《心理学报》审稿意见与作者回应

题目：多级计分测验中基于残差统计量的被试拟合研究
作者：童昊; 喻晓锋; 秦春影; 彭亚风; 钟小缘

第一轮

审稿人 1 意见：

意见 1：第 3 页第 6 行，“疲劳”、“粗心”和“焦虑”不是异常行为，而是导致异常行为的可能原因；另外，“粗心”或“焦虑”导致的错误是否属于“异常”？或许这正反映了学生对知识掌握不牢固，恰恰是最需要分析的数据。因此，该如何界定“异常”？这可能是引言部分需要强调的内容，类似于给“异常作答”下操作性定义；

回应：感谢您细致的审稿！

我们在原文中没有对异常行为解释清楚，在这里， 对它进行阐明。通常在测量模型中，我们都假设考生在考试过程中是依靠自身的能力作答，考生在每一道题目上的作答都是一个随机变量，它这里面暗含的假设是考生在这个题目上的得分存在随机因素的影响，比如猜测或失误（粗心）等随机因素，进而导致随机误差。

而我们在文章中提到的猜测、粗心等异常行为准确地说是指超出了正常范围的猜测或粗心等作答行为。这和很多已有的研究中考察的背景相同 (Cheng & Shao, 2020; Liu & Liu, 2021; Sinharay, 2016; Shao & Cheng, 2016; van Krimpen-Stoop & Meijer, 2002; Yu & Cheng, 2019, 2020)，它是能够导致系统误差的因素，因此需要进行检测和处理。对于这一点，我们在修改稿中进行了具体的阐明，希望能更准确地表达意思。

异常是相对于正常而言的。正常的作答，是考生能够依据考试规范，完整发挥自身相关领域能力的作答，但凡违反上述描述，均可以被认为是异常作答。异常作答是测验过程中很常见的现象。我们将导致异常作答的行为统称为 “异常作答行为”，它可以是具体物理行为，如考试前事先得到题目的信息(item preknowledge; Sinharay, 2017; Wang et al., 2018),考试过程中违反考虑的规定(cheating; Frary, Tideman, & Watts, 1977; Shu, Henson, & Luecht, 2013; Sinharay & Johnson, 2020)，考试过程中的加速作答行为(speededness; Cheng & Shao, 2021; Sinharay, 2016; Shao, Li, & Cheng, 2016; Yu & Cheng, 2020)等；也可以前面提到的其它异常作答行为，如因为粗心漏看题目的关键信息或低作答动机(粗心；carelessness; Meade & Craig, 2012; Yu & Cheng, 2019)，疲劳 (疲劳; Sinharay, 2016)等。

您的建议非常好！我们也意识到了这个问题，仅仅检测出数据的异常，从统计学的角度来看是不够的，还需要从心理学本身去探讨这些数据产生异常的真正原因。

就我们目前所掌握的知识来看，国内外对异常作答行为研究相关的文献一般也不刻意区分异常行为究竟是一种物理行为（作弊等）还是心理行为（粗心、焦虑），更多的是从数据本身的角度出发，如果涉及到心理因素，可能需要另开辟主题进行研究。为了使我们研究更聚焦，这里我们主要是从数据的角度来考虑统计量的检验效果。当然，导致这些行为背后的心理因素也是非常重要的，但同时它又是非常复杂 的，这些是本研究存在的不足之一，我们在讨论部分对这一点进行了阐明，我们期待未来在这方面开展进一步的研究。

意见 2：对统计量 R 的引出过于生硬、莫名其妙，根据已有文献回顾无法直接引出统计量 R，

需要完善补充；
回应：非常好的建议！我们对这部分内容进行了重写，详细阐述了基于加权残差统计量的提出：从基于 Snijders(2001)的通用指标，到已有加权残差统计量的参考，再到提出的全过程。

意见 3：公式 17 中对观测与期望差值求绝对值的做法与公式 13 中求平方的做法哪个更有理论优势？或许可以借鉴下为何方差是对标准差求平方而不是绝对值的思路；
回应：这仍然是一个非常好的建议！因为使用受到“污染”的数据来估计能力值，残差取平方带来的误差会大于取绝对值，从而影响检验力，因此取绝对值要优于取平方。这一点也在我们研究的前期模拟实验中得到了验证。我们对这部分内容进行了修改和补充。

意见 4：公式 17 中，为何加权函数要设定为公式 18？缘由(不是理由)是什么，需要做更清晰的交代；公式 18 中的 P(X|θ)和上文中的 Pij(θ)什么关系?
回应：感谢您细心的审稿！我们已在文中进行修改，加权函数的思路是放大可能异常的残差，抑制相对正常的残差，因此选择使用得分概率的倒数作为加权函数，这一点我们在修改稿中进行了更具体的说明。

P(X|θ)指的是在多级计分下（同样包括二级计分）获得 X 分的概率，前文中的 Pij(θ)是指在二级计分下考虑 *i* 答对题目 *j* 的概率，二者在背景和含义上有不同。我们进行了检查，均已在原文中进行修改，统一了符号和格式。

意见 5：公式 19 和 20 中的 M 是什么？如果 M 是有“随意”取值可能的，那么 R 显然是 U 的特例（M=1）,而不是作者说的在二级评分下 U 是 R 的特例；
回应：感谢您细致的审稿，非常好的意见！在这里，M 指的是测验长度，即测验包含的项目数量，我们对这个符号进行了明确，并详细了文中该部分的解释，希望不至于引起误解。

在二级计分的背景下，对于同一批数据，有以下关系：R=M*U，此时二者在临界值以及统计量的数值上均为整数倍 M 的关系，因此具有相同的检验效力。

而在多级计分背景下，由于 U 是针对二级计分的统计量，因此无法使用，R 仍可以适用于多级计分数据，从这个 角度可以说 U 是 R 统计量在二级计分背景下的特殊情况。

意见 6：模拟研究 1 (1)异常行为检测的现实应用场景是什么？并非李克特量表，为什么是五级计分是常见的？有研究或参考文献支持吗？ (2)区分度的生成设定是否符合实际？ (3)第 9 页第 13 行，“群体平均得分”是什么？ (4)当题目难度和能力都服从标准正态分布时（如此强假设）得到的临界值，是否具有实际可应用性？如果实际数据中被试能力不满足标准正态分布或题目难度不满足标准正态分布，使用该临界值会带来什么错误结果？【这是非常重要的问题，请作者谨慎回复】 (5)依据公式 23 生成数据不符合实际，请更换为 categorical distribution；(6)说图 2 中 lzp 负偏态是否太主观，如果正态分布检验表示其不符合正态分布，或许更有说服力【不重要的问题】;
回应：感谢您的意见和建议！

（1） 异常行为检测的现实应用场景包括一切评估某种心理和教育特质水平的测试，只要是对考生进行测试，就有可能出现导致系统误差的“异常作答行为”。
在实际应用中，通常关注作弊 cheating (Frary et al., 1977; Shu et al., 2013; Sinharay & Johnson, 2020)，加速作答 speededness (Sinharay, 2016; Shao, Li, & Cheng, 2016; Yu & Cheng, 2020; Cheng & Shao, 2021)，题目预知 preknowledge (Sinharay, 2017;

Wang et al., 2018)，粗心或低作答动机 carelessness(Meade & Craig, 2012; Liu & Liu, 2021; Yu & Cheng, 2019)等异常行为；下面列出部分参考文献：

Frary, R., Tideman, N., & Watts, T. (1977). Indices of cheating on multiple choice tests. *Journal of Educational Statistics, 2*(4), 235–256.

