# 《心理学报》审稿意见与作者回应

题目：利用游戏 log-file 预测学生推理能力和数学成绩——机器学习的应用
作者：孙鑫，黎坚，符植煜

---

## 第一轮

**审稿人 1 意见：**尽管传统心理测验方式在我们生活中有着广泛应用，但因其诱发测验焦虑和题目曝光等问题，在测量精确性及其对受测者的长期影响等方面存在一些弊端。随着社会发展和技术进步，越来越多的研究者试图探讨新的测量工具和方法，基于游戏的评估就是其中一种受欢迎的应用。作者基于推箱子游戏，通过在 log-file 提取数据生成相应指标，并借助机器学习这样比较新颖的技术进行分析，实现对学生推理能力和数学成绩较好的预测。研究选题较为前沿，综述比较详实，研究过程一些细节交代比较清楚，得出了一些比较有意义的结果。作者这种努力和尝试非常值得嘉许和鼓励，必将对推进国内心理测量领域的发展有所启示。以下有几个问题或建议，供作者修改时参考。

**意见 1：**特征类别及其个数确定的逻辑：文中抽取了 23 个特征用于模型训练，这些特征类别确定的主要依据是什么？是否有全局视角的考虑？毕竟不同特征的选择对模型的效用存在差异，而且特征越多，模型的预测率自然更高。希望作者能对此做些说明。
**回应：**感谢专家的建议，我们已在正文中(采用蓝色字体)对特征选取的逻辑进行说明，请您再次审阅，谢谢！

**意见 2：**过程分析：与瑞文推理测验和数学测试相比，推箱子是个更为复杂的认知和操作任务。尽管建立的模型可以较好预测个体在瑞文推理和数学测试上是否成功，但我们对推箱子任务本身到底测量了什么仍是不清晰。是否可以替代瑞文推理和数学测试？当然这是这类研究普遍存在的问题。希望作者对如何解释和推广模型结果做些讨论。
**回应：**感谢专家的建议，关于这个问题，我们是这样考虑的：相对于瑞文测验的任务，推箱子确实是个更为复杂的认知任务，它本身到底涉及了哪些心理加工过程还有待探索，未来研究可以考虑结合眼动或者 fMRI 进行进一步分解。我们目前的研究是一项尝试性研究，试图结合游戏化任务和机器学习技术来解决传统心理测验的一些问题。传统心理测验主要是基于结果的评价，而我们的理念是认知加工原本是一个过程，可以通过收集大量过程性指标对其进行全程视角的分析。通过本研究的尝试，我们初步证明通过游戏的形式，log-file 的整理分析，以及机器学习算法，有可能发展出一种新的能力测量的方法。因此，目前阶段还没有考虑用推箱子来替代瑞文推理测验和数学测试，未来拟进行更加深入和全面的研究。关于您的建议我们在讨论部分有所增加(采用蓝色字体)，请您再次审阅，谢谢！

**意见 3：**分组标准：按照心理测量上通常做法，若分数呈现正态分数，进行高低分数时，通常确定高低 27%作为分界点，既能覆盖尽可能多的受测者，又能保证高低分组间的差异尽可能大。本研究选择 25%作为分界点，若采用 27%作为分界点，预测结果是否存在差异？建议作者尝试一下。
**回应：**感谢专家的建议，我们曾采用 27%作为分界点进行建模，结果发现模型整体效果与现有研究结果相比略差，具体表现为：在研究方法和参数设置等方面相同的前提下，对于数

