文章编号:1000-5862(2016)04-0369-08

4 年级数学应用题 Q 矩阵的适宜性

康春花[1,2],杨亚坤[1,2],钟晓玲[1],曾平飞*
(1. 浙江师范大学教师教育学院,浙江 金华　321004;
2. 海云天教育测评有限公司,广东 深圳　518000)

摘要:在认知诊断评估中,Q 矩阵的界定和挑选非常重要,因其关系到诊断测验的质量和诊断评估的准确性. 在模拟研究中,Q 矩阵可以任意设定,但在实践研究中,Q 矩阵的界定和测验 Q 矩阵的选择确非易事. 该研究基于已有理论和模拟研究关于 Q 矩阵选择的原则,以小学 4 年级数学应用题为例,阐述如何在实践认知评估中选择适宜的测验 Q 矩阵,并通过实证和模拟研究验证所选测验 Q 矩阵的适宜性. 研究结果表明:测验 Q 矩阵在包含 R 矩阵的前提下,考核模式并非越多越好、测验长度并非越长越好,相比较而言,只包含 R 矩阵的测验 Q 矩阵均要好于考核模式太多的 Q 矩阵.

关键词:数学应用题;测验 Q 矩阵;R 矩阵;GRM-AHM-A 方法
中图分类号:B 841　　文献标志码:A　　DOI:10.16357/j.cnki.issn1000-5862.2016.04.08

0　引言

认知诊断评估(Cognitive Diagnostic Assessment,CDA)的目的在于对学生的认知结构进行诊断和分类,由此教师可根据学生的知识状态有针对性地实施补救教学. 在 CDA 中,诊断和分类精确性的影响因素众多,其中两点至关重要,即有效的认知诊断测验和适宜的认知诊断模型(Cognitive Diagnostic Model,CDM)[1]. 在认知诊断测验的开发中,Q 矩阵至关重要,Q 矩阵的适宜性是指该 Q 矩阵具有良好的质性和量性特征. 质性方面是指考核模式或题量适中,但同时应该每个属性应该得到多次考察. 量化特征主要指具有较好的测量学指标,如题目效度、结构效度、归类比率等. Q 矩阵是否适宜不仅关系着认知诊断测验的质量,也关系着认知诊断评估的准确性[2-4]. 模拟研究显示,Q 矩阵中属性缺失会高估失误参数或低估被试在剩余属性上的掌握概率,而属性冗余则会高估猜测参数或被试的属性掌握概率[5-6]. 当 Q 矩阵中的元素随机发生 30% 误差及改变 Q 矩阵所含属性个数时,会降低分类判准率[7].

按照 AHM 的思想,属性层级关系可表述为属性之间直接与间接关系的 R 矩阵,采用扩张算法或减法算法,可得到符合层级关系的减缩 Q 矩阵 Q_r,命题专家只能命拟符合这些层级的项目[8]. Q_r 是开发认知诊断测验的项目考核模式,在属性个数较多的情况下,Q_r 的考核模式可能较多,在实际情境中,考虑到学生做题的疲劳效应及测验的组织实施,测验长度不能太长,因此不可能命拟所有的考核模式,只能从中抽取有代表性的考核模式编制认知诊断测验,称之为测验 Q 阵 Q_t. 从 Q_r 中选择 Q_t 的组合方式有较多,怎样的 Q_t 才适宜呢?

J. S. Gorin[9] 强调 Q_t 中,每个项目至少包含 2 个属性. M. Gierl 等[10] 认为 Q_t 应包含 R 矩阵中的所有列,才能保证测验能够反映属性层级. 丁树良等[8,11] 从数理上证明了 Q_t 包含 R 就可实现期望反应模式(Expected Response Patterns,ERP)和属性掌握模式(Attribute Master Patterns,AMP)间的一一对应关系,并且在 Q_t 中包含的 R 矩阵个数越多,其模式判准率越高. 可见,R 矩阵在认知诊断测验编制中具有不可或缺的作用,为保证对被试知识状态的判准率,首要原则是 Q_t 必须包括 R. 然而在实证研究中,Q_t 中放入多个 R 的同时又包含其它模式,这样会造成测验长度太长,造成学生不情愿作答的疲劳效应. 那么 Q_t 中除了 R 矩阵以外的模式是不是任意抽取,其效果均一样呢? 涂冬波等[12] 认为一个完备的 Q_t 应该能够实现对每个属性的多次考察就可以,并非需要穷尽所有的模式. 康春花等[13] 认为还应考虑属性在

收稿日期:2016-04-17
基金项目:浙江省高校重大人文社科项目攻关计划(2013QN048)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.

项目间及项目在属性间的平衡问题.

