

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*

蔡 艳¹ 涂冬波¹ 丁树良²

(¹江西师范大学心理学院, 南昌 330022) (²江西师范大学计算机信息工程学院, 南昌 330022)

摘 要 本研究主要在 Leighton, Gierl 和 Hunka (2004) 的四种属性层级结构测验情景下, 考察不同被试知识状态(knowledge states)分布形态、不同样本容量和不同认知属性数 3 种实验条件下, 分别比较、分析五种常用认知诊断模型的属性诊断正确率(含边际判准率和模式判准率)及其影响因素, 从而深入探讨每种模型的计量性能及模型属性诊断正确率的影响因素等, 试图为实际应用在模型比较与选用上提供借鉴和指导。

关键词 认知属性; 认知诊断; 认知诊断模型; 属性诊断正确率

分类号 B841

1 引言

认知诊断(cognitive diagnosis, CD)技术是认知心理学与心理计量学相结合的产物, 其目的是探讨被试内部的心理加工机制, 诊断被试当前的认知状态, 找出被试的认知优势与不足, 为个体的因材施教及有针对性地开展补救提供服务。当今, 认知诊断受到了研究者和实践者们的广泛关注, 到 2008 年止就已开发出 60 多种认知诊断模型(Fu & Li, 2008)。

目前被国内外研究者和应用者广泛运用的认知诊断模型主要有: Tatsuoaka (1995, 2009) 的规则空间模型(Rule Space Model), Junker 和 Sijtsma (2001)的 DINA 模型, Leighton, Gierl 和 Hunka (2004)以及 Leighton 和 Gierl (2007)的属性层次模型(Attribute Hierarchy Model), Hartz, Roussos 和 Stout (2002)的融合模型(Fusion Model), 以及 von Davier (2005)的 GDM (general diagnosis model)等等。在国内, 研究者在上述模型的基础上进一步开发了若干新模型, 如孙佳楠、张淑梅、辛涛和包珏 (2011)在 AHM 的基础上开发了 GDD 模型, 并提高

了诊断正确率; 涂冬波、蔡艳和戴海琦(2013)开发了基于属性层级关系的 DINA 模型即 DINA_HC 模型, 该模型弥补了 DINA 模型因在诊断过程中不考虑属性间的层级关系而会估计出不合属性层级关系的知识状态的不足, 进一步增强了认知诊断结果的可解释性及属性诊断准确率。总之, 国内外学者为在实践中提供科学、准确的计量模型进行了大量探索, 开发了大量认知诊断模型供研究者及实践应用者选用。然而出现这么多认知诊断模型, 也给实践应用者和研究者带来一些困惑: 每个模型有什么性能? 模型的诊断准确率如何? 如何选用模型? 等一系列涉及模型比较的问题。对于这些问题的探讨, 国内外研究的较少, 且目前关于认知诊断模型的比较研究也存在一定的缺陷: 如陈德枝、戴海琦和赵顶位(2009)只比较了规则空间模型与属性层级模型的判准率, 研究中涉及的模型过少; De la Torre, Hong 和 Deng (2010)主要探讨了 DINA 模型参数估计的准确性及诊断准确率的影响因素, 同样研究中只关注了 DINA 一个模型。涂冬波等(2013)从属性层级关系入手, 比较分析了属性层级关系正误两种条件下不同模型的诊断正确率, 研究中模型

收稿日期: 2013-02-04

* 教育部人文社科基金(11YJC190002), 国家自然科学基金(31100756, 31160203), 高等院校博士点基金项目(20103604120001, 20123604120001), 江西省教育科学规划项目(13YB029, 12YB088), 江西师范大学青年英才培育资助计划项目资助。

通讯作者: 涂冬波, E-mail: tudongbo@aliyun.com

比较的条件比较单一,只考虑属性层级关系这一实验因素,因此研究结论的推广性存在一定的局限。故未来研究应从多个更为广泛角度对于多个模型进行综合分析比较,这对于实践应用者具有更重要的参考意义和借鉴价值。

有研究(孙佳楠等,2011;涂冬波,蔡艳,戴海琦,丁树良,2011)表明,认知诊断中属性层级关系的类型、失误比例、属性数、样本容量等因素均会不同程度地影响着认知诊断的效果和效率。本研究拟在前人研究基础上,重点探讨多种实验条件下,综合考察不同认知诊断模型的诊断准确率及其影响因素。采用国际上常用的 Monte Carlo 模拟方法,在不同实验条件下,比较当前国内外几种运用较广泛的认知诊断模型(RSM, AHM、GDD 模型, DINA 模型及 DINA_HC 模型)的属性诊断准确率,同时探查其影响因素,为相关认知诊断模型的选择提供参考与借鉴,为实际工作者服务。

2 研究设计

2.1 研究目的与内容

本研究主要在 Leighton 等 (2004)的四种属性层级结构(attribute hierarchy structure)测验情景下,考察不同被试知识状态分布形态(实验一)、不同样本容量(实验二)和不同认知属性数(实验三) 3 种实验情景,分别分析比较 5 种常用认知诊断模型属性诊断正确率,从而深入探讨每种模型的计量性能和模型属性诊断正确率的影响因素等,为实际应用者在模型比较与选用上提供参考。

2.2 涉及的认知诊断模型

主要比较 5 种常用的认知诊断模型,分别为 RSM、AHM、GDD 方法(孙佳楠等,2011)、DINA 模型以及 DINA_HC 模型(涂冬波等,2013)。

2.3 属性层级结构的类型

本研究采用 Leighton 等 (2004)提出的四种基本的属性层级关系结构——线型、收敛型、分支型和无结构型,详见图 1。图 1 是 8 个认知属性的层级关系类型,如果测验中只含 6 个属性,则只需将图 1 中的认知属性 A7 和 A8 去掉即可,其余类推。

