COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13036

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13036

Analogy Generation in Science Experts and Novices

Micah B. Goldwater,[a] Dedre Gentner,[b] Nicole D. LaDue,[c] Julie C. Libarkin[d]

[a]*School of Psychology, University of Sydney*
[b]*Department of Psychology, Northwestern University*
[c]*Department of Geology and Environmental Geosciences, Northern Illinois University*
[d]*Department of Earth and Environmental Sciences, Michigan State University*

Received 25 December 2019; received in revised form 16 July 2021; accepted 18 July 2021

Abstract

There is a critical inconsistency in the literature on analogical retrieval. On the one hand, a vast set of laboratory studies has found that people often fail to retrieve past experiences that share deep relational commonalities, even when they would be useful for reasoning about a current problem. On the other hand, historical studies and naturalistic research show clear evidence of remindings based on deep relational commonalities. Here, we examine a possible explanation for this inconsistency—namely, that remindings based on relational principles increase as a function of expertise. To test this claim, we devised a simple analogy-generation task that can be administered across a wide range of expertise. We presented common events as the bases from which to generate analogies. Although the events themselves were unrelated to geoscience, we found that when the event was explainable in terms of a causal principle that is prominent in geoscience, expert geoscientists were likely to spontaneously produce analogies from geoscience that relied on the same principle. Further, for these examples, prompts to produce causal analogies increased their frequency among nonscientists and scientists from another domain, but not among expert geoscientists (whose spontaneous causal retrieval levels were already high). In contrast, when the example was best explained by a principle outside of geoscience, all groups required prompting to produce substantial numbers of analogies based on causal principles. Overall, this pattern suggests that the spontaneous use of causal principles is characteristic of experts. We suggest that expert scientists adopt habitual patterns of encoding according to the key relational principles in their domain, and that this contributes to their propensity to spontaneously retrieve relational matches. We discuss implications for the nature of expertise and for science instruction and assessment.

Keywords: Analogy; Expertise; Science expertise; STEM education; Causal reasoning

Correspondence should be sent to Micah B. Goldwater, School of Psychology, University of Sydney, Brennan MacCallum Building (A18), Camperdown, NSW 2042, Australia. E-mail: micah.goldwater@sydney.edu.au

1. Spontaneous analogies often seed new discoveries

For example, Velcro was inspired by burrs stuck in a dog's fur, the Dyson vacuum was based in the mechanics of a sawmill, and Kepler's laws of planetary motion drew on an analogy to light. These advances were driven by the recognition of commonalities in the underlying principles across different contexts, despite a lack of superficial similarity. The historical record offers many such examples, in which scientists generated distant analogies that lead to new insights (Gentner, 2002; Gentner et al., 1997; Nersessian, 1984; Thagard, 1992). Thus, far analogical transfer occurs in scientific and technological innovation.

But there is a critical inconsistency in research in this arena, perhaps suggesting that these examples are uncharacteristic of typical analogical thinking. On the one hand, there is abundant evidence from laboratory research that people rarely experience spontaneous reminding based on deeper principle-based connections, even when it would be quite useful to do so, for example, in order to solve a problem (e.g., Gentner, Loewenstein, Thompson, & Forbus, 2009; Gick & Holyoak, 1980; Ross, 1997, 1999; Trench & Minervino, 2015). Instead, retrieval is typically driven by other forms of semantic similarity between a current context and prior experiences, which may be less useful for problem solving. However, in contrast to the laboratory work, naturalistic studies of real-world reasoning have reported higher levels of spontaneous analogical retrievals. For example, Dunbar and Blanchette's (2001) and Dunbar's (1999) studies of microbiologists working in their own labs documented abundant use of "close" analogies based on both shared deep principles and more superficial similarity, and some occurrence of "distant" analogies (which only share deep principles), especially in hypothesis generation. This naturalistic work suggests that the historical studies do paint a representative picture of analogical thinking among science experts.

Many attempts to explain this disparity have been made and these are reviewed below. First, however, we preview our own proposal. We suggest that a major determinant of principle-driven relational retrieval is expertise—specifically, that experts in a domain experience more relational retrievals useful for solving problems and generating explanations than do novices. This comes about because experts tend to represent domain phenomena in terms of domain principles. This promotes *uniform relational encoding* (Forbus, Gentner, & Law, 1995; Gentner et al., 2009).[1] Because both current phenomena and prior stored phenomena are likely to be represented according to the key principles that hold in the domain, the likelihood of a deep and useful match between a current case and a stored case is relatively likely. In contrast, novice representations are often idiosyncratic and context-driven, so that phenomena encoded at different times and in different contexts may fail to match in deep relational structure. The view that expertise favors principle-based relational retrieval is not new (e.g., Dunbar & Blanchette, 2001; Forbus, Carney, Sherin, & Ureel, 2005; Novick, 1988), and is fairly intuitive, but surprisingly, there is little direct evidence supporting it.

Our studies focus on scientific expertise—specifically on levels of expertise in geoscience. Like other scientific domains, geoscience emphasizes key abstract principles concerning cause and effect relationships that provide a unified explanation for disparate natural phenomena. Because many domain phenomena are encoded according to these principles, we suggest that in these domains, experts are likely to show spontaneous analogy generation and

retrieval. That is, the use of uniform relational principles across the domain makes is likely that even distant matches will share some causal commonalities (Gentner et al., 2009).

We next review the evidence that led us to our expertise account. First, we review evidence that domain experts are more likely to spontaneously encode domain examples in terms of domain relational principles than are novices. Then we go on to review research on analogical retrieval, including suggestive evidence that domain experts show higher rates of analogical retrieval than do novices. Because this evidence is incomplete, we then present an experiment that tests analogical generation across a broad range of expertise in geoscience.

1.1. Evidence for novice–expert differences in spontaneous encoding

There is considerable evidence that experts encode materials differently from novices. For example, Chi, Feltovich, and Glaser (1981) found that advanced physics students classified problems based on shared principles (such as conservation of momentum), while less advanced students classified by relatively superficial aspects of the problems (such as the presence of pulleys or ramps). Stains and Talanquer (2008) showed analogous findings in chemistry, with more advanced students classifying reactions by chemical mechanisms (such as an acid–base reaction), while less advanced students classified by more superficial aspects (such as the water as a product of the reaction). Using a sorting task, Rottman, Gentner, and Goldwater (2012) gave students example descriptions of natural phenomena that could be sorted either by domain (e.g., biology, economics, electronics) or by causal structure (e.g., positive feedback, negative feedback, causal chains). Physical science students from the "Integrated Science Program" at Northwestern University, who receive intensive training across science disciplines, sorted phenomena according to their causal structure, regardless of domain. In contrast, social science students sorted phenomena by their domain.

The evidence for novice–expert differences in encoding and classification suggests that science experts are likely to spontaneously encode phenomena according to the abstract relational principles of their domain. Over time, this pattern of encoding means that many of the expert's domain representations will be encoded according to principle-based relational structure. If the expert then encodes a new example according to these same principles, its relational structure is likely to match that of a prior example. Thus, experts will show a greater likelihood of spontaneous principle-based relational remindings than will novices; and this advantage will be specific to the experts' domain (rather than reflective of domain-general skills). This view is consistent with theories that postulate that expertise involves not only the accumulation of knowledge, but also changes in domain representation (Carey & Smith, 1993; Chi, 2006; Chi et al., 1981; Forbus et al., 1995; Glaser, 1984; Thagard, 1992).

