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

问卷分享链接（Credamo见数平台）

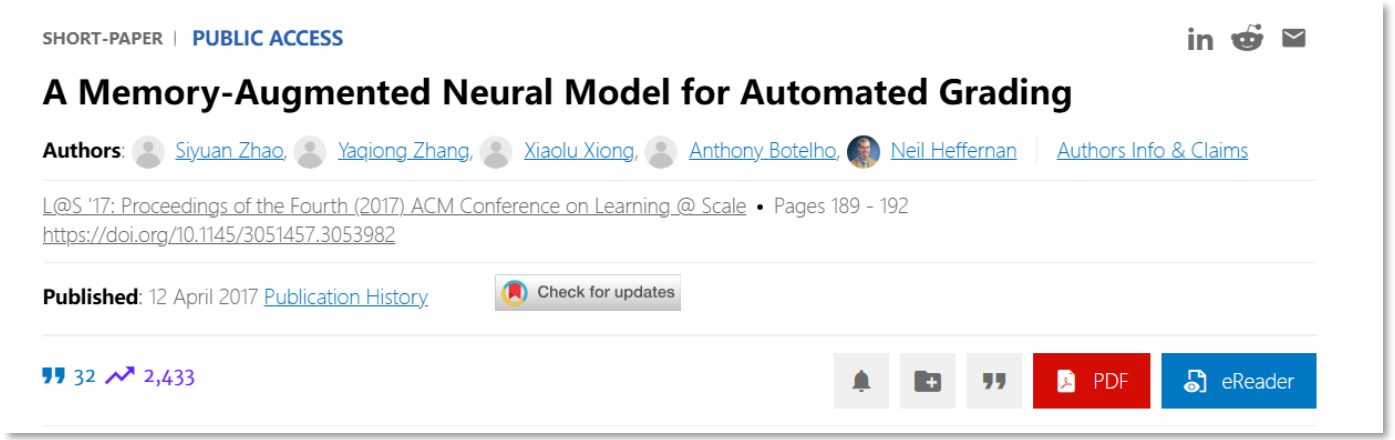
背景　人工智能+教育评价

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

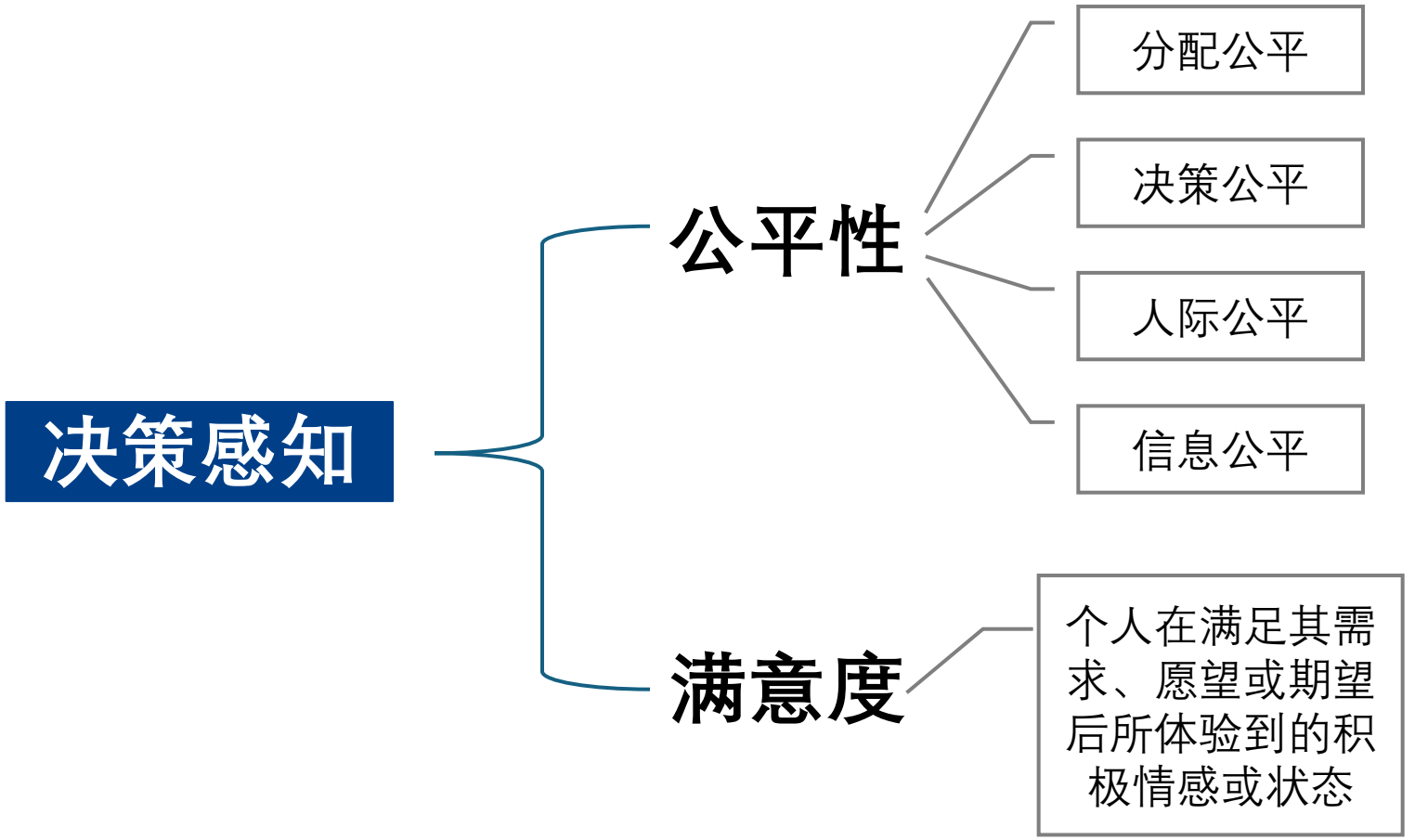
Published: 12 April 2017 Publication History　Check for updates

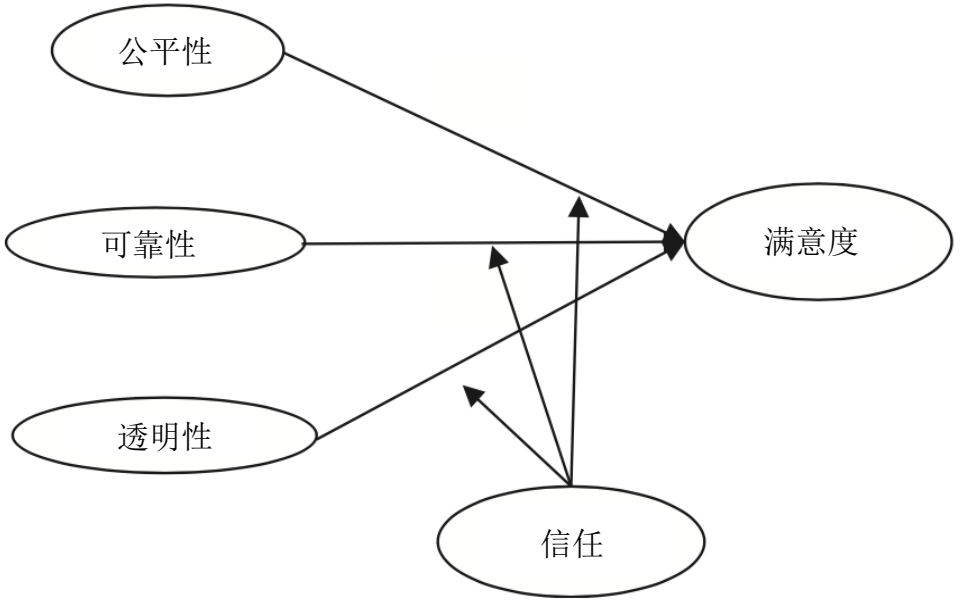
32　2,433　PDF　eReader

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》

背景　公平性与满意度感知





FAT概念模型，Shin & Park，2019

- **公平性感知与满意度之间存在显著的正相关关系**

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

问题与假设 哪些因素影响最终的分数感知?

评分主体

假设1： 公平性感知和满意度受评分者影响，被试对人类英语教师的公平性感知和满意度可能高于人工智能评分系统

期望分数与实际分数

假设2a： 实际得分是满意度的决定性因素，并且实际得分越高，满意度越高

假设2b： 公平性感知取决于预期得分和实际得分两者，当两者相符时，公平性感知最高

假设2c： 公平性感知受预期与实际差距的影响可能是不对称的，低于预期的分数被认为比高于预期的分数更加不公平

期望评分者和实际评分者

假设3： 评分者一致性组的公平性感知和满意度将显著高于评分者不一致性组

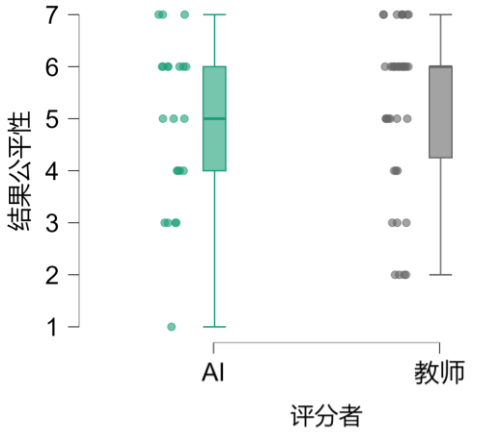
内隐态度

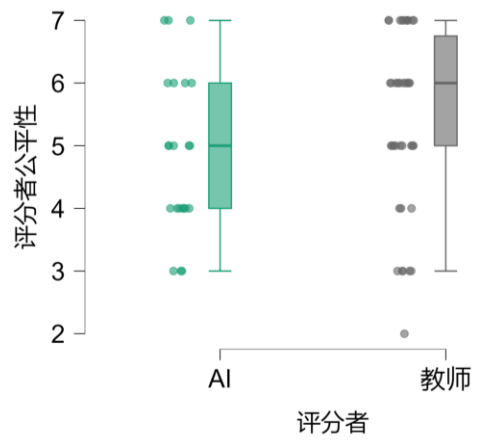
假设4a： 被试对“人类-点评”“AI-受评”以及“人类-控制”“AI-被控”这两组概念存在隐性偏好。

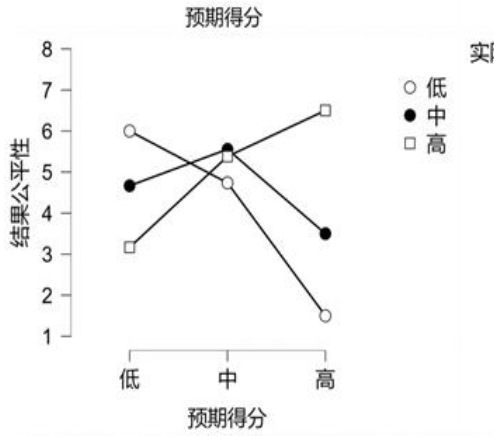
假设4b： 在点评和控制两种语境下的内隐态度具有基于个体的高度一致性

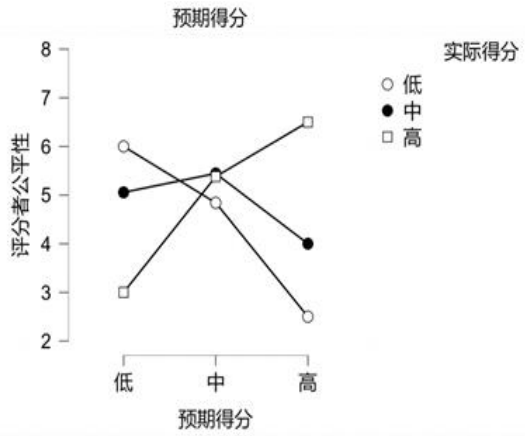
核心结果

- AI评分时被试的公平性和满意度感知均与教师评分时无显著差异。
- 公平性感知取决于预期得分和实际得分两者，且受二者差距影响具有不对称性。
- 当实际得分较低时，若被试希望得到人工评分却受到AI评分时，对满意度和公平性的感知反而会提高；反之亦然。
- 被试对“人类-上位”“AI-下位”概念存在隐性偏好，认为人类应该主导AI。



