心理学探新　2025,Vol.45,No.6,564－574
PSYCHOLOGICAL EXPLORATION

基于模型聚类与基于距离聚类在表现标准设定中的比较研究*

梁正妍[1],李　浩[2],张　潮[2],张敏强[2]
(1.广东技术师范大学教育科学学院,广州 510665;2.华南师范大学心理学院,广州 510631)

摘　要:传统表现标准设定方法(如 Angoff 法)依赖专家主观判断,在结构复杂的混合题型测验中易受质疑。本研究比较三种聚类方法——基于模型聚类、Modha－Spangler 聚类与 KAMILA 聚类在混合题型测验中的标准设定效果。模拟研究表明,类别概率与类间距对分类准确率有显著影响,Modha－Spangler 和 KAMILA 方法对类间距变化敏感,而基于模型的方法在小类间距且分布满足假定时更具适应性。实证研究显示,KAMILA 聚类在组内同质性和临界分数可解释性方面表现最优。建议在混合题型测验中优先采用 KAMILA 聚类,以提升标准设定的客观性与稳健性。

关键词:表现标准设定;混合题型测验;KAMILA 聚类;Modha－Spangler 聚类;基于模型的聚类
中图分类号:B841.2　　**文献标志码**:A　　**文章编号**:1003－5184(2025)06－0564－11

1　引言

表现标准(performance standard)是教育评估与心理测量领域中的核心概念,指依据考生在测验中的实际表现,对其是否达到特定能力水平或掌握预期知识与技能进行判定的过程(Cizek,2012)。表现标准的设定直接决定考生能力等级的划分,其结果广泛影响一系列重要教育决策,包括学生的升学与毕业资格认定、教师绩效考核与薪酬分配,以及学校层面的资源配置与问责机制。

在实践中,表现标准的设定仍主要依赖专家判断(Binici & Cuhadar,2022;李珍 等,2010)。以 Angoff 方法为代表,该方法要求学科专家(subject matter experts,SMEs)设想处于能力边界水平的“临界考生”,并估计其在每一道试题上的答对概率(Bruso et al.,2017;杨观惠,王晓慧,2023)。然而,大量研究表明,不同标准设定方法之间,甚至同一专家组在不同实施情境下,往往会产生显著差异的临界分数(Jaeger,1989)。此类方法也因此受到诸多批评:Glass(1978)认为其在本质上具有较强的武断性,而 Shepard 等人(1993)则指出,要求专家对试题逐一进行精细判断,实际上构成了一项“近乎不可能完成的认知任务”。上述观点从不同角度揭示了传统表现标准设定方法在方法论层面所面临的显著局限。

正如梁正妍和张敏强(2024)在探讨标准参照与常模参照并存的测量模型时所强调的,评价体系应当“从学生的表现出发”,这种以学生为中心的理念更符合现代教育教学评价与教育考试改革的初衷与本质。这一观点为突破高度依赖专家主观判断的传统标准设定路径提供了重要的理论指引。与此同时,标准化测验的题型结构也正在发生深刻变化:以往以选择题(multiple－choice,MC)为主的测验,正逐渐转向同时包含建构反应题(free－response,FR)(如填空、简答和论述题)的混合题型测验(mixed－format testing)(Ercikan et al.,1998)。该类测验形式已被 NAEP、AP 考试以及我国新高考等多项重要教育评价项目广泛采用。

尤其值得注意的是,在我国新高考制度背景下,学业水平等级性考试普遍采用混合题型设计,其评价机制同时融合了标准参照与常模参照的双重属性(张敏强 等,2025)。然而,针对这类兼具复杂题型结构与双重参照特征的测验形式,目前尚缺乏系统、整合的表现标准设定方法与相应的统计建模框架,这在一定程度上制约了教育评价结果的科学性与公平性。

鉴于此,本研究提出一种基于混合型数据(mixed－type data)聚类模型的表现标准设定框架。通过结合模拟研究与实证分析,系统考察不同聚类方法在分类效能、统计特性及实际可行性等方面的表现,旨在为推动表现标准设定向更加客观、稳定与可重复的方向发展提供方法论支持。

* **基金项目**:2025 年度国家社会科学基金教育学青年项目(CSA250341)。
通信作者:张敏强,E－mail:2640726401@qq.com。

1.1　聚类分析与表现标准的内在关联性

从本质上看，表现标准设定可被视为一种分类过程（Cizek，2012），即对连续的结果变量（如测试分数）进行离散化处理的操作。该过程与聚类分析在方法论上具有内在一致性——聚类分析依据样本自身的特征分布，通过特定统计准则将其划分为若干类别，使得同一类别内个体之间的差异最小化，不同类别间个体之间的差异最大化（卢燕，张颖，2010）。在表现标准设定中，聚类分析通常以考生的作答模式作为能力表征，对其潜在知识状态进行分类（汪存友，余嘉元，2010）。此类完全由算法驱动的分类方式有助于减少主观偏差，提高结果的可重复性与准确性（温红博 等，2024）。然而需注意，其结果的解释效力和效度仍高度依赖于测验本身的内容效度与结构效度。

在教育测量领域，已有研究尝试将潜在类别分析（Templin et al.，2007；Binici & Cuhadar，2022）、混合项目反应模型（Jiao et al.，2010；Templin et al.，2008）以及认知诊断模型（Henson & Templin，2008）应用于表现标准设定。例如，Brown（2007）直接将潜在类别分析用于标准设定；Templin 等人（2007）将专家评分作为先验信息嵌入潜在类别模型，提出“增强型潜在类别分析（the augmented LCA method）”以进行标准设定；Templin 等人（2008）开发了基于混合 Rasch 模型的标准设定流程，并以临界组法与对比组法中的专家评定作为贝叶斯框架中的先验信息；Jiao 等人（2010）则借助混合 Rasch 模型验证所规定的能力类别数目，并估计相应的表现标准划界分数。

尽管基于统计模型的聚类方法已在标准设定中得到较多探讨，但是目前对混合题型测验中其他聚类方法（如基于距离的聚类算法）的研究仍相对不足（梁正妍，2023），亟待进一步系统探索。

1.2　适用于混合题型测验的聚类分析方法

在聚类分析方法体系中，适用于混合型数据的处理技术主要可归纳为三类：数据转换方法、混合距离方法以及统计模型方法。其中，数据转换方法存在明显局限性：首先，若离散化过程中选用不恰当的分割点，可能导致分类准确性显著下降（Kerber，1992），其次，尽管可借助聚类过程选择最优分割点（Dougherty et al.，1995），该方法却可能引发循环论证的方法论问题（Foss et al.，2018）。鉴于上述局限性，下文未采用数据转换方法。下面将详细介绍本研究所采用的 3 种方法。

