

情境真实性究竟对小学科学试题难度是否 有影响：一项基于多维框架的测量学研究

杨 煌^{1*} 黄 真^{2,3} 曾淑婷⁴

(1. 珠海市香洲区教师发展中心, 珠海, 519000;

2. 北京师范大学未来教育学院, 珠海, 519000;

3. 北京师范大学珠海实验学校, 珠海, 519000;

4. 珠海市海湾小学, 珠海, 519000)

摘 要：一线教学常误将“情境真实性”等同于“试题难度”。为厘清二者关系，本研究构建了包含 16 个指标的“试题任务复杂度评级框架”，对 235 道小学科学试题进行编码，结合 12 万份作答数据，利用 Rasch 模型与结构方程模型（SEM）进行实证检验。研究发现：（1）构念分离，“情境真实性”与由信息支架、认知加工构成的“任务复杂度”是两个结构独立的构念。（2）功能差异，“情境真实性”与试题难度无显著相关（ $r \approx 0$ ），澄清了普遍误解；而“任务复杂度”维度与难度呈显著负相关，呈现独特的“脚手架效应”。（3）内在机制，SEM 证实高质量的信息支架能显著激活认知加工（ $\beta = 0.569$ ），从而在不增加客观难度的前提下提升思维层级。本研究证明了情境仅是载体，真正的“任务复杂度”通过信息支架支撑高阶思维。

关键词：科学素养；任务复杂度；Rasch 模型；情境真实性；脚手架效应

一、引言

（一）科学教育的“情境化”转向与挑战

当前，全球科学教育正经历着一场深刻的范式转移，其核心从关注碎片化的学科知识习得，转向聚焦学生在复杂、多元的现实世界中解决实际问题的核心素养获取。无论是国际学生评估项

* 杨煌, 33953035@qq.com。

目（PISA）所强调的“在个人、地区、国家及全球背景下的科学素养”，还是美国《下一代科学标准》（NGSS）所倡导的融合了学科核心概念、跨学科概念与科学与工程实践的三维目标体系，都毫无例外地将“真实情境”置于了教学与评价的核心位置。^[1-2]我国教育部颁布的《义务教育科学课程标准（2022年版）》（以下简称“新课标”）明确提出，倡导以探究和实践为主的多样化学习方式，让学生在真实情境中体验科学探究过程。^[3]

然而，理想的课程理念在向一线教学与评价实践转化的过程中，遭遇了具体的挑战与困境。在“情境化”命题改革的热潮中，一线教师普遍存在一种基于直觉经验的朴素认识——“情境越真实，试题就越难”。这种将“真实性”与“高难度”简单画等号的预设，在实践中导致了两种典型的消极倾向：

一是“畏难而退”，部分教师出于保护学生考试信心或担心教学“超纲”的考虑，有意回避那些复杂的真实科学与工程情境（如航天工程、生态治理、流行病调查等），仍旧偏好使用抽象的、理想化的“纯学科”问题进行训练。这使得科学教学脱离了其应有的现实土壤，退化为枯燥的知识点操练，背离了素养导向的初衷。

二是“形式主义”，部分命题者为了迎合“情境化”的政策要求，在试题中生硬地堆砌大量真实但与核心任务无关的背景信息（即所谓的“穿靴戴帽”），却未能设计出与之深度融合的有效认知任务。这种“假情境”不仅未能有效考查出学生的高阶思维，反而因增加了无效的阅读负担而招致师生的广泛诟病，甚至引发了对情境化命题方向的质疑。^[4]

（二）问题的症结：任务与情境概念的混淆

造成上述误解与实践困境的根源，在于学界和一线实践领域对“试题任务复杂度”这一高阶概念缺乏清晰的结构化认识，往往将“情境”这一单一维度特征的真实性，错误地等同于整个“任务”的复杂度。

事实上，任务是一个包含目标、条件、过程和结果的完整系统，而情境只是任务呈现的背景条件。一个看似极其“真实”且复杂的“火星基地建设”情境，如果其设计仅仅是要求学生写出“氧气”的化学符号或计算简单的加减法，那么对于掌握了该基础知识的学生而言，这个任务本质上仍是低难度的记忆提取或简单应用。反之，一个看似最普通的“测量小车速度”的实验室情境，如果要求学生自行设计控制变量实验、处理带有误差的数据并建立多因素的数学模型，其任务难度则可能极高。

显而易见，情境的“外衣”固然重要——正如许多一线教学研究反复强调的，其核心价值在于激发学习兴趣、调动参与积极性，并为学生理解抽象概念提供认知脚手架^[5-6]，但它并非决定学生能否成功解决问题（即任务难度）的决定性因素。真正决定任务难度的，是这件外衣下包裹的“内核”——认知任务。

将情境的“动机价值”或“理解辅助”功能，错误地等同于其对任务难度的贡献，正是教学改革中“为了情境而情境”误区的根源。因此，亟须引入严谨的测量学实证研究，将“情境”的特征从整体“任务”的难度来源中精准地剥离出来，量化不同维度对试题难度的独立贡献率。这不仅是一个亟待澄清的理论问题，更是指导未来科学教学与评价设计、提升命题质量的迫切实践需求。

（三）研究问题与核心目标

值得深思的是，尽管 PISA 等大型国际评价项目极力倡导情境的真实性，旨在测量学生应对“真实世界挑战”的能力，但一些实证研究却对这一理念的实际效果提出了疑问。例如，有学者在深入分析 PISA 任务后尖锐地指出，许多所谓的“真实生活挑战”在学生看来，仍然是人为编造的、“学校化”的问题，而非真正的现实挑战。^[7] 这揭示了命题者设定的“客观真实”与学生体验到的“主观真实”之间可能存在的鸿沟。

