自选赛道

人工智能与人类评分对结果满意度和公平性感知的影响

第二届全国大学生心理与行为在线实验精英赛

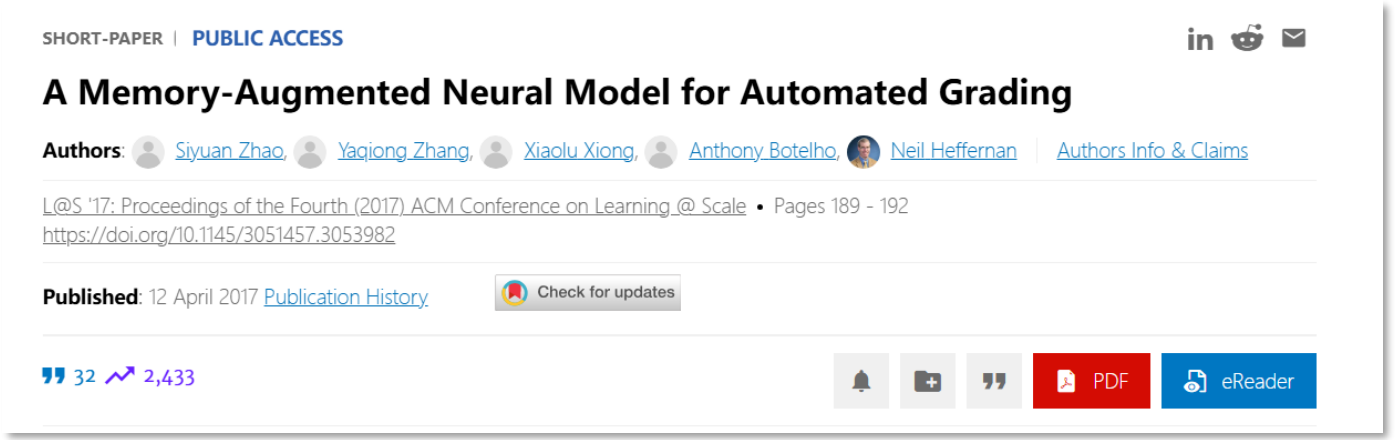
背景　人工智能+教育评价

ChatGPT出世、发展
→AI赋能教育

AI赋能教育：
“AI+教育”大模型应用成果显著，小度学习机人均使用时长提升1.25倍

AI评分系统

- AI评分系统正逐步渗透教育领域
- AI较传统教师存在诸多不同之处


SHORT-PAPER | PUBLIC ACCESS

A Memory-Augmented Neural Model for Automated Grading

Authors: Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, Neil Heffernan　Authors Info & Claims

L@S '17: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale • Pages 189 - 192
https://doi.org/10.1145/3051457.3053982

Published: 12 April 2017 Publication History　Check for updates

32　2,433　PDF　eReader

对AI评分结果与教师评分结果的感知是否存在差异?

注：数据来源于微软咨询，《“AI+教育”大模型应用成果显著，小度学习机人均使用时长提升1.25倍》

对AI评估的偏好？| 自我卷入不足与实际意愿缺失

| Chai等人（2024）在大学教育评估中发现，学生普遍认为AI评估系统相较于大学英语教师显得更公平且透明 |

自我卷入不足

仅要求学生想象特定情境而非实际体验，未能实现学生的深度参与和卷入

→ 以英语翻译题目为背景，模拟一个真实的考试评分场景

实际意愿缺失

关于学生对评估者选择的偏好与意愿的研究仍然较为稀缺

→ 以量表探究被评分者主观感知，特别关注于评价的公平性和满意度两个维度

提供对于AI评分系统在教育领域应用前景的深入见解
促进对传统评分方法的反思，调整和改进评分系统
满足教育评价的公平性和准确性要求
提升教育质量和效率

研究1 影响公平性和满意度的因素

探究评分者为不同角色（“AI评分系统”和“大学英语教师”）时、得分不同时，被评价者对于两类评价主体和评分结果的公平性和满意度感知的差异。

研究对象

- 通过Credamo平台共收集到有效问卷122份，其中男性被试37名，女性被试42名，年龄24.12 ± 6.09岁，平均作答时间7.23分钟。
- 实验开头与末尾设置自我卷入程度筛查，未通过的问卷将被拒绝。

