Theoretical model and quantitative assessment of scientific thinking and reasoning

Lei Bao,[1,*] Kathleen Koenig,[2] Yang Xiao,[3] Joseph Fritchman,[1] Shaona Zhou,[3] and Cheng Chen[4]

[1]*Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*
[2]*Department of Physics, University of Cincinnati, Cincinnati, Ohio 45220, USA*
[3]*School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, Guangdong 510006, China*
[4]*Teachers College, Jimei University, Xiamen, Fujian 361021, China*

(Received 2 November 2021; accepted 20 January 2022; published 23 February 2022)

Abilities in scientific thinking and reasoning have been emphasized as core areas of initiatives, such as the Next Generation Science Standards or the College Board Standards for College Success in Science, which focus on the skills the future will demand of today's students. Although there is rich literature on studies of how these abilities develop in students across grade levels, the research community has not reached consensus on their definition, modeling, or assessment. To advance research in this important area, a coherent theoretical model of scientific reasoning is needed for practically guiding instruction and assessment. For decades, the only instrument available for large-scale application was the Lawson's Classroom Test of Scientific Reasoning, but the instrument has demonstrated validity weaknesses and ceiling limitations, and its design is missing an explicit modeling framework for justifying the included skills. As a result, there is an urgent need for the development of a comprehensive modeling framework of scientific reasoning and a valid scientific reasoning assessment that targets the wide-ranging skills required for 21st century learners. This paper reports on the development of a modeling framework of scientific reasoning along with a new assessment instrument, adding to the research literature in a much needed area. The modeling framework integrates research in scientific and causal reasoning and operationally defines the skills and subskills that underlie the reasoning for knowledge development through scientific inquiry. Subsequently, this framework is used to guide the development of an assessment instrument on scientific reasoning. The validity and reliability of the instrument, which have been established based on large-scale testing, will also be discussed.

DOI: 10.1103/PhysRevPhysEducRes.18.010115

I. INTRODUCTION

The economy and future workforce call for a shift of education goals from content drilling towards fostering higher end skills including reasoning, creativity, and open problem solving [1]. In education of science, technology, engineering, and mathematics (STEM) initiatives on advancing 21st century learning, such as the Next Generation Science standards (NGSS) [2] or the College Board Standards for College Success in Science [3], focus on the skills the future will demand of today's students; i.e., both the STEM-disciplinary knowledge and attributes necessary to successfully contribute to the workforce and global economy [4–7]. Among the many skills emphasized in 21st century education, student abilities in scientific reasoning and critical thinking are the most commonly noted, which are highly connected with other cognitive skills needed for problem solving, decision making, and creative thinking [8–10]. As a result, they play a foundational role in defining, assessing, and developing the skills and learning outcomes emphasized in the 21st century science standards [2,11].

In the literature, there is extensive research on critical thinking [8,9,12–14], which is defined as the cognitive skills and strategies that aim for and support evidence-based decision making. It is the thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions [15,16], and is recognized as a way to understand and evaluate subject matter, produce reliable knowledge, and improve thinking itself [17,18].

Meanwhile, the notion of scientific thinking or reasoning is often used to label the set of skills that support critical thinking, problem solving, and creativity in STEM learning. In the literature, the terms scientific thinking and

*Corresponding author.
bao.15@osu.edu

scientific reasoning are often used interchangeably, and in this paper the term scientific reasoning will be used throughout. Broadly defined, scientific reasoning includes the thinking and reasoning skills involved in scientific inquiry for knowledge development and revision, such as the ability to systematically explore a problem, formulate and test hypotheses, manipulate and isolate variables, and observe and evaluate consequences [19,20].

Critical thinking and scientific reasoning share many common features, where both emphasize evidence-based decision making in multivariable causal conditions. Critical thinking can be promoted through the development of scientific reasoning in inquiry-based learning, which trains students' ability to identify a researchable question, formulate hypotheses, design and implement experiments, gather and analyze data, and evaluate the hypotheses. In this way, scientific reasoning can be viewed as a more domain-specific expression of critical thinking in the context of STEM learning. Therefore, targeting scientific reasoning in teaching is aligned with the goals emphasized in 21st century education and promoted through initiatives such as NGSS. Through development of scientific reasoning skills, student critical thinking, open-ended problem-solving abilities, and decision-making skills can be improved. The educational importance and benefits for students to develop scientific reasoning have been widely researched, which has documented favorable outcomes including positive correlation with course achievement [21–23], improvement on concept tests [24,25], engagement in higher levels of problem solving [26], and success on transfer to improve learning of STEM content [27,28].

Unfortunately, research has shown that college students lack essential skills in scientific reasoning, suggesting that these skills were not developed in K-12 or beyond. For example, Lawson [29] found that about half of introductory biology students lack the ability to develop hypotheses, control variables, and design experiments. Others demonstrated that undergraduates have difficulty making evidence-based decisions and differentiating between and linking evidence with claims [30–32]. In addition, research has shown that scientific reasoning skills are difficult to develop in traditional STEM curricula but can be effectively promoted with targeted inquiry-based instruction [20,33].

In order to develop scientific reasoning in formal and informal STEM education settings, it is important to have a guiding model and effective assessment tools to facilitate the teaching and evaluation of scientific reasoning under different educational settings. The research reported in this paper is about the development of a modeling framework as well as a new assessment instrument for scientific reasoning. The former is particularly important as it provides a theoretical foundation that can organize the concepts and learning outcomes emphasized in NGSS and the College Board Standards for College Success in Science into a coherent modeling framework and provide the cognitive underpinnings to support the teaching and learning of their respective goals. As an example, the 21st century skills emphasized in NGSS promote a cohesive understanding of science through three dimensions in science learning, including cross-cutting concepts, science and engineering practices, and disciplinary core ideas [2]. Although there are different views and arguments about the emphasis and organizing principles of NGSS [34,35], there are important concepts and reasoning skills that are commonly recognized as the learning outcomes for which scientific reasoning and causal explanation are highly emphasized. In addition, the modeling framework and the operationally defined reasoning skills can also provide practical means for the teaching and learning of the related concepts and skills emphasized in NGSS.

In terms of the assessment of scientific reasoning, the Lawson's Classroom Test of Scientific Reasoning (LCTSR) [36] has gained wide popularity in the STEM education community. Unfortunately, the design of the assessment lacks a theoretical framework other than the claim that the test is designed around formal operational reasoning, which was defined for purposes of the LCTSR to include abilities in control of variables, hypothesis testing, correlational thinking, probability, proportional reasoning, and conservation. In addition, a recent study, which thoroughly inspected the assessment features of LCTSR, has identified several validity weaknesses and a ceiling effect for college students [37]. Although Kalinowski and Willoughby addressed some of these issues through the design of an updated version of Lawson's instrument, which they refer to as the Montana State University Formal Reasoning Test (MSU-FORT) [23], they acknowledge that additional approaches are needed to define and measure scientific reasoning, and they call for a broader set of constructs to guide assessment development. Therefore, it is critical to develop a valid and updated assessment instrument on scientific reasoning that targets 21st century learners.

To contribute to the literature in the needed areas, this paper presents a new modeling framework along with the development and validation of a scientific reasoning assessment instrument that is grounded in this framework. In the following sections, a synthesis of existing models of scientific reasoning will lead to the development of the new model and the justification for the assessment design.

II. EXISTING MODELS OF SCIENTIFIC REASONING

Thinking and reasoning as a cognitive ability has been studied for many decades by psychologists and cognitive scientists. A comprehensive review of this area of work has been conducted by Zimmerman [19,38], including the landmark work by Piaget [39] on cognitive development to recent studies on reasoning in the STEM context by Lawson [40], Klahr [41], Kuhn [42], and many more.

In current education initiatives, scientific reasoning (or thinking) has been well established as a core ability for the 21st century learners.

Among the existing research, Lawson has done extensive work in assessment of scientific reasoning and in understanding how to teach these skills through inquiry-based science curricula [43,44]. Following Piaget's theory of formal reasoning and stages of development, Lawson identified 6 subskills as the basis for the assessment of scientific reasoning. Among these subskills, control of variables (COV) and hypothetico-deductive reasoning are highly emphasized as they provide the foundation for hypothesis testing, which is essential for scientific inquiry. In Lawson's studies, scientific reasoning is considered to play a central role in the generation of scientific knowledge. In his approach to science teaching, the scientific reasoning skills are incorporated into cycles of scientific inquiry, a process which has proven effective in helping students construct concepts and conceptual systems as well as develop more effective reasoning patterns [45].

In cognitive science, the research on thinking and reasoning is extensive. Within the topic of scientific reasoning, two threads of research are most related to this research, including Kuhn's research on multivariate causal inference and theory-evidence coordination [42] as well as Klahr's theoretical framework on scientific discovery as dual search (SDDS) and empirical studies regarding control of variable skills [41,46]. Both researchers have broadened the field of study on scientific reasoning to go beyond investigating student abilities in controlling variables and engaging in inductive causal inferences, which had previously been the primary areas of research.

Kuhn claims that scientific reasoning is a conscious and purposeful process for revising ideas and generating new understanding in light of evidence. This process, known as theory-evidence coordination [47,42], represents an integrative framework of reasoning that requires questioning existing theories, identifying alternative explanations, seeking and validating evidence (both supportive and contradictory), and evaluating and determining explanations based on evidence. New knowledge is constructed through the intersection of students' existing theories (including misconceptions), data-driven outcomes (covariation relations established through controlled experiments), and scientifically accepted theories. This coordination for building new knowledge involves a process for considering these types of evidence to form a network of meaningful connections among them and between evidence and explanations. It also involves the ability to consider the potential impact of unknown, but possible causal factors, as components and relations to form new evidence and explanations. These competencies are crucial as they represent the kinds of reasoning necessary for understanding the physical world as they allow for predictions, inferences, and explanations in cyclic inquiry processes. Kuhn's work emphasizes the multivariable nature of causal relations embedded within a variety of reasoning contexts and has demonstrated a lack of effective multivariable reasoning among children and post-college adults in coordinating evidence with explanations [42,48].

Klahr's research emphasizes the role of prior knowledge in scientific reasoning and provides a theoretical framework, the SDDS, for capturing and interpreting human behavior in a reasoning task [41]. The SDDS framework involves a hypothesis (theory) space, an experiment (data) space, and a set of possibilities for how the search in these two spaces are coordinated through evidence evaluation. The framework allows for movement back and forth between the hypothesis and experiment spaces based on students' prior knowledge, strategic preferences, evidence generated, and so on. In this way, the framework portrays the cognitive and developmental processes that scientists engage in for the purposes of generating new scientific knowledge, a process which is highly complex and does not necessarily proceed in a straightforward way.

The models proposed by Kuhn and Klahr provide important theoretical work for modeling scientific reasoning. Their work can be synthesized and combined with Lawson's research on subskills of scientific reasoning to form an operational framework for guiding the teaching and assessment of scientific reasoning. Here, we provide an example for how these ideas can be synthesized. Given a multivariable causal reasoning task, such as those found in Kuhn's study [42], students can begin their investigation from the theoretical side and identify possible explanations (hypothesized causal relations) with provided evidence. They can also approach their investigation from the experimental side based on a set of possible explanations to evaluate the consistencies between the evidence and explanations. Outcomes can then be used to propose new experiments or explanations through predictions and inferences. These pathways of exploration and discovery represent typical processes discussed in Klahr's SDDS framework [46], which also resonate with the hypothetico-deductive reasoning and inquiry activities emphasized in Lawson's learning cycles [44]. Regardless of where the process starts, which may include simultaneous thoughts from both sides, students need proficiency in moving between sides and in synthesizing all possible explanations and evidence to decide the best coordination outcome as their new understanding, which is the central element in Kuhn's work on theory-evidence coordination [42]. This process usually proceeds in multiple cyclic pathways that become the reasoning basis for what is commonly emphasized as the inquiry learning process. Subsequently, this process depends heavily on students' abilities associated with control of variables, data analytics, and causal decision-making. Building from these existing models, our work proposes an operationally defined framework of skills for the purposes of developing instruction and an

assessment instrument for scientific reasoning. The next section lays the foundation of this framework.

III. DEVELOPMENT OF A COMPREHENSIVE MODELING FRAMEWORK FOR SCIENTIFIC REASONING

From the existing literature on modeling scientific reasoning, there are several cognitive entities that are much related to and even entangled with the definition of scientific reasoning. These include scientific knowledge, scientific inquiry, and causal reasoning. The relations among these can be interpreted where scientific inquiry is a cognitive process supported by scientific reasoning to generate scientific knowledge which builds on causal relations. Therefore, both knowledge acquisition and reasoning have strong interactions in the cyclic process of inquiry.

Within the broadly defined scientific reasoning, causal reasoning appears to be a common and key element emphasized within existing models. However, causal reasoning has not been explicitly targeted in many studies on scientific reasoning, despite that most of the reasoning tasks on hypothesis testing in the related studies involve determining if an evidence-based causal relation exists in a multivariable setting [48,49]. As a result, the finer attributes of causal relations have not been explicitly addressed or integrated within current scientific reasoning models. Rather, studies on causal and scientific reasoning are somewhat developed as two independent tracks of work [47]. In addition, the causal reasoning involved is often broadly defined without attention to its specific properties and without explicit connections to scientific reasoning features and processes. Meanwhile, causal reasoning itself has been extensively studied in philosophy and cognitive science, generating a large body of work on its definitions and constructs. Since causal and scientific reasoning have substantial overlap for their functions and properties but have not been well connected in previous research, the following section takes on the important task of clearly identifying the roles and connections among the two schools of thinking. The section also integrates the research of causal reasoning with scientific reasoning studies.

A. Connections between causal and scientific reasoning

The relation between cause and effect is often considered the most fundamental component of thinking and reasoning for knowledge generation, especially in the science domains [50–54]. For example, conceptual knowledge is interpreted as an understanding of the essential parts and cause-effect relationships within a system [55] (p. 289). In addition, causal understanding is considered a reasoning primitive necessary for the development of naïve theories of physics, which enable very young children to understand the world around them [56,57]. Causality is thus integral to the development of knowledge in general.

1. Definition of causal relation

The definition of causality has a long history [58,50]. In more recent studies, three elements are explicitly emphasized for establishing a causal relation [51,59–63]. These include (i) the element of time order of potential cause and effect, where a cause must necessarily precede an effect temporally, (ii) the covariation element, which describes the quantitative covariation relations between variables of cause and effect that are typically established based on experimental observations of events, and (iii) the mechanism element. The latter refers to mechanistic understandings or models of processes linking a cause to its effect, often at a smaller scale or deeper level of a theory. Together, these three elements, which are explicitly defined and investigated in two recent studies [62,63], form the basis of causal reasoning. Typically, the time element is the *a priori* condition for a causal relation and is moderately established among college students [62,63]. In this paper, the focus is on the covariation and mechanism elements of reasoning given that a complete understanding of a causal relation is best established with both. However, in the process of scientific inquiry, it is common for covariation (experimental evidence) or mechanism (conceptual theorization) to take a temporary lead in the progression of knowledge development.

From the philosophical and epistemological perspective, humans are observers and make observation-based inferences on how certain events may be explained and predicted [62]. Such explanations and predictions are constructed with understanding of the underlying causal and noncausal relations. In general, the temporal element of causality is the foundation in defining a causal relation, which gives rise to the time-evolution covariations of temporal events. The understanding of a covariation process often starts with observing the initial and final states of the process and the changes between the states. Based on these observations, consistent trends and patterns are identified to form covariation relations, which can be further generalized to make inferences on the possible mechanistic processes underlying the changes. In knowledge formation, the observations of the changes between the initial and final states produce the understanding of covariation relations, while the inferences on the mechanistic processes lead to the understanding of mechanisms causing the covariations.

Figure 1 illustrates the three elements and the related processes involved in developing a causal understanding. Typically, people start the learning process by observing time-ordered events occurring in multivariable contexts. The observations produce descriptive data of the covariation behaviors, which can be further processed to extract

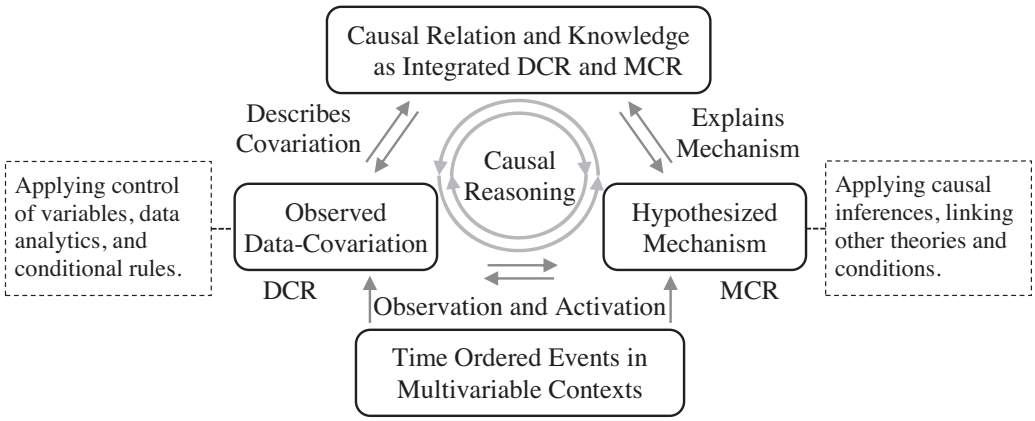


FIG. 1. Essential elements and processes contributing to developing an understanding of a causal relation, which describes the process underlying time ordered events. Through observation, covariation data patterns can be generalized to provide empirical evidence of causation in terms of the covariation attributes. Meanwhile, the hypothesized mechanism can provide mechanistic explanations of why and how the causal relation and the covariation may exist. Both are needed to form a complete understanding of a causal relation.

specific data patterns based on experimental conditions, such as control of variables. Valid covariation data can then be used to make inferences on the underlying mathematical and logical relations as well as possible mechanisms that can be used to explain the covariation data patterns and to make predictions on outcomes in extended contexts. Meanwhile, if the events and contexts are familiar, they can also activate a learner's prior knowledge for explaining the observations or cue additional processing when revisions of certain understandings are needed. Then, through cycles of coordination and integration of both covariation data and mechanistic explanations, a synthesized understanding of causality of the targeted events can be produced and integrated into one's knowledge system.