这一现象引出了一个亟待回答的关键问题：驱动试题难度的，究竟是情境的“真实性”标签本身（这在实践中常被误解为难度来源），还是其背后所承载的认知任务的内在复杂性？

因此，本研究的核心目标之一，正是试图通过量化模型，将这两种潜在的影响效应进行实证分离，以探明它们各自的独立贡献。

二、理论基础

（一）沙沃森（Shavelson）的任务复杂度模型

要厘清“情境”与“难度”的关系，首先需要在理论上对“任务”这一概念进行科学的解构。著名教育测量学家沙沃森等人在其关于科学表现性评价的经典研究中提出，“任务复杂度”是一个多维度的整合性构念，它不能被简单还原为单一因素。Shavelson 模型指出，至少有三个核心维度共同决定了一项科学任务的复杂程度^[8]：

1. 内容维度，涉及任务所包含的学科知识广度、深度以及知识元素间关联的复杂性，这是任务的“本体”。
2. 认知过程维度，指学生解决任务时大脑必须进行的信息加工操作类型和层级，如简单的检索、理解，或复杂的推理、评价、创造，这是任务的“灵魂”。
3. 情境维度，指任务呈现的背景特征，包括其真实性、熟悉度、结构化程度等，这是任务的“载体”。

沙沃森的这一理论模型为本研究提供了最核心理论：“情境”只是构成完整“试题任务”的一个侧面，而非全部。它与其他两个维度（内容、认知）既相互交织，又在理论上彼此独立。一个高难度的任务，完全可能是“低情境复杂度”（如抽象数学难题）与“高认知复杂度”的组合；其他亦然。这一理论框架为本研究在实证层面“剥离”情境效应提供了坚实的逻辑起点。

（二）认知负荷理论：任务难度的心理学机制

如果说 Shavelson 模型提供了任务的宏观结构，那么斯威勒（Sweller）提出的认知负荷理论（Cognitive Load Theory, CLT）则从微观心理学机制上解释了“复杂度”如何转化为学生感受到的“难度”^[9]。CLT 基于人类工作记忆容量有限这一基本前设，将学习者在解决问题时所承受的认知负荷区分为三类，这恰好能与沙沃森的维度产生精妙的对应：

1. 内在认知负荷，由学习材料本身固有的复杂性（即知识元素间的交互程度）决定，直接对应于任务的内容维度（如变量关系）。

2. 外在认知负荷, 由信息呈现的方式决定, 糟糕的情境设计 (如含大量无关“噪声”或信息呈现混乱) 会增加无效的外在负荷, 这与任务的情境维度与信息呈现方式密切相关。

3. 关联认知负荷, 指学习者用于构建图式、进行深度理解的有效认知努力, 这与任务的认知过程维度紧密相连。

基于 CLT 可以推论: 一个高难度的科学试题, 其难度来源是多样的。如果一个“真实情境”的设计, 仅仅是增加了无效的阅读负担 (外在负荷), 它自然会提升难度; 但如果一个“真实情境”的设计, 通过提供生动的图表和数据 (即“信息支架”), 激发了学生的学习动机, 并辅助其构建了科学图式 (关联负荷), 那么它反而可能降低内在认知负荷, 从而降低整体难度。本研究的核心目的之一, 正是要探明“情境真实性”究竟在多大程度上通过影响这些不同类型的负荷来左右最终的试题难度。

(三) “真实性”的再审视: 从“生活真实”到“实践真实”

学术界对“真实性”(Authenticity)的界定存在多种视角。一种观点认为“真实”(Real)即与学生日常生活经验的接近程度高^[10]; 另一种观点则认为“真实”并非指情境的生活化, 而是指任务在知识生成逻辑上与真实的科学实践或工程设计相似^[11]; 还有研究者提出“真实”不是“real”, 而是“authentic”, 其核心特征是模拟专业实践^[12]。

Authentic (实践真实) 强调的是能够引导正确的科学思维方式的情境, 而非表面场景细节的熟悉。例如, 一个高度“authentic”的情境 (如模拟科学研究、分析火星数据) 对小学生而言可能因缺乏经验而非常陌生, 这与高熟悉度的真实生活情境 (如校园生活) 所引发的认知过程截然不同。

国内学者毕华林、刘菊等人曾对化学教材中的情境进行类型学分析, 将其划分为生活、生产、科技、化学史等类型^[13-14], 这为理解情境的多样性提供了基础。本研究借鉴上述思路, 在 PISA 2025 的“个人、职业、社会、科学”四大真实情境基础上, 特意增加了“教学/学术情境”作为对比 (或作为最低层级), 从而构建了一个从“非真实”到“高阶真实”的纵向分级模型。

三、研究方法

(一) 研究过程

为确保研究的科学性与规范性, 本研究采用了“工具构建—大样本编码—模型检验”的定量研究路径。研究的探索之旅遵循以下几个递进的步骤: 首先, 在工具构建阶段, 本研究基于相关理论, 设计了包含 16 个二分指标的“试题任务复杂度评级框架 (TCRF)”。随后的样本编码阶段, 采集了 235 道小学科学试题项目, 并组织了 3 名经验丰富的科学教研员老师对所有试题的 16 个指标进行逐一编码 (0/1)。然后在难度标定阶段, 本研究收集了这 235 道试题的近 10 万名小学生的原始作答数据, 并运用 Rasch 模型分析, 计算出每道试题的客观 Rasch 难度值 (Logits)。最后, 在数据分析阶段, 本研究依次进行了相关分析、探索性因子分析 (EFA)、验证性因子分析 (CFA) 与结构方程模型 (SEM) 分析。相关分析初步检验了四个维度 (D1—D4) 总分与 Rasch 难度值之间的关系; EFA 检验了 TCRF 框架 (16 个指标) 的内部结构效度; 最终, 在 EFA 结果的基础上, 本研究构建并检验了一个更精简、稳健的 CFA 与 SEM 模型, 以揭示各潜变量对试题