2 × 3组间设计

得分高低分组

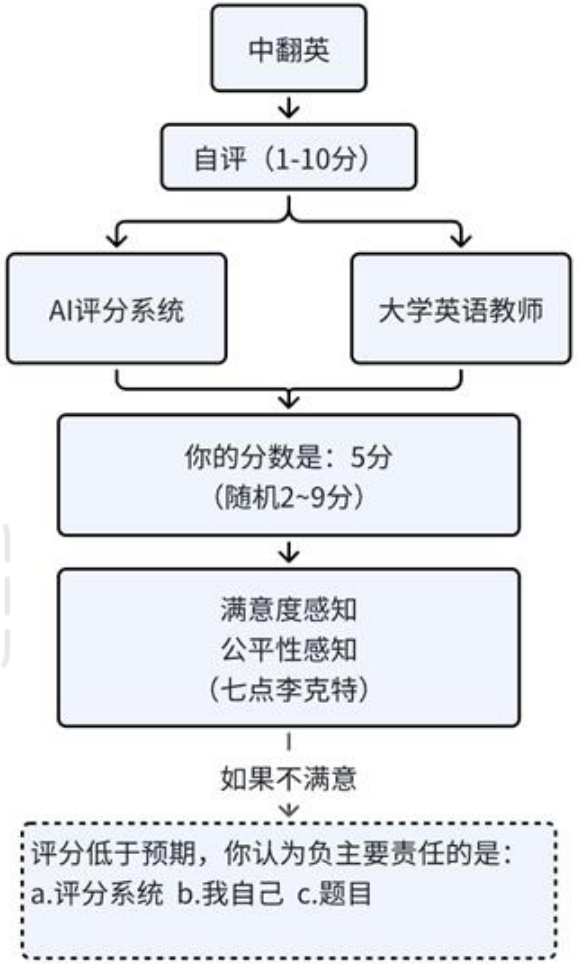
低得分（2分、3分）
中得分（4分、5分、6分、7分）
高得分（8分、9分）

自变量（组间）：

- 实际得分高低（高得分/中得分/低得分）
- 实际评分者（AI评分系统/大学英语教师）

因变量：

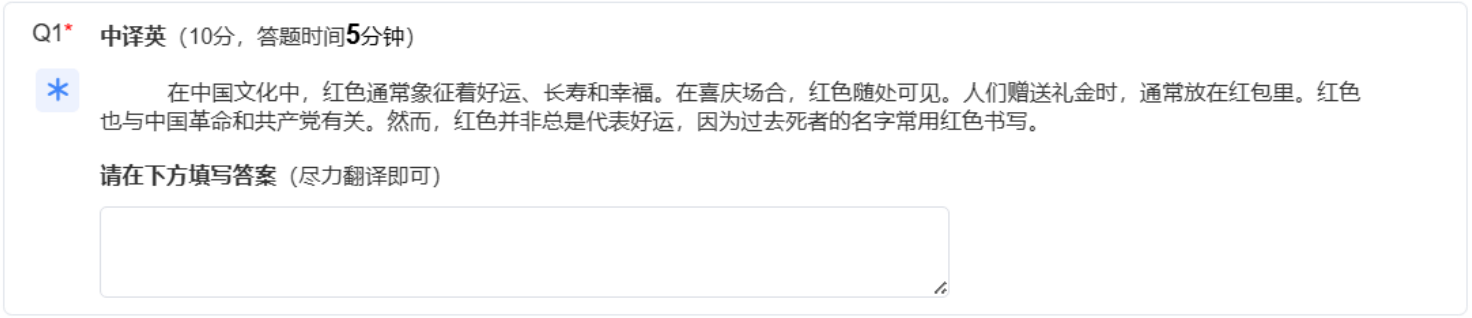
- 被试主观感知（公平性与满意度）



实验范式

英语四级中译英题目

高度自我卷入

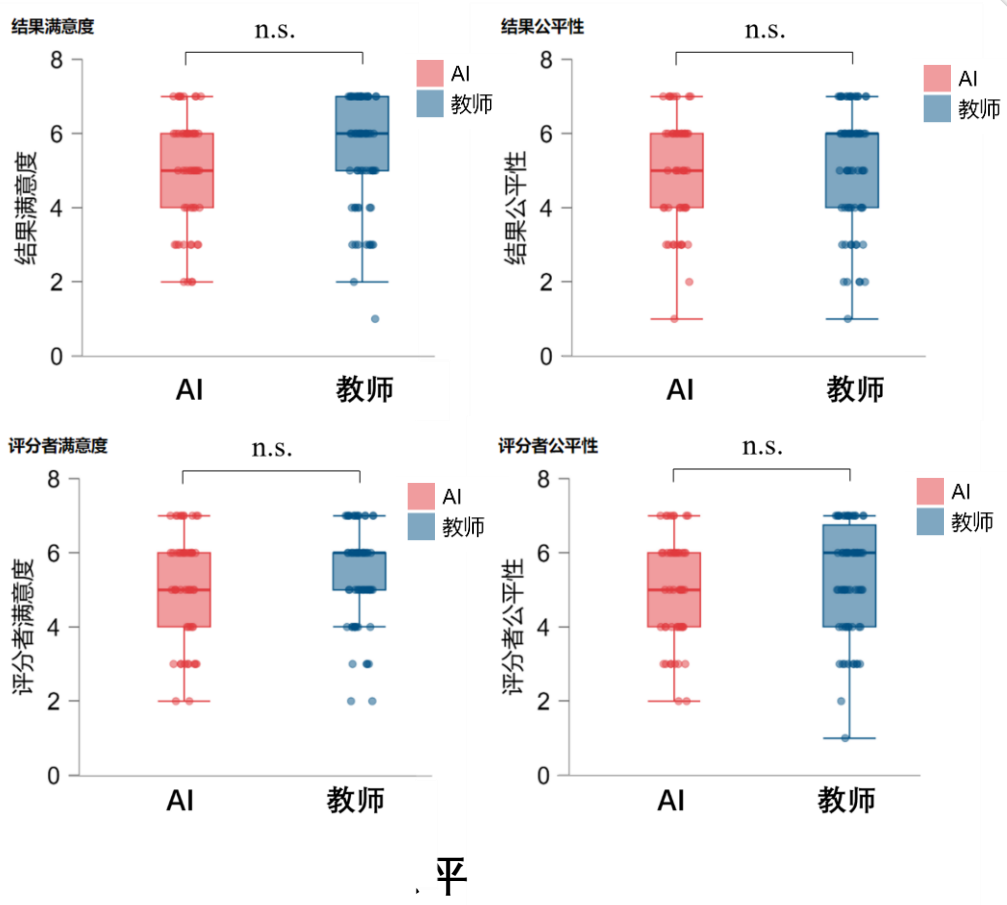


题目难易控制

- 三道题目改编自2016年12月大学英语四级考试中三份试卷，选取其中的中译英题目进行简化。

真实评分场景

研究1 不同类型评分者之间结果和对评分者的公平性感知和满意度感知无显著差异