学分类模型，以 25%为分界点，模型在精确率、查准率、查全率、F1 四个指标上的值分别是 69.44%、80.19%、69.67%、71.65%；以 27%为分界点，模型在精确率、查准率、查全率、F1 四个指标上的值分别是 67.47%、72.03%、59.12%、63.63%。对于瑞文推理分类模型，以 25%为分界点，模型在精确率、查准率、查全率、F1 四个指标上的值分别是 65.72%、76.11%、59.05%、64.22%；以 27%为分界点，模型在精确率、查准率、查全率、F1 四个指标上的值分别是 55.08%、59.62%、44.50%、48.91%。由此可以看出，以 27%为分界点所建立的模型整体上不如以 25%为分界点建立的模型效果好，这一结果与我们之前的想法有所吻合。在心理测量学研究中，采用 27%作为高低水平的临界点确实是常见做法，但也有很多研究会根据研究目的进行一定幅度的调整，其核心目标是既要让高低分组的被试在能力上有实质性的区分，又要有一定的被试量从而保证样本的代表性，例如，戴海琦老师的《心理与教育测量》认为“人数少可以用 50%为分界点”；Kaplan 的 Psychological Testing 中认为“可以使用三分之一作为分界点”等。在本研究中，由于受测者数量较多，并且来自同一所学校，能力水平较为接近。为了尽可能将高低能力组区分开，我们在分界点设置上要求的比较严格，选择了 25%作为分界点，以便使两组的差距增大。通过尝试不同的分界点进行建模，从结果可以看出，当不同能力组的受测者在能力相差相对较大的情况下，建立的模型具有更好的区分度。

**意见 4**：预测精度：目前通过研究结果主要是对受测者在瑞文推理和数学成绩是否处于前 25%和后 25%有比较高的预测率，远没有达到心理测量精确刻画个体差异的目的。建议在摘要在加入相应限定，如高低分组，以免读者误认为对分数全范围的预测。
**回应**：感谢专家的建议，已在摘要中加入相应限定(采用蓝色字体)，请您再次审阅，谢谢！
……

**审稿人 2 意见**：通过对游戏过程中的行为数据分析，识别学生的推理能力和数学成绩，数据采集及处理过程合理，结果恰当。
**意见 1**：在行为特征提取的时候，建议增加对时序特征的讨论。
**回应**：感谢专家的建议，已在特征提取部分增加了对特征类别的说明，以及时序特征的讨论(采用蓝色字体)，请您再次审阅，谢谢！

**意见 2**：在特征提取之后，可以计算特征与推理能力及数学成绩的相关性。
**回应**：感谢专家的建议，已计算特征与推理能力及数学成绩的相关性，具体结果如下表所示(由于特征数量较多，无法在一个页面显示全，因此将成功组和失败组特征分为两个表显示)。可以看出，对于瑞文能力和数学成绩，成功组和失败组的第一步用时/总时间、ln(第一步用时/总时间)、第一步用时/平均执行时间、ln(第一步用时/平均执行时间)这些特征均与它们有显著相关；对于数学成绩，失败组思考步数占比、失败组完成箱子的比例也与其有显著相关。在论文中由于篇幅限制没有加入表格，但增加了说明(采用蓝色字体)，请您再次审阅，谢谢！

表 1 成功组特征与推理能力和数学成绩的相关关系

| 变量 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.成功组 平均执行时间 | 1.00 | | | | | | | | | | | |
| 2.成功组 第一步用时/平均执行时间 | -0.13 | 1.00 | | | | | | | | | | |
| 3.成功组 ln(第一步用时/平均执行时间) | -0.06 | 0.89** | 1.00 | | | | | | | | | |
| 4.成功组 第一步用时/总时间 | -0.08 | 0.90** | 0.92** | 1.00 | | | | | | | | |
| 5.成功组 ln(第一步用时/总时间) | -0.03 | 0.79** | 0.92** | 0.95** | 1.00 | | | | | | | |
| 6.成功组 思考步数占比 | 0.18* | -0.07 | -0.11 | 0.05 | 0.08 | 1.00 | | | | | | |
| 7.成功组 执行间波动 | 0.35** | 0.04 | 0.11 | -0.07 | -0.06 | -0.44** | 1.00 | | | | | |
| 8.成功组 与最优步数相差 | -0.04 | -0.18* | -0.15 | -0.34** | -.367** | -0.32** | 0.19* | 1.00 | | | | |
| 9.成功组 重复步数占比 | -0.03 | -0.23** | -0.23** | -0.36** | -.393** | -0.25** | 0.12 | 0.82** | 1.00 | | | |
| 10.成功组 与最优路径重合步数占比 | 0.15* | 0.02 | -0.02 | 0.06 | 0.04 | 0.05 | 0.09 | -0.15* | 0.10 | 1.00 | | |
| 11.瑞文成绩 | 0.05 | 0.21** | 0.22** | 0.23** | .195** | -0.04 | 0.06 | -0.02 | -0.09 | -0.10 | 1.00 | |
| 12.数学成绩 | 0.02 | 0.42** | 0.46** | 0.43** | .416** | 0.03 | 0.02 | -0.06 | -0.14 | 0.02 | 0.44** | 1.00 |