难度的深层影响机制。

(二) 研究样本

本研究的分析单位是“试题项目”，研究样本共包含 235 个试题项目，来源于珠海市香洲区 2022 至 2025 学年间的 6 套小学科学素养监测与期末试卷（包括四年级、五年级和六年级）。

为获取可靠的试题难度参数，本研究收集了参与这 6 次监测的全部学生作答数据，总计 $N=98\,086$ 人次。数据采集过程均由区教师发展中心统一组织，在校标准化考场内完成，确保了数据来源的真实性与可靠性。

这 6 套试卷的一个显著特点是命题方式全面采用了情境化与综合化的理念，内容广泛覆盖了物质科学、生命科学、地球与宇宙科学、技术与工程等多个核心领域。样本量为 EFA/SEM 等高级统计模型的应用提供了必需的数据支持，确保了分析结果的统计功效与稳定性。

(三) 研究方法

1. 试题任务复杂度评级框架（TCRF）

基于前文所述的理论重构，本研究团队自主开发了 TCRF（表 1）。该框架包含 4 个维度和 16 个二分观测指标。为保证评级的客观性与一致性，编码过程由三名经验丰富的科学教研员独立完成。首先对所有指标的定义和典型题例进行深入研讨，达成一致理解；随后进行两轮试编码与反馈修正；最后对全部 235 道试题进行正式编码。对于评级结果存在分歧的条目（约占总条目的 8%），团队通过集体讨论的方式进行协商，直至意见完全一致。经计算，评定者间的信度（Cohen’s Kappa 系数）达到了 0.91，表明 TCRF 评级工具具有高度的稳定性和可靠性。

表 1 试题任务复杂度评级框架（TCRF）及其观测指标

维度	指标名称	操作定义	评分依据与典型题例
D1：信息 支架	多模态表征	题目是否提供了文字以外的视觉信息（图片、示意图、流程图）？	实证发现：图片往往起到“可视化辅助”作用，而非增加难度。例：2406 _ Q01 工程流程图。
	数据密集度	题目是否包含结构化的数据表、统计图或超过 3 个关键数值？	实证发现：表格化数据能帮助学生快速锁定规律，充当了认知的“脚手架”。例：2506 火星环境参数表。
	信息转化	解决问题是否需要将一种信息形式转化为另一种（如读图填空、绘图）？	强调信息处理的主动性。例：2305 _ Q21 绘制水温折线图。
	冗余/干扰信息	题目是否包含解决核心问题非必需的背景噪声或互斥条件？	这是区分“真复杂”与“假复杂”的关键。例：2506 蚯蚓实验中的“人为踩踏”干扰。
D2：系统 思维	单线性因果	题目是否仅考查单一的“条件—结果”对应关系？	基础逻辑。例：2304 _ Q05 凤仙花叶子交叉是为了吸收阳光。
	多变量控制	题目是否涉及对比实验设计，需识别自变量、因变量和控制变量？	科学探究的核心逻辑。例：2305 _ Q15 保温杯对比实验。
	交互与制约	变量间是否存在相互制约或非线性关系（如权衡取舍）？	高阶逻辑。例：2506 火星种植中“氧气产量”与“食物占比”的权衡。
	模型构建与迁移	是否需要建立抽象模型或将学过的模型应用到全新情境？	最高阶逻辑。例：2406 建立太阳系行星位置模型。

续表

维度	指标名称	操作定义	评分依据与典型题例
D3: 情境 真实性	具象化叙事	题目是否依托于一个具体的、特定的个案、事件或数据展开?	这是最基础的“人为情境”,对应传统的学科教学评价。例:2206 显微镜操作题。
	个人/生活情境	情境素材是否直接来源于学生熟悉的家庭生活、娱乐或个人健康?	对应“生活化科学”。例:2305 厨房里的小苏打和白醋。
	社会/职业情境	情境是否涉及社区环境、农业生产、公共设施建设等社会性议题?	对应“STSE(科学、技术、社会与环境)”教育。例:2406 港珠澳大桥白海豚保护。
	专家/科研实践情境	情境是否模拟了真实的科学家或工程师面临的“不确定性”和“复杂挑战”?	对应论文核心观点——提供科学实践的“场域”,必须包含真实约束或开放性决策。例:2506 火星基地生态系统设计。
D4: 认知 加工	提取与回忆	是否需要再认或回忆具体的事实、概念、术语?	基础认知。例:2406 昆虫识别(苍蝇)。
	理解与阐释	是否需要用自己的语言解释现象、归纳特征或说明理由?	中阶认知。例:2304 解释水位下降的原因。
	分析与推理	是否需要基于证据进行逻辑推断、预测趋势或论证观点?	高阶认知。例:2206 六瓶液体逻辑推理。
	创造与评价	是否需要设计新方案、构建新产品或对他人方案进行批判性评价?	最高阶认知。例:2305 改进小船设计方案。

2. Rasch 模型难度值

利用 Winsteps 5.1 软件对 6 套试卷共计 235 道题目的 98 086 名学生的作答数据进行了 Rasch 模型分析,从而获得了每道试题的 Rasch 难度值。该值为对数单位,其数值越大,代表试题的难度越高。这一客观难度值将作为后续相关分析与结构模型中的核心因变量。