1.2.1　Modha－Spangler 聚类方法

混合距离方法的核心挑战在于如何为不同数据类型分配合适的权重，以平衡其在聚类过程中的影响，从而确保最终生成的类别结构真实反映数据内在特征，而非被某一类型数据所主导。Modha 和 Spangler（2003）基于 K－prototypes 框架提出 Modha－Spangler 算法，该算法通过暴力搜索在不同权重参数下反复执行聚类，以确定最优加权方案。研究表明，该方法在多数情况下能有效协调不同数据类型的贡献，表现出良好的聚类效能（Markos et al.，2020；Velden et al.，2019）。Modha－Spangler 聚类算法的具体步骤如下：

（1）特征加权机制——该算法为全部特征集合分配单一权重参数 α，该参数决定着连续变量与分类变量在聚类过程中的相对影响力。通过系统调整 α 值，可有效平衡不同数据类型对聚类结果的贡献度。

（2）距离度量体系——该方法整合两种距离度量方式：

平方欧氏距离：适用于连续变量空间，用于量化两个数据点在连续维度上的差异幅度。其计算公式为：

$$d^2_{Eud}(x,c) = \sum_{i=1}^{n}(x_i - c_i)^2 \tag{1}$$

其中，x_i，c_i 分别表示数据点 x 与聚类中心 c 在第 i 个连续变量上的取值。

余弦距离：适用于分类变量空间，通过计算二元向量空间中两个数据点夹角余弦值来度量其相似程度。其特点在于将取值区间限定于[0，1]范围内。分类变量的余弦距离计算公式如下：

$$d_{\cos}(x,c) = 1 - \frac{x \cdot c}{\|x\| \, \|c\|} \tag{2}$$

其中，$x \cdot c$ 表示数据点 x 与聚类中心 c 的分类变量所对应二元向量的点积，$\|x\|$，$\|c\|$ 分别表示其 L2 范数（即欧几里得距离）。

（3）加权距离整合——Modha－Spangler 方法通过加权参数 α 平衡平方欧氏距离与余弦距离的贡献度，构建融合两种度量的复合距离指标。数据点 x 与聚类中心 c 之间的复合距离 $D(x,c)$ 可表示为：

$$D(x,c) = \alpha \cdot d^2_{Eud}(x,c) + (1-\alpha) \cdot d_{cos}(x,c) \tag{3}$$

（4）优化机制——通过系统优化权重参数 α 以实现类间分离最大化。该过程通常采用暴力搜索策略，通过评估不同 α 值下的聚类质量，选择能实现最佳聚类效果的参数值，其优化目标是在所有特征空间内同时实现类内距离最小化与类间距离最大化。

（5）聚类分配——根据数据点与各聚类中心的最小复合距离进行类别归属判定。

1.2.2 Model – based 聚类方法

统计混合模型方法建立在联合分布模型的基础上,属于一种“软聚类”范式。该方法无需设定权重,因其本质上支持混合型数据的直接建模,从而避免数据转换或近似处理带来的偏误(Choi et al.,2023;McParl & Gormley,2016)。该模型能够有效捕捉变量内部及变量间的依赖关系,表现出较强的适应性。这一灵活性使其成为适用于多样情境的通用聚类工具(McLachlan & Peel,2005;McNicholas,2016)。

基于模型的聚类方法以参数化联合分布框架为理论基础,通过融入先验知识,此类模型可被精确调整以针对性解决特定研究问题。最常用的基于模型的聚类方法当属正态 – 多项混合模型(Fraley & Raftery,2002),该模型通常表述如下:

$$P(X|\Theta) = \prod_{i=1}^{n}\left(\sum_{k=1}^{K}\lambda_k N(x_{ic}|\mu_{kc},\sum_{kc})\prod_{j=1}^{J}\pi_{kj}^{xij}\right) \quad (4)$$

此处,X 表示包含 n 个观测值的连续变量 x_{ic} 与分类变量 x_{ij} 组成的完整数据集。模型参数 Θ 包含以下组成部分:混合比例 λ_k(代表聚类 k 在总体数据集中的占比)、连续变量的均值 μ_{kc} 与协方差矩阵 $\sum$、以及跨 K 个聚类和 J 个类别的分类变量多项分布概率 π_{kj}。其中,连续变量分量服从均值为 μ_{kc}、协方差矩阵为 $\sum_{kc}$ 的正态分布。

1.2.3 KAMILA 聚类方法

Foss 等人(2016)出了一种半参数化方法——KAMILA(KAy – means for MIxed LArge data)聚类,该方法融合了 K – means 聚类与高斯 – 多项混合模型的建模思路(Ahmad & Khan,2018)。KAMILA 能够在协调连续与分类变量影响的同时,放宽对参数假设的依赖。实证研究显示,在处理偏态和厚尾分布时,KAMILA 聚类相比传统高斯混合多项式模型具有更优的表现(Foss et al.,2016),体现出其对复杂数据特性具有良好的鲁棒性,在传统模型难以适用的情况下仍能保持较高的估计精度与稳健性。

KAMILA 通过迭代优化过程持续改进聚类中心与成员分配:针对连续变量采用核密度估计方法,实现对连续数据分布的非参数表征;对于分类变量,则采用能够有效处理不同类别概率的多项分布模型。KAMILA 聚类假设数据集从 $(P+Q)$ 维随机向量 $(V^T,W^T)^T$ 中抽取的 N 个样本。其中 V 是 P 维连续型随机向量,服从由 G 个正态分布混合而成的高斯混合分布,G 为聚类个数,其概率密度函数为:

$$f_V(v) = \sum_{g=1}^{G}\pi_g f_{V,g}(v;\mu_g,\sum_g) \quad (5)$$

其中,π_g 表示数据来自第 g 个正态分布的先验概率,$f_{V,g}(\cdot)$ 表示第 g 个正态分布的密度函数,μ_g 表示第 g 类的质心,$\sum_g$ 表示第 g 类的协方差矩阵。W 是 Q 维离散随机向量,服从混合多项式分布,W 的第 q 个分量的取值范围为 $\{1,2,\cdots,Lq\}$,其密度函数为:

$$f_W(w) = \sum_{g=1}^{G}\pi_g\prod_{q=1}^{Q}m(w_q;\theta_{gq}) \quad (6)$$

其中,$m(w_q;\theta_{gq})$ 多项式概率密度函数,θ_{gq} 表示第 g 个簇中第 q 个分类变量的多项式参数向量。给定第 g 个簇的聚类样本,在局部独立性假设下 $(V^T,W^T)^T$ 的联合密度函数为:

$$f_{V,W,g}(V,W;\mu_g,\sum_g,\theta_{gq}) = f_{V,g}(v;\mu_g,\sum_g)\prod_{q=1}^{Q}m(w_q;\theta_{gq}) \quad (7)$$