Q_t是否适宜直接关系着认知诊断测验和认知诊断评估的质量. 然而,已有关于 Q_t选择和适宜性的研究,要么是数理证明要么是模拟研究,并未取得实证数据的支持. 本研究拟以 4 年级数学应用题为例,根据以上原则,选取适宜的 Q_t编制认知诊断测验,并通过实证和模拟 2 种方式验证所选 Q_t的适宜性,为 CDA 补充实证证据. 由于数学认知属性具有一定的逻辑关系及宜采用建构反应题,本研究采用多级计分的 AHM 方法(GRM-AHM-A 方法)[13] 对数据进行处理.

1 4 年级数学应用题测验 Q 阵的确定

采用自上而下的思路,先采用文献法与教材分析、专家小组讨论和学生出声思维[10,12-14] 确定 4 年级应用题的属性及其层级关系,然后按照 Q 矩阵充分性的原则[8,11] 选择测验 Q 矩阵编制认知诊断测验.

1.1 4 年级数学应用题的属性及其层级关系

R. Maye[15] 认为数学应用题解决包括问题表征和问题解答 2 个步骤. M. K. Enright 等[16] 指出影响学生数量推理问题解决过程的因素包括问题情境、数学复杂性(运算步骤数)和项目代数性. G. Arendasyr 等[17-18] 认为数量推理的难度也会受到隐含的数量关系的影响. 此外,国内外研究表明图式表征、图画表征、语词表征是学生问题解决的主要表征方式. 小学 4 年级数学教材中,应用题的典型问题情境有时间、金钱和距离等问题. 而在运算方面,包括基本运算和复杂运算. 对于复杂问题,学生需要识别题中的隐含条件,通过一些辅助方式对问题进行表征. 由此,小学 4 年级数学应用题问题解决包括 2 个认知成分:数学内容属性($A_1 \sim A_5$)和认知过程属性(A_6、A_7),见表 1.

表 1 4 年级数学应用题的认知属性

属性成分	属性	含义描述
数学内容	A_1	基本算术运算,基本的加、减、乘、除运算
	A_2	基本数量关系,如路程 = 速度 × 时间,总价 = 单价 × 数量
	A_3	单位换算,如长度单位、重量单位、时间单位等
	A_4	复杂运算,需采用四则运算法则的多步运算
	A_5	复杂知识图式,如路程 ÷ 速度 = 相遇时间
认知过程	A_6	识别隐含条件,需运用已知量求出另一关键量
	A_7	图式表征,借助线段图等手段理解数量关系

通过分析学生出声思维的流程,7 个属性间存在一定的逻辑关系. 在 7 个属性中,包括与计算相关的属性(A_1和 A_4)、与数学问题表征有关的属性(A_2、A_5、A_6、A_7)及单位换算(A_3),这 3 者之间相互独立. 此外,会不会基本运算(A_1)、是否知道基本数量关系(A_2)、能否单位换算(A_3)之间也相互独立,但 A_1和 A_2却是其它属性的先决条件,如复杂运算(A_4)要以基本运算(A_1)为基础,复杂知识图式(A_5)、识别隐含条件(A_6)和图式表征(A_7)要以 A_2为基础,由此构成了图 1 中的层级关系.

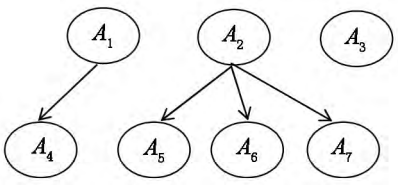


图 1 数学应用题 7 属性的层级关系

1.2 测验矩阵 Q_t

按照 AHM 的思想,属性层级可用矩阵的形式来表示,即属性之间直接与间接关系的 R 矩阵,通过采用扩张算法可得到项目考核模式(53 个)和属性掌握模式全集 A_{MP}(54 个)为

$$A_{MP}=\begin{pmatrix}
1001110000111111000000111111111100001111111111101111110\\
0101001111111110111111111111111111111111111111111111110\\
0010101000100001111000111100000001110111111000011111010\\
0000010000010001000000100011100000000111000111001110110\\
0000000100001000100110010010011011011001101101111011110\\
0000000010000100010101001001010110110101011011111011110\\
0000000001000010001011000100101101110010110111101111110
\end{pmatrix}.$$

A_{MP}中,去掉全 0 模式,其余 53 种即项目考核模式,或称减缩矩阵 Q_r,它们是测验编制的蓝图. 在实际编制测验中,要考虑到测验长度和测试时间与小学生注意力的关系,避免疲劳效应,及组织施测的便利. 在考核模式较多的情况下,一般不会穷尽所有的考核模式来编制测验,而是从 Q_r中选择典型的项目考核模式来编制测验,称其测验 Q 阵(Q_t),诊断测验的编制是基于 Q_t而非 Q_r. 按照文献[8,11] 的 Q