从 Rasch 分析的整体结果来看,项目信度高达 0.99,这表明构建的难度量尺具有极高的稳定性。在项目拟合度方面,大部分(198/235,占比 84.3%)项目的 Infit/Outfit MNSQ 值都落在了 0.7 至 1.3 这一理想区间内。对于其余拟合欠佳的项目(多为难度极高或极低的题目),考虑到删除项目可能会损害测验的内容效度,且本研究的分析单位是“试题”而非“学生”,保留所有项目以获取完整的难度分布是必要的,本研究最终决定保留所有 235 个项目进行后续的结构分析。

(四) 数据分析方法

本研究主要使用 SPSS 26.0 和 Mplus 8.3 软件进行数据分析,严格遵循现代测量学的规范路径。

1. 描述性统计与相关分析:首先对四个潜变量(D1—D4)的总分(通过对各自的 4 个指标得分求和得到)以及 Rasch 难度值进行描述性统计与皮尔逊相关分析,以初步检验本研究的核心假设。

2. 探索性因子分析(EFA):为检验 TCRF 框架的内部结构效度,本研究对 16 个二分指标(D1_Ind_1 至 D4_Ind_4)进行 EFA。由于数据是二分类的,故采用 Mplus 8.3,并指定 ESTIMATOR=WLSMV(稳健加权最小二乘法)和 ROTATION=GEOMIN(斜交旋转),以探索 1 至 4 个因子的模型拟合情况与因子结构。

3. 验证性因子分析（CFA）与结构方程模型（SEM）：基于 EFA 的探索结果，提炼出一个结构清晰、理论意义明确且拟合优良的测量模型。在 CFA 通过检验后，纳入 Rasch 难度值作为因变量，构建最终的 SEM。所有 CFA/SEM 分析均在 Mplus 8.3 中完成，以确保对类别变量和连续变量的混合建模处理得当。模型拟合指数采用 χ^2 、RMSEA、CFI 和 TLI 进行报告。

四、研究发现

本研究的发现部分严格按照“相关—EFA—SEM”的递进逻辑展开，旨在从不同层面、由浅入深地揭示任务复杂度各维度与试题难度之间的真实关系。

（一）描述性统计与变量间关系的初步探查

为初步检验本研究的核心假设，我们首先计算了每个试题在四个潜变量（D1—D4）上的总分，并将其与 Rasch 难度值（*Rasch_logints*）进行描述性统计与皮尔逊相关分析（表 2、表 3）。

表 2 主要变量（维度总分）的描述性统计（ $N=235$ ）

	<i>M</i>	<i>SD</i>
D1	1.83	0.926
D2	1.29	0.666
D3	2.11	0.687
D4	2.29	0.878
<i>Rasch_logints</i>	0.008	1.601

注：*M* 和 *SD* 分别代表平均数和标准差

表 3 主要变量（维度总分）相关矩阵（ $N=235$ ）

	D1	D2	D3	D4	<i>Rasch_logints</i>
D1	1				
D2	0.292**	1			
D3	-0.012	0.120	1		
D4	0.380**	0.567**	0.346**	1	
<i>Rasch_logints</i>	-0.200**	-0.066	-0.064	-0.159*	1

注：* $p < 0.1$ ，** $p < 0.05$ （双尾）

表 3 的相关矩阵揭示了三个具有颠覆性的核心发现：

第一，“情境无罪”得到实证。D3（情境真实性）与 Rasch 难度值之间完全不相关（ $r = -0.064$ ， $p = 0.327$ ）。这一结果以大样本数据有力驳斥了“情境越真实，试题越难”的传统迷思。

第二，构念成功分离。D1（信息支架）与 D3（情境真实性）的相关系数仅为 -0.012 （ $p = 0.850$ ），证明“信息呈现”与“情境背景”是两个独立的维度。这意味着高真实性情境并不必然导致高信息负荷，验证了框架的结构效度。

第三，发现显著的“脚手架效应”。D1（信息支架）和 D4（认知加工）均与试题难度呈显著

负相关 (r 分别为 -0.200 和 -0.159)。这一“反常”现象揭示了素养命题的内在机制：丰富的信息 (D1) 并非阅读负担，而是作为必要的“认知脚手架”，支持了学生的高阶思维 (D4)，从而降低了由死记硬背或线索缺失导致的无效难度。这表明，优秀的素养试题能够通过提供充分的信息支持，实现“高认知要求”与“适切难度”的统一。

(二) 试题任务复杂度框架的结构探索 (EFA)

相关分析虽然直观，但它建立在将 16 个指标简单求和的基础上。为检验构念效度，必须检验 TCRF 框架 (16 个指标) 的内在结构。我们采用 EFA 来探索“这 16 个指标在数据层面是如何聚合的”以及“它们是否能清晰地指向理论上的四个维度 (D1—D4)”。

我们对 16 个二分指标进行了 EFA (WLSMV 估计, GEOMIN 斜交旋转), 模型拟合指数的比较见表 4。

表 4 TCRF 框架 (16 指标) 的 EFA 模型拟合指数比较

Model	χ^2	df	p	RMSEA	CFI	TLI
1-因子模型	211.552	104	0	0.066	0.824	0.797
2-因子模型	160.547	89	0	0.058	0.883	0.843
3-因子模型	96.049	75	0.051	0.035	0.966	0.945
4-因子模型	74.62	62	0.131	0.029	0.979	0.96

从模型拟合指数来看, 1-因子和 2-因子模型的拟合度均不理想 ($CFI/TLI < 0.90$)。而 3-因子模型和 4-因子模型均达到了优秀的拟合标准 ($p > 0.05$, $RMSEA < 0.05$, $CFI/TLI > 0.95$)。4-因子模型 ($RMSEA = 0.029$) 的拟合度略优于 3-因子模型 ($RMSEA = 0.035$)。

