第一轮修改

意见一：研究议题属于心理测量中的成就类测验范畴，尤其是高利害的考试。
回应：感谢审稿专家的意见！本研究议题是基于教育数字化的大背景，心理测量和评估的数字化已成发展趋势。对于在计算机上实施的高利害测验，其测验过程和数据的安全性已然成为重要研究议题，本文正是在这样的时代背景下开展的研究，我们对这一点进行了阐明。

意见二：文献回顾不充分，一些有关该议题的研究均未被引用。建议作者进一步补充完善。
回应：感谢审稿专家的意见，这个意见非常有帮助！对于该议题有关的研究文献，我们进行了更加深入和细致的检索，补充了有关的文献，并进行相关的述评，我们已经对文献的回顾与综述作了进一步的补充和完善。

意见三：问题提出逻辑有待调整，作者已经在引言中提及了一些基于作答时间数据进行异常作答检测的研究；题目预知作为异常作答的一种类型，已有基于作答时间数据进行异常作答检测的研究有什么不足以至于作者开展一个新的研究？
回应：感谢审稿专家细致的审稿，非常好的意见！针对问题的提出逻辑，我们进行了调整，对于研究的动机进行了更加具体和明确的阐述。

题目预知作为最常见三种异常作答行为之一，一直以来受到研究人员的广泛关注，这在测量和评估数字化越来越成为趋势的今天尤为重要，研究者们针对这个议题展开了深入的探索和研究。

从以有研究来看，多数研究是基于作答得分数据展开的，有少数研究是基于作答时间数据开展的。由于作答时间数据容易获取，并且包含丰富的考生作答行为信息，本研究也是基于作答时间数据进行。

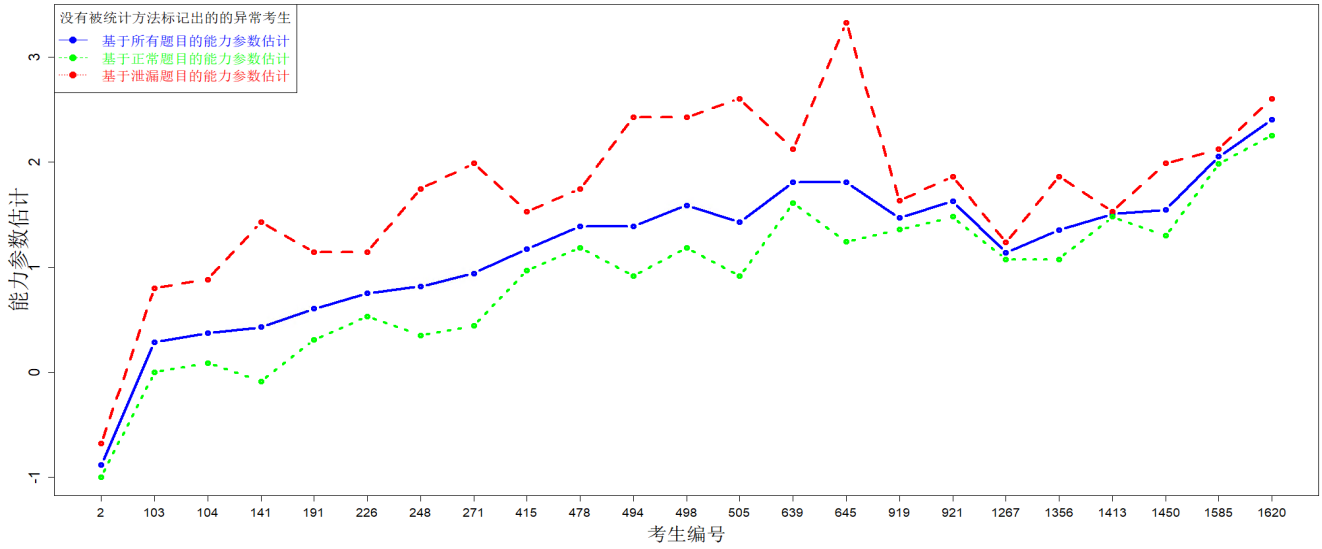
已有关于题目预知行为检测有关的研究表明，符号似然比检验相对于其它方法表现较好，但是缺点是它对于时间差异的敏感性较差，即考生在测验中正常题目集合和异常题目集合中表现出来的速度差异较小（即异常程度较低，表现在题目预知对于考生作答速度只有较小的提升）时检验力较低(Sinharay, 2017a, 2017b; Sinharay, 2020; Sinharay & Johnson, 2021)。在实际的测验情境中，题目预知对于不同考生所带来的速度差异水平有高有低，因此，需要探索不同速度差异水平下各检测方法的表现，尤其是速度差异水平较低时各方法的表现。基于以上的考虑，本研究构建基于作答时间数据的两个贝叶斯统计量：贝叶斯因子和后验概率，并探索当题目预知所带来的不同程度的异常水平时三种统计量（符号似然比，贝叶斯因子，后验概率）的表现。