则聚类样本的概率密度函数为:

$$f_{V,W}(V,W) = \sum_{g=1}^{G}\pi_g f_{V,W,g}(V,W,;\mu_g,\sum_g,\theta_{gq}) \quad (8)$$

与传统的有限混合模型相比,KAMILA 倾向于在区间变量和离散变量之间实现有利的平衡。但 KAMILA 算法的灵活性是以牺牲样本容量为代价的,如果在真实分布已知的条件下,建议使用正确的分布模型。

2 模拟研究

2.1 研究设计

本研究采用蒙特卡洛模拟方法,系统比较三种聚类方法在不同实验条件下的分类准确性。模拟设定类别数为两类(如“及格/不及格”),并重点考察以下自变量:类别概率($P=0.4/0.6$、$0.5/0.5$、$0.1/0.9$)、样本量($N=600,1800,3600,7200$)以及类间距($D=1SD,2SD,3SD$,其对应重叠率分别为 9%、20% 和 40%,详见表 2)。其中,类别概率参考 Brusco 等人(2017)的模拟研究设计,本研究的三种类别概率($P=0.5/0.5$、$0.4/0.6$、$0.1/0.9$)分别对应均衡分布(如课程结业考试)、轻度不平衡(如教师资格考试)和高度不平衡(如顶尖高校自主招生测验)三类现实情境。均衡分布是最理想的类别分布,而轻度、高度不平衡旨在检验聚类方法对稀少但真实群体的识别能力,即聚类方法是否会将小类别误判为噪声(Foss et al.,2016)。类间距设为 $1SD$、$2SD$ 和 $3SD$,分别对应测验区分度是低、中、高的现实情境,旨在检验聚类方法在类别边界高度模糊、部分重叠、明晰分离时的性能。

为提升研究的生态效度,测验结构参考中国新高考数学试卷,设定总题量为 22 题,其中包括二级计分题目 12 道、三级计分题目 4 道和六级计分题目

6 道。实验设计共计 3(类别概率) ×4(样本量) ×3(类间距) ×3(聚类方法) =108 种水平组合,每种组合重复 500 次。每次重复首先采用广义部分评分模型(Generalized Partial Credit Model, GPCM)生成三个数据集,再通过数据集的两两组合构建不同类间距条件(参见表 2 与图 1)。所有数据生成与分析过程均通过 R 语言完成,主要使用 *mirt* 包(Chalmers et al. ,2012)、*kamila* 包(Foss & Markatou,2018)和 *fpc* 包(Hennig,2015)。

表 1　参数设置表

数据集	能力参数(θ)	区分度(a)	难度(b)
Set1	$\theta \sim normal(-1.5,1.0)$	$a_{bi} \sim unif(0.3,2.5)$	$b_{bi} \sim normal(0.0,1.0)$
Set2	$\theta \sim normal(0.5,1.0)$	$a_{poly} \sim unif(0.3,1.5)$	$b_{poly} \sim normal(0.0,1.0)$
Set3	$\theta \sim normal(1.5,1.0)$		$\Delta b_{poly} \sim unif(-2.5,2.5)$

注:a_{bi},b_{bi}为二级计分题区分度、难度参数;a_{poly},b_{poly}为多级计分题区分度、难度参数;Δb_{poly}为多级计分题等级难度参数

表 2　类间距水平表

类间距等级	数据集组合	重叠率(overlapping)	类别间平均能力差
大	Set1 & Set3	9%	3SD
中	Set1 & Set2	20%	2SD
小	Set2 & Set3	40%	1SD

研究的因变量为调整兰德指数(Adjusted Rand Index,ARI),用于评估聚类结果与真实类别之间的一致性(Brusco, Shireman, & Steinley,2017)。其计算公式如下:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} \tag{9}$$

$$RI = \frac{TP + TN}{\binom{c}{2}} \tag{10}$$

其中,兰德指数(RI)是用于衡量两种聚类结果一致性的指标。E(RI)表示在随机标签分配下兰德指数的期望值,其计算基于两种聚类结果的边际分布。max(RI)为兰德指数可能取得的最大值,当所有样本对在两种聚类中都属于同一类别时取值为 1。真阳性(TP)指在两种聚类中均被划分为同一类的样本对数量;真阴性(TN)指在两种聚类中均被划分到不同类别的样本对数量;而 $\binom{c}{2}$ 则表示所有可能样本对的总组合数。ARI 的取值范围为 -1 至 1,其值越高,表明两种聚类的一致性越好;若值为 0,则说明一致性不低于随机分类结果。