Regarding the connections between causal and scientific reasoning, the existing models on scientific reasoning often make emphasis on covariation evidence in determining a causal relation [64,38], where evidence-based justifications are considered superior to theory-based justifications. Meanwhile, the importance of reasoning in causal mechanism has also been considered and studied [65], where learners were thought to incorporate information about both covariation and causal mechanism in their reasoning. In the work described in this paper, the integration of scientific and causal reasoning will emphasize both covariation and mechanism as a common foundation for modeling causal reasoning.

Synthesizing the literature and discussion here, the concept of causality can be operationally defined in terms of three essential components, including time order, data covariation, and mechanism. Regarding causal reasoning, once a temporal process is established, causal relations can be explored and determined through two reasoning pathways. One path goes through analyzing covariation data patterns to determine the data-covariation relations (DCR). The other explores mechanism-based explanations that connect cause and effect through a set of hypothetical conceptual (theoretical) claims and mathematical logical relations. These are referred to as mechanistic causal relations (MCR). DCRs provide the covariation patterns to imply or justify causality among concerned variables but do not (or lack the needed mechanistic understanding to) explain why and how certain causes may lead to the observed effects. The latter is the function of MCRs. In causal reasoning, DCRs provide the evidence for validating a hypothetical mechanism of a causal relation. Meanwhile, MCRs provide the explanatory mechanisms for how and why certain variables cause outcomes under certain conditions. The mechanisms can be purely hypothetical without any existing data-covariation evidence, such as a proposed new theory. On the other hand, they can be generalized and tested based on collections of data-covariation evidence, which is the process of inductive theorization and experimental testing of a theory or hypothesis.

2. Inquiry-based knowledge generation through scientific and causal reasoning

In the scientific inquiry process of knowledge generation, DCRs constitute most of the experimental evidence, while MCRs establish the basic components of the theoretical (conceptual) understanding. From the epistemological perspective, a DCR is an observation-based description of a possible causal phenomenon, while an MCR is a hypothetical mechanistic explanation of the underlying causal mechanism. The two processes and their outcomes are coordinated by learners to validate, revise, and develop each other in order to advance both experimental and theoretical understandings of a specific knowledge domain. A complete understanding of a causal relation for a specific topic would require both data-covariation evidence and mechanistic explanations, which provide the basic constructs of scientific knowledge. In most scientific areas, however, this kind of understanding is a moving

progression such that both experimental and theoretical studies develop simultaneously but in intertwined pathways advancing one after the other.

There are a few unique features contrasting the roles and functions of the two aspects of causal relations in knowledge development. DCRs are based on data patterns without explanatory understanding. Therefore, in their original forms they require a large amount of memory to store the wide-ranging patterns manifested by the causal relation in different conditions and contexts. As a result, it is typically inefficient to recall and transfer such knowledge and understanding. It is also difficult to predict outcomes in conditions and contexts out of the range of experimentally confirmed domains. In contrast, an MCR is often encoded with a simple explanatory rule, which is generalized based on many experimentally confirmed DCRs. For example, in the case of forces between electric charges, many experiments have been conducted to determine the DCRs at different conditions, including types of charges and distances between charges. The collection of DCRs are later generalized into a hypothetical relation, $F = kq_1q_2/r^2$, which is explained with a mechanistic understanding, claiming that interaction forces exist between two charges and their magnitudes follow this simple relation. The equation and its mechanistic explanation form the MCR for forces between charges. This MCR can then be applied to a wide range of conditions and contexts involving multiple charges and distances for accurate predictions or calculations of forces, most of which may have never been experimentally measured.

Obviously, generalization of DCRs to form MCRs significantly reduces the cognitive resources needed to encode such relations and allows this type of knowledge to be easily stored and readily transferred and applied to extended domains of contexts. In education, an MCR is also easier to be taught and learned between people and documented for future generations' learning. Therefore, much of what is typically defined as scientific knowledge, which is accumulated from prior scientific development, has its basis in the form of MCRs. Meanwhile, DCRs provide the experimental (observational) evidence to confirm hypothetical MCRs (hypotheses), to revise and further validate existing MCRs, and to generalize new MCRs as improved or new knowledge. Here, it is important to note that an MCR can include generalized mathematical and logical relations as well as the mechanisms that explain the mechanistic origins of the relations. There can also be situations in which ideas on mechanism and mathematical and logical relations exist before observable DCRs are available. These ideas represent hypothetical MCRs, which are commonly referred to as theoretical hypotheses and need to be validated by experiments to obtain related DCRs.

In developing DCRs, covariation data can also be processed and generalized to yield mathematical relations using analytic and modeling algorithms. Therefore, both MCRs and DCRs can include mathematical and logical relations, and containing them is not a feature that distinguishes between DCRs and MCRs. However, it is also helpful to compare the differences between DCR-based and MCR-based mathematical and logical relations. Those that are DCR-based typically represent local computation modeling (e.g., regression) outcomes of specific DCRs, which cannot be generalized beyond the specific context domain. In addition, these relations are not backed by mechanistic explanations, which further limit their general application. Because of the lack of supporting mechanism, DCR-based mathematical and logical relations are not mechanistically meaningful, making them difficult to be theoretically manipulated. Through accumulation of DCRs from a wide range of contexts, the involved mathematical and logical relations can be further validated and synthesized with mechanistic hypothesis to develop MCRs. When the DCRs and MCRs are validated to be in agreement, the mathematical and logical relations developed based on DCRs can be transformed into those that are MCR based, which are mechanistically explained and can be theoretically manipulated and generally applied as laws and principles.

3. A new comprehensive framework for modeling scientific reasoning

In the inquiry process of learning, both DCRs and MCRs are important building blocks for knowledge formation. As a result, the generation of knowledge can be modeled as a process for coordinating the DCRs and MCRs to construct understanding that is experimentally validated and mechanistically plausible. This process has a number of aspects as it often involves contexts with multiple variables and coordination processes between DCRs and MCRs to form a more consistent and synthesized causal understanding. Moving forward, this reasoning and knowledge formation process will be referred to as the data-covariation and mechanistic causal reasoning (DMCR) framework, which is based on a synthesis of the work described earlier. Since the DMCR process coordinates between DCRs and MCRs, it can be considered as a dual-pathway process for operationally modeling the reasoning involved in causal decision making and knowledge formation. From this perspective, the DMCR framework resonates with both Klahr's SDDS framework and Kuhn's theory-evidence coordination model. However, in the DMCR framework, constructs and functional processes of causal relations and causal reasoning are explicitly and operationally defined. In particular, reasoning is considered a process that directly targets a variety of types of relations in knowledge development, which will be discussed in detail in the following sections.

Putting all components together, Fig. 2 shows a schematic of a multivariable causal network and the associated reasoning processes for causal decision making and

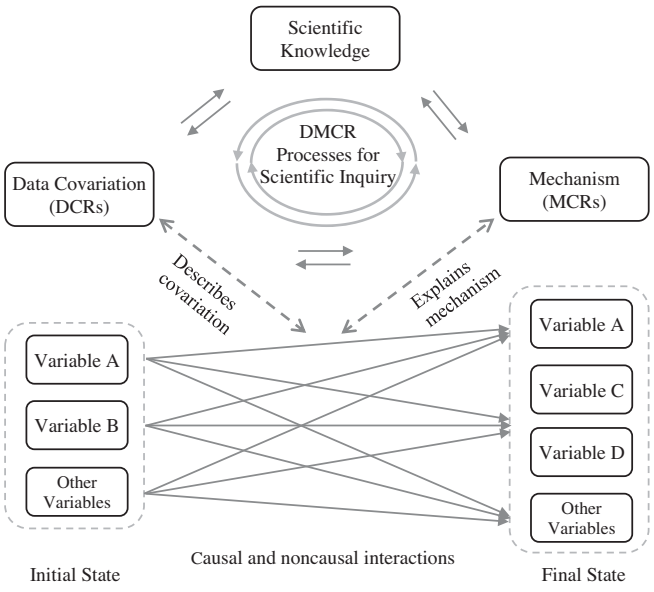


FIG. 2. Schematic of the DMCR framework for scientific reasoning to support causal decision making and knowledge formation. The variables are illustrative and do not represent any specific examples. Variable A represents features that exist in both initial and final states, which may have changed. Variable B exists only in the initial state, while variables C and D appear only in the final state. Other variables include those that are either controlled, ignored, or hidden (unknown). The solid arrows connecting the variables in the initial and final states represent the possible known and unknown temporal evolution and contextual interactions of the variables. The dashed arrows represent observations that contribute to developing MCRs and DCRs.

knowledge generation. The figure illustrates the relations among scientific knowledge, scientific inquiry, and causal relations and reasoning. In this representation of the DMCR framework, the notion of scientific reasoning can be interpreted as an umbrella definition which encapsulates all of the functions and processes that support causal reasoning, inquiry, and knowledge formation.

Since causal reasoning plays an essential role in inquiry and knowledge development, the framework gives explicit attention to its structures and processes, which are shown as a causal network in Fig. 2. The causal network represents generic time ordered causal events, which are described with an initial state, a final state, and processes connecting the initial and final states. The initial state contains the variables that constitute possible causal and noncausal factors, while the final state contains variables representing the outcomes of causal and noncausal processes. The processes connecting the initial and final state describe the mechanistic interactions and conditions that determine the possible causal and noncausal dynamics. Within the causal network illustrated in Fig. 2, variables in the initial state can be causal and noncausal, while variables in the final state can be consequential and nonconsequential. There are typically other variables involved in the process that are controlled, undetermined, and or hidden. In addition, the sets of variables describing the initial and final states are not necessarily identical due to conditions and constraints on measurements and environmental influences.

From the initial to the final state, collected data that describe the states and their changes under controlled conditions can be analyzed to identify DCRs. Here, DCRs are determined based on data that describe the covariation of variables from the initial state to the final state. However, DCRs do not describe the processes linking the initial and final state. These processes, which explain how and why the variables may covary, are described and explained by MCRs with causal mechanisms. Using the model of the casual relations shown in Fig. 2, theory and hypothesis can be considered as mechanism-based claims. This is because they mechanistically explain the structural and temporal network of causal relations connecting the concerned variables, which can be experimentally observed and quantitatively described in terms of DCRs. A complete causal theory is best supported with both DCRs and MCRs in integrated networks of variables and relations, which are referred to as DMCRs and are developed through causal reasoning processes for coordinating and restructuring the DCRs and MCRs. DMCRs represent fully developed causal relations that constitute the basis of a theory or a piece of established scientific knowledge in a specific content domain. For cross-cutting concepts and theories that span multiple knowledge domains, these often establish as integrated networks of multidomain DMCRs.

For example, the classical theory of gravitational forces involves several variables including two objects with mass, the distance between the two objects, and the observed outcomes of the gravitational interaction between the two objects in terms of forces. The covariation relations among the three variables can be experimentally determined as DCRs and generalized to form a mathematical relation, $F = Gm_1m_2/r^2$, which is mechanistically explained (MCR) such that the gravitational interaction between masses forms the origin of the observed forces. Taken together, the DCR and MCR form a complete descriptive and explanatory framework for the classical theory of gravity.

For scientific knowledge, most of the existing theories and hypotheses can be broken down into networks of causal relations among involved variables. Being a theory, it would involve causal mechanisms to explain the nature of the covariation relations (DCRs). For the example of gravity, the mechanism assumes that the gravitational interaction between masses leads to the presence of gravitational forces. However, it is still at best an observation-based hypothetical inference for which the actual deeper level mechanism is still not understood. From a philosophical point of view, the ultimate mechanism may never be understood or reached by an observer.

It is worth noting that in most cases DCRs are determined with covariation changes between the initial and final states, while MCRs explain the mechanistic processes that transform the system from the initial state to the final state. However, it is possible that in certain situations DCRs may be identified to describe the quantitative distributions of the processes connecting the initial and final states. Similarly, MCRs may also be established as the mechanisms underlying the identified variables of the initial and final states. These additional properties are considered when the causal network of a particular domain needs to be fundamentally restructured or extended to other networks, which is beyond the scope of this paper. Here, the focus of the discussion is on the primary functions of DCRs and MCRs in an established causal network of a specific knowledge domain.

The multivariate causal network in Fig. 2 also resembles the structure of a Bayesian network, which is commonly used in determining the probabilistic features of causal attributions. In quantitative causal decision making, Bayesian probabilities play a central role in drawing evidence-based conclusions. Therefore, understanding and reasoning with multivariate causal networks and Bayesian probabilities are included as key skills in the assessment of scientific reasoning, which will be discussed later in this paper.

In the following sections, a number of fundamental processes and elements that support the DCR and MCR reasoning pathways will be defined. Together, these form the theoretical basis of the DMCR framework for modeling scientific reasoning and knowledge development. This level of finer detail is necessary for the creation of an assessment framework to inform the related assessment design.

B. Complexity of causal networks and data-covariation relations

The complexity of a causal relation must be clearly mapped out to operationally support the development of an effective assessment. Using the causal network representation, the complexity of a causal relation can be modeled with the structure of its causal network. For a network of variables and connecting relations, two types of complexity can be considered. The first type is due to the network's structure, which primarily describes features of DCRs. Here, the complexity typically increases with the number of variables' interconnections.

The second type is the conceptual and computational complexity of individual variables and interactions within a network, which represent features of MCRs. For example, relations between two variables can be certain or uncertain. The involved mathematical nature can be simple, such as linear relations, moderately complex, such as quadratic and other nonlinear functions, or complex, such as recursive and noncontinuous. For a specific example in physics, consider the MCR of mechanical energy conservation. In classical mechanics, conservation of mechanical energy is expressed as a simple summation of classically defined kinetic and potential energies, while in quantum mechanics it is expressed within the Schrödinger equation applied to a wave function that describes the probabilistic nature of reality. The conceptual and computational complexities of the two explanatory mechanisms are vastly different, although the general ideas of the two are analogous within their respective domains.

In addition, the explicitness of variables and relations also plays an important role. Reasoning tasks with implicit (or hidden) variables and relations are usually more difficult than tasks in which all variables are explicitly provided with relations obviously indicated. The identification of explicit or implicit variables will typically involve both DCR and MCR types of reasoning. The implication on a possible variable often requires a hypothetical idea of mechanism for why and how such a variable may contribute to the outcome, as well as some existing or predicted covariation phenomenon that can be used to test the hypothesis. Therefore, in assessment design, manipulating the explicitness of variables and relations in a task can help to meaningfully control the difficulty level of test items. The design of items using hidden variables and mechanisms can provide a useful method to assess students' comprehensive reasoning that requires both MCRs and DCRs.

For our purposes, both types of complexity will be used in assessment design, which will be discussed in detail in the assessment section of this paper. The DCR-based complexity is manipulated by varying the structures of multivariable causal networks, while the MCR-based complexity is manipulated by involving different configurations of hidden variables and mechanisms as well as conditional settings in causal tasks. To help illustrate the complexity of causal relations, examples of typical causal network structures are reviewed next.

From simple to complex, the structure of a causal network can be represented in terms of different numbers and types of connections between covariation variables. A representative list may include (i) bivariate relation, (ii) multivariate relations, (iii) multivariate relational network (various levels as well), and (iv) complex systems of connected networks with complex coupling and feedback, in which nonlinear dynamics and uncertain phenomenon such as chaotic behaviors may exist in certain conditions. Several generic examples of these different types of causal networks are shown in Fig. 3.

In a research design to measure covariation data for DCR analysis, the common categories of variables often include independent variables (IV), dependent variables (DV), controlled variables (CV), mediating or intervening variables (MV), and environmental variables (EV) that are not controlled and can contribute to the covariation outcomes. In a specific design, these variables can also be explicit or implicit (hidden) with known or unknown mechanisms and/or covariations (or correlations).

