《心理学报》审稿意见与作者回应

题目：多维项目反应理论模型下分类准确性和分类一致性指标
作者：汪文义，宋丽红，丁树良

第一轮

审稿人 1 意见：研究采用模拟的方法，在各种决策规则下比较了两种计算分类一致性和分类准确性指标的方法，得到了一些有意义的结论。但仍存在一些问题，供作者参考。

意见 1：在各种决策规则下比较分类准确性和分类一致性指标是本研究有别于国外期刊同类研究的创新点，但在作者在文献综述部分内容中，却缺少对各种决策规则的研究综述，仅在后文中介绍了本文适用的决策规则。因此，建议作者最好在文献综述部分对决策规则加以介绍并明确提出这一创新点。

回应：专家见解十分精辟，谢谢专家的宝贵建议。

决策规则是影响分类效果的一个非常重要的方面。在国外期刊中，项目反应理论框架下的同类研究更多倾向于指标估计方法的研究，较少注意到不同决定规则对指标估计的影响。我们已经在摘要、引言和结论部分强调了这一创新点。特别在是在引言中倒数第二段加了一段话：

另外，对于学生有重要影响（如影响受教育的机会）的决策，教育与心理测量标准要求不能基于单个测验分数(Henderson-Montero, Julian, & Yen, 2003)，要求使用多重测量(multiple measures)结果做决策，以提高测量的信度、效度、公平性等(Chester, 2003; McBee, Peters, & Waterman, 2014)。在“中小学教育修正法”和“不让一个孩子掉队”法案推动下，一般采用合成分数(composite score)合成多重测量结果，合成方法常采用联合(conjunctive)、补偿(compensatory)、联合-补偿混合和验证(confirmatory)的合成规则，并应用于英语水平考试、通识教育发展考试和学业水平评价等(Abedi, 2004; Carroll & Bailey, 2015; Chester, 2003; Henderson-Montero, et al., 2003)。其中，联合规则要求在各个测量目标上达标，补偿规则允许测量结果之间补偿，验证规则用于一个测量去证实或评估另外独立测量所提供的信息。以上关于决策规则的研究基本是集中于经典测量理论。虽然多维项目反应理论非常适合分析多重测量结果，如分析标准参照测验时，能准确地反馈学生的多个方面内容、技能和能力的诊断信息(Chang, 2012; 康春花，辛涛, 2010)，但是至今尚没有研究在多维项目反应理论框架下比较各种决策规则下的分类准确性和分类一致性。

为突出本文的创新点，我们还在修改了 3.2 节第一自然段的第一句，如下：本节主要将单维模型下的 Guo 方法推广用于计算多维模型下分类一致性和分类准确性指标，可以用于 3.1 节介绍的各类决策规则下的指标估计，这是本文和前人研究不同之处。

意见 2：本文多次引用的 Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. Applied Psychological Measurement, 31(2), 83–105 中，模拟了 4 个维度，60 题的情境，与本研究的一种模拟条件类似，但是被试样本量分为 1000、3000、6000 三个水平。本研究对样本量只设

置了 1000、3000 两种水平，对于维度数为 4，题目数为 60 的情况，是否样本量显得较少，并且样本量两个水平对于表现结果的变化情况稍显不足。

回应：专家看的十分仔细，谢谢专家提出的问题。

我们这样设置样本量，主要是根据 Yao 和 Boughton (2007, p.102)的研究结论：样本量 1000 是 MIRT 参数估计的下限，样本量 3000 时参数估计精度大幅提高(提高幅度在小数点十分位)，样本量 6000 时参数估计精度提高幅度较小(提高幅度在小数点百分位)。

意见 3：3.本文有几处明显笔误请作者仔细修改，如：

（1）“4.4 决策规则”中第三行文字 “知测验满分为 30”。

回应：谢谢专家的细心审阅。此句已经修改为：例如，当测验长度为 15 且所有测验项目的最高等级分均为 2 时，测验满分为 30，此时划界分数设置为 15 和 24 分。

（2）“5.1 总分决策规则下的指标误差评价”中第二自然段最后一行“再使用公式(13)和(15)，得到 Guo 方法或 Lee 方法所估计的分类准确性指标”，公式(13)和(15)都是 Lee 方法，并未涉及到 Guo 方法。

回应：谢谢专家的提出的问题。我们仔细检查了全文的公式编号，发现一其中一处公式漏了编号（现在的公式(12)）。根据新的公式编号，此句已经修改为：然后由真实(模拟)或估计的项目参数分别使用极大似然法估计被试能力，再使用公式(16)和(30)，得到 Lee 方法和 Guo 方法所估计的分类准确性指标。另外，修正了全文中多处引用公式的编号。例如，修正了 4.3 节中“公式(13)和(16) ”、“公式(27) ”，修正了 4.4 节中“公式(24)和(25)”等。