Liu, Y., & Liu, H. (2021). Detecting noneffortful responses based on a residual method using an iterative purification process. *Journal of Educational and Behavioral Statistics*, 1-36. DOI: 10.3102/1076998621994366.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika, 78*(3), 481–497.

Sinharay, S. (2016). Person Fit Analysis in Computerized Adaptive Testing Using Tests for a Change Point. *Journal of Educational and Behavioral Statistics, 41*(5), 521–549.

Sinharay, S. (2017). Detection of Item Preknowledge Using Likelihood Ratio Test and Score Test. *Journal of Educational and Behavioral Statistics, 42*(1), 46 -68.

Sinharay, S., & Johnson, M. (2020). Detecting test fraud using Bayes factors. *Behaviormetrika, 47*, 339-354.

Yu, X., & Cheng, Y. (2019). A Change-Point Analysis Procedure Based on Weighted Residuals to Detect Back Random Responding. *Psychological Methods, 24*(5), 658-674.

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. https://doi.org/10.1080/00273171.2020.1809981.

Wang, C., Xu, G. J., Shang, Z. R., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics, 43*(4), 469 -501.

有很多涉及到多级计分的参考文献中都使用了 5 级计分作为模拟实验的条件，如 Chalmers (2020); Cohen et al.(1993); Dodd, Ayala & Koch, 1995; Emons, 2008; 李佳, 丁树良, 2018; 程小扬, 丁树良, 朱隆尹, 巫华芳, 2012; Sinharay, 2016; Yu & Cheng, 2019 等。下面列出部分参考文献

Chalmers, R. P. (2020). Partially and fully noncompensatory response models for dichotomous and polytomous Items. *Applied Psychological Measurement, 44*(6), 415 -430

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335–350.

程小扬,丁树良,朱隆尹, 巫华芳.(2012).等级评分模型下的最大信息量分层选题策略. 江西师范大学学报(自然科学版), 36(5),446-451.

Dodd, B, G., Ayala, R, J, De., & Koch, W, R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5-22.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247.

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model Selection Indices for Polytomous Items. *Applied Psychological Measurement, 33*(7), 499–518.

李佳,丁树良.(2018).基于 GRM 模型的 CAT 分层方法在校准误差中的应用研究. *江西师范大学学报(自然科学版), 42*(4),374-378.

Sinharay, S. (2016). Asymptotically Correct Standardization of Person-Fit Statistics Beyond Dichotomous Items. *Psychometrika, 81*(4), 992-1013.

Yu, X., & Cheng, Y. (2019). A Change-Point Analysis Procedure Based on Weighted Residuals to Detect Back Random Responding. Psychological Methods, 24(5), 658-674.

（2） 文章的区分度参数设置主要参考 Dodd 等 (1995)、Emons (2008)、罗芬等(2012)、Xiong 等(2020)、陈青等(2010)对 GRM 有关的研究，下面是对应的参考文献

Dodd, B, G., Ayala, R, J, De., & Koch, W, R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5-22.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. Applied Psychological Measurement, 32(3), 224–247.

罗芬,丁树良，王晓庆.(2012).多级评分计算机化自适应测验动态综合选题策略. 心理学报(03),400-412.

Xiong J., Ding S., Luo F., & Luo Z. (2020) Online calibration of polytomous items under the graded response model. *Frontiers in Psychology. 10*:3085, 1-11.

陈青,丁树良,朱隆尹, 许志勇.(2010).三参数等级反应模型及其参数估计. 江西师范大学学报(自然科学版), 34(2),117-122.

熊建华,罗慧,王晓庆, 丁树良.(2018).基于 GRM 的在线校准研究. 江西师范大学学报(自然科学版), 42(1), 62-66.

（3） 群体平均得分指的是各个题目在被试总体上的平均得分，它可以用来判断题目的相对难度，具体的计算方式如下：基于被试群体能力分布先验信息，抽取被试，得到该抽取被试的平均得分，并乘以其概率密度，最终整个式子的求和就是群体平均得分。可以简单理解为在该分布下模拟抽取大量被试作答，得到他们的平均分，这个平均分可以看作是题目的平均难度。这个操作是在实验中的一种技术性处理，不是必需的操作，因此，为了不产生不必要的误导，我们把这部分细节在修改稿件中移除了。

（4） 非常好的问题。由于本研究聚焦于统计量本身，因此是在已知项目参数的基础之上做的研究。若被试能力分布和项目参数分布不严格服从标准正态分布，对该方法检测效果部分无影响。文章涉及的异常检测方法并不依赖被试样本能力参数的分布，因此也不会受到分布形态的影响。模拟实验之所以如此设置，是在前人的假设和研究的基础上，力图方便快捷地生成一批可用于分析的模拟数据。实证数据中，我们只需要针对具体的真实数据进行参数估计，便可以继续使用文章提出的检测方法，因为临界值是基于 10000 个能力服从标准正态分布的考生得到的，用以代表正常考生群体的统计量分布，由此界定的临界值是用来判断某个考生是否偏离该正常群体。至于实证数据中真实的能力参数分布如何，并不影响本方法的使用。这个方法也是很多研究中所采用的方法，比如 Meijer (2002), Sinharay (2016), Shao, Li, & Cheng (2016), van Krimpen-Stoop & Meijer (2002), Worsley (1979), Yu & Cheng (2020)等。

可能影响的部分是涉及先验信息的参数估计方法——即 EAP 方法的估计结果，此时我们可以采用两种措施①放宽条件，改用更为宽泛的部分信息先验进行替代②使用不需要先验信息的极大似然估计方法(MLE)进行估计。

未来，当项目参数未知时，R 统计量的表现还需要进一步进行探索。在这种条件下，因为它会产生“掩蔽效应 masking effect (Fung, 1993; Yuan & Zhong, 2008)”，所以这种情况下通常需要对数据进行适当的处理 (Hong et al., 2020)。关于这一点，我们在讨论部分进行了说明.