Mann-Whitney检验结果显示：

（1）结果满意度在教师评分和AI评分之间无显著差异

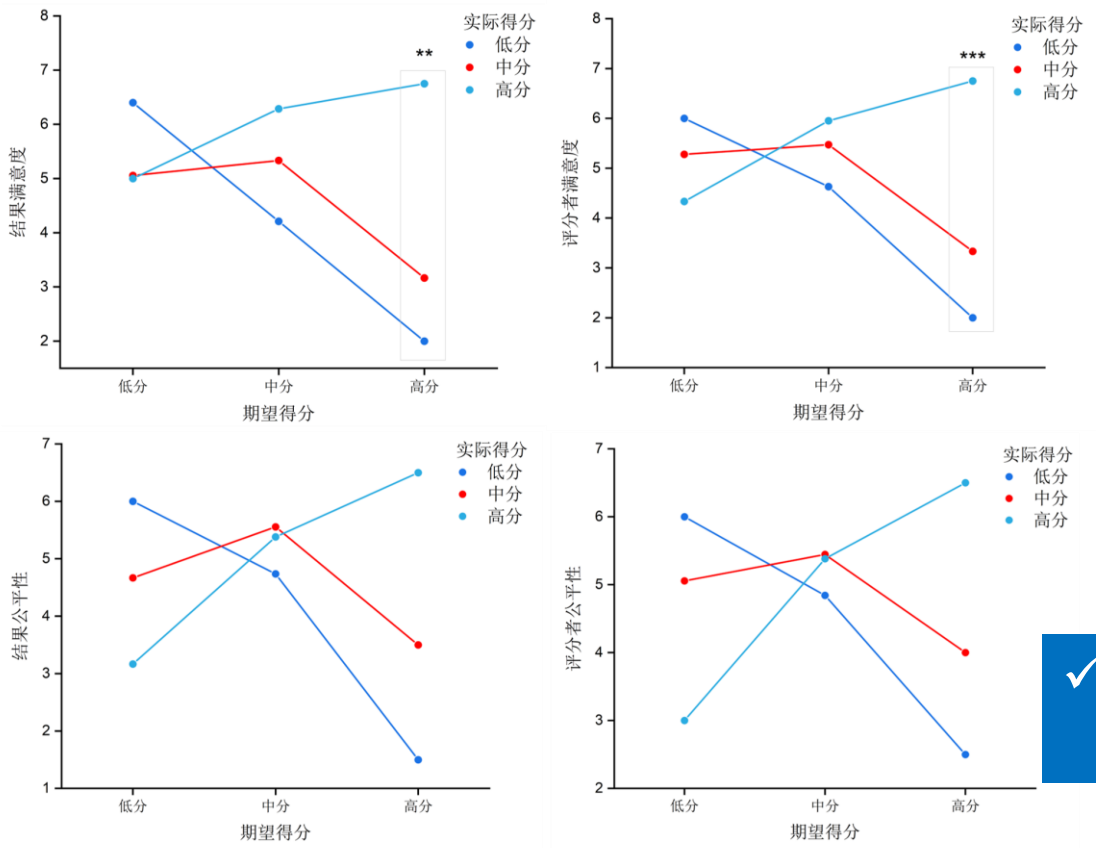
（2）结果公平性在教师评分和AI评分之间无显著差异

（3）对评分者的满意度在教师和AI之间无显著差异

（4）评分者公平性在教师和AI之间无显著差异

不同类型评分者之间，被试对结果和对评分者的公平性感知和满意度感知没有显著差异

研究1 实际得分高低对评分者的公平性感知和满意度感知影响显著，期望得分边缘显著



期望得分与实际得分对公平性和满意度的影响

多元方差分析（MANOVA）

2（评分者：AI，英语教师）
×3（实际得分：低、中、高）
×3（实际得分：低，中，高）
被试间因素
×4（显性感知：结果满意度，结果公平性，
被试内因素 评分者满意度，评分者公平性）