矩阵充分性原则，首先，Q_t 中必须包含 R 矩阵，其次从 Q_t 中抽取其它模式. 本研究考虑几种抽取方式：(i) 只包含 R 矩阵的 Q_{t1}；(ii) 包含 R，同时考虑属性个数和项目平衡，还要考虑测试时间以避免被试注意力和疲劳效应的 Q_{t2}；(iii) 包含 R，但随机选择，且项目长度在 Q_t 的 2/3 以上的 Q_{t3}；(iv) 全模式 Q_{t4}. 由于实际评估中不可能同时对同一批被试测试以上 4 种测验，本研究综合考虑到测验长度与测试时间对测试效果的影响，认为 40 min(1 节课) 的测试对于小学生来说是适宜的. 在考虑 Q_t 包含 R，测验能实现对每个属性的多次观察，长度不宜太长以 40 min为宜等原则，选用第 2 种抽取方式，测验 Q 阵包含 16 种项目考核模式 Q_{t2} 为

$$Q_{t2}=\begin{pmatrix}1010000111011110\\0101110111111111\\0000001000000110\\0010000111011010\\0001000100111000\\0000100011111011\\0000010001101001\end{pmatrix}.$$

2　4 年级数学应用题 Q_t 适宜性的实证效度

利用 GRM-AHM-A 方法，对 4 年级学生在数学应用题上的表现进行认知诊断评估，利用所得数据，对所选 Q_{t2} 矩阵的适宜性进行效度验证. 被试为中国浙江省金华和温州两地区的 4 年级学生，共 888 人. 测试时长为 1 节课(40 min). 采用团体施测，主试由心理系硕士研究生担任. 本研究采用按属性个数给分的多级计分方式，获得 888 名学生在 16 道题上的原始得分. 基于这个原始得分矩阵，通过以下方法分析 Q_t 的适宜性.

2.1　属性矩阵对项目难度的贡献

S. Embretson 等[19] 建议使用分层回归来验证认知属性的充分性. 本研究中，认知属性包含数学内容属性($A_1 \sim A_5$) 和认知技能属性($A_6 \sim A_7$) 2 个成分. 以 IRT 的最大难度等级为因变量，将数学内容属性的 5 个属性作为第 1 层，构成简洁模型 C，在此基础上，加入第 2 层认知技能属性 A_6 和 A_7，构成扩展模型 A，通过分层回归验证 Q_t 的充分性.

从表 2 可知，模型 C 和模型 A 都是有效的. 数学知识对项目难度的解释量为 68.3%，认知技能属性对项目难度的解释量为 12.1%，两者综合高达 80.3%，且模型 C 和模型 A 的 F 值都是显著的，这表明数学应用题的内容知识和认知技能两成分是合理的. 涂冬波等[12] 认为 R^2 达到 60% 以上就说明 Q_t 是充分的. 在表 2 中，不仅 R^2 达到 60% 以上(80.3%)，且效果量也在 60% 以上(63.1%)，这说明本实证研究所认定的 7 个认知属性是合理的，所选用的 Q_t 也是充分的.

表 2　难度对属性的分层回归

模型	属性	R	R^2	F	η^2	$\triangle R^2$
C	$A_1 \sim A_5$	0.826	0.683	4.301*	0.524	
A	$A_1 \sim A_5, A_6 \sim A_7$	0.893	0.803	4.668*	0.631	0.121

从表 2 可知，模型 C 和模型 A 都是有效的。数学知识对项目难度的解释量为 68.3%，认知技能属性对项目难度的解释量为 12.1%，两者综合高达 80.3%，且模型 C 和模型 A 的 F 值都是显著的，这表明数学应用题的内容知识和认知技能两成分是合理的。涂冬波等[12] 认为 R^2 达到 60% 以上就说明 Q_t 是充分的. 表 2 中，不仅 R^2 达到 60% 以上(80.3%)，且效果量也在 60% 以上(63.1%)，说明本实证研究所认定的 7 个认知属性是合理的，所选用的 Q_t 也是充分的.