（5） 对于这部分内容，我们重新进行了语言组织，并在文中添加了更详细的描述，希望能更明白地表达我们的意思。文中生成实际分数的模拟程序是编程逻辑的文字化表述，这也是通常的模拟多级计分数据的方法。具体过程是，由于各得分概率之和为 1，因此每种得分概率占据[0,1]区间上的某一段，长度等于其概率值大小，这样累计的等级得分概率填满整个[0,1]区间，在区间内生成均匀分布的随机数，观察随机数处于哪一段概率区间内，便可以生成其实际作答结果。

（6） 对于这个问题，我们在文中进行了补充说明，Lzp 在每个长度条件(20,40,60,80)下的分布，都经过 Jarque-Bera 检验和 Kolmogorov-Smirnov 检验否定其符合正态分布的假设，再计算其偏度，均小于 0，由此得出 Lzp 呈负偏态的结论。

意见 7：模拟研究 2 (7)对表 2 中一些设定的质疑（跟是否已有研究也同样使用过无关）：①作弊，作弊者为何只挑难题抄袭？ ②幸运猜测的操作定义是什么？跟猜测有什么区别？另外，按作者生成方式，作弊和幸运猜测的区别只是人数(或比例)上的差异，0.2*P(作弊)=P(幸运猜测)，比例少了就从作弊变成幸运猜测了？且为何幸运猜测的人只会去猜难题？类似问题，粗心-创造性作答。【现实中，往往会出现增益性猜测的都是能力较高的学生】 ③现实应用中学生出现随机作答的原因是什么？通常是由于考试时限，在最后面的题目出现该现象；另外，随机作答和猜测有什么区别？④粗心只是异常行为的可能原因，且为什么粗心导致的是异常作答而不是正常作答？ ⑤创造性作答的操作定义是什么？根据作者的模拟方法，这貌似是一个负性词； ⑥Z 是什么？

回应：

（1） 感谢您的意见！对于本研究中异常作答行为的定义，我们一方面确实是参考了已有研究中的做法，另一方面，由于不同研究中对于异常作答行为的定义也不完全相同。因此，我们是综合了已有研究和我们对数据的分析经验，最后选择的设定。

作弊者在面对难度远高于自身能力的项目上，极难获得高分作答，此时容易产生作弊行为（如看参考答案，使用通讯工具等），作弊行为通常能够达到虚高的得分（靠作弊直接获得答案），因此文章将其设置为获得该题的满分，而对于作弊者自身能够应付的低难度项目，他们可以选择更安全稳妥的方式——正常作答。所以作弊通常更可能发生在较低能力者身上，因为面对同一个测验，他们相比于中高能力者在更多的项目上感到难以应对。

（2） 操作定义表 2 中已有展示，幸运猜测者同样属于低能力群体，他们面对和作弊者一样的处境——诸多较困难的项目无法获得高分作答，此时他们采取的是不同于作弊的另一种手段，猜测，不同的是，靠猜测获得高分的概率远小于直接作弊，因此我们设定其仅有 0.2 的概率获得满分。猜测和作弊的区别主要在于考生对于异常项目的处理方式。

鉴于提到的能力较高的学生也会出现猜测行为，其背后的原因不外乎题目难度超越自身能力或者时间不够，这同样可以用文章的检测方法进行检验，这和前文的低能力者产生猜测行为的逻辑不会冲突。

因此，我们的研究中的模拟实验是综合了已有的研究，选择了几种较为典型的异常行为类型，确实难以顾及所有的异常行为类型，这是我们的一个不足之处，拟在在未来的研究中予以考虑。

（3） 一方面，随机作答的原因，可能是因为时间不够或者作答动机较低。对于作答时间不够，通常会使测验后面部分的项目受到影响。而对于作答动机较低，这种情况已有很多专门的研究进行讨论。

另一方面，鉴于模拟实验中的项目在生成时是随机的，因此选定测验后部分的 n 个项目和在测验中随机抽取 n 个项目对后续研究结果是等价的，在操作定义中没有特地选定后半部分项目。

随机作答和猜测的区别在于对项目的选择上，猜测发生在被试无法应对项目时，因此通常出现在难题上；而随机作答则不存在约束条件，可以发生在任意难度的项目上；如前文所言，选取后半部分固定项目和随机抽取项目在本研究的模拟实验中是等价的。

我们对上面的这些内容的措辞都进行了修改，希望能更准确地表达我们的意思 。

（4） 粗心使得被试处在非正常的答题状态，正常的状态下，被试能够准确获得项目的信息，发挥自身能力，进行作答。而粗心导致信息获取不完全，不能够发挥其在相关领域的能力，因此其作答结果视为异常作答。

（5） 文中的“创造性作答”源自异常行为操作定义的引用文献 Doval 和 Delicado(2020)中“Creative examinees”的直接使用，表示能力高的被试在容易的项目上答错，此时表达的是一种异常状态，已经在文中进行修改和标注。

（6） Z 指的是累积百分比，写法上借鉴了 Doval 和 Delicado(2020)的研究，如 θ<Z.375 指的是能力由低到高排序前 37.5%的人群。已在文中表格后进行标注解释。

意见 8：为何自变量选用的是受影响的题目的比例而不是出现异常行为的被试的比例？

回应：由于文章中项目参数是定为已知的，且临界值的获取是依靠模拟正常群体得到，此时改变异常被试人数的比例不对检测结果产生任何影响，如占总体 10%或 30%的异常被试，甚至设定全体被试均为异常，对检验结果均无影响。而异常项目比例是在被试水平上的，能够显著影响被试的能力估计和统计量计算结果，进而影响异常检测的结果，所以使用异常项目比例作为操纵检验力的自变量。

意见 9：实际上根据表 2 中对不同异常行为的模拟方法（指标本身并无法区分作者所赋予的名字，指标本身只是对异常作答的占比有所区分；比如作弊和幸运猜测、粗心和创造性作答对指标而言它们只是异常作答占比不同），在对结果进行解读时不应该主观引入“异常类型”的名字。比如，第 14 页最后一段，结果本身只能说明异常比例(或数量)大了(粗心→创造性作答：0.8*P → P)，检测率就高了；反之，异常比例(或数量)低了(作弊→幸运猜测：P →0.2*P)，检测率就降低了。

回应：非常好的意见！关于这一点，确实如审稿人所说，分析结果时不应该主观带入“异常类型”名字。模拟研究之所以如此设置，是希望能够模拟出几种典型的异常行为，来判断各个统计方法在模拟实验中的检测异常被试的效果。文章提出的统计量的主要目的是尽可能准确地筛选出存在异常的被试，是从无到有的过程，至于异常的被试究竟属于何种异常行为，不在文章的研究范围内，任何关于行为类型的言论都应该是一种初步的猜测，参考意义有限。我们对这些内容的表述进行了修改。后文的实证研究中，结论部分使用了较为谨慎的语言，表达了某种可能的推论。

