第二轮修改说明

首先非常感谢评审专家细致评审，根据专家们的意见，我们进行了认真修改并做了详细的修改说明，文章所有修改了的部分我们均用蓝色进行了标注。具体如下：

[审稿专家 1]：

谢谢作者对一审意见的详细的回应和修改。通读修改后的稿件，还有以下几点疑惑：

1.作者补充了项目质量高低两种实验条件，在结果呈现有个困惑如在报告一类错误率时，表 2 和表 3 具体的生成模型是什么，拟合模型是什么？如果是按照作者所说的，生成模型和拟合模型一致，那么为何不讨论生成模型和拟合模型不一致的情况？为何在讨论统计检验力时，讨论生成和拟合模型不一致的情况；请作者说明实验这么设计的缘由是什么。

答：感谢专家的问题，表 2 和 3 呈现的是题目拟合方法的一类错误率，一类错误率是指当生成模型和拟合模型一致的情况下，统计方法拒绝生成模型（拟合模型）的比例。例如，表 2 和 3 中的当用 DINA 模型生成数据，拟合模型也是 DINA，即生成模型和拟合模型一致；而统计检验力是指生成模型和拟合模型不一致时，统计方法拒绝拟合模型的比例，如当生成模型是 DINA 时，使用 DINO 和 ACDM 模型拟合数据，DINO 模型对应的统计检验力是指成功拒绝 DINO 的比例。

为了让读者更加清楚，在实验设计部分，我们详细解释了一类错误率和统计检验力的计算过程。

2.作者在小结和展望部分对各种题目拟合方法的应用做了补充。读后的困惑是题目拟合方法是为了探索题目模型是否拟合，为选择适当的模型做参考。从作者现有的表述来看，是先有了确定的模型再来考量题目拟合问题。这是否与实际应用相吻合，请作者解释。

答：感谢专家的建议。评估模型和数据的拟合通常从 2 个方面开展：（1）测验水平的拟合，测验拟合从总体水平上评估模型和数据的拟合：（2）题目拟合用于评估每个题目和模型的拟合度，题目拟合检验有助于识别异常题目，通过删除或修改异常题目将提高整个测验和模型的拟合水平。题目拟合检验是测验拟合的一个补充，因此，在实际的数据分析中，测验拟合和题目拟合检验在评估模型-数据

拟合中是必不可少的步骤，两者相互结合使用才能最大化地提高模型和数据的拟合，进而提高测验的诊断精度。

在“引言”的第二段，我们解释了测验拟合与题目拟合的关系，以及两者在实际应用的意义。

[审稿专家 2]：

感谢作者对于问题的仔细回答，通读全文后，但仍有一些疑惑：

1、对于 p(1)和 p(0)的解释并不能与所用的三个模型相契合。

答：感谢专家的问题。模拟 CDM 的参数通常有 2 种方式：一种是直接模拟题目的参数，例如，DINA 模型的 s 和 g 参数，然后将题目参数代入模型公式中计算答对概率；另一种是直接模拟答对概率： $P(\mathbf{1})$和 $P(\mathbf{0})$，这种模拟方法并不需要额外模拟每个CDM 的题目参数，已有许多研究采用了这种方法（例如，De la Torre & Lee, 2013; Ma et al., 2016）而不是模拟 CDM（如，DINA，DINO，ACDM）的原始参数，本研究正是采用了第 2 种模拟方法。为了让读者理解这种模拟方法，我们更详细地解释了该方法的具体过程。

De la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355-373.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200-217.

2、文中“采用一类错误和二类错误作为因变量”的说法是非常不严谨的措辞，应使用一类错误率（type-I error rate）和二类错误率 (type-II error rate)。并且对二类错误率的解释不清晰。建议作者反复通读全文，修改类似的表达上不清楚、啰嗦的地方。

答：感谢专家的建议，我们已修改原文“一类错误和二类错误”的表述，并认真通读全文，修改了表述不清和啰嗦的地方。

3、建议精简小结部分，用概括性的语言阐述不同自变量对不同拟合指标的影响。

答：根据专家的建议，我们已简化了小结部分对于实验结果的总结。

4、目前文中使用了 6 种拟合指标，但文中仍出现“5 种拟合指标”的说法。

答：感谢专家的细致审稿，已修改了相关表述。

第一轮修改说明

首先非常感谢评审专家细致评审，并为本文的进一步完善提出的宝贵意见及建议。根据专家们的意见，我们进行了认真修改并做了详细的修改说明，文章所有修改了的部分我们均用蓝色进行了标注。具体如下：

