Received: 6 May 2022 | Revised: 27 June 2023 | Accepted: 13 July 2023

DOI: 10.1002/tea.21892

RESEARCH ARTICLE

JRST | WILEY

A hypothetico-deductive theory of science and learning

Steven T. Kalinowski[1] | **Avital Pelakh**[2]

[1]Department of Ecology, Montana State University, Bozeman, Montana, USA

[2]Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence
Steven T. Kalinowski, Department of Ecology, Montana State University, Bozeman, MT 59717-2000, USA.
Email: skalinowski@montana.edu

Funding information
United States National Science Foundation, Grant/Award Number: 1432577

Abstract

This article presents a simple, cognitive theory of science and learning. The first section of the paper develops the theory's two main propositions: (i) A wide range of scientific activities rely heavily on one type of reasoning, hypothetical thinking, and (ii) This type of reasoning is also useful to students for learning science content. The second section of the paper presents a taxonomy of multiple-choice questions that use hypothetical thinking and the third section of the paper tests the theory using data from a college biology course. As expected by the theory, student responses to 24 scientific reasoning questions were consistent with a one-dimensional psychometric construct. Student responses to the scientific reasoning questions explained 36% of the variance in exam grades. Several directions for additional research are identified, including studying the psychometric structure of scientific thinking in more detail, performing randomized, controlled experiments to demonstrate a causal relationship between scientific thinking and learning, and identifying the relative contribution of other factors to success in college.

KEYWORDS

college, hypothetico-deductive, learning, reasoning, science

1 | INTRODUCTION

In 1910, John Dewey lamented "science has been taught too much as an accumulation of ready-made material... [and] not enough as a method of thinking" (Dewey, 1910). In a statement readers may find eerily prescient, Dewey warned "the future of civilization depends on the widening spread and deepening hold of the scientific habit of mind." A century later, the US National Research Council (e.g., NRC, 2012) and newspaper editorials are expressing similar sentiments. Despite these pleas for change, college faculty emphasize teaching science content over method—as if they assume "students will 'magically' obtain science process skills somewhere during their four years of study" (Coil et al., 2010).

Calling for faculty to teach scientific methods is easy, but specifying what this means is much harder. Historians and philosophers have devoted centuries of effort to elucidating how science works, and their analyses have painted a complex and nuanced picture of science (e.g., Faye, 2016; Okasha, 2002). Not surprisingly, educators have contrasting opinions regarding how scientific methodology should be taught. At one extreme, science textbooks often present a simple version of "the scientific method" that emphasizes hypothesis testing (e.g., Blachowicz, 2009; Johnson et al., 2017; Reece et al., 2020; Withgott & Laposata, 2018). At the other extreme, education researchers often emphasize science uses diverse methods to understand the world. For example, the National Research Council (2012, p. 44) identified eight important practices for doing science: asking questions, developing models, planning and carrying out investigations, analyzing and interpreting data, using mathematics, constructing explanations, engaging in argument from evidence, and obtaining/evaluating/communicating information. The National Research Council is the preeminent scientific institution in the United States, so we will consider this the consensus position among education researchers. However, other authors have taken different stances. For example, Klahr and coauthors (Klahr & Dunbar, 1988; Zimmerman & Klahr, 2018) have emphasized the importance of a few core practices for doing science.

The disagreement regarding how scientific methodology should be taught may be fueled by at least three potential causes. First, this disagreement may reflect the inherent nature of science: research may be so complicated that disagreement regarding how to teach its methods may be almost inevitable. However, we would like to consider two other possibilities. A second potential contributor to this disagreement may be how science has been studied (Klahr & Simon, 2001). The contemporary emphasis on the diversity of scientific *practice* emerged from historical and sociological analyses of scientific research (e.g., National Research Council, 2012, p. 43). In this article, we will propose that an alternative, and considerably simpler, view of science emerges if we examine the cognitive skills used to perform these practices. Third and last, some of the disagreement regarding which scientific skills should be taught may persist because empirical criteria have not been used to settle this debate. In this article, we will show how readily available data from classrooms can be used to select science thinking skills to teach.

The goal of this article is to develop a simple theory of scientific reasoning that will help instructors teach science (both method and content). Section 1 of this article describes a theory of how science answers research questions and how students learn science content. This theory emphasizes the role of hypothetical thinking in both types of learning. Section 2 presents a brief taxonomy of scientific reasoning tasks and multiple-choice questions that require students to use hypothetical thinking in different ways. Section 3 presents data from an introductory biology course that tests the theory in four ways. The last section of the paper, the Discussion, compares this investigation to previous research and proposes directions for additional work.

2 | SECTION 1: A THEORY OF SCIENCE AND LEARNING

This section of the paper will present a theory of science and learning that has implications for science instruction. Both of these topics, how science works and how students learn, have been studied by many research disciplines, including the philosophy of science, cognitive psychology, and education science. These disciplines have different goals, different theoretical perspectives, different methods, and different vocabulary, but have much to contribute to each other. We will weave together concepts from all three disciplines to develop a simple, unified theory of science and learning.

2.1 | How to identify scientific thinking skills to teach

An instructor intending to teach college students scientific thinking or reasoning[i] has many choices regarding what to teach: scientific questioning, experimental design, induction, deduction, abduction, eduction, propositional logic, critical thinking, analytical thinking, hypothetico-deductive (HD) reasoning, control-of-variables, causal reasoning, correlations, mathematical modeling, computer simulation, probability, statistics, graph interpretation, and so forth. Every science discipline uses different methods, so the list is long indeed.

Deciding which of these skills to teach is daunting. One strategy for resolving dilemmas like this is to ask "What will students be able to do if they learn some concept or skill?" (Wiggins et al., 2005). The answer to this question can help identify which concepts or skills are most useful for students. Therefore, let us ask this simple question of science: *What should students be able to do if they possess basic scientific thinking skills?* This is an important question, and we believe the answer is more expansive than is generally recognized (National Academies of Sciences, Engineering, and Medicine, 2016).

Science is a method for learning about the natural world (Okasha, 2002, p. 1), so possessing basic scientific thinking skills should help students learn about the natural world. This may seem like an uninteresting truism, but if we interpret the statement broadly, it has important implications for educators. Let us start with a narrow interpretation and then broaden it. To begin, scientific thinking skills are useful for designing, performing, and interpreting the results of experiments (National Research Council, 2012). Students, for example, might perform a controlled experiment to learn how fertilizer affects plant growth. Such investigations are time-consuming, so hands-on experiments can teach students only a small fraction of the content on a typical college syllabus. However, if students understand how to interpret scientific studies, they should also be able to learn from research conducted by others. For example, generations of faculty have tried to teach genetics by discussing the experiments of Mendel (e.g., Reece et al., 2020). We doubt any reader will question the value of scientific reasoning for such learning. However, this may be just the tip of the iceberg for how, why, and when scientific thinking is useful for learning science concepts. Accumulating empirical research has shown that students of all ages with strong scientific reasoning skills learn more chemistry (Cracolice & Busby, 2015), physics (Ates & Cataloglu, 2007; Ding, 2014; Moore & Rubbo, 2012), and biology (Cannady et al., 2019; Cavallo, 1996; Kalinowski & Willoughby, 2019; Lawson & Thompson, 1988) than classmates with less developed scientific reasoning skill. The explanation for this correlation has received little discussion (but see Lawson, 1992; Lawson & Thompson, 1988) but this body of research suggests some type of reasoning used to do science research might be useful to students for learning a wide variety of science concepts.

If this is true, it provides a strategy for resolving some of the debate regarding what scientific thinking skills should be taught. Scientific thinking skills that help students learn science content deserve some priority. We are not claiming this should be the only criterion for deciding how to teach scientific methodology, but we are arguing that knowing what students can do with different thinking skills is useful for designing curricula.

2.2 | A role for HD reasoning in learning

Now that we have proposed that students use scientific reasoning to learn a wide variety of science content, we will discuss a question that arises naturally: "What specific types of reasoning are most useful for learning science content?" To be clear, we are searching for a type of reasoning students might use to learn all sorts of science concepts during typical learning practices—for example, while listening to lectures, reading textbooks, studying notes, or discussing science concepts with classmates. This eliminates specialized methods for testing ideas such as statistical conjecturing (Dvir & Ben-Zvi, 2018). We will begin our search for such a skill by reviewing constructivism, the most important theory of learning in education.

Constructivism (Fosnot, 2013; Mintzes, 2020; Tobin, 1993) asserts students do not passively absorb knowledge presented by instructors, but must construct their own understanding of the world. In order to do this successfully, students must continuously evaluate their understanding of new concepts and revise erroneous beliefs they have about these new concepts or about their understanding of the world. Erroneous beliefs are common because students often come to the classroom with misconceptions regarding the natural world. For example, many college students believe growing trees get most of their raw material from the soil (Schneps & Sadler, 1997), moving objects have an inherent tendency to slow down (Halloun & Hestenes, 1985), and species evolve because individuals use or do not use body parts (Gregory, 2009). In order to successfully learn photosynthesis, Newton's laws, or natural selection, these students need to recognize their ideas are wrong and restructure their understanding. This process is called conceptual change and is difficult for most students (Chi, 2009; DiSessa, 1993; NRC, 2005; Posner et al., 1982).