注：*代表 $p<0.05$，**代表 $p<0.01$，以下同。

表 2 失败组特征与推理能力和数学成绩的相关关系

| 变量 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.失败组 平均执行时间 | 1.00 | | | | | | | | | | | | |
| 2.失败组 第一步用时/平均执行时间 | -0.15 | 1.00 | | | | | | | | | | | |
| 3.失败组 ln(第一步用时/平均执行时间) | -0.03 | 0.84** | 1.00 | | | | | | | | | | |
| 4.失败组 第一步用时/总时间 | -0.09 | 0.86** | 0.88** | 1.00 | | | | | | | | | |
| 5.失败组 ln(第一步用时/总时间) | -0.02 | 0.72** | 0.89** | 0.94** | 1.00 | | | | | | | | |
| 6.失败组 思考步数占比 | 0.10 | -0.20** | -0.18* | -0.01 | 0.04 | 1.00 | | | | | | | |
| 7.失败组 执行间波动 | 0.23** | 0.05 | 0.16* | 0.03 | 0.05 | -0.01 | 1.00 | | | | | | |
| 8.失败组 与最优步数相差 | 0.02 | -0.01 | -0.04 | -0.23** | -0.27** | -0.47** | 0.02 | 1.00 | | | | | |
| 9.失败组 重复步数占比 | 0.03 | -0.23** | -.23** | -0.40** | -0.41** | -0.33** | 0.01 | 0.84** | 1.00 | | | | |
| 10.失败组 与最优路径重合步数占比 | -0.19** | 0.16* | 0.15* | 0.17* | 0.13 | -0.06 | -0.06 | 0.09 | 0.04 | 1.00 | | | |
| 11.失败组 完成箱子的比例 | -0.03 | 0.24** | 0.23** | 0.12 | 0.09 | -0.39** | 0.13 | 0.10 | -0.06 | 0.25** | 1.00 | | |
| 12.瑞文成绩 | 0.07 | 0.16* | 0.22** | 0.22** | 0.19** | -0.06 | 0.06 | 0.01 | -0.14 | 0.12 | 0.11 | 1.00 | |
| 13.数学成绩 | 0.05 | 0.38** | 0.43** | 0.36** | 0.34** | -0.16* | 0.02 | 0.10 | -0.11 | 0.10 | 0.17* | 0.44** | 1.00 |

**意见 3：**目前的建模只使用了随机森林，建议尝试其他的建模方法，对它们的预测结果进行对比。

**回应：**感谢专家的建议，在随机森林之外，我们选取了 SVM 和 KNN 算法进行尝试，结果

发现，随机森林所建立的分类模型明显优于 SVM 和 KNN 算法所建立的模型，具体预测结果如下表所示。因此，在本研究的结果中，我们只报告了最优的模型。

表 3 不同建模方法在推理能力和数学成绩分类模型上的预测效果

| 算法 | F1 | 查准率 | 查全率 | 精确率 |
|---|---|---|---|---|
| **推理能力** | | | | |
| 随机森林 | 64.22% | 76.11% | 59.05% | 65.72% |
| SVM | 65.35% | 68.07% | 65.67% | 65.66% |
| KNN | 61.20% | 61.23% | 66.79% | 59.20% |
| **数学成绩** | | | | |
| 随机森林 | 71.65% | 80.19% | 69.67% | 69.44% |
| SVM | 61.50% | 65.23% | 59.17% | 63.39% |
| KNN | 56.80% | 60.51% | 55.07% | 58.93% |

