《心理科学进展》审稿意见与作者回应

题目：测量不变性检验方法的新进展：惩罚对齐法
作者：温聪聪

第一轮

审稿人 1 意见：

本文回顾了几种常用的测量不变性检验方法的原理、优势与不足，在此基础上引入了 Asparouhov 和 Muthén 在 2023 年提出的惩罚对齐法并详细介绍了该方法的优势，最后运用大学生职业价值观研究实例演示如何使用惩罚对齐法进行测量不变性检验和多组比较。为应用研究者进行测量不变性检验和多组比较提供了新的思路，对于国内心理学研究者了解该方法有很好意义，但文中还有一些问题需要改进。

意见 1：文中有部分参考文献与测量不变性的主题无关，建议作者通读全文进行精简。例如，“测量不变性检验方法进展回顾”部分，潘俊豪等人(2016)和潘俊豪等人(2017)。
回应：非常感谢您的建议。我对参考文献再次进行了通读与精简，删去了 MacCallum et al., 1996，Asparouhov & Muthén, 2009，潘俊豪等人(2016)，潘俊豪等人(2017)等参考文献。

意见 2：作者在介绍贝叶斯结构方程模型时，对信息先验和无信息先验的定义不够准确。
回应：非常感谢您的指正。我对贝叶斯结构方程模型综述部分进行了较大修改，也更正了信息先验和无信息先验的定义。

意见 3：作者在介绍惩罚对齐法进行测量不变性检验的原理时，介绍完公式（3）后突然出现公式（4）和（5）比较突兀，后面的参数求解过程应该是基于公式（3）才对？同样，公式（6）中的定义为何与 θ 无关？
回应：非常感谢您的细心指正。原本这样列出是想让求解过程介绍变得简单些，结果两位审稿人都发现了问题，这样介绍没有达到预期效果。我已经将这两个公式改为基于公式（3）的计算。公式（6）确实是一处错误，公式中定义的是$P(\theta)$，应为$P(\theta) = \sqrt[4]{\theta^2 + \varepsilon} \approx \sqrt{|\theta|}$，我已经在全文进行了修改。

意见 4：在 3.2.3 节，作者提到“该方法可以对模型参数设定满足正态分布的先验分布，有效估计模型中的交叉载荷、不等参数、残差相关”，“贝叶斯结构方程模型对模型参数设置的是正态先验分布”，针对这些参数的先验不一定都是正态分布；作者也提到“贝叶斯结构方程模型是基于贝叶斯估计法，而惩罚对齐法是基于极大似然估计法。基于极大似然估计法的模型拟合指标更多、更全面、更成熟，能更好地判断模型拟合数据的效果”，该结论是否适用于新提出的贝叶斯模型拟合指标？如 BCFI，BTLI 和 BRMSEA 等？另外，作者对敏感性分析的定义不太准确。
回应：非常感谢您的专业建议和指正。本人再次认真阅读了 Asparouhov & Muthén, 2023b 文献中关于 BSEM 设定正态先验分布的背景和文献中关于贝叶斯先验分布的相关文献，发现本文中的表述确实有误。BSEM 在进行测量不变性检验的时候检验的是模型的近似测量不变

性，其隐含的假设是模型中的参数跨组近似相等，偏离模型主结构的参数较少且较小，这种情况下适合采用正态先验分布。但在其他非测量不变性检验情况下，参数的先验分布可以是非正态的。

在您提出的关于贝叶斯模型拟合指标问题的指引下，我认真阅读了 Asparouhov & Muthén, 2021 关于贝叶斯模型拟合评价指标进展的文献，并且向 Mplus 团队询问了 Asparouhov & Muthén, 2023b 文献中“PML has the advantages of the ML framework which is more complete in terms of model fit evaluation”的含义，Mplus 团队回复“more complete in terms of model fit evaluation 不是指 CFI 和 TLI，而是进行测量不变性检验比较嵌套模型时极大似然估计法框架中可以使用似然比检验。”我已经在文中对相应内容进行了更正。

至于您提出的敏感性分析的定义，经认真阅读宋琼雅等, 2021 和 Muthén & Asparouhov, 2012 文献，发现确实不准确。我已经在文中对相应内容进行了更正。

意见 5：研究实例部分，注意到四类高校的学生数量差异较大，这些差异是否会影响测量不变性检验的结果？如果是，遇到这种情况应该如何处理？

回应：非常感谢您提出的专业建议。对于任意一整套的因子均值 α_g 和因子方差 ψ_g，总有对应的截距 τ_{pg} 和因子载荷 λ_{pg}，使得所得模型和形态不变模型 *M0* 具有相同的似然值。由于形态不变模型的 $\lambda_{pg,0}$ 和 $\tau_{pg,0}$ 为已知，λ_{pg} 和 τ_{pg} 可以运算得出。如此不断重复、迭代，直到以下总损失函数(total loss function)*F* 被最小化，模型中各组间的测量不等性(measurement noninvariance)也被最小化，此时就找到了最优近似测量不变模型。总损失函数 *F* 如下式所示：

$$F = \sum_{p} \sum_{g1<g2} W_{g1,g2} f(\lambda_{pg1} - \lambda_{pg2}) + \sum_{p} \sum_{g1<g2} W_{g1,g2} f(\tau_{pg1} - \tau_{pg2})$$

