第二届全国大学生心理与行为在线实验精英赛

（分区赛/总决赛• 研究报告）

研究题目	人工智能与教师评分对结果公平性和满意度感知的影响
团队名称	湖底水草队
参赛赛道	A 自选赛道 ☑　　B 揭榜赛道 ☐
问卷分享链接（Credamo 见数平台）	https://www.credamo.com/u/mNgPbRJBAdz https://www.credamo.com/u/vw3OKWkwwxQ https://www.credamo.com/u/WQ43lq0DJ8E https://www.credamo.com/u/vwRY6ODnxXn https://www.credamo.com/u/mNQRGkRGogO/

人工智能与教师评分对结果公平性和满意度感知的影响

【摘　要】 在教育评价领域，人工智能（Artificial Intelligence，AI）评分系统正逐渐成为一种替代传统人类教师评分的方案。然而，关于 AI 评分与人类教师评分在公平性和满意度感知上的差异尚缺乏深入探讨。本研究旨在比较 AI 评分系统与人类教师评分在教育评价中的公平性和满意度感知差异。通过采用在线实验方法，模拟考试评分场景并使被试高度自我卷入，通过问卷收集被试对评分的公平性和满意度感知，分别分析评分者类型（AI 或教师）以及期望评分者与实际评分者一致性对其的影响。研究发现，实际得分对满意度感知有显著的积极影响，而评分者类型（AI 或教师）并未显著影响满意度和公平性感知；当实际得分较低且期望与实际评分者不一致时，对公平性的感知和满意度的感知有一定程度的提高。此外，本研究还通过内隐联想测验（Implicit association test，IAT）测量了对 AI 评分的内隐态度，发现被试更多地将 AI 视为受控、辅助型角色。基于此，我们还进一步研究了 AI 与教师辅助协作的评分系统，验证了内隐偏好并探索了该协作模式的特征。这些发现为 AI 评分系统在教育评价中的应用提供了实证支持，并指出了在不同评分情境下感知的差异性，对教育评价改革具有重要的理论和实践意义。

【关键词】 AI 评分系统；期望评分者；实际得分；公平性感知；满意度感知

1　引言

2022 年，随着 ChatGPT 的横空出世，AI 评分系统正逐步渗透教育领域，引起教育界的广泛关注。部分教师已开始在课堂作业评估中尝试采用这些 AI 评分系统。预见在不久的将来，AI 评分系统有望在教育评估中扮演越发重要的角色。随着其应用范围的不断扩大，对 AI 评分结果与传统教师评分结果之间感知差异的研究显得尤为关键。这种感知差异不仅关乎评分结果的接受度，也触及到评分公平性和满意度的核心问题。当前，人工智能等新兴技术的应用为教育评价改革开辟了新的路径和解决策略。教育评价历来由教师或教育机构承担，AI 系统的介入是否能提升学生对评价结果的公平性或满意度感知尚不明确。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。然而，该研究并未深入探讨不同评价结果对学生感知的影

响，且在实验中仅要求学生想象特定情境，未能实现学生的深度参与和卷入。这种实验设计的局限性可能影响了研究结果的准确性和可靠性，学生在没有实际体验评分过程的情况下，其对 AI 评分系统的看法可能基于想象而非实际体验。此外，尽管我国正积极推进教育的数字化转型，但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。这一研究空白提示我们，未来的研究需要更真实、更具体地模拟评分环境，以便更准确地捕捉和理解学生对 AI 评分系统的真实感受和反应。

鉴于此，本研究旨在通过模拟一个真实的考试评分场景，深入探讨被评价者对 AI 评分结果与教师评分结果的感知差异，并特别关注评价的公平性和满意度两个维度。通过比较被评价者对两种评分方式的响应，我们旨在揭示 AI 评分系统的潜在优势和局限，以及这些系统在未来教育评价实践中的可行性和接受度。

此研究的意义在于，它不仅能够提供对于 AI 评分系统在教育领域应用前景的深入见解，同时也能够促进对传统评分方法的反思。通过理解被评价者对不同评分方法的感知，教育者和技术开发者可以更好地调整和改进评分系统，以满足教育评价的公平性和准确性要求，进而提升教育质量和效率。

2 文献综述

2.1 公平性感知和满意度感知

2.1.1 公平性感知

公平一直是人类社会追求的重要目标。在分配或决策过程中，公平性不仅仅涉及到最终结果的分配是否合理，还包括决策过程本身是否透明和公正。决策公平感的基本概念涉及个人对结果分配（分配公平）、决策过程（程序公平）、人际待遇质量（人际公平）以及所提供的有关结果和决策过程的信息和解释（信息公平）的看法（Colquitt & Rodell, 2015）。

影响公平性感知的因素很多。当投入与回报之间存在公正的对应关系时，员工会增强其对组织的承诺和满意度（Adams, 1965）。信息的透明度和解释的充分性也是公平性的重要组成部分（Thibaut & Walker, 1975；Bies & Moag, 1986）。组织的氛围，如包容性氛围，也可以调节对公平的感知。在一个高度包容的氛围中，员工可能对决策的公平性有更高的容忍度（Lee, 2018）；当员工感觉到组织关心他们的福祉时，他们可能会体验到更高的程序公平性（Eisenberger, Fasolo &

Davis-LaMastro, 1990）。另外，不同文化背景下的员工对组织公平性的感知可能存在差异。跨文化比较研究表明，文化价值观对员工对组织公平性的看法有显著影响（Zhu, Martens & Aquino, 2012）。总的来说，公平性感知是一个多层面的复杂过程，既受决策结果的影响，也受决策过程和环境特征等的影响。

在人们的日常生活中，程序公平的感知还受到具体情境的影响。Tyler 和 Bies（1990）以及 Lind 和 Tyler（1992）指出，在不同场合下，人们对于程序公平的归因是不同的，有时以决策背后的程序感知程序公平，有时以决策者的行为感知程序公平。在某些情况下，人们可能更注重决策背后的程序是否公正，而在其他情况下，他们可能更关心决策者的行为是否符合公正标准（Cobb, Vest & Hills, 1997）。

教育评估与组织决策在很多方面具有相似性，学生不仅关注评分结果是否公平合理，还关注评分过程是否透明公正。这种相似性表明，组织公平的分类同样适用于教育评分的公平性感知。此外，也有研究通过教育评价来探讨组织公平的适用性，这也为进一步研究教育评分中的公平性感知提供了理论支持（Colquitt, 2001）。

2.1.2 满意度感知

在绩效评估和决策领域，对结果的满意度感知和评估系统的满意度感知常被研究（Helberger, Karppinen & D'Acunto, 2018)。满意度是指个人在满足其需求、愿望或期望后所体验到的积极情感或状态。在组织行为学和管理学领域，满意度通常被定义为个人对某一特定结果、过程或系统的整体评估（Saifullah et al. 2015）。在教育领域，学生的满意度指他们对学习经历、教学质量及评估系统的整体感受（Douglas et al., 2015）。

影响满意度的因素多种多样。决策过程的透明度、信息的完整性、决策者的行为及人际关系质量等因素也会显著影响个体的满意度（Shin & Park, 2019）。在组织环境中，工作条件、薪酬待遇、发展机会及工作与生活的平衡等因素也会影响员工的满意度（AL-Omari, Alomari & Aljawarneh, 2020）。此外，文化背景、个体价值观、社会支持系统等外部因素同样会调节满意度的感知（Gomez et al., 2012）。

公平性感知与满意度之间存在显著的正相关关系。研究表明，当个体认为结果的分配、决策的过程以及所受到的待遇是公平的，他们往往会体验到更高的满意度（Shin & Park, 2019）。公平性感知的不同维度（分配公平、程序公平、人际公平和信息公平）都会对满意度产生影响。例如，当员工认为其工作表现得到了

公平的评价和报酬，他们的工作满意度通常会较高（Palaiologos, Papazekos & Panayotopoulou, 2011）。在教育环境中，当学生感到评分标准透明、公正，且评分过程公平时，他们对课程和教师的满意度也会显著提高。此外，公平性感知不仅直接影响满意度，还通过调节个体的情感和行为反应间接影响满意度。例如，高公平性感知可以缓解个体因不满意结果而产生的负面情绪，并减少他们的离职或退出意愿（Byrne, 2005）。

2.2 AI 决策评分系统

AI 决策系统（artificial intelligence decision making systems, ADMSs）正在逐渐进入人类生产生活相关领域之中。为了确保 ADMSs 能够更好地融入人类社会发挥作用，除了研究 ADMSs 本身运作的算法和机制，研究人们如何感知和评价 ADMSs 也是至关重要的（Narayanan & Devesh, 2024）。相关领域的研究多聚焦于人类对于 AI 决策评分系统的公平性感知以及对于决策评分系统的信任和满意度等维度。

2.2.1 AI 决策对公平性感知的影响

人们对于 AI 决策评分系统的公平性感知不仅受到算法本身的公正性影响，还与信息的透明度、解释的充分性以及个体差异等多种因素有关。

信息的透明度和解释的充分性是构成公平性的关键要素。透明度指用户能够洞察 AI 系统的决策逻辑和过程，而可解释性则指系统能够向用户提供其决策原因的清晰解释（Guidotti et al., 2018; Shulner et al., 2024）。当 AI 系统的决策过程和结果对用户而言是可理解的，用户对系统的公平性感知会相应提高（Arrieta et al., 2020）。透明度高、解释充分的 AI 决策系统能够增强用户的信任感，从而提升他们对系统公平性的感知（Abdollahi & Nasraoui, 2018; Ribeiro et al., 2016）。

AI 决策系统的公平性感知也受到个体差异的影响。研究发现，不同的用户特征，如年龄、教育水平、性别以及个性特质，都会影响他们对 AI 决策公平性的看法（Araujo et al., 2020; Helberger et al., 2020）。例如，年轻用户和受过更高教育的用户可能对 AI 系统的公平性有更高的期望和标准。

AI 决策系统的使用场景也会影响公平性感知。在涉及人力资源招聘等敏感领域，用户对 AI 决策的公平性感知尤为关键（Krishnakumar, 2019）。研究表明，用户对 AI 在这些领域的决策持有复杂的情感反应，既有可能因为算法的高效性而感到满意，也可能因为缺乏人类直觉和主观判断能力而感到不满（Lee et al., 2019）。

研究者们提出了多种方法来提高用户对 AI 决策系统的公平性感知，包括提供个性化的解释、进行系统审计以及引入公平性认证等（Shulner-Tal et al., 2022; Binns et al., 2018）。这些方法旨在通过增强系统的透明度和解释能力，以及通过独立的第三方的认证，来提升用户对 AI 决策的信任和接受度，从而增强用户对于该 AI 决策系统的公平性感知。

2.2.2 AI 决策的信任和满意度

公平性感知在建立用户对 AI 决策的信任和提升满意度方面起着至关重要的作用。研究指出，用户对算法决策的公平性评价直接影响他们对该系统的信任度和满意度。当算法被感知为公平时，用户更倾向于信任并满意其决策（Abdollahi & Nasraoui, 2018; Arrieta et al., 2020）。然而，如果用户认为算法存在偏见或不公平，不仅信任度降低，而且可能导致对整个系统的满意度下降，进而影响系统的广泛接受和使用（Došilović et al., 2018; Rai, 2020）。

算法透明度和结果的可解释性作为影响公平性感知的两大关键要素，其同样在影响用户对于 AI 评分系统的信任和满意程度上起着重要作用。研究表明，当 AI 评分系统能够提供易于理解的解释时，用户更有可能接受系统的决策，这不仅增强了用户的信任感，也提升了他们对系统的整体满意度（Abdollahi & Nasraoui, 2018; Arrieta et al., 2020）。此外，可解释的 AI 系统有助于用户更好地理解决策背后的逻辑，从而在使用过程中提高用户的控制感和参与度，以帮助用户获得更高的满意度。

用户对 AI 决策系统的情绪反应，如愤怒、失望或满意，是影响信任和满意度的另一个重要因素。研究发现，用户的情绪反应与他们对决策过程的公正性和透明度的认知密切相关（Lee et al., 2018）。用户对 AI 决策系统的期望和先前经验会影响他们的情绪反应，当用户感知到 AI 决策缺乏人性化元素，例如无法进行情感共鸣或主观判断时，可能会引发负面情绪，这些情绪反应会削弱用户对 AI 决策的信任和满意度（Shin, 2020）。

用户的个人特征，包括年龄、教育水平、性别以及个性特质等，也是影响他们对 AI 决策感知的重要因素。不同年龄和教育水平的用户对 AI 的信任和满意度有不同的预期和反应（Araujo et al., 2020; Helberger et al., 2020）。例如，年轻用户可能更愿意接受 AI 的决策，而年长用户可能对 AI 的决策持有更多的怀疑态度（Araujo et al., 2020）。此外，个性特质如开放性和尽责性也与用户对 AI 决策的信任度和满意度相关，开放性高的个体可能更愿意尝试 AI 系统提供的解决方案，并对结果持更开放和积极的态度，因此对 AI 决策系统的信任度和满意度可