（3）表 3 的表头，“抽样数量”和“项目参数”位置应该是放反了。

回应：谢谢专家提出的问题。“抽样数量”和“项目参数”位置确实放反了，已经进行了修改。

（4）表 3 中的第一列，分为“真实”和“估计”两种情况，建议参照本部分第二自然段中的介绍“给定真实(模拟)项目参数，由极大似然法估计被试能力，然后分别计算估计能力或观测总分与真实(模拟)能力所在类相同的比率，得到 Guo 方法或 Lee 方法的模拟分类准确性指标(Lathrop & Cheng, 2013) 。”将“真实”这一类别修改为“模拟”，且与后文表格中“模拟值”“估计值”这种称呼相一致。

回应：谢谢专家提出的建议。为保证上下文意思一致，已经将您指出的第二自然段中部分描述修改为：给定模拟项目参数，由极大似然法估计被试能力，然后分别计算估计能力或观测总分与模拟能力所在类相同的比率，得到 Guo 方法或 Lee 方法的模拟分类准确性指标(Lathrop & Cheng, 2013)分别为公式(34)和公式(35)。由模拟或估计的项目参数分别使用极大似然法估计被试能力，再使用公式(16)和(30)可得到 Lee 方法和 Guo 方法的估计分类准确性指标。

（5）“5.1 总分决策规则下的指标误差评价”中第二自然段最后一行“然后由真实(模拟)或估计的项目参数分别使用极大似然法估计被试能力，再使用公式(13)和(15)，得到 Guo 方法或 Lee 方法所估计的分类准确性指标。”表述不清楚，进一步明确什么是模拟分类准确性指标，什么是估计分类准确性指标。

回应：谢谢专家提出的问题。如您所说，我们仅在 4.1 节最后一句用文字叙述了“模拟分类准确性，是指所有被试中模拟能力与估计能力属于同一类的比率。其中，被试的估计能力是根据模拟的作答反应估计而来（这一句话和 4.1 节第一句话已经稍做了修改）”。根据您的建议，为了更加明晰什么是模拟分类准确性指标，我们在“5.1 总分决策规则下的指标误差评价”中第二自然段增加了两个公式，分别用来计算 Guo 方法或 Lee 方法的模拟分类准确性指标(Lathrop & Cheng, 2013)：

$$SCCR_{guo}(\theta)=\frac{\sum_{i=1}^{N}\sum_{h=1}^{H}W_{nh}}{N},\quad W_{nh}=\begin{cases}1 & \text{若 } \hat{\theta}\text{与}\theta\text{都属于}h\text{类中}\\ 0 & \text{其他}\end{cases} \tag{34}$$

$$SCCR_{Lee}(\theta_h)=\frac{\sum_{i=1}^{N}\sum_{h=1}^{H}W'_{nh}}{N},\quad W'_{nh}=\begin{cases}1 & \text{若 } \sum_{j=1}^{M}y_{ij}\text{与}\tau(\boldsymbol{\theta})\text{都属于}h\text{类中}\\ 0 & \text{其他}\end{cases} \tag{35}$$

由模拟或估计的项目参数分别使用极大似然法估计被试能力，再使用公式(16)和(30)可得到 Lee 方法和 Guo 方法的估计分类准确性指标。

意见 4：本文中总分决策规则下的指标误差评价（表 3）对真实或估计项目参数、三种抽样数量条件下两类分类准确性指标的误差进行比较，是否是模拟研究所有条件的平均值，还是仅选择了其中一种条件，请具体说明。

回应：谢谢专家提出的问题。表 3 是模拟研究所有条件的平均值，已经将表 3 的表头修改为“模拟研究所有条件两类分类准确性指标的三类误差指标的平均值”。

意见 5：实验结果部分均未呈现两维模型的结果，但仅在“5.2 总分决策规则下的指标估计”中作出说明“两维模型与单维模型的结果类似（结果未列出）”。但：

（1）两维模型随着能力间相关而变化等结果应当与四维模型的结果类似，文中并未给出说明；

回应：谢谢专家提出的问题。对照两维模型的实验结果（两维模型和四维模型结果基本类似，为节省篇幅，故两维模型结果未列出），我们对“5.2 总分决策规则下的指标估计”中表 4 的实验结果描述部分进行了修改。主要修改了涉及了两维模型结果的第(2)、(4)、(5)、(6)点，如下：(2)单维、两维和四维模型下，分类准确性随着测验长度增加而严格递增；(4)两维模型和四维模型下，分类准确性多数随样本量增加而有所提高。直观上，维数越大需要估计的项目参数数量更多，对样本量有更高要求；(5)两类方法的分类准确性均随着能力间相关增加而严格递增，并且四维模型与两维模型的结果类似；(6)单维模型和两维模型下，Guo 方法下的模拟或估计的分类准确性指标均高于 Lee 方法相应指标，两种方法得到的估计值对应的 Kappa 有类似的趋势。

（2）表 5 给出总分决策规则下分类一致性指标及估计值对应的 Kappa，未呈现两维模型结果，且该部分正文中未作说明；

回应：谢谢专家提出的问题。对照两维模型的实验结果，我们对“5.2 总分决策规则下的指标估计”中表 5 的实验结果描述部分进行了修改。主要修改了涉及了两维模型结果的第(4)点和第(5)点，如下：(4)两维模型和四维模型下分类一致性随着能力间相关增加而提高；(5)无论维数多少，Guo 方法比 Lee 方法的分类一致性高，特别估计值对应的 Kappa 值差异明显。

