文章编号: 1000-5862(2020) 01-0028-11

一种评判认知诊断方法有效性的新指标: 真实判准率

康春花,朱仕浩,龚 伟,俞向军,曾平飞*
(浙江师范大学教师教育学院,浙江 金华 321004)

摘要: 该文提出了 TPMR 和 TAAMR 2 种真实判准率指标,用于替代现有评判认知诊断方法有效性的指标 PMR 与 AAMR,并分析新旧指标的区别,随后比较参数和非参数方法在新旧指标上的表现以及高判、低判、错判情况,为认知诊断的实践提供理论依据. 研究发现: (i) TPMR、TAAMR 相比于 PMR、AAMR 而言,更能真实地反映认知诊断方法的有效性; (ii) TPMR、TAAMR 在不同条件下表现均较为稳定,相比 PMR、AAMR 更能区分认知诊断方法的优劣; (iii) 从高低错判的角度来看,PNN 更适合于实践应用.

关键词: 真实判准率; 方法有效性; 高低错判
中图分类号: B 841 **文献标志码**: A **DOI**: 10. 16357/j. cnki. issn1000-5862. 2020. 01. 06

0 引言

在现有认知诊断研究中,对认知诊断方法的研究较为丰富,根据文献[1]统计,目前研究者们已经开发了超过 100 种认知诊断方法. 针对这些方法,目前采用的评价指标主要是模式匹配率(Pattern Match Ratio,PMR) 和平均属性匹配率(Average Attribute Match Ratio,AAMR) [2-7]. 这 2 个指标分别从模式和属性 2 个层面验证认知诊断方法的有效性,因在实证研究中存在较多的不可控因素,从而这种检验主要在模拟研究中实现.

若要使模拟研究有效,则其生成的数据就需要与实际情况相符合,在现有认知诊断方法研究中较为常用的数据模拟方法有 3 种: (i) 传统数据模拟方法[8],该方法模拟的得分是以一定的滑动概率在理想反应得分上进行滑动,但是滑动的值只有 1 分,对于 0-1 计分的这种方法是可行的,但是对于多级计分的题目,该方法会存在一定的缺陷,如假设某一题目满分是 3 分,用该方法只能滑动为 2 分,不可能滑动为 1 分或 0 分,这显然与实际情况不符. (ii) 滑动数据模拟方法[9],该方法可以滑动任何数值,它会根据滑动概率计算出一个滑动概率矩阵,距离理想得分越近的分数滑动概率越大,而这显然更接近实际情况. (iii) DINA 模型及其相关衍生模型数据模拟方法[10-46],以 DINA 模型数据生成为例,模型设置 2 个参数: 失误参数 s 和猜测参数 g,根据公式计算出一个概率值,其数学表达式为

$$s_j = P(X_{ij} = 0 \mid \eta_{ij} = 1),$$

$$g_j = P(X_{ij} = 1 \mid \eta_{ij} = 0),$$

$$P_j(\alpha_i) = P(X_{ij} = 1 \mid \alpha_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}},$$

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}.$$

通过 s、g 的设定计算其变化的概率,并与随机数比较得到模拟作答数据. 这 3 种模拟数据的方法均有公共因子,在前 2 种中该因子是滑动概率,在第 3 种中该因子是失误参数和猜测参数,但无论是滑动概率还是失误和猜测参数均为概率值,即模拟的作答数据有可能与理想反应得分完全相同,而此时无论用任何方法都可以对被试者进行准确判别,而这种判准情况并不是认知诊断方法真实的有效状况.

目前,较为常用的 2 种评价认知诊断方法的指标主要是 PMR 和 AAMR,其计算公式为

$$P_{MR} = \sum_{i=1}^{N} N_{i\_correct}/N,$$

$$A_{AMR} = \sum_{i=1}^{N}\sum_{k=1}^{K} N_{ik\_correct}/(NK),$$

其中 N 为被试人数; K 为属性个数; $N_{i\_correct}$ 为被试 i 的属性掌握模式判断值,判对为 1,否则为 0; $N_{ik\_correct}$ 为被试 i 的属性 k 判断值,判对为 1,否则为 0.

上述 2 个指标可以直观地评价某种认知诊断方

收稿日期: 2019-09-12
基金项目: 教育部人文社会科学青年基金(19YJC880122) 和浙江省教育科学规划课题(2019SCG312) 资助项目.
通信作者: 曾平飞(1963-) ,男,广西荔浦人,教授,博士,主要从事心理测量与评价方面的研究. E-mail: zpf@ zjnu. edu. cn

法的判别效果好坏,但存在以下不足(以 PMR 指标为例,AAMR 与 PMR 情况相同):(i) 在计算 PMR 指标时,以属性掌握模式判对的被试人数除以总人数. 在模拟研究中通过设定滑动概率模拟在实践中学生失误或超常发挥情况,然而在模拟研究中无论采用何种滑动方法均会出现某些学生的理想反应模式(Ideal Response Pattern,IRP) 与观察反应模式(Observed Response Pattern,ORP) 完全相同的情况,这种情况在题量少、滑动概率小时更为突出. 而这些 IRP 与 ORP 完全相同的学生无论采用何种方法都必然会判对,即把这些学生判对的原因是数据本身而并非诊断方法有效. 然而,目前的 PMR 指标计算将这种情况同样归为方法的有效性,导致 PMR 指标与实际情况相比偏高,其高估了诊断方法的有效性. (ii) 现有 PMR 指标只给出判别结果,即每种诊断方法的判准率,对于每一名学生只有判对或判错 2 种情况,对于判错的学生是被判好了还是判差了并未做出任何评价. 为解决上述不足,本文提出比 PMR、AAMR 更精准的评判认知诊断方法有效性的指标,并对不同诊断方法判错情况比较高低错判比例,以便让一线教师可从多个角度选择认知诊断方法.

1　真实判准率指标的构建与定义

1.1　2 种真实判准率的构建

目前普遍采用 PMR 和 AAMR 指标作为认知诊断方法的评判标准. 然而,这 2 个指标都存在对诊断效果高估的倾向,因此,本文提出真实模式匹配率(True Pattern Match Ratio,TPMR) 和真实平均属性匹配率(True Average Attribute Match Ratio,TAAMR). 其具体计算公式为

$$T_{PMR} = \left(\sum_{i=1}^{N} N_{i\_correct} - T\right) /(N - T),$$

$$T_{AAMR} = \left(\sum_{i=1}^{N}\sum_{k=1}^{K} N_{ik\_correct} - TK\right) /((N - T)K),$$

其中 N 为被试人数; K 为属性个数; $N_{i\_correct}$ 为被试 i 的属性掌握模式判断值,判对为 1,否则为 0; $N_{ik\_correct}$ 为被试 i 的属性 k 判断值,判对为 1,否则为 0; T 为 IRP 与 ORP 相同的被试数量.

