

• 研究方法(Research Method) •

## 多项选择题认知诊断潜能的最大化\*

李 瑜<sup>1</sup> 丁树良<sup>2</sup> 唐小娟<sup>3</sup>

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

(<sup>3</sup> 南昌航空大学数学与信息科学学院, 南昌 330063)

**摘 要** 多项选择题是认知诊断中常用的测验项目形式, 其正确答案选项和经过特殊编制的干扰项都能提供诊断信息。为了提取干扰项的信息, 需要采用不同于传统认知诊断模型的心理计量模型。分类介绍能使干扰项提供诊断信息的多项选择题编制方法, 分析这类多项选择题的认知诊断模型, 认为未来应加强干扰项编制方法和能提取干扰项信息的认知诊断模型的选择和开发等方面的研究。

**关键词** 认知诊断; 多项选择题; 干扰项; 诊断信息

**分类号** B841

### 1 引言

多项选择题(Multiple Choice item, MC item)由题干和一个正确答案选项及若干个干扰项组成。在广泛的学科领域的测验中, 特别是在大规模标准测验中, 多项选择题被认为是最通用、最有效的测验项目形式(Airasian, 2001; Kubiszyn & Borich, 1987; Linn & Gronlund, 2000; Nitko, 2004)。多项选择题具有如下优点: 可以在相对短的时间里考察更多的主题; 信度高; 计分具有客观性, 且易于快速计分; 易于计算机登分等(Roediger & Marsh, 2005), 因而受到测验编制者和使用者的青睐(Tamir, 1991)。

认知诊断是现代测量理论发展的重要领域。与传统测量理论相比较, 认知诊断不再只是提供一个测验的总分, 它旨在测量学生具体的知识结构和加工技能, 以便能够提供学生的认知强项和弱项的信息(Gierl, Leighton, & Hunka, 2007), 使

教师可以根据这些信息给予学生有效的指导以满足每个学生的学习需要。用于认知诊断的心理计量模型称为认知诊断模型(Cognitive Diagnosis Models, CDMs)。CDMs 是用项目反应函数估计被试的认知属性, 识别被试对多个细粒度认知技能的掌握情况(Rupp & Templin, 2008)。认知诊断是根据被试在项目上的作答反应, 采用 CDMs 推断被试的属性掌握模式, 估计项目的诊断能力(Hartz, 2002)。到 2008 年为止, 心理测量学者们开发了 60 多种 CDMs (Fu & Li, 2008)。大多数 CDMs 在处理多项选择题反应数据时采用二级评分形式来估计被试的知识状态, 即如果被试掌握了项目所考查的所有属性, 那么该被试被分到掌握组; 如果被试至少有一个属性未掌握, 那么该被试被分到未掌握组。这样, 对于未掌握组被试而言, 显然忽视了未掌握一个属性的被试与未掌握多个属性的被试之间的差异。在理想反应情况下, 被试缺失一个属性与缺失多个属性在作答时做出的反应会不同, 只有一个属性未掌握的被试答对项目的概率要高于多个属性未掌握的被试。事实上, 如果对多项选择题的干扰项包含的信息作进一步诊断, 那么由干扰项得到的诊断信息会有助于提高估计被试知识状态的准确性。如何使干扰项尽可能多的提供诊断信息, 如何提取干扰项的诊断信息以提高被试知识状态估计的精度,

收稿日期: 2013-08-28

\* 国家自然科学基金(30860084, 31160203, 31100756, 31360237, 31300876)、国家社科基金(12BYY055, 13BYY087)、教育部人文社科青年基金(13YJC880060)、江西省教育厅科技项目(GJJ13207, GJJ13208, GJJ13209, GJJ13226, GJJ13227)资助。

通讯作者: 丁树良, E-mail: ding06026@163.com

是学者们在研究多项选择题诊断信息时最为关心的问题。本文将从多项选择题编制方法、多项选择题计分方式及相应的 CDMs 两个方面对前人的研究进行述评。最后, 对多项选择题未来的研究发展做出展望。

## 2 认知诊断多项选择题的编制

认知诊断测验提供的信息越丰富, 对被试知识状态的估计就会越准确, 因此测验要尽可能多地提供诊断信息。要使多项选择题的认知诊断测验提供尽可能多的信息, 不仅要最优设计测验蓝图(丁树良, 杨淑群, 汪文义, 2010; 丁树良, 汪文义, 杨淑群, 2011), 使测验尽量合理、科学、结构最具代表性, 还要使单个多项选择题提供尽可能多的诊断信息。多项选择题除了正确答案选项能提供诊断信息外, 干扰项也能提供附加诊断信息。要使干扰项能提供诊断信息, 编制系统的、有意义的干扰项是关键(Haladyna, 2004)。编制干扰项涉及到两个问题。第一, 干扰项编制方法问题: 是否存在干扰项的最优设计方案使其能够提供尽可能多的诊断信息? 第二, 干扰项的效度验证问题: 编制的干扰项能否提供诊断信息, 干扰项信息能否提高诊断效率? 下面从这两个方面来分析认知诊断多项选择题的编制问题。

### 2.1 多项选择题编制方法及评价

编制能够提供诊断信息的干扰项不是一个新鲜的研究问题, 在传统测量理论中, 为了充分发挥多项选择题干扰项的作用, 使干扰项尽可能多地提供诊断信息以提高诊断效率, 研究者开发出多项选择题的替代形式。如 Hestenes, Wells 和 Swackhamer (1992)以及 Sadler (1998)将干扰项与学生具体的错误类型结合编制多项选择题; Treagust (1995, 2006)构建了双层多项选择题(two-tier MC items); Liu, Lee 和 Linn (2011)在双层多项选择题的基础上改编成解释性多项选择题(explanation of MC items); Pearson 公司设计了干扰项原因分类法(distractor rationale taxonomy), 其目的是在知识的不同理解水平上编制干扰项以准确找到被试错误的原因(King, Gardner, Zucker, & Jorgensen, 2004)。这些多项选择题编制方法都是为了从干扰项中提取学生理解的诊断信息, 这些方法反映出编制能提供诊断信息的干扰项的思想, 为认知诊断测验中编制能提供诊断信息的干

扰项提供了可借鉴的思路。Briggs, Alonzo, Schwa 和 Wilson (2006)在这些编制方法的基础上开发出 OMC (Ordered-Multiple Choice, OMC)项目, OMC 项目的每个选项都能提供诊断信息(Gorin, 2007), OMC 项目是认知诊断多项选择题的开端。但是, OMC 项目编制方法不是完全意义上的认知诊断测验项目编制方法, de la Torre (2009)认为 OMC 项目编制方法是站在单维 IRT 的角度提出的, 他从认知诊断的角度类似地提出了多项选择题编制方法, 即将选项对应的属性向量(假设测验考察  $K$  个属性, 用一个  $K$  维 0-1 向量对应选项包含的属性, 如果选项考察了第  $j$  个属性, 这个 0-1 向量的第  $j$  个分量等于 1, 否则等于 0。称这样构成的  $K$  维 0-1 向量为相应选项的属性向量)与被试知识状态相结合编制多项选择题的方法。下面就 OMC 项目编制方法和将选项对应的属性向量与被试知识状态相结合编制多项选择题的方法进行述评。

#### 2.1.1 OMC 项目编制方法

学生对知识的理解是一个发展进程, 学生在不同的学习阶段对知识的理解程度不同。学生对知识的错误理解与学生所处的学习阶段有关, 而不是由学生的学习障碍导致的(Sadler, 1998)。Briggs 等人(2006)认为当学生对知识的错误理解与学生理解的发展进程之间存在联系时, 就可以从这种联系中获得更有效的诊断信息, 据此可以开发一种更可靠、更有效的测量工具。Briggs 等人(2006)使用学生认知发展模型与待测量的内容结构相结合的方法来建立学生对知识的错误理解与学生理解的发展进程之间的联系, 并由此开发出 OMC 项目。

OMC 项目的编制过程分为两个阶段。第一阶段是根据学生的认知发展模型建立结构图(construct map, National Research Council, 2001; Wilson & Sloane, 2000)。理解的有序水平(ordered levels of understanding, National Research Council, 2001)理论认为学生的学习过程不仅仅是习得更多的知识和技能, 还是当新的知识与已有知识结合后向更高水平能力发展, 在原有理解基础上形成更深水平的理解甚至取代原有的理解。因此, 当学习进程中获得更多的知识、习得更高阶的技能、理解水平更深时, 学习进程就存在一定的有序水平。如果每一个学习发展水平用一个层级表示, 那么各水平之间构成连续性的线性层级关系,

这种线性层级关系可以用结构图来表征(Wilson & Sloane, 2000; Briggs et al., 2006), 这就是结构图的层级与学生认知发展水平的结合。第二阶段是根据结构图编制 OMC 项目选项。将 OMC 项目的每个选项与结构图中不同层级对应, 这就是项目选项与结构图层级的结合。OMC 项目编制的具体操作过程如下: 第一步, 将待测验的知识按照学习发展进程分成几个水平阶段, 用定性描述的方法描述每个水平阶段的特征; 第二步, 对应每个水平编制选项, 使每个选项表征学习发展过程中的一个水平。

