《心理学报》审稿意见与作者回应

题目：基于作答时间的改变点分析在检测加速作答中的探索与研究
作者：钟小缘, 喻晓锋, 苗莹, 秦春影, 彭亚风, 童昊

第一轮

审稿人 1 意见：

意见 1：引言部分，问题引入慢，比如，第一段内容似乎并没有存在的必要性，文献回顾占篇幅过多；其实就是已有研究缺乏基于题目作答时间的 CPA 研究；
回应：第一段确实比较冗余。我们已对引言部分的内容进行了精简，并删除了一些内容。调整后的内容呈现在修改稿的 1-2 页。

意见 2：方法部分，公式 20 前一段提及“本研究中 Wald 检验是单侧检验，检测速度参数 τ 是否增大”，相当于本研究假设所提出方法适用于高风险测验或其他会导致被试加速的测验，请在文中适当位置说明；
回应：加速作答行为通常发生在有时间限制的考试中。当考试在中后期时，未完成作答的考生由于受到时间因素的影响会倾向于提高自己的答题速度，出现加速作答。

本研究通过检测考生在整个考试过程中的特质参数(即速度参数)是否发生显著性的变化来判断数据中是否存在改变点也就是考生是否出现加速作答。Wald 检验适用于双侧检验，但由于本研究的检测目标明确，也就是检测加速度作答的考生。当考生出现加速作答时，我们期望考生的答题速度会大于正常答题时的速度。因此这里采用单侧检验，当改变点k已知时 Wald 检验的虚无假设为考生i在前k个项目上的速度参数等于后$(J-k)$个项目上的速度参数，即$\hat{\tau}_{i,k-}=\hat{\tau}_{i,k+}$。备择假设为生$i$在前$k$个项目上的速度参数小于后$(J-k)$个项目上的速度参数，即$\hat{\tau}_{i,k-}<\hat{\tau}_{i,k+}$。

我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 7 页和第 9 页。

意见 3：在两种检验方法提出后，建议增加对两者的理论对比，比如为何要提出两种方法；
回应：常用的 CPA 统计量大致有四种，分别是基于似然比检验的统计量$(\nabla l_i, L_{max}, L_S)$，基于 Wald 检验的统计量$(W_{max}, W_i^{(j)})$，基于得分检验(score test)的统计量$(R_S, S_{max})$和基于残差检验的统计量$(R_{max})$。在 4 种统计量中， 前三种统计量都是通过检验虚无假设(考生的潜在特质没有发生显著的变化)来判断考生是否存在异常作答行为；残差统计量R_{max}对得分计算加权残差来检验考生在测验前期或后期是否存在异常作答行为。前三种检验量更适合用于高风险、大规模的教育测验，R_{max}更适合用于低风险的心理测验(张龙飞等，2020)。本研究是基于作答时间数据检测异常作答，这里选用基于似然比检验和 Wald 检验的统计量。

本研究聚焦加速作答行为的检测，故在研究中使用的是单侧检验。

我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 7 页。

意见 4：建议调整文章撰写逻辑，比如 3.1 和 3.4 应与本文新提出的两种检验放在不同的章节下，其中，3.4 应放在模拟研究数据生成部分；公式 24 是否应放在分析部分；等等

回应：非常好的建议，我们对文章的结构进行了调整，大致为 1 引言-2 改变点分析 CPA 技术-3 基于 CPA 的统计量-4 基于加速作答行为的作答时间模型-5 模拟研究-6 结论和讨论

本文提出的检验统计量放在修改稿的第 3 部分(详见第 7-9 页)，3.1 作答时间模型与 3.4 基于加速作答的作答时间模型放在修改稿的第 4 部分(详见第 9-10 页)，公式 24(修改稿中为公式 23)为ADL的计算公式，放在修改稿中第五部分的评价指标部分(详见第 12 页)。

意见 5：我不太理解公式 23 中为何还有 L；如何过没有 L，则没有必要呈现公式 21；

回应：十分抱歉这里存在笔误。公式 23 中不包含 L，更正后的公式为：

$$\ln(t_{ij}) = (\beta_j - \tau_i + \varepsilon_{ij}) \times \min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i}, \varepsilon_{ij} \sim N(0, \alpha_j^{-2})$$

我们已将上述修改更新到公式 20，并删除了公式 21。详见修改稿的第 10 页。

意见 6：第 11 页“由于...几乎相同，表 2 中只给出了似然比检验统计量的...”，而实证研究中说“这两个统计量的结果非常接近,...,只给出基于 Wald 检验统计量的结果”，这么随意吗？为何不统一？

回应：在模拟研究中似然比检验和 Wald 检验都是单侧检验，结果显示两种方法的检验力和 I 类错误率都非常接近，故为了节省篇幅模拟研究只呈现了似然比的结果。在实证研究中，研究者想为读者呈现另一种方法的检测效果。

意见 7：本研究未涉及“实验”，不需要用“实验结果”；

回应：我们将“实验结果”更改为“模拟研究结果”。详见修改稿第 12 页。

意见 8：基于图 3 无法判断是考生原因导致了 RT 下降还是因为题目原因导致的（比如后面题目对作答时间的要求本来就少）；图 3 和图 4 提供的信息重复，保留图 4 即可；

回应：非常好的建议。我们已对图 3 进行了删除，对图 4 的内容进行了补充，调整后的图 4 被标记为图 3，详见修改稿第 17 页。

意见 9：另外仅一名学生的结果为例似乎不够充分；建议作者将所有检测出的被试的作答时间均值和全体被试的作答时间的均值进行对比；尽管不同学生的改变点可能不同，但如果检测都有效，那么这些人的平均 RT 还是应该小于全体均值的。另外如果这些人的平均 RT 存在规律性，可以进一步有助于了解数据特性。

回应：我们对所有检测出的考生的作答时间均值和全体考生的作答时间的均值进行了对比，结果显示检测出的异常考生在测验中后期的平均项目作答时间是小于全体考生的平均项目作答时间的。这一结果是与猜想相符的。

我们将检测结果补充至修改稿的第 17 页。

意见 10：其他建议 (1)引言中明确本研究所要关注的“异常作答”具体是那种，所有的异常作答行为都可以使用该方法吗？

回应：本研究关注的异常作答行为是加速作答，基于作答时间数据使用 CPA 方法检测由加速作答行为造成的异常作答模式。CPA 方法本质上是检测异常数据的方法，因此它同样可用于检测由其他异常作答行为如题目预知，热身效应等造成的异常作答模式。本研究以加速作答为例是为了让读者更好地理解 CPA 方法的使用过程。

我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 2 页。

意见 11：文章整体不够简洁，除了两个新检测方法外还引入了一些细小的点

回应：我们已对文章内容进行了调整和删减。除了介绍两种检验统计量外，本研究还对比了基于蒙特卡洛模拟生成的经验临界值和 Sinharay (2016)所采用的近似临界值，目的是考察当改变点不是在接近测验中间的位置时，Sinharay (2016)所采用的近似临界值是否可以直接应用。

经验临界值与近似临界值的对比结果呈现在文章修改稿的模拟研究结果部分，详见修改稿的第 13 页。

意见 12：参考文献覆盖不全，国内没有学者做 RT 相关的研究吗？如果国内没有学者关注相关议题，那这篇文章写来给谁看，有何发表意义？

回应：我们已重新对国内相关文献进行检索，发现詹沛达(2019); 詹沛达, Jiao Hong, Man Kaiwen (2020)；张龙飞 等(2020) 等围绕 RT 进行了有关的研究。已在引言部分引用了国内相关的文献，详见修改稿第 1-2 页。

