《心理学报》审稿意见与作者回应

题目：用于处理不努力作答的标准化残差系列方法和混合多层模型法的比较
作者：刘玥，刘红云

第一轮

审稿人 1 意见：

文章对识别不努力作答的标准化残差系列方法和混合多层模型法进行随机模拟比较，展示了各方法在不努力作答规模、不努力作答严重性、反应时差异等指标下的表现，并给出不努力作答严重性高时，固定参数迭代标准化残差法最优的结论。本文的研究发现有助于研究者在存在不努力作答的测验模型中更精确的估计参数，具有较好的创新性和实用性。文章脉络清晰，逻辑通顺。但是，作为方法比较文章，作者应更加全面的讨论方法优劣，更加明确的给出方法选择指导。以下是具体意见和建议。

意见 1：表 1 报告了 TPR(正确识别不努力作答的比例)、FDR(错误识别不努力作答的比例)、Pr(识别出不努力作答比例)。尽管 CSRI 的 TPR 大，表现好，但 MHM 的 FDR 最小，表现好。作者在重点关注正确识别率的同时(如表 9)，也应同时讨论错误识别率及其带来的危害。比如将努力作答识别为不努力，并对其替换为缺失，可能会提高参数估计的标准误，是否能在参数估计章节进一步讨论。此外，当 pi 不为 0 时，是否也可以报告错误识别不努力作答占所有努力作答的比例（当真实情况为努力作答，而被识别为不努力的概率），以考察类似 I 类错误率；1-TPR 相当于 II 类错误率，读者对于熟悉的概率会有更直观判断。

回应：同意审稿专家的意见，错误识别不努力作答的确会带来一定的危害。但是，本研究关注不努力作答的处理，其目的主要是提高参数估计结果的准确性。通过本研究和同类研究发现，**TPR 对参数估计结果的准确性有重要影响，而 FDR 几乎不影响参数估计结果的精度**。具体来说，一方面，从表 2、表 5-表 7 可以看出，各条件下 MHM 方法的 FDR 均最小，CSRI 方法的 FDR 均最大。但在很多条件下 CSRI 方法得到的参数估计准确性要优于 MHM 方法。总体来说，各条件下参数估计准确性高的方法，其 TPR 也高。另一方面，同类研究也表明了 TPR 与参数估计结果准确性的相关。例如，在 Wang, Xu, Shang 和 Kuncel（2018）的研究中，混合多层模型方法和贝叶斯残差法的 FDR 都较小（都在 0.12 以下），两种方法在识别准确性方面的差异主要体现在混合多层模型方法的 TPR 明显高于贝叶斯残差法。这

也最终反映在混合多层模型方法参数估计准确性明显高于贝叶斯残差法。基于此，本研究主要从正确识别率方面评价各方法的识别准确性，错误识别率仅起到参考作用，因此不再采用多个指标反映错误识别的问题。审稿专家所提出的错误识别不努力作答占所有努力作答的比例（当真实情况为努力作答，而被识别为不努力的概率）的确是一个类似于第 I 类错误的概念。但是，该指标与 FDR 相比仅在分母上不同，其结果应当具有很大的一致性，同样对估计结果的准确性影响小。此外，该指标还存在一个问题，即在不同的条件下，计算的分母不同（真实的努力作答在整个数据中的比例不同），是否都可以参照常见的 I 类错误率的标准（0.05）进行解读还是值得考虑的问题。综上，基于同类研究（Wang, Xu, & Shang., 2018; Wang, Xu, & Shang et al., 2018）的评价指标，本研究主要参考 TPR 评价识别准确性，同时结合 FDR 指标反映各方法错误识别的情况。

另外，审稿专家提出关注错误识别率及其带来的危害，对我们很有启发。我们已经在修改稿中的表 12 中加入了有关错误识别率的结果总结，并在讨论中加入了错误识别率的总结及其影响，具体请参见讨论部分第 35 页第一自然段标红文字。本研究中，标准化残差法系列方法对于所识别出的不努力作答是采用替换为缺失的方式处理。这种处理方式类似于缺失值处理中忽略的方法，其背后隐含了一个假设，即不努力作答形成的缺失是随机缺失（missing at random, MAR）或者完全随机缺失（missing completely at random, MCAR）(e.g., De Ayala, Plake, & Impara, 2001; Rose, von Davier, & Xu, 2010)。也有研究证明，在缺失比例不太大的情况下，即使缺失机制是非完全随机缺失（missing not at random, MNAR），采用忽略方式得到的参数估计结果也是可以接受的(e.g., Custer, Sharairi, & Swift, 2012; De Ayala et al., 2001; Holman & Glas, 2005; Köhler, Pohl, & Carstensen, 2017; Rose, 2013; Rose et al., 2010)。并且，很多大规模教育测验都采用忽略的方式处理缺失值(e.g., OECD, 2009)。而 MHM 方法是采用对努力作答和不努力作答分别建模的方式处理，在识别出不努力作答比例不大的情况下，即使错误识别了不努力作答，可能会影响不努力作答部分模型的题目参数估计值（这部分可利用的数据本来就少，加入了错误的信息带来的参数估计偏差大），而努力作答部分模型的题目参数估计值和能力参数估计值受到影响较小。**综上，可以推测在本研究中，即使错误的删除了努力作答的数据（即 FDR 高），也应当不会带来较大的参数估计误差。**此外，根据审稿专家的建议，我们计算了 MHM 和 CSRI 在 FDR 上差异最大的条件（情境 1，π为 20%，π_i^{non}低，d_{RT}小）下两种方法参数估计的标准误，结果如表 R1 所示

表 R1 情境 1 中π为 20%，π_i^{non}低，d_{RT}小条件下两种方法参数估计标准误

	a	b	α	β	θ	τ
CSRI	0.109	0.054	0.389	0.016	0.283	0.087
MHM	0.117	0.052	0.370	0.017	0.279	0.094

从表 R1 中可以看出，两种方法各参数估计的标准误较为接近，因此可以推测，**在本研究中所识别出的不努力作答比例不大的情况下，错误识别率应当不会对参数估计的标准误造成较大影响。**

参考文献

Custer, M., Sharairi, S., & Swift, D. (2012, April). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation.* Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, BC, Canada.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of educational measurement*, *38*(3), 213–234.

Holman, R., & Glas, C.A.W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with Item Nonresponse in Large-Scale Cognitive Assessments: The Impact of Missing Data Methods on Estimated Explanatory Relationships. *Journal of Educational Measurement*, *54*(4), 397–419.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (PhD thesis). Friedrich-Schiller-University, Jena, Germany.

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), 1-53.

Wang, C., Xu, G., & Shang, Z. (2018). A two–stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254.

Wang, C., Xu, G., Shang, Z., & 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.

意见 2：在模拟研究中，作者考虑了不努力作答规模，不努力作答严重性，反应时差异。在主要结论中，文章围绕不努力作答严重性高低讨论方法选择。鉴于该指标在指导方法选择中的重要性，建议考虑更多的作答严重性比例。如固定其他模拟指标，随着严重性比例的升高，各方法表现的趋势图，为实际使用提供指导。如当研究者估计不努力作答严重性比例为 8% 时，应该如何选择方法？

回应：同意审稿专家的观点。不努力作答严重性确实是影响方法表现较重要的因素。因此，在情境 1 中，π为 40%， d_{RT}大条件下，增加不努力作答严重性的比例的水平，形成无不努力作答，不努力作答严重性低（$\pi_i^{non} \sim U(0, 0.25)$），中（$\pi_i^{non} \sim U(0.25, 0.5)$），高（$\pi_i^{non} \sim U(0.5, 0.75)$）共四个条件。在不同水平下比较四种方法得到的结果。首先，表 R2 呈现了不同方法的识别准确性结果。

表 R2 情境 1 中不同不努力作答严重性条件下各方法识别准确性结果

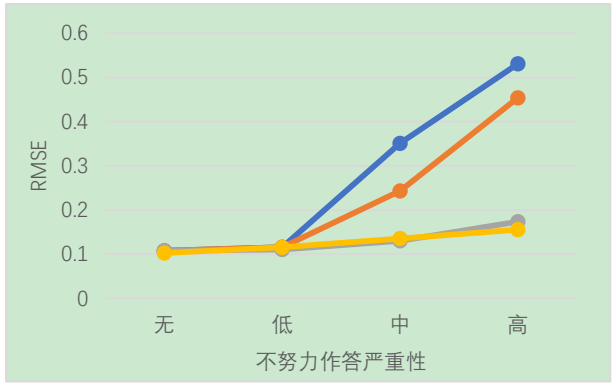
π_i^{non}	指标	OSR	CSR	CSRI	MHM
无	FPR	0.047	0.048	0.060	0.000
低（0.050）	TPR	0.870	0.865	**0.939**	0.913
	FDR	0.166	0.157	0.181	0.089
	Pr	0.052	0.051	0.057	0.050
中 (0.150)	TPR	0.473	0.734	**0.952**	0.945
	FDR	0.020	0.114	0.127	0.079
	Pr	0.072	0.124	0.163	0.154
高 (0.250)	TPR	0.167	0.494	0.931	**0.937**
	FDR	0.029	0.172	0.136	0.070
	Pr	0.043	0.149	0.269	0.253