（3）表 7 给出合成能力决策规则下分类准确性指标及估计值对应的 Kappa，未呈现两维模型结果，且该部分正文中未作说明。

回应：谢谢专家提出的问题。对照两维模型的实验结果，我们对“5.2 总分决策规则下的指标估计”中表 7 的实验结果描述部分进行了修改。主要修改了涉及了两维模型结果的第(1)点，如下：(1)两维模型和四维模型下，推广的 Guo 方法能很好地估计合成能力规则下的分类准确性和分类一致性。

意见 6：讨论部分的 6.1、6.2，内容较为宽泛，大部分内容在讨论多维 IRT 分析测验数据的优势，仅在 6.2 的后半部分回归到本研究主题“分类准确率和分类一致性评估方法”上来，讨论的针对性不强。建议对讨论部分进行重新调整和增补。

回应：谢谢专家提出的宝贵意见。根据两位专家的意见，我们重写了讨论部分，主要分为三个部分进行了叙述，分别是：(1)6.1 新方法提出的背景和意义。该小节在叙述新方法提出的背景基础之上，然后结合研究设计和研究内容、结果，作出讨论，并总结了新方法的应用领域；(2) 6.2 分类准确性和分类一致性的用处。根据另一位审稿专家所提意见，增加了这一部分内容；(3) 6.3 研究不足和有待进一步探讨的问题。参考了相关文献并结合了我们的思考写成了此小节内容。具体内容请参见第 6 节红色标注部分。

审稿人 2 意见：本文在 IRT 框架下，将计算分类一致性和分类准确性的 Guo 方法拓广至 MIRT 中，具有一定的应用前景。全文公式推导正确，模拟研究设计合理。审稿人有些小疑问和建议，如下：

意见 1：2.1 中$-\infty=\beta_{j0}<\beta_{j1}<\beta_{j2}<\ldots<\beta_{jK_j}<\beta_{j(K_j+1)}=+\infty$，在 GRM 下 β0 和 β(K+1)是“不存在”的，该模型假设 P0*=1 和 P(K+1)*=0，并不涉及其中的两个临界难度参数。

回应：谢谢专家提出的问题。原本想只需定义其中两个“不存在”的临界难度，自动得到 P0*=1 和 P(K+1)*=0。根据专家的意见，为了保持与已有文献的定义的一致性。修改了模型描述部分中两处：β_{jk} 表示与项目 j 第 k 等级难度有关的参数，并且 β_{jk} 是严格单调递增的，有 $\beta_{j1}<\beta_{j2}<\ldots<\beta_{jK_j}$；该模型假设 $P(y_{ij}\geq 0|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=1$ 和 $P(y_{ij}\geq K_j+1|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j)=0$，且项目 j 的各个等级难度是严格单调递增，公式(1)、必然事件的概率和不可能事件的概率成为项目 j 上作答反应的分布函数。

意见 2：6.2 中“简单结构”和“复杂结构”是否分别指“题目间多维”和“题目内多维”，如果是，则后者更为常用易于理解。

回应：谢谢专家提出的建议。十分同意专家的意见，后者在文献中比较常见且更易于理解，我们已经按照您的意见进行了相应的修改，如下：“题目间多维”(between item dimensionality)类型的多维项目反应理论模型下的领域分数报告研究较多(Yao, 2013; Yao & Boughton, 2007)，而“题目内多维”(within item dimensionality)类型下除了有研究报告能力领域分数(Yao, 2010)，鲜有原始总分量尺上领域分数报告研究。

意见 3：建议作者增加一小段内容介绍分类准确性和分类一致性的用处。比如分类一致性和

准确性，尤其是准确性只能在模拟研究中体现，那么其实际应用意义的什么呢？如果把分类准确性指标理解为能力参数估计值与能力“真值”之间的一种差异，那么该指统计指标的作用是否就相当于一个粗糙的 RMSE 指标值？

回应：谢谢专家提出的问题和建议。根据您所提的意见，我们在讨论部分增加了一小节（6.2 节），部分内容如下：

6.2 分类准确性和分类一致性的用处

测验的分类一致性和准确性，在实际应用中，如果条件允许可以通过重测一致性来评价分类一致性，而分类准确性可以通过模拟研究得到模拟的分类准确性。那么提出分类一致性和准确性的估计方法的实际应用意义是什么呢？如果研究结果难以应用于实际，这种研究结果的推广就比较困难，因此这是一个十分重要的问题，我们分如下三段进行叙述。

第一，由于重测条件十分苛刻而要获得重测数据不太可能(Lee, 2010)，而模拟分类准确性一般需借助估计能力模拟作答数据再估计能力并比较两者分类相同的比率，这正是为什么众多研究者要提出其他方法用于估计单个测验数据分类一致性和准确性的初衷。