Mainstream constructivism has not identified the cognitive skills students need to evaluate their understanding of science concepts (Fosnot, 2013; Mintzes, 2020), but philosophers, cognitive psychologists, and education researchers have highlighted the usefulness of HD reasoning for testing ideas (Evans, 2007; Hempel, 1966; Inhelder & Piaget, 1958; Lawson, 1995; Platt, 1964; Popper, 1959; Stanovich, 2009; Whewell, 1858). HD reasoning is a method for testing hypotheses that asks what a hypothesis predicts should be observed in some sort of test if the hypothesis were true (Blackburn, 2005; Butts, 2015). The results of the test are then compared to the prediction. If the results differ from the prediction, the hypothesis is rejected. HD arguments, therefore, often have the form: IF X is correct, THEN Y should be observed. BUT, Y was not observed. THEREFORE, X is not correct.

Predictions are an important element of HD arguments, and prediction-making is arguably the most important element of HD reasoning. This is a common source of confusion. A prediction in an HD argument is not what an investigator believes is true or what the investigator believes will happen in an experiment. A prediction is a description of what *will* be observed in a specific test *if* a hypothesis is correct. Predictions are logical consequences of hypotheses, hence the term hypothetico-*deductive*. For example, the hypothesis that a meteor impact caused non-avian dinosaurs to go extinct makes several predictions. These include (1) extinction would

be rapid, (2) extinction would take place on land and in the ocean simultaneously, and (3) sedimentary rock layers formed at the time of impact will contain high levels of iridium (Alvarez, 1997).

We hypothesize students use the IF/THEN/BUT/THEREFORE logic of HD reasoning to evaluate their understanding of science concepts. An example illustrates how students might do this. Many students erroneously believe plants get most of the raw material they use to grow by absorbing organic nutrients from the soil (Schneps & Sadler, 1997). IF this idea were true, THEN plants could not be grown without soil. BUT many plants grow perfectly well using hydroponic cultivation. THEREFORE, plants must not get most of their raw material from soil. This idea that HD reasoning is valuable for learning was an important component of Piaget's theory of development (Inhelder & Piaget, 1958; Piaget, 1972) and a consistent theme in Lawson's scholarship (e.g., Lawson, 1995; Lawson, 2003; Lawson & Thompson, 1988).

The HD method for testing hypotheses has several well-known limitations. First, HD reasoning can never prove a hypothesis is correct (Popper, 1959). This is because other hypotheses, including hypotheses that have not been imagined, might make the same predictions. Second, making predictions usually requires making auxiliary assumptions that may or may not be correct. If these assumptions are incorrect, the results of a test are likely to be interpreted incorrectly (Hempel, 1966). Third, evaluating whether the results of an experiment are consistent with predictions can be difficult when sampling error or other stochastic processes are at play. Statistical analyses may be needed. These limitations are real, but may be less significant to students learning science concepts than to scientists performing original research. This is because, unlike researchers, students have the great advantage of being presented with scientifically correct ideas. If a student can use HD reasoning to detect an error in their thinking, this will help them move toward a correct understanding of science content.

2.3 | An expansive role for hypothetical thinking in science

HD reasoning has long been considered an important element of scientific investigation (Hempel, 1966; Platt, 1964; Popper, 1959; Whewell, 1858), and for better or worse, many textbooks call some variation of it "the scientific method" (e.g., Reece et al., 2020; Withgott & Laposata, 2018). A potential objection to the theory we are proposing is that science is much more complex than the textbook description of hypothesis testing. For example, the National Research Council (2012) explicitly argued the notion of "a scientific method" was a misconception. This conclusion arose from historical and sociological observations of scientific research (National Research Council, 2012, p. 43) and from philosophical analyses of scientific reasoning (e.g., Blachowicz, 2009). We have no doubt science frequently does not follow the linear HD method described in textbooks. For example, Walter Alverez did not hypothesize a meteor caused the extinction of dinosaurs, identify what this hypothesis predicted, and then conduct the research required to test this prediction. Instead, he came to this conclusion after unexpectedly discovering high levels of iridium in a thin layer of clay at the Cretaceous–Paleogene boundary (Alvarez, 1997). However, focusing on the complex activities used to do science may overlook important similarities among the cognitive processes used to perform these activities. Alverez may not have followed "the scientific method" presented in textbooks but he clearly was using hypothetical reasoning to explain the iridium at the C-P boundary and the extinction of the dinosaurs.

Consider, as an example, two science practices that are frequently classified as distinct: designing experiments and interpreting data (National Research Council, 2012; Zimmerman & Klahr, 2018). These two practices have easily recognizable differences. For example, designing experiments often requires ensuring environmental effects are held constant, and interpreting data often involves applying statistical concepts that are not needed to design or perform an experiment. However, designing an experiment is well-nigh impossible without some awareness of how the results will be interpreted. This suggests these two practices both use a common skill. We propose this is the case for many scientific research practices, and that the widely used cognitive skill used throughout science is hypothetical thinking.

Hypothetical thinking is the ability to entertain an idea without accepting it as true (Amsel, 2011; Ball, 2020; Evans, 2007). Hypothetical thinking involves three related components: imagining possible worlds, exploring these imaginary worlds through mental simulation, and inferring the real-word consequences of these explorations (Amsel, 2011). Young children are capable of simple forms of hypothetical thinking—for example, pretending a banana is a telephone—but more powerful hypothetical thinking skills develop during adolescence (Inhelder & Piaget, 1958). Hypothetical thinking is useful for a wide variety of judgments. For example, a firefighter faced with a difficult rescue might use hypothetical thinking to evaluate whether a standard method for lowering a person to safety is likely to work (Gary, 1998). An office worker preparing to ask for a raise might use hypothetical thinking to evaluate different ways to make the request. And, as we will discuss below, scientists use hypothetical thinking to test ideas about the natural world. For such reasons, Stanovich (2011, p. 47) called hypothetical thinking "the foundation of rationality."

The role of hypothetical thinking in the HD method for testing hypotheses is obvious—it is, after all, in the name *hypothetico*-deductive. However, the extent to which hypothetical thinking supports other scientific practices may be less evident. We will make the case that hypothetical thinking is important for other important scientific practices in two ways. First, we will discuss the role of hypothetical thinking in a very simple description of science, and then we will discuss its role in three specific practices we believe are central for doing science.

Figure 1 provides a simple model of science. In this view, science is a search for the unseen processes or principles that give rise to observable features in the natural world. This model of science contains two components: observations and their explanations. For example, Darwin (1859) observed that several species of birds in the Galapagos Islands had distinctive beaks. He explained these beaks by proposing a flock of finches migrated to the Galapagos Islands long ago and gradually evolved into the species we see today. As in this evolutionary example, explanations are often difficult or impossible to directly observe (Table 1). Because explanations cannot be directly observed, they must be imagined and explored in the mind and evaluated by determining whether they are likely to give rise to what is observed. This is an application of hypothetical thinking. For this reason, we argue hypothetical thinking is fundamental for performing science and that science is largely an application of hypothetical thinking. Kuhn (2011) described science as "theory-evidence coordination" and this is a reasonable description of both science and hypothetical thinking.

We will now make the case that hypothetical thinking is useful for performing three specific practices that we argue are central to scientific investigation. These three practices are (1) constructing explanations, (2) designing experiments, and (3) interpreting data. These practices all rely heavily on hypothetical thinking but differ according to how or when hypothetical thinking is initiated. In its most classic form, HD reasoning (Butts, 2015) is initiated from a hypothesis to test. Hypothetical thinking is then used to determine what the hypothesis predicts will be

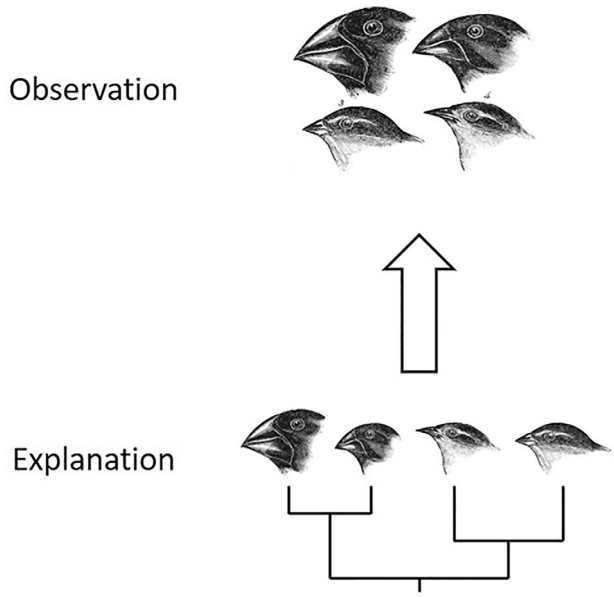


FIGURE 1 A simple model of science.

TABLE 1 Examples of observations and potential explanations.