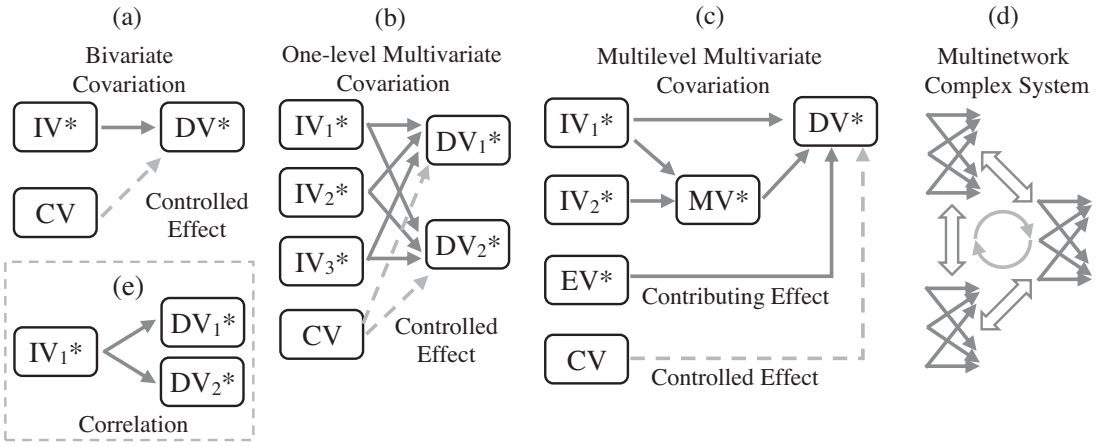


FIG. 3. Causal networks for relations among different variables. Variables marked with an asterisk (*) represent the ones that are varied and may exhibit covariation or correlation relations. The solid lines represent covariation relations and dashed lines represent controlled effects that do not impact the covariation. (a) bivariate relational network, (b) one-level multivariate relational network, (c) multilevel multivariate relational network, (d) a complex system of multiple interconnected networks with cross-coupling and recursive feedback. Note that (e) is a special case representing a correlation between DV_1 and DV_2.

For a DCR type covariation causal relation, IVs are typically studied as the hypothetical causes while DVs are the outcomes of the causes. A well-established IV can be manipulated to change in specific patterns, which may cause the DV to co-vary accordingly to form a covariation relation. Here, a necessary condition for covariation is the control of variables, without which the co-varying data pattern can only be interpreted as correlation instead of covariation [see Fig. 3(e) as an example]. Ideally, an IV should not be the DV of deeper unknown variables. However, this assumption usually cannot be achieved from the philosophical perspective but can be operationally controlled such that no known variables become the obvious causing variable of the IV.

Controlled variables can impact the DV and, therefore, should be held constant when the IV is varied. This is an essential concept in constructing covariation experiments. A typical flaw in drawing conclusions for covariation relations is the confusion between covariation and correlation [see Figs. 3(a) and 3(e) for a comparison of covariation and correlation]. Correlation is a covarying relation between two DVs with known or unknown IVs or between an IV and a DV without control of variables when additional variables may be contributing. A correlation cannot be interpreted as covariation and does not imply a causal relation. Simply being covarying does not warrant a covariation relation. To establish a covariation situation, the experiment must include designs with control of variables, with IVs being manipulated, and with DVs covaried. When the conditions are all satisfied, the covarying data patterns can then be validated as DCRs between IVs and DVs [see Fig. 3(b) for a simple multivariable situation].

For more complex situations [Fig. 3(c)], there can exist mediating or intervening variables acting as a middle layer linking the IVs and DVs. These MVs can also be latent and unmeasurable with the given technology. In addition, there can also be a range of extraneous variables innate to the experimental environment, which cannot be controlled but can influence various variables in the design. It is typical that this influence be kept at a small level or accounted for in the analysis in order to establish the validity of the measured covariation relations. Since these variables contribute to the variation of many involved variables, they are referred to as environmental contributing variables.

The most complex causal network is a system of interconnected networks with recursive feedback [Fig. 3(d)]. This kind of structure forms a complex system that can exhibit nonlinear or even chaotic behaviors. Within the complex system, each connected network maintains its relational patterns, which are also influenced by the inputs and outputs of other connected networks that can exist in multiple temporal recursive forms. In this situation, microscopic and macroscopic behaviors are often related in complex interactions that cannot be readily predicted or inferred.

C. Complexity of causal reasoning processes

When faced with a task for determining causal relations, a student's reasoning can occur through simple to complex processes. The initial simple reasoning often starts with identifying the relevant variables based on the features of the context. Relations among the variables are then developed by recalling similar existing understandings or generating hypothetical ones based on cued domain-relevant understandings. The relations exist from simple bivariate to complex multivariate forms and can be causal and noncausal, but in this work the causal relations are emphasized. Typically, bivariate relations are developed first as these are simple to validate. Next, a cluster of relevant bivariate

relations can be integrated to develop multivariate relations among multiple variables (IV/MV/CV) and outcomes (DV/MV). Further, multivariate relations can be integrated to construct networks of relations connecting multiple sets of multivariate relations and cross domain networks to form an increasingly complex web of relational networks [see, e.g., Fig. 3(d)], which can eventually evolve into a nonlinearly coupled complex system. Summarizing the reasoning processes and the corresponding structures of the causal networks, five general levels of complexity can be defined, as shown in Fig. 4.

The levels of reasoning complexity resonate with the theories in knowledge development, especially the knowledge integration models on learning domain-specific content [63,66,67]. For example, the taxonomy of structure of the observed learning outcomes (SOLO) [66] models a student's knowledge structure in five progression levels of complexity of connections including prestructural, unistructural, multistructural, relational, and extended abstract, all of which share structural similarities with the complexity levels discussed above. This model provides a broadly defined progression on the structure of learners' knowledge without attention to specific reasoning details. In our work, emphasis is placed on the specific reasoning skills that underlie knowledge development, which will also be applied to designing an assessment of the reasoning skills.

It is also important to note that the different levels illustrated in Fig. 4 do not represent a strict progression of development. In learning and problem solving, the actual reasoning processes often occur in parallel at multiple levels with substantial interactions in between. Regarding the learning of a specific content topic, student reasoning may behave with a trend of development from simple to more complex levels. However, branching out and recursive processes are common. For example, the very process of identifying variables will cue and interact in parallel with the relevant relations associated with the concerned variables. When the identified variables and relations do not form a satisfactory understanding or explanation for the task context, additional cycles of identification and evaluation will be conducted to achieve a better match between the constructed understanding and the task goal. Therefore, this kind of reasoning often occurs in multiple recursive loops at all levels of knowledge construction.

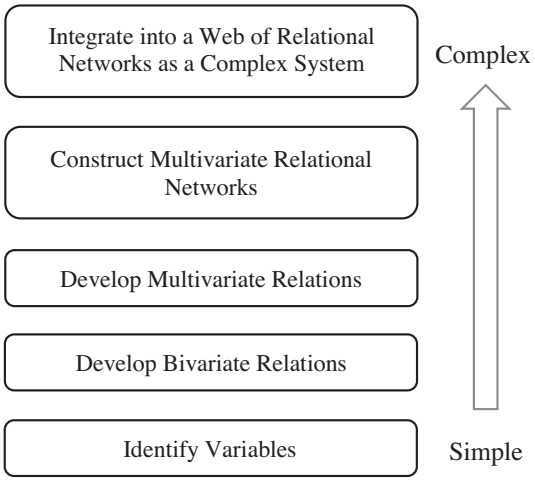


FIG. 4. Complexity of causal reasoning processes.

D. Operational processes of reasoning for knowledge generation

This section discusses the fine-grained operational reasoning functions necessary for problem solving and scientific inquiry, where different reasoning processes in the DMCR framework provide the fundamental cognitive support for knowledge development. For example, the hypothetico-deductive model [68,69], which describes similar reasoning processes, is widely accepted as being central to scientific inquiry and learning. In order to operationally model the reasoning underlying the DMCR framework, five general processes and operations are defined, including "I-process," "D-process," "evaluation-analysis," and "loop." Together, these form a concrete, functional basis for performing DMCR reasoning in specific tasks, such as those shown in Figs. 2 and 3. This level of detail is necessary to support the development of an assessment framework for scientific reasoning.

The I-process represents generalized inductive type reasoning such as the functions to induce, infer, imply, identify, etc. It is a process of creating or searching for new elements to be added to the current reasoning. The results of the I-process include a wide range of constructs, such as possible variables, relations, and mechanisms (MCR hypotheses), which are often new or unknown to the learner and can even be non-existing (new to the world). The validity, plausibility, and usefulness of the products of an I-process are often uncertain and need to be evaluated or validated through other processes.

The D-process represents generalized deductive type reasoning, such as the functions to deduce, derive, deploy, etc. It is a process that incorporates (plugs in) the contextual features (variables) into a given (existing) set of rules or functions to generate defined outcomes. Results of the D-process are often "determined." That is, although a result could be unknown to an individual, it is conceptually, mathematically, and logically warranted based on the given rules or functions.

In inquiry-based learning, the D-process usually operates with the elements created by the I-process to derive new predicted results, which can be further processed to evaluate the validity of the I-process outcomes. The evaluation-analysis (EA) processes serve to analyze and compare the outcomes of the I- and D-process in the context of the task and generate evidence-based decisions for the agreement between the outcomes and task goals. This kind of process typically goes through multiple cyclic

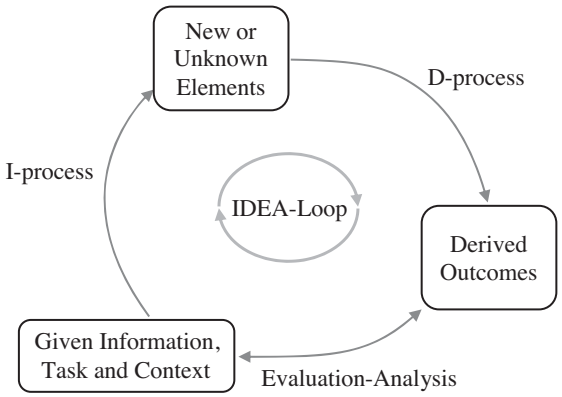


FIG. 5. A conceptual diagram of the IDEA-Loop model of reasoning functions.

loops. Thus, the complete process can be operationally understood as cycles of the inductive-deductive-evaluative-analytic-loop, which will be referred to as the IDEA-Loop model of reasoning moving forward in this paper.

Figure 5 shows a schematic diagram of the IDEA-Loop model of reasoning functions. The tasks, contexts, created elements, and outcomes of IDEA-Loop can be components or processes in all levels of cognitive operation. At the neural computing level, these can represent activation of clusters of neural networks and their inputs and outputs. At the behavioral level, these can represent observable cognitive states such as proposed hypotheses and predicted or derived outcomes from applying certain rules and principles in contexts.

In practical applications, certain parts of IDEA-Loop may serve as the primary function. For example, in problem solving that involves simple plug and chug, the D-process is often the primary operation followed with EA validation. However, when a task involves the I-process, the whole chain of IDEA-Loop is typically activated for the purpose of validating the I-process outcomes. That is, since the I-process create new elements, it naturally activates the D-process to apply the new elements to generate predicted outcomes, which then go through the EA process for validation of the newly created elements based on comparisons between predicted and observed or expected outcomes. If revisions are needed, further cycles of IDEA-Loop will be activated.

In teaching and learning, it is also possible to design tasks, in which parts of the DEA processes can be specifically inhibited. For example, a task focusing on the I-process may ask students to analyze a problem without solving it to search for a related principle as a proposed problem-solving strategy. It can be argued, however, that while conducting the inductive search and formulation, the student may still engage in some level of IDEA-Loop either explicitly or implicitly. This is because one would need to look for something that is plausible in a search, which means the DEA processes would be engaged throughout one's reasoning for validation and decision of outcomes from a search, creation, or prediction. In addition, the results of the I- and D-process can be in all forms of cognitive constructs at all levels, such as variables, relations, theories, hypotheses, new contexts, new knowledge domains, etc. Therefore, IDEA-Loop would occur among all levels of reasoning on elements with wide ranging complexity and abstraction.

At the behavioral level, the IDEA-Loop model of reasoning can be compared with several related models including hypothetico-deductive reasoning [68,69], theory-evidence coordination [42], and SDDS [46]. For the most part, the existing models bear much similarity on the general processes of reasoning and their cognitive products, such as determining valid evidence and testing a hypothesis. On the other hand, the IDEA-Loop model provides operational definitions of the related reasoning with finer elemental functions, which can directly inform the design of assessment and instruction. In our work, these functional elements will be extracted as assessment attributes for measurement of scientific reasoning skills, which are discussed later. The connections between IDEA-Loop and the existing models are discussed next.

The notion of hypothetico-deductive reasoning is a proposed description of scientific method in an inquiry process that proceeds by formulating a hypothesis in a form that can be tested based on observable data [68]. The reasoning aspects of this model have been studied by Lawson in terms of a set of scientific reasoning skills required for hypothesis testing. In assessment of these skills, students are often given experimental data with the task goal to identify a hypothetical causal mechanism that can produce outcomes consistent with the data (e.g., see questions 21–24 in the LCTSR). In this case, the I-process is primarily inductive thinking to identify or form a hypothesis based on given contexts and conditions. The identified hypothesis is then applied through the D-process to generate predicted data, which is further compared with the given data in the task for evaluation and analysis of the validity of the hypothesis.

Here, the hypothesis generation-identification (hypothetical part) can be considered as a special case of the I-process for producing a hypothesis, while the deductive reasoning can be considered as a special case of the D-process for producing a prediction. The comparison and validation are part of the EA processes. Therefore, the hypothetico-deductive reasoning can be represented as special-case processes of the IDEA-Loop model [Fig. 6(a)]. Comparing the two models, the IDEA-Loop model is more general and flexible than the hypothetico-deductive model. The I-process represents a generic creative process targeting many types of constructs rather than being limited to the hypothetical explanations in the hypothetico-deductive model. In addition, the IDEA-Loop model provides concrete definitions of

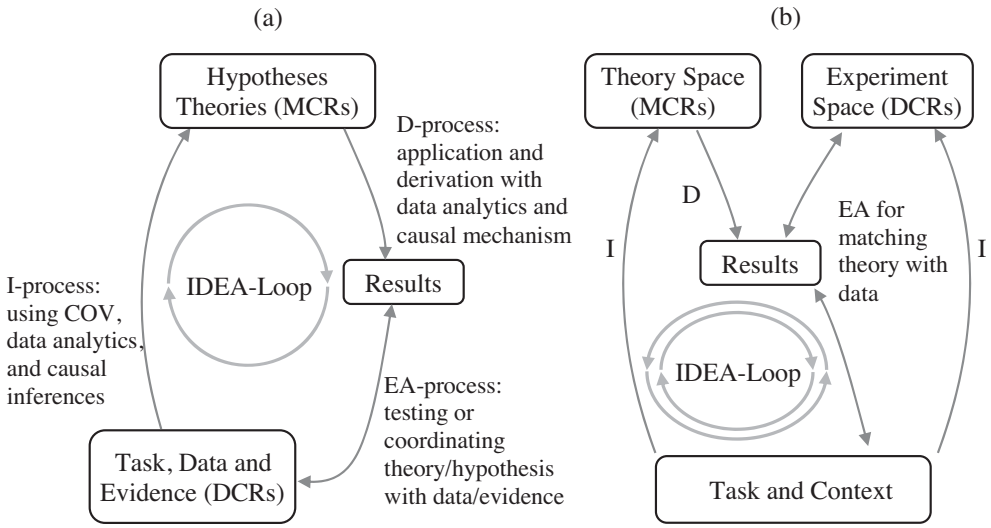


FIG. 6. Special cases of the IDEA-Loop reasoning model. (a) IDEA-Loop for hypothetical deductive reasoning and theory-evidence coordination. (b) SDDS as multiple parallel nested IDEA-Loop cycles. Double-headed arrows represent the interconnected dual-space cyclic pathways, which occur in both directions.

all the involved functions and processes, which are not explicitly defined in the hypothetico-deductive model.

For the theory-evidence coordination model, the central process is to achieve consistency between evidence (in terms of given or collected data) and a hypothesis or theory. The task usually involves matching datasets to given hypotheses or developing or revising a hypothesis to match data. These processes are similar to those in hypothetico-deductive reasoning, and therefore, can also be represented with IDEA-Loop as various types of contextualized applications, but may involve multiple cycles of IDEA-Loop at different processing levels depending on the tasks and contexts [Fig. 6(a)].

The SDDS model can be considered as another variation with the emphasis on reasoning occurring in both the experimental and theoretical pathways, and therefore, has a dual-space structure. The primary process is to search in both spaces to identify a consistent match between data and theory. This is also similar to theory-evidence coordination and hypothetico-deductive reasoning such that possible theories (MCRs) are identified through the I-process and applied through the D-process to produce outcomes comparable to experimental data. Meanwhile, the relevant data-covariation relations are also searched through a parallel I-process in the experimental space and compared with the predicted outcomes from the identified hypothesis in the theoretical space. Evaluation and analysis processes are conducted to determine whether the theory and/or the experimental data are consistent and if further cycles are needed. A unique emphasis in SDDS is that the searches in experimental and theoretical spaces often occur in multiple cycles, which can be represented with multiple parallel processes of IDEA-Loop [see Fig. 6(b)].

In summary, the DMCR model of scientific reasoning integrates causal reasoning by defining DCRs and MCRs as the causal basis for developing scientific knowledge, which is functionally modeled with the IDEA-Loop processes. Together, these can provide an operational framework with concretely defined functions to represent the existing models of scientific reasoning including the hypothetico-deductive, theory-evidence coordination, and SDDS models. The operational definitions of the specific scientific reasoning skills and functions can provide further utility on guiding assessment design.

E. Developmental progression of reasoning and knowledge formation

One of the challenges in developing an assessment framework of scientific reasoning is understanding what to include based on the level of students described by the framework, which for our purposes includes high school and college students. This section provides possible explanations for some of the difficulties in reasoning observed in high school and college students, while also providing justification for which reasoning tasks to include on an assessment of scientific reasoning.

