修改说明

第一次修改说明和对专家意见的逐条回复

尊敬的编辑：

非常感谢您专业的工作！很高兴我们有机会对本文进行修改，以提高它的质量。根据两位匿名审稿专家的意见和建议，我们对本文进行了仔细修改，把其中存在的问题和不足之处进行了一一修改和说明，并以高亮的方式显示，对专家的意见和建议做了逐条回复。同时也修改了一些书写方面和格式方面的问题，期待能够得到您们的回复。

再次感谢您和两位审稿专家的工作！

下面我们就专家们的意见和建议进行逐一回复。

审稿专家一的意见

作者提出了一种新的认知诊断中的被试拟合指标，并通过模拟数据和实证数据对新指标的有效性进行考察，具有一定的理论与应用价值。但存在的问题也较多，描述如下： 一、文章结构的问题 1. 文章第 2 与第 3 部分的篇幅似乎过短，4.1 中新指标的提出应当是本文的亮点，却淹没在模拟研究中。建议将现有的 2、3 以及 4.1 整合在一个部分。

回复：您的建议非常好，我们按照建议，对文章的结构进行了重组，并且对重要和关键的部分进行了更为详细的描述和介绍，希望能够做到详略得当。

2. 本文所提出的 R 指标似乎是适用于所有认知诊断模型的，但对它的介绍却被放置在 DINA 模型的介绍之后。或许 DINA 模型只应该是本文对其有效性探究过程中的一个条件而已，而不应该如现在这样。 二、引言部分存在的问题 1. 本研究基本是基于 DINA 模型的，但是全文，尤其是引言部分并未介绍为什么选择 DINA 模型。 三、"4.1 R 指标"部分存在的问题 1. 如前所述，本文所提出的 R 指标似乎是适用于所有认知诊断模型的，或许不应该像现在这样使得所有对新指标本身的讨论都被局限在了 DINA 模型上。

回复：的确，R指标并不基于具体的诊断模型，我们目前的安排有误导读者之嫌。我们在正文中补充了R指标的定义和理论意义的介绍，并且明确了本研究选择 DINA 模型的原因。

3. 作为本研究的主要贡献，建议增加理论层面上的对 R 指标性质的阐述。比如，分别讨论不同观测作答反应下，拟合和失拟如何影响 R 值的变化；还可以讨论失拟会导致不同参数如何变化，进而如何导致 R 值的变化。

回复：感谢您宝贵的建议，提出一个新的指标，明确其性质和理论价值非常重要，我们在正文中专门对R指标的性质和理论价值进行了补充介绍。

4. 对 R 指标的定义可以更清晰。 （1）E(X_ij | α_i)是条件期望，其本身是α_i 的函数，而与 X_ij 的取值无关。因此其实际含义应为属性掌握模式为α_i 的被试 i 在项目 j 上的期望得分，而不是所谓的"正确作答的期望得分"。 （2）对 P(X_ij |α_i)的定义或许可以更清楚，"属性掌握模式为α_i 的被试 i 在项目 j 上实际得分的概率"中实际得分"是什么值"的概率似乎没有说清楚？虽然可以理解作者的意思，但容易造成一定困扰。

回复：感谢您的意见！与上面一条建议类似，我们在介绍R指标时没有充分介绍R指标的性质和价值，对于有关符号的描述也不够准确，会导致读者对R指标的理解出现困扰。我们把这部分内容进行了重写。$E(X_{ij}|\alpha_i)$表示属性掌握模式为α_i的被试i在项目j上的期望得分，$P(x_{ij}|\alpha_i)$表示属性掌握模式为α_i的被试i在项目j上得x_{ij}分的概率。希望能够把困扰消除。

5.文章参考文献的格式也存在多处错误，比如：(1) 每篇文献后是否呈现期数，需要统一按照要求； (2) 参考文献中页码之间的连接符不统一； (3) 第八条参考文献中缺失一个空格； (4) Santos, de la Torre 和 von Davier(2019)在参考文献中的顺序不对； (5) 三名作者的文献在文中首次出现是全部列出还是使用 et al 不统一。 可能还有其他问题，希望作者再逐一核对参考文献的格式，保证准确。

回复：感谢您细致的审稿，对于出现这些格式上的问题我们非常抱歉。我们对全文进行

了细致的检查和反复的阅读，主要修改包括：对每篇文献后面补充了期数，页码之间的连接符也进行了统一，Santos, de la Torre 和 von Davier(2019)的顺序也进行了更正,对三名及以上作者在文中的引用在首次出现统一成‘et al’，中文文献用‘等’。对有关的书写和格式方面的问题进行了消除。

