《心理学报》审稿意见与作者回应

题目：两种新的计算机化自适应测验在线标定方法
作者：陈 平

第一轮

首先非常感谢三位审稿专家及编委会专家的辛勤劳动，感谢您们提出这么多宝贵的建设性意见与建议。接下来我们对三位审稿专家提出的意见逐条进行回复，并希望修改后的稿件能够得到各位专家的认可。修改部分见黄色高亮部分。

审稿人 1 意见：

意见 1：在线标定是一种为计算机化考试题库更新进行新项目实时预试技术。在信息技术飞速发展的今天，基于计算机的考试方式已经成为越来越普及的方式。因此，在线标定的研究具有很大的实际应用价值。本研究将两种估计误差方法与在线估计的常用方法之一 Method A 结合，解决了 Method A 中不能对能力估计进行实时修改的问题，从而提高了在线项目参数估计的精确性。因此，本研究具有较高的理论与实践贡献。同时，为了其他研究者更好借鉴本研究，有些问题需要进一步说明。优势：

第一，Method A 是在线估计最常用的方法之一，本研究提出的新方法改进了 Method A 的固有问题，因此具有较高的理论与实践意义。

第二，本文对全功能极大似然估计方法 (FFMLE) 与“利用充分性结果”估计方法 (ECSE) 以及如何将它们运用到 Method A 进行了细致的介绍，行文非常清晰。

第三，这两种新方法不仅可以提高参数估计精度，并且可以缩减参数估计的时间。这对实施实时化的计算机化考试具有重要的意义。

回应：非常感谢您对本研究给予的正面的、积极的评价，也感谢您对计算机化自适应测验在线标定研究领域的支持。

意见 2：MEM 法实际上也在处理本文涉及的能力估计误差问题。MEM 的优势是方法简单直观，它利用 EM 多次循环的结构，自动修正能力估计误差，从而提高了项目估计的精度。因此，可以说它与本文提出的方法殊途同归。从文中的结果看来，新方法的精度提高与 MEM 相当；在运算时间角度，比 MEM 下降了很多。但是从结果看，我不能肯定这种时间提高是否具有显著的实际意义。

回应：非常同意您的观点。首先，MEM 在 M 步中是通过最大化对数边际似然函数来估计新题的题目参数，而边际似然函数是在联合似然函数的基础之上通过积分把能力 θ 积掉而得到。所以从本质上讲，MEM 在标定新题时通过积掉 θ 也能控制能力的估计误差，因此新题

参数的估计值不受能力估计误差的影响。Ban, Hanson, Wang, Yi 和 Harris (2001)[1] 的研究也表明：相对于 Method A、Method B、OEM 和 BILOG/Prior 等方法，MEM 的表现最优。

其次，本文提出的两种新方法较 Method A 可以改进标定精度且与 MEM 的表现相当，而且标定效率 (用时不到 0.02 秒) 远高于 MEM (用时范围在 6.0827 秒与 21.0330 秒之间)。正如您所评论，从用时结果来看，这种时间上的提高没有太大的实际意义，因为即使采用算法最为复杂的 MEM 也只需 22 秒不到的时间即可完成标定任务。但是当将这些标定方法推广到多维情境时 (比如多维 CAT)，两种新方法较 MEM 的时间优势就开始突显。我们在一项预研究 (考虑 3 个能力维度、3600 名被试、总共 900 个旧题和 18 个新题、每名被试作答 6 个新题、测验长度为 40) 中发现，两种新方法的多维版本只需 2 秒以内的时间即可完成标定，而 MEM 的多维版本则需要 1 至 2 个小时的运行时间，这在实践当中可能会难以接受。

我们已经将上述观点或内容补充到文章的修改稿中，详见第 17 页和第 18 页。

意见 3：新方法需要大样本才能发挥出了较好的效果 (MLE 方法都需要大样本才能保证效果)，因此这是新方法的局限性之一。将来应当重点考虑在小样本中的方法改进。

回应：完全同意您的建议。我们已经在修改稿中增加相关内容讨论这个问题，详见第 20 页。非常感谢！

审稿人 2 意见：

意见 1：选题有意义也有创新性。研究结果较为可靠，但结果可靠并不意味着就是好结果。本文提出的两种新方法对 *b* 参数的估计精度在多数条件下还不如原有的 Method A 方法，这个结果就不是个好的结果。建议将这两种新方法用于某些只有 *a* 参数而没有 *b* 参数的特殊 IRT 模型上，看看是不是有好的效果。详情见附件。

