《心理科学进展》审稿意见与作者回应

题目：认知建模中模型比较的方法
作者：郭鸣谦，潘晚坷，胡传鹏

编委初审

意见 1：综述内容基本上是已知的内容，作者自身方面的贡献太少，也太单薄，深度和广度方面需要加强；
回应：感谢专家的建议，本次修改中已经按照专家的建议进行了相应地修改。

意见 2：数据为模拟数据，且只有 10 例数据，模拟量似乎偏少。此外，建议作者可以考虑增加实际数据作为参考；
回应：感谢专家的建议，修改后的版本使用真实数据进行后续的分析。具体而言，我们使用了 Raab & Hartley (2020)的公开数据，包含 61 名被试的真实数据。具体数据地址为：https://osf.io/4h6ne/。

意见 3：作者提出了模型比较的新进展，也提供了新的可能方法，如贝叶斯模型的平均，建议作者利用模拟数据可以对此方法和原有方法进行直接比较，对读者而言也许更有价值。
回应：感谢专家的建议。已经增加了此方法与经典方法的直接比较。见第 585 行～ 第 609 行新增的内容。

第一轮

审稿人 1 意见：

意见 1：在本文的写作方面，建议作者重新考虑调整文章的结构，可以先介绍各种模型比较指标的原理、适用情况和优缺点，然后把示例数据分析部分单独出来陈述。而不是像现在穿插在介绍不同方法的过程中。另外，建议单独出来的示例数据分析部分，更加详细地陈述在该实验背景下，各种模型比较指标是如何计算，并适当地提供示例的分析程序，供感兴趣的读者参考和使用。
回应：感谢审稿人提出的宝贵意见，上一版手稿将示例放在介绍各个指标的做法令文章结构略显散乱。我们现将示例数据部分单独放置出来作为文章的第五节，并对各个指标的使用有了更清晰的介绍。

意见 2：第 101 行，作者提到“每个条件下的正负反馈都是概率的”，这样的表述比较像英文直译，建议作者注意行文的流畅性和可读性，并通读全文，修改类似的表述。
回应：感谢您的细致阅读和对行文流畅性的建议，我们再次通读全文并对相关内容进行修改。针对本条意见，我们重新“每个条件下的正负反馈都是概率的”这一句，修改后的内容如下（见第 505-511 行）：该范式是 2×2 的被试内实验设计，其中第一个变量是刺激反应动作：Go 和 No Go；第二个变量是行为反应后的反馈类型：获得奖励和避免惩罚。刺激反应动作

和反馈类型两个条件结合起来共形成四种实验条件：Go-获得奖赏、Go-避免惩罚、No Go-获得奖赏和 No Go-避免惩罚。每种条件下的反馈均非 100%确定性的事件，在“Go-避免惩罚”条件下，正确反应（即 Go）有 80%的概率避免惩罚，但有 20%的概率无法避免；而错误反应（即 No-Go）则有 80%的概率受到惩罚，20%的概率避免惩罚。试次开始第一屏的图片在该范式中被称作提示符号 cue，共有四种，与实验条件一一对应。

意见 3：文中图 1 的注释没有显示完全，另外建议调整图 1 的位置，避免把图插入在一个段落中间的情况。
回应：感谢您的细致审阅和意见。我们将上一版手稿的图 1 修改为了图 2，并调整了位置，并补全了的注释。

意见 4：文中的图 1-图 3 在正文中都没有被引用，建议作者补充。
回应：感谢您的细致审阅。我们在正文里修改并引用了这三张图。上班手稿的图 1(现为图 2)，上一版手稿中的图 2 和图 3 则是此版的图 4 和图 5。它们具体的位置在正文第 59 行，第 503 行和第 523 行。

意见 5：第 121 行，作者陈述贝叶斯参数估计的优点时，请补充相关参考文献。例如，“贝叶斯参数估计的先验为有信息且合适的先验时，能降低模型过拟合的程度”的出处。同时，如果是使用正则化先验，那么它是属于无信息先验？但这种先验理论上也能降低过拟合的程度。
回应：感谢您的细致审阅。正则化方法如 L1 正则化与贝叶斯参数估计使用拉普拉斯先验等价，而 L2 正则与正态分布先验等价。这二者都是弱信息先验。不仅仅有信息先验均能降低模型的过拟合，上述提到的弱信息先验也能降低模型过拟合。因此我们在修改了这句话，并引用了 Bishop, C. M. (2006). Pattern recognitionand machine learning. springer.作为参考文献。修改之后这句话的位置在正文第 74 行。具体修改内容如下：贝叶斯参数估计里的先验分布能起到正则化的作用，从而减少模型的过拟合 (Bishop, 2006)。