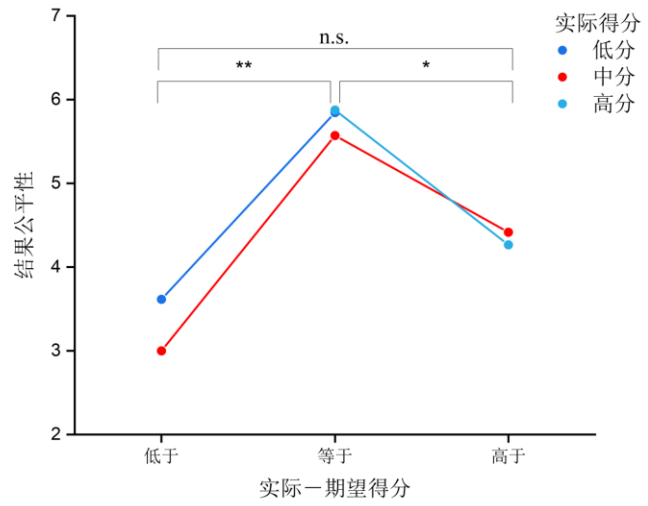
多变量方差分析结果显示：

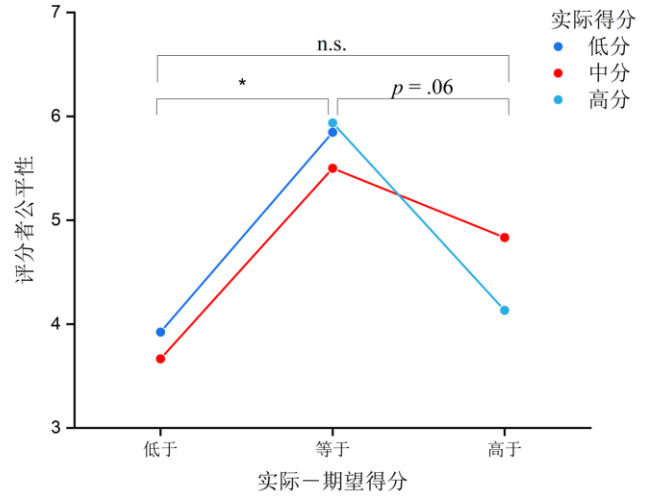
- 实际得分显著影响显性感知

✓ *假设 2a*：**实际得分是满意度感知的重要影响因素，实际得分越高，满意度感知得分越高。**

- 实际得分和期望得分交互作用显著

研究1 当期望得分与实际得分相等时，公平性评分达到最高





实际－期望得分与实际得分对公平性感知的影响

多元方差分析事后比较

结果公平性：
期望&实际得分相符＞不相符

评分者公平性：
期望&实际得分相符＞不相符

✓ ***假设 2b：*公平性感知取决于期望得分和实际得分两者的差异，当两者差异最小时，公平性感知最高**

研究2 期望与实际评分者一致性对公平性和满意度感知的影响

① 探究被试选择AI评分系统或大学英语教师作为期望评分者的选择比例差异以及期望评分对其的影响。
② 探究期望评分者和实际评分者对公平性和满意度感知的影响。

- 通过Credamo平台共收集到有效问卷120份，其中男性被试41名，女性被试79名，年龄24.30 ± 7.64岁，平均作答时间8.02分钟。

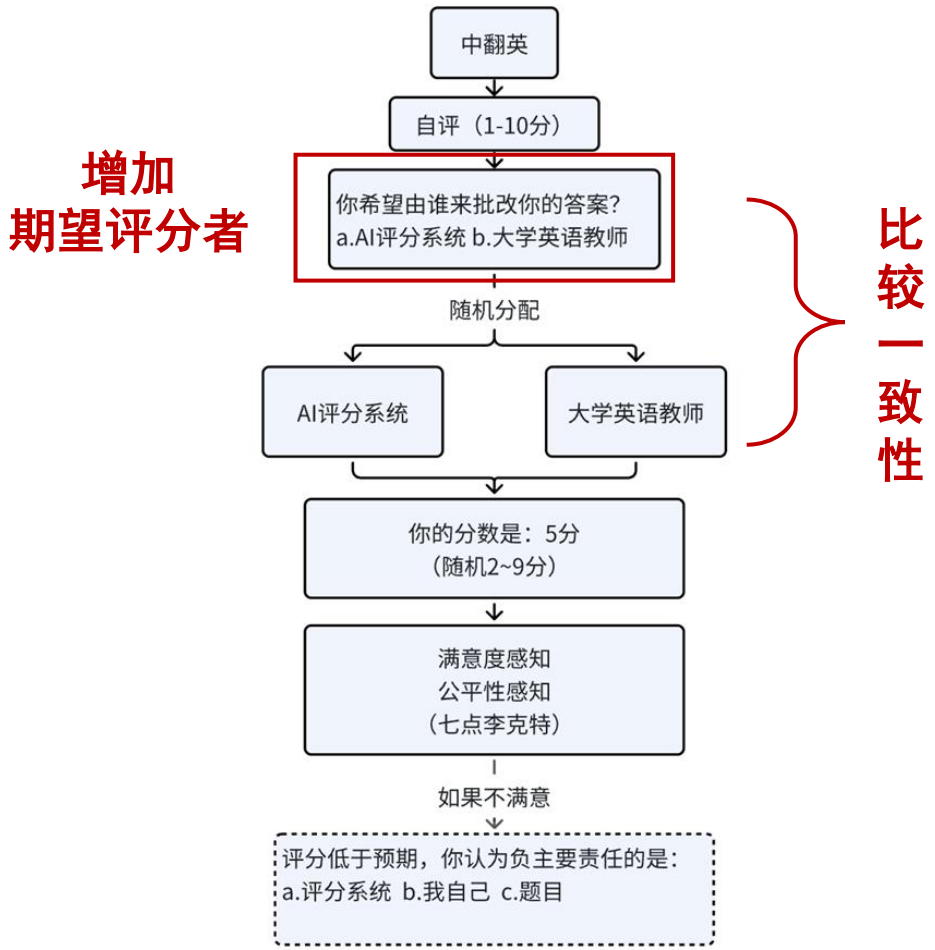
3 × 2 × 2 × 2 × 3组间设计

自变量（组间）：

- 期望得分（高期望得分/中期望得分/低期望得分）
- **期望评分者（AI评分系统/大学英语教师）**
- 实际评分者（AI评分系统/大学英语教师）
- **期望-实际评分者一致性（一致/不一致）**
- **实际得分（高/中/低得分）**

因变量：

- 被试选择偏好（选择AI评分和教师评分的比例）
- 被试主观感知（公平性与满意度）



研究2 期望与实际评分者一致性对公平性和满意度感知的影响

表 2-1 不同期望得分的期望评分者选择频数

期望得分	期望评分者	
	AI 评分系统	大学英语教师
低分（1～3 分）	18	12
中分（4～7 分）	39	40
高分（8～10 分）	3	8
总计	60	60