注：固定的模拟条件为π=40%，d_{RT}=大。TPR 表示正确识别率，FDR 表示错误识别率，FPR 表示误检率，Pr 表示努力作答占所有作答的比例。π_i^{non}一列中括号内数字表示真实不努力作答的百分比。加粗的结果表示每种条件下 TPR 最高的结果。

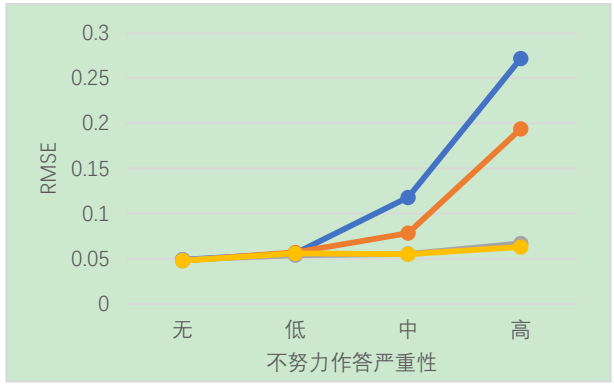
从表中可以看出，随着不努力作答严重性增加，在 TPR 指标上 MHM 和 CSRI 方法相对于另外两种标准化残差法的优势不断增加，并且，MHM 方法与 CSRI 方法之间的差距也不断缩小。并且，总体来说，MHM 方法的 FDR 较低。

其次，图 R1 呈现了随着不努力作答严重性增加，各方法参数估计 RMSE 的趋势图。各方法在各条件下参数估计偏差和 RMSE 的具体结果见表 R3。从图中可以看出，当不存在不努力作答时，MHM 方法得到的时间区分度参数估计值 RMSE 明显小于其他方法；各方法得到的其他参数估计值 RMSE 较为接近。随着不努力作答严重性增加，OSR 和 CSR 方法得到的参数估计值 RMSE 增加，而 CSRI 和 MHM 方法得到的 RMSE 基本稳定，它们和另外两种方法的差异逐渐增大。

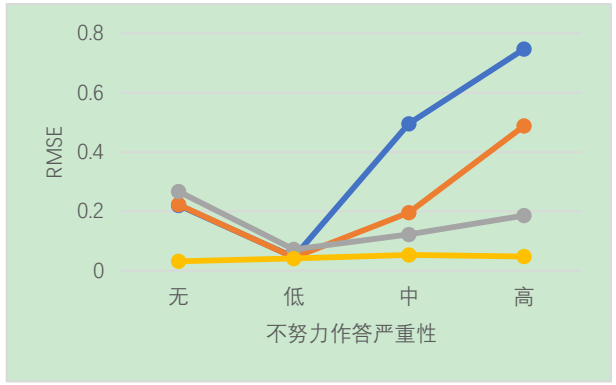
这些结果与原文中模拟研究的规律是一致的。



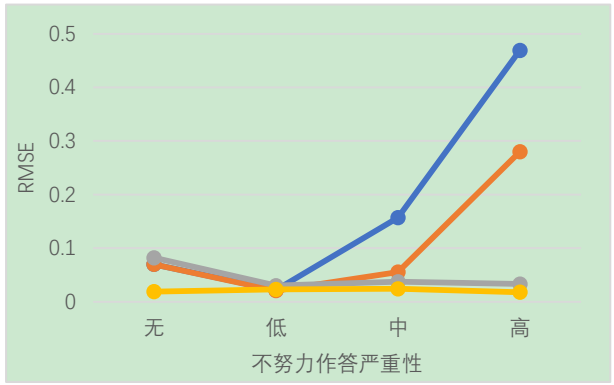
区分度参数



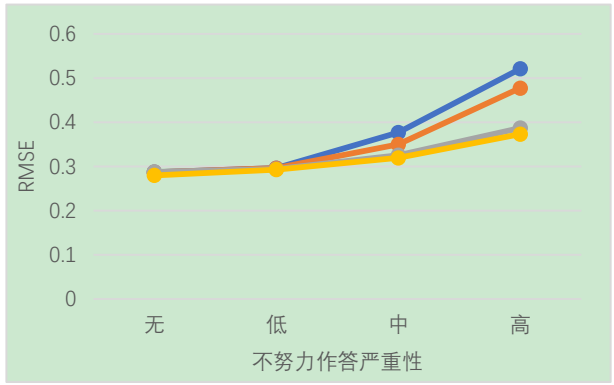
难度参数



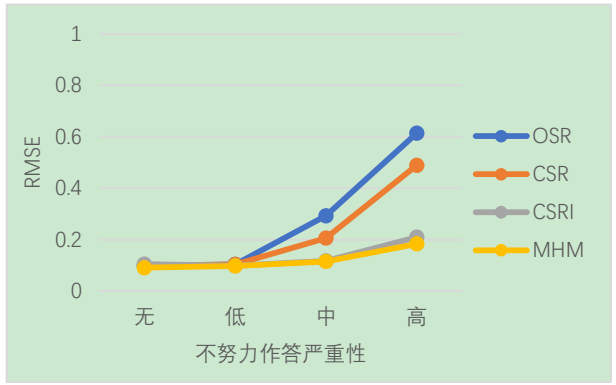
时间区分度参数



时间密度参数



能力参数



速度参数

图 R1 不同不努力作答严重性条件下各方法参数估计 RMSE

注：固定的模拟条件为π=40%， d_{RT}=大。

表 R3 不同不努力作答严重性条件下参数估计准确性

π_i^{non}	方法	bias						RMSE					
		a	*b*	*α*	*β*	*θ*	*τ*	*a*	*b*	*α*	*β*	*θ*	*τ*
无	OSR	-0.012	0.001	-0.214	-0.068	-0.004	-0.014	0.108	0.049	0.22	0.07	0.287	0.099
	CSR	-0.012	0.001	-0.217	-0.069	-0.005	-0.014	0.108	0.048	0.223	0.07	0.287	0.099
	CSRI	-0.012	0.001	-0.26	-0.081	-0.005	-0.014	0.108	0.049	0.267	0.082	0.288	0.105
	MHM	0.000	0.000	0.000	0.000	0.000	0.000	0.103	0.048	0.032	0.019	0.28	0.091
低	OSR	0.034	-0.011	-0.021	-0.017	-0.005	-0.015	0.116	0.056	0.046	0.022	0.296	0.104
	CSR	0.036	-0.011	-0.016	-0.016	-0.005	-0.015	0.116	0.057	0.045	0.021	0.296	0.104
	CSRI	0.016	-0.007	-0.059	-0.027	-0.005	-0.015	0.111	0.054	0.072	0.03	0.294	0.099
	MHM	-0.028	-0.007	-0.018	-0.02	-0.005	-0.015	0.116	0.056	0.041	0.023	0.293	0.098
中	OSR	-0.275	0.085	-0.475	-0.155	0.004	0.015	0.351	0.118	0.495	0.157	0.377	0.293
	CSR	-0.181	0.045	-0.182	-0.052	0.005	0.015	0.243	0.079	0.196	0.055	0.350	0.206
	CSRI	-0.046	0.009	0.113	0.034	0.005	0.015	0.131	0.055	0.122	0.037	0.325	0.118
	MHM	0.053	0.007	0.033	0.020	0.005	0.015	0.136	0.055	0.053	0.024	0.319	0.116
高	OSR	0.351	-0.215	0.719	0.468	-0.004	-0.016	0.531	0.272	0.747	0.469	0.521	0.614
	CSR	0.333	-0.153	0.473	0.278	-0.004	-0.015	0.454	0.194	0.488	0.28	0.477	0.489
	CSRI	0.098	-0.032	-0.176	-0.029	-0.004	-0.014	0.174	0.067	0.186	0.033	0.387	0.21
	MHM	-0.063	-0.024	-0.012	0.009	-0.005	-0.015	0.156	0.063	0.048	0.018	0.373	0.184

综上，随着不努力作答严重性增加，CSRI 和 MHM 方法的优势增大。因此，当不努力作答严重性为中或高时，建议选用 CSRI 和 MHM 方法，尤其当不努力作答严重性为高时这两种方法的优势更强。已经在文中讨论部分加入了相关结果及相应的方法选择建议。具体请参见讨论部分表 12 上方标红文字，以及表 12 中标红文字。

