

## 修改说明

### 编辑部决定意见

题目功能差异（DIF）检测是非常具有实践应用价值的方法学议题，是保证测评公平性的基本条件之一。本研究所提出的方法本身具有一定创新性，对已有方法做了进一步推进。研究结果合理可靠。作为测量方法类研究，具有较高的发表可能性。但为了进一步增加研究的实践应用价值，并突出实践应用与仿真研究之间的相互支撑性（实证研究为仿真研究提供模拟基础，仿真研究补充实证研究中测验条件不足的局限性），请作者重新调整文章撰写逻辑，将两个实证研究放在模拟研究之前（即强调实证发现和方法的实践可应用性），然后再根据实证研究中的测验条件开展模拟研究（比如，DIF 题目占比；如果当前模拟研究条件与实证研究不匹配，还需要修改模拟研究条件）；限于篇幅原因，可将模拟研究全部或模拟研究结果放在附录部分。另外，请作者核对全文图表中的小错误，比如附录表 3 中 RCD 方法的前两行。

#### 回复：

尊敬的编辑：

衷心感谢您抽出宝贵时间审阅我们的稿件，并对我们的研究提出宝贵的意见和建议。我们对于编辑部提出的肯定和认可感到非常荣幸。

特别感谢您对于研究逻辑结构的建议。根据审稿意见，我们将两个实证研究调整到模拟研究之前，以更好地突出实证研究的实践价值和方法的应用性；并将模拟研究全部移至网络附录 C，以适应篇幅的限制，同时确保研究内容的完整性不受影响。我们认为这种结构上的调整使得研究的流程更加合理。

此外，非常感谢您的细心审稿。关于您指出的图表中的小错误，我们已经进行了仔细的核对和修正，特别是您提到的附录表 3（在新版本中为表 2）中 RCD 方法的前两行的错误，我们已经更正。

为了更清晰地呈现我们所做的修改，我们已在文中用蓝色字体标出了所有修改的部分。再次感谢您的专业指导和细致反馈，我们真诚地期望这些修改能够满足您的期望。

# 无需先验信息的两步项目功能差异检验方法

## 第二轮审稿意见

尊敬的主编、责编、审稿专家：

我们衷心感谢您对我们的文章“无需先验信息的两步项目功能差异检验方法”（稿件编号：psysci22-916）给予的专业与细致的评审与建议。我们也非常感激您给予我们重新修改与完善稿件的机会。审稿专家提出的所有意见和建议都是极其宝贵的，这对我们来说无疑将极大地推动和完善我们的研究论文。

我们已经仔细研究了每一条审稿意见，并且尽我们所能地按照这些建议修改了稿件。为了更清晰地呈现我们所做的修改，我们已在文中用蓝色字体标出了所有修改的部分。以下，我们会一一解释我们是如何针对每一条意见和建议进行修改的。

再次表达我们对主编、责编和所有审稿专家投入时间与努力的深深感激，您的专业建议无疑极大地加强了我们的研究质量。我们真诚地期望这些修改能够满足您的期望，并使得本研究能够达到《心理科学》的发表标准。

### 审稿专家 1

对于项目功能差异的检验非常有必要，我们国家地域辽阔，发展不平衡，而又经常出现大规模测验，很有必要特别重视 DIF 的研究和应用。文章研究 DIF 检验方法，提出两步项目功能差异检验方法，试图解决人为设定锚题，而锚题一旦设定错误，将产生误导的难题，这有一定意义。

但是文章还存在一些问题，主要是实证研究，不知道 PISA 数据是不是拟合 Rasch 模型，因为审稿专家没有看到模型资料拟合检验，而文章的研究，包括模拟研究都是建立在数据满足 Rasch 模型基础之上。如文章中所说的“对于真实数据，DIF 和非 DIF 题的真实参数是未知的，因此不能断言哪种方法表现最佳”，实证研究没有办法去评估获得结果的准确性，这也是很遗憾的事情。如果作者能够获得国内某一个大规模测验的数据进行分析，可能作者对具有 DIF 的项目的侦查以及请相关专家对分析结果的认定都会比较容易做，得到的结果也比较容易得到读者的认可。当然测验的得分数据集的获得是相当困难的事情，但是值得努力去搜集。