计算 TPMR 与 TAAMR 2 个指标时,需剔除 IRP 与 ORP 相同的学生样本,并在此基础上计算判准率,此时的结果才是诊断方法真实的有效性. 而数据模拟的方式主要分成 3 种,其中传统数据模拟方法和滑动数据模拟方法都只涉及到 1 个参数,即滑动概率,而 DINA 模型的数据模拟方法涉及到 2 个参数,即失误参数和猜测参数. 然而,在实际数据模拟过程中,在计算未产生滑动的学生样本比例时,这 3 种数据模拟方法是相同的,具体说明如下:

假设某次模拟实验题量为 10,滑动概率为0.1,在使用传统数据模拟方法或滑动数据模拟方法时,在每一题上学生作答情况与真实情况相同的概率为 1 - 0.1 = 0.9,即0.1的可能性得分发生滑动,而0.9的可能性得分不变,故在整个测验上学生作答情况与真实情况相同的概率为 0.9 的 10 次方(约为 0.35),即有 35% 的学生 IRP 与 ORP 相同. 在使用 DINA 模型模拟数据时,设定失误参数和猜测参数均为0.1,原有公式计算为

$$P_j(\alpha_i) = P(X_{ij} = 1 \mid \alpha_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}} = 0.1^{1-\eta_{ij}}0.9^{\eta_{ij}}, \quad (1)$$

由(1) 式可知,当理想反应得分为 0 分时,概率值为 0.1,当理想反应得分为 1 分时,概率值为 0.9; 用该概率值与随机数比较可知,当理想得分为 1 分时,有 0.9 的概率不发生变化,而有 0.1 的概率从 1 分变为 0 分; 当理想得分为0分时,有0.1的概率从0分变为 1 分. 因此,无论理想反应分数是 0 分还是 1 分,其变化的概率均为 0.1. 故在模拟数据过程中,当计算其未产生滑动的学生样本比例时,其计算结果与传统数据模拟方法或滑动数据模拟方法相同.

如表 1 所示,给出在不同题量与滑动概率(失误参数和猜测参数) 情况下 IRP 与 ORP 相同的学生人数占总人数的比例. 由表 1 可知,当滑动概率较小、题量较少时,IRP 与 ORP 相同的学生占比较高; 当滑动概率较大、题量较多时,IRP 与 ORP 相同的学生占比较低.

表 1　不同题量、滑动概率下未滑动人数占比

题量	滑动概率					
	0.05	0.10	0.15	0.20	0.25	0.30
5	0.77	0.59	0.44	0.33	0.24	0.17
6	0.74	0.53	0.38	0.26	0.18	0.12
7	0.70	0.48	0.32	0.21	0.13	0.08
8	0.66	0.43	0.27	0.17	0.10	0.06
9	0.63	0.39	0.23	0.13	0.08	0.04
10	0.60	0.35	0.20	0.11	0.06	0.03
11	0.57	0.31	0.17	0.09	0.04	0.02
12	0.54	0.28	0.14	0.07	0.03	0.01
13	0.51	0.25	0.12	0.05	0.02	0.01
14	0.49	0.23	0.10	0.04	0.02	0.01
15	0.46	0.21	0.09	0.04	0.01	0.00
16	0.44	0.19	0.07	0.03	0.01	0.00
17	0.42	0.17	0.06	0.02	0.01	0.00
18	0.40	0.15	0.05	0.02	0.01	0.00

表 1(续)

题量	滑动概率					
	0.05	0.10	0.15	0.20	0.25	0.30
19	0.38	0.14	0.05	0.01	0.00	0.00
20	0.36	0.12	0.04	0.01	0.00	0.00
21	0.34	0.11	0.03	0.01	0.00	0.00
22	0.32	0.10	0.03	0.01	0.00	0.00
23	0.31	0.09	0.02	0.01	0.00	0.00
24	0.29	0.08	0.02	0.00	0.00	0.00
25	0.28	0.07	0.02	0.00	0.00	0.00
26	0.26	0.06	0.01	0.00	0.00	0.00
27	0.25	0.06	0.01	0.00	0.00	0.00
28	0.24	0.05	0.01	0.00	0.00	0.00
29	0.23	0.05	0.01	0.00	0.00	0.00
30	0.21	0.04	0.01	0.00	0.00	0.00

1.2 高判、低判及错判的定义

在以往研究中,评价认知诊断方法只给出方法的判准率情况,通过判准率的高低来评价诊断方法的好坏,未考虑诊断错误的学生属于高低错判中的哪一种. 为使一线教师可以从多方面地评价一种诊断方法,本研究在原有判准率基础上,对于诊断错误的学生,分析其为高判(High discriminant,HD)、低判(Low discriminant,LD)还是错判(Error discriminant,ED)中的哪一种. 3 个指标的具体计算公式为

$$H_D = h_n / \sum_{i=1}^{N} I(N_{i\_correct} = 0),$$

$$L_D = l_n / \sum_{i=1}^{N} I(N_{i\_correct} = 0),$$

$$E_D = e_n / \sum_{i=1}^{N} I(N_{i\_correct} = 0) = 1 - H_D - L_D,$$

其中 N 为被试人数;I 为指示函数,满足函数内条件为 1,不满足为 0;$N_{i\_correct}$ 为被试 i 的属性掌握模式判断值,判对为 1,否则为 0;h_n、l_n、e_n 分别为高判、低判、错判的个数.

以表 2 为例说明上述 3 种情况,假设某学生真实属性掌握情况为(11100),判对是将学生判别为与真实属性掌握情况相同,即判为(11100);高判是在学生真实属性掌握基础上增加了掌握的属性,在表 2 中只给出其中一种情况,还包括(11101)、(11111),而(11011)并不是高判,而是错判,因其真实掌握的 3 个属性并未完全判准;低判是在学生真实属性掌握基础上减少了掌握的属性,在表 2 中也只给出其中一种情况,还包括(10100)、(01100)、(10000)、(01000)、(0010)、(00000);除上述情况外,其余情况皆为错判,在表 2 中只给出了一种错判情况.

表 2 高判、低判与错判情况

学生	属性 1	属性 2	属性 3	属性 4	属性 5
真实属性掌握情况	1	1	1	0	0
判对	1	1	1	0	0
高判	1	1	1	1	0
低判	1	1	0	0	0
错判	1	1	0	1	0

2 非参数诊断方法的新旧指标比较

2.1 研究目的

研究通过比较 PNN、KNN、MDD、EDD 4 种非参数诊断方法在 PMR、AAMR、TPMR、TAAMR 4 种新旧指标上的结果,探讨 TPMR、TAAMR 与 PMR、AAMR 相比的优势,以及在评判认知诊断方法时采用 TPMR、TAAMR 指标的可行性与必要性,并分析在诊断错误时,4 种方法更倾向于高判、低判还是错判.

2.2 实验设计

实验采用 6×5×8×4 的 4 因素混合实验设计,自变量分别为滑动概率(0.05,0.10,0.15,0.20,0.25,0.30)、题量(10,15,20,25,30)、被试人数(100,200,300,500,1 000,2 000,3 000,5 000)和判别方法(PNN,KNN,MDD,EDD). 每个实验条件重复进行 30 次. 采用 PMR、AAMR、TPMR、TAAMR 以及 HD、LD、ED 作为诊断方法的评价指标,其具体计算公式与定义见前一节. 从公式中可知,PMR、TPMR 与 AAMR、TAAMR 相比能更敏感地反映分类准确率,而 TPMR、TAAMR 与 PMR、AAMR 相比更能客观地评价诊断方法.