第二，尽管标准参数测验的分类误差还可通过其他指标来衡量，如条件标准误等指标(戴海琦，罗照盛，2010)。由于条件标准误只能反映能力参数估计值与能力“真值”之间的一种差异，并不能直接以“百分比”的形式反映测验上所有被试的分类准确率。不过，在单维项目反应理论模型和误差分布为正态分布条件下，有研究者发现能力估计的标准误与分类准确性指标存在着一种较为复杂的非线性转换关系(Cheng, Liu, & Behrens, 2015)。在正态分布假设下，理论上这种关系应该可以推广到多维项目反应理论模型，但仍需要进行相关研究。

第三，一般来讲，模拟研究的逻辑是，如果模拟条件下结果不好，那么在错综复杂的真实情况下结果一般更加差，即模拟研究至少可以取到淘汰作用。结合本文来说，进行模拟研究的条件是相当理想的，如果在模拟条件下，测验分类准确率和分类一致性都比较低，那么在更加复杂的实际情况中，该测验的分类一致性和分类准确率就不可能比模拟情景更高。本文提出的指标可用于评价复杂决策规则下领域分数报告质量。

意见 4：另外，对于分类准确性指标，如果我们仅仅简单的统计下判准率，这样得到的结果和分类准确性指标相比有何差异或优劣？

$$判准率(\theta_h)=\frac{\sum_{n=1}^{N}W_{nh}}{N}，\ W_{nh}=\begin{cases}1 & \text{if } \hat{\theta}与\theta都属于h类中\\ 0 & \text{otherwise}\end{cases}$$

判准率相对更为简单，只需要事先设定好分类指标，然后把真值和估计值按指标分类即可，求一下所有被试的分类正确的比例。

回应：谢谢专家提出的问题。专家所提到的判准率指标，其实就是文中计算的分类准确性指标的模拟值。然而在真实测验情景中，是无法得到真值。只能得到“实际”分类，再通过估计该“实际”分类的正确分类概率，从而得到本文研究的新指标。原来文中并没有列出模拟判准率的计算公式，只是用文字进行了相应的叙述。请参见 4.1 节最后一句：模拟分类准确性，是通过模拟一批被试在一份测验上的作答反应，然后计算模拟能力与估计能力属于同一类的比率（稍做了修改）。根据两位专家的意见，为了形式化给出定义，本次修改中增加了公式(34)和公式(35)。公式(34)类似于您上面给出的公式，用于计算模拟判准率或模拟分类准确性指标。

意见 5：为评价不同决策规则下指标的表现，采用三种决策规则：(1)基于整个测验上原始总分的决策规则，划界分数设置为测验满分的 50%和 80%。例如，测验长度为 15 和所有测验项目的最高等级分为 2，知测验满分为 30，因此划界分数分别为 15 和 24 分。(2)基于各维

度能力分数的决策规则，各划界分数采用各能力维度下子测验满分的 50%和 80%。如四维模型下测验长度 30 的测验，各维度能力分数的划界分数分别为 10 和 16 分。

回应：谢谢专家提出的问题。我们对该部分描述进行了如下修改：为评价不同决策规则下指标的表现，采用三种决策规则：(1)基于整个测验上原始总分的决策规则，划界分数设置为测验满分的 50%和 80%。例如，当测验长度为 15 且所有测验项目的最高等级分均为 2 时，测验满分为 30，此时划界分数设置为 15 和 24 分。(2)基于各维度能力分数的决策规则，各划界分数采用各能力维度下子测验满分的 50%和 80%。如四维模型下测验长度为 30 的测验，每个能力维度上有 10 个项目（含测量两个维度的项目），共计 20 分，对应的划界分数设置为 10 和 16 分。

意见6：2.2多维模型下Lee方法

在多维模型下，下面介绍基于 Lee 方法(Lee, 2010)的分类一致性和分类准确性指标(Yao, 2013)。基于 Lee 方法(Lee, 2010)，推广用于计算多维模型下分类一致性和分类准确性指标。

回应：谢谢专家提出的问题和修改。我们按照了您的意见，对这一句进行了修改。

意见7：些指标主要分为两类：一类是以Lee方法为代表的基于观察分数(测验总分)的决策指标；另一类是以Guo方法为代表的基于能力分数的决策指标(Lathrop & Cheng, 2013)。

2.2 多维模型下 Lee 方法

在多维模型下，下面介绍基于 Lee 方法(Lee, 2010)的分类一致性和分类准确性指标(Yao, 2013)。记 $g(\boldsymbol{\theta})$ 表示能力分布的密度函数。假设根据测验总分将被试分为 c 类(或表现水平)，设置划界分数或划界点(cutting point)： $s_0, s_1, ..., s_c$,满足 $0 = s_0 < s_1 < ... < s_{c-1} < s_c = +\infty$ 且 $s_{c-1} < \sum_j^J K_j$ 。当被试观察总分小于 s_1 时，被试判断为第一类；当被试观察总分大于等于 s_1 且小于 s_2 时，被试判断为第二类；依次类推，当被试观察总分大于 s_{c-1} 时，被试判断为第 c 类。

本文主要关注项目反应理论模型下单个测验的指标估计，这也是该领域中一个十分重要的研究热点(Guo, 2006 ; Lathrop & Cheng, 2013; Lee, 2010; Rudner, 2005; Wyse & Hao, 2012)。这些指标主要分为两类：一类是以Lee方法为代表的基于观察分数(测验总分)的决策指标；另一类是以Guo方法为代表的基于能力分数的决策指标(Lathrop & Cheng, 2013)。是否应该改为Rudner's approach，Guo (2006)只是对Rudner's approach的改进。 为什么不采用Rudner's approach及其拓广到多维？如果非要引用GUO方法，后面引用是否应该改为Guo (2006)。

而(Lathrop & Cheng, 2013).Within the framework of item response theory (IRT), there are two recent lines of work on the estimation of classification accuracy (CA) rate. One approach estimates CA when decisions are made based on total sum scores, the other based on latent trait estimates. The former is referred to as the Lee approach, and the latter, the Rudner approach, each after its representative contributor.

