you like to play with me and my wagon?" The walrus said, "Yes." The seagulls and the walrus had a great time pulling each other around in the seagull's wagon.

When the cat came back and found the walrus playing with someone else he got very angry. He shouted, "I'll never play with you again!" The cat was so angry that he jumped into the seagull's wagon. But the wagon began to roll faster down a steep hill. The car was very scared. The seagull jumped up and chased after the wagon so that cat wouldn't crash. The seagull stopped the runaway wagon and saved the cat's life.

Moral [b]: In the end, the cat realized that being jealous only got him into trouble. It is better to have two friends instead of one.

[a] *Setting*, non-systematic version: There was once a very strong cat who was friends with a walrus

[b] *Moral* is omitted in non-systematic version

Story / Long-Text Problem-Solving Tasks (Gick)

The Gick problem set was created to examine the role of analogy in problem solving. It centers around the radiation problem from Duncker (1945) in which reasoners are asked to solve the following problem:

> Suppose you are a doctor faced with a patient who has a malignant tumor in his stomach. It is impossible to operate on the patient, but unless the tumor is destroyed, the patient will die. There is a kind of ray that can be used to destroy the tumor. If the rays reach the tumor all at once at a sufficiently high intensity, the tumor will be destroyed. Unfortunately, at this intensity, the healthy tissue that the rays pass through on the way to the tumor will also be destroyed. At lower intensities, the rays are harmless to healthy tissue, but they will not affect the tumor either. What type of procedure might be used to destroy the tumor with the rays, and at the same time avoid destroying the healthy tissue?

In order to evaluate how analogy aids human reasoners in problem solving, the dataset includes a series of stories that vary in the degree that they present (1) analogous situations and (2) analogous solutions to the radiation problem. Here is an example of a story that presents both an analogous situation and an analogous solution to the radiation problem:

> A small country fell under the iron rule of a dictator. The dictator ruled the country from a strong fortress. The

fortress was situated in the middle of the country, surrounded by farms and villages. Many roads radiated outward from the fortress like spokes on a wheel. A great general arose who raised a large army at the border and vowed to capture the fortress and free the country of the dictator. The general knew that if his entire army could attack the fortress at once it could be captured. His troops were poised at the head of one of the roads leading to the fortress, ready to attack. However, a spy brought the general a disturbing report. The ruthless dictator had planted mines on each of the roads. The mines were set so that small bodies of men could pass over them safely, since the dictator needed to be able to move troops and workers to and from the fortress. However, any large force would detonate the mines. Not only would this blow up the road and render it impassable, but the dictator would then destroy many villages in retaliation. A full-scale direct attack on the fortress therefore appeared impossible.

The general, however, was undaunted. He divided his army up into small groups and dispatched each group to the head of a different road. When all was ready he gave the signal, and each group charged down a different road. All of the small groups passed safely over the mines, and the army then attacked the fortress in full strength. In this way, the general was able to capture the fortress and overthrow the dictator.

This set contains eight stories in total. Three of these stories are three similar versions of the above story that vary slightly in wording but that all present both analogous situations and solutions to the radiation problem. Two other stories use the same first paragraph of the above story and so present analogous situations to the radiation problem, but they present different, disanalogous solutions. One other story presents a disanalogous situation but an analogous solution to the radiation problem. A final pair of stories presents both disanalogous situations and solutions to the radiation problem.

Gick and Holyoak (1980) presented undergraduates with different combinations and subsets of these stories across several experiments. One experiment generated data on the rates at which participants proposed one of three different solutions to the radiation problem after having read one of three analogous stories, each featuring a different solution. Another experiment generated data on the completeness of participants' generated solutions to the radiation problem after either having read one of two stories or no story at all. A third experiment generated data on the rates at which participants were able to provide solutions to the radiation problem without having read any story, after having read at least one story, and after having read at least one story and receiving a hint to use it.

Conclusion

Verbal analogy tests are extremely varied in form and content, and we have offered a rough taxonomy to capture and systematize some of this variability (see Tables 1 and 2). Here, we provide a database consolidating analogy problem sets across education, AI research, and cognitive psychology research. Education offers a useful source of test sets with *A:B::C:D* multiple-choice problems tailored to specific age groups. AI research has provided large test sets with *A:B::C:D* evaluative problems. Finally, cognitive psychology research has generated highly varied test sets often accompanied by human performance data.

As a procedural note, we urge readers and users to cite original papers when using analogy sets accessed from the present inventory, and to respect copyright claims when reproducing test items. Our aim here is to provide a centralized source of analogy problem sets so as to facilitate future research on analogical reasoning in both humans and machines.

Acknowledgements Preparation of this paper was supported by NSF Grant BCS-1827374.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238–247).

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

Catrambone, R. & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*, 1147–1156.

Center for Talented Youth, Johns Hopkins University (2013a). School and college ability tests: Sample questions for 2nd and 3rd graders. Baltimore, MD: Johns Hopkins University.