2.4 评价指标

采用属性的边际判准率(Marginal Match Ratio, MMR)和模式判准率(Pattern Match Ration, PMR),这两个指标评价不同认知诊断模型的诊断准确率。

3 实验一:被试知识状态的分布形态对诊断准确率的影响

3.1 实验目的

该实验主要在控制其它因素的条件下,重点比较不同知识状态(knowledge states)分布形态(见图 2,即每种属性掌握数的人数分布形态)对 5 种诊断模型诊断正确率的影响。为了便于说明问题,被试数定为 1000 人,测验项目数定为 60,认知属性数为 6,考察 4 种基本的属性层级关系。

3.2 实验数据模拟过程

用 Monte Carlo 模拟方法进行,具体过程为:

(1)被试知识状态真值的模拟

研究中被试知识状态的分布形态主要分四种:正态分布、正偏态分布、负偏态分布和均匀分布。具体模拟过程为,将被试掌握属性个数的分布分别定义为以上四种分布形态,从而确定每种分布形态下每种属性掌握个数的人占总人数百分比,同时对掌握属性个数相同的不同知识状态平均分配人数。

(2)测验 Q 矩阵的模拟

为了便于比较和结果的一般化,在四种属性层级关系下,测验 Q 矩阵随机生成,其中每个属性被考核的概率均为 0.5 且每个属性独立模拟,但对于模拟出的项目考核模式若不符合属性层级关系的,则重新模拟直至符合属性层级关系,且模拟时 Q 矩阵中包含 R 矩阵。

(3)被试作答反应矩阵模拟

对于每个实验,根据步骤(1)生成的被试掌握模式真值及步骤(2)的测验 Q 矩阵,在没有任何猜测及失误的情况下,模拟被试在测验项目上的理想作答;然后采用 Leighton 等(2004)的模拟方法,在理想作答基础上,模拟作答反应失误概率(即 slip)分别为 5%, 10%, 15%, 20%四种的情况下被试的作答反应矩阵。

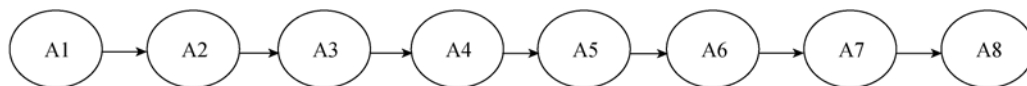
(4)每种测验情景均重复实验 30 次。

3.3 研究结果

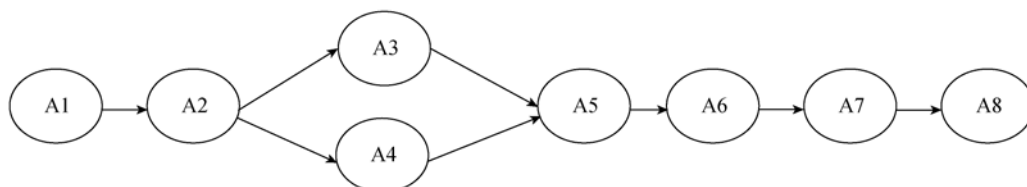
表 1 为正态分布、正偏态分布、负偏态分布和均匀分布四种知识状态分布下,5 种认知诊断模型诊断在四种属性层级关系下的属性诊断正确率(含 MMR 和 PMR),其中表 1 中数据均是 5%, 10%, 15%和 20%四种 Slip 条件下的平均判准率(表 2 和表 3 同)。

表 1 和图 3 可知,5 个模型诊断正确率会受到知识状态分布形态的影响,其中对于 RSM 而言,不

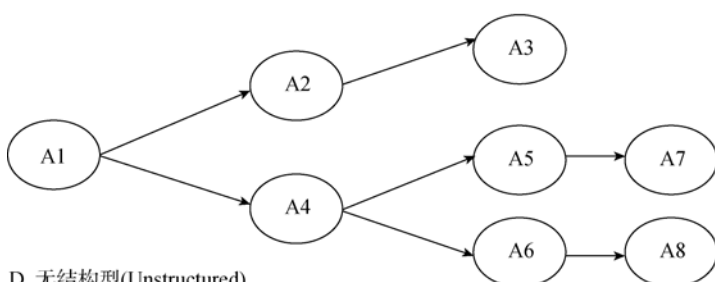
A. 线型(Linear)



B. 收敛型(Convergent)



C. 分支型(Divergent)



D. 无结构型(Unstructured)

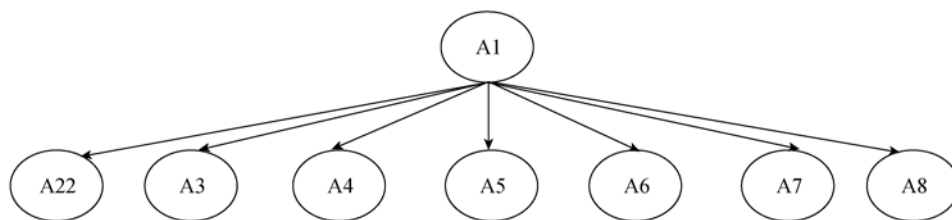


图 1 四种基本的属性层级关系(Leighton, et al., 2004)