**回复：**首先感谢审稿专家对研究议题的肯定。在上次的修改稿中，我们对使用 Rasch 模型分析 PISA 数据的前提假设（单维性和试题之间局部独立性）以及模型-资料拟合都进行了检验。结果发现，数据满足使用 Rasch 模型进行分析的前提，并且在测验水平上数据和 Rasch 模型之间的不拟合程度是可以忽略的，在试题水平上，每道题目都能较好的拟合 Rasch 模型。由于篇幅的限制，我们把这部分内容呈现在了附录 B。

在这次的修改中，根据审稿专家的宝贵意见，我们增加了一个对国内大型测验数据集的 DIF 分析实例。在此例中，我们对我国某地区初中一年级学生入学语文能力测验的实测数据进行了 DIF 分析，并邀请资深中学语文教师、语言学专家对检验结果进行了分析。结果发现，在大规模数据分析中，几种方法的检验结果比较一致，相对来说，传统的 Wald 方法可能存在漏判风险，而 RCD 方法则可能检验出相对较多的存在 DIF 的试题，可能导致分析成本的增加。由于篇幅限制，我们把这部分内容呈现在了附录 C。

另外，书写方面存在一些小问题，审稿专家发现输入有重复之处，比如：

- 该方法首先通过难度差异的 QQ 图（以下简称 D-QQ 图）选择选择参照点（选择输入了两次），

- 分别估计两组数据对应模型的参数，得到两组难度参数差异值。这句话输入了两次。
- 另外，3.1 节第一段最后一句说“当测验存在 DIF 试题 0 时”，请问这是什么意思？  
建议作者仔细审读文章，改正输入错误，特别回应审稿专家提出的实证研究问题。  
修改以后再审。

**回复：**我们非常感谢审稿专家的细心审稿，并对文章中的错误表示歉意。我们首先修改了审稿专家指出的问题，并且对全文进行了仔细检查，以确保文章表述的正确。

## 审稿专家 2

作者详细地回复了上次意见，并对文章内容进行了相应修改。仅有两条小建议：

(1) 值得注意的是，从模拟结果来看，MH 和 Wald 仍在一定条件下（如 DIF 题目较少时）表现优于两步法。

**回复：**感谢审稿专家的认真审稿。我们在第 11 页，表 2 上方补充说明了这一结果：“最后，当样本量不足 2000 时，在平衡条件下，可能出现传统的 MH 和 Wald 方法的统计检验力略高于相应的两步法的情况。”

(2) 虽然结果中描述了样本量对各方法表现的影响，但在摘要和结论中似乎并未提及。

**回复：**谢谢审稿专家的宝贵意见。由于本刊对中文摘要字数的严格控制，我们在英文摘要与结论中都对此内容进行了补充，如第 27 页，结论与讨论部分第一段：“

针对测验公平性分析中选择不含 DIF 的锚题的挑战，本研究提出了先采用 D-QQ 图选择锚题再使用传统 DIF 检验方法进行分析的两步 DIF 检验法，并在不同样本量、测验长度、DIF 模式和 DIF 值条件下对原始的 MH 方法、Wald 方法以及使用 D-QQ 图选择锚题的两步 MH 方法、两步 Wald 方法和 RCD 方法进行了综合比较。首先，借助 D-QQ 图不但可以辅助选择锚题，而且可直观判断测验是否包含 DIF 试题以及 DIF 的模式。其次，样本量和实际 DIF 水平对于各种 DIF 检验方法的平均经验 I 类错误率的影响并不明显，但对统计检验力有较大影响，即样本量越大，DIF 值越高，DIF 检验方法的平均统计检验力也越高。最后，基于 D-QQ 图的两步 MH 和 Wald 方法在各种条件下表现优异，特别适用于测验中有一半题目包含 DIF 的情况，在平衡 DIF 条件下其对经验 I 类错误的控制优于 RCD 方法，在非平衡 DIF 条件下其对经验 I 类错误的控制明显优于原 MH 和 Wald 方法，且统计检验力也高于原 MH 和 Wald 方法。不过，当各组样本量不足 2000 人时，在平衡条件下两步法的统计检验力也可能略低于原 MH 和 Wald 方法。”