意见 3：在实证研究中，5.2.2 节的估计结果比较，作者考虑各方法与基于原始数据得到的参数估计值的差异。但是，基于原始数据得到的参数估计值既不是真值，也不是最优值，那么比较的意义是什么？随后，作者说因为使用原始数据会高估或低估参数，推测方法对偏差有一定的修正。但是，在实证研究中，我们并不知道不努力作答的真实情况和比例，那么是否会有过度修正的情况？注意，模拟研究的图 1-6 均只考虑了不努力作答比例最高的情况。此外，是否可以根据自陈量表估计不努力作答比例，推荐合适的方法？

回应：同意审稿专家的观点，实证研究中真值未知，无法通过各方法结果与原始数据的比较得到修正偏差的推测，个别方法的确有可能存在过度修正的情况。实证研究的目的主要有两个：**一是利用自陈量表的结果对各方法不努力作答识别的结果进行效度验证（见 5.2.2）；二是与模拟研究对应，考察和比较各方法参数估计结果的差异。**由于模拟研究结果发现，当数据中存在不努力作答时，使用原始数据拟合模型会得到有偏差的估计结果，而使用标准化残差系列方法与混合多层模型法能够在大部分情况下减小参数估计的偏差。因此，我们选择原始数据作为比较的基线（可能是误差最大的情况），考察各方法参数估计结果和原始数据的差异。根据审稿专家的意见，我们在修改稿中首先根据测验属于低利害测验，测验长度较长的特征，预估数据中可能存在较为严重的不努力作答行为。然后，统计了数据中自陈量表得到的随机猜测比例，与模拟研究设置的条件比较，认为实证数据中的不努力作答应当略大于模拟研究中严重性低的情况。其次，考察发现所有被试在所有题目上的对数反应时分布都呈现出双峰分布的特点，说明在所有题目上都普遍存在不努力作答（Meyer, 2010; Wang, Xu, & Shang et al., 2018; Wise & Kong, 2005）。因此，总的来说数据中应当存在略为严重的不努力作答现象。此时，各方法得到的结果差异应当略大，采用 CSRI 或 MHM 方法可能是较好的选择。具体请参见“5.2 实证研究结果”下方标红文字。然后，在参数估计结果部分删除了有关偏差修正的描述，仅关注各方法参数估计结果与基于原始数据估计结果的差异。最后，通过 MHM 方法的识别结果，计算两种作答反应时差异，结合之前随机作答比例的报告结果，综合推测实证研究的数据接近模拟研究中π_i^{non}高d_{RT}大的条件。参考模拟研究结果，此时 MHM 和 CSRI 方法表现应优于 CSR，OSR 方法。这一点也可以从实证研究中效度验证的结

果，参数估计差异的结果上得到印证。具体请参见实证研究最后一段标红文字。

参考文献

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538.

Wang, C., Xu, G., Shang, Z., & 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.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

意见 4：作者在表 9 比较了各方法的所需时间，这也是选择合适方法的重要指标。是否能根据测验规模等给出更具体的所需时间或计算复杂度，帮助研究者了解方法耗时，以便选择合适方法。文章在讨论部分给出 CSRI 的耗时是 MHM 的 1/3，如针对文中的实证研究案例各方法计算复杂度和耗时具体是多少？

回应：感谢审稿专家的建议。我们以情境 1 中π为 40%，π_i^{non}高，d_{RT}大的条件为例，在不同测验长度条件下，重新模拟生成数据，并使用四种方法处理不努力作答，统计了各方法的耗时情况，结果如表 R4 所示，以帮助研究者了解方法耗时，便于选择合适方法。并且，统计了文中实证研究中各方法具体耗时（见表 R4 最后一行）。应用方法所使用的计算机处理器为 Intel(R)Core(TM)i7-9700，内存为 32GB。从表中可以看出，首先，随着测验长度增加，各方法耗时都增加，并且 MHM 方法耗时的增加最为明显。其次，整体来说，MHM 方法的耗时最长，为 CSRI 方法的 1.7 倍以上，其次是 CSRI 方法，OSR 和 CSR 方法耗时相对较少，但与 CSRI 方法耗时相差不大。最后，在实证研究中 MHM 是 CSRI 方法耗时的约 2 倍。已经在修改稿中的讨论部分加入了更详细的关于方法耗时的讨论。具体请参见修改稿第 34 页第 2 自然段中标红文字。

表 R4 不同测验长度条件下各方法耗时统计（单位：分钟）

测验长度	OSR	CSR	CSRI	MHM
10	92	63	78	240
20	181	105	168	546
30	261	195	245	622
40	362	396	492	831
50	526	510	630	1160
实证数据	226	256	359	604

注：以情境 1 中π为 40%，π_i^{non}高，d_{RT}大的条件为例。

意见 5：结合意见 3 和 4，文章通过随机模拟已经比较了各方法的优劣，在实证研究章节中应更加关注实际应用。如根据估计的不努力作答比例，测验规模等给出具体方法的选择，并基于该方法进行不努力作答识别、替换、模型参数估计等，解决实际问题。

回应：感谢审稿专家的建议。在修改稿中已经参考审稿专家的建议修改了实证研究章节。首先，根据测验属于低利害测验，测验长度较长的特征，预估数据中可能存在较为严重的不努力作答行为。然后，根据自陈量表估计的不努力作答严重性，以及根据所有题目上反应时呈现双峰分布认为测验中普遍存在不努力作答行为，因此建议选择 CSRI 和 MHM 方法。最后基于这两种方法以及本研究中涉及的另外两种方法进行了不努力作答的处理和模型参数估计，并对结果进行比较，以解决实际问题中的识别和参数估计问题。具体请参考对问题 3 的回复，以及实证研究中标红部分。

意见 6：其他意见：

（1）.建议文章题目包含“不努力作答”，明确本文研究的应用。

（2）.公式(2)及其上面一段，二元正态各参数所代表的含义未解释。

（3）.公式(3)下面一行，e_ij<-1.645 怎么来的？正态分布 5%分位数？

（4）.4.1.2 节情形 1，不努力作答是否也跟题目难度及被试能力相关，当然这两者也反应在速度上，但建议说明。

（5）4.1.3节，请给出先验分布的具体设置，请给出burn-in, thinning rate, Rhat(reduction factor?)的含义。

（6）.表 1 中，请给出 M 和 SD 的具体意义，若代表均值和标准差，是否为 L=30 次重复的均值和标准差？这里报告 SD 是否有必要？

（7）.表 2 中，请说明 pi=0, dRT 大时，没有均方误差值的原因。

（8）.在报告图 1-6 前一段，请更具体的说明图 1-6 是怎样生成的？如图 1 中大概有 30 个点，是否对应 L=30 次抽样？每次的横坐标 true 真值又是如何确定的？图中纵坐标代表什么？图 5-6 又是如何得到的？

（9）.表 4 题目不明确，由前文可知自陈量表包括 3 方面内容。这里题目建议具体为“…认真完成测验重要性评价以及完成测验努力程度评价…”

（10）.请检查参考文献中的大小写问题，部分文章题目每个单词首字母都大写，而部分没有，请按照 APA 格式修改。特殊单词如 Bayesian 首字母应大写。Ulitzsch et al.文献已

有具体刊号，请更新。

回应：

（1）感谢审稿专家指出这个问题，已经在题目中加入了“用于处理不努力作答的”，修改后的题目为“用于处理不努力作答的标准化残差系列方法和混合多层模型法的比较”。

（2）感谢审稿专家的意见，已经在正文中公式（2）上面一段文字中加入了对二元正态分布中各参数含义的解释，具体请参见公式（2）上面的标红文字。

（3）审稿专家的理解是正确的，已经在正文中加入了说明，“基于显著性水平为 0.05 的标准正态分布左侧检验”。具体请参见公式（3）下方标红文字。

（4）感谢审稿专家指出的重要推断。的确有一些研究者关注了不努力作答与题目难度或被试能力的关系。例如，Wise（2006）发现，题目难度与考生的努力没有显著关系。但一些学者认为，如果题目相对于被试能力太难，在题目上差的表现可能就是由努力程度不足导致的（Asseburg & Frey, 2014; Pastor, Ong, & Strickman, 2019）。又例如，一些学者的研究支持了动机（与考生是否努力作答有关）和考生的能力是有关的（Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005）。另外，有的学者基于 PISA 的实证数据，认为能力和努力程度之间存在正相关，而混合多层模型忽略了这种相关，因此提出了使用能力和速度来预测被试是否由于不努力而缺失作答的模型（Ulitzsch, von Davier, & Pohl, 2020）。**综上，关于不努力作答是否和题目难度有关，学术界尚未得出一致的结论。而很多学者普遍认为低能力被试倾向于不努力作答（Rios, Guo, Mao, & Liu, 2017; Wise, 2017）。**此外，在很多关于多层模型的研究中，都将速度和能力设为正相关(e.g., Fan, Wang, Chang, & Douglas, 2012; Wang & Xu, 2015; Wang, Xu, & Shang., 2018; Wang, Xu, & Shang et al., 2018)，也就是说速度较慢的被试倾向于低能力，也倾向于有更多不努力作答。基于此，并参照 Wang 等人研究中的模拟设置，在本研究的模拟研究中设置为速度较慢的被试倾向于猜测作答（不努力作答）。

参考文献：

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92.

Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189-212.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on

aggregated-scores: To filter unmotivated examinees or not?. *International Journal of Testing*, *17*(1), 74–104.

Sundre, D. L., & Wise, S. L. (2003). *"Motivation filtering": An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*, 83-112.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254.

Wang, C., Xu, G., Shang, Z., & 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.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114.

Wise, S. L. (2017). Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

（5）感谢审稿专家的意见。已经在 4.1.3 中给出先验分布的具体设置及参考，具体请参见文中标红文字。其中，被试参数的设置采用了与产生值相同的分布，这是因为经过前期与先验分布有关的敏感性分析，发现以不含有不努力作答的条件为例，被试参数先验分布设置为变量之间不相关的标准二元正态分布和与产生值相同的分布，得到的参数估计值的 Bias 和 RMSE 都非常接近（差异在小数点第三位），并且 Bias 普遍较小（在–0.002 到 0.018 之间）。可以推测，只要使用共轭的先验分布，能力和速度的先验分布对参数估计结果基本没有影响，不同先验分布几乎都能得到准确的参数估计结果。因此，在后面的研究中，被试参数先验分布采用与产生值相同的分布。并且，在 4.1.3 中还加入了 burn-in, thinning rate, Rhat(即 potential

scale reduction factor，潜在量尺缩减因子)的含义，具体请参见第 11 页标红文字。

（6）这里的 M 和 SD 是 30 次重复的均值和标准差。同意审稿专家的意见，这里重复 30 次报告的 SD 意义不大，故在修改稿中删除 SD，仅报告均值。

（7）在表 3 中，当$\pi = 0$时，表示数据中没有不努力作答，此时没有π_i^{non}和d_{RT}的其他水平。为避免误解，已将原表拆分为两个表格（见表 5 和表 6），分别表示没有不努力作答和含有不努力作答的情况。

（8）感谢审稿人的建议。图中横坐标代表每种条件下参数的真值，由于每种条件下参数真值不变，因此在每个条件下，所有点的横坐标是固定的。每个点代表每道题目（图 1——图 4）或每名被试（图 5——图 6），纵坐标代表每道题目或每名被试的参数偏差，即，在 30 次重复中，每次参数真值减去其估计值所得到的偏差的均值。已经在图 1 前面的一段加上了图的解释。具体请参见这一段标红的文字。

（9）感谢审稿人仔细的审阅。已经在修改稿中将表 8 的标题改为“实证研究不同方法 RTE 指标与认真完成测验重要性评价以及完成测验努力程度评价的相关”。

（10）感谢审稿人仔细的审阅。已经按照 APA 格式修改了参考文献格式。具体请参见修改稿中参考文献部分。

……

审稿人 2 意见：

如何识别被试不努力作答是心理测量的重要课题，也是一个难题。《标准化残差系列方法和混合多层模型法的比较》一文，将标准化残差系列方法和混合多层模型方法进行了模拟与实证比较，研究结果对实践应用提供了很好的参考。通读全文，有以下疑惑：

意见 1：CSR 和 CSRI 两种方法是文中要比较的重要方法，请再添笔墨做些介绍

回应：感谢审稿专家的建议。已经在修改稿中加入了一些对这两种方法的补充介绍，以及两种方法之间关系的说明。具体请参见 2.2 和 2.3 中标红的文字。另外，由于篇幅限制，关于这两种方法的原理和更详细的研究，感兴趣的读者可以参考 Liu 和 Liu（2021） 的文章，也已经在文中说明。

参考文献

Liu, Y., & Liu, H. (2021). Detecting non-effortful responses based on a residual method using an iterative purifying approach. *Journal of Educational and Behavioral Statistics, online*.

意见 2：混合多层模型方法的介绍，难以理解如何识别努力还是不努力作答，如其中的不努力作答实践 C 参数如何求取；其次，模拟研究的两个情景均是从混合多层模型假设来界定，请在模型介绍部分，详细、清晰并强调混合多层模型的基本假设是什么，又有哪些是不满足其基本假设的情况。

回应：感谢审稿专家的意见。首先，我们在修改稿中混合模型方法介绍部分加入了对不努力作答部分模型的限定，及其估计方法。例如，不努力作答的答对概率g_j会小于努力作答的答对概率，在使用贝叶斯估计时，可以对该参数设置一个均值较小的先验分布。通过这种方式可以实现不努力作答部分模型的参数估计。然后，在模型介绍的最后一段，加入了混合多层模型的假设以及可能违背假设的情况说明。具体请参见“3 混合多层模型法”中标红文字。

意见 3：作者在研究方法部分详细介绍了模拟数据的生成等，但缺少模拟作答过程的说明，请补充说明。其次，在模拟比较条件的说明中，作者多处采用现有研究，如被试量等，作者采用现有研究的条件的目的是什么，是便于比较还是其它。因为被试量或许也是影响模型参数估计的一个因素，项目量或许是影响不努力作答比率的一个因素等。另外，文中提到“数据不符合混合模型假设”，如何在实践中判断是否符合，请作者在讨论部分从实际情景及应用适当提及讨论该问题。

回应：感谢审稿专家的意见。在数据生成部分，我们是基于模型直接模拟作答反应数据和反应时数据，并未通过模拟作答过程来生成数据。这一生成数据的方式与前人研究类似（Lu et al., 2020; Wang & Xu, 2015; Wang, Xu, & Shang, 2018; Wang, Xu, & Shang et al., 2018）。具体步骤如下：（1）利用题目参数和被试参数的真值，基于 van der Linden（2007）的多层模型模拟生成努力作答的作答反应和反应时。具体来说，对于作答反应，先将题目参数和被试参数的真值代入公式（1）中的作答反应模型，计算得到每名被试在每道题目上答对的概率值$P\left(Y_{ij}=1\middle|\theta_i\right)$。然后在[0,1]的区间内产生一个随机数，比较这个随机数和计算的概率值的大小，如果概率值大于等于随机数，则作答反应为 1（答对），如果概率值小于随机数，则作答反应为 0（答错）。对于反应时，代入参数真值，对于每名被试在每道题目上的反应时，生成一个均值为$\beta_j-\tau_i$，方差为α_j^{-2}的正态分布，然后从这个正态分布中随机抽取一个点，作为取对数后的反应时（$\ln\left(t_{ij}\right)$）。（2）生成不努力作答的作答反应和反应时。这包括首先在每种条件下，生成标记了原始数据中哪些被试对哪些题目的作答是不努力作答的矩阵。然后基于不努力作答的模型，产生不努力作答的作答反应和反应时。产生方式与（1）中努力作答相同。（3）使用不努力作答的作答反应和反应时替换原有数据中相应位置的数据。结合

审稿专家的意见，我们对数据生成部分进行了补充，详见修改稿中“4.1.2 数据生成”中标红文字。

本研究在模拟研究中，被试量和题目数固定为 2000 和 30，有两方面原因。一是前人类似研究（Liu & Liu, 2021; Wang & Xu, 2015; Wang, Xu, & Shang, 2018; Wang, Xu, & Shang et al., 2018）都采用了固定题目数和样本量的方式，操纵与不努力作答相关的因素，以保证研究结果清晰、简练（keep the scope of the study manageable，Wang, Xu, & Shang et al., 2018）。二是我们也曾经尝试在不同样本量和题目数的情况下考察 CSRI、CSR 和 OSR 三种方法的表现，发现总的来说，样本量越大，使用各方法得到的题目参数估计结果准确性提高越大，题目数越多，使用各方法得到的被试参数估计结果准确性提高越大。而在每种条件下，不同方法之间的优劣关系基本没有变化。因此，我们认为可以将被试量和题目数固定为前人研究中设置的水平（Liu et al., 2020; Liu & Liu, 2021），该水平也代表了实际的教育与心理测验中被试量和题目数的普遍水平，重点考察与不努力作答相关的因素的影响。此外，本研究在模拟研究中对不努力作答规模、严重性和两种作答反应时差异这三个因素的水平设置参考了前人研究（Liu et al., 2020; Liu & Liu, 2021; Wang, Xu, & Shang, 2018; Wang, Xu, & Shang et al., 2018），是因为这些水平代表了这三个因素下的典型情况，感兴趣的读者也可以将本研究结果与前人研究结果进行对比。

最后，由于本研究中所讨论的混合多层模型假设是针对不努力作答的特征建立的，在实际数据中由于哪些作答是不努力作答是未知的，无法具体探讨这些作答是否符合模型所假设的特点。本研究在研究局限和未来研究展望部分提出，未来研究可以基于一些不含强假设方法的初步识别结果，尝试构建指标用于检验数据是否符合混合多层模型假设，从而指导实践研究者根据指标反映出的情况选择合适的方法。例如，可以基于标准化反应时残差系列方法的识别结果，计算所识别出的不努力作答正确率是否在不同被试间相等，从而初步判断数据是否违背了混合多层模型中关于不努力作答反应模型的假设。具体请参见第 36 页正文中标红文字。