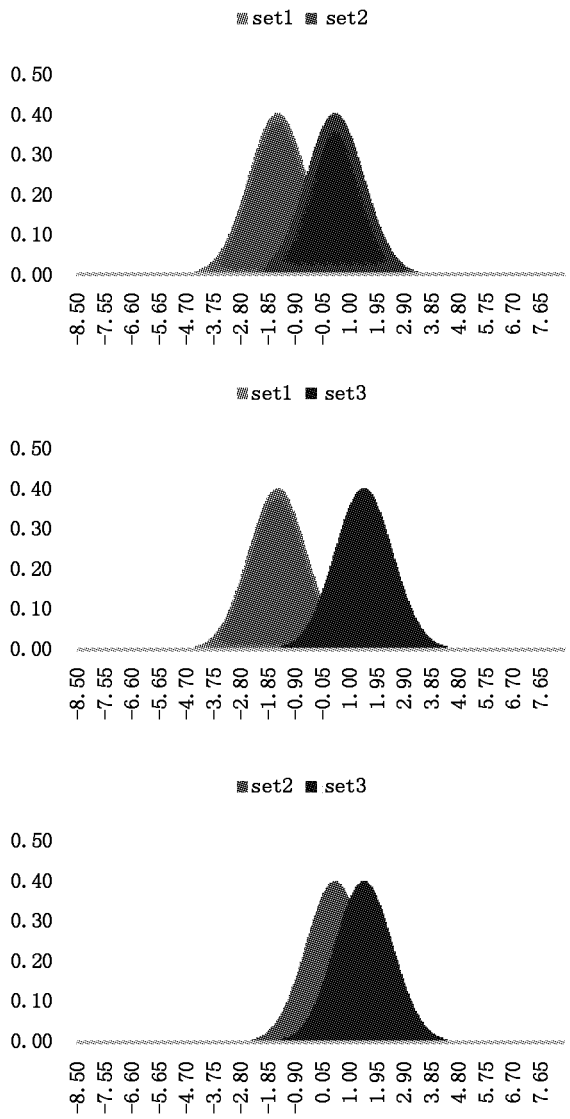

图 1　类间距水平图

2.2　研究结果

表 3 与表 4 中呈现了基于 ARI 指标的重复测量方差分析结果。组间变量为聚类方法(包括 KAMILA 聚类、Modha - Spangler 聚类和基于模型的聚类),组内变量包括样本量(N)、类别概率(P)及类间距(D)。分析结果显示,聚类方法、样本量、类别概率和类间距的主效应均显著,部分交互项也达到统计显著性水平。这表明各变量不仅单独对聚类准确性具有重要影响,变量间的交互作用也对分类结果产生了不可忽视的实质性影响。

在组内变量的主效应分析中,类间距(D)的效应量最大($\hat{\eta}^2 = 0.961$)。随着类间距缩小,分类准确率显著降低,当类间距为 1 个标准差(两类别重叠率到达 40%)时,ARI 均值仅为 0.063。类别概率(P)的主效应次之($\hat{\eta}^2 = 0.758$),其中在概率为

0.9/0.1 时分类精度最低(ARI = 0.25)。样本量(N)的主效应量相对较小($\hat{\eta}^2$ = 0.009),其对总变异的贡献未超过 1%。成对比较结果显示,样本量仅在 600 与 3600、600 与 7200 两组比较中呈现统计显著差异。表明样本量存在一个不低于 600 的临界水平,低于该阈值可能影响分类标准的稳健性。

在交互效应方面,所有组内变量交互项的效果量($\hat{\eta}^2$)均大于 0.01。其中类别概率(P)与类间距(D)的交互效应最强。具体而言,当类别概率均衡(0.5/0.5)或略不均衡(0.6/0.4),且类间距较大(3 SD)时,分类准确性最高,ARI 超过 0.70;而在类别概率高度不均衡(0.9/0.1)时,即便类间距达到 3 SD,ARI 也仅为 0.45,说明类别概率失衡对分类性能影响显著。

就组间变量(聚类方法)而言,其主效应显著,可解释约 1.7% 的总变异。除 Methods × N 与 Methods × N × D 外,其余交互项效应量均大于 0.01。其中 Methods × P($\hat{\eta}^2$ = 0.368)与 Methods × D($\hat{\eta}^2$ = 0.359)的交互效应最为突出,Methods × P × D 的三元交互效应量为 0.242。由表 5 可知,在 1 SD 类间距下各方法分类准确性普遍较差;在 2 SD 与 3 SD 条件下,不同方法的优劣随类别概率变化而异:概率均衡(0.5/0.5)时,基于模型的聚类方法最优,KAMILA 次之;概率略不均衡(0.6/0.4)时,KAMILA 表现最佳;概率高度不均衡时(0.9/0.1),KAMILA 仍保持最优,Modha - Spangler 方法位列第二。

表 3 组内变量重复测量方差分析结果

变异来源	SS	df	MS	F	p	$\hat{\eta}^2$
Sample size(N)	0.07	3	0.02	3.47	0.02*	0.009
Cluster Probabilities(P)	22.34	2	11.17	1722.85	0.00***	0.758
Cluster Distance(D)	173.13	2	86.56	13350.17	0.00***	0.961
$N \times P$	0.48	6	0.08	12.35	0.00***	0.063
$N \times D$	0.11	6	0.02	2.84	0.01**	0.015
$P \times D$	4.93	4	1.23	189.92	0.00**	0.409
$N \times P \times D$	0.63	12	0.05	8.16	0.00**	0.082

注:*p < 0.05;**p < 0.01;***p < 0.001

表 4 组间变量重复测量方差分析结果

变异来源	SS	df	MS	F	p	$\hat{\eta}^2$
Methods	0.05	2	0.02	18.90	0.00***	0.017
Methods × N	0.02	6	0.00	2.18	0.04*	0.006
Methods × P	1.54	4	0.39	319.96	0.00***	0.368
Methods × D	1.48	4	0.37	307.60	0.00***	0.359
Methods × N × P	0.04	12	0.00	2.51	0.00***	0.014
Methods × N × D	0.02	12	0.00	1.50	0.12	0.008
Methods × P × D	0.85	8	0.11	87.73	0.00***	0.242
Methods × N × P × D	0.10	24	0.00	3.56	0.00***	0.037

注:*p < 0.05;**p < 0.01;***p < 0.001

表 5 各水平分类准确性(ARI)表

方法			KA	MB	MS
N	P	D			
600		3SD	0.760(0.041)	0.765(0.035)	0.752(0.032)
	0.5/0.5	2SD	0.432(0.040)	0.462(0.068)	0.424(0.041)
		1SD	0.137(0.063)	0.188(0.119)	0.135(0.064)
		3SD	0.768(0.047)	0.750(0.080)	0.761(0.045)
	0.6/0.4	2SD	0.514(0.060)	0.514(0.088)	0.504(0.057)
		1SD	0.074(0.040)	0.135(0.074)	0.074(0.038)
		3SD	0.599(0.070)	0.356(0.109)	0.548(0.063)
	0.9/0.1	2SD	0.386(0.092)	0.318(0.135)	0.352(0.097)
		1SD	-0.027(0.009)	-0.002(0.014)	-0.027(0.010)