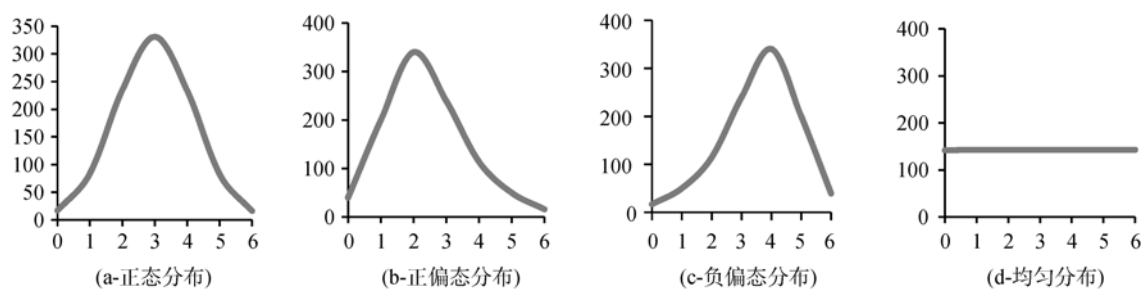


图 2 知识状态的四种分布形态(1000 人 6 个认知属性)

备注: 图中横坐标代表认知状态的属性个数, 纵坐标为人数。

同知识状态分布下整体诊断正确率(即将四种属性层级关系平均处理, 下同)由高到低的分别是: 负偏态、正偏态、正态、均匀分布; 对于 AHM 和 DINA 模型, 整体诊断正确率由高到低的分别是: 负偏态、正态、正偏态、均匀分布; 对于 GDD 和 DINA_HC 模型, 整体诊断正确率由高到低的分别是: 负偏态、

均匀、正态、正偏态分布; 也即对于不同模型, 知识状态的分布特征对不同模型诊断正确率的影响各不相同, 但对 5 种模型而言, 当知识状态为负偏态时, 5 个模型的诊断正确率均达最高, 这可能与负偏态时绝大部分被试为高能力(即掌握更多的属性数)有关。表 1 还可知, 对于同一模型, 在不同属

表 1 不同知识状态分布形态下 5 种个模型的属性诊断正确率(30 次实验的平均)

分布形态	属性层级 关系	RSM		AHM_A		GDD		DINA_HC		DINA	
		MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR
正态分布	线型	96.64	81.04	98.44	91.6	99.53	97.26	99.57	98.03	99.57	97.53
	收敛型	91.37	60.66	97.2	85.51	99.43	96.68	99.58	97.97	99.58	97.6
	分支型	80.15	34.89	96.3	83.99	97.82	88.5	98.65	93.82	98.65	93.01
	无结构型	64.79	13.24	90.46	67.25	93.89	72.6	93.96	79.61	93.96	76.25
	Mean	83.24	47.46	95.6	82.09	97.67	88.76	97.94	92.36	97.94	91.1
正偏态分布	线型	95.12	72.83	98.17	90.01	99.53	97.28	99.67	98.03	99.67	98.11
	收敛型	92.65	64.87	97.15	85.18	99.46	96.83	99.56	97.97	99.56	97.44
	分支型	82.93	41.36	94.81	77.52	97.78	87.88	98.44	91.41	98.44	92.39
	无结构型	69.03	16.04	87.44	56.22	93.86	71.46	94.19	76.26	94.19	73.9
	Mean	84.93	48.78	94.39	77.23	97.66	88.36	97.97	90.92	97.97	90.46
负偏态分布	线型	96.27	80.59	98.41	91.65	99.6	97.71	99.55	98.45	99.55	97.38
	收敛型	92.38	64.54	97.01	84.6	99.55	97.37	99.56	98.15	99.56	97.45
	分支型	78.29	36.82	97.1	87.55	98.3	91.18	98.84	95.32	98.84	94.12
	无结构型	67.13	18.62	93.79	77.82	96.05	81.4	96.56	87.89	96.56	85.72
	Mean	83.52	50.14	96.58	85.41	98.38	91.92	98.63	94.95	98.63	93.67
均匀分布	线型	88.67	54.9	95.63	77.11	99.69	98.25	99.71	98.63	99.71	98.33
	收敛型	98.38	56.02	94.13	71.69	99.51	97.1	98.98	96.8	98.98	93.97
	分支型	72.12	32.37	90.46	66.03	97.99	89.24	97.59	92.03	97.59	88.44
	无结构型	64.74	22.19	85.1	57.87	95.53	79.84	93.79	84.14	93.79	76.52
	Mean	80.98	41.37	91.33	68.18	98.18	91.11	97.52	92.9	97.52	89.32

注：单位为%

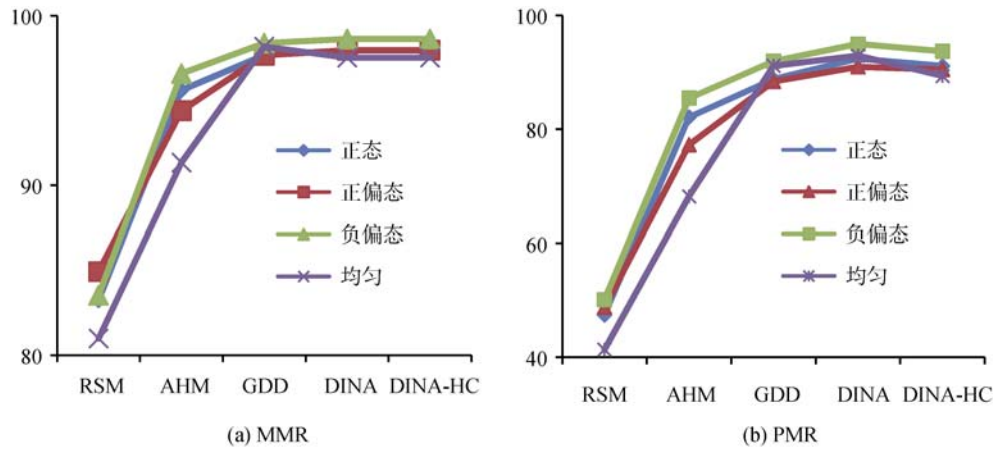


图 3 知识状态分布形态对诊断正确率(%)的影响

性层级关系下模型的诊断准确率也不尽相同。如对于 RSM 下,整体上(即将四种属性层级关系平均处理)正偏态分布的诊断正确率高于正态分布的诊断正确率,但当属性层级关系是线型的情况下,正态分布的结果要好于正偏态分布,对于这些现象读者应加以注意,当然限于篇幅本研究更多地是从整体上把握知识状态等因素对诊断结果的影响(研究二