一种方法是 Lee 方法，另外一种是 Rudner approach，Guo (2006) discussed a modification of Rudner's approach that evaluates the area under the posterior of u Directly. Guo (2006)只是采用 a latent distribution method While the latent distribution method relaxes several of the assumptions needed to apply Rudner's method, both approaches yield extremely comparable results.

所以，"另一类是以 Guo 方法为代表的基于能力分数的决策指标(Lathrop & Cheng, 2013)。"这样陈述不妥当，建议作者考虑如何妥当处理？为什么采用将 Guo 方法拓展到多维，推广用于计算多维模型下分类一致性和分类准确性指标。为什么不把非常普遍使用的

Rudner approach 进行多维拓广？

Expected Classification Accuracy using the Latent Distribution

Fanmin Guo

Graduate Management Admission Council

Rudner (2001, 2005) proposed a method for evaluating classification accuracy in tests based on item response theory (IRT). In this paper, a latent distribution method is developed. For comparison, both methods are applied to a set of real data from a state test. While the latent distribution method relaxes several of the assumptions needed to apply Rudner's method, both approaches yield extremely comparable results. A simplified approach for applying Rudner's method and a short SPSS routine are presented.

回应：谢谢专家提出的问题和细心审阅。"另一类是以 Guo 方法为代表的基于能力分数的决策指标(Lathrop & Cheng, 2013)。"这样陈述的确不妥。针对您提到的问题和建议，进行了修改和增加了补充说明，如下：另一类是以 Rudner 方法为代表的基于能力分数的决策指标(Lathrop & Cheng, 2013；Rudner, 2005)。Guo 方法作为 Rudner 方法的改良，不像 Rudner 方法需要借助正态性假设(Guo, 2006；Wyse & Hao, 2012)，因此本研究中暂不考虑 Rudner 方法。不过在讨论部分，针对 Rudner 方法进行了相关讨论。

意见 8：讨论部分有点脱离主题，作者在讨论部分应该注重实践方面，可是作者论文中并没有采用真实数据进行不同的方法比较，建议作者应该针对研究设计和研究内容、结果，作出深入讨论，作者是否考虑参照 Lathrop & Cheng (2013)，Lathrop & Cheng (2014)等论文讨论所涉及的方面和角度。

回应：谢谢专家提出的问题并提供相关参考文献。根据两位专家的意见，我们重写了讨论部分，主要分为三个部分进行了叙述，分别是：(1)6.1 新方法提出的背景和意义。该小节在叙述新方法提出的背景基础之上，然后结合研究设计和研究内容、结果，作出讨论，并总结了新方法的应用领域；(2) 6.2 分类准确性和分类一致性的用处。根据您上面提的意见，增加了这一部分内容；(3) 6.3 研究不足和有待进一步探讨的问题。参考了您提供的文献并结合了我们的思考写成了此小节内容。具体内容请参见第 6 节红色标注部分。

最后，非常感谢各位评审专家的宝贵意见和建议，谢谢！

第二轮

审稿人 1 意见：作者较好地解决了稿件中的问题，文章质量较初稿有了明显的改进。建议做一下检查：

意见 1：文章中公式较多，建议仔细检查公式中的上下标及符号前后表达的一致性。

回应：谢谢专家提出的问题。修改如下：

为保持公式 6 和 7 的下标的一致性，公式 6 中 P 角标及其相关叙述进行了修改。

修改了公式 8，增加了下式中间一部分，以清晰地给出概率公式，如下：

$$p_{\boldsymbol{\theta}}(h) = P_J(s_{(h-1)} \le X < s_h \mid \boldsymbol{\theta}) = \sum_{\{x: s_{(h-1)} \le x < s_h\}} P_J(X = x \mid \boldsymbol{\theta})$$

将公式 $\tau \in [\tau_h, \tau_{h+1})$ 修改为 $\tau(\boldsymbol{\theta}) \in [\tau_h, \tau_{h+1})$

将公式 $I_c = [s_{c-1}\ s_c)$ 修改为 $I_c = [s_{c-1}, s_c)$。

意见 1：文章整体读起来不太好懂，有些句子有明显翻译的痕迹，建议仔细通读语言文字，将句子修改通顺。

回应：谢谢专家提出的问题。我们仔细对通读了文章，并进行了仔细修改，部分修改内容如下：

为更好地表述愿意和衔接上下文，对引言中第一段的第二句进行了修改："标准参照测验的广泛应用或需求，很好地体现了其在教育评价中的重要性：…美国前教育部长阿恩•邓肯(Arne Duncan)曾表示"一旦建立和采用新的标准，就需要创建新的测试，测量学生是否满足这些标准"（Duncan, 2009）"；