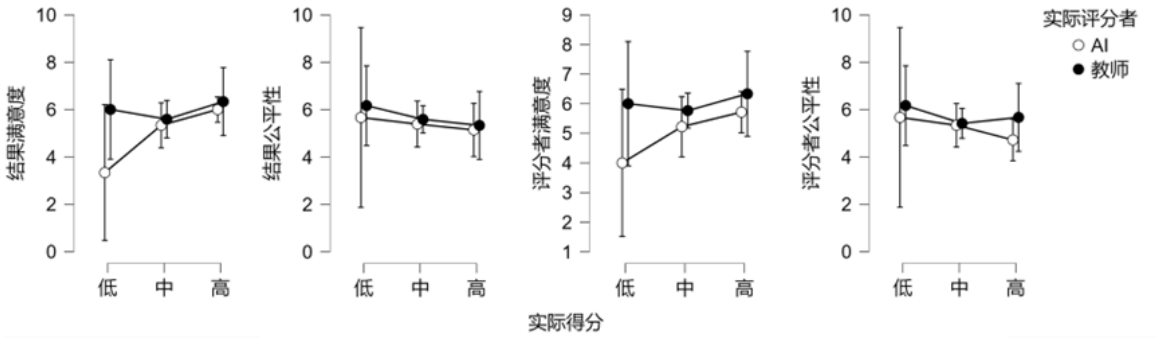


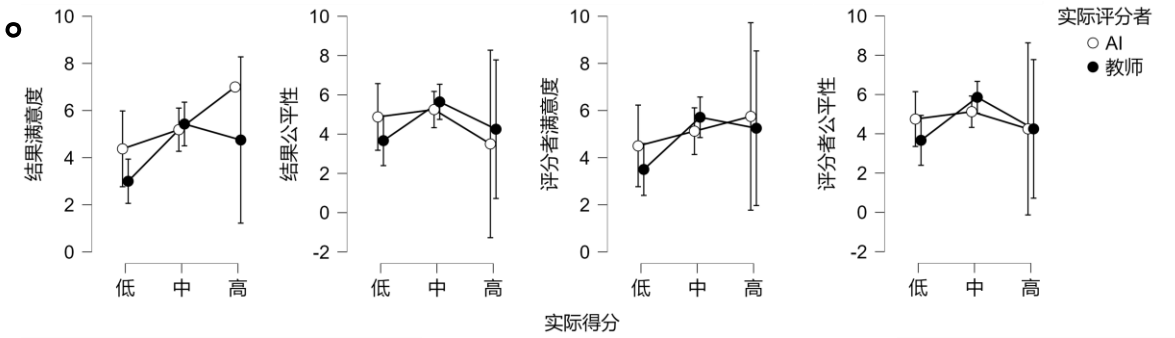


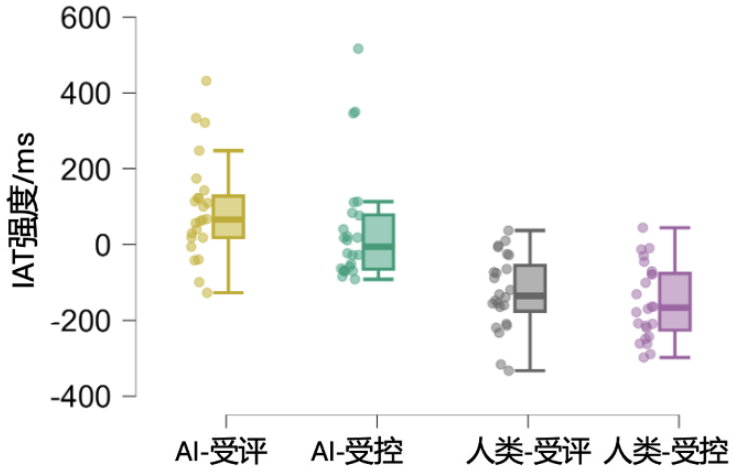


核心结果

- AI评分时被试的公平性和满意度感知均与教师评分时无显著差异。
- 公平性感知取决于预期得分和实际得分两者，且受二者差距影响具有不对称性。
- 当实际得分较低时，若被试希望得到人工评分却受到AI评分时，对满意度和公平性的感知反而会提高；反之亦然。
- 被试对“人类-上位”“AI-下位”概念存在隐性偏好，认为人类应该主导AI。







研究1 评分者类型对公平性和满意度的影响

探究评分者为不同角色（“AI评分系统”和“大学英语教师”）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60份，其中男性被试18名，女性被试42名，年龄24.93±6.58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。

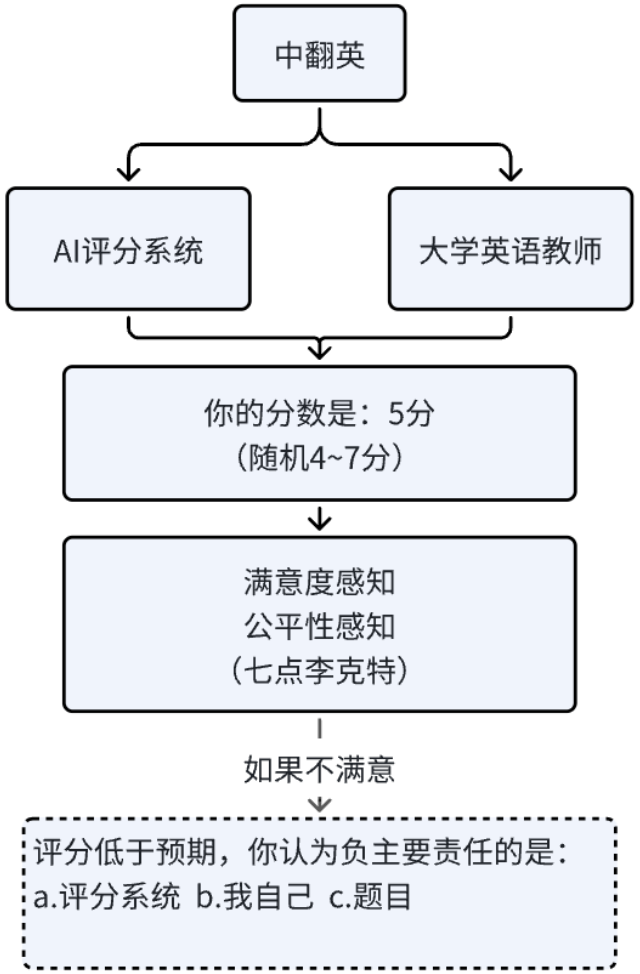
单因素完全随机设计

自变量（组间）：

- 评价者主体类型（AI评分系统/大学英语教师）

因变量：

- 被试主观感知（对评分结果和评价主体的公平性与满意度感知）



研究1 实验材料及评估

自变量控制

- 被试会被随机分为2组，即评分者为“AI评分系统”或者是“大学英语教师”

被试随机分配一种评分者类型

更具象化地感知“大学英语老师”的存在：

评分界面设置大学老师具体信息展示与延时跳转

被试作答时间控制：

被试作答时间控制在早上10点到下午6点

评分制度控制：

在评分过程中，评分制度是固定的，则评分者如何执行评分标准（如是否一致、是否合理）成为了影响学生对公平性感知的关键因素

研究1 实验材料及评估

实验材料

- 通过Credamo在线实验平台收集数据。
- 题目改编自2016年12月大学英语四级考试中的中译英题目。
- 2016年12月份共有三份四级试卷，选取其中的中译英题目进行简化。

被试随机接收其中一道题目进行作答

Q1* **中译英**（10分，答题时间**5**分钟）

*

在中国文化中，红色通常象征着好运、长寿和幸福。在喜庆场合，红色随处可见。人们赠送礼金时，通常放在红包里。红色也与中国革命和共产党有关。然而，红色并非总是代表好运，因为过去死者的名字常用红色书写。

请在下方填写答案（尽力翻译即可）

OR

Q1* **中译英**（10分，答题时间**5**分钟）

*

虽然现在白色是纯洁的象征，但在中国传统文化中，白色常用于葬礼，象征死亡和哀悼。因此，白色不应用于祝福他人康复的场合，尤其不宜送给老年人或病重者。同样，礼金也应装在红色而非白色的信封里。

请在下方填写答案（尽力翻译即可）

OR

Q1* **中译英**（10分，答题时间**5**分钟）

*

在中国文化中，黄颜色是一种很重要的颜色，因为它具有独特的象征意义。在封建社会中，它象征统治者的权力和权威。那时，黄色是专为皇帝使用的颜色，皇家宫殿全都漆成黄色，皇袍总是黄色的，而普通老百姓是禁止穿黄色衣服的。

请在下方填写答案（尽力翻译即可）

研究1 实验材料及评估

因变量评估

- 两道七点李克特量表分别对评分结果和评价主体的**满意程度**进行评分
- 两道七点李克特量表分别对评分结果和评价主体的**公平性感知**进行评分
- 当评分结果的满意度低于4分时，呈现附加选择题调查被试**低分归因**

Q2* 请问您对该**评分结果**是否满意?

非常不满意 非常满意

1 2 3 4 5 6 7

Q3* 请问您对**做出评分的AI评分系统**是否满意?

非常不满意 非常满意

1 2 3 4 5 6 7

Q4* 请问您认为该**评分结果**是否公平?

非常不公平 非常公平

1 2 3 4 5 6 7

Q5* 请问您认为该**AI评分系统**是否公平?

非常不公平 非常公平

1 2 3 4 5 6 7

当评分结果的满意度低于4分时出现：

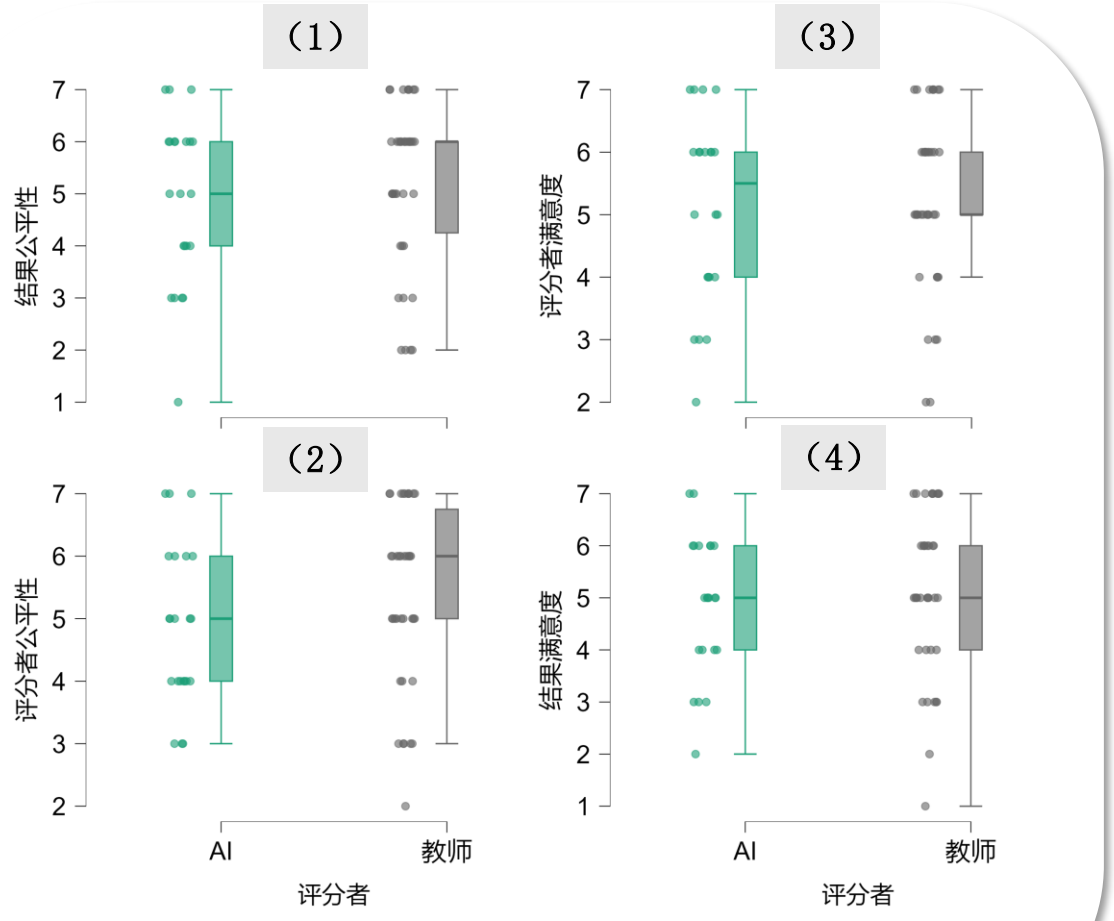
显示此问题：

请问您对该评分结果是否满意? 小于 4

Q2.1* **评分结果满意度**偏低，您认为负主要责任的是：

- ○ 评分者
- ○ 题目
- ○ 我自己

研究1 数据分析与实验结果



评分者对公平性和满意度的影响

Mann-Whitney检验结果显示：

（1）对评分者的满意度在教师和AI之间无显著差异：
$U = 394.00$，$Z = -0.38$，$p = 0.71$；

（2）评分者公平性在教师和AI之间无显著差异：
$U = 321.5$，$Z = -1.51$，$p = 0.13$。

（3）结果满意度在教师评分和AI评分之间无显著差异：
$U = 364.00$，$Z = -0.35$，$p = 0.40$；

（4）结果公平性在教师评分和AI评分之间无显著差异：
$U = 363.00$，$Z = -0.87$，$p = 0.521$；

虽然AI评分时被试的公平性和满意度感知都会低于教师评分，但是AI评分和教师评分之间，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2 实际得分和期望得分对公平性和满意度的影响

评分者类型对最后的公平性和满意度感知没有显著影响

被试的实际得分较为居中，且离散程度较低（4～7分）

- 增大实际得分的范围（2～9分）
- 添加被试分数自评环节（1～10分）

① 探究实际得分高低对不同评价主体的评分满意度和公平性感知的影响。

② 探究实际评分和期望得分的差异对结果满意度和公平性感知的影响。

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）

研究对象

- 通过Credamo平台共收集到有效问卷117份，其中男性被试35名，女性被试82名，年龄24.17±6.34岁，平均作答时间7.85分钟。
- 实验开头与末尾设置自我卷入程度筛查，未通过者将被主试操作拒绝。

