**2025年度浙江省自然科学基金**

**项目申请书正文撰写提纲**

二、其他浙江省自然科学基金项目申请书正文撰写提纲

1．项目名称

2．研究工作的科学意义与拟解决的关键科学问题

3．本项目研究目标，以及与申请者研究工作长期目标的关系

4．项目研究内容，研究方案和进度安排

5．项目创新之处

6．工作基础与工作条件

7．预期研究结果

8．参考文献

# 1．项目名称

教育领域中人工智能与人类评分感知差异及其影响因素的实证研究

# 2．研究工作的科学意义与拟解决的关键科学问题

本研究工作的科学意义主要体现在以下几个方面：

（1）评价方法的比较研究：

本研究对AI评分系统与传统教师评分进行比较分析，为教育评价方法提供了一种新的视角。这不仅有助于理解AI在教育评价中的应用前景，也为教育评价方法的创新提供了科学依据。

（2）评价感知的深入理解：

通过探讨被评价者对AI评分和教师评分的感知差异，本研究深化了我们对评价公平性和满意度感知的理解，揭示了评分体系选择对评价感知的影响，这对于设计更加公平、透明和被广泛接受的评价系统具有重要意义。

（3）人机交互的社会心理学分析：

研究AI评分系统在教育评价中的使用，可提供关于人机交互在实际应用场景中的社会心理学见解，有助于优化AI系统设计，使其更好地适应用户的需求和期待。

本研究工作拟解决的关键科学问题如下：

（1）评分结果的感知差异：

在AI与教师评分中，被评价者对评分结果的公平性和满意度有何差异？这一问题探讨了技术介入背景下评价感知的变化，为设计和应用更符合公平性要求的评价系统提供了实践指导。

（2）内隐态度的差异探究：

被评价者在面对AI与教师评分时，其内隐态度存在何种差异？此问题旨在探讨被评价者对AI评分系统的深层次接受度，对评价系统的普及和接受具有重要影响。

（3）个体和环境因素的影响：

个体人格特征和所处环境因素如何影响评分结果的外显感知？这一问题有助于揭示个体差异和环境背景如何共同作用于评价结果的感知过程，为开发个性化的教育评价系统提供了宝贵的启示。

综上所述，本研究不仅在理论上具有重要的科学价值，还通过解答这些关键科学问题，为教育评价领域提供实际的指导和建议，推动AI评分系统在教育评价中的健康发展和广泛应用。

# 3．本项目研究目标，以及与申请者研究工作长期目标的关系

**本项目的研究目标包含以下几个方面：**

本项目旨在深入探讨人工智能（AI）与传统教师评分在教育评价领域的应用，特别是对被评价者感知的影响。具体研究目标包括：

（1）外显感知分析：

考察被评价者在AI评分和教师评分情境下，对评分结果公平性和满意度的外显感知差异。分析影响这些差异感知差异的其他因素，以揭示评分主体类型对评价结果接受度的影响。

（2）内隐态度探究：

基于外显感知，进一步探索被评价者对AI与教师担任评分者时的内隐态度差异。该部分研究将提供更深层次的见解，评估人们对AI评分系统的潜在接受度和看法。

（3）个体特质的影响：

探讨个体人格特征，尤其是大五人格模型各维度，是如何影响被评价者对AI和教师评分结果的外显感知与反应。

（4）情景因素的影响：

深入探讨外部环境因素对被评价者的外显感知的影响，尤其是不同的情景因素（如焦虑水平、他人评价）如何调节被评价者对评分结果的感知和态度。

**与申请者研究工作长期目标的关系：**

申请者的长期研究目标集中在利用AI技术优化教育评价和实验教学，提升教育质量和效率。本项目作为申请者的一个重要研究议程，与其长期目标紧密相关，具体体现在以下几点：

（1）提升教育评价的公平性和准确性：

通过深入研究AI评分系统与传统评分方法的感知差异，本项目将为设计更加公平、透明和被广泛接受的教育评价系统提供理论支持，直接助力申请者优化教育评价的长期目标。

（2）深化AI在教育领域的应用研究：

申请者此前的项目已经探索了AI技术在心理学研究与教学中的应用，本项目进一步扩展了这一研究方向，将其应用至教育评价领域，为AI技术在教育中的全面应用奠定了理论和实证基础。