1.2. Analogical retrieval and generation

As noted above, studies of analogical retrieval have found mixed results. A large number of laboratory studies have found that deep relational retrievals are rare. In these studies, subjects study a set of examples and are later given new examples. Sometimes, the dependent measure is just whether they retrieve the prior examples; in other studies, it is whether they apply the prior example to solve a novel problem (e.g., Catrambone & Holyoak, 1989;

Gentner et al., 2009; Gentner, Rattermann, & Forbus, 1993; Gick & Holyoak, 1980, 1983; Holyoak & Koh, 1987; Loewenstein, Thompson, & Gentner, 2003; Olguín, Trench, & Minervino, 2017; Ross, 1997, 1999; Trench & Minervino, 2015). The patterns across these two measures (e.g., what affects their frequency), however, are quite similar, which is why retrieving useful prior cases is considered a primary hurdle in problem solving (see Goldwater & Jamrozik, 2019, and Trench & Minervino, 2020 for reviews). Purely relational retrievals that match the novel case or problem in the most relevant and useful manner have been extremely rare; rather, participants tend to retrieve prior examples that share a collection of semantic similarities, such as superficial features and contextual features less relevant to the target case or problem, or overall similar examples, which share both these more superficial features and deeper relational structure.[2]

In contrast, naturalistic studies report a higher rate of retrieving relational matches useful for reasoning. For example, in Dunbar and Blanchette's (2001) and Dunbar's (1999) studies of research in microbiology labs, it was found that relational analogies constituted 25% of the retrieved cases and were an important source of new hypotheses. Although some of these were "overall-similar" (i.e., they shared both structural and superficial commonalities), there were also some purely relational remindings, especially when trying to develop new explanations. Naturalistic studies of teams of designers and scientists have also documented frequent use of analogies that show little superficial similarity between analogs (e.g., Chan & Schunn, 2015; Chan et al., 2011; Fu et al., 2013).

How do we account for the disparity in relational retrieval between laboratory and naturalistic studies? One possibility is that laboratory tasks underestimate people's actual rate of useful analogical retrieval in everyday life (Dunbar & Blanchette, 2001). Given that laboratory tasks have typically tested retrieval of material learned earlier in the study, subjects' poor analogical retrieval could stem from lack of familiarity and/or interest in the material. Another way in which laboratory studies may underestimate the natural rate of relational retrieval is that they often use the somewhat artificial task of asking subjects to retrieve material given a probe, rather than providing participants with motivating goals (Blanchette & Dunbar, 2001; Blanchette & Dunbar, 2000). In support of this claim, Blanchette and Dunbar (2001) ran a different kind of laboratory study, in which typical experimental participants were asked to retrieve a novel persuasive analogy from their own knowledge. These participants produced many more distant relational retrievals than did participants who were asked simply to retrieve prior information.

Blanchette and Dunbar's (2001) findings generated considerable interest among analogy researchers. Their case for the importance of participants' goals and of their prior knowledge of and interest in the material dovetailed with the views of others who were skeptical concerning the laboratory findings (e.g., Hofstadter & Sander, 2013). However, further research showed that these factors do not fully account for the pattern of rare relational retrieval found in laboratory studies. In an ingenious study, Trench and Minervino (2015) tested whether people would show purely relational retrieval in a domain that is both highly familiar and highly interesting to them—namely, popular movies that the subjects had seen many times. They devised probe problems that were analogous to key plots in the movies, and asked subjects how they would solve the problem. When the probes shared both superficial and structural

information with the past event (the movie plot), subjects readily accessed the movie plot and transferred it to the new situation (70% of trials). But when the probe shared only structural information with the movie plot, subjects largely failed to retrieve it (15% of trials). Despite the fact that subjects had seen the movies many times, purely relational similarities were not enough to lead to spontaneous remindings. In a second study, Trench and Minervino (2015) found a similar pattern when participants were asked to generate persuasive analogies in scenarios that were designed to relationally match their own autobiographical knowledge. In both of these studies, the materials were highly familiar and interesting to the participants, and the task involved a motivating goal—to solve a dilemma or to persuade a character. Yet, purely relational similarities were not enough to lead to spontaneous remindings. Thus the "rare relational retrieval" pattern does not seem to be simply an artifact of laboratory studies.

Thus, we need to look further for the explanation of the disparate results across analogical retrieval studies. In many of these studies, expertise is a possible factor. For example, in Dunbar's (1999) research on microbiology laboratories, which showed 25% relational analogies (Dunbar & Blanchette, 2001), the participants were highly expert in their domains. Likewise, the groups studied by Schunn and colleagues were professionals from a range of design-related disciplines (e.g., industrial design and mechanical engineering; Chan & Schunn, 2015; Chan et al., 2011; Christensen & Schunn, 2007; Fu et al., 2013). It is possible that the high rate of deep relational retrievals found in these studies resulted at least in part from the expertise of the participants.

The most direct evidence for the claim of a novice–expert difference in relational retrieval comes from a landmark study by Novick (1988). She gave Stanford students mathematics problems to study, and later presented them with a new problem to solve. Students with high mathematical ability (as assessed by the MSAT mathematics aptitude test) were significantly more likely to retrieve prior problems based on the same solution principle as the given problem than were students with lower mathematical ability, who typically retrieved surface-similar problems.

Novick's findings are the best evidence to date for the idea that experts are better able to spontaneously retrieve prior examples that share key domain principles. However, they leave important questions unanswered. Here, we expand on Novick's research in four ways. First, we studied people's retrievals from their own prior knowledge bases, rather than using a study-and-test design. This allows us to focus on our central claim: that years of systematic engagement in a field leads to a more principle-based knowledge representation in long-term memory. Second, we greatly extend the range of expertise tested. In Novick's study, the novices and experts were undergraduates at different levels of math achievement. In our study, the groups range from people with little or no science background to intermediate scientists to active researchers who are many years past completing their PhD.

Third, we compare spontaneous relational retrieval to guided retrieval, allowing us to distinguish people's *propensity* to spontaneously retrieve causally similar examples from their *ability* to do so when asked. This is a critical distinction because our key predictions concern spontaneous retrieval, and because spontaneous retrieval is critical in real-world scientific reasoning and discovery.[3] A fourth point is that the use of the MSAT to separate

novice and experts in Novick's studies leaves open the possibility that the two groups differed in aptitude as well as (or instead of) experience in a specific domain. We have tried to separate these factors by using amount of education[4] rather than scores on a test and by testing scientists from different domains on materials both inside and outside of their domain, allowing us to test our prediction that the effects of expertise on relational retrieval should be domain-specific.

To test this claim, we compare four groups of people who vary in their level of science and geoscience expertise on the analogy generation task (the AGT). In the AGT, participants are given an example and asked to produce another example that is like it. The examples they are given are common events familiar to nonscientists. However, when that explanation is based in a principle that is prominent in geoscience, we predict that spontaneous principle-based analogy generation will be most likely for expert geoscientists, less likely for intermediate geoscientists, and least likely for nonscientists and scientists from other fields. When that explanation is from another scientific domain, we do not expect any advantage for geoscientists. Our account of expertise in analogical retrieval posits that the advantage is driven by domain-specific knowledge, not differences in more general reasoning abilities.

We focus particularly on analogical retrieval (or generation[5]) based on shared causal explanations. A central goal of scientific inquiry is to identify the causes of phenomena (Mackie, 1974; Thagard, 1978). In the words of Imre Lakatos, "A proposition might be said to be scientific only if it aims at expressing a causal connection…" (Lakatos, 1970, p. 102). Thus, it seems reasonable to expect that for expert scientists, encodings in the domain are influenced by key causal relational principles. Encoding the explanation for the phenomena is part and parcel with encoding the phenomena themselves. One possibility is that when expert scientists encounter a new phenomenon in their domain, part of how they encode the phenomenon is its causal explanation. If so, this means that (a) over time expert scientists will amass a long-term knowledge base in which phenomena have been construed in terms of causal structures; and (b) when presented with a new example, expert scientists are likely to encode the example in terms of these explanatory causal principles. This pattern of *uniform relational representation* (Forbus et al., 1995) makes it likely that the example's representation will match some previously encoded phenomena in the domain—thus increasing the possibility of a spontaneous principle-based relational retrieval.[6]

We chose the field of geoscience for this investigation. Geoscientists routinely reason about the causes underlying Earth's structure and features. Often, the task of geoscience is to describe the present state of the world and explain this state in terms of the causal forces that made it this way, further fostering the prediction of future changes or retrodicting past changes. For example, a geoscientist who sees a U-shaped formation of rock layers may encode it as a syncline (description), along with the possible explanation that the formation was caused by tectonic compression (causal explanation).