能会更高（Huo et al., 2022）。

2.3 期望结果和实际结果

当我们处于被评价的场景中时，对于该评价结果以及评价者的感知会受到期望结果和实际结果间差异的影响。在现实情境中，当个体接收到的成果超出预期时，他们往往会体验到积极的情绪反应，并对结果给予较高的满意度评价。相反，当实际得分低于期望时，可能会引发不满和对公平性的质疑。这种情绪和认知的评估过程不仅受到结果本身的影响，还受到个体对结果的期望以及场景真实性的影响。

个体在评估结果的公平性时，会将实际得分与自己的期望得分进行比较。如果实际得分与期望相符，个体倾向于认为结果是公平的；如果存在偏差，无论是超出还是未达到期望，都可能被感知为不公平（Cherry et al., 2003; van den Bos et al., 1997）。因此，公平性的感知更多地受到期望匹配的影响而非结果的实际价值：当期望得到满足时，个体更可能认为结果是公平的。然而，值得注意的是，当个体在自然环境中接收到实际成绩时，他们对公平性的感知可能与在实验室环境中基于假设情境的感知存在差异。在现实课堂环境中，即使成绩超出了期望，学生也可能认为这是公平的，这与实验室环境中的发现不同（Cherry et al., 2003）。

满意度的评估则更多地与个体的价值观相关，而不仅仅是期望是否得到满足。根据 Locke（1976）的价值匹配假设，个体对于与自己价值观相符的结果感到满意，无论这一结果是否符合先前的期望。这意味着，如果个体的价值观为得到更好的结果，那么即使结果超出了个体的期望，，也能带来高度的满意度。在这一点上公平性和满意度感知存在较大差异。

2.4 对 AI 的内隐态度

内隐联想测验（Implicit association test, IAT）用于测量人们潜在的、无意识的态度和信念。IAT 通过评估人们在不同类别之间的反应时间，揭示他们的隐性偏见和刻板印象（Greenwald et al., 1998）。

IAT 的基本原理是人们在对与他们潜在信念一致的配对做出反应时速度更快，而在不一致的配对上速度较慢。例如，如果一个人潜在地将“AI”与“家庭”相关联，而将“人”与“职业”相关联，那么在要求将“AI”与“职业”配对时，他们的反应时间会更长。这里的反应时差距就反应了内隐态度中对于 AI 与职业相连结的偏见（Rezaei, 2011）。

近年来，IAT 在研究人类态度、偏见和认知之间的关系方面取得了显著进展。研究者们利用 IAT 来探究种族、性别、年龄等社会身份因素对个体态度的影响，部分情况下，内隐态度和外显态度还会产生差异（Colledani & Ciani, 2021; Greenwald et al., 1998）。研究发现，这种分离现象也在人对 AI 的态度中被观察到，Wei 等人（2021）指出，虽然人们对类人机器人持有积极的显性态度，但隐性态度往往表现出负面倾向。

由于 IAT 在内隐态度的测量上比主观陈述题更具优势，本研究将采用 IAT 探究人们对 AI 评分的内隐态度，作为外显感知的补充。

3 问题提出及假设

3.1 评分主体

过往研究发现，不同评分主体对公平性感知和满意度感知存在显著影响，Chai 等人（2024）从信息透明度和可解释性等角度入手来探究不同评分主体对公平性感知的影响。由于前人采用“纸笔人研究”的实验方法，被试的情境卷入程度可能较低，故本研究旨在通过在线模拟真实作答情境来探讨不同评分主体对公平性感知和满意度的影响，其中评分主体包括人工智能（AI）评分系统与人类英语教师。

***假设 1：*公平性和满意度感知受评分者类型影响，被试对人类英语教师的公平性和满意度感知可能高于 AI 评分系统。**

3.2 期望得分与实际得分

本研究旨在探讨期望得分与实际得分对公平性感知和满意度感知的影响。根据先前研究，个体对教育评估的公平性感知和满意度感知受期望结果与实际结果差异的显著影响（Cherry et al., 2003; van den Bos et al., 1997）。个体在评估结果的公平性时，会将实际得分与自己的期望得分进行比较，而满意度的评估则与实际结果的价值紧密相关（Locke, 1976）。

根据 Locke（1976）的价值匹配假设，个体对于与自己价值观相符的结果感到满意，无论这一结果是否符合先前的期望。因此我们提出假设 2*a*。

***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。**

根据 Cherry et al. (2003) 和 van den Bos et al. (1997) 的研究，公平性感知更

多地受到期望匹配的影响，因此我们提出假设 2*b*。

***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（实际得分与期望得分的相符程度），当两者差异最小时，公平性感知最高。**

3.3 期望评分者和实际评分者

本研究旨在探讨不同评分者对公平性和满意度感知的影响，其中评分者包括 AI 评分系统与人类英语教师。根据先前研究，公平性和满意度感知受期望结果与实际结果之间差异的显著影响（Cherry et al., 2003; van den Bos et al., 1997）。这种差异不仅仅包括得分差异，还包括了评分者的差异。基于此，本研究引入一个新的变量，即期望评分者与实际评分者之间的差异，以考察其对公平性和满意度感知的潜在影响。

我们认为，当实际评分者与期望评分者相符时，参与者可能对评分结果及评分者表现出更高的宽容度和积极情绪，进而影响其对公平性感知和满意度的评价。据此，我们提出假设 3。

***假设 3：*实际评分者与期望评分者一致组被试的公平性感知和满意度感知将显著高于评分者不一致组。**

3.4 内隐态度

过往研究揭示了一个引人注目的现象：人们对 AI 的内隐态度往往与他们的外显态度不完全一致（Wei et al., 2021）。这种差异可能源于多种因素，包括但不限于对 AI 的熟悉度、个人经验、文化背景以及对 AI 技术潜力和局限的认识。

本研究旨在深入探讨在特定语境下，被试对人类与 AI 之间上下位关系的内隐态度。具体来说，我们将分析被试在不同权力和评价结构中对人类和 AI 角色的无意识偏好。这些语境包括评价（如“人类-点评”与“AI-受评”）和控制（如“人类-控制”与“AI-被控”）的场景。

为了系统地检验这些内隐态度，我们提出了以下假设：

***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：*对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，即“人类-点评”和“人类-控制”的 IAT 效应高度相关，“AI-受评”和“AI 被控”的 IAT 强度高度相关。**

4 实证研究

4.1 研究 1 评分者类型对公平性和满意度感知的影响

4.1.1 研究目的

① 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。

② 探究实际得分高低对不同评价主体的评分公平性和满意度感知的影响。

③ 探究实际得分和期望得分的差异对结果公平性和满意度感知的影响。

4.1.2 被试

使用 Gpower 3.1 计算被试量，选择 *F* 检验，f = 0.25， α = 0.05，1- β = 0.80 计算得到最小被试量为 125。通过 Credamo 平台招募被试，要求非英语专业，右利手，年龄在 18~55 岁，高中以上学历。由于本次收集时间较短，且由于线上问卷回答质量波动较大，目前共收集到有效问卷 122 份，其中男性被试 37 名，女性被试 85 名，年龄 24.12±6.09 岁，平均作答时间 7.85±3.23 分钟。

4.1.3 研究设计

本研究采用 2（评价者主体类型：“AI 评分系统”与“大学英语教师”）×3（实际得分：低分、中分、高分）的完全随机设计。具体而言，低分包括 2 分和 3 分，中分涵盖 4 分至 7 分，而高分则包括 8 分和 9 分。每位参与者在完成翻译任务后，需对其答案进行自我评分，评分范围为 1 至 10 分。研究 1 包含两个因变量，分别是对评分结果和评价主体的公平性与满意度感知。

4.1.4 工具及材料

通过 Credamo 在线实验平台收集数据。答题任务选用一道中译英题目，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，被试在作答过程中，将随机接收其中一道题目进行作答。这一选择旨在构建一个客观与主观兼备、具有一定难度且适用于各专业学生的考试任务场景。在进行实际评分时，评分系统会随机给出 2~9 的整数分数。其中，定义 2~3 分为低分，4~7 分为中分，8~9 分为高分。

4.1.5 研究过程

通过在线问卷收集平台上发布问卷，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”。为防止被试在中译英的时候使用辅助工具，实验开始前被试选择“我承诺在答题过程中独立完成题目”方可进入实验，如果选择“无法保证不使用辅助工具”，则实验结束且无法再次作答。在正式答题前，被试需要在文本框中正确输入“我承诺答题过程中不使用任何辅助工具，如复制粘贴，在线翻译等”才能继续实验。

实验开始，被试首先要求在 5 分钟内完成一道中译英的翻译题。答案提交后，被试需要对自己的答案进行评分，评分范围为 1~10 分。自评完成后，“AI 评分系统”或“大学英语教师”将随机给出 2～9 分的评分。得到评分后，被试需要通过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，而后被试需要通过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。需要注意的是，当评分结果的满意度低于 4 分时，会有一个附加选择题。题目为“评分结果满意度偏低，您认为应负主要责任的是”，选项为“评分者”、“我自己”、“题目本身”三个选项。为了确保被试能够更具象化地感知“大学英语老师”的存在，会给出这个大学老师姓名，性别和年龄（如李*敏老师，女，34 岁）。同时，被试作答时间控制在早上 10 点到下午 6 点。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。

对公平性和满意度进行评价之后，被试填写自己的个人信息，包括性别、年龄、就读情况（在读学生/毕业工作）。最后，有一道事后检验题，“在中译英过程中，您是否使用了任何辅助工具？”，选择“我使用了辅助工具”的数据将会被剔除。研究 1 的实验流程详见图 4-1。

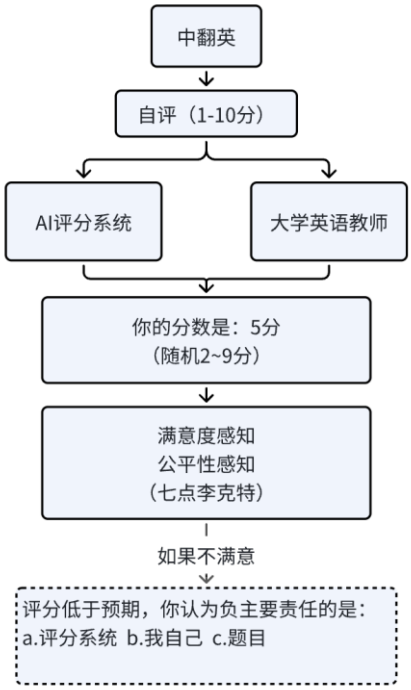

图 4-1　研究 1 实验流程图

4.1.6 分析及结果

（1）评分者影响分析

按照评分者类型进行分类，被试对结果满意度（Score Satisfaction, SS）、结果公平性（Score Fairness, SF）、评分者满意度（Evaluator Satisfaction, ES）和评分者公平性（Evaluator Fairness, EF），这四种显性感知的分数汇总见表 4-1。

表 4-1　不同实际评分者在不同的显性感知上的分数汇总

评分者	*N*	显性感知	平均值	标准差
AI 评分系统	52	结果满意度	4.98	1.57
		结果公平性	4.96	1.52
		评分者满意度	5.12	1.49
		评分者公平性	4.96	1.43
大学英语教师	70	结果满意度	5.49	1.51
		结果公平性	5.19	1.64
		评分者满意度	5.40	1.29
		评分者公平性	5.29	1.50

注：*N*=被试人数

Shapiro-Wilk 正态性检验表明显性感知得分不服从正态分布，故采用 Mann-Whitney 非参数检验。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，$U=2178.50$，$Z=1.901$，$p=0.057$；（2）结果公平性在教师评分和 AI 评分之间无显著差异，$U=2005.00$，$Z=0.980$，$p=0.327$；（3）对评分者的满意度在教师和 AI 之间无显著差异，$U=1992.50$，$Z=0.917$，$p=0.359$；（4）评分者公平性在教师和 AI 之间无显著差异，$U=2072.50$，$Z=1.335$，$p=0.182$。结果详见图 4-2。

