

创客教育中同伴互评分组策略和任务难度对大学生高阶思维的影响研究

A Study on the Effects of Peer-Assessment Grouping Strategies and Task Difficulty in Maker

Education on University Students' Higher-Order Thinking

殷蔓¹, 胡启慧^{1*}, 殷成久²

¹ 浙江师范大学教育学院

² 日本九州大学信息基础研究开发中心

* huqihui@zjnu.edu.cn

【摘要】 本研究在大学生创客教育课程中采用准实验 2×2 混合设计, 比较同质化分组互评(实验班)与随机分组互评(对照班)在不同难度任务中的思维结构表现。以 Scratch(低负荷任务)与 Arduino(高负荷任务)评语为材料, 按 SOLO 五级对评语进行编码, 并以线性混合模型检验组别、任务类型及其交互对 SOLO 得分的影响。结果显示: 组别方面, 实验班 SOLO 得分总体显著高于对照班; 任务方面, 任务难度的提升导致了两组学生互评思维结构水平的整体显著下降; 二者的交互效应显著: 任务由 Scratch 迁移到 Arduino 后两组均下降, 但实验班下降幅度更小, 说明同质化分组在高难度任务中具有缓冲作用。

【关键词】 同伴互评; 同质化分组; 创客教育; SOLO 分类; 线性混合模型

Abstract: This study adopted a quasi-experimental 2×2 mixed design in an undergraduate maker-education course to compare the thinking-structure performance of peer assessment under homogeneous grouping (experimental class) versus random grouping (control class) across tasks with different difficulty levels. Peer-assessment comments from a Scratch task (low-load) and an Arduino task (high-load) were coded using the five-level SOLO taxonomy, and a linear mixed-effects model was employed to examine the main effects of class (grouping strategy), task type, and their interaction on SOLO scores. The results indicated that, in terms of class effects, the experimental class obtained significantly higher overall SOLO scores than the control class. In terms of task effects, increased task difficulty led to a significant overall decline in both groups' peer-assessment thinking-structure levels. A significant interaction was also observed: as tasks shifted from Scratch to Arduino, SOLO scores decreased in both groups, but the decline was smaller in the experimental class, suggesting that homogeneous grouping exerts a buffering effect under high-difficulty tasks.

Keywords: Peer assessment, homogeneous grouping, maker education, SOLO taxonomy, linear mixed-effects model

1. 前言

综合来看, 学术界已形成基本共识: 同伴互评具备促进学习与认知发展的潜力, 但其效果取决于设计质量与情境条件 (Double et al., 2020; Panadero et al., 2023)。同伴互评的有效性不仅取决于“谁在评”, 更取决于评价者是否具备理解受评作品所需的背景知识。以往研究多关注基于学生学业水平的“能力同质化”, 认为相近水平的学生互动更对等 (Bi et al., 2019)。在创客与项目式学习中, 作品的多样性带来了新的挑战: 如果评价者完全不熟悉受评者的设计意图与规则, 即便两人能力相当, 评价者也需消耗大量认知资源去“读懂”作品, 而非“评价”作品。因此, 本研究提出的“同质化分组”并非指分数的同质, 而是指“任务情境的同质化”。根据认知负荷理论, 当互评双方处于相似的任务情境(如都开发赛车类游戏)时, 他们拥有共享的领域知识。

任务难度提升通常意味着认知负荷增加、评价标准更抽象、质量差异更细微。研究显示, 任务复杂度会调节同伴评价的信度与效度表现, 说明互评并不是对所有任务都同样有效或同

样稳定 (Tong et al., 2023)。创客教育中包含软件编程、硬件连接等多种任务。若忽视任务难度,可能出现“简单任务下差异不显著、复杂任务下差异被放大”的现象。但遗憾的是,现有研究较少以严谨的重复测量建模方式直接检验“任务难度 $\times$ 分组策略”的交互效应,尤其缺少以思维结构为因变量的证据链条。因此,有必要在更贴近课堂的任务迁移情境中,用适当的统计模型对交互作用进行检验。