意见四：建议作者进一步清晰阐述 3.3 节的例子所表达的意思。
回应：感谢审稿专家的意见。3.3 节的这个简单的描述性实例是想表达题目预知对于考生作答速度，并进一步对于作答时间所带来的影响，最后又是如何反馈到三个统计指标上的。可以看出，考生的速度差异越大（出现异常行为的程度更高），三个统计指标的值都更能够有效检测出异常，可以在此基础上进行进一步的实证研究与更复杂的模拟研究。我们对这个内容进行了重新表述，希望能够更准确和清晰地表达含义。

意见五：三个指标在实证研究中的检测结果与“施测机构广泛的调查”结果存在差异的原因是什么？这种差异是否反映了新提出的三个指标缺乏实用价值？

回应：感谢审稿专家的意见，我们针对实证数据进行了更深入的剖析，对分析结果进行了更为广泛的讨论。对于三个指标在实证研究中的检测结果与机构调查结果之间差异的原因进行了分析。

首先，根据实证数据的说明文档，施测机构是在通过汇集多方信息后，对泄漏的题目进行了标注，并对涉及的考生也进行了标注。通过对标注考生在正常题目集和泄漏题目上的速度分析表明，多数考生在泄漏题目上表现出了更快的做题目速度。但是有一部分考生并没有表现出这个特点。我们猜测这可能还跟考生本身的答题风格有关，即这类考生虽然事先获得了题目的信息，但是其在考生过程中仍然会采用相对“保守”的时间策略，因此其并未在时间上明显体现出来。为了验证我们的这一猜想，我们对这部分考生在正常题目集合和泄漏题目集合上的能力进行了估计，得到了如下图的结果。



结果进一步证实了我们的猜想，这部分考生虽然没有显示出速度差异，但是他们显示出了能力差异，将他们在泄漏题目上的能力估计值$\hat{\boldsymbol{\theta}}_1$减去他们在正常题目上的能力估计值$\hat{\boldsymbol{\theta}}_2$，得到能力差异$\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2$的平均值为 0.818，能力差异的区间为[0.05, 2.09]，即基于他们在泄漏题目上的能力参数估计或多或少地都大于在正常题目上的能力参数估计。这里面的原因有可能是考生的作答策略或认知风格所导致的(王超, 2018)，有的考生更倾向于使用更保守的时间策略，或者考生可能有意识地对抗题目预知的影响。这提醒我们在实证数据分析中往往需要综合多种信息来源综合来判断考生的异常和异常类型，对于有意隐藏“异常行为的考生”，基于单一信息来源的判断不足以对他们做出准确的判断。

上述结果的出现并不意味着本文所涉及的三种检测方法没有实用价值，因为统计检验往往都是基于某些假设前提之下做出的，没有任何一种统计方法可以适用所有的场景。如果数据违反了假设，检验的结果肯定会受影响。因此一方面我们需要针对典型异常行为开发相应的检测方法，另一方面我们也需要联合多种数据来源（比如结合作答得分和作答时间），开发适应性更广的方法。

第二轮修改

意见一：引言中需要进一步阐述清楚符号似然比指标的不足，即在实践中使用该指标带来什么具体危害以至于需要提出新的指标；比如，作者所谓的“并不能保证是相同的增加或增加到相同值”，是否有证据支持作者的说法（即对已有方法的批评）？

回应：感谢审稿专家的意见！已有研究中通常认为题目预知对于不同考生的影响程度相同，比如 van Krimpen-Stoop 和 Meijer(2001)及 Sinharay(2016)都认为考生在作答预知题目时会表现出“异常高能力”，将考生在作答预知题目时的能力值增加某个具体值(比如 1 或 2)来体现题目预知的影响；而 Wang 等人(2017)，Sinharay 等人(2017a, 2017b)将考生在预知题目上的正确作答概率设置为固定的 0.9，从而体现考生在预知题目上的高正确率。Zhu 等人(2023)基于作答时间数据，认为考生在作答预知题目时会表现出“异常快的速度”，将考生在作答预知题目时的作答速度添加固定增量的方式来体现题目预知带来的影响。这样的假设在实际的数据中可能并不成立，因为题目预知只能让考生表现出题目正确作答概率或作答速度的增加，但并不能保证是相同的增加或增加到相同值。