意见 10：实证研究 (10)实证研究与模拟研究之间相互匹配相互支持吗？比如， ①模拟研究为何模拟 5 个选项，而实证研究中却只有 4 个选项（还合并了其中两个）； ②模拟研究中模拟生成的异常行为在李克特量表中会出现吗？比如，李克特量表中如何幸运猜测？如何创造性作答？等； ③缺失不属于异常作答？【连接问题 1】 ④模拟研究中用题目参数真值，而实证研究中却用估计值；用含有异常作答的数据得到的题目参数估计值来生成数据，对指标

的影响是什么？【重要问题，需要通过模拟研究来回答】； ⑤两个案例被试的解读，都用随机作答来解释，是否有些让人迷惑？

回应：感谢您细致的审稿！

（1） 在设计模拟研究时，主要是希望构建一个中等评分等级数量的实验条件，来比较统计量之间的效果，主要是参考了众多多级计分研究中的常见设置。因此没有考虑和后文的实证研究数据的评分等级进行匹配。

另外我们还使用 R 统计量和 Lzp 在实证数据下进行模拟实验，即 3 级计分下，实验结果和 5 级计分下的结论基本一致。我们在文中进行了补充说明。其他计分等级下拟在未来的研究中进行考虑。

（2） 模拟研究主要是模拟能力选拔测验中可能出现的异常情况，因此李克特量表中不一定会出现类似猜测，作弊等行为，更可能出现的是作答动机不足如 carelessness, inattentiveness, noneffortful response, low motivation 等。基于我们已有的知识，李克特量表类的数据是能够很好的拟合本文中使用的 GRM 的。并且异常检测的基本原理是探测数据-能力不匹配，只要李克特量表中存在一定程度的异常行为（如特质水平与其作答严重不匹配的情况），文章的方法同样可以使用的，我们对这部分内容也进行了进一步的说明。

（3） 的确，数据缺失也是一种异常。如果数据缺失是由于被试个人原因未进行作答，那么该种缺失数据确实有一定的解读空间，但是由于无法判定数据缺失发生在哪一个阶段(被试作答阶段、数据录入阶段、数据传输阶段等)。本研究本质上是一种数据驱动的统计方法，难以对缺失数据进行归因和推论，因此存在缺失的数据不在文章的考虑范围内。

（4） 如何影响项目参数的估计取决于被试在项目上的异常行为，如部分人群在某个难题上采取作弊行为，这势必导致对该项目难度的低估，从而降低拟合统计指标对该项目上作弊人群的探测能力；又或者是测验最后一个项目，如果真实难度并不高，但由于部分被试时间预留不充分，只能匆忙作答获得较低的得分，这将导致对该项目难度的高估，同样会降低拟合统计指标对该部分异常人群的检测能力。模拟实验结果也证明了这一点，使用项目参数估计值时，不论是 R 统计量还是 Lzp，相同一类错误率下的检测力都有一定程度地下降。因此，使用项目参数估计值必然在某种程度上限制了拟合统计指标的发挥，因为这会产生“掩蔽效应 masking effect” (Fung, 1993; Yuan & Zhong, 2008), 这种情况下我们拟在未来的研究中进行进一步的考虑，比如考虑更稳健的参数估计方法，或者对数据中的“噪音”进行适当的处理之后再进行异常检测(Hong, Steedle, & Cheng, 2019)。另一方面，根据 Rupp(2013)的研究，异常人数比例通常在 10%左右，如此比例的异常人数对项目参数估计的影响较小，因此实证研究中拟合统计指标量仍具有较大的使用价值。我们对这有关的内容进行了补充说明。

（5） 已在原文中进行修改，采用了更恰当的措辞，并且补充了一个异常案例。由于研究的主要目的是筛选出异常被试，关于异常类型的结论仅为研究者对数据的简单推测，参考意义有限，因此在原文的相关部分修改为使用更为谨慎、恰当的说法。

意见 11：在多级评分条件下发现的 R 与 lzp 的对比结论，与已有研究在二级评分下发现的 W（U）和 lz 的对比结论是否相符合？

回应：感谢您的细致的审稿！结论一致。我们开展的研究是从二级计分开始，向多级计分进行拓展。二级计分下同样对比了 U、W 以及 Lz 的检测效果，其结论是一致的。我们将部

分结果作为附录放在了修改说明的最后面。

由于二级计分下，R 与 U 是完全等价的，所以附录的结果表示的是二级计分 下 R 或 U 与 lz 的对比结果。而 W 的表现是这些统计量中最差的，故这部分结果没有放在附录里。

意见 12：第 4 页第 13 行，“高区分度项目可以有效甄别具有不用能力的被试”这句话描述不准确，反映作者对 IRT 模型的特性的理解不够。实际上，根据 ICC 曲线，对整个被试群体而言，无论是低区分度还是高区分度题目都不能实现对不同能力被试的区分，前者无法区分所有被试，后者仅能区分能力与难度一致的窄范围被试，无法区分远离题目难度的其他所有被试。因此，为避免给读者尤其是新手误导，请对这句话进行调整，用“一定范围内的”或“适当的”等形容词或许更好。

回应：感谢您的建议，我们在原文中进行了修改，用更准确和恰当的说法解释了区分度的含义。

意见 13：公式符号可以再统一下，比如 P 和 P(θ)等；

回应：已在原文中进行修改，统一了各个公式的写法格式。

意见 14：第 5 页第 20 行，“会随着 θ 值的变化而变化”，建议适当说明下是如何变化的；

回应：非常好的建议！经过对文章第二部分的调整修改，已删去有关该部分的原文。改用更精炼的说法对 Lz 进行了介绍。我们参考了文献 Drasgow, Levine 和 Williams (1985)《Appropriateness measurement with polychotomous item response models and standardized indices》 p. 73，进行了改写。

意见 15：第 7 页第 11~12 行，明确给出 U 和 W 的加权函数，这是本研究提出方法的前提，相比于回顾一堆无所谓的 GRM 等，U 和 W 的回顾应该是更重要的。

回应：这还是一个非常好的建议，我们对原文中的这部分内容进行修改，详细地介绍了 U 和 W 以及 R 统计量提出的由来，希望能够让读者更清晰地了解 R 统计量的构造过程。

审稿人 2 意见：

该研究聚焦于成就测验和问卷中异常作答行为的检测，该领域是保障教育与心理测量结果有效性的重要环节，因此，该研究问题很有意义，而且该研究从文献调研，设计指标，模拟研究，实证研究多个方面探究指标的特性和检验效果，研究方法恰当完整，模拟异常作答的条件丰富，结果呈现清晰，研究结论妥当。现存如下问题有待探讨：

意见 1：新方法的理论意义和必要性

不论从“1 引言”部分，还是“2 现有的多级计分 PFS 方法”部分，均未看到对于新方法提出的必要性的论述。

具体来讲，首先，该方法是否是一种解决已有问题的全新方法，是否有效拓展了对异常作答检测的思路和方向，从新的视角来理解和评估异常作答？如果不是全新方法或思路，那么第二，虽然文中第 6 页 15~20 行的内容简要说明了 lz 指标的局限，但是本文是否可以对上述局限进行修正和改进，是怎样从理论上考虑新指标能够与之互补使用的？是因为新指标的分布形态更标准？还是新指标用到的 θ 值更有效，不会对指标产生严重影响？还是针对 lz 的适用范围，分析了某种可以改进的空间？我们从结果部分可以看出新方法在某些异常作答检测中的优越性，但是在提出新方法之前，为什么考虑到要对 lz 进行补充和