意见 6：第 124 行，“使得模型拟合的结果更少出现极端值”这个陈述具体应该如何理解
回应：感谢您的细致审阅和意见。层级贝叶斯估计较少是有参数的极端值结果是因为组水平先验的引用。我们修改了该陈述，具体修改内容为：贝叶斯估计十分利于构建层级贝叶斯模型(Hieraricial Bayesian Model)引入了组水平(Group level)先验，不同被试的参数均是从组水平参数所形成的分布中抽取的，而组水平参数的估计本身也受到单个被试参数的约束，因此，单个被试的参数值会通过组水平的参数间接受到其他被试数据的影响，向组水平参数均值方向偏移，从而减少了被试中极端数据对其参数值的影响 (Ahn et al., 2017; Gelman, Carlin, et al., 2013)。（见第 74-80 行）

意见 7：第 140 行，Mean squared error 一般翻译成均方误差，另外，建议作者通读全文，保持术语名词前后的一致。
回应：感谢您的细致审阅和宝贵建议。我们将“Mean squared error”统一翻译为“均方误差”，并统一文中出现的“均方误差”这一术语，确保其在全文中的表述前后一致。

意见 8：表 1 和表 2 在文中没有被引用，请作者补充。
回应：感谢您的细致审阅和宝贵建议。我们在正文的第 90 行和第 340 行分别引用了表 1 和表 2。

意见 9：第 154 行，作者对嵌套模型的表述不够准确。
回应：经过进一步查阅相关资料，我们完善了嵌套模型的定义，修改后的表述如下：嵌套模型指的是一个模型相对于另一个模型具有更少的参数或者某些参数被限制（例如固定为特定值）。在嵌套模型中，一个模型（简单模型）是另一个模型（更为完整模型）的子集，它在更完整模型的基础上降低了复杂性。（见第 101 行到 103 行）

意见 10：第 167 行，“RSS (Residual sum of squares)与为残差平方和”表达不通顺。
回应：感谢您的指正。我们已经修改为： RSS(Residual sum of squares)为残差平方和（见 115 行）。

意见 11：第 178 行，空模型应该如何理解？针对文章中的实例，应该如何表述？
回应：感谢您的宝贵意见。我们增加了对于空模型的解释。增加的内容位于正文的第 115-117 行。增加的内容如下：空模型(null model)认为实验刺激对观测数据没有任何的影响, 观测数据是均匀分布的。此处空模型指的是参数为(1/选项数量)的二项式分布或者多项式分布模型。例如，在本文的案例里，可能的选项有两个，因此二项式分布的参数为 12，即观察到两个选项的可能性相同，而空模型的似然函数为试次数量乘上 log 0.5。

意见 12：第 186 行，公式（8）的表述不准确。
回应：感谢审稿人的宝贵意见。我们修改了对数似然函数的描述，修改内容位于正文 132-133 行，具体修改后的内容为:似然函数是在给定观测数据的情况下，各模型参数产生该观测数据的概率。对似然函数求对数即得到对数似然函数，可以用来评估模型参数与实际数据拟合度。

意见 13：第 187 行-189 行，对数似然函数通常是某种分布的这种表述不准确。
回应：感谢您的指正。我们将修改原文中的表述，明确指出对数似然函数是基于数据分布的对数概率密度函数，而不是直接“是”某种分布。我们还将检查全文，确保在讨论对数似然函数时，术语的使用是一致和准确的。修改后的位置在正文第 123-125 行，内容如下：不同任务的数据分布不同，因此对数似然函数的形式也有所区别。对于选项数据，对数似然函数通常基于伯努利分布或多项式分布来构建；而对于反应时或肌电等连续数据，对数似然函数则一般基于高斯分布来构建。

意见 14：第 220 行，“Precision-recall curve”一般为精确率-召回率曲线？建议作者通读全文，对涉及的英文术语翻译进一步斟酌。
回应：感谢您的细致审阅和宝贵建议。我们将全文的相关术语进行修改，以确保术语的一致性和准确性。修改的位置包括：正文 167 行。