与三同)。同时,属性层级关系的类型均会影响 5 个模型的诊断正确率,在四种知识分布形态下,属性间的逻辑关系越强烈或紧密(如线型层级关系),则模型诊断正确率越高;属性间的逻辑关系越松散(如无结构型层级关系),则模型诊断正确率越低。

表 1 还表明,不论在何种知识状态分布下,诊断正确率最高的模型是 DINA_HC 和 DINA 模型

(PMR 达 90%以上), 其次为 GDD 模型(PMR 达 85%以上), AHM (PMR 达 65%以上)诊断正确率一般, RSM 的诊断正确率最低, 且其 PMR 基本不足 50%。

4 实验二: 样本容量对诊断准确率的影响

4.1 实验目的

采用 3×3 实验设计, 考察被试样本容量和项目样本容量二个因素下, 比较 5 种认知诊断模型的判准率。被试样本容量取 3 个水平: 100, 1000, 5000 人, 分别代表小, 中, 大样本容量; 同时, 项目样本容量也取 3 个水平: 20, 60, 100 题, 分别代表小, 中, 大样本容量, 考察四种基本的属性层级关系。

4.2 实验数据模拟过程

采用 Monte Carlo 模拟方法进行, 知识状态的分布形态固定为正态分布, 认知属性的个数固定为 6 个。其被试知识状态、Q 矩阵及被试得分矩阵的模拟与实验一基本相同, 实验重复 30 次。

4.3 研究结果

表 2 和图 4 可知, 5 个模型的属性诊断正确率会受到样本容量的影响。当被试数固定, 试题数越大时, 则属性诊断正确率越高。当试题数固定, 被试数越大时, 属性诊断准确率也会略增加, 但影响不大。当被试为小样本容量(100)时, 模型的属性诊断正确率由高到低依次为 GDD、DINA_HC、DINA、AHM 和 RSM 模型; 而当被试为中或大样本容量时(1000 或 5000), 模型的属性诊断正确率由高到低则依次为 DINA_HC、DINA、GDD、AHM 和 RSM 模型。值得一提的是: 当试题为小样本容量时(20), DINA_HC、DINA、GDD 和 AHM 四模型的表现相当, 而当试题为中或大样本容量时(60 或 100), 除无结构型属性层级关系, DINA_HC、DINA 的 PMR 均超过了 90%。

表 2 还表明, RSM、GDD、DINA 和 DINA_HC 模型的诊断正确率受试题样本容量的影响较大, 当试题样本容量从小样本增加到中等样本容量再到大样本容量时, 其 PMR 增幅从 15%到 20%, 而 AHM 模型受试题样本容量的影响相对要小, 其 PMR 的最大增幅在 10%左右。因此在实践中, 我们建议在采用 DINA 或 DINA_HC 时, 应保证一定的试题样本容量(如 60 题以上), 以进一步提升该模型的诊断正确率; 若样本容量较小时, 则建议采用 GDD 模型更佳。

同样, 不论在何种样本容量下, 属性层级关系

的类型也会影响 5 个模型的诊断正确率, 属性间的逻辑关系越强烈或紧密(如线型层级关系), 则模型诊断正确率越高; 属性间的逻辑关系越松散(如无结构型层级关系), 则模型诊断正确率越偏低。

5 实验三: 认知属性个数对诊断准确率的影响

5.1 实验目的

该实验主要比较不同认知属性数下, 5 种认知诊断模型的判准率。认知属性的个数分别为 5、6、7 和 8 个, 共 4 个水平, 以充分考察属性数的变化对不同模型诊断正确率的影响。

5.2 实验数据模拟过程

采用 Monte Carlo 模拟方法进行, 知识状态的分布形态为正态分布, 测验项目数为 60, 考察四种基本的属性层级关系。其被试知识状态、被试得分矩阵的模拟和 Q 矩阵与实验一基本相同, 实验重复 30 次。

5.3 研究结果

表 3 和图 5 可知, 5 个模型的属性诊断正确率会受到认知属性个数的影响, 属性个数越多, 其判准率越低。整体来看, 不论在何种认知属性数下, DINA_HC 和 DINA 模型的诊断正确率最高, 其次为 GDD 和 AHM, RSM 最低。同时, 5 个模型诊断准确率受属性个数的影响大小相差不大。同时, 当认知属性数增至 8 个时, DINA_HC 和 DINA 模型的 PMR 仍高达到 85%以上(MMR 高达 97%), 表现出较好的诊断性能。

同样, 不论在何种样本容量下, 属性层级关系的类型也会影响 5 个模型的诊断正确率, 属性间的逻辑关系越强烈或紧密(如线型层级关系), 则模型诊断正确率越高; 属性间的逻辑关系越松散(如无结构型层级关系), 则模型诊断正确率越偏低。