总损失函数 *F* 中的 $W_{g1,g2}$ 是用来反映组样本量大小和特定组参数估计的确定性，较大的组样本量比较小的组样本量对总损失函数的贡献更大，如下式所示：

$$W_{g1,g2} = \sqrt{N_{g1} N_{g2}}$$

因此，您提出的疑问对本研究实例中的数据分析有很大启发意义。数据的总样本量为 90416，而 G3 普通高校学生的样本量为 63585，C9 高校学生的样本量只有 1415，这会使 G3 组的测量不变性程度决定模型的测量不变性结果。我认为遇到这种情况有两种解决办法，一是对 G3 随机重新抽样，减少 G3 的组样本量；二是显著增加 G4 和 G2 等较小组的样本量。在本例中，显然做法一比较可行。我在研究实例中相应削减了 G1、G2、G3 的组样本量，G4 的样本量保持不变，使总样本量变为 29162 进行了分析，分析效果很好。

意见 6：研究实例部分，建议作者可以结合 3.1 节原理部分进行介绍，如在该例子中，M1 代表什么？M1 代表什么？θ 又代表什么？让读者对方法的原理有更好的理解。

回应：非常感谢您的悉心指导。根据您的意见，我在文中加入了一个段落用来说明实例分析中对应的 M1、M0、θ 所代表的具体内容。

……

审稿人 2 意见：

测量不变性的检验是一直是测量领域关注的重要话题之一，对于实际应用中测量工具的应用都具有重要意义。本文关注主题契合对量化研究方法的实际需求，紧跟前沿进展，介绍了测量不变性检验的一种新方法——在惩罚结构方程模型框架下实现的惩罚对齐法，并综述

了测量不变性检验的几种较新的方法，说明了惩罚对齐法相比于多组探索性因子分析、对齐法、贝叶斯结构方程模型等已有方法所继承的优势与进一步克服的不足，之后以一个实例说明惩罚对齐法的实际应用。文稿有助于国内的方法与应用研究者及时了解测量不变性检验的前沿动态，并且提供的实例对方法应用具有参考意义。但个人阅读下来感觉文章还有少数不清晰或不甚准确的地方，希望作者对此进行修改或做出解释，使读者能更好地了解和把握惩罚对齐法。

意见 1：公式(6)应该有误，等号左侧的括号中的参数是 θ，右侧的参数是 x，请检查修正。

回应：非常感谢您的细心指正。公式（6）确实是一处错误，公式中定义的是$P(\theta)$，应为

$P(\theta) = \sqrt[4]{\theta^2 + \varepsilon} \approx \sqrt{|\theta|}$，我已经在全文进行了修改。

意见 2："2.4 二水平随机效应法"第一段中，对三种模型的表述不准确。首先，二水平模型中，第二（组间）水平的因子载荷本身不存在随机性，只有第一（组内）水平可以存在载荷和截距的随机性。并且，参照 Muthén & Asparouhov (2018)，只有第一种模型（因子载荷和截距都随机）是假定近似测量等价性的，即各组参数的参数可以变化，但是服从相同的分布。原文表述"第二种，设置随机截距，限定每个水平的因子载荷近似不变，但不同水平的因子载荷不等；第三种，完全限定因子载荷和截距参数的近似测量不变性，即同水平和不同水平的因子载荷都近似不变"不准确。

回应：非常感谢您的专业建议和细心指正。我又认真阅读了 Muthén & Asparouhov，2018，Kim et al., 2017, Jak et al., 2013 等人的研究，在文中对相应内容进行了修改。

意见 3：为了与公式(3)保持一致，公式(4)中 L(θ2，θ3)是否漏了下标 w，参见 Asparouhov & Muthen (2023b)的公式(5)。

回应：非常感谢您的细心指正。我在文中相应之处进行了修改。

意见 4：研究实例中，量表题目使用四点 Likert 计分，作者直接将其视为连续变量，并在文中表述"利他因子和自我实现因子个包含 3 个连续型观测指标"。数据应作为分类顺序变量处理更为合适，而且即使是 5~7 点计分的 Likert 题目也不应将其表述为"连续型"指标。

回应：非常感谢您的专业建议和指导。本研究实例将 4 点 Likert 计分量表看作连续型指标确实不妥。我已经遵照您的建议，将这些量表看作分类变量进行分析。由于每个题项 4 个类别需要 3 个 thresholds，会造成一些填答类别有很少的样本量(empty cell)，使本研究的模型无法识别。所以，我将 4 点计分量表转换为了 3 点计分等级多分类变量，1,2 分记为"1"分，3,4 分重新分别记为"2"分和"3"分。这样模型得到了较好的识别效果和分析效果。具体结果和 Mplus 语法示例详见文中标红之处。

意见 5：研究实例的惩罚对齐法 Mplus 语句中，model 部分定义了所有交叉载荷，而在分组模型中只定义了主载荷和 EFA 确定的交叉载荷，并在先验设置部分，只对这些主载荷与 3 个交叉载荷的组间差值设定了对齐先验。这样看起来惩罚对齐法是需要先通过 EFA 确定具有交叉载荷的因子结构。而 3.2.1 部分也提到"惩罚对齐法继承了多组探索性因子分析可以精确估计模型中交叉载荷和近似为零因子载荷的优势。"ESEM 无需人为设定交叉载荷结构，直接基于数据估计，并可以通过目标旋转法得到尽量简洁的因子结构。建议作者检查这部分语句是否有误。