然而, 在检查 4-因子模型的 GEOMIN 旋转载荷矩阵 (表 5) 后, 我们有了第二个核心发现: 理论上清晰的四维 (D1—D4) 划分, 在实证数据中并不成立, 而是呈现出高度混淆的交叉负载。

表 5 EFA 四因子模型 (GEOMIN 斜交旋转) 的因子载荷矩阵 (N=235)

指标 ID	指标名称	因子 1 (认知加工)	因子 2 (信息支架)	因子 3 (复杂实践)	因子 4 (情境噪声)
D4_Ind_2	理解与阐释	0.981			
D4_Ind_3	分析与推理	0.777		0.409	
D4_Ind_1	提取与回忆	-0.742			
D2_Ind_1	单线性因果	0.781			
D1_Ind_1	多模态表征		0.826		
D1_Ind_2	数据密集度		0.62		
D1_Ind_3	信息转化		0.542	0.424	
D3_Ind_2	个人/生活情境		-0.885		0.57
D2_Ind_2	多变量控制			0.562	
D2_Ind_3	交互与制约			0.313	
D3_Ind_4	专家/科研情境			0.373	
D4_Ind_4	创造与评价			0.41	

续表

指标 ID	指标名称	因子 1 (认知加工)	因子 2 (信息支架)	因子 3 (复杂实践)	因子 4 (情境噪声)
D1_Ind_4	冗余/干扰信息				0.578
D3_Ind_1	教学/学术情境				
D3_Ind_3	社会/职业情境				-0.526
D2_Ind_4	模型构建与迁移				

注：为清晰起见，仅显示>0.3 的载荷，并按理论维度重新排序

EFA 的结果极其深刻，它否定了理论上“干净”的四维划分，揭示了实证数据中更复杂的结构。

1. 构念效度的证据：理论上的 D1（信息）和 D4（认知）在数据上形成了两个非常清晰的因子。

因子 2（信息支架）：D1_Ind_1、D1_Ind_2、D1_Ind_3 三个指标聚合良好。值得注意的是，D3_Ind_2（个人/生活情境）在该因子上呈强烈的负向载荷（-0.885），这说明“生活化”与“信息支架化”在实证上是对立的，这为“形式主义假情境”提供了数据画像。

因子 1（认知加工）：D4_Ind_2（理解与阐释）和 D4_Ind_3（分析与推理）聚合良好。最精彩的是，D4_Ind_1（提取与回忆）呈强烈的负向载荷（-0.742），这完美地证明了该因子测量的是“高阶认知”，而高阶认知的对立面（即“非高阶认知”）正是“提取与回忆”。

2. 构念混淆的证据：理论上的 D2（系统思维）和 D3（情境真实性）未能形成独立的因子，而是散落在各个因子中，呈现严重的交叉负载。

例如，D1_Ind_3（信息转化）同时在因子 2 和因子 3 上有载荷。D4_Ind_3（分析与推理）同时在因子 1 和因子 3 上有载荷。因子 3 和因子 4 更是 D2、D3、D1 指标的“大杂烩”。

EFA 分析至关重要。它证明了，如果试图在 SEM 中强行拟合一个包含全部 16 个指标的四因子（D1—D4）模型，必将失败。EFA 的结果迫使我们必须对模型进行修正，即“提纯”——保留那些结构最清晰、载荷最高、交叉负载最低的“黄金指标”，并剔除那些带来“噪声”的指标（如 D1_Ind_4，D3_Ind_1，D3_Ind_3 等）。

（三）任务复杂度对试题难度的影响：结构方程模型（SEM）分析

基于 EFA 的探索，我们采取了“模型修正”的科学策略，构建了一个更精简、稳健的 7 指标、双因子模型。该模型保留了 EFA 中结构最清晰、理论意义最强的两个构念。为清晰起见，下图仅展示潜变量路径。所有 7 个观测指标及其载荷均在 Mplus 中进行了估计。

潜变量 1：Info（信息支架），包含 3 个最纯粹的 D1 指标测量（D1_Ind_1，D1_Ind_2，D1_Ind_3）。

潜变量 2：Cog（认知加工），包含 EFA 中“认知—逻辑”因子中载荷最高的 4 个指标测量（D2_Ind_2，D2_Ind_3，D4_Ind_2，D4_Ind_4）。

1. 测量模型（CFA）的拟合优度

该 SEM（包含 CFA 测量模型）的拟合指数达到了极高的程度（表 6），卡方检验 $p = 0.312$ （不显著），表明模型与观测数据之间无统计学差异。RMSEA=0.024（远低于 0.05 的优秀标准），CFI=0.990，TLI=0.984（均高于 0.95 的优秀标准）。证明了该测量模型具有极高的构念效度，是一个测量学上稳健的、可复制的潜变量模型。

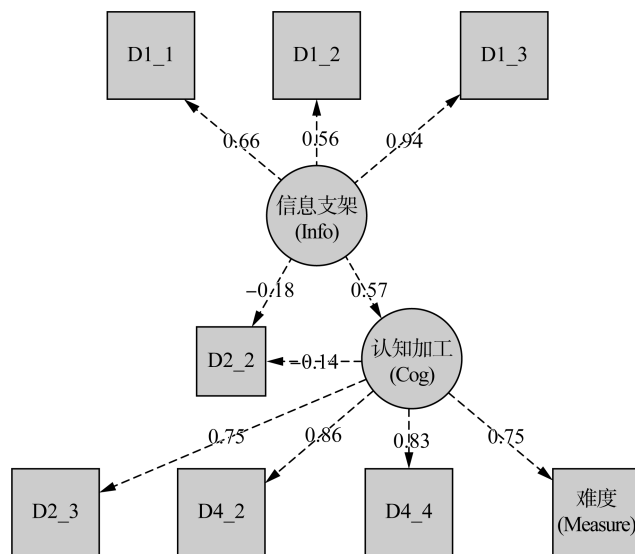


图 1 信息支架、认知加工与试题难度的结构方程模型

表 6 SEM 模型的拟合指数

指标	χ^2	df	p	RMSEA (90% C. I.)	CFI	TLI
拟合值	20.375	18	0.312	0.024 [0.000, 0.065]	0.99	0.984