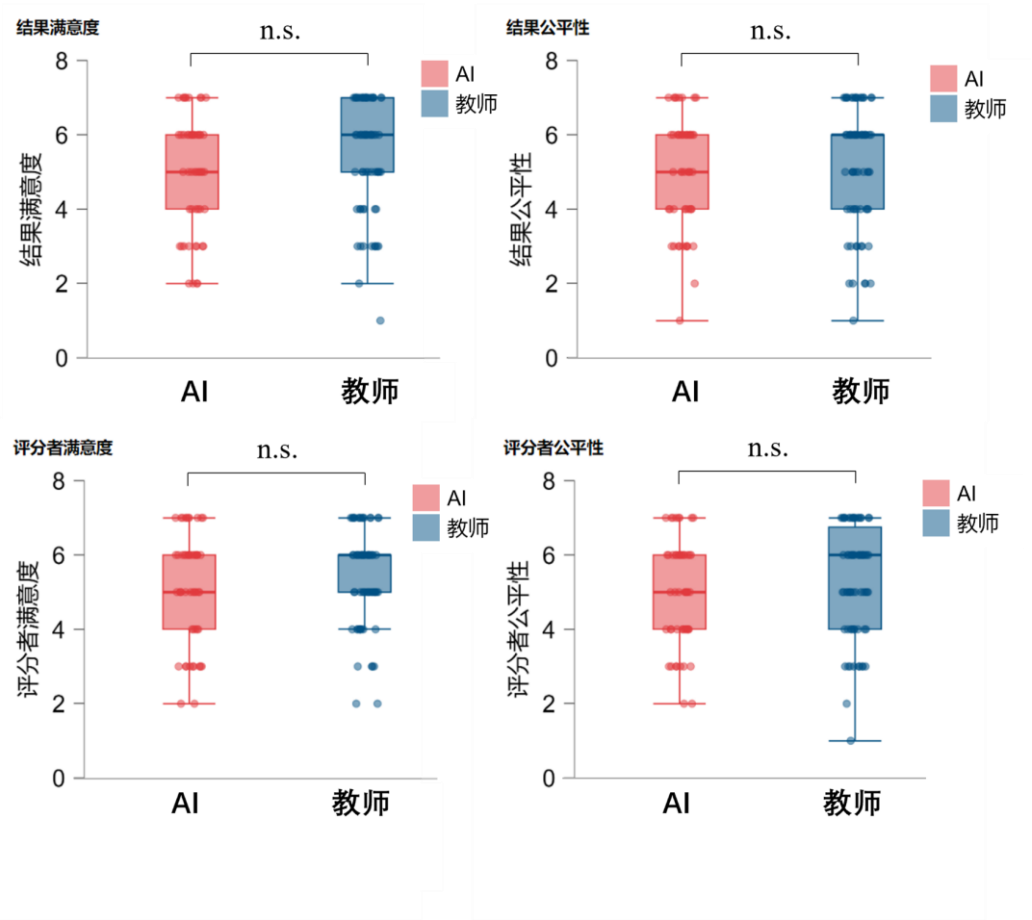

图 4-2　不同评分者在不同显性感知上的 Mann-Whitney 非参数检验结果

注：n.s. = no significance

结果表明，AI 评分和教师评分之间，被试对评分结果和评分者的公平性感知和满意度感知没有显著差异。

（2）实际得分影响分析

按照实际得分高低和评分者类型进行分类，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，这四种显性感知分数汇总见表 4-2。

表 4-2　不同评分者与不同实际得分组在不同显性感知上的得分汇总

评分者	显性感知	实际得分		
		低分组 *M(SD)*	中分组 *M(SD)*	高分组 *M(SD)*
AI 评分系统	结果满意度	4.33（1.88）	4.92（1.32）	5.85（1.28）
	结果公平性	4.73（1.39）	4.96（1.57）	5.23（1.64）
	评分者满意度	4.60（1.50）	5.13（1.45）	5.69（1.44）
	评分者公平性	4.93（1.53）	4.96（1.27）	5.00（1.68）
大学英语教师	结果满意度	5.09（1.58）	5.18（1.59）	6.30（0.98）
	结果公平性	5.18（1.25）	5.26（1.60）	5.05（1.93）
	评分者满意度	5.27（1.27）	5.26（1.39）	5.75（1.07）
	评分者公平性	5.18（1.25）	5.41（1.45）	5.10（1.74）

注：M=平均数；SD=标准差

进行 2（评分者：AI、教师）×3（实际得分高低：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）多变量方差检验，其中实际评分者和实际得分高低是被试间因素，显性感知是被试内因素。结果表明，评分者类型对显性感知影响不显著，$F(4, 113) = 1.126$，$p = 0.348$，$\eta_p^2 = 0.038$，实际得分高低对显性感知影响显著，$F(8, 226) = 3.227$，$p = 0.002$，$\eta_p^2 = 0.103$，评分者类型和实际得分高低交互作用不显著，$F(8, 226) = 0.659$，$p = 0.727$，$\eta_p^2 = 0.023$。具体结果见图 4-3。

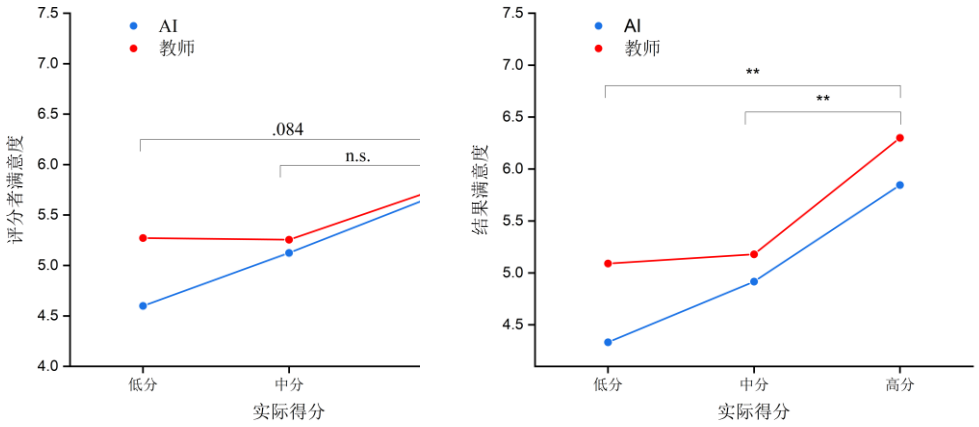

图 4-3　实际得分和实际评分者对显性感知的方差检验结果

后续双因素变量方差分析表明，实际得分高低对结果满意度的影响显著 $F(2,$

119) = 8.45，$p < 0.001$，$\eta_p^2 = 0.124$，对评分者满意度的影响边缘显著 $F(2, 119) = 2.98$，$p = 0.054$，$\eta_p^2 = 0.048$。而实际得分高低对结果公平性和评分者公平性感知没有显著影响。分析变量之间的相关性发现（图 4-4），得分高低和结果满意度呈现显著的正相关（Kendall's tau = 0.336，$p < 0.001$），得分高低和对评分者满意度呈现显著的正相关（Kendall's tau = 0.228，$p = 0.001$ ），得分高低与公平性感知之间无显著相关。

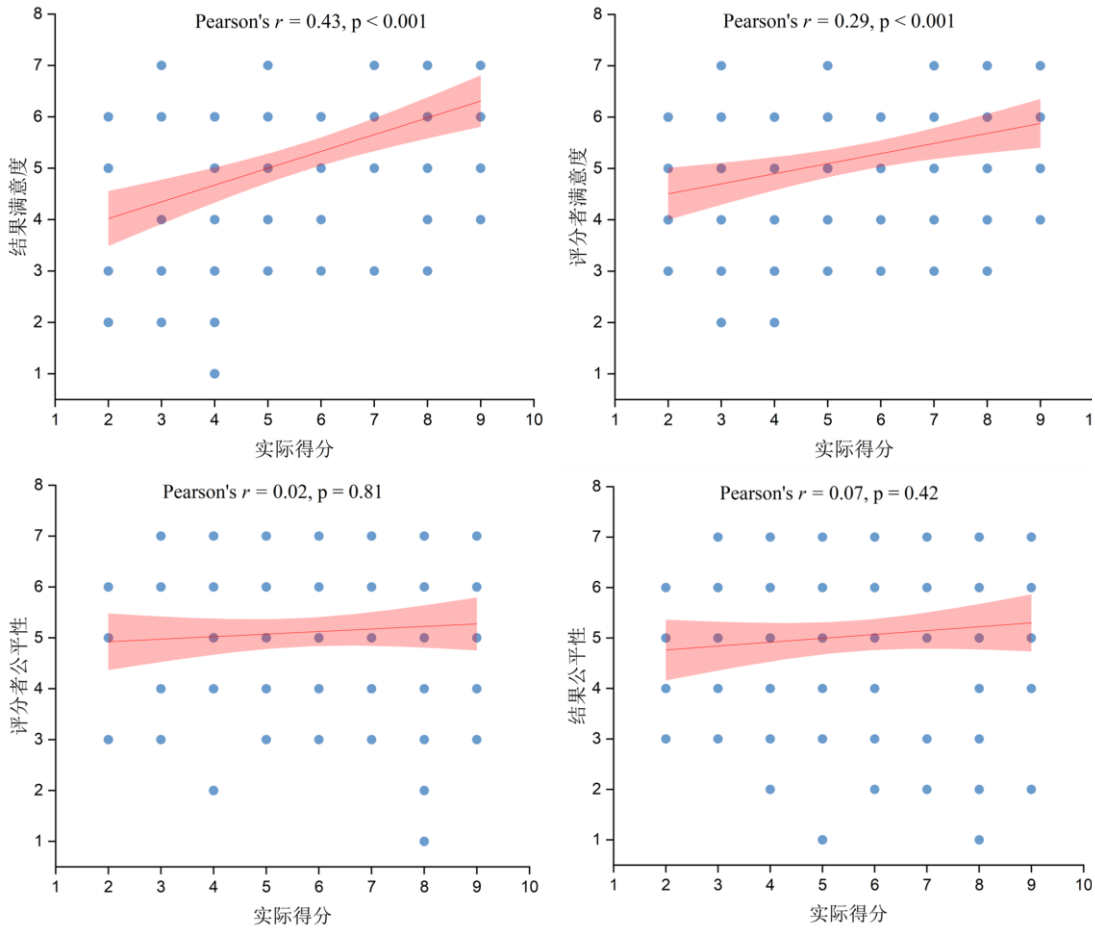

图 4-4　实际得分与满意度和公平度相关性结果

该结果表明，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。

（3）实际得分和期望得分差异分析

进行 2（评分者：AI、教师）×3（期望得分：低、中、高）×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，期望得分和实际得分为被试者间变量，显性感知是被试内变量。结果显示实际得分会显著影响显性感知，$F(8, 206) = 3.053$，$p = 0.003$，$\eta_p^2 = 0.106$; 期望得分对显性感知的影响边缘显著，$F(8, 206) = 1.859$，$p = 0.068$，

η_p^2 = 0.067; 实际评分者不会影响显性感知，$F(4, 103) = 0.423$，$p = 0.792$；实际得分和期望得分交互作用显著，$F(12, 273) = 2.38$，$p = 0.003$，η_p^2 = 0.092。为了进一步解释实际得分与期望得分之间的交互作用是如何影响公平性和满意度感知的，将期望得分作为横轴，实际得分作为分隔线，显性感知作为因变量，绘制图 4-5。

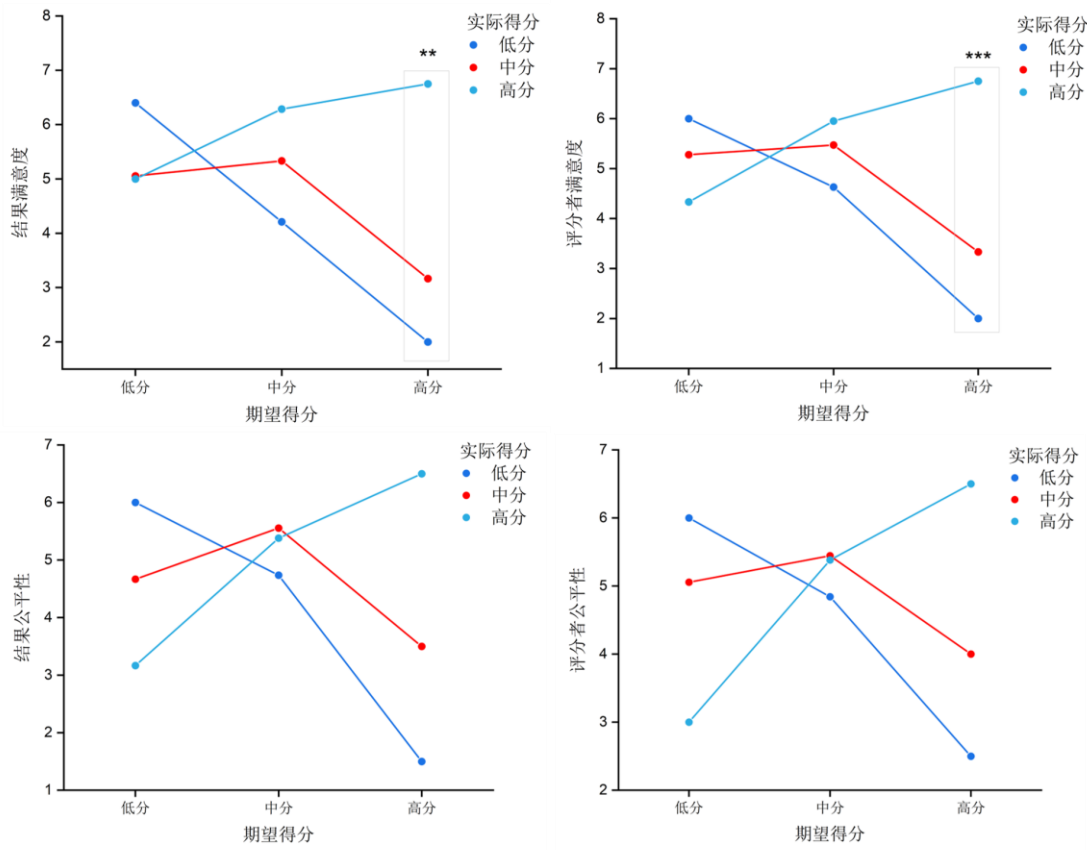

图 4-5　期望得分与实际得分在显性感知的交互作用图

图 4-5 显示了期望得分与实际得分在结果满意度，评分者满意度，结果公平性，评分者公平性上的交互作用情况。

可以发现无论是结果满意度还是评分者满意度，在期望得分为中分组和高分组时，其评分均随着实际得分的增高而增高；其主效应显著，结果满意度 $F(2, 106) = 6.89$，$p = 0.002$，η_p^2 = 0.115，评分者满意度 $F(2, 106) = 3.682$，$p = 0.028$，η_p^2 = 0.065。然而，满意度与期望得分之间并不存在相关。基于此，假设 2*a* 得到了有力的支持。

