**人工智能与教师评分对结果公平性和满意度感知的影响研究方案**

1. 研究背景与目标

2022年，AI评分系统如ChatGPT开始进入教育领域，被一些教师用于作业评估，预示着其在教育评估中的重要性将日益增加。然而，AI评分与传统教师评分之间的感知差异，包括公平性和满意度，尚未得到充分研究。现有研究表明，学生可能认为AI评分更公平透明，但这些研究通常基于假设情境，缺乏实际体验的支持。本研究旨在通过模拟真实考试评分场景，探讨学生对AI评分与教师评分的感知差异，特别关注公平性和满意度。我们的目标是揭示AI评分系统的优劣，评估其在教育评价中的可行性和接受度，以期为教育者和技术开发者提供改进评分系统的见解，从而提高教育评价的质量和效率。

1. 研究方案
   1. 实验设计

2.1.1 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。该研究以评分者组间变量，采用单因素完全随机设计。自变量包括评价者主体类型，分别为：“AI评分系统”和“大学英语教师”。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。

2.1.2 研究2 实际得分和期望评分对公平性和满意度的影响

在研究2中，我们增大了实际得分的范围（2～9分），并添加了被试分数自评环节（1～10分），以探究实际评分和预期评分是否为影响被试感知的关键因素。研究2的研究目的如下：

① 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。

② 探究实际评分和期望得分的差异对结果公平性和满意度感知的影响。

研究2在研究1的基础之上增加实际得分高低这一因素，包括三个水平：低评分（2分、3分），中评分（4分、5分、6分、7分），高评分（8分、9分）。除此之外，被试在完成翻译后还需要对其答案进行自我评分（1～10分），其余部分与研究1相同。

2.1.3 研究3期望与实际评分者一致性对公平性和满意度感知的影响

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之间的相互作用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3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。

和研究2相比，研究3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。同时研究3将实际得分控制在中等分数（4～7分）。研究3有4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），期望得分（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。

2.1.4 研究4 不同实际得分下，期望与实际评分者一致性对公平性和满意度感知的影响

研究３中，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。受到研究2的启发，我们猜想不同实际得分下，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。因此，研究4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。

和研究3相比，研究3增加了1个自变量，即实际得分高低，水平为分别是低实际得分、中实际得分、高实际得分，其他部分和研究3相同。

2.1.5 研究5 对AI评分的内隐态度

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5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。这种上下位关系的差别可能是本项目结果与先前的AI决策研究不同的本质原因。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

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。

* 1. 被试招募
     1. 研究一

使用GPower3.1计算被试量，选择独立样本*t*检验，*d* = 0.5，*α* = 0.05，1 - *β* = 0.8，计算得到最小被试量为102。利用Credamo平台，至少收集102名满足以下条件的中国本土长大的被试：年龄在18至40岁之间，汉语为母语，学历高中及以上，矫正视力正常，没有心境障碍、精神分裂症谱系障碍或其他正在接受治疗的精神障碍，也没有酒精、药物等物质滥用史；男女各占一半。

* + 1. 研究二

使用GPower3.1计算被试量，选择*F*检验，计算得到最小被试量为125。计划通过Credamo平台收集有效问卷125份。利用Credamo平台，至少收集125名满足以下条件的中国本土长大的被试：年龄在18至40岁之间，汉语为母语，学历高中及以上，矫正视力正常，没有心境障碍、精神分裂症谱系障碍或其他正在接受治疗的精神障碍，也没有酒精、药物等物质滥用史；男女各占一半。

* + 1. 研究三

使用GPower3.1计算被试量，选择*F*检验，计算得到最小被试量为125。计划通过Credamo平台收集有效问卷125份。利用Credamo平台，至少收集125名满足以下条件的中国本土长大的被试：年龄在18至40岁之间，汉语为母语，学历高中及以上，矫正视力正常，没有心境障碍、精神分裂症谱系障碍或其他正在接受治疗的精神障碍，也没有酒精、药物等物质滥用史；男女各占一半。

* + 1. 研究四

使用GPower3.1计算被试量，选择*F*检验，计算得到最小被试量为125。计划通过Credamo平台收集有效问卷125份。利用Credamo平台，至少收集125名满足以下条件的中国本土长大的被试：年龄在18至40岁之间，汉语为母语，学历高中及以上，矫正视力正常，没有心境障碍、精神分裂症谱系障碍或其他正在接受治疗的精神障碍，也没有酒精、药物等物质滥用史；男女各占一半。

* + 1. 研究五

使用GPower3.1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。利用Credamo平台，至少收集27名满足以下条件的中国本土长大的被试：年龄在18至40岁之间，汉语为母语，学历高中及以上，矫正视力正常，没有心境障碍、精神分裂症谱系障碍或其他正在接受治疗的精神障碍，也没有酒精、药物等物质滥用史；男女各占一半。

* 1. 实验材料
     1. 实验一、二、三、四

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。

* + 1. 实验五

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 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 |

2.4实验任务

2.4.1 研究一

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，“AI评分系统”或“大学英语教师”将给出中等分数（随机4分、5分、6分、7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1-1。

|  |
| --- |
|  |
| 图 1-1 研究1流程图 |

2.4.2 研究二

在研究1的基础上，被试在完成翻译任务后，“AI评分系统”或“大学英语教师”将随机给出2～9分的评分。外显态度感知部分和研究1相同。研究2的实验流程详见图2-1。

|  |
| --- |
|  |
| 图 2-1 研究2流程图 |

2.4.3 研究三

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4～7分）。外显态度感知部分分为公平性和满意度感知。研究3的实验流程详见图3-1。

|  |
| --- |
|  |
| 图 3-1 研究3流程图 |

2.4.4 研究四

在研究3的基础上，被试在完成自评和意愿选择完成后，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）。研究4的实验流程详见图4-1。

|  |
| --- |
|  |
| 图 4-1 研究4流程图 |

2.4.5 研究五

具体过程如表1-1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。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“控制组”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）。