[审稿专家 1]：

《认知诊断模型中题目拟合评估的研究》一文通过模拟研究和实证数据分析，讨论了各种不同模拟条件下的五种拟合指标。通读全文，有以下疑惑：

1. 在模拟研究中，作者基于三种不同的认知诊断模型 DINA，DINO 和 ACDM 生成，但是在结果呈现部分，如表 2 是基于哪个模型生成的，又是用哪些模型进行拟合的，请作者做清晰表述。另外，在讨论检验力时，为何不将生成模型进行比较。如用 DINA 生成时，为何不将 DINA 做为拟合模型进行比较。

答：感谢专家的问题，专家的疑惑主要是由于我们未在原来的正文中详细解释因变量，即一类错误率和统计检验力的计算过程。

实验采用了 DINA，DINO 和 ACDM 作为生成模型，拟合模型也是 DINA，DINO 和 ACDM。例如，采用 DINA 模型来生成数据，然后，分别用 DINA，DINO 和 ACDM 来拟合数据，当用 DINA 模型拟合数据，即用真模型拟合数据时，如果统计方法拒绝 DINA 模型，此时，就属于犯了一类错误，在 100 次重复实验中，统计题目拟合方法拒绝零假设（真模型）的比例，即为该题目拟合方法所犯的一类错误率；而当用 DINO 或 ACDM 拟合数据，即用一个不正确（假）的模型拟合数据时，统计检验力为题目拟合方法拒绝零假设，即假模型（DINO 和 ACDM）的比例。因此，当生成模型是 DINA 模型时，用 DINA 模型来拟合数据可以计算统计方法的一类错误率，当用 DINO 或 ACDM 拟合数据时，此时，DINO 和 ACDM 在每个题目拟合指标下都可以计算一个统计检验力，即正确拒绝假模型的概率。

为了便于读者理解，我们在文中补充了一类错误率和统计检验能力的详细计算过程。详见正文第 6 页

2. 模拟结果的比较指标，作者采用了一类错误率和统计检验力，请补充并说明为何采用这两类指标，是否有其它比较指标。

答：感谢专家的建议。查阅已有的相关研究发现，在评价题目拟合统计方法的效果时，几乎所有研究都会采用一类错误率和统计检验力，或是这两个指标的等价变量作为评价指标。例如，已有研究（Köhler, Robitzsch, & Hartig, 2020; Su, Wang, & Weiss, 2021）采用一类错误和统计检验力来评价题目拟合方法的效果；涂冬波等人（2014）采用一类错误和二类错误作为因变量，统计检验力和第二类错误是等价的，即统计检验力等于 1 减去二类错误率。Zhang 等人（2018）采用假阳性率（False Positive Rates, FPRs）和真阳性率（True Positive Rates, TPRs）作为评价指标，事实上，FPRs 和 TPRs 等价于统计中的一类错误率和统计检验力。因此，借鉴已有的研究，本实验因变量采用了一类错误率和统计检验力。

我们已按照专家的建议，在实验设计部分，补充说明了为何采用一类错误率和统计检验力作为评价指标。详见正文第 6 页

涂冬波，张心，蔡艳. (2014). 认知诊断模型一资料拟合检验统计量及其性能. *心理科学*, *37*(1), 205–211.

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*(3), 251–273.

Su, S., Wang, C., & Weiss, D. J. (2021). Performance of the S− χ^2 Statistic for the Multidimensional Graded Response Model. *Educational and Psychological Measurement*, *81*(3), 491-522.

Zhang, X., Wang, C., & Tao, J. (2018). Assessing item-level fit for higher order item response theory models. *Applied psychological measurement*, *42*(8), 644-659.