从图 4-5 中可以初步发现无论是结果公平性还是评分者公平性，当实际得分与期望得分相等时，其公平性评分达到最高，而无论实际得分是低于还是高于期望得分，公平性评分均下降。为了使这一特征更为明显，以实际-期望得分为横坐标，公平性评分为纵坐标绘图，结果见图 4-6。

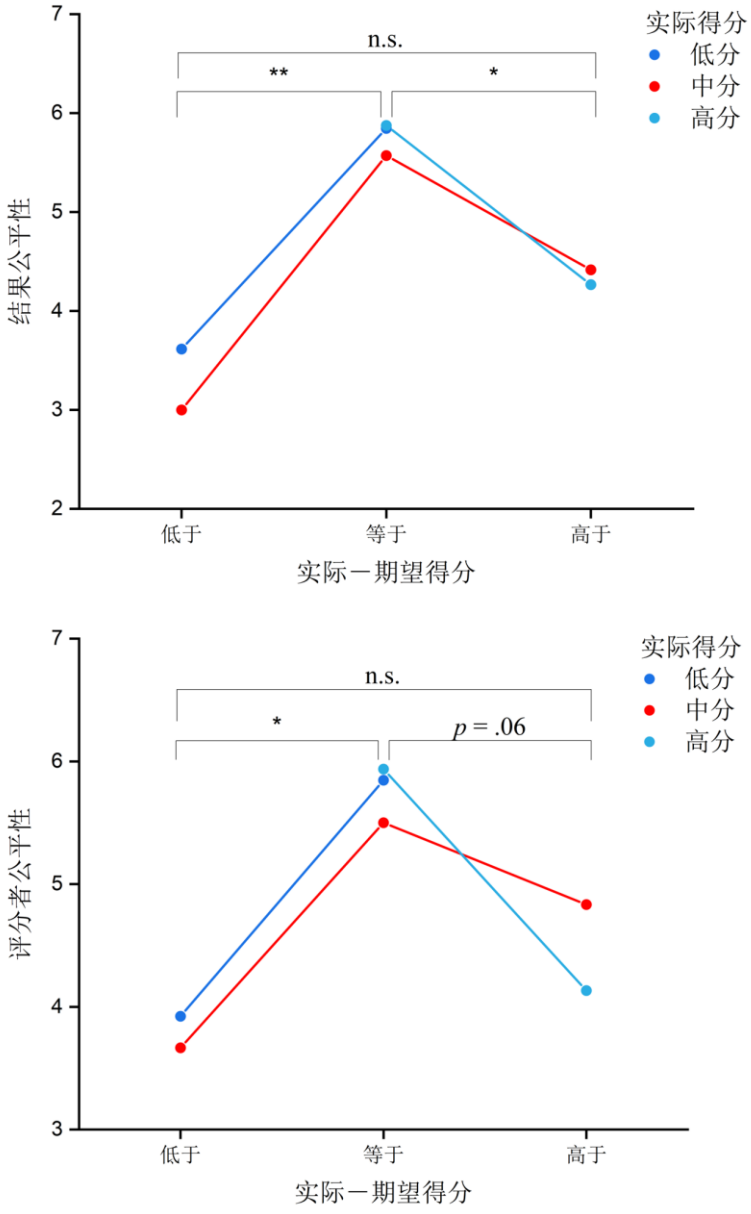

图 4-6　实际-期望得分与实际得分对公平性感知的影响

注：实际-期望得分：相等=期望得分与实际得分相符组；低于=实际得分低于期望得分组；高于=实际得分高于期望得分组

从图 4-6 可以清晰发现无论是结果公平性还是评分者公平性，当实际得分与期望得分相等时，其公平性评分达到最高，而无论实际得分是低于还是高于期望得分，公平性评分均下降。为了进一步精确比较不同实际-期望得分组别的差异，进行多元方差分析事后比较，结果见表 4-3。

表 4-3　结果公平性和评分者公平性的事后比较结果

	结果公平性			评分者公平性		
	平均差值	标准差	*p*	平均差值	标准差	*p*
相等-低于	1.42	0.45	0.002	1.06	0.42	0.012
相等-高于	0.94	0.42	0.027	0.75	0.39	0.061
高于-低于	0.48	0.32	0.134	0.31	0.30	0.291

注：相等=期望得分与实际得分相符组；低于=实际得分低于期望得分组；高于=实际得分高于期望得分组

多元方差分析事后比较的结果表明，当期望得分与实际得分相符时，结果公平性显著大于实际得分与期望得分不相符的情况（实际得分高于期望得分 $p = 0.027$，实际得分低于期望得分 $p = 0.002$）；当期望得分与实际得分相符时，评分者公平性显著大于实际得分低于预期得分的情况（$p = 0.012$），与实际得分高于期望得分的情况差异边缘显著（$p = 0.061$）。假设 2*b* 得到了有力支持。

研究 1 旨在探究不同评分主体对满意度和公平性感知的影响以及实际得分和期望得分对被试公平性和满意度感知的影响。研究发现，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。研究还发现，实际得分的高低显著影响了被试的满意度感知，其中高得分组的满意度得分显著高于低得分组。此外，实际得分与期望得分的差异也显著影响了公平性感知。当实际得分与期望得分相符时，公平性感知最高，而无论实际得分是低于还是高于预期，公平性感知均有所下降。这表明，被试的满意度得分更多地受到实际得分的影响，而公平性感知不仅受到实际得分的影响，还受到期望得分的影响。

4.2 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

4.2.1 研究目的

在研究 1 中，我们发现实际评分者对公平性和满意度并没有显著影响，反而是实际得分以及实际得分与期望得分的差异会显著影响被试的公平性和满意度感知。为排除实际得分的影响，我们将会探究被试在得到实际评分之前，是否会对 AI 评分或者是教师评分有不同的偏好。进一步，预期评分者和实际评分者的一致性是否会影响公平性和满意度感知也十分关键。基于这样的猜想，研究 2 的

目的如下：

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。

4.2.2 被试

使用 Gpower 3.1 计算被试量，选择 *F* 检验，f＝0.25，α＝0.05，1-β＝0.80，计算得到最小被试量为 125。通过 Credamo 平台招募被试，要求非英语专业，右利手，年龄在 18~55 岁，高中以上学历。由于本次收集时间较短，且由于线上问卷回答质量波动较大，目前共收集到有效问卷 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 8.73±2.23 分钟。

4.2.3 研究设计

和研究 1 相比，研究 2 增加了 2 个自变量，分别是期望评分者和期望与实际评分者一致性。研究 2 有 5 个自变量：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），期望与实际评分者一致性（一致、不一致），实际得分（高期望得分、中期望得分、低期望得分）。3 个因变量为：选择 AI 评分和教师评分的比例，以及对公平性和对满意度的感知。

4.2.4 工具及材料

同研究 1。

4.2.5 研究过程

在研究 2 的基础上，被试在完成自评后，可选择自己期望的评分者是“AI 评分系统”还是“大学英语教师”。意愿选择完成后，系统将会随机分配由“AI 评分系统”或“大学英语教师”进行评分（随机 2～9 分）。外显态度感知部分分为公平性和满意度感知。研究 2 的实验流程详见图 4-7。

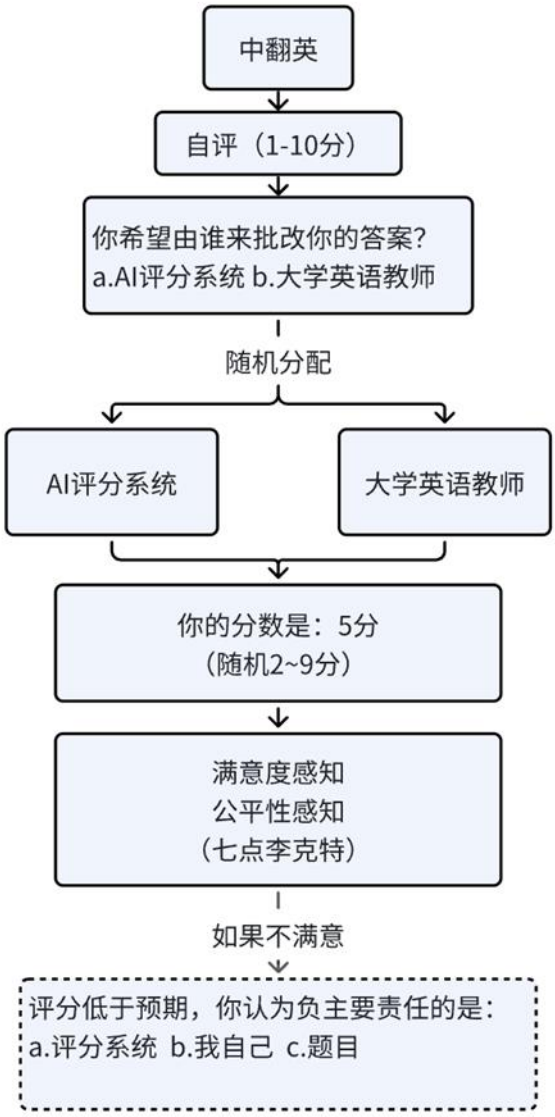

图 4-7　研究 2 实验流程图

4.2.6 分析及结果

卡方检验的结果表明，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，$\chi^2(1) = 0.00$，$p = 1.00$。进一步，将期望得分为三档，小于等于 3 分为低分，4～7 分为中分，大于等于 8 分为高分，研究期望得分高低与期望评分者之间的关系。2×3 交叉表卡方分析结果表明，期望得分并未显著影响期望评分者的选择比例，$\chi^2(2) = 3.48$，$p = 0.18$，Cramer's $V = 0.17$。具体见表 4-4。观察到期望得分低时，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。这里选取高期望评分（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。结果表明：期望得

分对期望评分者选择频数的影响边缘显著，$\chi^2(1) = 3.45$，$p = 0.06$，Cramer's $V = 0.29$。

表 4-4　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

按照期望评分者和实际评分者进行分类，对结果满意度、结果公平性、评分者满意度和评分者公平性的显性感知分数见表 4-5。

表 4-5　不同期望评分者和实际评分者的显性感知分数汇总

期望评分者	显性感知	实际评分者	
		AI 评分系统 *M(SD)*	大学英语教师 *M(SD)*
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）

注：M=平均数；SD=标准差

为探究期望和实际评分者的一致性是否会影响显性感知，进行 2（期望评分者：AI、教师）×2（实际评分者：AI、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。结果表明：期望评分者主效应不显著，$F(4, 113) = 1.45$，$p = 0.22$，$\eta_p^2 = 0.05$；实际评分者主效应不显著，$F(4, 113) = 1.58$，$p = 0.18$，$\eta_p^2 = 0.05$；实际评分者和期望评分者交互作用

不显著，$F(4, 113) = 0.54$，$p = 0.71$，$\eta_p^2 = 0.02$。说明期望评分者和实际评分者一致性并不会影响显性感知。

为探究实际得分与期望与实际评分者一致性的交互作用，进行 3（实际得分：高、中、低）×2（期望与实际评分者一致性：一致、不一致）×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。结果表明，一致性主效应不显著，$F(4, 111) = 2.57$，$p = 0.04$，$\eta_p^2 = 0.08$；实际得分、期望评分者、实际评分者三阶交互作用显著，$F(4, 111) = 6.68$，$p < 0.001$，$\eta_p^2 = 0.19$。说明在不同实际得分情况下，期望与实际评分者的一致性对显性感知的影响不同。

为进一步探究不同实际得分下，期望与实际评分者一致性的影响，我们将分别探究期望评分者为 AI 评分系统（见图 4-8）和大学英语教师（见图 4-9）两种情况下，期望与实际评分者一致性对显性感知的影响。