（3）促进教育技术的创新发展：

申请者在教育技术创新方面已有多项成果，本项目的研究成果将进一步推动教育评价技术的创新发展，有助于实现个性化和高效的教育评价系统。这与申请者提升教育质量和效率的长期目标完全一致。

综上所述，本项目不仅是申请者在AI评分系统研究领域的深化和扩展，也是其长期目标实现过程中的一个重要里程碑，对推动教育评价体系的创新和发展具有重要的战略意义。

# 4．项目研究内容，研究方案和进度安排

结合本项目的研究目标，本项目的计划进行四个系列研究：

* 研究一：AI与人类评分主体对满意度和公平性感知的影响因素研究
* 研究二：AI与人类评分主体的内隐与外显态度的研究
* 研究三：个体人格特征对AI与人类评分主体感知的研究
* 研究四：外部情景因素对AI与人类评分主体感知的研究

4.1 研究一：AI与人类评分主体对满意度和公平性感知的影响因素研究

本研究设计了五个子实验，旨在探讨以下几个方面对评分主体满意度和公平性感知的影响：

（1）被试选择意愿；

（2）实际评分者与期望评分者之间的一致性；

（3）实际得分的高低；

（4）实际评分与期望评分之间的差异大小；

4.1.1研究一实验*a*

1.研究目的

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分主体的结果满意度和公平性感知的影响。

2.研究假设

被评价者对于不同评分主体的满意度感知和公平性感知存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，采取独立样本*t*检验，采用*d* = 0.5，*α* = 0.05，1-*β* = 0.8，计算得到最小被试量为102。考虑到线上问卷回答质量波动较大，以10%拒绝率计算，利用在线实验平台共收集115名中国本土长大、母语为汉语、学历高中及以上的18～55岁被试。

3.2研究设计

该研究以评分者为组间变量，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”。实验包含两个因变量，分别是对评分结果和评分主体的满意度与公平性感知。

3.3工具及材料

通过在线实验收集平台进行实验。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，我们简化了三道中译英题目，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。

3.4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查手机，实验开始前先进行眼动校准，在正式实验中，被试的注视点不得离开屏幕超过10秒，否则判定为作弊或分心，其数据不予录用。

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”将随机给出中等分数（4分、5分、6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分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分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18点。

4.1.2研究一实验b

1.研究目的

①探究被试对评分者的选择意愿；②探究实际评分者和期望评分者一致性对评分结果满意度和公平性感知的影响。

2.研究假设

①被试对选择“AI评分系统”和“大学英语教师”进行评分的人数比率存在差异；②实际评分者和期望评分者一致与否对评分结果的满意度和公平性感知存在差异。具体表现为：当两者一致时，被试对评分主体的满意度和公平性感知要优于不一致条件；当两者不一致时，被试对“大学英语教师”为评分主体的满意度和公平性感知要优于“AI评分系统”。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择F检验，计算得到最小被试量为128。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集140名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

实验*b*增加了一个自变量为实际评分者和期望评分者的一致性程度：实际评分者和期望评分者相一致与不一致，其余部分和实验*a*相同。

3.3工具及材料

同实验*a*。

3.4研究过程

在实验*a*的基础上，被试在完成翻译任务后，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。外显态度感知部分和实验*a*相同。

4.1.3研究一实验c

1.研究目的

探究实际得分高低对不同评分主体的评分满意度和公平性感知的影响。

2.研究假设

实际得分高分、中分与低分之间，被试对不同评分主体的评分满意度和公平性感知存在差异。具体表现为：当被评价者得分偏低时，“大学英语教师”的满意度和公平度感知要优于“AI评分系统”；但是，当被评价者得分偏高时，“AI评分系统”与“大学英语教师”的满意度和公平性感知并不存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择F检验，计算得到最小被试量为158。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集175名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

实验*c*在实验*a*的基础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）。其余部分和实验*a*相同。

3.3工具及材料

同实验*a*。

3.4研究过程

在实验*a*的基础上，被试完成一道中译英的翻译题并提交后，“AI评分系统”或“大学英语教师”将随机给出2～9分的评分。评分后被试对不同评分主体的结果满意度和公平性感知进行评价。