意见 13：无论是模拟研究还是实证研究，似乎都没有得出两种方法之间差异性的结果；外加作者也没有理论阐述两者的异同，因此，审稿人不清楚为何要提出两种方法，且实践应用者应该选用哪个呢？

回应：从表 1 中可知，常用的 CPA 统计量大致有四种，分别是基于似然比检验的统计量$(\nabla l_i, L_{max}, L_S)$，基于 Wald 检验的统计量$(W_{max}, W_i^{(j)})$，基于得分检验的统计量$(R_S, S_{max})$和基于残差检验的统计量$(R_{max})$。在 4 种统计量中，前三种统计量都是通过检验虚无假设(考生的潜在特质没有发生显著的变化)来判断考生是否存在异常作答行为；基于残差检验的统计量R_{max}可直接检验考生在测验前期或后期是否存在异常作答行为。前三种检验量更适合用于高风险、大规模的教育测验，R_{max}更适合用于低风险的心理测验(张龙飞等，2020)。本研究是基于作答时间数据检测异常作答，这里选用基于似然比检验和 Wald 检验的统计量。

本研究关注的异常作答行为是加速作答行为，因此在研究中使用单侧检验。我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 7 页。

意见 14：研究局限阐述不全面，比如所提出方法仅局限于单维测验，是否可用于多维测验？Van der Linden (2006)的模型并没有考虑加工速度的区分度参数也未考虑考生自然而然可能存在的变速(Fox & Marianti, 2016; doi: 10.1080/00273171.2016.1171128)等

回应：使用 CPA 方法检测多维测验下的异常作答反应是一项具有可行性和价值的研究。

首先多维测验多维量表的开发是当前趋势所在，例如在基于英文语言的数学测验中，每道题上同时考察英语与数学两个维度的能力，如果某考生存在加速作答，那么在加速之后，对应于其英语和数学能力都可能降低(张龙飞 等, 2020)。

另外，现在 RT 的多维模型也越来越多，例如詹沛达, Jiao Hong, Man Kaiwen (2020)开发了多维对数正态作答时间模型，其研究表明在多维测验中，潜在加工速度具有与潜在能力相匹配的多维结构。并在模拟研究中实现了对被试的潜在加工速度的估计。

我们拟在未来进行进一步的研究，将此部分内容加入到修改稿中，详见修改稿第 18 页。

意见 15：在正常测验中，加速真的会带来什么问题吗？这种情况难道不是测验设计者所设计出的吗？既然是设计出的，为何要检测？

回应：加速不是我们设计出来的，它是心理和教育测验中存在的由来已久的问题，由于它会扭曲模型的估计参数，甚至对测验的信效度带来严重的负面影响，有很多研究者都对此展开了研究(Oshima, 1994; Shao, 2016; Shao et al., 2016; Suh et al., 2012; Wollack et al., 2004 等)。对加速作答行为进行检测，有很多作用，下面我们列出几条：(1)一般来说，举办测验的目的都是为了准确评估考生的能力。基于考生认真且独立完成作答时产生的数据评估出的能力更能代表考生真正的能力。当考生出现加速作答时，考生的能力明显下降，这时产生的作答数据就很难代表考生真实的能力。对考生的加速作答进行检测，以及处理相关的作答数据，有利于我们更加精确地评估考生的能力。

(2)一般来说，考试时间的设置都是经过深入分析的研究而设置的。对加速作答行为进行检测，有利于我们判断测验的时间设计是否合理。例如当测验中出现加速作答的考生数量很多时，我们就要去思考测验长度，难度和测验时间设置的是否合理，是否需要对测验设计进行优化。

(3)多项研究证明考生的加速作答行为会对项目参数的估计产生负面影响(Oshima, 1994;Bolt, Cohen, & Wollack, 2002; Suh, Cho, & Wollack, 2012),检测加速作答行为有利于对项目参数进行校准，提高参数估计的精度。

意见 16：检测出这些加速的考生之后又怎样？让他们重新考试吗？取消成绩吗？
回应：非常好的问题。当我们检测出加速作答的考生，通常会将考生在加速作答行为下产生的数据进行处理，比如 list-wise deletion 或 partial deletion 等 (Patton, Cheng, Hong, & Diao, 2019; Shao, Li, & Cheng, 2016)，这样我们可以得到一个质量更高的数据，进行后续的分析(如参数估计等)。

意见 17：本研究中依旧是基于所有人的作答（即把所有人都视为正常被试）来进行参数估计，然后基于参数估计结果再马后炮找到加速的被试。这样的逻辑不奇怪吗？这些加速被试的作答数据不会影响之前的参数估计吗？
回应：正如您所考虑的那样，考生加速作答时产生的数据会对作答数据造成污染，基于这些数据对参数进行估计以及检测加速作答的考生，结果并不会那么的纯粹。因此在检测过程中，当 CPA 检测出第一批加速作答的考生后，可以对这批考生加速作答时产生的数据进行处理（Patton et al., 2019），处理之后 重新进行参数估计以及加速作答行为的检测。重复上述过程，直到 CPA 方法没有检测到加速作答的考生。通过这种方法提高项目参数和能力估计的精度等。

意见 18：请勿将 response time 与 reaction time 相混淆，建议参考詹沛达等(2020；10.3724/SP.J.1041.2020.01132)的翻译“作答时间”；
回应：我们已对自检报告以及文章中 94 处相应名词进行修改，全部更改为“作答时间”。

意见 19：第一句话将外国作者姓名进行翻译的做法似乎是一种很古老的做法，建议直接使用英文原名；
回应：我们已对文章的引用格式进行了检查并修改。
……

审稿人 2 意见：

意见 1：题目文章提出的方法是针对加速作答行为的识别，其实，加速作答只是异常作答模

式中的一种，例如，异常作答还应包括被试疲劳之后反应时增加的作答。因此，这里是否能概括为“检测异常作答模式”还希望作者推敲。

回应：本研究的目的是为了阐述如何基于作答时间数据检测测验中的异常作答行为。本研究中我们考虑的异常作答行为是加速作答，这样能让读者可以在具体的情境下更好地理解 CPA 的原理及其检测异常作答数据的过程。

CPA 本质上是检测异常数据的方法，也就是说 CPA 方法的功能是检测异常作答模式，不论异常作答模式出现的原因是热身效应、题目预知或者其他等等。因此它也可以应用于其他异常作答行为引起的异常作答模式。

我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 2 页。

意见 2：引言（1）引言第 1、2 自然段建议简写，表述可以更精炼一些。因为这些内容只是为了引出改变点分析这一概念，与文章主题相关不大。

回应：我们已对第 1,2 自然段以及引言部分的其他内容进行了精简，调整后的内容呈现修改稿的 1-2 页。

意见 3：第 2 页，第 2 自然段中提到“答题能力显著下降(突然下降和逐渐下降)”，其实已有模型也可以分为基于突然下降的假设和基于逐渐下降的假设两类。然而在本段最后，作者仅回顾了逐渐下降的文献，没有列出关于突然下降的假设。并且，最后一句话“热身效应(warm-up effect; Shao et al., 2016)，预知试题(item pre-knowledge, Wang & Liu, 2020; Zhang, 2014; Choe, Zhang, & Chang, 2018)和粗心作答(Patton, Cheng, Hong, & Diao, 2019; Yu & Cheng, 2019)等。”不是一个完整的句子。