基于以上研究不足,本研究聚焦以下研究问题:同质化分组的同伴互评是否比随机分组更能促进大学生高阶思维 (SOLO 思维结构水平),当任务由低负荷的 Scratch 迁移到高负荷的 Arduino 时,学生互评文本中的思维结构水平是否会整体下降,任务难度与同质化分组是否存在显著交互作用。

据此提出可检验假设:

H1: 使用同质化分组的同伴互评,其高阶思维显著高于随机分组互评。

H2: 当任务从低难度迁移到高难度任务时,学生互评思维结构水平的整体显著下降。

H3: 任务难度与分组策略存在显著交互作用:任务由低向高难度迁移时,随机分组的思维结构下降更明显,而同质化分组具有缓冲作用。

2. 研究设计与方法

2.1 研究设计

本研究在大学生创客课程情境中采用准实验设计,以两个平行班级作为研究对象,在不打乱原有班级的前提下实施不同的同伴互评分组策略。研究形成一个 2 (组别:同质化分组/随机分组) $\times$ 2 (任务难度:T1 Scratch 低负荷任务/T2 Arduino 高负荷任务)的混合设计,其中组别为组间变量,任务难度为组内重复测量变量。因变量为互评文本的 SOLO 结构层级得分 (1-5 分)。考虑到同一学生可能产生多条互评文本,数据具有重复测量与潜在嵌套特征。因此,本研究以线性混合模型作为主分析框架,在同一模型中同时估计组别、任务难度及其交互作用,并据此获得估算边际均值用于解释交互效应。

2.2 研究对象与分组方式

研究对象为某高校《创客基础教育》课程的两个平行班级。学生均为教育技术学专业本科生,但在 Scratch 编程与 Arduino 硬件开发方面均无前驱经验,属于初学者,确保了样本在初始知识水平上的一致性。其中 1 班为实验班 (N=35),在互评环节实施小范围同质化分组;2 班为对照班 (N=36),实施常规随机分组为控制初始差异,实验前对两班学生进行高阶思维问卷前测,结果显示无显著差异 ($p>0.05$)。

同质化分组的实施以作品类型与任务要求为主要依据。在互评前,教师根据学生提交的作品主题 (如“动作游戏类”、“教育游戏”)将学生聚类。若干大组内包括 3-4 个的小组,以保证组内成员面对相似的技术难点与设计逻辑,具备较高的情境共识;对照班系统随机抽取形成互评小组,完成同样的互评任务。

2.3 任务与过程

研究设置两个阶段任务以操控任务难度并模拟从低到高的迁移过程。T1 为 Scratch 图形化编程任务,该任务主要涉及逻辑积木拼接与基本事件控制,语法门槛相对较低,元素交互性较小;T2 为 Arduino 硬件系统开发任务,要求学生同时处理线路连接、传感器输入、控制逻辑等多重约束,元素交互性更高,认知负荷显著上升。两阶段均采用同伴互评流程:学生完成作品后,依据课程提供的互评要求对同伴作品进行评价并提交文字评语。为尽量降低“不会评”带来的噪音,研究在实施互评前对评价标准进行统一讲解,并提供典型例子帮助学生理解评价维度。

2.4 数据收集与编码

研究收集两阶段所有同伴互评文本,并以“评语”为分析单位进行编码。从数据量看,实验班在每个阶段共收集评语 72 条,对照班每个阶段共收集评语 85 条,总计每阶段 157 条 (见

表 2)。编码框架采用 SOLO 分类理论为框架进行量化表征：按照前结构、多点结构、关联结构与抽象拓展结构五个层级进行编码，分别赋值 1-5 分。

编码过程采用双盲方式，两名编码员在不知晓组别信息的条件下独立完成编码，两名编码员的一致性检验结果为 Cohen's kappa=0.837 (≥ 0.70)，编码结果具有良好的一致性，针对存在分歧的条目，两名编码员共同研讨以达成一致形成最终编码结果。数据整理与统计分析在 SPSS26.0 中完成。