审稿专家二的意见

审稿人认为，本文写作逻辑和规范都没有问题，内容也比较充实，但审稿人对本研究中提出的 R 指标，其理论意义不强，实践意义和作用更是存疑。具体意见如下。

1.引言部分，只是叙述被试拟合研究的重要性，以及简单提及认知诊断中被试拟合研究的指标有哪些，并没有对这些指标进行述评，引出本研究的研究问题，对于为什么要构建 R 指标，R 指标的特点与作用有哪些更未涉及。

回复：非常感谢您的意见！我们补充了定义*R*指标的动机，并对性质和理论价值的进行了介绍，对其特点与作用进行了详细的描述，希望不再会给读者造成困扰。

2.第二部分的认知诊断模型，可以删除，不仅因为 DINA 简单，几乎人人皆知，而且这一部分放在文章，影响了文章的连续性和逻辑性。

回复：的确，DINA 模型放的位置可能会对读者造成困扰。因为*R*指标是不基于特定诊断模型的，我们增加了*R*指标性质和意义的介绍。并且把 DINA 模型的介绍这部分内容改用最简洁的文字描述并结合参考文献的方式， 希望能最大限度不影响文章的内在连续性和逻辑性。

3.模拟研究中，是先模拟“创造性作答、随机作答、疲劳、睡眠、作弊、随机作弊”这些类型的被试，然后用各种指标去检测，发现不同指标在不同条件下，表现不一。则 R 指标的突出作用和贡献是什么？

回复：感谢专家的意见。据我们了解，不同的异常行为有其独特的产生机制，很多已有的统计量在不同的异常行为检测上也是存在表现不一的情况，目前还没有通用的统计量，可以在不同的异常行为检测上表现都好，实际应用中我们需要结合检测的主要目的和各统计量的特点来选择和使用。对于 R 指标的突出作用和贡献，主要是体现在：定义的 R 指标可以较好地完成诊断测验中的被试拟合检验，在很多条件下都有较好的表现，更重要的是，该指标很具有推广价值，可以推广到项目拟合检验，模型拟合检验当中。我们补充了对*R*指标的定义和性质的介绍，并且在讨论部分增加了对 R 指标的作用和未来研究方向的介绍。

4.指标在实证研究中存在一定的局限性或逻辑问题：（1）实证中，只能检测出哪些被试存在异常反应模式，并不能给出是哪一类的异常（模拟研究中 6 类中的哪一类）。而实际上，除了作弊，是不允许的，存在不诚实或道德问题，对于自己考试中存在哪一种状况，考生自己最清楚。那么，这就引发一个问题，R 指标的研究，存在的意义或必要性在哪里？主要是作弊检测？（这是有意义的）还是其它？（2）被试的作答异常，在实践中，并不是一定不受欢迎的，比如，超出发挥就是很好的事情。对于考试状况，考生自己最清楚，则本研究对于教师、对于实践者，意义何在？（3）项目质量，或者说测验质量，本身也是影响考生作答的一个重要因素，认知诊断的目的是把考生的 ORP 与 IRP 比较，找到其 AMP，则判别方法或模型会影响其准确性。所以，出现了异常反应，即 ORP 和 IRP 不一致的情况，其原因很多，实践中无法排除。（4）实践中，检测异常反应，还得倒回去进行模拟研究，找到一个临界点，靠这个临界点来判别哪个存在异常，对于使用者（尤其是非测量学研究者）来说，不好理解，也不便于使用。（5）靠临界点划分出的异常反应，是不是真的是异常反应，其效果没有验证。

回复：感谢您宝贵的意见！（1）的确，因为本实证数据不是我们自己收集的数据，我们没有可能得到考生的详细信息，因此也无法对*R*指标检验的结果进行进一步的分析。

对于考生异常类型的判断，目前有一些研究试图在这方面进行探索，比如 Wang 等(2018)。基于我们所了解的信息，当前有关异常作答检验的模式通常是用多种方法结合，多阶段综合分析的方法去完成。因为基于任何单一的指标对被试做出异常的标识是很有风险的，尤其是在一些高风险测验中。这个也是我们希望未来能够进一步进行研究的方向。这个不足之处，我们在讨论中进行了更详细的分析和说明。

（2）这是个非常好的意见，我们在本研究中确实没有关注“超常发挥的考生”。超常发挥的考生在整个测验中应该都是高水平发挥的，因此，多数检测方法（包括*R*指标）会把这个考生“处理”成高水平的考生，从而不会检测出异常，对于这类考生，可能需要结合其平时的成绩，进行纵向数据分析才能够检测出来。我们在本文所考虑的情况主要是考生在部分题目上出现了异常的作答行为。对于“超常发挥的考生”，可能需要采用其它或一些新的方法才能够检测出来。