回应：非常感谢您对本研究在选题意义、创新性及结果可靠性方面的认可，也谢谢您提供的宝贵建议与意见。由于您这里提到的意见在如下呈现的附件内容中也有完整体现，所以接下来我们逐条回复附件内容中的意见。

附件内容：

意见 2：对新题的项目参数进行在线标定，是计算机化自适应测验中的重要研究内容，近年来《心理学报》也曾经发过这个方面的研究论文。对于先前学者提出的 Method A 方法，本文找到了它在理论上的不足之处，并加以改进。选题有理论意义和实践意义，具有创新性，研究设计合理。

回应：非常感谢您对本文在研究意义、研究原创性以及研究设计合理性等方面给予的正面评价。

---

[1] Ban, J. -C., Hanson, B. H., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item—calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

意见 3：对于这篇论文，最大的问题在于研究结果和结论。从表 4 和表 5 看，本文提出的两种新方法对 b 参数的估计精度在多数条件下还不如原有的 Method A 方法。那么，在结论部分说两种新方法比原有的 Method A 方法有改进，就是不合适的。从这篇论文的结果看，似乎说“新方法对 a 参数的估计精度上优于 Method A 方法，而在 b 参数的估计精度上不如 Method A 方法”更合理。既然如此，研究意义就要大打折扣了。

回应：非常抱歉我们对这部分结果与结论的相关描述还不够准确，因此进行以下修改：(1) 对于表 4 的结果，我们进行了更为细致的分析，并以脚注形式增加“在各样本量下，相对于 M2 方法，M3 与 M4 两种新方法分别在 a 和 b 参数上的 RMSE 值降低或增加的比例”的内容，目的是想说明：如果将 a 和 b 的标定精度看成同等重要时，两种新方法总体上优于 M2 (详见修改稿的第 15 页)；(2) 对于表 5 的结果，我们同样以脚注形式增加相关内容，并更为明确地指出：如果把 a 和 b 的标定精度看成同等重要的话，只要当样本量较大 (N = 2000 和 3000) 时，M4 的表现才在总体上优于 M2 (详见第 16 页)；(3) 在文章的相应部分更为准确地描述两种新方法的表现，比如“总体上，两种新方法在大多数实验条件下可以改进 Method A 的标定精度”，详见修改稿的“论文自检报告”第 1 页、正文的第 1、18 和 19 页。

另外，“在大多数实验条件下，新方法能够提高 a 参数的标定精度，而且 a 的标定精度改进程度都大于 b 的标定精度降低程度”已经非常重要，因为 a 参数本质上是模型中 θ 的回归系数，非常容易受到 θ 的测量误差的影响。

意见 4：以上缺陷足够使得这篇论文被拒稿。但鉴于文章的选题很有意义，我想给作者一次修改的机会：尝试将这两种新方法用于某些只有 a 参数而没有 b 参数的特殊 IRT 模型上，看看是不是有好的效果。如果效果好，可以将文章大幅修改以后重新送审。

回应：非常感谢您给我们再次修改的机会。也谢谢您为我们出主意——“尝试将这两种新方法用于某些只有 a 参数而没有 b 参数的特殊 IRT 模型上”，对此我们想做如下解释：(1) 本文提出的两种新方法是基于标准形式的逻辑斯蒂克回归 (LR) 的理论发展而构建，而本文讨论的两参数逻辑斯蒂克模型 (2PLM) 即可视为含能力 θ 的标准形式 LR 模型。对于其他的模型，FFMLE 和 ECSE 是否仍具有优良的统计特性还值得进一步考证；(2) 经过查阅文献书籍、征求同事意见后，我们发现“只有 a 参数而没有 b 参数、且要满足 LR 模型标准形式的特殊 IRT 模型”确实非常难找，所以没有考虑新的模型。但是对于文中表达不准确或不精确的地方，我们都进行了相应修改。希望您能理解并满意我们的决定。

此外，还发现了一些细节上的问题，其中多数是文字表述上的问题，列举如下：

意见 5：第一段“传统纸笔测验……高能力被试需要作答较多的容易题，低能力被试需要作答较多的难题”这个表述有歧义，容易让读者以为不同能力的被试作答的题目不一样。请修改。