3. 数据分析与结果

3.1 评语思维结构的描述性统计

首先，对实验班与对照班在 T1（低负荷任务）和 T2（高负荷任务）两个阶段的 SOLO 编码得分进行描述性统计。各组均值与标准差如图 1 所示。从图 1 可见，两组在 T1（Scratch 任务）阶段的平均得分均处于较高水平；当任务迁移到 T2（Arduino 任务）阶段时，两组得分均出现下降，提示任务难度上升可能对互评中高阶思维表现产生抑制作用。

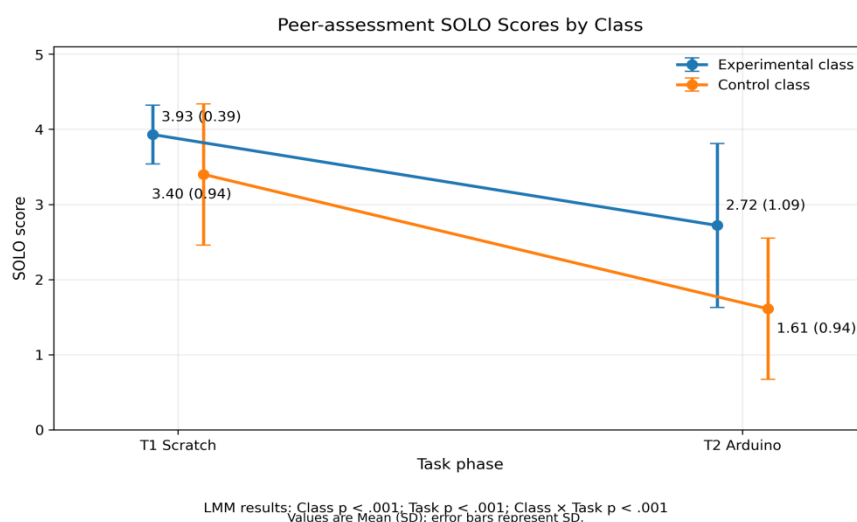


图 1 实验班与对照班同伴互评 SOLO 得分描述性统计

3.2 线性混合模型结果

为在同一统计框架内同时检验组别效应、任务难度及其交互作用，本研究以 SOLO 得分为因变量构建线性混合模型。结果可知，任务难度主效应显著， $F(1,425.033)=362.437$, $p < .001$, $\eta^2=0.460$ ，表明从 T1 到 T2，互评文本的 SOLO 得分整体显著下降；组别主效应同样显著， $F(1,425.033)=135.496$, $p < .001$, $\eta^2=0.242$ ，说明在整个研究周期内实验班总体得分高于对照班。更关键的是，组别 $\times$ 任务难度交互效应显著， $F(1,425.033)=25.999$, $p < .001$, $\eta^2=0.058$ ，意味着任务难度上升对两组的影响幅度并不一致：同质化分组在高难度阶段表现出更强的“维持效应”，能够在一定程度上缓冲得分下降。

3.3 假设检验小结

综合固定效应检验与估算边际均值解释，可以得到以下结论：

(1) Class 主效应显著，实验班（同质化分组互评）的 SOLO 水平总体高于对照班（随机分组互评），支持 H1。(2) 从低负荷的 Scratch (T1) 任务迁移到高负荷的 Arduino (T2) 任务中，实验班和对照班的 SOLO 得分整体出现显著下降，支持 H2。(3) Class $\times$ Task 交互作用显著，但考虑到分组策略与班级层面变量不可完全分离，以及任务难度与时间顺序的耦合，本研究更谨慎地将该结果解释为：研究发现与“同质化分组在高难度任务中可能具有优势”的理论预期一致，但尚不足以支持严格的因果结论。