2.3 实验流程

第 1 步 被试的属性掌握模式确定与测验 $\boldsymbol{Q}$ 矩阵模拟. 研究使用高阶 DINA 模型模拟被试的属性掌握模式. 采用蔡艳等[17]的方法,在保证测验 $\boldsymbol{Q}$ 矩阵至少包含 1 个可达矩阵的情况下,使测验 $\boldsymbol{Q}$ 矩阵中其他元素考核的概率为 0.5,对不符合属性层级结构的考核模式重新模拟直至符合层级结构.

第 2 步 根据属性个数与属性层级结构确定所有被试可能的知识状态(Knowledge State,KS),并与测验 $\boldsymbol{Q}$ 矩阵相乘得 IRP 矩阵,再模拟作答反应数据. 作答反应数据生成的具体步骤为:(i)产生一个服从均匀分布 $U(0,1)$ 的随机数矩阵,其维度为 $N\times$

J,其中 N 表示被试数量,J 表示题量; (ii) 设定滑动概率大小; (iii) 采用张淑梅等[9] 提出的滑动模拟方法得到滑动矩阵,并使滑动矩阵内得每个分数的概率与随机数矩阵对应位置的 r_{ij}进行比较,根据滑动规则将不同的 r_{ij}分别滑动到不同的得分,即得到模拟被试的 ORP.

第 3 步　分别利用 PNN、KNN、MDD、EDD 4 种诊断方法对被试的 ORP 进行判别,并与真值进行比较,得到 4 个评价指标 PMR、AAMR、TPMR、TAAMR,并分析错判的被试,得出 HD、LD、ED 的比例. 数据模拟程序、PNN、MDD、EDD、4 个评价指标、3 种错误判别分类均通过自编 R 语言程序实现,KNN 判别由自编 R 语言程序调用 CLASS 包实现.

2.4　实验结果

由于数据结果较多,且前人已有研究表明非参数方法基本不受被试人数影响[18-20],故表 3 只列出了在被试人数为 1 000 时的 PMR 与 TPMR 的统计结果,其余人数情况与 1 000 人情况趋势大致相同. 由表 3 可以看出,在不同滑动概率、题量和方法情况下,TPMR 指标一直低于 PMR 指标,且当滑动概率越小、题目数量越少时,2 个指标之前的差异越大,这说明 TPMR 指标能有效剔除 IRP 与 ORP 相同的学生,得到诊断方法的真实判准率. AAMR 指标与 TAAMR指标对比情况与上述相同.

表 3　不同条件下 4 种方法的 TPMR 与 PMR 比较

滑动概率	题量	PNN		KNN		MDD		EDD	
		PMR	TPMR	PMR	TPMR	PMR	TPMR	PMR	TPMR
0.05	10	0.90	0.76	0.91	0.79	0.94	0.84	0.90	0.75
	15	0.96	0.92	0.96	0.92	0.99	0.99	0.95	0.91
	20	0.97	0.96	0.98	0.97	1.00	1.00	0.97	0.96
	25	0.99	0.98	0.99	0.99	1.00	1.00	0.99	0.99
	30	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99
0.10	10	0.78	0.66	0.81	0.70	0.85	0.78	0.79	0.67
	15	0.87	0.84	0.90	0.87	0.97	0.96	0.88	0.84
	20	0.92	0.91	0.94	0.93	0.99	0.99	0.93	0.92
	25	0.95	0.94	0.96	0.96	1.00	1.00	0.95	0.95
	30	0.97	0.96	0.98	0.97	1.00	1.00	0.97	0.97
0.15	10	0.65	0.56	0.71	0.63	0.73	0.67	0.67	0.58
	15	0.78	0.76	0.81	0.80	0.91	0.90	0.78	0.76
	20	0.84	0.83	0.88	0.87	0.97	0.97	0.84	0.84
	25	0.88	0.87	0.92	0.92	0.99	0.99	0.89	0.89
	30	0.91	0.91	0.93	0.93	1.00	1.00	0.92	0.92
0.20	10	0.56	0.51	0.60	0.55	0.61	0.56	0.55	0.49
	15	0.65	0.64	0.71	0.70	0.82	0.81	0.67	0.66
	20	0.73	0.73	0.79	0.78	0.93	0.92	0.74	0.74
	25	0.79	0.79	0.84	0.84	0.96	0.96	0.79	0.79
	30	0.83	0.83	0.87	0.87	0.98	0.98	0.84	0.83
0.25	10	0.44	0.40	0.48	0.45	0.52	0.49	0.46	0.43
	15	0.55	0.54	0.60	0.59	0.71	0.71	0.55	0.55
	20	0.63	0.62	0.69	0.69	0.85	0.85	0.64	0.64
	25	0.69	0.69	0.73	0.73	0.91	0.91	0.68	0.68
	30	0.73	0.73	0.77	0.77	0.95	0.95	0.73	0.73
0.30	10	0.33	0.31	0.38	0.37	0.39	0.37	0.35	0.33
	15	0.45	0.45	0.49	0.49	0.60	0.60	0.44	0.44
	20	0.52	0.52	0.57	0.57	0.74	0.74	0.52	0.52
	25	0.57	0.57	0.63	0.63	0.82	0.82	0.58	0.58
	30	0.61	0.61	0.65	0.65	0.88	0.88	0.62	0.62

具体 PMR、TPMR，AAMR、TAAMR 之间的对比结果如图 1 ~ 图 4 所示. 4 张图分别为各水平下 4 种诊断法评判结果的新旧指标对比. 由此可知以下结论: (i) 4 种方法受滑动概率影响较大，滑动概率越大，4 种方法的评判结果越差; 而滑动概率越小，TPMR 与 PMR 指标之间的差异越大，这与前一节中的理论相符. (ii) 4 种方法受题目数量影响较大，题目数量越多，4 种方法的评判结果越好; 而 TPMR 指标受题目数量的影响更大，即随着题目数量的变化，TPMR 指标的变化比 PMR 指标更大，这同样与前一节中的理论相符. (iii) 新指标在 4 种方法中表现相似，这说明新指标对在不同方法上均适用. (iv) 从图 2 ~ 图 4 可知 MDD 诊断法的诊断结果最好，而在 MDD 诊断法的 4 个结果指标中，TPMR 与 PMR 差距也是相对更小的，这说明好方法在新指标上的表现更加稳定，而相对差一些的方法在新指标上的表现变化较大. 4 张图中同样呈现了各水平情况下的 TAAMR、AAMR，大致趋势与 TPMR、PMR 一致，可以看到 AAMR 指标没有 PMR 敏感，与之前判断一致.