利用 OMC 项目的测验结果对学生的诊断解释遵循一个清晰的推理链条: 先根据学生对 OMC 项目的作答来确定学生在结构图中所处的层级, 即将作答选项与结构图的层级建立联系; 再根据学生在结构图中所处的层级确定学生的学习理解水平阶段, 即将结构图的层级与学生的认知发展模型建立联系; 最后根据学生学习的理解水平阶段可以诊断学生掌握知识情况, 即将认知发展模型与测量内容建立联系。

OMC 项目有以下优点: (1)保留了传统多项选择题客观计分和信度高的优势的同时, 还结合了开放性问题作答反应丰富的优点。(2)根据学生认知发展的有序水平设计选项, 先将选项与不同认知水平对应, 再根据学生选择的选项来诊断学生对知识的掌握程度(Briggs et al., 2006)。它不仅仅关注学生能否正确作答, 在学生错误作答时, 还关心诊断学生所处的理解水平, 准确找到学生具体的弱项(Lin, Chu, & Meng, 2010)。这样, OMC 项目提供的关于学生学习发展的诊断信息会更多, 这些信息能够快速可信地反馈给教师、学校和学生。OMC 项目的不足之处有: (1)建立结构图是编制 OMC 的关键, 但是只能对具有层级发展进程性质的学习过程建立结构图, 比如地球与太阳系关系的学习发展过程适合建立结构图(Briggs et al., 2006; Briggs & Alonzo, 2009)。很难对不具备层级发展进程的学习过程建立结构图。越是定性的和整体性的学习发展过程就越不适合建立结构图, 比如力和运动的学习发展过程很难建立结构图(Alonzo & Steedle, 2009)。因而, 对于不能建立结构图的测量内容不适合编制成 OMC 项目。(2)对于能够建立结构图的测量内容, 根据结构图所表征的多个层级水平编制项目选项是一项艰难的

任务(Briggs et al., 2006), 它需要为测验编制者提供建构结构图和根据结构图编制项目选项的训练, 还需要为教师提供根据结构图解释项目反应的训练, 这样使得测验成本很高, 因此其应用受到限制。(3)根据结构图编写 OMC 项目选项可能会出现“天花板效应”(ceiling effect)和“地板效应”(floor effect, Briggs et al., 2009)。所谓“天花板效应”是指学生的能力可能会超过选项所描述的最高水平。Briggs 等人(2006)假设能力超过选项所描述的最高水平的被试会选择最高水平的选项, 结果会导致低估这类被试的能力。所谓“地板效应”是指学生的能力可能会低于选项所描述的最低水平。Briggs 等人(2006)假设能力低于选项所描述的最低水平的被试会猜测, 结果会导致高估低水平被试的能力。(4)可能出现两个或两个以上选项对应同一个认知发展水平的情况, 这会增加学生选择某个理解水平的选项的猜测概率(Briggs et al., 2009)。理想情况下, 一个项目的每个选项只对应一个认知发展水平。

OMC 项目的编制思想是将学生认知发展的有序水平与选项结合, 利用干扰项提供的诊断信息, 以达到提高诊断准确性的目的。但是, OMC 项目不是根据认知诊断的目的来编制, 它不能对干扰项进行认知解释, de la Torre (2009)认为 OMC 是站在单维 IRT 的角度编制项目, 他提出选项对应的属性向量与被试知识状态相结合的多项选择题编制方法, 这种方法使干扰项具有认知解释功能。

### 2.1.2 选项对应的属性向量与被试知识状态相结合的多项选择题编制方法

de la Torre (2009)构建了选项对应的属性向量与被试知识状态相结合的多项选择题编制方法, 这种方法的思路是将干扰项与具体的潜在类相对应, 这种思想与 OMC 项目编制思想类似。与 OMC 方法不同的是, de la Torre (2009)的方法是根据认知诊断的目的来编制多项选择题。需要说明的是, Luecht (2007)也指出认知诊断评估最常用的测验设计方法之一就是明确地设计跨项目的干扰项以吸引有系统错误概念和错误策略的被试。如果根据被试理解能力的不同状态设计跨项目的干扰项, 那么这些跨项目的干扰项就可以诱导出特定的被试观察反应模式; 反之, 被试的认知机制和引起被试选择特定干扰项的不同能力水平能从实质上

解释被试的观察反应模式, 达到诊断的目的。但是 Luecht (2007)只提供了一种测验开发思路, 在文献中未给出项目编制方法。

de la Torre (2009)将多项选择题的选项分为编码选项(coded option)和非编码选项(non-coded option)。如果选项对应的属性向量是  $2^K-1$  个潜在类中的一类( $K$  是测验所考查的属性个数), 那么该选项称为编码选项, 否则, 称为非编码选项。编码选项包括正确答案选项和编码干扰项, 对每个项目而言, 考查属性个数最多的编码选项是正确答案选项。de la Torre (2009)的多项选择题编制方法要求干扰项能够区分缺少正确答案选项必需的 1 个属性或多个属性的被试类, 也就是要求干扰项反映的被试知识状态是正确答案选项反映的知识状态的子集。基于这个要求, 在编制干扰项时, 该方法根据正确答案选项对应的属性向量的子集来编制编码干扰项。例如, 编制一份考查 5 个属性的测验, 这 5 个属性的层级关系是独立结构, 一道多项选择题的正确答案选项对应的属性向量是  $(1,1,1,0,0)'$ , 根据干扰项反映的被试知识状态是正确答案选项反映的知识状态的子集的要求, 编码干扰项对应的属性向量可以是  $(1,1,0,0,0)'$ 、 $(1,0,1,0,0)'$ 、 $(0,1,1,0,0)'$ 、 $(1,0,0,0,0)'$ 、 $(0,1,0,0,0)'$  或  $(0,0,1,0,0)'$ 。

de la Torre (2009)的多项选择题编制方法将选项对应的属性向量与  $2^K-1$  个潜在类结合, 因此可以根据被试在项目  $j$  上的选择将所有潜在类(latent class)的被试分为  $H_j^*+1$  组(group), 其中  $H_j^*$  是项目  $j$  的编码选项个数, 1 指的是属性掌握模式与任何一个编码选项对应的属性向量都不匹配的潜在组。将被试在项目  $j$  上进行分组的目的是为了推断出被试是掌握了项目的所有属性、部分属性、还是全部属性都没有掌握。对未掌握正确答案选项考查的所有属性的被试进一步分类, 提高了分类的准确性。下面介绍具体的分组方法。

记项目  $j$  的编码选项  $h$  对应的属性向量为  $q_{jh}$ , 非编码选项对应的属性向量记为  $q_{j0}$ , 约定  $q_{j0}=0$ 。以  $g_{ij}$  表示被试  $i$  在项目  $j$  上被分到的潜在组,  $g_{ij}$  定义为

$$g_{ij} = \arg \max_{h'} \{ \alpha'_i q_{jh'} \mid \alpha'_i q_{jh'} = q'_{jh'} q_{jh'} \} \quad (1)$$

$h'=0, \dots, H_j$ ,  $H_j$  是项目  $j$  的选项个数。注意到  $q_{j0}=0$ , 由公式(1)知  $g_{ij}=0$  当且仅当被试的属性掌

握模式与任何一个编码选项对应的属性向量都不匹配。下面以一道考查 4 个属性中的前 3 个属性, 有 4 个选项的多项选择题为例来说明分组方法, 其选项与属性的关联关系如表 1 所示。

表 1 选项与属性的关联关系

|       | A | B | C | D |
|-------|---|---|---|---|
| $A_1$ | 0 | 0 | 0 | 1 |
| $A_2$ | 1 | 1 | 1 | 1 |
| $A_3$ | 1 | 1 | 0 | 1 |
| $A_4$ | 1 | 0 | 0 | 0 |