1.3. The analogy generation task

We needed a task that could be applied over a broad range of expertise, from expert geoscientists to people with little or no scientific expertise. The task needed to be brief, not only

because experts have limited time to engage in such a task, but because a brief task would be easier to adapt to classroom activities (see Section 4). Therefore, we devised a novel method—the AGT—to elicit remindings from long-term memory. The basic idea is to give people a target phenomenon and ask them to produce a similar phenomenon and to explain their rationale. The idea is that what people produce as analogous tells us something about how their knowledge is encoded. Further, by framing the task to the participants as about analogy generation, and not memory retrieval, we maximized the chances that our participants (novice or expert) would use relational information (as argued by Dunbar & Blanchette, 2001).

An example question is, "A balloon floating is like _____because _____".

The cause of balloons floating is explained by Archimedes' principle. The upward force exerted in a balloon is equal to the weight of the air displaced by the balloon. In the case of a balloon filled with helium, for example, the helium plus the balloon material is significantly less dense than the surrounding air. The explanation for floating balloons that focuses on density differences is introduced in school at an early age (even if not referred to as "Archimedes principle" and the deeper explanation of how forces work is not offered until years later). However, we see little reason to assume that nonscientists spend much time considering Archimedes' principle as a causal mechanism for the movement of materials. In contrast, this principle underlies much of plate tectonics, a central paradigm in geoscience, which explains Earth's major topographic features. Thus, if geoscience experts are more likely to encode phenomena in terms of causal principles from their field, then we would expect that they are likely to generate analogies in which the two phenomena under consideration have common explanatory causes—for example, a balloon floating in the air is like oil rising above water because each is less dense than its surroundings. In contrast, geoscientists at an intermediate level of experience, novice nonscientists, and scientists from other domains are likely to generate analogies based on similar objects, associations, or first-order events—for example, relating a helium balloon floating to a leaf floating because both involve objects suspended in air.

The AGT allows us to address three related predictions. First, when asked to produce a similar example, more experience in science will overall lead to producing more examples that share a causal explanation. Expert scientists should generate more principle-based analogies than novice scientists. Second, prompting people to focus on causal explanations should overall lead to more analogies generated based on common causal explanations. While intuitive, this pattern is only really possible if (a) people have some causal knowledge of the phenomenon in question (such as floating balloons), but (b) this knowledge is not a highly prominent aspect of their understanding, so they might not think about it without a cue.[7] This, however, leads us to the critical third prediction. Our claim is not simply that with more scientific experience, people increase their *amount* of scientific knowledge, but the scientific expertise leads to a change in *how knowledge is represented*. Specifically, our contention is that for expert scientists, causal encodings are the default mode of representation; therefore, they will *spontaneously* use the causal principles from their domain without any prompting. Thus, the third prediction is that the effect of prompts to focus on causal explanations when generating analogies will be smaller (or even nonexistent) for expert scientists. Importantly, this only applies to principles that are prominent within the expert's domain of expertise. This

stems from our claim that this pattern depends mostly on knowledge representation, rather than on, for example, general intelligence. Scientists may have causal knowledge outside of their domain, especially when it concerns familiar common events (e.g., the cognitive scientists reading this paper probably are aware of the causal explanation for floating balloons), but this causal explanation may not be considered without prompting, because cognitive scientists do not spend much time thinking about density gradients and Archimedes' principle (and thus do not encode events in these terms).

To test these predictions, we developed two versions of the AGT. The open-ended AGT (exemplified above and in the Section 2) does not guide the participant toward causal principles, or any other specific basis for an analogy; it simply asks them to retrieve a phenomenon similar to the example shown, and to say why they are similar. This assesses participants' spontaneous patterns of retrieval and use of past experiences. In addition, we developed the prompted AGT in which participants were explicitly instructed to focus on common-cause relationships (see Section 2.1). We expected participants to produce more common-cause analogies in the prompted AGT than in the open-ended AGT, based on Dunbar and Blanchette's finding that task goals influence patterns of analogical retrieval.

We distinguish between the effects of overall scientific knowledge and experience from the effects of domain-specific expertise in two ways. First, we collect data from four groups with different kinds of science experience: novice nonscientists, intermediate geoscientists (who have some experience in geoscience, but less than experts), expert geoscientists, and expert scientists from another domain (here, vision scientists). Second, we used one AGT item where the explanation is from geoscience (floating balloons), and a second item where the causal explanation is common knowledge, but does not draw on geoscience principles (i.e., catching the common cold; see Section 2.1).

Specifically, our first prediction is that within geoscience, the production of spontaneous causal analogies will be highest for expert scientists, followed by intermediate geoscientists, followed by novices. Second, overall, the prompted AGT will elicit more principle-based analogies than the open-ended AGT. Critically, the third prediction is for interactions between the condition of AGT, the type of science experience, and the specific item. Expert geoscientists should show the smallest effect of prompting, but only for the balloon item. So, for example, expert vision scientists and geoscientists should show similar increases in principle-based analogies in the prompted condition for the "catching a cold" item, but only vision scientists should show a similar increase for the balloon item.

2. The experiment

We present our research as a single experiment because it leads to the most coherent way of statistically analyzing our results, but we note that data collection happened in multiple waves over multiple years. We collected data from attendees at a professional geoscience conference in two different years, recruiting both intermediate and expert geoscientists. We collected data from experts in vision science from a professional email listserve. Last, we collected data from novice nonscientists from Amazon's Mechanical Turk. We used two AGT

Table 1
n for each level expertise across for each item. The geoscientists who answered the cold question comprise a distinct group from those who answered the balloon question. The other groups consist of a single sample each

AGT Condition	Item	Expert Geoscience	Intermediate Geoscience	Expert Vision Science	MechTurk Novices
Open-ended	Balloon	42	52	28	63
	Cold	24	39		
Prompted	Balloon	33	56	21	64
	Cold	22	51		

items, the balloon and cold items mentioned above. The cause of catching a cold (i.e., from a virus) is common knowledge, no one in our study was an expert in medicine or virology. This created a 4 (Science Experience: Expert Geoscience, Intermediate Geoscience; Expert Vision Science; Novice) X 2 (AGT Condition: Open-Ended, Prompted) X 2 (Item: Balloon, Cold) design. Open-ended versus prompted was manipulated between subjects. The same vision scientists and the novices each answered the balloon and cold AGT items. Different sets of geoscientists (both intermediates and experts) answered the balloon and cold items. Our mixed-effects logistic regression models account for this asymmetry in design.

2.1. Methods

2.1.1. Participants

Four levels of science expertise were examined (see Table 1): expert geoscientists, intermediate geoscientists, expert vision scientists, and novice nonscientist recruits from Mechanical Turk. We excluded 28 of the initial 155 Mechanical Turk subjects from all analyses because they reported having completed a bachelors or graduate degree in a science—technology–engineering–mathematics field.

The first group of geoscientists (who answered the balloon item) were 183 attendees at a professional geoscience conference held in 2010, who volunteered to participate in the study at a research booth (see Kalafatis & Libarkin, 2019). Based on their answers to a questionnaire, the professional participants were divided into two groups: expert geoscientists ($n = 75$), whose education was master's level or above, and intermediate ($n = 108$), whose education was below the master's level. Among the respondents, five were excluded from analyses because they did not indicate their level of education; another five intermediate-level respondents were excluded for failure to complete the task. Among the experts, 47 participants had PhDs and 28 had masters in geoscience.

The second group of geoscientists (who answered the cold item) were 146 attendees at a geoscience conference held in 2011, who volunteered to participate in the study at a research booth. As above, the professional participants were divided into two groups based on their answers to a questionnaire: experts ($n = 46$), whose education was master's level or above, and intermediate ($n = 90$), whose education was below the master's level. Among the experts,

22 participants had PhDs and 24 had masters. An additional 10 scientists were excluded from the analyses for not completing the questionnaire or the AGT.