Based on the work in developmental psychology [39], in early stages of development, children observe the environment around them and develop an understanding of the world. This knowledge is primarily in the form of observed data along with simple generalizations of covariation patterns of objects and events. Therefore, at this stage of development, knowledge and the associated reasoning are mostly in the form of data-covariation casual relations of basic daily life events and phenomena. As children's language develops, linguistic descriptions and explanations

of mechanistic causal relations start to develop and can also be directly taught to children without the need for them to make their own observation and generalization. This can provide efficiency in learning established knowledge, especially on things that cannot be conveniently experienced or observed. At this stage, the development of knowledge and reasoning mostly involves learning MCRs. The variables and relational networks of domain-specific knowledge are often memorized as facts, and children's reasoning is trained to process and validate the logical and computational consistency of the involved relations. Some of the reasoning functions are domain general and can be applied in other domains to form different sets of domain-specific knowledge.

As children grow with more experience from their real-world environment, as well as formal and informal learning, they also develop their own versions of MCRs about the world. For instance, most students develop their own conceptions in physics, such as believing a force is needed to maintain motion. These naïve conceptions are developed intuitively from experience of the physical world (which creates DCRs) and attempts to explain the experience (which create MCRs). For example, a common misconception on force and motion attributes an applied force as the cause of an object's motion. It provides satisfactory explanations in the world with friction, which is often a latent mediating factor not explicitly or correctly interpreted by most non-Newtonian thinkers. Without explicitly including friction, the force-motion relation is difficult to be generalized into the Newtonian understanding. This example shows that learners naturally develop both DCRs and MCRs as part of their knowledge system and they tend to generalize their own MCRs, which become a meaningful part of their understanding of the real world. As development continues, these DCRs and MCRs can be further integrated to form more complex networks of casual understanding (DMCR Net).

Meanwhile, in the current formal education system, students obtain a large portion of their knowledge of science through formal education, in which the teaching and learning focuses on delivering established scientific knowledge (i.e., scientifically validated MCRs). Through many years of such schooling, many students learn to rely on memorization, with reasoning skills trained primarily as incorporating, manipulating, and evaluating the computational and logical relations provided within the MCRs. Many students rarely experience the inquiry type of learning that requires them making their own observations to develop DCRs from data, generalizing DCRs to form hypotheses (MCRs), and evaluating the DCRs with the MCRs to test hypotheses. This is likely an important factor that leads students to inappropriately use prior knowledge or belief biases, instead of drawing evidence-based conclusions from data, on inquiry-based learning tasks. In a way, the traditional education strategies train students to learn in modes involving passively receiving and memorizing information. This can lead them away from the natural curiosity of young children who would otherwise observe the world and develop their own observation-based conceptual understanding.

In Fig. 7, a flowchart is shown to depict the representative features of the developmental pathways of

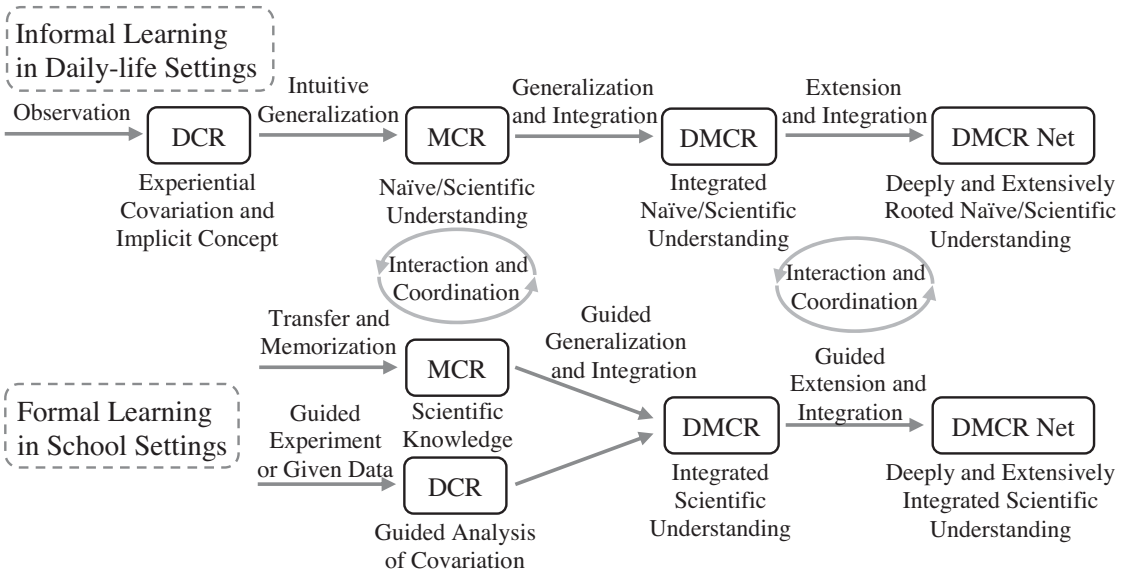


FIG. 7. The development of knowledge and reasoning in formal and informal education settings within a specific content domain. A learner at a specific age (with some variations) can be in different stages for different content domains, while learners at different ages can also be in similar or different stages for the same or different content domains.

knowledge and reasoning in formal and informal education settings. The process starts from observing the world and forming DCRs at a young age, which can be further generalized by the learners to form intuitive concepts and understandings (naïve versions of MCRs, DMCRs, and networks of DMCRs). As language develops and communication increases, especially in formal education settings, learning and reasoning often transition to learning MCR-type knowledge during most years in school. Students' reasoning capability in forming DCRs and integrating them with scientific concepts becomes limited due to a lack of persistent training. As a result, many students have limited ability in asking researchable questions, designing controlled experiments, processing, and analyzing data, and drawing evidence-based conclusions, all of which are crucial skills needed in scientific inquiry. Many learners stay at this stage well into adulthood. Students in advanced academic tracks, who receive further research training, may eventually develop the needed scientific reasoning abilities and achieve integrated networks of scientifically based causal understandings. These learners are capable of constructing scientifically sound understanding through inquiry-based learning and investigation. However, as shown by several studies, a significant portion of college students are still lacking the necessary skills in conducting effective scientific inquiry and integrating DCRs with MCRs to construct scientific concepts [29,70,71,33]. In response to the underdeveloped reasoning skills among students, the assessment of scientific reasoning should emphasize skills in developing and coordinating DCRs and MCRs, which will be discussed next.

IV. iSTAR ASSESSMENT FRAMEWORK AND INSTRUMENT

In this section, a new assessment instrument on scientific reasoning is introduced, which is referred to as Inquiry in Scientific Thinking, Analytics, and Reasoning (iSTAR). The term inquiry is used here to indicate that iSTAR's main purpose is to support inquiry learning as it provides an operational framework for learning objectives and assessment. The framework can be used to guide the development and evaluation of inquiry-based instruction designed to foster skills in scientific thinking and reasoning. This assessment framework is based on the DMCR model of scientific reasoning and the IDEA-Loop processes previously discussed.

A. Defining an operational assessment framework and skill dimensions

From the developmental progression of learning and reasoning discussed in Fig. 7, students going through formal education often lack proper training on the reasoning skills necessary to develop DCRs and to integrate DCRs with MCRs to draw evidence-based causal conclusions. Therefore, the assessment framework of scientific reasoning is designed to emphasize these skills. For example, the related skills for developing DCRs are measured through evaluation of students' ability to conduct effective data analysis in simple to complex settings (see, e.g., Figs. 3 and 4). Meanwhile, reasoning skills for developing MCRs are assessed based on students' ability to handle biases from prior knowledge as well as their ability to identify possible mechanisms with explicit and implicit variables in simple to complex contexts. The complexity of these assessment tasks can be manipulated using different numbers of variables, types of relations, causal conditions, and so on. Additionally, given that it is essential for effectively coordinating theory and evidence to draw valid causal conclusions, the ability to integrate DCRs and MCRs for evaluation and analysis of the relations between evidence and hypothesis is also emphasized in the assessment framework. These abilities are measured through data-analytic and causal decision making tasks.

Aligned with the DMCR model of scientific reasoning, the iSTAR assessment framework is established with three primary dimensions of reasoning skills and processes, where each involves multiple subskills. The three main dimensions include control of variables, data analytics (DA), and causal decision making (CDM). The framework is illustrated in Fig. 8. A list of subskills and developed test items is shown in Table I and will be discussed in the next section. This list represents the core skills in the DMCR processes using control of variables to design covariation experiments, data analytics to identify valid DCRs, and causal decision making to coordinate DCRs and MCRs for drawing valid conclusions and developing new knowledge. In this way, the identification of these subskills operationally defines how DMCR is functionally carried out to support scientific reasoning.

In modeling the scientific reasoning skills and processes, it is useful to clarify the relations among the three modeling frameworks discussed earlier. The DMCR framework

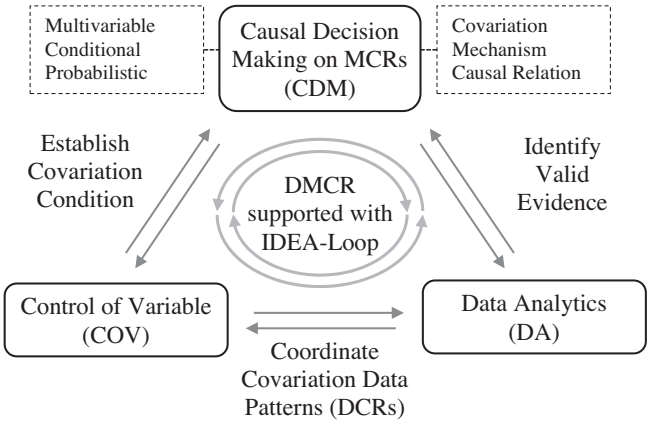


FIG. 8. iSTAR assessment framework of scientific reasoning.

TABLE I. iSTAR assessment dimensions, contexts of subskills, and question distribution.

Skill dimensions	Contexts of subskills	Questions
Control of variables	• Identify or design COV conditions with multiple testable and untestable variables • Real life and STEM contexts • With or without experimental data • Simple to complex relations • Extension to DA and CDM dimensions	9 COV questions: 1, 4, 5, 10, 21, 24, 28, 29, 30
Data analytics	• Multivariable linear proportion • Combinations • Conditional probabilities (wide variety) • Multivariable correlation and covariation • Fundamental statistics such as weighted average and random sampling concept • Bayesian probability	15 DA questions: 2, 3, 6, 7, 8, 13, 14, 22, 23, 25, 26, 27, 32, 33, 35
Causal decision making	• Prior knowledge and bias in causal decision • Correlation and covariation causal decision • Bayesian inference and causal decision • Conditional rules for causal decision • Conditional probability and basic statistics for causal decision	11 CDM questions: 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 31, 34

provides the fundamental concept that scientific knowledge is developed through scientific inquiry using scientific and causal reasoning (see Fig. 1). Meanwhile, the IDEA-Loop model shown in Fig. 5 describes the functions of scientific and causal reasoning in dynamic cycles of scientific inquiry. Finally, the iSTAR assessment framework shown in Fig. 8 outlines the structural components and interactive relations among the different areas of reasoning skills that can be isolated and measured. These models together provide a complete theoretical framework for describing, modeling, and measuring scientific reasoning skills.

Among these skills, control of variables is the fundamental first step in setting up controlled experiments so that covariation data and DCRs between independent and dependent variables can be obtained (see Fig. 3). Ample research has studied COV, which has informed the establishment of its subskills and item designs [49,72]. As described earlier, the complexity of COV assessment tasks can be controlled through manipulating the familiarity of contexts, number of variables, data representation, and relations among variables [72].

The dimension on data analytics is a broad category including a wide range of data analysis and interpretation skills, which enable students to identify meaningful data patterns and trends and evaluate their validity in order to develop and validate DCRs among variables. In addition, evaluation of conditional probability is particularly emphasized since it is considered a key factor in causal decision making [73]. Overall, the subskills in the data analytics category consist of a wide range of computational and analytic skills including the evaluation and interpretation of proportion, correlation, covariation data patterns, conditional probabilities, and Bayesian probabilities. These skills support the I, D, and EA processes in the IDEA-Loop model for identifying, deriving, and evaluating valid evidence as well as conditional constraints of hypotheses for coordination between theory (MCRs) and evidence (DCRs) (see, e.g., Figs. 5 and 6).

Causal decision making is another broad category focusing on the ability to comprehensively analyze DCRs and MCRs by applying results from the COV and DA processes for determining valid causal relations. This is the key step in DMCR reasoning, as it integrates the results from the previous two (COV and DA) to carry out causal coordination between DCRs and MCRs.

Within causal decision making, four subcategories of skills can be defined. The first is the ability to distinguish between covariation and correlation, which students often have difficulty with as they tend to interpret correlation results as covariations to draw causal conclusions [74,75]. This ability is typically measured using tasks showing correlations without proper COV setups. The contexts may also involve hidden variables and other confounding factors and relations, which may be ignored by students.

The second subcategory involves a range of conditional and Bayesian probability evaluation skills that are frequently employed in both DCR and MCR tasks. Conceptual understanding and computational manipulation of these probabilities are essential in correctly predicting a probabilistic outcome, inferring a possible cause or contributing factor, and determining causal relations.

The third subcategory includes the conditional and logical rules in determining causal relations. These are often common in MCR tasks in which certain causal premises and outcomes are given or assumed, and students need to apply the conditional logic rules in order to determine the evidence that would correctly match the claims. Examples of these rules include those handling conditions that are sufficient, necessary, contributing (neither sufficient nor necessary), and unrelated. The skills also include the abilities to translate these rules in both forward causal prediction, such as in a D-process with given cause to derive or predict effect, and backward causal inference, such as in an I-process with observed effect to infer possible causes. For example, if A is the sufficient condition of B, the forward logic can be described as "if A exists then B exists." The corresponding reverse logic warrants that "if B doesn't exist then A cannot exist either." These conditional rules are important logical reasoning skills that students need to know and apply properly in the IDEA-Loop process to identify the correct causal relations in multivariable settings.

The fourth subcategory includes the reasoning skills in manipulating MCRs, which are typically domain-specific knowledge. In assessment, mechanism-based reasoning skills can be measured through two types of subskills. The first is the students' ability in evaluating the consistency between a hypothetical claim and its supporting evidence, i.e., the ability to synthesize DCRs and MCRs for causal decisions. For example, LCTSR includes four questions measuring hypothetical deductive reasoning, which is along the same line of mechanism-based causal reasoning. To answer these questions, students need to either produce consistent predictions (mostly through the D-process) on the outcomes of differently conditioned experiments based on a given hypothesis or identify experimental outcomes through cycles of IDEA-Loop for validation of certain hypotheses. The second type of subskill for handling causal mechanisms focuses on students' ability in understanding and handling covariation situations under the influence of their prior knowledge and beliefs, which may bias their reasoning in processing DCRs. Students with underdeveloped understanding and ability in reasoning with DCRs may rely primarily on mechanism-based knowledge, rather than covariation data, as evidence to support a claim. For example, in the correlation questions (mice questions) of the LCTSR, students are expected to explain the correlation based on analytic skills to evaluate the number of mice having different features. However, students lacking the needed data analysis skills may answer with a mechanism-based belief: "there may be a genetic link between mouse size and tail color" (LCTSR Q20), claiming that a genetic mechanism can be the cause for having darker tails, which is not supported by the data presented.

It is expected that in completing reasoning tasks, many of these skills may be used in combination to support multiple pathways of IDEA-Loop. The interactive relations among the three main categories of reasoning skills are also illustrated in Fig. 8, which shows the primary functions of different reasoning in supporting the IDEA-Loop for coordination between DCRs and MCRs. For example, in a typical task, the control of variables skills are applied to establish controlled trial conditions as the experimental basis for collecting covariation data. These conditions are evaluated by the causal decision making skills for their covariation or correlation nature and the outcomes are used as part of the evidence for causal decision making. Based on the established control of variables conditions, the data analytics skills can be used to analyze the collected data to identify unique covariation patterns that can be used as evidence in causal decision making. Working together, these skills serve to generate, evaluate, and synthesize both DCRs and MCRs to determine the validity of a causal claim. For question design considerations, the difficulty of a causal reasoning task can be manipulated with the complexity of the underlying causal network. The causal network can range from simple few-variable systems to complex many-variable systems, while the embedded relations can be simple linear, conditional, and complex probabilistic (see Figs. 3 and 4).

When working with complex reasoning tasks, the three areas of reasoning shown in Fig. 8 often feed into each other in dynamic cycles of IDEA-Loop. For example, when a satisfactory conclusion is not reached, the tentative outcomes of the causal decision making process can reinitiate or manipulate the control of variables process to modify the experimental conditions. Such modifications often include controlling or changing different or additional variables in order to obtain a specific measurement setting or to modify the current covariation condition. The causal decision making outcomes can also provide cues to guide the data analytics process to identify new or alternative data patterns and relations or to use a different set of data analytic algorithms. Results of the data analytics process can then feed into the control of variables operation for purposes such as altering the covariation conditions and improving the reliability of observed data patterns. The fundamental reasoning elements supporting these functions and processes go through multiple IDEA-Loop cycles at different levels of complexity and abstraction in both inductive and deductive pathways. Together, these functions and processes provide a theoretically based operational framework that can concretely model and assess scientific and causal reasoning.

B. Development of an assessment instrument on scientific reasoning

Following the creation of the iSTAR assessment framework, an assessment instrument on scientific reasoning was

developed, which is referred to as the iSTAR test. The current version contains 35 multiple choice questions populated over the three main skill dimensions, including control of variables, data analytics, and causal decision making. In order to facilitate the implementation of the test, a study is underway to split the full length iSTAR test into two short parallel tests that each contain approximately 20 questions. The idea is that the short versions can be randomly delivered to a population to generate outcomes equivalent to the full-length test. The measured skill dimensions, subskills, and question distributions of the iSTAR test are summarized in Table I.