3. 作者针对模拟研究结果，概括性的认为 Z(l)是这五种方法中最好的指标。但是从模拟研究结果来看，Z(l)的一类错误率表现的并不好，如在样本量 n=1000 时，表 2 中其一类错误率均值近乎最大(0.078、0.463)。为何作者还要认为它是最好的。该结论也影响到文中实证研究部分直接采用了 Z(l)，请作者做详细解释。

答：感谢专家提出的问题。$z(r)$和 $z(l)$是基于题目对（Item pairs）的统计量，在计算过程中要涉及到多次比较，因此，为了降低多次比较引起的一类错误膨胀，Chen 等人（2013）建议对 p 值进行校正。之所以会出现专家提到的，在某些实验条件下，$z(l)$的一类错误率会偏高的情况，这是由于我们在原来的实验中，并

未对原始 p 值进行校正。同时，基于另一位审稿专家的建议：校正 p 值和完善实验设计。我们改进了实验设计并重新开展了实验，在新的实验设计中，我们使用 Holm 调整法（Holm's adjustment）对 $z(r)$和 $z(l)$的 p 值加以校正。新的实验结果发现，在所有实验条件下，$z(r)$和 $z(l)$的一类错误率是最低的，最大一类错误率不超过 0.05。

另外，在实证研究部分，通过整体拟合检验发现 ACDM 是拟合更好的模型。模拟实验的结果显示当生成模型是 ACDM 时，在中等样本容量下（N=1000），综合考虑一类错误率和统计检验力，$z(r)$和 $z(l)$的效果相似，并且是所有方法中表现最好的。通过比较 2 个指标的实际分析结果也发现，它们筛选出拟合不佳的题目是一致的。因此，在实证数据的题目拟合检验中，我们使用了调整 p 值的 $z(l)$ 统计量。详见正文第 18 页

4. 对结果的讨论，建议作者从列举的一类错误率和统计检验力两个指标分别进行讨论。并在第 5 部分，结论和讨论部分，针对权衡实际需要的利弊对各拟合指标的选用，做详细的讨论和可行性建议。

答：感谢专家的建议。在结果呈现部分，我们已将一类错误率和统计检验力的结果，使用单独的小标题加以呈现并讨论。详见正文第 7 和 10 页

另外，在“结论与讨论”部分，重新为使用者提供了更详细的方法选择建议。详见正文第 19 和 20 页

5. 文中还有表述不清，参考文献有误或不存在等(如 Sorrel et al., 2016)，具体详见附件。

修后再审。

答：确实我们遗漏了这篇文献，现已在正文后增补了完整的文献。另外，专家在正文中批注的意见和建议，我们也一一进行了回答和认真的修改。感谢专家的细致审稿。

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506-532.

[审稿专家 2]：

研究方面：

1、既然研究聚焦不同题目拟合检验方法的系统比较研究，但在自变量的选择中并未考虑与题目有关的自变量，例如属性个数和题目长度、题目质量等。这样选择的理由是什么？此外，关于题目拟合检验方法的选择上，Wang 等人（2015）的研究结果显示，结合 Stone's 方法的拟合检验方法效果较好，但本研究未考虑 Wang 等人（2015）中所提到题目拟合检验方法的理由是什么？请进一步丰富实验设计。

答：感谢专家提出的建议。根据专家的建议，我们重新完善了实验设计，并将题目长度和测验质量作为 2 个新的自变量，题目长度设为 2 个水平：30 题和 60 题，其中 60 题通过重复 30 题 2 次而生成；题目质量分为：高质量和低质量。在新的实验设计中，我们并未将属性个数作为自变量，这是因为涂冬波等人（2014）的研究发现，在不同属性个数的条件中，传统题目拟合方法（如，χ^2 和 G^2）的表现是一致的，即方法之间的优劣不会随着属性个数的变化而变化。因此，为了节约篇幅，我们并未将属性个数作为一个新的自变量，并在“展望”部分提出未来的研究可以考虑属性个数对于题目拟合方法的影响，这会是我们后续研究需要继续探讨的内容。详见正文第 5 页

另外，原先的实验并未将 Wang 等人（2015）结合 Stone’s 方法的拟合指标纳入比较，是出于以下考虑：结合 Stone’s 拟合检验方法是借助蒙特卡洛重复抽样技术获得统计量的抽样分布（Bartholomew & Tzamourani,1999; Tollenaar & Mooijaart, 2003），从而完成拟合检验。但由于该方法需要模拟多个数据集，且需对每个数据集，重新进行参数估计和拟合统计量估计，例如，Wang 等人（2015）Stone’s 拟合方法设置为重复抽样 500 次，即首先，基于已估计的参数和后验概率模拟一批作答数据，然后重新进行估计参数和计算题目拟合统计量，这个过程重复 500 次，从而产生一个经验的抽样分布。因此，结合 Stone’s 拟合检验方法在实际使用中需要耗费大量的时间，该方法在实际使用中并不被研究者推荐，而是作为其它方法的基础（Rupp et al., 2010）。

根据专家的建议，我们也将 Wang 等人（2015）中表现较好的 Stone-Q_1 作为

一个新的题目拟合方法，并通过重新模拟实验和其他方法进行了比较，从而为使用者在题目拟合方法选用上提供更多的参考。详见正文第 7-16 页

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research, 27*, 525–546.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology, 56*, 271–288.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.