完全随机设计

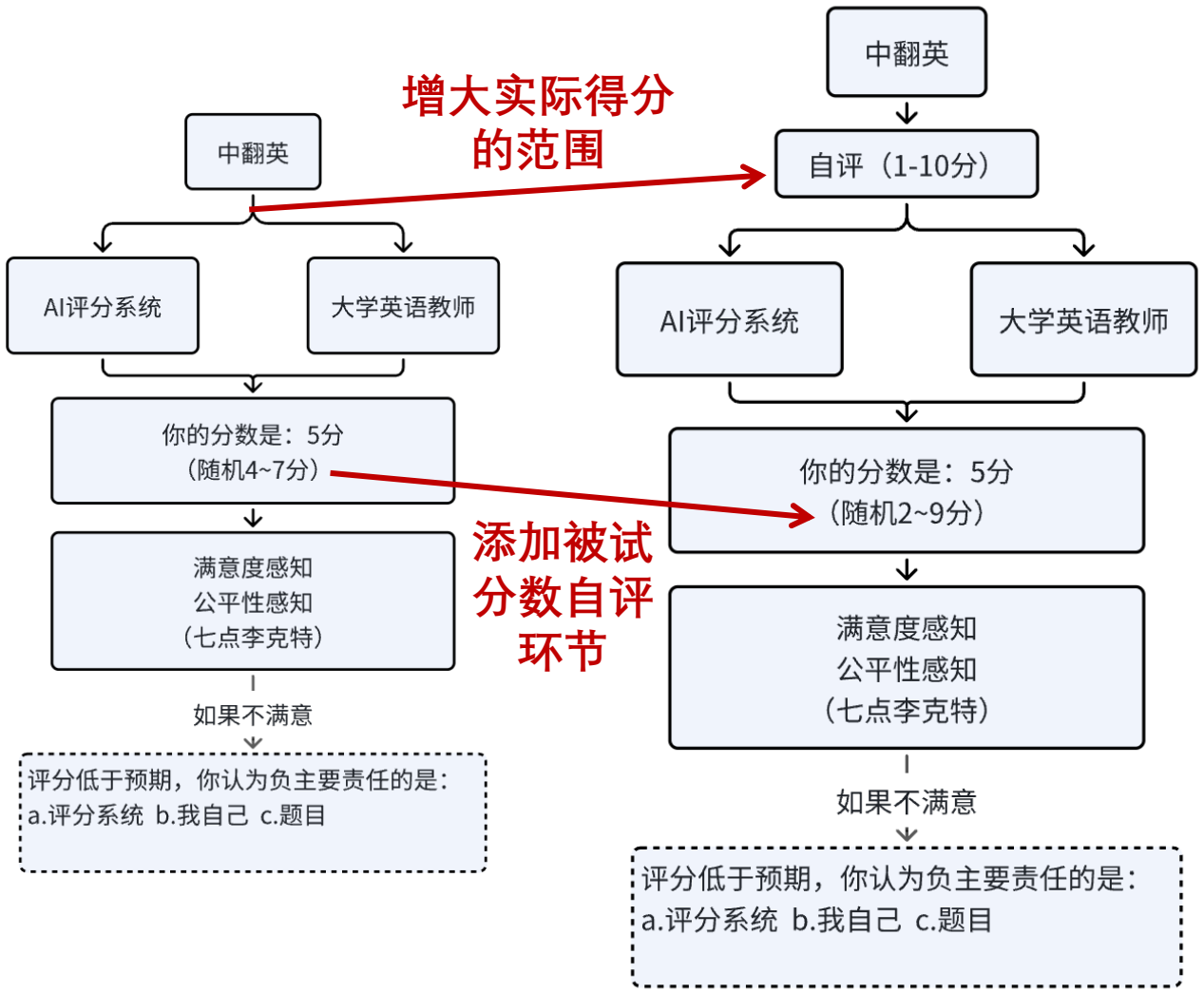
自变量（组间）：

- 实际得分高低（高得分/中望得分/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）

研究2 实际得分和期望得分对公平性和满意度的影响



研究1　　研究2

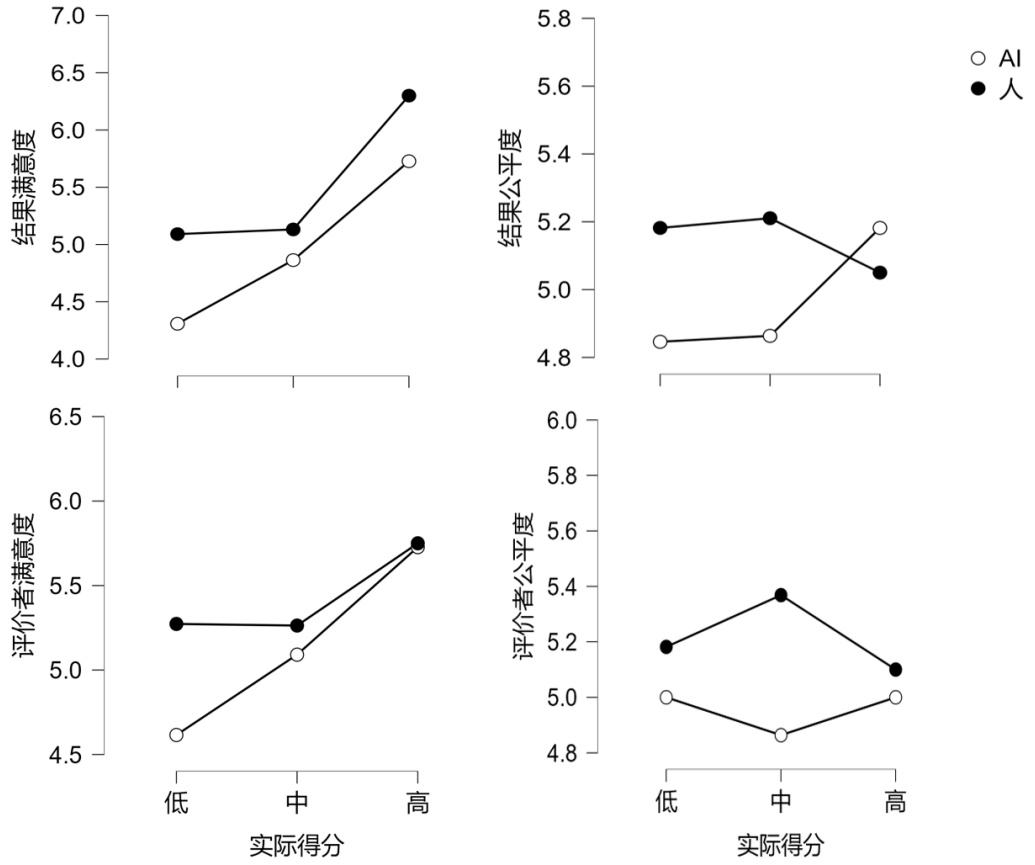
按照实际得分高低和实际评分者进行分类，被试对结果满意度、结果公平性、评分者满意度和评分者公平性感知的结果

表 2-1 实际得分者和实际评分对公平性和满意度的影响

实际评分者	显性感知	实际得分		
		低分	中分	高分
AI评分系统	结果满意度	4.31（1.93）	4.86（1.36）	5.72（1.35）
	结果公平性	4.85（1.41）	4.86（1.61）	5.18（1.66）
	评分者满意度	4.62（1.50）	5.09（1.51）	5.73（1.56）
	评分者公平性	5.00（1.53）	4.86（1.28）	5.00（1.79）
大学英语教师	结果满意度	5.09（1.58）	5.03（1.58）	6.40（0.98）
	结果公平性	5.18（1.25）	5.11（1.62）	5.05（1.93）
	评分者满意度	5.27（1.27）	5.17（1.42）	5.75（1.07）
	评分者公平性	5.18（1.25）	5.26（1.44）	5.10（1.74）

其余同研究1

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分影响



实际得分者和实际评分对公平性和满意度的影响

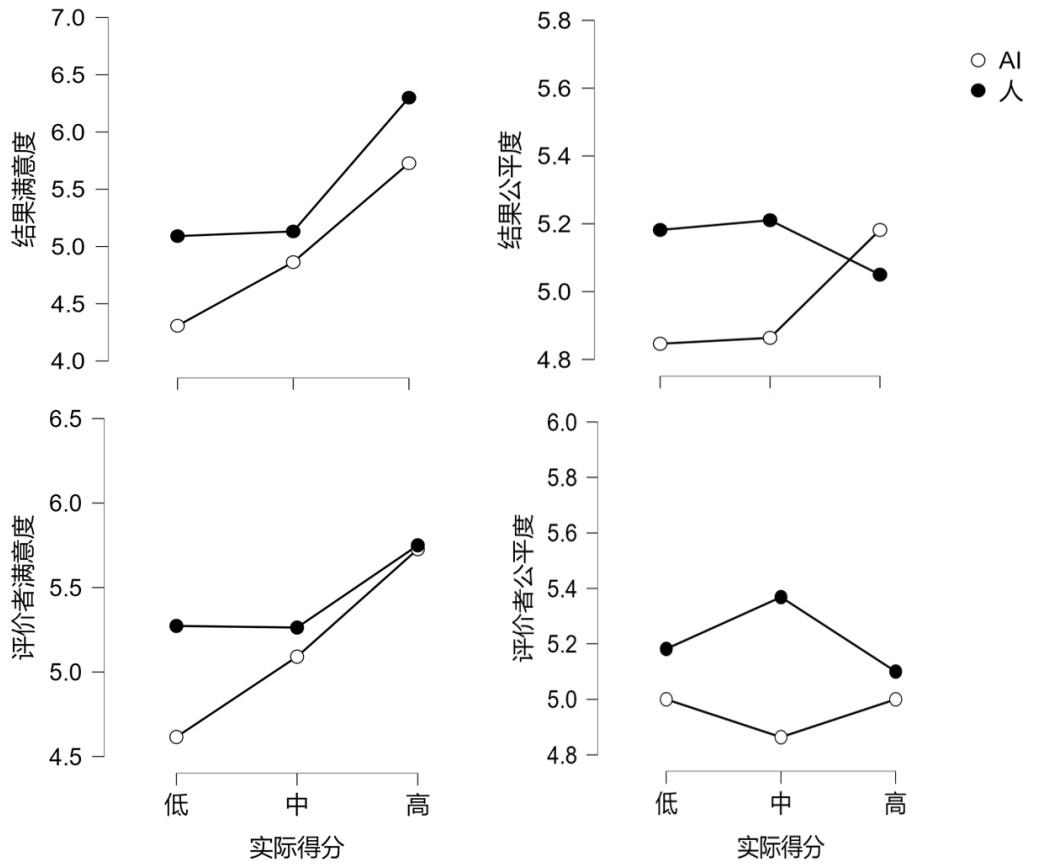
多元方差分析（MANOVA）

2（评分者：AI，英语教师）
×3（实际得分：低，中，高）} 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：

- 评分者类型对显性感知影响不显著
  $F(4, 106) = 1.298$，$p = 0.442$，$\eta_p^2 = 0.047$
- 实际得分高低对显性感知影响显著
  $F(8, 212) = 4.411$，$p = 0.007$，$\eta_p^2 = 0.094$
- 评分者类型和实际得分高低交互作用不显著

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分影响



实际得分者和实际评分对公平性和满意度的影响

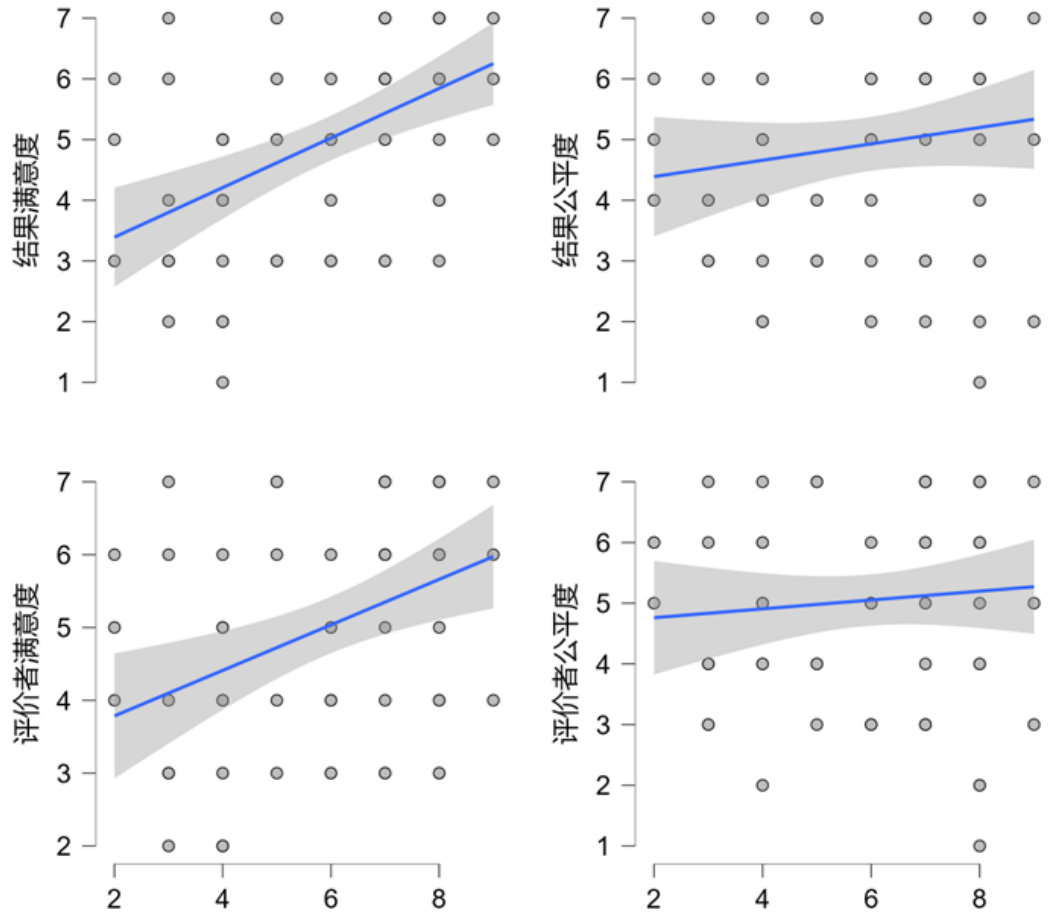
多变量方差分析结果显示：

- 评分者类型对显性感知影响不显著
  $F(4, 106) = 1.298$，$p = 0.442$，$\eta_p^2 = 0.047$
- 实际得分高低对显性感知影响显著
  $F(8, 212) = 4.411$，$p = 0.007$，$\eta_p^2 = 0.094$
- 评分者类型和实际得分高低交互作用不显著