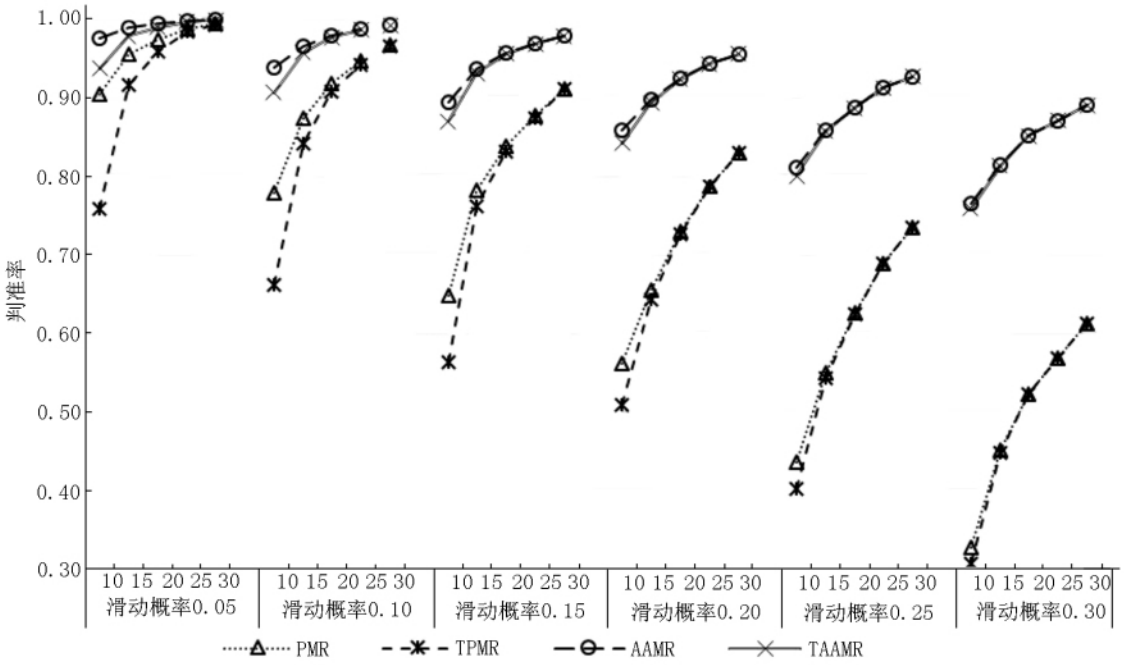

图 1　各水平下 PNN 诊断法新旧指标对比图

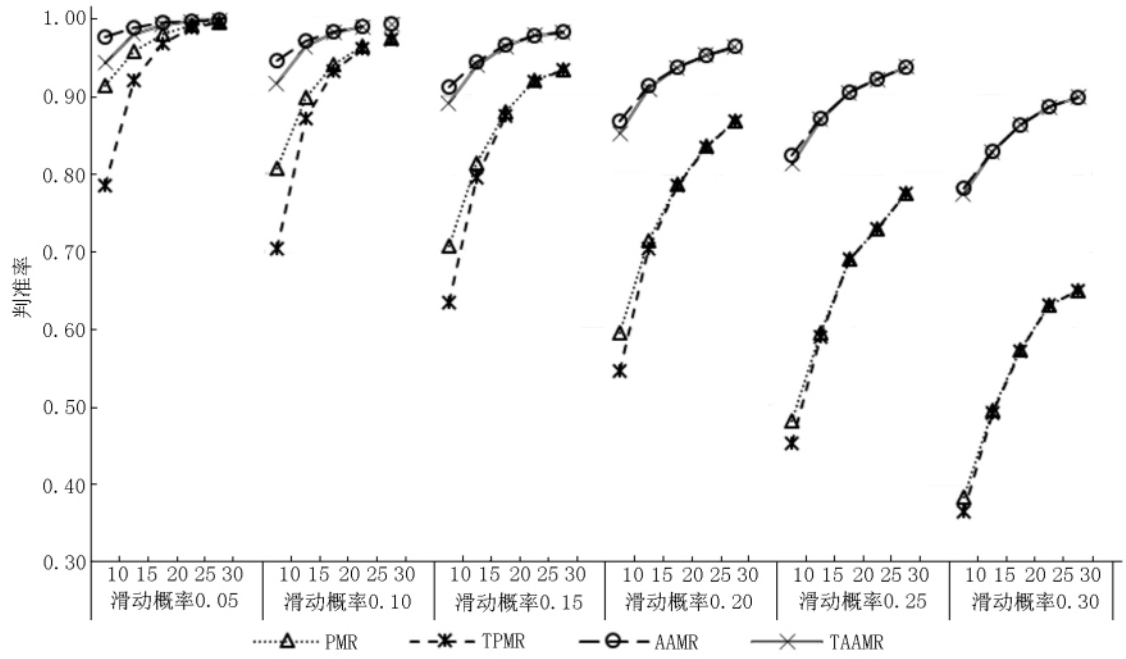

图 2　各水平下 KNN 诊断法新旧指标对比图

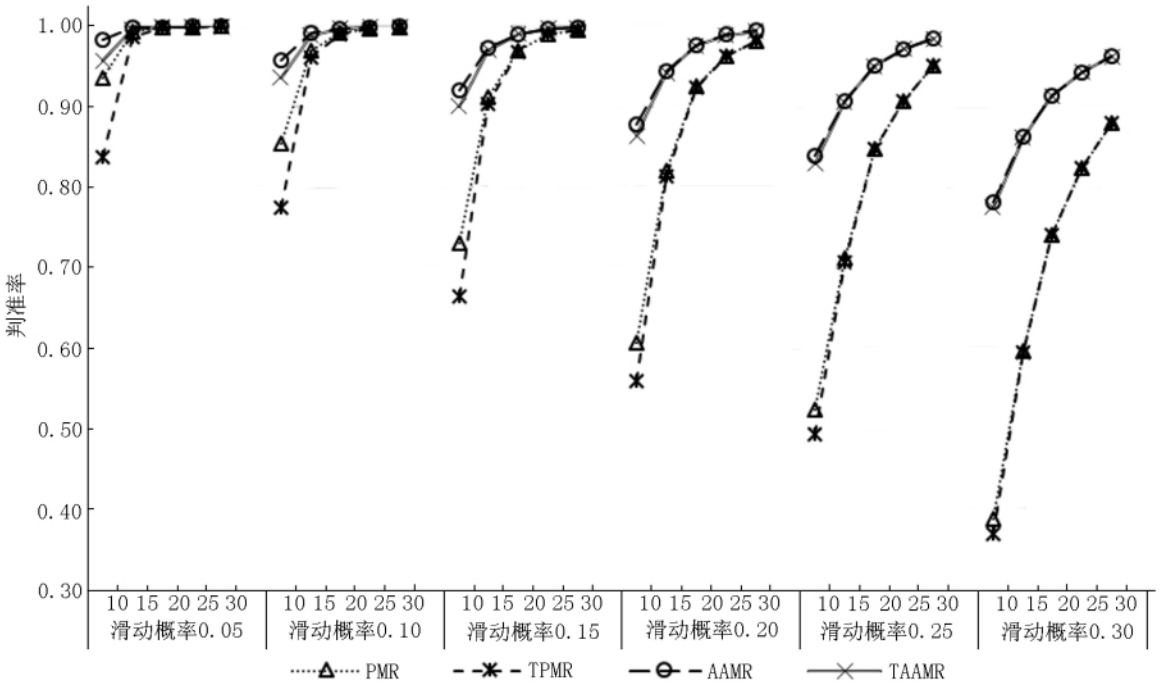

图 3　各水平下 MDD 诊断法新旧指标对比图

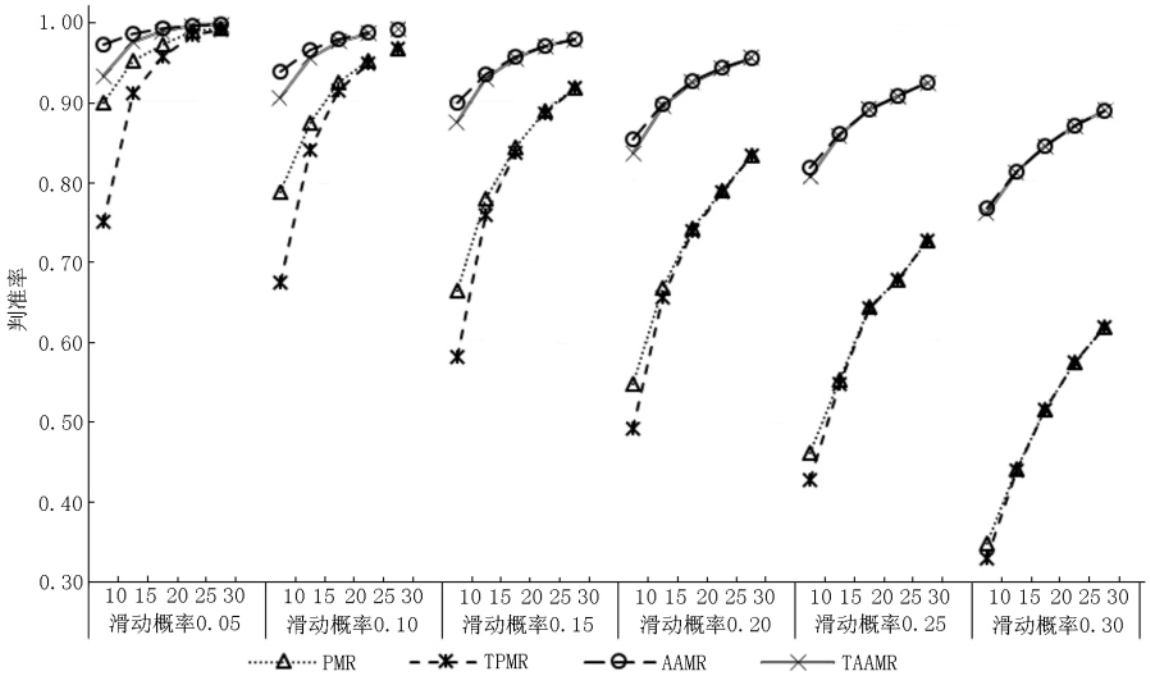

图 4　各水平下 EDD 诊断法新旧指标对比图

4 种诊断方法的高低错判情况如图 5 所示. 若只考虑每个条件下的数据结果,会出现偶然情形,故呈现结果为所有情况下 4 种方法的高低错判的均值. 由图 5 可知: (i) 4 种诊断方法在对被试错误分类时,高低错判情况有略微差距,但并不悬殊. (ii) 4 种诊断方法在高低错判情况下表现各异,KNN 在诊断错误时 3 种判别情况比例基本相同,EDD 在诊断错误时较少地把被试诊断为错判,MDD 在诊断错误时较多地把被试诊断为错判,而 PNN 在诊断错误时较多地把被试诊断为低判.

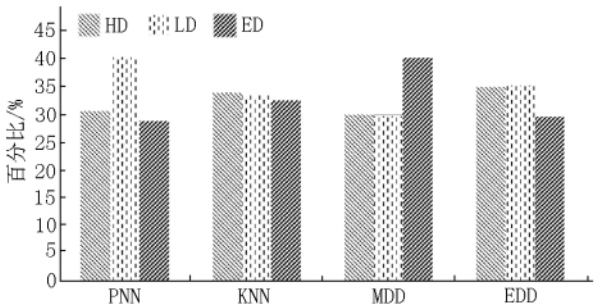

图 5　4 种认知诊断方法高判、低判、错判对比图

3 参数诊断方法的新旧指标比较

3.1 研究目的

与上一节研究类似,本节通过比较 DINA 模型在 PMR、AAMR、TPMR、TAAMR 这 4 种新旧指标上的情况,探讨 TPMR、TAAMR 与 PMR、AAMR 相比的优势,以及在评判认知诊断方法时采用 TPMR、TAAMR 指标的可行性与必要性,并分析在诊断错误时,DINA 模型更倾向于高判、低判还是错判.

3.2 实验设计

实验设计基本沿用上一节研究的设计,采用6 × 5 ×8 的 3 因素被试间实验设计,自变量分别为滑动概率(0.05,0.10,0.15,0.20,0.25,0.30)、题量(10,15,20,25,30)、被试人数(100,200,300,500,1 000,2 000,3 000,5 000)和判别方法(DINA 模型).每个实验条件重复进行 30 次.采用 PMR、AAMR、TPMR、TAAMR 以及 HD、LD、ED 作为诊断方法的评价指标.