意见 4：文中还有些表述需再清晰，如混合多层模型参考文献，简写等。尤其在结果呈现部分，各表格中的英文缩写分别表示什么，请作者在表格备注中加以说明。

回应：感谢审稿专家指出的问题。首先在引言部分，混合多层模型第一次出现的地方，加入了简写的说明以及参考文献。然后，在结果呈现部分的表格后以备注的形式加入了英文缩写的解释。同时，检查了文中有其他类似问题的地方进行修改。具体请参见引言和结果等部分

标红文字。

意见 5：结果呈现部分，模拟研究采用的是实验设计的方式，建议按照多因素实验设计的结果进行报告，如各主效应交互效应的情况，以便更可知各类方法的统计结果，也可适当增加可读性。另外，如 P19-20 页的结果说明部分，涉及多方法多情景多参数的说明，也建议用表格的形式进行说明以增强可读性。

回应：感谢审稿专家的建议。我们已经在修改稿中增加了表 3、表 4，分别表示情境 1、2 中以各参数 RMSE 为因变量，以各模拟因素为自变量的方差分析结果。并在表上方的文字中对结果进行了总结。具体请参见 4.2.3 中标红的第一段文字和表 3、表 4。但是，由于方差分析仅能说明各因素及其二次交互项的显著性，无法比较各方法在不同条件下的优劣变化情况，因此仍保留了表 5——表 7 的具体结果及其描述。关于图 1——图 6 的结果说明部分，是按照每个参数（6），每种情境（2），原始数据结果和处理不努力作答后的结果（2）来说明，一共形成了 6*2*2=24 种交叉。如果用表格形式说明，有两个问题。一是有的结果在不同情境下相同，有的不同，造成单元格合并还是分开的形式不够统一。二是这部分说明中还分析了基于原始数据得到不同偏差结果的原因，文字量较大，放到表格中显得冗长，不够美观。但是这一段说明的确存在可读性不强的问题。因此，我们对这部分内容进行了归类和凝练。每个参数部分的说明分为（a）（b）两点，（a）表示原始数据的情况，（b）表示使用 CSRI 和 MHM 方法得到的结果。如果在两种情境下结果不一致，就在（a）（b）下一级再分别说明。具体请参见图 6 后结果说明文字中标红部分。

……

审稿人 3 意见：

文章系统比较了标准化残差系列方法和混合多层模型法，对两类方法不努力作答识别和参数估计结果的准确性进行了比较，研究结论可以为实际应用者在选用这 2 类方法上提供一定的借鉴，具有一定的理论创新和实践参考价值。但文章仍有以下问题需要作者完善：

意见 1：Wang 等人（2018）曾对贝叶斯残差法和混合多层模型法进行过比较研究，作者提出标准化残差系列方法是更新的方法，在此基础上，作者对标准化残差系列方法和混合多层模型法进行了比较研究。为了让研究结论更具普适性，建议作者在实验中将贝叶斯残差法也加入比较，通过比较这 3 类方法在不同实验情境下的效果，可以为使用者在方法选用上提供更多的指导。

回应：感谢审稿专家的建议。本研究没有将贝叶斯残差法纳入比较，主要出于以下三个方面的考虑。**第一，贝叶斯残差法本身计算较为复杂且在实际中很少应用。**自 2008 年 van der Linden 和 Guo（2008）提出该方法以来，除了在 Wang 等人（2018）的研究中与混合多层模型法进行比较，据作者所知鲜有应用。贝叶斯残差法的原理是推导出反应时的后验预测密度函数，再根据参数分布，求出临界值作出判断。具体来看，用t_{ij}^{*}表示被试 i（i = 1,…, I）在题目 j（j = 1,…, J）上的实际反应时通过自然对数转换后的数值，$\boldsymbol{t}_{(i/j)}^{*}$和$\boldsymbol{y}_{i/j}$分别表示除去被试 i 在题目 j 上作答的反应时和作答反应后到的反应时矩阵和作答反应矩阵。使用 van der Linden(2007)的高阶模型对数据进行拟合，可以得到模型预测反应时$\tilde{t}_{ij}^{*}$的后验预测密度

$$\mathrm{f}\left(\tilde{t}_{ij}^{*}\middle|\boldsymbol{t}_{i/j}^{*},\boldsymbol{y}_{i/j}\right)=\int f\left(\tilde{t}_{ij}^{*}\middle|\tau_{\mathrm{i}}\right)f\left(\tau_{i}\middle|\boldsymbol{t}_{i/j}^{*},\boldsymbol{y}_{i/j}\right)d\tau_{\mathrm{i}}, \tag{1}$$

其中，τ_{i}表示被试 i 的速度，$f\left(\tilde{t}_{ij}^{*}\middle|\tau_{\mathrm{i}}\right)$表示基于高阶模型估计得到的$\tilde{t}_{ij}^{*}$的对数正态分布的密度。从以上公式可以看出，在计算反应时的后验预测密度时，需要依次除去每个数据点上的反应时和作答反应数据进行估计，计算量非常大。这一点也已经在正文中进行了说明，详见修改稿引言部分第 4 自然段标红文字。**第二，在 Wang 等人（2018）的研究中已经对贝叶斯残差法与混合多层模型法进行比较，发现贝叶斯残差法表现差。**即，在异常作答比例较低时，贝叶斯残差法表现就比混合多层模型差，随着异常作答比例增加，贝叶斯残差法表现显著变差，而混合多层模型法相对稳定。即使数据是基于残差模型生成，当个人异常作答的比例产生于 U(0.5, 0.75)的均匀分布时（相当于本研究中不努力作答严重性高的情况），贝叶斯残差法的正确识别率只有 0.301，而混合多层模型法高达 0.953。前面两点原因说明，贝叶斯残差法计算复杂且在各条件下表现均不如混合多层模型法，因此基于效率的考虑，在不努力作答严重性各水平下都应当选择混合多层模型法，故而本研究没有再加入贝叶斯残差法的比较。**第三，标准化残差法相对于贝叶斯残差法计算和原理都更为简单，并且，Liu 和 Liu（2021）基于该方法，新提出了固定参数标准化残差法和固定参数迭代标准化残差法，得到了较好效果，因此标准化残差系列方法是本研究重点比较的方法。**已有研究证明新方法在不努力作答严重性较高的条件下也有较好的表现，这与混合多层模型法已有的结论是一致的。并且，与混合多层模型法相比，标准化残差法及其发展出来的新方法不含强假设，目前，尚没有研究在混合多层模型法假设是否满足的条件下，比较该方法和标准化残差系列方法的表现。因此，本研究关注的重点是标准化残差系列方法与混合多层模型法的比较。**从方法改进的思路上看，贝叶斯残差法与本研究所指的标准化残差系列方法有明显不同，故出于研究重点和篇幅限制，在本研究中暂不考虑。**

参考文献

Liu, Y., & Liu, H. (2021). Detecting non-effortful responses based on a residual method using an iterative purifying approach. *Journal of Educational and Behavioral Statistics, online*.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365–384.

Wang, C., Xu, G., Shang, Z., & 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.

意见 2：文章提出“当不努力作答严重性低时和当不努力作答严重性高时。”这些方法的表现有一些差异，但并没有说明，在实际使用中该如何量化评价不努力作答的严重性，即根据不努力作答的程度来选择合理的方法。

回应：审稿专家提出了很重要的问题。在实际中确实需要一些指标来量化评价不努力作答严重性，以指导方法的选择。我们已经在研究局限和未来研究展望部分提出“从提高方法使用效率的角度考虑，未来研究可以基于一些不含强假设方法的初步识别结果，尝试构建一些指标，用于测量整个数据中不努力作答严重程度，或检验数据是否符合混合多层模型假设，从而指导实践研究者根据指标反映出的情况选择合适的方法。”实际上，我们也已经参考结构方程模型中残差相关拟合指标的思路，尝试使用 CSRI 方法中第一次迭代中识别的不努力作答，在第一、二次迭代中标准化反应时残差差异均方根构建评价指标。经过模拟研究发现，该指标与数据中不努力作答严重性程度有较强的相关，并且能够显著预测使用 CSRI 方法后，相对于使用原始数据估计，各参数 RMSE 减少的程度。我们将在未来深入该项研究，完善并发表与这个指标相关的研究结果。

意见 3：表 4 中不同方法 RTE 指标和认真完成测验重要性评价之间的相关系数都低于 0.09。尽管通过假设检验显示相关显著，但如此低的相关系数是否真能说明 2 种变量之间存在关联吗？