改进，怎样做才能够实现上述改进？都没有做具体的描述，这会导致新方法的提出，没有足够的基础和依据，也会降低新方法提出的必要性和创新性。

因此，建议补充论述新方法提出的理论意义和必要性。

回应：为了更充分地表达新方法的理论意义和必要性，我们对有关的内容进行了重新组织语言和修改表述。对于新方法提出的理论意义和必要性这一点，可以看作是加权残差统计量在多级计分下的一种拓展，而非是对 lz 统计量的补充改进（这是文章初稿让读者产生的误解的地方）。现已对方法介绍部分进行大幅度的修改，修改后的方法部分着重介绍新方法的提出背景和构建缘由，并且进行了一定程度的理论分析，再将 lz 方法部分放在介绍方法后进行相对简单的介绍，目的是使用一个学界较为认可的“标杆”统计量来与文章中的方法进行比较。

意见 2：新方法提出的位置

引言和文献综述部分，应用于阐述研究背景和现有研究结果，如果要提出新指标，应当在文献综述部分充分论证现有方法的有待改进的部分之后，在文献综述后的问题提出部分，或另起一节专门介绍新方法的具体思路、算法、可能的适用范围、期待解决的问题。

因此，建议将引言部分和文献综述部分对于 R 统计量的描述，重新整理到下一节的内容中。

回应：非常好的建议！

已对原文的第二部分进行了修改调整，改变了部分章节结构，首先介绍了 R 统计量提出的背景，阐述了 R 统计量的来龙去脉，并在之后简要介绍了用以对比的 Lz 方法。

相比修改之前，删减了一些 *lz* 的介绍篇幅。这是因为：①*R* 统计量的提出并非是对 *lz* 短板的弥补或改进，而是从加权残差的角度出发，构建的新统计量。②需要突出介绍 *R* 统计量的来龙去脉，后续使用 Lz 进行比较仅因为 Lz 具有广受认可的检测能力，并且也经常被其他研究者用以进行统计量间的比较。

意见 3：语言表述需要进一步凝练和准确

回应：感谢您的建议！已对有关部分进行修改，尽可能保持了语言的准确和精炼。

意见 4：原文摘要：作弊和猜测行为在诸多异常作答行为中广受重视”，搭配不当，能否改为“更为普遍”。

回应：谢谢！已对摘要相关部分重新措辞，将这句话修改为“考虑到实际应用中异常作答行为存在多样性，并且相对于高能力考生的焦虑、疲劳等异常行为，低能力考生的作弊和猜测行为更为普遍和受到关注，因此 R 统计量在实际应用中具有良好的应用价值”。

意见 5：原文 3 页 13 行：能否增加一词，改为“部分”，否则整个句子的含义将会被理解为，所有参加测验和问卷的被试，实际作答过程都不一定能够体现被试的真实水平。这样的理解是你想表达的原意吗？

回应：感谢您的建议！文章原先想表达的意思是：正常情况下，所有被试的作答都会受到随机因素的影响，这是正常的情况。我们的意思是有一些被试的作答中包含“系统误差”，即受到了严重的异常行为的影响。

听取了审稿人的意见，并且考虑到后文承接的是对异常被试的举例，因此对原文进行了修改，改为部分被试，以避免读者产生误解。

意见 6：原文 6 页 21 行：lz 的一类错误率往往偏小，这是弊端吗？还是表述有误？

请仔细检查全文的用词和表述，尽可能做到凝练和准确。

回应：感谢您细致的审稿。对于 lz 一类错误率偏小的问题，其解释如下：如果贸然使用标准正态分布下的截断点来区分被试，lz 的实际一类错误率往往会偏小(如当设置一类错误率为 5%时，标准正态分布下的截断点为-1.645，lz 经验分布的截断点通常在-1.4 至-1.5 之间)。这会导致实际采用更严格的临界值，进而导致更小的一类错误率和检测力。这是因为能力参数估计值往往不是服从标准正态分布，对于 lz 的影响所造成的。

考虑到详细阐述该部分信息对本研究的帮助不大，为了聚焦我们的研究，并且为了不对读者产生误导，我们在修改版中删去了该部分的具体介绍。

第二轮

审稿人 1 意见：

作者在上一轮修改中回复了我的一些问题，但仍有问题没有修改。本次修改意见不再区分大或小，依次列出。

意见 1：建议修改题目，以便更好地体现出本文的主要创新点；

回应：非常好的建议！的确，经过前一轮的修改，论文的标题不能完全体现本文的工作，因此，我们拟将论文的题目修改为“多级计分测验中基于残差统计量的被试拟合研究”，希望标题能更好地体现本文所做的工作。

意见 2：文中赘述较多，不够精简；比如，正文第一句，直接引入潜变量建模的作答真实性假设即可；再比如，用较多篇幅描述“非参数化的 PFS 如 Molenaar(1991)...”有什么必要性？可以考虑用“(e.g., Molenaar, 1991; XXXX)”这样的方式来精简内容；比如，一个 GRM 用了 4 个公式的篇幅来介绍；一个 lzp 统计量的介绍需要用这么多篇幅吗？

回应：感谢您提出的建议，本文在书写方面确实不够精简。我们对所涉及到的部分进行了精简，对其它的一些不够精简的地方也进行了修改。此次修改已经将引言至方法部分进行了大篇幅重写，重新组织了引言和已有方法及模型的介绍。

意见 3：引言部分仍未很好地引入研究问题： (1)为何要提出 R 统计量；换句话，为何不提出基于别的方法的统计量，而恰恰是基于残差的？ (2)为何要与 lzp 进行比较？lzp 为何有资格作为一种对比方法？ (3)为什么要做多级评分题目的？就因为“针对多级计分的个人拟合研究不多”？不多很可能是不重要，没有研究的必要性； (4)上述内容都没有在引言中得到较好的阐述，给审稿人一种“就是做这个，没有理由”的读后感；

回应：感谢您提出的建议，我们重新梳理了行文的逻辑，将异常检测研究的重要性，已有研究的缺陷和提出的目的进行了重新阐述。希望能够更明确地表达本研究的理论和应用意义。

意见 4：文中仍未明确 M 是什么；类似问题在文中多次出现，比如公式 5~7 中的 w_0, w_j 和 v_j 是什么？什么是“适宜的函数”？感觉作者的攥写逻辑有些混乱，比如在 2.3 节中突然又蹦出公式 20 和 21 来跟 2.2 节中的公式 5 进行联系；