3.3 实验流程

实验流程整体思路与上一节研究类似,有别之处在于:(i) 本节的研究因采用参数方法 DINA 模型,故采用 0-1 计分方式模拟数据;(ii) 模拟数据过程不同于上一节研究,上一节的研究采用滑动数据模拟方法进行模拟,本节研究采用 DINA 模型数据模拟方法,即设定失误参数与猜测参数进行数据模拟.其余步骤一致,具体流程如下:

第 1 步　采用与上一节研究相同的方式生成被试的属性掌握模式与测验 $\boldsymbol{Q}$ 矩阵.

第 2 步　根据属性个数与属性层级结构确定所有被试可能的 KS,并与测验 $\boldsymbol{Q}$ 矩阵相乘得 IRP 矩阵,该 IRP 为多级计分,将多级计分转化为 0-1 计分,以在每一题上得满分的被试转为 1 分,其余情况皆为 0 分的原则进行转化,得到 0-1 计分 IRP.再模拟作答反应数据,作答反应数据生成的具体步骤为:首先,利用设定的失误参数和猜测参数,根据 DINA 模型公式计算出每一个得分产生滑动的概率,得概率矩阵 $\boldsymbol{p}$;然后产生一个服从均匀分布 $U(0,1)$ 的随机数矩阵,将概率矩阵 $\boldsymbol{p}$ 与随机数矩阵进行比较,小于随机数则发生滑动,大于则不变,得到模拟被试的 ORP.

第 3 步　使用参数方法 DINA 模型对被试的 ORP 进行判别,并与真值进行比较,得到 4 个评价指标 PMR、AAMR、TPMR、TAAMR,并通过分析错误判别的被试,得出 HD、LD、ED 的比例.

3.4 实验结果

由于数据结果较多,但前人研究表明参数方法受被试人数影响较大,对被试人数的需求一般需要在 1 000 人以上[14,18],故表 4 列出了在被试人数分别为 100、500、1 000、5 000 人时 PMR 与 TPMR 的统计结果.由表 4 可以看出,在不同滑动概率、题量和人数情况下,TPMR 指标一直低于 PMR 指标,且当滑动概率越小、题目数量越少时,2 个指标之前的差异越大,这一结果与上一节研究结果基本一致,这说明 TPMR、TAAMR 指标在参数方法中同样适用.

具体 PMR、TPMR,AAMR、TAAMR 之间的对比结果如图 6 ~ 图 9 所示.4 张图分别为不同人数情况下各水平 DINA 模型评判结果的新旧指标对比.