回应：Yu 等(2019)曾回顾了加速作答考生可能存在的两种潜在加速作答机制，即作答速度突变和作答速度逐渐改变，并使用两种模型来表示这两种作答机制，分别是混合模型(the hybrid model, HM)和逐渐变化模型(the graduate change model, GCM)。HM 假设考生出现加速作答时的作答速度会发生突变；而 GCM 认为每位考生有自己独特的加速点，并且考生在加速点之后的题目上的答对概率会逐渐下降。

由于对引言部分内容进行了删减，我们将加速作答模型相关的内容调整至修改稿第 4 部分(基于加速作答行为的作答时间模型),详见修改稿第 9 页。

第 2 页第 2 自然段中的最后一句话，我们将它改为测验过程中常见的异常作答行为有热身效应、加速作答、题目预知等等(张龙飞等, 2020)。详见修改稿第 1 页。

意见 4：第 2 页第 3 自然段，“因此检测测验中是否存在改变点或者异常作答行为是非常重要和关键的。”这句话逻辑不太顺，应当是检测是否存在改变点，进而是否表示存在异常作答行为。

回应：由于对引言部分的内容进行了调整，我们将上述内容修改为：

考生出现异常作答行为后，考生的作答数据称为异常作答数据或异常作答模式，它与正常作答时的数据有着显著的不同。测验数据中包含异常作答数据会降低其自身及整体测验数据的质量，从而对后续的分析结果产生一系列的不良影响，例如造成模型与数据的失拟、被试与题目参数估计的扭曲(Stefan, Dietrich, Wolfgang, & Michael, 2016)，影响考试的信度和效度(Guo, Tay, & Drasgow, 2009)等等。因此，检测测验中的异常作答行为或异常作答数据是非常重要和关键的，研究者们也一直在寻找相关的解决方法(e.g., Bejar, 1985; Evans & Reilly, 1972; Shao, Li, & Cheng, 2016; Bradlow, Weiss, & Cho, 1998; McLeod, Lewis, & Thissen, 2003; Wise & Kong, 2005; Yu & Cheng, 2019, 2020)。

详见修改稿第 1 页。

意见 5：改变点 CPA 分析技术（1）第 4 页第 16 行，“题目信息量”应为“题目信息量之和”更加准确？
回应：我们已将“题目信息量”改为“题目信息量之和”，详见修改稿第 4 页。

意见 6：第 4 页第 23 行，第一次出现 3PLM 的缩写应标明含义。
回应：我们已将 3PLM 改为三参数 Logistic 模型(3PLM; Birnbaum,1968)。详见修改稿第 4 页。

意见 7：建议作者将已有的 CPA 方法按照原理分类，例如，是否使用似然比，是否加权（信息量加权，残差加权）等。然后在每个类别里介绍一个有代表性的方法。同时，总结不同方法适用的情境。根据以上内容可以丰富表 1，将方法按原理归到不同类别，并列出方法的适用条件等，将参考文献放到最后一列，这样对读者来说提供的信息更有用，更容易加工。
回应：由于整理后的内容较多，表格中不好呈现，我们将总结的内容放在了修改稿第 3 节“基于 CPA 的统计量”中，具体如下：

从表 1 中可知，常用的 CPA 统计量大致有四种，分别是基于似然比检验的统计量$(\nabla l_i, L_{max}, L_S)$，基于 Wald 检验的统计量$(W_{max}, W_i^{(j)})$，基于得分检验(score test)的统计量$(R_S, S_{max})$和基于残差检验的统计量$(R_{max})$。在 4 种统计量中， 前三种统计量都是通过检验虚无假设(考生的潜在特质没有发生显著的变化)来判断考生是否存在异常作答行为；基于残差检验的统计量R_{max}可直接检验考生在测验前期或后期是否存在异常作答行为。前三种检验量更适合用于高风险、大规模的教育测验，R_{max}更适合用于低风险的心理测验(张龙飞等，2020)。本研究是基于作答时间数据检测异常作答，这里选用基于似然比检验和 Wald 检验的统计量。

详见修改稿第 7 页。

意见 8：第 5 页最后一个自然段，作者认为“采用 CPA 分析心理测量数据，还没有基于反应时的尝试”。其实，有一些其他利用反应进行序列分析的方法，和 CPA 很类似。例如，在 CAT 中识别从哪一名被试开始，出现了题目泄露的情况。可参见 Choe, E. M., Zhang, J., & Chang, H.H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. Psychometrika, 83(3), 650-673.
回应：我们已在修改稿中的第二部分“改变点分析 CPA 技术”中补充了相关文献。具体内容为：

表 1 中虽然没有基于作答时间数据采用 CPA 方法进行检测的研究，但已有类似研究基于作答时间数据检测测验中的异常项目，例如 Choe, Zhang 和 Chang(2018)使用序列分析方法分别基于作答数据、作答时间数据以及结合两种数据对泄露试题进行检测。该研究结果证明了在相同的 I 类错误率情况下，方法的检验力呈现：(1)仅基于作答时间数据的检验力要比仅使用作答数据的检验力高得多；(2)结合两种数据对泄露试题进行检测的方法有两种，其中一种方法的检验力略大于仅基于作答时间数据的检验力，第二种方法的检验力小于仅基于作答时间数据。并且基于作答时间数据进行检测的探测点的识别延迟是 Choe 等(2018)所有方法中最小的。

从 Choe 等(2018)的研究结果上可以发现相比于作答数据，作答时间数据的确可以提供更多的测验信息，从而使检验力有实质性的提高。因此基于作答时间数据检测异常作答行为是具有非常好的研究前景的。

详见修改稿第 5 页。

意见 9：第 5 页最后一个自然段，作者在阐述基于反应时识别比作答得分数据拥有先天的优势时举例，“比如从得分上不容易判断考生是否事先预知题目的信息，但是从反应时数据上则容易判断这一点”，这里的例子可以换成加速作答的，更加贴合文章主题。

回应：我们已经文中的例子改为加速作答，具体内容为：

作答时间数据的获取已经越来越容易，并且作答时间数据在检测考生的异常作答行为上比作答得分数据拥有先天的优势。比如从同样是得 80 分的考生，相对于得分数据，结合作答时间数据则更容易判断考生的异常作答作为。

详见修改稿第 5 页。

意见 10：第 5 页最后一个自然段，提到了加速作答的危害，应当放在引言部分，说明为什么要采用各种 CPA 的方法去识别加速作答。

回应：非常好的建议。我们已将这部分内容调整至文章的引言部分，具体内容如下：

由于加速作答是众多异常作答行为中最常见和普遍的(Goegebeur, De Boeck, Wollack, & Cohen, 2008)，对于测验数据质量有非常大的负面影响，受到很多研究者的关注 (比如 Bolt, Cohen, & Wollack, 2002; Oshima, 1994; Suh, Cho, & Wollack, 2012; Yu et al., 2020 等)。因此本研究拟聚焦于基于作答时间数据使用 CPA 方法检测由加速作答行为造成的异常作答模式。

详见修改稿第 2 页。

意见 11：第 5 页最后一个自然段，应当阐述为什么使用反应时数据识别异常作答具有优势。例如，反应时是连续变量，相比于二分变量，能够提供更多信息（后文提到过）。其次，反应时对于测验设计效应不敏感，不会受题目难度等影响，基于反应时识别异常作答指标的检验力在整个测验中几乎能保持稳定。

回应：非常好的建议。由于对文章内容进行调整和删减，我们分别在引言部分和第 5 页中阐述了使用作答时间数据识别异常作答的优势，具体内容如下：

作答时间数据是一种连续数据，同时包含了考生能力信息和题目信息(Marianti, Fox, Avetisyan, & veldkamp, 2014)，对于提高考生能力估计的精度与优化测验设计有很大的帮助；如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。