卡方检验

- 期望得分对期望评分者的选择比例的影响不显著

AI评分系统-60人 大学英语教师-60人

期望得分低时，被试倾向于选择AI评分系统进行打分；而期望得分高时，被试倾向于选择大学英语教师进行打分

卡方检验

- **高/低期望得分边缘显著影响**期望评分者的选择比例

研究3 对AI评分的内隐态度——IAT范式

未能在评价者偏好上得到普遍的外显结果

上下位的关系

① 探究个体对于AI评分的内隐态度
② 探究“评分”关系与“上位”关系的一致性

研究对象

- 通过Credamo平台共收集到有效问卷24份，其中男性被试11名，女性被试13名，年龄21.88 ± 3.43岁。

被试内设计

自变量:
- 概念图（人/AI）
- 属性词（点评词“评分”/受评词“受评”）

因变量:
- IAT效应

＋

被试内设计

自变量:
- 概念图（人/AI）
- 概念图（上位词“控制”/下位词“受控”）

因变量:
- IAT效应

实验材料

人的图片

类人机器人图片

点评词 评价 评分 点评 考查 审核

受评词 受议 被议 被评 受查 受审

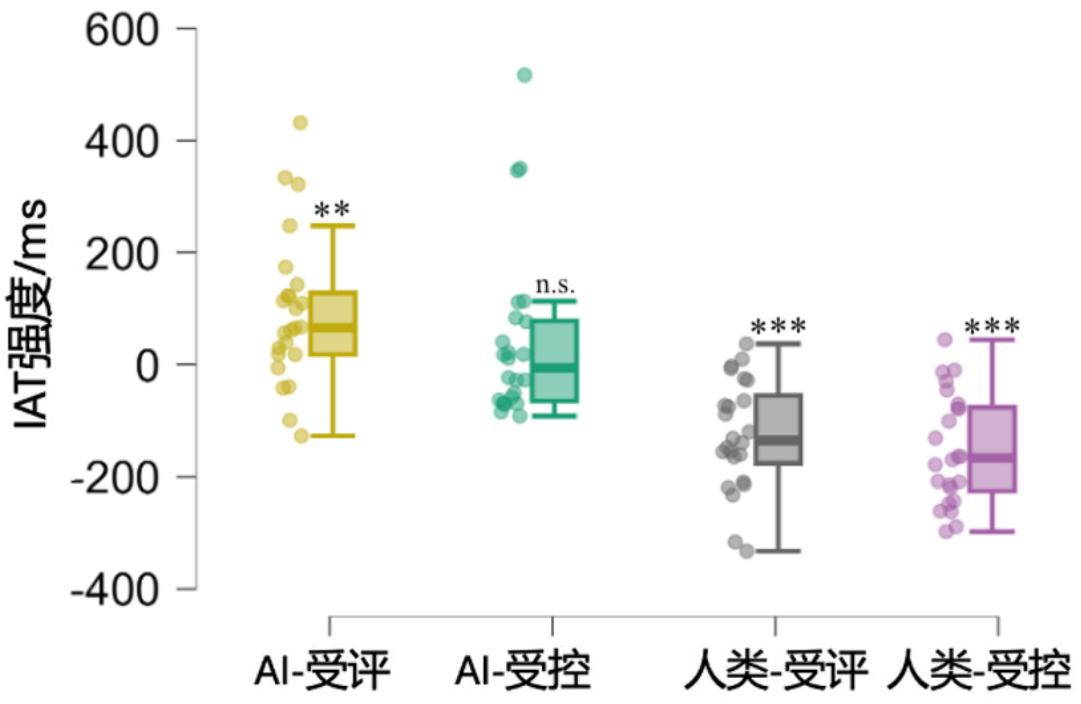
研究3 对AI评分的内隐态度——IAT范式

表 5-1 IAT 实验序列

组别	任务性质	任务类型	试验次数	功能	左键对应项目	右键对应项目
B1	相容	属性词分类	20	练习	点评	受评
B2		人/AI图分类	20	练习	人	AI
B3		联合分类	20	练习	人或点评	AI或受评
B4		联合分类	40	正式	人或点评	AI或受评
B5	不相容	人/AI图分类	20	练习	AI	人
B6		联合分类	20	练习	AI或点评	人或受评
B7		联合分类	40	正式	AI或点评	人或受评

- 共有六种刺激材料：人的图片、类人机器人图片、受评词、点评词、上位词、下位词
- 每个实验都包含相容和不相容两种情况
- 共包含7个实验组块
- 共计180个试次

研究3 对AI评分的内隐态度——IAT范式



对AI和人类受评和受控的IAT强度

以相容与否为自变量

对AI的IAT强度进行独立样本t检验

（1）评分组IAT强度显著大于0

被试对“AI-受评”的概念存在隐性偏好

（2）控制组IAT强度不显著

被试对“AI-受控”的概念无显著隐性偏好

对AI的IAT强度进行独立样本t检验

（1）评分组IAT强度显著大于0

被试对“人-点评”的概念存在隐性偏好

（2）控制组IAT强度显著大于0

被试对“人-控制”的概念存在隐性偏好

研究4 教师与AI协作评分模式的偏好与感知

描述1

Q1* 为了更好地推进AI评分系统进入校园与工作场景，我们尝试将大学教师与AI系统相结合，其一作为**主导评分者**，另一作为**辅助评分者**，对二者的描述如下