单变量方差分析结果显示：

- 实际得分高低对结果满意度的影响显著
  $F(2, 114) = 8.96$，$p < 0.001$，$\eta_p^2 = 0.136$
- 实际得分对评分者满意度的影响显著
  $F(2, 114) = 15.69$，$p = 0.024$，$\eta_p^2 = 0.064$
- 实际得分高低对结果的满意度对结果公平性和评分者公平性的感知没有显著影响。

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分影响



实际得分与满意度和公平度相关性

相关性分析结果显示：

- 得分高低和结果满意度呈现显著的正相关
  Kendall's tau = 0.373, $p < 0.001$
- 得分高低和对评分者满意度呈现显著的正相关
  Kendall's tau = 0.270, $p < 0.001$
- 得分高低与公平性感知之间无显著相关

在实际任务中，评分者是AI还是教师并不会显著影响被试的满意度感知，而实际得分高低对最后的满意度感知有着显著的影响。

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分和期望评分差异分析

实际得分并没有显著影响公平性感知，且二者之间并未发现显著的正相关

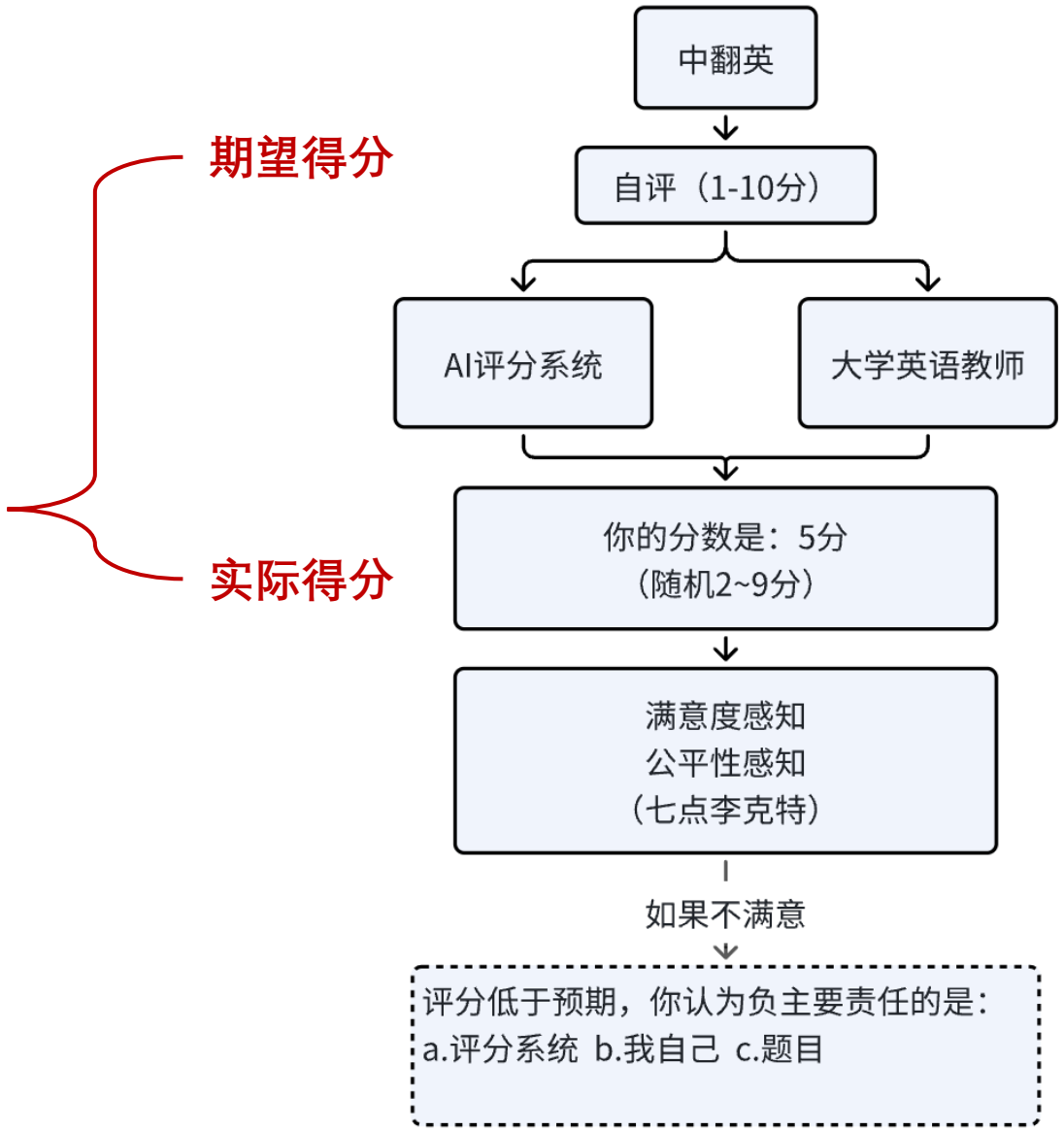
可能由实际得分与期望得分共同影响

- 将期望得分也作为满意度和公平性的影响因素

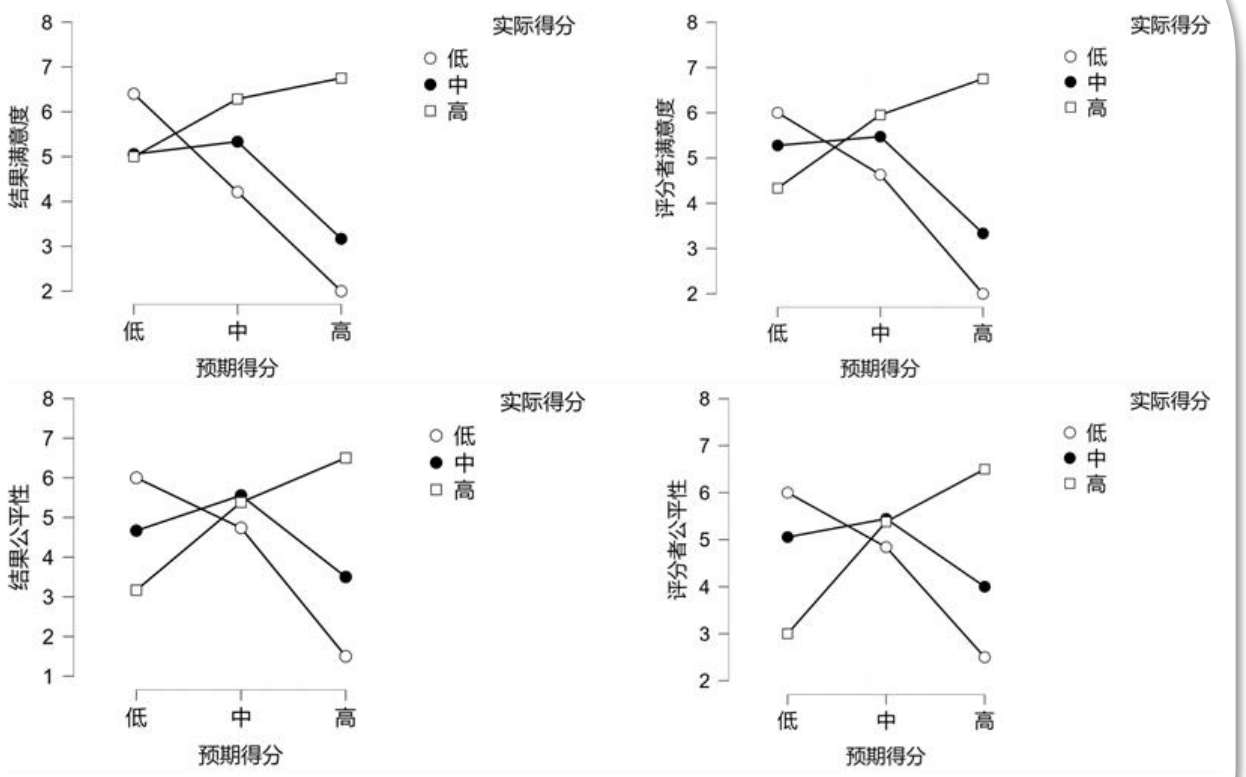
探究不同评分者类型之间，实际得分和期望得分的差异对结果满意度和公平性感知的影响

研究对象

- 通过Credamo平台共收集到有效问卷65份，其中男性被试25名，女性被试40名，年龄24.77 ± 8.59岁，平均作答时间7.69分钟。
- 实验开头与末尾设置自我卷入程度筛查，未通过者将被主试操作拒绝。



研究2 实际得分和期望得分对公平性和满意度的影响——实际评分和期望评分差异分析



显性感知作为预期得分的函数图

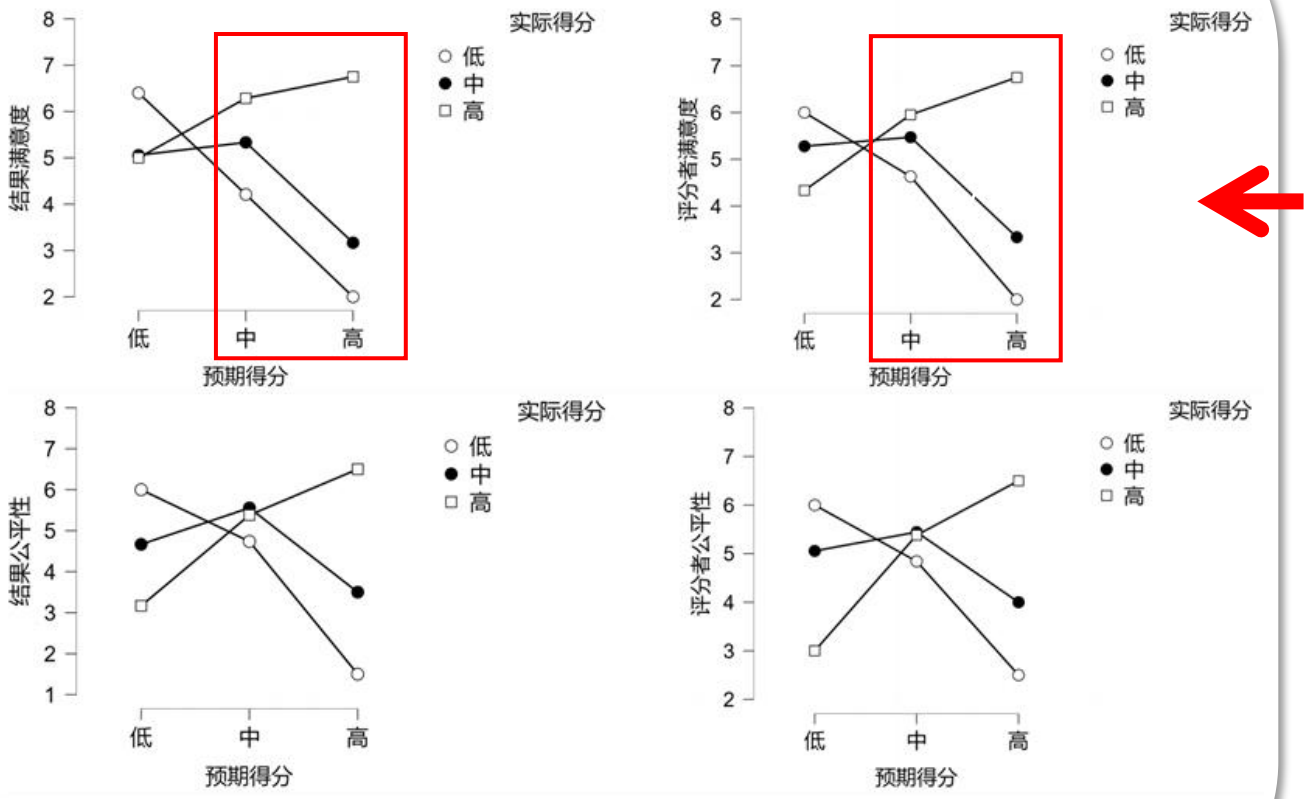
多元方差分析（MANOVA）

2（评分者：AI，英语教师）
×3（预期得分：低，中，高）
×3（实际得分：低，中，高）
} 被试者间变量
×4（评分维度：结果满意度，结果公平性，评分者满意度，评分者公平性）
被试内变量

多变量方差分析结果显示：

- 结果显示实际得分会显著影响显性感知 $F(8,190)=3.992$，$p<0.001$，$\eta_p^2=0.144$;
- 预期得分会显著影响显性感知 $F(8,190)=2.492$，$p=0.014$，$\eta_p^2=0.095$;
- 实际评分者不会影响显性感知 $F(4,95)=0.485$，$p=0.747$；
- 实际得分和预期得分交互作用显著 $F(16,291)=2.377$，$p=0.002$，$\eta_p^2=0.090$。

