教育测量与评估 . 研究论文

科学思维的理论模型与量化评估

(Theoretical Model and Quantitative Assessment of Scientific Thinking and Reasoning)

包雷[*1], Kathleen Koenig[2], 肖洋 [3], Joseph Fritchman[1], 周少娜 [3], 陈诚 [4]

[1] 俄亥俄州立大学物理系, 哥伦布, 俄亥俄州 43220, 美国
[2] 辛辛那提大学物理系, 辛辛那提, 俄亥俄州 45220, 美国
[3] 华南师范大学物理与电信工程学院, 广州, 广东 510631, China
[4] 集美大学教育学院, 厦门, 福建 361021, 中国
[*] 通讯作者：bao.15@osu.edu

(收稿日期: 06/15/2022; 录用日期: 07/15/2022; 发布日期: 07/15/2022)

DOI：https://doi.org/10.37906/realcn.2022.1

摘要：科学思维能力一直被各种教育改革举措强调为一个核心领域，如“下一代科学标准（NGSS）”或美国大学理事会制定的“大学科学成功标准（CBSCSS）”等等，这些标准侧重于当代学生在未来所需的技能。虽然已有丰富的文献研究这些能力是如何在不同年级的学生身上发展的，但学界还未就其定义、模型或评估方式达成共识。为推进这一重要领域的研究，需要一个连贯的科学思维理论模型来指导实践教学和评估。几十年来，唯一可用于大规模应用的工具是“劳森科学思维课堂测试（LCTSR）”，但该工具已经被证明在有效性和可测量上限等层面存在一定短板，且其设计缺乏一个明确的建模框架来自证其所包含之技能的合理性。因此，当前迫切需要开发一个全面的科学思维建模框架和有效的科学思维评估工具，以满足“21 世纪的学习者”更为多样的能力需要。本文报告了一种新创立的科学思维理论模型框架及相应的科学思维评估工具，推动这一亟需领域的研究发展。该理论框架整合了科学思维和因果思维方面的研究从而建立一个完整融合的理论模型，并从操作上定义了在科学探究的过程中为达到知识发展的目标所需的思维技能和子技能。随后，该框架被用于指导科学思维能力评价量表的开发，并在大规模测试的基础上对量表的信效度进行了讨论。

关键词：科学思维、因果思维、科学探究、知识发展、评估

I. 引言

未来的经济和劳动力（发展）所需要的教育目标已经由内容钻研转向高层次能力的培养，包括思维能力、创造力和解决开放性问题的能力[1]。在 STEM 教育中，各种致力于推进“21 世纪的学习”的教育改革举措，如“下一代科学标准（the Next Generation Science standards, NGSS）”[2]或美国大学理事会制定的“大学科学成功标准（the College Board Standards for College Success in Science, CBSCSS）”[3]，都在基于未来社会的需求培养当代学生的能力，换言之，在 STEM 学科中，无论是“知识”还是

“能力”，都是带动未来经济和劳动力发展所不可或缺的[4-7]。在“21 世纪教育”强调的众多能力中，学生的科学思维和批判性思维能力是最受关注的，这与问题解决、决策和创造性思维所需的其他认知技能高度相关[8-10]。因此，它们在定义、评估和发展那些在 21 世纪科学标准[2，11]中被反复强调的技能和学习成果方面起着基础性作用。

既有文献已经对批判性思维[8，9，12-14]进行了广泛的研究，批判性思维被定义为“旨在并支持基于证据的决策的认知技能和策略”。它是一种包括问题解决、制定推论、计算可能性和做出决定的思维[15，16]，且被认为是理解和评估主题、产生可靠知识和改进思维本身的一种方式[17，18]。

同时，科学思维或科学推理的概念也经常被用来标记那些能够为“STEM 学习中的批判性思维、问题解决能力和创造力”提供支持的能力的集合。在既有文献中，“科学思维”和 “科学推理”这两个概念术语经常互换使用，（为免混淆）本文将通篇使用“科学思维”一词，并将其定义为一种广义的“在以发展知识为目标的科学探究过程中涉及的思维和推理技能，包括对问题进行系统性探索、建立和检验假设、控制和分离变量，以及观察并评估结果等等[19，20]。

批判性思维和科学思维有诸多异曲同工之处，它们都强调多变量因果条件下的基于证据的决策。在探究性学习中发展科学思维能力，可以促进批判性思维，培养学生识别可研究问题、提出假设、设计和实施实验、收集分析数据以及评估假设的能力。故在 STEM 学习场景中，科学思维可以被视为批判性思维在科学领域中的特定表达。因此，在教学中针对性地培养科学思维能力，与 NGSS 等倡议和推动的“21 世纪教育”所强调的目标是一致的。培养科学思维能力可以提升学生的批判性思维、开放性问题解决能力和决策能力。已有广泛的研究探讨了在教育过程中培养科学思维能力的重要性和益处，研究发现科学思维能力与课程成绩正相关[21-23]。提升科学思维能力可以帮助提升概念测试的成绩 [24，25]，获得更高水平的问题解决能力[26]，以及通过正迁移提升 STEM 内容学习的效果[27，28]。

然而，有研究显示，一部分大学生在科学思维方面依旧缺乏基本的技能。这表明这些技能在 K-12 或更高阶段可能没有得到发展。例如，劳森[29]发现，大约一半的生物学专业新生缺乏建立假设、控制变量和设计实验的能力。另一系列研究表明，本科生很难做出基于证据的决策，难以区分证据和主张并将其联系起来[30-32]。此外，也有研究提出，科学思维的技能在传统的 STEM 课程中很难发展，但有针对性的探究式教学可以有效地促进其发展[20，33]。

为了在正式和非正式的 STEM 教育场景中进行针对性的科学思维教学和评估，以培养学生的科学思维能力，学者和教师急需一套指导性的理论模型和有效的评估工具。为此，本文提出了一套科学思维过程的理论模型框架及相应的科学思维能力评估工具。其中，前者尤其重要，因为它提供了一个理论基础，使 NGSS 和 CBSCSS 强调的概念和学习成果得以整合到一个连贯的理论框架中，并为他们各自教学目标提供理论支撑和认知基础。例如，在 NGSS 强调的 21 世纪能力标准，主张通过三个维度的科学学习来建立对于科学的整合理

解，包括交叉概念、科学和工程实践以及学科核心思想[2]。尽管对 NGSS 标准的侧重点和组织原则尚有不同的看法和争论 [34，35]，但其中重要的概念和思维技能被公认为是科学思维和因果解释的学习成果。此外，对于那些在 NGSS 中被强调的相关概念和技能而言，在统一的理论框架和操作性定义中确立的思维技能也可以为其对应的教与学提供实用手段。

在评估方法上，劳森科学思维课堂测试(LCTSR)[36]在 STEM 教育界获得了广泛的使用。但问题在于，这一评估工具在设计之初就缺乏可靠的理论框架。它虽然声称测试是围绕皮亚杰的“形式操作思维”设计的，但 LCTSR 实际上根据自身的目的对其进行了重新定义，认为它包括“控制变量、假设检验、相关思维、概率思维、比例思维和守恒思维”。此外，一项近期的研究深入检查了 LCTSR 的评估特征，并最终发现了它在信效度上的数个短板，以及在以大学生为测试对象时的天花板效应[37]。虽然卡利诺夫斯基（Kalinowski）和威洛比（Willoughby）通过设计劳森测试的更新版(他们称之为蒙大拿州立大学正式思维测试(MSU-Fort)[23])解决了其中的一些问题，但他们承认需要更多的方法来定义和评估科学思维，并呼吁开发更具包容力的框架来指导评估的发展。因此，针对 21 世纪的学习者开发一套有效且现代化的科学思维能力评估工具至关重要。

为填补这一重要且亟需的研究空白，本文提出了一个整合的科学思维理论框架，并在此框架基础上开发并检验了一套科学思维能力的评估工具。在接下来的几部分，介绍这个新的理论模型，并且通过回顾和总结现有的科学思维方面的研究来论证其设计合理性。

II. 现有的科学思维模型

在当前的教育活动中，科学思维已被确立为“21 世纪学习者”的核心能力。而思维和推理作为一种认知能力，已经被心理学家和认知学家研究了几十年。齐默尔曼（Zimmerman）[19，38]曾对这方面的研究进行全面的回顾，并综述了从皮亚杰[39]开始的在认知领域的相关研究，包括劳森（Lawson）[40]、克拉尔（Klahr）[41]、库恩（Kuhn）[42]等人关于 STEM 场景下的思维过程的研究等一系列研究的发展谱系。

在现有的研究中，劳森在两方面做了大量工作：测量科学思维能力，以及帮助人们理解如何在基于探究的科学课程中培养相关的技能 [43，44]。借鉴皮亚杰的形式思维和发展阶段论理论，劳森确定了 6 个子技能作为评估科学思维能力的基础。在这些技能中，控制变量(COV)和假设演绎推理受到高度重视，因为它们为假设检验提供了基础，而假设检验正是科学探究的关键。在劳森的研究中，科学思维被认为在科学知识的产生中起核心作用。在他的科学教学方法中，科学思维技能被融入到科学探究的循环中，这一过程已被证明在帮助学生构建概念系统以及发展更有效的思维模式方面是有效的[45]。

在认知科学中，对思维的研究都已经非常广泛。在科学思维的主题下，有两个研究线索与本研究最相关，一是库恩关于多变量因果思维和理论-证据协调 [42]的研究，二是克拉尔的科学发现的双路搜索模型(SDDS)的理论框架和关于控制变量(COV)技能的实证研究

[41，46]。这两位研究人员都拓宽了科学思维的研究领域，对过去单纯以调查形式研究学生控制变量和因果思维能力的研究范式做出了突破。

库恩认为，科学思维是根据证据有意识、有目的地修正观点并产生新的理解的过程。这个过程被称为理论-证据协调[42，47]，库恩基于这一解释建立了一个综合性的推理过程框架，其主要结构包括：质疑现有的理论，确定可供选择的假设解释，寻找和确认证据(既有支持的，也有矛盾的)，并根据证据来评估和确定假设。新知识（这里的新知识应该不是指广义上的知识，而是特指个体建构的知识，译者注）是在学生既有的理论认知(包括误解)、基于数据的结果呈现(通过对照实验建立的协变关系)和科学接受的理论的交汇之中构建。这一建立新知识的协调过程包括通过思考各种类型的证据以在证据和解释之间形成一个有意义的联系网络，以及考虑未知但可能存在的因果因素的潜在影响，进而将其发展为新证据和新解释的组成部分或者是它们之间的联系。这些能力是至关重要的，因为它们代表了理解物理世界所需的各种推理。它们使学生得以在一个循环的探究过程中进行预测、推理并做出解释。库恩的工作强调了嵌入在各种推理场景中的因果关系的多变量性质，并证明了儿童和大学毕业后的成年人都在协调证据与解释方面缺乏有效的多变量推理能力[42，48]。

克拉尔的研究强调了先验知识在科学思维中的作用，并为捕捉和解释推理任务中的人类行为提供了一个理论框架，这一框架被命名为作为科学发现的双路搜索模型 (SDDS) [41]。SDDS 的框架包括假设（理论）空间、实验（数据）空间，以及包含（通过证据评估）协调这两个空间中的各种可能的思维过程和通路。该框架允许基于学生的先验知识、策略偏好和探究过程中生成的证据等要素，在假设和实验空间之间来回迁移。通过这种方式，该框架描绘了科学家在产生新的科学知识的过程中伴随的认知发展过程，这一过程非常复杂，不一定以直通的方式进行。

库恩和克拉尔提出的模型为更为有效的科学思维理论的建立提供了重要的研究基础。通过综合他们的研究，并进一步与劳森的科学思维子技能研究相结合，就可以形成一个指导科学思维教学和评估的操作性框架。我们在此提供了一个如何综合这些想法的例子：比如，对于一个给定的多变量因果思维任务，比如在库恩的研究[42]中发现的那些，学生既可以从理论角度开始他们的调查，并用既有的证据确定可能的解释(假设的因果关系)，也可以基于一套可能的理论解释，从实验的角度来开展他们的调查，以评估证据和解释之间的一致性。而无论是基于哪种角度的调查获得的结果，都可以进一步通过预测和推论提出新的实验路径或理论解释。这些探索和发现的途径与克拉尔的 SDDS 框架[46]中讨论的典型过程相似，也与劳森的学习周期理论中强调的假设-演绎推理和探究活动有着共鸣[44]。无论这个过程从哪里开始，或者从这两个角度同时出发，学生都需要熟练地在两者之间迁移，综合所有可能的解释和证据，以决定最佳的协调结果作为他们的新理解，而这正是库恩关于理论-证据协调的研究的中心要素[42]。这个过程通常以多个循环路径进行，并构成我们通常强调的探究性学习过程的思维推理基础。另外，这一过程也在很大程度上依赖于学生在控制变量、数据分析和因果决策方面相关的能力。在这些现有模型的基

础上，我们整合了当前的研究形成一个新的更完整的理论并以此定义了一个可操作的技能框架，用于开发针对科学思维的教学模块和评估工具。下一节将具体介绍这一理论框架。

III. 科学思维综合性理论模型框架的开发

从现有的科学思维建模的文献来看，有一些认知实体与科学思维的定义有着密切的关联甚至交叉关系。这些认知实体包括科学知识、科学探究和因果思维。它们三者之间的关系可以解释为，科学探究是一种由科学思维作支撑并建立在因果关系基础上的可以激发科学知识的认知过程。因此，在科学探究的循环过程中，知识的获取和思维之间有着强烈的交互作用。

在科学思维的众多宽泛定义中，因果思维是现有模型中强调的较为普遍和关键的元素。在多数关于检验假设的推理任务中都涉及在多变量条件下是否存在基于证据的因果关系这一命题。但是，在许多科学思维研究中因果思维却并未明确作为目标进行讨论。故而，因果关系中更精细具体的属性尚未明确解决或整合到当前的科学思维模型中。相反，对因果思维和科学思维的研究在某种程度上是作为两个相对独立的领域平行发展着。此外，目前研究所涉及的对于因果思维的定义通常很宽泛，并没有注意到它的特定属性，也没有确定科学思维和因果思维过程的联系。同时，因果思维本身已在哲学和认知科学领域中得到了广泛的研究，越来越多的科研人员投入到对其定义和结构的探究中来。由于因果思维和科学思维在其功能和性质上存在大量相通之处，而在以往的研究中并没有将二者很好的联系起来，为此接下来的内容将整合因果思维与科学思维的研究，厘清两种思维研究领域之间的区别和联系。

A. 因果关系和科学思维之间的联系

因果关系通常被认为是生成知识的思维和推理过程中最本质的组成成分，尤其是在科学领域。例如，概念性知识可以理解为对系统中基本组成部分和其因果关系的描述。此外，因果理解被认为是朴素物理学理论发展所必需的推理，使幼儿能够理解他们周围的世界。总之，因果关系理解对于知识的发展是不可或缺的。

A1. 因果关系的定义

因果关系的定义由来已久。在最近的研究中，明确强调了建立因果关系的三个要素，其包括：（1）潜在因果的时间顺序元素，其中原因必须在时间上先于效果；（2）协变量元素，它描述了基于对事件的实验观察而建立的因果变量之间的定量协变量关系；（3）机制元素，指将因果联系起来的机制理解和模型解释，是更精细化和更深层次的理论模型。这三个要素在最近的两项研究中得到了明确的定义[62,63]，共同构成了因果思维的基础。通常情况下，时间元素是因果关系的先决条件，人们在日常生活通常具有一定的基本认识。因此，本文将重点研究思维过程中的协变量和机制元素，这两个元素是构成完整因果关系的基本条件。通常在科学探究的发展过程中，人们对协变(实验证据)或机制元素(概念理论)的因果认识可能暂时领先对方，而它们的交替发展促进正是推动知识发展进程的基本过程。

从哲学和认识论的角度来看，人类是观察者，并基于观察对如何解释和预测某些事件做出推断。这种解释和预测是在理解潜在的因果关系和非因果关系的基础之上构建的。一般来说，因果关系的时间因素是定义因果关系的基础，导致事件因时间因素的变化发生协变。对协变量过程的描述通常以观察过程的初始和最终状态以及状态之间的变化过程作为基础。基于这些观察，一致的趋势和模式被识别出来，形成协变关系，这些关系可以被进一步归纳，从而对变化背后可能的机制过程做出推断。在知识生成过程中，要对初始状态和最终状态之间的变化进行观察从而产生对协变关系的理解，此外，对机制过程的推断也有助于对协变机制原因的理解。

图 1 说明了形成因果关系所涉及的三个要素及相关过程。通常来讲，人们的认知过程开始于观察多变量场景下发生的顺序事件。观察结果产生了协变行为的描述性数据，这些数据可以用来进一步提取以实验条件（例如控制变量）为基础的特定数据模式。然后，可以使用有效的协变关系数据来推断潜在的数学和逻辑关系，并且这些数据可用于归纳协变数据模式中存在的可能机制，还能对更多场景下的结果进行预测。同时，如果学习者遇到的事件和问题场景是熟悉的，它们还可以激活学习者已有的知识，从而帮助他们解释观察结果，或者在产生认知冲突时提示额外的处理方案。然后，通过反复的对协变数据和对应机理的解释进行协调和整合，可以产生对目标事件因果关系的综合理解，并将其整合到一个人的知识系统中。