主导评分者：

责任描述：

拥有更高的决策权力。
对最终结果负有主要责任
负责制定策略、方向和标准

职责描述：

根据题目本身及试批情况，设置评分标准
对于辅助者的不确定之处进行定夺，抽取部分答卷进行二次评分

辅助评分者：

责任描述：

执行主导者制定的策略和标准。
负责协助主导者，但不对最终结果承担主要责任
遵循主导者的指导和指令

职责描述：

依据评分标准进行初步评分
对于无法确定之处向主导者提出疑问

经过多轮测试，可以认为该AI英语写作评分系统和大学英语老师水平接近

Q2* 我已经清楚知晓了**主导评分者**和**辅助评分者**之间的关系

- ○ 我已清楚
- ○ 我不清楚

研究4 教师与AI协作评分模式的偏好与感知

表 4-1 不同期望得分的评分模式选择频数

期望得分	期望评分者	
	教师主导、AI 辅助	AI 主导、教师辅助
低分（1～3 分）	15	3
中分（4～7 分）	52	13
高分（8～10 分）	7	5
总计	74	21

选择频数卡方检验

“教师主导评分、AI辅助评分”

显著高于“AI主导评分、教师辅助评分”

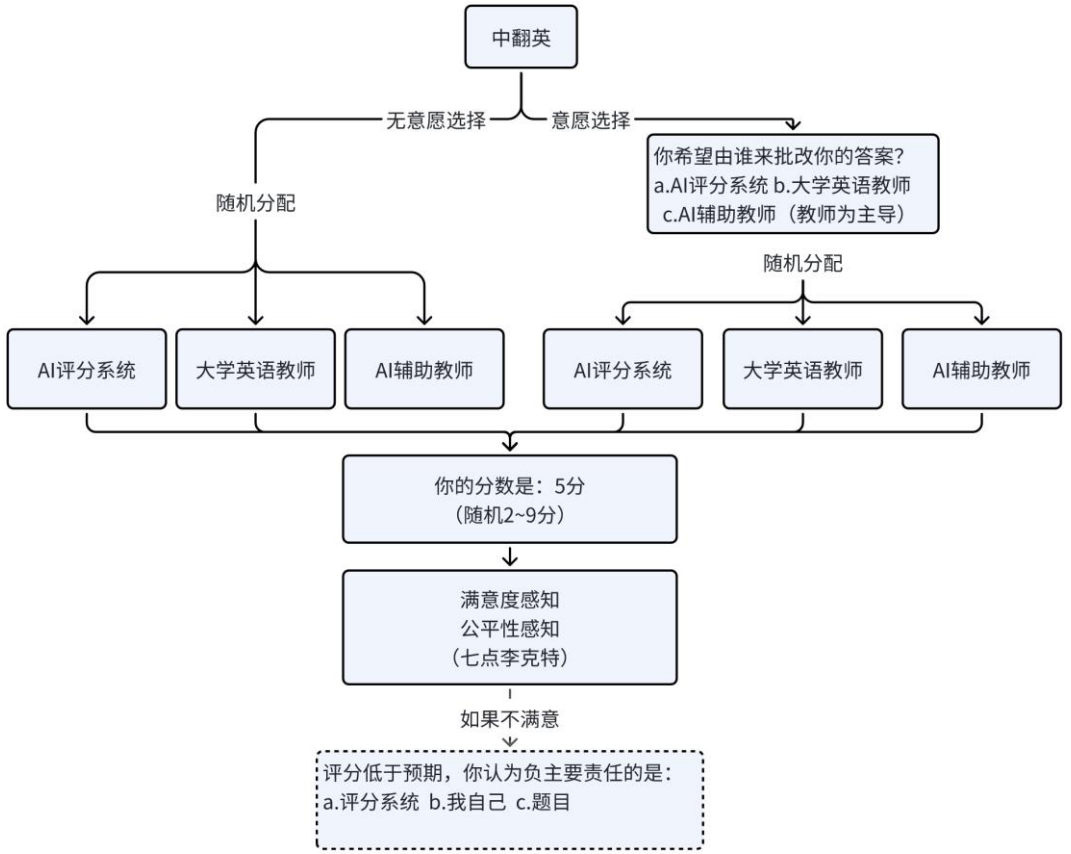
独立性检验

期望得分高低没有显著影响对“教师主导、AI辅助”评分模式的偏好

研究5 AI辅助教师评分系统的探索

①探究AI辅助教师评分系统相比单独的AI或教师评分的选择偏好和感知差异
②探究AI辅助教师评分系统和实际分数、有无期望选择以及期望一致性之间的交互作用

- 通过Credamo平台共收集到有效问卷440份，其中无意愿选择实验201份，男性被试70名，女性被试131名，年龄23.37 ± 5.73岁；有意愿选择实验239份，男性被试88名，女性被试151名，年龄21.86 ± 4.36岁。