令  $A=0$ ,  $B=1$ ,  $C=2$ ,  $D=3$ 。这道题的编码选项个数是  $H_j^*=3$  (选项  $A$  为非编码选项, 选项  $B$ 、 $C$  和  $D$  为编码选项),  $q_{j0}=(0,0,0,0)'$ ,  $q_{j1}=(0,1,1,0)'$ ,  $q_{j2}=(0,1,0,0)'$ ,  $q_{j3}=(1,1,1,0)'$ 。譬如, 对于掌握了前 3 个属性的被试  $i_1$ , 有  $\alpha_{i_1}=(1,1,1,0)'$ , 此时  $\alpha'_{i_1} q_{j0} = q'_{j0} q_{j0} = 0$ ,  $\alpha'_{i_1} q_{j1} = q'_{j1} q_{j1} = 2$ ,  $\alpha'_{i_1} q_{j2} = q'_{j2} q_{j2} = 1$ ,  $\alpha'_{i_1} q_{j3} = q'_{j3} q_{j3} = 3$ , 由公式(1)知  $g_{i_1 j} = 3$ , 即被试  $i_1$  被分到潜在组 Group 3。又如, 对于掌握了前两个属性的被试  $i_2$ , 有  $\alpha_{i_2}=(1,1,0,0)'$ , 此时  $\alpha'_{i_2} q_{j0} = q'_{j0} q_{j0} = 0$ ,  $1 = \alpha'_{i_2} q_{j1} \neq q'_{j1} q_{j1} = 2$ ,  $\alpha'_{i_2} q_{j2} = q'_{j2} q_{j2} = 1$ ,  $2 = \alpha'_{i_2} q_{j3} \neq q'_{j3} q_{j3} = 3$ , 由公式(1)知  $g_{i_2 j} = 2$ , 即被试  $i_2$  被分到潜在组 Group 2。对于没有掌握任何一个属性或只掌握了  $A_1$  或  $A_3$  的被试, 其知识状态与任何一个编码选项对应的属性向量都不匹配, 此时  $g_{ij}=0$ , 即被试被分到潜在组 Group 0。

de la Torre (2009)的多项选择题编制方法是真正意义上的认知诊断多项选择题编制方法。这种编制方法在理论上更合理, 实践上更具操作性。de la Torre (2009)和 Gu (2011)在项目考查的属性多于 1 个, 每个项目的编码选项个数等于正确答案选项所需的属性个数, 且编码选项对应的属性向量之间构成一个详尽的层级线性结构(exhaustive and hierarchical linear structure), 即各编码选项对应的属性向量构成一种真包含关系的条件下, 研究编码干扰项能否提供诊断信息。他们的模拟研究证明: 在上述实验条件下, 编码干扰项可以提供附加的诊断信息, 因而更具有诊断能力。需要指出的是, de la Torre (2009)和 Gu

(2011)的模拟研究结论是在一个特定的模拟条件下得到的,在其它的实验条件下是否有相同的结论?这是一个有待验证的问题。再者,de la Torre (2009)的方法要求根据正确答案选项对应的属性向量的子集来编制编码干扰项。当正确答案选项的属性个数多于1个时,此时正确答案选项对应的属性向量的子集有多个,这就需要从这些子集中选择一个或多个作为编码干扰项对应的属性向量。编码干扰项个数的确定和由哪几个子集来编制编码干扰项,这些问题其实是如何使干扰项提供尽可能多的诊断信息的最优设计方案问题,该问题有待进一步探讨和研究。

## 2.2 干扰项的效度问题

根据一定的原理和方法编制好干扰项后,还需要检验干扰项的诊断能力(diagnostic capacity of item distractors, Gu, 2011, P.38),也就是要检验干扰项的效度。检验干扰项的效度通常采用的方法是比较法。所谓比较法,就是使用 CDMs 来推断被试的知识状态时,将考虑干扰项信息和不考虑干扰项信息两种情况下得到的被试知识状态的估计精度(比如估计精度用知识状态的模式判准率或者属性的平均判准率来描述)进行比较。如果考虑了干扰项信息后对被试知识状态有着较为准确的估计,那么说明干扰项确实提供了有效的信息,其具有较好的诊断能力,即干扰项的效度高。反之,如果考虑了干扰项后诊断的准确性不优于未考虑干扰项的情形,则说明干扰项未能够提供有效的诊断信息,此时干扰项的效度低。导致干扰项效度低的原因主要有以下3种情况:(1)所选择的 CDMs 过分稳健,造成模型对信息的变化不敏感;(2)属性间的层级关系导致干扰项不能提供较多的有效信息;(3)项目干扰项编制不合理,如正确答案选项考查的属性较多,但干扰项考查的属性过少,以致干扰项提供的信息不足以对被试的知识状态的估计产生较为显著的影响。Templin, Henson, Rupp, Jang 和 Ahmed (2008)、de la Torre (2009)以及 Gu (2011)等的研究都用了这种比较法。Templin 等人(2008)在补偿性的 RUM (compensatory Reparameterized Unified Model, cRUM, Hartz, 2002)下,de la Torre (2009)和 Gu (2011)在 DINA (Deterministic Inputs, Noisy "And" gate, Macready & Dayton, 1977; Haertel, 1989; Junker & Sijtsma, 2001)模型下都研究了干扰项的

效度问题;Templin 等人(2008)将带干扰项信息的 cRUM 称为称名反应 cRUM (Nominal Response cRUM, NR-cRUM);de la Torre (2009)将带干扰项信息的 DINA 模型称为多项选择题的 DINA 模型 (Multiple Choice DINA, MC-DINA),他们的模拟研究表明干扰项能提供有效的诊断信息。

在实证研究中,由于不知道被试真实的知识状态,因而无法判断干扰项是否提供了有效的信息,为此,需要建立检验干扰项效度的评估指标。Gu (2011)开发了 MC-DINA 模型的项目选项的效度评估指标以评估项目选项的诊断能力。Gu (2011)采用的指标是选项区别指数(option discrimination index)和选项诊断能力指标(diagnostic capacity of item options),这两个指标既可以用于评估正确答案选项的诊断能力也可以用于评估干扰项的诊断能力。

选项区别指数是 DINA 模型的项目区别指数(item discrimination index, de la Torre, 2008; Rupp, Templin, & Henson, 2010)的拓展。选项区别指数表示掌握选项所有属性的被试选择该选项的概率与选项中至少有一个属性没有掌握的被试选择该选项的概率之间的差异,其数学表达式是

$$P_j(h_j | g_i, H_{h_j} = G_{g_i}) - P_j(h_j | g'_i, H_{h_j} \neq G_{g'_i}) \quad (2)$$

上式中  $h_j$  表示项目  $j$  的选项  $h$ ,  $H_{h_j}$  是  $h_j$  的属性集,  $g_i$  表示第  $g$  组中的被试  $i$ ,  $G_{g_i}$  是  $g_i$  掌握的属性集。类似地可以解释  $g'_i$  和  $G_{g'_i}$ 。对于项目  $j$  而言,若  $g_i$  掌握的属性与  $h_j$  的属性完全匹配,即  $H_{h_j} = G_{g_i}$ ,则  $h_j$  是  $g_i$  在项目  $j$  上的期望反应选项(Gu, 2011),其它选项是  $g_i$  在项目  $j$  上的非期望反应选项。记  $g_i$  选择期望反应选项  $h_j$  的概率为  $P_j(h_j | g_i, H_{h_j} = G_{g_i})$ ,  $g_i$  的期望反应选项是  $h_j$  但实际选择的选项不是  $h_j$  的概率为  $P_j(h'_j | g_i, H_{h_j} = G_{g_i})$ ,显然有  $P_j(h'_j | g_i, H_{h_j} = G_{g_i}) = 1 - P_j(h_j | g_i, H_{h_j} = G_{g_i})$ 。若  $g'_i$  掌握的属性与选项  $h_j$  的属性不完全匹配,即  $H_{h_j} \neq G_{g'_i}$ ,则  $g'_i$  选择选项  $h_j$  的概率为  $P_j(h_j | g'_i, H_{h_j} \neq G_{g'_i})$ 。易知,要使选项  $h_j$  具有较大的区别指数,须要求  $P_j(h_j | g_i, H_{h_j} = G_{g_i})$  的取值较大,同时要求  $P_j(h'_j | g'_i, H_{h_j} \neq G_{g'_i})$  的取值较小,即  $P_j(h_j | g_i, H_{h_j} = G_{g_i}) - P_j(h'_j | g'_i, H_{h_j} \neq G_{g'_i})$  的值越大说明选项  $h_j$  区分掌握该选项的所有属性

的被试和没有完全掌握所有属性的被试的能力越强,反之,越弱。区分能力越强的选项提供的诊断信息越丰富,而区分能力弱或没有区分能力的选项提供很少的诊断信息或不能提供诊断信息。

选项诊断能力指标是用来检验被试是否是随机地选择期望反应选项的指标。假设项目  $j$  有 4 个选项,那么被试随机地选择到期望反应选项  $h_j$  的概率是 0.25。这样,判断选项  $h_j$  是否具有诊断能力的问题等价于检验  $P_j(h_j | g_i, H_{h_j} = G_{g_i}) > 0.25$  是否成立。若这个假设检验显著地成立,则认为选项  $h_j$  具有诊断能力;否则就认为选项  $h_j$  不具有诊断能力。

选项区别指数和诊断能力指标分别从被试组间和被试组内两个方面评估项目选项的诊断能力。这两个指标都可以用于检验已有测验的项目选项是否具有诊断能力,也可用于指导编制具有诊断能力的项目选项。但是,对于选项区别指数  $P_j(h_j | g_i, H_{h_j} = G_{g_i}) - P_j(h_j | g_i, H_{h_j} \neq G_{g_i})$ , 其阈值取多大时才认为选项  $h_j$  具有区分能力,这是一个非常值得研究的问题。其次,如果需要对选项的区分能力进行定性的描述,那么需要对区分能力作进一步的分类,例如将区分能力分为具有较强的区分能力、一般的区分能力和较弱的区分能力三种,那么各种区分能力的界限如何从量上进行界定?这也是一个亟需解决的问题。值得注意的是,对于选项诊断能力指标,若假设  $P_j(h_j | g_i, H_{h_j} = G_{g_i}) > 0.25$  通过了检验,我们只能认为选项  $h_j$  具有一定的诊断能力,但是并不能得到选项  $h_j$  诊断能力的大小。这样,为评估选项的诊断信息,我们先检验该选项是否具有诊断能力,若检验表明该选项具有诊断能力,在选项具有诊断能力的前提下,再计算该选项的选项区别指数以确定其诊断能力的大小。