Observation	Potential explanations
The sun rises in the east and sets in the west	• The sun orbits the Earth • Earth rotates on an axis
A column of mercury in a sealed barometer stands 32 inches tall	• A vacuum force pulls the mercury up • Atmospheric pressure pushes the mercury up
Student grades in college courses vary	Differences in: • Motivation • Self-efficacy • HD reasoning skill
The prevalence of lung cancer in the United States increased in early 20th century	• Air pollution from automobiles • Cigarette smoking
Mesosaurus fossils are present in South America and Africa	• South America and Africa were once connected by a land bridge • South America and Africa were once connected to each other like pieces of jigsaw puzzle
Male peacocks have colorful feathers	• Colorful feathers frighten predators • Colorful feathers attract females
Heart pumps blood into aorta	• Heart continuously creates blood • Heart circulates blood through body
Rate of violent crime in the United States dropped in the 1990s	• Lead eliminated from gasoline • Political/economic/social explanations
A burning candle in a sealed bell jar soon goes out	• Air in the jar becomes saturated with phlogiston • Air in the jar runs out of oxygen

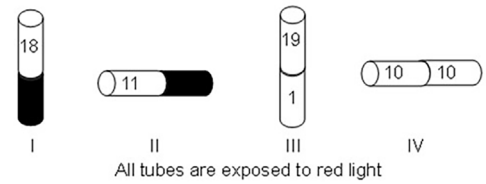
observed in some sort of test. Such prediction-making is essential for designing experiments (Item 2 in the list at the beginning of this paragraph). Alternatively, hypothetical thinking can be initiated from a phenomenon requiring explanation. In this case, hypothetical thinking is used to find a principle or concept that can explain the observation of interest (Item 1 in our list). These two practices, explanation seeking and prediction making, can be considered two sides of the same coin (Okasha, 2002): to say hypothesis X explains phenomenon Y is to say that if X were true, we would predict Y would be observed (Hempel, 1965). Lastly, hypothetical thinking can be used to interpret data from an experiment or other tests. The results of experiments are observations, and hypothetical thinking can evaluate how well a hypothesis predicts the data that was observed.

An example illustrates the relationship between scientific reasoning, hypothetical thinking, and learning that we believe exists. Consider the "Flies" question (Figure 2) from Lawson's Classroom Test of Scientific Thinking (Lawson, 2000). Students who answer this question correctly tend to learn more science content than classmates who do not (Results; See also Ding, 2014; Kalinowski & Willoughby, 2019; Lawson & Thompson, 1988). A conventional view of this question is that it assesses control-of-variables (Kalinowski & Willoughby, 2019; Lawson, 1978), a type of scientific reasoning that is widely viewed as important for students to

(a) Flies

Twenty fruit flies are placed in each of four sealed glass tubes. Tubes I and II are half covered with black paper. Tubes III and IV are not covered. The tubes are arranged as shown and are exposed to red light from all directions. The number of flies in the uncovered part of each tube is shown in the drawing. *This experiment shows the location of flies is affected by:*

a. Gravity but not red light.
b. Red light but not gravity.
c. Both red light and gravity.
d. Neither red light nor gravity.
e. The experiment has too many variables to tell.

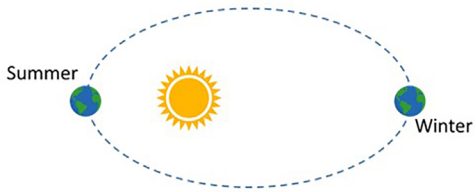


(b) Chiropractor

A chiropractic clinic wants to know if their standard 10-week treatment program for low back pain is effective. The clinic treated over a thousand men with low back pain, and 77% of them reported they felt "much better" after completing the ten weeks of treatment.

Does this data demonstrate the chiropractic treatment was effective?

a. Yes. The chiropractor treated a large number of patients, and more than three-quarters of them reported they felt much better after treatment.
b. Yes, but the clinic should confirm the treatment actually helped the patients by performing strength or flexibility tests at the start and end of each patient's 10 week treatment.
c. No. This data does not demonstrate the chiropractic treatment had any effect, and testing the strength and flexibility of patients will not change this.



(c) Seasons

Earth's orbit around the sun is elliptical (see diagram). This is a potential explanation for Earth's seasons: summer might be caused by Earth being close to the sun, and winter by Earth being far from the sun.

If this explanation for Earth's seasons is correct, what season would it be in Australia when it is SUMMER in the United States?

a. Spring or Fall
b. Summer
c. Winter
d. This would depend on other factors.

FIGURE 2 Three examples of hypothetico-deductive reasoning questions.

learn (Chen & Klahr, 1999; Inhelder & Piaget, 1958; Schwichow et al., 2016). However, how or why understanding control-of-variables would be useful for learning most science content is not clear (Ding, 2014; Kalinowski & Willoughby, 2019; Lawson & Thompson, 1988). We propose students use hypothetical thinking to answer the question. Students mentally simulate where flies should be if they respond to red light or gravity and then compare these predictions to the actual location of the flies. This is why we believe students who answer the "Flies" question correctly tend to learn more science content than classmates who do not answer the question correctly.

We have argued three core scientific practices rely heavily on hypothetical thinking and that hypothetical thinking is a single skill. So far, we have supported this claim with philosophical arguments, but this claim can and should be tested psychometrically (American Educational Research Association, 2014; Price, 2016). If these three scientific practices rely on different cognitive skills, students may be skilled at one practice and not another. This is equivalent to saying scientific thinking has multiple psychometric dimensions. If this is true, instructors will need to teach and assess each practice separately. On the other hand, if our argument is correct, students who are skilled at generating hypotheses will tend to be skilled at designing experiments and interpreting data. This is equivalent to saying scientific thinking is a one-dimensional psychometric construct. This can be tested by examining how students perform a variety of scientific reasoning tasks.

2.4 | Divergent validity and the potentially confounding effect of intelligence

We have hypothesized students use HD reasoning to learn science content in school. If this is true, scores on a test of HD reasoning will have predictive validity: they will be positively correlated with science grades. Such correlations, however, need to be interpreted with caution because scores on almost all cognitive tests are correlated (Cattell, 1963; Jensen, 1998; Lim, 1988; Spearman, 1904). This includes tests of skills as diverse as vocabulary, mathematics, language fluency, and musical ability (Spearman, 1904). This empirical phenomenon is called the positive manifold (Pluck & Cerone, 2021), and it is attributed to variation in general cognitive ability, also known as fluid intelligence. Fluid intelligence is the ability to solve problems and discern relationships without training on how to do so. It is often contrasted with crystallized intelligence, which is the ability to recall learned knowledge such as vocabulary (Cattell, 1963).

Because of the positive manifold, scores on a scientific reasoning test are likely to be correlated with grades and other measures of academic achievement—even if scientific reasoning plays no role in learning. Fluid intelligence is difficult to change (Jensen, 1998), so if variation in fluid intelligence explains the correlation between scientific reasoning and science learning that has frequently been observed (e.g., Lawson & Thompson, 1988), teaching students scientific reasoning would not help them learn science content. Contemporary education research does not seem interested in the role of intelligence in learning, but statistically controlling for its potential effect is essential if we want to understand how other traits, such as scientific reasoning skills, affect learning (Opitz et al., 2017). Stated another way, demonstrating that HD reasoning is distinct from fluid intelligence is important for establishing its divergent validity.

2.5 | Dual process theory

Scientific thinking is performed by the human mind and, therefore, is affected by all the strengths and weaknesses of human cognition. We will summarize one theory that has important implications for teaching scientific thinking. Dual process theory (Evans & Frankish, 2009; Stanovich, 2011; see Kahneman, 2011 for an easy-to-read introduction) describes human cognition as being performed by two collaborating cognitive systems: the autonomous mind (sometimes called System 1) and the deliberative mind (sometimes called System 2). The autonomous mind is responsible for the majority of decision-making and judgment. It is fast, effortless, unconscious, and can handle multiple tasks at the same time. Examples of tasks performed by the autonomous mind include recognizing a person is angry, driving in light traffic, or performing a task that has been rehearsed many times. In contrast, the deliberate mind is slow, effortful, and conscious. It can only perform one task at a time and requires a conscious decision to engage and sustain.

The autonomous mind uses a variety of biases and heuristics to make sense of the world. We will discuss five that are relevant to science education. The first is that the autonomous mind considers only the information at hand and does not ask if relevant information is missing. Because of this, the autonomous mind will not search for an alternative explanation if one is already available. Second, the autonomous mind is quick to see causation. Correlation, for example, is automatically interpreted as causation. Third, and fourth, the autonomous mind is confident in its conclusions when it can put together a coherent account of some event and can do this quickly. Lastly, the autonomous mind tends to answer difficult questions by replacing them with easier questions.

Engaging the deliberative mind is an important component of scientific thinking. This might sound obvious, but the autonomous mind is perfectly capable of generating wrong answers to many science questions. Consider the "Chiropractor" question (Figure 2). The question presents students with a research study that lacks a control but should have one. College students universally recognize this when it is pointed out, but the vast majority of students fail to recognize this on their own (see Section 4.2). Dual process theory offers an explanation. The autonomous mind: (1) considers only the information provided, (2) is quick to see causation, and is confident in its judgment because a (3) coherent explanation came to mind (4) quickly.

The deliberative mind is responsible for scientific reasoning and relies heavily on hypothetical thinking (often called "cognitive simulation") to reach its judgments. Hypothetical thinking can be cognitively challenging because the mind must decouple what it knows about the real world from the hypothetical world being explored, lest the two become confused. Consider the "Seasons" question (Figure 2). About half of introductory biology students at MSU answer this question incorrectly. Dual process theory offers an explanation: students struggle with this question because they do not decouple what they know about Earth's seasons in the real world from the seasons in the hypothetical world presented. Instead of answering the question presented to them, they answer an easier question "What season is it in Australia (in the real world) when it is summer in the United States?"