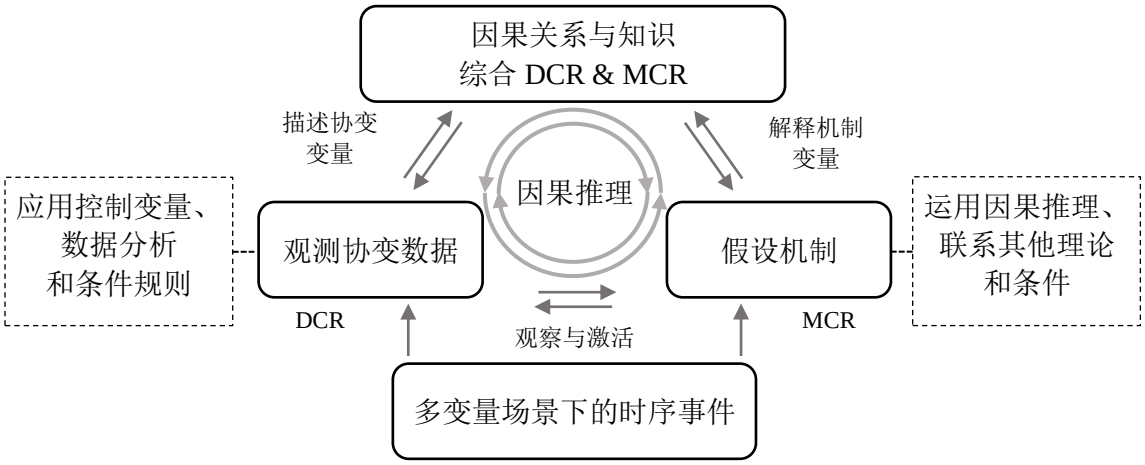


图 1. 有助于理解因果关系的基本要素和过程，它描述了按时间顺序发生的事件内在联系。通过观察，协变数据可以进行概括整合以便为因果关系提供经验证据。同时，假设的机制可以对因果关系和协变信息存在的原因和方式进行机制解释。二者都是对因果关系形成全面理解所必须的。

对于因果关系和科学思维之间的联系，现有的关于科学思维的模型通常强调确定因果关系时的协变证据，基于证据的判断往往被认为要优于基于理论的判断[38, 64]。同时，也有人考虑和研究过科学思维在因果机制中的重要性，认为学习者在思维过程中既包含了

协变信息，也包含了因果机制信息[65]。在本文的描述中，通过对科学思维和因果思维的整合，把协变和机制信息作为因果思维模型建立的共同基础。

这里综合有关文献和讨论得出，因果关系概念的操作性定义有三个基本组成部分，其中包括时间顺序、数据协变和机制。在因果思维方面，一旦建立了一个时间过程，就可以通过两条推理途径来探究和确定因果关系。一种途径是通过分析协变数据模式来确定数据协变关系的因果关系(DCR)。另一种则通过一系列假设的概念(理论)和数学逻辑，将因果联系起来探索基于机制的解释，这些被称为机制解释的因果关系(MCR)。DCRs 提供协变模式来启示或证明相关变量之间的因果关系，但不能(或缺乏必要的解释机制)解释导致观察结果的方式和原因。后者是 MCRs 的功能。在因果思维中，DCRs 为验证某个假设的因果关系机制提供了证据。同时，MCRs 为某些变量在特定条件下导致结果的方式和原因提供了解释机制。这些机制可以是纯粹的假设，没有任何现有的数据协变证据，如提出的新理论。另一方面，它们可以基于汇总的数据协变证据进行归纳和检验，这是对一个理论或假设进行归纳理论化和实验检验的过程。

A2. 通过科学思维和因果思维促进探究性知识的生成

在生成知识的科学探究过程中，DCRs 构成了大部分的实验证据，而 MCRs 则构成了理论(概念性)理解的基本组成部分。从认识论的角度来看，DCR 是基于观察对可能的因果现象的描述，而 MCR 是对潜在因果机制的假设和解释。这两个思维过程及其结果由学习者经历协调，相互验证，修改和发展等过程以推进对特定知识领域实验和理论的理解。若要完全理解某个特定主题的因果关系，则同时需要数据协变和机制解释，这二者提供了科学知识的基本结构。然而，在大多数科学领域，这种理解是一个不断发展的过程，实验和理论研究同时发展并在相互交织的发展过程中递次推进。

因果关系的两个方面在知识发展过程中的作用和功能有一些独特的特征。DCRs 是基于数据模式而缺少解释性的理解。因而，它们的原始形态需要大量的内存来存储不同条件和背景下由因果关系所表现出的各自数据模式。因此，学习者回忆和迁移以及理解这些知识往往是低效的。如果条件和背景超出实验所能验证的领域范围，其结果也是难以预测的。相比之下，MCR 通常以简单规则进行编码提供机制解释，这些规则是在大量实验验证的 DCRs 基础之上进行归纳所得。例如，计算电荷之间的作用力，为了确定不同条件下（包括电荷种类和电荷间距离）的 DCRs，研究人员进行了大量实验，这些 DCRs 的集合被统一概括为一种假设关系，即，$F = kq_1q_2/r^2$，这个方程可以对作用机制进行解释，它表明两个电荷之间存在相互作用力，并且他们的大小遵循这个简单的关系，该方程及其力学解释形成了电荷间力的 MCRs。这样的 MCR 可以被广泛应用于不同的条件和背景之下，例如多电荷和多距离问题，通过该方程也可以准确的预测或计算力的大小，其中大部分可以是未通过实验测量和验证的情况。

显然，将 DCRs 泛化为 MCRs 显著地减少了对此类关系进行编码存储所需的认知资源，并使这种类型的知识更易于迁移和存储，从而应用于拓展情境之中。在教育方面，被记录下来的 MCR 更方便在人与人之间进行传授,并记录下来供后代学习。因此，许多通常被定

义为科学知识的东西，大部分内容都以 MCRs 的形式作为基础，从先前的科学发展中积累起来的。同时，DCRs 提供了实验（观察）证据来证实假设的 MCRs（假设），对现有的 MCRs 进行进一步的修改和验证，并将新的 MCRs 归纳概括为改进的或新的知识。在这里需要注意的是，MCR 可以包括广义的数学和逻辑关系以及对有关于机理起源进行的解释。在某些情况下，对于某些问题的机制以及数学和逻辑关系的想法可以产生于可观测的 DCRs 之前。这些想法代表了纯理论性假设的 MCRs，通常被称作理论假设，需要通过实验进行验证以获得相关的 DCRs。

在获取 DCR 的过程中，通常可以使用分析和建模算法对协变数据进行处理和数字建模，以产生数学关系。因此，MCRs 和 DCRs 都可以包含数学和逻辑关系，因此数学和逻辑关系本身并不是区分 DCRs 和 MCRs 的特征。然而，基于 DCR 和基于 MCR 的数学和逻辑关系之间存在一些基本的差异，值得厘清。基于 DCR 的数学模型通常代表特定 DCRs 的局部计算建模（例如，回归）结果，不能在特定场景之外进行推广。此外，这些关系没有得到机制解释的支持，这进一步限制了它们的一般性应用。由于缺乏支持机制，基于 DCR 的数学和逻辑关系在机制解释上没有意义，因此很难在理论层面上进行操作。通过从广泛的场景中积累 DCRs，可以进一步验证所涉及的数学和逻辑关系，并与机制假设相结合，以开发 MCRs。当 DCRs 和 MCRs 被验证为一致时，基于 DCR 建立的数学和逻辑关系可以转换为基于 MCR 的数学和逻辑关系，这些关系可以进行机制解释，可以在理论上进行操作，并作为定律和原则普遍应用。

A3. 科学思维模型新框架的构建

在学习探究的过程中，DCRs 和 MCRs 都是知识生成过程中的重要基石。因此，知识的生成过程可以理解为一个整合 DCRs 和 MCRs 而构建经过实验验证并在机制上得到合理解释的因果认知的过程。这个过程包括很多方面，因为它通常涉及具有多个变量的场景、背景以及 DCRs 和 MCRs 之间的协调过程，以形成更加一致和全面的因果理解。我们将这种思维和知识形成的过程称为基于数据协变和机制解释的因果思维框架(DMCR) ，该框架的建立基于前文所述工作的整合提升。由于 DMCR 过程在 DCRs 和 MCRs 之间进行整合协调，因此可以将其视为一个双路径过程，用于对涉及因果决策和知识形成涉及的推理进行操作建模。从这个角度来看，DMCR 框架与 Klahr 科学发现的双重搜索模型 (SDDS) 框架和 Kuhn 的理论证据协调模型具有共鸣。然而， DMCR 框架有其进步的特点，尤其是因果关系和因果思维的结构和功能过程是明确的且可进行操作性定义的。此外，科学思维被认为是在知识发展过程中对各种类型关系进行直接描述和建立的过程，这将在以下内容中进行详细讨论。

综合考虑所有的组成成分，形成了如图 2 所示的多变量因果关系的网络示意图，以及因果决策和知识生成的相关推理过程。该图说明了科学知识、科学探究、因果关系和科学思维之间的关系。在 DMCR 框架的表述中，将科学思维的概念进行概括性定义，它涵盖了支持因果思维、探究学习和知识形成的所有功能和过程。

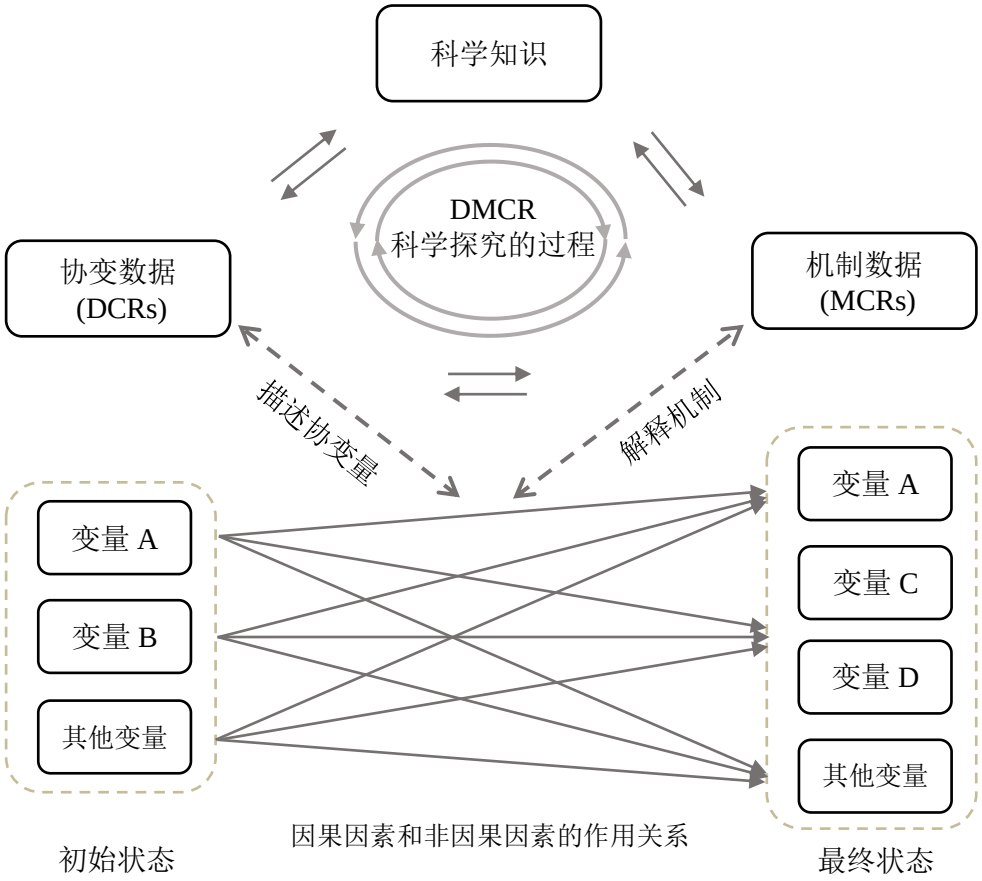


图 2. DMCR 科学思维的框架示意图，用来描述因果决策和知识的生成。设置的变量只是为了更形象的进行说明，并不代表任何具体实例。变量 A 表示初始状态和最终状态下都存在的特征，这些特征可能已经发生改变。变量 B 只存在于初始状态，而变量 C 和 D 仅在最终状态下出现。其他变量包括可能的受控、忽略或隐藏（未知）的变量。连接初始状态和最终状态变量的实线箭头表示可能的已知和未知变量之间的时间演化和场景交互，虚线箭头表示生成 MCRs 和 DCRs 的认知过程。

由于因果思维在科学探究和知识发展过程中起着至关重要的作用，因此该框架尤其注重其结构和过程的描述，如图 2 中的因果网络图所示。因果网络图表示一般时间顺序的因果事件，这些因果事件用初始状态、最终状态以及连接初始和最终状态的过程来进行描述。初始状态包含构成事件可能的因果和非因果因素的变量，而最终状态包含表示因果和非因果过程结果的变量。连接初始状态和最终状态的过程动态的描述了可能的因果和非因果动力学的条件及它们之间的作用机理。在图 2 所示的因果网络图中，初始状态的变量可以是因果的和非因果的，而最终状态的变量也可以是因果的和非因果的。过程中通常还涉及受控变量、未知或隐藏的其他变量。此外，由于测量条件的限制和外界环境的影响，描述初始状态和最终状态的变量集并不一定相同。

从初始到最终状态，可以分析描述在受控条件下的状态和收集发生变化的数据来识别和筛选 DCRs。在这里，DCRs 基于数据进行确认，描述从初态到终态的协变关系。然而，DCRs 没有描述连接始末状态的机制过程。这些过程可以解释变量如何以及为什么可能发生协变，通过 MCRs 用因果机制进行描述和说明。使用如图 2 所示的因果关系模型，可以将理论和假设视为基于机制的解释。这是因为它们从机制上解释了连接相关变量的因果关系的结构和时间演变网络，可以通过实验观察并根据 DCRs 进行定量描述。完整的因果理论是由变量及其关系所形成的复合网络中的 DCRs 和 MCRs 作为支撑进行综合描述的，这被称为基于数据协变和机制解释的因果思维 (DMCRs)，并通过因果思维过程来协调和重组 DCRs 和 MCRs。DMCRs 代表充分发展的因果关系，构成特定内容领域中理论或已建立科学知识的基础。对于跨越多个知识领域的交叉概念和理论，它们通常建立为多领域 DMCRs 的集成网络。

例如，引力的经典理论涉及多个变量，包括两个有质量的物体，两个物体之间的距离，以及观察到的两个物体之间相互作用力的结果。三个变量之间的协变关系可以通过实验确定为 DCRs 并抽象形成数学关系，$F = Gm_1m_2/r^2$，以及对该方程机的机制解释 (MCR)，即质量之间存在引力相互作用从而形成可观察到的力。总之，DCR 和 MCR 构成了经典引力理论的完整描述和解释框架。

对于科学知识，现有的大多数理论和假设都可以分解为相关变量之间的因果关系网络。一个成熟的理论将包括因果机制（MCRs）来解释协变关系（DCRs）的本质。以引力为例，该机制假设质量之间的引力相互作用导致引力的存在。然而，它充其量仍是一种基于观察的假设推断，其实际更深层次的机制仍未被理解。从哲学的角度来看，观察者可能永远无法理解或达到最终机制。

值得注意的是，在大多数情况下，DCRs 是由初始状态和最终状态之间的协变变化确定的，而 MCRs 解释了将系统从初始状态转换为最终状态的机制过程。然而，在某些情况下，DCRs 也可以是用来描述连接初始状态和最终状态的过程的特征定量分布。类似地，MCRs 也可以是用于确定初始和最终状态变量的基础机制。当特定领域的因果网络需要从根本上重构或扩展到其他网络时，会考虑这些附加属性，这部分内容超出了本文讨论的范畴。在这里，我们重点讨论的是 DCRs 和 MCRs 在特定知识领域的已建立因果网络中的主要功能。

图 2 中的多元因果网络也类似于贝叶斯网络的结构，该网络通常用于确定因果属性的概率特征。在定量因果决策中，贝叶斯概率在得出基于证据的结论中起着核心作用。因此，对多元因果网络和贝叶斯概率的理解和推理被列为评估科学思维的关键技能，本文稍后将对此进行详细讨论。

在以下内容中，将定义一些支持 DCR 和 MCR 推理路径的基本过程和元素。这些共同构成了构建科学思维和知识发展的 DMCR 框架模型的理论基础。这种更精细的细节对于创建一个评估框架并对相关的评估设计提供指导是至关重要的。

B. 因果网络和数据协变量关系的复杂性

因果关系的复杂程度必须进行清楚地描述，以便在操作上支持有效评估的发展。使用因果网络表示法，因果关系的复杂性可以用其因果网络的结构来进行建模。对于变量网络和连接关系，可以考虑两种类型的复杂性。第一种类型是由于网络结构而导致的复杂性，主要描述 DCRs 的功能。在这里，复杂性通常随着变量互连的数量而增加。

第二类是网络中单个变量及交互关系的概念和计算的复杂性，它们代表了 MCRs 的特征。例如，两个变量之间的关系可以是确定的，也可以是不确定的。所涉及的数学性质可以是简单的，如线性关系，中等复杂的，如二次函数和其他非线性函数，或复杂的，如递归和非连续函数。参考物理学中的一个具体例子，考虑机械能守恒的 MCR。在经典力学中，机械能守恒表示为经典定义的动能和势能的简单求和，而在量子力学中，机械能守恒表示为薛定谔方程，该方程适用于描述现实概率性质的波函数。这两种解释机理在概念上和计算上的复杂性大不相同，尽管这两种机制的总体思路在各自的领域内是相似的。

此外，明确变量和变量之间的关系也十分重要。具有隐式（或隐藏）变量和关系的推理任务通常比所有变量都显式提供有明显指示关系的任务更困难。显性或隐性变量的识别通常涉及 DCR 和 MCR 类型的推理。对一个可能变量的设定通常需要一个假设性的想法，即该变量为什么会影响和如何影响结果的机理，以及一些可用于检验该假设的现有或预测的协变现象。因此，在评估设计中，控制任务中变量并明确变量之间的关系有助于有意义地控制测试项目的难度水平。更复杂的设计可以使用隐藏变量和机理，此类任务需综合使用 MCRs 和 DCRs，因此可以有效评估学生的综合推理能力。

出于我们的目的，这两种类型的复杂性都将用于评估设计，这将在本文的评估部分详细讨论。基于 DCR 的复杂性主要是通过改变多变量因果网络的结构来控制的，而控制基于 MCR 的复杂性主要表现为涉及因果任务中隐藏变量和机理的不同配置以及条件设置来操纵的。为了帮助说明因果关系的复杂性，接下来将进行举例说明，以回顾典型的因果网络结构。