The vision scientists were recruited through an email listserve for professional vision scientists via a link to the survey contained in an email sent in January 2016. While 157 people clicked on the survey link, 50 completed the survey. Of those 50, one had completed only a bachelor's degree, seven had completed a master's, and the remaining 42 had completed a PhD. To match the qualifications of the geoscience experts, we considered the 49 responses of the participants with a master's degree or PhD.

All participants were randomly assigned the open-ended or prompted AGT (see Table 1).

2.1.2. Materials and procedure

The first group professional geoscientists (2010) received a packet with an instruction page and a page with the Analogy Generation question,[8] "A balloon floating is like _____ because_____". There was ample blank space below the question for their selection and their explanation of why the two phenomena were similar.

For the open-ended task, the instructions were "Analogies are based on similarities between two things. Because things may be similar in many ways, there are many ways to draw an analogy and no single correct way. We are studying students' analogies. As part of this study, we would like to compare students' analogies with those drawn by experts. Please complete the analogies on the back side of this page."

For the prompted task, participants were similarly told about the comparison between students and expert analogies, but had different instructions concerning the kind of analogies to focus on: "Science is mostly about understanding what causes things to happen. Often, this understanding is built through analogical reasoning. In this survey, you will be asked to build analogies based on causal similarities." The instructions further included these illustrative examples:

> *An example of a causal analogy is*:
>
> Getting in an auto accident is similar to tripping on a step because they both can be caused by not paying attention.
>
> *An example of a non-causal analogy is*:
>
> Getting in an auto accident is similar to tripping on a step because they both can result in getting hurt.
>
> The first example is causal because it relates an underlying and common reason for two distinct events. This second example is non-causal because the similarity is a common outcome of two distinct events.

Expert and Intermediate geoscientists sat at a table in the research booth at the professional meeting and filled out the packet with a pen. After this first batch of data collection, some changes were made to the survey. First, we added the "Cold" AGT item, "Catching a cold is like_____because_____." As said above, the second batch of geoscientists did not respond to the balloon item, but the novices and vision scientists responded to both. Like the first batch of geoscientists, the second completed the task at a booth with paper and pen.

In addition, for the second batch of geoscientists and the novices, there were subtle changes to the instructions. The words "noncausal analogy" were changed to "an analogy that does not focus on the cause" because we realized that it was misleading to label an analogy that is based on a common result as "noncausal."

For the final group of participants, the vision scientists, we modified the instructions further. On the advice of a consultant vision scientist (who helped with the recruitment), we modified the instructions of the tasks to give more guidance in both the open-ended and the prompted tasks. This consultant was concerned that the vision scientists would not complete the survey otherwise. But as you see below, there is no reason to think that rates of causal analogies would specifically be affected by these changes.

The open-ended instructions on the survey sent to vision scientists read as follows:

Analogies are based on similarities between two things. Because things may be similar in many ways, there is no single correct answer as to what is similar to something else.
For example, let us say an example started like this "An auto accident is like _____ because_____"
One could say "Getting into an auto accident is like being a pool ball getting knocked around because everything is out of your control"
We are studying the analogies that students and vision scientists produce from their experience. This is not to see whether people can produce the "correct" analogy. Rather, we are interested in what naturally comes to mind.
With this in mind, please complete the analogies on the next few pages (and give your explanations).

The prompted instructions on the survey sent to vision scientists read as follows.

Science is mostly about understanding what causes things to happen. Often, this understanding is built through analogical reasoning. In this survey, you will be asked to build analogies based on causal similarities—specifically, on things that have similar, or even the same causes.
Here is an example of an analogy based on common causes:
Getting in an auto accident is similar to tripping on a step because they both can be caused by not paying attention.
Here is an example of an analogy that is not based on common causes:
Getting in an auto accident is similar to tripping on a step because they both can result in getting hurt.
The first example is a common-cause analogy—it is based on a common underlying cause (not paying attention) for the two distinct events. This second example is not a common-cause analogy—the similarity is about the result of an event, not about what causes it. Your task is to build analogies based on common causes, not common results.

The task took between 5 and 10 min for all groups of participants.

Table 2
Examples of each kind of analogy scored for "A balloon floating is like____ because____"

Kind of Analogy	Example Response
Correct explanatory (Common cause)	Hot water in a cold sea. Both rise in a dense fluid as a result of having a lower density
Incorrect causal	A leaf in the wind. Main propulsion is due to wind.
Social causal	A mother taking her child to college for the first time. They both have to be let go of.
Noncausal	Flying a kite. Both are in the air.

Table 3
Examples of each kind of analogy scored for "Catching a cold is like _____ because____"

Kind of Analogy	Example Response
Common cause	Catching the flu, both are caused by a virus
Result	Being hungover, both make you feel terrible.
Noncausal	Falling asleep; it is bound to happen.

3. Results

3.1. Scoring

Balloon: Detailed scoring rubrics were developed, and responses were scored by an expert in geoscience (see Online Appendix A). After discussion of the rubric and practice analogies, a second scorer, also an expert in geoscience, scored a subset (99) of the responses. Both scorers were blind to level of expertise, and the second scorer was also blind to whether the task was open-ended or prompted. The two scorers achieved a Cohen's Kappa for interrater reliability of 0.78, which is considered "substantial" agreement. The data reported here are based on the scores of the primary scorer.

Responses were coded as belonging to one of four main categories: common-cause (i.e., common explanatory cause), result (common result); and noncausal and uncertain. Within the common-cause category, there were three subtypes: correct cause, social cause, and incorrect cause (see Tables 2 and 3 for examples of each). *Correct cause* responses drew upon established physical or chemical causal principles that could serve as causes of both analogs. These support abductive inference in geoscience. *Social cause* responses drew on human activities instead of on physical and chemical mechanisms. *Incorrect cause* responses drew on physical causes but were inaccurate. *Result* responses stated a common result. These were rare ($\sim$2% of responses). *Noncausal* responses included relating the topic to directly observable entities and substances and simply restating all or part of the process in the prompt.

Cold. The coding was adapted (by the first author) for this question because there was a greater variety of common-cause responses, as there were multiple perspectives to focus

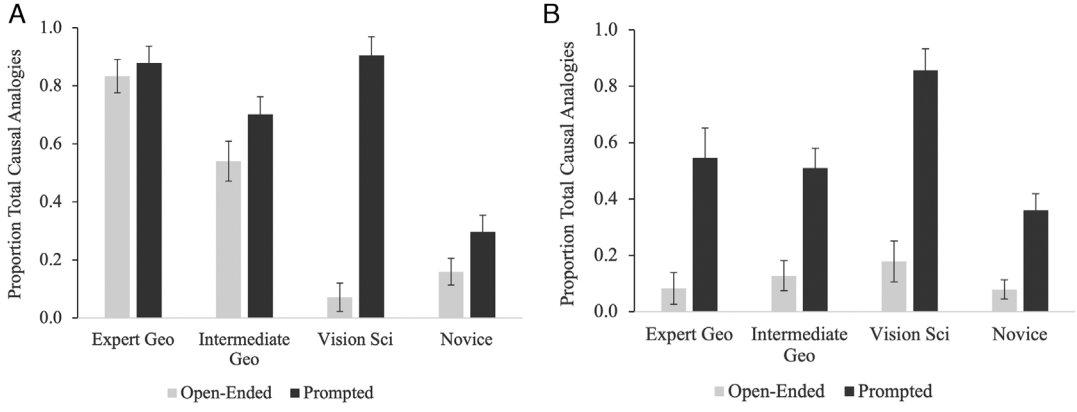

Fig 1. (A) Proportions and standard errors of the proportion of total common cause responses across conditions and science experience levels for the Balloon item. (B) Proportions and standard errors of the proportion of total common-cause responses across conditions and science experience levels for the Cold item.

on. For example, some subjects focused on how a cold is caused by a viral infection and drew comparisons to other illnesses caused by a virus, such as the flu or acquired immunodeficiency syndrome. Others took a more metaphorical tack, comparing the cold virus to an external invader, such an army attacking a fortress. Others focused on the idea that the cause of catching a cold is due to contagion from other people. Finally, some focused on the internal processes (of the cold recipient): for example, on a weakened immune system, which, in turn, could be caused by a stressful environment. Because these all focus on the causes of colds, we grouped them together in a single common-cause category.