2.2　归类百分比

GRM-AHM-A 将 888 名被试全部归类到 53 种属性掌握模式中的 44 种(见图 2)，归类比达到 100%，按照 K. K. Tasuoka 等[20] 的观点，Q 矩阵是充分的. 图 2 表明，人数主要集中在 1(1000000)、2(0100000)、7(0110000)、12(1101000)、28(1101010)、40(1111101)、49(1110110)，人数比例分别为 6.64%、11.15%、6.19%、6.08%、10.25%、9.80%、7.32%，其总和为 57.43%，超过 50%. 从这 7 种 A_{MP} 的人数比例，可以发现大部分学生都掌握了 A_1(基本运算) 和 A_2(基本数量关系) 这 2 个基本知识属性，而在 A_3(单位换算)、A_4(复杂运算)、A_5(复杂数量关系) 上的掌握比例明显偏低，在 A_6(识别隐含条件) 和 A_7(图式表征) 这 2 个认知技能属性上掌握比例最低. 可见，学生的认知错误和认知属性本身的难度性质是匹配的. 这表明本研究所认定的属性是与学生的认知规律相符的，Q_t 是充分有效的.

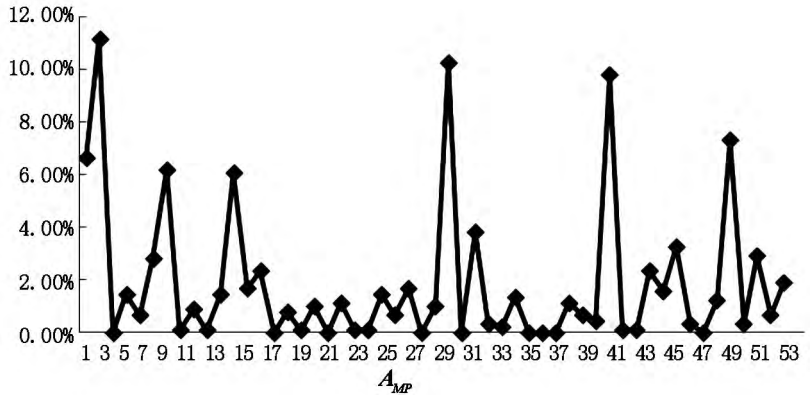


图 2 GRM–AHM–A 的归类结果

2.3 属性的层级一致性

HCI 是反应属性层级一致性的指标[10,21]，按其思想，每个被试都可计算出一个 HCI，然后以全体被试的平均 HCI 值表示属性层级的一致性高低. 然而，HCI 有一个天然的缺陷，就是有可能出现分母为 0 的情形. 本研究中有 2 名被试的 HCI 值无法计算，若将 2 名被试的 HCI 值取值为 0，则 888 名被试的平均 HCI 为 0.62; 若剔除这 2 名被试，则 886 名被试的平均 HCI 为 0.63.

丁树良等[22]指出，上述 HCI 公式中，$S_j=\{g \mid g$ 为项目，$AS_g \subseteq AS_j, g \neq j\}$，若项目 i 仅仅包含 1 个属性 k，且同类试题仅此 1 题，被试 j 只在该试题上正确作答，此时 HCI 的分母为 $N_{cj}=0$，公式无意义. 为了避免这种错误，丁树良等对 S_j 做了这样的修改，$S_j=\{g \mid g$ 为项目，$AS_g \subseteq AS_j\}$，修改后的 HCI 称为 MHCI. 本研究中 888 名被试的平均 MHCI 为 0.74.

本研究中，HCI 和 MHCI 的计算，都经过由多级计分严格转换为 0-1 计分得到，尽管计算精度受到一定影响，但无论是 HCI 均值 0.62(0.63) 还是 MHCI 均值 0.74，它们都表明本测验的属性层级具有良好的合理性. 由此，基于所选择 Q_t 编制的认知诊断测验能反映学生的认知结构.

3 4 年级数学应用题 Q_t 适宜性的模拟证据

模拟研究旨在使用 GRM–AHM–A 方法比较实证研究中提到的 4 种测验 Q 阵的判准率，以进一步验证依据上述原则所选择的 Q_{t2} 的适宜性. 由此，本部分的 4 个测验 Q 阵分别为 1.2 中的 Q_{t1}、Q_{t2}、Q_{t3}、Q_{t4}. 4 个测验 Q 阵见图 3，各测验 Q 阵对各属性考核次数见表 3.

$$\boldsymbol{Q}_t=\begin{pmatrix}1001000\\0100111\\0010000\\0001000\\0000100\\0000010\\0000001\end{pmatrix},\boldsymbol{Q}_{t2\_16items}=\begin{pmatrix}1010000111011110\\0101110111111111\\0000001000000110\\0010000111011010\\0001000100111000\\0000100011111011\\0000010001101001\end{pmatrix},\boldsymbol{Q}_{t3\_30items}=\begin{pmatrix}11101101111101100001010110101 0\\01111101101111111111111111111111\\00100111101001001011100011111 0\\01100001010101000000000101000\\01001001001011101010011011 0011\\00001101001101011000101000 1000\\01001000101001100110101100 0111\end{pmatrix},$$

$$\boldsymbol{Q}_{t4\_53items}=\begin{pmatrix}10011100001111110000001111111111000011111111110111111\\01010011111111101111111111111111111111111111111111111111111\\00101010001000011110001111000000111011111100001111101\\00000100000100010000001000111000000011100011100111011\\00000001000010001001100100100110110110011011011110111\\00000000100001000101010010010101101101010110111101111\\00000000010000100010110001001011011100101101111011111\end{pmatrix}.$$