回应：非常感谢您的细心指正。经过您的提示，我又认真学习了 PSEM 的相关内容，并且在 Mplus 官网找到了在惩罚对齐法中设置 ESEM 的语句示例，参考链接如下：

https://www.statmodel.com/download/Version%208.9%20and%208.10%20Addendum

我已经对研究实例的 Mplus 语法示例进行了相应修改，研究结果也在文中做出了修正并标红。

意见 6：研究实例的惩罚对齐法分析部分，作者对 Mplus 软件输出的各组结果进行了多次 Z 检验进行 4 个组的两两比较，是否考虑了两两比较时 I 类错误率的控制问题？

回应：非常感谢您的专业建议。如您所说，4 个组 2 个潜因子均值的两两比较次数为 $2\times C_4^2$=12 次，两两比较的次数越多，犯 I 类错误的概率越大。我查阅了一些 I 类错误率控制的做法，包括降低 *p* 值、降低样本量等做法。遵照您提出的第 4 条审稿意见和审稿人 1 提出的第 5 条审稿意见，我把研究实例中的 4 点 Likert 计分量表转换为了 3 点计分等级多分类变量，1,2 分记为“1”分，3,4 分重新分别记为“2”分和“3”分。为了降低样本量以控制两两比较时 I 类错误率，平衡各组对测量不变性的贡献，得出较科学的研究结果，我把 C9 院校学生的 1415 样本量保持不变，985 非 C9 院校、211 院校、普通高校的样本量做相应重新抽样，大幅减少了样本量。新的分析结果和 Mplus 口令已在文中标红。

意见 7：第 2 部分“测量不变性检验方法进展回顾”第 1 段第 8 行，“邓丽芳和 Yuan（Deng&Yuan,2016）”文内引用不规范。

回应：非常感谢您的细心指正。原本把 Deng 写成中文名是想刻意体现是中国研究者提出了该方法，现已改正。

意见 8：文章所介绍的新方法以及文内综述和比较的方法都是测量不变性检验方法，而现在的文章标题无法显示出该文与测量不变性检验的任何联系，读者无法从标题中获取关于惩罚对齐法的应用目的信息，建议修改文章标题。

回应：非常感谢您的悉心指导。我已经将文章标题改为“惩罚对齐法——整合多组探索性因子分析，对齐法和贝叶斯结构方程模型的测量不变性检验新方法”。

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。但文中还有一些细节问题需要注意。

意见 1：作者在研究背景部分提到“使用模型修正指数来修正模型的临界值需要研究者自己确定”，这句话中的临界值应该如何理解？

回应：感谢您的提示与启发。模型修正指数代表的是将模型中的一个限制固定的参数变为自由估计，所减少的 1 个自由度对应的卡方统计量减小值。当限制固定的参数的修正指数较大时，说明需要将其改设为自由估计，逐渐修正模型。但目前尚未有研究提出修正指数达到多大时需要将限制固定的参数改设为自由估计。文中“使用模型修正指数来修正模型的临界值需要研究者自己确定”想表达的是研究者依据多大的修正指数来修正模型，修正到什么程度，需要自己来决定。文中表述可能不太清晰，我进行了修改。

意见 2：在测量不变性检验方法进展回顾部分，“Deng 和 Yuan(Deng & Yuan, 2016)”应该为“Deng 和 Yuan(2016)”。其次，BSEM 应该先提供全称，再给出缩写。第三，作者提到“这些先验信息设定了参数差异后验分布的不确定性量的多少”的表述不太准确，先验分布中的方差决定的是先验信息的不确定性，并不是后验分布。第四，“先验分布可以依据数据分析的实际情况设置”中根据数据分析的实际情况应该怎么理解？第五，“在正态先验分布情况下 95%置信区间的 Z 分数边界约为±1.96SD”这句话的表述也不够清晰准确。

回应：感谢您的细心指正和专业指导。我已经将引用改为“Deng 和 Yuan(2016)”。

BSEM、AESEM 和 PSEM 遵照您的意见在小标题处都先写了全称，再写出缩写。

文中关于先验信息的表述确实不够准确，我又查阅相关资料，进一步理解，对文中的表述进行了修正。

感谢您的提示与启发。“先验分布可以依据数据分析的实际情况设置”是想表达研究者一般情况下先为这些参数差异设置 0 均值和 0.01 的先验方差，“数据分析的实际情况”在此处就是指参数差异作为模型额外设定的限定参数的估计结果。如果这些参数差异的估计值明显没有落在[-0.196, 0.196]的区间里，那么就需要改动这些限定参数的先验方差，再次运行新模型拟合数据。文中表述并不清晰，我已经做出修改，感谢您的专业指导！

审稿人 1 和审稿人 2 都对文中的这一表述提出了疑问，我已经将该表述改为“即通过设置 0 均值正态先验分布预测其估计值在以 0 为均值基线的[-1.96SD, 1.96SD]范围内变动(此处 SD 为先验标准差，在先验方差不为 1 的非标准正态先验分布情况下 95%置信区间的边界约为±1.96SD)”，并且进行了举例说明，让读者更加清楚这段表述的含义。

意见 3：在对齐法和对齐探索性结构方程模型部分，建议作者补充说明 g1 的含义。

回应：感谢您的建议。我已经在文中进行了补充说明。

意见 4：PSEM 处也应该先提供全称，再给出缩写。

回应：感谢您的建议。我已经在研究背景部分首次提到 PSEM 的地方标明了全称和缩写，并且标红。

意见 5：在实证研究中，惩罚函数的权重的不同，例如 0.1 或者 0.01，对结果会不会有什么显著的影响？

回应：感谢您的提示与专业指导。经查阅 Asparouhov & Muthén, 2023b，一般来说，惩罚函数的权重越小，模型分析就需要更低的收敛标准。我已经在文中做出相应说明。