### 3 多项选择题的计分方法和 CDMs 及评价

编制好能使干扰项提供诊断信息的多项选择题后,接下来的任务是为这类多项选择题选择合适的计分方式以达到提取干扰项诊断信息的目的。传统上,多项选择题采用的是二级计分方式,例如 1-0 计分,如果被试选择正确答案,那么该项目记为 1 分,此时可以认为被试掌握了该项目所

考查的属性;如果被试未能选择正确答案,那么该项目记为 0 分,此时可以认为被试对该项目所考查的属性至少有一个未能掌握(这里只考虑属性间的非补偿性情形)。这种二级计分方式忽视了选择干扰项的被试之间的差异,丢弃了干扰项的信息。为了能从干扰项中提取诊断信息,必须采用“恰当”的计分方式,目前使用较多的计分方式有两种——数据扩张法(data augmentation method, Luecht, 2007)和多级计分法。多级计分法对各选项给予不同的计分以达到对选择不同干扰项的被试进行区分的目的。数据扩张法是 1-0 计分方式的一种扩展形式,一方面将被试选择正确答案和未选择正确答案采用 1-0 计分来诊断被试“强”项的诊断信息,另一方面对被试的“弱”反应干扰项(最吸引低分被试选择的干扰项)也采用 1-0 计分以诊断被试的“弱”项的诊断信息。

不同计分方式的数据需要采用不同类型的 CDMs 来分析,关于 CDMs 的分类可以参考 Rupp 和 Templin (2008)以及 Gu (2011)。分析多项选择题数据的 CDMs 的相关研究并不多,到目前为止,国内尚没有找到相关研究,国外的研究见 Templin 等(2008)、Briggs 和 Alonzo (2009)、de la Torre (2009)以及 Gu (2011)等。

#### 3.1 干扰项得分反应的数据扩张法

Luecht (2007)认为任何可以用来计算诊断得分的附加数据源(additional data sources)都可以被利用,从附加数据源中提取诊断信息的方法称为数据扩张法。附加数据源有很多种,多项选择题干扰项中有干扰作用的测量信息就是其中的一种,我们将利用干扰项信息的数据扩张法称为干扰项得分反应(distractor-based scored responses)的数据扩张法。Luecht (2007)认为如果传统的单维测验(如单维 IRT 测验)被认为是诊断被试的“强”项的诊断信息,那么利用干扰项提供的信息采用传统的单维测验的分析方法也能够诊断出被试的“弱”项的诊断信息,这样就达到了对被试的“强”项和“弱”项的诊断目的。

记正确反应得分矩阵为  $U$ , 干扰项得分反应矩阵为  $V$ , 这里  $V$  称为扩张数据矩阵,两个矩阵都是 1-0 记分的反应矩阵。 $U$  和  $V$  分别用不同的计分器(scoring evaluators)计算得到(Luecht, 2001, 2005)。计分器是实现记分的规则, $U$  和  $V$  的计分器使用简单的模式匹配规则。 $U$  通常用正确答案

(Correct Answer Key, CAK)计分器计算得到, CAK 计分器公式是

$$u_{ij} = f(r_{ij}, \alpha_j) \quad (3)$$

其中  $r_{ij}$  是被试  $i$  在项目  $j$  上的反应,  $\alpha_j$  是项目  $j$  的正确反应。如果被试  $i$  在项目  $j$  上的反应是正确反应, 那么  $r_{ij} = a_j$ , 此时令  $u_{ij} = 1$ ; 否则  $u_{ij} = 0$ 。V 可以用常见错误答案法 (Popular Incorrect Key, PIK) 或高度负相关法 (Strongest Negative Correlation, SNC) 计算得到。PIK 假设每个项目至少有一个干扰项对低分被试特别有吸引力 (低分被试一般是按照标准测验划分得到)。如果被试的不正确反应不是由于随机作答或猜测导致的, 而是由于被试缺乏相应的知识 (如错误理解概念、使用错误的解题策略等) 导致的 (被试会同时在多个项目间表现出其缺乏该知识), 那么没有考查相应知识的干扰项对低分被试特别有吸引力, 从而产生独特的反应模式。这种独特的反应模式, 对低分被试而言能提供有用的诊断信息。PIK 计分器公式是

$$v_{ij} = f(r_{ij}, b_j) \quad (4)$$

其中  $b_j$  是最可能吸引低分被试的干扰项。如果被试  $i$  在项目  $j$  上选择最可能吸引低分被试的干扰项, 那么  $r_{ij} = b_j$ , 此时记  $v_{ij} = 1$ ; 否则  $v_{ij} = 0$ 。SNC 假设干扰项与总分有高度的负相关性, 其计分器公式是

$$v_{ij} = f(r_{ij}, c_j) \quad (5)$$

其中  $c_j$  是与总分有最大负点二列相关的干扰项。如果被试  $i$  在项目  $j$  上选择与总分有最大负点二列相关的干扰项, 那么  $r_{ij} = c_j$ , 此时令  $v_{ij} = 1$ ; 否则  $v_{ij} = 0$ 。采用上述方法计算得到  $U$  和  $V$  后, 利用  $U$  可得到被试“强”项特质的诊断信息, 利用  $V$  可得到被试“弱”项特质的诊断信息。

干扰项得分反应的数据扩张法是一种从多项选择题中分离干扰项诊断信息的方法, 该方法采用不同计分器将正确反应和干扰项的反应分开计分、分开计算, 分别得到正确反应得分矩阵和干扰项得分反应矩阵, 最后利用正确反应得分矩阵和干扰项得分反应矩阵分别得到被试的“强”项特质的诊断信息和“弱”项特质的诊断信息。干扰项得分反应的数据扩张法的优点在于 (1) 方法直观。它是单维测验 1-0 计分方式的一种推广, 它既可以对正确反应选项采用 1-0 计分, 也可以对干扰项采用 1-0 计分。 (2) 应用灵活, 可以根据诊断目标来确定计分方式。例如, 如果仅仅关心被试的

“强”项特质的诊断信息, 那么只需对正确反应选项采用 1-0 计分即可, 如果只关心被试的“弱”项特质的诊断信息, 那么只需对“弱”项的干扰项采用 1-0 计分。实际上, 对所有的干扰项和正确答案选项都可以分别采用这种计分方式。 (3) 适用范围比较广。几乎所有的多项选择题都可以采用这种方法进行计分, 如 OMC 项目和选项对应的属性向量与被试知识状态相结合编制的多项选择题都可以采用该方法。干扰项得分反应的数据扩张法的不足之处在于 (1) CAK、PIK 和 SNC 是简单的计分器, 适合从单维测验中提取多维诊断信息, 但边际可靠性 (marginal reliability, Luecht, 2007) 比较差。认知诊断测验是多维测验, 应用该方法分析认知诊断测验时, 需要重新开发复杂的计分器, 复杂的计分器可以很容易从概念上扩展 CAK、PIK 和 SNC 得到 (Luecht, 2001, 2005)。 (2) 干扰项得分反应的数据扩张法对项目的属性及属性层级结构不做要求, 用该计分方法适合用 IRT 来诊断信息, 用 CDMs 来诊断信息时一般不用此方法。 (3) 在对“弱”项的干扰项计时需要先确定哪一个干扰项为“弱”项, 为确定“弱”项需要考查多个项目。这样, 这种计分方法需要较多的测验项目, 这一点在测验中往往是不现实的。

### 3.2 多项选择题的多级计分形式及其 CDMs

如果能包含干扰项诊断信息的多项选择题采用的是多级计分方式 (包括称名反应计分, 漆书青, 戴海崎, 丁树良, 2002), 那么需要用多级计分的 CDMs 来分析反应数据 (Rupp & Templin, 2008; Gu, 2011)。例如, Briggs 和 Alonzo (2009) 对 OMC 项目采用的是多级计分, 因此需要选择适用于分析多级计分观察反应数据的 CDMs 来分析 OMC 项目; Templin 等 (2008)、de la Torre (2009) 和 Gu (2011) 对选项对应的属性向量与被试知识状态相结合编制的多项选择题采用的是称名反应计分方式, 因此需要用称名反应认知诊断模型 (称名反应认知诊断模型是一种用于分析称名反应数据的验证性潜在分类模型, Templin et al., 2008) 来分析此类数据。下面分别介绍适合分析 OMC 项目和选项对应的属性向量与被试知识状态相结合编制的多项选择题的认知诊断模型。

#### 3.2.1 分析 OMC 项目反应的 CDMs

编制 OMC 项目的目的是将多项选择题的选项与待测量的被试的知识学习发展水平对应起来以诊断被试的学习发展水平所处的阶段, 进而为

教师能有的放矢地进行教学提供有用的信息。Briggs 和 Alonzo (2009)选择 AHM 来分析 OMC 项目的观察反应数据。一方面是因为 OMC 项目的选项反映了学习发展水平之间的层级关系,这一点与属性层级方法(AHM, Leighton, Gierl, & Hunka, 2004)的理论出发点吻合。AHM 的理论出发点是测验所考查的属性之间是有序的层级结构关系。另一方面是因为 AHM 适用于分析观察反应是二级计分或多级计分的数据(Rupp & Templin, 2008)。需要指出的是 AHM 分析 OMC 项目的过程比分析传统的二值计分的多项选择题的过程复杂。