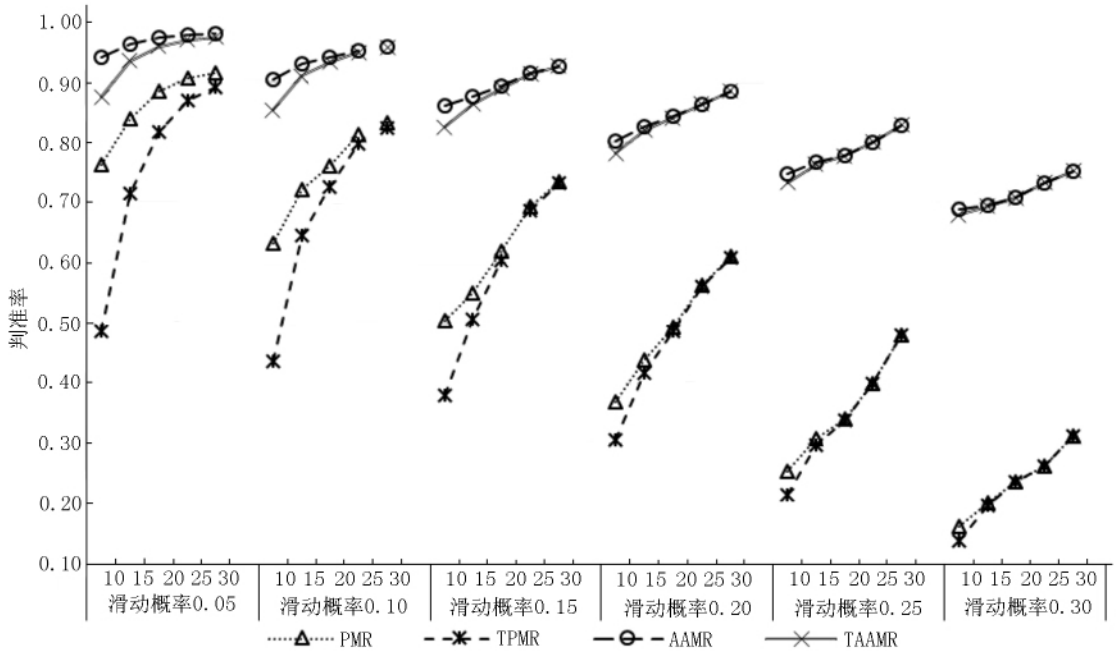

图 6　DINA 模型新旧指标对比图(被试人数为 100 人)

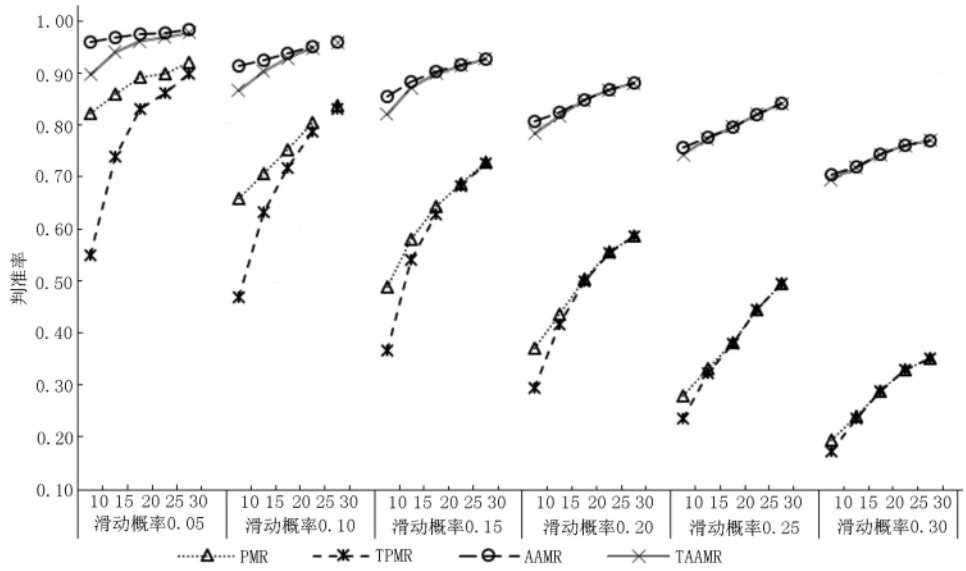

图 7　DINA 模型新旧指标对比图(被试人数为 500 人)

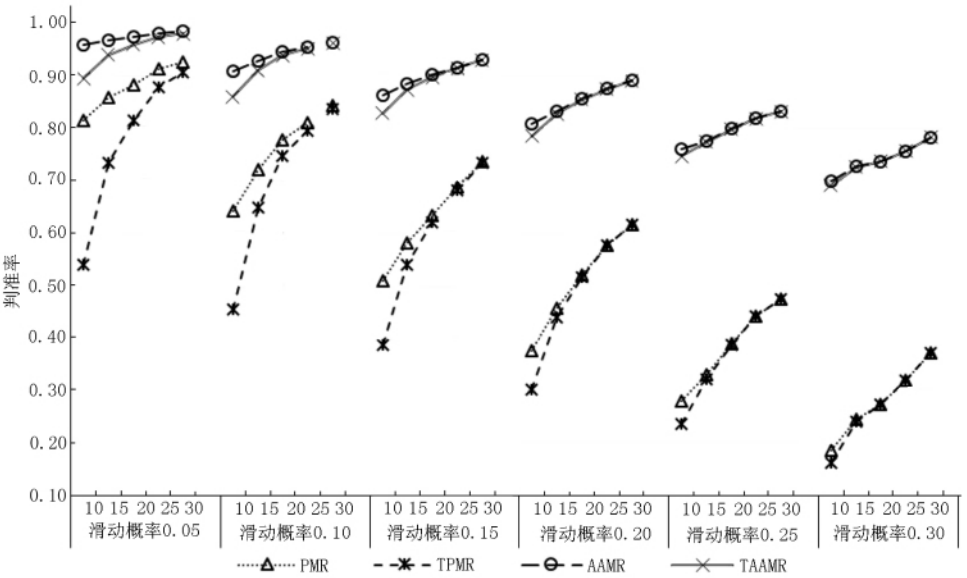

图 8　DINA 模型新旧指标对比图(被试人数为 1 000 人)