4.1.4研究一实验*d*

1.研究目的

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。

2.研究假设

实际评分和期望得分的差值大小对评分结果满意度和公平性感知是否会因评分主体不同而存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择线性回归分析，计算得到最小被试量为68。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集75名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

实验*d*在实验*a*的基础上以评分者的实际得分与期望得分两者的差异作为自变量，其余部分和实验*a*相同。

3.3工具及材料

同实验*a*。

3.4研究过程

在实验*a*的基础上，被试在完成翻译任务后，对自己的答案进行自我评分。自评完成后，系统将会自动分配由“AI评分系统”或“大学英语教师”进行评分。外显态度感知部分和实验*a*相同。

4.1.5研究一实验e

1.研究目的

综合探究选择意愿、实际评分者和期望评分者是否一致、实际得分高低以及实际评分和期望得分的差异大小对得分评价和评分主体的满意度和公平性感知的影响。

2.研究假设

实际评分者和期望评分者是否一致、实际得分高低与否以及实际评分和期望得分的差异大小对得分评价和评分主体的满意度和公平性感知存在交互作用。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择线性回归分析，计算得到最小被试量为196名。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台收集216名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

实验*e*是将实验*a、b、c*和*d*中涉及的各个自变量进行综合考察其交互作用。

3.3工具及材料

同实验*a*。

3.4研究过程

被试在完成翻译任务后，对自己的答案进行自我评分。自评完成后，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，分数范围落在2～9分之间。评分后被试对不同评分主体的结果满意度和公平性感知进行评价。

4.2 研究二：AI与人类协作关系的内隐态度与外显感知的研究

4.2.1研究二实验*a*

1.研究目的

①探究个体对于AI评分的内隐态度；②探究个体对AI评分与人类评分的内隐态度差异。

2.研究假设

①个体对于AI上位评分持有相对负面的内隐态度；②个体对于人类上位持有相对正面的内隐态度。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择配对样本*t*检验（双尾），dz = 0.5，*α* = 0.05，1-*β* = 0.8，计算得到最小被试量为34。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集40名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

该IAT研究有两个自变量，采用被试内设计。第一个自变量为概念图，包括两个水平：第一个是AI，第二个是人。第二个自变量为属性词，包括两个水平：第一个是点评词（“评分”），第二个是受评词（“受评”）。该研究的因变量为IAT效应。

3.3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四种：人的图片、类人机器人图片、受评词、点评词（具体词表见附录），共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 1-1 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 |

3.4研究过程

具体过程如表1-1所示。通过在线问卷收集平台上发布问卷，被试被随机分为两组。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，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，按错的试次额外增加500ms的反应时。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。

4.2.2研究二实验*b*

1.研究目的

探究个体对AI辅助人类教师（人类教师为主）和人类教师辅助AI（人工智能为主）两种评价方式的外显偏好程度。

2.研究假设

①个体对于AI辅助人类教师（人类教师为主）的评价方式有更多的选择偏好；②两种评价方式的评分满意度和公平性感知不存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择卡方检验，计算得到最小被试量为68。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集75名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

研究二实验*b*以评分者为组间变量，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辅助人类教师（人类教师为主）评分系统”和“人类教师辅助AI（人工智能为主）评分系统”。其余部分和研究一实验*a*相同。

3.3工具及材料

同研究一实验*a*。

3.4研究过程

在研究一实验*a*的基础上，被试在完成翻译任务后，可选择自己期望的评分者是“AI辅助人类教师（人类教师为主）评分系统”或“人类教师辅助AI（人工智能为主）评分系统”。意愿选择完成后，系统将会被试意愿分配进行评分。外显态度感知部分和研究一实验*a*相同。

4.2.3研究二实验*c*

1.研究目的

探究个体对AI、教师与AI辅助人类教师三种评价方式的外显偏好程度以及对这三类评分主体的结果满意度和公平性感知。

2.研究假设

①个体对于AI辅助人类教师的评价方式有更多的选择偏好；②三种评价方式的评分满意度和公平性感知不存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择卡方检验，计算得到最小被试量为86。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集95名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