回应：再次感谢您的意见！这里 M 是指测验长度。此次修改对方法介绍部分进行了较大规模的修订，重新调整了公式有关的说明和介绍。

意见 5：第 10 页，什么是“零分布”？
回应：零分布(null distribution)指的是在满足零假设(即考生为正常作答)的条件下，检验统计量的分布。

意见 6：请使用正确的参考文献格式，自行查阅投稿指南；
回应：感谢您的建议，已经对参考文献部分进行了全面的检查，修改了一些不规范的地方。

意见 7：公式 27 更换为类别分布，这并不涉及程序，只是替换表达方法。当前这种表达方法“不专业”，明明一个简单的类别分布就可以表达清楚，还要费劲占用一个公式和篇幅
回应：您的建议非常好，修改后的该部分采用了更简洁的语言来进行表达。

意见 8：表 2 中为何选取前后 37.5%，而不是 CTT 中计算难度区分度常用的 27%？
回应：感谢您细致的审稿！我们这里是参考了前人研究(Doval & Delicado, 2020)。
Doval, E., & Delicado, P. (2020). Identifying and classifying aberrant response patterns through functional data analysis. *Journal of Educational and Behavioral Statistics, 45*(6), 719-749.

意见 9：上一轮审稿意见中已经提及，表 2 中一些所谓的“类型”不过是比例上的不同，审稿人觉得没有必要区分这么清楚（易误导读者，让读者认为本文所提出的方法可以区分这些类型）；在实践研究中，数据分析者能根据结果区分出“作弊”和“幸运猜测”吗？
回应：感谢您的建议，我们对原文该部分进行了重写，将这些异常类型总结为了三个类别，并且为了避免让读者产生误导，后续文中不再对数据分析的结果进行类似解读。

意见 10：结果解释部分需要加强，比如，对研究 2 结果，不是简单描述下“但是它们对于探测其他类型异常行为的检测力则没有这么高”，尽可能从方法本身或其他方面解释这个结果。
回应：您的建议非常好，经过修改，我们在文章结果部分中对结果的描述都尽量附上客观的解读和解释。

意见 11：实证研究中，作者需要思考所提出方法的实践应用可行性，或者是该方法的实践适用范围（谁来用？谁会在用统计量前还要先做下模拟来计算 PFS 的临界值？）
回应：R 可以适用于分析任何测量潜在特质的测验，包括能力选拔测验和心理评估测验。

大部分的拟合检验统计量都并非完全标准化，即使是l_z也并非服从渐进正态分布，因此我们在使用前都需要获得其分布和临界值。这也是未来的一个研究方向，开发 R 的改良版本，得到其理论分布或渐进理论分布，力求省去这一步骤。

审稿人 2 意见：

本研究内容充实，过程清楚，结论可靠，将基于加权残差的统计量由二级计分扩展到多级计分，非常有意义，是很有价值的研究，可以看出作者付出了很多努力。不过该文在内容呈现和文章结构编排上尚存在问题，承接上次意见中提到的不同内容的位置顺序，对于整体行文的结构再次做出详细的说明和建议。

本文内容丰富，但是叙述不够紧凑，逻辑不够连贯，重点不够突出，容易让读者很难把握作者的意图和重点。下面我将详细对前言部分的行文结构提出具体建议，请作者对整体结构充分考虑，并参照已发表论文的行文逻辑，尝试学习和迁移。并对后文进行同样的

反思和修改。这一步无疑是有困难的，但也是作为科研人员必要的过程，也是科技论文的规范。这并不是根据建议补充或修改具体的位置就可以完成的，需要您充分梳理要表达的重点内容，进而做到删繁就简，围绕主题，前后呼应，逻辑连贯。本人下面的建议针对引言部分，仅用作抛砖引玉，以表明对文章结构可能的理解和编排过程，后文其他部分，也建议梳理清楚结构，更加清晰地表达。具体采用何种编排方式，尊重作者原意。

回应：感谢您耐心的审稿，我们对全文的很多内容进行了精简和重写，对全文所涉及到的批注部分进行了一一修改和回复，希望能使逻辑更清晰，大幅度提高书写质量。

意见 1：引言部分一般用于引出本文的主题，例如本研究的引言部分包括的主要内容及其逻辑序列应当如下：

异常作答是什么，为什么会出现异常作答（简单精准直接的概述）。可以参照如下方式进行简洁的叙述：教育和心理测量的最主要目的，就是评估受测者能力、兴趣和态度等内隐特质，然而实际测验中，往往不可避免地会有其他因素影响考生的作答反应，进而威胁测验结果的有效性。影响因素一般包括粗心、作弊、疲劳、焦虑等。

这部分内容，您的论文第一页 12-18 行中都已经涉及到了，但是语言不够凝练，信息分散，用语专业性有待提高。而且，上述影响因素论述完毕之后，在 27 行之后，间隔了其他内容，再补充说明影响因素是系统误差的内容，容易让人产生思维跳跃、逻辑混乱的感觉。个人建议断尾的补充是不需要在文中赘述的。

回应：非常好的建议，我们对全文进行了细致的修改和重写，希望能够在语言的凝练和专业性上有大的提高。

意见 2：异常作答的严重程度和危害（简要说明情况即可，表明研究的必要性）

可以包含如下两部分内容：异常人数比例（严重程度）+决策偏差（危害），这两点可以引证充分一些，才能表现当前情况的严重性和紧迫性，突出研究的必要性。（第一页 18-25 行）

回应：感谢您的建议，我们进行了对应修改。对已有研究中对异常作答的流行程度（prevalence）及其负面影响等内容进行了补充。

意见 3：概括性说明学术领域开展了怎样的工作检测异常作答。利用总分总的结构，先说有几类方法，然后逐一介绍，最后汇总其贡献和不足。第一页 25-26 行，是承接句（“因此，如何将包含异常作答行为的个体筛选出来，一直是教育和心理测量领域所广泛关注的问题”），然后缺少了该部分内容的总领句，总领句大致如下：“对此，众多研究者提出了两类个人拟合指标，包括参数个人拟合指标和非参数个人拟合指标。”然后依次介绍两类指标的含义和举例。

这部分内容的功能是，介绍异常检测方法的基本思路和原理，然后用几个指标举例，最终落脚点在于当前研究的有待发展的方面。

然后详细的方法介绍放在下一节的内容中。

回应：非常感谢您的建议，我们对已有的工作进行了更为概括的综述，力图使综述具有广度和一定的深度，尽力去让读者明晰本研究有关领域的概貌。并在此基础上提出当前研究的不足，及本研究的动机和目的。

意见 4：此外，第 3 页 15-29 行的内容，叙述的是关于多级计分的个人拟合研究，但是这样简单罗列无助于突出本研究的研究问题。阐述以往研究时，一方面要详略得当，与本研究关联不大的指数，则统一根据原理进行分类，然后简单列举即可；与本研究有关的指数，必须要详细说明原理、效果和局限，这是非常重要的。对于您的研究来说，必须要详细说