图 3 模拟研究中的 4 个测验 Q 阵

模拟研究假设，在 Q_t 包含 R 的情况下，测验长度并非越长越好，而是只要每个属性被考察多次，且考虑属性在项目间平衡就好，即认为 Q_{t2} 的效果要好于其它.

表 3　各属性在各个 Q_t 上被考察的次数

Q_t	各属性被考察次数						
	A_1	A_2	A_3	A_4	A_5	A_6	A_7
Q_{t1}	2	4	1	1	1	1	1
Q_{t2}	9	13	3	7	5	8	5
Q_{t3}	18	27	16	8	15	11	14
Q_{t4}	36	48	27	18	24	24	24

3.1　模拟过程

已有研究结果表明,随着失误率的增大,判准率将降低,因而该研究不考虑失误率的影响,即 4 种 Q 矩阵下的失误率均统一设定为 5%,每种条件模拟 10 次以减少误差. 模拟过程为:(i) 采用矩阵乘法 $((A_{MP})_{m\times k}\times Q_{k\times n})$ 计算期望反应模式(ERP);(ii) 模拟观察反应模式. 首先,将 ERP 总分从小到大排序,使具有这些知识状态的被试人数满足标准正态分布,总分相同的 ERP 平均分配人数,产生 1 000 个被试进行分配; 其次,发生 5% 的失误,对每一个 ERP 中每个项目上的得分先产生一个服从 $U(0,1)$ 的随机数 r. 若 $r>0.975$,ERP 为满分,则减 1 分,否则,加 1 分; 若 $r<0.025$,ERP 不为 0 分,则减 1 分,否则,加 1 分; 若 $0.025\leqslant r\leqslant 0.975$ 时,则 ERP 的项目得分不变.

3.2　参数估计

将 ERP 和 ORP 合并,采用 Multilog7.3 中的 GRM 参数估计程序对项目参数和被试能力参数进行估计,得到项目的区分度、等级难度和被试能力等参数. 自编 Excel 2003 的 VB 程序,采用 GRM-AHM-A 方法对模拟被试的知识状态进行判别归类.

3.3　评价指标

模式判断率(Pattern Match Ratio,P_{MR}),即掌握模式判准的被试比例. 若失误前后 2 个 A_{MP} 完全一致,$f_i=1$,否则 $f_i=0$,通过 $\sum_{i=1}^{N} f_i/N$ 可求得模式判准率.

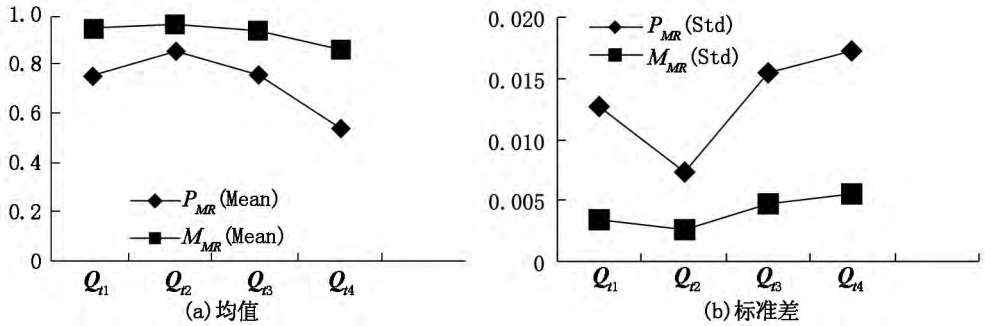
属性判准率(Marginal Match Ratio,M_{MR}),对每一个被试而言,若被试 i 在 2 种 A_{MP} 中的第 k 个属性上相同,记 $g_k=1$,否则 $g_k=0$. 被试属性判准率,即 k 个属性中判准的属性比例,通过 $\sum_{k=1}^{K} g_k/K$ 可求得.