续表5

方法			KA	MB	MS
N	*P*	*D*			
1800		3*SD*	0.763(0.046)	0.801(0.068)	0.758(0.047)
	0.5/0.5	2*SD*	0.481(0.052)	0.512(0.063)	0.476(0.054)
		1*SD*	0.122(0.038)	0.221(0.100)	0.117(0.037)
		3*SD*	0.771(0.044)	0.754(0.055)	0.760(0.040)
	0.6/0.4	2*SD*	0.529(0.049)	0.550(0.079)	0.524(0.048)
		1*SD*	0.042(0.022)	0.044(0.030)	0.040(0.023)
		3*SD*	0.520(0.046)	0.316(0.119)	0.481(0.044)
	0.9/0.1	2*SD*	0.350(0.072)	0.301(0.158)	0.335(0.075)
		1*SD*	−0.035(0.009)	−0.007(0.014)	−0.036(0.010)
3600		3*SD*	0.799(0.020)	0.803(0.029)	0.784(0.022)
	0.5/0.5	2*SD*	0.528(0.033)	0.563(0.048)	0.500(0.032)
		1*SD*	0.112(0.040)	0.183(0.084)	0.110(0.037)
		3*SD*	0.750(0.030)	0.732(0.042)	0.740(0.028)
	0.6/0.4	2*SD*	0.459(0.038)	0.441(0.065)	0.450(0.029)
		1*SD*	0.050(0.025)	0.110(0.067)	0.051(0.025)
		3*SD*	0.510(0.070)	0.275(0.076)	0.468(0.063)
	0.9/0.1	2*SD*	0.348(0.071)	0.257(0.155)	0.329(0.074)
		1*SD*	−0.035(0.010)	−0.002(0.010)	−0.035(0.011)
7200		3*SD*	0.776(0.034)	0.766(0.046)	0.766(0.029)
	0.5/0.5	2*SD*	0.492(0.052)	0.545(0.063)	0.467(0.041)
		1*SD*	0.107(0.045)	0.193(0.231)	0.104(0.043)
		3*SD*	0.758(0.033)	0.745(0.051)	0.749(0.032)
	0.6/0.4	2*SD*	0.515(0.042)	0.517(0.072)	0.517(0.044)
		1*SD*	0.057(0.048)	0.078(0.076)	0.056(0.049)
		3*SD*	0.531(0.069)	0.302(0.080)	0.494(0.056)
	0.9/0.1	2*SD*	0.343(0.054)	0.250(0.077)	0.324(0.050)
		1*SD*	−0.032(0.007)	0.000(0.011)	−0.032(0.007)

注：KA = KAMILA 聚类；MB = Model − based 聚类；MS = Modha − Spangler 聚类

3　实证研究

3.1　研究设计

基于前述模拟研究的结果，选取在实际数据中兼具简约性与适用性的聚类模型进行比较。实证数据来自国内某大型考试，试卷总分150分，题型涵盖选择题、填空题与开放式问答题。测验共包含22道题目，计分方式包括：12道二值计分题、4道三级计分题、1道五级计分题和5道六级计分题。

研究采用分层随机抽样方法，从全国考生中抽取80,000人作为样本。差异检验结果显示，样本与总体在关键指标上无显著差异，具有良好的代表性。

考虑 Modha − Spangler 聚类方法在模拟研究的各条件下均未有最优的准确性，且在大样本数据分析中非常耗时，实证研究仅选取了 KAMILA 聚类与基于模型的聚类方法（采用混合项目反应理论模型，MixIRT）在实证数据中的分类效果，分别考察将考生划分为2类、3类及4类的情形。所有分析均在 R 语言环境中完成，使用了 *mirt*、*kamila* 和 *fpc* 等包。

3.2　研究结果

表6　两类别条件下 KAMILA 聚类和 MixIRT 模型的测验分数描述性统计结果

方法	类别	类别人数	类别比例	平均数	标准差
MixIRT	class1	38095	47.62%	77.54	19.30
	class2	41905	52.38%	34.80	15.42
KAMILA	class1	46098	57.62%	74.10	19.22
	class2	33902	42.38%	29.38	11.55

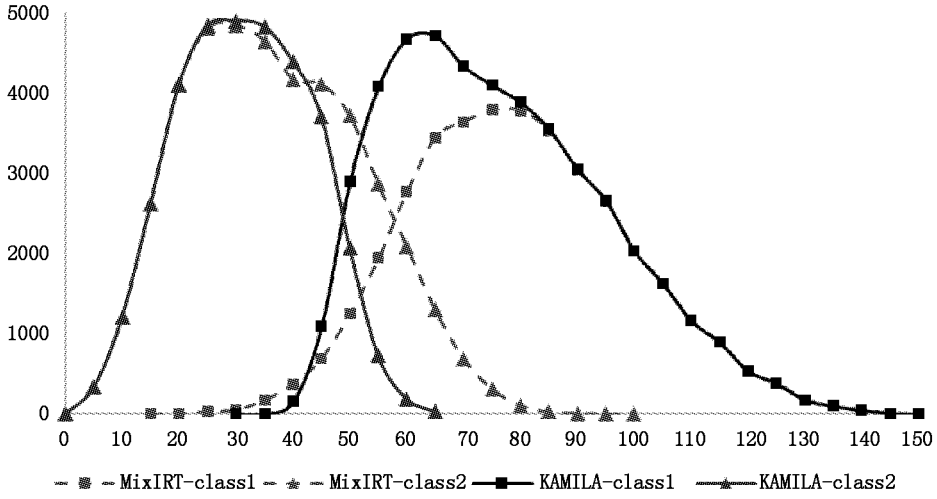

图 2　两类别条件下 KAMILA 聚类和 MixIRT 模型的测验分数分布图

表 6 与图 2 分别呈现了采用 KAMILA 聚类与 MixIRT 模型将考生划分为两个类别后的测验分数描述统计量及分布情况。分析结果表明，两种方法在表现标准设定方面存在明显差异，具体体现在以下几个方面：

首先，在划界分数方面，MixIRT 模型设定的划界分数为 60 分，而 KAMILA 聚类分析设定的划界分数较低，为 50 分。

其次，在群体分布上，KAMILA 聚类中的高能力组（class1）人数少于低能力组（class2），MixIRT 模型则呈现相反趋势。此外，KAMILA 聚类中两个组别的平均分数均高于 MixIRT 模型所对应的组别。

最后，从组内变异程度来看，两种方法在高能力组（class1）的标准差相近；而在低能力组（class2）中，KAMILA 方法的标准差小于 MixIRT，说明该组别在 KAMILA 聚类下具有更高的同质性。

表 7　三类别条件下 KAMILA 聚类和 MixIRT 模型的测验分数描述性统计结果

方法	类别	类别人数	类别比例	平均数	标准差
MixIRT	class1	35737	44.67%	78.96	18.80
	class2	27481	34.35%	41.21	14.60
	class3	16782	20.98%	27.27	14.37
KAMILA	class1	23076	28.85%	89.37	14.08
	class2	32867	41.08%	53.46	11.41
	class3	24057	30.07%	24.63	9.58