意见 15：第 264 行，n 代表什么？
回应： 感谢您的细致审阅。这里的 n 代表了试次的数量。我们在正文第 231 行增加了如下陈述：其中 n 表示为试次的数量。

意见 16：第 299 行-300 行，公式（13）和（14）需要做进一步解释。
回应： 感谢您的宝贵意见。对于赤池权重的计算，我们在正文 245-247 行增加了如下解释：上述两个公式中第一个代表了各模型与最优模型之间的差异，最优差异则会通过公式(14)映

射到 0-1 区间之中，代表不同模型的权重。公式(14)被称作 softmax 公式，公式中Δ AIC 乘上 -0.5 则是为了保证 AIC 更小的模型占据的权重更高。

意见 17：第 366 行，i 应该为第 i 个样本数据点，建议类似的表述要完整。
回应： 感谢您的细致审阅和宝贵建议。我们在正文第 297 行将这部分表述改为了：其中，i 是第 i 个样本数据点，S 是 MCMC 采样的后验分布的样本的数量。

意见 18：第 388 行，“因为交叉验证类的指标更容易确认复杂模型的为最优模型，这使得它们在心理学研究的应用格外的广泛。”该陈述应该如何理解？
回应： 感谢您的细致审阅和宝贵建议。这句陈述是笔误。我们在正文中删去了这句话。

意见 19：第 402 行，如何对不同的模型进行 Wald 检验？
回应： 感谢您提出的问题。我们在正文 542-550 行增加了 Wald 检验的介绍，增加的内容如下：

Devine et al. (2023)建议使用 Vehtari et al. (2017)里采用的方法，对基于贝叶斯模型的指标，例如 DIC，WAIC 和 PSIS-Loo-CV，研究者可以对不同模型进行 Wald 检验。Wald 检验具体流程是分别计算模型指标差异的均值和标准误，如果均值大于 1.96 个标准误时，就判断为模型之间的差异显著。 根据 Vehtari et al. (2017)，单个模型比较指标的标准误计算公式为：

$$se(elpd) = \sqrt{N\sum_{i=1}^{N}\left(elpd_n - \overline{elpd}\right)^2} \quad (36)$$

其中i是样本数据点，N在心理学实验里即为所有被试的所有试次，$\overline{elpd} = \frac{\sum_{i=1}^{N} elpd_i}{N}$是指标的均值。计算两个模型的模型比较指标之差的标准误的公式为：

$$se(elpd_A - elpd_B) = \sqrt{N(\sum_{i=1}^{N}\left(elpd_{A_i} - \overline{elpd_A}\right)^2 - \sum_{i=1}^{N}\left(elpd_{B_i} - \overline{elpd_B}\right)^2} \quad (37)$$

Wald 检验将模型指标的不确定性考虑在内，其假阳性的概率更低。

意见 20：第 406 行开始，建议补充解释各符号的含义。

回应： 感谢您提出的问题。我们在正文 349-350 行增加了如下解释：公式的左侧$p(\theta|y)$为参数的后验分布，右侧的第一项$p(\theta)$是参数的先验分布，而第二项$p(y|\theta)$则是似然函数。

意见 21：第 430 行，“先验分布对参数估计的结果不恰当的先验分布会对边际似然的计算结果产生很大的影响 (Boehm et al., 2018)”该表述不太通顺。
回应： 感谢您提出的问题。我们将对相关内容进行修改，以提高文章的清晰度和可读性。

我们也检查全文其他部分，确保在讨论先验分布对边际似然影响的表述都是一致和准确的。修改后的内容位于正文 366-368 行，具体内容如下：先验分布对边际似然的计算结果具有重要影响。不恰当的先验分布，尤其是在数据点较多的情况下，可能会对参数估计的结果产生显著影响，进而对边际似然的计算结果产生很大的影响（Boehm et al., 2018）。

意见 22：第 485 行，phi 这个符号代表什么？

回应： 感谢您的细致审阅和宝贵建议。这里我们因为疏忽而没有提及ϕ的含义。ϕ是核密度估计里的带宽，我们在正文 418-420 行将这段陈述改为了：核密度估计方法使用了非参统计方法中的核密度估计计算参数的后验概率$p(\hat{\theta}|y) = k(\hat{\theta}|\theta, \phi)$。其中，$k$为密度核函数，通常为高斯分布(Wasserman, 2006)，而ϕ是密度核的带宽(Band width)。