3 | SECTION 2: A TAXONOMY OF SCIENTIFIC THINKING QUESTIONS

In the first section of this article, we argued hypothetical thinking is useful for generating hypotheses, designing experiments, and interpreting data. This theory cannot be tested, nor can these practices be taught or assessed, unless they are accompanied by a detailed description of what these practices look like when performed by students. Therefore, in this section of the

paper, we will present a taxonomy (Table 2) of multiple-choice questions that ask students to generate hypotheses, design experiments, and interpret data. Three of the questions are shown in Figure 2; the rest are published online as Supporting Information. A thorough understanding of how we operationally defined scientific reasoning will require examining the questions in the Supporting Information, but we have written this section so that readers should be able to gain a reasonable understanding of each task without consulting the supplement.

3.1 | Category I: Generating hypotheses

The first category in the taxonomy includes questions that ask students to generate hypotheses. The most direct way to assess this skill is to present students with an interesting observation of the world and ask students to come up with potential explanations (Taxon IA). This is easily done during class discussions or with open-response questions but is difficult with multiple-choice questions. This is because if a multiple-choice question provides a list of potential explanations for some phenomenon,

TABLE 2 A taxonomy of scientific reasoning tasks and 32 multiple-choice questions.

Scientific reasoning tasks	Multiple choice questions
I. Generating hypotheses	
A. From observations with no explanations suggested.	(None)
B. Recognize an unmentioned explanation is plausible.	Chiropractor, Attendance, Seagulls, Potatoes, Lungs, Video Games, Pornography, Physics
II. Make a prediction from a hypothesis	Seasons, Mice Inbreeding, Candy, Full Moon
III. Designing tests or experiments	
A. Design experiment using HD logic (and not control-of-variables)	Seagulls, Keyboard
B. Control of variables	Fishbowls, Pendulums, Variables XYZ
IV. Interpret experiments or observations	
A. Reject hypothesis when results ≠ prediction	Aspens, Permian, Tree Roots, Trees on Mountain
B. Multiple hypotheses presented	Wolves & Elk, Pigeons
C. Interpret a sound, controlled experiment	
1. One independent variable	Tomatoes, Zinc
2. Multiple variables	Flies, Sportsade
D. Interpret an uncontrolled experiment	
1. Two variables explicitly varying	Apple Pie, Seeds, Tee Shirts, Teddy Bears
2. Control missing	Chiropractor, Seagulls, Potatoes, Lungs, Hospital
3. Hidden variable	Attendance, Video games, Physics
E. Correlation is not causation	
Hidden variable is plausible	Attendance, Video games, Physics, Pornography
Reverse causation is plausible	Video games, Pornography

Note: All the questions listed here are provided in the Supporting Information. The 24 questions given to BIOB 170 students during the fall semester of 2021 are listed in Table 3.

the question becomes an exercise in hypothesis evaluation rather than hypothesis generation. Therefore, Taxon IA has no multiple-choice questions associated with it. However, hypothesis-generation skill can be assessed by testing whether students recognize alternative explanations exist for observations presented in a question and Taxon IB is composed of questions that do that.

3.2 | Category II: Making predictions

This category includes questions that ask students to make a prediction from a hypothesis. No data interpretation is required nor must alternative hypotheses be considered. The "Seasons" question (Figure 2) is an example of such a task.

3.3 | Category III: Designing tests and experiments

The third category in the taxonomy includes questions that ask students to design experiments to test hypotheses. This category includes two classes of questions: questions that require HD reasoning without using control-of-variables and questions that apply control-of-variables.

3.4 | Category IV: Interpreting data

The largest category in the taxonomy is composed of questions that ask students to interpret observations or the results of experiments. This category includes five classes of questions (A–E).

3.4.1 | Reject a hypothesis when results differ from prediction

This class of questions presents students with *one* hypothesis and observations that do not agree with what the hypothesis predicts should be observed. The canonical interpretation of such observations is that a hypothesis is wrong, but these questions can be difficult if distractors mention irrelevant concepts (Tree Roots), other hypotheses ("Trees on Mountain"), or present data that agrees with some predictions but not others (Permian).

3.4.2 | Multiple hypotheses presented

The next class of question requires students to interpret observations when multiple hypotheses are explicitly mentioned. For example, the "Wolves and Elk" question asks students to figure out what caused a population of elk to decline: wolf predation or harsh winter weather. This is an example of abduction, also known as reasoning to the best explanation.

3.4.3 | Interpret a sound, controlled experiment

This class of questions presents students with a sound controlled experiment and assesses whether students can evaluate the results. It includes two types: questions with one independent variable and questions with multiple independent variables.

One independent variable

This taxon presents students with a randomized, controlled experiment with one independent variable and assesses whether students interpret this as evidence of causation. These questions can be difficult if students are distracted by a modest treatment effect (Tomatoes) or the mention of potential differences between treatment and control groups (Zinc).

Multiple variables

Another challenging type of questions for many students is a controlled experiment with multiple independent variables ("Flies"; Figure 2).

3.4.4 | Interpret an uncontrolled experiment

This class includes questions that ask students to interpret uncontrolled experiments. It includes three distinct types of uncontrolled experiments.

Two variables explicitly varying

These questions present students with an experiment that has two independent variables and assess how students interpret the results. The difficulty of these questions varies considerably depending on how the question is asked ("Apple Pie", "Tee Shirts," and "Teddy Bears").

Control missing

This group of questions describes experiments that should have a control but do not (e.g., "Chiropractor"; Figure 2).

Hidden variable

This group of questions describes experiments that appear to be controlled experiments but may be affected by hidden variables.

3.4.5 | Correlation is not causation

The adage "correlation is not causation" will be familiar to readers. The final class in the taxonomy includes questions that apply this concept. All the questions in this class belong to other taxa, but we have included this class because the concept "correlation is not causation" is valuable and we want to highlight its applicability.

4 | SECTION 3: TESTING THE THEORY

In the preceding two sections of this article, we described a HD theory of science and learning (Section 1) and a taxonomy of different ways HD reasoning can be applied (Section 2). In this section, we test the theory using data from an introductory college biology course. We will begin by presenting four predictions made by the theory and then determine whether data from the biology course is consistent with these predictions.

Prediction 1: *Unidimensionality*. Student responses to multi-choice questions that assess students' ability to generate hypotheses, design experiments, and interpret data will be consistent with a one-dimensional psychometric model.

Prediction 2: *Predictive validity of HD reasoning*. Scores on a HD reasoning test will predict exam grades in an introductory biology course.

Prediction 3: *Controlling for fluid intelligence*. HD reasoning will predict grades when fluid intelligence is controlled for.

Prediction 4: *Predictive validity of HD reasoning test questions*. Individual HD questions will be correlated with grades.

4.1 | Methods

4.1.1 | Student population

The data presented here was collected in BIOB 170—Principles of Biological Diversity at Montana State University. BIOB 170 is an introductory biology course on ecology, evolution, and organismal diversity intended for biology majors. It was taught by three instructors, including STK, with each instructor teaching one of the aforementioned topics. Class met twice a week for a 75-min lecture and once a week for a two-hour lab. Four exams were administered: three midterms and a comprehensive final. Almost all of the exam questions were multiple-choice. All four exams had equal weight and the lowest exam score for each student was dropped at the end of the semester. If a student missed one exam, this exam score was dropped. Enrollment at the end of the semester was 259 students.

Demographic information was obtained from university records. These records showed 55% of the students were female and 92% were Caucasian. Most of the students were Freshman (70%) or Sophomores (18%) and most (82%) were science majors. The median age in the class was 19 years old, with only 7% of the students being 22 years old or older. Consent was obtained from students to participate in this research following MSU's Institutional Review Board policy.

4.1.2 | HD reasoning test

A test of HD reasoning was created by using questions from published sources and writing new questions (Table 3). Development of the test followed an iterative cycle of revision during seven semesters prior to this study. Candidate questions were given to students on pretests, exams, and nongraded posttests. Questions that were too easy, too difficult, or had low discrimination were dropped or revised. Semi-structured interviews were conducted during spring semesters of 2018 and 2019 ($N = 31$ students), and these interviews were used to revise the wording of questions. Twenty-four questions were selected for the HD instrument. This number of questions was chosen in order to maximize instrument reliability without creating too much testing fatigue.

The data presented here were collected during the fall semester of 2021. The HD reasoning test was presented to students as a pretest of scientific reasoning and was administered during the first full week of the semester as a regular online quiz. Students were not told whether their responses were correct and did not have access to the test questions after they answered them.

4.1.3 | Psychometric analysis

Item response theory (IRT) was used for psychometric analysis (De Ayala, 2009), and all calculations were performed in the R statistics environment (R Core Team, 2020). The *mirt* statistics package (Chalmers, 2012) was used for model fitting and score estimation. A one-dimensional, three-parameter (3PL) IRT model was fit to student responses to the HD reasoning test. This

TABLE 3 Source and summary statistics for questions on the HD reasoning test: Published source of question (Source), Proportion of students answering the question correctly ($P_{Correct}$), item response theory difficulty coefficient (IRT Diff.), item response theory discrimination coefficient (IRT Disc), item response theory guess rate (IRT Guess), and biserial correlation between responses to each question and average exam grades (Biserial corr).