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分和期望评分差异分析



显性感知作为预期得分的函数图

在预期得分为中和高的维度上时，结果满意度与评分者满意度均随实际得分的增高而增高

- 主效应显著

结果满意度：

$F(2,98) = 13.978,\ p < 0.001,\ \eta_p^2 = 0.222$

评分者满意度：

$F(2,98) = 10.43,\ p = 0.002,\ \eta_p^2 = 0.123$

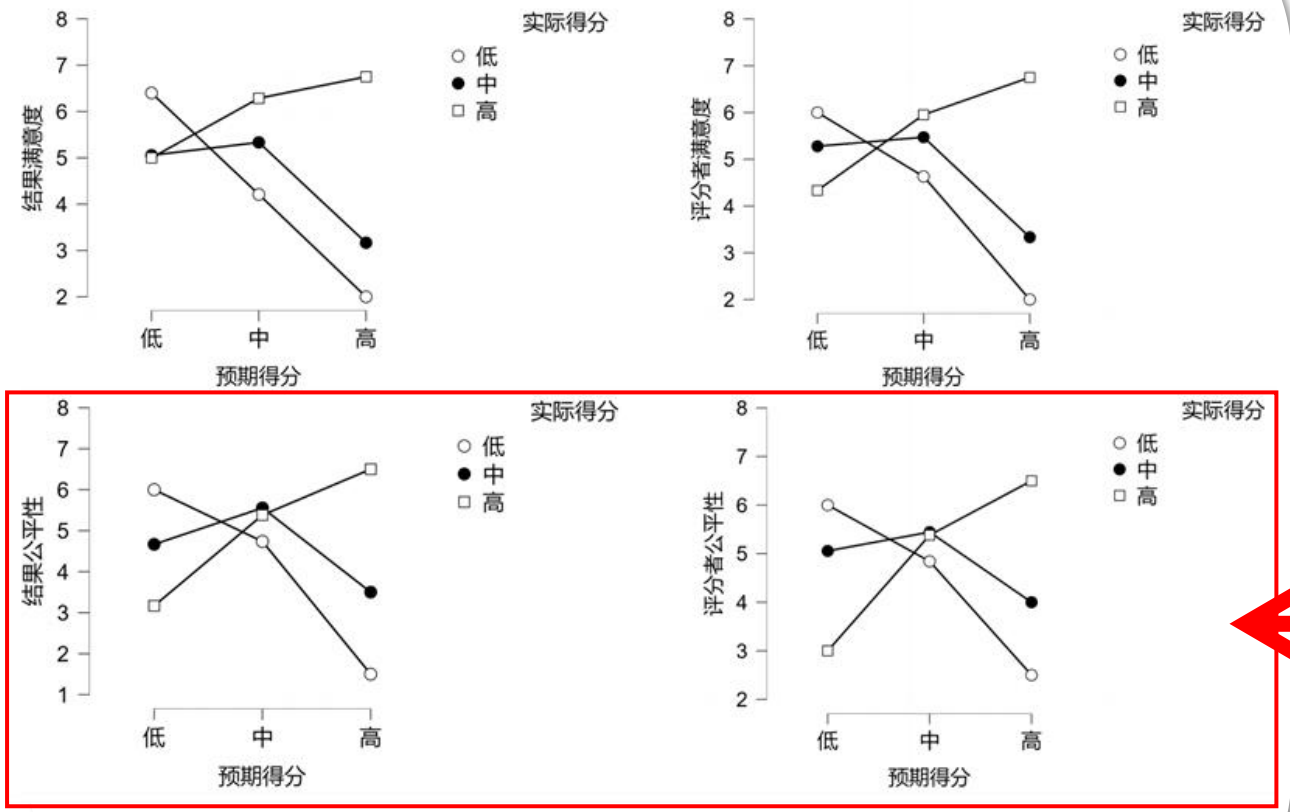
满意度与预期得分的关系并不如与实际得分那样显著，仅有结果满意度评分随预期得分的增高而降低

- 主效应显著

$F(4,98) = 4.462,\ p = 0.014,\ \eta_p^2 = 0.083$

实际得分是满意度的决定性因素，并且实际得分越高，满意度越高

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分和期望评分差异分析

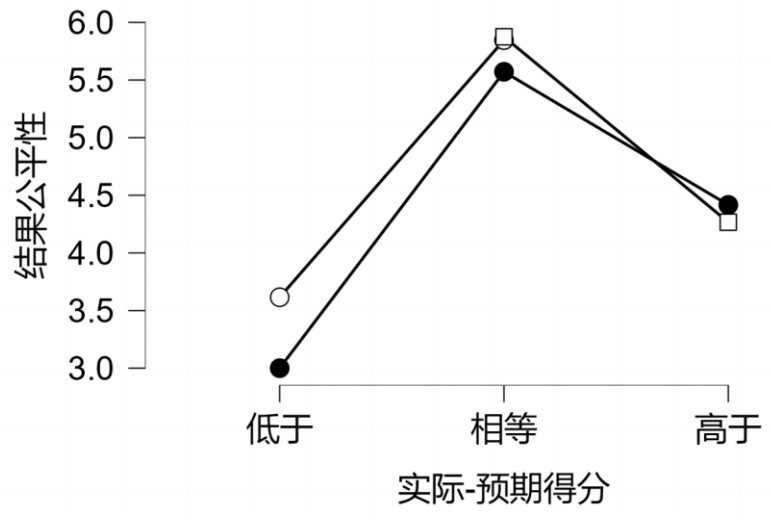


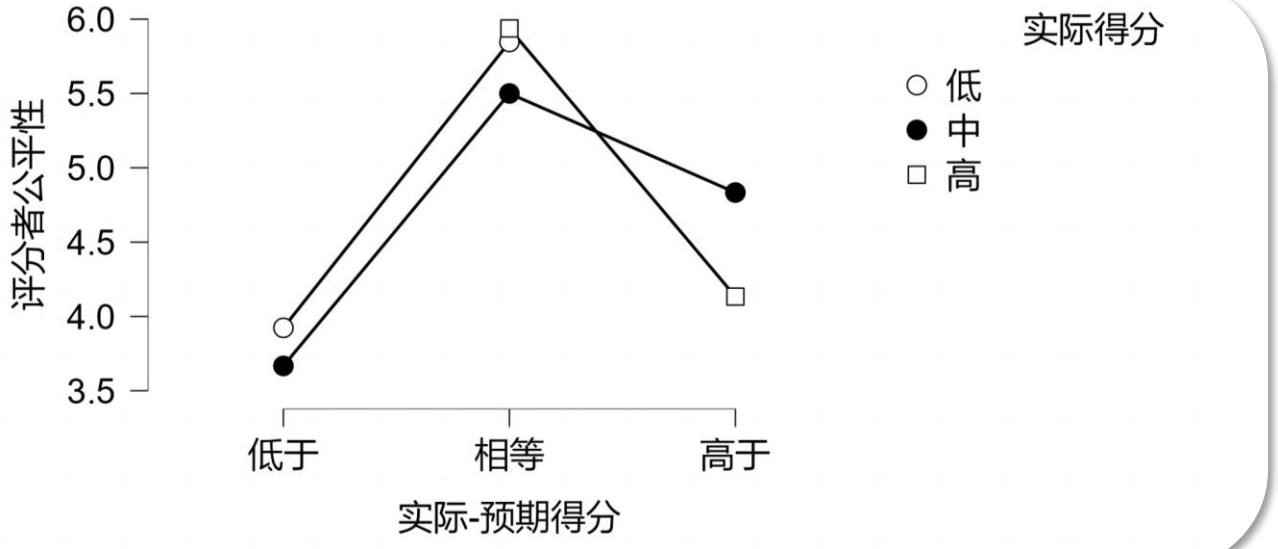
显性感知作为预期得分的函数图

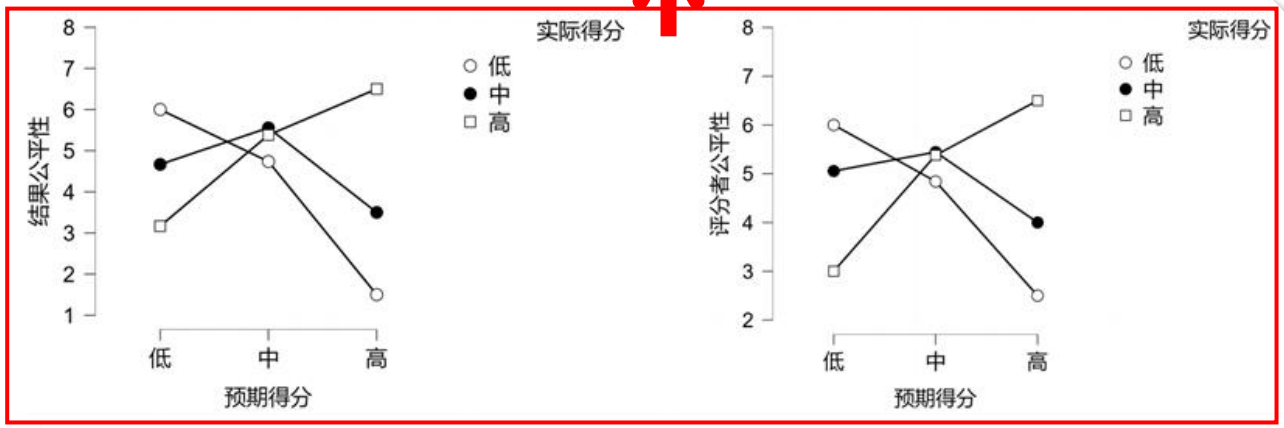
预期得分与期望得分相等时，其公平性评分达到最高

- 无论是结果公平性还是评分者公平性，当预期得分与期望得分相等时，其公平性评分达到最高；
- 无论实际得分是低于还是高于预期得分，公平性评分均下降

研究2 实际得分和期望得分对公平性和满意度的影响——实际评分和期望评分差异分析







显性感知作为预期得分的函数图

预期得分与期望得分相等时，其公平性评分达到最高

- 无论是结果公平性还是评分者公平性，当预期得分与期望得分相等时，其公平性评分达到最高；
- 无论实际得分是低于还是高于预期得分，公平性评分均下降

研究2 实际得分和期望得分对公平性和满意度的影响——

实际评分与期望评分差异分析

进一步精确比较不同实际-预期得分组别的差异

多元方差分析的事后比较

结果公平性和评分者公平性事后比较表

	结果公平性			评分者公平性		
	平均值	标准差	*p*	平均值	标准差	*p*
相等-低于	3.111	0.353	<0.001	1.925	0.336	<0.001
相等-高于	1.357	0.309	<0.001	1.218	0.295	<0.001
高于-低于	1.754	0.410	<0.001	0.7076	0.390	0.073

- 当预期得分与实际得分相符时，公平性评分均**显著大于**不相符（实际低于预期&实际高于预期）的情况

公平性感知取决于预期得分与实际得分两者，当两者相符时，公平性感知最高

- 实际得分高于预期的情况下公平性得分**显著高于**实际得分低于预期的情况

二者不相符时，实际得分高于预期得分时公平性感知较高

研究3 期望评分者和实际评分者对公平性和满意度的影响

实际得分以及实际得分与期望得分的差异显著影响被试的公平性和满意度感知

预期评分者和实际评分者之间的相互作用？

- 增加自变量：
期望评分者+期望-实际评分者一致性

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比率差异以及期望评分者对其的影响。
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。

研究对象

- 通过Credamo平台共收集到有效问卷65份，其中男性被试25名，女性被试40名，年龄24.77±8.59岁，平均作答时间7.69分钟。
- 实验开头与末尾设置自我卷入程度筛查，未通过者将被主试操作拒绝。