从简单到复杂，因果网络的结构可以用协变量之间不同数量和类型的联系来表示。具有代表性的列表可能包括（1）二元关系，（2）多元关系，（3）多元关系网络（也包括各级），以及（4）具有复杂耦合和反馈的连接网络的复杂系统，其中在某些条件下可能存在非线性动力学和混沌行为等不确定现象。图 3 显示了这些不同类型因果网络的几个通用示例。

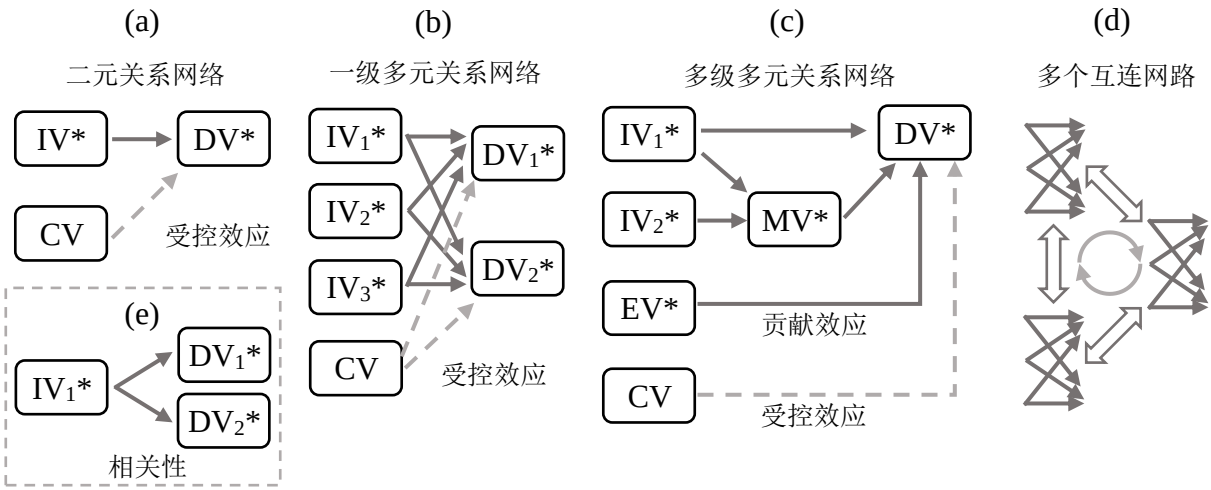


图 3. 不同变量之间关系的因果网络示意图。标有星号 (*) 的变量表示变化的变量，可能表现出协变或相关关系。实线代表协变关系，虚线代表不影响协变的受控效应。(a) 二元关系网络，(b)一级多元关系网络，(c)多级多元关系网络，(d)具有交叉耦合和递归反馈的多个互连网络的复杂系统。(e)是表示 DV1 和 DV2 之间具有相关性但是没有因果协变关系的特例。

在为分析 DCR 而测量协变数据的研究设计中，常见的变量类别通常包括自变量 (IV)、因变量 (DV)、控制变量 (CV)、中间变量 (MV) 和环境变量 (EV)。其中有些变量可能不可控制，对协变量结果会产生影响。在特定的研究设计中，所有这些变量可以是显式或隐式（隐藏）的存在于设计中，也可以具有已知或未知的机制和协变或相关关系。

对于 DCR 类型的协变因果关系，通常将自变量作为假设的原因，而因变量作为原因的结果。可以通过操纵自变量以形成特定的变化模式，导致因变量相应地发生协变，形成协变关系。在这里，协变的一个必要条件是控制变量，没有控制变量，协变数据模式只能被解释为具有相关性而不是协变关系（参见图 3e 作为示例）。在理想情况下，自变量（IV）不应该是更深层次未知变量的因变量（DV）。然而，这个假设通常无法从哲学的角度实现，但可以在操作上进行控制，使任何已知变量都不会成为 IV 的更深层次的自变量。

控制变量（CV）通常会影响 DV 的协变表达，因此，当 IV 发生变化时，应控制 CV 保持恒定或在不影响协变的范围内。这是构建协变实验的一个最基本原则。通常学生们对协变关系理解的一个典型缺陷是在协变和相关之间存在混淆（协变和相关关系的比较见图 3a 和 3e）。相关关系是两个具有已知或未知 DV 之间的共变关系，或者是在未控制变量情形下一个 IV 和一个 DV 之间的共变关系。相关关系不是协变关系，也不能用来支撑因果关系，即仅仅是共同变化并不能保证存在协变关系。为了建立协变的条件，实验必须包括控制变量、并获得可操作的 IV 和 DV 间协变的设计。当所有条件都满足时，共变数据方可以用来提炼和验证 IV 和 DV 之间的 DCR（例如简单的多变量情况见图 3b）。

对于更复杂的情况（如图 3c），则可能存在中间变量（MV），充当连接 IV 和 DV 的中间层。这些 MV 也可能是潜在的,且无法用给定的技术测量。此外，实验环境中还可能存在一系列固有的内在和外在变量，通常无法控制，但会影响研究设计中的各种变量。典型的情况下，这种影响需要保持在较小的水平或在分析中进行补偿，以便保证测量到的协变关系的有效性。由于这些变量影响许多相关变量的变化，因此它们被称为环境因素（EV）。

最复杂的因果关系是具有递归反馈的复杂网络关系系统（如图 3d）。这种结构通常是一个复杂系统，可以表现出非线性甚至不确定的混沌行为。在复杂系统中，每个连接的网络都保持其关系模式，这些模式还受到其他连接网络输入和输出的影响，而这些连接网络可以具有多种递归模式。在这种情况下，微观和宏观行为往往存在复杂的相互作用和相互关联，因此这样的关系具有本质的不确定性。

C. 因果思维过程的复杂性

当面对确定因果关系的任务时，学生的思维有一个从简单到复杂的发展过程。最初的简单思维通常是从根据场景特征识别相关变量开始。然后，通过回忆相似的已有理解或基于相关领域知识生成的新假设，从而构建变量之间的关系。这些关系从简单的二元形式到复杂的多元形式，可以是因果关系，也可以是非因果关系，本文的研究中强调因果关系。通常，二元关系是最先建立的，因为它们相对容易进行提炼和验证。接下来，若干相关的二元关系可以整合发展成多变量（IV/MV/CV）和结果（DV/MV）之间的多元关系。继续深入则可以通过整合多元关系进而构建连接多组多元关系的跨域网络，从而形成一个日益复杂的关系网络（例如，见图 3d），它最终可以演化为一个非线性耦合的复杂系统。总结思维过程和因果网络的相应结构，可以宽泛地定义五个复杂性层级，如图 4 所示。

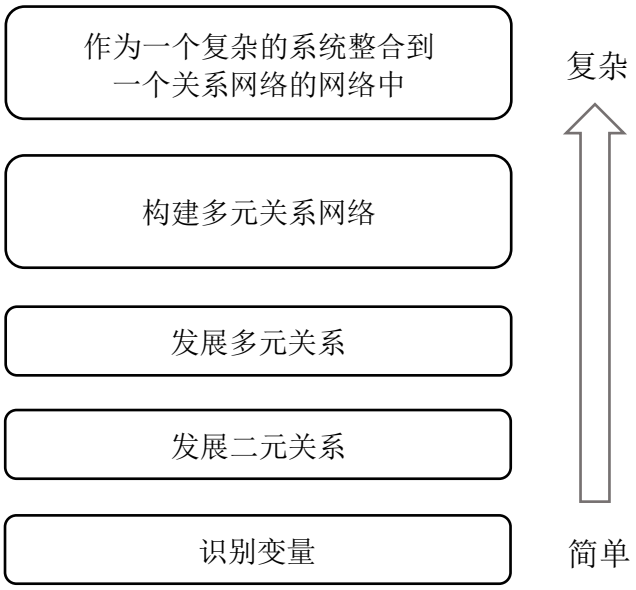


图 4. 因果思维过程的复杂性。

思维复杂性的层级与知识发展理论，尤其是学习领域特定内容的知识整合模型有许多共通之处[63,66,67]。例如，观察学习成果结构分类法（SOLO）[66]将学生的知识结构分为 5 个层次，包括前结构、单结构、多结构、多元关系和扩展抽象，所有这些都与上面讨论的思维复杂程度在结构上有相似之处。然而，知识整合模型是对学习者的知识结构上提供了一个宽泛的结构描述，而不关注具体的思维细节。在本研究中，我们的重点放在以知识发展为目标的思维技能上，并应用于针对这些思维技能评估设计。

还需要重点注意的是，图 4 所示的不同级别并不代表严格的发展过程。在学习和问题解决的过程中，实际的思维过程往往在多个层面并行进行，其间存在大量的交互作用。关于特定内容主题的学习，学生的思维可能表现为从简单到更复杂的发展趋势。然而，分支和递归过程很常见。例如，识别变量的过程将提示与相关变量的关系，并与之平行互动。当确定的变量和关系不能对任务场景形成令人满意的理解或解释时，将进行额外的周期识别和评估，使构建的理解和任务目标之间更好的匹配。因此，这种思维通常发生在知识构建的各个层次的多个递归循环中。

D. 知识生成的思维过程

DMCR 框架中的不同思维过程为知识开发提供了基本的认知支持，本节则讨论在问题解决和科学探究中所需的更细致层面的思维功能。例如，假设演绎模型[68,69]描述了类似的思维过程，被广泛认为是科学探究和学习的核心。为了对 DMCR 框架下的思维进行操作性建模，在这里定义了五个类型的思维过程和操作，包括“I-过程”、“D-过程”、“评估分析（EA）”和“循环（Loop）”。这些共同构成了在特定任务中执行 DMCR 思维具体功能的基础，如图 2 和图 3 所示。这种详细程度对于支持科学思维评估框架的开发是必要的。

I-过程代表广义的发现型思维，例如归纳、推断、发现等功能。它是一个创建或搜索要添加到当前思维中的新元素的过程。I-过程的结果包括广泛的认知内容，例如可能的变量、关系和机制（MCR 假设），这些通常对学习者来说是新的或未知的，甚至可能是先前不存在的（对世界来说是新的）。I-过程结果的有效性、合理性和有用性通常是不确定的，需要通过其他过程进行评估或验证。

D-过程代表广义的推演型思维，例如用于推理、推导、应用等的功能。它是一个将场景特征（变量）合并（导入）到给定（现有）规则或函数集中以生成确定结果的过程。D-过程的结果通常是“确定的”。也就是说，尽管一个结果可能对某个人来说是未知的，但它在概念上、数学上和逻辑上是保证有一个确定结果的。

在探究式学习中，D-过程通常与 I-过程创建的元素一起运行，以得出新的预测结果，对这些结果进一步处理可以评估 I-过程结果的有效性。这里评估分析（EA）过程用于分析和比较任务背景下的 I-和 D-过程的结果，并为结果和任务目标之间的一致性生成基于证据的决策。这类过程通常会经历多个循环。因此，整个过程在操作上可以理解为“发现-推演-评价-分析-循环”，故称之为发展思维的 IDEA-Loop 模型。

图 5 显示了思维功能的 IDEA-Loop 模型的示意图。IDEA-Loop 的任务、场景、发现的元素和推演分析等等的结果可以是认知操作中宏观微观各个层面的组成部分和过程。例如在神经计算层面，这些可以代表神经网络集群及其输入和输出的激活过程。在宏观行为层面，这些可以代表可观察到的认知状态，比如提出的假设，以及在场景中应用某些规则预测或导出结论。

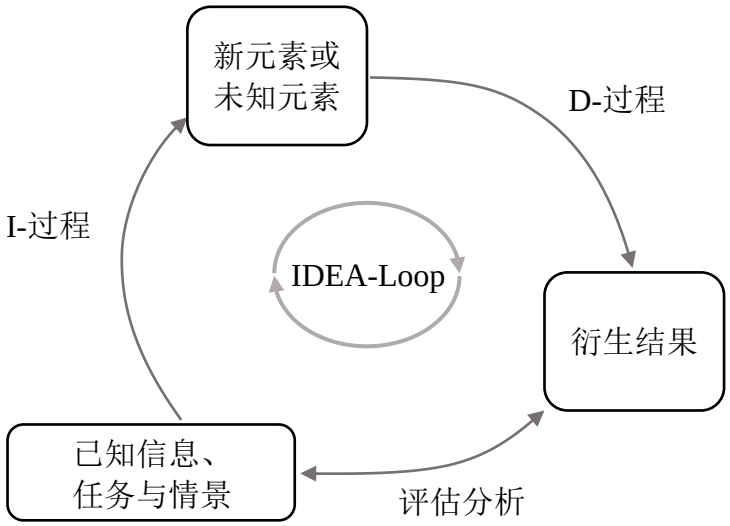


图 5. 思维功能的 IDEA-Loop 模型的概念图。

在实际应用中，IDEA-Loop 的某些部分可以是被关注的主要功能。例如，在解决涉及简单代入运算问题时，D-过程和之后的 EA 验证通常是主要操作。然而，当一项任务涉及 I-过程时，为了验证 I-过程的结果，整个 IDEA-Loop 通常会被激活，也就是说，由于 I-过程创建了新的元素，它便会自动激活 D-过程，应用新的元素来生成预测结果，然后通过 EA 过程，根据预测和观察结果之间的比较来验证新创建的元素有效性。如果需要修改，IDEA-Loop 的进一步循环将被激活。

在教学和学习中，也可以设计 DEA 过程之外的任务。例如，针对于 I-过程的任务可能会要求学生通过分析问题去寻求可作为问题解决的策略，但不需要完整地解决问题。然而，通常来说，在进行归纳搜索的过程中，学生可能仍然会直接或间接的参与到 IDEA-Loop 中。这是因为一个人需要在搜索中寻找合理的东西，因此需要 DEA 提供验证，这意味着 DEA 过程将贯穿于搜索、创造以及预测结果的验证和决策的整个思维过程。此外，I-和 D-过程的结果可以是所有层次上的所有形式的认知构型，包括变量、关系、理论、假设、新场景、新知识域等。因此，IDEA-Loop 会出现在所有具有广泛复杂性和抽象性元素的思维中。

在行为层面，思维的 IDEA-Loop 模型可以与几个相关模型进行比较，包括假设演绎推理模型[68,69]、理论证据协调[42]和科学发现的双路搜索模型（SDDS）[46]。在大多数

情况下，现有模型在思维的一般过程及其认知结果上有很大的相似性，例如确定有效证据和检验假设。另一方面，IDEA-Loop 模型提供了有关思维过程具有更精细的基本功能的的操作性定义，可以直接为评估和教学设计提供指导。在我们的工作中，这些功能元素将被提取出来作为用于测量科学思维技能的测评维度，这将在后文进行讨论。接下来将讨论 IDEA-Loop 和现有模型之间的联系。

假设演绎思维模型是对探究过程中科学方法的一种描述，其核心是提出假设并以数据为基础对假设进行检验[68]。劳森根据假设检验所需的一套科学思维技能对该模型的思维方面进行了研究。在评估这些技能时，通常会向学生提供实验数据，以确定一种假设的因果机制可以产生与数据一致的结果（例如，参见 LCTSR 中的问题 20-24）。在这种情况下，I-过程主要是基于给定场景和条件发现或建构假设的归纳思维。然后，通过 D-过程应用新建构的假设来生成预测结果，并将其与任务中的给定数据进行进一步比较，以评估和分析假设的有效性。

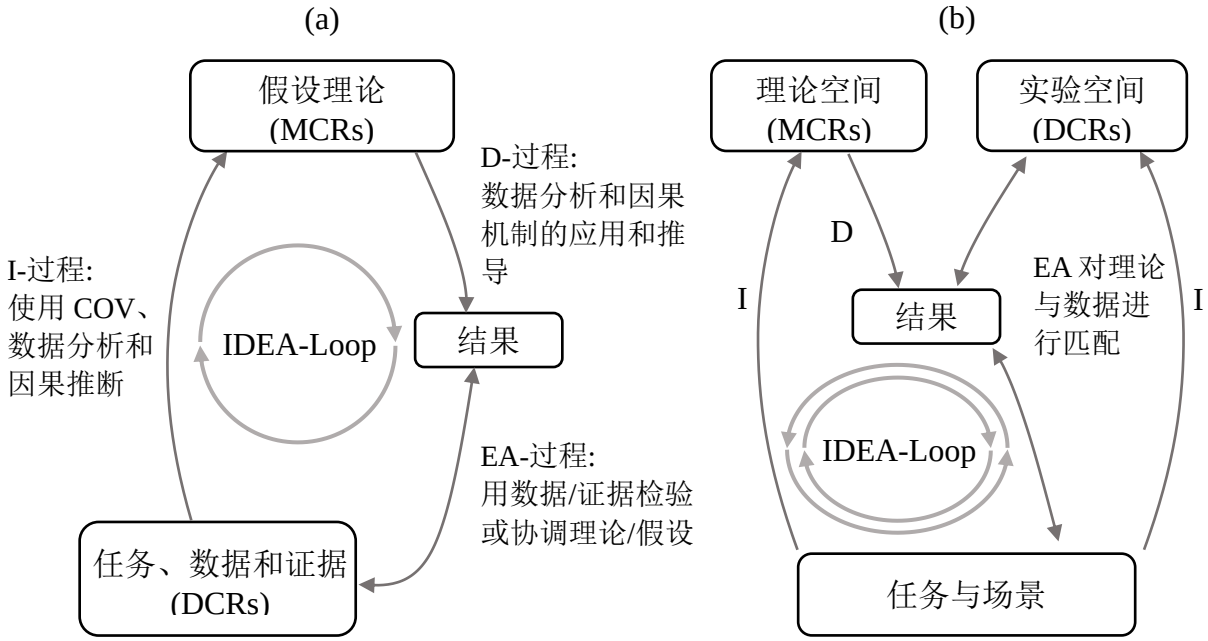


图 6. IDEA-Loop 思维模型的特殊情况。（a） 假设演绎思维和理论证据协调的 IDEA-Loop。（b）SDDS 作为多个并行嵌套的 IDEA-Loop。双头箭头代表双向发生的相互连接的双空间循环路径。

这里，假设生成部分可以被视为 I-过程针应用于产生假设的特例，而演绎思维可以被视为 D-过程应用于产生预测结果的特例。比较和验证是 EA 过程的一部分。因此，假设-演绎思维模型可以表示为 IDEA-Loop 模型的特例过程（图 6a）。比较两种模型，IDEA-Loop 模型比假设演绎模型更具通用性和灵活性。I-过程代表了一个通用的发现过程，是

