《心理学报》审稿意见与作者回应

题目：四参数 Logistic 模型潜在特质参数的 Warm 加权极大似然估计
作者：孟祥斌 陶剑 陈莎莉

第一轮

审稿人 1 意见：

意见 1：公式(19)既然采取迭代，迭代的初值如何取？终止规则是什么，应该简略描述。
回应：首先感谢您的修改建议。没有对 N-R 迭代初值的选取和迭代终止规则进行说明是我们的疏漏。根据您的建议，修改稿中对 N-R 迭代初值的选择策略和迭代终止规则进行了详细说明。

意见 2：(28)式 EAPE 的结点个数 K 似乎有推荐值，可以参见 Baker(1992)，或者 Baker& Kim(2004)的 IRT 参数估计技术一书，或者漆书青，戴海琦，丁树良(2002)的书，高等教育出版社。
回应：感谢您的修改建议和提供的参考资料。认真阅读您提供的资料，收获很大，但没有查找到对 EAPE 数值计算所需节点个数的推荐值。在 Baker& Kim（2004）的代码部分，K 默认取为 10，但没有给出理由。通过进一步查找资料，我们在下面这本书中，“IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT” (Edited by Mathilda du Toit) 查找到一个关于 K 的取值策略：“为了节省计算时间，求积节点数可取为项目总数平方根的 2 倍。”于是，修改稿中对这一策略及相关出处进行了介绍。此外，考虑到 Baker& Kim（2004）和漆书青，戴海琦，丁树良（2002）这两本书对 EAPE 的理论及其它参数估计方法都进行了非常详细的介绍，有必要把它们推荐给读者，于是我们在文中进行相应的引用和说明。

意见 3：什么是估计的有效性，为什么 RMSE 可以用来衡量估计的有效性，文章没有交代。
回应：首先，感谢您的建议。通过讨论和查阅文献，我们确定“估计的有效性”这一称呼是不恰当的，应使用“估计的返真性能”，我们已对此进行修改，请见正文。RMSE 是评价估计返真性能的常用指标，在很多文献都有详细的介绍，文中不必对此进行详细说明。

意见 4：另外两个评价估计好坏的指标 ME（第 32 式）和 ABME（第 33 式），显然 ABME 是 ME 的绝对值，有必要用两个指标描述估计的偏差吗？通常是用 ABS 描述估计的绝对偏差的平均值，而不是作者使用的 ABME。
回应：（1）ME 是偏差（bias）的评价指标，它能够反映偏差的方向（正或负）和大小。因为有“正负”方向，使用它对不同估计的偏差大小进行比较，不够清晰和直接。本研究关心的重点是加权似然方法是否具有更小的偏差，权函数的纠偏效果是否有效。为此，我们决定使用 ME 的绝对值(即 ABME)对三种估计的偏差大小进行比较。本文使用 ME 描述估计偏差的方向变化，ABME 用于比较估计偏差的大小,相关解释已在文中给出。（2）ABS 与 RMSE 类似，是一种参数估计值与真值平均距离的度量，ABS 基于欧氏距离，RMSE 基于平方距离，所以 ABS 是评价估计返真性的，不是评价估计无偏性。本研究已使用了 RMSE 作为估计返真性的评价指标，也就没有必要再使用 ABS 了。为了解释的更为直观，我们选择以下

例子是对 ABS 进行补充解释：“假设某参数真值为 0，进行 1000 次模拟，得到该参数的 1000 个估计值，其中 500 个1,500 个 –1 。”这 1000 个估计值等距离均匀分布在参数真值的两侧，通过计算可得 ME=0, 即估计的平均值与真值相等，该估计是无偏的。但 ABS=1，与 ME 不同。ABS=1 表示参数的估计值与真值的平均欧氏距离为 1，是参数估计对参数真实值反应能力的度量。