3.4　结果

表 4 是基于 GRM-AHM-A 方法所得 4 个 Q_t 的 P_{MR} 和 M_{MR} 描述统计量,从表 4 可以看出 M_{MR} 整体较高且很稳定,P_{MR} 稍差,但总体均值也达 0.728 3. 从数值上看,4 个 Q_t 的 P_{MR} 和 M_{MR} 均值及标准差并不一样,方差分析表明,4 个 Q_t 的 P_{MR} 和 M_{MR} 均值存在显著差异($F_{(3,36)}=939.1$,$P<0.001$,$\eta^2=0.986$,$F_{(3,36)}=1\ 257.748$,$P<0.001$,$\eta^2=0.990$). Scheffe 结果表明(见表 5 和表 6),就 P_{MR} 而言,$Q_{t2}>Q_{t3}$,$Q_{t1}>Q_{t4}$,就 M_{MR} 而言,$Q_{t2}>Q_{t1}>Q_{t3}>Q_{t4}$,这也可以从图 4 的均值变化图清晰地看出. 此外,从稳定性来看,图 4 的标准差趋势图也表明,Q_{t2} 的标准差最小,Q_{t1} 和 Q_{t3} 次之,而 Q_{t4} 最大.

表 4　4 种条件下的 P_{MR} 和 M_{MR} 均值及标准差

MR	Q	Mean	Std	n
P_{MR}	Q_{t1}	0.752 2	0.012 7	10
	Q_{t2}	0.857 8	0.007 4	10
	Q_{t3}	0.761 0	0.015 4	10
	Q_{t4}	0.542 4	0.017 3	10
	Total	0.728 3	0.117 3	40
M_{MR}	Q_{t1}	0.945 4	0.003 5	10
	Q_{t2}	0.968 0	0.002 6	10
	Q_{t3}	0.936 1	0.004 8	10
	Q_{t4}	0.858 3	0.005 5	10
	Total	0.927 0	0.042 0	40



(a) 均值　　(b) 标准差

图 4　4 种条件下的 P_{MR} 和 M_{MR} 的均值与标准差变化趋势

表 5 P_{MR} 的比较结果

Q_t	Subset			
	n	1	2	3
4.00	10	0.542 3		
1.00	10		0.752 2	
3.00	10		0.761 0	
2.00	10			0.857 8
Sig.		1.000 0	0.565 0	1.000 0

表 6 M_{MR} 的比较结果

Q_t	Subset				
	n	1	2	3	4
4.00	10	0.858 3			
3.00	10		0.936 1		
1.00	10			0.945 4	
2.00	10				0.968 0
Sig.		1.000	1.000 0	1.000 0	1.000 0

4 讨论与结论

在 CDA 的模拟研究中,认知模型或 Q 矩阵可以任意设定,然而在实证研究中,认知模型或 Q 矩阵的界定却是一件非常复杂的事情. 本文以 4 年级数学应用题为例,按照前人关于认知属性和认知模型的界定方法及 Q_t 选择的原则:(i) Q_t 矩阵要包含 R 矩阵;(ii) 考虑属性考核次数在项目中的平衡;(iii) 考虑时间疲劳效应. 选择 Q_{t2} 作为实践评估中认知诊断测验编制的蓝图,并通过实证和模拟验证 2 个角度探讨了 Q_{t2} 的适宜性,为 CDA 实践中 Q 矩阵的选择提供了有用信息.

4.1 Q_{t2} 具有良好的实证效度

关于 Q_t 矩阵的选择,已有研究表明 Q_t 必须包含 R,才能保证 ERP 和 AMP 一一对应,否则容易出现一个 AMP 对应多个 ERP 的情况. 在数学应用题 Q_t 选择中,从同一个 Q_r 中选择 Q_t 有无数种组合,本研究从考核模式的多少、测验长度、属性被考察次数的多少等方面,考虑了相对较为典型的 4 种方案:只包含 R 的 Q_{t1}、包含 R 但考虑属性个数平衡和学生作答疲劳效应的 Q_{t2}、包含 R 但测验长度更长的 Q_{t3}、全考核模式 Q_{t4}. 为吻合实践情境,选择 Q_{t2} 作为最终认知诊断测验编制的蓝图.

888 名学生的实证数据表明,Q_{t2} 具有较好的实证效度. 首先,难度对属性的回归结果表明数学内容属性和认知技能属性两成分所包含的 7 个认知属性是充分的,其对难度的解释量及效果量均达到了测量学要求. 采用 GRM-AHM-A 对被试的模式判别的归类比达到了 100%,且 HCI 和 MHCI 均表明 Q_{t2} 能较好地表征数学认知属性的层级关系. 尽管实证研究不能对各个 Q_t 的效果进行比较,因为不可能对同一批被试重复 4 次测量内容相同但长度不同的考题,但至少表明本研究所选择的 Q_{t2} 具有较好的实证效度,Q_{t2} 是适宜的且能较好地对不同知识状态的学生进行诊断和分类.