明l_{zp}和加权残差统计量的特点、优势和局限，为你研究中提出新指标做铺垫。另一方面介绍多个指标要具有一定的逻辑连贯性，诸多指标之间是什么关系，是否每个指标都是在前一个指标基础上改进而来的，然后发展到现在，尚有未解决的问题。这样才能看到作者对问题的理解和思考过程，简单罗列是无法体现这个过程的。

回应：非常好的建议，我们对这部分内容进行了重写，尽力把我们对有关问题的思考和理解，提出我们方法的动机和过程体现出来。

意见 5：最后顺势引出本研究的目的和要解决的问题。借着以往研究未能覆盖的范围或未能实现的功能，提出为了达成上述目的，结合 lzp 和加权残差统计量，合成某种有助于凸显异常作答的、适用于低能力者的、结构简单的新统计量。

每一部分内容都不是简单罗列现有研究，而是紧紧围绕本研究要解决的问题来论述，基于本研究的问题，删除冗余，保留重点。可以分段逐一阐述，使得结构清晰，易于理解。前文已经清楚叙述过的内容，如无必要，不再赘述。强烈建议下载《心理学报》心理测量领域的其他论文，以及作答异常检测的其他论文，仿照其文章结构编排，结合本文内容，对文章其他部分也请逐一梳理，使叙事结构更加清晰。

其他建议请见文中批注。

回应：非常感谢您的建议，我们充分考虑的方法的提出逻辑，并对其介绍部分进行了修改，包括引言、已有模型介绍、方法部分，重新设置了子章节，力求语言准确干练，逻辑清晰，条理通顺。

批注 1：第 1 页 7 行：可考虑改为“焦虑、疲劳引发的异常作答”

回应：我们在修改稿中进行了对应修改，并且对原有摘要进行了重新阐述，删去了一些冗余的部分。

批注 2：第 3 页 9 行：本研究是否要通过指标分析不同异常作答的类型？如果不是，则没必要针对以往研究的这项不足开展论证。

回应：非常好的建议，在考虑了引言部分的行文逻辑后，这部分关于异常类型的文本描述确实没有必要，所以删去了这部分。

批注 3：第 4 页 14 行：这部分的第三个标题“2.3***Lz*** 统计量”，且在该部分说明了 LZ 统计量是二级计分统计量，放在这个大标题下不合适。建议重新编排该部分内容，具体来说：将 LZ 统计量放在二级计分的介绍段落中介绍，将其在多级计分题目上的变式放在 Lzp 放在本段中介绍。

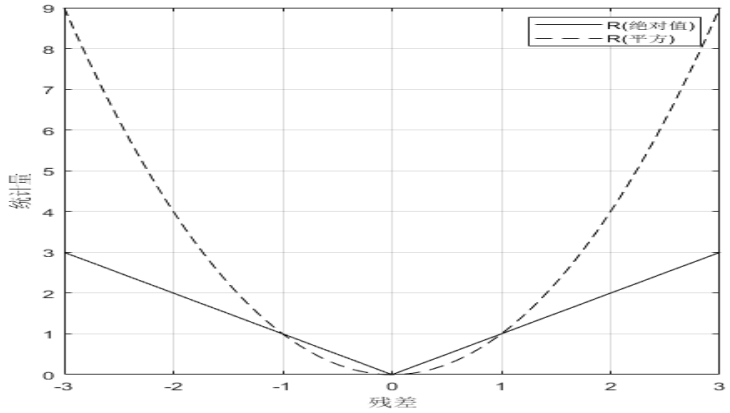
回应：我们在修改版本中重新编排了已有方法和模型的介绍部分，该部分介绍了多级计分下 IRT 模型和另外两种个人拟合统计量，对于二级计分有关的内容进行大幅度的降低。

批注 4：第 6 页 19 行：如果这一节是文献综述部分的话，仅需介绍和阐述现有指标和方法，并且点名现有方法尚未能解决的问题是什么。包括下面介绍 lz 的过程也是相同的。提出新方法的部分要放在后面单独的专门一节来介绍。

回应：感谢您的建议，我们在修改后的版本中将已有方法和提出的方法两部分更清晰地区分开来，第二章介绍已有方法，第三章介绍提出我们的方法。

批注 5：第 7 页 28 行：建议最好以图形方式呈现出来。

回应：感谢您的建议！为了使文章的表达更聚焦和凝练，我们对这部分的表达进行了较大规模的重写，在修改后的版本中删去了该部分内容。下图给出了残差统计量取绝对值和平方时，统计量值随着残差变化的散点图。



为了使内容更加聚焦，在修改后的版本里，我们删除了原版本的第 7 页 28 行。

批注 6：第 8 页 12 行：为避免歧义，建议改为“会进一步掩盖实际作答与真实作答的差异”，您看是否是您要表达的含义。
回应：感谢你的建议！修改后改用更规范和通用的“掩蔽效应”(masking effect; Fung, 1993; Yuan & Zhong, 2008)来进行表达。

批注 7：第 8 页 13 行：对于 R 指标的两种形式的选择和探究，需要详细数据或图表表示出来，建议最好作为“预研究”或“研究一”来呈现，详细描述和解释结果，并尝试引用其他研究说明该测试结果产生的依据。

同时，建议详细解释，为什么 U 指数和 W 指数能够采用平方的形式，而 R 指数使用平方的形式却效果不好？
回应：非常好的问题！R 统计量的提出参考了 Snijders(2001)提出的二级计分下个人拟合统计量中常用线性形式，这种形式在个人拟合统计量中得到广泛认可，该线性模型认为统计量应该只与得分本身以及能力值 θ 有关，统计量与得分呈线性关系(项目水平上)，所以 R 统计量仍然沿用了该线性形式。

U 和 W 之所以能够使用平方的形式，是因为它们在二级计分下，得分(0-1)的平方等于其本身，因此平方之后的 U 和 W 同样满足上述的线性形式。

为了使得内容更加聚焦，避免混淆，文章除引言部分，修改后的版本对 U 和 W 不做详细介绍。

批注 8：第 8 页 18 行：这段内容应该紧跟在 R 指数公式的前后，放在这里跟上下文不连贯。请对全文其他部分逻辑进行梳理，划分不同的内容模块，每部分内容尽可能做到聚焦和集中。
回应：非常好的建议，修改后重新编排了内容，力求做到紧凑和集中。

批注 9：第 24 页 2 行：19 个测量情绪状态的题目是否测量了同一维度的情绪？这里能否详细说明是什么情绪状态的题目？如果不是同一维度的情绪的话，那么在情绪维度一上得高分，维度二上得低分，也是合理情况，不算异常作答。