回应：非常感谢您的建议。为了消除歧义，我们将这段话改为“……于是，题目对高能力水平被试而言大多比较容易、对低能力水平被试来说大多比较难，不利于对被试能力水平的估计”，详见修改稿的第 1 页。

意见 6：在 2.1 节里，为什么选用 2PLM 模型，文章给出的理由是：“由于本文提出的新方法

(FFMLE-Method A 和 ECSE-Method A) 是基于逻辑斯蒂克回归 (*Logistic Regression*, LR) 的理论发展而构建并且两参数逻辑斯蒂克模型 (*Two-Parameter Logistic Model*, 2PLM) (Birnbaum, 1968) 可视为包含一个潜变量 θ 的 LR 模型，所以本文选择 2PLM 作为 IRT 模型。”但这个理由不是很充分，因为有其他模型（如 3PLM）也同样符合以上理由。

回应：非常抱歉我们在原文中可能没有描述清楚这一点。因为本文提出的两种新方法 (FFMLE-Method A 和 ECSE-Method A) 是基于 Stefanski 和 Carroll (1985)[2] 的 FFMLE 和 ECSE 方法而构建，而 FFMLE 和 ECSE 方法又是基于标准形式的逻辑斯蒂克回归 (LR) 框架而开发 (详见其文章的第 1 页，标准形式的 LR 形如：$\Pr\{y_i = 1 | x_i\} = (1 + \exp(-x_i^T \beta_0))^{-1}$)，又因为两参数逻辑斯蒂克模型 (2PLM) 的项目特征函数 (ICF) 与 LR 模型的标准形式完全相符，所以本文选择 2PLM 作为 IRT 模型。至于其他模型，比如 3PLM，它的 ICF 是在 2PLM 的 ICF 的基础上乘以 (1-*c*) 后再加上 *c* 而得到；对于这种对标准形式的 LR 模型进行简单变换后的模型，FFMLE 与 ECSE 是否仍具有优良的统计特性还有待进一步的考证，所以本文暂时没有考虑 3PLM。希望您满意我们的解释。

相应地，我们对文章中的相关描述进行了修改，请参见修改稿的第 3 页。

意见 7：第 3 节第二段，“本模拟研究采用 3×3 的实验设计”的说法不当。在这个研究里，在线标定方法也是一个自变量，它具有 5 个水平，所以这里是 $3\times3\times5$ 设计。

回应：非常感谢您的建议。对于本模拟研究所考虑的因变量 (衡量新题标定精度的 RMSE 等评价指标)，确实受到 3 个不同因素的影响，分别是有 3 个水平的样本大小、3 个水平的测验长度和 5 个水平的在线标定方法。我们已经在修改稿的第 7 页和第 8 页中调整了相关描述。

意见 8：第 3 节第二段，即使配置完全相同，9 台虚拟机本身的运行速度也会有一些差异。直接将它们视为无速度差别，可能是不妥的。可以先让 9 台虚拟机运行一个相同的程序，以找到 9 台虚拟机本身速度的关系。

回应：非常感谢您指出这一点。根据您的建议，我们在这 9 台配置完全相同的虚拟机 (Virtual Machine，简记为 VM) (配置是：2.60 GHz 的双核 AMD 处理器、8GB 的内存以及 64 位的操作系统) 上运行相同的一个 Matlab 程序。该程序的功能是计算从 1 加到 1 千万的求和值，程序代码 (也可简化程序，用向量运算代替 for 循环) 以及运行时间结果呈现如下：

程序代码：

```
tic;
sum = 0;
for i = 1:10000000
    sum = sum + i;
end
Running_Time = toc;
```

[2] Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics, 13*, 1335-1351.