4.2 Q_{t2} 具有较好的模拟效度

为进一步考察 Q_{t2} 的适宜性,通过模拟研究比较本研究所提出的 4 种 Q_t 的优劣,以考察是否 Q_t 包含的模式越多,以及由此导致的测验长度越长越好,或属性被考察的次数越多越好,评价指标为 PMR 和 MMR 结果表明,无论是 PMR 还是 MMR、Q_{t2} 和 Q_{t1} 的均值都明显高于 Q_{t3} 和 Q_{t4},这说明在包含 R 的情况下,Q_t 并非越长越好. 在稳定性方面,Q_{t2} 也是最好的,其标准差最小. 为更进一步说明 4 个 Q_t 判准率的差异实质,把 4 个 Q_t 对被试的归类结果与模拟时设定的知识状态(真值)进行比较,如图 5 所示. 图 5 中,真值趋势线反映模拟设计时设定的各类知识状态的人数比例,Q_{t1} ~ Q_{t4} 趋势线反映在 Q_{t1} ~ Q_{t4} 的考核模式下被试知识状态的判归趋势. 从图 5 可知,Q_{t2} 几乎与真值线重合,Q_{t1} 次之,Q_{t3} 和 Q_{t4} 则较差. 具体而言,Q_{t1} 在 A_{MP} 的 16、54 处多判,而在其他模式稍有低判. Q_{t3} 在模式 6、10、22、28、31、38、41 处高判,而在 1、7、20、24 稍有低判. Q_{t4} 在 6、9、34、36、48、50 处高判,在 1、15、24、26、51 处低判.

4.3 Q_t 的考核模式并非越多越好

在 CDA 实践研究中,Q_t 的选择是非常重要的,因其关系到测验的质量优劣及判准率高低. Q_t 的选择必须符合一些原则,本研究结果表明,在包含 R 的情况下,Q_t 的考核模式并非越长越好,每个属性被考察的次数并非越多越好,否则,由此导致太长的测验长度反而会降低判准率,即使是只包含 R 的 Q_t 也要好于太长的 Q_t.

本研究的目的是要验证数学应用题 Q_t 的适宜性,无论是实证还是模拟研究都表明所选 Q_t 是适宜的,这在一定程度上验证了理论研究中关于 Q_t 选择原则的合理性,也为 CDA 实证研究提供了经验和借鉴. 即在 CDA 实践中,在 Q_r 考核模式较多的情况下,Q_t 的选择原则为:(i) 包含 R;(ii) 实现每个属性的多次考察;(iii) 考虑属性个数在项目中的平衡;(iv) 考虑测验长度对被试的疲劳效应,就小学生而

言，测试时间不宜太长，以 40 min 为宜.

当然，本研究只考虑了数学应用题中一种层级关系下的 Q_t 适宜性问题，实证研究中也只用了一种 Q_t 的题目. 未来研究可以考虑在不同学科不同内容领域、不同层级关系、不同失误率的前提下，Q_t 的不同选择方式对判准率的影响，看其反应趋势与本研究是否一致，为本研究的发现提供更多证据支持.

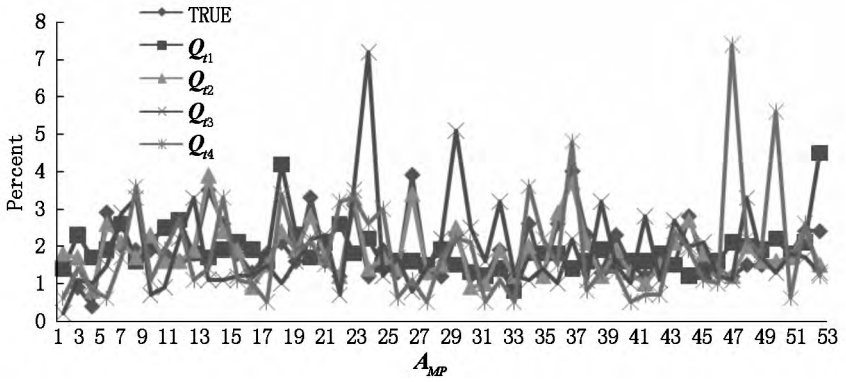


图 5　4 种 Q_t 对被试知识状态的归类与真值的比较

5　参考文献

[1] Borsboom D，Mellenbergh G J，van Heerden J. The concept of validity [J]. Psychological Review，2004，111(4)：1061.

[2] de la Torre J. An empirically based method of Q-matrix validation for the DINA model：development and applications [J]. Journal of Educational Measurement，2008，45(4)：343-362.

[3] De Carlo L T. On the analysis of fraction subtraction data：the DINA model，classification，latent class sizes，and the Q-matrix [J]. Applied Psychological Measurement，2011，35(1)：8-26.