以及文末英文摘要第四段：“

The results of the simulation study indicated that the impact of sample size and actual levels of DIF on the average empirical Type I error rate is not substantial across various DIF detection methods, but it does have a significant impact on statistical power. That is, The larger the sample size and the greater the DIF values, the higher the average statistical power of the DIF detection methods. The study also proofed that when the test length was between 20 and 40, selecting the middle four items on the  $x = y$  line of the D-QQ plot as anchor items resulted in desirable outcomes. And the two-step DIF procedures performed optimally in regards to empirical type I error rate

and statistical power, even when half of the items exhibited DIF. The RCD method performed well under most conditions, though its type I error rates were slightly inflated when the DIF items favored both reference and target groups (the balanced condition). Meanwhile, the MH and Wald methods with purification were ineffective in detecting DIF items when 10 out of the 20 total items favoring one group. However, when the sample size for each group is less than 2000, under balanced conditions, the statistical power of the two-step methods might be slightly lower than that of the original MH and Wald methods.

# 无需先验信息的两步项目功能差异检验方法

## 第一轮审稿意见

尊敬的责编、审稿专家：

非常感谢您在百忙之中对本文进行审阅并提出宝贵建议。我们尽可能地基于各位的意见对文章进行了修改（已使用蓝色底纹标出），希望本次修改能够得到您们的认可，也希望修改稿能达到《心理科学》的发表水平。当然，如果我们对您的审稿意见有不准确的理解或还有进一步完善的地方，也烦请您再次指正，我们愿意做进一步的修改与完善。下面是修改列表及我们对审稿意见的逐一回复。

### 审稿专家 1

实际测验（特别是高厉害大规模测验）必须探查项目功能差异（DIF）以增强考试公平性。因此对 DIF 的检验是一个很有必要研究的课题。文章将探查 DIF 的 D-QQ 图与传统 DIF 检验方法相结合，提出两步 DIF 检验方法，有一点新意。

1. 但是关于实证研究，却显得比较粗糙，比如参加 PISA2012 的数学领域测验的这两国学生的得分数据是否遵从 Rasch 模型？

**回复：**首先感谢您的审阅和肯定。我们在修改稿中更换了实证数据，并根据您的建议对两组数据使用 Rasch 模型进行分析的前提进行了检验，包括测验单维性检验、试题局部独立性检验、测验水平拟合检验以及试题水平拟合检验。结果显示测验具有单维性、试题之间满足局部独立性假设、测验数据与模型的不拟合程度是可忽略的，所有试题均与模型拟合良好。因此，我们认为更新的实证数据可以采用 Rasch 模型进行分析。具体的前提检验方法与结果请见修改稿文末的附录 B。

2. 得分数据导出的被试能力分布是否等方差？因为文章前面的研究（包括模拟数据）结果是在一定条件下获得的，实际的得分数据符合这些条件吗？如果不符合，为什么能够套用前面的结果？建议作者认真修改以后再重审。