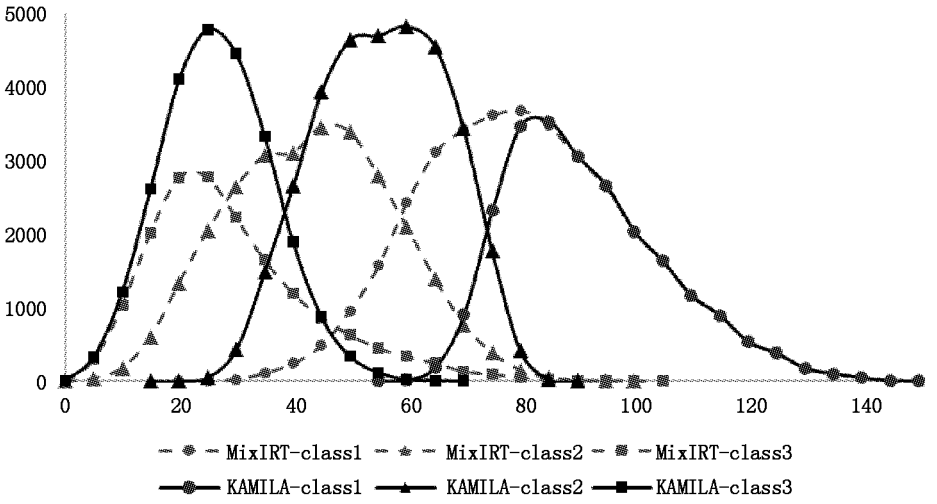

图 3　三类别条件下 KAMILA 聚类和 MixIRT 模型的测验分数分布图

表7与图3呈现了将考生划分为三个组别时，两种聚类方法在各组的描述性统计量及分数分布情况。分析结果显示，MixIRT模型设定的两个划界分数分别为60分和30分，而KAMILA方法所确定的划界分数分别为75分和40分。从组内变异程度来看，MixIRT模型中各类别的标准差均大于KAMILA方法，说明KAMILA所形成的类别内部同质性更高。此外，MixIRT所得class1与class2的平均分均低于KAMILA方法下对应组别的均值，进一步反映出两种方法在分类标准与群体区分上的系统性差异。

表8　四类别条件下KAMILA聚类和MixIRT模型的测验分数描述性统计结果

方法	类别	类别人数	类别比例	平均数	标准差
MixIRT	class1	13420	16.78%	94.68	15.95
	class2	28809	36.01%	66.50	13.82
	class3	21811	27.26%	37.14	12.98
	class4	15960	19.95%	26.03	13.02
KAMILA	class1	9678	12.10%	99.18	14.24
	class2	24796	31.00%	74.29	11.49
	class3	24179	30.22%	45.88	9.46
	class4	21347	26.68%	23.45	9.22

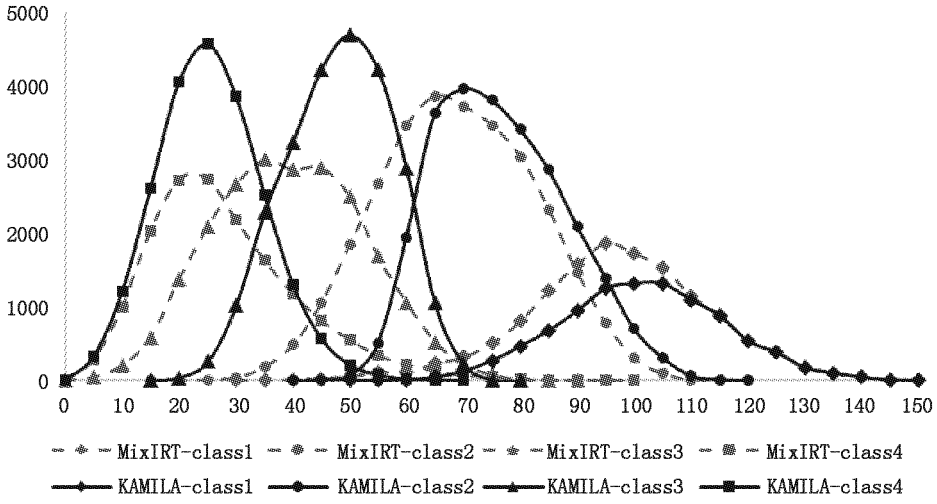

图4　四类别条件下KAMILA聚类和MixIRT模型的测验分数分布图

表8与图4展示了将考生划分为四个组别时的测验分数描述统计量与分布特征。MixIRT模型设定的三个划界分数分别为90、55和30，而KAMILA聚类方法设定的划界分数分别为95、65和35。在平均分方面，KAMILA方法中class1、class2与class3的平均分均高于MixIRT模型对应组别，仅class4的平均分略低于MixIRT模型同类组。从组内变异程度来看，MixIRT模型四个类别的标准差均高于KAMILA方法，尤其在class3与class4中差异更为明显。

综合上述实证研究结果可见，在表现标准设定中，KAMILA聚类方法整体优于MixIRT方法。多项统计指标一致表明，KAMILA方法所构建的类别内部同质性更高，分类效果更为稳定。

4　讨论

本研究旨在系统甄别适用于表现标准设定的聚类分析方法。通过文献梳理，初步筛选出三种具有潜力的方法：Modha－Spangler聚类、KAMILA聚类与基于模型的聚类。研究进一步通过模拟与实证分析，从方法严谨性、分类性能与实践适用性等方面进行比较，以确定最适合表现标准设定的算法。

模拟研究结果与已有文献一致，类别概率与类间距离对分类准确性具有显著影响。当两类群体能力差异较小（如1个标准差）时，所有方法的分类准确性均明显下降，提示在潜在能力分布重叠较高的情况下，聚类结果可能难以形成具有实质差异的类别。因此，在实际设定表现标准前，应审慎考察群体能力分布，以确保所划分的类别具备实际意义并反映真实能力结构。

在三种类别概率条件（0.5/0.5、0.6/0.4、0.9/0.1）中，群体分布严重不平衡（如0.9/0.1）时，所有方法的分类准确性均显著降低。聚类分析作为依赖

统计估计的方法，在极端分布条件下参数估计稳定性下降，因此不建议在明显不平衡的群体中直接采用此类方法进行标准设定。

从方法原理来看，Modha – Spangler 聚类属于混合距离方法，其核心在于采用合适的距离度量，以衡量不同数据类型相对于聚类中心的差异；基于模型的聚类方法则假设数据服从某种参数化联合分布（如有限混合模型），并通过模型拟合程度进行聚类（McLachlan & Peel，2005；McNicholas，2016）；KAMILA 方法则结合 K – means 与高斯混合模型理念，无需预设变量权重即可协调不同类型变量的作用。