Center for Talented Youth, Johns Hopkins University (2013b). School and college ability tests: Sample questions for 4th and 5th graders. Baltimore, MD: Johns Hopkins University.

Center for Talented Youth, Johns Hopkins University (2013c). School and college ability tests: Sample questions for 6th graders and all higher grades. Baltimore, MD: Johns Hopkins University.

Chan, J., & Schunn, C. D. (2015). The importance of iteration in creative conceptual combination. *Cognition*, *145*, 104–115.

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

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R.J. Sternberg, & J. Davidson (Eds.). *Mechanisms of insight* (pp. 365–395). Cambridge MA: MIT press.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, *58* (Whole No. 270).

EnglishForEveryone.org (n.d.). Analogies Worksheets. https://www.englishforeveryone.org/Topics/Analogies.htm .Accessed 9 Apr 2018

Enrollment Management Association (2017a). *The Official Study Guide for the Elementary Level Grade 3*.

Enrollment Management Association (2017b). *The Official Study Guide for the Elementary Level Grade 4*.

Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., & Smith, N. (2015). Sparse overcomplete vector representations. *arViv preprint: 1506.02004*.

Gentner, D., & Jeziorski, M. (1993). The shift from metaphor to analogy in western science. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed.) (pp. 447–480). Cambridge, UK: Cambridge University Press.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524–575.

Gentner, D, & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277–300.

Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gladkova, A. & Drozd, A. (2016). Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations NLP* (pp. 36–42).

Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pp. 47–54, San Diego, California.

Green, A. E., Kraemer, D. J. M., Fugelsang, J., Gray, J. R., & Dunbar, K. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *10*, 70–76.

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

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning*. New York: Oxford University Press.

Holyoak, K. J. & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15,* 323–340.

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

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T., Sageman, B., Ormand, C. J., & Tikoff, B. (2010). Analogical thinking in geoscience education. *Journal of Geoscience Education*, *58*(1), 2–13.

Jurgens, D. A., Mohammad S. M., Turney P. D., & Holyoak K. J. (2012) SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (**SEM*), pp. 356–364.

Keane, M. (1987). On retrieving analogues when solving problems. *Quarterly Journal of Experimental Psychology Section A, 39*(1), 29–41.

Kmiecik, M. J., Brisson, R. J., & Morrison, R. G. (2019). The time course of semantic and relational processing during verbal analogical reasoning. *Brain and Cognition*, *129*, 25–34.

Kotchian, V., & Simmons, C. (2012) *SSAT and ISEE For Dummies*.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologica*, *46*, 2020–2032.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lu, H., Liu, Q., Ichien, N, Yuille, A, & Holyoak, K. J. (2019a). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Montreal, Canada: Cognitive Science Society.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019b). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*, *116*(10), 4176–4181.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Mikolov, T., Chen, K, Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*, 3111–3119.

Mikolov, T., Yih, W., & Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751.

Mindware (2007). *Analogy Challenges (Beginner Level)*.

Morrison, R.G., Krawczyk, D., Holyoak, K.J., Hummel, J.E., Chow, T., Miller, B., & Knowlton, B.J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*, 260–271.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109–130.

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722–745.

Richland, L.E., Zur, O., & Holyoak, K.J. (2007) Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128–1129

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 298–307, Lisbon, Portugal: Association for Computational Linguistics.

Seifert, C. M., McKoon, G., Abelson, R. P., & Ratcliff, R. (1986). Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *12*, 220–231.

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

Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, *51*, 27–38.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*, 259–310.

Turney, P.D. (2008), A uniform approach to analogies, synonyms, antonyms, and associations, In *Proceedings of the 22nd International Conference on Computational*, pp. 905–912.

Turney, P. D. (2013). Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics*, *1*, 353–366.

Turney, P. D., Littman, M.L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pp. 482–489.

Turney, P. D., & Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*(1-3), 251-278.

Tohill, J. M., & Holyoak, K. J. (2000). The impact of anxiety on analogical reasoning. *Thinking and Reasoning*, *6*(1), 27-40.

Varsity Tutors (2007–2019a). SSAT elementary level verbal: Analogies. https://www.varsitytutors.com/ssat_elementary_level_verbal-help/analogies . Accessed 1 Jul 2018

Varsity Tutors (2007–2019b). SSAT middle level verbal: Analogies. https://www.varsitytutors.com/ssat_middle_level_verbal-help/analogies . Accessed 1 Jul 2018

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(3), 1–6

Weinberger, A., Iyer, H., & Green, A. E. (2016). Conscious augmentation of creative state enhances "real" creativity in open-ended analogical reasoning. *PLoS ONE*, *11*, e0150773.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., & Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, *26*, 64–101.

Zhai, M., Tan, J., & Choi, J. D. (2016). Intrinsic and extrinsic evaluations of word embeddings. In *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 4282–4283.

Open Practices Statement The materials contained in our Verbal Analogy Inventory are available as supplementary information. We offer no original data or experiments, and so preregistration is not applicable to the present work.