完全随机设计

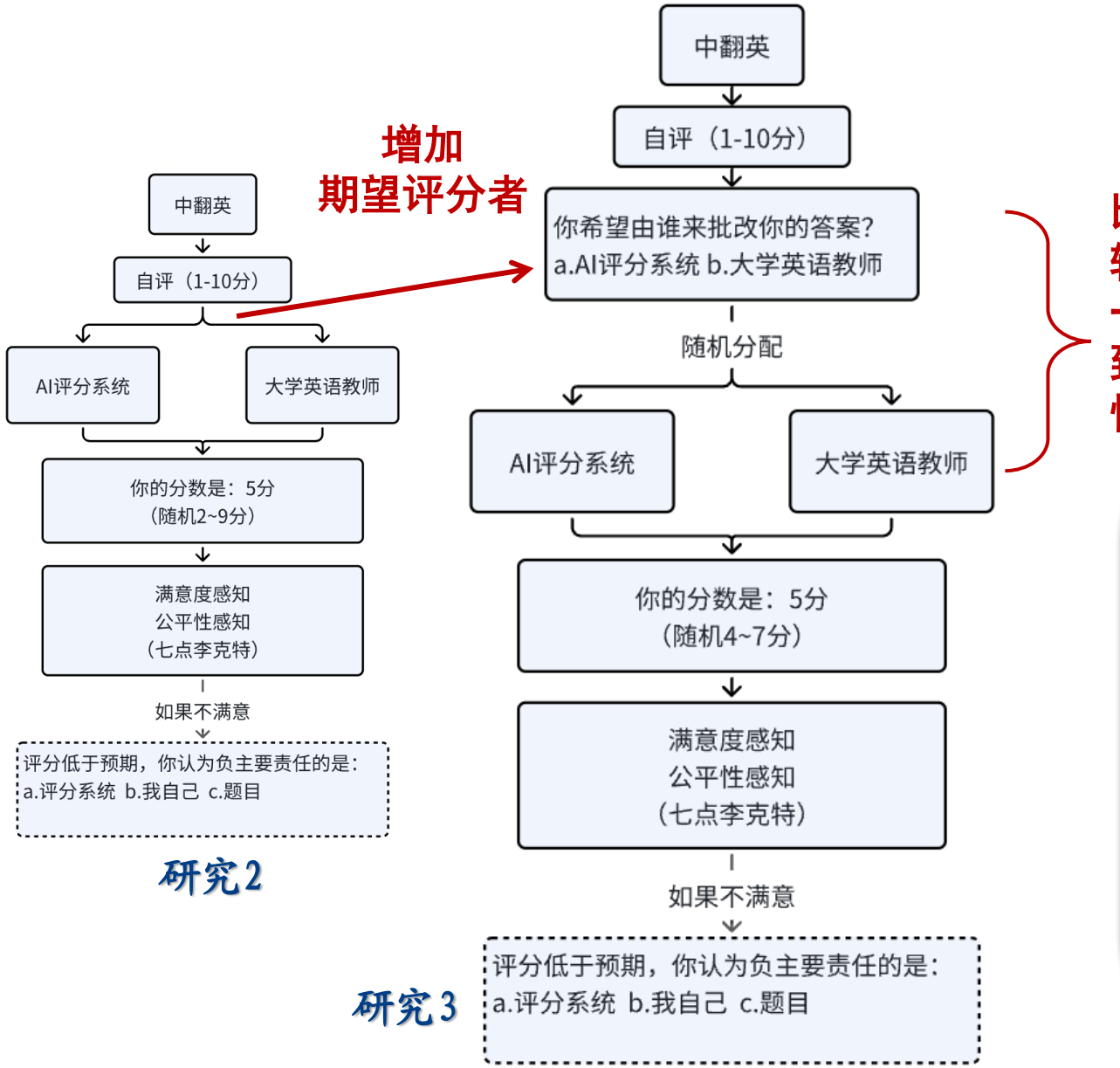
自变量（组间）：

- 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）**

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3 期望评分者和实际评分者对公平性和满意度的影响



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比率差异以及期望评分者对其的影响。

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。

完全随机设计

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）**

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3 期望评分者和实际评分者对公平性和满意度的影响

表 3-1 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）	13	3
中分（4～6分）	20	23
高分（7～10分）	2	4
总计	35	30

期望得分分组

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）

当期望得分较低时，被试倾向于选择AI评分系统；而当期望得分较高时，被试倾向于选择大学英语教师进行评分。

卡方检验

- 被试选择“AI评分系统”和“大学英语教师”进行评分的**频数没有差异**

$\chi^2 = 0.39$，$p = 0.53$

AI评分系统-35人 大学英语教师-30人

卡方检验

- **期望得分会显著影响**期望评分者的选择比例

$\chi^2 = 6.78$，$p = 0.03$

研究3 期望评分者和实际评分者对公平性和满意度的影响

按照期望评分者和实际评分者进行分类

表 3-2 期望评分者和实际评分者对显性感知的影响

期望评分者	显性感知	实际评分者	
		AI评分系统	大学英语教师
AI评分系统	结果满意度	5.33（1.91）	5.59（1.54）
	结果公平性	5.39（1.94）	5.59（1.12）
	评分者满意度	5.22（2.05）	5.77（1.15）
	评分者公平性	5.33（1.85）	5.41（1.23）
大学英语教师	结果满意度	5.19（1.72）	5.43（1.60）
	结果公平性	5.25（1.73）	5.64（1.55）
	评分者满意度	5.13（1.86）	5.71（1.49）
	评分者公平性	5.13（1.50）	5.86（1.41）

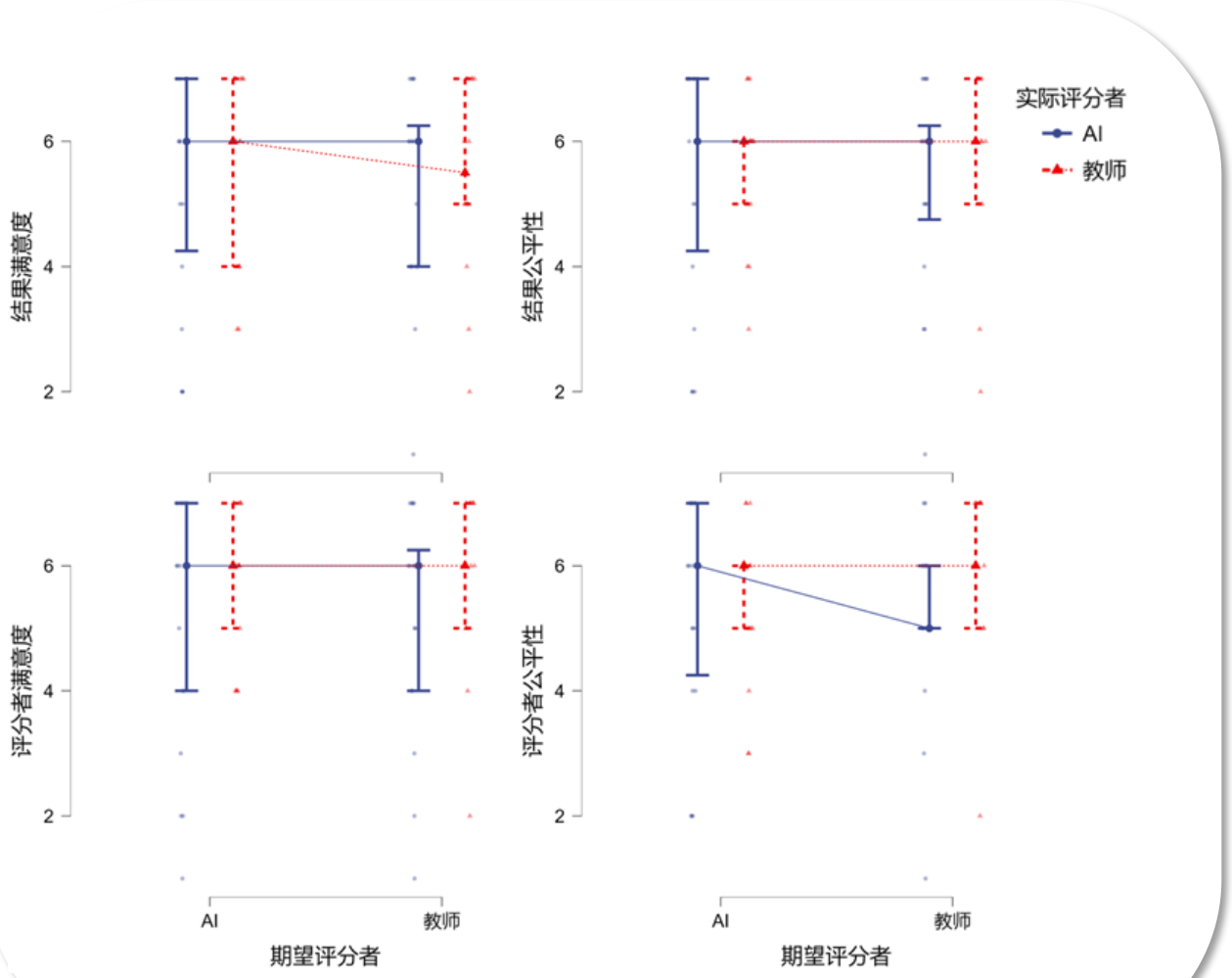
Mann-Whitney检验结果显示：

虽然AI评分时，公平性和满意度感知都会低于教师评分，但是AI评分和教师评分之间，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**。

表 3-3 实际评分者对显性感知的影响

显性感知	Mann-Whitney *U*	*Z*	*p*
结果满意度	491.00	-0.49	0.63
结果公平性	515.50	-0.16	0.88
评分者满意度	463.00	-0.87	0.39
评分者公平性	469.00	-0.79	0.43

研究3 期望评分者和实际评分者对公平性和满意度的影响



评分者对公平性和满意度的影响

多变量方差分析结果显示：

- 期望评分者对显性感知**没有**显著影响
  $F(4, 58) = 0.38$，$p = 0.82$，$\eta_p^2 = 0.03$
- 实际评分者对显性感知**没有**显著影响
  $F(4, 58) = 1.04$，$p = 0.40$，$\eta_p^2 = 0.07$
- 实际评分者和显性感知交互作用**不显著**
  $F(4, 58) = 0.80$，$p = 0.53$，$\eta_p^2 = 0.05$

单因素方差分析：

- 期望评分者和实际得分者**不会**对公平性或满意度感知产生显著的影响
- 期望评分者和实际得分者之间**没有**交互作用

Mann-Whitney检验：

- 期望-实际评分者一致性**不会**影响被试的显性感知

研究3 期望评分者和实际评分者对公平性和满意度的影响

尽管被试在选择期望评分者时存在一定的偏好，尤其是低期望得分者更倾向于选择AI评分系统，但期望评分者和实际评分者均未显著影响被试对评分结果的公平性和满意度感知。

期望评分者与实际评分者的一致性也未对显性感知产生显著影响。

与研究1一致

当实际评分为4～7分时，评分者的类型并非影响公平性和满意度感知的主要因素

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

当实际得分为4～7时，期望评分者和实际评分者的一致性并不会影响被试对公平性和满意度的感知

不同实际得分下，期望与实际评分者一致性可能对显性感知产生不一样的影响

→

- 增加自变量：实际得分高中低

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者对其的影响。
② 探究不同实际得分下，实际得分期望评分者和实际评分者对公平性和满意度感知的影响。

研究对象

- 通过Credamo平台共收集到有效问卷106份，其中男性被试38名，女性被试68名，年龄24.09 ± 7.75岁，平均作答时间8.02分钟。
- 实验开头与末尾设置自我卷入程度筛查，未通过者将被主试操作拒绝。

完全随机设计

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（高实际得分/中实际得分/低实际得分）

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4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

表 4-1 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）	17	9
高分（8～10分）	3	8
总计	20	17

期望得分分组

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）

当期望得分较低时，被试倾向于选择AI评分系统，而当期望得分较高时，被试倾向于选择大学英语教师进行评分

选取高期望评分（8～10分）和低期望评分（1～3分）的数据进行2×2交叉表卡方检验

卡方检验

- 期望得分会**显著影响**期望评分者的选择比例

$\chi^2 = 4.52$，$p = 0.03$

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

多元方差分析（MANOVA）

2（期望评分者：AI，英语教师）
×2（实际评分者： AI，英语教师）
×4（显性感知：结果满意度，结果公平性，评分者满意度，评分者公平性）

多变量方差分析结果显示：

- 期望评分者主效应不显著
  $F(4, 99) = 1.09$，$p = 0.36$，$\eta_p^2 = 0.04$；
- 实际评分者主效应不显著
  $F(4, 99) = 1.51$，$p = 0.21$，$\eta_p^2 = 0.06$
- 实际评分者和期望评分者交互作用不显著
  $F(4, 99) = 0.76$，$p = 0.56$，$\eta_p^2 = 0.03$。

探究实际得分与期望-实际评分者一致性的交互作用

多元方差分析（MANOVA）

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

多变量方差分析结果显示：

- 期望评分者、实际评分者二阶交互作用显著
  $F(4, 91) = 4.98$，$p = 0.001$，$\eta_p^2 = 0.180$；
- 实际得分、期望评分者、实际评分者三阶交互作用显著
  $F(8, 182) = 2.45$，$p = 0.015$，$\eta_p^2 = 0.097$。

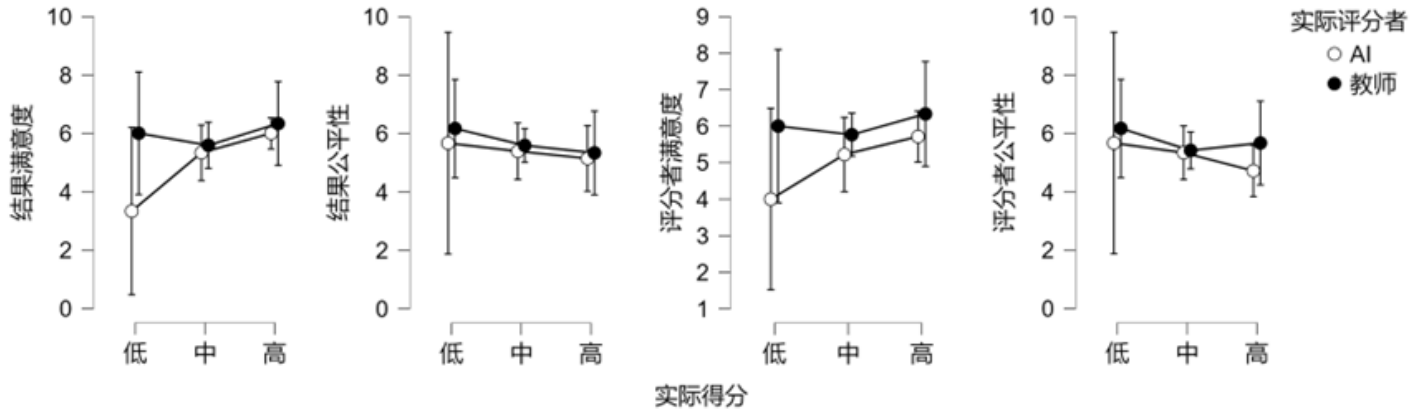
当控制实际得分后，期望-实际评分者一致性会显著影响显性感知，且不同实际得分情况下，期望-实际评分者一致性对显性感知的影响不同