Two other classes of analogies were coded. The first were results-based analogies that focused on the resulting symptoms of the cold, or on negative effects on one's life that a cold can cause. The second were noncausal analogies that did not focus on the cause or the result of catching a cold, for example, "catching a cold is like slipping on ice because they can happen to anyone." The first author explained the coding rubric to a research assistant who then coded all of the responses. The first author additionally coded a random sample of forty responses, on which they achieved an interrater reliability of Cohen's Kappa = 0.82.

3.2. Analysis

Our coding schemes were developed to capture a potential variety in responses, however, to allow for coherent analyses where we can compare items, and conditions together that capture the essential pattern in our results, we focus on a binary: a common-cause analogy = 1, or not = 0 (see Fig. 1). In Tables 4 and 5, we present the full breakdown of every category of response. For the balloon item, this puts results and noncausals as 0, and correct cause, incorrect cause, and social cause as 1. Tables 4 and 5 show that the vast majority of responses were correct cause and noncausal (the other categories were quite rare), and the patterns of results are the same if we just analyzed those two categories. However, we grouped them all together to have more consistency with the cold item where distinguishing between correct

Table 4
Counts (proportions) for the Balloon Item

Response Type	AGT Condition	Expert Geoscience	Intermediate Geoscience	Expert Vision Science	MechTurk Novice
Common cause	Open-ended	33 (0.79)	25 (0.48)	1 (0.04)	2 (0.03)
	Prompted	26 (0.79)	30 (0.54)	15 (0.71)	10 (0.16)
Incorrect causal	Open-ended	2 (0.05)	3 (0.06)	1 (0.04)	8 (0.13)
	Prompted	3 (0.09)	10 (0.18)	4 (0.19)	7 (0.11)
Social causal	Open-ended	0	0	0	0
	Prompted	0	0	0	2 (0.03)
Noncause	Open-ended	6 (0.14)	19 (0.37)	20 (0.71)	48 (0.76)
	Prompted	3 (0.09)	15 (0.27)	1 (0.05)	45 (0.70)
Result	Open-ended	0	1 (0.02)	3 (0.11)	5 (0.08)
	Prompted	1 (0.03)	0	0	0
Uncertain	Open-ended	1 (0.02)	4 (0.08)	3 (0.11)	0
	Prompted	0	1 (0.02)	1 (0.05)	0
Total	Open-ended	42	52	28	63
	Prompted	33	56	21	64

Table 5
Counts (proportions) for the cold Item

Response Type	AGT Condition	Expert Geoscience	Intermediate Geoscience	Expert Vision Science	MechTurk Novice
Common	Open-Ended	2 (0.08)	5 (0.13)	5 (0.18)	5 (0.08)
Cause	Prompted	12 (0.55)	26 (0.51)	18 (0.86)	23 (0.36)
Noncause	Open-Ended	7 (0.29)	15 (0.38)	15 (0.54)	33 (0.52)
	Prompted	3 (0.14)	12 (0.24)	1 (0.05)	22 (0.34)
Result	Open-Ended	15 (0.63)	19 (0.49)	8 (0.29)	25 (0.40)
	Prompted	7 (0.32)	13 (0.25)	2 (0.10)	19 (0.30)
Total	Open-Ended	24	39	28	63
	Prompted	22	51	21	64

and incorrect causes was more difficult (so we did not), and where result responses were much more frequent (because catching a virus has more salient consequences than floating a balloon).

Fig. 1A and B shows the proportion of common-cause analogies across the items, conditions, and participant groups. A clear pattern emerges. In the open-ended AGT, the only participants to produce over 50% common-cause analogies are the intermediate geoscientists (51%), and the expert geoscientists (83%) for the balloon item only, and no one else was even over 20%. No one produced over 20% common-cause analogies of the cold item in the open-ended condition. On the other hand, in the prompted condition, only the novices recruited from Mechanical Turk failed to produce over 50% common-cause analogies for both items. This figure clearly demonstrates two findings. The first finding is relatively obvious that

Table 6
Fixed effects parameter estimates and odds ratios (exp(*B*)) for Models 1–3

Model	Effect	Estimate	exp(B)	*p*
Model 1Expert versus Intermediate Geoscientists	Expert Geoscientist − Intermediate Geoscientist	0.65	1.92	.025
	Prompted − Open-Ended	1.39	4.01	<.001
	Balloon − Cold	2.17	8.75	<.001
	(Expert Geoscientist − Intermediate Geoscientist) ✻ (Balloon − Cold)	1.345	3.86	.020
	(Prompted − Open-Ended) ✻ (Balloon − Cold)	−1.54	0.21	.008
	Expert Geoscientist − Vision Scientist	4.11	61.09	.027
Model 2Expert Geoscientist versus Vision Scientist	Prompted − Open-Ended	14.84	2.80e+6	<.001
	Balloon − Cold	4.76	116.66	<.001
	(Expert Geoscientist − Vision Scientist) ✻ (Prompted − Open-Ended)	−12.68	3.11e−6	<.001
	(Expert Geoscientist − Vision Scientist) ✻ (Balloon − Cold)	11.15	6.95e+4	<.001
	(Prompted − Open-Ended) ✻ (Balloon − Cold)	−5.07	0.006	.041
	(Expert Geoscientist − Vision Scientist) ✻ (Prompted − Open-Ended) ✻ (Balloon − Cold)	−22.29	2.09e−10	<.001
Model 3Vision Scientist versus Novice	Prompted − Open-Ended	3.84	46.33	<.001
	Vision Scientist − Novice	2.15	8.57	<.001
	(Prompted − Open-Ended) ✻ (Vision Scientist − Novice)	4.11	60.87	<.001

scientists have more scientific explanatory knowledge than novice nonscientists, even about familiar events from outside their domain of expertise (as shown by the prompted AGT). The second finding is key to our account—that scientists may only *spontaneously* apply their explanatory knowledge specific to their domain of expertise.

To analyze the pattern further, we conducted a series of logistic regression analyses. We included Science Experience, AGT Condition (Prompted, Open-Ended) and Item (Balloon, Cold) as fixed factors, and where appropriate, subjects as a cluster variable with a random intercept (making a mixed-model analysis). We present three models because each contrasts a distinct pair of science experience levels to answer our research questions. All analyses were conducted with the *GAMLj* linear model module (Gallucci, 2019) for Jamovi (2020) statistical analysis software. The most critical results for all three models are shown in Table 6—the parameter estimates, exp(*B*) and the *p* value for each fixed factor. Exp(*B*), which refers to the exponentiation of the *B* coefficient, is an odds ratio. This is the key effect size measure, showing the ratio between the odds of generating a correct causal analogy in one level of the factor (e.g., the prompted condition) compared to the other (the open-ended condition). The Online Appendix presents the parameter estimates and odds ratios for the simple effects for

the statistically significant interactions for the three models. In addition, the Online Appendix presents global model fit metrics.

The first analysis contrasts expert and intermediate geoscientists. Here, we examine how analogy generation changes as expertise grows within a domain. Because all geoscientists only answered a single item, a logistic GLM was conducted (no random cluster variables to necessitate a mixed-effects model as below). Model 1 includes fixed effects of science experience, AGT condition, and Item, with interactions between Science Experience * Item, and AGT condition * Item. Science Experience *AGT interactions was not included in the model because it did not improve model fit.