回应：感谢审稿专家指正。数据显示 RTE 指标与认真完成测验重要性评价之间的相关系数确实较低。我们查找了 Pastor 等人（2019）使用相同数据发表的论文，发现这个结果与他们关于这个变量的研究结果类似。他们基于被试在测验上是否努力作答的情况，使用潜在类别模型将被试分为两类，一类表示努力作答行为较多的被试，一类表示努力作答行为较少的被

试。最后发现这两类被试在认真完成测验重要性评价上的均值尽管差异显著，但是效应量并不大。原文如下“Although the mean difference on the importance variable was statistically significant, Cohen’s d was not as large for this variable, indicating only small to moderate differences between Classes 1 and 2 in the extent to which they felt it was important to do well on the assessments.”因此，为保证结果解释的准确性和客观性，我们在修改稿中加入了对相关系数较低的说明，具体请参见表 8 上方标红的文字。我们推测认真完成测验重要性评价这一指标可能并不能很好地反映被试是否努力作答的真实情况，即可能有被试认为认真完成测验很重要，但却仍不努力作答。今后的研究可以在自陈量表中加入其他更能真实反映被试努力作答水平的变量以进行效度验证。

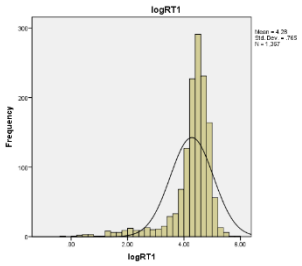
参考文献

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189–212.

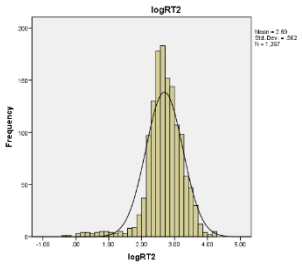
意见 4：混合多层模型法的效果依赖于考生作答时间分布。因此，在实证研究中，建议作者报告所有考生的作答时间分布图，方便读者更直观的了解时间分布是否符合混合多层模型法的假设。

回应：感谢审稿专家的建议。根据我们的研究结果，混合多层模型法由于包含强假设，因此它的表现依赖于其假设是否满足。具体来说，第一，该模型假设异常作答的正确率为g_j，即所有被试在同一道题上不努力作答的答对概率是相同的。第二，该模型假设不努力作答行为的反应时服从均值和标准差恒定的对数正态分布（见公式（9））。具体请参见“3 混合多层模型法”中最后一段标红文字。但是，要考察是否满足假设，必须知道哪些作答是不努力作答，并计算其作答准确性和反应时的分布。目前，尚没有研究提出如何检验其假设是否满足，我们已经在研究局限性和未来研究展望部分对这个问题进行了讨论，具体请参见该部分标红文字。如果仅对所有被试的作答时间分布进行观察，由于被试是否努力作答未知，并不能严格检验混合多层模型法的假设。因此我们仅能从整体分布是否符合双峰分布来大致判断数据中是否存在不努力作答（Meyer, 2010; Wang, Xu, & Shang et al., 2018; Wise & Kong, 2005）。在正文的图 7 中，我们以一道题目为例，展现了所有被试的反应时分布，发现呈现出双峰分布的特点，说明作答反应中可能同时混合了努力作答和不努力作答。此外，由于正文篇幅所限，下面的图 R2 呈现了在所有题目上所有考生的反应时分布。从图中可以看出，所有被试在所有题目上的反应时分布几乎都呈现出双峰分布的特点，说明在所有题目上都普遍存在不

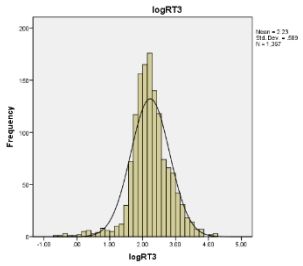
努力作答（Meyer, 2010; Wang, Xu, & Shang et al., 2018; Wise & Kong, 2005）。结合低利害测验的特点和学生在自我报告的随机猜测比例各选项上的情况，认为数据中应当存在略为严重的不努力作答现象，此时使用 CSRI 和 MHM 方法得到的结果会明显优于另外两种方法。这一点我们也在正文中加入了说明，具体请参见修改稿中“5.2 实证研究结果”中第一段标红文字。



