面向测试开发者、研究者及教师的试题编写技术

迈克尔·罗德里格兹　安东尼·阿尔巴诺　托马斯·哈拉代那(美)

本论文是 2010年在美国科罗拉多州丹佛市(Denver, Colorado)举行的全美教育测量学会 (National Council on Measurement in Education)年会上关于试题编写技术的培训课程(Training Sessions)的节选部分,论文原名为《从艺术到科学:测试开发员、研究员及教师的试题编写技术》。本文分为四部分,前两部分对单项选择题(multiple-choice items)及建构反应题(constructed-response items)的各种具体形式进行概述,第三部分对两种题型的试题编写准则进行评论,最后一部分内容包括试题的效度介绍,试题开发过程中收集定性的效度信息的模型,开发新题型以增强对测试构念的测量,以及提高测试的参与性,让所有的学生,包括那些有一般及严重认知缺陷的学生也能参加测试。本文旨在说明如何通过认真设计试题(这是进行测试的基础),完善试题编写技术,最大程度上增强测试内容的效度及测试分数的意义。

许多目前看来对测试命题质量至关重要或非常有效的试题编写准则都是应用于不同的考试情境中,并往往和大型测试项目相关。但是,这些编写准则几乎都没有经过实证研究。1917年美国陆军甲种测验、陆军乙种测验(the U. S Army Alpha and Army Beta tests)以及 20世纪 20年代的美国大学入学考试都开始使用单项选择题题型(DuBois, 1970)。单项选择题测试在具体的内容范围内可以获得简单而且有针对性的回答,由于在命题及管理方面的简单性,单项选择题成为与其他题型,如建构型笔试或口试(constructed written or oral responses)的交替使用的一种题型。在《教育测量》(Educational Measurement, 1951)第一版中,埃贝尔(Ebel)探讨了 20世纪上半叶试题编写技术的研究和发展,其中特别强调了单项选择题的广泛应用。

20世纪初以来,随着技术的发展,特别是计算机测试技术的应用,越来越多的新型试题模式开始出现,例如视频、声音、与考生的互动等。在很多方面,技术的发展带来了测试的改进,因为新的题型使得试题能更好地反映其要测量的构念(construct)。例如,要测试健康科学专业的学生以及医学从业人员是否掌握了进行复杂手术需要的技术,就

作者简介　迈克尔·罗德里格兹(Michael C. Rodriguez),博士,美国明尼苏达大学教育心理系,美国 ETS的 GRE考试的技术顾问主任。安东尼·阿尔巴诺(Anthony D. Albano),博士,美国明尼苏达大学教育心理系;明尼阿波利斯,明尼苏达州,美国,55455-0364。托马斯·哈拉代那(Thomas M. Haladyna),博士,美国亚利桑那州立大学荣誉退休教授,凤凰城,亚利桑那州,美国,85028。

本文由南开大学外国语学院王丽华翻译,本刊编辑部校译。

可以用视频,甚至某一解剖部位的三维模拟仿真来进行测试(Shanedling Van Heest, Rodriguez, Putnam, & Agel, 2010);而在进行数学或统计学测试时,试题中可以出现电子数据表或图表,从而更好地测量学生对相关知识与能力的掌握情况。

本文旨在讨论试题开发的几个重要方面,以提高测试以及测试分数的质量。前两部分对普通题型,如单项选择题及建构反应题进行概述,第三部分对两种题型的命题准则进行评论,最后一部分的内容包括试题的效度介绍,试题开发过程中收集定性的效度信息的模型,开发新题型以增强对测试构念的测量,以及提高试题的参与性,让所有的学生,包括那些有一般及严重认知缺陷的学生也能参加测试。所有内容的论述旨在通过精心设计试题(这是进行测量的基础),达到完善试题编写技术,最大程度上增强测试内容的效度及测试分数的意义。

一、单项选择题题型(MC Formats)

单项选择题的形式有很多种。最常见的单项选择题形式如下所述(Haladyna, Downing & Rodriguez, 2002)。这些是具有代表性的形式,还有一些试题编写者利用信息技术及计算机测试(computer-enabled testing)的优势,创造出很多新题型。

1. 传统单项选择题(ConventionalMC)

当描述分布时,标准偏差告诉我们:

A. 大多数分数的位置

B. 分布是否正常

C. 分数的分布范围

2. 二选一题(Alternate-Choice)

如果原始分数分布为正偏态,转化成 T分数会导致哪一类型分布?