意见 23：第 488 行-489 行，这些表述是否放错了位置？

回应： 感谢您的细致审阅和宝贵建议。由于我们在上一版上传的手稿并未添加正文行数，所以非常抱歉的是，在本次修改中未能定位出 488 和 489 这两行可能放错位置的句子。我们对全文进行了通读，避免表述位置不当的情况。如果仍然存在这个 问题，请您不吝继续赐教。

意见 24：模型选择指标的使用建议的部分，是否可以考虑使用流程图说明如何选择合适的模型比较指标？

回应： 感谢您的宝贵意见。我们尝试使用流程图归纳总结选择模型比较指标的流程。但是在过程中发现用流程图总结归纳较为困难。其中的原由主要是模型比较的指标中，很难说哪一个是绝对最优的，多数情况之下，需要根据问题与其他条件（如模型拟合的方法）进行综合选择，按这种方式绘出的流程图与上班手稿图 1(现今图 2)基本类似。此外，增加过于明确的模型指标选择标准可能会造成刻板印象误导，使读者误认为只有符合这些特定标准的模型才是有效的或者最佳的。这种做法可能会忽视了模型的灵活性和多样性，因此我们未呈现流程图。

意见 25：第 707 行，Random effect Bayesian model selection, RE-BMS 的翻译不准确。

回应： 感谢您的细致审阅和宝贵建议。我们将 Random effect Bayesian model selection 的翻译改为了：随机效应的贝叶斯模型比较。并且我们将对第 707 行及全文的相关术语进行修改，以确保术语的一致性和准确性。

审稿人 2 意见：

本文整体逻辑通顺，行文流畅，系统总结了当前认知心理学中计算建模的各种模型比较的方法和特点。我对整体的行为逻辑并没有什么意见。但是文中有一些写作和表达上的细节，希望和作者商榷，进一步提高本文的精确度。

意见 1：72 行提到了“误差项”，但是后面没有对这一项进行介绍，只介绍了方差和偏差。不是很明确“误差项”在这里的意义。

回应：感谢您的建议。关于您指出的第 72 行中“误差项”未详细介绍的问题，在认知建模的语境下，误差项通常指的是观测数据与模型预测之间的差异。在统计建模中，误差项代表模型未能解释的变异部分，它可以是随机误差，也可以是系统误差。在本文中，当我们提到方差和偏差时，实际上是在描述模型预测的不确定性（方差）和模型预测与真实值之间的一致性偏差（偏差）。这两者都是误差项的组成部分，但上一版本中我们未明确地阐述这一点。为了增强文章的清晰度和完整性，我们进行了如下修改：在介绍方差和偏差的部分之前，增加对误差项的定义和解释，明确指出误差项是模型预测值与实际观测值之间的差异。我们将说明方差和偏差是如何从误差项中衍生出来的，并强调这两者对于理解模型性能的重要性。我们在正文第 33-39 行添加了如下内容：泛化误差可以被分为方差(Variance)、偏差(Bias)和误差项(Irreducible error)。偏差衡量的是模型预测的期望值与真实数据之间的偏差。一个高偏差的模型通常意味着模型过于简单，无法捕捉到数据中的复杂关系，从而导致欠拟合。方差衡量的是模型在不同训练数据集上的预测结果的变异程度。一个高方差的模型通常意味着模型过于复杂，对训练数据中的随机噪声也进行了学习，从而导致过拟合。误差项是指数据本身所包含的不可减少的噪声和不确定性。这部分误差是由于数据本身的复杂性或者是测量过程中的误差造成的，任何模型都无法预测或消除这部分误差。

意见 2：119 行，文中提到“拟合认知模型的方法主要有点估计的极大似然法(Maximum likelihood estimation, MLE)，最大化后验概率法(Maximum a posterior estimation, MAP)和贝叶斯参数估计(Bayesian estimation)”这样的表达读起来贝叶斯参数估计也属于点估计。建议改成“拟合认知模型的方法主要有点估计的极大似然法(Maximum likelihood estimation, MLE)和最大化后验概率法(Maximum a posterior estimation, MAP)，以及不基于点估计而是估计整个后验分布的贝叶斯参数估计(Bayesian estimation)”
回应：感谢您的建议。我们在正文 69-72 行将拟合方法的表述改为了您建议的陈述。