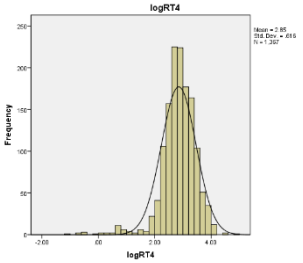
第1题



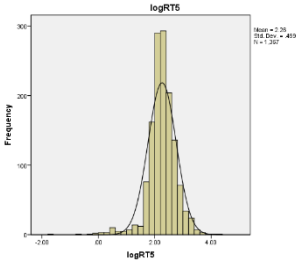
第2题



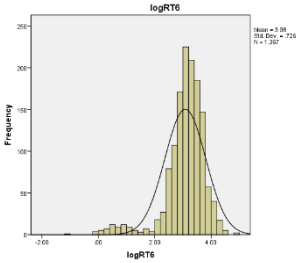
第3题



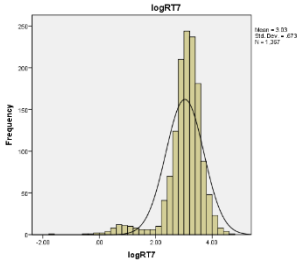
第4题



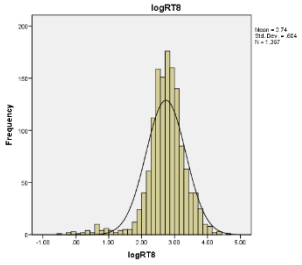
第5题



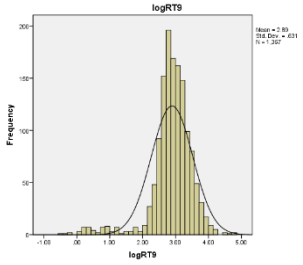
第6题



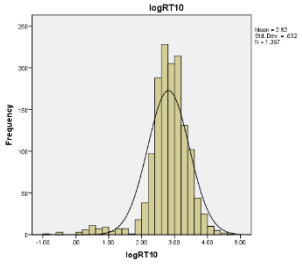
第7题



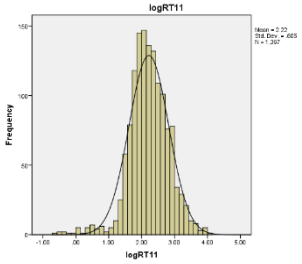
第8题



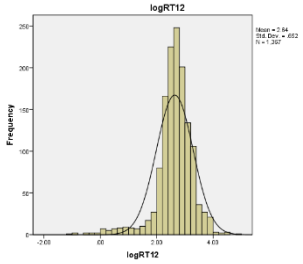
第9题



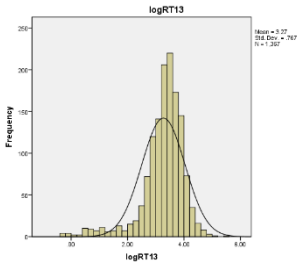
第10题



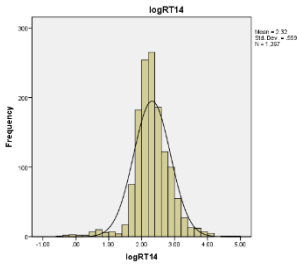
第11题



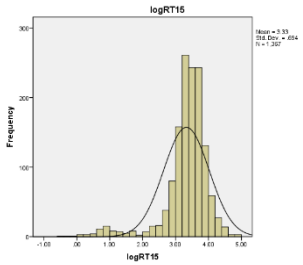
第12题



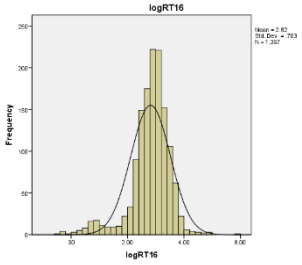
第13題



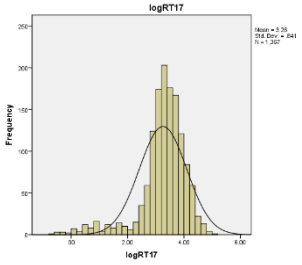
第14題



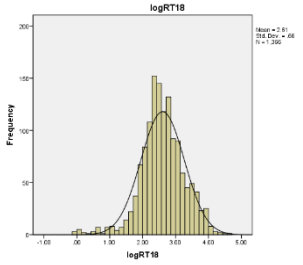
第15題



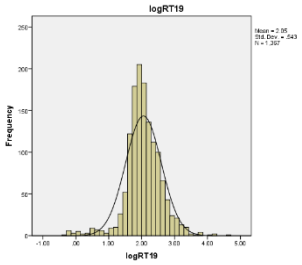
第16題



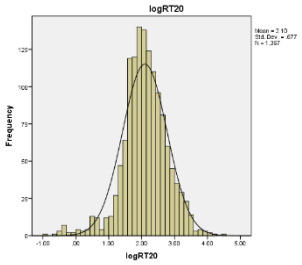
第17題



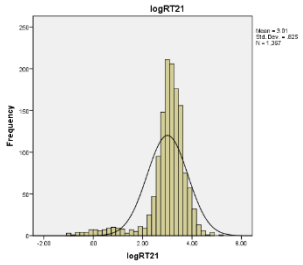
第18題



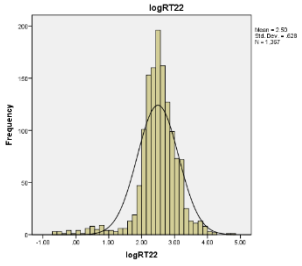
第19題



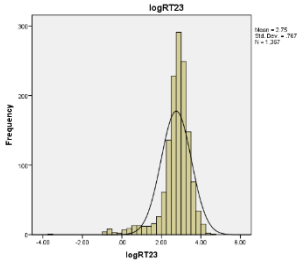
第20題



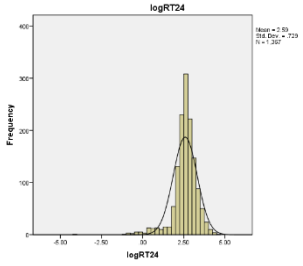
第21題



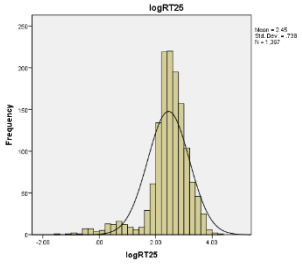
第22題



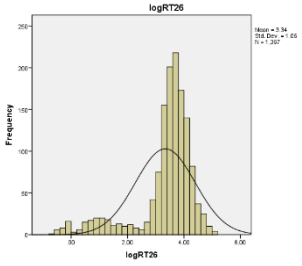
第23題



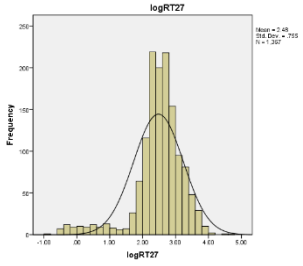
第24題



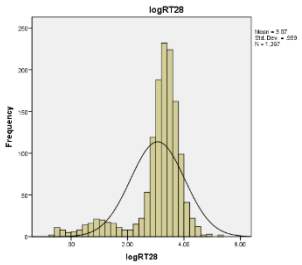
第25題



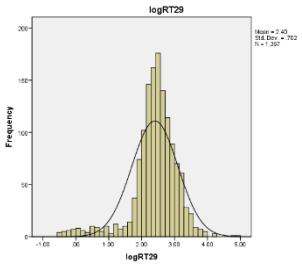
第26題



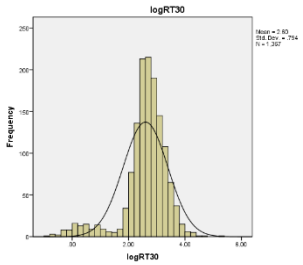
第27題



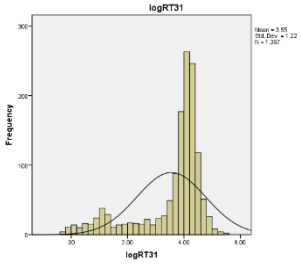
第28題



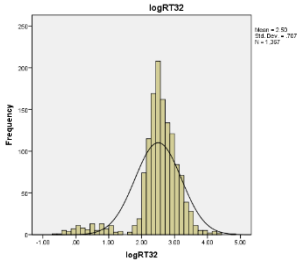
第29題



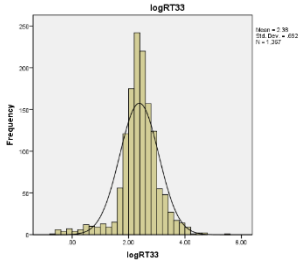
第30題



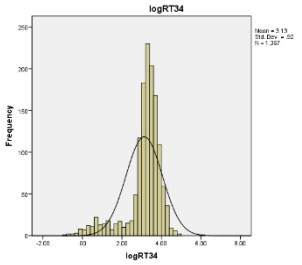
第31題



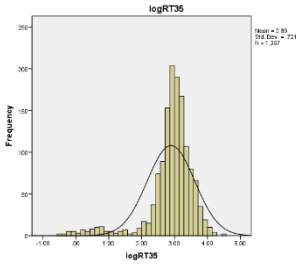
第32題



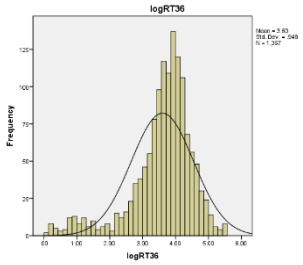
第33題



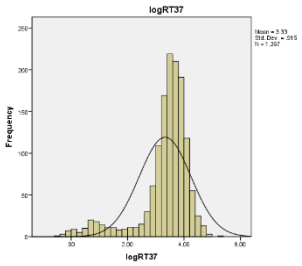
第34題



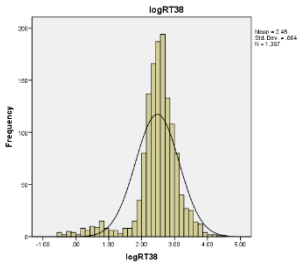
第35題



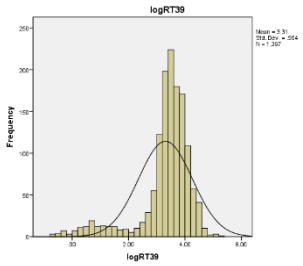
第36題



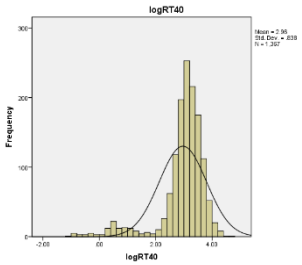
第37題



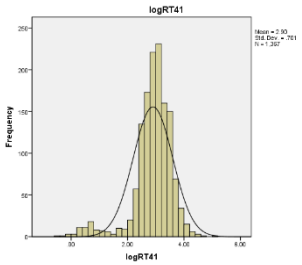
第38題



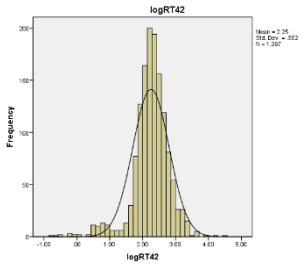
第39題



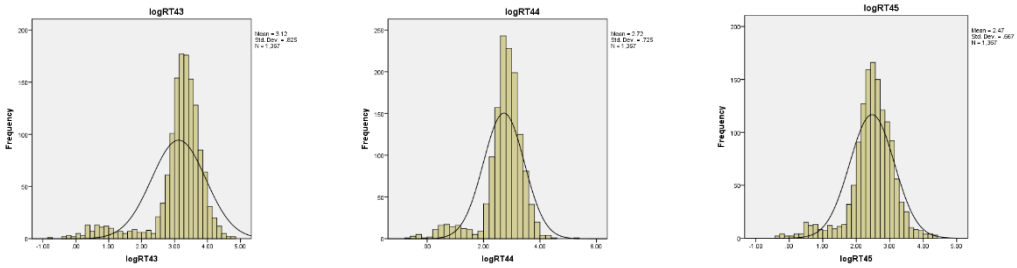
第40題



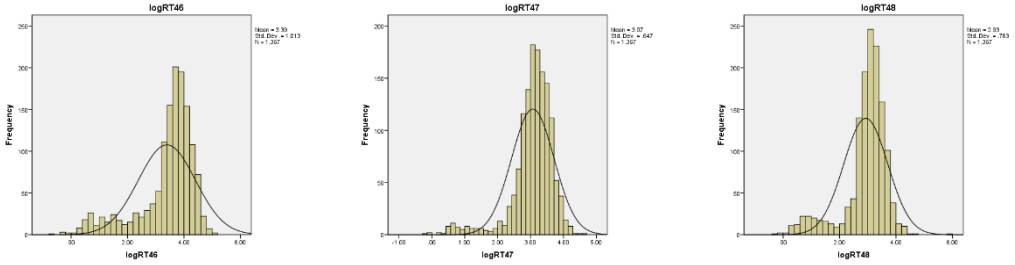
第41題



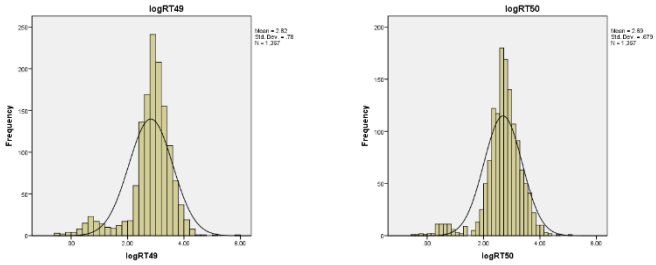
第42題



第43题　　第44题　　第45题



第46题　　第47题　　第48题



第49题　　第50题

图R2 实证研究所有被试在所有题目上反应时分布图

注：横坐标是反应时取自然对数，纵坐标是频率。

参考文献

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538.