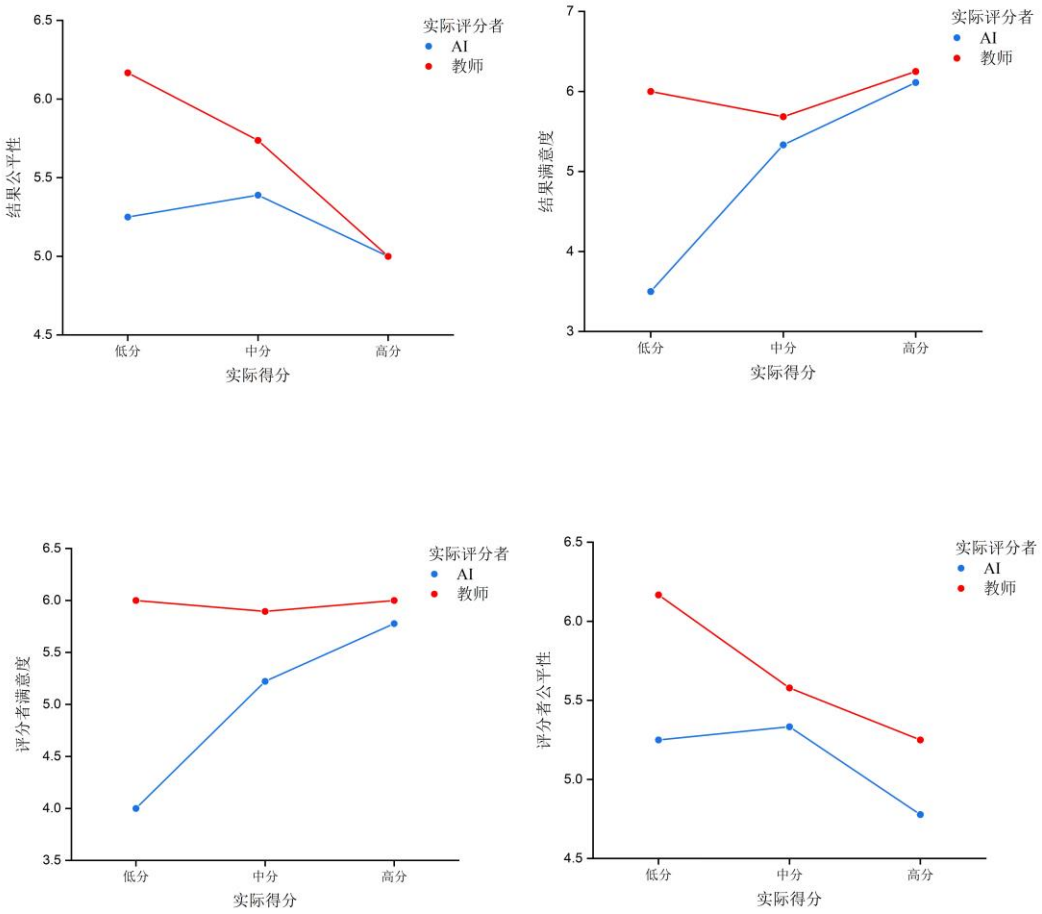

图 4-8　实际评分者对显性感知的影响（期望评分者为 AI）

当期望评分者为 AI 评分系统时，以显性感知为因变量，以期望-实际评分者

一致性、实际得分为自变量，进行双因素方差分析。结果表明：期望-实际评分者一致性对结果满意度和评分者满意度主效应显著（结果满意度 $F(1, 54) = 4.23$，$p = 0.04$，$\eta_p^2 = 0.07$；评分者满意度 $F(1, 54) = 4.10$，$p = 0.05$，$\eta_p^2 = 0.07$）；实际得分和期望-实际评分者一致性交互作用不显著。

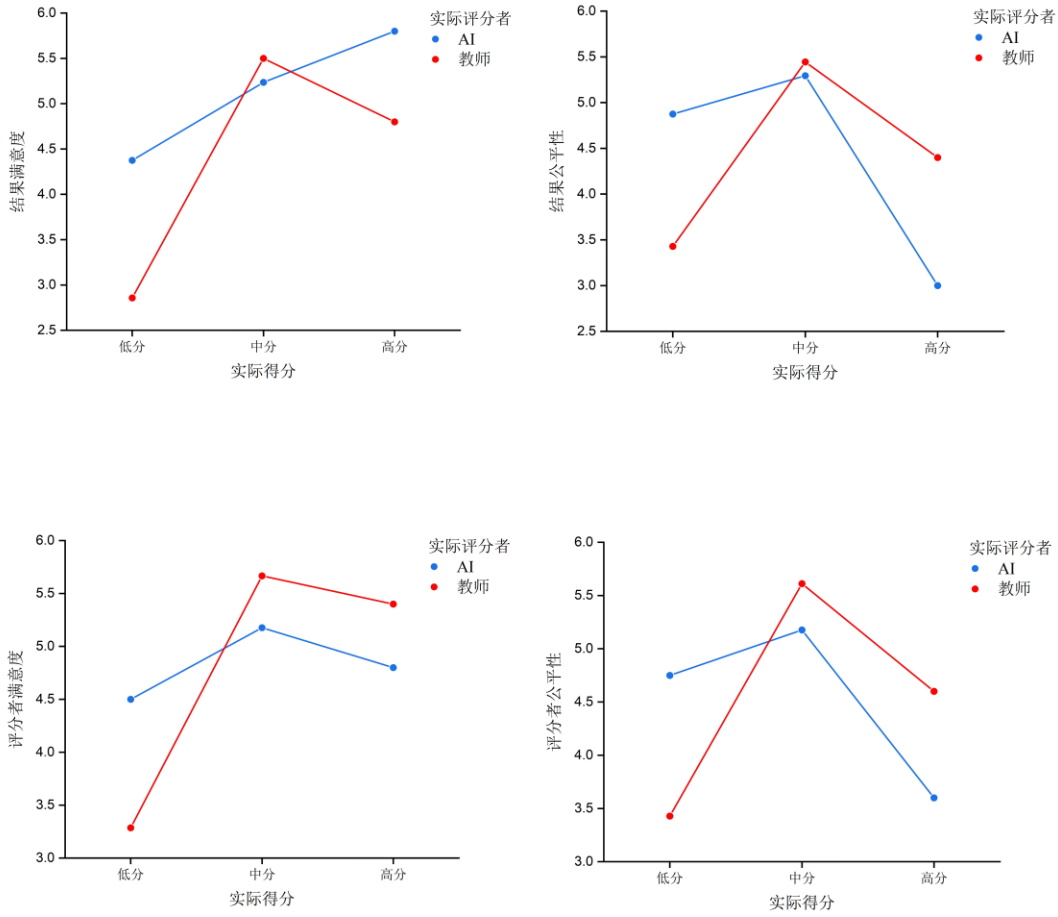

图 4-9　实际评分者对显性感知的影响（期望评分者为教师）

当期望评分者为大学英语教师时，进行 2（实际评分者：AI、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。结果表明：期望-实际评分者一致性主效应不显著，$F(4, 51) = 2.01$，$p = 0.09$，$\eta_p^2 = 0.14$；期望-实际评分者一致性和实际得分交互作用不显著，$F(8, 102) = 1.70$，$p = 0.11$，$\eta_p^2 = 0.12$。后续双因素方差分析的结果和多元方差分析的结果相一致。

研究 2 旨在探究被试在得到实际评分之前，是否会对 AI 评分或者是教师评分有不同的偏好，以及这种偏好是否会影响其后的公平性和满意度感知。进一步，探究期望评分者和实际评分者的一致性是否会影响公平性和满意度感

知。研究 2 的结果表明，被试对评分者的显性偏好并不显著，即在没有实际评分结果影响的情况下，被试并没有表现出对 AI 评分系统或教师评分的明显偏好。此外，期望评分者与实际评分者一致性的高低也并不显著影响被试的公平性和满意度感知。这可能说明，在评分结果未知的情况下，被试对于评分者的期望并不强烈，或者他们对 AI 评分和教师评分持有相似的态度和预期。

4.3 研究 3　对 AI 评分的内隐态度

4.3.1 研究目的

在上述研究中，我们未能在评价者偏好上得到普遍的外显结果，这与先前的研究结果存在一定的差异。因此在研究 3 中，我们尝试使用 IAT 范式，通过“人-AI”和“评价-受评”两组关系词，对评价者内隐偏好进行进一步探究。此外，我们注意到，当 AI 进行评分任务时，其相对人处于上位，表现一种“控制”“指导”关系；但在我们生活中，AI 大多进行决策任务，其相对人则处于下位，表现出一种“辅助”关系。为了验证这一解释的合理性，我们额外设置了一组“控制-受控”属性词，用于表示经典上下位关系，对比“评价”关系与典型的“上位”关系是否表现出一致性。本研究目的如下：

① 探究个体对 AI 评分与对人评分的内隐态度差异。

② 探究“评分”关系与“上位”关系的一致性。

4.3.2 被试

使用 Gpower 3.1 计算被试量，选择配对样本 t 检验（单尾），dz = 0.5，α = 0.05，1-β = 0.8, 计算得到最小被试量为 27。通过 Credamo 平台招募被试，要求非英语专业，右利手，年龄在 18~55 岁，高中以上学历。由于本次收集时间较短，且由于线上问卷回答质量波动较大，目前共收集到有效问卷 24 份，其中男性被试 11 名，女性被试 13 名，年龄 21.88±3.43 岁。

4.3.3 研究设计

该 IAT 研究包含两个子实验，每个实验有两个自变量，采用被试内设计。第一个自变量为概念图，包括两个水平：AI，人。第二个自变量为属性词，包括两个水平，在实验一中为：点评词（“评分”），受评词（“受评”）；在实验二中为上位词（“控制”），下位词（“受控”）。该研究的因变量为 IAT 效应。

4.3.4 工具及材料

本研究通过在线问卷收集平台进行，以对 AI 评分的内隐态度为主题，为了让 AI 点评的行为更加真实，这里使用类人机器人的图片作为 AI 的概念词。IAT 实验中呈现的刺激材料共有六种：人的图片、类人机器人图片、受评词、点评词、上位词、下位词（具体词表见附录），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。

表 4-6　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

注：B=block

4.3.5 研究过程

具体实验序列如表 4-6 所示。首先，通过在线问卷收集平台上发布问卷，被试被随机分为两组，一组先进行“评价组”再进行“控制组”，另一组顺序相反。在每一个实验中，一半的被试先进行相容任务，后进行不相容任务；另一半被试先进行不相容任务，后进行相容任务，相容与不相容的顺序在被试间被平衡。被试先进行练习，以“评价组”为例，相容任务（或不相容任务）的练习共有两部

分组成：（1）属性词（点评和受评）的分类任务；（2）概念图（人与 AI）的分类任务。在进行相容任务练习时，首先进行概念词（点评与受评）的分类任务。此时，屏幕上方一左一右分别呈现一个属性词，左侧为“点评”，右侧为“受评”。而后在屏幕下方中央呈现一个词（如“评分”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。属性词分类任务结束后，进入概念图分类任务。同样，屏幕上方一左一右分别呈现一个类别词，左侧为“人”，右侧为“AI”，而后在屏幕中央呈现一个图（如一张亚洲人的照片），被试的任务时判定该图属于左侧类别（按“F”键）还是右侧类别（按“J”键）。在进行不相容任务的练习时，首先进行概念图（人与 AI）的分类任务，屏幕上方一左一右分别呈现概念词，但与相容任务相反：左侧为“AI”，右侧为“人”。而后在屏幕下方中央呈现一张图（如“AI”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。概念图的分类任务结束后，进入属性词的分类任务。同样，屏幕上方一左一右分别呈现一个属性词，与相容任务一致：左侧为“点评”，右侧为“受评”。而后在屏幕中央呈现一个词（如“评分”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。

正式实验则要求被试对概念图和属性词进行联合反应。在相容任务条件下，屏幕上方一左一右分别呈现两个词，左侧为“人或点评”，右侧为“AI 或受评”，而后在屏幕下方中央呈现一个词（如“评分”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）；在不相容任务条件下，同样，屏幕上方一左一右分别呈现两个词，左侧为“AI 或点评”，右侧为“人或受评”，而后在屏幕下方中央呈现一个词（如“评分”），被试的任务时判定该词属于左侧类别（按“F”键）还是右侧类别（按“J”键）。

上述每一反应会持续到被试按键为止，反应时和正确率由计算机自动记录。参考过往研究，按错试次的反应时额外增加 500ms 的惩罚反应时。数据处理时不删除任何数据。接下来对所有反应时数据进行对数转换，再对相容组（AI-受评）和不相容组（AI-点评）分别计算其平均反应时。最后，把不相容组的平均反应时减去相容组的平均反应时，这样，所得到的分数便为相对于 AI-点评而言，把 AI 与受评相联的程度，即内隐态度对于 AI 应该被点评的强度。同理可以得到内隐态度对于人应该被点评的强度。

“控制组”在任务流程上和“评价组”一致。“控制组”采用和“评价组”中一样的人类和 AI 的概念图，但“控制组”的概念词采用和“控制”和“受控”相关的词语（具体词表见附录 1）。