Question	Source	$P_{Correct}$	IRT diff.	IRT disc.	IRT guess	Biserial corr.
1. Apple pie	1[a]	0.42	1.19	14.21	0.34	0.31***
2. Seasons		0.30	1.15	2.80	0.16	0.43***
3. Chiropractor		0.05	2.31	2.50	0.03	0.30**
4. Aspens		0.69	−0.95	1.04	0.00	0.35***
5. Fishbowls	1	0.74	−0.61	1.64	0.22	0.43***
6. Mice inbreeding		0.53	0.13	1.62	0.12	0.39***
7. Pendulums	2	0.78	−1.36	1.17	0.00	0.32***
8. Candy		0.78	−0.71	1.70	0.28	0.42***
9. Attendance		0.10	2.72	1.14	0.03	0.04
10. Wolves and elk	3	0.64	−0.60	1.26	0.00	0.35***
11. Flies	2	0.53	−0.15	0.77	0.00	0.25***
12. Full Moon		0.37	0.79	0.78	0.00	0.26***
13. Tee-shirts	1[a]	0.32	1.14	1.32	0.10	0.22**
14. Red blood cells	2	0.69	−1.36	0.65	0.00	0.18*
15. Tree roots	3	0.20	2.07	1.29	0.10	0.32***
16. Tomatoes		0.17	2.16	2.59	0.14	−0.04
17. Seagulls	3	0.45	1.35	1.28	0.31	0.16*
18. Potatoes	3	0.38	0.91	2.88	0.20	0.39***
19. Permian		0.34	1.11	0.65	0.00	0.16*
20. Teddy bears	1[a]	0.20	1.45	2.90	0.11	0.24**
21. Trees on Mtn		0.63	0.63	1.40	0.43	0.14
22. Zinc		0.50	−0.01	0.68	0.00	0.13
23. Pigeons	3a	0.57	0.59	1.45	0.34	0.34***
24. Variables XYZ	1[a]	0.68	−0.82	1.12	0.00	0.49***

Note: Questions are listed in the order presented on the HD reasoning test. See the Supporting Information for the actual questions.

[a]Variation of a published question.

*p value <0.05. **p value <0.01. ***p value <0.001.

Source: (1) American Association for the Advancement of Science (2020); (2) Lawson (2000); (3) Kalinowski and Willoughby (2019).

model includes three parameters for each question: difficulty, discrimination, and a "guessing" rate. The discrimination parameter is particularly informative for test development. It quantifies how strongly responses to each question are correlated with scores on the entire test. A high discrimination coefficient indicates a question assesses the same skill as the other questions on the test. Marginal maximum likelihood (Bock & Aitkin, 1981) was used to estimate IRT parameters for each question. This widely used method assumes the HD reasoning abilities of test-takers are normally distributed with a mean of zero and a standard deviation of one.

We used the Bayesian EAP (expected a priori) method with a standard normal prior (Bock & Aitkin, 1981; Kim & Nicewander, 1993) to estimate student scores. We standardized scores to have a variance of 1.0. The reliability of test scores, R_{xx}, was estimated using the formula $R_{xx} = 1 - \widehat{MSE}$, where $\widehat{MSE}$ is the mean squared standard error of each student's estimated score (Kim, 2012). As the formula shows, the reliability of test scores is a measure of estimation error for test scores. More specifically, this reliability coefficient is an estimate of the correlation (squared) between test scores (which are estimates) and the unknown abilities being estimated (Kim, 2012).

Test 1: Unidimensionality

The fit of the one-dimensional psychometric model was assessed using a chi-squared test (Orlando & Thissen, 2000) with a Bonferroni correction for multiple comparisons and a graphical method (Kalinowski & Willoughby, 2019).

Test 2: Predictive validity of HD reasoning test scores

We used average exam grades as a measure of academic performance. Students who missed more than one of the four exams were not included in the analysis.

We quantified the predictive validity of HD reasoning in two ways. In the first analysis, we calculated the average exam grade for each raw score (number of questions correct) on the HD reasoning test. We performed this simple analysis because the results can easily be visualized with a graph.

In the second analysis of predictive validity, we used linear regression to quantify how scores on the HD reasoning test predicted exam grades:

$$\overline{Exam} = Int + B_{HD}\widehat{HD} \tag{1}$$

where $\overline{Exam}$ is the average exam grade of a student, Int is the intercept, $\widehat{HD}$ is the HD test score of a student, and B_{HD} is the regression coefficient quantifying how HD reasoning scores related to exam scores. Scores on the HD reasoning test had a mean of zero, so the intercept is the expected exam grade for a student with an average score on the HD reasoning test.

Confidence intervals for the regression coefficient in this model (and in other models described below) were estimated using Wild bootstrapping with a Rademacher distribution of residuals and 50,000 bootstrap samples (Wu, 1986). This nonparametric method for estimating confidence intervals for regression coefficients works much better than parametric methods when heteroscedasticity is present (which is common with this sort of data) and only slightly less well when heteroscedasticity is not present (Hodoshima & Ando, 2010).

Test 3: Controlling for fluid intelligence

As discussed above, scores on cognitive tests tend to be correlated due to variation in fluid intelligence. In order to control for this potential effect, we estimated students' fluid intelligence using 20 questions from Raven's Advanced Progressive Matrices (RAPM; Raven, 1990). RAPM is a nonverbal test of general cognitive ability that assesses students' ability to detect patterns in matrices of geometric figures. The test was administered online with a time limit of 25 min. Students who took the test received three points of extra credit on the first exam. The test was scored using item response theory (as described above) and a two-parameter (2PL) model. Estimates of fluid intelligence were included in two regression models:

$$\overline{Exam} = Int + B_g\widehat{RAPM} \tag{2}$$

$$\overline{Exam} = Int + B_{HD}\widehat{HD} + B_g\widehat{RAPM} \tag{3}$$

where $\widehat{RAPM}$ represents scores on Raven's Advanced Progressive Matrices. The regression model shown in Equation 2 was not strictly necessary to fulfill the goals of this investigation, but was included to provide a comparison with the other regression models.

Test 4: Predictive validity of HD reasoning test questions

Lastly, the predictive validity of individual HD questions was measured by calculating the biserial correlation coefficient between responses to each question and the average exam grade for each student. Values of p for biserial correlation coefficients were estimated using a randomization test (10,000 randomizations).

4.2 | Results

4.2.1 | Summary statistics

The HD reasoning test was completed by 264 students. Raw scores on the test had an approximately symmetric, unimodal distribution with a mean of 46% correct (Figure 4). The percentage of students answering each question correctly ranged from 5% to 78% (Table 3). The median discrimination coefficient for the questions on the HD reasoning test was 1.3 (Table 3). The reliability of HD reasoning test scores was estimated to be 0.79. The Raven's Advanced Progressive Matrices (RAPM) questions were answered by 185 students. Scores on the RAPM had an estimated reliability of 0.81. Scores on both tests were correlated with each other and with exam grades (Table 4).

Test 1: Unidimensionality

Student responses to questions on the HD reasoning test were consistent with a one-dimensional psychometric model. No statistically significant problems with item fit were detected with the chi-squared test; and no obvious problems with item fit were observed in the graphs for each question (Figure 3).

TABLE 4 Pearson correlation coefficients between scores on the hypothetico-deductive (HD) reasoning test, Raven's advanced progressive matrixes (RAPM), and average exam grades.

	HD	RAPM	Exams
HD	–	0.44	0.56
RAPM	0.44	–	0.40
Exams	0.56	0.40	–

Test 2: Predictive validity of HD reasoning test scores

Students who had high scores on the HD reasoning test tended to have high exam grades. The relationship was easily seen in the graph comparing scores on the HD reasoning test with exam grades (Figure 4). The trend was also evident in the regression analyses (Table 5). In a univariate regression model (Equation 1), increasing HD reasoning skill by one standard deviation increased expected exam scores by 6.7 points (on a 100-point scale). This one-variable model explained 36 percent of the variance in grades.

Test 3: Controlling for fluid intelligence

Including scores from Raven's Advanced Progressive Matrices in the regression model did not change these results substantially (Compare Equation 1 with Equation 3 in Table 5): the adjusted R-squared value increased by only 2% and the regression coefficient for HD reasoning decreased from 6.70 to 5.79.

Test 4: Predictive validity of HD reasoning test questions

Biserial correlation coefficients (Table 3) showed that 20 out of the 24 questions on the HD reasoning test were correlated (at the 0.05 level of significance) with exam grades.

5 | SECTION 4: DISCUSSION

5.1 | Summary

Section 1 of this article combined ideas from the philosophy of science, constructivism, conceptual change theory, Piaget's theory of development, and cognitive psychology to create a novel theory of science and learning. Stated simply, this theory claims that the ability to generate, explore, and evaluate hypothetical scenarios is useful for a wide variety of scientific research practices and for students learning science content.

We tested this theory four ways, and the results of all four tests were consistent with predictions made by the theory. In the first test, student responses to 24 HD reasoning questions were consistent with a one-dimensional psychometric construct. This is consistent with our theory's claim that a single cognitive skill is used to generate hypotheses, design experiments, and interpret results. In the second test, scores on the HD reasoning test explained 36% of the variance in exam grades in an introductory biology course. In the third test, controlling for fluid intelligence had little effect upon the predictive validity of the HD reasoning test. Lastly, 20 out of the 24 HD reasoning questions were correlated with exam grades.