研究二实验*c*以评分者为组间变量。自变量包括三类评价者主体类型，分别为：“AI”、“人类教师”与“AI辅助人类教师”。其余部分和研究一实验*e*相同。

3.3工具及材料

同实验*a*。

3.4研究过程

在研究一实验*a*的基础上，被试在完成翻译任务后，可选择自己期望的评分者是“AI”、“人类教师”或“AI辅助人类教师”。意愿选择完成后，系统将会被试意愿分配进行评分。外显态度感知部分和研究一实验*e*相同。

4.3 研究三：个体人格特征对AI与人类评分主体感知的研究

4.3.1研究三

1.研究目的

探究个体人格因素（开放性、责任心、外向性、宜人性和神经质）对评价结果和评分主体的满意度和公平性感知的影响。

2.研究假设

个体人格因素尤其是神经质和开放性得分高低对评价结果满意度和评分主体的公平性感知存在显著差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择多元回归分析，计算得到最小被试量为352。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集380名中国本土长大的、汉语为母语的18～55岁学历高中及以上的被试。

3.2研究设计

研究三是在研究一实验*e*的基础上增加了大五人格测验（开放性、责任心、外向性、宜人性和神经质）的前测。

3.3工具及材料

在研究一实验*e*的基础上增加大五人格因素测验前测。

4.4 研究四：外部情景因素对AI与人类评分主体感知的研究

4.4.1研究四实验*a*

1.研究目的

探究个体所处的焦虑情景（焦虑唤起和缓解）是否对于不同评分主体的满意度和公平性感知存在影响。

2.研究假设

个体所处的不同焦虑情景会影响被试对不同评分主体的满意度和公平性感知。具体变现为：焦虑唤起组的被试，相较于焦虑缓解组，更有可能选择人类教师进行评价；同时，焦虑唤起组在评价结果的满意度和公平性感知上，人类教师要优于AI；焦虑缓解组和对照组则不存在差异。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择F检验，计算得到最小被试量为216。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集240名中国本土长大的、汉语为母语的18～35岁学历高中及以上的被试。

3.2研究设计

研究四实验*a*通过操作情景因素以改变被试的焦虑水平。该研究共有两个自变量，第一个是评分者，包含两个水平：AI评分和教师评分。第二个自变量是情景因素，包含3个水平：对照组、焦虑唤起组和焦虑缓解组。以这两个自变量为组间变量，采用双因素完全随机设计。因变量有三个，分别是AI的选择比例，以及对评分结果和评分主体的满意度与公平性感知。

3.3工具及材料

本研究在研究一实验*e*的基础上，增加了唤起或缓解状态焦虑的操作。多项研究指出，舒缓音乐和自然声音都可以帮助缓解焦虑和压力（Saadatmand et al., 2013; Aghaie et al., 2014），这里采用自然白噪音作为缓解被试状态焦虑的背景音乐。

3.4研究过程

通过在线实验收集平台上发布问卷，被试会被随机分为3组，即对照组、焦虑缓解组和焦虑唤起组。对照组的实验流程与研究一实验*e*完全相同，而焦虑缓解组则在完成中译英任务之后，获知最终评分结果前，会播放5分钟的自然白噪音，白噪音结束后继续评价任务；焦虑唤起组组则在完成中译英任务之后，获知最终评分结果前，需要完成一个5分钟的快速反应时任务，要求被试以极快的速度做出反应，如果不做反应或做出反应错误，则会有红色报警。以激发被试的焦虑水平，反应时结束后继续相应任务。

4.4.2研究四实验*b*

1.研究目的

探究他人评价对被试选择意愿的影响，以及对结果满意度和公平性感知的影响。

2.研究假设

当受到身边他人对于评分主体（AI或教师）的不同评价（积极评价或消极评价）时，被试对AI和教师的选择偏好可能会受影响，对于评价结果满意度和公平度感知也会因此存在差异。具体表现为：当个体接收到对某一评分者的积极评价时，其会对该评分者产生更多的选择偏好，并对结果有更高的满意度和公平性感知，反之，则会产生更少的选择偏好与更低的满意度和公平性感知；此外，考虑到可能存在的内隐偏见，他人评价与不同评分主体间可能存在交互作用，消极评价可能对AI（相较于人类教师）产生更多不利的影响。