The subskills in control of variables form a progression from simple identification and design of valid COV experiments to more complex situations in which experimental data is given and students are asked to determine if certain variables are causally influential [72]. Additional contextual features such as real-life and STEM based contexts are also blended into the question design to control the task difficulty. A total of nine questions are included to assess the COV dimension and its subskills.

Data analytics contains the largest number of subskills, which are measured with 15 questions. These subskills have more independence from each other and were not designed in any order of progression. As evident from Table I, the DA skills have a focus on a variety of probability concepts and evaluation skills. In particular, conditional probability and Bayesian probability are heavily emphasized, since these often play an essential role at both the conceptual level for understanding the needs and purposes of probability conditions and at the operational level for calculating the quantitative weighing in causal decision making [73].

As discussed earlier, causal decision making is a key ability in drawing valid conclusions based on evidence and hypothesis. Five subskills are included and assessed with 11 questions, among which the correlation versus covariation and causal conditional rules are most emphasized. Ample research has shown that students often treat correlation as covariation-based causal relation [74,75]. Developing abilities in this area will improve students' ability for making valid interpretations of data from not only scientific experiments but also reports in public media. Meanwhile, causal conditional rules provide the logical computing functions that appropriately link evidence, claims, and conditions in reasoning to identify consistent and logically cohesive causal relations. These are essential skills that enable students to correctly align (or coordinate) claims with evidence under different conditions.

C. Sample questions from iSTAR test

1. Questions on control of variables

Reasoning skills in control of variables are likely the most studied area in scientific reasoning [49,72]. On the LCTSR, six questions (of 24) were designed to measure COV skills. However, a recent study reveals that four of the six COV questions appear to have design issues [37]. Together, these research outcomes have informed the development of the iSTAR questions for COV.

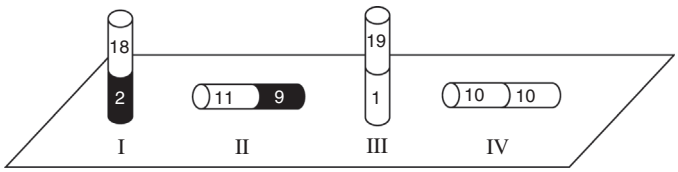
The iSTAR test includes nine questions to measure abilities in COV, out of which three have been adapted from LCTSR for equating between iSTAR and LCTSR results. These revised versions are based on the outcomes of the validity evaluation [37]. For example, Fig. 9 shows the modified version of the fruit fly questions from the LCTSR, with three main changes from Lawson's version. First, the original pictures show tubes covered by overlapping black dots, which have been replaced with smooth black covers in the new version. From students' comments during interviews, the black dots in the original pictures were often misinterpreted as flies rather than black paper: "*I thought the black dots on the tubes were a mass of flies*." After changing the figure to that shown in Fig. 9, these comments did not come up in later interviews.

Second, the layout of the four tubes in the original version sometimes confused students as if all four tubes were placed horizontally on a table: "*I didn't realize tubes I and III were vertical. I thought they were all lying flat on the table*." In addition, the arrows representing incoming light in the original version misled some students to think that light only came from directions along the arrows: "*I thought those arrows were the light beams, and the flies fly towards the light in tubes I and III*." To remedy these issue, the new version makes use of a transparent horizontal table to more clearly show the spatial positions of the four tubes. The arrows representing the light have been removed, and a note has been added stating that the light comes from all directions. After these revisions, students did not report any issues in subsequent interviews.

Third, the original version of this question only labels the number of flies in the noncovered portion of the tubes. Based on a suggestion from a middle school teacher, this may cause students with weak math skills to make a mistake when interpreting the numbers. Therefore, to avoid miscalculation of the number of flies, the new version clearly labels the numbers of flies in both the covered and noncovered parts to help students make explicit comparisons of the conditions and results without the need to do a calculation. Several wording clarifications have also been made in the new version.

In addition, the questions in LCTSR were designed with a two-tier structure, such that the first question in a two-question group asks for a relation-based conclusion, and the second question asks for the reasoning that explains the answer to the first question. The four fly questions on LCTSR were designed in two two-tier sets. Since the two LCTSR questions asking for reasoning were found to be unclear and sometimes misleading for students in a previous study [37], the iSTAR test does not include them. Rather, the contexts of the LCTSR fly questions were

(Fly1) Four glass tubes that each contain twenty fruit flies are sealed shut. Half of Tubes I and II are covered with black paint including the end; Tubes III and IV are not covered. These tubes are placed on a clear horizontal glass table with tubes I and III standing up on their ends and tubes II and IV lying flat as shown. For five minutes the tubes are exposed to red light coming from all directions. At the end of the five minutes, the number of flies in the covered and uncovered parts of each tube are counted as shown below. Note that a total of 20 fruit flies are still in each tube.

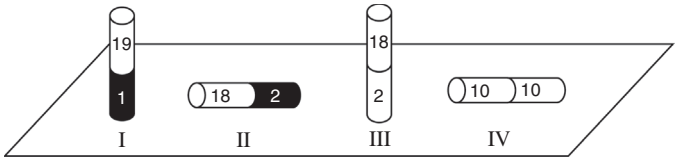


Red light comes from all directions. Table is made of clear glass.

This experiment shows that flies respond to ("respond to" means move towards or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity
e. cannot be determined from the information given

(Fly2) In a second experiment, blue light and a different kind of fly is used. The results are shown in the diagram.



Blue light comes from all directions. Table is made of clear glass.

Results from which tube(s) are needed to determine whether or not these flies respond to blue light?

a. Tube I or tube II
b. Both tubes I and II
c. Both tubes I and III
d. Both tubes I and IV
e. Both tubes II and III
f. Both tubes II and IV
g. Both tubes III and IV
h. Tubes I, II, and III
i. All four tubes
j. None of the above

%	A	B	C	D	E	F	G	H	I	J
Fly1	18.7	48.0*	21.3	4.8	7.2					
Fly2	6.0	20.8	4.1	1.7	2.6	32.8*	0.8	6.7	22.5	2.0

FIG. 9. Modified versions of the fruit fly questions from LCTSR, both of which were adapted for iSTAR. The percentage distribution of answers are based on the college population discussed in the next section. The correct answer is marked with an asterisk (*).

altered using a staggered two-tier structure such that the second question still builds from the first one, but instead of explaining the answer to the first question, the second question extends the reasoning to identify additional COV strategies for generating appropriate covariation evidence that can support a possible DCR claim (see Fig. 9). That is, in the new version, the answer choices were redesigned to target explicit comparisons among different COV conditions and outcomes, allowing them to directly measure the core process of COV reasoning.

For example, for the first fly question shown in Fig. 9, many students at the college level were able to select the correct answer (choice b). However, a significant fraction of the students did not know how to compare the results of the different tubes [37]. Among the students who answered incorrectly, they tended to focus on tube I in their comparisons. These students compared Tubes I and II and concluded that flies respond to the red light since the majority of the flies were in the unshaded part. For the effect of gravity, many of these students made comparisons using tubes I and III, but their conclusions on the influence of gravity were split between effective and noneffective. Some considered that flies going against gravity to the top of the tube was a sign of responding to the effect of gravity and picked choice c. Meanwhile others considered that flies going with the gravity to the bottom of the tube was a sign of responding to the effect of gravity and picked choice a. In both cases the students appeared to base their responses

TABLE II. Designs of assessment questions on their targeted skills mapped to the modeling and assessment frameworks. The six example questions were designed to probe three main areas of skills for COV, DA, and CDM with emphasis on different causal reasoning components (DCRs and/or MCRs) and IDEA-Loop processes. Mapping to specific casual relations and the complexity of related reasoning processes defined in Figs. 3 and 4 is also included.

Questions	iSTAR reasoning subskills	IDEA-Loop reasoning processes	DMCR components (DCR and MCR)	Causal networks (Fig. 3)	Complexity of reasoning processes (Fig. 4)
Fly 1 (Fig. 9)	COV	ID	DCR	COV condition	Two-variable simple relation
Fly 2 (Fig. 9)	COV	ID	DCR	COV condition	Two-variable simple relation
Dice (Fig. 10)	DA	EA	DCR	Bayesian probability	Bayesian complex relation
Market share (Fig. 10)	DA	EA	DCR	Conditional probability	Conditional complex relation
Giraffe (Fig. 11)	CDM	IDEA-Loop	DMCR	Hidden mechanism	Hidden complex relation
Card (Fig. 11)	CDM	IDEA-Loop	DMCR	Conditional logic link	Complex conditional logic

on prior beliefs or knowledge rather than the data patterns and COV conditions. In addition, in their comparisons to determine the effect of a variable, they tended to focus only on the cases in which the concerned variable varied but often ignored the need for controlling the possible confounding variable. Therefore, for the effect of red light, these students compared tubes I and II. For the effect of gravity, they compared tubes I and III. Such reasoning clearly demonstrates a lack of basic understanding for the need for COV in creating a covariation measurement.

The second fly question builds on the first question to explicitly target how students make comparisons in a slightly varied situation. Since this question targets the explicit comparison process in COV reasoning, it is expected to be harder than the first fly question where students may have a "gut feeling" of the answer but are not able to explain it clearly. The correct answer for the second fly question involves the comparison of tubes II and IV (choice f), in which gravity is controlled. As expected, fewer students picked the correct answer, and many selected either choice b or choice i. Similar to the first fly question, students who picked choice b tended to focus on the condition in which the concerned variable varied but ignored the need for controlling the possible confounding variable. Meanwhile, choice i was included as a general distractor to identify students with little understanding of the purpose of COV.

As shown in Table II, which has been provided at the end of this section to map the selected iSTAR questions into the assessment framework, both fly questions target the COV subskill in a causal DCR setting. In addition, the reasoning task is to identify valid DCRs, which in this case include 2-variable simple relations (i.e., simple effects of light and gravity). To identify and validate the relations, students need to engage in inductive reasoning (I-process) to identify possible causes and then apply the hypothesis to generate an outcome using deductive reasoning (D-process), which guides the selection of a possible choice. Simple processes for evaluation and analysis (EA-processes) are also engaged to compare the outcomes of the deductive reasoning with the choices of a question to determine or validate an answer. Since these simple EA processes are not the essential part of the reasoning here, they are not listed in Table II as the main targeted skills of the COV questions.

In this section, the fly questions were used to demonstrate the process of question development and revision, both of which depended, in part, on the use of student interviews and teacher feedback. These questions went through several iterations to eliminate possible design issues. In addition, six other COV questions are included in the iSTAR test. All six are new designs that are not based on the LCTSR. They involve both real life and STEM contexts. The validation of these other six questions has been reported elsewhere, and they have been shown to follow a progression of COV skills [72]. In the remaining discussions, the development process of the questions will not be detailed, but rather emphasis will be placed on introducing the new features of the iSTAR test.

2. Questions on data analytics

Data analytics is an extended category including a wide range of computational and analytic skills. The basic set of skills overlaps with the ones measured in the LCTSR, including simple probability, proportion, and correlation. Extending the basic skill set, data analytics in iSTAR also targets more advanced computation and reasoning. These include combination, conditional probability, multivariable correlation and covariation, fundamental statistics such as weighted average and random sampling concept, and Bayesian probability. These latter skills are important for students to analyze data and identify valid evidence, which can be further applied in causal decision making.

The iSTAR test includes 15 questions to measure data analytics subskills. For equating purposes, one of the LCTSR questions on correlation (LCTSR Q19) has been adapted for the iSTAR test. The remaining 14 questions are new designs that measure the range of data analytics subskills listed in Table I. Two examples are given in

(Dice) One day, while travelling in another country, you observed a group of people engaged in a game tossing a six-sided die. The die was hand carved into a rough six-sided cube that has three sides painted black and the other three painted white. You counted that when the die was tossed 1000 times, white sides turned up 720 times and black sides turned up 280 times. If the die was tossed another 100 times, about how many times would white sides most likely turn up?

a. about 30 times
b. about 50 times
c. about 70 times
d. There is no way to know the number because it is an uncertain event. The die will turn up either white or black randomly.
e. None of the above

(Market Share) A recent report about the quality of a type of product shows that 80% of confirmed poor quality issues are from products made by company A. Based on this result, which of the following statements can be concluded?

a. When a customer buys this type of product made by company A, the customer is likely to encounter poor quality issues.
b. This type of product made by company A is more likely to have poor quality issues than those made by other companies.
c. The data shows that 80% of this type of product on the market may have quality issues.
d. a and b
e. a and c
f. b and c
g. a, b, and c
h. None of the above can be concluded based on the report.

%	A	B	C	D	E	F	G	H
Dice	7.2	16.8	43.2*	31.3	1.4			
Market Share	17.3	24.1	4.9	26.9	13.8	1.6	6.1	5.4*

FIG. 10. Example questions on data analytics in iSTAR. The percentage distribution of answers are based on the college population discussed in the next section. The correct answer is marked with an asterisk (*).

Fig. 10. The first question measures students' understanding of Bayesian probability, and the second question measures reasoning with conditional probabilities. Through interviews conducted in the development of iSTAR, it has been found that college level students have sufficient understanding of the concept of independence and equal probability in uniform random processes with cases such as coin flipping and die tossing. However, a weakness has also been identified such that students tend to overextend this uniform (or equal) probability to all random processes, including nonuniform conditions. The concept that probabilities are conditional, such that probabilistic states may not always be evenly distributed, is not well established among students. In real world situations, conditional and nonuniform random processes are common, and therefore, it is important to assess whether or not students can understand and reason with conditional and nonuniform probabilities.

The first question in Fig. 10 targets Bayesian probability, which measures if students can handle nonuniform random processes based on observation outcomes. The six-sided cube used in the question is not a perfect die and may generate uneven probabilities for showing the different faces. The reasoning, which uses observed data to make an inference of the die's intrinsic feature and probability in turning up certain faces, is a Bayesian type decision process. From interview and open-ended survey results, many students appeared to understand the randomness and independence idea such that each toss of the die is a random event independent of other tosses. However, students' thinking seemed to be dominated by the randomness and/or the equal probability idea, and they applied this reasoning indiscriminately to this possible nonuniform situation. As a result, these students typically chose choice b for uniform probability or choice c for indicting that each tossing event is independent of previous results.

The second question measures students' understandings of base rate in a conditional probability situation. The market share of product A is not given and the statistical outcome of defects from the market sampling depends on both the defect rate and market share. Therefore, the results cannot be extended to estimate the actual defect rate as a quality index of product A without knowing the market share. Results from student interviews and quantitative responses show that many students cannot properly analyze this type of probability. They often selected choices a, b, or c, or their combinations, suggesting a lack of proper understanding of the concept of base rate in conditional probability.

The skills targeted by these two questions are also mapped to the components and processes of the modeling and assessment frameworks shown in Table II. As

discussed above, both questions emphasize the EA processes for evaluation and analysis of observed probabilistic data, which represent DCR based causal understanding. In particular, reasoning skills on manipulating and conceptualizing Bayesian and conditional probabilities at varying complexities are the main targeted areas of these two questions.

Compared to the related questions in LCTSR, which involve only three subskills on proportion, simple probability, and correlation, the iSTAR questions cover a much wider range of data analytics subskills. More importantly, the data analytics skills on iSTAR are designed based on the new modeling framework so that they serve explicit purposes for connecting COV skills to form valid evidence that support causal decision making. On the technical side, the LCTSR questions appear to be relatively easy for senior high school and college students, which can lead to significant ceiling effect when testing these students [37]. In contrast, the difficulty levels of the iSTAR questions are designed to distribute over a wide range. The goal is to make the test effective at measuring broad student populations from middle school to graduate levels. The validation of the test with middle school and high school populations is underway and will be presented in future publications.

3. Questions on causal decision making

In the iSTAR assessment framework, causal decision making represents a core step for drawing valid evidence-based conclusions in a reasoning task. For example, in a task to test a hypothesis, a learner needs to use control of variables and data analytics skills to establish valid covariation conditions and identify DCRs and MCRs that are computationally and logically sound and mechanistically plausible. The evidence and hypothetical causal claims are then evaluated for their consistencies and conditional validities, which are incorporated into a series of decision-making processes to determine the most probable conclusion regarding the validity and confidence of both the involved theoretical claims and evidence.

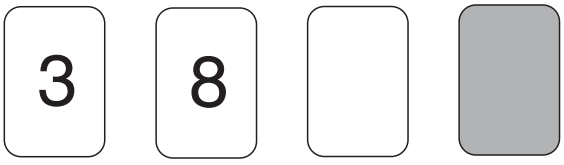
For LCTSR, the skill dimension for hypothetical-deductive reasoning is comparable to the CDM subskills assessed in the iSTAR test. However, the CDM skill dimension involves an extensive set of explicitly defined subskills, while the hypothetico-deductive assessment questions in LCTSR focus primarily on evaluating the consistency between evidence and predicted outcomes based on a given hypothesis. The designs of the hypothetico-deductive questions in LCTSR have also been critiqued for their content validity due to the involvement of implausible assumptions [37]. Because of the validity issues of the

(Giraffe) A visitor traveled to Africa to tour the natural breeding environment of giraffes. While there, the visitor noticed a type of tall tree that grew fruit only at the top of the tree. The visitor also noticed that giraffes that frequently ate this fruit appeared to be stronger and taller than those that could not reach the fruit. Based on these observations, which of the following statements can be concluded?

a. When a giraffe frequently eats this fruit, it grows stronger and taller.
b. The nutrients in the fruit can help the giraffe grow stronger and taller.
c. Both choices a and b above.
d. The result cannot be used to show that eating the fruit causes a giraffe to grow stronger and taller.
e. The result doesn't matter. The height of a giraffe is determined by its genes.
f. None of the above is a good conclusion.