2、请结合以往研究，阐述本研究的创新之处和意义。

答：感谢专家的建议。当前，仅有少量研究试着将 IRT 中的题目拟合检验指标拓展到 CDM 中，例如，涂冬波、张心和蔡艳（2014）比较了 χ^2 和 G^2 统计量在 DINA 模型的效果；Wang 等人（2015）将 IRT 中的题目拟合指标： $Q1$ 和 PD（power-divergence）等应用于 DINA 模型中；Sorrel 等人（2017）将 S-χ^2 应用于 CDM 中，并通过模拟实验比较了其效果；Chen，de la Torre 和 Zhang（2013）将基于题目对的统计量应用于 CDM 中。然而，一方面，已有的研究主要集中在 DINA 模型下比较传统题目拟合方法的效果，而这些题目拟合方法在其他模型下的效果如何，仍值得探讨；另一方面，上述这些题目拟合方法都属于绝对题目拟合（Absolute item fit）指标，绝对题目拟合的评价是通过比较每个组的题目表现与拟合模型预测的表现来进行的，并且绝对题目拟合指标在实际应用中也是最常用的一类模型拟合评价方法，如在 IRT 的应用中有大量的研究使用 S-χ^2 指标来评估题目拟合（例如，Acevedo-mesa et al., 2020; Flens et al., 2020; Sunderland et al., 2020）；尽管这些绝对题目拟合方法已被初步应用于 CDM 中，但是，这些方法在 CDM 的效果仍缺乏系统的比较，在 CDM 的题目拟合检验中，这些指标的效果如何？面对不同的测验情境，该如何选择最佳的题目拟合检验指标？查阅相关文献发现，这些问题仍没有得到有效的解决。因此，本研究旨在不同的实验条件

下，系统比较这些绝对题目拟合方法在 CDM 的表现，从而为实际使用者在题目拟合方法的选用上提供有价值的参考。

根据专家的建议，我们在正文中重新解释了本研究的创新之处和意义。详见正文第 2-3 页

Acevedo-Mesa, A., Tendeiro, J. N., Roest, A., Rosmalen, J. G., & Monden, R. (2020). Improving the measurement of functional somatic symptoms with item response theory. *Assessment*, 1073191120947153.

Sunderland, M., Afzali, M. H., Batterham, P. J., Calear, A. L., Carragher, N., Hobbs, M., ... & Slade, T. (2020). Comparing Scores From Full Length, Short Form, and Adaptive Tests of the Social Interaction Anxiety and Social Phobia Scales. *Assessment*, *27*(3), 518-532.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a computerized adaptive test for anxiety based on the Dutch–Flemish version of the PROMIS item bank. *Assessment*, *26*(7), 1362-1374.

3、请解释模拟研究中的一类错误率和统计检验力的计算方法

答：感谢专家建议，的确原来对于实验因变量的解释过于简单。因此，为了让读者更加了解因变量的计算，在实验设计部分，我们增加了一段话，更加详细地解释了一类错误率和统计检验力的计算步骤。新增加的内容如下：

“为了便于读者理解，现以一个生成模型为例，说明题目拟合指标的一类错误率和统计检验能力的计算步骤。例如，当生成模型是 DINA 时，分别用 DINA、DINO 和 ACDM 分析数据，当用 DINA 模型拟合数据，即用真模型拟合数据，一类错误率是指在 100 次重复实验中，每个题目拟合指标拒绝原假设（真模型）的比例；而当用 DINO 或 ACDM 分析数据，即用一个不正确（假）的模型拟合数据时，统计检验力为题目拟合方法拒绝零假设，即假模型（DINO 和 ACDM）的比例。” 详见正文第 6 页