详见修改稿第 2 页。

如今作答时间数据的获取已经越来越容易，并且作答时间数据在检测考生的异常作答行为上比作答得分数据拥有先天的优势。比如从同样是得 80 分的考生，相对于得分数据，基于作答时间数据则更容易判断考生是否出现了异常作答行为。因此基于作答时间数据检测异常作答行为是具有非常好的研究前景的。

详见修改稿第 5 页。

意见 12：反应时模型与 CPA 检测加速作答（1）3.1 反应时模型下的第 1 自然段应当放在前面，和第 5 页最后一段整合在一起，说明反应时在异常作答识别中的应用。

回应：我们已对此部分内容进行了整合，具体内容如下：

作答时间数据是一种连续数据，同时包含了考生能力信息和题目信息(Marianti, Fox, Avetisyan, & veldkamp, 2014)，对于提高考生能力估计的精度与测验设计的优质性有很大的帮助；如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利。作答时间数据逐渐获得国内学者的关注。例如，詹沛达(2019)和詹沛达，Jiao Hong, Man Kaiwen (2020)分别提出了关于作答时间数据的多维模型，研究结果皆显示引入作答时间数据可提高或精确估计模型的参数等等。

详见修改稿第 2 页。

意见 13：第 7 页第 19 行，“有两类方法来解决这个问题”，应当引用具体方法及文献，例如“将异常行为作为模型参数，通过参数去体现考生的作答行为”，可以举 mixture model(Wang & Xu, 2015)的例子等。

回应：非常感谢您的建议。由于文章内容调整，我们在修改稿中删除了这句话所在的自然段，在 CPA 统计量一节中直接提出本研究是通过检测考生在整个考试过程中的特质参数(即速度参数)是否发生显著性的变化来判断数据中是否存在改变点，也就是考生是否出现加速作答的。考生出现加速作答行为后，考生的答题速度会增大。当改变点k已知时，两个统计量的虚无假设为考生i在前k个题目上的速度参数等于后$(J-k)$个题目上的速度参数，即$\hat{\tau}_{i,k-}=\hat{\tau}_{i,k+}$。备择假设为考生$i$在前$k$个题目上的速度参数小于后$(J-k)$个题目上的速度参数，即$\hat{\tau}_{i,k-}<\hat{\tau}_{i,k+}$。

具体内容详见修改稿第 7 页。我们对文章的其他部分进行了检查。

意见 14：第 8 页说明本研究获得零分布和临界值的方法是基于模拟的方法，这部分应当综合总结前面几种获得临界值方法的缺陷，再提出本研究方法的优势。

回应：我们将此部分调整至“CPA 统计量临界值的获取”一节中，具体内容如下：

参考表 1 中的信息，统计量临界值可通过置换分布、经验临界值和近似临界值获得。由于置换分布方法的计算量非常大，需要很长的时间获得临界值；而近似临界值比较适合改变点出现在测验中间位置(比如中间 70%的位置)的情况(Sinharay, 2016)。在实际情况中加速作答更容易出现在测验中后期阶段，因为考生在测验的中后期阶段更容易感受到时间的压力。因此本研究采用经验临界值。

详见修改稿第 8-9 页。

意见 15：第 9 页，“渐进临界值”是否就是“近似临界值”，建议统一术语。

回应： “渐进临界值”就是“近似临界值”，我们已经文章中的相关术语进行检查和更改，统一为“近似临界值”。

意见 16：“3.4 基于加速作答行为的反应时模型”其实是在讲数据生成方法，应当放在模拟研究中介绍。

回应：由于文章内容的调整，修改稿的文章结构大致为“1 引言-2 改变点分析 CPA 技术-3 基于 CPA 的统计量-4 基于加速作答行为的作答时间模型-5 模拟研究-6 结论和讨论”。我们将“基于加速作答行为的反应时模型”单独设置了一节，主要阐述模型构建的思路，在“5 模拟研究”中的增加数据生成部分阐述数据的生成方法。具体内容详见修改稿第 9-11 页。

意见 17：第 9 页，公式 21 下面，“L 表示由于加速作答增加导致作答速度增加的部分”，（第一个“增加”应去掉），如果要直接体现导致速度增加，应把公式改为 β_j-(τ_i+L)+ε_ij，如果使用现在的公式，可以改为“L 表示由于加速作答导致期望反应时减少的部分”。

回应：由于本研究中构建的模型没有使用到“L”，为了精简内容，我们在修改稿中删除了此公式，并将内容改为：

为模拟考生加速作答行为下的作答时间，以往的研究提出了两种方法，第一种是将考生在加速作答行为下的作答时间设置为固定的几个水平，比如 10s, 20s, 30s(van der Linden & Guo, 2008)；第二种是在对数正态作答时间模型的参数τ_i上增加一个正数L，表示加速作答对考生答题速度产生的影响。在 van der Linden 和 van

Krimpen-Stoop (2003)的研究中，L被设置为.375 和.750。

具体内容详见修改稿第 9 页。再次感谢您的提醒，我们对文章中的其他公式进行了检查。

意见 18：第 9 页 24-25 行，与固定效应相对的应当是所有考生加速行为表现出来的反应时减少是不一样的，而与作答速度“逐渐改变”相对的应该是突然改变的模型。

回应：非常感谢您的建议。加速作答考生可能存在两种潜在的加速作答机制，即作答速度突变和作答速度逐渐改变，这两种作答机制可用两种模型来进行表示，即混合模型(the hybrid model, HM)和逐渐变化模型(the graduate change model, GCM)。在 HM 中模型假设考生出现加速作答时的作答速度会发生突变；而 GCM 认为每位考生有自己独特的加速点，并且考生在加速点之后的题目上的答对概率会逐渐下降。

以往的研究中常把加速作答设置成固定效应，即所有加速作答的考生都会出现相同的作答时间或受到相同大小的影响，这其实不太符合实际情况。因此在本研究中，我们拟采用更可能出现的作答速度“逐渐改变”的方式(对应逐渐改变模型)来模拟数据。

我们已将上述观点和内容补充到修改稿中，详见修改稿第 9-10 页。

意见 19：公式 23 是否不应当有 L。当 j/J≤η_i，没有到加速点，此时应当为正常反应时公式？

回应：这里存在笔误。公式 23 中不包含 L，更正后的公式为：

$$\ln(t_{ij}) = (\beta_j - \tau_i + \varepsilon_{ij}) \times \min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i}, \varepsilon_{ij} \sim N(0, \alpha_j^{-2})$$

当测验没有进行到η_i所表示的阶段时$\frac{j}{J}$将小于η_i，故而$\left[1 - \left(\frac{j}{J} - \eta_i\right)\right]$的值会大于 1，$\min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i} = 1$，这说明考生的作答时间依旧使用对数正态作答时间模型模拟。当测验进行到η_i所表示的阶段时，即$\frac{j}{J} > \eta_i$，则$\min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i}$小于 1；此时$ln(t_{ij})$的值将小于其正常的作答时间，表示考生在$\eta_i$所表示的阶段上出现了异常的加速作答行为。

我们已将上述公式更新至修改稿中的公式 20。内容详见修改稿的第 10 页。

意见 20：研究设计（1）模拟条件的设置是否参考了前人研究，请列出。

回应：本研究的模拟条件参考了 Shao(2016)和 Yu &Cheng(2020)的设置。

Shao, C. (2016). *Aberrant response detection using changepoint analysis* (Doctoral dissertation). University of Notre Dame.

