《心理科学进展》审稿意见与作者回应

题目：测验模式效应：来源、检测和应用
作者：陈平，代艺，黄颖诗

第一轮

审稿人 1 意见：

意见 1：文章介绍了 TME 的来源，然后探讨了 TME 的检测方法，可以为测量研究者了解 TME 的来源、检测方法和研究思路提供参考。文章层次分明，行文流畅，文献综述完整。
回应：谢谢您对本文的肯定和认可。

意见 2：通读全文，审稿人认为以下几个方面有待完善：在题目类型对于 TME 的影响中，作者只提到了多选题，审稿人认为题目类型会是造成 TME 的一个重要因素，因此，作者可以考虑引用已有研究来详细地介绍不同类型的题目，例如，客观题（选择题、判断题），简答题，写作题等会引起什么样的 TME。此外，不同题型涉及的认知过程（cognitive processes）也可能会引起 TME。
回应：非常感谢您指出这一点。我们已在“2.2 题目层面”的“（2）题目类型”中增加/补充对相关内容的介绍，详见修改稿的第 4 页和第 5 页。

意见 3：除了文章介绍的几种 TME 来源之外，审稿人认为还有一种是评分者效应，对于 PBT 和 CBT 来说，客观题的评分程序应该是相同的，答案是唯一的，所以较少或不会受到测验模式的影响。相反，对于主观题而言，目前仍需要通过人类主观进行评分，更有可能受到测试模式的影响。
回应：谢谢您的宝贵建议。我们已经将评分者效应添加为 TME 的来源，详见修改稿第 5 页和第 6 页的“2.4 评分者层面”小节。

意见 4：如果还能像文中表 2 那样，将 TME 的来源总结成一个表格，说明每一种来源的 Mode Effect，可能更容易让读者快速提取信息。
回应：谢谢您的建议。我们已在修改稿的第 6 页添加表 1（原文中表 1 和表 2 分别变为表 2 和表 3），以总结 TME 的来源。

意见 5：TME 的检测方法中介绍了几种侦测方法都用到了 R package，对于每一种方法的使用，为了方便实践者的应用，建议作者提供一个代码示例来说明其使用过程。
回应：谢谢您的宝贵意见。我们将实现各种 TME 检测方法的 R 代码示例放在附录部分，详见修改稿的第 22 页和第 23 页。

审稿人 2 意见：

意见 1：基于传统的纸笔测验（PBT）或者基于计算机测验（CBT）这样的不同的测验模式

可能引出测验模式效应（TME），测验模式效应对测验公平、选拔标准和测验等值都有一定的影响。作者对 TME 的来源、检测和研究结果进行了比较系统的介绍。文章选题有意义，写作条理清楚，重点突出，是一篇比较好的综述文章。
回应：谢谢您对本文给予的积极的、正面的评价。

意见 2：但是审稿人认为测验模式效应的来源相当复杂，比如：传统“纸笔测验（Paper-based Testing, PBT）”，“计算机测验（Computer-based Testing, CBT）”有一个区别：CBT 是机器评分，PBT 是专家（人工）评分；机器评分没有主观因素影响，没有疲劳效应和光环效应，而人工评分受到疲劳效应和光环效应的影响，所以这可能是那些评估测验模式效应（TME）的方法难以解决的问题；审稿人对这一点似乎有点儿担心。
回应：我们非常认同您的观点“机器评分不受主观因素影响，没有疲劳效应和光环效应，而人工评分受到疲劳效应和光环效应的影响”。但是，“CBT 和 PBT 是机器评分还是人工评分”可能取决于题型。比如，CBT 中的部分建构题（如中文作文题）目前仍以人工评分为主，而 PBT 中的选择题借助答题卡也能实现机器评分。我们在修改稿的第 5 页对此问题进行了一些分析。另外，我们非常同意您关于“评估 TME 的方法难以解决人工评分的影响”的担心，对此我们在修改稿第 16 页纳入对“如何避免评分者影响”的讨论。谢谢您指出这一点。

意见 3：审稿人只是了解比较早的 CBT，那时候几乎每一个项目都配置了一个“进度条”，对作答时间倒数，这对被试压力很大，PBT 就不存在这样的问题；现在的 CBT 系统是否仍然存在这个问题，审稿人不得而知。是否还有其他的一些影响因素，研究 TME 的人员应该去了解，文章的针对性就更强。
回应：谢谢您的宝贵意见。我们在“2.1 测验层面”小节增加“（4）测验计时与选题方式”部分，专门对 CBT 中的计时设计以及选题方式可能引起的 TME 进行探讨，详见修改稿的第 3 页和第 4 页。

意见 4：测验的题型可能是造成 TME 的比较大的影响因子，如造句、短文写作，机器评分和专家（人工）评分差别就比较大；而选择题（包括多选题）两者差别就可能比较小；这一点文章阐述得比较清楚，但是如何评估题型和测验形式的交互效应，还是比较困难。
回应：谢谢您指出这一点。如果我们没有理解错的话，审稿专家本条意见的最后一句话应该是指“如何评估题目和评分者间的交互效应，还是比较困难”。我们在“2.4 评分者层面”小节对此问题进行了述评，详见修改稿的第 5 页和第 6 页。

意见 5：另外审稿人想和作者商量如下问题，即对于认知诊断测验的 PBT 和 CBT（CD-CAT）的 TME 似乎应该提一笔，因为认知诊断测验日益受到实际工作者和研究人员的重视，所以至少在文章的小结与讨论中应该提一笔。审稿人认为检测认知诊断测验中的 TME 可能更加麻烦一些，但是作为今后的可能遇到的问题应该交代一句。
回应：我们完全同意您的观点。根据您的建议，我们在讨论部分展望了基于认知诊断测验的 TME 研究方向，详见修改稿的第 15 页。

意见 6：另外文字表达中可能存在一些笔误，比如：摘要中“全面展示有效检测 TME 的方法论”这个“论”字是否应该删除？
回应：谢谢您指出这一点。已将原文修改为“全面展示 TME 研究的方法论”，详见修改稿的第 1 页和第 24 页。

意见 7：审稿人认为进行一些修改以后可考虑录用。
回应：谢谢您对本文的肯定与认可。

第二轮

审稿人 1 意见：

作者介绍了测验模式效应（Test Mode Effect, TME）的来源，侦测方法，文章写作逻辑清晰，有述有评，对于关注 TME 的研究者具有重要的借鉴价值。经过第一轮的修改，作者已完全解答了审稿人提出的几点疑问，文章内容显得更加完整和丰富，建议发表。
建议作者在附录的代码中添加注解，解释每一行的功能，方便不熟悉 R 语言的读者使用。
回应：谢谢您对文章的认可和肯定。根据您的建议，我们已为每一行代码添加注释，详见修改稿的第 22~23 页。

第三轮

编委 1 意见：

这篇文章经过两轮审稿和修改，已经有很大改进。同意两位审稿人意见，建议发表。因为文章较长，建议作者将可有可无的内容删减。
回应：谢谢您的积极反馈。根据您的建议，我们对文章中可有可无的内容进行了删减。

编委 2 意见：同意发表。

主编意见：同意发表