**回复：**非常感谢您指出的问题。为了另实证数据与模拟研究条件相匹配，我们在修改稿中更换了实际数据。更新的数据仍来自 PISA 2012 的数学领域测试，数据替换为加拿大和西班牙学生在第 4 个题册中 0-1 计分试题上的作答数据。我们分别对两组考生数据进行了能力估计，发现其能力分布方差分别为 1.484 和 1.517，可以近似为等方差。此外，考生人数与测验长度均在模拟研究条件范围内。从实证数据的 D-QQ 图初步判断，存在试题散点落在  $x = y$  的辅助线上，可被选做锚题，且测验 DIF 模式为平衡型（即分别有 DIF 试题偏向两组学生），这些情况均在模拟研究考虑条件范围内，因此可对照模拟研究结论对实证数据结果进行分析。详见修改稿第 12 至 14 页的实证研究部分。

## 审稿专家 2

本研究 (psysci22-916) 应用 Yuan 等人(2021)的 D-QQ 图选择锚题再结合传统 DIF 检验方法而提出了两步 DIF 检验法, 新方法具有一定新颖性和应用价值, 但仍有存在一些小问题:

1、应结合实验结果, 着重分析传统方法和新方法的适用条件, 并在摘要和结论中进行叙述, 以供读者参考;

回复: 非常感谢您的建设性建议。首先, 我们在修改稿的模拟研究的结果部分, 对不同条件下各种 DIF 检验方法的优劣进行了更加详细的比较, 详见第 9 页的“4.2 经验 I 类错误率”部分和第 10 页的“4.3 统计检验力”部分。其次, 我们在修改稿的摘要部分更具体地描述了模拟研究结果: “模拟研究表明当测验中有一半试题存在 DIF 时, 若 DIF 试题仅偏向一组, 则两步法有更高的统计检验力, 且比原方法能更好地控制 I 类错误; 若 DIF 试题分别有利于两组, 则其在 I 类错误控制上优于 RCD 方法。”最后, 在修改稿的结论部分对方法间的比较进行了总结: “结果显示, 基于 D-QQ 图的两步 MH 和 Wald 方法在各种条件下表现优异, 特别适用于测验中有一半题目包含 DIF 的情况, 在平衡 DIF 条件下其对经验 I 类错误的控制优于 RCD 方法, 在非平衡 DIF 条件下其对经验 I 类错误的控制明显优于原 MH 和 Wald 方法, 且统计检验力也高于原 MH 和 Wald 方法。并且, 借助 D-QQ 图可直观判断测验是否包含 DIF 试题以及 DIF 的模式。”

2、实验中也对比了最新的 RCD 方法, 在第 2 节中最好简要介绍其过程;

回复: 根据您的建议, 我们在修改稿第 3 页最后一段增加了 RCD 方法的介绍:

“Yuan 等 (2021) 提出的 RCD 方法也可以被视为一种两步 DIF 检验法。其基本思路如下: 在采用 D-QQ 图选定参照点 (锚题) 后, 对于实际数据, 计算相对难度差异  $\hat{\delta}_{(j)} = \hat{a}_{(j)} - \bar{a}_{(ref)}$ , 其中  $\bar{a}_{(ref)}$  是所选参照点在两组实际数据上难度差异的均值; 对于模拟数据, 同样计算相对难度差异  $\hat{\delta}_{(j)}^{(k)} = \hat{a}_{(j)}^{(k)} - \bar{a}_{(ref)}^{(H_0)}$ , 其中  $\bar{a}_{(ref)}^{(H_0)}$  是所选参照点在 K 次重复中的两组平均难度差异的均值。则对于每个试题有 K 个  $\hat{\delta}_{(j)}^{(k)}$ , 分别统计其均值  $\bar{\delta}_{(j)}^{(H_0)} = \sum_{k=1}^K \hat{\delta}_{(j)}^{(k)} / K$ , 2.5%分位点  $L_{(j)}^{(H_0)}$  和 97.5%分位点  $U_{(j)}^{(H_0)}$ 。最后, 比较由实际数据获得的  $\hat{\delta}_{(j)}$  和由模拟数据获得的 95%置信区间  $(L_{(j)}^{(H_0)}, U_{(j)}^{(H_0)})$ , 若  $\hat{\delta}_{(j)}$  落在区间外, 则判断第 (j) 道试题存在 DIF, 反之, 无充分理由说明第 (j) 道试题存在 DIF。”