引言中第 5 段："伴随着测量理论与实践需求和多维项目反应理论的发展"修改为"伴随着多维项目反应理论的发展"；"并指出忽视多维数据下使用单维模型会导致指标估计有偏"修改为"并指出使用单维模型分析多维数据会导致指标估计有偏"；

删除了公式 2.2.1 节第一段中一个多余的句子"多维等级反应模型已经定义了给定能力为$\boldsymbol{\theta}$的被试在项目 j 上作答反应 y_j 的条件分布。"；

公式(7)下面的一句修改为"该式表示总分为 x 的各个得分向量 $\mathbf{y}_j$ 的联合概率之和。"；

2.2.1 和 2.2.2 节的小标题分别修改为"基于 Lee 方法的分类一致性指标"和"基于 Lee 方法的分类准确性指标"；

3.2 节中"被试能力的极大似然估计为$\boldsymbol{\theta}$" 修改为"根据被试能力的极大似然估计$\boldsymbol{\theta}$"；

3.3 节中第一句话后面一部分修改为"需要统一两者的决策区域或建立两者之间的一一对应关系"；

4.3 节中第一句话中间增加了"即服从多元正态分布，其中"；

对"6.2 分类准确性和分类一致性的用处"一节内容进行了修改；

还修改了结论中第(2)条中的内容。

意见 2：仔细核对文献的引用。

回应：谢谢专家提出的问题。核对了文后各文献是否在文中引用，核对了文中各引用是否在文后文献中列出，并检查了引用格式和文后文献的格式。主要修改如下：

结论中第(2)条中的内容参考文献引用"Lathrop 等人(2013)"修改为"Lathrop 和 Cheng(2013)"。虽为多次出现，但是只有两位作者。

修改了文后参考文献：Carroll, P. E., & Bailey, A. L. (2015). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. Language Testing, 33(1) , 23–52.删除了 doi，增加了最新出版的卷期和页码。

增加了一条参考文献：Duncan, A. (2009, June 14). Address by the Secretary of Education at the 2009 Governors Education Symposium: States will lead the way towards reform. Washington, DC: U.S. Department of Education. Retrieved May 10, 2016, from http://www2.ed.gov/news/speeches/2009/06/06142009.pdf.

另外，还修改了文献列表中不符合要求的短划线等。

审稿人 2 意见： 建议引用 Between and within item multidimensionality 的初始参考文献, Adams, Wilson, & Wang (1997). The multidimensional random coefficients multinomial logit model. APM.

回应：谢谢专家提供的第一手参考文献。我们在文中和文后引用并添加了该文献，还根据文献中的描述，在文中简要地叙述了以上两个概念：根据项目与潜在维度之间的关系，多维模型或测验主要分为两类："题目间多维"(between-item multidimensionality) 和"题目内多维"(within-item multidimensionality)，其中题目间多维测验的各个项目仅能测量多个潜在维度中一个；而题目内多维测验允许每个项目考察多个潜在维度(Adams, Wilson, & Wang,

1997)。

意见 2：公式 6 和 7 中 P 角标的一致性；

回应：谢谢专家提出的问题。为保持公式 6 和 7 的下标的一致性，公式 6 中 P 角标及其相关叙述修改如下：

在项目反应理论的局部或条件独立性假设下，对于含 J 个项目的测验，能力为 $\boldsymbol{\theta}$ 的被试的测验总分为 x 的条件概率的递推公式为：

$$P_J(X=x|\boldsymbol{\theta})=\sum_{k=0}^{\min(K_J,x)}P_{J-1}(X=x-k|\boldsymbol{\theta})P_{Jk}(\boldsymbol{\theta}) \tag{6}$$

$P_{Jk}(\boldsymbol{\theta})$ 由公式(2)定义，表示能力为 $\boldsymbol{\theta}$ 的被试在项目 J 恰得 k 分的概率，$P_{J-1}(X=x-k|\boldsymbol{\theta})$ 表示含前 $J-1$ 个项目上总分为 $x-k$ 的概率。公式(6)也可以写成容易理解的公式：

$$P_J(X=x|\boldsymbol{\theta})=\sum_{y_1,y_2,...,y_J:\sum_{j=1}^{J}y_j=x,0\le y_j\le K_j,j=1,2,...,J}\prod_{j=1}^{J}P_{jy_j}(\boldsymbol{\theta}) \tag{7}$$

该式表示总分为 x 的各个得分向量 $\mathbf{y}_j$ 的联合概率之和。

意见 3：文章题目范畴大于实质内容，建议题目聚焦于研究问题或创新点上；
论文阐述逻辑仍有问题，文中的分类一致性和准确性主要应用于模拟研究之中，因此需要在文献综述或方法提出部分强调。