**意见 4：**在论文的讨论部分，建议增加对特征的讨论，目前的特征更多的是根据经验设计，能否通过对特征的讨论发现一些新的内容。

**回应：**感谢专家的建议，已在讨论部分增加了对特征的讨论(采用蓝色字体)，请您再次审阅，谢谢！

---

## 第二轮

**主编意见：**作者基于游戏任务，通过提取 log-file 中的相应数据指标，并应用机器学习技术较好地预测了学生推理能力和数学成绩。研究具有一定的创新思想，这种尝试有利于推进国内心理学测量工具的发展。以下有仍一些建议和问题，供作者参考并修改完善论文。

**意见 1：**建议作者在前言部分就说明推箱子任务所选择的特征指标，并理清推箱子任务与所测量的推理能力以及数学成绩之间关系。

**回应：**感谢主编的建议，已在前言部分说明推箱子任务所选择的特征指标，并对推箱子任务与所测量的推理能力以及数学成绩之间的关系进行了论述(采用红色字体)，请您再次审阅，谢谢！

**意见 2：**作者计算了特征和推理能力以及数学成绩之间的相关性，是所有被试的数据还是其中提取的以 25%为分界点高低分组被试的数据？两者之间是否有显著差别？

**回应：**感谢主编的建议，文章中呈现的特征与推理能力和数学成绩之间的相关性是基于提取的以 25%为分界点高低分组被试的数据，此外，作者曾对所有被试的数据进行相关分析，结果略差于以 25%为分界点高低分组被试的数据结果，具体表现为：两类数据中，均是同样的特征指标与结果变量存在显著相关关系，对于以 25%为分界点高低分组被试的数据，成功组和失败组的第一步用时/总时间、ln(第一步用时/总时间)、第一步用时/平均执行时间、ln(第一步用时/平均执行时间)等特征均与瑞文测验成绩和数学成绩有显著相关，相关系数在

0.19~0.46 之间，数学成绩还与失败组思考步数占比、失败组完成箱子的比例显著相关，相关系数分别是 0.16 和 0.17；对于所有被试的数据，成功组和失败组的第一步用时/总时间、ln(第一步用时/总时间)、第一步用时/平均执行时间、ln(第一步用时/平均执行时间)等特征均与瑞文测验成绩和数学成绩有显著相关，相关系数在 0.15~0.37 之间，数学成绩还与失败组思考步数占比、失败组完成箱子的比例显著相关，相关系数均是 0.12。这在一定程度上说明文章以 25%为分界点构造高低分组被试的方法是合理的。

**意见 3**：作者在行为特征提取时，计算了特征和推理能力以及数学成绩之间的相关性，但发现其中有一些特征指标与推理能力和数学成绩之间相关不显著，且部分指标相关系数较小，相关性大小是否会影响作者考虑在模型中纳入哪些指标？作者能否进一步说明选用这些指标的有效性呢？如果减少所选特征指标构建模型，预测结果是否会更好？

**回应**：感谢主编的建议，在特征提取时，一方面会从理论层面选取我们认为可能有效的特征，另一方面会参考相关系数，作者曾尝试减少特征进行建模(剔除部分相关不显著的特征)，结果发现，现有的 23 个特征的模型在预测推理能力和数学成绩上具有最好的效果，因此最终呈现的结果即为含有 23 个特征的预测模型。关于某些特征指标与推理能力和数学成绩之间相关不显著或者相关系数较小的问题，作者是这样考虑的：目前传统的相关系数都是基于变量的线性相关假设而计算出的，然而实际中，两个关联密切的变量极有可能呈现出非线性的相关，因此不能只根据积差相关系数来判定是否应该将某个特征纳入随机森林模型。

**意见 4**：建议作者参考《心理学报》的文献引用和图表规范，规范引用文献的格式和图表的呈现方式。

**回应**：感谢主编的建议，已参考《心理学报》的文献引用和图表规范，对引用文献的格式和图表的呈现方式进行了修改，请您再次审阅，谢谢！