针对多种类型的构建，而不是仅局限于假设-演绎模型中的假设解释。此外，IDEA-Loop 模型提供了所有相关功能和过程的具体定义，而这些在假设演绎模型中没有明确地定义。

对于理论证据协调模型，核心过程是实现证据（可以是给定或实验收集的数据）与假设或理论之间的一致性。这项任务通常涉及将数据集与给定的假设相匹配，或构建和修改假设以匹配数据。这些过程与假设-演绎模型中的过程类似，因此也可以被表达成 IDEA-Loop 在各种特定情形下的应用，但可能涉及不同处理层级的多个 IDEA-Loop，具体形式取决于特定的任务和场景（图 6a）。

SDDS 模型可以被视为另一种 IDEA-Loop 的变化形式，强调在实验和理论中发生的思维路径，因此具有双重空间结构。主要的过程是在两个空间中搜索，以确定实验证据和理论假设之间的一致性匹配。本质上，SDDS 也类似于理论证据协调和假设演绎模型的基本过程，即通过 I-过程寻找可能的理论假设（MCRs），目标是找到的假设可以通过 D-过程产生与实验数据相当的结果。同时可以并行的是在实验空间中通过 I-过程进行搜索数据协变关系（DCRs），并与理论空间中已识别假设的预测结果进行比较，以寻求获得与理论一致性的数据结果。在这些搜索比较的过程中，评估和分析是一个核心要素，被用于确定理论或实验数据是否一致，以及是否需要进一步循环迭代以获得更加一致的结果。SDDS 模型中比较独特的一点是，在实验和理论空间中的搜索通常以多个循环的形式出现，这可以用 IDEA-Loop 的多个并行过程来表示（见图 6b）。

总之，科学思维的 DMCR 模型通过定义 DCR 和 MCR 作为构建科学知识的因果基础，并在此基础之上整合了因果思维，利用 IDEA-Loop 循环过程进行功能性的建模，从而形成一个综合的理论模型。该模型同时也可以发展出一个操作性框架，提供具体功能的定义，用以表达现有的科学思维模型，包括假设演绎模型、理论证据协调模型和 SDDS 模型。具体化的科学思维技能和功能的操作性定义可以进一步指导评估设计。

E. 思维和知识形成的发展进程

构建科学思维评估框架的挑战之一是理解基于该框架所描述的学生思维能力应该包括哪些具体内容。本节为在学生思维中观察到的一些困难提供了可能的解释（主要针对人群是高中生和大学生），同时也为科学思维评估中应该包括哪些思维任务提供了依据。

根据发展心理学[39]的研究，在发展的早期阶段，孩子们观察周围的环境，并形成了对世界的理解。这些知识主要是以观测数据的形式，以及对物体和事件的协变模式的简单概括。因此，在这个发展阶段，知识和关联的思维大多是基于数据协变的因果关系（DCR）的形式出现，用于描述日常生活事件和现象。随着儿童语言的发展，基于机制解释的因果关系（MCR）的语言描述和解释开始发展，这种形式可以让他们无需自己进行观察和概括而进行直接交流和传授。因此，MCR 的出现可以大大提高学习已有知识的效率，尤其是那些不方便亲自体验或观察的事情上。在这个阶段，知识和思维的发展主要涉及学习 MCR。特定领域知识的变量和关系网络通常被通过记忆方式学习并被认为是已知事实，

而儿童的思维训练主要针对处理和验证知识构建中用到的关系、逻辑和计算等等。其中一些思维功能是跨领域的，可以应用于其他领域，以形成不同的领域特定知识集。

随着孩子们从生活环境中获积累经验，以及正式和非正式的学习经历，他们也会构建自己版本的 MCR 来理解这个世界。例如，大多数学生在物理方面会形成自己的认知概念，比如认为需要施加一个力来维持运动。这些素朴的概念是从学生们对物理世界的观察（DCR）和对观察的直觉解释（MCR）中发展出来的。例如，关于力与运动的一个常见误解是认为施加的力是物体运动的原因。这个观点是可以直观解释在有摩擦的世界中通常观察到的运动方式，而摩擦力本身则往往是大多数人没有明确或正确理解的潜在中介因素。如果不明确包括摩擦力，力与运动的关系就很难推广到牛顿力学的理解中。这个例子表明，学习者在生活经历中会自然而然地构建 DCR 和 MCR 作为其知识系统的一部分，并且这样的认知过程会自动地归纳 MCR，成为他们理解现实世界运作机制的一种有意义的解释。随着学习的发展，这些 DCR 和 MCR 可以进一步整合，形成更复杂的因果理解网络（DMCR-Net）。

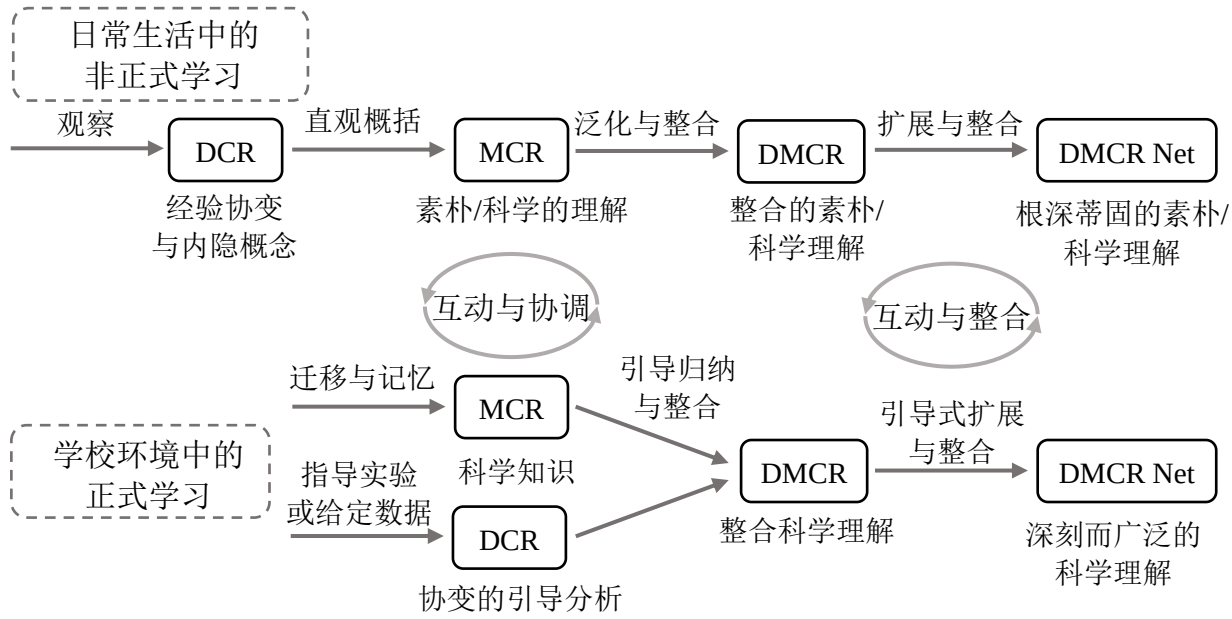


图 7. 知识和思维在特定内容领域内正式和非正式教育环境中的发展。特定年龄段的学习者（有一些差异）可能处于不同内容领域的不同阶段，而不同年龄段的学习者也可能处于相同或不同内容领域的类似或不同的阶段。

同时，在目前的教育体系中，学生很大部分科学知识是在正式教育中获得的，其教与学的重点是传授已确立的科学知识（即经过科学验证的 MCR）。通过多年接受学校教育，许多学生学会了主要依靠记忆获得知识，而思维技能主要是针对传授的 MCR 中所包含的计算和逻辑关系而训练学生的整合、操作和评估技能。因此，大部分学生很少经历探究式学习，而探究学习要求学生能通过数据进行观察从而构建 DCR，然后进一步归纳 DCR 形成假

设（MCR），并综合分析 DCR 和 MCR 进而评估测试假设的可信性。重视记忆性学习而缺乏探究的训练可能导致学生在基于探究的学习任务中不恰当地使用先前的知识或形成思维定势，而不是从数据中得出基于证据的结论。这样长期以往，可以表现为学生擅长死记硬背，但缺乏科学思维或批判性思维以及创造力等高阶思维能力。在某种程度上说，虽然每个孩子天生是具备自然的科学探究潜力并且在早期的自我学习中也是这么做的，类似一个小小科学家。但是传统的教育中的训练往往使学生习惯于被动地接受和记忆信息。这会使他们逐渐失去天生的好奇心，进而忘记怎样从自己的视角去观察周围世界，并建构基于观察的概念理解。如果不从根本上改变传统的灌输式的教育模式，学生则会因经历长期这样的训练而逐渐失去他们本就具有的探究和思考能力。

图 7 所示的流程图，描述了正规和非正规教育环境中知识和思维发展路径的典型特征。这个过程起始于年幼时期的观察和感知世界并形成 DCR，从而进一步概括归纳形成直觉性的概念和理解（MCR、DMCR 和 DMCR 网络的素朴版本）。随着语言的发展和交流的增加，尤其是在正式教育环境中的大部分时间里，学习和思维通常会过渡到学习 MCR 类型的知识。与此对应，学生构建 DCR 并将其与科学概念相结合的思维能力因为缺乏持续的训练将会逐渐缺失。因此，许多学生在提出可研究的问题、设计对照实验、处理和分析数据以及得出基于证据的结论方面的能力非常薄弱，而这些都是科学探究所需的关键技能。许多学习者一直处于这个阶段，直到成年。接受进一步研究培训的高级学术领域的学生最终可能会发展获得所需的科学思维能力，并在他们的知识构建中实现深度整合的因果理解网络（DMCR-Net）。具备这样能力的学习者能够在全新的领域通过探究的学习和研究方法构建新的创造性科学的知识。然而，有研究表明，相当一部分大学生仍然缺乏进行有效科学探究以及将 DCR 与 MCR 结合起来构建科学概念的必要技能[29,33,70,71]。针对学生思维能力不发达的问题，科学思维评估应重点关注构建和协调 DCR 和 MCR 方面的技能，这将在下一步讨论。

IV. iSTAR 评估框架和工具

本节将介绍我们开发的一个新的科学思维评估工具 iSTAR（Inquiry in Scientific Thinking, Analytics, and Reasoning）。这里使用“探究（Inquiry）”一词是为了表明 iSTAR 的主要开发目的是为学习和评估提供一个操作框架，以支持探究性学习。该框架可用于指导开发和评估旨在培养科学思维的探究式教学。此评估框架基于前文讨论的科学思维的 DMCR 理论模型和 IDEA-Loop 思维过程。

A. 定义操作评估框架和技能维度

从图 7 中讨论的学习和思维能力的发展进程来看，处于传统教育环境中的学生往往缺乏适当的思维技能训练，这些技能是发展 DCR、将 DCR 与 MCR 结合起来以得出基于证据的因果结论所必需的。因此，科学思维的评估框架是为了强调这些技能。例如，以发展 DCR 为目的的相关技能，是通过评估学生在从简单到复杂的情景中进行有效数据分析的能力来衡量的（如图 3、图 4）。与此对应，构建 MCR 的技能则根据学生处理来自先前知识偏见的能力，以及他们在不同场景中应对显性和隐性变量识别可能机制的能力进行评估。这些

评估任务的复杂程度可以通过调控变量数量、关系类型、因果条件等加以实现。此外，鉴于“有效协调理论和证据以得出有效的因果结论”这一能力至关重要，评估框架中还强调了整合 DCR 和 MCR 以评估和分析证据和假设之间关系的能力。这些能力是通过数据分析和因果决策任务来测量的。

基于科学思维的 DMCR 理论模型，iSTAR 评估框架定义了三个主要维度以描述思维技能和过程，其中每个维度也涉及多个子技能。这三个主要维度分别是控制变量（COV）、数据分析（DA）和因果决策（CDM），如图 8 所示。子技能和测试项目列表如表 1 所示，并将在下一节中讨论。该列表展示了 DMCR 理论中的核心技能：使用控制变量设计协变实验、分析数据以提取有效 DCR，以及通过因果决策协调 DCR 和 MCR 以得出有效结论和构建新知识。基于此，我们在定义子技能的同时，也是在对“DMCR 是如何在功能上运作以支持科学思维”的过程进行操作性定义。

在对科学思维的过程和相关技能进行建模时，需要理清之前讨论的三个建模框架之间的关系。DMCR 理论模型提供了思维技能的概念基础，即科学知识是在使用科学思维和因果思维的科学探究过程中发展起来的（见图 1）。同时，图 5 所示的 IDEA-Loop 模型描述了科学思维和因果思维在科学探究的动态循环中的功能。最后，图 8 所示的 iSTAR 评估框架概述了不同思维技能领域的结构组件和交互关系，而这些都是可以定义和测量的。这些模型共同为描述、建模和测量科学思维技能提供了一个完整的理论和操作框架。

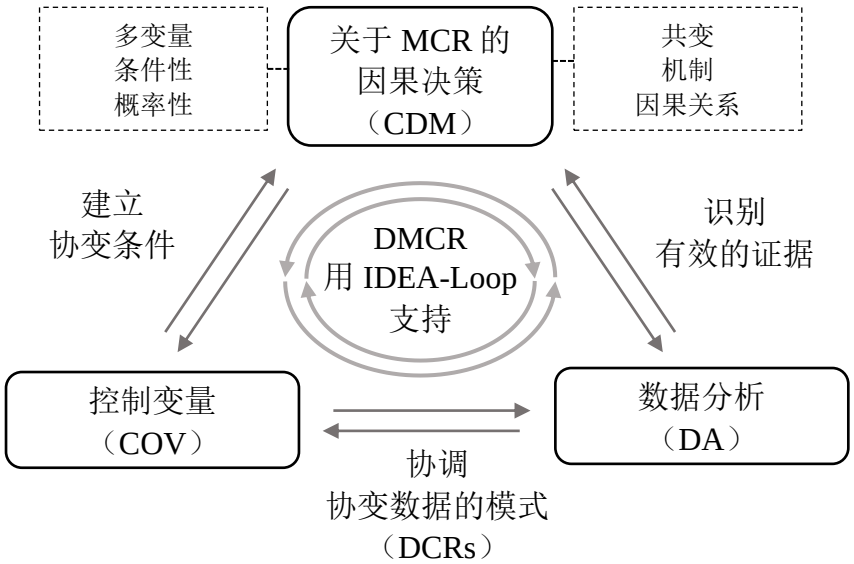


图 8. iSTAR 的科学思维评估框架

在这些技能中，控制变量（COV）是建立受控实验的第一步，从而获得自变量和因变量之间的协变数据，并形成 DCR（见图 3）。现有关于 COV 的研究较多，这为确定其子技能和项目设计提供了依据 [49, 72]。如前所述，COV 评估任务的复杂性可以通过调控场景熟悉程度、变量数量、数据呈现方式和变量间关系来控制 [72]。

数据分析（DA）是一个宽泛定义的维度，包括各种数据分析和解释技能，使学生得以识别有意义的数据模式和趋势，并评估其有效性，并构建和验证变量之间的 DCR。此外，这一维度在测评中也特别强调对条件概率的评估，因为条件概率被认为是因果决策中的一个关键因素［73]。总的来说，数据分析类别的子技能包括对比例、相关性、数据协变模式、条件概率和贝叶斯概率的评估和解释。这些技能支持 IDEA-Loop 中的 I-、D-和 EA-过程，用于发现、推导和评估有效证据和假设的约束条件，广泛用于协调假设（MCR）和证据（DCR）的过程中（如图 5、6）。

因果决策（CDM）也是一个宽泛定义的维度，侧重于“通过应用 COV 和 DA 过程的结果来综合分析 DCR 和 MCR，并最终得出有效的因果关系”的能力。这是 DMCR 的关键步骤，因为它整合了前两个过程（COV 和 DA）的结果，并试图在 DCR 和 MCR 之间进行因果关系的协调和确认。

在因果决策中，进一步定义了四个子类别的技能。首先是区分协变和相关关系的能力，学生们经常在这方面遇到困难，因为他们倾向于把相关结果解释为协变从而得出因果关系的结论［74, 75]。这种能力通常使用“在没有设置适当的控制变量的情况下表现出相关关系”的任务进行测量。这些场景中也可能涉及到隐藏变量和其他混杂的因素或关系，而这些都有可能被学生忽视从而使得任务变得更为困难。

第二个子类别涉及一系列的条件概率和贝叶斯概率评估技能，这些技能在涉及 DCR 和 MCR 的任务中都经常使用。这些概率方面的概念理解和计算操作对于正确预测概率结果，推断可能的原因或促成因素，以及确定因果关系至关重要。

第三个子类别包括确定因果关系的条件规则和逻辑规则。这些规则常见于给定或假设了某些因果前提和结果的 MCR 任务中，学生需要应用条件化的逻辑规则，以确定能正确匹配这些要求的证据。这些规则包括处理充分、必要、有贡献和不相关的条件。另外，在这一类的技能中，还包括在前向因果预测和后向果因推断中转化这些规则的能力，前者如基于给定原因的 D-过程中得出或预测结果，后者如基于观察结果的 I-过程中推断可能的原因。例如，如果 A 是 B 的充分条件，前向逻辑可以描述为“如果 A 存在，则 B 一定存在”。相应的反向逻辑则是“如果 B 不存在，那么 A 也不可能存在”。这些条件规则是重要的逻辑思维技能，学生需要知道它们并在 IDEA-Loop 过程中正确应用，以在多变量环境中的确定正确的因果关系。

第四个子类别包括构建和修改 MCR 的思维技能，这通常会涉及到特定领域的知识。在评估中，这一类别的技能可以进一步分成两个细类。一是分析确定假设和其支撑证据之间一致性的能力，即在 MCR 的基础上综合考虑 DCR 进行因果决策的能力。例如，LCTSR 包括四个测量有关假设验证的演绎推理能力的问题，其测量思路与基于机制的因果思维相同。为了回答这些问题，学生需要根据给定的假设，对不同条件的实验结果做出相应的预测（主要是通过 D-过程），或者通过 IDEA-Loop 来确定实验结果，以验证某些假设。二是理解和处理协变情况时的能力，在学生们原有知识的影响下，他们在处理 DCR 时的推理可能会出现偏差，在这种情况下 DCR 理解和推理能力发展不足的学生可能转而依靠基于 MCR