意见 5：公式（4）的加减符号请斟酌。
回应：谢谢您指出的错误。已经修正，请见正文。

意见 6：公式（9）到底对 i 还是 j 相加？
回应：公式（9）是对 *j* 相加。已经修正，请见正文。

意见 7：公式（24）K 到底是对 i 还是 j 求和？
回应：公式（24）K 是对 *j* 求和。已经修正，请见正文。

意见 8： 附录（a8）求和符号对 i 还是对 j 求和？
回应：附录（a8）求和符号是对 *j* 求和。已经修正，请见正文。

意见 9： 图 1 的第 2 行第 2 列的标注 m=15 应该是 m=30;
回应：感谢您的建议。已做修正。

意见 10：作者说“模型包含的参数越多，理论就更具一般性”，似乎应该是：模型描述的现象越广泛；
回应：已按照您的建议进行修改，请见正文。

意见 11：错别字问题：渐近线，还是渐进线？
回应：已经修正，请见正文。

意见 12： 第 5 面（15）式上方第 3 行漏字。
回应：已补充漏字“进行”，请见正文。

意见 13：建议作者再仔细推敲所有公式，包括求和的变量，变量取值的范围，以及文字，进行修改。
回应：感谢您的建议，我们已对全文进行了认真的检查和校对，针对本文的文字和公式中的不妥之处进行了相应的修改。

审稿人 2 意见：
意见 1. 四参数模型提出的背景是，高能力的被试在低难度项目上不一定做出正确反应，因为高能力的被试有可能在低难度项目上犯错误，因此增加了一个参数。问题是无论是二参数或者三参数模型，高能力的被试在低难度上的的正确反应概率其实都不可能等于 1，只是接近于 1。已经包含了四参数模型所说的情况。而且即使概率为 1 也不是说就一定做出正确反应.因此增加一个参数其实是不必要的，本审稿人不认为 4PL 模型将成为未来主流的 IRT 模型；

回应：感谢您的评论。项目反应模型所刻画的是项目反应概率，即某种潜在特质水平下的作答概率，与作答失误无关。从模型的假设角度看，1PL,2PL 和 3PL 就是没有包含作答失误的可能，因为反应概率无限趋近与 1。作答失误的定义是指，能力再强的被试都会存在一定概率的作答错误，即是能力无穷大，反应概率也不能为 1，所以上渐近线要小于 1。按照您的理解和逻辑，3 参数模型的猜测度参数也是无需引入的，因为 1 参数和 2 参数模型所刻画的反应概率不等 0，而是无限接近 0，说明 1 参数和 2 参数已经包含了低能力被试猜测的情况。4PL 模型的理论价值已毋庸置疑，近年来的一些研究也不断验证 4PL 模型的应用价值，本文的第一部分已给出大量相关文献，您可查阅。

意见 2. 作者在参数估计中选用的算法是 N-R 算法，然而理论与实践都表明，这一算法要求目标函数必须是凸函数，否则迭代点列可能发散，在这种条件下，无论是 MLE 或者 WMLE，都不能保证满足凸函数条件，除非初始点非常靠近真值，在实践中这一点难以做到。
回应：虽然 N-R 算法存在一定的不足，但它因收敛速度快而广受青睐。不可否认，N-R 迭代仍是 IRT 领域极大化的常用算法。初值的选取对 N-R 迭代非常重要，也是 N-R 迭代的不足之一，但如果采取恰当的取值策略，结合恰当的终止规则，由初值所带来的不收敛现象还是能在很大程度上避免的，并非完全不能解决。我们已在文中给出了本研究所采用的初值选取策略和终止规则条件，请见正文。

意见 3. WMLE 与贝叶斯估计颇为类似，只是先验概率选取方法不同。而且 WMLE 计算复杂，作者只是模拟了项目参数已知的条件下能力参数的估计，回避了项目参数估计，实际上用 N-R 算法估计项目参数会很麻烦，迭代点列很多都会发散，由于这一原因，现在 IRT 参数估计很少使用 N-R 算法，通常选用 MCMC 算法
回应：项目反应理论模型的参数估计分成两部分：项目参数估计（项目参数的标定）和被试潜在特质参数估计（为被试评分）。本研究是仅对被试能力参数估计进行研究，提出适用于 4PL 模型潜在特质参数估计的 WMLE 方法，所以我们并未回避项目参数估计，而且论文的题目已清晰说明是“潜在特质参数的加权极大似然估计”。在 IRT 模型参数估计方面，MCMC 的确比较流行，因为模型的复杂度到达一定程度，模型的参数估计按照常规的数值计算方法难以实现，唯有求助于 MCMC 算法。MCMC 算法的优势是：即使模型复杂度增加，它算法依然可以顺利实现。但它也有很多不足，例如，对先验的过分依赖、不能收敛到平稳状态、需要很高的计算量、模型可识别性的影响，等等。所以，MCMC 算法并非完美，不能说它已成为 IRT 领域的主流方法。最后，要强调的是本研究是对被试潜在特质参数的估计，在这一阶段没有必要使用 MCMC 算法，也使用不了 MCMC 算法。