Wang, C., Xu, G., Shang, Z., & 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.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

第二轮

审稿人 1 意见：

作者详尽的回答了我的主要问题，文章也有了较大的进步。针对作者的回应和对文章的修改，我还有以下意见和建议：

意见 1：在上一轮对意见 1 的回应中，作者提出即使错误删除了努力作答的数据（即 FDR 高），也应当不会带来较大的参数估计误差。同时作者也提到，标准化残差法将不努力作答替换为缺失，并且忽略的方式。那么 FDR 较高必然导致更多的作答被忽略，样本信息（样本量）会减少，参数估计的误差理应上升。甚至，错误删除的努力作答数据可能会产生估计偏差（比如错误删除正确率低的作答可能会低估难度参数？）。希望作者能够再明确的解释参数估计误差变化不大的逻辑。

回应：同意审稿专家的意见。FDR 高会导致错误忽略更多作答，增加参数估计的误差。在本研究中，FDR 未带来参数估计误差明显变化的原因，主要是由于识别的不努力作答是在作答层面，而非被试层面。因此，**只是将某些被试在某些题目上的作答当成缺失处理，并不会导致样本量明显减少。**另外，**在模拟研究中，识别出的不努力作答（即记为缺失的作答）的比例均不高。**从表 2 可以看出，各条件各方法识别出的不努力作答比例在 0.4%到 26.9%之间（Pr 结果）。根据 Rose（2010, 2013）的模拟研究结果，无论缺失机制如何（i.e., MCAR, MAR 或 MNAR），当整体数据中的缺失比例在 **30%以下**时，采用忽略的方式得到的参数估计结果是具有稳健性的。本研究中识别出的不努力作答比例均小于 30%，可以推测，在这些条件下即使错误忽略较多作答，也不会带来较大的参数估计误差。最后，**在实际测验中，不努力作答所占比例一般不高**（Hong, Steedle, & Cheng, 2020; Niessen, Meijer, & Tendeiro, 2016; Steedle, Hong, & Cheng, 2019）。例如，研究发现，实际测验中不努力作答的发生比例分布在 2%到 50%之间（Curran, Kotrba, & Denison, 2010; Johnson, 2005），其中大部分在 10%或以下（Meade & Craig, 2012; Maniaci & Rogge, 2014）。审稿专家提出的问题对我们很有启发，FDR 的结果应当结合不努力作答识别比例一起评价，如果不努力作答识别比例很高（i.e., >30%）且 FDR 很高，那么有理由相信采用忽略的方法会带来一定程度的参数估计误差，此时需要考虑基于模型的方法（e.g., 混合多层模型法）处理。我们将以上考虑加入到了讨论与结论部分，具体请参见该部分 35 页标红文字。

参考文献：

Curran, P. G., Kotrba, L., & Denison, D. (2010, April). *Careless responding in surveys: Applying traditional techniques to organizational settings*. Paper presented at the 25th annual conference of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and psychological measurement*, *80*(2), 312-345.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*(1), 61–83.

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

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. Journal of Research in Personality, 39, 103-129.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use?. *Journal of Research in Personality*, *63*, 1-11.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (PhD thesis). Friedrich-Schiller-University, Jena, Germany.

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), 1-53.

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*(2), 101-111.

意见 2：针对上一轮的意见 2，作者在文章第 35 页表 12 上面一段增加了讨论，其中，“在模拟研究的情境 1 中…，增加了不努力作答严重性的比例水平“，这里容易让人误解是文中模拟部分增加了更多比例水平，但这四种比例仅使用在讨论部分。请修改表述方式，明确只是在讨论中考虑了更多水平。

回应：感谢审稿专家细致的意见。我们已经修改表述方式，改为“为进一步探讨不努力作答严重性对各处理方法的影响，基于已有模拟研究，在模拟研究的情境 1 中，固定π = 40%，d_{RT}为大，增加了……”。具体请参见讨论与结论部分第 35 页标红文字。

意见 3：作者对我上一轮的意见 3、4、5 做了清晰的回应和修改。我赞同这些回应，作者也在表 12 中总结了四种方法的优劣。这里我的建议是，将对研究者如何使用方法，使用何种方法的建议单独放到一个小节（如 5.3 节），而不是放在最后讨论或穿插在别的章节。如之前意见 3、4、5 所提到的，根据研究者不同的实际问题，实际数据，实际需求，给他们明确的建议应该使用什么方法。这将提高文章的实用性。

回应：感谢审稿专家的建议。在修改稿中我们结合模拟研究和实证研究结果，提出了一些方法应用的建议，并独立放到一个小节。这些应用建议是基于模拟和实证研究得到的，因此放在了最后。具体请参见“7 实际应用建议”。

意见 4：作者应对一些语句表述再做检查和修改。1）第 2 页，最后一行。“针对标准化残差法这一缺陷…“，由于作者在这句话之前新加入了贝叶斯残差法的讨论，此句不连贯。2）第 9 页 4.1.2 节第一段，“协方差矩阵为 Sigma=…”，这里协方差矩阵 Sigma 的主对角线应都为 1。3）第 11 页 4.1.4 标题上面红字部分，“检验每个参数是否收敛”，应为“检验每个链条是否收敛”。4）第 29 页，红字修改部分，“计算了 1367 名学生在自我报告的随机猜测比例各选项上的情况”，建议改为“计算了 1367 名学生在自我报告各选项上随机猜测比例的情况”。

回应：感谢审稿专家细致的审阅。我们已经对上述语句表达进行了检查和修改。

（1）改为了“针对标准化残差法和贝叶斯残差法在不努力作答严重时表现差的缺陷”。

（2）这里协方差矩阵确实为为$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$，这是参考了前人研究（Wang & Xu, 2015; Wang, Xu, & Shang., 2018; Wang, Xu, & Shang et al., 2018），速度参数分布的方差为 0.25。

（3）已修改为“检验每个链条是否收敛”。

（4）这里不是指自我报告各选项上随机猜测的比例，而是对于整个测验题目的随机猜测比例各选项的情况。为避免歧义，改为了“计算了 1367 名学生自我报告的在所有题目上随机猜测比例的情况”。

最后，通读全文检查并修改了类似问题。具体请参见文中标红部分。

……

审稿人 2 意见：作者对审稿意见做了详细的回复，并对论文做了相应的修改。建议发表。

……

审稿人 3 意见：

作者在回复中非常详细的解释了贝叶斯残差法的缺点，建议作者在文中也解释为什么没

有对贝叶斯残差法进行比较，以便读者更好地了解这些方法的优缺点。

回应：感谢审稿专家的意见。已经在文中加入了贝叶斯残差法的缺陷并简述了没有比较的原因。具体请参见引言倒数第二段中标红文字。

第三轮

编委意见：该研究有理论意义和实际价值，但目前文稿冗长，用 5 号字体有 40 多页，不少内容没有必要，在考虑发表之前需要大修。大的修改意见如下： 1. 凡是可有可无的内容和句子都删除。 2. 请删除模拟研究中的方差分析及相关内容，这部分内容无关紧要，通常的模拟研究不做方差分析的。 3. 删除模拟研究部分的全部图以及说明。 具体评论和修改建议参考审改稿。

回复：感谢编委提出的修改意见。已经按照意见对全文进行了精简。具体请参见修改稿。

第四轮

编委意见：作者已经按上一次的修改意见做了大篇幅删减，建议发表。公式后面第一行是否空两格，要视乎是否另起一段而定。

主编意见：

经审阅，同意刊发。但需要修改后，我再看一下。

第一， 研究二使用 James Madison 大学的自然界管理测验测试数据。其中，“自然界管理测验” 不明白是什么意思，是不是一门课？建议再补充一下。

回复：感谢主编的建议。自然界管理测验是由 James Madison 大学教师开发的，基于网络施测的测验，旨在考察学生对与保护环境相关的管理原则、问题和实践应用的了解程度。已经在原文中简要加入解释，并且引用了使用这个测验数据的参考文献。具体请参见“5.1 数据和设计”中标红部分文字。

第二，6 讨论与结论。建议将此分开，如果在讨论部分加上小标题，更有利于读者理解。

回复：感谢主编的建议。已经按照主编要求，将讨论与结论部分分开，并在讨论部分加上小标题，具体请参见修改稿相应标红部分。

第三，7 实际应用建议。建议此部分内容合并到讨论之中。

回复：感谢主编的建议。已经按照主编要求，将这部分内容合并到讨论的“6.2 方法总结和建议”中，具体请参见该部分标红内容。

第五轮

主编意见：同意刊发