的知识，而不是协变数据，来作为支持结论的证据。例如，在 LCTSR 的相关关系问题（小鼠问题）中，学生被要求运用分析技能来评估具有不同特征的小鼠的数量是否具有相关性。然而，缺乏必要的数据分析技能的学生可能会用基于机制的理解来回答：例如“小鼠的大小和尾巴的颜色之间可能存在遗传联系”（LCTSR 第 20 题），即声称遗传机制可能是拥有深色尾巴的原因，但事实上这个机制性解释和题目所问问题的出发点是不相干的。

在完成具体推理任务时，这些技能会被组合使用，以支持 IDEA-Loop 的多途径运行。图 8 也展示了三个维度的思维技能之间的交互关系，并示意了不同思维技能在支持 IDEA-Loop 来协调 DCR 和 MCR 方面的主要功能。例如，在一个典型的任务中，控制变量维度的技能会被应用于建立受控的试验条件，这是收集协变数据的实验基础。这些条件由因果决策维度的技能来评估其协变或相关性质，其结果则被用作因果决策的证据（之一）。确定了控制变量条件，数据分析维度的技能则可用于分析所收集的数据，以确定特征的协变模式，作为因果决策的证据。这些技能共同作用于生成、评估和综合 DCR 和 MCR，以确定因果主张的有效性。出于对问题设计的考虑，因果思维任务的难度可以通过因果关系网的复杂性来控制。因果关系网的范围可以从简单的少数变量系统到复杂的多变量系统，而嵌入的关系可以是简单的线性关系、条件关系和复杂的概率关系（如图 3 和 4）。

在处理复杂的推理任务时，图 8 所示的三个维度的思维技能经常在 IDEA-Loop 的动态循环中相互补充。例如，当没有得出满意的结论时，因果决策环节的暂定结果可以重启实验，或操纵控制变量环节以修改实验条件。这种修改往往包括控制或改变不同的或附加的变量，以获得特定的测量设置或修改当前的协变条件。因果决策环节的结果也可以提供线索来指导数据分析环节，以确定新的或替代性的数据模式和关系，或使用一套不同的数据分析算法。然后，数据分析环节的结果可以反馈到控制变量环节的操作中，以达到改变协变条件和提高观察的数据模式的可靠性等目的。支持这些功能和过程的基本推理元素在归纳和演绎路径中以不同的复杂性和抽象级别经历多个 IDEA-Loop 循环。最终，这些功能和过程的组合提供了一个基于理论的操作框架，可以具体地评估科学思维和因果思维。

B. 科学思维测评工具的开发

在 iSTAR 评估框架的指导下，本研究进一步开发了测量学生科学思维的评估工具即 iSTAR 测试。当前的测试版本包含 35 道单选题，涉及三个技能维度：控制变量（COV），数据分析（DA），以及因果决策(CDM)。为了方便测试的开展，另一项研究也在进行当中，即把完整长度的 iSTAR 测试分成两个短版本的平行测试，每个短版本包含约 20 道题目。目的是在同一个被试群体中随机地使用两个短版本测试并能产生与完整版本等价的结果。表 1 总结了 iSTAR 的技能维度、子技能和问题的分布。

控制变量的子技能具有一个从易到难的进阶特征，包括对控制变量实验的简单识别和设计，到“给出实验数据并且要求学生确定某些变量是否具有因果影响”的更为复杂的情况[72]。问题的设计通过融入场景特征，例如真实生活情景和基于 STEM 的场景来控制任务的难度。总共有 9 个题目来评估 COV 维度及其子技能。

数据分析维度包含了数量最多的子技能，总共涉及 15 个问题。这些子技能彼此之间相对独立，且没有进行难度梯度设计。从表 1 中可以看出，DA 技能侧重于各种概率概念和评估技能，特别是条件概率和贝叶斯概率，它们对于在在概念层面上理解概率性条件的目的和需求，以及在操作层面对因果决策中的定量权重的计算，都发挥着重要的作用[73]。

如前所述，因果决策维度是基于证据和假设得出有效结论的关键能力。该能力包含 5 个子技能，总共 11 题，其中，理解相关与协变之间的差异，掌握因果条件规则是最为重要的。大量的研究表明，学生经常将相关关系视为基于协变的因果关系[74, 75]。培养这方面的能力将提高学生对科学实验和公共媒体报道的数据的解读能力。同时，因果条件规则提供了逻辑计算能力，它能使学生恰当地链接推理中的证据、观点、条件，以识别逻辑正确且证据一致的因果关系。这些都是使学生能够在不同条件下协调主张与证据的基本技能。

表 1 iSTAR 的评估维度、子技能的场景以及问题分布

技能维度	子技能的场景	问题
控制变量（COV）	• 识别或设计具有多个可检验和不可检验变量的 COV 条件 • 真实生活和 STEM 背景 • 有或没有实验数据 • 从简单到复杂的关系 • 扩展到 DA 和 CDM 维度	9 个 COV 问题：1, 4, 5, 10, 21, 24, 28, 29, 30
数据分析(DA)	• 多变量线性比例 • 组合 • 条件概率 （包括变式） • 多变量相关和协变 • 基础统计学，如加权平均和随机抽样的概念 • 贝叶斯概率	15 个 DA 问题：2, 3, 6, 7, 8, 13, 14, 22, 23, 25, 26, 27, 32, 33, 35
因果决策(CDM)	• 因果决策中的先验知识和偏见 • 相关和协变的因果判断 • 贝叶斯推理和因果决策 • 因果判断的条件规则 • 条件概念和用于因果判断的基本统计	11 个 CDM 问题：9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 31, 34

C. iSTAR 测试样题

C1. 控制变量维度下的问题

控制变量维度（COV）的思维技能可能是科学思维研究最集中的领域[49, 72]。在LCTSR 测试中，有 6 道问题（总共 24 道）用以测量 COV 技能。然而，最近的一项研究显示这 6 道题中有 4 道题的设计存在问题 [37]。但无论如何，这些研究成果为 COV 维度的 iSTAR 问题的开发提供了基础。

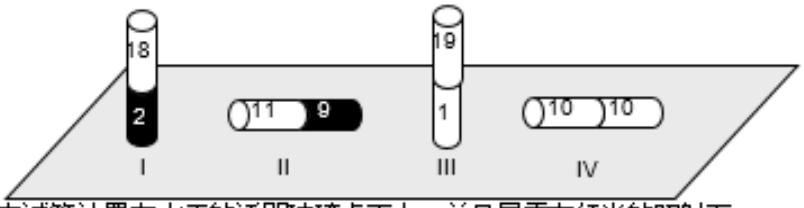
iSTAR 测试中有 9 个问题涉及 COV 能力的测量，其中 3 个问题自 LCTSR 中改编，以等效 iSTAR 和 LCTSR 的测试结果。这些问题的修订是基于效度评估的研究进行的[37]。例如，图 9 呈现了 LCTSR 中的果蝇问题的修改版本，与 LCTSR 中的版本相比，修订版有三个主要变化。首先，在原题附图中，管子被重叠的黑点覆盖，而在新版本中，这些黑点被光滑的黑色取代。根据采访时的学生评论，原始图片中的黑点往往被误解为果蝇，而不是黑色的纸:“我还以为管子上的黑点是一群果蝇呢。”在将原始图片修改为图 9 后，随后的采访中没有学生出现类似误解。

其次，原版中四个管子的布局有时会让学生感到困惑，好像所有四个管子都水平放置在一张桌子上：“我不知道 I 和 III 管是垂直于桌面的。我还以为它们都平躺在桌子上。”此外，在原版中，代表入射光的箭头误导了一些学生，使他们认为光只来自箭头方向: “我以为那些箭头是光束，管子 I 和 III 中的果蝇会朝光的方向飞去。”为了解决这些问题，新版本画了一个透明的水平桌子，以更清楚地显示四个管的空间位置，删除代表光线的箭头，添加了注释说明灯光来自各个方向。经过修改，学生在随后的采访中也没有继续报告相关问题。

第三，这个问题的原始版本只标注了管子中未覆盖部分的果蝇数量，根据一位中学老师的建议，这可能会导致数学能力较差的学生在解读数字时出错。因此，为了避免学生计算能力对答题的影响，新版本在有遮盖和无遮盖部分都标注了果蝇数量，帮助学生明确比较条件和结果，而无需进行计算。一些语言措辞也在新版本中进行了优化。

此外，LCTSR 的问题设计为两层结构，两题组中的第一个问题要求给出基于关系的结论，第二个问题则要求解释第一个问题的推理过程。LCTSR 中的四个果蝇问题也基于这一原则被设计成两组，每组包含两个不同层次的问题。因为在之前的研究中发现，两个要求解释推理过程的 LCTSR 问题不清晰，有时会误导学生[37], 所以 iSTAR 测试并未将这两个问题包含其中。相反，我们改变了 LCTSR 中的果蝇问题的场景，使用了一个叠加的两层结构，即第二个问题仍然是基于第一个问题而构建的，但不是解释第一个问题的答案，而是要求利用拓展性的推理来识别新的 COV 策略，以生成适当的协变证据来支持可能的 DCR 主张（如图 9 所示）。换言之，为了明确地比较不同的 COV 条件和结果，新版本中将答案选项重新设计，使它们得以直接测量 COV 推理的核心过程。

（果蝇1）在四支试管内各放入20只果蝇并密封。 试管 I 和试管 II 有二分之一（包括顶端）被包裹了黑色的纸。试管 III 和试管 IV 没有用任何东西包裹。四个试管按下图所示放置于水平的透明玻璃桌面上 （I 和 III 为竖直放置， II 和 IV 为水平放置），再用红色的光从四面八方照射5分钟。照射后果蝇在每支试管中的分布数量如下图所示。

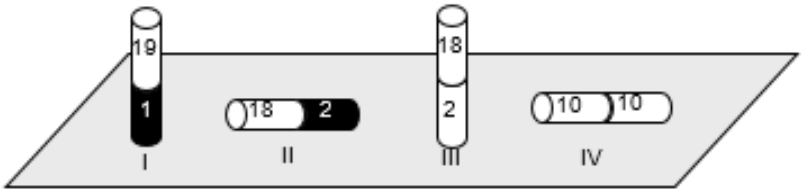


四支试管放置在水平的透明玻璃桌面上，并且暴露在红光的照射下

这个实验显示，果蝇在试管中的分布会受什么因素的影响?

a. 红光照射，而非重力。
b. 重力，而非红光。
c. 红光照射和重力都产生影响。
d. 红光照射和重力都不产生影响。
e. 根据所给信息无法确定影响因素。

（果蝇2）在第二个实验中采用另一种果蝇，并用蓝光照射5分钟。照射后果蝇的分布数量如下图所示。



四支试管放置在水平的透明玻璃桌面上，并且暴露在蓝光的照射下

为了研究果蝇在试管中的分布是否会受蓝光照射的影响，你需要比较哪些试管的实验结果?

a. 试管 I 或者试管 II 。
b. 试管 I 和试管 II 。
c. 试管 I 和试管 III 。
d. 试管 I 和试管 IV 。
e. 试管 II 和试管 III。
f. 试管 II 和试管 IV 。
g. 试管 III 和试管 IV 。
h. 试管I、试管II和试管III。
i. 所有试管。
j. 以上均不对。

%	A	B	C	D	E	F	G	H	I	J
果蝇1	18.7	48.0*	21.3	4.8	7.2					
果蝇2	6.0	20.8	4.1	1.7	2.6	32.8*	0.8	6.7	22.5	2

图 9. 果蝇问题的修改版本。答案的百分比分布是基于下一节讨论的大学生群体的结果。正确答案标有星号 (*)。

例如，对于图 9 所示的第一个果蝇问题，许多大学水平的学生能够选择正确答案(选项 b)。然而，相当一部分学生不知道如何比较不同管子的结果[37]。回答错误的学生倾向于在比较过程中把注意力集中在管 I，并通过对管 I 和管 II 进行比较从而得出结论:

果蝇对红光有反应，因为大多数果蝇都在无阴影的部分。对于重力的影响，许多学生用 I 和 III 管进行了比较，但他们得出的结论却分为了有影响和无影响两类。部分学生认为，果蝇逆着重力飞到管子顶部是对重力影响做出反应的标志，于是选择了 c。而其他学生则认为果蝇在重力作用下飞到管子底部是受重力影响的标志，因此选择了 a。在这两种情况下，学生的回答可以认为是基于先前的知识理解，而不是数据模型和控制变量的条件。此外，学生们在通过数据比较来确定某一变量影响时，往往倾向于只关注有关变量变化的情况，但忽略对可能的混淆变量进行控制的必要。因此，对于红光的影响，这些学生比较了管 I 和管 II。对于重力的影响，他们比较了管 I 和管 III。这种推理方式清晰的表明，此类学生在创建协变测量时缺乏对必要的 COV 条件的基本理解。

第二道果蝇问题建立在第一个问题的基础上，以针对性地衡量学生对略有变化的前后环境进行比较的能力。由于这个问题针对的是控制变量环节中更为显性的比较过程，因此预计会比第一个果蝇问题更难（对于第一个问题，学生或许能通过“直觉”选择答案，但这种“直觉”并不能提供明确的解释）。第二个果蝇问题的正确答案涉及在控制重力的情况下，对管 II 和管 IV 的比较（答案 f）。正如预期的那样，选择正确答案的学生更少，很多人选择 b 或 i。与第一个果蝇问题相似，选择 b 的学生倾向于关注相关变量变化的情况，但忽略了控制可能的混淆变量的必要性。同时，选项 i 被作为一个干扰项，以发现对控制变量的目的缺乏理解的学生。

本节末尾表 2 中提供了将所挑选的 iSTAR 问题对应到评估类型的表格。如表 2 所示，两道果蝇问题都针对以得出因果性 DCR 为目标的 COV 子技能。此外，推理的目的是识别有效的 DCR，在本例中包括 2 个简单的变量关系(即光和重力的简单效应)。为了识别和验证这种关系，学生需要进行发现思维(I-过程)来识别可能的原因，然后通过推演思维(D-过程)来应用假设，产生结果，以指导对一个可能答案的选择。评估和分析的简单过程（EA-过程）也运用于对推演所得结果与问题选项间的比较，其目的是确定或验证某一答案。由于这些简单的 EA-过程在这里并不是必要的思维过程，他们没有在表 2 中被列为 COV 问题的主要目标技能。

在本节中，果蝇问题被用来演示问题的开发与修订的过程，这两者都在一定程度上依赖于学生访谈和教师反馈。这些问题都经过几轮迭代的修订以消除可能的设计问题。此外，iSTAR 测试还包括另外 6 个 COV 问题。这 6 个问题都是全新设计的问题，它们涉及真实的生活或 STEM 场景，且遵循 COV 能力的进阶顺序进行设计。其中一些已经在其它研究中报告了其有效性[72]。在余下的讨论中，本文将不对问题的开发过程做详细说明，而着重介绍 iSTAR 测试的新特点。

C.2 数据分析的试题

数据分析(DA)是一个宽泛定义的维度，包括了广泛的计算和分析技能。其基本技能集与 LCTSR 中测量的数个技能重叠，包括简单概率、比例和相关性。在此基础上 iSTAR 的数据分析部分进一步扩展了更高级的计算和思维技能，包括组合、条件概率、多变量相关和

协变、加权平均和随机抽样等基本统计概念以及贝叶斯概率。这些技能在学生分析数据和识别有效证据过程中起着重要作用，并可进一步应用于因果决策。

iSTAR 测试中有 15 个问题是针对数据分析子技能的测量。为了方便于 LCTSR 做等效关联，iSTAR 测试中有一个相关关系的问题改编自 LCTSR (LCTSR Q19)，剩下的 14 个是新设计的问题，用于测量表 1 中数据分析所包括的子技能。图 10 给出了两个例子，第一个问题测量的是学生对贝叶斯概率的理解，第二个问题测量的是进行条件概率推理的能力。在开发 iSTAR 过程中的访谈中，我们发现，大学生对均匀随机过程下的掷铜板、掷骰子等案例所涉及的随机事件和独立性等概率概念有一个基本的理解，但他们也存在将这种均匀概率或等概率规律过度扩展到所有随机过程的倾向，包括非均匀条件的问题。概率是有条件的，概率状态通常不是均匀分布的，但学生并未建立起对该概念的深度理解。在现实世界的情境中，有条件的和非均匀的随机过程很常见，因此，评估学生是否能理解并在条件概率和非均匀概率下进行推理，就显得尤为重要。

图 10 中的第一个问题考查的是贝叶斯概率，测量学生是否能够根据观察结果处理非均匀随机过程。问题中使用的六面立方体不是完美的骰子，骰子的不同面的呈现可能产生不均匀的概率。这种推理是一种贝叶斯决策过程，它使用观察到的数据来推断骰子的内在特征和出现某些面的概率。访谈和开放式调查结果显示，许多学生似乎理解随机性和独立性的概念，知道每次掷骰子都是独立于其他掷骰子的随机事件。然而，学生们的推理似乎被随机性或等概率的想法所主导，他们将这种想法不加区分地应用于非均匀情况。因此，这些学生通常选择 b 选项表示概率一致，或选择 c 选项表示每个抛掷事件与之前的结果无关。

（骰子）某天你正在国外旅游，你看到哪里有一群人在玩掷骰子游戏。这颗骰子是手工雕刻的六面体，三面印有黑色图案，另外三面印有白色图案。你发现，这颗骰子抛了1000次后，出现了720次白色图案和280次黑色图案。如果再投掷这颗骰子100次，出现白色图案的次数最可能是多少？请从下列的选项中选出最恰当的一个。

a. 大约 30 次。
b. 大约 50 次。
c. 大约 70 次。
d. 由于这是不确定事件，我们不能预测出现白色图案的次数，只知道每一次出现的不是白色就是黑色图案。
e. 以上均不对。