同时，两个潜变量的测量指标（表 7）均呈现出高（ $\beta > 0.55$ ）且统计显著（ $p < 0.001$ ）的标准化载荷，表明 Info 和 Cog 两个构念具有极佳的收敛效度。

表 7 CFA 测量模型的标准化载荷 (STDYX)

潜变量	观测指标	β	S. E.	Est. /S. E.	P-value	R^2
Info (信息支架)	D1_Ind_1 (多模态表征)	0.665	0.101	6.549	0	0.442
	D1_Ind_2 (数据密集度)	0.556	0.09	6.173	0	0.309
	D1_Ind_3 (信息转化)	0.935	0.117	7.965	0	0.874
Cog (认知加工)	D2_Ind_2 (多变量控制)	0.75	0.081	9.279	0	0.562
	D2_Ind_3 (交互与制约)	0.861	0.097	8.883	0	0.741
	D4_Ind_2 (理解与阐释)	0.827	0.12	6.872	0	0.685
	D4_Ind_4 (创造与评价)	0.754	0.094	8.009	0	0.568

2. 结构模型 (SEM) 的路径检验

在验证了测量模型的可靠性之后，我们检验了核心的结构路径（表 8），试图回答“信息”“认知”与“难度”之间的因果关系。

表 8 SEM 结构模型标准化路径系数 (STDYX) 与 R^2

结构路径	β	S. E.	Est. /S. E.	P-value
Info→Cog (支架→认知)	0.569	0.105	5.399	0
Info→measure (支架→难度)	-0.181	0.123	-1.472	0.141

续表

结构路径	β	S. E.	Est. /S. E.	P-value
Cog→measure (认知→难度)	-0.136	0.122	-1.122	0.262
模型 R^2	Cog 的 R^2	measure 的 R^2		
解释率	0.323 (32.3%)	0.079 (7.9%)		

SEM 的结果揭示了本研究的第三个也是最深刻的核心发现。

1. “脚手架机制”的路径确认：“信息支架”（Info）对“认知加工”（Cog）具有高度显著的正向预测作用（ $\beta=0.569$ ， $p<0.001$ ），Info 潜变量解释了 Cog 潜变量 32.3% 的变异。这在因果模型上证实了相关分析中的推论——提供信息支架是激活学生高阶认知加工的关键驱动力。

2. “难度来源”再厘清：当 Info 和 Cog 同时进入模型后，它们对 Rasch 难度值的直接预测路径均不显著（ $p=0.141$ 和 $p=0.262$ ），整个模型对试题难度的总解释率（ R^2 ）也仅为 7.9%。

这并非模型的失败，而是一个极其重要的发现——相较于试题所需的具体“学科知识内容”（未进入本模型的变量），试题的“任务形式复杂度”（如信息、认知）对客观难度的直接贡献是微弱的。

这与相关分析（ $r=-0.200$ ）的结论并不矛盾。相关分析显示了 Info 与难度的总关系为负；而 SEM 则进一步证明，这种负向关系并非通过 Info 或 Cog 的直接路径产生。这再次印证，无论是提供信息支架（Info）还是要求认知加工（Cog），其本身并不会系统性地让题目变得更难或更易。

五、讨论

本研究在克服了原有研究的方法论缺陷后，基于一个更大数据集和更规范的测量模型，为“情境真实性”与“试题难度”的关系提供了更清晰、深刻的解答。

（一）“情境无罪”：为“真实情境”正名

本研究最核心的结论是：“情境真实性”本身并不驱动试题难度。

这一结论由两个强有力的证据支持：一是相关分析（表 3）显示，D3（情境真实性）总分与 Rasch 难度值完全不相关（ $r=-0.064$ ， $p=0.327$ ）；二是 EFA（表 5）显示，D3（情境真实性）的各个指标（如生活情境、科研情境）未能聚合为一个有意义的因子，而是呈现出混乱的交叉负载，表明“真实性”本身不是一个能稳定预测难度的同质构念。

这一发现有力地证伪了一线教学中“情境越真实=难度越高”的直觉误解。该误解的根源在于混淆了“任务”与“情境”的边界。在以往的经验中，高真实性的情境（如 PISA 试题）往往伴随着极高的“认知加工复杂度”（如复杂的证据推理），导致人们错误地将难度归咎于“情境”本身。

当本研究利用统计模型将“情境真实性”（D3）与“认知加工”（D4）、“信息支架”（D1）等维度进行实证分离后，真相便浮出水面：“情境”本身是无辜的，它不应为“高难度”背锅。一个情境是否贴近生活，或者是否模拟了“高大上”的科研，其本身并不会直接让题目变得更难或

更易。

（二）“脚手架效应”：信息与认知的共生机制

如果说“情境”不是难度的来源，那么什么才是？相关分析（表3）给出了一个反直觉的答案：“信息支架”（D1）和“认知加工”（D4）与难度呈显著负相关。

这一“脚手架效应”揭示了素养导向命题的真正机制。SEM（表8）的路径分析为此提供了因果解释：“信息支架”是“认知加工”的强力“助推器”（ $\beta=0.569$, $p<0.001$ ）。