用 AHM 分析 OMC 项目反应的具体步骤如下。

第一步,将学习发展各个水平的知识的定性描述标定成属性集,并确定属性之间的层级关系。确定属性间的层级关系的推理过程如下,因为在 OMC 项目中每个学习发展水平对应的属性集构成一个属性模式,又因为学习发展水平之间具有层级关系,所以各学习发展水平对应的属性模式之间必然具有层级关系,根据属性模式之间的层级关系就可以确定属性间的层级关系。下面用一个例子来说明确定属性间的层级关系的过程。

例 1 (Briggs & Alonzo, 2009) 某知识的学习有 5 个发展水平,这 5 个水平的知识可标定成 4 个属性集。假设这 4 个属性集包含 4 个属性,分别记为  $A_1, A_2, A_3$  和  $A_4$ ,又假设各发展水平对应的属性集分别为水平 1:没有相应的属性(未掌握任何相关知识);水平 2: $A_1$ ;水平 3: $A_1, A_2$ ;水平 4: $A_1, A_2, A_3$ ;水平 5: $A_1, A_2, A_3, A_4$ 。显然,这 5 个水平的属性模式之间具有层级关系,根据属性模式之间的层级关系可以确定这 4 个属性之间是直线型层级结构关系(Leighton et al., 2004): $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4$ ,箭头方向表示从低层级属性到高层级属性, $A_1$ 是 $A_2$ 的先决属性, $A_1$ 和 $A_2$ 是 $A_3$ 的先决属性, $A_1, A_2$ 和 $A_3$ 是 $A_4$ 的先决属性,被试只有掌握了低层级的属性才能掌握更高层级的属性。

第二步,根据属性间的层级关系得到反映属性之间直接关系的邻接矩阵  $A$  和反映属性之间直接和间接关系的可达矩阵  $R$ 。例 1 的邻接矩阵  $A$  和可达矩阵  $R$  分别是

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, R = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

第三步,由可达矩阵  $R$  经过布尔(Boolean)运算得到简化矩阵  $Q_r$ (Leighton et al., 2004)。 $Q_r$  的每一列表示两层含义,一是表示一类项目,二是表示一类被试。 $Q_r$  阵的列加上全零列是学生矩阵  $Q_s$ (丁树良,汪文义,罗芬,2012)。

第四步,确定合适的测验矩阵  $Q_t$ 。Leighton 等(2004)以及 Briggs 和 Alonzo (2009)将  $Q_r$  矩阵作为测验矩阵。但是,由于  $Q_r$  阵的列数取决于所测属性个数以及属性间的层级关系,当属性个数很多时,测验长度会很长。因此在实际测验中不可能根据  $Q_r$  阵对应的所有列编制项目,而是用  $Q_r$  阵中部分列组成测验矩阵  $Q_t$ (丁树良,祝玉芳,林海菁,蔡艳,2009)。在二级计分的多项选择题测验中, $Q_r$  矩阵是一个  $K \times J$  矩阵( $K$  为属性数, $J$  为项目类个数)。OMC 项目采用的是多级计分(根据选项计分,属于称名反应计分),若将每个选项与属性的关联关系单独列为一列(Briggs & Alonzo, 2009),即将每个选项看成是一个项目,那么其属性向量对应  $Q_t$  阵中的一列,这样  $Q_t$  是一个  $K \times (\sum_{j=1}^J H_j)$  矩阵( $H_j$  是项目  $j$  的选项个数)。由于 OMC 项目将每一个选项都视为一个项目,因此它可以减少测验项目数量(Templin et al., 2008)。表 2 是例 1 的一个测验  $Q_t$  阵的一部分,项目 2 和项目 3 各有 5 个选项,2A 表示项目 2 的选项 A,2A 的属性向量是(1,1,1,0)'。

表 2 例 1 的  $Q_t$  阵的一部分<sup>1</sup>

|       | 2A | 2B | 2C | 2D | 2E | 3A | 3B | 3C | 3D | 3E |
|-------|----|----|----|----|----|----|----|----|----|----|
| $A_1$ | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  |
| $A_2$ | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0  | 1  |
| $A_3$ | 1  | 0  | 0  | 1  | 1  | 0  | 1  | 0  | 0  | 0  |
| $A_4$ | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| 水平 4  | 4  | 3  | 3  | 5  | 4  | 2  | 4  | 3  | 1  | 3  |

注:列是一部分的 OMC 项目,最后一行是选项对应的知识水平,其它行表示属性。

(资料来源:Briggs & Alonzo, 2009)

第五步,计算被试在每一个 OMC 项目上的

<sup>1</sup> 原表头是“ESS 属性层级的  $Q_r$  矩阵的一部分”,此处改为“例 1 的  $Q_t$  阵的一部分”。原文作者认为  $Q_r$  是测验矩阵,其实不然。丁树良等(2009)建议用  $Q_t$  表示测验  $Q$  阵,以免混淆。故此处作了相应的修改。

期望反应的概率分布。设测验矩阵为  $Q$ , 被试  $i(1 \leq i \leq I)$  的属性掌握模式为  $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})'$ , 项目  $j(1 \leq j \leq J)$  的选项  $h(1 \leq h \leq H_j)$  的属性向量为  $q_{jh} = (q_{jh1}, q_{jh2}, \dots, q_{jhK})'$ 。记  $\alpha_i = q_{jh}$ , 若  $\alpha_{ik} = q_{jhk}, k=1, 2, \dots, K$  成立; 记  $\alpha_i < q_{jh}$ , 若  $\alpha_{ik} < q_{jhk}$  至少对某个  $k$  成立。下面分三种情况来说明如何计算被试在每一个 OMC 项目的期望反应的概率分布。情况一:  $\alpha_i < q_{jh}$ , 对所有的  $h(1 \leq h \leq H_j)$ 。此时出现“地板效应”, 被试  $i$  的期望反应是等可能地从  $H_j$  个选项选择一个选项作为答案, 因此这时被试  $i$  在项目  $j$  上的期望反应的概率分布为  $(H_j^{-1}, \dots, H_j^{-1})$ 。下面的表 3 中的被试 1 和被试 2 在项目 2 上的反应就属于这种情况。情况二: 存在  $h$  使得  $\alpha_i = q_{jh}$ 。在理想情况下被试  $i$  能够正确作答项目  $j$  (即选择选项  $h$ , 因为选项  $h$  的属性向量与被试  $i$  的属性掌握模式匹配)。假如项目  $j$  的选项中有  $\tilde{H}_j$  个选项与被试  $i$  的属性掌握模式匹配, 此时被试  $i$  会等可能地从  $\tilde{H}_j$  个(不妨设为前  $\tilde{H}_j$  个)选项选择一个选项作为答案, 这样被试  $i$  在项目  $j$  上的期望反应的概率分布为  $(\tilde{H}_j^{-1}, \dots, \tilde{H}_j^{-1}, 0, \dots, 0)$ 。例如表 3 中的被试 3 在项目 2 和项目 3 上的反应都属于这种情况。情况三: 存在  $h$  使得  $\alpha_i > q_{jh}$ , 但不存在  $h'$  使得  $\alpha_i = q_{jh'}$ 。此时出现“天花板效应”, 被试  $i$  的期望反应是从  $\hat{H}_j$  个(不妨设为前  $\hat{H}_j$  个)同时为最高水平的选项选择一个作为答案, 这样被试  $i$  在项目  $j$  上的期望反应的概率分布为  $(\hat{H}_j^{-1}, \dots, \hat{H}_j^{-1}, 0, \dots, 0)$ 。表 3 中被试 5 在项目 3 上的反应属于情况三的情形。

第六步, 采用 AHM 的分类方法对被试的观察反应模式进行诊断分类, 得到被试的知识状态。AHM 主要有两大类分类方法, 一种是参数方

法, 如最大相似概率法和根据观察反应模式是否逻辑包含期望反应模式的分类法(Leighton et al., 2004), 另一种是非参数方法, 如神经网络方法(Gierl, Wang, & Zhou, 2008)。