3、MH 方法, Wald 方法, RCD 方法, 两步 MH 方法, 两步 Wald 方法所采用的错题 (及题数) 有何差异, 模拟研究中各方法所提取的错题 (及题数) 并未列出;

回复: 感谢您的细心审阅。需要指出的是, 在模拟研究中, 为了保证结果的稳定性与可靠性, 我们在每种组合条件下都重复了 100 次, 因此无法将各种 DIF 检验方法错选的锚题一一列出。不过, 根据修改稿中的



实证研究部分，我们仍可深入了解各种方法对锚题的选择。其中，两步 MH、两步 Wald 和 RCD 方法根据 D-QQ 图选择锚题 t3、t14、t18 和 t26。而 Wald 方法首先由 Wald2 算法寻找锚题，然后用 Wald1 算法逐题检验 DIF (Cao et al., 2017)，在修改稿的实证研究中 Wald2 算法选择的锚题为 t25、t8、t34、t2、t15、t20、t12、t32、t3、t14、t18、t26、t28、t4、t10、t1、t31、t33、t13，这些题目在第 14 页表 3 的 DIF 检测结果中用“...”表示。而 MH 方法采用了试题提纯程序，我们仅能得到被判断为包含 DIF 的试题，即 t19、t17、t21、t5、t7、t16、t30、t23。对于实证研究中各种 DIF 检验方法结果的全面比较详见修改稿第 12 页至 14 页的“5.2 研究结果”部分。

4、如何根据图 1 和图 2 确定 4 个锚题，叙述过于简洁，是否需要借助实证研究中类似的红色直线作为辅助观察线：

回复：感谢您指出这一问题，根据您的建议，我们在图 1 和图 2 中都增加了红色辅助线。但由于图 1 和 2 的每张子图中都同时展示了 4 种不同水平的 DIF 条件，因此在添加辅助线时综合考虑了这 4 种情况。此外，我们还对如何利用 D-QQ 图选择锚题，以及样本量和试题数量等条件的影响进行了更加详细的阐述，详见第 5 页中第 3 和第 4 段：

“我们还在每张图中增加了一条浅红色的  $x=y$  的辅助直线以方便观察，落在这条辅助线上的试题可被视为可能不存在 DIF 的锚题。当样本量增加时，DIF 试题与非 DIF 试题的区别变得更加明显，特别是在样本量为 2000 时，即使 DIF 值较小，D-QQ 图也可以明显分组，即非 DIF 的试题落在辅助线周围，而存在 DIF 的试题明显偏离辅助线。并且，随着 DIF 值的增加，分组也更加明显，DIF 值更大的试题偏离辅助线更远。当测验中的试题仅偏向一组被试时，非 DIF 试题分布在 D-QQ 图的一侧（如图 1 和图 2 中的前两列），而当测验中的试题同时偏向不同组别的被试时，非 DIF 试题集中在 D-QQ 图的中间（如图 1 和图 2 中的第三列）。相比于测验包含 20 题的情况（图 1），当测验包含 40 题时（图 2），可以更好的对试题分组，这是因为存在 DIF 的试题的比例变低了。

在采用 D-QQ 图选择锚题时，也可能出现试题分组不明显的情况，特别是在样本量较小（如 500 人），且 DIF 值也较小（DIF=0.4）的情况下。尽管如此，D-QQ 图仍可用于选择锚题。可以选择近似落在  $x=y$  直线上的点作为锚题；或者，当确认测验中有一半以下的试题存在 DIF 时，选择落在 D-QQ 图中间的点作为锚题。”