6 研究结论与讨论

到目前为止, 心理测量学者们开发了 60 多种认知诊断模型(Fu & Li, 2008), 每种认知诊断模型各具特点, 那么实际应用者在实践中应如何来选用模型呢? 这个问题一直以来被学界忽视, 因此更多的基础性研究应为实践者在模型选用提供基础、指导和借鉴。本文从不同角度比较了目前国内外几种常用的认知诊断模型, 从计量学的角度(如模型诊断的正确率)为实际应用者及研究者在模型选用及比较上提供依据和基础。

表 2 不同样本容量下 5 种认知诊断模型的属性诊断正确率(30 次实验的平均单位: %)

样本容量	属性层级关系	RSM		AHM_A		GDD		DINA_HC		DINA	
		MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR
100 人/20 题	线型	83.54	38.62	96.54	82.32	96.91	83.7	96.48	82.62	96.46	82.15
	收敛型	85.53	48.77	95.38	78.22	95.94	80.38	95.65	79.07	95.65	79.3
	分支型	74.68	28.2	92.06	68.48	93.05	70.53	92.28	68.28	92.28	68.87
	无结构型	70.99	23.4	88.77	60.37	90.67	62.37	89.35	60.07	89.35	58.85
	Mean	78.69	34.75	93.19	72.35	94.14	74.25	93.44	72.51	93.44	72.29
100 人/60 题	线型	96.35	80.18	98.65	92.43	99.56	97.4	99.35	97.43	99.35	96.22
	收敛型	92.46	66.68	97.35	86.05	99.36	96.27	99.01	95.03	99.01	94.28
	分支型	79.41	33.1	95.38	80.65	97.72	88.07	97.19	91.35	97.19	86.62
	无结构	64.68	11.65	87	57.75	93.79	72.75	92.82	72.97	92.82	71.78
	Mean	83.23	47.9	94.6	79.22	97.61	88.62	97.09	89.2	97.09	87.23
100 人/100 题	线型	97.82	87.4	98.07	89.47	99.95	99.72	99.18	97.33	99.18	95.38
	收敛型	91.89	67.02	97.46	86.47	99.8	98.83	99.16	96.1	99.16	95.25
	分支型	79.62	36.15	96.97	84.96	98.56	91.88	98.85	94.16	98.85	93.91
	无结构	65.84	13.2	89.46	61.47	96.65	82.95	95.14	79.07	95.14	79.17
	Mean	83.79	50.94	95.49	80.59	98.74	93.35	98.08	91.67	98.08	90.93
1000 人/20 题	线型	84.15	38.41	96.49	81.98	96.95	84.04	97.08	85.16	97.08	85.11
	收敛型	86.46	52.84	95.73	79.54	96.2	81.25	96.22	81.55	96.22	81.58
	分支型	73.98	29.6	92.62	70.69	93.59	72.39	93.92	74.27	93.92	74.36
	无结构	71.58	24.78	89.98	63.28	91.35	64.47	91.36	65.95	91.36	65.94
	Mean	79.04	36.41	93.71	73.87	94.52	75.54	94.65	76.73	94.65	76.75
1000 人/60 题	线型	96.64	81.04	98.44	91.6	99.53	97.26	99.57	98.03	99.57	97.53
	收敛型	91.37	60.66	97.2	85.51	99.43	96.68	99.58	97.97	99.58	97.6
	分支型	80.15	34.89	96.3	83.99	97.82	88.5	98.65	93.82	98.65	93.01
	无结构型	64.79	13.24	90.46	67.25	93.89	72.6	93.96	79.61	93.96	76.25
	Mean	83.24	47.46	95.6	82.09	97.67	88.76	97.94	92.36	97.94	91.1
1000 人/100 题	线型	97.73	86.68	97.89	88.74	99.9	99.42	99.51	99.69	99.51	97.28
	收敛型	91.24	63.85	96.94	84.14	99.78	98.7	99.41	99.27	99.41	96.83
	分支型	79.33	35.55	97.19	86.38	98.86	93.53	99.03	97.94	99.03	94.64
	无结构	67.53	16.83	91.77	69.47	96.85	83.83	95.93	89.67	95.93	82.54
	Mean	83.96	50.73	95.95	82.18	98.85	93.87	98.47	96.64	98.47	92.82
5000 人/20 题	线型	84.62	39.3	96.52	82.15	96.94	84	97.12	85.3	97.12	85.31
	收敛型	86.77	53.47	95.76	79.77	96.24	81.37	96.31	81.77	96.31	82.02
	分支型	74.25	30.04	92.61	70.76	93.63	72.39	94	74.72	94	74.8
	无结构	71.36	24.47	90.16	63.88	91.47	64.97	91.53	66.75	91.53	66.71
	Mean	79.25	36.82	93.76	74.14	94.57	75.68	94.74	77.14	94.74	77.21
5000 人/60 题	线型	96.5	80.84	98.62	92.48	99.54	97.32	99.69	98.23	99.69	98.22
	收敛型	91.41	60.9	97.52	87.21	99.46	96.85	99.69	98.3	99.69	98.2
	分支型	80.41	35.37	96.68	85.43	97.78	88.39	98.77	93.66	98.77	93.58
	无结构	65.08	13.45	90.98	68.67	94.04	73.33	95.04	80.38	95.04	79.49
	Mean	83.35	47.64	95.95	83.45	97.71	88.97	98.3	92.64	98.3	92.37
5000 人/100 题	线型	98.15	89.67	98.46	91.46	99.99	99.87	99.99	99.99	99.99	99.98
	收敛型	91.41	65.18	97.26	85.7	99.76	98.58	99.89	99.66	99.89	99.34
	分支型	79.15	35.12	97.36	87.23	98.86	93.55	99.66	98.06	99.66	98.08
	无结构型	68.17	18.04	92.57	72.42	96.87	84.03	97.97	90.06	97.91	90.17
	Mean	84.22	52	96.41	84.2	98.87	94.01	99.38	96.94	99.36	96.89