（质量问题）国家质量监察部门对某类产品的调查报告显示，过去一年人们反映有质量问题的这类产品中，80%都是由 A 公司生产的。根据上述信息，下列选项中最恰当的是？

a. 当购买由A公司生产的该类产品时，顾客很可能遇到质量问题。
b. A公司生产的该类产品比其他公司的更可能有质量问题。
c. 上述数据显示，市场中该类产品80%可能有质量问题。
d. “a” 和 “b” 。
e. “a” 和 “c” 。
f. “b” 和 “c” 。
g. “a” 、 “b” 和 “c” 。
h. 根据报告中的数据得不到以上任意结论。

%	A	B	C	D	E	F	G	H
骰子	7.2	16.8	43.2*	31.3	1.4			
质量问题	17.3	24.1	4.9	4.9	13.8	1.6	6.1	5.4*

图 10. iSTAR 中关于数据分析的例题。答案的百分比分布是基于下一节讨论的大学生群体的测试结果。正确答案标有星号 (*)。

第二个问题测量学生条件概率中局部归一化概念的理解。问题没有给出产品 A 的市场份额，而产品缺陷的市场抽样统计结果则取决于缺陷率和市场份额两要素。因此，在不知道市场份额的情况下，不能将题中给的百分比推广为实际缺陷率，并将其作为产品 A 的质量指标。学生访谈和定量研究的结果表明，许多学生不能正确地分析这种类型概率。他们经常选择选项 a、b 或 c 或它们的组合，这表明学生对条件概率中的局部归一化概念缺乏正确理解。

这两个问题所针对的技能，也与表 2 中所示的建模和评价框架的成分和过程相对应。如上所述，这两个问题都强调了用于评估和分析观察到的概率数据（即 EA 过程），这些数据往往代表了基于因果协变的 DCR。类似题目的主要的测量目标就是针对在不同场景下学生理解和处理贝叶斯概率和条件概率的思维技能。

相较于相关试题只涉及比例、简单概率和相关性三个子技能的 LCTSR，iSTAR 涵盖了范围更广的数据分析子技能。更为重要的是，iSTAR 中数据分析技能的设计基于了新的建模框架，这使得这些技能拥有一个共同的明确目的，即连接 COV 技能以形成支持因果决策的有效证据（DCR）。在技术层面上，LCTSR 的问题对于高中生和大学生来说相对容

易，在测试这些学生时可能会产生显著的天花板效应[37]。相比之下，iSTAR 问题的难度区间更广。其目标是有效地测量从中学到研究生水平的广大学生群体。对初高中人群的测试验证正在进行中，其结果将在未来的报告中发表。

C.3 关于因果决策的问题

在 iSTAR 评估框架中，因果决策(CDM) 代表了在推理任务中基于证据得出有效因果结论的核心环节。例如，在一个假设验证任务中，学习者需要使用控制变量和数据分析技能来建立有效的协变条件，并识别合乎逻辑和计算的 DCR 结果，以及在机理上可成立的 MCR 机制，然后评估证据和假设的因果关系之间的一致性和有效性，最后将其纳入一系列决策过程，以确定关于所涉的假设和证据的有效性和可信度的最可能的结论。

LCTSR 中假设演绎推理的技能维度与 iSTAR 测试中的 CDM 维度相对应。然而，CDM 技能维度涉及一套定义明确且领域更为宽广的子技能，而 LCTSR 中的假设演绎推理主要侧重于衡量基于给定假设的证据和预测结果之间的一致性。LCTSR 中，假设演绎问题的设计也因其内容的有效性而受到批评，因为其中包含了不合理的假设[37]。由于 LCTSR 中假设演绎题的效度问题，iSTAR 测试中的 11 道 CDM 题目均为新设计的题目。图 11 呈现了两个示例。

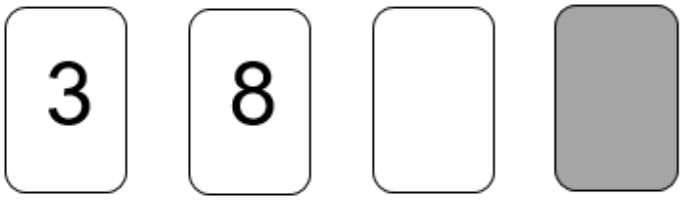
图 11 中的第一个问题测量的是学生在 CDM 过程中区分相关性和因果关系的推理能力。这个问题的场景与现实世界中的许多例子类似，比如特定的饮食习惯是否与特定的健康状况有关。在这些情况下，一个通常的错误思维是将混淆相关和因果关系，简单地将不同变量之间的相关关系认同为某种因果关系。同样，在这个问题中，给定的观察结果显示了长颈鹿的身高和力量与它是否吃某种水果之间的相关性。然而，由于缺乏对变量的控制，该数据不能形成有效的协变设计，因为只有高大的长颈鹿才能达到吃掉水果的高度，这是问题以一种相对隐含的形式给出的潜在不确定因素。因此，选项 d 是正确答案，即不能得出基于协变的因果关系。选项 a、b 和 c 代表了将相关性视为因果关系而不了解有效协变所需条件的思维模式。选择 e 代表了基于机制的先验知识对推理的影响，即观察到的数据被忽略或没有形成有意义的认知，故而学生没有对其进行解释，其决策完全基于现有的知识和主观想法。这种类型的推理表明学生在 DCR 和 MCR 之间缺乏理解和综合，意味着因果决策依赖于其先验知识。

（长颈鹿）一位游客到非洲去观察长颈鹿生活的自然生态环境。在那里他看到一种长得很高的树，这种树的果实只结在树顶的位置。他发现，经常吃这种果实的长颈鹿比那些够不着这些果实的长颈鹿长得更壮更高大。基于上述现象，下列选项中最恰当的是？

a. 如果长颈鹿常常吃这种果实，长颈鹿将长的更壮更高。
b. 这种果实的营养能促进长颈鹿的生长，使长颈鹿更壮更高。
c. “a” 和 “b” 都对。
d. 上述现象并不足以证明是吃了这种果实使长颈鹿长的更壮更高。
e. 上述现象并没有意义，因为长颈鹿的身高是由其基因所决定的。
f. 以上任意选项均不对。

（纸牌）小张和小赵一起玩一种新的纸牌游戏。每张纸牌的一面是整数，另一面则是灰色或者白色。过了一会儿，小赵说出了他的发现：“如果纸牌的一面为偶数，那么翻过来的那面对应的一定是灰色。”
假设小张随机抽出四张牌，分别显示：3，8，白色和灰色（如下图所示）。
如果想验证小赵的推断是错误的，应该翻哪张牌或哪几张牌？

a. 只翻 3。
b. 只翻 8。
c. 翻 3 和白色那张。
d. 翻 3 和灰色那张。
e. 翻 8 和白色那张。
f. 翻 8 和灰色那张。
g. 四张牌都需要翻开。
h. 以上均不对。



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

图 11. iSTAR 中有关因果决策的示例问题。答案的百分比分布是基于下一节讨论的大学生群体的测试结果。正确答案标有星号 (*)。

图 11 中的第二个问题是基于 Watson 的选择任务[76]，该任务测量了“在条件逻辑规则下判断证据和结论之间的一致性”的推理能力。问题中的假设性主张代表了“如果…那么…”关系的充分条件；换言之，如果纸牌的正面有偶数，则纸牌的背面必须是灰色的。对于这个题目，可以用两种逻辑思路的操作来展示题目中的假设可能是错误的。一种是正向的思维，通过翻开偶数纸牌检查其背面颜色来测试题中假设是否符合“如果…那么…”关系。第二种则是用逆向思维来测试相反的逻辑。既然纸牌数字为偶数是其背面为灰色的充分条件，如果一张纸牌的背面不是灰色（是白色），这个纸牌数字就不会是偶数（一定是奇数）。因此，为了测试题目中给出的假设，应该翻开纸牌数字为偶数的纸牌和背面为白色的纸牌（选项 e）。其他纸牌的结果对评估假设的有效性没有提供任何有用的信息。从对学生的访谈以及他们的测试结果来看，学生们倾向于遵循正向的演绎推理（一种确认型的推理）来关注证据，从而导致他们选择偶数纸牌和灰色纸牌（选项 b 和 f）。此外，

许多学生只是简单地想使用所有纸牌，但却没有意识到纸牌在条件逻辑中的作用。这些结果说明大部分大学阶段的学生在因果决策中缺乏对条件逻辑的充分理解。

表 2. 映射到建模和评估框架中针对目标技能的所设计的评估问题。这六个示例问题旨在探究 COV、DA 和 CDM 的三个主要维度下的各种技能，重点在于不同的因果推理成分（DCR 和/或 MCR）以及 IDEA-Loop 过程。图 3 与图 4 中定义的特定因果关系与相关推理流程复杂度之间的映射也被包含其中。

问题	iSTAR 思维子技能	IDEA-Loop 推理流程	DMCR 组成成分 (DCR & MCR)	因果网络 (图3)	推理过程的复杂性(图4)
果蝇问题 1 (图 9)	COV	ID	DCR	COV 条件	2-变量 简单联系
果蝇问题2 (图9)	COV	ID	DCR	COV 条件	2-变量 简单联系
骰子问题 (图10)	DA	EA	DCR	贝叶斯概率	贝叶斯 复杂关系
市场份额问题 (图10)	DA	EA	DCR	条件概率	条件复杂联系
长颈鹿问题 (图11)	CDM	IDEA-Loop	DMCR	隐藏机制	隐藏复杂关系
卡片问题 (图11)	CDM	IDEA-Loop	DMCR	条件逻辑链	复杂的条件逻辑

在现有的文献中，关于理论与证据之间的推理已经得到了充分的研究[38]。CDM 维度的问题是针对支持“理论与证据间的协调操作”的基本思维技能而专门设计的。这些基本思维技能例如评估协变关系的有效性的技能，或是在以 DCR 为基础的证据和以 MCR 为基础的理论之间达成一致所需的技能。如表 2 所示，这两个 CDM 问题都包括完整的 IDEA-Loop 循环，用于在 DCR 和 MCR 之间进行协调，以形成一种综合性的 DMCR 类型的因果理解。这两个问题根据学生的相关子技能，例如识别隐藏的变量和关系，或是处理条件规则，做了针对性的设计，以从任务所涉及的因果关系的不同结构和复杂性方面评估学生的思维水平。

V. iSTAR 测评工具的有效性和可靠性

iSTAR 测评工具是历经十多年，通过广泛的定性定量研究逐渐发展起来的，其开发和验证也反应了一个逐渐完善科学思维工具和建模框架的过程。此处讨论的最新版本是 2018 年完成的稳定版本，可以通过附录中提供的线上系统使用。附录中的表 A1 还提供了题项级别的描述性统计数据。以下部分将重点介绍建立 iSTAR 测试的基本评估属性和有效性。

A. iSTAR 测评特点以及与 LCTSR 的比较

在本节中，将介绍不同人群 iSTAR 的描述性统计数据，以建立评估结果的基线。因为 LCTSR 是使用最广泛的科学思维评估工具，拥有庞大的用户群和数据库，因此将结果与 LCTSR 的测量值进行比较，以作为当前文献中现有结果的参考。同时，两者之间的比较也将有助于解释两种测评之间的异同。

如上一节所述，iSTAR 包含 3 个通用技能维度，每个维度都包含不同复杂程度的多个子技能。比较来看，LCTSR 则包含 6 个有限的技能维度。除了物质守恒维度，其他的 LCTSR 技能都可以对应到 iSTAR 维度上。因为物理守恒维度对初中及其以上水平的学生来说过于容易，所以该维度并未包含在 iSTAR 中 [37]。技能维度和相应的问题如表 3 所示。

表 3. iSTAR 和 LCTSR 中技能维度和问题的对照

iSTAR 技能维度 & 问题		LCTSR 技能维度 & 问题	
控制变量	1, 4, 5, 10, 21, 24, 28, 29, 30	控制变量	9, 10, 11, 12, 13, 14
数据分析	2, 3, 6, 7, 8, 13, 14, 22, 23, 25, 26, 27, 32, 33, 35	线性比例 概率计算 相关关系	5, 6, 7, 8, 15, 16, 17, 18, 19, 20
因果决策	9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 31, 34	假设演绎	21, 22, 23, 24
		物质守恒	1, 2, 3, 4

为了比较 iSTAR 和 LCTSR 的基线评估特征，对来自中西部郊区高中的高中生以及来自中西部综合性大学的大学生和研究生进行了随机 A-B 测试。其中，参与调查的大学生是正在上第一学期大学物理入门课程的新生，研究生是物理学专业博士二年级的学生。在高中和大学的测试中，学生们被随机分配 iSTAR 或 LCTSR 试卷（每个人只做一份）。研究生则采用了不同的测试流程，每个学生在同一周的两个不同时间都分别测试了 iSTAR 和 LCTSR，每个学生的两次测试顺序是随机的。iSTAR 和 LCTSR 的平均得分见表 4。

表 4. 随机 A-B 测试中 iSTAR 和 LCTSR 总分的比较

	iSTAR			LCTSR			差异	t 检验	
年级	人数	均值	标准差	人数	均值	标准差		P 值	效应量
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

结果表明，所有年级学生的 LCTSR 分数始终高于 iSTAR 分数，说明 iSTAR 比 LCTSR 难度更高。iSTAR 解决了在测试大学生群体时 LCTSR 的天花板效应问题[37]。从结果可以看到，一年级大学生的差异最大，这可能是该年龄段技能发展和人群差异（即从高中进入大学的筛选过程）的结果。同时，研究生水平的较小差异可能是由于两个测试都具有的天花板效应。从不同年级平均分可以发现，iSTAR 测量的技能开始在大学和研究生阶段得到更多发展，而 LCTSR 测试的技能似乎主要在高中到大学早期阶段得到发展。

为了比较同一类学生人群在两种测试上的表现，表 4 和图 12 分别列出和绘制了大学生的分数分布。结果表明 iSTAR 分数以 50%左右为中心，接近正态分布。同时，LCTSR 分数以 80%左右为中心，分布偏高，天花板效应明显。

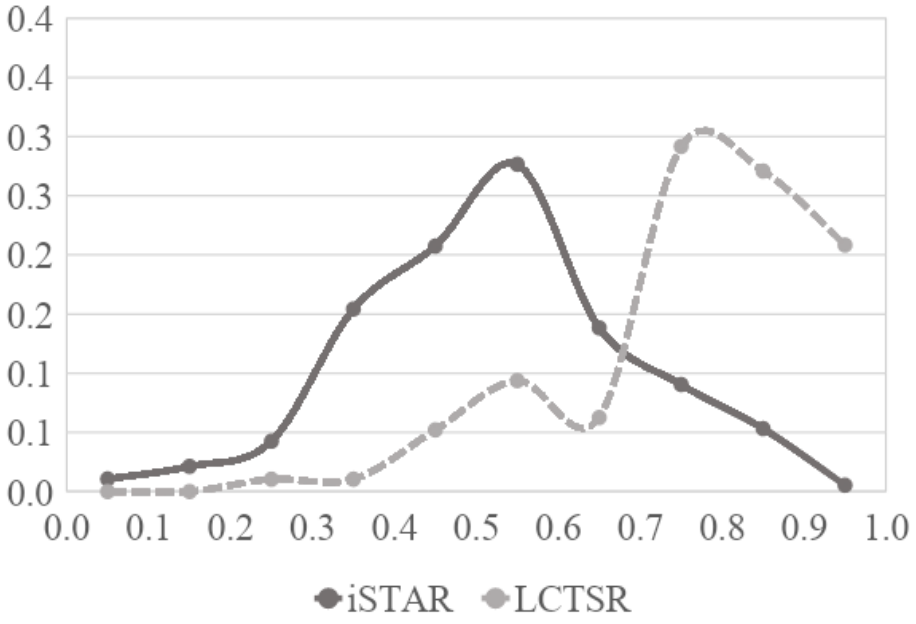


图 12. 大学生在 iSTAR 和 LCTSR 上的分数分布

对于同一组大学生，本研究对他们在 iSTAR 和 LCTSR 上的各个维度的得分进行了比较，如图 13 所示。结果表明，在三个常见维度上，LCTSR 的维度得分显着高于 iSTAR（$p<0.001$ ）。对于 LCTSR，守恒维度的分数接近 90%，这证实了这个维度对于大学生来说非常简单。在三个共同维度上，iSTAR 呈现了难度从 COV→DA→CDM 维度的持续增加，这个结果验证了设计的预期，即控制变量是用于设置协变的基础，数据分析用于中间处理和分析，因果决策用于思维的整合。相比之下，LCTSR 中因果决策维度问题的难度也是该测试题目中最高，平均分数为 60%；然而，控制变量和数据分析维度的问题似乎对大学生来说比较容易，平均分数接近 80%。尤其是数据分析与控制变量维度的问题处于相似水平，表明 LCTSR 缺乏对一些高级数据分析技能的测量。

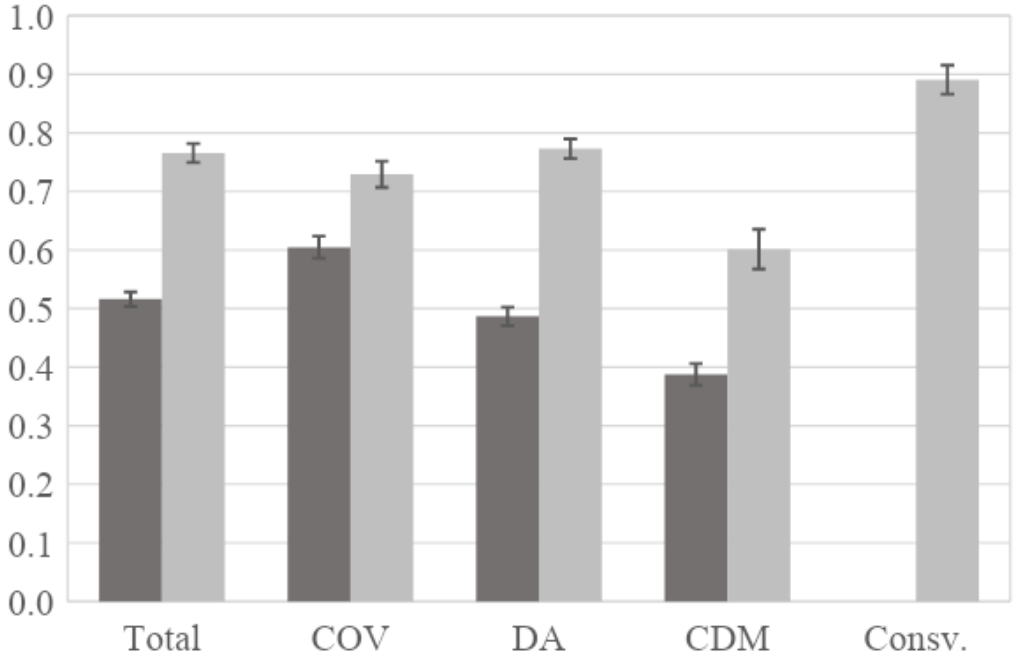


图 13. 大学生在 iSTAR 和 LCTSR 上的维度分数

通过对描述性统计的总结和比较，可以看出 iSTAR 的评估特点更适合大学生。学生在三个技能维度上的表现也呈现出了预期的难度等级进阶，这些进阶始终遵循前面讨论的数据协变与因果机制思维（DMCR）模型的预期设计。下一节将对 iSTAR 评估工具的有效性和可靠性作进一步的分析。