运行时间结果 (单位是秒，保留 4 位小数)：

编号	VM 1	VM 2	VM 3	VM 4	VM 5	VM 6	VM 7	VM 8	VM 9
时间	8.3140	8.4510	8.7361	8.3432	8.5561	8.4430	8.5341	8.4130	8.6268

从表中可以看出，即使在配置完全相同的多个虚拟机上运行同一个程序，也会得到稍微不同的运行时间结果 (差异主要体现在小数点后第 1 位或第 2 位上)。另外，为了满足好奇心，作者还将相同的 Matlab 程序在一台性能稍好的笔记本电脑 (2.67 GHz 双核 i7 处理器、4GB 内存以及 64 位操作系统) 上重复运行 5 次，也得到不同的运行时间，分别是 7.1955 秒、7.3066 秒、7.3344 秒、7.2718 秒以及 7.3511 秒。究其原因，可能是因为电脑的硬件运行环境在不同时间点都会不同，比如电脑在运行一段时间后，CPU、内存以及硬盘等的状态 (比如温度) 都会发生细微变化，从而影响运行速度。

注意上述“配置相同虚拟机运行同一程序”产生的微小时间差异并不影响本文的主旨。因为本文主要是想在各种 CAT 测验情境 (样本量与测验长度的不同组合) 内比较不同在线标定方法的运行效率，而不想比较不同样本量或不同测验长度对在线标定方法运行时间的影响。所以从严格意义上讲，原文对表 7 结果的部分描述——“另外，对表中数据进行观察，还可发现当样本量增大时，平均运行时间严格单调递增，这与期望也是一致的”不是特别精确，因为没有排除上述时间差异，所以在修改稿中删除了这句话。另外，我们还在修改稿第 8 页中修改了相关语句并增加脚注内容，以使描述更为准确。

意见 9：表 1 里的相关系数，不列进表里可能更合适一些，改为在正文里用文字表述。

回应：完全同意。我们把关于相关系数的描述部分从表 1 中移除，并在正文中增加文字描述“而且模拟生成的区分度与难度之间的皮尔逊积差相关系数等于 0.2507，与预设的真值 0.25 非常接近”，详见修改稿的第 8 页和第 9 页。

意见 10：第 3.3 节第二段，积分结点为什么选 61 个？请说明理由。积分结点的数量可能需要反复尝试以找到最佳值。

回应：非常同意“积分结点数量需要反复尝试以找到最佳值”的观点。本研究之所以选取 61 个积分结点主要是因为，我们在预研究中考查了 3 种不同数量的积分结点数 (分别为 21 个、41 个和 61 个) 对 EAP 估计精度的影响，结果表明：使用 21 个积分结点的精度最差，采用 41 个积分结点已经可以得到较高的估计精度，但为了保险起见，我们还是选取估计精度最高的 61 个积分结点 (这部分解释已以脚注的形式添加到修改稿第 9 页)。另外，为什么选取 61 个积分结点而不是 60 或 62 呢？这主要是为了方便运算，此时步长刚好等于 0.1。

意见 11：第 3.3 节第二段，最后一句，“不管是采用 EAP 方法还是 MLE 方法，都将最终的能力估计值截取在 [-3, 3] 之间。” 如果估计出的能力越过这个区间，是强行把估计结果赋值 3 或-3 吗？请说得更清楚一些。

回应：为了使得表达的含义更加清楚，我们在这句话后面增加了一些解释，即“当得到的能力估计值大于 3 时，将其赋值为 3；当得到的能力估计值小于-3 时，将其赋值为-3”，详见修改稿的第 9 页和第 10 页。非常感谢您指出这一点。

意见 12：第 3.4.1 节，“从由 20 个新题组成的新题题集中随机选择 5 个新题 (即 $D = 5$)”。这里的 D 和 2PLM 模型里的 D 不是同一个变量，却用了相同的符号，能否换个符号以示区分？

回应：完全同意。我们在修改稿中使用另外的符号 (即大写的字母 *C*) 来表示每名被试作答的新题个数 (即 *C* = 5)，详见第 10 页。

意见 13：行间距没有统一。

回应：谢谢您的建议。在修改稿中，我们将行间距统一设置为“1.25 倍行距”。

意见 14：破折号的写法是“——”，而不是三个连字符“---”。

回应：我们将文章中表示“破折号”的 8 处连字符“---”全部修改为“——”，详见修改稿的“论文自检报告”页、第 2、5、10、11、19 和 21 页。谢谢。

意见 15：公式编号到 (21) 就结束了，后面的公式都没有编号。

回应：谢谢您指出这一点。根据您的建议，我们对“3.5 评价指标”部分呈现的 11 个评价指标也进行了编号，编号从 (22) 到 (32)，详见修改稿的第 11 页和第 12 页。