注: 单位为%

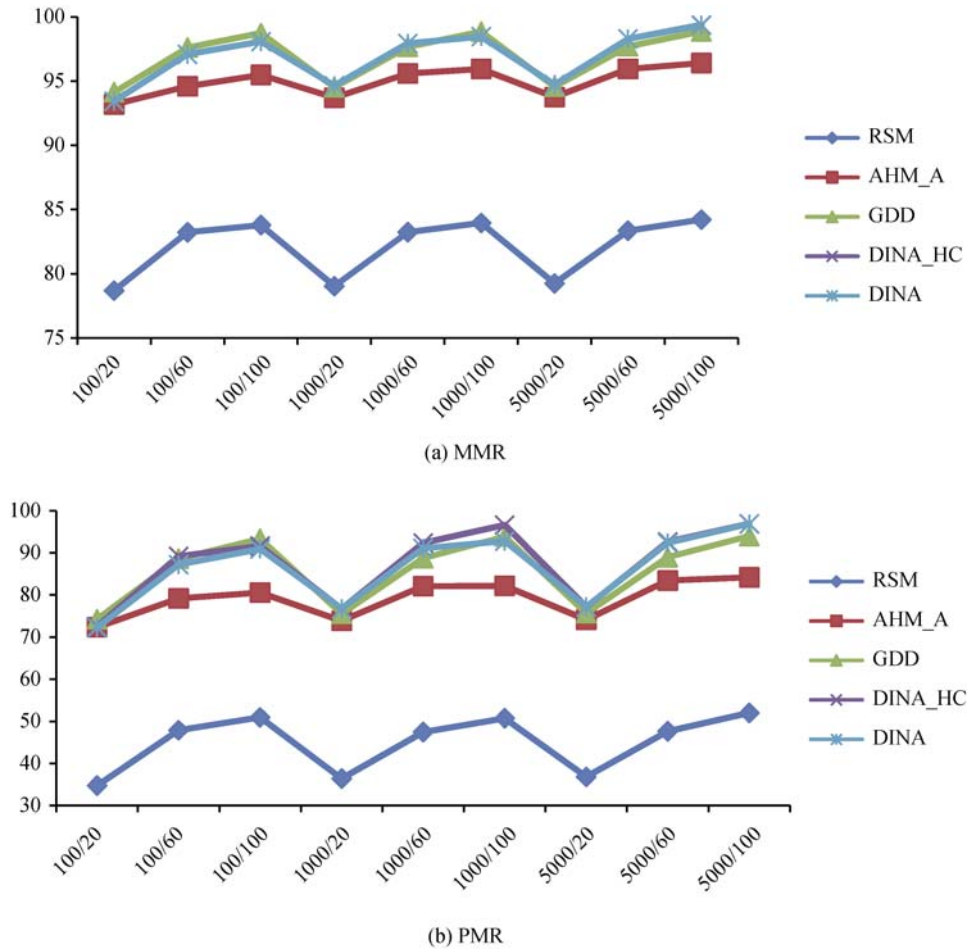


图 4 样本容量对诊断正确率的影响

表 3 不同属性个数下 5 种认知诊断模型的属性诊断正确率(30 次实验平均)

属性 个数	属性层 级关系	RSM		AHM_A		GDD		DINA_HC		DINA	
		MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR	MMR	PMR
5K	线型	95.33	77.45	98.61	93.46	99.67	98.42	99.73	98.87	99.73	98.67
	收敛型	89.45	65.7	98.07	91.14	99.64	98.23	99.72	99.04	99.72	98.61
	分支型	79.41	43.93	97.37	89.07	98.37	92.5	99.22	96.36	99.22	96.38
	无结构型	68.59	18.99	94.6	79.34	97.49	88.57	98.41	93.56	98.41	93.21
	Mean	83.2	51.52	97.16	88.25	98.79	94.43	99.27	96.96	99.27	96.72
6K	线型	96.64	81.04	98.44	91.6	99.53	97.26	99.57	98.03	99.57	97.53
	收敛型	91.37	60.66	97.2	85.51	99.43	96.68	99.58	97.97	99.58	97.6
	分支型	80.15	34.89	96.3	83.99	97.82	88.5	98.65	93.82	98.65	93.01
	无结构型	64.79	13.24	90.46	67.25	93.89	72.6	93.96	79.61	93.96	76.25
	Mean	83.24	47.46	95.6	82.09	97.67	88.76	97.94	92.36	97.25	91.1
7K	线型	95.61	73.23	98.3	89.81	99.6	97.35	99.56	98.17	99.56	97.13
	收敛型	92.31	60.47	97.67	86.38	99.47	96.45	99.35	98.12	99.35	95.91
	分支型	81.01	31.54	96.38	81.29	97.45	85.74	98.05	90.18	98.05	88.95
	无结构型	64.1	7.89	88.57	60.71	92.57	65.68	92	69.91	92	67.95
	Mean	83.26	43.28	95.23	79.55	97.27	86.31	97.24	89.1	97.24	87.49
8K	线型	93.79	57.95	97.88	85.26	98.87	91.34	98.91	92.38	98.91	91.62
	收敛型	92.5	58.73	97.37	83.61	99.48	96.13	99.28	97.35	99.28	94.92
	分支型	70.7	13.72	94.4	74.1	96.5	78.77	96.98	84.78	96.98	82.75
	无结构型	62.59	5.6	90.3	63.14	93.43	65.78	93.55	73.17	93.55	71.59
	Mean	79.9	34	94.99	76.53	97.07	83.01	97.18	86.92	97.18	85.22