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

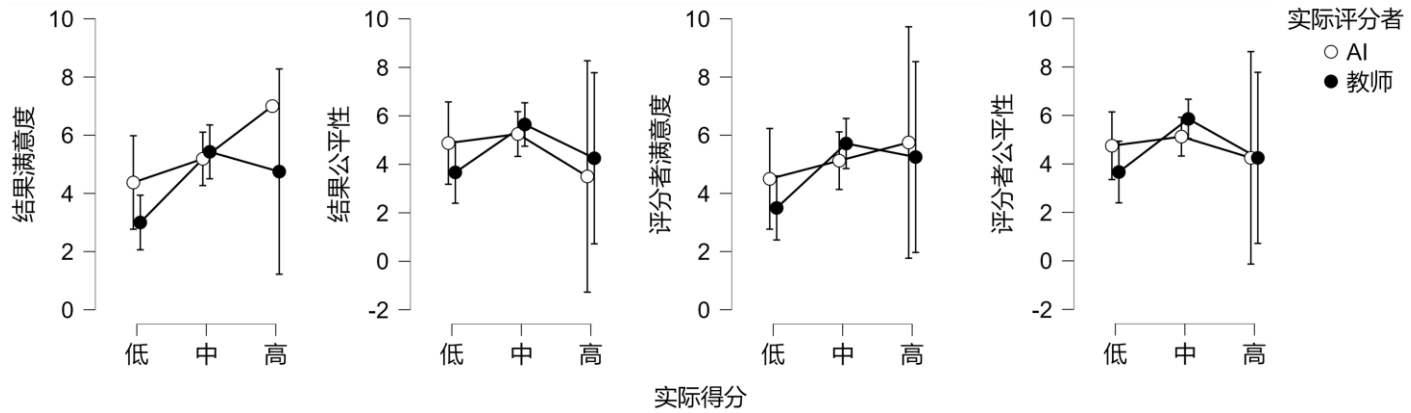
探究期望-实际评分者一致性对显性感知的影响

期望评分者

AI评分系统

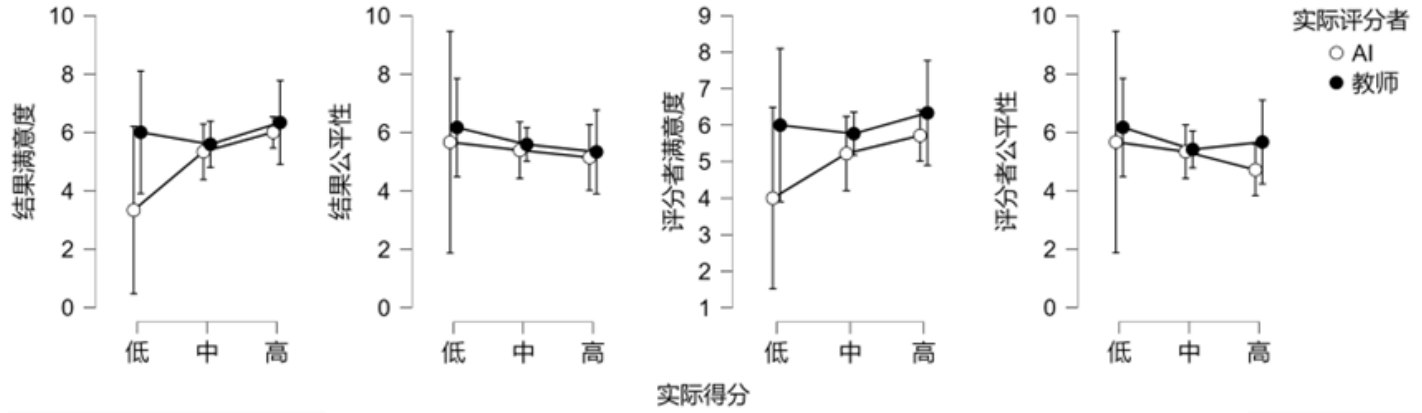


大学英语教师



研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

AI评分系统



多元方差分析（MANOVA）

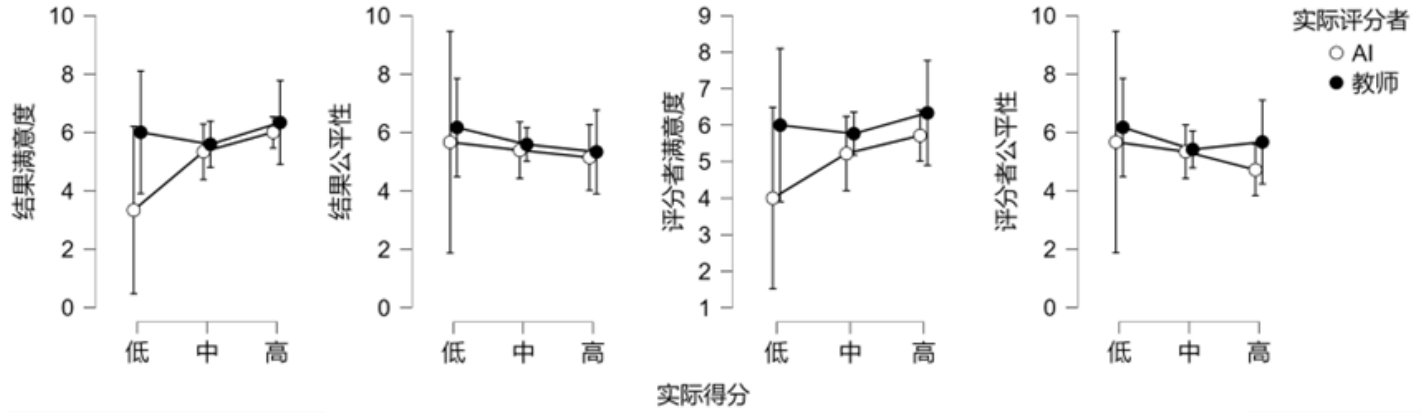
2（实际评分者：AI，英语教师）
×3（实际得分：低，中，高）
×4（显性感知：结果满意度，结果公平性，评分者满意度，评分者公平性）

多变量方差分析结果显示：

- 实际评分者主效应**边缘显著**
  $F(4, 45) = 2.15$，$p = 0.090$，$\eta_p^2 = 0.160$
- 实际评分者和实际得分交互作用**边缘显著**
  $F(8, 90) = 1.80$，$p = 0.087$，$\eta_p^2 = 0.138$

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

AI评分系统



多变量方差分析结果显示：

- 实际评分者主效应边缘显著
  $F(4, 45) = 2.15$，$p = 0.090$，$\eta_p^2 = 0.160$
- 实际评分者和实际得分交互作用边缘显著
  $F(8, 90) = 1.80$，$p = 0.087$，$\eta_p^2 = 0.138$

单因素方差分析：

- 实际评分者对结果满意度和评分者满意度的影响**边缘显著**（$p = 0.060$）

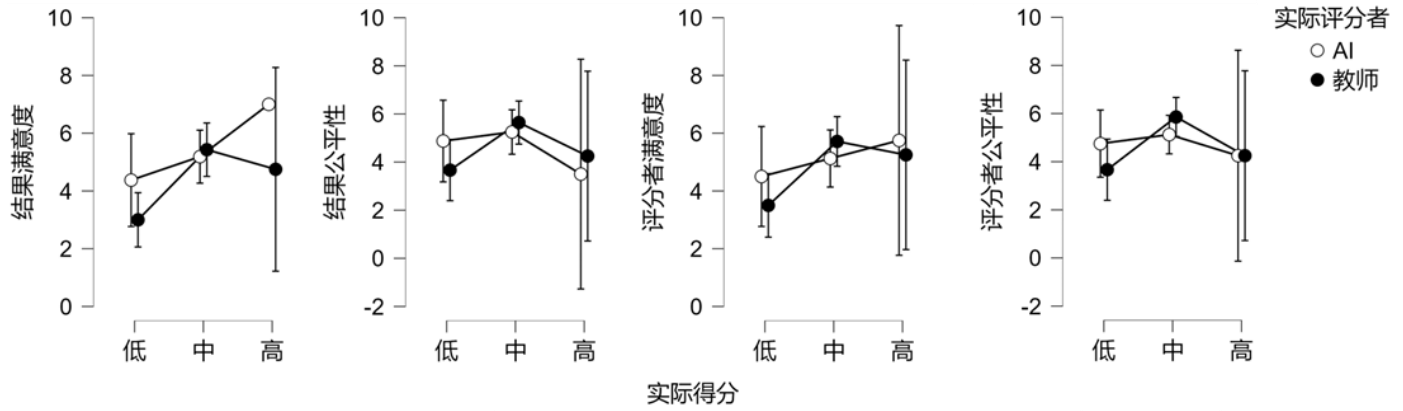
事后检验（Tukey矫正）：

- 当实际得分为低分时，教师评分会比AI评分带来更高的结果满意度（$t = 2.333$，$p = 0.201$，$d = 1.650$）和更高的评分者满意度（$t = 1.802$，$p = 0.474$，$d = 1.274$）

当期望评分者为AI评分系统且实际得分较低时，如果实际评分者也为AI评分系统（期望-实际一致），则对结果满意度和对评分者满意度的感知会低于实际评分者为大学英语教师（期望-实际不一致）

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

大学英语教师



多元方差分析（MANOVA）

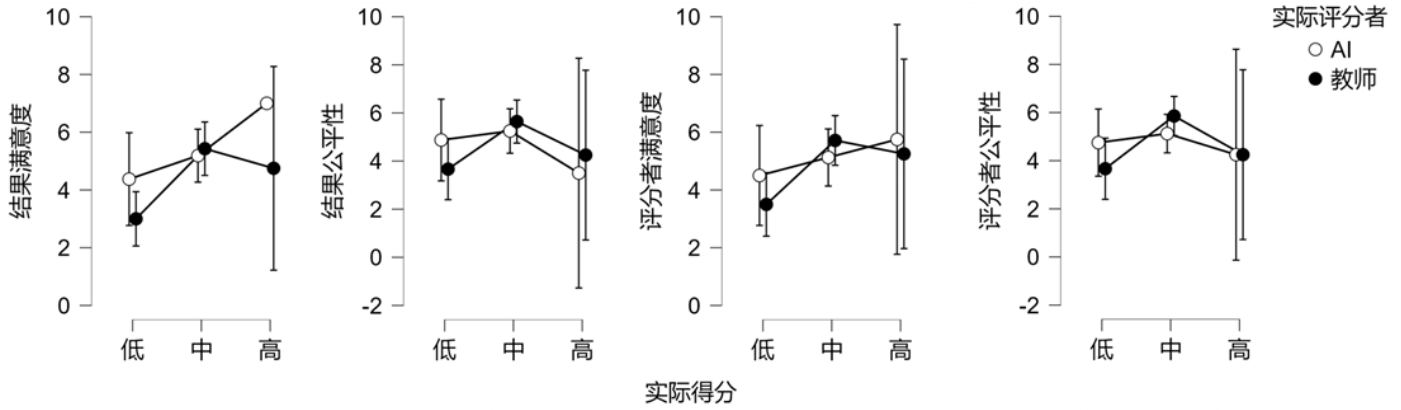
2（实际评分者：AI，英语教师）
×3（实际得分：低，中，高）
×4（显性感知：结果满意度，结果公平性，
评分者满意度，评分者公平性）

多变量方差分析结果显示：

- 实际评分者主效应**显著**
  $F(4, 43) = 3.57$，$p = 0.013$，$\eta_p^2 = 0.249$
- 实际评分者和实际得分交互作用**显著**
  $F(8, 86) = 4.63$，$p = 0.028$，$\eta_p^2 = 0.301$。

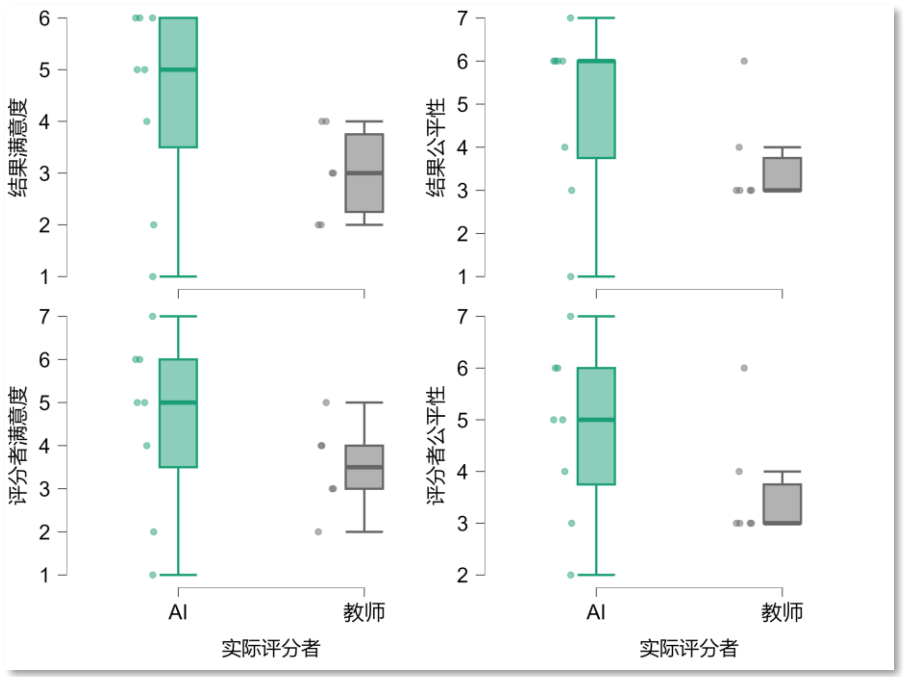
研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