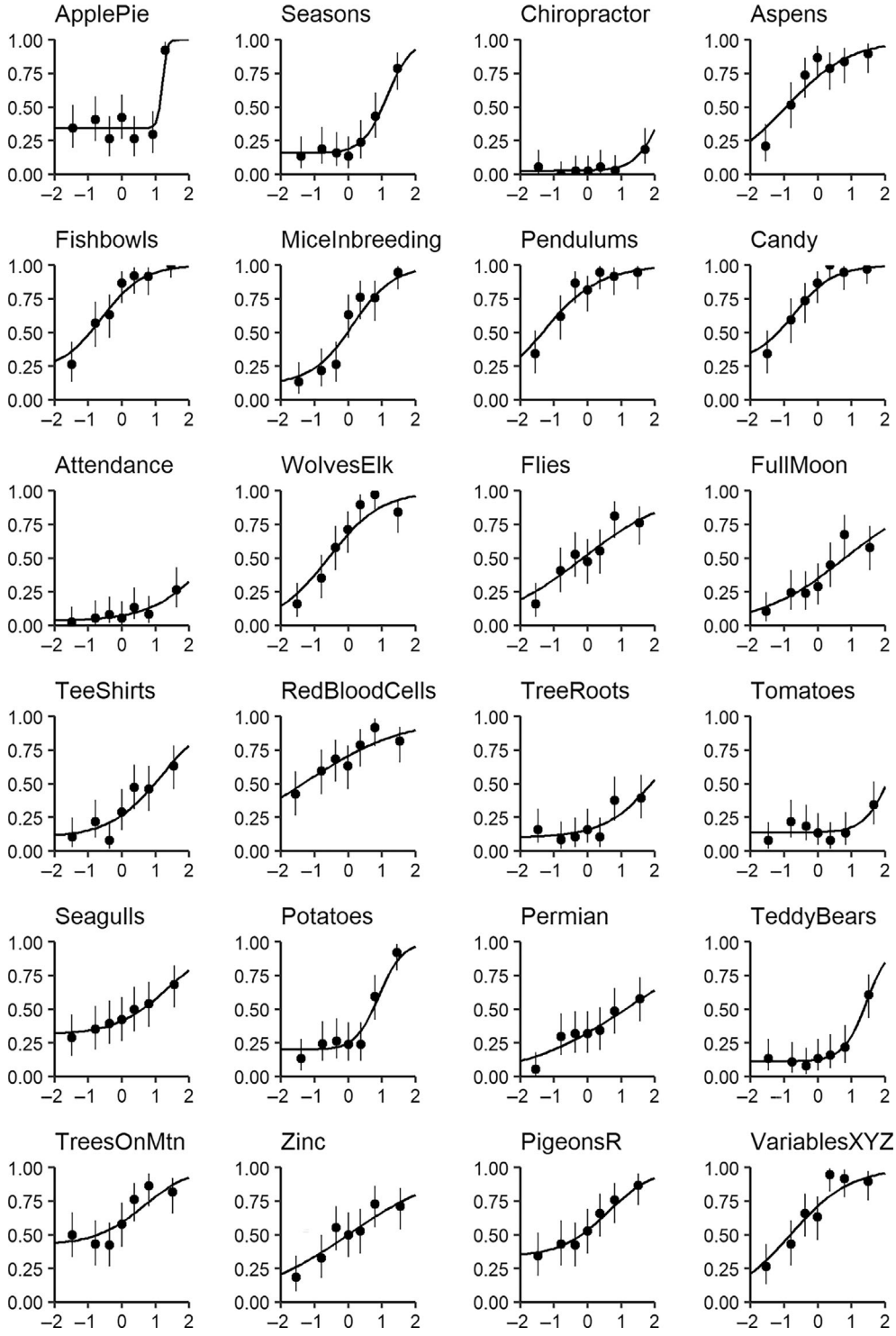


FIGURE 3 Legend on next page.

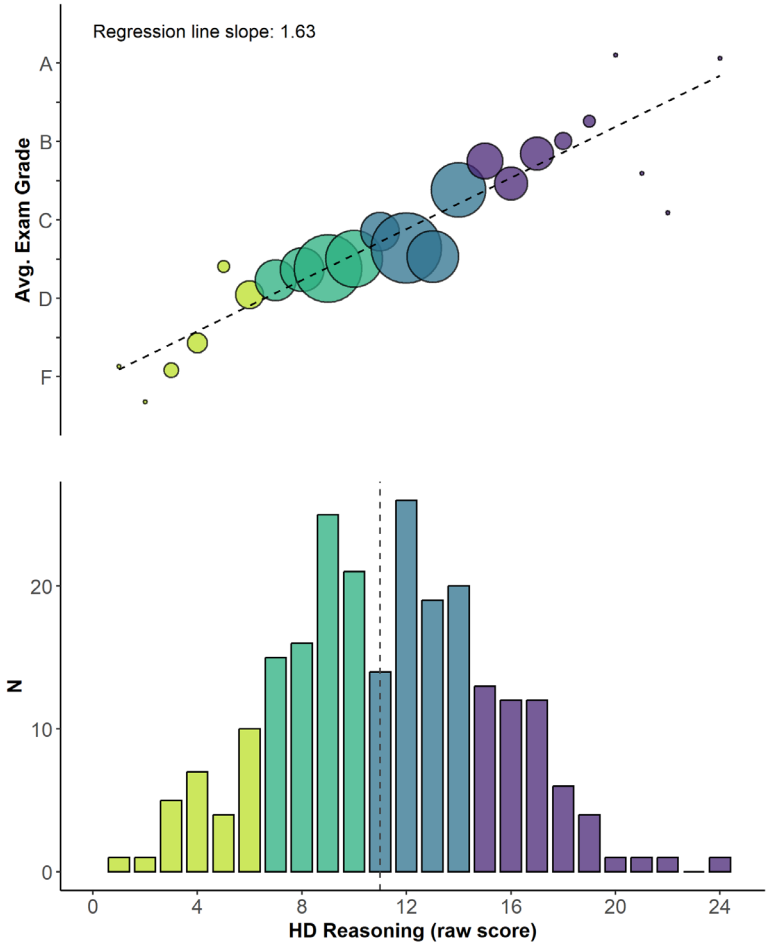


FIGURE 4 Average exam grades for each raw score on the HD reasoning test. The lower panel shows a histogram for raw scores (number correct) on the HD reasoning test. The median score is depicted with a vertical dashed line. The upper panel shows the average exam score for each raw score on the HD reasoning test. The size of each "bubble" is proportional to the sample size. The dashed line in the upper panel depicts a regression line fit to the data.

5.2 | Comparison to previous research

This article adds to a growing body of empirical research that shows students with strong scientific reasoning skills learn more science content than classmates with less developed skills (Ates & Cataloglu, 2007; Cannady et al., 2019; Cavallo, 1996; Cracolice & Busby, 2015;

FIGURE 3 IRT trace curves for 24 scientific thinking questions among 264 students in an introductory biology course. Each panel shows results for one question. The horizontal axis measures student ability (measured in standard deviations from the mean), and the vertical axis measures the probability of students answering the question correctly. The 3PL IRT curve for each question is shown with a solid black line. Filled circles depict proportions for seven groups of students binned by scores on the entire test (vertical lines show 95% confidence intervals).

TABLE 5 Coefficients in three regression models for average exam grades in an introductory college biology course.

	Equation 1	Equation 2	Equation 3
Intercept	73.4	74.0	73.5
HD reasoning	6.70***	–	5.79***
Raven's APM	–	4.87***	2.13*
N (Number of students)	171	171	171
Adjusted R^2	0.36	0.17	0.38

Note: Exam grades were measured on a conventional 100-point scale where 95 is an A, 85 is an B, and so forth. Scores on the HD reasoning test and the fluid intelligence test (Raven's APM) were calculated using item response theory and had a mean of zero and a standard deviation of one. All three regression models used the same data.

*95% confidence interval does not contain zero.

***99.9% confidence interval does not contain zero.

Ding, 2014; Kalinowski & Willoughby, 2019; Lawson & Thompson, 1988; Moore & Rubbo, 2012; Niaz, 1985). This investigation advances this research in three ways. First, it is one of three investigations (Cracolice & Busby, 2015; Lawson & Thompson, 1988) that has controlled for general intelligence. Second, this investigation provided a more detailed explanation of why scientific reasoning is useful for learning science content. Third, and last, this project developed a taxonomy of scientific thinking tasks and multiple-choice questions to assess them.

This article also contributes to the literature seeking to understand how science makes sense of the world. As discussed above, the National Research Council (2012) has emphasized science is complex and has advocated teaching students eight different scientific practices. This perspective originated from historical and sociological analyses of science (Collins & Pinch, 1993; Latour & Woolgar, 1986; Pickering, 1995). In this article, we presented an alternative and simpler view of science that emphasizes the role of hypothetical thinking.

We used psychometric data to support our claim that three core scientific practices rely on the same cognitive skill. This type of evidence is highly relevant to this question (American Education Research Association, 2014) but has often been absent from discussions of the cognitive skills used to do science. For example, Zimmerman and Klahr (2018, table 7.1) argued that generating hypotheses, designing experiments, and interpreting data were distinct cognitive processes but did not support their claim with psychometric data. Nor did Kind and Osborne (2017) when they argued science uses six styles of scientific reasoning.