Briggs 和 Alonzo (2009)指出与 IRT 模型相比较, 采用 AHM 分析 OMC 项目利用了干扰项的附加诊断信息, 提高了诊断的准确性, 其方法为开发能同时分析正确答案选项和干扰项的数据的认知诊断模型提供了思路。用 AHM 分析 OMC 项目有其局限性。(1) AHM 需要在编制测验前确定细粒度的属性及属性间的层级关系, 再根据属性及其层级关系指导测验的编制。OMC 项目编制前确定的是待测验知识的水平阶段, OMC 项目的选项与知识的水平阶段对应。当用 AHM 分析 OMC 项目时, 需要将 OMC 项目选项对应的学习发展各水平所考查的知识转化成属性, 但是并非所有的学习发展过程的知识都可以具体化成细粒度的属性, 因此, 对于不能将学习发展过程的知识具体化成细粒度属性的 OMC 项目就不适合用 AHM 来分析, 也不适合用其它的认知诊断模型来分析。(2) 尽管 Briggs 和 Alonzo (2009)的研究表明用 AHM 分析 OMC 项目的诊断准确性比较高, 但是当 OMC 项目测验的属性个数比较多时, AHM 的诊断准确率会很低。理论上, 其它注重层级关系且适用于分析观察反应是多级计分数据的认知诊断模型也可以用于分析 OMC 项目。如徐冬波、蔡艳、戴海琦和丁树良(2010)开发的多级计分 DINA (Polytomous DINA, P-DINA)模型、张淑梅(2012)提出的广义多级计分认知诊断模型(Generalized Polytomous Cognitive Diagnosis Model, GPCDM)以及 Sun, Xin, Zhang 和 de la Torre (2013)提出的多级评分广义距离判别法(Polytomous of the Generalized Distance Discriminating, GDD-P), 模

表 3 OMC 项目选项的期望反应

| 被试 | 期望反应                                                            |                                   | 属性掌握模式<br>[A <sub>1</sub> A <sub>2</sub> A <sub>3</sub> A <sub>4</sub> ] | 水平 |
|----|-----------------------------------------------------------------|-----------------------------------|--------------------------------------------------------------------------|----|
|    | 项目 2                                                            | 项目 3                              |                                                                          |    |
| 1  | $[\frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5}]$ | [00010]                           | [0 0 0 0]                                                                | 1  |
| 2  | $[\frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5}]$ | [10000]                           | [1 0 0 0]                                                                | 2  |
| 3  | $[0 \frac{1}{2} \frac{1}{2} 0 0]$                               | $[0 0 \frac{1}{2} 0 \frac{1}{2}]$ | [1 1 0 0]                                                                | 3  |
| 4  | $[\frac{1}{2} 0 0 0 \frac{1}{2}]$                               | [01000]                           | [1 1 1 0]                                                                | 4  |
| 5  | [00010]                                                         | [01000]                           | [1 1 1 1]                                                                | 5  |

(资料来源: Briggs & Alonzo, 2009)

拟研究表明这些多级评分 CDM 有较高的诊断准确率, 特别是 GPCDM 和 GDD-P 都有很高的诊断准确率, 但是应用到 OMC 测验分析的实际效果如何有待验证。

### 3.2.2 MC-DINA 模型

MC-DINA 模型(de la Torre, 2009)是用于分析多项选择题的称名反应认知诊断模型。MC-DINA 模型具有非补偿性和连接性的特征。所谓非补偿性特征是指被试未掌握某个属性不能由其掌握的其它属性来补偿。所谓连接性是指被试掌握项目所有属性, 可以较高的概率正确作答, 而即使只有一个项目属性未掌握, 也将大大降低答对概率。MC-DINA 模型的连接性适用于项目的所有编码选项, 被试选择某个编码选项需要被试掌握了该编码选项涉及的所有属性。

在 MC-DINA 模型框架下, 属性掌握模式是  $\alpha_i$  的被试在项目  $j$  上选择选项  $h$  的概率是

$$P_{jh}(\alpha_i) = P(X_{ij} = h | \alpha_i) = P(X_{ij} = h | g_{ij} = g) \triangleq P_j(h | g) \quad (6)$$

$P_j(h|g)$ 也可说成是第  $g$  组被试在项目  $j$  上选择选项  $h$  的概率。对于有  $J$  个项目, 且项目  $j$  有  $H_j$  个选项的测验, MC-DINA 模型待估计参数有  $\sum_{j=1}^J H_j(H_j^* + 1)$  个 ( $H_j^*$  为编码选项个数)。注意到约束条件  $\sum_{h=1}^{H_j} P_j(h | g) = 1$ , 故实际的待估参数有  $\sum_{j=1}^J H_j^* + J$  个。采用极大似然法得到  $P_j(h|g)$  的估计是

$$\hat{P}_j(h | g) = \frac{I_j(h | g)}{\sum_{h=1}^{H_j} I_j(h | g)} \quad (7)$$

其中  $I_j(h|g)$  是第  $g$  组被试在项目  $j$  上选择选项  $h$  的期望被试量。需要说明的是用  $\hat{P}_j(h | g)$  估计  $P_j(h|g)$  时, 需要充分大的样本量才能准确估计。

实际上, MC-DINA 模型是 DINA 模型的拓展。MC-DINA 模型分类考虑了干扰项的附加诊断信息, 而 DINA 模型则将正确选项外的选项视为没有任何区别, 不考虑其为诊断提供的信息。更具体地讲, DINA 模型将被试在每个项目上仅分为掌握者与未掌握者, MC-DINA 模型则利用干扰项信息将未掌握者再进一步分组以达到将未掌握者更准确地分类的目的。下面以一道有 2 个编码干扰项和 1 个非编码干扰项的多项选择题为例来说明 MC-DINA 模型是如何将未掌握者再分类的。

对于正确答案选项, 要估计的参数有 4 个: 期望反应选项是正确答案选项的被试选择正确答案选项的概率是  $P_{master}(key)$ ; 期望反应选项是两个编码选项的被试选择正确答案选项的概率分别是  $P_{D1}(key)$  和  $P_{D2}(key)$ ; 属性掌握模式与任何一个编码选项对应的属性向量都不匹配的被试选择正确答案选项的概率是  $P_0(key)$ 。  $P_{master}(key)$  相当于 DINA 模型中的  $1-s_j$  ( $s_j$  是项目  $j$  的失误参数),  $P_{D1}(key)$ ,  $P_{D2}(key)$  和  $P_0(key)$  的加权和相当于 DINA 模型的猜测参数。由此可见, MC-DINA 模型将 DINA 模型的猜测参数分解成 3 个参数, 也就是将未掌握组被试分成了 3 类, 这种分类较 DINA 模型更准确。由于  $P_0(key)$ 、 $P_{D1}(key)$ 、 $P_{D2}(key)$  和  $P_{master}(key)$  分别是不同知识状态被试组选择正确答案选项的概率, 这 4 个概率值可能不等, 这说明 MC-DINA 模型允许不同知识状态的被试组选择同一选项的概率不同, 且掌握了选项属性向量中更多属性的被试组比掌握较少属性的被试组选择该选项的概率高。

de la Torre (2009) 和 Gu (2011) 的模拟研究结果显示, MC-DINA 模型的判准率明显地高于 DINA 模型的判准率, 究其原因 MC-DINA 模型包含了干扰项的附加诊断信息。当 MC-DINA 模型的判准率高于 DINA 模型的判准率时, 一方面说明增加干扰项的附加诊断信息后提高了学生属性掌握模式的估计精度, 另一方面也说明编制具有诊断能力的干扰项, 应用 MC-DINA 模型从干扰项中获得附加诊断信息提高诊断效率是可能的, 也是可行的。对于 MC-DINA 模型在实证方面的研究, Gu (2011) 分别用 DINA 模型和 MC-DINA 模型分析了大型成就测验的实测数据 (2006 年 Ontario 的 6 年级数学评估), 但因为实测数据的测验不是根据认知诊断目标编制, 实测数据的分析结果显示干扰项提供的信息很少, 只有部分干扰项的诊断能力达到可接受程度, 这说明对于分析不是结合选项属性向量与知识状态编制的多项选择题测验, MC-DINA 模型效果不理想。

de la Torre (2009) 和 Gu (2011) 的结论是当 DINA 模型和 MC-DINA 模型不含层级关系时研究得到的。如果对 DINA 模型和 MC-DINA 模型考虑属性间的层级关系或对属性间关系进行约束 (涂冬波, 蔡艳, 戴海琦, 2013; 颜远海, 丁树良, 汪文义, 2011), 那么在各种模拟条件下, MC-DINA 模型的模

拟研究结果是否好于 DINA 模型?增加干扰项的附加诊断信息能否提高学生属性掌握模式的判准率?这些问题都需要研究验证。

### 3.2.3 称名反应对数线性认知诊断模型

MC-DINA 模型是一种适用于分析选项对应的属性向量与知识状态结合编制的多项选择题的称名反应认知诊断模型。由于 MC-DINA 模型只适用于属性间的非补偿性情形,且由于参数数量有限,模型不能直接解释每个属性与每个反应选项之间的关系,从而限制了模型参数解释的一般性。因此,有必要构建一个更有弹性、更广义的多项选择题认知诊断模型。Templin 等人(2008)以广义诊断模型(the General Diagnostic Model, Von Davier, 2005)和对数线性认知诊断模型(Log-Linear Cognitive Diagnostic Model, LCDM, Henson, Templin, & Willse, 2009)为基础开发出称名反应对数线性认知诊断模型(Nominal Response LCDM, NR-LCDM)。NR-LCDM 是 MC-DINA 模型的拓展(Templin et al., 2008),它既适用于属性间的非补偿性情形也适用于补偿性情形。