审稿人 3 意见：

意见 1：计算机化自适应测验 (CAT) 有许多优点，得到越来越广泛的应用。实施 CAT 需要一个题库。建立题库的成本高，技术要求高，有时候等值相当困难，这些是实施 CAT 的障碍。 “两种新的计算机化自适应测验在线标定方法”试图提供克服这些困难的一种有理论支持的新方法。文章研究的问题很有意义，逻辑清晰，文字表达通畅，结果比较好。

回应：非常感谢您对本文的积极评价。

审稿人认为，文章在某些问题的表述上，值得商榷，这主要表现在：

意见 2：正文 p.14 的如下表述“(2) 在所有 3 种样本量下，M3 和 M4 在参数 *a* 上的 RMSE 值都比 M2 的稍微低一些，然而它们在参数 *b* 上的 RMSE 值都要比 M2 的稍微高一些。这主要是因为 2.1 节所示的 2PLM 属于补偿性模型，要同时提高 *a* 和 *b* 的标定精度可能会比较困难，而且能够提高 *a* 参数的标定精度已经非常重要，因为它本质上是模型中 θ 的回归系数，非常容易受到 θ 的测量误差的影响。总体上，M3 和 M4 的表现还是优于 M2”。

这一段文字中，单维 2PLM 为什么是补偿模型没有交代清楚，一般认为多维情形才讨论是否补偿。审稿人认为作者这种说法难以成立；事实上，论文引用的文献 (游晓锋等人, 2010) 相关的研究中，使用的是单维 2PLM，游晓锋等人 (2010) 的结果表明 *a* 和 *b* 参数的估计精度同时都可以改进，这似乎表示作者所说的理由有一点勉强；反过来请作者想一想，为什么作者提出的新方法不能够做到这一点？

回应：非常感谢您指出这个错误。单维 2PLM 确实不能称为补偿性模型，而且“是否补偿模

型”一般只在多维 IRT 领域或认知诊断领域中才加以讨论，对此我们在修改稿中进行了纠正(详见第 15 页)。另外，至于为什么本文提出的两种新方法没能像游晓锋等人 (2010) 的方法(即 CMLE 及多重迭代 CMLE 方法) 一样可同时改进a和b的估计精度，我们在修改稿中描述了可能的原因：“a参数本质上是 2.1 节所示模型 (即 2PLM) 中θ的回归系数，非常容易受到θ的测量误差的影响；M3 与 M4 对$\hat{\theta}$中蕴含的测量误差进行校正，从而可提高a的标定精度，但是并未采取类似于“夹逼平均法” (游晓锋等人, 2010) 的任何措施以提高b的标定精度” (详见第 15 页)。

意见 3：正文讨论部分第二段指出“(2) 当 CAT 测验长度较短 (比如 10 题) 时，新方法对 Method A 的改进程度最大”，这一点和正文 p.6 步骤 6 下面关于题目量 *t* 的阐述有什么关系，能否进一步解释；

回应：非常感谢您的建议。对于两者的关系，我们在修改稿中增加了描述与解释，详见第 19 页的脚注部分。谢谢！

意见 4：另外作者固定了 *a* 与 *b* 的一个相关系数，是不是可以考查相关系数变化对 *a* 和 *b* 估计精度的影响？即将相关系数作为一个因素加以考虑，看一看是不是相关系数比较大时，*a* 的估计精度的提高能否带动 *b* 的精度的提高。

回应：非常感谢您提供的好建议。根据您的建议，我们又考查了当 *a* 和 *b* 之间的相关系数分别为 0.5 和 0.75 时 (分别代表中等相关和高度相关)，9 种 CAT 测验情境下 (3 种样本量与 3 种测验长度的组合) 5 种在线标定方法的标定结果。限于篇幅，以下仅呈现相关系数为 0.50 且测验长度为 20 时，各种方法的标定结果：