Yu, X., & Cheng, Y. (2020). A comprehensive review and comparison of CUSUM and Change-Point-Analysis methods to detect test speededness. *Multivariate behavioral research*, 1–22. Advance online publication.

意见 21：第 10 页，“假定 η 服务贝塔分布”，此处“服从”应为笔误，应记为 beta 分布，且列出具体分布的表达式。

回应：这里存在笔误。我们已在修改稿相应位置进行修改，具体内容为：改变点位置η_i按照 Shao 等(2016)的处理，即假定η_i服从 beta 分布，改变点的中值为 0.6 和 0.7，方差为$\sigma_{\eta_i}^2 = 0.001$和 0.04 共四种情况。对应的 beta 分布具体形式为：beta(143.367, 95.689)，beta(2.970, 2.091)，beta(146.345, 62.910)和 beta(3.033, 1.490)。详见修改稿第 11 页。

意见 22：图 1 中应列出不同的模拟条件。

回应：我们已在模拟研究设计部分增加“表 2 模拟条件”。表格如下

表 2 模拟条件

因素	水平
考生数量	1000
测验长度	40,60,80
加速作答考生的比例	10%,20%,30%
改变点的位置参数η_i	Median (0.6,0.7)×σ_η^2(0.04,0.001)

详见修改稿第 11 页。

意见 23：对于不同的临界值，我们除了关心数值差异，更关心基于不同临界值下得到的结果，其第一类错误率和检验力如何？目前模拟研究结果显得较为单薄，可以适当补充。

回应：模拟研究中 36 种模拟条件下的 I 类错误率和检验力呈现在表 4 中，详见修改稿第 15 页。

意见 24：作者在“4 研究设计”下又包括了实验结果，建议改为“4 模拟研究”。

回应：我们已将“研究设计”改为“模拟研究”。详见修改稿第 10 页。

意见 25：建议在研究设计部分加入评价指标，包括检验力和第 I 类错误率，绝对的延迟(ADL)等。

回应：我们已在模拟研究部分增加“5.3 评价指标”一节内容，具体内容如下：

使用 I 类错误率和检验力评价 CPA 方法的性能，I 类错误率和检验力的最终结果为每种条件下的均值。并计算在给定时间内未完成测验的学生比例($\%NF$)以及检测到的改变点位置与真实改变点位置之间绝对的延迟(absolute detection lag, ADL)指标的均值和标准差。I 类错误率、检验力和ADL的计算公式分别如下

$$\text{I 类错误率} = \frac{\text{错误标记加速考生的数量}}{\text{正常考生的总数}} \tag{21}$$

$$\text{检验力} = \frac{\text{正确标记加速考生的数量}}{\text{加速作答考生总数}} \tag{22}$$

$$ADL = \frac{\sum_{i=1}^{N}|\hat{p}_i - p_i|}{N} \tag{23}$$

其中，$\hat{p}_i$和p_i表示考生 i 由 CPA 方法探查到的改变点的位置和真实的改变点位置，N是考生人数。

计算在给定时间内未完成测验的学生比例($\%NF$)是为了考察测验时间、长度和题目难度等设置的是否合理，为测验设计提供一些有用的信息。

详见修改稿第 12 页。

意见 26：检验力和 I 类错误率的分母是否都是在规定时间完成测验的所有考生？应当说明。

回应：检验力的分母为加速作答考生总数，在模拟研究中当考试结束，考生还未完成所有的试题时测验直接终止，没有做完的题目的作答时间设置为 0，未完成作答的

考生直接被标记为具有加速作答行为。检验力的分母中包含这些考生。I 类错误率的分母为正常考生(即没有出现加速作答的考生)的总数。我们已在修改稿中进行补充和说明。详见修改稿第 12 页和公式 21,22。

意见 27：作者统计“在给定时间内未完成测验的学生比例(%NF)”的意义是什么，这其实是与模拟条件设置有关的，并不是应用检测方法得到的结果。
回应：非常感谢您的提问。计算在给定时间内未完成测验的学生比例($\%NF$)是为了考察测验时间、长度和题目难度等设置的是否合理，为测验设计提供一些有用的信息。一般情况下测验时间，长度等设置的都是比较合理的。但当“%NF”值比较大时，我们就需要重新思考测验的设计是否合理以及是否是要改进。

我们已将上述观点和内容补充到修改稿中，详见修改稿第 12 页。

意见 28：建议把结果部分的 3 个表合成一个表，可以清晰的比较研究结果。
回应：。我们已将表格进行合并，详见修改稿第 15 页。

意见 29：实证数据分析（1）第 17 页，“我们也删除了那些在测验末期题目上的反应时间为 0 的考生”，是指没有做完就退出的？还是到时间没有做完的？
回应：在测验末期题目上的作答时间为 0，表明该考生属于“较严重”的加速行为，这里我们主要是想考察文中的方法对于“不那么严重”加速作答行为的考生的“检验力”，因此，我们把“较严重”加速行为的考生删除了。我们在修改稿第 16 页进行了阐明。

意见 30：为什么要抽取 5000 名考生，而不取和模拟研究类似的样本量（1000）？
回应：模拟研究中的样本量是参考 Yu 等(2020)的设置，在实证数据分析中，对数据进行整理后，一共保留了 33000 名考生的数据，我们从中随机抽取 5000 名考生的数据是为了在这样一个相对较大样本量条件下，更好地考察 CPA 在检测加速作答考生上的表现。

意见 31：第 18 页的几个阈值，1000 名考生得到的阈值可以用于 5000 名考生的情况？之前说明阈值具有跨测验长度的稳定性，这里也假设具有跨样本量的稳定性？
回应：统计量的阈值通过蒙特卡洛方法获得，它是通过模拟大量正常被试，根据小概率原理，将统计量超过 95%的人数对应的值作为临界值。

其步骤大致如下：

(1)通过公式 11，在测验长度为 40,60,80 的条件下随机生成 10000 个正常的作答时间模式。(2)基于前面介绍的似然比统计量和 Wald 统计量的计算公式，分别得到$\Delta l_{max,i}$和$W_{max,i}$的 10000 个值；(3)将它们按从大到小排序，得到它们第 500、第 100 和第 10 个最大值$c_{0.05}$，$c_{0.01}$和$c_{0.001}$，分别近似对应检验水平为 0.05，0.01，0.001 时的临界值;(4)每种实验条件重复 100 次，取平均的$c_{0.05}$，$c_{0.01}$和$c_{0.001}$值作为后面实验中用到的经验临界值。

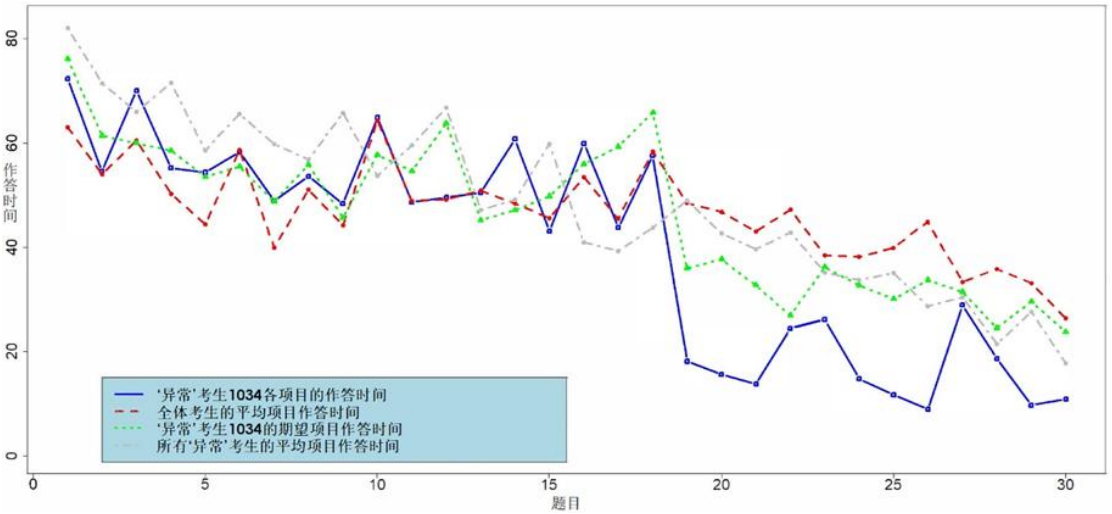
各条件下的阈值都是基于 10000 个正常的作答时间模式、重复 100 次取均值计算得到的；经过重复实验发现在$\alpha = 0.05$和 0.01 时阈值的方差都比较小，说明阈值是比较稳定的，因此既可以作为模拟研究中的阈值，也用于实证研究中。