回应：谢谢专家提出的问题。文章题目的确没有聚焦于创新点上，根据本文的出发点和审稿专家们之前的意见，现将文章题目聚焦于创新点上。文章题目修改为：复杂决策规则下 MIRT 的分类准确性和分类一致性。英文标题修改为：Classification Consistency and Accuracy Indices for Complex Decision Rules in Multidimensional Item Response Theory。

您在本条意见中提到“论文阐述逻辑仍有问题，文中的分类一致性和准确性主要应用于模拟研究之中，因此需要在文献综述或方法提出部分强调。”。并且，您还在意见 4 中提到“审稿人认为作者可以更多地强调该方法主要适用于模拟研究，而非实证研究，所以“缺少”实际用途。”。由于您提到的这两个方面关联性较大，由于涉及 6.2 节的修改，我们在意见 4 中一并进行说明。

意见 4：另外，讨论部分增加的“6.2 分类准确性和分类一致性的用处”一节内容中好像只有第 3 点是相关的，因此需要作者再次思考这个问题。审稿人认为作者可以更多地强调该方法主要适用于模拟研究，而非实证研究，所以“缺少”实际用途。

回应：谢谢专家提出的问题。

针对您在意见 3 和意见 4 中提出的问题，我们认真检查文中的内容，发现在上一次增加的内容中有几段话的确阐述逻辑存在问题，特别是“6.2 分类准确性和分类一致性的用处”一节内容描述不当，内容不太相关，尤其是过多地方提到模拟的分类一致性和分类准确性，易形成新方法好像只能用于模拟研究的错觉。

需要特别强调的是：本文推广前人的方法，可用于复杂决策规则下多维项目反应理论模型的分类一致性和分类准确性指标估计。由于在真实测验情景下，被试真实能力未知，本文开展的模拟研究只是为了验证新指标的表现。**新方法或指标并不仅仅能用于模拟研究，更为重要是可以应用于实证研究，理由如下：**

(1)从文中叙述的方法和条件来看，新方法或指标完全可用于真实测验情景。本文提出的复杂决策规则下多维项目反应理论模型的分类一致性和分类准确性指标的估计方法，只要将相关算法嵌入到相应的多维项目反应理论模型参数估计程序中，基于测验作答数据、参数估计(中间)结果和决策规则（或划界分数），就可计算得到真实测验的分类结果的分类一致性和分类准确性指标，用于反映分类结果的信度和效度。

(2)相关文献研究显示，有些分类一致性和分类准确性指标估计方法已经应用于真实测

验。例如，在单维项目反应理论模型或其他统计模型下，Lathrop 和 Cheng(2014)在其文中的引言中提到(pp. 318-319)，前人提出的分类一致性和分类准确性估计方法，例如L&L approach和 Lee 方法(Lee 方法正是本文中提到的方法之一)，现在已经在许多实际测验中有所应用，用于评价的真实测验的分类结果质量，并且已经开发了专门商业或免费软件供用户使用。

结合您的意见和我们的回应，我们已经对“6.2 分类准确性和分类一致性的用处”一节内容进行了修改。为了便于您审阅，下面列出了该节修改后的内容：

众多研究者和本文提出了分类一致性和分类准确性的估计方法，这些方法实际用处到底是什么、是否有替代方法、这些方法如何应用于真实测验情景和是否已经有应用的例子、以及在什么情景下需要使用新方法？这些问题十分重要，直接决定这类方法或新方法的推广性。为了清晰地阐明分类一致性和分类准确性或新方法的用处，下面对这些问题分别进行说明。

第一，新方法可用于估计单个测验的分类一致性和分类准确性指标，无需进行重测、能力模拟和估计。一方面，尽管测验的分类一致性可以通过重测结果来计算，但是由于重测条件十分苛刻而要获得重测数据不太可能(Lee, 2010)，因此，实际应用中较难直接通过重测获得测验的分类一致性。另一方面，由于在实际应用中真实能力并不知道，估计分类准确性的模拟方法不甚合理且需要模拟并估计能力。具体来说，在模拟方法中，需要先根据估计的能力和项目参数，模拟作答数据再估计能力并比较两者分类相同的比率，即模拟的分类准确性。由于估计能力并非被试的真实能力，模拟方法仍有不足之处。以上两个方面的考虑，正是为什么众多研究者提出了其他方法用于估计单个测验数据分类一致性和准确性的初衷。

第二，条件标准误指标并不能直接反映测验的分类准确性。尽管标准参数测验的分类误差还可通过其他指标来衡量，如条件标准误等指标(戴海琦，罗照盛，2010)。由于条件标准误只能反映能力参数估计值与能力“真值”之间的一种差异，并不能直接以“百分比”的形式反映测验上所有被试的分类准确率。不过，在单维项目反应理论模型和误差分布为正态分布条件下，有研究者发现能力估计的标准误与分类准确性指标存在着一种较为复杂的非线性转换关系(Cheng, Liu, & Behrens, 2015)。在正态分布假设下，理论上这种关系应该可以推广到多维项目反应理论模型，但仍需要进行相关研究。