3.研究方案

3.1参与者/被试

使用GPower3.1计算被试量，选择F检验，计算得到最小被试量为252。考虑到线上问卷回答质量波动较大，以10%的数据拒绝率计算，利用在线实验平台至少收集280名中国本土长大的、汉语为母语的18～35岁学历高中及以上的被试。

3.2研究设计

研究包含两个自变量，第一个是评分者，包含两个水平：AI评分与教师评分；第二个自变量是社会评价的内容，包括两个水平：积极评价与消极评价。以这两个自变量为组间变量，采用双因素完全随机设计。因变量有三个，分别是AI的选择比例，以及对评分结果和评分主体的满意度与公平性感知。

3.3工具及材料

在研究1实验*e*的基础上进行了修改，同时加入了社会评价变量，通过设置一定数量的“假被试”来实现。考虑到任务类型与实际情况的特点，初步选用三名“假被试”，通过较为情绪化（如惊喜、愤怒）的语句和对话，传达对某一评分主体的积极或消极评价。

3.4 研究过程

研究采用线下实验的方式，被试三人一组，先在实验室外等候，此时，三位“假被试”假装刚刚进行完实验，在离开时进行谈论，形成对某一评分主体的评价。根据被试听到谈论内容和对象，可分为四组，分别是：AI积极评价组，AI消极评价组，教师积极评价组，教师消极评价组。随后，被试进行中译英任务及后续的选择和感知评价，其流程与研究1实验*e*保持一致。

# 5．项目创新之处

本项目在多个方面展现了研究创新性，主要体现在以下几个方面：

（1）跨学科研究方法的应用

本项目采用跨学科的研究方法，结合心理学、教育学和人工智能技术，深入探讨AI评分系统与传统教师评分在教育评价领域的应用及其影响。这种跨学科的方法不仅能够深入分析评价过程中的人机互动问题，还能够综合考虑评价感知的心理学和社会文化因素，为教育评价领域提供新的研究视角和方法论。

（2）评价感知的内隐与外显差异研究

本项目通过设定高度自我卷入的评价场景，真实地且深入地探究被评价者对AI与教师评分的外显感知差异。同时，进一步比较了被评价者内隐态度的差异，这在现有研究中较为罕见。通过结合外显感知和内隐态度的双重分析，本项目能够更全面地理解不同评价系统对被评价者的影响，提供了评价公平性和满意度感知研究的新视角。

（3）个体与环境因素的综合分析

本项目不仅关注评价结果的感知差异，还考虑了个体人格特征和情景因素如何影响这些感知。通过这种综合分析，本项目能够揭示不同个体和情景下评价感知的变化规律，从而为设计更为个性化和适应性强的评价系统提供理论支持和实践指导。

（4）促进AI在教育评价领域的应用研究

本项目通过探索AI评分系统在教育评价中的应用及其影响，推动了AI技术在教育评价领域的应用研究。这不仅为AI技术未来在教育领域的进一步应用提供了理论和实证基础，还为未来教育技术的发展开辟了新的路径，助力教育评价的技术革新。

（5）强调教育评价系统设计的人文关怀

本项目强调在教育评价系统设计中融入人文关怀，重视被评价者的心理感受和需求。通过深入分析评价感知的差异及其影响因素，本项目为构建更加公平、透明且有利于学习者发展的评价系统提供了新的思路。这种以人为本的设计理念体现了对教育公平性和学习者体验的高度关注。

总体而言，本项目的创新之处在于其跨学科的研究方法、对评价感知差异的深入探讨、对个体与环境因素的综合分析以及在教育评价中对人文关怀的高度重视。这些创新为AI评分系统在教育评价领域的应用和发展提供了新的视角和深入的理解。