Debates regarding the nature of scientific reasoning are worth having because they may inform curriculum design. However, completely resolving these debates may not be necessary for improving science instruction. If the most rudimentary interpretation of our theory is correct, teaching students scientific reasoning should increase how much science content they learn. Available research suggests this is likely to occur. Several research projects have shown that scientific thinking skills can be taught (e.g., Brownell et al., 2015; Erlina et al., 2018; Schwichow et al., 2016). Only a few investigations have examined whether teaching scientific thinking skills improves academic performance in subsequent courses, but the available results are promising. The most relevant experiment might be that of Shayer and Adey (1993) who showed that teaching HD reasoning to middle school students in the United Kingdom improved standardized test scores 3 years later. More recently, faculty at the University of Washington (Buchwitz et al., 2012; Dirks & Cunningham, 2006) showed that biology students who received training in science process skills earned higher grades in an introductory biology course.

5.3 | Limitations and future research

The most important limitation of this investigation is that it was an observational study and does not demonstrate causation. Our data showed a correlation between HD reasoning and exam grades in a college biology course. Our theory's explanation for this correlation is that students use HD reasoning to learn science. However, an alternative explanation is that we have inferred the direction of causation backward: students might develop scientific reasoning skill by learning science content. This possibility deserves additional study, but two lines of evidence argue against it. First, improving science content knowledge does not seem to increase scientific reasoning skill (Bao et al., 2009), and second, the relationship between scientific reasoning skill and learning documented in this investigation was also present in a similar study that controlled for pre-instructional content (Kalinowski & Willoughby, 2019). A third possible explanation for the correlation presented in this investigation is that a hidden (unmeasured) variable influences both scientific reasoning and grades. Possible candidates for such a hidden variable include executive function (Zelazo et al., 2016), metacognitive skill (de Boer et al., 2018), or some type of motivation (Glynn et al., 2011). Testing for the presence of a hidden variables is not difficult if it can be included in a regression analysis, and we hope future research tests candidate variables. The strongest way to demonstrate a causal relationship between HD reasoning and learning would be to perform a randomized, controlled experiment testing whether improving students' scientific reasoning improves learning in subsequent science courses. This is a high priority for future research.

Additional research is also needed on the psychometric structure of scientific thinking. We argued scientific reasoning is a unidimensional skill, but this is surely an oversimplification. Some sort of hierarchy of cognitive skills is likely, with hypothetical thinking at the top of the hierarchy and more specific skills lower in the hierarchy. We hope future work identifies these skills and quantifies how they contribute to scientific thinking, problem-solving, and learning. We hope such research also explores the role of deliberation in scientific thinking. Scientific reasoning skill is not useful unless students apply the skill they possess. The disposition to apply System 2 mental processes for answering scientific questions may be just as important for scientific thinking as possessing well-developed reasoning skills. This is an important concept in dual process theory (Stanovich, 2011) and a central concept in Lawson's (1992) multiple-hypothesis theory of scientific reasoning. Students might become better scientific thinkers and more effective learners by learning to deliberate more. This deserves study.

On a related topic, we have argued three core scientific practices—generating hypotheses, designing experiments, and interpreting data—rely heavily on hypothetical thinking. We did not investigate other scientific practices. For example, the National Research Council (2012) identified five other practices that are important for doing science: asking questions, developing models, using mathematics, engaging in arguments from evidence, and obtaining/evaluating/communicating information. Some of these practices—for example, asking questions, developing models, engaging in argument from evidence—seem likely to use hypothetical thinking. We hope future research will determine whether this is true or not. This will require operational definitions of these practices so they can be assessed. We also hope future research will quantify how mastery of these practices helps students learn science content. In this investigation, differences in HD reasoning skill explained 36% of the variance in exam grades. This is an impressive amount for one variable, but 64% of the variance remains to be explained, and we hope future research will develop more comprehensive models of academic performance.

We hope future research will examine the role of scientific reasoning for learning different types of science content. Previous research has shown scientific reasoning to be correlated with

test scores in chemistry (Cracolice & Busby, 2015), physics (Ates & Cataloglu, 2007; Ding, 2014; Moore & Rubbo, 2012), as well as biology (Cannady et al., 2019; Cavallo, 1996; Kalinowski & Willoughby, 2019; Lawson & Thompson, 1988). This suggests scientific reasoning may be broadly useful for learning a wide variety of science concepts. However, scientific reasoning may be more useful for learning some concepts than others. We hypothesize scientific reasoning is especially useful for learning concepts that require conceptual change or understanding phenomena that cannot be directly observed. We further hypothesize scientific reasoning is less useful for learning science content that can be memorized. Scientific reasoning might also be useful for learning history or economics, and we hope future research explores these questions.

Another limitation of this investigation is that it used exam grades as a measure of academic performance. Exam grades have the advantage of being a de facto measure of academic performance in universities. However, they have the disadvantage of being affected by pre-instructional knowledge: grades reflect how much students learned in a course *and* prior to a course. A similar study to this one (Kalinowski & Willoughby, 2019) controlled for pre-instructional knowledge and had similar results to those presented here; therefore, our lack of a pre-instructional measure of student knowledge may not have had a substantial impact on our results. Nevertheless, this deserves additional study.

Another limitation of this research is that most of the students were white. We hope future research will examine the scientific reasoning skills of other demographic groups. In particular, we are concerned some populations of students may receive educations that do not fully develop their reasoning skills (National Research Council, 2003). If this is the case, it may explain some of the disparity among demographic groups in American post-secondary education (Espinosa et al., 2019). Increasing access to training in scientific thinking, therefore, could promote equity and diversity in science (Dirks & Cunningham, 2006).

5.4 | Concluding comments

For over a century, science reformers have exhorted college faculty to teach students the scientific process (e.g., American Association for the Advancement of Science, 2020; Dewey, 1910; National Research Council, 2012). Yet, college faculty continue to focus on teaching science content (Addis & Powell-Coffman, 2018; Coil et al., 2010; Momsen et al., 2010; Petersen et al., 2020). The best-supported explanation for this focus is that instructors fear teaching science process skills will interfere with teaching science content (Coil et al., 2010; Petersen et al., 2020). The results of this investigation suggest such a tradeoff may be illusionary. Teaching students scientific reasoning skills will take time, and therefore, will presumably reduce how much science content instructors can *cover*. However, teaching students how to reason more effectively might increase how much science content students *learn*. This is exactly what happens when instructors make room in traditional lectures for active learning exercises (Freeman et al., 2014). If this is the case, it provides additional motivation to teach students scientific reasoning.

ACKNOWLEDGMENTS

We thank over 10,000 students who answered scientific reasoning questions during the past 7 years and three anonymous reviewers for comments that improved this article. The National Science Foundation provided funding for this work (grant number 1432577).

ORCID

Steven T. Kalinowski https://orcid.org/0000-0001-8504-4923

ENDNOTE

[i] In most circumstances, the distinction between scientific "thinking" and scientific "reasoning" is not consequential and we shall mostly use the terms interchangeably. The difference is that thinking includes both conscious and unconscious mental processes while reasoning involves the deliberate (conscious) application of logic.

REFERENCES

Addis, E. A., & Powell-Coffman, J. A. (2018). Student and faculty views on process of science skills at a large, research-intensive university. *Journal of College Science Teaching*, *47*(4), 72–82.

Alvarez, W. (1997). *T. Rex and the crater of doom*. Princeton University Press.

American Association for the Advancement of Science. (2020, August 1). Project 2061. Science Assessment Website. http://assessment.aaas.org/

American Educational Research Association. (2014). *Standards for educational and psychological testing*. AERA.

Amsel, E. (2011). Hypothetical thinking in adolescence: Its nature, development, and applications. In *Adolescence: Vulnerabilities and opportunities* (pp. 86–113).

Ates, S., & Cataloglu, E. (2007). The effects of students' reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics. *European Journal of Physics*, *28*(6), 1161–1171. https://doi.org/10.1088/0143-0807/28/6/013

Ball, L. (2020). Hypothetical thinking. In A. Abraham (Ed.), *The Cambridge handbook of the imagination (Cambridge handbooks in psychology)* (pp. 514–528). Cambridge University Press. https://doi.org/10.1017/9781108580298.031

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., Li, L., & Wu, N. (2009). Learning and scientific reasoning. *Science*, *323*(5914), 586–587. https://doi.org/10.1126/science.1167740

Blachowicz, J. (2009). How science textbooks treat scientific method: A philosopher's perspective. *The British Journal for the Philosophy of Science*, *60*(2), 303–344. https://doi.org/10.1093/bjps/axp011

Blackburn, S. (2005). *The Oxford dictionary of philosophy*. Oxford University Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Chandler Seawell, P., Conklin Imam, J. F., Eddy, S. L., Stearns, T., & Cyert, M. S. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE—Life sciences Education*, *14*(2), ar21. https://doi.org/10.1187/cbe.14-05-0092

Buchwitz, B. J., Beyer, C. H., Peterson, J. E., Pitre, E., Lalic, N., Sampson, P. D., & Wakimoto, B. T. (2012). Facilitating long-term changes in student approaches to learning science. *CBE—Life Sciences Education*, *11*(3), 273–282. https://doi.org/10.1187/cbe.12-01-0011

Butts, R. E. (2015). Hypothetico-deductive method. In R. Audi (Ed.), *The Cambridge dictionary of philosophy* (3rd ed.). Cambridge University Press.