我们将蒙特卡洛模拟的过程调整至修改稿中的“CPA 统计量临界值的获取”部分，详见修改稿第 9 页。

意见 32：图 4 中的两条线代表什么，应当加入图例。另外，能否加一条线表示这名考生的期望反应时。文中提到“各题的反应时都小于平均反应时。这些表明将编号为 1034 的考生标

记为异常考生是合适的”。其实，并不是各题反应时都小于平均反应时就能被标记，应当是和自己的期望反应时比，显著小的应该被标记。

回应：我们已在图 4 中增加相应的图例，其中红色线表示测验中各题目的平均作答时间，蓝色线表示的是 1034 号考生各题目的作答时间，绿色线表示 1034 号考生各题目的期望作答时间，灰色线表示的是样本中所有“异常”考生的平均项目作答时间。修改后的图如下：



详见修改稿第 17 页。

意见 33：结论和讨论（1）作者认为“相对于基于作答数据对加速作答考生的检测(Shao et al., 2016; Sinharay, 2016)，基于反应时数据的检测具有更高的检验力”，这个结论从何而来？前人研究和本研究模拟条件不同无法比较，有可能是本研究模拟条件的设置在反应时上差异更大，因此导致比作答反应上的差异更容易被检测出来。

回应：Choe, Zhang 和 Chang(2018)中分别基于作答数据、作答时间数据以及结合两种数据对泄露题目进行了检测。基于这篇文章的研究结果，我们将讨论部分中的内容更改为：

已有研究证明与基于作答数据的异常项目检测相比，基于作答时间数据的检测具有更高的检验力，且检验力的提高是来自于作答时间数据所提供的额外信息(Choe et al., 2018)，本研究的结果也与这一结论相符，即基于作答时间数据检测异常考生会比基于作答数据具有更高的检验力，出现这种现象的原因也可能正如引言中所说，作答时间是连续数据，同时包含了题目信息和考生能力信息，因此它相对于常见的二级计分作答数据有更高的检验力。

详见修改稿第 18 页。

意见 34：作者对检验力更大的归因为“很明显，检验力的提高来自于反应时数据所提供的额外信息。”作者的方法并没有同时结合作答反应和反应时信息，何来额外信息一说？

回应：这里的表述确实不太准确，我们对结论进行了更改，具体为：

已有研究证明与基于作答数据的异常项目检测相比，基于作答时间数据的检测具有更高的检验力，且检验力的提高是来自于作答时间数据所提供的额外信息(Choe et al., 2018)。

详见修改稿第 18 页。

意见 35：第 19 页第 13 行，“连续类型的测验数据”是指什么？与“二级计分测验数据”相对，每道题都是连续计分的测验吗？

回应： “连续类型的测验数据”指的是像作答时间数据这样的连续数据，它是相对于传统的二级计分测验数据或多级计分测验数据来说的。

原文的阐述不够清晰，我们已在修改稿中进行修改，详见修改稿第 18 页。

第二轮

审稿人 1 意见：

意见 1：作者对一些问题的回复及思考过于简单，比如，对于问题“检测出这些加速的考生之后又怎样？让他们重新考试吗？取消成绩吗？”的回答，现实中怎么可能仅仅因为考生加速作答了就删除她/他的成绩呢？谁有删除考生分数的权利？

回应：非常好的问题。已有的研究中关于异常作答数据的处理方式，主要包括 partial deletion、listwise deletion (Patton, 2015; Patton et al., 2019)或者构建稳健的参数估计方法 (Schuster & Yuan, 2011)。异常作答数据不但会影响项目参数的估计，而且也会影响考生特质参数的估计。更重要的是，项目参数会影响所有考生的能力参数估计。因此，在估计项目参数时，需要将异常的数据“排除”在外，可以采用 partial deletion, 即删除那些被识别为异常的作答数据；或者采用 listwise deletion，即删除被识别为异常作答数据对应的考生。研究表明，无论是 partial deletion 还是 listwise deletion，都可以提高项目参数的估计精度，因而可以进一步提高考生的能力参数估计精度 。需要注意的是，采用 partial deletion 或 listwise deletion 估计得到更准确的项目参数估计之后，就可以对考生进行能力参数估计，对于正常作答考生的数据，能力参数是基于全体题目的作答数据估计得到；而对于包含异常作答数据的考生，可以只基于那些正常作答数据估计得到其能力，因为这样可以得到更准确的能力估计(Patton, 2015; Patton et al., 2019)。

正如您所疑惑的那样，Patton (2016)和 Shao 等(2016)也在文中指出删除部分作答数据（即异常数据）的做法是需要慎重的，需要结合多方面的信息做出决策。我们提出的统计分析方法是帮助测验管理人员去推断哪些考生的作答数据异常，以及哪些数据是异常的，及可能的异常作答行为。删除部分作答数据的目的是为了降低这些异常数据所导致的“遮罩效应(masking effect; Fung, 1993; Yuan & Zhong, 2008)”，从而提高项目参数和被试能力参数的估计精度。

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*(422), 515–519.

Patton, J. M. (2015). *Some consequences of response time model misspecication in educational measurement* (Doctoral dissertation). University of Notre Dame.

Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics, 44*(3), 309-341.

Schuster, C., & Yuan, K-H. (2011). Robust estimation of latent ability in item response models, *Journal of Educational and Behavioral Statistics, 36* (6), 720-735.

Yuan, K. H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. Sociological Methodology, 38(1), 329–368.