样本量 (*N*)	标定方法	RMSE		Bias		*r*		WMSE
		a	*b*	*a*	*b*	*a*	*b*	
1000	M1	0.1355	0.2022	0.0024	0.0150	0.9499	0.9804	0.0013
	M2	0.1522	**0.2096**	-0.0353	0.0454	0.9415	0.9799	**0.0015**
	M3	0.1475	0.2140	0.0152	0.0487	0.9410	0.9798	**0.0015**
	M4	**0.1465**	0.2138	0.0002	0.0470	0.9413	0.9798	**0.0015**
	M5	0.1469	0.2147	0.0160	0.0498	0.9414	0.9799	**0.0015**
2000	M1	0.1007	0.1552	0.0062	0.0112	0.9720	0.9890	0.0006
	M2	0.1169	**0.1612**	-0.0310	0.0326	0.9664	0.9885	0.0007
	M3	0.1129	0.1670	0.0239	0.0368	0.9670	0.9885	**0.0008**
	M4	**0.1100**	0.1670	0.0069	0.0354	0.9670	0.9885	**0.0008**
	M5	0.1126	0.1680	0.0256	0.0375	0.9674	0.9886	**0.0008**
3000	M1	0.0843	0.1390	0.0024	0.0131	0.9823	0.9913	0.0004
	M2	0.1080	**0.1478**	-0.0419	0.0517	0.9784	0.9912	**0.0006**
	M3	0.0944	0.1549	0.0139	0.0553	0.9783	0.9912	**0.0006**
	M4	0.0943	0.1548	-0.0035	0.0539	0.9783	0.9911	**0.0006**
	M5	**0.0941**	0.1560	0.0156	0.0559	0.9785	0.9912	**0.0006**

注：(1) M1 = Method A(True), M2 = Method A(Original), M3 = FFMLE-Method A, M4 =

ECSE-Method A, M5 = MEM; (2) 表中加粗数字代表 4 种方法中 (除基准方法 M1 以外) 表现最好者。

从上表中可以看出：当相关系数比较大 (0.50) 时，*a* 的估计精度的提高**并不能**带动 *b* 的精度的提高。而且，当相关系数为 0.75 时，也得到类似的结果。所以，修改稿中暂未将相关系数作为一个因素进行考虑，非常感谢！

第二轮

再次感谢两位审稿专家、编委会专家以及编辑部老师的辛勤劳动，感谢您们提出的宝贵意见。接下来我们对两位审稿专家的意见或评论逐条进行回复。修改部分详见红色高亮部分。

审稿人 1 意见：

意见 1：看完了作者的修改稿。作者对审稿人提出的所有意见都进行了认真回答：有些地方根据审稿人的意见做了认真修改，有些实在改不了的地方也作出了认真而诚恳的解释。经过修改，论文的质量已经有了很大提高，而且在文字表述上也变得更加规范。我同意发表。

另外，如果这篇文章真的能发表，我建议各位读者向本文作者学习“如何回答审稿人的问题”。

回应：非常感谢您对修改稿给予的正面的、积极的评价。

审稿人 2 意见：

意见 1：作者比较充分地回应了审稿人的意见。建议图 1 中的矩阵采用 Kronecker 积表达，可以节省篇幅，而且可以将文章中 750 个相同的矩阵纵向合并的表达式简洁明了地表示。建议接受发表。

回应：谢谢您对修改稿的积极评价。我们已在修改稿中采用克罗内克积 (*kronecker product*) 表达图 1 中的矩阵并提供“由基本矩阵单元 $\mathbf{V}_B$ 得到随机矩阵 $\mathbf{V}$ ”的简洁表示，详见第 10 页的红色高亮部分。非常感谢您指出这一点。

第三轮

再次感谢编委会专家及编辑部老师提出的宝贵意见与建议。接下来对编委会专家的建议或评论进行回复。修改部分详见灰色高亮部分。

编委专家意见：

意见 1：The current article, excluding abstract and reference is 16000 words long. Will recommend to trim down to around 12000 words.

回应：We sincerely appreciate your suggestion. Under the premise of not affecting the readability and integrity, we have tried our best to shorten the manuscript (including deleting the redundant expressions and reducing the redundant simulation details, etc.) from 16027 words to 14162 words (including seven tables). Hope you are satisfied with our efforts.

意见 2：There are 60+ references, can reduce to around 40??

回应：Yes, we have reduced the number of references to 40 (see pages 19~21). Actually, the original manuscript includes 43 references rather than 60+ references, because the Chinese references should be followed by their English versions according to the format requirement of *Acta Psychologica Sinica*. Many thanks.

意见 3：Figure 1 is not a figure, it is an equation???

回应：Totally agree. It is an equation rather than a figure, and we have revised it on page 9 in the revised manuscript. Thank you for pointing this out.

意见 4：English abstract slightly polished for the authors' consideration (see attached).

回应：Thank you very much for helping us polish the English abstract. The English abstract has been greatly improved by following your advice, see pages 21~22 for details.