这说明，命题者在设计高阶思维题目（高D4）时，往往会同时提供丰富的信息支架（高D1）。学生之所以能够完成这些高阶任务（如分析、评价），并不是因为他们天生具备了了不起的思维能力，而是因为试题所提供的图表、数据和流程（D1）为他们的思维运转提供了必要的抓手。

这一发现具有重要的教学启示，它反对两种错误的教学倾向：一是“纯知识灌输”，即只要求学生“回忆”（低D1、低D4），这会导致学生在面对真实问题时因缺乏“支架”而无从下手；二是“纯思维训练”，即在不提供充分信息材料的情况下，要求学生进行空中楼阁式的探究（低D1、高D4），这会因过高的内在负荷而导致学生产生挫败感。

本研究的实证数据倡导的是第三种模式——“支架式探究”：即通过提供结构化的“信息支架”（高D1），来支持和激活学生的高阶“认知加工”（高D4）。

（三）核心发现：情境与任务的“实证解耦”

本研究的终极价值在于回答了引言中的核心关切——如何厘清情境与任务的关系？结构方程模型（SEM）显示，“任务复杂度”潜变量对Rasch难度值的解释率仅为7.9%，且直接路径不显著。这一结果并非模型的失效，而是本研究的核心发现——它在测量学层面首次实现了情境与任务的实证解耦。

结构上的独立性：数据证明，不能用同一个维度来衡量“情境”和“任务”。SEM结果表明，“任务复杂度”与“试题难度”是两个在统计上几乎正交（独立）的构念。这意味着，单纯增加任务形式的复杂度，并不会系统性地导致试题变得更难。

功能上的异质性：情境复杂度（D3）与难度无相关性，其价值在于赋予问题以现实意义，充当认知的背景。任务复杂度（D1/D4）与难度呈负相关，其价值在于通过“信息支架”助推高阶思维，充当认知的机制。

这一结论对于纠正一线教学观念至关重要，教师不应再混淆二者，误以为引入真实情境（增加情境复杂度）就会导致学生无法完成任务。事实上，只要任务复杂度设计得当（即提供充足的信息支架），无论情境多么真实复杂，试题的难度都是可控且适切的。

（四）重构评价观：TCRF框架作为“试题质量”而非“试题难度”的标尺

本研究发现“任务复杂度”与“试题难度”在统计上相互独立，这一结果凸显了TCRF框架作为“试题内容效度（Content Validity）”诊断工具的独特价值。TCRF不再重复Rasch模型对“难度”的标定，而是提供了一把独立于分数的“质量标尺”。

甄别“伪难题”：识别那些高难度（Rasch值高）但低复杂度（低D1、低D4）的试题。这类题目往往依赖机械记忆或缺乏必要支架，属于“劣质难度”，虽能区分分数，却无法考查素养。

确证“素养题”：验证那些高复杂度但难度适中的试题，证明其通过“信息支架”成功实现了对高阶思维的有效测评。这是新课标所倡导的“让学生在解决问题中表现素养”的命题方向。

指导命题设计：SEM 中“信息支架”对“认知加工”的强路径系数 ($\beta=0.569$) 提示命题者，要设计高认知水平 (Cog) 的试题，必须同步设计高质量的信息支架 (Info)。

(五) 深层启示：从“探究”到“实践”的情境设计观

既然真实性不直接增加难度，那么我们为何还要费尽心机创设真实情境？这涉及科学教育更深层次的目标转向——从强调步骤流程的“探究”转向强调思维论证的“科学实践”。在这一新范式下，“真实情境”的核心价值不在于作为增加难度的“砝码”，而在于提供“科学实践”得以发生的“场域”。

一个模拟科研 (D3_Ind_4) 的“火星”情境，其价值在于包含了科学界真实面临的不确定性和复杂性。只有在这样的情境中，学生才必须像科学家一样去经历“定义问题、建立模型、基于证据辩论”等真实的认识论认知过程。如果仅仅是为了考一个知识点，虚构一个理想情境确实更简单直接，但要培养学生的科学思维习惯和创新能力，“真实性”则是不可或缺的手脚架。因此，教师应成为主动的“情境设计师”，大胆引入高真实性情境以确保“实践”属性，同时利用本研究的 TCRF 框架，通过精细调节信息量、变量关系复杂度和认知层级来控制绝对难度，从而设计出既有“素养高度”又符合学情“难度适切”的优质教学方案。

六、结论与展望

(一) 主要结论

通过对 235 道小学科学试题的大样本实证研究，本文得出以下主要结论：

1. 概念厘清与解耦。本研究证实，“情境真实性”与“任务复杂度”是两个在结构和功能上均相互独立的构念，不能简单地将情境的真实程度等同于任务的复杂程度或难度。
2. 情境的中性特征。“情境真实性”与试题客观难度之间无显著相关 ($r=-0.064$, $p=0.327$)。这一结果有力地反驳了“情境越真实，试题越难”的直觉假设，证明情境是素养的载体而非难度的来源。
3. 任务的支架功能。本研究定义的“任务复杂度”（特指信息支架与认知加工）表现出显著的“脚手架效应”，与试题难度呈负相关。这说明科学的任务设计能有效降低认知负荷，支持学生的问题解决。
4. 内在驱动机制。SEM 分析证实，“信息支架”是“认知加工”的强劲预测因子 ($\beta=0.569$, $p<0.001$)。这揭示了高质量任务设计的核心逻辑——用丰富的信息去驱动高阶的思维。
5. 命题实践启示。试题的难度控制不应通过“阉割”情境来实现，而应通过优化“任务设计”来达成。教师应致力于在真实情境中，通过设计高质量的信息支架，消除通往核心素养的“形式壁垒”。