审稿人 3 意见：
意见 1. 本文没有将 WMLE 方法与常用的后验最大估计（MAPE）进行比较，原因是什么？算法上应该没有困难。作者在展望中也提到，WMLE 与 MAPE 可能有一定的关系，之后可以进一步研究。请作者给出解释。
回应：近年来，一些学者重点关注了 WMLE 与 MAPE 的关系研究，并得到了一些非常有价值的发现。可见，WMLE 与 MAPE 的关系是密切而复杂的，已有研究都是在统计理论上进行论述的，仅通过模拟研究难以得到一般性的结论。先验选择不同，MAPE 就不同，而模拟只能考虑一种或几种先验的 MAPE，得到的结论不具有一般性。本研究加入 MAPE，难以把问题论述透彻，如果过分详细讨论 WMLE 与 MAPE 的关系，又对文章整体行文有很大影响，会导致文章重点不够突出。此外，我们目前也正在进行相关研究，并完成了一些重要工

作，发现了一些有价值的结论。综合考虑之后，本研究没有加入 MAPE。我们在修改稿中对相关研究进行了综述，论述了这一问题的研究意义和必要性，并给出了本研究没有加入 MAPE 的理由。

意见 2. 图 1 中的第 7 张图似乎有点问题，它与其他 8 张图规律不同。
回答：感谢您发现这一重要的图表问题。因对其进行了修正，请见正文。

意见 3. 推导公式的过程中，有一些小错误，并且符号也有前后不一致的情况，比如，项目参数的下标有时错写成 *i*，需仔细改正。
回答：感谢您的建议，我们已对全文进行了认真的检查和校对，针对公式中的符号和下标等不妥之处进行了相应的修改。

第二轮

审稿人 1
意见 1. 第 11-12 页这段表述“通过计模拟比较得到的结论是，与其它常用的初值选取方法（例如，被试的测验得分与失分比的自然对数，被试测验总分的标准分数）相比，对于 4PL 模型 WMLE 的求解，使用间隔取点确定初值的 N-P 迭代表现出更加优良的收敛效果。由于该模拟不是本文的核心内容，故对次模拟结果不做详细汇报。”中标黄的两个字应删除。
回应：已进行修正，请见正文。

意见 2. 第 15-16 页，“关于 $\mathrm{EAPE}(\theta_i)$ 的计算，本研究采用如下数值积分计算方法：首先，把区间 $[-4,4]$ 分成 K 等分；然后，以每个小区间的中点作为求积节点，记为 $o_1, o_2, \ldots, o_K$，以 $\Lambda(o_k) = f(o_k) \times \frac{8}{K}, k = 1, \ldots, K$，作为求积系数；最后，$\mathrm{EAPE}(\theta_i)$ 可近似计算为，

$$\mathrm{EAPE}(\theta_i) \cong \frac{\sum_{k=1}^{K}\left[o_k l(\mathrm{x}_i \mid o_k)\Lambda(o_k)\right]}{\sum_{k=1}^{K}\left[l(\mathrm{x}_i \mid o_k)\Lambda(o_k)\right]}.$$

”的涂黄色的公式中，函数 $f(o_k)$ 中的函数应为标准正态分布的密度函数，根据上下文，应该用符号 $\phi(o_k)$，请作者检查一下。
回应：感谢专家的修改建议。已进行修正，请见正文。

意见 3. 第 22 页第 1-3 行 “不过，Magis（2012）的研究表明，3PL 模型下 WMLE 与 JMAPE 不存在等价关系，WMLE 要比 JMAPE 稍大一些，但 3PL 模型的 MAPE 是否与其它先验下的 MAPE 存在等价关系并未提及。” 中，涂黄的 MAPE 应修改为 WMLE。
回应：已进行修正，请见正文。

意见 4. 另外，这里的“WMLE 要比 JMAPE 稍大一些”的含义是指什么？是 WMLE 的 ABME 稍大一些还是 WMLE 的 RMSE 稍大些？请作者表述再明确些。
回应：该句“WMLE 要比 JMAPE 稍大一些”所在的段落是在论述 WMLE 与 JMAPE 之间估计值的大小关系（等价或是有差异），而不是它们的偏度或返真性等性质之间的差别。所以，此处要表达的是：同水平潜在特质的 WMLE 估计值稍大于 JMAPE 估计值，而不是两种估计 ME，ABME 或是 RMSE 的大小比较。您的问题让我们意识到此处表达含义不清，容易让读者混淆。已对此进行了修正，请见文章正文。