意见 6：实证研究中，进行两两比较的时候，是否需要进行类似于事后比较的校正以控制一类错误的水平？另外，重新抽样后，普通高校的学生人数仍然为 10346，这是出于什么考虑？

回应：感谢两位审稿人提出了事后比较校正以控制一类错误率水平的问题和相关提示。我在本例中采用了 Bonferroni 校正方法得出新的显著性水平以控制一类错误率水平，得出了新的两两比较结果。我对研究实例的两两比较表格中的内容进行了修改。

总损失函数 F 中的 $W_{g1,g2}=\sqrt{N_{g1}N_{g2}}$是用来反映组样本量大小和特定组参数估计的确定性，较大的组样本量比较小的组样本量对总损失函数的贡献更大。原始样本数据的总样本量为 90416，其中 985 高校学生的样本量为 16167，211 高校学生的样本量为 9249，普通本科高校学生的样本量为 63585，C9 高校学生的样本量为 1415。根据$W_{g1,g2}=\sqrt{N_{g1}N_{g2}}$的公式

可知，在比较 G3 普通本科高校组和其他组参数的不等性时，G3 对$W_{g1,g3}$的贡献是 G1 的$\sqrt{\frac{N_{g3}}{N_{g1}}} \approx 2.0$倍，G3 对$W_{g2,g3}$的贡献是 G2 的$\sqrt{\frac{N_{g3}}{N_{g2}}} \approx 2.6$倍，G3 对$W_{g3,g4}$的贡献是 G4 的$\sqrt{\frac{N_{g3}}{N_{g4}}} \approx 6.7$倍。

根据您的提醒和点拨，为了防止对齐法测量不等性优化计算中各组样本量差距过大，较大组对参数差异不等性的贡献过高，本轮修改稿我对原始样本中普通本科高校组和 985 高校组各重新随机抽样随机选取 9249 个个案保留。这样一来，四个高校组的样本量比数变为 6.5:6.5:6.5:1。985 高校、211 高校、普通高校组的参数在进行测量不等性优化计算时，它们之间的样本量权重是相等的。而这三个组在测量不等性优化计算中样本量影响的权重是 C9 院校组的$\sqrt{\frac{N_{g3}'}{N_{g4}}} = \sqrt{\frac{9249}{1415}} \approx 2.6$倍。通过这样的重新随机抽样策略，平衡了各组样本量在测量不等性优化计算中的权重，明显缩小了各组样本量的差距。

……

审稿人 2 意见：

目前修改后的文稿中还存在一些个人觉得不清晰或有误的地方，希望作者对此进行修改或解释。

意见 1：第二章“测量不变性检验方法进展回顾”中第一段中着重强调国内的测量不变性检验方法进展。综述类文章应回顾的是整个领域中相关话题的进展，而且文中所介绍的方法也并不是专门针对国内研究，建议根据文章定位修改这一段的表述。

回应：感谢您的专业指导和建议。经过您的点拨，我认为在文中回顾国内外测量不变性检验领域的研究进展非常必要。我已经在文中加入相关综述并标红。

意见 2：“Deng 和 Yuan(Deng & Yuan, 2016)创造性地提出了投影法”应为“Deng 和 Yuan(2016)…”

回应：感谢您的细心指正。我已经将引用改为“Deng 和 Yuan(2016)”。

意见 3：“2.1 贝叶斯结构方程模型(BSEM)”的最后一段“使其估计值在以 0 为均值基线的[-1.96SD, 1.96SD]范围内变动(此处 SD 为标准差，在正态先验分布情况下 95%置信区间的 Z 分数边界约为±1.96SD)”在这个语境下表意模糊，容易对读者产生误导。标准正态分布中，中间 95%区域对应的临界值 Z 分数就是±1.96，而非±1.96SD，没有必要做这句解释。而且这里的 SD 应当指的是先验标准差，不注明的前提下反而容易让读者产生迷惑。

回应：感谢您的专业建议。审稿人 1 和审稿人 2 都对文中的这一表述提出了疑问，我已经将该表述改为“即通过设置 0 均值正态先验分布预测其估计值在以 0 为均值基线的[-1.96SD, 1.96SD]范围内变动(此处 SD 为先验标准差，在先验方差不为 1 的非标准正态先验分布情况下 95%置信区间的边界约为±1.96SD)”，并且进行了举例说明，让读者更加清楚这段表述的含义。

意见 4：“2.2 等效性检验(Equivalence test)”第一段中“将卡方变化值、对应的自由度变化值与置信水平α 相结合计算出对应 p 值，根据 p 值的显著性判断测量不变性程度更强的嵌套模型是否被拒绝”，表述有多处不严谨之处。首先，卡方值的 p 值根据自由度所确定的卡方分布就可以得出，与 α 水平无关。其次，是将 p 值与 α 水平进行比较后，获得该统计量的显著性，而非 p 值的显著性。第三，α 是显著性水平，而非置信水平（后文“4.2 多组验证性因子分析(Multiple-Group CFA)”中也出现了类似问题）。
回应：感谢您的细心指正。我已经在文中对相关表述进行了修改。

意见 5：公式(2)下面的段落中，“成分损失函数 $f(\theta) \approx \sqrt{|\theta|}$高估中等不等性惩罚，低估高不等性和低不等性惩罚的特点”表述有误，应做适当修改。事实上，当 θ <1 时, $f(\theta)>\theta$, 参数不等性是都被放大，没有低估，但是较大的组间差值比较小的差值对模型不等性的放大程度更小;而 θ >1 时, $f(\theta)< \theta$， 参数不等性此时才是被低估，而较大的不等参数比较小的不等参数使模型不等性衰减的程度更大。