意见 2：参考文献标注请使用 APA 7th 格式；

回应：我们已将文中的参考文献标注改为 APA 7th 格式。

意见 3：请统一表述，比如“反应时”、“作答时间”等；

回应：我们已将文中的专业术语进行了检查，例如“反应时”、“作答时间”等全部统一为“作

答时间”。

意见 4：请修改题目，使其与研究内容匹配，比如，明确什么类型的异常作答；
回应：我们已将题目修改为“基于作答时间的改变点分析在检测加速作答的探索与研究”。

意见 5：讨论部分需要增加一些探讨，比如，当基于 RT 数据检验出的结果和基于传统作答结果数据的检验结果出现差异时，应该以哪个检测结果为准？实际上为使得文章更加完整，且更体现出研究的价值，审稿人建议作者额外增加研究，对比本文所提出的方法和已有基于作答结果检测方法之间的一致性或差异性，并尝试基于此为实践应用者提供使用建议。
回应：非常好的问题。不论我们基于何种作答数据（得分数据、作答时间数据或者结合两种作答数据）使用统计分析方法检测异常数据时，我们都需要保持谨慎的态度，因为统计分析方法只能做出推断性的结论。当统计方法显示考生的数据异常时，我们最好结合其他信息例如考场中的摄像记录，学习和测评记录等综合判断，尤其是当我们认为考生作答数据异常是由加速作答以外的抄袭、作弊、题目预知等行为造成时。

关于基于作答结果的检验，Shao 等人 (2016), Yu & Cheng (2020)采用的实验条件与本研究中的实验条件基本相同，可以看出，相同的实验条件下本研究中的结果优于这些研究中的结果，因此，本研究只是聚焦基于作答时间的检测。

我们已将上述内容补充到修改稿中，详见修改稿第 19 页。

……

审稿人 2 意见：

作者较好地回答了我上一次提出的问题，并进行了较为有效的改进，文章质量有了较大提高。但是，仍有一些地方不够清楚，在此提出与作者探讨。

意见 1：作者描述两种潜在加速作答机制的模型时提到“分别是混合模型(the hybrid model, HM)和逐渐变化模型(the graduate change model, GCM)。HM 假设考生出现加速作答时的作答速度会发生突变；而 GCM 认为每位考生有自己独特的加速点，并且考生在加速点之后的题目上的答对概率会逐渐下降。”根据这部分的主题和关于 HM 的介绍，是否在介绍 GCM 时应当先说加速点之后的答对概率逐渐下降这一特点，再说每位考生的加速点可以不同。
回应：我们已对这部分内容进行修改，具体如下：

一方面，关于加速作答行为对于考生做题的影响机制，HM 假设考生出现加速作答时的作答速度会发生突变；而 GCM 认为考生在加速点之后的题目上的答对概率会逐渐下降；另一方面，有加速作答行为的考生，他们出现加速作答行为的位置是随机变量，即加速点各不相同。

详见修改稿第 10 页。

意见 2：修改稿中“例如造成模型与数据的失拟、被试与题目参数估计的扭曲(Stefan, Dietrich, Wolfgang, & Michael, 2016)，影响考试的信度和效度(Guo, Tay, & Drasgow, 2009)等等。因此，检测测验中的异常作答行为或异常作答数据是非常重要和关键的，研究者们也一直在寻找相关的解决方法”。（1）有一些表述不太符合常见的表达方式，如“失拟”建议改成“不拟合”，“扭曲”建议改成“偏差”；（2）最后“检测测验中的异常作答行为或异常作答数据”，到底是检测行为还是数据，似乎直接检测到的是异常的数据模式。
回应：我们已对这部分内容进行了修改，具体内容如下：

（1）测验数据中包含异常作答数据会降低其自身及整体测验数据的质量，从而对后续

的分析结果产生一系列的不良影响，例如造成模型与数据的不拟合、被试与题目参数估计的偏差(Stefan, Dietrich, Wolfgang, & Michael, 2016)，影响考试的信度和效度(Guo, Tay, & Drasgow, 2009)等等。

（2）考生出现异常作答行为时产生的数据称为异常作答数据或异常作答模式。因此我们将“检测测验中的异常作答行为或异常作答数据”对应的内容改为：

因此，检测测验中的异常作答数据是非常重要和关键的，研究者们也一直在寻找相关的解决方法(e.g., Bejar, 1985; Evans & Reilly, 1972; Shao, Li, & Cheng, 2016; Bradlow, Weiss, & Cho, 1998; McLeod, Lewis, & Thissen, 2003; Wise & Kong, 2005; Yu & Cheng, 2019, 2020)。

详见修改稿第 1 页。

意见 3：修改稿中“前三种统计量都是通过检验虚无假设(考生的潜在特质没有发生显著的变化)来判断考生是否存在异常作答行为”，建议改成“通过检验是否能拒绝虚无假设来判断……”。

回应：我们已对相应内容做出修改，具体内容如下：

这几种统计量都是通过检验是否能拒绝虚无假设(考生的潜在特质没有发生显著的变化或考生的作答数据没有异常变化)来判断考生是否存在异常作答行为。

详见修改稿第 7 页。

意见 4：作者对于一审问题 3（5）的修改“作答时间数据的获取已经越来越容易，并且作答时间数据在检测考生的异常作答行为上比作答得分数据拥有先天的优势。比如从同样是得 80 分的考生，相对于得分数据，结合作答时间数据则更容易判断考生的异常作答作为。”这里仍然没有体现举例是加速作答的情况。建议改成“考生加速作答的异常作答行为”。

回应：我们已将这个例子修改为：

如今作答时间数据的获取已经越来越容易，并且作答时间数据在检测考生的异常作答行为上比作答得分数据拥有先天的优势。比如当某位考生的作答模式为[1111101010]时仅从分数上不容易判断考生是否出现了加速作答，但是结合作答时间数据[57, 48, 51, 36, 42, 23, 18, 13, 7, 6]则更容易判断，因为加速作答的直接体现就是在作答时间上。

详见修改稿第 5 页。

意见 5：修改稿中“如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。例如，詹沛达(2019)和詹沛达, Jiao Hong, Man Kaiwen (2020)分别提出了关于作答时间数据的多维模型，研究结果皆显示引入作答时间数据可提高或精确估计模型的参数等等。”这部分的论证和举例，如果聚焦于反应时在异常作答识别中的应用，以及相关研究，是否更有针对性，更有说服力？

回应：我们已对这部分的内容进行修改，具体如下：

如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。例如，van der Linden (2006), Wang 和 Xu (2015), 郭小军和罗照盛 (2019), 詹沛达(2019)和詹沛达, Jiao Hong, Man Kaiwen (2020)等基于不同的应用场景，构建了作答时间模型，对有关的理论和应用展开了深入研究。

详见修改稿第 2 页。

意见 6：模拟研究设计中，考生数量固定就不应作为模拟因素，出现在表 2 中，应作为固定参数设置介绍，并且也不应出现在计算条件数的公式中。

回应：我们已对相应内容做出修改，具体内容如下：

模拟研究共$3 \times 3 \times 4 \times 2 = 72$种条件，每种条件重复 50 次。模拟研究使用 R 程序完成。

表 2 模拟条件

因素	水平
测验长度	40,60,80
加速作答考生的比例	10%,20%,30%
改变点的位置参数η_i	Median (0.6,0.7) ×σ_η^2(0.04,0.001)
项目参数	已知, 未知

详见修改稿第 11 页。

意见 7：在介绍评价指标之前应当是分析过程，说明模拟研究中比较的方法及分析流程。

回应：非常感谢您的建议，我们已在文中评价指标之前增加了“异常作答数据的检测过程”一节，具体内容如下：

基于前文对 CPA 统计量的分析，研究拟使用似然比统计量与 Wald 统计量依次对每位考生的作答时间数据进行检测。大致过程如下：(1)计算每位考生在每道题上的似然比值,选取最大的似然比值作为似然比统计量的值，Wald 检验类似；(2)将两种方法的统计量与各自对应条件下的临界值进行比较，当统计量值超出临界值时，将考生的作答数据标记为异常作答数据；(3)当考生的作答数据标记为异常时，统计考生异常作答行为出现的位置；(4)用预定的评价指标对 CPA 方法的检测效果进行评价。

详见修改稿第 12 页。

意见 8：作者在讨论中提到“已有研究证明与基于作答数据的异常项目检测相比，基于作答时间数据的检测具有更高的检验力，且检验力的提高是来自于作答时间数据所提供的额外信息(Choe et al., 2018)，本研究的结果也与这一结论相符，即基于作答时间数据检测异常考生会比基于作答数据具有更高的检验力”该研究并没有对比基于反应时和基于作答反应的方法。并且，模拟数据的生成也只模拟了异常作答的反应时，没有在模拟作答反应时体现异常作答。因此，何以得到“本研究的结果也与这一结论相符，即基于作答时间数据检测异常考生会比基于作答数据具有更高的检验力”这一结论？