# 6．工作基础与工作条件

已具备的实验条件

**（1）实验环境**：浙江大学工业心理学国家专业实验室提供了一流的实验环境，包括先进的实验室设施和丰富的研究资源。实验室设有专门的实验区域，能够进行复杂的心理学实验和数据分析。

**（2）实验设备**：实验室配备了多种高端科研设备，如脑电波(EEG)记录仪、眼动追踪器、生理信号测量设备等，能够满足认知与脑研究、人因工程分析等领域的实验需求。

**（3）研究资源**：实验室拥有庞大的心理学文献数据库和心理行为大数据资源，便于开展基于数据的研究工作。

**（4）人才支持**：实验室汇聚了国内外一流的心理学、人工智能领域的专家学者，如长江学者沈模卫教授、青年长江学者周吉帆教授，他们为本研究提供了强大的学术支持和合作平台。

**（5）科研基金**：实验室在多年发展中积累了丰富的科研项目经验，获得了国家、省部级等多级科研基金支持，为项目的研究提供了稳定的资金保障。

尚缺少的实验条件

**（1）特定AI软硬件资源**：尽管实验室设备齐全，但针对本项目的特定需求，可能需要进一步引进最新的AI评分系统软硬件资源，以及相关的开发和测试工具。

**（2）跨学科合作平台**：本项目需要心理学、教育学与人工智能等多学科的紧密合作，目前可能还需要进一步构建和完善跨学科合作的机制和平台。

拟解决的途径

**（1）科研基金申请**：通过国家、省部级等科研基金申请，为购置特定AI软硬件资源和开发工具提供资金支持。

**（2）建设合作网络**：利用浙江大学及工业心理学国家专业实验室的资源优势，积极构建与国内外高校、研究机构的合作关系，特别是在人工智能领域，以获取最新的技术支持和共享资源。

**（3）依托重点学科**：通过依托浙江大学的重点与优势学科，如人工智能、心理学等，集聚学科交叉的研究力量和创新思维，推动项目的研究工作。

**（4）利用国家、省部级重点实验室**：积极与国家、省部级重点实验室建立合作关系，借助其先进的研究平台和丰富的科研资源，共同推进项目的研究工作。

通过上述途径，可以有效补充和完善项目所需的实验条件，为实现项目目标提供有力的支持。

本项目预期将达到以下研究结果：

# 7．预期研究结果

（1）AI评分系统与传统教师评分的比较分析

**定量数据分析**：通过实证研究，收集并分析AI评分系统与传统教师评分在准确性、一致性和公平性等方面的表现数据，提供定量的比较结果，以评估两种评分方式的差异。

**定性感知差异**：深入探讨学生和教师对于AI评分和人工评分的感知差异，包括满意度、接受度、信任度等维度上的差异，以便更好地理解两者在外显感知与内隐态度方面的差异。

（2）评价感知差异的影响因素分析

**个体因素**：揭示个体差异（如大五人格特征、个体焦虑特质等）如何影响对AI评分系统的感知和接受度，明确个体差异对评价系统使用效果的作用。

**环境因素**：分析外部环境（如焦虑情景、他人评价等）对评价系统感知差异的影响，探讨如何通过环境调整来改善被评价者对评价过程的体验与接受度。

（3）优化教育评价模型的构建

**综合评价模型**：基于研究结果，构建一个融合AI和教师评分优势、同时考虑个体与环境因素的教育评价模型，提升评价的公平性、准确性和学生的学习体验。

**策略与实践建议**：提出针对不同教学场景如何有效实施AI评分系统的策略和建议，包括技术应用、教师培训、学生引导等方面的具体措施，以确保AI评分的有效性与可操作性。

（4）理论与实践的贡献

**理论贡献**：通过跨学科研究，丰富并深化心理学、教育学和人工智能领域对教育评价的理解，尤其是在评价感知和技术接受度方面，进一步完善相关的理论框架。

**实践指导**：研究成果将为教育机构实施和优化AI评分系统提供科学依据和操作指南，助力提升教育评价质量和效率，同时促进学生的公平、全面发展。

（5）政策建议与未来展望

**教育政策建议**：基于研究结果，向教育相关政策制定者提供关于AI在教育评价中应用的政策建议，涵盖技术规范、伦理指引以及监管框架，确保AI技术合理并安全的使用。

**未来研究方向**：鉴于本研究的局限性，为未来研究者指出潜在的研究路径与方法，探索更广泛的教育评价领域中的科技应用。

通过这些预期成果，本项目旨在为教育评价领域的科技应用提供理论和实践的支持，推动教育评价向着更加科学、合理和人性化的方向发展。

# 8．参考文献

[1] ABDOLLAHI B, NASRAOUI O. Transparency in fair machine learning: the case of explainable recommender systems [J]. Human and machine learning: Visible, explainable, trustworthy and transparent, 2018: 21-35.