大学英语教师



多变量方差分析结果显示：

- 实际评分者主效应**显著**
  $F(4, 43) = 3.57$，$p = 0.013$，$\eta_p^2 = 0.249$
- 实际评分者和实际得分交互作用**显著**
  $F(8, 86) = 4.63$，$p = 0.028$，$\eta_p^2 = 0.301$。

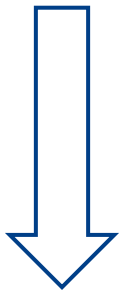


当实际得分为低分时（右图），可以观察到AI评分（期望-实际不一致）会使公平性和满意度的显性感知比教师评分（期望-实际一致）更高

研究4 不同实际得分下，期望-实际评分者一致性对公平性和满意度的影响

当控制实际得分后，期望-实际评分者一致性会显著影响显性感知

期望-实际评分一致性的影响在不同实际得分情况下有所不同



当实际得分较低时，期望-实际评分者不一致条件下，被试对满意度和公平性感知的反而会提高

研究5 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

- IAT范式

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5 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 • IAT范式

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 2

人的图片

类人机器人图片

控制词 控制 支配 领导 统领 权威

受控词 受控 服从 顺从 依从 顺服

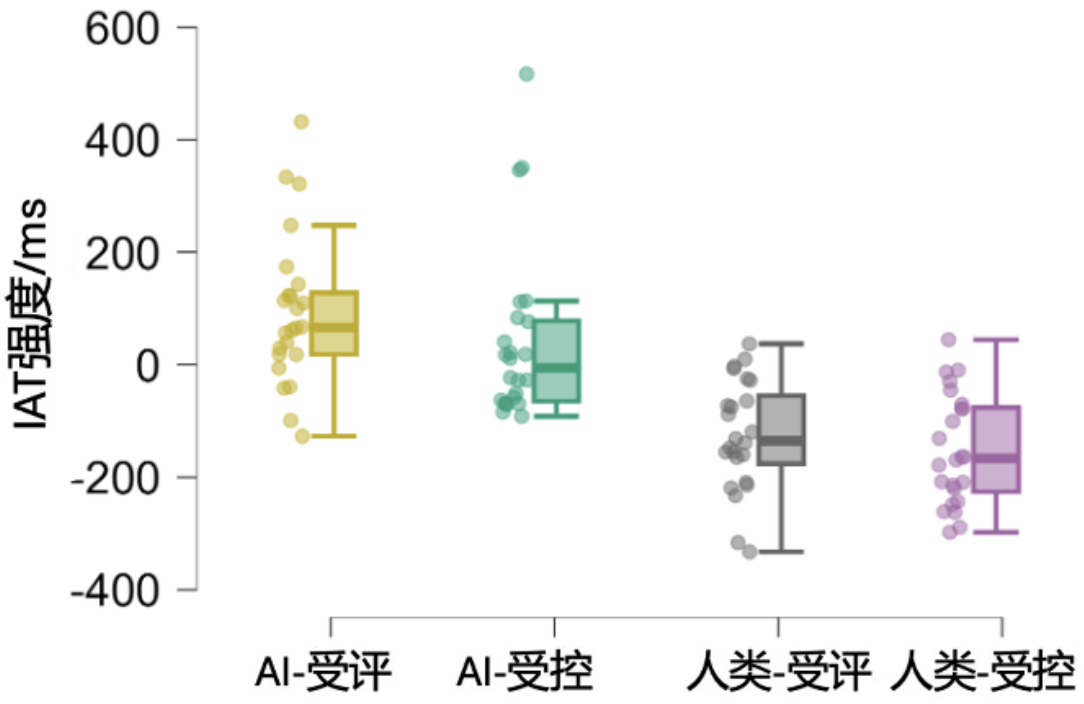
研究5 对AI评分的内隐态度——IAT范式

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5 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

$t = 3.48$，$p = 0.002$，$d = 0.71$

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

$t = 1.35$，$p = 0.19$，$d = 0.27$

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

$t = 6.18$，$p < 0.001$，$d = 1.02$

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

$t = 7.50$，$p < 0.001$，$d = 1.58$

被试对“人-控制”的概念存在隐性偏好

研究5 对AI评分的内隐态度——IAT范式

配对样本t检验

对AI应该被点评的内隐态度强度和对人类应该进行点评的内隐态度强度进行配对样本t检验

- 被试将人-点评联结在一起的内隐强度和将AI-受评联结在一起的内隐强度**无显著差异**

  $t = 0.98$，$p = 0.34$，$d = 0.25$

配对样本t检验

对AI应该被控制的内隐态度强度和对人类应该控制的内隐态度强度进行配对样本t检验

- 人类专家IAT强度**显著高于**AI,说明被试对AI的内隐态度更强，即“AI-受控”概念比“人-控制”隐性偏好更强

  $t = 2.91$，$p = 0.008$，$d = 0.83$

被试间的内隐态度中观察到了一致性，但在被试内的相关性并未显著

皮尔逊相关分析

点评组和控制组的AI内隐态度：*Pearson's r* = 0.21，$p = 0.30$

点评组和控制组的人类内隐态度： *Pearson's r* = -0.12，$p = 0.55$

可能是由于单个被试在单个客体上的试次数较少，导致结果的稳定性不足

研究结果

评分主体

*结论1：*虽然AI评分时被试的公平性和满意度感知都会**低于**教师评分，但是AI评分和教师评分之间，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**

期望分数与实际分数

*结论2a：*评分者并不会显著影响被试的满意度感知，**实际得分是满意度的决定性因素**，并且实际得分越高，满意度越高

*结论2b：*公平性感知取决于**预期得分和实际得分两者**，当两者相符时公平性感知最高

*结论2c：*公平性感知受预期与实际差距的影响是**不对称**的，低于预期的分数被认为比高于预期的分数**更加不公平**

期望评分者和实际评分者

*结论3a：*当实际评分处于中档时，评分者的类型**并非**影响公平性和满意度感知的主要因素

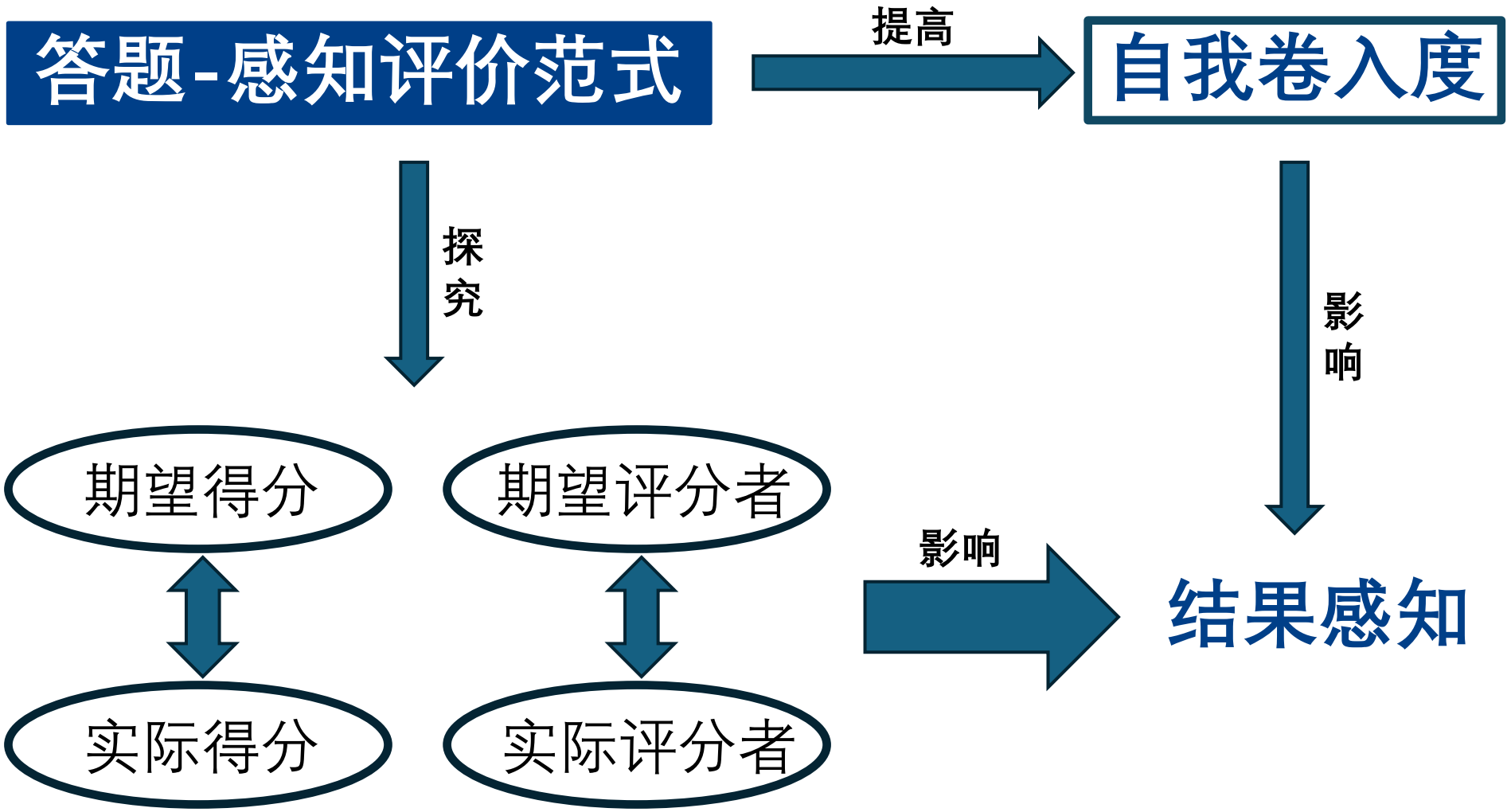
*结论3b：*当实际得分较低时，期望-实际评分者不一致条件下，被试对满意度和公平性感知的**反而会提高**

内隐态度

*结论4a：*被试对“人类-点评”“AI-受评”以及“人类-控制” 概念存在**隐性偏好**

*结论4b：*在点评和控制两种语境下,被试间的内隐态度中观察到了**一致性**，但在被试内的相关性**并未显著**

讨论



讨论

AI与人的一致性

在本实验中，无论是对AI评分的满意度还是公平性感知，都与人类评分者无显著差异。这可能是由于大学生和就业青年对AI的使用频率较高，且对其具有较高的熟悉度和信任度。此外，实验中将AI与教师的评分水平设定为一致，这也可能导致了结果的显著性差异不明显。

预期落差

当预期得分和期望得分一致时，公平性感知得分表现为最高，当二者不一致时，公平性感知则会显著降低。与先前满意度和公平性相关领域的研究基本保持一致，验证了预期落差在AI评分中的影响效果。

- 值得注意的是，在现实情境中，个体通常只要获得较高的分数就会体验到较高的满意度。然而，在实验条件下，部分被试由于持有“测试AI性能”的心态，更倾向于以客观的视角看待评分者。

讨论

能力决定态度

本实验中，不同群体对AI评分的认同情况存在差异，高评分的被试可能对传统教师的信任程度更高，而低评分的被试大多因为作答质量有限而对“真人”评价感到有压力。以下是被试的部分典型观点：

- “AI没有能力准确评估我的答案”
- “我觉得我糟糕的答案可能会让老师抓狂”
- “想测试AI评估极端答案的能力”

情境化感知

当实际得分较低时，期望与实际评分者一致会导致更低的满意度，这一结果在期望评价者为AI时表现的尤为显著。由此可见，个体对于AI和人的评分并没有绝对而普遍的偏好情况，而是更多地受到情境因素的影响。

潜意识中人与AI的关系

我们发现，在个体的潜意识中，普遍存在将AI视为下位关系的倾向，即AI更多地被赋予“协助”和“辅佐”的职能，而人类则被视为评分者和控制者的角色。这在一定程度上解释了本研究结果与以往研究之间的差异。

讨论 未来研究方向

- AI更偏向于下位关系的内隐态度未能表现为外显态度
- 不同实际得分与其他不同情境下，被试表现出对AI的不同感知情况

→ 进一步的研究深层机制

- 选取的被试大多为**大学生**，年龄跨度和离散程度有限
- 答题情境下，所得结论与得分等实际情境密切相关，缺乏普适性，难以推广试表现出对AI的不同感知情况

→ 扩大被试年龄范围

比较不同年龄人群对于AI评价系统的态度以及关注点

- 开展进一步研究，对本实验疑虑进行解释
- 扩大被试年龄范围，为人工智能引入各年龄群体多情境评估提供支持

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

讨论 研究意义与应用价值

研究意义

- 研究结果补充了**AI评分感知领域**的理论框架
- 挖掘了**期望与实际**、**得分和评价者**等变量的影响及其交互作用

为人工智能评分系统在实际应用中的引入提供了重要支持

评估指导

- 本研究强调了在设计基于人工智能的评估系统时，**整合认知和情境因素**的必要性
- 要更好地理解和提高**学生对人工智能在教育评估**中应用的接受度，不仅需要关注评估的准确性，还应考虑期望与认知偏差如何影响学生的感知

有助于促进更公平且更能令学生满意的教育体验

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）