回应：感谢您的细心指正。我已经在文中加入了$f(x) = \sqrt{|x|} - x$的函数图，用较为直观的方式呈现并解释成分损失函数$f(x) \approx \sqrt{|x|}$的特点。

意见 6：“2.4 二水平随机效应法(two-level random effects model)”第一段修改后的介绍中逻辑稍显混乱。Muthén & Asparouhov (2018)提出的前两种模型实际都是随机截距模型，在他们文章的实例分析中参照 Jak 等人(2013)推荐的三步拟合，第一步和第三步用到的就是这两个随机截距模型。建议作者调整这一段的说明，帮助读者有更清楚的理解。
回应：感谢您的专业指导和点拨。我再次认真阅读了 Muthén & Asparouhov (2018)中提出的随机效应模型，对涉及这些内容的段落进行了修改。

意见 7：“3.1 惩罚对齐法进行测量不变性检验的原理”中，作者目前的介绍基本是 PSEM 自身原理的介绍，而基本没有涉及到测量不变性的内容，并且，后续提及的一些方法细节和名词也没有提前交代清楚。例如，对齐法在检验测量不变性时是会在成分损失函数里构建参数差异，从而将小的参数差异最小化，大的参数差异最大化，从而获得参数的等价与否的结果。类似地，应用惩罚对齐法检验测量不变性时，公式 4-6 中哪个部分是和参数的组间差异有关呢？并且，在应用中需要对参数设定对齐先验分布，这一先验分布的形式是什么，在测量不变性检验中一般应对哪些参数进行设定？这部分内容是本文的核心，有必要交代清晰，使得读者能准确理解惩罚对齐法在测量不变性检验中的应用。
回应：感谢您的细心指正。在介绍惩罚对齐法部分确实有些关键内容没有写清楚。根据您的指正与点拨，我在文中更详细地描述了惩罚对齐法是如何通过总损失函数 F 和惩罚函数 $P(\theta)$ 进行测量不变性检验，具体介绍了对齐先验分布的形式和其设定的参数。相关内容已经在文中标红。感谢您的专业建议。

意见 8：实证分析中作者对样本和方法都进行了调整，但目前来看存在几个关键问题：
a) 对各组样本进行清洗和重新抽样的目的是使得各组样本量尽量均衡，但是现在第三组仍是其他三组样本量的 4-10 倍。所以得到目前这个样本量的标准和手段是什么，依据又是什么？
回应：感谢您提出的疑问和点拨。第一次修改稿调整样本量的时候是对除了 C9 院校样本的 1415 个个案外的其他三个院校类型个案都做了相应的缩减，普通本科院校由于样本量比较

多，所以减少得比较多，变为只有 10346 个个案。但与此带来的问题是 985 院校和 211 院校的个案也大幅减少，除了和 C9 院校样本量的比数有明显降低外，普通本科院校样本量相对 985 院校和 211 院校样本量的比数还是比较高。

总损失函数 F 中的 $W_{g1,g2}=\sqrt{N_{g1}N_{g2}}$是用来反映组样本量大小和特定组参数估计的确定性，较大的组样本量比较小的组样本量对总损失函数的贡献更大。原始样本数据的总样本量为 90416，其中 985 高校学生的样本量为 16167，211 高校学生的样本量为 9249，普通本科高校学生的样本量为 63585，C9 高校学生的样本量为 1415。根据$W_{g1,g2}=\sqrt{N_{g1}N_{g2}}$的公式可知，在比较 G3 普通本科高校组和其他组参数的不等性时，G3 对$W_{g1,g3}$的贡献是 G1 的$\sqrt{\frac{N_{g3}}{N_{g1}}}\approx 2.0$倍，G3 对$W_{g2,g3}$的贡献是 G2 的$\sqrt{\frac{N_{g3}}{N_{g2}}}\approx 2.6$倍，G3 对$W_{g3,g4}$的贡献是 G4 的$\sqrt{\frac{N_{g3}}{N_{g4}}}\approx 6.7$倍。

根据您的提醒和点拨，为了防止对齐法测量不等性优化计算中各组样本量差距过大，较大组对参数差异不等性的贡献过高，本轮修改稿我对原始样本中普通本科高校组和 985 高校组各重新随机抽样随机选取 9249 个个案保留。这样一来，四个高校组的样本量比数变为 6.5:6.5:6.5:1。985 高校、211 高校、普通高校组的参数在进行测量不等性优化计算时，它们之间的样本量权重是相等的。而这三个组在测量不等性优化计算中样本量影响的权重是 C9 院校组的$\sqrt{\frac{N_{g3}'}{N_{g4}}}=\sqrt{\frac{9249}{1415}}\approx 2.6$倍。通过这样的重新随机抽样策略，平衡了各组样本量在测量不等性优化计算中的权重，明显缩小了各组样本量的差距。

b) 作者在做 MG-CFA 检验时，根据表 3 的 df 结果来看，应是未将数据作为分类变量处理。另外在使用 WLSMV 等分类数据的估计方法时，嵌套模型的卡方差值不应直接相减，需要进行校正，具体校正方式在 Mplus 官网上有做介绍。表 3 的结果应做相应修改。