4.3.6 分析及结果

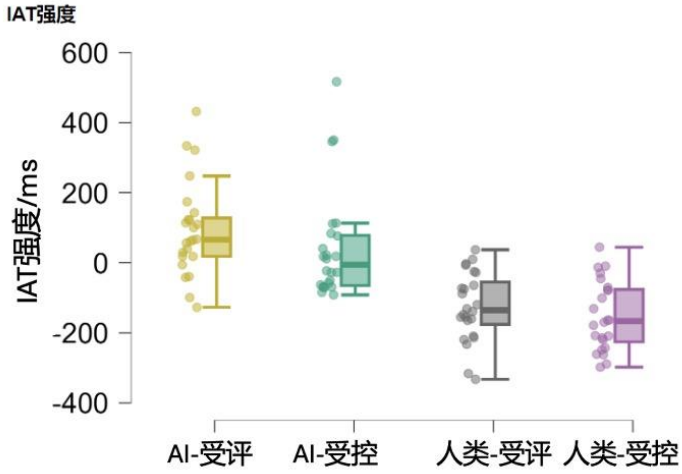

图 4-10　对 AI 和人类受评和受控的 IAT 强度

以相容与否为自变量，对 AI 的 IAT 强度进行独立样本 *t* 检验。结果显示：（1）评分组 IAT 强度显著大于 0，$t = 3.48$，$p = 0.002$，$d = 0.71$，说明被试对“AI-受评”的概念存在隐性偏好。（2）控制组 IAT 强度不显著，$t = 1.35$，$p = 0.19$，$d = 0.27$，说明被试对“AI-受控”的概念无显著隐性偏好。

以相容与否为自变量，对人类专家的 IAT 强度进行独立样本 *t* 检验。结果显示：（1）评分组 IAT 强度显著大于 0，$t = 6.18$，$p < 0.001$，$d = 1.02$，说明被试对“人-点评”的概念存在隐性偏好。（2）控制组 IAT 强度显著大于 0，$t = 7.50$，$p < 0.001$，$d = 1.58$，说明被试对“人-控制”的概念存在隐性偏好。

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，结果表明，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，$t = 0.98$，$p = 0.34$，$d = 0.25$。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。结果显示人类专家 IAT 强度显著高于 AI，$t = 2.91$，$p = 0.008$，$d = 0.83$，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。

尽管在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。对点评组和控制组的 AI 内隐态度进行了皮尔逊相关分析，结果显示 *Pearson's r* $= 0.21$，$p = 0.30$。对人类的内隐态度作相同分析得到 $r = -0.12$，$p = 0.55$。这可能是由于单个被试在单个客体上的试次数较少，导致结果的不显著。后续可以进一步增大样本容量以考察其结果的稳定性。

4.4 研究 4　教师与 AI 协作评分模式的偏好与感知

4.4.1 研究目的

研究 3 的结果发现，人们更倾向于将 AI 与“受评”“受控”等关键词联系起来，这与生活中 AI 通常充当辅助角色（例如 ChatGPT）的情况相符。但这一结果却未在外显选择和偏好上展现出来，这可能是由于 AI 评价较为少见，被试对评价者角色的理解不充分有关。因此，研究 4 聚焦于“辅助”这一常用的关系进行设计，并加入文字说明强化被试的理解，探究内隐偏好是否能映射到外显表现上。研究目的如下：

① 探究 AI 辅助教师和教师辅助 AI 作为评价者的选择偏好。

② 探究 AI 辅助教师或教师辅助 AI 作为评价者时，评价者与实际得分的交互作用。

4.4.2 被试

使用 Gpower 3.1 计算被试量，选择 *t* 检验，d＝0.5，α＝0.05， 1-β＝0.80，计算得到最小被试量为 102。通过 Credamo 平台招募被试，要求非英语专业，右利手，年龄在 18~55 岁，高中以上学历。由于本次收集时间较短，且由于线上问卷回答质量波动较大，目前共收集到有效问卷 95 份，其中男性被试 26 名，女性被试 69 名，年龄 22.67±4.98 岁，平均作答时间 8.73±2.23 分钟。

4.4.3 研究设计

研究 4 采用被试间设计，研究 4 有 3 个自变量：期望得分（高期望得分、中期望得分、低期望得分），期望评分模式（AI 为主导教师为辅助、教师为主导 AI 为辅助），实际得分（高期望得分、中期望得分、低期望得分）。3 个因变量为：选择 AI 辅助教师评分和教师辅助 AI 评分的比例，以及对公平性和对满意度的感知。

4.4.4 工具及材料

同研究 1。

4.4.5 研究过程

在研究 1 的基础上，被试在完成自评后，会首先看到对评分过程中“主导评分者”和“辅助评分者”的责任和职责描述。清楚了主导和辅助评分者之间的关系之后，被试可选择自己期望的评分模式是“AI 为主导评分者、教师为辅助评

分者”还是“教师为主导评分者、AI 为辅助评分者”。意愿选择完成后，系统将会按照选择意愿，由对应的评分模式进行评分（随机 2～9 分）。外显态度感知部分分为公平性和满意度感知。研究 4 的实验流程详见图 4-11。

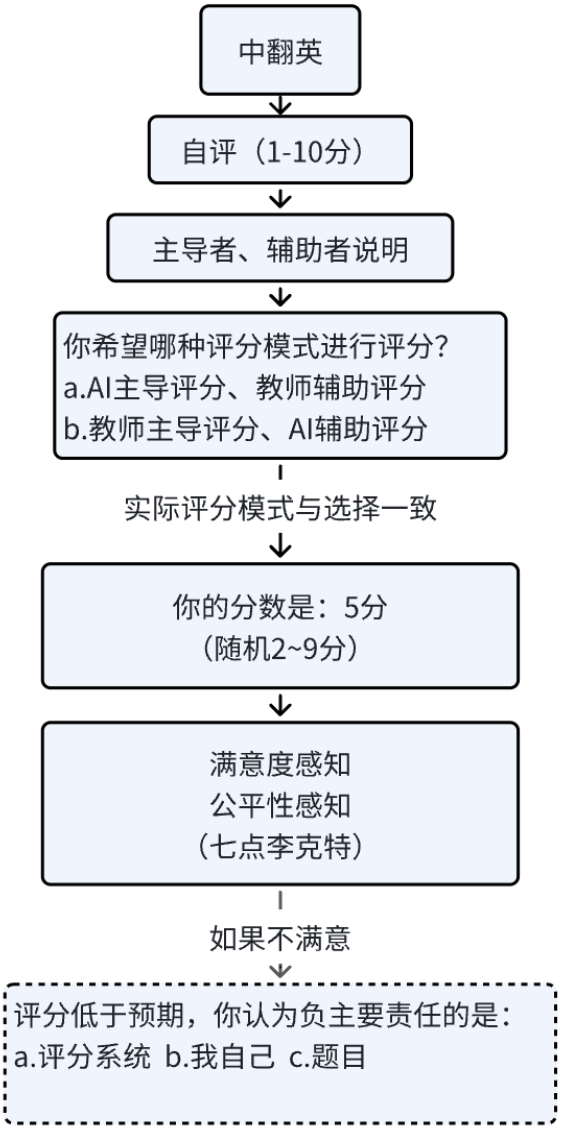

图 4-11　研究 4 实验流程图

4.4.6 分析及结果

为探究被试对不同评分模式的偏好，对选择频数进行卡方检验。结果表明，被试选择“教师主导评分、AI 辅助评分”进行评分的频数（$f = 74$）差异高于选择“AI 主导评分、教师辅助评分”的频数（$f = 21$），$\chi^2(1) = 29.57$，$p < 0.001$。为探究期望评分模式的选择偏好是否会受到不同期望得分的影响，将期望得分为三档，小于等于 3 分为低分，4～7 分为中分，大于等于 8 分为高分，并对期望得分和

期望评分模式进行独立性检验（见表 4-7）。结果表明，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，$\chi^2(2) = 3.14$，$p = 0.21$，Crame's $V = 0.18$。

表 4-7　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

将实际得分分为低分（2～3 分）、中分（4～7 分）、高分（8～9 分）。为了探究评分模式、实际得分对显性感知的影响，以公平性和满意度为因变量，评分模式和实际得分为组间变量进行多变量方差分析（见图 4-12）。结果表明：评分模式对显性感知没有显著影响，$F(4, 86) = 1.49$，$p = 0.21$，$\eta_p^2 = 0.06$；实际得分主效应显著，$F(8, 174) = 4.86$，$p < 0.001$，$\eta_p^2 = 0.18$。

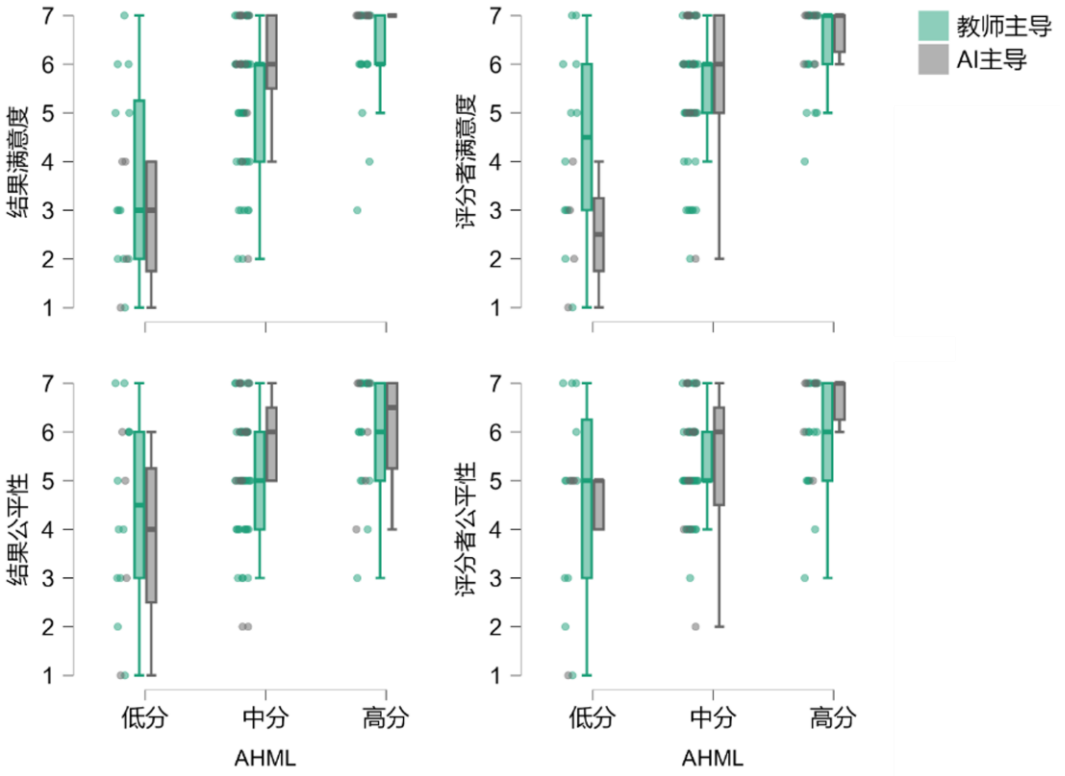

图 4-12　评分者对公平性和满意度的影响

后续单因素方差分析也表明，评分模式不会对结果满意度、结果公平性、评分者公平性感知产生显著的影响，且评分模式和实际得分之间没有交互作用。然而，在评分者满意度的感知上，评分模式和实际得分交互作用显著，$F(2, 89) = 3.59$，$p = 0.03$，$\eta_p^2 = 0.08$；然而事后检验表明，在不同实际得分的情况下，评分模式对评分者满意度均没有显著差异。我们推测此处的交互作用显著是由于数据量较少，出现了抽样误差。

4.5 研究 5　AI 辅助教师评分系统的探索

4.5.1 研究目的

在研究 4 中，我们发现被试对于教师主导，AI 辅助的评分系统具有明显的选择偏好。为了进一步探索该评分模式，在本实验中，我们参考研究 1 和研究 2 的范式和设计，将其与单独的 AI 或教师评分进行对比，通过强制分配和意愿选择的方式，测定被试的选择偏好，并比较结果的满意度和公平度感知。因此，研究目的如下：

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。

4.5.2 被试

使用 Gpower 3.1 计算被试量，选择 *F* 检验, f＝0.25，α ＝0.05，1-β ＝0.80，计算得到最小被试量为 196。通过 Credamo 平台招募被试，要求非英语专业，右利手，年龄在 18~55 岁，高中以上学历。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，平均作答时间 8.33±2.56 分钟；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。平均作答时间 8.24±1.96 分钟。

4.5.3 研究设计

研究 5 的设计在研究 1 和研究 2 的基础上增加了部分因素，分为无意愿选择和有意愿选择两个子实验。无意愿选择实验中，在评分者分配环节增加了 AI 辅助教师评分（教师主导）选项；有意愿选择实验中，在评分者意愿选择和实际评分者中增加了 AI 辅助教师评分（教师主导）选项。

研究 5 无意愿选择实验的自变量为评分者（AI 评分系统、大学英语教师、AI 辅助教师）和实际得分（高期望得分、中期望得分、低期望得分），因变量为对公平性和对满意度的感知。有意愿选择实验在无意愿实验的基础上增加了两个自变量，分别为期望评分者（AI 评分系统、大学英语教师、AI 辅助教师）和期望与实际评分者一致性（一致、不一致），其因变量除了满意度和公平性感知，还有选择 AI 评分、教师评分或 AI 辅助教师评分的比例。