There were significant main effects of: (a) Expertise—the odds that an expert generated a causal analogy were 1.92 times higher than the intermediates; (b) Condition—the odds that the prompted condition elicited a causal analogy were 4.01 times the odds for the open-ended condition; and (c) Item—the odds that the balloon item elicited a causal analogy was 8.75 times higher than the cold item. Importantly, there were also two interactions. (a) Expertise * Item—the odds that an expert generated a causal analogy were 3.71 times higher than an intermediate for the balloon item, but the two groups were not significantly different for the cold item ($exp(B) = 0.98$). (b) Condition * Item—the odds that the prompted condition elicited a causal analogy were 4.64 times the odds the open-ended condition did for the cold item, but the two conditions were not significantly different for the balloon item. The overall pattern shown by Model 1 is that there is minimal effect of prompting for scientists within their domain regardless of experience, but experts do have more useable causal knowledge.

The next analysis contrasts expert geoscientists and vision scientists. These two groups are matched for their science experience but differ in their domain of expertise. Here, the best fitting model, Model 2, contains all three fixed factors, and all two-way and three-way interactions (and the cluster variable of subject with a random intercept.) Every main effect and two-way interaction is statistically significant; however, the three-way interaction here is critical for understanding the pattern. The odds an expert geoscientist generated a causal analogy was 2.40e+9 times the odds a vision scientist did for the balloon item in the open-ended condition, but these two groups showed no significant difference with the cold item (either condition), or the balloon item in the prompted condition. Likewise, the prompted condition had greater odds to elicit a causal analogy than the open-ended condition for both groups for the cold Item (expert geoscientists, $exp(B) = 1.64e+7$; vision scientists, $exp(B) = 7.60e+7$), but for the balloon item, the prompted condition only elicited greater odds than the open-ended condition for the vision scientists ($exp(B) = 3.31e+10$); and not the expert geoscientists. Last, there was a single significant Item contrast—The balloon item elicited more causal analogies than the cold Item but only for the expert geoscientists in the open-ended condition ($exp(B) = 1.02e+8$). Overall, Model 2 shows that the key point of how scientists may certainly have encoded causal explanations outside of their domain, but there are unlikely to spontaneously use it. In contrast, they will spontaneously use the causal knowledge from their own domain.

Our final Experience group contrast is between the two nongeoscientist groups, the vision scientists and novices. Here, the model is not improved by including Item as a fixed

factor, nor any interactions between Item and the other fixed factors. Both items are outside any expertise this group has, and so, the item contrast makes no statistical contribution. Model 3 has the other two fixed factors, Science Experience and AGT Condition, and their interaction (along with participant as a cluster variable with a random intercept). Overall, the odds that a vision scientist generated a causal analogy was 8.57 times the odds a novice did, and the odds the prompted condition elicited a causal analogy was 46.33 times the odds that the open-ended condition did. Critically, the two factors interacted. The odds that the prompted condition elicited a causal analogy was 5.94 times higher than the open-ended condition for the novices, and 361.44 times higher for the vision scientists. In addition, the odds the vision scientists generated a causal analogy was 66.88 times higher than the novices for the prompted condition, but the two groups were not significantly different in the open-ended condition. This pattern shows that neither group's causal knowledge was ready to apply spontaneously, and although prompting could elicit some knowledge for both groups, novices had less causal knowledge to bring to bear even when prompted.

4. General discussion

There are three main findings. First, for the item explained by a geoscientific principle, we found a gradient of likelihood of spontaneously producing causal analogies, with expert scientists highest followed by intermediate geoscientists, and then by novice nonscientists.[9] Second, not surprisingly, the prompted AGT elicited more principle-based analogies than the open-ended AGT. However, the third finding is that prompting interacts with expertise. The prompting effect was minimal for expert and intermediate geoscientists when analogizing to the balloon example. This fits with the assumption that their level of spontaneous retrieval of the density principle is sufficiently high as to largely negate the advantage of prompting. This high level of spontaneous principle-based analogies applies only for principles in the scientists' domain. Both expert geoscientists and expert vision scientists showed significant increases in principle-based analogies in the prompted condition for the "catching a cold" item; but only vision scientists also showed a similar increase for the balloon item.

It is worth noting that there were two kinds of interactions between science experience and prompting. Novices showed effects of prompting with both items, showing that they do have some scientific causal knowledge in both domains, even if they do not use it spontaneously. In comparison to novices, scientists show a smaller effect of prompting within their domain (reflecting greater spontaneous use of their causal knowledge), but a larger effect of prompting outside of their domain (reflecting more total causal knowledge; see Model 5).

In sum, these results support the claim that expertise is a major predictor—perhaps *the* major predictor—of *spontaneous* deep relational remindings. Science experts amass knowledge, but beyond these knowledge gains, scientific expertise entails a focus on key relational principles in their domain. Because these relational patterns are used repeatedly in interpreting domain phenomena, experts will tend to have uniform relational representations across the domain, promoting relational retrieval (Forbus & Gentner, 1986; Gentner et al., 2009).

4.1. The role of goals in relational retrieval

Dunbar and Blanchette (2001) have argued that meaningful goals are important in achieving relational retrieval, and that the low levels of relational retrieval found in laboratory studies result in part from the absence of such goals. Our findings bear out their contention that goals (here manipulated by task instructions) can have sizable effects on rates of analogy production. Novices produced many more common-cause analogies when told to seek them.[10] However, there is clearly also a large role for relational knowledge. The expert geoscientists showed no effect of prompting on the geoscience examples, but did show prompting effects outside of their domain as did the expert vision scientists.

These findings bear on the issue of spontaneous discovery. The distant analogies generated by Kepler to explain planetary motion, or by de Mestral, who invented Velcro after pulling burrs out of his dog's fur, are possible because of a prepared mind. We view our findings as complementary to the findings of Ball, Ormerod, and Morley (2004) who showed how expert analogies during the design process are rooted in abstract principles, while novice analogies are based in cases. Future work should examine interactions between task goals and knowledge representation in spontaneous use of analogies (and see Chan & Schunn, 2015; Chan et al., 2011; and Fu et al., 2013).

4.2. Concerns and limitations

That the expert geoscientists scored equally high with and without prompting on the balloon example is consistent with our claim that their default representations incorporate the relational principles of the domain. However, it could be argued that there was a ceiling effect: given a spontaneous common-cause analogical generation rate of 0.83, it might be hard to detect a prompting effect; though certainly not logically impossible. Regardless, if experts *were* at ceiling without prompting, which is actually quite consistent with our primary claim of how readily these principles come to mind with expertise.

A second issue to consider is whether other factors, such as age or intelligence, could have led to the retrieval patterns, rather than domain expertise. The marked domain specificity of scientists' relational retrieval patterns suggests, to the contrary, that domain knowledge is the determining factor. However, one might wonder whether the difference between scientists and nonscientists was due to the fact that the scientists took the AGT at a geoscience conference—a context in which geoscience knowledge is likely to be generally activated. However, even if there was a priming effect from the geoscience context, the difference between the expert geoscientists and the intermediate geoscientists (both tested at the same conference) still attests to an important role for expertise.

A further issue is the cognitive explanation for our pattern. We have suggested that expertise tends to confer uniform relational representation, and that this is the underlying reason that expertise leads to superior relational retrieval. However, there are other ways in which frequently used relational principles could be "ready to access." For example, the principle could be connected to more things in memory, it could have stronger connections to the same number of things, these connections could be more differentiated, or the level of detail encoded with the causal principle could vary.

4.3. Implications for learning and development

These findings are consistent with research on novice–expert differences in learning and cognitive development. Gaining expertise in a domain is often marked by a *relational shift* from a focus on objects and concrete contextual details to a focus on higher order relational patterns that connect objects and events (Gentner & Rattermann, 1991; Gentner & Toupin, 1986). This pattern is seen both in cognitive development (Gentner, 1988; Kotovsky & Gentner, 1996; Richland, Morrison, & Holyoak, 2006) and in science learning (Ball et al., 2004; Chi et al., 1981; Rottman et al., 2012; Stains & Talanquer, 2008). Of course, concrete aspects of phenomena continue to be important in later development; but they are encoded and stored within larger relational structures (see, e.g., Clement, 1988, 1993; Dunbar & Blanchette, 2001; Nersessian & Chandrasekharan, 2009 for examinations of analogical reasoning in working scientists).