(二) 研究局限与展望

本研究虽然在方法论上进行了系统的重构和改进，但仍存在一定的局限性，这也为未来的研

究指明了方向。

首先，EFA 与 CFA 的张力。本研究的 EFA（表 5）揭示了一个复杂的因子结构，而最终的 SEM（表 6）采用了一个“提纯”后的精简模型。虽然该模型拟合优异，但它也牺牲了对 D2（系统思维）和 D3（情境真实性）中部分指标（如 D3_Ind_4 “专家/科研情境”、D2_Ind_4 “模型构建与迁移”）的深入分析。未来的研究可以尝试构建更高阶、复杂的测量模型，以更全面地容纳这 16 个指标所反映的复杂信息。

其次，学生因素的缺失。本研究分析单位是“试题项目”，旨在探究试题特征对试题难度的客观影响。然而，难度最终是通过学生的主观体验和认知负荷来实现的。未来的研究应采用多层线性模型，将“学生”作为更高层级，引入学生的“阅读能力”“先备知识”“空间想象能力”等个体差异变量，探究这些学生因素与试题特征的交互作用如何共同影响问题解决的成功率。

最后，“难度”的再定义。本研究以 Rasch 难度值作为因变量，这是一种经典的定义。然而，在素养导向的评价中，“难度”或许不应仅仅是“对错”，还应包括“解决问题所需的时间”或“调用认知资源的强度”。未来的研究可以结合眼动、脑电等认知神经科学手段，测量学生在解决不同复杂度任务时的真实认知负荷，从而更立体地解开“复杂度”与“难度”之间的黑箱。

作者简介

杨煌，珠海市香洲区教师发展中心小学科学教研员。

黄真，北京师范大学未来教育学院教育学博士生，北京师范大学珠海实验学校金凤校区教师。

曾淑婷，珠海市海湾小学教师。

参考文献

- [1] OECD. PISA 2018 assessment and analytical framework [R]. Paris: OECD Publishing, 2019: 14-153.
- [2] NGSS Lead States. Next generation science standards: For states, by states [M]. Washington, D. C.: The National Academies Press, 2013. 2.
- [3] 中华人民共和国教育部. 义务教育科学课程标准（2022 年版）[S]. 北京：北京师范大学出版社，2022：2-9.
- [4] 王磊，郭玉英. 核心素养立意下 PISA 科学试题的情境特征分析 [J]. 化学教育，2018，39（12）：15-20.
- [5] Herrington J, Oliver R. An instructional design framework for authentic learning environments [J]. Educational technology research and development, 2000, 48（3）：23-48.
- [6] 郭玉英，辛涛. PISA 科学素养测评中“情境”的特征、功能及启示 [J]. 教育研究，2011，32（10），88-94.
- [7] Eilks I, Ralle B. Participatory action research within chemical education [J]. Research in chemical education-What does this mean, 2002: 87-98.
- [8] Shavelson R J, Ruiz-Primo M A, Wiley E W. Windows into the mind [J]. Higher education, 2005, 49（4）：413-430.
- [9] Sweller J. Cognitive load theory [M] // Psychology of Learning and Motivation. Vol. 55. San Diego, CA: Academic Press, 2011: 37-76.
- [10] Gilbert J K. On the nature of “context” in chemical education [J]. International Journal of Science Education, 2006, 28（9），957-976.

- [11] Roth W M. Authentic school science: The enactment of curriculum in a science classroom [M]. Dordrecht: Kluwer, 1995. 3-6.
- [12] Bennett J, Lubben F. Context-based chemistry: The Salters approach [J]. International Journal of Science Education, 2006, 28 (9), 999-1015.
- [13] 毕华林, 刘知新. 化学教材中“化学与社会”情境的类型学分析 [J]. 化学教育, 2005, (12), 1-4.
- [14] 刘菊, 王磊. “核心素养”视野下化学情境素材的选择与呈现 [J]. 化学教育, 2016, 37 (20), 10-14.

Does Contextual Authenticity Really Affect the Difficulty of Primary Science Items?

Yang Huang¹, Huang Zhengzhen^{2,3}, Zeng Shuting⁴

- (1. Teacher Development Center, Xiangzhou District, Zhuhai, 519000;
2. College of Education for the Future, Beijing Normal University, Zhuhai, 519000;
3. Zhuhai Experimental School Affiliated to Beijing Normal University, Zhuhai, 519000;
4. Bay Primary School, Zhuhai, 519000)

Abstract: Frontline teaching often mistakenly equates “contextual authenticity” with “item difficulty.” To clarify the relationship between the two, this study constructed a “Task Complexity Rating Framework” containing 16 indicators, coded 235 primary school science test items, and combined this with response data from 120, 000 samples. The study utilized the Rasch model and Structural Equation Modeling (SEM) for empirical verification. The findings indicate: (1) Construct Separation: “Contextual authenticity” and “task complexity,” which consists of information scaffolding and cognitive processing, are two structurally independent constructs. (2) Functional Difference: There is no significant correlation between “contextual authenticity” and item difficulty ($r \approx 0$), falsifying the common misconception; meanwhile, the “task complexity” dimension is significantly negatively correlated with difficulty, presenting a unique “scaffolding effect.” (3) Internal Mechanism: SEM confirms that high-quality information scaffolding can significantly activate cognitive processing ($\beta=0.569$), thereby elevating thinking levels without increasing objective difficulty. This study proves that context is merely a carrier, while true task complexity supports higher-order thinking through information scaffolding.

Key Words: Scientific Literacy; Task Complexity; Rasch Model; Contextual Authenticity; Scaffolding Effect