4、在使用基于题目对的拟合检验方法时，是否对显著性水平进行了校正？

答：非常感谢专家的提示。在原来的实验中，我们并未对 z(r)和 z(l)的 p 值进行校正，所以，在某些实验条件下会出现，z(r)和 z(l)的一类错误率偏大的情况。为了降低多次比较引起的一类错误膨胀，Chen 等人（2013）建议对 p 值进行校正。因此，在重新开展的实验中，我们采用了较常用的 Holm 调整法（Holm’s adjustment）对 z(r)和 z(l)的 p 值加以校正。在正文中，我们也对此进行了解释。详见正文第 5 页

5、对 P(1)和 P(0)的解释表述混乱

答：谢谢专家的建议。我们重新调整了这部分的描述，调整后的内容如下：

“（6）测验质量，包括高质量和低质量 2 种水平。题目参数的模拟参考 Gao 等人（2020）的方法，即对于高质量的题目，1-*P*(**1**)和 *P*(**0**)从均匀分布 *U*（0.05, 0.15）中随机生成。对于低质量的题目，1-*P*(**1**)和 *P*(**0**)从均匀分布 *U*（0.15, 0.25）中随机抽取。其中，*P*(**1**)表示被试已掌握题目所有属性的正确答对概率，*P*(**0**)表示被试未掌握题目任何属性的猜对概率。”详见正文第 6 页

6、表 2 下面的一段对于一类错误率的结果解释中，在不同显著性水平下的结果应该与各自的显著性水平进行对比。

答：感谢专家的建议，我们已按照两个显著性水平分别对表 2 的结果进行了重新解释。详见正文第 8-9 页

7、z(l)是基于题目对的拟合检验方法，那么表 8 中呈现的每一个题目的 p 值是如何计算得到的？

答：感谢专家的问题。基于题目对的拟合检验方法是通过度量观测数据和期望数据中题目对相关的差异来进行拟合检验，并统计每一题对应的最大题目对相关差异值，根据最大差异值对应的 *p* 值来检验是否拟合。因此，尽管是基于题目对的拟合检验方法，但每个题目只有一个最大题目对相关差异值对应的 *p* 值。

同时，为了让读者更清楚该方法的原理，我们在方法介绍部分，更加详细地解释了基于题目对的拟合检验的具体步骤。新增加的内容如下：

“由于每个题目会和测验其余题目组成一组题目对，因而，每个题目会对应有 *J*-1（*J* 表示题目总数）个 *z* 分数，对于有 *J* 个题目的测验，则会有 *J*(*J* - 1)/2（对称矩阵对角线以上或以下的元素）个 $r_{jj'}$ 和 $l_{jj'}$ 统计量需要评估。为了减少计算量，Chen 等人（2013）提出选取每个题目中相应统计量的最大 *z* 分数，通过最大 *z* 分数对应的 *p* 值来评价题目是否拟合。同时，为了降低多次比较引起的一类错误膨胀，Chen 等人（2013）建议对 *p* 值进行校正，本研究使用 Holm 调整法（Holm’s adjustment）对 *p* 值加以校正。” 详见正文第 5 页

8、本文几乎都是聚焦在不同显著性水平上不同拟合检验方法的结果，并未对样本量如何影响这五种拟合检验方法进行描述。考虑到样本量是本研究的一个自变量，请补充相关结果。

答：确实如专家所言，样本量是一个影响题目拟合效果的重要因素。我们在结果解释部分，着重对不同样本量下的结果进行了阐述。详见正文第 8-16 页

写作规范：

1、 更正全文中卡方的书写方式，不能用 X 替代 chisquare 符号

答：感谢专家的建议，我们已用公式编辑器改写了卡方符号。

2、 文中出现第一人称“我们”

答：已按专家建议进行了调整，删掉了第一人称“我们”。

3、 第一次出现 CDM 应给出全称

答：谢谢专家的细致审稿，已给出了 CDM 的全称（Cognitive Diagnosis Model）。详见正文第 1 页

4、摘要中“本研究通过模拟实验比较了 X2， G2， S-X2， z(r)和 z(l)等几种题目拟合统计量”的描述，已将所使用的题目拟合检验方法尽数写出，不需要在使用：“等几种”的说法。此外，“模拟实验结果显示， 综合 alpha=0.01 和 alpha=0.05 两种显著性水平下的表现， G2 在一类错误率和统计检验力方面的效果最优，而在 alpha=0.01 的条件下，z(l)的一类错误率和统计检验力的综合表现最好”的说法不明确，不能清楚的让读者了解究竟哪种方法更好。

答：感谢专家的建议，已在摘要中重写了实验结果的介绍。详见正文第 1 页