B. iSTAR 的有效性评估

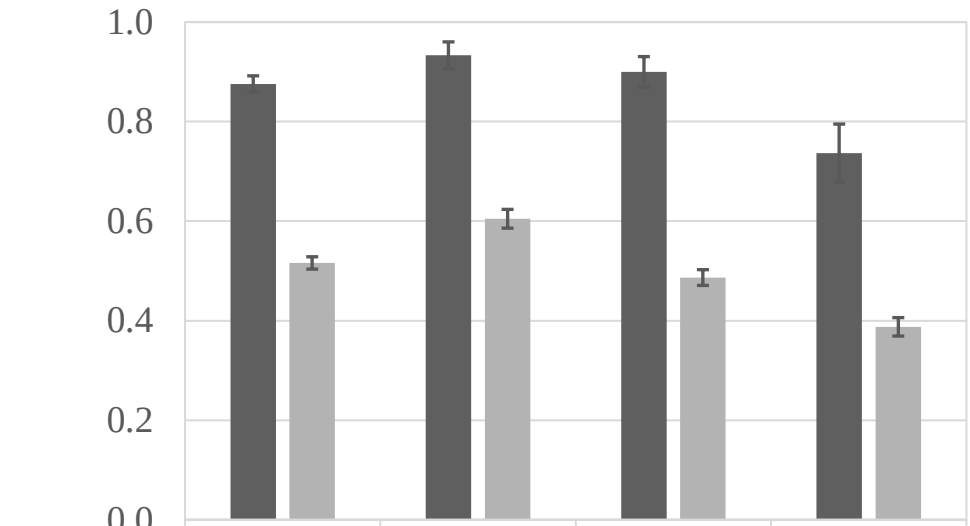
在教育研究中，效度证据的典型形式包括内容效度、标准效度和结构效度[77, 78]。内容效度是通过保证评估内容充分、恰当地覆盖目标内容来建立，可以由该领域的一组专家进行定性判断，并根据专家级应试者的回答进行定量评估。标准效度是检验新测评工具与已建立的类似测评工具之间的一致性，通常根据两个测试之间的相关性进行评估。结构效度是指一个测验实际测到所要测量的认知理论框架中设定的目标能力或潜在特质的程

度。iSTAR 这个测试中，能力结构包括 iSTAR 评估框架中定义的 COV、DA 和 CDM 三个技能维度。可以通过多种方法建立结构效度，例如传统的因子分析方法[79, 78]和基于 Rasch 模型的方法[77, 80]。在本研究中，将使用 Rasch 模型分析。

B.1 iSTAR 的内容有效性

在开发过程中，iSTAR 中的所有项目都经过了一个由科学教育研究人员和教师组成的专家团队检验评估。试测的过程中还对学生进行访谈，以收集他们解题的思维过程的详细信息。专家组在小组会议上对访谈结果和题目设计进行评估，分析学生的理解情况并完善问题设计。这个开发过程经历了大量的试验和修订周期，直到专家团队中的所有研究人员一致认为该测试经过适当和有效地设计，达到了探测特定科学思维技能的目标。

作为内容效度评估的一部分，来自同一所中西部大学的另一组 30 名研究生被用作外部专家组，以检查他们的答案是否与 iSTAR 问题的预期设计一致。这个小组由物理学专业三年级或四年级的博士研究生构成。他们在 iSTAR 及其子技能上的分数如图 14 所示，并与本科生的分数一起比较，后者数据取自表 4。



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

图 14. 研究生和本科生在 iSTAR 及其子技能上的分数。 误差线反映标准误差。

如图 14 所示，研究生和本科生的子技能相对难度趋势相似，研究生的分数达到了上限。结果表明，专家级学生对子技能的理解与设计团队的理解一致。本科生和研究生之间的成绩差异进一步表明，随着学习的进展，学生在子技能上的能力趋向于专家状态。因此，图 14 中的结果可以提供额外的定量证据来证明 iSTAR 的内容有效性。

B.2 iSTAR 的标准效度

在现有研究中，LCTSR 长期以来一直被用作科学思维的评估[36]。因此，在本研究中，标准相关效度的证据是基于 iSTAR 和 LCTSR 得分之间的相关性来评估的。

表 4 中显示的大学生的数据是在第一学期物理入门课程的第四周测量的。在 283 名学生中，约三分之一 (96) 参加了 LCTSR，其余学生参加了 iSTAR。测试对象的数量不均匀是因为设计了其他平行研究。每个学生随机参加两项测试中的一项，称为 LCTSR-4 和 iSTAR-4，以标记他们的测试类型和时间。对于所有学生，他们还在课程的第一周测试了 iSTAR，标记为 iSTAR-1。通过这种设计，可以获得 iSTAR-1 和 LCTSR-4 之间以及 iSTAR-1 和 iSTAR-4 之间的相关性。尽管第 1 周和第 4 周测试之间有三周的时间，但通常在这么短时间内学生科学思维能力的差异不会有显着变化[20]。

对测试数据计算学生分数之间的 Pearson 相关性，结果显 iSTAR-1 和 LCTSR-4 之间有中等的相关性 0.589 ($p<0.001$)，同时 iSTAR-1 和 iSTAR-4 之间也具有稍强的中等相关性 0.680 ($p<0.001$)。正如预期的那样，重复 iSTAR 测试之间的相关性强于 iSTAR 和 LCTSR 之间的相关性。这些结果表明，iSTAR 的重复测试以及 iSTAR 和 LCTSR 之间具有基本的一致性。此外，表 4 所列的研究生也参加了 iSTAR 和 LCTSR。他们两次考试成绩的相关性为 0.750（$p=0.002$），略高于大学生中测得的相关性。研究生更高的相关性进一步证实，随着学生接近专家水平，两项测试的对科学思维能力的反应更趋于一致。

为了检查两个测试中技能维度测量值之间的一致性，本研究还计算了 iSTAR-1 和 LCTSR-4 之间三个常见技能类别的相关性，如表 5 所示。结果显示 COV 的相关性为中等，CDM 的相关性较低，而 DA 的相关性最小。从相关性的结果可得，与其他技能相比，这两项测试更一致地测量了 COV 技能。另一方面，涉及 DA 技能的测量在两个测试之间的设计则有很大差异。LCTSR 只涉及到比例和基本概率的几个简单的 DA 技能，而 iSTAR 设计了 15 个题目，涵盖了从简单到复杂的 DA 技能。同样，两个测试之间的 CDM 设计也有很大不同。尽管如此，依据总分的总体相关性可以认为两项测试在衡量三种技能对科学思维综合的一维特征的测量具有良好的一致性。

表 5. iSTAR 与 LCTSR 学生技能维度得分的相关性。除 $r = 0.172$ ($p = 0.093$) 外，所有相关性在 $p < 0.01$ 水平上均具有统计学意义。

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

B.3 iSTAR 的结构有效性

如前所述，iSTAR 测试设计了 COV、DA 和 CDM 的三个领域的思维技能并假设该设计具有从 COV→DA→CDM 维度难度递增的进阶。结构效度的评估将重点分析 iSTAR 数据是否揭示了一个三维结构，以及三类思维技能的难度级别是否遵循研究所设计的进阶。

首先，通过与一维构造模型进行比较,对 iSTAR 的三维构造模型进行检验。 为此，iSTAR 数据分别拟合到三维和一维 Rasch 模型。然后使用似然比检验比较两个模型之间的拟合优度。如果这个测试的结果有利于三维模型，那么假设的三维结构就被认为得到确认和验证。

其次，可以根据学生在三项技能上的平均能力来评估三项思维技能的难度进阶。如果设计是有效的，学生应该在 COV 上表现出高能力，在 DA 上表现出中等能力，在 CDM 上表现出低能力。使用 Rasch 分析，可以计算和比较学生在三个技能维度上的平均能力来验证设计的有效性。此外，还可以用被试-项目图（怀特图）显示个人能力和题目难度在常见 logit 尺度上的分布(Bond & Fox, 2015)，以比较该分布是否适当地覆盖了广泛的能力范围和难度等级。适当的分布表明测试题可以有效区分不同水平的学生。

对于 Rasch 分析的这一部分，iSTAR 数据来自于同一所大学的另一组样本数量更大的大学生人群，总共 378 名学生。首先，对数据进行一维和三维 Rasch 模型拟合。两个模型之间的比较表明模型拟合参数有利于三维模式。似然比检验表明，与一维模型相比，三维模型在模型偏差方面具有统计学意义的改善($\chi^2 = 65.793, df = 5, p < 0.001$)。结果表明，iSTAR 中三个技能维度的结构设计与评估数据的 Rasch 分析一致。这部分分析总结性地给出 Rasch 建模的主要结果，而 Rasch 模型拟合的其他细节在附录的表 A2 中提供。

接下来，如表 6 所示，本研究计算了学生在三个技能维度上的能力估计均值，以及可靠性度量，后者将在下一节中讨论。计算结果与学生在 COV、DA 和 CDM 方面表现出由高到低的能力预测非常吻合，并且差异具有统计学意义($t_{COV-DA}(377) = 15.421, p < 0.001, d = 0.793$; $t_{COV-CDM}(377) = 34.943, p < 0.001,\ d = 1.797$; $t_{DA-CDM}(377) = 26.759, p < 0.001,\ d = 1.376$).

表 6 三维 Rasch 模型各维度的学生能力均值、信度和相关矩阵。对角线值是 EAP/PV 可靠性。 对角线下方的值是潜在相关性。

子技能	平均值	标准误差	COV	DA	CDM
控制变量(COV)	0.658	1.337	(0.770)		
数据分析 (DA)	-0.042	0.687	0.807	(0.720)	
因果决策 (CDM)	-0.815	0.963	0.794	0.819	(0.697)

此外，iSTAR 的怀特图也绘制在附录的图 A1 中，该图表面测试题目覆盖了整个 logit 量表的大范围的难度级别（-3.428 到 3.045）。学生对这三种技能的估计能力也很好地覆盖了广泛的 logit 尺度范围并呈近正态分布。结果表明，iSTAR 能够对不同水

平的学生在三个技能维度中的每一个方面进行有效区分。综上所述，Rasch 分析结果表明，iSTAR 的三维结构设计是完善的，并且测试项目对学生在三个技能维度上的能力范围都有很好的覆盖。

C. iSTAR 的信度评估

信度是指测评结果具有可靠的一致性，即在类似人群中重复应用该工具应能产生类似的结果[77]。在经典测试理论（CCT）中，信度系数通常基于两种等效测试工具的分数之间的相关性来确定。在实践中，这一概念被扩展到将一个工具的每一个项目都视为一种等效测试单元，进而衍生出使用 Cronbach's α 内部一致性系数作为信度的衡量标准。对于在上文 Rasch 模型分析中的大学生 iSTAR 数据，Cronbach's α 的计算结果为 0.737，达到可接受的信度要求（>0.7）。

此外，Rasch 模型以及项目反应理论体系的其他模型都是利用信息函数来评估测量信度 [81]。这些函数反应了观察到的成绩可以用来估计每个学生在单个项目或整个测试中的潜在特征值的精确度 [82]。使用这种带有 Rasch 建模的方法，可以根据项目信息函数和群体中潜在性状的分布来估计类似于传统信度系数的指数。在这个评估中，使用预期后验与合理值 (EAP/PV)信度的比率来衡量三个子量表的信度，这些子量表与三个维度之间的潜在相关性记录于表 6。

如表 6 所示，三个技能维度的 EAP/PV 信度为：控制变量（COV）为 0.770，数据分析（DA）为 0.720，因果决策（CDM）为 0.697。通常，0.65-0.70 的信度被认为是“最低限度可接受的”，而 0.70-0.85 的信度对于研究目的来说是“相当好的”[83]。结果表明，三个技能维度的信度是足够的，特别是考虑到学生潜在特征能力测量的复杂性质和每个维度的项目数量较少的情形。结合 Cronbach's α 值和 Rasch 分析的结果，iSTAR 的信度可以在整套测试层面和技能子维度层面得以确立。此外，三个维度之间的潜在关联度在 0.794 到 0.819 之间，表明技能维度之间有很强的关联性，它们构成了整体科学思维能力的共同基础。

综上所述，效度和信度的评估表明，iSTAR 是一种有效、可靠的大学生科学思维评估工具。然而，由于该工具的复杂性，其设计具有多个子技能和广泛的项目难度分布，需要进一步研究以确定不同学生群体的效度与信度。尽管如此，本研究的结果为 iSTAR 在与大学新生相近水平群体的科学思维评估中的有效性提供了基本证据。

VI. 总结与讨论

科学思维作为 21 世纪教育的核心能力已经得到了广泛的研究。然而，现有的文献资料中对于如何建构一个整合的科学思维理论模型尚没有形成共识。因此，虽然科学思维被广泛认为是在 NGSS 或“大学科学成功标准”等教育改革举措中有着重要地位的能力，其教学和评估在一定程度上仍然缺乏相应的理论指导。此外，目前还没有基于整合的理论模型和评估框架而开发的精细评估科学思维技能的实操工具。研究层面的这一不足，会大大限制为有效提高学生的科学思维能力而开展的各种教育实践的设计、实施和评估。

本文在既有文献的基础上，提出了一个集科学思维和因果思维于一体的理论框架(DMCR)，并根据处理和协调基于数据协变和机制解释的因果关系(即 DCR 和 MCR)所需的技能，操作性地定义了科学思维。在 DMCR 模型框架的基础上，定义了控制变量 (COV)、数据分析(DA)和因果决策(CDM)三个推理技能域，它们共同组成科学思维评估框架的基础技能集。其中，控制变量技能 COV 和数据分析技能 DA 是发展基于数据协变的因果思维 DCR 的基础，同时，数据分析 DA 又与因果决策技能 CDM 共同协调基于数据协变的因果思维 DCR 和基于机制的因果思维 MCR 以形成适当的因果理解。在评估框架的指导下，开发了科学思维评估工具(即 iSTAR)，该工具主要测评包括控制变量(COV)、数据分析(DA)和因果决策(CDM)等技能及其子技能。通过对大规模测试的数据分析，检验了 iSTAR 作为评估工具的基本特征，并与流行的 LCTSR 的结果进行了比较。结果表明，iSTAR 对推理技能的三个领域提供了一致的测量，并显示出从控制变量 COV 到数据分析 DA 再到因果决策 CDM 的难度梯度，证实了基于 DMCR 评估框架的设计是可行的。此外，对 iSTAR 的效度和信度进行了经典统计和 Rasch 分析，表明 iSTAR 对大学生科学思维能力的测量是有效且可靠的。

本研究在几个方面对相关领域做出了贡献。在理论方面，现有文献中关于科学思维和因果思维的研究有着不同的定义和侧重点，处于相对独立地发展状态[47]。然而，这两种类型的推理都是知识形成的基本要素，它们的目标、过程和具体的推理技能之间有很大重叠。因此，将这两个思维框架连接起来，可以综合不同的推理和学习模型之间的关系。通过整合这些模型，可以帮助我们解释思维和知识发展中的所历经的结构和过程之间的关系，并形成更为全面的理解。

此外，现有的科学思维研究往往过分强调基于证据的假设检验中的数据协变关系。然而，正如一些研究人员所建议的那样，因果思维的机制解释部分应该被视为科学思维的另一个核心要素。正因为机制解释是因果思维的两个基本要素之一，随着因果思维与科学思维的结构的整合，将机制解释加入到科学思维当中就变得理所当然了。DMCR 模型在现有因果思维文献的基础上，明确定义了因果思维的两个基本要素，即基于数据协变的因果关系(DCR)和基于机制解释的因果关系(MCR)，以及协调于这两个要素之间众多推理过程。基于这些新的定义，因果思维和科学思维将得以集成到一个整合的框架中，以一种更具操作性的形式对相应能力的评估和教学起到指导作用。

在操作层面来看，许多现有的研究中，科学思维技能的定义往往是一种行为的描写性定义，及基于对于推理的一般过程和结果的认知行为描述，如 “理论和证据之间的协调”、“确定一个假设”、“寻找合适的证据或备选假设”等。这些定义普遍存在一个问题，即缺乏足以构建推理过程的可操作性结构细节。例如，“理论和证据之间的协调”描述了一个思维过程的行为或任务，但是缺乏对具体怎样进行这个协调行为的技能和思维过程的定义。在本研究中，DMCR 模型框架明确定义了实际的结构、关系和过程，以及理解、应用和评估这些结构、关系和过程所需的理论模型和具体操作。这些要素共同构成了具体的细节模块和结构，在操作上定义了科学思维和因果思维中涉及的各种技能。这一模型框架将可以为测评的开发和实施提供明确的指导，从而促进科学思维和因果思维中涉及

的特定技能的发展。将科学思维和因果思维整合于一个整体性的理论模型可以为更好地理解思维和知识发展提供理论基础，同时也为更多的理论和实证研究开辟了新的空间。

综合本研究的理论工作，可以给科学思维下一个全面的定义。在现有的文献中，科学思维通常被宽泛而隐含地定义为支持科学探究学习的各种技能，可以被视作一种认知行为的描述性定义。基于本研究开发的理论模型，科学思维的定义现在可以扩展到概念定义和操作定义。这三个部分共同建构了更为完整的对“科学思维”的定义，包括：