Cannady, M. A., Vincent-Ruz, P., Chung, J. M., & Schunn, C. D. (2019). Scientific sensemaking supports science content learning across disciplines and instructional contexts. *Contemporary Educational Psychology*, *59*, 101802. https://doi.org/10.1016/j.cedpsych.2019.101802

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22.

Cavallo, A. M. (1996). Meaningful learning, reasoning ability, and students' understanding and problem solving of topics in genetics. *Journal of Research in Science Teaching*, *33*(6), 625–656.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Chi, M. T. (2009). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In *International handbook of research on conceptual change* (pp. 89–110). Routledge.

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE—Life Sciences Education*, *9*(4), 524–535. https://doi.org/10.1187/cbe.10-01-0005

Collins, H., & Pinch, T. (1993). *The golem: What everyone should know about science*. Cambridge University Press.

Cracolice, M. S., & Busby, B. D. (2015). Preparation for college general chemistry: More than just a matter of content knowledge acquisition. *Journal of Chemical Education*, *92*(11), 1790–1797. https://doi.org/10.1021/acs.jchemed.5b00146

Darwin, C. (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. John Murray.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford.

de Boer, H., Donker, A. S., Kostons, D. D., & van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review*, *24*, 98–115. https://doi.org/10.1016/j.edurev.2018.03.002

Dewey, J. (1910). Science as subject-matter and as method. *Science*, *31*, 121–127.

Ding, L. (2014). Verification of causal influences of reasoning skills and epistemology on physics conceptual learning. *Physical Review Special Topics - Physics Education Research*, *10*(2), 023101. https://doi.org/10.1103/PhysRevSTPER.10.023101

Dirks, C., & Cunningham, M. (2006). Enhancing diversity in science: Is teaching science process skills the answer? *CBE—Life Sciences Education*, *5*(3), 218–226. https://doi.org/10.1187/cbe.05-10-0121

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2–3), 105–225.

Dvir, M., & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM Mathematics Education*, *50*(7), 1183–1196. https://doi.org/10.1007/s11858-018-0987-4

Erlina, N., Susantini, E., Wicaksono, I., & Pandiangan, P. (2018). The effectiveness of evidence-based reasoning in inquiry-based physics teaching to increase students' scientific reasoning. *Journal of Baltic Science Education*, *17*(6), 972–985. https://doi.org/10.33225/jbse/18.17.972

Espinosa, L. L., Turk, J. M., Taylor, M., & Chessman, H. M. (2019). *Race and ethnicity in higher education: A status report*. American Council on Education.

Evans, J. S. B. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Psychology Press.

Evans, J. S. B., & Frankish, K. E. (2009). *In two minds: Dual processes and beyond*. Oxford University Press.

Faye, J. (2016). *The nature of scientific thinking: On interpretation, explanation and understanding*. Springer.

Fosnot, C. T. (2013). *Constructivism: Theory, perspectives, and practice*. Teachers College Press.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gary, K. (1998). Sources of power: How people make decisions. *Nature*, *392*(6673), 242–242.

Glynn, S. M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching*, *48*, 1159–1176. https://doi.org/10.1002/tea.20442

Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach*, *2*(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, *53*(11), 1056–1065. https://doi.org/10.1119/1.14031

Hempel, C. G. (1965). *Aspects of scientific explanation* (Vol. 3). Free Press.

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice Hall.

Hodoshima, J., & Ando, M. (2010). Bootstrapping stochastic regression models under homoskedasticity: Wild bootstrap vs. pairs bootstrap. *Journal of Statistical Computation and Simulation*, *80*(11), 1225–1235. https://doi.org/10.1080/00949650903014971

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. Basic Books, Inc.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Johnson, C., Affolter, M. D., Inkenbrandt, P., & Mosher, C. (2017). *An introduction to geology*. Salt Lake Community College www.opengeology.org/textbook

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.

Kalinowski, S. T., & Willoughby, S. (2019). Development and validation of a scientific (formal) reasoning test for college students. *Journal of Research in Science Teaching*, *56*(9), 1269–1284. https://doi.org/10.1002/tea.21555

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*(4), 587–599.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*(1), 153–162.

Kind, P. E. R., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education*, *101*(1), 8–31. https://doi.org/10.1002/sce.21251

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48. https://doi.org/10.1016/0364-0213(88)90007-9

Klahr, D., & Simon, H. A. (2001). What have psychologists (and others) discovered about the process of scientific discovery? *Current Directions in Psychological Science*, *10*(3), 75–79. https://doi.org/10.1111/1467-8721.00119

Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 497–523). Wiley-Blackwell.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lawson, A. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. *International Journal of Science Education*, *25*(11), 1387–1408. https://doi.org/10.1080/0950069032000052117

Lawson, A. E. (1978). Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24. https://doi.org/10.1002/tea.3660150103

Lawson, A. E. (1992). What do tests of "formal" reasoning actually measure? *Journal of Research in Science Teaching*, *29*(9), 965–983. https://doi.org/10.1002/tea.3660290906

Lawson, A. E. (1995). *Science teaching and the development of thinking*. Wadsworth.

Lawson, A. E. (2000). *Development and validation of the classroom test of formal reasoning* (Revised ed.). Arizona State University.

Lawson, A. E., & Thompson, L. D. (1988). Formal reasoning ability and misconceptions concerning genetics and natural selection. *Journal of Research in Science Teaching*, *25*(9), 733–746. https://doi.org/10.1002/tea.3660250904

Lim, T. K. (1988). Relationships between standardized psychometric and Piagetian measures of intelligence at the formal operations level. *Intelligence*, *12*(2), 167–182. https://doi.org/10.1016/0160-2896(88)90014-1

Mintzes, J. J. (2020). From constructivism to active learning in college science. In J. J. Mintzes & E. M. Walter (Eds.), *Active learning in college science: The case for evidence-based practice* (pp. 3–12). Springer.

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE—Life Sciences Education*, *9*(4), 435–440. https://doi.org/10.1187/cbe.10-01-0001

Moore, J. C., & Rubbo, L. J. (2012). Scientific reasoning abilities of nonscience majors in physics-based courses. *Physical Review Special Topics - Physics Education Research*, *8*(1), 010106. https://doi.org/10.1103/PhysRevSTPER.8.010106

National Academies of Sciences, Engineering, and Medicine. (2016). *Science literacy: Concepts, contexts, and consequences*. The National Academies Press.

National Research Council. (2005). *How students learn*. The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

National Research Council, Hawley, W. D., & Ready, T. (2003). *Measuring access to learning opportunities*. The National Academies Press.

Niaz, M. (1985). Evaluation of formal operational reasoning by Venezuelan freshmen students. *Research in Science & Technological Education*, *3*(1), 43–50.

Okasha, S. (2002). *Philosophy of science: A very short introduction* (Vol. 67). Oxford University Press.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning–a review of test instruments. *Educational Research and Evaluation*, *23*(3-4), 78–101.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64. https://doi.org/10.1177/01466216000241003

Petersen, C. I., Baepler, P., Beitz, A., Ching, P., Gorman, K. S., Neudauer, C. L., Rozaitis, W., Walker, J. D., & Wingert, D. (2020). The tyranny of content: "Content coverage" as a barrier to evidence-based teaching approaches and ways to overcome it. *CBE—Life Sciences Education*, *19*(2), ar17. https://doi.org/10.1187/cbe.19-04-0079

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development*, *15*(1), 1–12.

Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. University of Chicago Press.

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

Pluck, G., & Cerone, A. (2021). A demonstration of the positive manifold of cognitive test inter-correlations, and how it relates to general intelligence, modularity, and lexical knowledge. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43). Cognitive Science Society.

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Price, L. R. (2016). *Psychometric methods: Theory into practice*. Guilford Publications.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raven, J. C. (1990). *Advanced progressive matrices*. Oxford Psychological Press.

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Jackson, R. B. (2020). *Campbell biology*. Pearson.

Schneps, M., & Sadler, P. M. (1997). *Lessons from thin air. Minds of our own*. Annenberg Foundation.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001

Shayer, M., & Adey, P. S. (1993). Accelerating the development of formal thinking in middle and high school students IV: Three years after a two-year intervention. *Journal of Research in Science Teaching*, *30*(4), 351–366. https://doi.org/10.1002/tea.3660300404

Spearman, C. (1904). 'General intelligence', objectively determined and measured. *The American Journal of Psychology*, *15*, 201–293.

Stanovich, K. (2011). *Rationality and the reflective mind*. Oxford University Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St, B. T. Evans, & K. Frankish (Eds.), *Two minds: Dual processes and beyond* (pp. 55–88). Oxford University Press.

Tobin, K. G. (1993). *The practice of constructivism in science education*. Psychology Press.

Whewell, W. (1858). *Novum organon renovatum* (3d ed.). J. W. Parker and Son. [in English, with large additions].

Wiggins, G., Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Pearson.

Withgott, J., & Laposata, M. (2018). *Environment: The science behind the stories*. Pearson.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295.

Zelazo, P. D., Blair, C. B., & Willoughby, M. T. (2016). *Executive function: Implications for education [NCER 2017-2000]*. National Center for Education Research.

Zimmerman, C., & Klahr, D. (2018). Development of scientific thinking. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *4*, 1–25.

SUPPORTING INFORMATION

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

How to cite this article: Kalinowski, S. T., & Pelakh, A. (2023). A hypothetico-deductive theory of science and learning. *Journal of Research in Science Teaching*, 1–27. https://doi.org/10.1002/tea.21892