现介绍 NR-LCDM。设测验矩阵  $Q_t = (q_{kj})_{K \times J}$ ,  $K$  为属性个数,  $J$  为项目个数。 $K$  个属性把被试分成  $C=2^K$  个潜在类,若设第  $c$  类被试占总被试量的比例是  $\eta_c$ ,则有  $\sum_{c=1}^{2^K} \eta_c = 1$ 。以  $M_j$  表示项目  $j$  的选项个数,  $\pi_{j,c,m_j}$  表示第  $c$  类被试在项目  $j$  上选择选项  $m_j$  的概率,则  $\sum_{m_j \in M_j} \pi_{j,c,m_j} = 1, \forall j, c$ 。易知被试在测验中的作答反应为  $x = (x_1, \dots, x_J)$  的概率是

$$P(X=x) = \sum_{c=1}^C \eta_c \prod_{j=1}^J \left[ \prod_{m_j \in M_j} \pi_{j,c,m_j}^{I(x_j=m_j)} \right] \quad (8)$$

式中  $I(\cdot)$  是指示函数。将  $\pi_{j,c,m_j}$  参数化,

$$\pi_{j,c,m_j} = P(X_j = m_j | \alpha_c) = \frac{\exp(\lambda_{0,j,m_j} + \lambda_{j,m_j}^T h(\alpha_c, q_j))}{\sum_{m_j \in M_j} \exp(\lambda_{0,j,m_j} + \lambda_{j,m_j}^T h(\alpha_c, q_j))} \quad (9)$$

其中,

$$\lambda_{j,m_j}^T h(\alpha_c, q_j) = \sum_{k=1}^K \lambda_{1,j,k,m_j} (\alpha_{ck} q_{kj}) + \sum_{k=1}^{K-1} \sum_{l=k+1}^K \lambda_{2,j,k,l,m_j} (\alpha_{ck} \alpha_{cl} q_{kj} q_{lj}) + \dots \quad (10)$$

$\alpha_c = (\alpha_{c1}, \dots, \alpha_{cK})'$  是第  $c$  类被试的属性掌握模式;项目  $j$  的每个选项有三类参数:截距参数  $\lambda_{0,j,m_j}$ , 属性  $k$  的主效应参数  $\lambda_{1,j,k,m_j}$ , 属性  $k$  与属性  $l$  的交互效应参数  $\lambda_{2,j,k,l,m_j}$ ;省略号表示略去了两个以上属性的交互效应。对上述各参数的解释可参考 Templin 等(2008)和 Gu (2011, P.20)。由于在将  $\pi_{j,c,m_j}$  参数化时采用了对数线性模型的方法,故该认知诊断模型被称为 NR-LCDM。为了使模型可识别,还需要一些必要的约束条件,如要求  $\sum_{m_j \in M_j} \lambda_{0,j,m_j} = 0, \forall j$ ;  $\sum_{m_j \in M_j} \lambda_{1,j,k,m_j} = 0, \forall j, k$  等。

NR-LCDM 是一个框架十分宽广的模型。当需要考虑属性间关系时,模型除了包括含主效应参数的项外,还需要包括含交互效应参数的项。考虑了含交互效应参数项的模型是非补偿性模型。MC-DINA 模型是 NR-LCDM 框架下的一种具有非补偿性特征的模型。由于考虑了交互效应参数项,这使得非补偿性 NR-LCDM 的参数很多,要估计这些参数需要有很大的样本总量,而且需要有足够的具有不同属性掌握模式的被试量,这在实际测验中是个很难的问题。基于这一点,非补偿性的 NR-LCDM 的实用性受到了局限。当不需要考虑属性间关系时,模型只包含主效应参数的项,不包含交互效应参数的项,这种模型是补偿性模型,具有较大的实用性,如 NR-cRUM。

NR-LCDM 通过建立称名反应类别以获得多项选择题的附加诊断信息,其诊断信息比二级评分 CDMs 更丰富,对被试的正确诊断率更高。NR-LCDM 克服了 MC-DINA 模型的一些缺陷,其参数的可解释性更强,是一个更广义的模型,从理论上讲, NR-LCDM 适用范围更广。但是实践上, NR-LCDM 对大样本量的要求导致其适用范围仅局限于大型测验,不适合于小样本测验和随堂测验。Templin et al. (2008)的模拟研究证明,当样本量很小的时候,二级评分诊断模型比称名反应诊断模型的诊断效果更好,但当样本量增大(如 5000 个被试)时,两类模型的差异就非常小,这说明只有在更大样本量的时候, NR-LCDM 才能提高分类判准率。因此, Templin 等人(2008)建议只有在很大样本的情形时应用 NR-LCDM。

#### 4 多项选择题认知诊断研究的展望

在认知诊断测验中采用多项选择题而不仅仅对是否选择正确答案选项进行评分, 这虽然简便易行, 但是容易丢失信息, 导致不能准确地估计被试的知识状态。为了使多项选择题的认知诊断测验能准确地估计出被试的知识状态, 需要多项选择题能尽可能多地提供诊断信息, 还需要能选择合适的计分方式和心理计量模型将这些诊断信息提取出来。近十几年来, 虽然有学者探讨了能尽可能多地提供诊断信息的多项选择题编制方法和适用于多项选择题的计分方式以及 CDMs 的开发, 但是还可以从以下几个方面进一步完善。

##### 4.1 能提取多项选择题干扰项信息的认知诊断测验蓝图设计问题

Gorin (2007) 强调认知诊断测验构造的重要性不亚于构造单个项目的重要性, 因为认知诊断测验蓝图的优劣直接影响认知诊断评估的分类效果。丁树良、杨淑群和汪文义(2010)以及丁树良、汪文义和杨淑群(2011) 探讨了 1-0 评分方式的认知诊断测验蓝图设计问题, 认为只要测验矩阵中包含可达矩阵就可以实现理想掌握模式与理想反应模式的一一对应, 达到有效区分所有被试类的目的。Ding, Wang 和 Luo (2013) 以及丁树良、罗芬和汪文义(2013) 还探讨了二级计分和多级计分认知诊断测验蓝图设计问题, 当测验矩阵是完备 Q 矩阵(所谓完备 Q 矩阵是使知识状态和期望反应一一对应且所含列数最少的测验 Q 矩阵)时, 能以最少的项目提供最丰富的诊断信息。在多项选择题的认知诊断测验中, 当把干扰项对应的属性向量与知识状态相结合编制多项选择题时, 干扰项相当于项目, 因而, 在理论上, 使用完备 Q 矩阵进行测验蓝图设计可以减少多项选择题的个数。因此设计包括干扰项在内的多项选择题的认知诊断测验蓝图是一个有重要理论价值和实用价值的研究课题。

##### 4.2 干扰项编制方法研究

特殊编制干扰项是能从干扰项中提取诊断信息的前提。当前关于干扰项编制方法多以被试因处于不同认知发展阶段而掌握的知识不完整 (Briggs et al., 2006; Tamir, 1991) 或干扰项能够提取被试的错误知识 (Luecht, 2007) 为理论基础, 结合认知发展水平或知识状态类型编制干扰项。这

些编制干扰项的方法都有适用范围和限制条件, 因此探索一种更一般化、能适用于更多学科甚至所有学科问题的多项选择题和干扰项的编制方法是今后研究的重点之一。

##### 4.3 干扰项个数问题

许多学者对传统测验中多项选择题的选项个数问题展开研究, 一方面是因为一个项目编写 3 个以上的有实际意义的干扰项很困难, 也很耗时, 另一方面是因为多项选择题的选项数量会影响到测验的难度、信度和效度。有关传统测验中多项选择题选项个数的研究证明, 当把多项选择题的选项从 4 个减到 3 个时, 项目难度就会稍微变小, 信度会稍微提高, 此说明选项的减少对测验的相关指标影响很小 (Rodriguez, 2005)。也有研究证明, 减少选项数量可以在固定时间内呈现更多的项目, 因而涵盖的内容更广, 信度也更高, 对测验分数的心理计量也不会产生影响 (Haladyna & Downing, 1989, 1993; Haladyna, Downing, & Rodriguez, 2002)。还有研究证明, 每个项目选项数量减少而测验长度成比例加长会提高高水平被试的测试效率, 同时会降低低水平被试的测试效率 (Lord, 1977)。然而, 在认知诊断测验中很少有学者研究多项选择题选项的个数问题。有关传统测验的研究结论是否可以推论到认知诊断测验的情形? 在认知诊断评估中, 是否存在用最少的干扰项提供尽可能多的诊断信息的最优设计方案? 这些应是今后研究的问题。

##### 4.4 多项选择题的选项与解题策略结合问题

影响认知诊断估计准确性的因素很多, 能否最大化地提取被试的诊断信息和是否考虑解题策略的多样性是其中两个重要的影响因素。有研究证明, 利用多项选择题干扰项的诊断信息可以提高对被试知识状态的估计准确性 (Templin et al., 2008; Briggs & Alonzo, 2009; de la Torre, 2009; Gu, 2011); 也有研究证明, 考虑解题策略的多样性也可以提高对被试知识状态的估计准确性 (de la Torre & Douglas, 2008; 涂冬波, 蔡艳, 戴海琦, 丁树良, 2012)。但是, 如果在多项选择题的认知诊断测验中综合考虑干扰项的诊断信息和解题策略的多样性这两个影响因素, 那么就需要将解题策略与多项选择题的选项结合以寻求既能使干扰项提供诊断信息又能诊断被试的解题策略的多项选择题编制方法。因此, 将多项选择题的选项与解题策略结合