回应：感谢您的明察秋毫和细心指正。我已经将数据作为分类变量重新进行多组验证性因子分析，并且检验了形态不变模型和强测量不变模型。进行嵌套模型似然比检验时，我也计算了校正后的卡方变化值。

c) 作者根据第一轮的问题对实证分析部分惩罚对齐法的代码进行修改，使用 ESEM 与 PSEM 结合分析，这对于实践者是非常有帮助的实证操作示例。但结合目前的介绍框架稍显不连贯。个人查阅 PSEM 相关资料后，对它的理解是该方法应该可以与 ESEM、EFA、CFA 等结合使用，因此具备 ESEM 的优势，但实证分析时并不一定必须做这一应用。并且实证分析的目的是将其与传统测量不变性分析流程作对比。结合文章逻辑，建议作者可考虑如下三个实证分析步骤：① 基于 CFA 的传统测量不变性检验；② 惩罚对齐法分析，与前一步使用相同的测量模型结构，而且在使用相同因子结构的前提下，这一步获得的自由度应与传统测量不变性分析中形态等价模型的 df 是相同的，卡方值应是很接近的，因此对于这一稿和初稿中得到 df=16 的结果，我个人稍存疑；③使用结合 ESEM 的惩罚对齐法，对比结果可以显示出该方法继承 ESEM 的优势，克服 CFA 交叉载荷问题的局限。

回应：感谢您为文章内容的撰写和质量提升指明了方向。我按照您提出的 3 个步骤对研究实例分析部分进行了较大修改，结果显示采用基于探索性结构方程模型的惩罚对齐法分析数据应是更好的选择。

正如您所说，传统多组验证性因子分析的卡方值为 1407.14，基于验证性因子分析的惩罚对齐法分析的卡方值为 1403.83，卡方值是很接近的，自由度都是 32。

至于您所提出的初稿和第一轮修改稿 df=16 的结果，计算过程如下：

初稿连续型指标 ML 估计法 H_1 模型

已知参数：观测指标的方差 6 个，观测指标的协方差 5+4+3+2+1=15 个，观测指标的均值 6 个，4 个组×(6+15+6)=108 个，

自由估计参数 92 个，df=16

第一轮修改稿等级多分类指标 WLSMV 估计法 H_1 模型

已知参数：观测指标间的相关 5+4+3+2+1=15 个，阈值 12 个，4 个组×(15+12)=108 个，

自由估计参数 92 个，df=16

d) 表 6 的两两比较导致 I 类错误率膨胀的问题不是控制样本量的问题，最简单的方式是使用 Bonferroni 方法调整每两组比较时使用的 α 水平。并且，“为了降低样本量以控制两两比较时 I 类错误率”这一表述有误，应删除。

回应：感谢您的专业指导和建议。我已将文中“为了降低样本量以控制两两比较时 I 类错误率”删除，并且查阅了相关文献和资料，采用 Bonferroni 方法调整的 α 水平得出两两比较结果，在文中做出了修改。

第三轮

审稿人 1 意见：作者对第二次的审稿意见做了认真的修改，文章质量有了较大的提升，建议发表。

审稿人 2 意见：

作者针对审稿意见都在文中做了相应修改，也在回复中进行了细致的说明和解释，文章的质量具有较大提升。但文中还有一些细节问题需要修改。

意见 1：第二章“测量不变性检验方法进展回顾”中第五行“使用部分测量不变模型拟合数据如上文所述也存在一些缺点”中“如上所述”一词指代不清，应做直接的说明。

回应：感谢您的指导。我已经将“如上所述”一词指代的内容写清楚。

意见 2：第二章“测量不变性检验方法进展回顾”中第二和第三段开头的“这些方法”、“该方法”都在指代上一段的方法内容，在行文上非常不连贯。建议修改为类似与“上一段提及的 ESEM、BSEM、二水平随机效应法等方法”以及“对齐法”，做明确指定。

回应：感谢您的建议。我已经将第二和第三段开头的“这些方法”、“该方法”等指代的内容写清楚。

意见 3：第四段“国内心理统计与测量领域的研究者们与时俱进，发表了一些关于测量不变性检验方法的研究…..”与目前的前面三段内容有重合之处，应再做梳理和修改。

回应：感谢您的建议。基于前面三段的内容，我大概评估了国内关于测量不变性检验方法的研究对领域的独特贡献，对相关内容进行了梳理和修改。

意见 4：章节 2.1 和 2.3 标题上的 BSEM、AESEM 等模型的英文全称在前文新加入的内容中已有提及，理应无需重复。
回应：感谢您的细心指正。我已在文中对模型缩写和全称进行了修改。

意见 5：章节 2.1 的第二段末尾“相反，如果对参数值有较为明确的把握，可以确定是在某个小的范围内变动，则可以为该参数的后验分布设置一个较小的先验方差”中，应是为先验分布设置方差。
回应：感谢您的细心指正。我已改为“为先验分布设置方差”。

意见 6：对于图 1 下方第三段的内容“计算第一组的潜因子均值可以分为固定识别(fixedidentification)和自由识别(free identification)两种算法….”与 Alignment 方法提出文章中的阐述稍有出入。根据 Asparouhov & Muthén (2014)的文章(P3. 公式(10)处)，各组因子方差乘积为 1 的限定是对齐法必须的限定，应与自由优化或固定优化算法无关。而自由或固定优化的区别主要是在于第一组因子均值是否限定为 0。建议作者对此再做确认。
Asparouhova, T. & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment.Structural Equation Modeling: A Multidisciplinary Journal, 21(4), 495-508.
回应：感谢您明察秋毫的阅读和指正。确实如您所说，各组因子方差乘积为 1 的限定是对齐法必须的限定，我已经在文中对相关表述进行了修改。