注: 单位为%

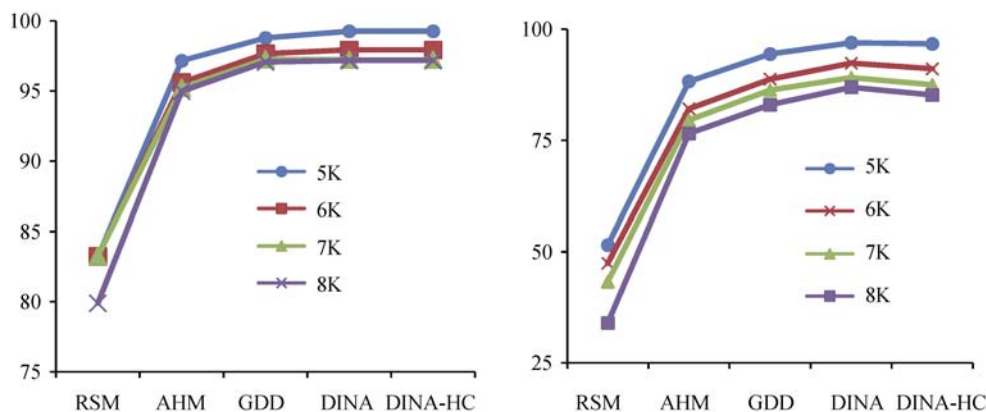


图 5 认知属性数对诊断正确率(%)的影响

通过以上 3 个实验研究, 可以得出以下结论:

(1)知识状态的分布形态会影响 5 个认知诊断模型的诊断正确率, 但对不同模型影响方式不尽相同。对于 RSM 而言, 不同知识状态分布下诊断正确率由高到低的分别是: 负偏态、正偏态、正态、均匀分布; 对于 AHM 和 DINA 模型, 诊断正确率由高到低的分别是: 负偏态、正态、正偏态、均匀分布; 对于 GDD 和 DINA_HC 模型, 诊断正确率由高到低的分别是: 负偏态、均匀、正态、正偏态分布; 也即对于不同模型, 知识状态的分布特征对不同模型诊断正确率的影响各不相同, 但对 5 种模型而言, 当知识状态为负偏态时, 5 个模型的诊断正确率均达最高。同时, 不论在何种知识状态分布下, 诊断正确率最高的模型是 DINA_HC 和 DINA 模型, 其后为 GDD 和 AHM 模型, 而 RSM 模型的诊断正确率最低, 其 PMR 不到 50%。

(2)在样本容量中, 被试样本容量的大小对诊断准确率的影响不大。当被试为小样本容量(100)时, 模型的属性诊断正确率由高到低依次为 GDD、DINA_HC、DINA、AHM 和 RSM 模型; 而当被试为中或大样本容量时(1000 或 5000), 模型的属性诊断正确率由高到低则依次为 DINA_HC、DINA、GDD、AHM 和 RSM 模型。

(3)在样本容量中, 试题样本容量是影响 5 个认知诊断模型的诊断正确率的重要因素, 随着试题样本容量的增加, 属性诊断正确率也在增加, 但对不同模型其增加的幅度不同, 其中 DINA、DINA_HC、GDD 和 RSM 模型的诊断正确率受试题样本容量的影响较大, 而 AHM 的影响相对要小。在小试题样本容量下(20), GDD 模型的属性诊断正确率最高; 而在中等或大试题样本容量下(60 或 100), DINA_HC、DINA 的属性诊断正确率最高

(其 PMR 均超过了 90%)。综上, 在实践中, 我们建议在采用 DINA 或 DINA_HC 时, 应保证一定的试题样本容量, 以进一步提升该模型的诊断正确率。若试题或被试样本容量较小时, 则建议采用 GDD 模型更佳。

(4)认知属性的个数会影响 5 个认知诊断模型的诊断正确率。属性个数越多, 模型的判准率越低。整体来看, 不论在何种认知属性数下, DINA_HC 和 DINA 模型的诊断正确率最高, 其次为 AHM, RSM 最低。同时, 5 个模型诊断准确率受属性个数的影响大小相差不大。

(5)3 个实验均表明: 第一, 属性层级关系的类型也会影响模型的诊断正确率。一般来讲, 属性间的逻辑关系越强烈或紧密(如线型层级关系), 则模型诊断正确率越高; 属性间的逻辑关系越松散(如无结构型层级关系), 则模型诊断正确率越低; 第二, 3 个实验均表明 DINA、DINA_HC 和 GDD 3 个模型整体上优于 RSM 和 AHM。

上述结论还进一步显示: 在实际工作中, 实践工作者可以根据实际情况从 DINA、DINA_HC 和 GDD 3 种模型中选择合适的模型。对这 3 种模型的优势程度做一步的比较分析发现:

(a)在 DINA_HC 与 DINA 模型的边际判准率(MMR)基本一致, 且 DINA_HC 的模式判准率(PMR)略优于 DINA 模型(最大优势为 1.5%);

(b)DINA_HC 和 GDD 都是基于属性层级结构的认知诊断模型, DINA 模型是不含属性阶层结构的认知诊断模型(认知诊断过程中无需用到属性层级结构)。实践工作中, 属性间层级关系错综复杂, 属性层级关系的认定也比较困难, 从而导致属性间层级关系被错误界定的现象时常发生(Tatsuoka, 2009), 因此, 如果属性层级结构难以确定或拿不