(Card) Albert is playing a new card game with his friend Jon. Each card contains a number on one side and is either white or gray on the other side. After a while, Jon makes a claim: "If a card has an even number on one side, then it is gray on the other side."
Four cards are picked and presented to Albert as shown below.
Which card or cards, when turned over, are useful to determine whether Jon's claim is false?

a. the 3 card only
b. the 8 card only
c. the 3 card and the white card
d. the 3 card and the gray card
e. the 8 card and the white card
f. the 8 card and the gray card
g. all four cards
h. none of the above



%	A	B	C	D	E	F	G	H
Giraffe	2.2	5.7	19.6	42.8*	18.9	10.8		
Card	1.0	7.9	3.3	2.4	3.9*	32.5	47.6	1.3

FIG. 11. Example questions on causal decision making in iSTAR. The percentage distribution of answers are based on the college population discussed in the next section. The correct answer is marked with an asterisk (*).

hypothetico-deductive questions in LCTSR, all of the 11 CDM questions in the iSTAR test are new designs. Two examples are shown in Fig. 11.

The first question in Fig. 11 measures students' reasoning in distinguishing between correlation and causation in CDM processes. The scenario of this question is similar to many real-world examples such as whether a specific eating habit can be correlated with certain health conditions. In these cases, there are often confounding factors that lead to the observation of correlated outcomes, which warrants a correlation but not causation. Similarly, in this question the given observation shows a correlation between the height and strength of a giraffe and whether it eats a type of fruit. However, this data does not form a valid covariation design due to a lack of control of variables, since only the tall giraffes can reach and eat the fruit, which is the confounding factor given semiexplicitly in the question. Therefore, choice d gives the correct answer, which indicates that a covariation based causal relation cannot be concluded. Choices a, b, and c represent the tendency in reasoning that treats correlation as causation without understanding the needed conditions for a valid covariation. Choice e represents the influence on reasoning from mechanism-based prior knowledge such that the observed data are ignored or disregarded without explanation, and the decision is made purely based on an existing belief or prior knowledge. This type of reasoning indicates the lack of understanding and synthesis between DCRs and MCRs and suggests a reliance on prior knowledge in causal decisions.

The second question in Fig. 11 is based on the Watson's selection task [76], which measures reasoning with conditional logic rules in determining consistency between evidence and claim. The hypothetical claim in the question represents a sufficient condition of "if…then" relation; that is, if the front side of a card has an even number, the back side of the card must be gray. In order to show that the claim can be false, there are two logical pathways that need to be tested. One is forward pathway to test the deductive "if … then" relation by turning over the even-numbered cards to check their back side colors. The other is the backward pathway to test the reverse logic. Since the even number is a sufficient condition for its back to be gray, if the back of a card is not gray (hence it is white), the card cannot have an even number (must have an odd number). Therefore, to test the hypothesis given in the question, the card with an even number and the card with a white back should be turned over (choice e). The results of the other cards do not provide any useful information for evaluating the validity of the hypothesis. From student interviews and test results, it appears that students have a tendency to focus on evidence following the forward deductive reasoning path (a confirmation type reasoning), which leads to the selection of the even numbered card and the gray card (choices b and f). In addition, many students simply wanted to use all cards without seeing their roles in the conditional logic. These outcomes suggest that most college level students lack a sufficient understanding of the conditional logic in causal decision making.

In the existing literature, reasoning on the coordination between theory and evidence has been well studied [38]. The CDM questions are designed to specifically target the essential reasoning skills that support the operation of theory evidence coordination, such as the skills needed to evaluate valid covariation relations and form consistent alignment between DCR-based evidence and MCR-based theory. As summarized in Table II, the two CDM questions will each engage the complete IDEA-Loop cycle for coordinating between DCRs and MCRs to form an integrated DMCR type of causal understanding. Subskills, such as identifying hidden variables and relations and handling conditional rules, are targeted in these two questions for evaluation of students' levels of reasoning in terms of the structures and complexity of the involved causal relations.

V. iSTAR ASSESSMENT VALIDITY AND RELIABILITY

For a little over a decade, the iSTAR assessment has been gradually developed through extensive qualitative and quantitative research and evaluation. The development and validation of the assessment have been an on-going process that continuously refines the instrument and the modeling framework of scientific reasoning. The current version discussed here reflects a stable release completed in 2018, which can be accessed and used in practice through an online delivery system described in the Appendix. Item-level descriptive statistics are also provided in Table VII in the Appendix. The following sections will focus on establishing the basic assessment attributes and validity of the iSTAR test.

A. iSTAR assessment properties and comparison with lctsr

In this section, descriptive statistics of iSTAR with different populations will be introduced to establish the baseline of assessment outcomes. The results will be compared with measures of LCTSR as a reference to existing results established in the current literature, since LCTSR is the most widely used assessment instrument on scientific reasoning with a large user base and data library. The comparison will also help in interpreting the similarities and differences between the two assessments.

As discussed in the previous section, iSTAR contains three general skill dimensions that each include multiple subskills at different levels of complexity. Meanwhile, LCTSR contains six narrowly defined skills dimensions. All of the LCTSR skills can be mapped onto the iSTAR dimensions except for the dimension on conservation of

TABLE III. Cross references of skills dimensions and questions in iSTAR and LCTSR.

iSTAR skill dimensions and questions		LCTSR skill dimensions and questions	
Control of variables	1, 4, 5, 10, 21, 24, 28, 29, 30	Control of variables	9, 10, 11, 12, 13, 14
Data analytics	2, 3, 6, 7, 8, 13, 14, 22, 23, 25, 26, 27, 32, 33, 35	Proportion	5, 6, 7, 8,
		Probability	15, 16, 17, 18,
		Correlation	19, 20
Causal decision making	9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 31, 34	Hypothetico-Deductive	21, 22, 23, 24
		Conservation	1, 2, 3, 4

matter, which is too simple for students in middle school and above and is not included in iSTAR [37]. The skill dimensions and the corresponding questions are listed in Table III.

To compare the baseline assessment features of iSTAR and LCTSR, randomized A-B testing was conducted with high school students from a Midwestern suburban high school as well as college and graduate students from a Midwestern comprehensive university. The college students were freshmen from a first-semester calculus-based introductory physics course. The graduate students were in the second year of their physics Ph.D. program. During the testing at the high school and college, students were randomly given the iSTAR and LCTSR, and they each completed only one test. The graduate students were tested with a different procedure, where the testing of iSTAR and LCTSR was conducted during two separate times in the same week, and the order of the two tests was randomized for each student. The average scores of iSTAR and LCTSR are listed in Table IV.

The results show that the LCTSR scores are consistently higher than the iSTAR scores for all grade levels, revealing that iSTAR is more difficult than LCTSR. This improves the concerns regarding the ceiling effect of LCTSR when testing college level students [37]. The difference is largest for the first-year college students, which is likely the result of both the development of skills at this age and the population differences (i.e., college vs high school). Meanwhile, the smaller difference at the graduate student level is likely due to the ceiling effect in both tests. The scales of the mean scores at different grade levels suggest that the skills measured in iSTAR start to get more developed in college and graduate levels, while skills tested in LCTSR appear to have been mostly developed during high school to early college years.

To compare students' performances on the two tests within a grade level, the score distributions of the college students listed in Table IV are plotted in Fig. 12. The results show that the iSTAR scores are centered around 50% with a near normal distribution. Meanwhile, the LCTSR scores are centered around 80%, and the distribution is skewed to the high end with obvious ceiling effect.

For the same group of college students, their dimensional scores on iSTAR and LCTSR are also calculated and compared in Fig. 13. The results show that on the three common dimensions, the dimensional scores of LCTSR are significantly higher than those of iSTAR ($p < 0.001$). For LCTSR, scores on the conservation dimension are nearly 90%, which confirms that this dimension is too easy for college level students. On the three common dimensions, iSTAR demonstrates a consistent increase of difficulty from COV to DA and to CDM, which confirms the expected design with COV for setting the foundation of covariation, DA for the intermediate processing and analytics, and CDM for the advanced synthesis and causal decision making. In comparison, LCTSR also has its CDM questions being the most difficult with an average score at 60%, however, the COV and DA questions appear to be on the easier side for college students with average scores close to 80%. In particular, the DA questions are at a similar level as the COV questions, indicating that some advanced DA skills are lacking in LCTSR.

Summarizing the descriptive statistics and comparisons, the assessment features of iSTAR appear to properly target

TABLE IV. Comparison of iSTAR and LCTSR total scores from randomized A-B testing.

	iSTAR			LCTSR			*t* test	
Grade	N	Mean	SD	N	Mean	SD	Difference	*p* value	Effect size
9	88	0.315	0.156	88	0.393	0.142	0.078	0.016	0.520
10	110	0.396	0.171	89	0.485	0.242	0.089	0.004	0.428
13	187	0.516	0.170	96	0.766	0.156	0.250	<0.001	1.507
18	20	0.848	0.087	20	0.921	0.073	0.073	0.007	0.891

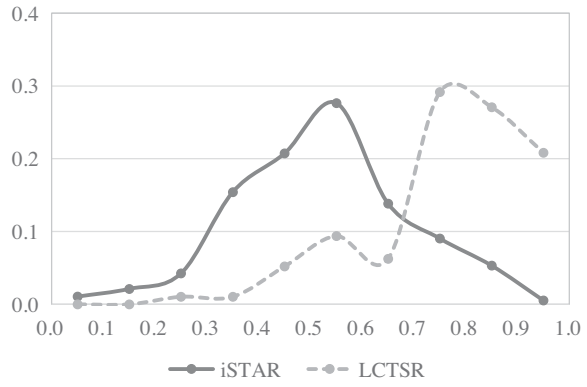
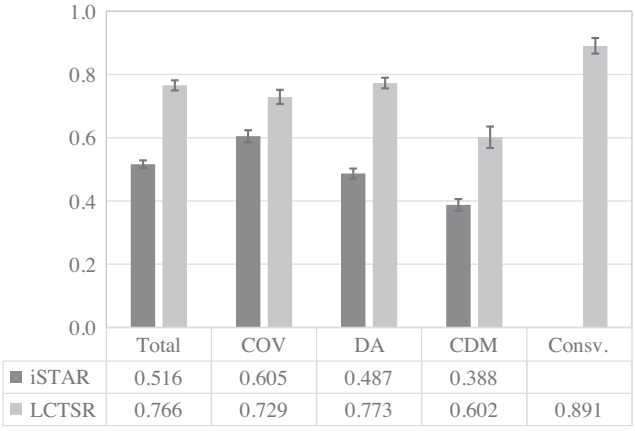


FIG. 12. Score distributions of college students on iSTAR and LCTSR.



	Total	COV	DA	CDM	Consv.
iSTAR	0.516	0.605	0.487	0.388	
LCTSR	0.766	0.729	0.773	0.602	0.891

FIG. 13. College students' dimensional scores on iSTAR and LCTSR.

college level students. Students' performances on the three skill dimensions also demonstrate a desired progression of difficulty levels that consistently follow the expected designs based on the DMCR model outlined in the previous sections. In the next section, additional analysis will be discussed to evaluate the validity and reliability of iSTAR.

B. Validity evaluation of iSTAR

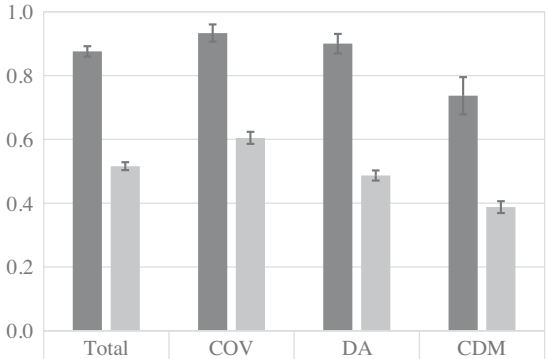
In education research, typical forms of validity evidence include content, criterion, and construct validity [77,78]. Content validity is established by ensuring the assessment adequately and properly covers the targeted content, which can be qualitatively judged by a group of experts in the field and quantitatively evaluated based on responses from expert level test takers. Criterion-related validity examines the consistency between a new assessment and an established similar measure, which is often evaluated based on the correlation between the two tests. Construct validity refers to how well an assessment measures the subject matter in terms of categories of targeted abilities (or latent traits) defined in a theoretical framework underpinning the cognitive attributes of the assessment. For iSTAR, the ability constructs include the three skill dimensions of COV, DA, and CDM, as defined in the iSTAR assessment framework. The construct validity can be established with a number of methods, such as the traditional approaches of factor analysis [79,78] and the Rasch model based approaches [77,80]. In this study, the Rasch model analysis will be used.

1. iSTAR content validity

In the development process, all items in iSTAR were examined by a team of experts who were science education researchers and teachers. Pilot versions of the test were also given to students in think-aloud interviews to collect detailed information on students' reasoning. The interview results and the designs of the items were evaluated by the expert team in group meetings to analyze students' understanding and refine question designs. This development process went through a large number of cycles of piloting and revision until all researchers in the expert team agreed that the instrument was properly and effectively designed to probe the targeted scientific reasoning skills.

As a part of the evaluation on content validity, another group of 30 graduate students from the same Midwestern university was used as an external expert group to examine if their answers agreed with the intended designs of the iSTAR questions. This group involved 3rd or 4th year physics Ph.D. graduate students, whose scores on iSTAR and its subskills are shown in Fig. 14, along with scores from undergraduate students for comparison, with the latter taken from Table IV.

As shown in Fig. 14, the trend of the relative difficulty among the subskills is similar for both the graduate and undergraduate students, with the graduate students reaching the ceiling. The results demonstrate that the expert-level students' understandings of the subskills agree with that of the design team. The difference in performance between



	Total	COV	DA	CDM
Graduate	0.876	0.933	0.900	0.737
College	0.516	0.605	0.487	0.388

FIG. 14. Scores on iSTAR and its subskills for graduate and undergraduate students. The error bars reflect standard errors.

undergraduate students and graduate students further indicates that as learning progresses, students' abilities on the subskills converge towards expertlike states. Therefore, the results in Fig. 14 can provide additional quantitative evidence to demonstrate the content validity of iSTAR.

2. iSTAR criterion validity

In existing research, LCTSR has long been used as a standard assessment for scientific reasoning [36]. Therefore, in this study, the evidence of criterion-related validity is evaluated based on the correlations between scores on iSTAR and LCTSR.

The results from the college students shown in Table IV were taken during the fourth week of their first semester introductory physics course. Among the 283 students, about one-third (96) took the LCTSR and the remaining students took the iSTAR. The uneven number of test subjects was due to the design of a parallel study conducted with the same population. Each student randomly took one of the two tests, which are referred to as LCTSR-4 and iSTAR-4 to label their time of testing. For all students, they were also given iSTAR during the first week of the course, which is labeled as iSTAR-1. With this design, correlations between iSTAR-1 and LCTSR-4 and between iSTAR-1 and iSTAR-4 could be obtained. Although there were three weeks of time between the week-1 and week-4 testing, the differences in students' scientific reasoning skills were not expected to change significantly [20].

Using the test results, Pearson correlations between students' scores were calculated, which show a moderate correlation of 0.589 ($p < 0.001$) between iSTAR-1 and LCTSR-4 and a moderately strong correlation of 0.680 ($p < 0.001$) between iSTAR-1 and iSTAR-4. As expected, the correlation between repeated iSTAR testing is stronger than the correlation between iSTAR and LCTSR. The results demonstrate good consistency of repeated testing and moderately good agreement between iSTAR and LCTSR. In addition, the graduate students listed in Table IV also took both iSTAR and LCTSR. The correlation between their scores on the two tests is 0.750 ($p = 0.002$), which is slightly higher than the correlation measured with the college students. The higher correlation from graduate students further confirms that performance on the two tests would converge as students approach the expert level.

To examine the agreement between measurements of skill dimensions of the two tests, correlations on the three common skill categories between iSTAR-1 and LCTSR-4 were also calculated and are listed in Table V. The results show a moderate correlation on COV, a low correlation on CDM, and a minimal correlation on DA. As indicated by the correlations, COV skills are more consistently measured by the two tests compared to other skills. On the other hand, measurements involving the DA skills are designed quite differently between the two tests. LCTSR only involves a few simplistic DA skills on proportion and basic probability, while iSTAR is designed with 15 items on a wide range of simple to complex DA skills. Similarly, the designs of the CDM measures are also quite different between the two tests. Nevertheless, the moderate overall correlation of the total scores demonstrates good agreement between the two tests for measuring the projections of the three skills onto a unidimensional trait of scientific reasoning.

TABLE V. Correlations of students' skill dimension scores between iSTAR and LCTSR. All correlations are statistically significant at $p < 0.01$ level except for $r = 0.172$ ($p = 0.093$).

iSTAR-LCTSR	COV	DA	CDM
COV	0.597	0.288	0.467
DA	0.298	0.293	0.262
CDM	0.354	0.172	0.317

3. iSTAR construct validity

As discussed earlier, iSTAR test is designed with three areas of reasoning skills on COV, DA, and CDM. The design is hypothesized to have a progression of increasing difficulty from COV to DA and to CDM. The evaluation of construct validity will then focus on analyzing whether the iSTAR data reveals a three-dimensional construct and whether the difficulty levels of the three categories of reasoning skills follow the designed progression.