KAMILA 与 Modha – Spangler 方法均以距离函数为基础，因而对类间距离尤为敏感。当类间距较小时，聚类中心稳定性下降，影响分类效果。在计算效率方面，Modha – Spangler 方法的运算时长约为 KAMILA 的三倍，主要因其权重优化过程计算复杂，在大规模数据中尤为耗时。

基于模型的聚类方法虽也受类间距影响，但其聚类依据更侧重于模型与数据的拟合优度。只要数据分布符合模型假设，即使类间距较小，该方法仍具一定适用性。然而，其效果高度依赖模型与数据的适配程度。对模型收敛率的评估显示（见表 9），在分布极端（如类别概率严重不平衡或类间距较小）条件下，该方法的收敛率显著下降，极端情况下低于 20%。因此，使用该方法时必须进行严格的拟合诊断，确保模型假设与数据结构一致。

表 9　基于模型的聚类方法收敛率结果

样本量	类间距	0.5/0.5	0.4/0.6	0.1/0.9
600	L	92%	72%	42%
	M	96%	74%	26%
	S	24%	68%	64%
1800	L	70%	94%	40%
	M	76%	88%	38%
	S	36%	82%	56%
3600	L	72%	90%	42%
	M	94%	76%	18%
	S	36%	56%	70%
7200	L	94%	88%	42%
	M	86%	68%	32%
	S	30%	68%	68%

注：收敛率是指同一条件下 500 次重复分析中，模型成功收敛并输出结果的次数占比。

此外，需要注意的是，当样本量不足时，要谨慎决定使用 KAMILA 方法。Foss 等人（2016）对 KAMILA 的模拟研究指出，当总样本量低于 500 时，KAMILA 的分类准确率明显下降。本研究中最小样本量设为 N = 600，正是基于该建议设定的经验下限，以确保 KAMILA 在各实验条件下具备基本可靠性。尽管如此，样本量仅在 600 与 3600、600 与 7200 两组比较中呈现统计显著差异。

本研究还将 KAMILA 方法与 MixIRT 模型应用于某大型测试的表现标准设定实证分析中。结果表明，无论在将考生划分为 2 类、3 类或 4 类的条件下，KAMILA 方法所构建的类别均显示出更高的组内同质性。结合模拟研究中各方法在稳定性与计算效率方面的表现，以及实证分析中的分类效果，可以认为 KAMILA 聚类方法在表现标准设定中具备更优的综合适用性。KAMILA 能够生成既具有统计可靠性，又能有效反映考生能力实际分布的特征标准。

在混合题型测验广泛使用的当下，设定表现标准是教育测评中的关键环节。传统设定表现标准的方法（如 Angoff 法）依赖专家对“最低合格考生”在每道题上答对概率的主观判断，虽然操作简便、理论成熟，但在面对结构复杂、题型混合的测验时，存在主观性强、一致性差、难以反映真实考生群体结构等缺陷。

与此相比，本文探究的适用于混合数据类型的聚类方法具有明显优势。一是基于实际作答数据，以数据驱动方式识别能力群体，能提升标准设定的客观性并打下实证基础；二是能直接处理连续与分类变量，无需将简答题分数强行二值化，从而更完整保留测验信息。然而，这些方法也存在局限，即对数据质量敏感，若简答题评分信度低或单选题猜测率高，噪声可能扭曲聚类结果，影响标准设定的准确性。

5　结论与展望

5.1　研究结论

基于系统的模拟比较与实证分析，可得出以下主要结论：

首先，在表现标准设定的方法选择上，KAMILA 聚类方法展现出显著优势。模拟研究表明，其在多种数据条件（包括不同类别概率、样本量与类间距）下均具有良好的分类准确性与稳健性。

其次，实证分析进一步验证了该方法的适用性。KAMILA 方法在不同分类数目（2 类、3 类与 4 类）下均能生成更具同质性的考生分组，且所设定的划界分数更具可解释性。

综上所述，研究结果为表现标准设定提供了一种有效、可靠的数据驱动解决方案。KAMILA 方法不仅具有较强的统计稳健性，其输出结果也更贴合实际教育评价情境的需要，可作为传统标准设定方法的有益补充，推荐在具备相应数据条件的教育测量实践中推广使用。

5.2　研究展望

本研究通过系统比较不同聚类分析方法在表现标准设定中的性能，为相关领域提供了有益的方法论参考。在此基础上，未来研究可从以下几个方向进一步深化和拓展本研究的成果。

首先，建议将聚类分析方法与 Angoff 等传统标准设定方法进行直接、系统的比较。传统方法如 Angoff 法依赖专家判断设定划界分数，虽广泛应用却存在主观性强、评分者一致性难保证等固有局限。聚类分析作为一种数据驱动的替代方法，具有客观、可重复的优势。未来研究可在不同测验情境下对比两类方法在标准设定准确性、稳健性及效率方面的表现，从而为教育测量实践提供更具针对性的方法选择依据。

其次，为进一步增强聚类结果的教育有效性和实践适用性，后续工作可引入学科专家参与分类结果的验证与诠释。例如，可通过专家问卷、结构化访谈或共识会议等方式，对聚类所得的能力类别进行内容合理性、教学对应性等方面的评估。提升表现标准设定的科学性与实效性。

综上，本研究为聚类分析在教育测量标准设定中的应用提供了实证基础，也凸显出其在方法融合与效度拓展方面的广阔前景。未来可通过结合传统方法与专家智慧，逐步构建更加完善、可靠的表现标准设定体系。

参考文献

李珍，辛涛，陈平.（2010）. 标准设定：步骤、方法与评价指标. *考试研究*，*6*(2)，83 – 95.

梁正妍.（2023）. *标准与常模并存的表现标准研究：基于KAMILA 聚类与潜变量建模方法*（博士学位论文）. 华南师范大学，广州

梁正妍，张敏强.（2024）. 促进教育测量模型研发护航高考改革：标准参照与常模参照并存的教育测量模型探究. *浙江考试*，(10)，11 – 14.

卢燕，张颖.（2010）. 使用聚类分析验证 Angoff 专家判断法有效性的研究：以医师资格考试医学综合笔试临床执业类别考试为例. *中国考试*，(5)，18 – 22. https://doi.org/10.19360/j.cnki.11 – 3303/g4.2010.05.003.

汪存友，余嘉元.（2010）. 标准参照测验中标准设定的聚类分析法. *南京师大学报*（社会科学版），(1)，103 – 108.