意见 3：120-125 行两次提到了贝叶斯参数更适合构建分层模型，逻辑比较混乱，建议仔细理顺逻辑
回应：感谢您的建议。我们对相关内容进行重新组织以确保逻辑更加清晰。修改后的内容位于正文第 72-80 行，具体为：首先，贝叶斯估计能够提供参数的后验分布，这不仅便于进行后续分析，而且对于构建分层模型特别有利。贝叶斯参数估计里的先验分布能起到正则化的作用，从而减少模型的 (Bishop, 2006)。此外，贝叶斯方法在处理多个被试数据时表现出其独特优势。贝叶斯估计十分利于构建层级贝叶斯模型(Hieraricial Bayesian Model)引入了组水平(Group level)先验，不同被试的参数均是从组水平参数所形成的分布中抽取的，而组水平参数的估计本身也受到单个被试参数的约束，因此，单个被试的参数值会通过组水平的参数间接受到其他被试数据的影响，向组水平参数均值方向偏移，从而减少了被试中极端数据对其参数值的影响 (Ahn et al., 2017; Gelman, Carlin, et al., 2013)。

意见 4：113 行，笔误“后延概率”
回应：感谢您的细致阅读。我们进行了修正，并检查了正文类似的地方。

意见 5：表 1，这里把平均平方误差(MSE), 决定系数(r2)和对数似然函数列成三个不同的方法。事实上，MSE 和 r2 是一种当 error distribution 为高斯函数的特殊对数似然函数的特例。所以其实第三个包括了前两个。我理解从使用角度来说这三者可能不太一样。但是建议写清楚这一点，更加的严谨。
回应：感谢您指出这个重要的问题。我们进行了以下修改：在表格的注释中明确指出，均方

误差（MSE）和决定系数（R^2）实际上是对数似然函数在高斯误差分布假设下的特例。
为了保持表格的简洁性，我们将保留这三个指标作为独立的列，但会在对数似然函数的“优点”一栏中补充说明，它包含了 MSE 和 R^2作为特殊情况。我们将检查全文，确保在其他地方提到这些指标时，也能够准确反映它们之间的关系。以下是修改后的内容（见正文第 92 页）。

表 1. 各拟合度指标的优缺点以及适用的参数估计范围

	适用的参数估计方法	优点	缺点
均方误差 (MSE)	极大似然法、最小二乘法	直观简单，易于计算和解释	不适用于分类问题，未考虑模型复杂度对过拟合的影响
决定系数（r^2）	极大似然法、最小二乘法	衡量模型变量变异性占比，提供模型拟合的可解释性	对模型的复杂性敏感，无法比较特征数目不同的模型
对数似然函数	极大似然法，最大后验概率法，贝叶斯参数估计	反映模型预测与实际数据的匹配程度，可用于模型比较和参数估计；MSE 和r^2是残差为正态分布时对数似然函数的特例	不适用于非概率、非参数模型；对异常值敏感
ROC 曲线	极大似然法，最大后验概率法，贝叶斯参数估计	用于评估模型对实际数据的预测能力。	不适用于数据为多选项的情况；对于不平衡数据，结果不够准确
后验预测检查	贝叶斯参数估计	考虑参数不确定性和模型复杂性；可检查对新数据样本的预测能力	需要领域专业知识对先验和后验分布进行假设；计算复杂度较高

意见 6：其中 279 行的 AIC 和 450 行的 BIC，都提到“logL(theta_hat | y)是使用极大似然法估计或者最大化后验概率估计求得最优参数 theta_hat 的对数似然函数值”。我个人的理解，这里的 theta_hat 只能由极大似然估计法求得，不能由最大化后验概率估计求得。因为如果是后者，理论上还应该包括 prior 的惩罚项。希望作者反复确认这一点，如果认为这是对的，麻烦给出参考文献。同样，表 2 也提到了这一点。
回应：感谢您的宝贵意见。我们重新查阅了引文 Pattern recognition and machine learning, 2006, Springer。根据该书的第 217 页的公式（4.139），最大化后验概率法得到的$\hat{\theta}$是可以用于计算 AIC 和 BIC 的。最大化后验概率的优化目标是$log\ L(\theta|y)\ +\ log\ L(\theta)$，在找到最大化这个目标函数的 theta 值，即 theta_hat 后，我们只保留目标函数的第一项$log\ L(\hat{\theta}|y)$ 用于计算 AIC 和 BIC。