- 行为定义：支持科学探究活动及相应过程的能力，通常包括系统地分析问题、确定可研究的问题、制定和评估假设、预测、设计和评估实验、分析数据、识别证据、验证假设和基于证据的决策，等等。
- 概念定义：在知识形成和修正时，建构和利用基于数据协变和机制解释的因果关系的认知过程。
- 操作定义：控制变量、数据分析和因果决策所需的一系列特定的推理技能。

本研究也推动了科学思维能力评估领域的进步。现有评估工具是基于“衡量一系列数目有限且松散联系的技能”的目的设计的，这些技能之间缺乏连贯的理论基础，使得相关评估结果的解释受到限制。相比之下，iSTAR 评估工具是根据 DMCR 理论框架专门设计，用于测量一套渐进的技能。这些技能共同组成了科学思维和因果思维的基本结构。在能够明确且操作性地被 DMCR 框架定义的基础上，这些技能还在验证性研究被证实可以通过 iSTAR 评估工具得到针对性测量。因此，评估结果可以直接对应到特定的技能集，并与 DMCR 模型的组件相关联，这有助于解释结果，并对学生的推理能力提供有意义的理解。这样的理解可以直接指导教学，解决目标技能的教学问题以及相关的学习困难。此外，iSTAR 评估工具的有效性和可靠性已经在大学人群中得到了验证，因此该工具也可以在研究和教学中直接使用。

本研究仍具有一定的局限性。整合科学和因果思维的理论框架的开发不可能在一项研究中完成，还需要进一步的研究来对其进行验证和完善。本文提出的 DMCR 模型是在综合现有模型的基础上提出的，这些作为基础的模型都有各自的实证研究支持。因此，新模型的有效性目前在一定程度上是由现有的实证研究支撑的，这些实证研究分别对应已整合到新模型中的先前模型的各个组成部分。此外，一同开发的评估工具及其测评结果是本研究建立的模型的具体表达，因此可以通过相应评测结果来验证模型本身的有效性。在本研究中，评估结果与模型的预期相符，从而为 DMCR 模型的有效性提供了额外的验证证据。因此，根据既有文献和实证评估结果，有理由认为新模型具有足够的有效性。

尽管如此，这项研究也只是提供了一个可以进一步发展的初步基础，并将有限的重点放在开发技能集的操作性定义上。在未来的研究中，还需要进一步更新模型，以便与更广泛的既有理论和实证研究建立联系。另外，由科学思维和因果思维所支持的学习目标，即知识发展和探究学习之间的详细联系，也还需要更多的研究。同时，建立该理论研究与科学和因果思维能力的教育实践之间的联系，以及扩展的评估研究，也都是必要的。尤其需要增加来自不同年龄段和教育背景的样本群体以进一步建立 iSTAR 评估的信效度。

附件：iSTAR 测试的统计评估

1.访问 iSTAR 测试

iSTAR 测试可以通过 https://istartest.com/home 上的在线测试系统访问和使用。有关使用测试的查询，请联系通讯作者获取更多信息。

2.统计分析结果

表 A1 提供了 iSTAR 试题的描述性统计数据，包括经典测量理论定义的试题难度（正确答案的分数）、试题区分度（高 30% 和低 30% 学生之间的分数差异）和单个试题的分数与测试总分之间的点二列相关（试题与总分的相关性）。计算结果与 Rasch 分析中用于评估结构效度的数据集相同。iSTAR 的测试信度由 Cronbach’s α 评估，包含所有 35 个试题，结果为 0.737。其中两个试题（30 和 33）与测试总分相关性不显着。如果删除这两项，则 Cronbach’s α 变为 0.750。此类分析表明 iSTAR 具有足够的可接受的信度。这两个试题保留在测试中，因为不同的人群可能会做出不同的反应，并且这两个试题是相关内容场景完整性所需的试题组的一部分。

表 A1. iSTAR 的基本描述性统计（N=378）

试题	难度	区分度	r_{pb}	试题	难度	区分度	r_{pb}
1	0.889	0.177	0.226	18	0.862	0.191	0.213
2	0.441	0.481	0.346	19&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 分析

为了调查这三个子技能是否代表学生科学思维和推理的不同维度，将三维模型与一维模型进行了，这里的一维模型假设数据背后只有一个潜在构造（即整体科学思维和推理）。两种模型的比较表明模型拟合参数有利于三维模型。似然比检验表明，与一维模型相比，三维模型在模型偏差方面显示出统计学上的显著改善($\chi^2 = 65.793, df = 5, p < 0.001$)。

为证实上述模式拟合的结果，我们进一步探讨了三维模式的模型参数。在表 A2 中列出的加权和未加权均方残差 (MNSQ) 用于检查学生对 iSTAR 的反应在项目级别与 Rasch 模型的拟合程度。如表 A2 所示，iSTAR 的所有项目，除了项目 10、12 和 15，似乎都符合项目拟合标准（0.7<MNSQ<1.3）。对于第 10、12 和 15 项，虽然未加权 MNSQ 略微超出建议范围，但加权 MNSQ 很好地符合标准。因此，项目 10、12 和 15 没有从以下分析中删除。基于三维模型，在图 A1 中绘制了怀特图。

表 A2. 由 iSTAR 的 Rasch 模型估计的项目难度和拟合统计量（Infit 和 Outfit MNSQ）的度量。请注意，平均项目难度被限制为零。参数估计旁边的星号表示它受到约束。由于项目与总分相关性不显着，项目 30 和 33 未包括在该模型拟合分析中。

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&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

Logit Scale	Dimensions COV	DA	CDM	Generalized Item Thresholds
4				
	X			
	X			
	X			
	X			
	X			
3	X			14_DA 15_CDM
	X			12_CDM
	X			
	XX			
	XX			35_DA
	XX			21_COV
2	XXX			32_DA
	XX	X		24_COV
	XXXX	X		5_COV
	XXXX	X	X	27_DA
	XXXX	X	X	3_DA
1	XXXX	XXX	X	
	XXXX	XXX	XX	26_DA 31_CDM
	XXXXX	XXXXXX	XX	4_COV
	XXXXX	XXXXXXX	XX	
	XXXXXXX	XXXXXXX	XXX	17_CDM
	XXXXXX	XXXXXXXX	XXXX	2_DA 13_DA
0	XXXXXX	XXXXXXXXX	XXXX	11_CDM
	XXXXX	XXXXXXXXXX	XXXXX	19-20_CDM 29_COV
	XXXXX	XXXXXXXXXX	XXXXXXX	8_DA
	XXXX	XXXXXXXX	XXXXXX	9_CDM
	XX	XXXXXXX	XXXXXX	34_CDM
	XX	XXXXXX	XXXXXXX	
-1	XX	XXXX	XXXXXXX	28_COV
	X	XX	XXXXXXX	
	X	X	XXXXXX	6_DA
	X	X	XXXXXX	7_DA
	X		XXXXX	
	X		XXXX	
-2	X		XXXX	1_COV 25_DA
			XX	
			XX	23_DA
			X	
-3			X	10_COV 16_CDM 18_CDM
				22_DA
-4				

Each X represents 4.0 cases

图 A1. iSTAR 怀特图

表 A2 中的结果表明，项目难度的测量涵盖了从-3.428（容易）到 +3.045（困难）的足够广泛的范围。图 A1 中显示的怀特图为项目在三个技能组中的难度分布提供了进一步的细节。怀特图在相同的 logit 量表上显示单个学生的能力测量和单个项目的难度测量，以便在项目难度和学生表现之间进行清晰对应。iSTAR 数据使用三维 Rasch 模型进行分析，该模型衡量学生的三个技能集，包括控制变量 (COV)、数据分析 (DA) 和因果决策 (CDM)。如图 A1 所示，项目难度广泛分布于三个技能维度，很好地涵盖了不同技能。对学生能力的测量也显示出在所有三个技能维度上理想的近正态分布，并且在能力量表上具有足够宽的跨度。三个技能维度上的能力分布中心也显示了预期的难度进阶，其中 COV 是最简单的（学生平均能力最高），DA 是中级，CDM 是最难的（学生平均能力最低）。总体而言，怀特图中显示的结果表明 iSTAR 测试对三个技能维度的覆盖分布令人满意，并且提供的测量结果与设计的预期一致。

参考文献:

[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] N. R. 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, 2012a.

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

[7] National Science & 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, Millbrae, CA: California Academic Press, 1990.

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

[10] M. Lipman, Thinking in Education (2nd Ed.), Cambridge: Cambridge University Press, 2003.

[11] M. Binkley, O. Erstad, J. Herman, S. Raizen, M. Ripley and M. Rumble, Draft White Paper Defining 21st century skills, Melbourne: ACTS, 2010.

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

[13] R. H. Johnson and B. Hamby, "A Meta-Level Approach to the Problem of Defining 'Critical Thinking'," *Argumentation,* vol. 29, no. 4, pp. 417-430, 2015.

[14] P. A. Facione and C. A. Gittens, Think critically, 3rd ed., Boston: Pearson, 2016.

[15] D. F. Halpern, Critical thinking across the curriculum: A brief edition of thought & knowledge, Routledge, 2014.

[16] R. H. Ennis, "Critical Thinking: A Streamlined Conception," in *The Palgrave Handbook of Critical Thinking in Higher Education*, New York, Palgrave Macmillan, 2015, pp. 31-47.

[17] H. Siegel, Educating reason: Rationality, critical thinking and education, New York: Routledge, 1988.

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

[19] C. Zimmerman, "The development of scientific reasoning skills," *Developmental Review*, vol. 20, no. 1, pp. 99-149, 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,* vol. 323, no. 5914, pp. 586-587, 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?," *Journal of Research in Science Teaching,* vol. 35, no. 1, pp. 89-103, 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," *Journal of College Science Teaching,* vol. 33, no. 3, pp. 18-22, 2003.

[23] S. T. Kalinowski and S. Willoughby, "Development and validation of a scientific (formal) reasoning test for college students," *Journal of Research in Science Teaching,* vol. 56, no. 9, pp. 1269-1284, 2019.

[24] V. P. Coletta and J. A. Phillips, "Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability," *American Journal of Physics,* vol. 73, no. 12, pp. 1172-1182, 2005.

[25] H. She and Y. Liao, "Bridging scientific reasoning and conceptual change through adaptive web-based learning," *Journal of Research in Science Teaching,* vol. 47, no. 1, pp. 91-119, 2010.

[26] M. S. Cracolice, J. C. Deming and B. Ehlert, "Concept Learning versus Problem Solving: A Cognitive Difference," *Journal of Chemical Education,* vol. 85, no. 6, pp. 873-878, 2008.

[27] S. Ates and E. Cataloglu, "The effects of students' reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics," *European Journal of Physics,* vol. 28, no. 6, pp. 1161-1171, 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 Sciences Education,* vol. 10, no. 1, pp. 64-73, 2011.

[29] A. E. Lawson, "The development of reasoning among college biology students - a review of research," *Journal of College Science Teaching,* vol. 21, no. 1, pp. 338-344, 1992.

[30] D. Kuhn, "Thinking as argument," *Harvard Educational Review,* vol. 62, no. 2, pp. 155-178, 1992.

[31] V. F. Shaw, "The cognitive processes in informal reasoning," *Thinking & Reasoning,* vol. 2, no. 1, pp. 51-80, 1996.

[32] A. Zeineddin and F. Abd-El-Khalick, "Scientific reasoning and epistemological commitments: Coordination of theory and evidence among college science students," *Journal of Research in Science Teaching,* vol. 47, no. 9, pp. 1064-1093, 2010.

[33] K. Koenig, K. E. Wood, L. J. Bortner and L. Bao, "Modifying traditional labs to target scientific reasoning," *Journal of College Science Teaching,* vol. 48, no. 5, pp. 28-35, 2019.

[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," *Journal of Research in Science Teaching,* vol. 55, no. 7, pp. 962-981, 2018.

[35] T.-R. Sikorski and D. Hammer, "Looking for coherence in science curriculum," *Science Education,* vol. 101, no. 6, pp. 929-943, 2017.

[36] A. E. Lawson, "Lawson Classroom Test of Scientific Reasoning," 2000. [Online]. Available: 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," *Physical Review Physics Education Research,* vol. 14, no. 2, pp. 020106-1-020106-19, 2018.

[38] C. Zimmerman, "The development of scientific thinking skills in elementary and middle school," *Developmental Review,* vol. 27, no. 2, pp. 172-223, 2007.

[39] J. Piaget, Construction of reality in the child, London: Routledge, 1954.

[40] A. E. Lawson, "The Nature and Development of Scientific Reasoning: A Synthetic View," *International Journal of Science and Mathematics Education,* vol. 2, no. 3, pp. 307-338, 2004.

[41] D. Klahr, Exploring science: The cognition and development of discovery processes, Cambridge, MA: MIT Press, 2002.

[42] D. Kuhn, M. Pease, Wirkala and Clarice, "Coordinating the effects of multiple variables: a skill fundamental to scientific thinking," *Journal of Experimental Child Psychology,* vol. 103, no. 3, pp. 268-284, 2009.

[43] A. E. Lawson, "Development and validation of the classroom test of formal reasoning," *Journal of Research in Science Teaching,* vol. 15, no. 1, pp. 11-24, 1978.

[44] A. E. Lawson, Science Teaching and the Development of Thinking, Belmont, CA: Watsworth Publishing Company, 1995.

[45] A. E. Lawson, "Using the learning cycle to teach biology concepts and reasoning patterns," *Journal of Biological Education,* vol. 35, no. 4, pp. 165-169, 2001.

[46] D. Klahr and K. Dunbar, "Dual space search during scientific reasoning," *Cognitive Science,* vol. 12, no. 1, pp. 1-48, 1988.

[47] D. Kuhn and D. J. Dean, "Connecting scientific reasoning and causal inference," *Journal of Cognition and Development,* vol. 5, no. 2, pp. 261-288, 2004.

[48] D. Kuhn, S. Ramsey and T. S. Arvidsson, "Developing multivariable thinkers," *Cognitive Development,* vol. 35, pp. 92-110, 2015.

[49] Z. Chen and D. Klahr, "All other things being equal: Acquisition and transfer of the Control of Variables Strategy," *Child Development,* vol. 70, no. 5, pp. 1098-1120, 1999.

[50] M. Bunge, Causality. The place of the causal principle in modern science, Cambridge, MA: Harvard University Press, 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," in *Les explications causales*, Paris, PUF, 1971.

[53] R. Harré, The philosophies of science, Oxford: Oxford University Press, 1972.

[54] J. Ogborn, "Approche théorique et empirique de la causalité," *Didaskalia,* no. 1, pp. 29-47, 1993.

[55] R. K. Guenther, Human cognition, Upper Saddle River, NJ: Prentice Hall, 1998.

[56] R. Corrigan and P. Denton, "Causal understanding as a developmental primitive," *Developmental Review,* vol. 16, no. 2, pp. 162-202, 1996.

[57] F. C. Keil, Concepts, kinds, and cognitive development, Cambridge, MA: MIT Press, 1989.

[58] A. Michotte, La perception de la causalité, Paris: Vrin, 1946.

[59] H. H. Kelley, "The process of causal attribution," *American Psychologist,* vol. 28, pp. 107-128, 1973.

[60] M. Bullock, R. Gelman and R. Baillargeon, "The development of Causal Reasoning," in *The developmental psychology of time,* New York, Academic Press, 1982, pp. 209-254.

[61] W. Hung and D. H. Jonassen, "Conceptual Understanding of Causal Reasoning in Physics," *International Journal of Science Education,* vol. 23, no. 13, pp. 1601-1621, 2006.

[62] C. Chen, L. Bao, J. C. Fritchman and H. Ma, "Causal Reasoning in Understanding Newton's Third Law," *Physical Review Physics Education Research,* p. in review, 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," *Physical Review Physics Education Research,* vol. 17, no. 2, p. 020116, 2021.

[64] D. Kuhn, E. Amsel and M. O'Loughlin, The development of scientific thinking skills, Orlando, FL: Academic Press, 1988.

[65] B. Koslowski, Theory and evidence: the development of scientific reasoning, Cambridge: MIT Press, 1996.

[66] J. Biggs and K. Collis, Evaluating the Quality of Learning: the SOLO taxonomy, New York: Academic Press, 1982.

[67] M. C. Linn, "The Knowledge Integration Perspective on Learning and Instruction," in *The Cambridge handbook of: The learning sciences* , New York, Cambridge University Press, 2005, pp. 243-264.

[68] W. Whewell, The Philosophy of the Inductive Sciences, Founded Upon Their History, New York: Johnson Reprint, 1840.

[69] A. E. Lawson, "Hypothetico-deductive Method," in *Encyclopedia of Science Education,* Dordrecht, Springer, 2015.

[70] J. C. Moore and L. J. Rubbo, "Scientific reasoning abilities of nonscience majors in physics-based courses," *Physical Review Special Topics Physics Education Research,* vol. 8, no. 1, p. 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 and Creativity,* vol. 27, pp. 101-113, 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 and Creativity,* vol. 19, pp. 175-187, 2016.

[73] P. W. Cheng, "From covariation to causation: A causal power theory," *Psychological Review,* vol. 104, no. 2, pp. 367-405, 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 of Science,* vol. 12, no. 2, pp. 123-145, 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," *Journal of Experimental Psychology: Applied,* vol. 23, no. 1, pp. 1-14, 2017.

[76] P. C. Wason, "Reasoning about a rule," *The Quarterly Journal of Experimental Psychology,* vol. 20, no. 3, pp. 273-281, 1968.

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

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

[79] B. Thompson and L. G. Daniel, "Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines," *Educational and Psychological Measurement,* vol. 56, no. 2, pp. 197-208, 1996.

[80] T. G. Bond and C. M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences, Third Edition (3rd ed.), New York: Routledge, 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," *Applied Psychological Measurement,* vol. 37, no. 3, pp. 201-225, 2013.

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

[83] R. F. DeVellis, Scale Development: Theory and Applications, vol. 26, SAGE Publications, 2012.