First, the hypothesized three-dimensional construct model of iSTAR is examined by comparing it with an alternative unidimensional construct model. To do this, the iSTAR data is fitted to both a three-dimensional and a unidimensional Rasch model separately. The goodness of fit is then compared between the two models using the likelihood ratio test. If the result of this test is in favor of the three-dimensional model, the hypothesized three-dimensional construct is then confirmed and validated.

Second, the progression of the difficulty of the three reasoning skills can be evaluated based on students' mean abilities on the three skills. If the design is valid, students should demonstrate high abilities on COV, intermediate abilities on DA, and low abilities on CDM. Using Rasch analysis, students' mean abilities on the three skill dimensions were calculated and compared. In addition, the person-item map (Wright map) was used to show the distributions of person ability and item difficulty on a common vertical logit scale (Bond and Fox [80]) to compare if the distributions spanned properly over a wide range of ability and difficulty scales. A proper distribution indicates an appropriate discrimination for students from various performance levels.

For this part of the Rasch analysis, iSTAR data were collected with another larger group of 378 college students from the same college population listed in Table IV. In this section, the main outcomes of Rasch modeling are

TABLE VI. Students' ability means, reliability, and correlation matrix for the dimensions of the three-dimensional Rasch model. Diagonal values are EAP/PV reliabilities. Values below the diagonal are latent correlations.

Subskills	Mean	SD	COV	DA	CDM
Control of variables	0.658	1.337	(0.770)		
Data analytics	−0.042	0.687	0.807	(0.720)	
Causal decision making	−0.815	0.963	0.794	0.819	(0.697)

summarized, while additional details of Rasch model fitting are provided in Table VIII in the Appendix. First, the data were fitted with a one-dimensional and a three-dimensional Rasch model. A comparison between the two models shows that the model fit parameters are in favor of the three-dimensional mode. Likelihood ratio tests indicate that the three-dimensional model shows statistically significant improvement in model deviance compared to the one-dimensional model ($\chi^2 = 65.793$, df $= 5$, $p < 0.001$). The results suggest that the construct design of the three skill dimensions in iSTAR is consistent with the Rasch analysis of the assessment data.

Next, the means of students' estimated abilities on the three skill dimensions were calculated and are listed in Table VI, along with measures on reliability, which will be discussed in the next section. The results agree well with the prediction that students would demonstrate high to low abilities on COV, DA, and CDM, and the differences are statistically significant [$t_{\mathrm{COV-DA}}(377) = 15.421$, $p < 0.001$, $d = 0.793$; $t_{\mathrm{COV-CDM}}(377) = 34.943$, $p < 0.001$, $d = 1.797$; $t_{\mathrm{DA-CDM}}(377) = 26.759$, $p < 0.001$, $d = 1.376$].

In addition, the Wright map of iSTAR is also plotted in Fig. 15 in the Appendix, which shows that the items span a wide range of difficulty levels across the entire logit scale (−3.428 to 3.045). The students' estimated abilities on the three skills are also well spanned with near-normal distributions across a wide range of logit scale. The results suggest that iSTAR establishes an appropriate discrimination on each of the three skill dimensions for students from various performance levels. Altogether, the results of Rasch analysis demonstrate that the three-dimensional construct design of iSTAR is well established, and that the test items present good coverage on the range of students' ability regarding the three skill dimensions.

C. Reliability evaluation of iSTAR

Reliability is the consistency of an assessment such that repeated applications of the instrument on similar populations should produce similar results [77]. In classical test theory (CCT), the reliability coefficient is defined as the correlation between scores on two equivalent forms of an instrument. Practically, the concept is extended to consider every item of an instrument as an equivalent form, which leads to using the Cronbach's α coefficient of internal consistency as a measure of reliability. With the iSTAR data from college students used in Rasch modeling discussed above, the Cronbach's α is measured to be 0.737, which is adequate for acceptable reliability (>0.7) for the entire test.

Additionally, the Rasch model, as well as other models of the item response theory family, addresses the basic issue of reliability using information functions [81]. These functions indicate the precision with which the observed performance can be used to estimate the value of a latent trait for each student on a single item or the test as a whole [82]. Using this approach with Rasch modeling, indices analogous to traditional reliability coefficients can be estimated from the item information functions and distributions of the latent trait in a population. For this evaluation, the ratio of expected-*a-posteriori* over plausible-value (EAP/PV) reliability is used to measure the reliabilities of the three subscales, which are listed in Table VI along with the latent correlations among the three dimensions.

As shown in Table VI, the EAP/PV reliabilities for the three skill dimensions are 0.770 for COV, 0.720 for data analytics, and 0.697 for causal decision making. Typically, a reliability of 0.65–0.70 is considered "minimally acceptable", and a reliability between 0.70 and 0.85 is "respectable" for research purposes [83]. The results suggest that the reliabilities of the three skill dimensions are adequate, especially when considering the complex nature of the latent traits and the small number of items per dimension. Combining the results of Cronbach's α and the Rasch analysis, the reliability of iSTAR can be established at both the instrument level and the skill dimension level. Additionally, the latent correlations between the three dimensions range from 0.794 to 0.819, indicating strong correlations between the skill dimensions, which together contribute to a common basis of the overall scientific reasoning ability.

In summary, the evaluations of validity and reliability have demonstrated that iSTAR is a valid and reliable instrument for assessing scientific reasoning at the college student level. However, due to the complexity of the instrument, which is designed with multiple subskills and wide-ranging item difficulties, further research is warranted for establishing the validity and reliability for different student populations. Nevertheless, the results from this study should provide sufficient evidence for the validity of using iSTAR in the assessment of scientific reasoning for populations similar to entry level college students.

VI. SUMMARY AND DISCUSSION

Scientific reasoning is emphasized as a core ability in 21st century education and has been extensively researched. However, the existing literature lacks a consensus on a coherent model of scientific reasoning that can guide instruction and assessment, which is particularly

important for initiatives such as the NGSS or College Board Standards for College Success in Science. In addition, there are no existing instruments that measure scientific reasoning skills at a fine-grained level based on a coherent modeling and assessment framework. This gap in the literature can significantly limit the development, implementation, and evaluation of educational practices for effectively advancing scientific reasoning abilities among students.

Grounded in the literature, this paper presents a modeling framework (DMCR) that integrates scientific reasoning with causal reasoning and operationally defines scientific reasoning in terms of skills needed to process and coordinate data-covariation and mechanistic causal relations (i.e., DCRs and MCRs). Building from the DMCR modeling framework, three areas of reasoning skills, which include control of variables, data analytics, and causal decision making, have been defined to form the fundamental skill sets of an assessment framework of scientific reasoning, where the COV and DA skills provide the basis to develop DCRs, and the DA and CDM skills serve to coordinate DCRs with MCRs to form appropriate causal understandings. Guided by the assessment framework, an assessment instrument for scientific reasoning (i.e., iSTAR) has been developed, which targets a wide range of skills and subskills for COV, DA, and CDM. Through large scale testing, the assessment features of iSTAR have been examined and compared with the popular LCTSR. The results reveal that iSTAR provides consistent measurement of three areas of reasoning skills and demonstrates a progression of difficulty from COV to DA and to CDM, which confirms the design based on the assessment framework. In addition, the validity and reliability of iSTAR have been evaluated using classical statistics and Rasch analysis, indicating that iSTAR is valid and reliable to measure scientific reasoning abilities of college level students.

This research contributes to the literature in several ways. On the theoretical side, two schools of work on scientific reasoning and causal reasoning in the existing literature have been conducted rather independently in parallel with different definitions and emphases [47]. However, both types of reasoning are the essential elements in knowledge formation and have strong overlap among their goals and processes as well as specific reasoning skills. Therefore, connecting these two reasoning frameworks provides a synthesis for how different models of reasoning and learning are related. The integration of these models can form a more comprehensive view on the relations among constructs and processes that underlie reasoning and knowledge development.

In addition, the existing work on scientific reasoning often overwhelmingly emphasizes the data-covariation relation in evidence-based hypothesis-testing. However, as also suggested by a number researchers, the mechanism side of causal reasoning should be included as another core element of scientific reasoning. With the integration of causal reasoning, the involvement of mechanism becomes natural, since mechanism is one of the two fundamental elements of causal reasoning, i.e., covariation and mechanism. Building on the existing work in the literature on causal reasoning, the DMCR model explicitly defines two fundamental elements of causal reasoning, the data-covariation causal relation and the mechanistic causal relation, along with a number of reasoning processes that operate within and between these elements. Using these new definitions, causal reasoning and scientific reasoning can then be integrated into a single framework that is also operational for guiding assessment and instruction.

On the operational side, definitions of scientific reasoning skills in many existing studies are often broadly defined based on descriptions of the general processes and products of the reasoning. Examples of such definitions include "coordinate between theory and evidence," "identify a hypothesis," "search for appropriate evidence or alternative hypotheses," etc. These definitions lack the concrete operational constructs that constitute the targeted reasoning processes. In this research, the DMCR modeling framework explicitly defines the actual constructs, relations, and processes as well as a coherent theoretical model and concrete operations that are needed to understand, manipulate, and evaluate the constructs, relations, and processes. These together form the concrete building blocks and structures to operationally define the various skills involved in scientific and causal reasoning. This modeling framework can then provide explicit guidance on the development and implementation of instruction and assessment to promote specific skills involved in scientific and causal reasoning. The integration of scientific and causal reasoning in a single modeling framework provides a needed advancement in the research literature towards a better understanding of reasoning and knowledge development, while opening new venues for additional theoretical and empirical studies.

Synthesizing the theoretical work in this study, a comprehensive definition of scientific reasoning can be developed. In the existing literature, scientific reasoning is often broadly and implicitly defined as skills to support a range of learning activities involved in scientific inquiry, which can be considered as a type of behavioral definition. Based on the modeling framework developed in this study, the definition of scientific reasoning can now be extended with a conceptual component and an operational component. Together, these three components provide a more complete definition of scientific reasoning, including:

- Behavioral definition: the ability to support the activities and processes in scientific inquiry, which often involve efforts to systematically analyze a problem, identify researchable questions, formulate and evaluate hypotheses, make predictions, design

and evaluate experiments, analyze data, identify evidence, validate hypotheses, and make decisions based on evidence.
- Conceptual definition: a cognitive process that develops and manipulates data-covariation and mechanistic causal relations in knowledge formation and revision.
- Operational definition: the specific reasoning skills needed for control of variables, data analytics, and causal decision making.

This research also contributes to the literature on assessment of scientific reasoning. The existing assessment tools have been designed to measure a limited number of loosely connected skills that lack a coherent theoretical base, which limits the interpretation of the assessment outcomes. In comparison, the iSTAR assessment instrument has been specifically designed based on the DMCR modeling framework to measure a progressive set of skills that are defined as the essential constructs of scientific and causal reasoning. The validation study suggests that the assessment can provide well-targeted diagnostics of wide-ranging skills and subskills that are explicitly and operationally defined based on the modeling framework. Therefore, the results of the assessment can be directly mapped to specific skill sets and linked to components of the DMCR model, which facilitate the interpretation of the outcomes and provide meaningful insights on students' reasoning abilities. Such understanding can then directly guide instruction to address the teaching and learning of the targeted skills and the associated learning difficulties. In addition, the validity and reliability of the iSTAR instrument have been solidly established with a college population, such that the instrument can be readily implemented in research and teaching.

Because of the limited scope of a single study, this research has a number of limitations. The development of a modeling framework to integrate scientific and causal reasoning cannot be completed in one study and will need additional work for its validation and further refinement. The DMCR model proposed in this paper is based on an integrative synthesis of the existing models, all of which are supported with their own empirical studies. Therefore the validation of the new model is currently established in part by the existing empirical studies, which support the various components of the previous models that have been integrated into the new model. In addition, the assessment instrument and outcomes provide tangible manifestations of the model's concepts and ideas, which can be evaluated to indicate the validity of the model itself. In this research, the assessment results are in agreement with the model's expectation, which can provide additional evidence to support the validity of the DMCR model. Therefore, based on the literature and assessment outcomes, it is reasonable to consider that the new model has sufficient validation for its initial introduction.

Nevertheless, this research only provides an initial glimpse of what can be further developed and has a narrow focus on developing the operational definitions of skill sets that can provide a model-based framework to guide assessment design. In future studies, research is needed to further update the model to make connections to a much wider range of existing theoretical and empirical work in the literature. Additional research is also needed to establish detailed connections to knowledge development and inquiry learning, which are the learning goals supported by scientific and causal reasoning. Connections of the model to educational practices in developing skills and abilities in scientific and causal reasoning, as well as additional assessment studies, are also warranted. In particular, the validity and reliability of the iSTAR assessment need to be further established with additional populations from different age groups and education backgrounds.

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation Grants No. DUE-1712238 and No. DUE-1431908. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

APPENDIX: STATISTICAL EVALUATIONS OF THE iSTAR TEST

1. Access to the iSTAR test

The iSTAR test can be accessed and delivered though an online testing system at https://istartest.com/home. For inquiries regarding using the test, please contact the corresponding author for more information.

2. Statistical analysis results

Table VII provides the classical descriptive statistics of iSTAR items, which includes the classical item difficulty (fraction of correct answers), item discrimination (score difference between upper 30% and lower 30% of students), and point biserial correlation (r_{pb}) between an item's score and the total score of the test (item-total correlation). The results are calculated with the same dataset used in the Rasch analysis for the evaluation of the construct validity. The test reliability of the iSTAR is evaluated by the Cronbach's α, which is found to be 0.737 with all 35 items included. Two of the items (30 and 33) have insignificant item-total correlations with this college population. If the two items are removed, the Cronbach's α becomes 0.750. The analysis indicates that iSTAR has adequate acceptable reliability. The two items are kept in the test because different populations may respond differently, and the two items are part of item groups needed for the completeness of the related context scenarios.

TABLE VII. Basic descriptive statistics of iSTAR ($N = 378$).

Item	Difficulty	Discrimination	r_{pb}	Item	Difficulty	Discrimination	r_{pb}
1	0.889	0.177	0.226	18	0.862	0.191	0.213
2	0.441	0.481	0.346	19 and 20	0.352	0.665	0.505
3	0.243	0.405	0.399	21	0.238	0.627	0.601
4	0.476	0.575	0.434	22	0.960	0.085	0.211
5	0.331	0.575	0.498	23	0.902	0.177	0.241
6	0.770	0.245	0.197	24	0.294	0.686	0.594
7	0.799	0.335	0.330	25	0.870	0.224	0.232
8	0.578	0.509	0.411	26	0.296	0.321	0.283
9	0.439	0.351	0.320	27	0.204	0.343	0.346
10	0.947	0.092	0.180	28	0.791	0.352	0.349
11	0.320	0.392	0.378	29	0.643	0.450	0.385
12	0.037	0.044	0.128	30	0.035	0.040	0.035
13	0.426	0.259	0.219	31	0.194	0.431	0.451
14	0.053	0.127	0.276	32	0.143	0.162	0.179
15	0.032	0.063	0.135	33	0.067	0.044	0.067
16	0.854	0.290	0.333	34	0.457	0.596	0.481
17	0.259	0.368	0.361	35	0.100	0.181	0.303
				Average	0.462	0.313	0.309

3. Rasch analysis

To investigate whether the three subskills represent distinct dimensions of students' scientific thinking and reasoning, a three-dimensional model was compared to a one-dimensional model, assuming only one latent construct underlying the data (namely scientific thinking and reasoning as a whole). A comparison between the two models shows that the model fit parameters are in favor of the three-dimensional model. Likelihood ratio tests indicate that the three-dimensional model shows statistically significant improvement in model deviance compared to the one-dimensional model ($\chi^2 = 65.793$, df $= 5$, $p < 0.001$).

The model parameters for the three-dimensional model were further explored to substantiate the above results on mode fit. Here, the weighted and unweighted mean square residuals (MNSQ), which are listed in Table VIII, were used to examine the extent to which students' response to iSTAR fit with the Rasch model at the item level. As shown

TABLE VIII. Measures of item difficulty, and fit statistics (infit and outfit MNSQ) estimated by Rasch model for iSTAR. Note that the average item difficulty is constrained to be zero. An asterisk (*) next to a parameter estimate indicates that it is constrained. Items 30 and 33 were not included in this model fitting analysis due to insignificant item-total score correlations.