2 × 3组间设计

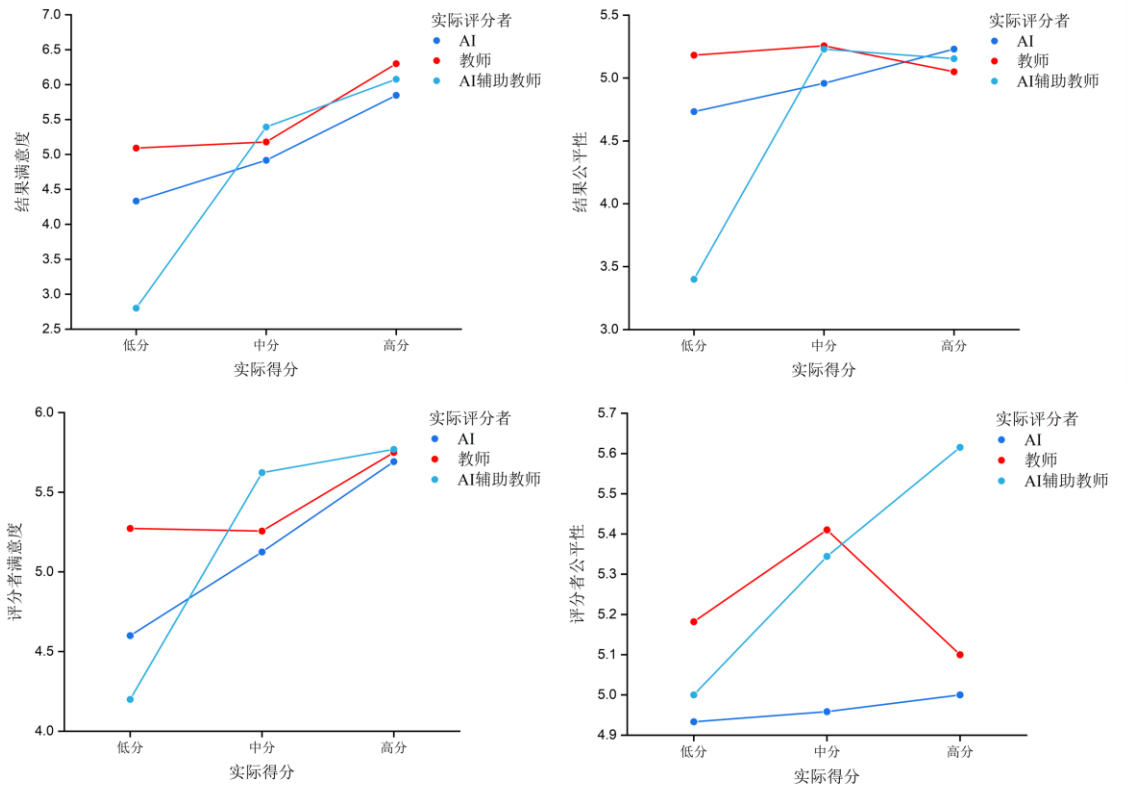
自变量（组间）：

- 评分者（AI评分系统、大学英语教师、AI辅助教师）
- 实际得分（高/中/低得分）

因变量:

- 被试主观感知（公平性与满意度）

研究5 无意愿选择实际低分时，AI辅助教师评分系统的结果满意度显著偏低



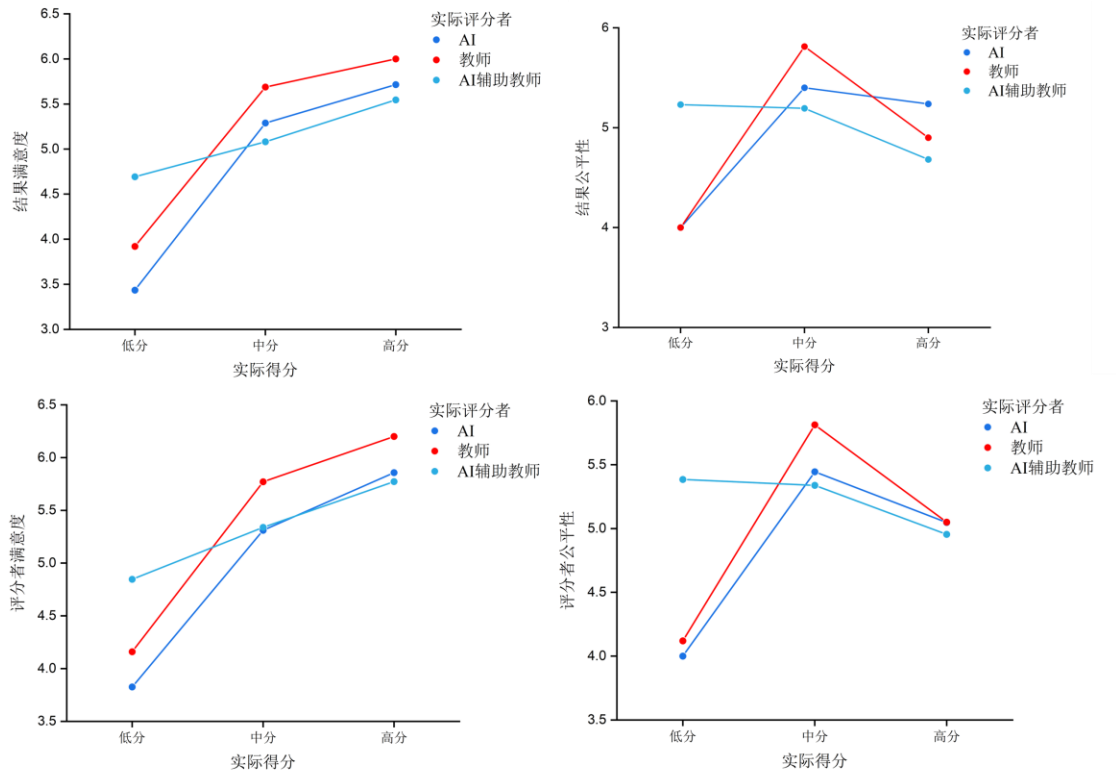
无意愿选择方差分析结果

单因素方差分析

评分者主效应边缘显著

在实际低分时，AI辅助教师评分系统的结果满意度显著偏低

研究5 有意愿选择实际低分时，AI辅助教师评分系统的结果满意度显著偏低



实际评分者和实际得分对显性感知的影响

卡方检验

- 被试存在显著偏好
  “AI辅助教师”＞AI / 教师
- 选择AI或教师评分的频数无显著差异

方差分析

- 实际评分者主效应不显著
- 期望-实际评分者一致性主效应不显著
- **实际得分**主效应显著
- 与研究3发现的内隐偏好一致，即被试更能接受“**人主导，AI辅助**”的关系