意见 7：公式 15-18，DIC 中的 theta_杠是 theta 的均值么？貌似没有定义，建议定义清楚。
回应： 感谢您的宝贵意见。我们对公式 15-18 及相关内容进行以下修改：
在公式中明确补充 DIC 中的 θ_杠（theta_杠）的定义为参数 θ 后验分布的均值。也增加对 DIC 公式的各部分更加详细和准确的描述，包括偏差的后验均值和有效参数的计算方法。解释了这些部分如何反映模型的拟合度和复杂度，并说明它们在模型比较中的作用。修改后的

内容位于正文第 266-271 行，具体内容如下：DIC 的计算公式为$DIC = -2D(\bar{\theta}) + 2 \times p_D$。

其中$\bar{\theta}$为参数后验分布的均值，而$D(\theta)$则是真实数据与模型预测分布之间的偏差（Deviance），用以衡量模型的性能。偏差的公式为：

$$D(\theta_s) = \log\ L(y|\theta_s) \tag{15}$$

其中s代表了 MCMC 的样本，因此θ_s是 MCMC 样本的参数值。DIC 的公式的第一项是-2 乘上参数后验分布上的均值的偏差，代表了模型拟合的程度，第二项p_D被称作为有效参数(effective number of parameters)，是模型拟合的复杂度的惩罚项，计算公式为：

$$p_D = \bar{D}(\theta) - D(\bar{\theta}) \tag{16}$$

$$\bar{D}(\theta) = -2 \times \left(\frac{1}{S}\sum_{s=1}^{S} \log L(y|\theta_s)\right) \tag{17}$$

意见 8：480 行，海森矩阵建议加上英文翻译(hessian matrix)帮助理解。
回应： 感谢您的建议。我们增加了相应的英文翻译，见 414 行。

意见 9：316 行，其中提到”DIC 用模型分布于真实模型分布的偏差”。这里模型分布和真实模型分布貌似定义不清，希望检查一下措辞表达更加准确。
回应： 感谢您的建议。这句话出于笔误，我们在正文里删去了这句话。

意见 10：126-128 行，文中提到，“一是以马尔科夫链蒙特卡洛采样(Markov Chain Monte Carlo, MCMC)为主的采样方法，另一种则是如变分推断(Variational inference, VI)的近似方法”。准确的讲，MCMC 也是一种近似方法，MCMC 和 VI 都是近似贝叶斯推断。所以建议把措辞改成“一是以马尔科夫链蒙特卡洛采样(Markov Chain Monte Carlo, MCMC)为主的采样近似方法，另一种则是如变分推断(Variational Inference, VI)通过近似后验分布求解的近似方法”。
回应： 感谢您指出我们在表述上的不准确之处。MCMC 和 VI 确实都是近似贝叶斯推断的方法，MCMC 更侧重于通过采样来近似后验分布。我们调整了介绍 MCMC 和变分推断的位置。修改后的内容位于正文第 259-263 行，具体内容如下：贝叶斯参数的近似估计通常有两种实现途径，一是以马尔科夫链蒙特卡洛采样(Markov Chain Monte Carlo, MCMC)为主的采样近似方法，另一种则是如变分推断(Variational Inference, VI)通过近似后验分布求解的近似方法。采样近似方法的计算量更大、速度更慢，但通常得到的结果也更为准确。

第二轮

审稿人 1 意见：作者基于审稿意见进行了仔细的修改，阐述较为清晰，使得文章质量有了较大的提升。但文中还有一个细节问题请作者澄清：

P24，公式(37)，在计算两个模型的模型比较指标之差的标准误时，应该是先计算每个数据点上模型比较指标之差，然后再计算 N 个差异的标准误？

回应： 感谢您的对本文公式的细致检查，确实，原公式出现了一些笔误， 本次修改了公式并增加了解释。修改后内容在正文第 545 行：同理，当计算两个模型比较指标之差的标准误时，先计算每个数据点上模型比较指标之差，然后再计算 N 个差异的标准误，其公式为：

$$se(elpd_A - elpd_B) = \sqrt{\frac{N}{N-1}\left(\sum_{i=1}^{N}\left((elpd_{A_i} - elpd_{B_i}) - \overline{(elpd_A - elpd_B)}\right)^2\right)} \qquad (37)$$

其中$\overline{(elpd_A - elpd_B)}$是两个模型比较指标之差的均值。Wald 检验将模型指标的不确定性考虑在内，其假阳性的概率更低。

编委 1 复审意见：同意发表。

编委 2 复审意见：同意发表。

主编意见：本文经过多位专家的审稿，作者进行了认真的修改，达到发表水平，同意发表。