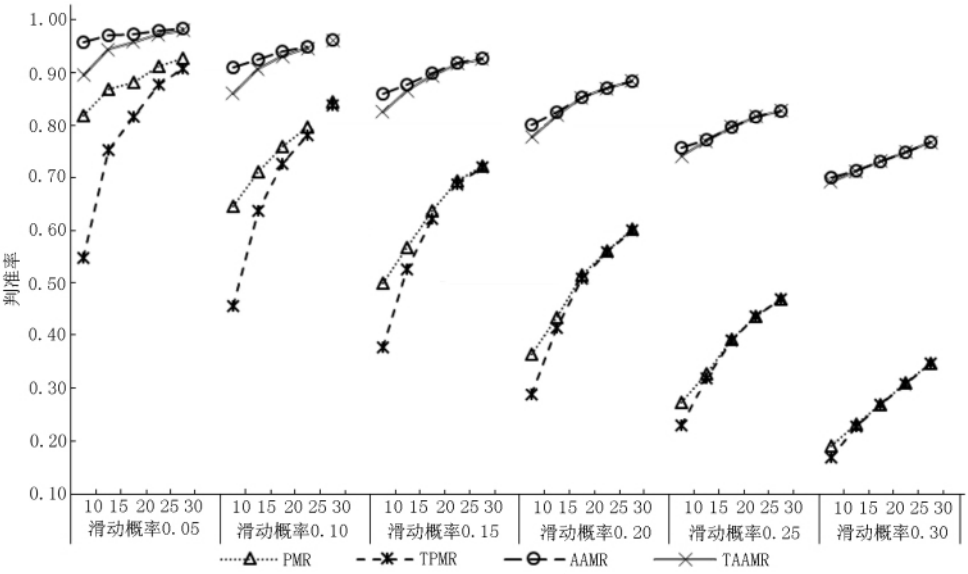

图 9　DINA 模型新旧指标对比图(被试人数为 5 000 人)

由图 6 ~ 图 9 可知:(i) DINA 模型受滑动概率影响较大,滑动概率越大,评判结果越差;而当滑动概率较小时,TPMR 与 PMR 指标之间的差异越大,这一结果与上一节研究结果一致.(ii) DINA 模型受题目数量影响较大,题目数量越多,评判结果越好;相比于 PMR,TPMR 受题目数量的影响更大,即随着题目数量的变化,TPMR 的变化比 PMR 更大,这一结果同样与上一节研究结果一致.(iii) 被试人数对 DINA 模型判别造成影响,但并不影响新旧指标之间的差值. 在 4 张图中同样呈现了各水平情况下的 TAAMR、AAMR,大致趋势与 TPMR、PMR 一致.

DINA模型的高判、低判、错判情况如图10所示. 与上一节研究相同,本节研究呈现结果为所有情况下 4 种方法的高判、低判、错判情况的均值. 由图 10 可得一个主要结果是,相比于低判和错判,DINA 模型更倾向于高判.

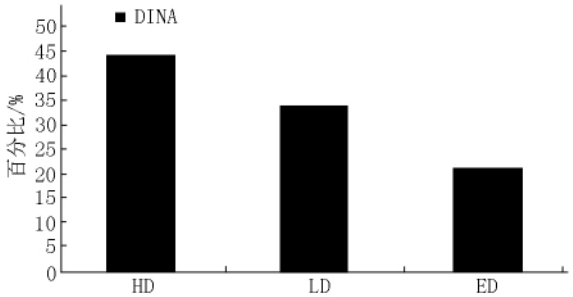

图 10 DINA 模型高判、低判、错判对比图

表 4 不同条件下 DINA 模型的 TPMR 与 PMR 比较

滑动概率	题量	DINA(100 人)		DINA(500 人)		DINA(1 000 人)		DINA(5 000 人)	
		PMR	TPMR	PMR	TPMR	PMR	TPMR	PMR	TPMR
0.05	10	0.76	0.49	0.82	0.55	0.81	0.54	0.82	0.55
	15	0.84	0.72	0.86	0.74	0.86	0.73	0.87	0.75
	20	0.89	0.82	0.89	0.83	0.88	0.81	0.88	0.81
	25	0.91	0.87	0.90	0.86	0.91	0.88	0.91	0.88
	30	0.92	0.89	0.92	0.90	0.92	0.90	0.93	0.91
0.10	10	0.63	0.44	0.66	0.47	0.64	0.45	0.65	0.46
	15	0.72	0.65	0.71	0.63	0.72	0.65	0.71	0.64
	20	0.76	0.73	0.75	0.72	0.78	0.75	0.76	0.73
	25	0.81	0.80	0.80	0.79	0.81	0.79	0.80	0.78
	30	0.83	0.83	0.84	0.83	0.84	0.84	0.84	0.84
0.15	10	0.50	0.38	0.49	0.37	0.51	0.39	0.50	0.38
	15	0.55	0.51	0.58	0.54	0.58	0.54	0.57	0.53
	20	0.62	0.60	0.64	0.63	0.63	0.62	0.64	0.62
	25	0.69	0.69	0.69	0.68	0.69	0.68	0.69	0.69
	30	0.74	0.73	0.73	0.73	0.74	0.73	0.72	0.72
0.20	10	0.37	0.31	0.37	0.30	0.38	0.30	0.37	0.29
	15	0.44	0.42	0.44	0.42	0.46	0.44	0.44	0.41
	20	0.49	0.49	0.50	0.50	0.52	0.51	0.51	0.51
	25	0.56	0.56	0.56	0.55	0.58	0.58	0.56	0.56
	30	0.61	0.61	0.59	0.59	0.62	0.62	0.60	0.60
0.25	10	0.25	0.21	0.28	0.24	0.28	0.24	0.27	0.23
	15	0.31	0.30	0.33	0.32	0.33	0.32	0.33	0.32
	20	0.34	0.34	0.38	0.38	0.39	0.39	0.39	0.39
	25	0.40	0.40	0.45	0.44	0.44	0.44	0.44	0.44
	30	0.48	0.48	0.50	0.50	0.47	0.47	0.47	0.47
0.30	10	0.16	0.14	0.20	0.17	0.19	0.16	0.19	0.17
	15	0.20	0.20	0.24	0.24	0.24	0.24	0.23	0.23
	20	0.24	0.24	0.29	0.29	0.27	0.27	0.27	0.27
	25	0.26	0.26	0.33	0.33	0.32	0.32	0.31	0.31
	30	0.31	0.31	0.35	0.35	0.37	0.37	0.35	0.35

4　讨论

4.1　TPMR、TAAMR 比 PMR、AAMR 更为敏感

在不同实验条件下,无论选用参数方法还是非参数方法,采用滑动模拟数据方法还是 DINA 模型模拟数据方法,TPMR、TAAMR 都比 PMR、AAMR 更为敏感. 如在非参数方法中,当滑动概率为 0.05,被试人数为 1 000 人时,题量由 10 升至 30,PNN 诊断方法的 PMR 变化为 0.09,而 TPMR 变化为0.23,其余 3 种非参数诊断方法情况类似; 在参数方法 DINA 模型中,当滑动概率为 0.05,被试人数为1 000人时,题量由 10 升至 30,DINA 模型判别结果的 PMR 变化为 0.11,而 TPMR 变化为 0.36,其余人数情况也类似. 综合上述情况可知,TPMR 指标比 PMR 指标更能反映认知诊断方法的有效性,可以达到避免高估认知诊断方法的效果.

4.2　判准率高的方法在新旧指标上的变化更小

在 PNN、KNN、MDD、EDD 4 种非参数方法与 DINA 模型一种参数方法中,MDD 的判准率最高,PNN、KNN、EDD 略次之,DINA 模型最低; 而这些方法在新旧指标上的表现也各不相同,MDD 变化也最小. 如在 0.05 滑动概率题目数为 10 的条件下,MDD 新旧指标的变化为 0.10,PNN、KNN、EDD 新旧指标的变化约为 0.15,而相同条件下 DINA 模型新旧指标的变化为 0.27. 综合上述情况可知,选用 TPMR、TAAMR 可以更有效地甄别诊断方法的优劣,好的方法受到的影响较小,而差的方法受到的影响较大.