准时, 建议选用不含属性层级关系的 DINA 模型。这一研究结论也进一步印证了涂冬波等人(2013)的实验。

(c)如果测验样本容量较小时, 则建议采用 GDD 模型; 若欲在诊断中采用 DINA 或 DINA_HC 模型, 则测验的样本容量应至少达 1000 人。

当然, 由于研究时间及能力, 本研究也存在一定的局限, 本研究更多是考察比较了 0-1 评分的认知诊断模型, 对于多级评分、对于多策略认知诊断模型并未考虑; 同时对于 FM、GDM 及 HO-DINA 等模型并未考察。本研究更多是从计量学特征来比较不同模型的诊断正确率, 未来研究还可进一步比较不同模型项目属性作用机制(即属性是如何影响项目作答的机制, 如属性是否可以补偿); 另外, 当被试知识状态为负偏态时, 其诊断准确率总体来讲相对较高, 但这出现这种原因到底是所采用的认知诊断模型导致, 还是由其它理论原因导致, 则有待未来进一步深入等等, 这些都有待今后进一步完善。

参 考 文 献

- Chen, D. Z., Dai, H. Q., & Zhao, D. W. (2009). Comparing the diagnostic accuracy between the rule-space approach and the attribute hierarchy method. *Psychological Science*, 32(2), 414-416.
- [陈德枝, 戴海琦, 赵顶位. (2009). 规则空间方法与属性层次方法的诊断准确性比较. *心理科学*, 32(2), 414-416.]
- de la Torre, J., Hong, Y., & Deng, W. L. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, 47(2), 227-249.
- Fu, J., & Li, Y. (2008). Cognitively diagnostic psychometric models: An integrative review. *ETS research report*, Princeton, NJ: ETS.
- Hartz, S., Roussos, L., & Stout, W. A. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Illinois, Urbana.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258-272.
- Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and Applications* (pp. 242-274). Cambridge, UK: Cambridge University Press.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41(3), 205-236.
- Sun, J. N., Zhang, S. M., & Xin, T., & Bao, Y. (2011). A cognitive diagnosis method based on Q-Matrix and generalized distance. *Acta Psychologica Sinica*, 43(9), 1095-1102.
- [孙佳楠, 张淑梅, 辛涛, 包珏. (2011). 基于 Q 矩阵和广义距离的认知诊断方法. *心理学报*, 43(9), 1095-1102.]
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-361). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge, Taylor & Francis Group.
- Tu, D. B., Cai, Y., & Dai, H. Q. (2013). Comparison and selection of five noncompensatory cognitive diagnosis models based on attribute hierarchy structure. *Acta Psychologica Sinica*, 45(2), 243-252.
- [涂冬波, 蔡艳, 戴海琦. (2013). 几种常用非补偿型认知诊断模型的比较与选用: 基于属性层级关系的考量. *心理学报*, 45(2), 243-252.]
- Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. (2011). A research on MCMC parameter estimation and the properties of the High Order DINA model. *Journal of Psychological Science*, 34(6), 1476-1481.
- [涂冬波, 蔡艳, 戴海琦, 丁树良. (2011). HO-DINA 模型的 MCMC 参数估计及模型性能研究. *心理科学*, 34(6), 1476-1481.]
- von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS research Report*, Princeton, NJ: ETS.

A Simulation Study to Compare Five Cognitive Diagnostic Models

CAI Yan¹; TU Dongbo¹; DING Shuliang²

(¹ School of Psychology, Jiangxi Normal University, Nanchang 330027, China)

(² School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330027, China)

Abstract

Cognitive diagnosis is an important topic in modern psychometric area. Now more than 70 cognitive diagnosis models (CDMs) are developed. There are some questions among these models: (1) When the attribute hierarchy structure is known, how to choose the most suitable model? (2) When the attribute hierarchy structure is unknown, and cognitive diagnosis is required, how to do it? These problems seems especially more puzzled for the practice workers.

This paper only paid main attention on three international popular models. Therefore, five cognitive diagnosis models (RSM, AHM_A, GDD, DINA and DINA_HC) were compared corresponding to the above questions from psychometric opinion. In this paper, Monte Carlo simulation study was used. Although the number of slips and the hierarchy structure are two important factors that affect the performance on corrected match ratio of cognitive diagnosis, this study would pay attention on other three factors: the distribution of cognitive pattern, the sample size, the number of attributes. The findings identified:

(1) When the characteristic of data was known, focusing on specific factor, the five methods had different advantages.

a) For the distributions of cognitive pattern, although they have different effects on different methods, the same conclusion could find that the performance on negative bias distribution was the best, and that of DINA_HC and DINA were better than the rest methods on any discussed distributions.

b) Considering the sample size, the performance of GDD with small scale assessment (100/20, persons/items) was the best one; with medium and large scale assessment (1000/60, 5000/100, persons/items), the performance of DINA_HC and DINA were better than the rest

c) For the number of attributes, the more the attributes are the worse the performance will be. But for the methods, the performance of DINA_HC and DINA were also better than the rest.

All these reflected that the most suitable method could be adapt from the three methods: GDD, DINA_HC and DINA, corresponding to the real scenario. And the RSM was the worst cognitive diagnosis method.

(2) When the characteristic of data was unknown, an unstructured attribute hierarchy is treated as a coarsened version of a structured one, and the DINA method has a similar performance under unstructured hierarchy with the GDD, DINA_HC methods under structured hierarchy. Thus if the hierarchy structure could not be identified clearly, and the test Q matrix was clear, then the DINA could be adapt.

Key words cognitive attributes; cognitive diagnosis; cognitive diagnosis models; attributes Match Ratio