实证数据的分析可以进一步支持我们的观点，即题目预知对于不同考生的影响并不相同，将题目预知对考生的影响按固定效应处理的方式是不恰当的，我们对这一点进行了补充说明。

意见二： 研究使用的是单参数 LRTM，如果使用两参数 LRTM 呢？这是否可能是一个研究局限？

回应：感谢审稿专家的意见！这应该是我们没有交代清楚，本研究中采用的是双参数的对数正态作答时间模型，请见正文中的公式 1。每个题目包含两个参数，分别是时间区分度α_i和时间强度β_i，则该模型下考生n在题目i上花费时间的概率密度可以表示为：

$$f(t_i;\tau_n,\alpha_i,\beta_i)=\frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_i-(\beta_i-\tau_n)\right)\right]^2\right\}, \qquad (1)$$

其中，t_{i}是该考生在i题上的作答时间；τ_{n}表示第n位考生的作答速度；β_i是题目的时间强度参数，β_i越大，考生花费在题目i上的时间就越多；α_i是题目的时间区分度参数，α_i越大，意味着第i题的作答时间分布的离散程度越小，该题在不同速度参数水平的人之间的区分性越好。对于这一点我们在正文中进行了补充说明。

意见三：3.3. 一个描述性的实例”，这为什么是“实例”？来自于哪个测验的什么参与者群体？该“实例”并未展现出已有的符号似然比指标的局限，没有引出开展研究的必要性。

回应：感谢审稿专家细致的审稿，非常好的意见！3.3 中的描述性实例并非是真实存在的实际数据，而是我们在研究中为了探究项目预知对作答时间的影响，以及其在三个统计指标上的不同反馈所模拟出的一个简单化的数据例子，主要的目的是让读者能够直接观察从考生作

答时间数据上的异常变化反馈到统计量上的变化，让读者对统计量有更直观的理解。我们对这一点进行了明确阐明。

针对专家所提到的“实例”可能会引发歧义的问题，我们考虑将“3.3 一个描述性的实例”改为“3.3 一个描述性的模拟例子”。模拟例子的结果表明当考生在测验的不同部分存在速度差异时，上面的三个统计量都能够给出反馈，并且随着异常程度的增大，有更大可能支持做出考生存在异常的判断。

由于这个例子中的数据比较“理想”或者说是“噪音比较明显”的数据，它只是为了描述三个统计量的使用，虽然分析结果没有体现出在三种统计量上的差异，但分析的结果初步表明提出的基于贝叶斯因子和后验概率的统计量在这个数据中的分析和经典的似然比统计量同样有效。而针对符号似然比指标的局限以及三种统计指标间的差异会在更复杂全面的模拟实验中展示。

意见四：文中一些解释用词还需严谨。比如，对表 3 的解释，“pp 法在判定考生是否异常时‘最严格‘，而 BF 相对’最宽松‘”，严格更好？还是宽松更好？再比如，讨论中提及的“认知风格和作答策略”，作者已经开展相关研究表明纳入作答精度后就能够提高不同认知风格和作答策略考生在题目预知上的检验力？如果没有，避免这种毫无根据的话语。
回应：感谢审稿专家的意见，非常好的意见！关于考生是否存在题目预知的判定标准，到底是“严格”更好，还是“宽松”更好，这没有统一的界定，可能还需要结合测验的性质和目的、及其它指标综合来考虑，因为这些指标都只能对考生的作答行为作出统计上的判断，如果用于实际的测评情境，给任何考生做出存在题目预知的决策需要谨慎，本文中涉及到的方法可以得到需要重点关注的考生，还要把这些考生的其它更多的信息（比如其过往的成绩，作答过程中的其它行为数据等）结合起来综合考虑，才能最终得出结论。

关于讨论中提到的“认知风格和作答策略”，主要是基于在实证数据上的分析结果，因为“检测出”的多数考生体现了“速度加快”和“正确作答概率升高”，但是也有少数“检测出”的考生没有体现出明显的速度变化，我们初步认为这部分考生的作答时间可能还受到其“认知风格和作答策略”的影响，导致结果与我们的预期存在偏差，这个考虑还有待于我们在未来的研究中进行探索。我们对讨论中涉及的有关内容进行了修改，以使表达更准确。