（3）这仍然是一个非常好的意见，非常感谢。项目质量会影响统计方法的表现，*R*指标也不例外，本研究为了聚焦我们的研究目的，没有把项目质量作为因素纳入考虑，但这不代表不需要重视。我们考虑在未来的进一步研究中考虑项目质量的影响，在讨论部分补充了这个说明。

（4）对于统计检验，我们目前通常有三种方法确定临界值，第一种是统计量有明确的理论零分布，这一种是最简单的，但是对于很多新构建的统计量不适用；第二种用近似临界值(Andrews, 1993)，这也只是适用一部分统计量； 第三种是用本研究中采用的蒙特卡罗模拟，具体有两种方式：①是采用类似我们的方法，即经验分布，②是采用置换分布（Permutation distribution; Shao, Li, & Cheng, 2016）的方式，置换分布方式的缺点是需要非常长的时间。因此，采用经验分布的方式在很多研究中被使用(比如 Sinharay, 2016, 2017; Shao et al., 2016; Shao, 2016; van der Linden & Xiong, 2013; van Krimpen-Stoop & Meijer, 2001)，更受欢迎。我们也会进一步考虑 R 指标的统计性质，去探索它的理论分布等特征，希望它能够在实际应用中发挥作用。

（5）基于临界值划分的异常反应，对于模拟实验中，我们可以通过一类错误率和统计检验力来评价其表现，对于实证数据，由于数据不是我们收集的，我们是通过比较和分析被划分的“异常被试”的观察作答数据和其能力是否相符的方式来评价，目前还做不到进一步的验证。这也是我们未来需要进一步改进和探索的地方。

5.讨论部分过于简单粗超，只有一篇参考文献。

回复：感谢审稿专家的建议，为了能给读者关于本研究的创新和不足之处有更清楚的认识，我们对讨论部分进行了重写。把我们对*R*指标的认识、理解和未来的研究方向进行了更细致的描述。

第二次修改说明和对专家意见的逐条回复

尊敬的编辑：

非常感谢您的工作！很高兴收到了专家对我们所做修改的意见和建议，也感谢专家对我们的鼓励，使我们有机会对本文进行进一步修改，以提高它的质量。

根据两位匿名审稿专家的意见和建议，我们对全文进行了阅读，把其中存在的问题和不足之处进行了一一修改和说明，并以高亮的方式（R1 的修改是黄色，R2 的修改是绿色）显示，对专家的意见和建议做了逐条回复。同时也修改了一些书写方面和格式方面的问题，期待能够得到您们的回复。

再次感谢您和两位审稿专家的工作！

下面我们就专家们的意见和建议进行逐一回复。

审稿专家一的意见

修改稿的质量有较大提升，但仍存在进一步改进的空间。主要的意见如下： 1. 作者专门增加标准化残差的部分用以解释 R 指标想法的来源，这很好。但是，就目前的表述来看，还可以进一步说明 R 指标与标准化残差的关系，尤其是应对分母部分的替换做出解释。

回复：非常好的建议，我们对于所构建的指标进行了更详细的解释和说明。对于对分母的替换部分，我们在设计的时候尝试了很多种方式，最后经过实验发现用目前的方式（观察得分的概率）会有更好的效果（检验力更高），主要原因是因为被试拟合研究关注的是考生的观察作答要与模型的预测作答的一致性， 当观察作答与模型的预测之间存在严重的不一致时（表现在观察得分出现的概率很小，并且由于它处于分母的位置，是一个逆向的权重），就会导致 R 指标异常的高。因此，观察作答概率的倒数可以作为被试拟合统计量的权重。

2. 在基于实证数据的研究部分，希望作者也可以给出其他指标的判断结果，对比分析所提 *R* 指标的优劣，并给出更多解释。

回复：这还是一条非常好的建议，我们同时对比了 *RCI*，*lz* 指标，将结果放在了附录部分，并就结果进行了更多的说明和解释，希望能够给读者提供更多使用上的参考。

其他的意见如下： 1. 第一部分最后一行的 l_z 没有斜体，以及英文摘要中的 R 指标和 RCI 是否需要斜体？

回复：感谢您细致的审稿，我们全文中对涉及到的统计量都进行了检查，对不规范的地方统一修改成斜体显示。

2. 参考文献第二条中的期数应为(1)。

回复：感谢您细致的审稿，我们进行了对应的修改。

审稿专家二的意见

建议作者文中表述，有些地方要注意，“我们”这种第一人称，在文中出现太多了。

回复：感谢您细致的审稿，我们对全文进行了梳理，将不是非常有必要使用“我们”的地方都进行了处理。将一些使用“我们”的地方换成了“本研究”，将一些“我们”进行了删除。