意见 7：请将文中的公式(4)~(6)以及下面对各符号的解释使用 word 中的公式格式进行编辑，以便清晰标明每个参数的上下标，以便读者理解。
回应：感谢您对文章规范性提出的建议。我按照您的建议将公式(4)~(6)用公式格式进行了编辑。

意见 8：“4.2 多组验证性因子分析”一节中，作者对于分类数据等价性的检验还是应用不够准确。根据 Wu & Estabrook (2016)的文章，并非是载荷等价限定会导致模型无法识别，而应是分类数据的特殊性导致基线模型的不同识别方式下再继续增加参数的等价限定会导致不等价的模型出现，使得不同识别方式的模型产生了不同的结果，而理论上不同的识别方式得到的应该是等价的模型。Wu & Estabrook (2016)的文章中也提出了基本的等价性检验逻辑，根据个人的理解，对于三分类及以上的分类数据，先对基线模型进行等价检验后，先进行阈值的等价性检验，在此基础上再直接对感兴趣的参数进行等价性检验，而不同于连续数据进行顺序步骤的检验。建议作者再重新阅读 Wu & Estabrook (2016)以及相关文章，对这一部分的等价性检验内容再作更正。
回应：感谢您的悉心教导和指正。我已经遵照您的意见对文中的内容进行了相应修改，并采用检验形态不变模型→阈值不变模型→强测量不变模型的顺序进行研究实例中等级多分类指标数据的测量不变性检验。

意见 9：实例部分同时呈现了基于 CFA 和 ESEM 的惩罚对齐法分析，而附录部分目前只有一个惩罚对齐法语句，建议提供的语句与文中实例部分相对应，并做好对应的注释。
回应：感谢您对文章提出的建议。我按照您的建议提供了多组验证性因子分析强测量不变模

型、基于 CFA 的惩罚对齐法和基于 ESEM 的惩罚对齐法的 Mplus 语法，并且做好相应的注释。

第四轮

审稿人 2 意见：

意见 1：对于对齐法的识别，作者并未做出实质修改。根据 Asparouhov & Muthén (2014)的文章(P3. 公式(10)处)，对齐法的算法中本身就能够识别(2G-1)个参数，包括 G 个组的均值和 G-1 个因子方差。根据各组因子方差乘积为 1 的限定，实现剩余一个因子方差的识别。实际中常常限定第一组方差为 1 实际上是起到标准化的作用。而并不是作者文中所述的固定识别需要限定一组方差，而对齐法用的连乘为 1 的限定方式来识别方差。 自由优化或固定优化算法仅仅区别在于第一组的因子均值是自由估计还是限定为 0。 作者应对方法内容做出准确描述，以免误导读者。

回应：非常感谢您的细心指正和指导。本人对对齐法运算的个别细节确实理解有误，并且未发现，感谢您的细心指正！我已经将文中相关内容改正并标红。

意见 2：关于 CFA 等价性的部分还是存在一点小问题。

a) 作者提及 Wu & Estabrook, 2016 的文章，那么建议应尽量参照这一文章提出的等价性检验方式来实现。 目前作者根据上一轮的建议增加了阈值等价的检验，然后在此基础上进行阈值和载荷的等价性检验。但是根据 Wu & Estabrook, 2016， 在只有阈值和载荷等价限定的时候，为了识别，所有组的因子均值理论上是固定为 0 的，在这一模型基础上是无法进行因子均值比较的， 应进一步实现潜在反应变量的截距等价，才可做均值比较的检验。 因此个人也不建议将阈值和载荷等价称之为强等价性检验，阈值和截距等价是不同的，这么定义似乎不太严谨。对于如何根据 Wu & Estabrook (2016)提出的识别方式在 Mplus 中实现，作者可以参考 Svetina 等(2019)的文章。

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling: A Multidisciplinary Journal, 27(1), 111-130. http://doi.org/10.1080/10705511.2019.1602776

回应：非常感谢您的耐心建议和方向性指导。结合 Wu & Estabrook(2016)和 Svetina 等(2019)的文章，明确了 Wu 和 Estabrook 提出的多分类变量测量不变性检验的步骤和识别条件，以及如何在 Mplus 软件中实现这些检验步骤。文中多组验证性因子分析部分我也根据 Wu 和 Estabrook 提出的检验步骤进行了对照修改，相关研究结果和 Mplus 语法的改动已经标红。非常感谢您的耐心建议和方向性指导！

b) 关于结果评价。当前研究的样本量相对很大， 由于卡方检验对样本量的敏感性，结果显著是完全符合预期的，应用这一结论作为不等价的证据说服力很弱。建议补充考虑 CFI、 RMSEA 等相对拟合指标的变化帮助评价结果。同时这一实操上的问题也是传统等价性检验的缺点之一，可以在文中阐明以凸显更先进的等价性检验方法的必要性。

回应：非常感谢您的专业建议和指导。我在文中实例补充考虑了ΔRMSEA 和ΔCFI 等拟合指标的变化协助描述评价结果，同时指出了卡方检验对样本量的敏感性和拟合指标标准不统一

的问题。

c) 另外，若途中某一步骤的等价性不成立，理论上不应在此基础上继续做下一步的等价限定。
回应：感谢您的指正。实例中多组验证性因子分析的阈值不变性模型和形态不变性模型自由参数数目和自由度相等，非嵌套模型，所以还是报告了形态不变、阈值不变、阈值和载荷不变模型的结果。我已经将三种模型的语法都列在了附录供核对。