对于专家所提到的问题我们做了严格的自检，优化了一些不够严谨的词语表达并增加了对于部分表达的阐释。

意见五：最后一个问题也想听作者的想法，在这种异常作答检测研究中，尤其是实证研究中（比如本文中的数据），各种检验方法均会得到不同的检验结果，那么到底哪个是正确

的？依据什么来判断“正确”？如果没有“正确”标准，那所谓的各种方法不就是自说自话？

回应：感谢审稿专家，非常好的问题！在这里首先对实证研究的数据作个补充说明，正文中使用的这批实证数据是由大型测评公司收集，并结合考生、考点的多方面信息，对数据中的“泄漏的题目”，“受到影响的题目预知考生”打了标签，这批数据由于存在“标签”，并且质量也比较高，因些，很多有关的研究都会对这批数据进行分析，并将分析的结果与数据中的标签进行比较。当然需要说明的是，原数据中的标签并不一定是“真值”，所以它也只能做为对比的参考。但是它至少缩小了需要重点考察的考生范围。在开展的以往研究中，包括我们的研究，都是将三个统计量检测出的结果与原数据中的标签进行比较，以上来判断不同统计量的检测结果。

关于“标准”的问题，专家提出了一个非常关键的问题，这里我们简要谈谈我们的理解。正如上一个问题中的回答，对于考生是否存在异常的判断，并不存在统一的界定标准，采用不同的检验方法会得到不同的结果，将结果与数据中预先存在的“标签”（如果是实证研究，则标签是通过其它途径得到，如果是模拟研究，则标签则可以生成）进行比较，由此各检验方法也就产生了“更稳健”或“更敏感”的特性。一切的检验方法或理论指标最终都要为实践所服务。实际的操作者需要根据测验的性质和他的研究目的来选择合适的方法，很可能是多种方法的综合。

第三轮修改

意见一：对意见 5 的回答没有体现在正文讨论中；且回答仍不够有说服力。如果机构能够根据非方法学手段对被试的异常作答打标签，那么为什么需要本研究，为什么需要这种异常检测研究？

回应：非常好的意见，感谢审稿专家！我们在全文的讨论部分补充了阐明本文研究的作用和意义。

本文的实证数据来自基于计算机施测的职业认证测试，随著作《Handbook of Detecting Cheating on Tests》一起对外公布。由于这个数据的质量相对较高，包含的信息也比较完整，它被很多研究所分析。正如著作的作者之一 Jim Wollack 教授所提到的“被打上标签的异常作答行为者是通过多种手段来确定的”，这个标签可以作为后续研究者进行异常行为检测时的参考对象。

但是需要重点说明的是（1）这个实证数据中的标签可以用来与各种统计方法的检测结果进行比较，用来评价统计方法的检测结果。

（2）正如数据收集者所强调的，被打上标签的考生有很大可能存在异常行为（题目预知是其中的一类），即这些考生并没有明确全部都属于“题目预知”类型，这也可能是很多统计方法检测的结果与原数据中的标签结果存在较大差异的一个原因。

（3）采用非方法学的方法检测考生作答行为存在的最大问题是效率低，在很多时候（尤其是大规模测验中）不容易实施，并且在高风险测验中，将考生标记为异常作答行为是需要非常慎重的操作，因此通常是需要将多种方法结合起来使用，综合多方面的证据才能做出判断。因此，新的统计方法和一些非方法学的方法并不是二选一的存在，在实际的应用中是需要综合起来考虑。

（4）由于实际的测验数据并不包含考生作答行为的标签，基于本文所提出方法在模拟和实证数据中的表现，表明它们是可以在其它的实际测验数据中使用，可以为研究者提供有用的参考信息的。

综合来看，在实际的应用中，多数情况下通过单一的证据或方法来判定考生是否出现作弊行为可能是不够严谨的，尤其是在高风险测验中，需要采用多种方法来进行综合判断与检测。因此，本研究构建了新的研究方法，提供了新的检测手段，并且研究结果表明新方法在考生存在速度差异的检测上更为敏感。

意见二： 进一步补充国内相关参考文献，比如，对各类异常作答检测的研究、对作答时间开展研究的研究等；考虑到篇幅问题，国内几本主流心理学期刊的近 5 年研究还是要囊括的。

回应：感谢审稿专家的意见！我们对参考文献进行了进一步的引用和补充，希望能给读者一个更全面的介绍，有利于阅读和理解本文的研究问题和研究思路。