温红博，刘先伟，姜有祥.（2024）. K – means 聚类方法在中考标准设定中的信度分析. *中国考试*，(8)，69 – 78. https://doi.org/10.19360/j.cnki.11 – 3303/g4.2024.08.008.

杨观惠，王晓慧.（2023）. 基于 IRT 框架采用 Angoff 法进行合格标准设置的探索. *考试研究*，*19*(4)，59 – 66.

张敏强，梁正妍，姚敏.（2025）. 破解中国高考复杂困局：教育测量学创新与协同改革路径. *教育与考试*，(4)，5 – 10. https://doi.org/10.16391/j.cnki.jyks.2025.04.002.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508 – 600). Washington, DC: American Council on Education.

Ahmad, A., & Khan, S. S. (2018). Survey of State – of – the – Art Mixed Data Clustering Algorithms. *IEEE Access*, *7*, 31883 – 31902. https://doi.org/10.1109/ACCESS.2019.2903568.

Brown, R. S. (2007). Using latent class analysis to set academic performance standards. *Educational Assessment*, *12* (3 – 4), 283 – 301. https://doi.org/10.1080/10627190701578321

Binici, S., & Cuhadar, I. (2022). Validating Performance Standards via Latent Class Analysis. *Journal of Educational Measurement*, *59*(4), 502 – 516. https://doi.org/10.1111/jedm.12325

Brusco, M. J., Shireman, E., & Steinley, D. (2017). A comparison of latent class, K – means, and K – median methods for clustering dichotomous data. *Psychological Methods*, *22* (3), 563 – 580. https://doi.org/10.1037/met0000095.

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations*. New York, NY: Routledge.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1 – 29. https://doi.org/10.18637/jss.v048.i06

Choi, Y., Ahn, S. H., & Kim, J. (2023). Model – Based Clustering of Mixed Data With Sparse Dependence. *IEEE Access*, *11*, 75945 – 75954. https://doi.org/10.1109/ACCESS.2023.3296790

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *In Machine Learning: Proceedings of the Twelfth International Conference* (pp. 194 – 202). Tahoe City, CA, USA: Morgan Kaufmann.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. B. (1998). Calibration and Scoring of Tests With Multiple – Choice and Constructed – Response Item Types. *Journal of Educational Measurement*, *35* (2), 137 – 154. https://doi.org/10.1111/j.1745 – 3984.1998.tb00531.x

Foss, A., & Markatou, M. (2018). kamila: clustering mixed – type data in R and Hadoop. *Journal of Statistical Software*, *83* (13), 1 – 44. https://doi.org/10.18637/jss.v083.i13

Foss, A., Markatou, M., & Ray, B. K. (2018). Distance Metrics and Clustering Methods for Mixed – type Data. *International Statistical Review*, *87*, 109 – 180. https://doi.org/10.1111/insr.12274

Foss, A. J. E., Markatou, M., Ray, B. K., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, *105* (3), 419 – 458. https://doi.org/10.

1007/s10994 - 016 - 5575 - 7

Fraley, C., & Raftery, A. E. (2002). Model - based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97* (458), 611 - 631. https://doi.org/10.1198/016214502760047131

Glass, G. V. (1978). Standard and criteria. *Journal of Educational Measurement*, *15* (4), 237 - 261. https://doi.org/10.1111/j.1745 - 3984.1978.tb00072.x

Henson, R., & Templin, J. (2008, March). *Implementation of standards setting for a geometry end - of - course exam.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hennig, C. (2015). *fpc: flexible procedures for clustering* (pp. 1 - 10). https://CRAN.R - project.org/package = fpc. R packageversion 2.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485 - 514). Macmillan Publishing Co, Inc; American Council on Education.

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2010, April). *Exploring using the Mixture Rasch Model for standard setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Kerber, R. (1992). Chimerge: discretization of numeric attributes. *In Proceedings of the Tenth National Conferenceon Artificial Intelligence* (pp. 123 - 128). San Jose, CA, AAAI: AAAI Press.

Markos, A., Moschidis, O., & Chadjipantelis, T. (2020) Sequential dimension reduction and clustering of mixed - type data. *International Journal of Data Analysis Techniques and Strategies*, *12* (3), 228 - 246. https://doi.org/10.1504/ijdats.2020.108043

McLachlan, G. J., & Peel, D. (2005). *Finite Mixture Models.* In Wiley series in probability and statistics. Wiley. https://doi.org/10.1002/0471721182

McNicholas, P. D. (2016). *Mixture Model - Based Classification.* In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9781315373577

McParland, D., & Gormley, I. C. (2016). Model based clustering for mixed data: ClustMD. *Advances in Data Analysis and Classification*, *10*, 155 - 169. https://doi.org/10.1007/s11634 - 016 - 0238 - x

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k - means clustering. *Machine Learning*, *52* (3), 217 - 237. https://doi.org/10.1023/A:1024016609528

Templin, J., Cohen, A., & Henson, R. (2008, March). *Constructing tests for optimal classification in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Templin, J., Poggio, A., Irwin, P., & Henson, R. (2007, April). *Latent class model based approaches to standard setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago, IL.

van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance - based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11* (3), 1456. https://doi.org/10.1002/wics.1456.

A Comparative Study of Model - Based and Distance - Based Clustering in Performance Standard Setting

Liang Zhengyan[1], Li Hao[2], Zhang Chao[2], Zhang Minqiang[2]

(1. School of Education Science, Guangdong Polytechnic Normal University, Guangzhou 510665;
2. School of Psycholog, South China Normal University, Guangzhou 510631)

Abstract: Traditional performance standard - setting methods (e.g., the Angoff method) rely heavily on subjective expert judgments, which often raises concerns when applied to tests with complex structures and mixed item types. This study compares the effectiveness of three clustering approaches—Model - based clustering, Modha - Spangler clusteringand KAMILA clustering—in performance standard setting for mixed - format tests. Simulation results indicate that both categorical probability and inter - class distance significantly affect classification accuracy. While Modha - Spangler and KAMILA methods are sensitive to changes in inter - class distance, the Model - based approach shows greater adaptability when the class distance is small and distributional assumptions are met. Empirical analysis further demonstrates that KAMILA clustering outperforms the others in terms of within - class homogeneity and the interpretability of cut - off scores. It is recommended that KAMILA clustering be prioritized in mixed - format testing contexts to enhance the objectivity and robustness of performance standard setting.

Key words: performance standard setting; mixed - format tests; KAMILA clustering; Modha - Spangler clustering; Model - based clustering