4.5.4 工具及材料

同研究 2。

4.5.5 研究过程

无意愿选择实验基本与研究 1 一致，在评分者分配中，被试可能会被分配到“AI 评分系统”、“大学英语教师”或“AI 辅助教师（教师为主导）”三个选项中。

有意愿选择实验基本与研究 2 一致，被试可选择“AI 评分系统”、“大学英语教师”或是“AI 辅助教师（教师为主导）”作为评分者，意愿选择完成后，被试同样会被随机分配至这三个选项中，详见图 4-13。

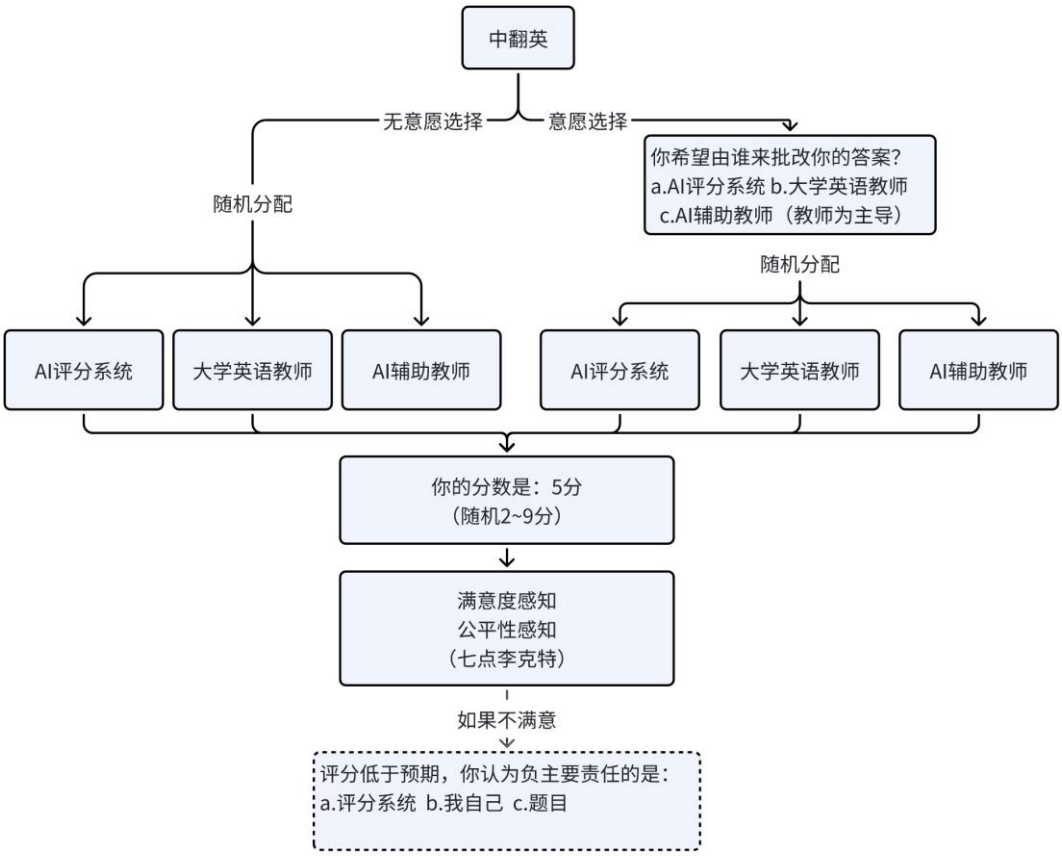

图 4-13　研究 5 流程图

4.5.6 分析及结果

（1）无意愿选择结果分析

以满意度、公平性感知为因变量，实际评分者、实际得分为自变量进行多因素方差分析。结果发现，实际得分在结果满意度和评分者满意度上效应显著或边缘显著；并且仅在结果满意度变量上，评分者效应边缘显著（$p = 0.06$，$\eta_p^2 = 0.02$），实际得分与评分者交互作用边缘显著，（$p = 0.07$，$\eta_p^2 = 0.04$）。

表 4-8　无意愿时外显感知方差分析结果

自变量	因变量	F	p	η_p^2
评分者	结果满意度	2.88	0.06	0.00
	结果公平性	1.27	0.28	0.01
	评分者满意度	0.60	0.55	0.01
	评分者公平性	0.68	0.51	0.01
实际得分	结果满意度	14.63	< 0.001	0.12
	结果公平性	2.18	0.12	0.02

	评分者满意度	4.59	0.01	0.05
	评分者公平性	0.20	0.82	0.00
评分者×实际得分	结果满意度	2.21	0.07	0.04
	结果公平性	1.19	0.32	0.02
	评分者满意度	0.89	0.47	0.02
	评分者公平性	0.28	0.89	0.01

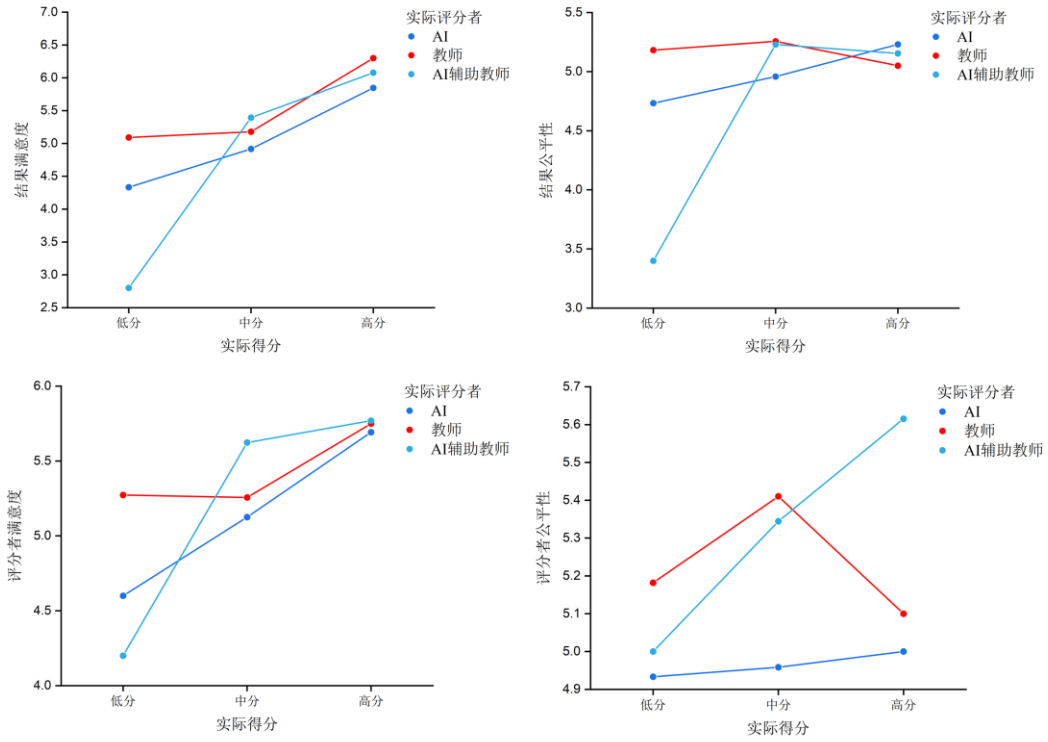

图 4-14　无意愿选择方差分析结果

根据初步分析的结果，选择实际得分的低分组，对结果满意度进行单因素方差分析，发现评分者主效应边缘显著，$F(2,28)=2.97$，$p=0.068$，$\eta_p^2=0.02$。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。

（2）有意愿选择结果分析

表 4-9　外显感知方差分析结果

自变量	因变量	F	p	η_p^2
实际评分者	结果满意度	1.90	0.152	0.02
	结果公平性	0.53	0.589	0.004
	评分者满意度	1.93	0.143	0.02
	评分者公平性	1.57	0.210	0.01

期望一致性	结果满意度	0.89	0.346	0.003
	结果公平性	0.23	0.631	0.00
	评分者满意度	0.38	0.536	0.001
	评分者公平性	0.26	0.612	0.001
实际得分	结果满意度	16.72	< 0.001	0.06
	结果公平性	6.84	0.001	0.05
	评分者满意度	8.41	< 0.001	0.07
	评分者公平性	5.74	0.004	0.05

对期望评分者选择情况进行卡方检验，结果发现被试存在显著偏好（χ^2 = 16.18，p < 0.001），选择“AI 辅助教师”的频数（f = 108）显著高于选择 AI（f = 72）或选择教师的频数（f = 59）（χ^2 = 14.38，p < 0.001；χ^2 = 7.20，p = 0.007），而选择 AI 或教师评分的频数无显著差异（χ^2 = 1.29，p = 0.26）。

以满意度、公平性感知为因变量，实际评分者、期望一致性、实际得分为自变量进行方差分析。结果表明，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。方差分析结果见表 4-9。

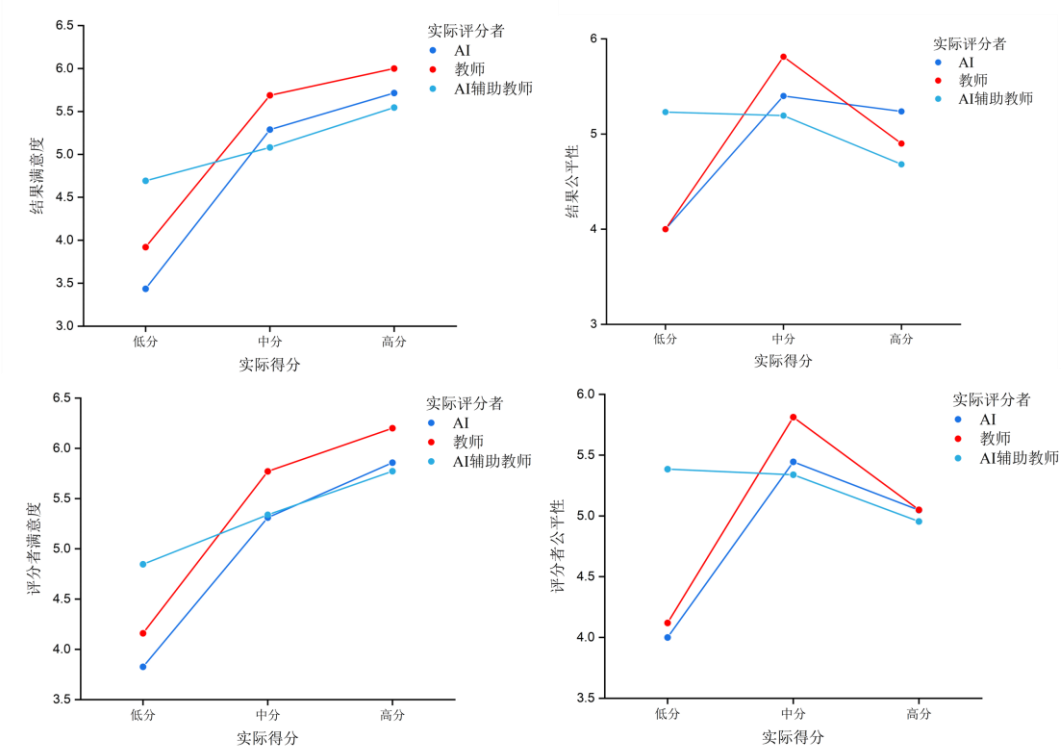

图 4-15　实际评分者和实际得分对显性感知的影响

研究 5 旨在从外显角度探究“AI 辅助教师”和“教师辅助 AI”两种评分系统的选择偏好与感知情况。结果发现，被试对于“AI 辅助教师”评分系统具有明

显的选择偏好，该结果与研究 3 发现的内隐偏好一致，即被试更能接受“人主导，AI 辅助”的关系模式。对满意度和公平性的分析结果发现，评分模式和实际得分存在交互作用，但事后检验未发现特定变量水平上的效应，该感知结果还需要进一步的实验验证。

5 综合讨论

本项目聚焦于人工智能与教师评分对结果满意度和公平性感知的影响，针对大学生这一特定群体，通过线上心理实验平台，深入探究了期望得分、实际得分、期望评价者、实际评价者等多个自变量与结果感知的关系，同时，我们从上下位关系角度，对于 AI 与人的感知和偏好差异提出了一种新的解释，并通过内隐测验进行了验证。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。

在本项目中，我们使用了“答题-感知评价”的线上实验方法，通过一道难度适当的翻译试题，提高被试的自我卷入程度，尝试模拟现实中作业批改或是考试改卷的场景。根据对实际得分的分析，我们发现实际得分的高低对满意度有显著影响，而对公平度的影响不明显；对期望得分分析发现，期望得分越高，满意度偏低，但这一结果较为模糊；而分析期望和实际得分的交互作用，我们发现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（偏高或偏低），公平性感知则会显著降低。该结果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，此实验结果可能与现实生活存在一定区别，在实际的考试和评分中，只要得到较高的分数，学生就会有较高的满意度，并认为该结果公平，即便此分数远高于预期。这可能是因为，个人利益在现实中具有不可忽视的影响，而在实验中，许多被试秉持“测试 AI 性能”的态度，因而得到的结果更加理性客观。