讨论

大学生被试

“答题-感知评价”线上实验

自我卷入程度↑

个人利益∝满意度、公平性

AI使用频率较高

熟悉&信任

实验设定AI与教师水平一致

- 被试对评分者无明显的选择偏好
- 评分者对主观感知影响不显著

IAT内隐测验

AI与人类上下位关系潜意识感知

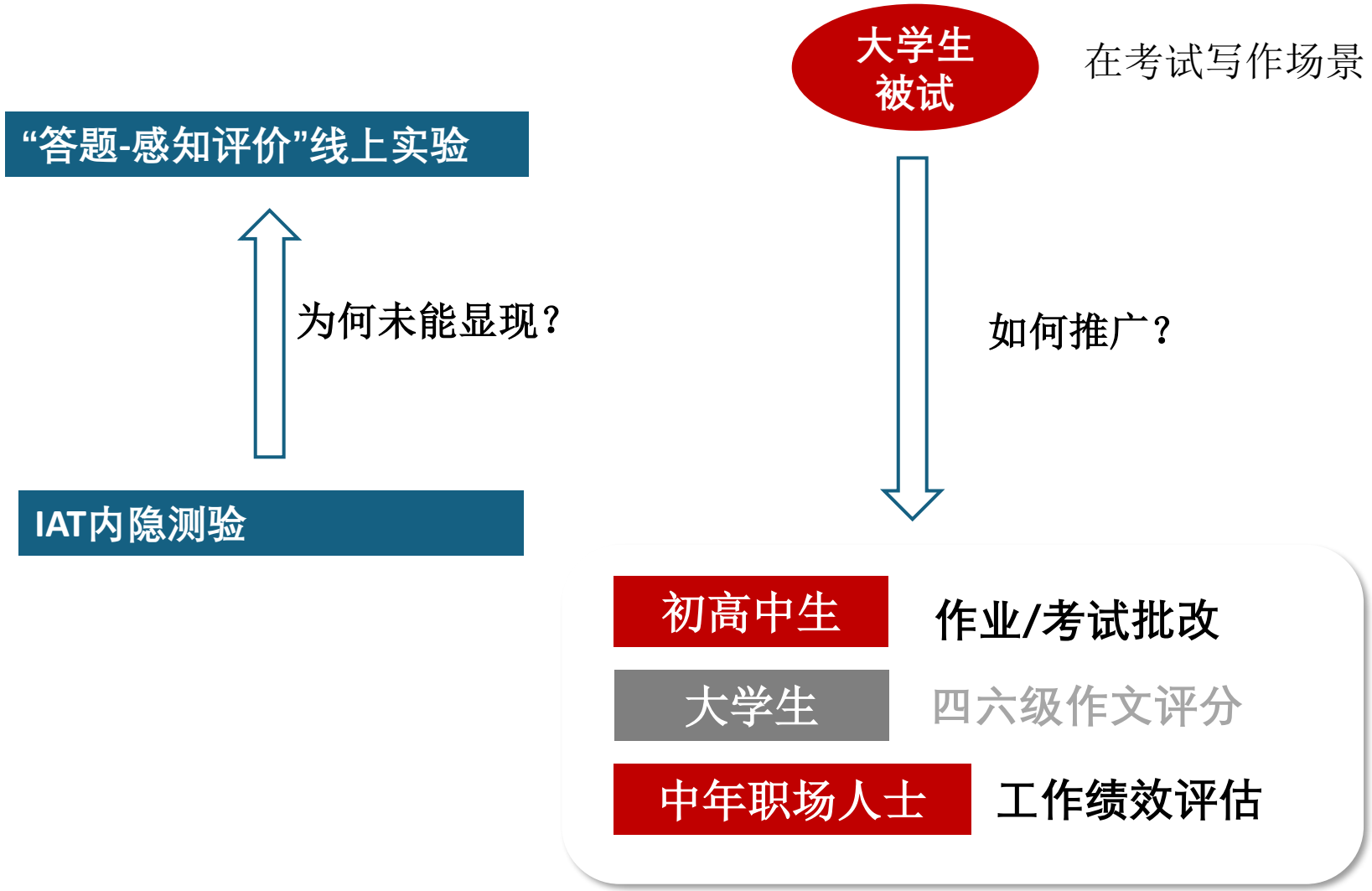
- AI更偏向于下位关系，执行“协助”“辅佐”的职能
- 解释了本项目与以往研究的差异

协作评分模式感知

倾向于AI辅助教师而非仅教师

- 对AI辅助功能的积极评价

讨论　未来研究方向


大学生
被试
在考试写作场景
“答题-感知评价”线上实验
为何未能显现？
如何推广？
IAT内隐测验
初高中生
作业/考试批改
大学生
四六级作文评分
中年职场人士
工作绩效评估

自选赛道

恳请老师批评指导！

第二届全国大学生心理与行为在线实验精英赛

问卷分享链接（Credamo见数平台）