Item	Item difficulty	Unweighted MNSQ	Weighted MNSQ	Item	Item difficulty	Unweighted MNSQ	Weighted MNSQ
1	−1.998	1.16	1.12	17	0.397	1.18	1.03
2	0.213	1.01	1.02	18	−2.944	1.09	0.98
3	1.192	1.01	0.97	19 and 20	−0.118	0.94	0.96
4	0.723	1.11	1.06	21	2.145	0.65	0.80
5	1.538	1.02	1.04	22	−3.428	0.80	0.97
6	−1.370	1.12	1.05	23	−2.436	0.94	0.97
7	−1.555	0.88	0.94	24	1.776	0.73	0.82
8	−0.400	0.94	0.95	25	−2.109	0.90	0.97
9	−0.551	1.13	1.08	26	0.898	1.10	1.02
10	−2.898	1.42	1.06	27	1.440	0.99	1.00
11	0.050	1.02	1.00	28	−1.110	0.96	1.03
12	2.836	1.30	1.08	29	−0.176*	1.02	1.04
13	0.275	1.09	1.08	31	0.832	0.94	0.94
14	3.045	1.22	1.02	32	1.900	1.08	1.08
15	3.003	1.70	1.09	34	−0.632*	0.94	0.96
16	−2.872	0.95	0.94	35	2.334*	0.99	1.00

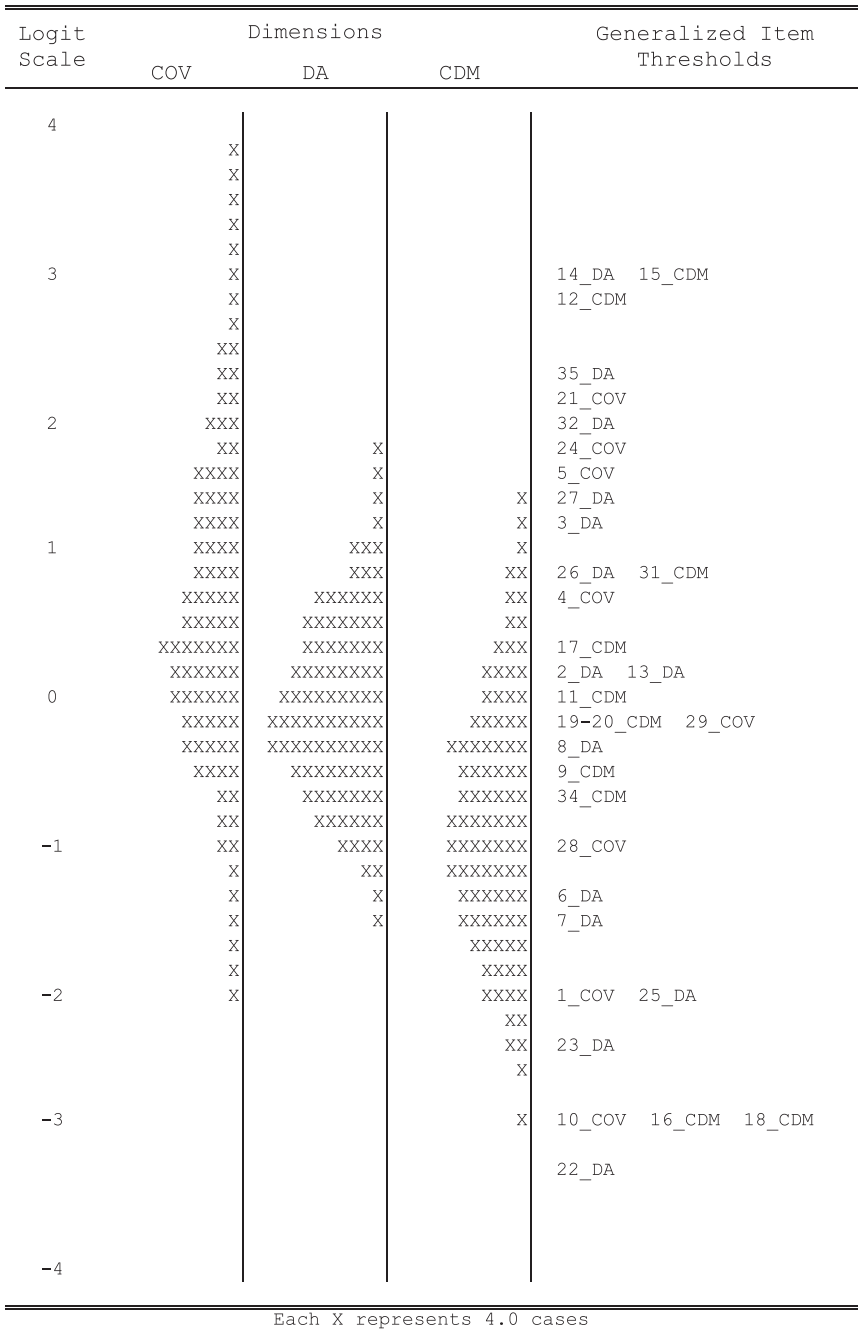


FIG. 15. Wright map of iSTAR.

in Table VIII, all items of iSTAR, except items 10, 12, and 15, appear to meet the standards for item fit ($0.7 < \text{MNSQ} < 1.3$). For items 10, 12, and 15, while the Unweighted MNSQ slightly exceed the suggested range, the weighted MNSQ meets the standards well. Hence, items 10, 12, and 15 are not removed from the following analysis. Based on the three-dimensional model, a Wright map is plotted in Fig. 15.

Results in Table VIII reveal that the measures of item difficulty cover a sufficiently broad range from −3.428 (easy) to +3.045 (hard). The Wright map shown in Fig. 15 provides further details for how the item difficulty is distributed among items of the three skill sets. A Wright map shows the ability measures of individual students and the difficulty measures of individual items on the same logit scale to allow clear mapping between item difficulty and student performance. The iSTAR data were analyzed with a three-dimensional Rasch model, which measures student abilities alone the three skill sets including control of variables, data analytics, and causal decision making. As shown in Fig. 15, item difficulties are broadly distributed across the three skill dimensions, which provide good

coverage on the different skills. The measures of student abilities also show desired near-normal distributions in all three skill dimensions with sufficiently wide spans on the ability scale. The centers of the ability distributions on the three skill dimensions also demonstrate the expected difficulty progression with COV being the easiest (highest average student ability), DA being the intermediate, and CDM being the hardest (lowest average student ability). Overall, the results shown in the Wright map suggest that the iSTAR test has a satisfactory coverage of the three skill dimensions and provides measurement outcomes in agreement with the expectations of the design.

[1] United States Chamber of Commerce, Bridging the Soft Skills Gap: How the Business and Education Sectors are Partnering to Prepare Students for the 21st Century Workforce, *Center for Education and Workforce* (U.S. Chamber of Commerce Foundation, Washington, DC, 2017).

[2] NGSS Lead States, *Next Generation Science Standards: For States, By States* (The National Academies Press, Washington, DC, 2013).

[3] Science Standards Advisory Committee, *College Board Standards for College Success: Science* (College Board, New York, 2009).

[4] National Research Council, *Assessing 21st Century Skills: Summary of a Workshop* (National Academies Press, Washington, DC, 2011).

[5] National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Academies Press, Washington, DC, 2012).

[6] National Research Council, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century* (National Academies Press, Washington, DC, 2012).

[7] National Science and Technology Council, *Charting a Course for Success: America's Strategy for STEM Education* (Office of Science and Technology Policy, Washington, DC, 2018).

[8] P. A. Facione, *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction—The Delphi report* (California Academic Press, Millbrae, CA, 1990).

[9] A. Fisher, *Critical Thinking: An Introduction* (Cambridge University Press, Cambridge, England, 2001).

[10] M. Lipman, *Thinking in Education*, 2nd ed. (Cambridge University Press, Cambridge, England, 2003).

[11] M. Binkley, O. Erstad, J. Herman, S. Raizen, M. Ripley, and M. Rumble, *Draft White Paper Defining 21st Century Skills* (ACTS, Melbourne, 2010).

[12] E. M. Glaser, *An Experiment in the Development of Critical Thinking* (Teachers College, Columbia University, New York, 1941).

[13] R. H. Johnson and B. Hamby, A meta-level approach to the problem of defining 'critical thinking', Argumentation **29**, 417 (2015).

[14] P. A. Facione and C. A. Gittens, *Think Critically*, 3rd ed. (Pearson, Boston, 2016).

[15] D. F. Halpern, *Critical Thinking Across The Curriculum: A Brief Edition of Thought & Knowledge* (Routledge, London, 2014).

[16] R. H. Ennis, Critical thinking: A streamlined conception, *The Palgrave Handbook of Critical Thinking in Higher Education* (Palgrave Macmillan, New York, 2015), pp. 31–47.

[17] H. Siegel, *Educating Reason: Rationality, Critical Thinking and Education* (Routledge, New York, 1988).

[18] R. Paul, *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World* (Center for Critical Thinking and Moral Critique, Rohnert Park, CA, 1990).

[19] C. Zimmerman, The development of scientific reasoning skills, Dev. Rev. **20**, 99 (2000).

[20] L. Bao, T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, L. Ding, L. Cui, Y. Luo, Y. Wang, L. Li, and N. Wu, Learning and scientific reasoning, Science **323**, 586 (2009).

[21] M. A. Johnson and A. E. Lawson, What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes?, J. Res. Sci. Teach. **35**, 89 (1998).

[22] A. M. L. Cavallo, M. Rozman, J. Blickenstaff, and N. Walker, Learning, reasoning, motivation, and epistemological beliefs: Differing approaches in college science courses, J. Coll. Sci. Teach. **33**, 18 (2003), https://my.nsta.org/resource/6003.

[23] S. T. Kalinowski and S. Willoughby, Development and validation of a scientific (formal) reasoning test for college students, J. Res Sci Teach. **56**, 1269 (2019).

[24] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[25] H. She and Y. Liao, Bridging scientific reasoning and conceptual change through adaptive web-based learning, J. Res Sci Teach. **47**, 91 (2010).

[26] M. S. Cracolice, J. C. Deming, and B. Ehlert, Concept learning versus problem solving: A cognitive difference, J. Chem. Educ. **85**, 873 (2008).

[27] S. Ates and E. Cataloglu, The effects of students' reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics, Eur. J. Phys. **28**, 1161 (2007).

[28] J. L. Jensen and A. E. Lawson, Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology, CBE Life Sci. Educ. **10**, 64 (2011).

[29] A. E. Lawson, The development of reasoning among college biology students—a review of research, J. Coll. Sci. Teach. **21**, 338 (1992).

[30] D. Kuhn, Thinking as argument, Harv. Educ. Rev. **62**, 155 (1992).
[31] V. F. Shaw, The cognitive processes in informal reasoning, Think. Reas. **2**, 51 (1996).
[32] A. Zeineddin and F. Abd-El-Khalick, Scientific reasoning and epistemological commitments: Coordination of theory and evidence among college science students, J. Res. Sci. Teach. **47**, 1064 (2010).
[33] K. Koenig, K. E. Wood, L. J. Bortner, and L. Bao, Modifying traditional labs to target scientific reasoning, J. Coll. Sci. Teach. **48**, 28 (2019), https://my.nsta.org/resource/117343.
[34] J. Osborne, S. Rafanelli, and P. Kind, Toward a more coherent model for science education than the crosscutting concepts of the next generation science standards: The affordances of styles of reasoning, J. Res. Sci. Teach. **55**, 962 (2018).
[35] T.-R. Sikorski and D. Hammer, Looking for coherence in science curriculum, Sci. Educ. **101**, 929 (2017).
[36] A. E. Lawson, Lawson Classroom Test of Scientific Reasoning (2000), http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Mathematics Assessments/Scientific Reasoning Test.pdf.
[37] L. Bao, Y. Xiao, K. Koenig, and J. Han, Validity evaluation of the Lawson classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020106 (2018).
[38] C. Zimmerman, The development of scientific thinking skills in elementary and middle school, Dev.Rev. **27**, 172 (2007).
[39] J. Piaget, *Construction of Reality in the Child* (Routledge, London, 1954).
[40] A. E. Lawson, The nature and development of scientific reasoning: A synthetic view, Int. J. Sci. Math. Educ. **2**, 307 (2004).
[41] D. Klahr, *Exploring Science: The Cognition and Development of Discovery Processes* (MIT Press, Cambridge, MA, 2002).
[42] D. Kuhn, M. Pease, Wirkala, and Clarice, Coordinating the effects of multiple variables: A skill fundamental to scientific thinking, J. Exp. Child Psychol. **103**, 268 (2009).
[43] A. E. Lawson, Development and validation of the classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978).
[44] A. E. Lawson, *Science Teaching and the Development of Thinking* (Watsworth Publishing Company, Belmont, CA, 1995).
[45] A. E. Lawson, Using the learning cycle to teach biology concepts and reasoning patterns, J. Biol. Educ. **35**, 165 (2001).
[46] D. Klahr and K. Dunbar, Dual space search during scientific reasoning, Cogn. Sci. **12**, 1 (1988).
[47] D. Kuhn and D. J. Dean, Connecting scientific reasoning and causal inference, J. Cognit. Dev. **5**, 261 (2004).
[48] D. Kuhn, S. Ramsey, and T. S. Arvidsson, Developing multivariable thinkers, Cognit. Dev. **35**, 92 (2015).
[49] Z. Chen and D. Klahr, All other things being equal: Acquisition and transfer of the control of variables strategy, Child Development **70**, 1098 (1999).
[50] M. Bunge, *Causality. The Place of the Causal Principle in Modern Science* (Harvard University Press, Cambridge, MA, 1959).
[51] F. Halbwachs, Réflexions sur la causalité physique. Causalité linéaire et causalité circulaire in *Les théories de la causalité* (Paris, PUF, 1971).
[52] J. Piaget, Causalité et opérations, *Les explications causales* (PUF, Paris, 1971).
[53] R. Harré, *The Philosophies of Science* (Oxford University Press, Oxford 1972).
[54] J. Ogborn, Approche théorique et empirique de la causalité, Didaskalia **1**, 29 (1993).
[55] R. K. Guenther, *Human Cognition* (Prentice Hall, Upper Saddle River, NJ, 1998).
[56] R. Corrigan and P. Denton, Causal understanding as a developmental primitive, Dev. Rev. **16**, 162 (1996).
[57] F. C. Keil, *Concepts, Kinds, and Cognitive Development* (MIT Press, Cambridge, MA, 1989).
[58] A. Michotte, *La perception de la causalité* (Vrin, Paris, 1946).
[59] H. H. Kelley, The process of causal attribution, Am. Psychol. **28**, 107 (1973).
[60] M. Bullock, R. Gelman, and R. Baillargeon, The development of Causal Reasoning, in *The Developmental Psychology of Time* (Academic Press, New York, 1982), pp. 209–254.
[61] W. Hung and D. H. Jonassen, Conceptual understanding of causal reasoning in physics, Int. J. Sci. Educ. **28**, 1601 (2006).
[62] C. Chen, L. Bao, J. C. Fritchman, and H. Ma, Causal reasoning in understanding Newton's third law, Phys. Rev. Phys. Educ. Res. **17**, 010128 (2021).
[63] L. Bao and J. C. Fritchman, Knowledge integration in student learning of Newton's third law: Addressing the action-reaction language and the implied causality, Phys. Rev. Phys. Educ. Res. **17**, 020116 (2021).
[64] D. Kuhn, E. Amsel, and M. O'Loughlin, *The Development of Scientific Thinking Skills* (Academic Press, Orlando, FL, 1988).
[65] B. Koslowski, *Theory and Evidence: The Development of Scientific Reasoning* (MIT Press, Cambridge, MA, 1996).
[66] J. Biggs and K. Collis, *Evaluating the Quality of Learning: The SOLO Taxonomy* (Academic Press, New York, 1982).
[67] M. C. Linn, The knowledge integration perspective on learning and instruction, *The Cambridge Handbook of The Learning Sciences* (Cambridge University Press, New York, 2005), pp. 243–264.
[68] W. Whewell, *The Philosophy of the Inductive Sciences, Founded Upon Their History* (Johnson Reprint, New York, 1840).
[69] A. E. Lawson, Hypothetico-deductive method, *Encyclopedia of Science Education* (Springer, Dordrecht 2015).
[70] J. C. Moore and L. J. Rubbo, Scientific reasoning abilities of nonscience majors in physics-based courses, Phys. Rev. ST Phys. Educ. Res. **8**, 010106 (2012).
[71] J. S. Woolley, A. M. Deal, J. Green, F. Hathenbruck, S. A. Kurtz, T. K. Park, S. V. Pollock, M. B. T., and J. L. Jensen, Undergraduate students demonstrate common false scientific reasoning strategies, Thinking Skills Creat. **27**, 101 (2018).
[72] S. Zhou, J. Han, K. Koenig, A. Raplinger, Y. Pi, D. Li, H. Xiao, Z. Fu, and L. Bao, Assessment of scientific reasoning: The effects of task context, data, and design on student

reasoning in control of variables, Thinking Skills Creat. **19**, 175 (2016).

[73] P. W. Cheng, From covariation to causation: A causal power theory, Psychol. Rev. **104**, 367 (1997).

[74] S. P. Norris, L. M. Phillips, and C. A. Korpan, University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty, Public Understanding Sci. **12**, 123 (2003).

[75] R. C. Adams, P. S. Sumner, S. Vivian-Griffiths, A. Barrington, A. Williams, J. Boivin, C. Chambers, and L. Bott, How readers understand causal and correlational expressions used in news headlines, J. Exp. Psychol. Appl. **23**, 1 (2017).

[76] P. C. Wason, Reasoning about a rule, Quarterly J. Exper. Psychol. **20**, 273 (1968).

[77] X. Liu, *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (Information Age Publishing, Charlotte NC, 2010).

[78] P. Kline, *A Handbook of Test Construction (Psychology Revivals): Introduction to Psychometric Design*, 1st ed. (Routledge, London, 2015).

[79] B. Thompson and L. G. Daniel, Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines, Educ. Psychol. Meas. **56**, 197 (1996).

[80] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd ed. (Routledge, New York, 2015).

[81] S. A. Culpepper, The Reliability and Precision of Total Scores and IRT Estimates as a Function of Polytomous IRT Parameters and Latent Trait Distribution, Appl. Psychol. Meas. **37**, 201 (2013).

[82] F. B. Baker and S.-H. Kim, *The Basics of Item Response Theory Using R* (Springer International Publishing, New York, 2017).

[83] R. F. DeVellis, *Scale Development: Theory and Applications* (Sage Publications, Thousand Oaks, CA, 2012), Vol. 26.