5、实验结果分析偏简洁，应该全面分析传统方法和新方法的优劣。例如，新提出的两步 MH 方法和两步 Wald 方法，在测验长度 20 且含有 10 个 DIF 题时，在一类错误和检验力方法均优于 MH 方法和 Wald 方

法;

回复: 感谢您的建议。根据您的建议, 我们对模拟研究的结果进行了更加详细的描述, 以在不同条件下比较各种 DIF 检验方法的优劣。在第 9 页的“4.2 经验 I 类错误率”部分, 我们将结果描述修改如下:

“各种条件下五种 DIF 检验方法的平均经验 I 类错误率如表 1 所示。首先, 本研究所提出的两步 MH 检验法和两步 Wald 检验法的平均 I 类错误率在所有条件下均未超过 5%。其次, RCD 方法在多数情况下 I 类错误率控制较好, 但在测验长度为 20 题且 DIF 模式为平衡条件下, 会出现 I 类错误率膨胀, 即高于 7.5%的情况, 这与 Yuan 等 (2021) 研究中, RCD 方法在平衡条件下对经验 I 类错误率的控制不理想相一致。最后, MH 和 Wald 检验法在测验长度为 20 题, DIF 模式为非平衡 10 题时, 其经验 I 类错误率远高于 7.5%, 无法满足 DIF 检验的实际需求。总之, 样本量和实际 DIF 水平对于各种 DIF 检验方法的平均经验 I 类错误率的影响并不明显, 主要影响来自于测验长度和 DIF 试题的数量与模式, 特别的, 在测样长度为 20 题且有一半试题存在 DIF 时, 若包含 DIF 的试题仅偏向其中一组被试, 则 MH 方法和 Wald 方法的平均经验 I 类错误率过高; 若包含 DIF 的试题同时偏向两组被试, RCD 方法对于平均经验 I 类错误率的控制可能不理想。而在这些条件下, 两步 MH 检验法和两步 Wald 检验法均能较好的控制经验 I 类错误率。”

在第 10 页的“4.3 统计检验力”部分, 我们将结果描述修改如下:

“表 2 展示了 5 种 DIF 检验方法在不同条件下的平均统计检验力。首先, DIF 检验方法的平均统计检验力主要受到样本量和 DIF 大小的影响, 样本量越小, DIF 值越低, DIF 检验方法的平均统计检验力也越低。例如, 无论测验长度与 DIF 试题数量, 当样本量为 500, DIF 值为 0.4 时, 所有方法的统计检验力都未达到 80%; 而当样本量达到 2000 时, 除个别情况, 大部分 DIF 检验方法的平均统计检验力都大于 90%。其次, 测验长度与 DIF 试题数量和模式会对 DIF 检验方法的统计检验力产生不同的影响, 特别是在测验长度较短 (20 题), 并且有一半试题偏向一组被试时, MH 和 Wald 方法在大部分情况下的统计检验力都未达到 80% (除了 MH 方法在样本量为 2000, 且 DIF 值为 0.6 和 0.8 的条件下)。最后, 就各方法平均统计检验力的整体结果来看, Wald 方法的平均统计检验力在更多的条件下低于 80%, 而两步 MH、两步 Wald 和 RCD 方法除了在样本量为 500, DIF 值为 0.4 的条件下, 它们的平均统计检验力在多数情况下大于 80%, 并且两步 MH 和两步 Wald 方法在各条件下的平均统计检验力最小值接近 50%, 而其余三种方法的平均统计检验力均出现过低于 30%的极端情况。”

6、在实证研究中, 表 5 中 RCD 方法为何将 t3 和 t15 分为类 N?