意见 3：Mplus 语句惯例是在注释部分的开头用感叹号，建议作者对此进行修改。例如：
FILE = C:\Users\wenco\Desktop\PSEM\3dele9249.csv; !指示待分析数据的路径
回应：感谢您的细心指正。我已经在语法中将感叹号的位置进行修改。

第五轮

编委 1 意见：
意见 1：建议将题目改为“测量不变性检验方法的新进展：惩罚对齐法”。
回应：感谢您的建议，已经修改了题目。

意见 2：建议将“强测量不变”直接修改为“载荷和截距不变”，因为不同检验方法和不同文献中关于强假设的说法不太一致。本文中的尺度等价也需要在“metric invariance”的前提。不如直接说载荷和截距。后面请参照修改。
回应：感谢您的专业建议。我已经在全文核对，将“强测量不变”改为“因子载荷和截距严格相等”。

意见 3：这一段 2.1 前面这一段文字进行适当压缩，尽量减少和后面详细介绍的重复。
回应：感谢您的耐心指导。我已经重新审读第二节刚开始的几段文字，进行了删减。

意见 4：我删掉了原来这里的表述，是因为在贝叶斯估计中，CI 的概念已经发生变化，作者的理解有些问题。
回应：非常感谢您的指正和修改。

意见 5：在等价性检验中是交叉载荷，还是交叉载荷的差异？如果是想说单组 BSEM 模型，建议这部分的写法修改为先写单组的 BSEM 模型，交叉载荷的设置，再谈等价性检验中交叉载荷差异的设置。
回应：感谢您的指正，我已经分别对单组 BSEM 和多组 BSEM 的内容进行介绍。

意见 6：设置 0.01 小先验方差的理由是什么？建议仔细阅读这方面的文献，这部分的表述不准确的地方比较多。
回应：感谢您的指导。我认真阅读了相关文献，补写了设置 0.01 小先验方差的原因。

意见 7：应删去组序数中的“序”字。
回应：感谢您提出的疑问。这里的序数是指第几组的意思，这里的 g1,g2 都是组编号，代表特定的组序数。

意见 8：后面所有内容需要仔细核对，有些与原文献介绍不一致的地方请修改，另外语言表述请尽量准确。
回应：感谢您的建议。我再次认真阅读了文章中的内容，将不准确的地方进行了修改。

意见 9：因子载荷矩阵中空的数字应该是很小的值，建议补充；另外，探索性因子分析中也需要将观测指标 Q1-Q6 定义为分类变量，请核实是否修改。另，为了语句的完整性，请在附录中增加 EFA 的语句。
回应：感谢您的建议和指导。遵照您的意见，已经加入了所有载荷的数值。Mplus 口令中的观测指标之前已经是分类变量，EFA 的语句已经加入附录中。

意见 10：将“定序”修改为更常用的“等级”或“顺序”。
回应：感谢您的指正。已经在全文进行修改。

意见 11：阈值不变自由参数的个数等拟合指数，为什么和形态不变模型完全相同？请检查标准数据，仔细核对。
回应：感谢您提出的疑问。形态不变和阈值不变模型的设定参照了 Svetina et al,. 2020 的研究，运用到本例中阈值不变模型中的阈值参数被设为测量不变，待估计参数相比形态不变模型减少 12*3 组=36 个。但相比形态不变模型，潜在反应变量 v1-v6 在 g2,g3,g4 中的 18 个截距和观测指标 y1-y6 在 g2,g3,g4 中的 18 个残差方差变为自由估计，所以形态不变模型和阈值不变模型的待估计参数数目相等，自由度也相等，因而拟合指数相同。本例我也发送给了 Mplus support 团队协助确认，本例中的形态不变模型和阈值不变模型恰好是相等的自由估计参数数目和自由度，无法得到 DIFFTEST 输出结果。

意见 12：修改表的格式，尽量使其少占空间，比如，可以将载荷和阈值分不同列呈现。
回应：感谢您的指导。已经按照您的意见进行修改。

意见 13：这里是 CFA 方法？另外附录中读入数据、变量等这些相同的语句，不必重复写，只需写与 XX 相同；另外，尽量精简语句的写法，减少不必要的重复。
回应：感谢您的建议和指导。我修改了附录语法的标题，让其变得更清晰。语法中重复的语句改为“同附录 X”，精简了不少篇幅。

编委 2 意见：
请作者按照编委复审意见修改，修后发表。

第六轮

编委 1 意见：
意见 1：中文名称上次我建议作者修改了，目前已修改，但是英文名称没有做对应修改；
回应：感谢您的细心指正。目前英文题目也进行了相应修改。

意见 2：附录中的 Mplus 语句有一处建议作者加注释，但是作者加的不对，“savedata:DIFFTEST=deriv3.csv; !指示保存 H1 阈值和载荷不变模型中的导数”,请作

者理解 DIFFTEST 方法，重新修改语句的注释。
回应：感谢您的指导与指正。我找到了 Mplus 软件使用 WLSMV 估计法进行 DIFFTEST 的过程文件，并且认真阅读。目前已经重新修改语句的注释。

意见 3：文章的格式，尤其是附录中 Mplus 语句的格式还有修改完善的空间，例如，“v1 by y1@1; v2 by y2@1; v3 by y3@1; v4 by y4@1; v5 by y5@1; v6 by y6@1;”可以写在同一行，而不用分六行呈现。请作者进一步修改完善。
回应：感谢您的修改意见。根据您的意见我举一反三，对附录中的语句进一步压缩合并，减少了一些篇幅，感谢您的指导！

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