4.3　TPMR、TAAMR 在不同条件下表现较为稳定

由 2 个模拟实验可知,TPMR、TAAMR 在不同人数、题量、滑动概率条件下,分别使用 4 种非参数诊断方法和一种参数诊断方法,其结果表现均较为稳定,即不同条件对新旧指标的影响趋势基本一致,而 TPMR、TAAMR 会让好的认知诊断方法表现更好,差的表现更差.

4.4　不同认知诊断方法的高低错判情况各不相同

在 4 种非参数诊断方法中,高判、低判、错判情况下表现虽相差不大,但细看其具体比例还是略有差异的,其中 KNN 在诊断错误时高低错判比例基本相同,EDD 较少地把被试诊断为错判,MDD 较多地把被试诊断为错判,而 PNN 较多地把被试诊断为低判; 在 DINA 模型的高低错判结果中,高判比例最高,其次是低判,错判比例最低. 结合实践,在对学生进行诊断时,一个基本的理念是宁可低判也不可高判或错判,因此比较这些判错的情况,MDD 低判比例最低为 0.30,KNN 为 0.33,EDD 为 0.35,DINA 模型为 0.34,PNN 比例最高(为 0.40) . 因此,从这一角度来看,非参数方法 PNN 相比较于其他几种方法更适于实践.

5　结论

本文提出了 2 个真实判准率指标: TPMR 和 TAAMR,用于替代现有评判认知诊断方法有效性的指标 PMR 与 AAMR,分析新旧指标的区别,并比较参数和非参数方法在新旧指标上的表现以及高判、低判、错判情况,为认知诊断的实践提供理论依据. 本文先构建 2 个指标并定义了 3 种错判情况,随后通过 2 个模拟研究对比了新旧指标,得到以下结论: (i) 无论参数方法还是非参数方法,TPMR、TAAMR 相比于 PMR、AAMR 而言,更能真实地反映认知诊断方法的有效性. (ii) 无论参数方法还是非参数方法,TPMR、TAAMR 在不同条件下表现均较为稳定,相比 PMR、AAMR 更能区分认知诊断方法的优劣. (iii) 参数方法的高低错判情况比例较为不均衡,其中高判比例最高,低判次之,错判最少; 而非参数方法的高低错判结果虽然有略微差异,但整体较为平均,结合实际,在对学生进行认知诊断时,为使得学生更好的发展,一般采用宁可低判也不高判或错判,从这一点来看,非参数方法中的 PNN 更适合于实践应用.

6　参考文献

[1] 辛涛,乐美玲,张佳慧. 教育测量理论新进展及发展趋势 [J]. 中国考试,2012(5) : 3-11.

[2] 汪文义,丁树良,宋丽红. 认知诊断中基于条件期望的距离判别方法 [J]. 心理学报,2015,47(12) : 1499-1510.

[3] 康春花,杨亚坤,曾平飞. 一种混合计分的非参数认知诊断方法: 曼哈顿距离判别法 [J]. 心理科学,2019,42(2) : 455-462.

[4] 康春花,张淑君,李元白,等. KNN 认知诊断法及其应用 [J]. 江西师范大学学报: 自然科学版,2019,43(2) : 29-35,53.

[5] 祝玉芳,王黎华,丁树良,等. 多策略的多级评分认知诊断方法的开发 [J]. 江西师范大学学报: 自然科学版,2015,39(4) : 371-376.

[6] 蔡艳,涂冬波. 基于属性层级关系的 rRUM 模型优化: 模型解释力及判准率的提升视角 [J]. 江西师范大学

学报: 自然科学版,2016,40(1):47-55.
[7] 李娟,丁树良,罗芬. 基于等级反应模型的广义距离判别法 [J]. 江西师范大学学报: 自然科学版,2012,36(6):636-639.
[8] 罗欢,丁树良,汪文义,等. 属性不等权重的多级评分属性层级方法 [J]. 心理学报,2010,42(4):528-538.
[9] 张淑梅,包钰,郭文海. 一种多级评分的广义认知诊断模型 [J]. 心理学探新,2013,33(5):444-450.
[10] Junker B W,Sijtsma K. Cognitive assessment models with few assumptions,and connections with nonparametric item response theory [J]. Applied Psychological Measurement,2001,25(3):258-272.
[11] Templin J L,Henson R A. Measurement of psychological disorders using cognitive diagnosis models [J]. Psychol Methods,2006,11(3):287-305.
[12] De la Torre J D L. The generalized DINA model framework [J]. Psychometrika,2011,76(2):179-199.
[13] Ma W,de la Torre J. A sequential cognitive diagnosis model for polytomous responses [J]. British Journal of Mathematical Statistical Psychology,2016,69(3):253-275.
[14] 涂冬波,蔡艳,戴海琦,等. 一种多级评分的认知诊断模型: P-DINA 模型的开发 [J]. 心理学报,2010,42(10):1011-1020.
[15] 蔡艳,赵洋,刘舒畅,等. 一种优化的多级评分认知诊断模型 [J]. 心理科学,2017,40(6):1491-1497.
[16] 涂冬波,蔡艳,戴海琦. 基于 HO-DINA 模型的多级评分认知诊断模型的开发 [J]. 心理科学,2013,36(4):984-988.
[17] 蔡艳,涂冬波,丁树良. 五大认知诊断模型的诊断正确率比较及其影响因素: 基于分布形态、属性数及样本容量的比较 [J]. 心理学报,2013,45(11):1295-1304.
[18] Chiu C Y,Douglas J. A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns [J]. Journal of Classification,2013,30(2):225-250.
[19] 康春花,任平,曾平飞. 非参数认知诊断方法: 多级评分的聚类分析 [J]. 心理学报,2015,47(8):1077-1088.
[20] 康春花,任平,曾平飞. 多级评分聚类诊断法的影响因素 [J]. 心理学报,2016,48(7):891-902.

The New Index to Evaluate the Effectiveness of Cognitive Diagnosis: True Accuracy

KANG Chunhua,ZHU Shihao,GONG Wei,YU Xiangjun,ZENG Pingfei*
(College of Teacher Education,Zhejiang Normal University,Jinhua Zhejiang 321004,China)

Abstract: Two kinds of real equality rate that are TPMR and TAAMR are put forward,which are used to replace the existing evaluation effectiveness index PMR and AAMR cognitive diagnosis methods,and analyze the difference between the old and the new index,then compare parameters and nonparametric method on the performance of the old and the new indicators and the cases of high discriminant,low discriminant and error discriminant,as to provide theoretical basis for the practice of cognitive diagnosis. The findings are as follows. Compared with PMR and AAMR,TPMR and TAAMR can more truly reflect the effectiveness of cognitive diagnosis. The performances of TPMR and TAAMR are relatively stable under different conditions,and compared with PMR and AAMR,they can better distinguish the advantages and disadvantages of cognitive diagnosis methods. From the perspective of high discriminant,low discriminant and error discriminant,PNN is more suitable for application in practice.

Key words: true accuracy rate; method validity; high discriminant,low discriminant and error discriminant

(责任编辑: 冉小晓)