回应：我们的表述确实不太恰当，已修改相应内容，具体如下：

已有研究证明与基于作答数据的异常项目检测相比，基于作答时间数据的检测具有更高的检验力，且检验力的提高是来自于作答时间数据所提供的额外信息(Choe et al., 2018)。本研究也进一步表明，相对于已有研究：Shao et al. (2016)，Sinharay (2016)，Yu 和 Cheng (2020)，基于作答时间对包含加速作答行为数据的检测有更高的检验力。出现这种现象的原因正如引言中所说，作答时间是连续数据，同时包含了题目信息和考生能力信息，因此它相对于常见的离散的得分数据可以导致更高的检验力。

详见修改稿第 18 页。

第三轮

审稿人 1 意见：

作者仅是简单地说基于 RT 的检测率高于基于 RA 的，这只是一种结果，并没有思考我的问题。建议作者再进一步思考并补充，当基于 RT 的检测结果与基于 RA 的检测结果出现矛盾或不一致时，应该以哪个为准；尤其是实证研究中，如果有对比标准应该依据哪个标准

来判断，如果没有判断标准，如果出现了矛盾应如何处理。

回应：非常好的问题。这确实是实际应用中很可能出现的问题，基于我们目前所掌握的方法和技术，当基于 RT 的检测结果与基于 RA 的检测结果出现矛盾或不一致时，我们需要针对出现矛盾的数据进行具体分析或者引入更多的信息才能做出判断。

一方面，基于 RA 和 RT 的分析各有其特点，比如 RA 数据，在教育测量领域通常需要满足单调性假设，如果严重违反这个假设，则可以得到考生异常作答的证据，比如低能力考生在容易题上大量做错，在难题上反而大量正确的情形，可以通过分析测验题目的具体内容得到进一步的信息，考生可能出现了作弊（cheating）或提前了解了题目的信息（preknowledge）。

而 RT 数据，通常需要结合对应题目在试卷上出现的位置来判断，比如考生在测验部分试题上的作答时间明显偏离群体在该题上的时间分布，并且使用的时间显著偏小，如果这部分题目出现在测验中后期，则有可能出现了加速作答（speededness）等。

因此，RA 和 RT 数据虽然都可以检测异常作答，但它们检测异常作答类型和特点是不同的。

另一方面，当基于作答时间数据与基于得分数据的检测结果出现矛盾，仅从统计分析结果不容易判断哪种数据的检测结果是准确的时候，我们需要引入更多的信息(包括对测验内容的具体分析，其它统计量的分析，甚至是考场中的摄像记录和历史数据等)来谨慎地对这种数据做出综合评估(Wang et al., 2018)。

我们将对应的内容进行了重写和补充，详见修改稿第 17 页。

……

审稿人 2 意见：

意见 1：第二轮审稿意见 5 指出，修改稿中“如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。例如，詹沛达(2019)和詹沛达, Jiao Hong, Man Kaiwen (2020)分别提出了关于作答时间数据的多维模型，研究结果皆显示引入作答时间数据可提高或精确估计模型的参数等等。”这部分的论证和举例，如果聚焦于反应时在异常作答识别中的应用，以及相关研究，是否更有针对性，更有说服力？——作者在修改后加入了一些新的文献，并概括“构建了作答时间模型，对有关的理论和应用展开了深入研究”。这里仍然没有明确说明这些研究是应用反应时来识别异常作答。

回应：我们对这部分内容进行了重写，希望能更清楚和准确地表达。

如今随着新技术的发展，计算机测验与在线评估越来越多，作答时间数据的获取也变得更加便利，逐渐获得学者们的关注。例如 van der Linden 和 van Krimpen-Stoop (2003)使用作答时间数据检测考生预知试题以及加速作答；van der Linden 和 Guo (2008)，Pan 和 Wollack (2021)等使用作答时间数据检测测验中考生预知试题的情况等等。还有研究者基于不同的应用场景构建作答时间模型，结果显示引入作答时间数据有助于模型的参数估计等，拓宽了作答时间数据的使用范围(Wang & Xu , 2015; 詹沛达, 2019;詹沛达 等, 2020)。

详见修改稿第 2 页。

意见 2：第二轮审稿意见 6 请作者修改模拟研究设计的相关表述，作者在这一稿中总结了 4 个模拟条件。表 2 将不同因素和水平称为模拟条件，那么文中的表述“模拟研究共 3×3×4×2=72 种条件，每种条件重复 50 次”，这里的“条件”就应当指模拟条件的组合，每种模拟条件的组合下，数据重复生成并分析 50 次。另外在这一条的回复中，作者的回答是“3×3×4=36 种条件”（是错的），修改稿中是 72 种条件（是对的），存在不一致。

回应：我们已将第二轮回复中的内容进行了修改。具体内容如下。
模拟研究共$3 \times 3 \times 4 \times 2 = 72$种条件，每种条件重复 50 次。模拟研究使用 R 程序完成。

表 2 模拟条件

因素	水平
测验长度	40,60,80
加速作答考生的比例	10%,20%,30%
改变点的位置参数η_i	Median (0.6,0.7) ×σ_η^2(0.04,0.001)
项目参数	已知, 未知

意见 3：第二轮审稿意见 8，质疑了“基于作答时间数据检测异常考生会比基于作答数据具有更高的检验力”这一结论。作者在回答中指出，“相对于已有研究：Shao et al. (2016)，Sinharay (2016)，Yu 和 Cheng (2020)，基于作答时间对包含加速作答行为数据的检测有更高的检验力。”这里作者是将本研究的检验力结果与这些研究比较过从而得出这一结论吗？本研究和这些研究的模拟条件和产生数据方式都相同吗？是否能做这样的比较呢？作为一篇严谨的学术论文，作出这样的结论需要有充分的理由。 修改后发表。
回应：针对这个问题，我们之前的表述的确存在不严谨之处，在仔细思考之后，我们把这句话在修改后的版本中进行了删除，对涉及的内容进行了重新改写，希望能更准确地表达我们的意思。

因为本研究基于的是作答时间数据，Shao et al. (2016), Sinharay (2016), Yu 和 Cheng(2020)等研究基于的是作答得分数据，所以模拟条件和产生数据的方式不能完全一样。实际上，这两类数据，它们在检测异常作答行为上有各自的特点，比如在检测 preknowledge 时，作答时间数据就有其优势 (Hong et al., 2016; Sinharay, 2020; Sinharay & Johnson, 2019)，而在一些根本没有收集作答时间数据的考试就只能基于得分数据进行分析等。未来研究结合作答时间数据和得分数据进行异常作答行为的检测值得深入研究。修改后的论述如下：

本研究进一步表明基于作答时间数据在异常作答行为检测上具有很高的检验力。实际上，基于得分数据和作答时间数据在检测异常作答行为时各有其特点，将它们结合起来则有可能进一步对考生的异常作答行为的类型进行分析和探索。

详见修改稿第 19 页。

编委意见：这篇论文提出了基于反应时数据的改变点分析方法以识别加速作答，具有一定的创新价值。作者采用模拟研究验证了提出方法的有效性，整个文章逻辑清楚。经审稿人挑剔性阅读，作者修改，论文达到发表要求。建议发表。

主编意见：修改满意。建议发表。