问题应该是今后一个有价值的研究内容。

#### 4.5 多项选择题认知诊断模型的选择与开发

为了能从多项选择题中提取尽可能多的诊断信息,选择或开发合适的 CDMs 很关键。选择 CDMs 分析多项选择题测验数据时要综合考虑 CDMs 的适用条件。其一,根据观察反应数据类型选择二级计分 CDMs、多级计分 CDMs (包含称名反应 CDMs)。比如, AHM 可以用于分析二级计分数据和多级计分数据, MC-DINA 模型和 NR-LCDM 用于分析称名反应数据(Gu, 2011)。其二,根据项目属性间关系选择补偿性 CDMs 或非补偿性 CDMs。比如, AHM 和 MC-DINA 是非补偿性 CDMs, NR-LCDM 既可以用于补偿型情形,也可以用于非补偿型情形,但是用于非补偿型情形时要估计的参数太多(Rupp & Templin, 2008; Gu, 2011)。其三,根据属性间层级关系选择“含层级关系的认知诊断模型”或“不含层级关系的认知诊断模型”(涂冬波,蔡艳,戴海琦,2013)。比如, AHM 是“含层级关系的认知诊断模型”, MC-DINA 和 NR-LCDM 是“不含层级关系的认知诊断模型”。一般情况下,当属性层级关系不能正确界定(或属性间层级关系不容易界定)时,建议选择“不含层级关系的认知诊断模型”。但是,当属性层级关系能正确界定(或属性间层级关系比较容易界定)时,是否“含层级关系的认知诊断模型”比“不含层级关系的认知诊断模型”好?比如, OMC 测验的属性是从学生理解的发展水平对应的知识转化得到,因而比较容易界定属性间层级关系, AHM 分析 OMC 项目数据是否是较好的选择?这些问题目前都无定论。

目前分析多项选择题数据的 CDMs 都只适用于单策略解题情形,而考虑解题策略多样性的 CDMs 又没有考虑干扰项的诊断信息,如果综合考虑干扰项的诊断信息和解题策略的多样性,则需要开发适用于分析此类测验数据的 CDMs,开发多策略多项选择 CDMs 应该是今后一个有价值的研究内容。

#### 参考文献

- 丁树良,罗芬,汪文义.(2013). 认知诊断分类中心的确定. *心理学探新*, 33(5), 396-401.
- 丁树良,汪文义,罗芬.(2012). 认知诊断中 Q 矩阵和 Q 矩阵理论. *江西师范大学学报(自然科学版)*, 36(5), 441-445.

- 丁树良,汪文义,杨淑群.(2011). 认知诊断测验蓝图的设计. *心理科学*, 34(2), 258-265.
- 丁树良,杨淑群,汪文义.(2010). 可达矩阵在认知诊断测验编制中的重要作用. *江西师范大学学报(自然科学版)*, 34(5), 490-494.
- 丁树良,祝玉芳,林海菁,蔡艳.(2009). Tatsuoka Q 矩阵理论的修正. *心理学报*, 41(2), 175-181.
- 漆书青,戴海琦,丁树良.(2002). *现代教育与心理测量学原理*. 北京:高等教育出版社.
- 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.
- 涂冬波,蔡艳,戴海琦,丁树良.(2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. *心理学报*, 42(10), 1011-1020.
- 涂冬波,蔡艳,戴海琦,丁树良.(2012). 一种多策略认知诊断方法: MSCD 方法的开发. *心理学报*, 44(11), 1547-1553.
- 颜远海,丁树良,汪文义.(2011). 影响 AHM 与 DINA 诊断准确率的因子研究. *江西师范大学学报(自然科学版)*, 35(6), 640-645.
- 张淑梅.(2012). *多级评分的认知诊断模型研究*. 博士学位论文. 北京师范大学.
- Airasian, P. W. (2001). *Classroom assessment* (4th ed.). Boston: McGraw-Hill.
- Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, 93, 389-421.
- Briggs, D. C., & Alonzo, A. C. (2009, June). *The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression*. Paper presented at the the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1), 33-63.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.
- de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33, 163-183.
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73(4), 595-624.
- Ding, S. L., Wang, W. Y., & Luo, F. (2013). *How to design polytomous cognitive diagnostic test*. Paper presented at the International Meeting Psychometric Society (IMPS),

- Arnhem, the Netherlands.
- Fu, J., & Li, Y. (2008). Cognitively diagnostic psychometric models: An integrative review. ETS research report.
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). Cambridge, UK: Cambridge University Press.
- Gierl, M. J., Wang, C. J., & Zhou, J. W. (2008). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *The Journal of Technology, Learning and Assessment*, 6(6). Retrieved June 8, 2009 from <http://www.jtla.org>
- Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton (Ed.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173–201). Cambridge, UK: Cambridge University Press.
- Gu, Z. M. (2011). *Maximizing the potential of multiple-choice items for Cognitive Diagnostic Assessment*. Unpublished Doctoral Dissertation, University of Toronto.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 333–352.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37–50.
- Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53, 999–1010.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice items-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 308–313.
- Hartz, S. M. (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.
- Henson, Robert. A., Templin, Jonathan L., & Willse, John T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191–210.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141–151.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- King, Kelly. V., Gardner, Doug A., Zucker, Sasha, & Jorgensen, Margaret. A. (2004). The Distractor Rationale Taxonomy: Enhancing Multiple-Choice Items in Reading and Mathematics. Pearson assessment report.
- Kubiszyn, T., & Borich, G. D. (1987). *Educational testing and measurement: Classroom application and practice* (2nd ed.). Glenview, IL: Scott, Foresman, and Company.
- 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.
- Lin, Jie, Chu, Kang lee, & Meng, Ying. (2010, April). *Distractor Rationale Taxonomy: Diagnostic Assessment of Reading with Ordered Multiple-Choice Items*. Paper presented at the the annual meeting of American Educational Research Association, Denver, Colorado.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall, Inc.
- Liu, Ou Lydia, Lee, Hee Sun, & Linn, Marcia C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, 16, 164–184.
- Lord, F. M. (1977). Optimal number of choices per item—A comparison of four approaches. *Journal of Educational Measurement*, 14(1), 33–38.
- Luecht, R. M. (2001, April). *Capturing, codifying and scoring complex data for innovative, computer-based items*. Paper presented at the the annual meeting of the National Council on Measurement in Education, Seattle.
- Luecht, R. M. (2005). Item analysis. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. London: Wiley.
- Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319–340). Cambridge, UK: Cambridge University Press.
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2, 99–120.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., & Glaser, R., (Eds.), Board on Testing and Assessment, Center for Education. Division of the Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Nitko, A. J. (2004). *Educational assessments of students* (4th ed.).

- Englewood Cliffs, NJ: Prentice Hall.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of Research. *Educational Measurement: Issues & Practice*, 24(2), 3–13.
- Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1155–1159.
- Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic models: A review of the current state of the art. *Measurement*, 6, 219–262.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.
- Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor driven assessment instruments. *Journal of Research in Science Teaching*, 35, 265–296.
- Sun, J. N., Xin, T., Zhang S. M., & de laTorre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*, 37(7), 503–521.
- Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education*, 19, 188–191.
- Templin, J. L., Henson, R. A., Rupp, A., Jang, E., & Ahmed, M. (2008, March). *Diagnostic models for nominal response data*. Paper presented at the Annual Meeting & Training Sessions of National Council for Measurement in Education, New York City.
- Treagust, D. F. (1995). Diagnostic assessment of students' science knowledge. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 327–346). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Treagust, D. F. (2006, September). Diagnostic assessment in science as a means to improving teaching, learning and retention. Paper presented at the the 2006 National UniServe Conference (Symposium of Science Teaching and Learning Research), Sydney, Australia.
- Von Davier, M. (2005). A general diagnostic model applied to language testing data. Educational Testing Service, Research Report, RR-05--16.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13, 181–208.

## Maximizing the Potential of Multiple-choice Items for Cognitive Diagnosis

LI Yu<sup>1</sup>; DING Shuliang<sup>2</sup>; TANG Xiaojuan<sup>3</sup>

(<sup>1</sup> School of Psychology, Jiangxi Normal University, Nanchang 330022, China)

(<sup>2</sup> College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

(<sup>3</sup> College of Mathematics and Information Science, Nanchang Hangkong University, Nanchang 330063, China)

**Abstract:** Multiple-Choice (MC) item format is commonly used in cognitive diagnosis assessment. To obtain much more diagnostic information from MC items, including the diagnostic information of the keys and the distractors, the methods to construct MC items are firstly introduced. To extract the diagnostic information of the keys and the distractors, some corresponding cognitive diagnosis models are also proposed. Finally, the article suggests that the researches on the methods to construct distractors and cognitive diagnosis models to extract diagnostic information from the distractors should be strengthened.

**Key words:** cognitive diagnosis; multiple choice items; distractors; diagnostic information