回应：谢谢您的意见！关于这 19 个题目的单维性，我们在后文中进行了补充了参考文献和说明，该 19 个项目均测量同一维度的情绪状态。

批注 10：第 25 页 4 行：能否在下一段具体分析一下，同时违反两个指标的 253 名考生的作答有何特点？单独违反 R 统计量的考生有何特点？单独违反 lzp 统计量的考生有何特点？能否尝试用数据方式说明。如果不能，尝试文字说明。

回应：修改了关于实证数据分析结果的解释部分，用表格形式列出了几名三种情况下被标记的异常考生案例，并对其进行分析。具体结果如下：

为了更一步考察两个统计量的表现，我们对被检测出的考生的作答进行分析。分别选取了三类根据统计量标记为异常的被试共 15 人的作答模式，如下表所示

表 9 标记为异常的被试作答模式

类型	被试序号	作答向量
共同	15	(2,1,0,0,0,0,0,0,0,0,0,2,0,0,2,0,0,0,0)
	23	(0,1,0,2,1,0,0,1,2,0,2,2,1,0,2,2,2,0,2)
	48	(2,0,2,2,1,2,0,2,2,1,2,1,2,2,2,2,2,0,2)
	49	(1,0,0,2,0,0,0,0,0,0,0,0,0,0,0,1,0,0,2)
	50	(2,2,2,0,2,2,2,1,0,0,1,2,0,0,1,0,0,0,0)
R	43	(0,0,2,2,0,0,1,1,0,1,1,0,0,0,0,0,0,1,0)
	45	(0,0,0,2,1,2,0,0,0,1,0,0,0,0,0,1,0,0,0)
	99	(1,0,0,1,0,1,1,0,0,2,1,2,0,0,0,0,1,1,0)
	108	(0,0,0,1,0,0,0,2,0,0,1,0,0,0,2,2,0,0,0)
	114	(0,2,0,0,1,0,0,0,0,0,0,0,0,2,0,0,0,0)
lzp	6	(1,0,0,1,2,0,2,0,1,0,2,1,0,0,1,0,1,0,1)
	36	(0,0,2,2,1,2,0,2,1,0,2,1,2,0,2,1,0,0,0)
	52	(1,0,2,2,2,0,2,0,1,2,2,1,2,0,2,2,2,1,0)
	55	(2,2,2,1,2,2,1,2,0,2,1,0,2,0,1,2,0,1,0)
	101	(0,1,0,1,0,2,1,2,1,2,2,0,1,0,2,1,1,1,1)

注：“共同”表示同时被 R 和 lzp 标记为异常，“R”表示仅被 R 标记为异常，“lzp”表示仅被 lzp 标记为异常

可以发现：(1)同时被两种统计量标记异常的被试，都具有较为明显的异常状况，得分较为极端，如第 15、49 号被试，其估计能力较低，然而他们都在不止一个项目上作答为 2，不符合低能力者的正常作答模式。再如第 48 号被试，其在大量项目上作答为 2，估计能力应该偏高，但是却在第 2、7、18 题上作答为 0，也同样不符合高能力者应有的作答模式；(2)仅被 R 标记的异常被试群体，具有较为明显的特点，即作答中包含大量的 0，这说明这部分被试以低能力者居多，但是却能在少数项目上获得 2，属于异常高能力表现。R“专注”于将这部分被试筛选出来，也与前文中介绍的 R 的特点相一致；(3)对于仅被 lzp 标记的被试，他们的作答结果较为“均匀”，这可以理解为不同于 R 对低能力者的异常高能力表现的特异性，lzp 对不同异常类型检测的覆盖面更广。

第三轮

审稿人 1 意见：经过再次修改，作者较好地回复了本人的问题。文章质量有所提高，建议作者进一步在讨论中补充实践应用建议，即站在实践者而非研究者的角度思考该方法的使用建议。

回应：感谢您一直以来的建议！我们对讨论部分进行了补充，添加了有关实践应用的建议和说明，以丰富其内容和对使用者的参考价值。

审稿人 2 意见：

本文在针对上次意见修改之后，整体行文更加顺畅，内容连贯，语句凝练，可以看出作者付出了大量努力进行修改，整体不存在明显问题。只有几个极小的问题，已标记在文中，请作者查看。

意见 1：适应性测量指的是什么？在文中和摘要中好像都没有体现。

回应：非常感谢您的建议，适应性测量 (appropriateness measurement; Meijer & Sijtsma, 2001)是一些文献中用到的术语，是用来指代被试的拟合检验，即 person-fit。

文章修改前曾在引言部分介绍过适应性测量这个概念，修改后则删去了这部分内容，因此关键词中的“适应性测量”也不再适合出现在目前版本的文章中，现已将其删除，并对关键词进行了重写。

意见 2：引言是针对背景和现状的介绍，其中需要铺陈与本研究有关的内容，但通常并不在引言中直接表明本研究的内容。因此，这句可以改为类似于如下的表述：“因此，亟待开发一种针对异常高能力敏感的多级计分 PFS，用于避免严重的测验安全问题。”

回应：非常好的建议，我们修改了该部分的表述，从背景和现状的角度出发，说明本文研究的意义所在，并且将与本研究有关的内容都调整至第三章进行介绍。

意见 3：基于残差的 PFS 具有对更高评分等级的敏感性，这句话的依据是什么，能否在文中作出解释，如何发现这一特点的？

回应：这句话的依据在于正式研究前，对现有的部分 PFS（包括二级计分下和多级计分下）进行了大量的模拟实验，比较相同条件下它们的检测性能，从结果中总结出的规律。

为了更准确地阐述这一点，我们在文中添加了说明，与此有关的表述被安排在第三章。

意见 4：这一段整体都属于问题提出和研究框架的介绍，建议放在：3 基于加权残差的多级计分 PFS 开发。”

回应：非常好的建议，我们将该部分按照内容拆分，进行更为合理地配置。有关方法提出的内容放置在第三章开头，有关研究框架的内容则放置在第四章开头，希望新的结构能让读者更容易理解。

意见 5：能否在“掩蔽效应”这句后面，简要解释一两句具体的含义。

回应：非常感谢您的建议，为了让读者更加清晰地了解，我们在掩蔽效应后添加了解释。

意见 6：长代表 60，较长代表 80，是这样吗？一般态度测量问卷中，多级计分的几个选项规律是：不同意，比较不同意，不清楚，比较同意，同意。仅供参考。
回应：非常感谢您的建议，为了避免歧义，更改了描述中长和较长的位置，此时较长代表 60 个项目，长代表 80 个项目。研究之所以选用五等级计分，一方面是为了选取一个相对适中的计分等级进行后续研究；另一方面也确实参考了常见的李克特量表的计分特点。

意见 7：心理学上一般叫做“观测分数”，或“观测得分”，供作者参考。
回应：非常感谢您的建议，现已将文中涉及的名词进行了修改，统一为观测得分。

意见 8：分布的一端通常叫做“单侧”
回应：非常感谢您的建议，使用“单侧”是更为专业简练的表达，已对原文中该部分进行修改。