Our findings bear out this kind of expertise-related shift toward encoding relational structure—in this case, causal patterns. Experts retrieved phenomena that shared causal structure with the probe phenomenon not only in directed retrieval but also in free reminding. This is noteworthy because failure of relational retrieval is the bête noir of analogical learning and problem solving. It is also a chief contributor to the *inert knowledge* problem in education (Whitehead, 1929, as cited in John & Bransford, 1989): the distressingly frequent phenomenon whereby students fail to retrieve previously learned material when in a new context.

Our findings suggest that retrieval via domain-related causal structure is not only possible but also natural with sufficient expertise—that is, for experts, domain knowledge is habitually encoded (and therefore readily retrieved) in terms of the key relational principles of the domain. Scientific principles are explanations for disparate events, and we suggest that experts encode these events in terms of their explanation. In structure-mapping terms, experts have achieved a degree of *uniform relational representation* (Forbus et al., 1995; Gentner et al., 2009; Goldwater & Jamrozik, 2019; Jamrozik & Gentner 2020). Through habitually processing causal patterns, and repeatedly comparing cases in terms of their causal structure, experts gradually form long-term representations in which key causal patterns are prominently represented at a sufficiently general level as to facilitate retrieval via common causal explanatory structure. Uniformity and abstractness foster a frequency of use because of the expanding number of events to which these causal principles are recognized as relevant. In case-based reasoning terms, experts have acquired causal indexing—they readily access memory on the basis of common causal patterns (Kolodner, 1994; Schank, Kass, & Reisbeck, 1994; Seifert, 1989).

We suggest that this habitual causal encoding may help explain why the even highly familiar movies from Trench and Minervino (2015) may not have been readily retrieved from relational cues alone. On the one hand, scientific principles are used by scientists to explain many phenomena, but even if you have watched Jurassic Park five times, you may not start to see the world in terms of the plot structure of Jurassic Park (at least without active imaginative elaboration). However, further research is needed to directly test the cognitive processes beyond frequency of encoding that support abstraction and spontaneous relational retrieval.

4.4. Implications for education

Our findings are consistent with recent treatises that have emphasized the importance of guiding students to focus on relational patterns, including key explanatory principles, in science instruction (Chinn & Malhotra, 2002; Forbus et al., 2005; Sandoval, 2003; Sandoval & Reiser, 2004; Tabak, Smith, Sandoval, & Reiser, 1996)). Indeed, there is general agreement that science education needs improvement (e.g., Alberts, 2011; Olson & Loucks-Horsely, 2000). Students need to go beyond mere learning of facts and procedures without conceptual understanding (e.g., Abrahams & Millar, 2008). A focus on causal relations is reaching prominence in U.S. national policy. The Framework for K-12 Science Education (2011; and its more recent updates) emphasizes the importance of fostering understanding of causal principles as a "style of scientific reasoning" (Osborne, Rafanelli, & Kind, 2018, p. 962). The results presented are consistent with the idea that an emphasis on discovering causal principles will further the goal of educating for excellence in science.

However, to implement the new standards effectively in K-12 schooling—and to best foster causal understanding among undergraduates—adequate assessment tools must be developed (Pellegrino, Wilson, Koenig, & Beatty, 2014). We need both formative assessments and fully validated research quality assessments. We suggest that the AGT is potential candidate as a basis, and already is quite in line with a well-known formative assessment "approximate analogies" (Angelo & Cross, 1988). Its advantages include ease of completion, applicability to a wide range of knowledge levels, and potential for adaptation to different areas of science. Its ease of completion makes it possible to use without taking much class time. Our results suggest that either directly using the AGT or for instructors already using the "approximate analogies" assessment, that particular focus on causal analogies may be a useful way to track developing scientific competence.

We note that in a pilot study of the use of the AGT as a formative assessment, the prompted AGT was given to an undergraduate introduction to geoscience class for nonmajors. The data are in the Online Appendix. They produced more causal analogies than the novices from Mechanical Turk, but fewer than the intermediate geoscientists. This was an easy and fast online exercise for a large cohort of students.

Our results suggest that the AGT—perhaps, in concert with existing measures of knowledge organization such as concept maps—may be better suited for assessing what is prominent in student's knowledge than are lists of multiple-choice questions (also see Scouler, 1998). Instructors could use the AGT with different kinds of prompts to discover whether students are focusing on the important principles that the instruction seeks to communicate. Of course, we are not suggesting that the AGT is ready to be incorporated into formal evaluation standards. There would need to be empirical evidence that the AGT is both valid and reliable. Beyond that, research would have to examine whether it can be adapted to the high-school level (or below). However, we do suggest that the AGT could be useful to undergraduate instructors as an aid to formative evaluation of their courses and as a way of tracking student progress.

Moving forward, we speculate that the AGT could also be useful to foster students' causal understanding. It could be effective for students to consider the responses of other students

and of domain experts after they generate their own analogies. Such experience could serve as a form of feedback. Further, considering causal analogies offered by experts or other students should engage comparison processes that will support the abstraction of causal principles (see, e.g., Alfieri, Nokes-Malach, & Christianson, 2013; Gentner et al., 2009; Gick & Holyoak, 1983). This idea fits with evidence that a focus on domain-relevant relational patterns in science can be promoted by the use of well-chosen analogies (Clement, 1993; Gentner, 2010; Jee et al., 2010; Kastens & Rivet, 2008; Richland, Zur, & Holyoak, 2007; Sibley, 2009).

4.5. Conclusion

Our findings suggest that expert scientists fluently traverse their knowledge bases via paths of shared causal relational patterns. Past research suggests that expert scientists can retrieve domain phenomena based on causal similarities when the task calls for it. Equally importantly, experts are highly likely to retrieve causal analogs even in an open-ended task—consistent with the idea that key domain-relevant relational principles are habitually encoded. Future work should continue to explore how this focus arises and how it can be fostered.

Notes

1 Representing different events or cases with the same relational representations is readily characterized by symbolic models of cognition, such as in these cited papers, but the same idea could potentially be accounted for in other cognitive architectures, such as the symbolic-connectionist hybrid models of relational thinking (e.g., Doumas, Hummel, & Sandhofer, 2008) with a sufficient degree of overlap in distributed representations.

2 It is important to note that when overall-similarity matches are included in the materials, they are often more likely to be retrieved than superficial-only matches, for example, Gentner et al. (1993), indicating that common deep relational structure does often contribute to retrieval. Further, in everyday life, things that share superficial features often also share relational structure. Gentner (1989) has called this the *kind world* phenomenon.

3 In real-world discovery, there is no one with the correct answer to give you a hint or a prompt. Further, given the right cue, deep relational remindings occur much more frequently (Gick & Holyoak, 1980). Our working hypothesis is that spontaneous retrieval is the true marker of expertise.

4 We acknowledge that differences in aptitude may influence whether people persevere in a field; this is one motivation for comparing experts from different fields.

5 As discussed by Trench and Minervino (2015, p. 1299), in tasks where the source is participants' own knowledge, it is in general impossible to distinguish *retrieval* of a prior analogous example from *generation* of a new analogous example. Further, many cases are probably intermediate, as when an initial partially overlapping retrieval is tweaked to make it a better relational match. Hereafter, we will refer to this task as an *analogy generation task*, with the understanding that this task includes retrieval.

6 Uniform relational representation may be the result of intentional encoding strategies or reflect a process that begins as a case-by-case encoding of causality, which might give rise to a more explicit causal encoding strategy over time. However, either an intentional or implicit encoding of causality has potential to produce spontaneous principle-based remindings.
7 In Tulving and Pearlstone's (1966) terms, the information is available but not readily accessible for spontaneous retrieval.
8 Both the geoscientists and vision scientists had additional AGT items that concerned phenomena specific to their respective field of study that no other group received. Those items are part of another research project. Here, we present data from "everyday phenomena," about balloons and the common cold, because nonscientists have knowledge of these examples and therefore make meaningful contrasts. We do not learn much from asking nonscientists about phenomena they never heard of, and we do not explain to them.
9 In unreported analyses, every pairwise contrast showed a significant effect.
10 We note that our data do not reveal whether people given the causal goal actually experienced more relational remindings, or whether the goal acted to filter out nongoal-relevant remindings (see discussion in Trench and Minervino, 2015).