4 讨论

4.1 主要发现与研究问题回应

整体来看，研究得到三点清晰结论：同质化分组的互评质量整体更高；任务难度提升会让

两组的互评质量都下降；而且难度越高，两种分组方式之间的差距越明显。首先看分组差异：无论在哪个阶段，同质化分组的 SOLO 得分都高于随机分组，说明分组方式并不是“流程小细节”，它会直接影响学生能否在互评中做到“抓住关键点—解释理由—给出可执行建议”。再看任务影响：当任务从 Scratch 迁移到 Arduino 后，两组 SOLO 都明显下降，说明任务难度上升会挤压互评中用于分析与推理的空间。更关键的是交互效应：实验班从 T1 到 T2 下降约 1.13 分，而对照班下降约 1.96 分；两组差距从 T1 的 0.53 扩大到 T2 的 1.36。这表明在高难度任务中，同质化分组更能“稳住”互评思维结构水平，形成缓冲效应。

4.2 主结果机制解释：为什么高难度任务会放大分组差异

要理解“难度越高差距越大”，先明确互评的本质：互评不是表态，而是标准去找证据、解释原因、提出改进方案。任务越复杂，越依赖对作品的理解和更有结构的推理；一旦学生对作品理解不稳，评语就容易退化为零散建议。这一结果可以从认知负荷理论与“认知距离”视角得到更稳健的解释。认知负荷理论指出，当任务复杂度提升时，学习者的工作记忆资源更容易被低层次加工占据，从而抑制高阶思维活动（Sweller et al., 2011；Sultanova, 2025）。在 Arduino 任务中，学生需要同时处理硬件连接、逻辑控制与调试反馈，这种高元素交互性任务显著提高了内在认知负荷。

在此情境下，分组方式通过“认知距离”影响互评过程：随机分组可能导致评价者面对不熟悉的作品结构，从而额外增加外在认知负荷；而同质化分组由于共享任务情境与问题类型，降低了理解成本，使认知资源能够更多分配给因果分析与原理迁移。

4.3 研究局限与未来方向

本研究的局限性主要体现在研究设计的内在约束。首先，分组策略在班级层面实施（每种条件仅一个班级），因此无法完全排除班级差异带来的混淆效应，这一问题在教育情境中的准实验研究中较为常见（Campbell & Stanley, 1963）。其次，任务难度与课程顺序耦合，难以区分“难度效应”与“时间/疲劳效应”。第三，样本量较小（两个班级），限制了多层模型的复杂建模与外推性。第四，未控制被评价作品质量的潜在影响，而已有研究指出，反馈内容在一定程度上受被评对象特征所制约（Sadler, 1989）。

未来研究可通过以下方式改进：（1）在同一班级内部随机分配分组策略，以实现真正的随机对照；（2）采用任务难度的对称或交叉设计以分离顺序效应；（3）扩大样本并构建三层线性模型，以更精确估计各层级效应。（4）将作品质量或评分作为协变量纳入模型，以更准确分离评价者思维水平与任务材料特征的影响。

参考文献

- Bi, Z., Tan, M., Yuan, D., & Pan, L. (2019). Investigating the validity of different peer groupings in the assessment of English writings. *International Education Studies*, 12(12), 55–61. <https://doi.org/10.5539/ies.v12n12p55>.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.
- Double, K.S., McGrane, J.A. & Hopfenbeck, T.N. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educ Psychol Rev* 32, 481–509. <https://doi.org/10.1007/s10648-019-09510-3>.
- Panadero, E., Alqassab, M., Fernández-Ruiz, J., & Ocampo, J. C. G. (2023). A systematic review on peer assessment: Intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education*, 48(8), 1053–1075.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
- Sultanova, G. (2025). Assessment tasks in STEM education: integrating Bloom's taxonomy and cognitive load theory. *Research in Science & Technological Education*. Advance online publication. <https://doi.org/10.1080/02635143.2025.2586564>

- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- Tong, Y., Schunn, C. D., & Wang, H. (2023). Why increasing the number of raters only helps sometimes: Reliability and validity of peer assessment across tasks of different complexity. *Studies in Educational Evaluation*, 76, 101233. <https://doi.org/10.1016/j.stueduc.2022.101233>.