在研究 1 中，实际评分者（AI 或人）对结果没有显著影响，我们猜测可能是因为实际评分过程中，评分结果效应过于显著，弱化了被试对评分者的感知。为了更深入探究评分者偏好这一问题，在研究 2 中引入了期望评分者这一变量，希望在结果呈现之前，探究对于 AI 评分或教师评分的偏好。从总体上来看，被试对评分者无明显的选择偏好，并且期望评分者、实际评分者以及二者之间的一致性都不会显著影响满意度和公平度的得分情况。这可能是因为大学生及就业青年使用 AI 的频率较高，对其有较高的熟悉程度和信任程度，再加上实验中设定 AI 与教师的水平一致，因此结果无显著变异情况。另外，我们也发现有些反馈提到，选择 AI 的原因是“好奇”“想尝试一下新的方式”，这反映了我们并没有真正实现被试的高度自我卷入，使得被试以开放的态度去选择期望评分者。

但值得注意的是，当聚焦于低期望评分和高期望评分的被试时，评价者的意愿选择出现了差异。低期望评分的被试更倾向于选择 AI 评分，而高期望评分的

被试倾向于教师评分。这一结果说明了不同群体对 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，其中有人认为“AI 没有能力准确评估我的答案”；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，而 AI 评价则会让他们感到更放松，有人提到“我觉得我糟糕的答案可能会让老师抓狂”。此外，也有一些低期望评分的被试出于好奇心而选择 AI，以测试其评估极端答案的能力。

通过以上 2 个研究，我们从外显角度得到了一系列关于满意、公平感知的结论。而在研究 3 中，我们从内隐态度入手，对个体的真实想法进行了验证，确定了上下位关系为本项目中 AI 评价和生活中 AI 决策的本质区别，说明在个体的潜意识中，AI 更偏向于下位关系，执行“协助”“辅佐”的职能，这从某种程度上解释了本项目与以往研究的差异。这表明了我们对于 AI 的外显态度和内隐态度存在一定程度的分离，外显感知是在高级认知活动加工后的结果，其中带有的道德评价、人格特质等因素都会对外显感知带来影响。

当我们通过强调教师和 AI 的交互关系（主导、辅助）后，他们表现出对“教师主导、AI 辅助”的评分模式的显著偏好。这可能反映了一种认知上的和谐，即人们在内隐层面上对 AI 的辅助角色有积极态度，而在外显选择上也倾向于这种模式。

在研究 5 中，我们观察到一个有趣的现象：尽管内隐联想测验（IAT）的结果表明存在一种隐性的偏好，倾向于将人类置于评价者的位置，而将 AI 视为被评价的对象，但当被试在显性选择中必须在“AI 辅助教师”和“教师”评分模式之间做出选择时，他们表现出了对“AI 辅助教师”评分模式的明显偏好。这种偏好与内隐态度中“人本位”的倾向似乎相悖。然而，这种选择实际上可能反映了一种对 AI 辅助功能的积极评价。当 AI 被定位为辅助者时，人们似乎更愿意接受它，因为它可以增强教师的评分能力，而不是完全取代教师评分者的角色。这种偏好可能源于对 AI 技术潜在优势的认识，如提高效率、减少重复性劳动、提供更一致的评分标准等。

另外，在以往的研究中，年龄也是影响 AI 感知的关键因素（Schepman & Rodway, 2020），本研究中，选取的被试大多为大学生，年龄跨度和离散程度有限，而所得到的结论又与得分等实际情境密切相关，缺乏普适性，可能存在同辈效应，这也是本研究的局限所在。在未来的研究中，可以尝试扩大年龄范围，针对初高中生或是中年职场人士进行探索，比较不同年龄人群对于 AI 评价系统的态度以及关注点，为人工智能引入作业/考试批改和工作绩效评估提供支持。

总而言之，本研究探讨了不同类型评分者（人工智能与人类教师）对结果公

平性感知和满意度感知方面的影响。我们发现，实际得分显著影响了满意度得分，而期望得分与实际得分的一致性对公平性感知起到了关键作用。此外，当期望得分与实际得分相符时，公平性得分最高；当两者不一致时，公平性得分显著下降。在实际评分中，对公平性感知和满意度感知有显著影响的是实际得分和期望得分，评分者类型对显性感知影响不大。这些发现表明，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。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，这有助于促进更公平且更能令学生满意的教育体验。

6 结论

本研究通过一系列在线实验，深入探讨了人工智能（AI）评分系统与人类教师评分在教育评价中的公平性和满意度感知差异。研究发现，实际得分对满意度感知有显著的积极影响，而评分者类型（AI 或教师）并未显著影响满意度和公平性感知。此外，当实际得分较低且期望与实际评分者不一致时，对公平性的感知和满意度的感知有一定程度的提高。内隐联想测验（IAT）结果则显示，人们对 AI 受评存在隐性的偏好。

以下是本研究得出的主要结论：

1. 评分者类型对满意度和公平性感知的影响不显著。无论是 AI 评分还是教师评分，学生对评分结果的满意度和公平性感知没有显著差异。

2. 实际得分对满意度感知有显著影响。高实际得分能够显著提高学生的满意度感知，而低实际得分则可能导致满意度感知降低。

3. 期望得分与实际得分的一致性对公平性感知至关重要。当期望得分与实际得分相符时，对评分的公平性感知最高；而当两者不一致时，公平性感知会显著降低。

4. 内隐态度测验揭示了对 AI 评分的隐性偏好。尤其在 AI 处于受评地位时，对 AI 评分的内隐态度更为积极。

5. 我们对于“教师主导、AI 辅助”的评分模式有显著的选择偏好，且

本研究为 AI 评分系统在教育评价中的应用提供了实证支持，并指出了在不同评分情境下感知的差异性，对教育评价改革具有重要的理论和实践意义。研究结果强调了在设计基于 AI 的评估系统时，需要综合考虑认知和情境因素，以促进更公平且更能满足学生需求的教育评价体系。同时，本研究也提示了未来研究需要关注不同年龄、文化背景学生对 AI 评分系统的态度和偏好，以及如何在实际应用中平衡期望与实际得分，优化 AI 评分系统的设计与实施。

7 参考文献

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: the case of explainable recommender systems. *Human and machine learning: Visible, explainable, trustworthy and transparent*, 21-35.

Adams, J. S. (1965). Inequity in social exchange. In *Advances in experimental social psychology* (Vol. 2, pp. 267-299). Elsevier.

Al-Omari, Z., Alomari, K., & Aljawarneh, N. (2020). The role of empowerment in improving internal process, customer satisfaction, learning and growth. *Management Science Letters, 10*(4), 841-848.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society, 35*(3), 611-623.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.

Bies, R. J. (2013). 3 Are Procedural Justice and Interactional Justice Conceptually Distinct? *Handbook of organizational justice*, 85-112.

Bies, R. J., & Tyler, T. R. (1993). The "litigation mentality" in organizations: A test of alternative psychological explanations. *Organization Science*, *4*(3), 352-366.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. Proceedings of the 2018 Chi conference on human factors in computing systems,

Byrne, Z. S. (2005). Fairness reduces the negative effects of organizational politics on turnover intentions, citizenship behavior and job performance. *Journal of Business and Psychology*, *20*, 175-200.

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, *15*, 1221177.

Cherry, B., Ordóñez, L. D., & Gilliland, S. W. (2003). Grade expectations: The effects of expectations on fairness and satisfaction perceptions. *Journal of Behavioral Decision Making*, *16*(5), 375-395.

Cobb, A. T., Vest, M., & Hills, F. (1997). Who delivers justice? source perceptions of procedural fairness 1. *Journal of Applied Social Psychology*, *27*(12), 1021-1040.

Colledani, D., & Camperio Ciani, A. (2021). A worldwide internet study based on implicit association test revealed a higher prevalence of adult males' androphilia than ever reported before. *The Journal of Sexual Medicine*, *18*(1), 4-16.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology*, *86*(3), 386.

Colquitt, J. A., & Rodell, J. B. (2015). Measuring justice and fairness. *The Oxford handbook of justice in the workplace*, *1*, 187-202.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO),

Douglas, J. A., Douglas, A., McClelland, R. J., & Davies, J. (2015). Understanding student satisfaction and dissatisfaction: an interpretive study in the UK higher education context. *Studies in Higher Education*, *40*(2), 329-349.

Eisenberger, R., Fasolo, P., & Davis-LaMastro, V. (1990). Perceived organizational support and employee diligence, commitment, and innovation. *Journal of applied psychology*, *75*(1), 51.

González-Gómez, F., Guardiola, J., Rodríguez, Ó. M., & Alonso, M. Á. M. (2012). Gender differences in e-learning satisfaction. *Computers & Education*, *58*(1), 283-290.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456.

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, communication & society*, *21*(2), 191-207.

Krishnakumar, A. (2019). Assessing the Fairness of AI Recruitment systems.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-26.

Lind, E. A., Tyler, T. R., & Huo, Y. J. (1997). Procedural context and culture: Variation in the antecedents of procedural justice judgments. *Journal of personality and social psychology*, *73*(4), 767.

Locke, E. A., Sirota, D., & Wolfson, A. D. (1976). An experimental case study of the successes and failures of job enrichment in a government agency. *Journal of applied psychology, 61*(6), 701.

Narayanan, D., Nagpal, M., McGuire, J., Schweitzer, S., & De Cremer, D. (2024). Fairness perceptions of artificial intelligence: A review and path forward. *International Journal of Human–Computer Interaction, 40*(1), 4-23.

Palaiologos, A., Papazekos, P., & Panayotopoulou, L. (2011). Organizational justice and employee satisfaction in performance appraisal. *Journal of European Industrial Training, 35*(8), 826-840.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137-141.

Rezaei, A. R. (2011). Validity and reliability of the IAT: Measuring gender and ethnic stereotypes. *Computers in Human Behavior, 27*(5), 1937-1941.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Saifullah, N., Alam, M., Zafar, M. W., & Humayon, A. A. (2015). Job satisfaction: A Contest between human and organizational behavior. *International Journal of Economic Research, 6*(1), 45-51.

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in human behavior reports, 1*, 100014.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277-284.

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2022). Fairness, explainability and in-between: understanding the impact of different explanation methods on non-

expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*, *24*(1), 2.

Shulner-Tal, A., Kuflik, T., Kliger, D., & Mancini, A. (2024). Who Made That Decision and Why? Users' Perceptions of Human Versus AI Decision-Making and the Power of Explainable-AI. *International Journal of Human–Computer Interaction*, 1-18.

Tyler, T. R. (1989). The psychology of procedural justice: A test of the group-value model. *Journal of personality and social psychology*, *57*(5), 830.

Van den Bos, K., Lind, E. A., Vermunt, R., & Wilke, H. A. (1997). How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. *Journal of personality and social psychology*, *72*(5), 1034.

Wei, Z., Chen, Y., Ren, J., Piao, Y., Zhang, P., Zhao, Q., Zha, R., Qiu, B., Zhang, D., & Bi, Y. (2021). Behavioral and neural evidence that robots are implicitly perceived as a threat. *bioRxiv*, 2021.2008. 2013.456053.

Zhu, L., Martens, J. P., & Aquino, K. (2012). Third party responses to justice failure: An identity-based meaning maintenance model. *Organizational Psychology Review*, *2*(2), 129-151.

Shin, D. (2020). User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. Journal of Broadcasting & Electronic Media, 64(4), 541–565. https://doi.org/10.1080/08838151.2020.1843357

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Society, 35, 611–623. https://doi.org/10.1007/s00146-019-00931-w

Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. Computers in Human Behavior, 132, 107253.

https://doi.org/10.1016/j.chb.2022.107253

ABSTRACT

This study explores the differences in perceived fairness and satisfaction between AI-based scoring systems and human teacher scoring in educational assessments. Using an online experimental method, the study simulates exam scoring scenarios to analyze the impact of AI and teacher scoring on perceptions of fairness and satisfaction. The findings indicate that actual scores have a significant positive effect on perceived satisfaction, while the type of scorer (AI or teacher) does not significantly influence perceptions of fairness and satisfaction. Further investigation into the consistency between expected and actual scorers revealed that when actual scores are low and the expected and actual scorers do not match, there is a moderate increase in perceived fairness and satisfaction. Additionally, the study measured implicit attitudes toward AI scoring using the Implicit Association Test (IAT) and found an implicit preference for AI scoring, particularly when AI was in the position of the evaluator. These findings provide empirical support for the application of AI scoring systems in educational assessments and highlight the perceptual differences in various scoring contexts, offering important theoretical and practical implications for educational assessment reform.

附录 1

点评：评价 评分 点评 考查 审核

受评：受议 被议 被评 受查 受审

控制：控制 支配 领导 统领 权威

受控：受控 服从 顺从 依从 顺服