[4] Henson R，Douglas J. Test construction for cognitive diagnosis [J]. Applied Psychological Measurement，2005，29(4)：262-277.

[5] Rupp A A，Templin J L. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model [J]. Educational and Psychological Measurement，2008，68：78-96.

[6] Im S，Corter J E. Statistical consequences of attribute misspecification in the rule space method [J]. Educational and Psychological Measurement，2011，71(4)：712-731.

[7] Kunina-Habenicht O，Rupp A A，Wilhelm O. The Impact of model misspecification on parameter estimation and item-fit assessment in log - linear diagnostic classification models [J]. Journal of Educational Measurement，2012，49(1)：59-81.

[8] 丁树良，杨淑群，汪文义. 可达矩阵在认知诊断测验编制中的重要作用 [J]. 江西师范大学学报：自然科学版，2010，34(5)：490-494.

[9] Gorin J S. Test construction and diagnostic testing [A]. Leighton J P. Cognitive diagnostic assessment for education：theory and applications [M]. Cambridge：Cambridge University Press，2007.

[10] Gierl M，Wang Changjiang，Zhou Jianwen. Using the attribute hierarchy method to make diagnostic inferences about examinees´cognitive skills in algebra on the SAT(c) [J]. Journal of Technology，Learning，and Assessment，2008，6(6)：1-53.

[11] 丁树良，汪文义，杨淑群. 认知诊断测验蓝图的设计 [J]. 心理科学，2011(2)：258-265.

[12] 涂冬波，漆书青，戴海琦，等. 教育考试中的认知诊断评估 [J]. 考试研究，2008(4)：4-15.

[13] 祝玉芳，丁树良. 基于等级反应模型的属性层级方法 [J]. 心理学报，2009(3)，267-275.

[14] 康春花，辛涛，田伟. 小学数学应用题认知诊断测验编制及效度验证 [J]. 考试研究，2013(6)：24-43.

[15] Mayer R E. Different problem-solving strategies for algebra word and equation problems [J]. Journal of Experimental Psychology：Learning，Memory，and Cognition，1982，8(5)：448-462.

[16] Enright M K，Morley M，Sheehan K M. Items by design：the impact of systematic feature variation on item statistical characteristics [J]. Applied Measurement in Education，2002，15(1)：49-74.

[17] Arendasy G，Sommer M. Using psychometric technology in educational assessment：the case of a schema-based isomorphic approach to automatic generation of quatitative reasoning items [J]. Learning and Individual Differences，2007，17：366-383.

[18] Arendasy M，Sommer M，Gittler G，et al. Automatic genera-

tion of quantitative reasoning items [J]. Journal of Individual Differences,2006,27(1):2-14.

[19] Embretson S, Gorin J. Improving construct validity with cognitive psychology principles [J]. Journal of Educational Measurement,2001,38(4):343-368.

[20] Tatsuoka K K, Corter J E, Tatsuoka C. Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries [J]. American Educational Research Journal,2004,41(4):901.

[21] Cui Ying, Leighton J P. The hierarchy consistency index: evaluating person fit for cognitive diagnostic assessment [J]. Journal of Educational Measurement,2009,46(4):429-449.

[22] 丁树良,毛萌萌,汪文义,等. 教育认知诊断测验与认知模型一致性的评估 [J]. 心理学报,2012(11):1535-1546.

The Suitability of *Q*-Matrix on the Primary School Grade Four Students' Arithmetical Word Problem

KANG Chunhua[1,2], YANG Yakun[1,2], ZHONG Xiaoling[1], ZENG Pingfei[1*]

(1. College of Teacher Education, Zhejiang Normal University, Jinhua Zhejiang 321004, China;
2. CN Test Company, Shenzhen Guangdong 518000, China)

Abstract: The definition and selection of *Q*-matrix are very important in cognitive diagnostic assessment(CDA), because these concern the quality of a test and accuracy of CDA. The *Q*-matrix of simulation study can be set arbitrarily, but it not always the case in practical research. Based on the principles of existing theory and related simulation studies, the primary school grade four students´ arithmetical word problem is taken as an example to illustrate how to choose a suitable testing *Q*-matrix in practice. Empirical and simulation studies are used toverify the appropriate of selected testing *Q*-matrix. The results suggested that increasing the pattern and number of test items not always improve the pattern match ratio (PMR) and marginal match ratio (MMR) when testing *Q*-matrix contains the reachability matrix; instead, a *Q*-matrix with reachability matrix is better than the *Q*-matrix which includes too many test patterns.

Key words: mathematical word problem; *Q*-matrix; reach-ability matrix; GRM-AHM-A method

(责任编辑:冉小晓)