Acknowledgments

Duncan Sibley is thanked for the initial conception and design of this project. In addition, he developed the initial scoring rubrics and was the primary scorer. We thank colleagues who assisted with the project across our three institutions. This study was supported by NSF-DUE grants to JL (grant no. 0941492) and to DG, (grant no. 0942099), and ONR Grant (N00014-92-J-1098) to DG.

References

Abrahams, I., & Millar R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education*, *30*, 1946–1969.

Alberts, B. (2011). Getting education right. *Science*, *333*, 919.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113.

Angelo, T. A., & Cross, K. P. (1988). *Classroom assessment techniques. A handbook for faculty*. Washington, DC: Office of Educational Research and Improvement.

Ball, L. J., Ormerod, T. C., & Morley, N. J. (2004). Spontaneous analogising in engineering design: A comparative analysis of experts and novices. *Design Studies*, *25*(5), 495–508.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition*, *28*(1), 108–124.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition*, *29*(5), 730–735.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist*, *28*(3), 235–251.

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

Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., & Kotovsky, K. (2011). On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design*, *133*(8), 081004.

Chan, J., & Schunn, C. (2015). The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science*, *39*(1), 126–155.

Chi, M. T. H. (2006). Methods to assess the representations of experts' and novices' Knowledge. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 167–184). Cambridge: Cambridge University Press.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*, 175–218.

Christensen, B. T., & Schunn, C. D. (2007). The relationship of analogical distance to analogical function and preinventive structure: The case of engineering design. *Memory & Cognition*, *35*(1), 29–38.

Clement, J. (1988). Observed methods for generating analogies in scientific problem solving. *Cognitive Science*, *12*(4), 563–586.

Clement, J. (1993). Using analogies to deal with students' preconceptions in physics. *Journal of Research in Science Teaching*, *30*(10), 1241–1257.

Committee on a Conceptual Framework for New K-12 Science Education Standards. (2011). National Research Council, 270 pp.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43.

Dunbar, K. (1999). The scientist in vivo: How scientists think and reason in the laboratory. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery*. New York, NY: Plenum Press.

Dunbar, K., & Blanchette, I. (2001). The in vivo/in vitro approach to cognition: The case of analogy. *Trends in Cognitive Sciences*, *5*, 334–339.

Forbus, K. D., Carney, K., Sherin, B. L., & Ureel, L. C. (2005). Vmodel - A visual qualitative modeling environment for middle-school students. *Ai Magazine*, *26*(3), 63–72.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Toward a theoretical framework. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. *2*, pp. 311–348). Los Altos, CA: Kaufmann.

Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., & Wood, K. (2013). The meaning of "near" and "far": The impact of structuring design databases and the effect of distance of analogy on design output. *Journal of Mechanical Design*, *135*(2), 021007.

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. Retrieved from https://gamlj.github.io/.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59*, 47–59.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). London: Cambridge University Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*, 752–775.

Gentner, D., Brem, S., Ferguson, R. W., Wolff, P., Markman, A. B., & Forbus, K. D. (1997). Analogy and creativity in the works of Johannes Kepler. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 403–459). Washington, DC: American Psychological Association.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*, 1343–1382.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London: 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*(3), 306–355.

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

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, *39*, 93–104.

Goldwater, M. B., & Jamrozik, A. (2019). Can a relational mindset boost analogical retrieval? *Cognitive Research: Principles and Implications*, *4*(1), 1–16.

Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. New York, NY: Basic Books.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

The Jamovi Project. (2020). *jamovi*. (Version 1.2) [Computer Software]. Retrieved from https://www.jamovi.org.

Jamrozik, A., & Gentner, D. (2020). Relational labeling unlocks inert knowledge. *Cognition*, *196*, 104–146.

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*, 2–13.

John, I., & Bransford, J. J. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470–496). Cambridge: Cambridge University Press.

Kalafatis, S. E., & Libarkin, J. C. (2019). What perceptions do scientists have about their potential role in connecting science with policy? *Geosphere*, *15*(3), 702–715. https://doi.org/10.1130/GES02018.1

Kastens, K., & Rivet, A. (2008). Multiple modes of inquiry in the earth sciences. *The Science Teacher*, *75*(1), 26–31.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797–2822.

Kolodner, J. L. (1994). *Case-based reasoning*. San Mateo, CA: Kaufmann.

Lakatos, I. (1970). Falsifiability and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge: Cambridge University Press.

Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education*, *2*(2), 119–127.

Mackie, J. L. (1974). *The cement of the universe; a study of causation*. Oxford, England: Clarendon Press.

Nersessian, N. J. (1984). Aether/or: The creation of scientific concepts. *Studies in History and Philosophy of Science Part A*, *15*(3), 175–212.

Nersessian, N. J., & Chandrasekharan, S. (2009). Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research Journal, Special Issue: Integrating Cognitive Abilities*, *10*, 178–188

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 510–520.

Olguín, V., Trench, M., & Minervino, R. (2017). Attending to individual recipients' knowledge when generating persuasive analogies. *Journal of Cognitive Psychology*, *29*, 755–768. https://doi.org/10.1080/20445911.2017.1304942.

Olson, S., & Loucks-Horsley, S. (Eds.). (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Research Council, 224 pp.

Osborne, J., Rafanelli, S., & Kind, P. (2018). Toward a more coherent model for science education than the cross-cutting concepts of the next generation science standards: The affordances of styles of reasoning. *Journal of Research in Science Teaching*, *55*(7), 962–981

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*. Washington, DC:National Academies Press.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*, 249–273

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

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory & Language*, *37*(2), 240–267.

Ross, B. H. (1999). Post-classification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(3), 743–757.

Rottman, B., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*, 919–932.

Sandoval, W. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of Learning Sciences*, *12*, 5–51.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic supports for science inquiry. *Science Education*, *88*, 345–372

Schank, R. C., Kass, A., & Riesbeck, C. K. (1994). The explanation process: Explanation questions and explanation patterns. In R. C. Schank, A. Kass, & C. K. Riesbeck (Eds.), *Inside case-based explanation* (pp. 27–69). Hillsdale, NJ: Erlbaum.

Scouler, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examinations vs. essay assignment. *Higher Education*, *35*, 453–472.

Seifert, C. M. (1989). Analogy and case-based reasoning. In K. J. Hammond (Ed.), *Proceedings: Second workshop on case-based reasoning* (DARPA; pp. 125–129). San Mateo, CA: Morgan Kaufmann.

Sibley, D. F. (2009). A cognitive framework for reasoning with scientific models. *Journal of Geoscience Education*, *57*(4), 255–263.

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. *Journal of Research in Science Training*, *45*, 771–793.

Tabak, I., Smith, B., Sandoval, W., & Reiser, B. (1996). Combining general and domain-specific strategic support for biological inquiry. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.), *Intelligent tutoring systems* (pp. 288–296). Berlin/Heidelberg: Springer.

Thagard, P. (1992). Analogy, explanation, and education. *Journal of Research in Science Teaching*, *29*(6), 537–544.

Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.

Trench, M., & Minervino, R. A. (2015). The role of surface similarity in analogical retrieval: Bridging the gap between the naturalistic and the experimental traditions. *Cognitive Science*, *39*(6), 1292–1319.

Trench, M., & Minervino, R. A. (2020). *Distant connections: The memory basis of creative analogy*. SpringerBriefs in Psychology. Cham: Springer.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*, 381–391.

Whitehead, A. N. (1929). *The aims of education and other essays*. New York: The Free Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information