回复: 感谢您的提问。为了另实证数据与模拟研究条件相匹配, 我们在修改稿中更换了实际数据。更新的



数据仍来自 PISA 2012 的数学领域测试，数据更换为加拿大和西班牙学生第 4 个题册中 0-1 计分试题上的作答数据。并且，我们结合模拟研究结论对实证研究结果进行了更加详细的阐述，详见第 13 页的实证研究结果描述部分：

“对于真实数据，DIF 和非 DIF 题的真实参数是未知的，因此不能断言哪种方法表现最佳。所有方法对于 DIF 试题的检验结果基本一致，即存在 DIF 的试题大多分布在两组试题难度差异值的最小和最大值两端，其中试题 19, 17, 21 被所有方法判断为有利于西班牙组（这些题目对于加拿大组而言难度值更高），且具有中等以上的效应；而试题 30 和 23 被所有方法判断为有利于加拿大组，且具有中等以上的效应。在所比较的方法中，RCD 方法检测出最多的存在 DIF 的试题，这与模拟研究中平衡 DIF 模式下，该方法可能存在的 I 类错误膨胀相符。不过，尽管 RCD 方法在 RCD 值较小那一端判断试题 22、5、27、6、25、8、34 可能存在 DIF，但从效应量来看，其 DIF 大小是可以忽略不计的。相对而言，两步 MH 方法比 MH 方法略为保守（未判断出试题 5 存在 DIF），从模拟研究结果来看，样本量不足 2000 人时，平衡 DIF 模式下两步 MH 方法的统计检验力略低于 MH 方法。

D-QQ 图可用来辅助判断 DIF 检验结果的合理性。在 D-QQ 图中，试题离参考线越远，就越有可能存在 DIF，则若第 21 题被判断存在 DIF，那么比它距离参考线更远的第 9 题也是可疑的，在实际的 DIF 分析工作中，为保证测验公平性，也应重新审视该题。”

## 7、应适当增列些中文参考文献。

答复：感谢您的建议。国内学者对于 DIF 议题进行了深入而广泛的研究，不仅涉及 DIF 检验方法在心理与教育测验中的应用研究（曹亦薇, 2003; 关丹丹等, 2019; 刘文等, 2010; 郑蝉金等, 2011），还提出了新的 DIF 检验方法（余跃等, 2016），并且对各种复杂情景下的 DIF 检验问题进行了探讨（郭聪颖 & 边玉芳, 2013; 骆方 & 张厚粲, 2006; 魏丹等, 2020; 张龙 & 涂冬波, 2015）。我们在修改稿中增加引用了这些中文文献：

“

曹亦薇. (2003). 项目功能差异在跨文化人格问卷分析中的应用. *心理学报*, 35(1), 120–126.

关丹丹, 乔辉, 陈康, & 韩奕帆. (2019). 全国高考英语试题的城乡项目功能差异分析. *心理学探新*, 39(01), 64–69.

郭聪颖, & 边玉芳. (2013). 题组项目功能差异(dif)检验方法的应用探索. *心理学探新*, 33(5), 423–429.

刘文, 边玉芳, 陈玲丽, & 马文超. (2010). 马洛-克罗恩社会赞许性量表在跨文化研究中的项目功能差异检验. *心理科学*, 33(6), 1473–1476.

骆方, & 张厚粲. (2006). 检验项目功能差异的两类方法——cfa 和 irt 的比较. *心理学探新*, 26(1), 74–78.

- 魏丹, 张丹慧, & 刘红云 (2020). 基于多维题组反应模型的项目功能差异检验探究. *心理科学*, 43(1), 206–214.
- 余跃, 杜文久, 周娟, & 秦菊香. (2016). Lp 方法及其与三种常用 dif 检测方法的比较. *心理科学*, 39(3), 720–726.
- 张龙, & 涂冬波. (2015). 多级计分题项目功能差异常用检测方法比较. *江西师范大学学报: 自然科学版*, 39(5), 441–448.
- 郑蝉金, 郭聪颖, & 边玉芳. (2011). 变通的题组项目功能差异检验方法在篇章阅读测验中的应用. *心理学报*, 43(7), 830–835.
- ”