第三，新方法或指标并不仅仅能用于模拟研究，更为重要是可以应用于实证研究。首先，在真实测验情景下，由于被试真实能力未知，无法得到分类准确性真值，本文开展的模拟研究只是为了验证新指标的表现。一般来讲，模拟研究的逻辑是，如果模拟条件下结果不好，那么在错综复杂的真实情况下结果一般更加差，即模拟研究至少可以取到淘汰作用。结合本文来说，进行模拟研究的条件是相当理想的，如果在模拟条件下，新指标不能很好地估计真实（模拟）的分类准确率和分类一致性，那么在更加复杂的实际情况中，新指标就不可能应用于实际测验情景。其次，从文中叙述的方法和条件来看，新方法或指标完全可用于真实测验情景。本文提出的复杂决策规则下多维项目反应理论模型的分类一致性和分类准确性指标的估计方法，只要将相关算法嵌入到相应的多维项目反应理论模型参数估计程序中，基于测验作答数据、参数估计(中间)结果和决策规则（或划界分数），就可计算估计真实测验的分类结果的分类一致性和分类准确性指标，用于反映分类结果的信度和效度。另外，相关文献研究显示，有些分类一致性和分类准确性指标估计方法已经应用于真实测验。例如，在单维项目反应理论模型或其他统计模型下，Lathrop 和 Cheng(2014)在其文中的引言中提到(pp. 318-319)，前人提出的分类一致性和分类准确性估计方法，包括本文中用到的 Lee 方法，现在已经用于评价许多实际测验的分类结果质量，并且已经开发了专门商业或免费软件供用户使用。

第四，新方法或指标可用于复杂决策规则下多维测验的领域分数报告质量评价。领域分数(domain scores)主要反映学生在一组代表掌握某个内容领域所需的内容和技能的试题(领域)上的表现，这比量表分或测验总分更直接更能让大众理解和接受(辛涛，谢敏, 2010)。基于项目反应理论模型的领域分数更具有优势。根据题目与潜在维度之间的关系，多维模型或测验主要分为两类："题目间多维"(between-item multidimensionality) 和"题目内多维"(within-item multidimensionality)，其中题目间多维测验的各个题目仅能测量多个潜在维度中一个；而题目内多维测验允许每个题目考察多个潜在维度(Adams, Wilson, & Wang, 1997)。题目间多维测验的领域分数报告研究较多(Yao, 2013; Yao & Boughton, 2007)，而题目内多维测验仅有报告能力领域分数(Yao, 2010)。新指标可用于评估这两类测验的分类准确率和分类一致性，从而丰富分数报告内容。

最后，非常感谢各位评审专家的宝贵意见和建议，谢谢！

第三轮

编委专家意见：我个人认为，从学术质量的角度来看，这篇稿件经过专家评审和作者修改，基本上达到了发表的水平。
回应：谢谢编委专家的意见和肯定，非常感谢。

主编意见：综合各个方面的意见，考虑到应该使得学报的资源（包括篇幅）与效应（包括读者群和读者能利用的信息），学报主编建议作者在审就的定稿的基础上，压缩到一万字以内，即保留论述和实验部分，具体推演还可以请有兴趣的读者和作者做深入沟通（实际上，对于不熟悉推演的读者，仅凭目前的篇幅，也难以完全明白），在学报发表。如果作者觉得，删减有伤原文主题，或者拒绝删减，只好请作者再投其他刊物，我们可以将审稿的全部过程，包括学报主编的最后判断，一并转过去，以利清晰情况。
回应：谢谢主编的意见。已经对篇幅进行了较大的删减，至少删减了 3, 000 字。根据您的意见，主要进行了以下修改：（1）简化了引言中最后两段；（2）精简了节"2.1 多维等级反应模型"；（3）简化了节"2.2 多维模型下 Lee 方法"第一段的描述；（4）为了便于阅读，删除了原来文章中一些相对不太重要的公式及相关描述，如公式(5)、(6)、(17)、(18)、(19)、(20)、(21)、(22)、(24)、(29)、(31)、(32)、(33)等；（5）简写了部分常用的测量术语，如标准参照测验(CRT)、经典测验理论(CTT)、项目反应理论(IRT)、单维 IRT(UIRT)、多维 IRT(MIRT)；（6）简化了"4.2 研究设计"中文字描述，由于部分内容与表 1 中内容重复；（7）简化了"4.4 决策规则"的公式及相关描述；（8）简化了"5.2 总分决策规则下的指标估计"结果描述中的第二段，因为分类一致性的结果与分类准确性指标的结论基本一致；（9）简化了"6.1 新方法提出的背景和意义"第一段的描述；（10）删除了原附录中的数学证明，只在文章中进行了说明"在真分数决策规则下，根据贝叶斯定理和乘法公式，在能力先验分布 $g(\boldsymbol{\theta})$ 为均匀分布下，当 $N \to \infty$，Lee 方法的分类准确性指标的估计值 $\hat{\gamma}_{Lee}$ 和 Guo 方法的分类准确性指标的估计值 $\hat{\gamma}_{Guo}$ 均依概率收敛于 γ (由于篇幅限制，将另文叙述)"；（11）正文中"分类一致性和分类准确性"简述为"分类一致性和准确性"。

最后，非常感谢各位评审专家、编委和主编的宝贵意见和建议，谢谢！