[2] ADAMS J S. Inequity in social exchange [M]. Advances in experimental social psychology. Elsevier. 1965: 267-99.

[3] AGHAIE B, REJEH N, HERAVI-KARIMOOI M, et al. Effect of nature-based sound therapy on agitation and anxiety in coronary artery bypass graft patients during the weaning of mechanical ventilation: A randomised clinical trial [J]. International Journal of Nursing Studies, 2014, 51(4): 526-38.

[4] AL-OMARI Z, ALOMARI K, ALJAWARNEH N. The role of empowerment in improving internal process, customer satisfaction, learning and growth [J]. Management Science Letters, 2020, 10(4): 841-8.

[5] ARAUJO T, HELBERGER N, KRUIKEMEIER S, et al. In AI we trust? Perceptions about automated decision-making by artificial intelligence [J]. AI & society, 2020, 35(3): 611-23.

[6] ARRIETA A B, DíAZ-RODRíGUEZ N, DEL SER J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [J]. Information fusion, 2020, 58: 82-115.

[7] BIES R J. 3 Are Procedural Justice and Interactional Justice Conceptually Distinct? [J]. Handbook of organizational justice, 2013: 85-112.

[8] BIES R J, TYLER T R. The “litigation mentality” in organizations: A test of alternative psychological explanations [J]. Organization Science, 1993, 4(3): 352-66.

[9] BINNS R, VAN KLEEK M, VEALE M, et al. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions; proceedings of the Proceedings of the 2018 Chi conference on human factors in computing systems, F, 2018 [C].

[10] BYRNE Z S. Fairness reduces the negative effects of organizational politics on turnover intentions, citizenship behavior and job performance [J]. Journal of Business and Psychology, 2005, 20: 175-200.

[11] CHAI F, MA J, WANG Y, et al. Grading by AI makes me feel fairer? How different evaluators affect college students’ perception of fairness [J]. Frontiers in Psychology, 2024, 15: 1221177.

[12] CHERRY B, ORDóñEZ L D, GILLILAND S W. Grade expectations: The effects of expectations on fairness and satisfaction perceptions [J]. Journal of Behavioral Decision Making, 2003, 16(5): 375-95.

[13] COBB A T, VEST M, HILLS F. Who delivers justice? source perceptions of procedural fairness 1 [J]. Journal of Applied Social Psychology, 1997, 27(12): 1021-40.

[14] COLLEDANI D, CAMPERIO CIANI A. A worldwide internet study based on implicit association test revealed a higher prevalence of adult males’ androphilia than ever reported before [J]. The Journal of Sexual Medicine, 2021, 18(1): 4-16.

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

[16] COLQUITT J A, RODELL J B. Measuring justice and fairness [J]. The Oxford handbook of justice in the workplace, 2015, 1: 187-202.

[17] DOŠILOVIĆ F K, BRČIĆ M, HLUPIĆ N. Explainable artificial intelligence: A survey; proceedings of the 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), F, 2018 [C]. IEEE.

[18] DOUGLAS J A, DOUGLAS A, MCCLELLAND R J, et al. Understanding student satisfaction and dissatisfaction: an interpretive study in the UK higher education context [J]. Studies in Higher Education, 2015, 40(2): 329-49.

[19] EISENBERGER R, FASOLO P, DAVIS-LAMASTRO V. Perceived organizational support and employee diligence, commitment, and innovation [J]. Journal of applied psychology, 1990, 75(1): 51.

[20] GONZáLEZ-GóMEZ F, GUARDIOLA J, RODRíGUEZ Ó M, et al. Gender differences in e-learning satisfaction [J]. Computers & Education, 2012, 58(1): 283-90.

[21] GREENWALD A G, MCGHEE D E, SCHWARTZ J L. Measuring individual differences in implicit cognition: the implicit association test [J]. Journal of personality and social psychology, 1998, 74(6): 1464.

[22] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models [J]. ACM computing surveys (CSUR), 2018, 51(5): 1-42.

[23] HELBERGER N, ARAUJO T, DE VREESE C H. Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making [J]. Computer Law & Security Review, 2020, 39: 105456.

[24] HELBERGER N, KARPPINEN K, D’ACUNTO L. Exposure diversity as a design principle for recommender systems [J]. Information, communication & society, 2018, 21(2): 191-207.

[25] KRISHNAKUMAR A. Assessing the Fairness of AI Recruitment systems [J]. 2019.

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

[27] LEE M K, JAIN A, CHA H J, et al. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation [J]. Proceedings of the ACM on Human-Computer Interaction, 2019, 3(CSCW): 1-26.

[28] LIND E A, TYLER T R, HUO Y J. Procedural context and culture: Variation in the antecedents of procedural justice judgments [J]. Journal of personality and social psychology, 1997, 73(4): 767.

[29] LOCKE E A, SIROTA D, WOLFSON A D. An experimental case study of the successes and failures of job enrichment in a government agency [J]. Journal of Applied Psychology, 1976, 61(6): 701.

[30] NARAYANAN D, NAGPAL M, MCGUIRE J, et al. Fairness perceptions of artificial intelligence: A review and path forward [J]. International Journal of Human–Computer Interaction, 2024, 40(1): 4-23.

[31] PALAIOLOGOS A, PAPAZEKOS P, PANAYOTOPOULOU L. Organizational justice and employee satisfaction in performance appraisal [J]. Journal of European Industrial Training, 2011, 35(8): 826-40.

[32] RAI A. Explainable AI: From black box to glass box [J]. Journal of the Academy of Marketing Science, 2020, 48: 137-41.

[33] REZAEI A R. Validity and reliability of the IAT: Measuring gender and ethnic stereotypes [J]. Computers in human behavior, 2011, 27(5): 1937-41.

[34] RIBEIRO M T, SINGH S, GUESTRIN C. " Why should i trust you?" Explaining the predictions of any classifier; proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, F, 2016 [C].

[35] SAADATMAND V, REJEH N, HERAVI-KARIMOOI M, et al. Effect of nature-based sounds' intervention on agitation, anxiety, and stress in patients under mechanical ventilator support: a randomised controlled trial [J]. International journal of nursing studies, 2013, 50 7: 895-904.

[36] SAIFULLAH N, ALAM M, ZAFAR M W, et al. Job satisfaction: A Contest between human and organizational behavior [J]. International Journal of Economic Research, 2015, 6(1): 45-51.

[37] SCHEPMAN A, RODWAY P. Initial validation of the general attitudes towards Artificial Intelligence Scale [J]. Computers in human behavior reports, 2020, 1: 100014.

[38] SHIN D, PARK Y J. Role of fairness, accountability, and transparency in algorithmic affordance [J]. Computers in Human Behavior, 2019, 98: 277-84.

[39] SHULNER-TAL A, KUFLIK T, KLIGER D. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users’ perceptions of fairness toward an algorithmic system [J]. Ethics and Information Technology, 2022, 24(1): 2.

[40] SHULNER-TAL A, KUFLIK T, KLIGER D, et al. Who Made That Decision and Why? Users’ Perceptions of Human Versus AI Decision-Making and the Power of Explainable-AI [J]. International Journal of Human–Computer Interaction, 2024: 1-18.

[41] TYLER T R. The psychology of procedural justice: A test of the group-value model [J]. Journal of personality and social psychology, 1989, 57(5): 830.

[42] VAN DEN BOS K, LIND E A, VERMUNT R, et al. How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect [J]. Journal of personality and social psychology, 1997, 72(5): 1034.

[43] WEI Z, CHEN Y, REN J, et al. Behavioral and neural evidence that robots are implicitly perceived as a threat [J]. bioRxiv, 2021: 2021.08. 13.456053.

[44] ZHU L, MARTENS J P, AQUINO K. Third party responses to justice failure: An identity-based meaning maintenance model [J]. Organizational Psychology Review, 2012, 2(2): 129-51.