A. 正态

B. 正偏态

3. 对错题(True-False)

如果 30%的考生对一个问题都答出了正确的答案,这道题的难度指数为.70。(正确或错误)

4. 多项对错题(Multiple True-False)

考虑以下提高测试分数信度的措施,并判断其正误。

A. 增加更多类似测试试题会增加测试分数的信度

B. 增大取样范围会增加测试分数的信度

C. 获得更多有关测试分数变异性的取样会增加测试分数的信度

D. 平衡肯定式及否定式题干的比重会增加测试分数的信度

5. 搭配题(Matching)

为右侧的术语在左侧找到对应描述项。

(1) score consistency	A. 系统误差
(2) test-wiseness	B. 随机误差
(3) score accuracy	C. 题目难度
(4) proportion correct	D. 题目区分度
(5) point-biserial correlation	E. 信度
	F. 效度

6. 根据上下文答题(Context-Dependent Item Set)

(1)哪一道试题的区分度最好? _____

(2)识别一道可以很容易改成判断正误题的题目。_____

(3)哪一道试题最简单? _____

(4)确定一道可以列入 3道最有效干扰项的题。_____

(5)确定一道最有可能有两个正确答案的题。_____

7. 复杂选择题(K型题)Complex Multiple-Choice (Type K)

以下哪些选项是有关标准参照测试中的

分数解释?

(1)约翰的分数高于班级平均分3个标准差。

(2)玛丽回答对了80%的试题。

(3)班级中80%的学生得分超过T分数45分。

(4)阿灵顿高中的数学平均分与区平均分持平。

(5)安东尼奥的5年级阅读达到熟练水平。

A. 第(1)、(3)项

B. 第(1)、(3)和(4)项

C. 第(2)、(5)项

D. 第(2)、(4)以及(5)项

E. 所有5项。

二、建构反应题题型(CR Formats)

建构反应题的形式多种多样,在以往文献中尚未发现连续、系统的研究。建构反应题与单项选择题不同,因为建构反应题要求考生组织或建构一份答案。Osterlind与Merz(1994年),Haladyna(1997年)描述了建构反应型试题的20种形式,包括短文写作(essays)、表格填空(grid-in responses)、研究报告(research papers)、简短回答(short answer items)、口头报告(oral reports)等;也可以是填空(fill-in-the-blank)和完形填空(cloze),但后两种形式的题型通常不推荐在测试中使用。建构反应题还是一些表现性评价中经常使用的形式,如学生档案袋、行为表现、表演以及实验等。这些评价要求有更广泛的评分准则,需要花费更多的时间去设计和准备。因此,这些题型不适合那些一经要求即可提供的测试(on-demand testing)。

建构反应题题型分类方式有多种,因为不同建构反应题在回答方式或评分过程上差异很大。在大规模的学业成就测验中,建构反应题通常有表格填空题、简答题以及论述题等形式。很多情况下,这些题型的答案都可以得到客观的评分,特别是在使用自动评分系统(automated scoring)后(Attali & Burstein, 2006)。在替换性评价(alternate assessments)中,可以允许对这些有约束的建构反应题以替换性方式作出回答,包括口头回答、图画、词汇表或构建示意图回答(construct maps)等。

三、试题编写的准则(Item Writing Guidelines)

在许多教育测量的教科书中,都有一个或几个章节对试题编写加以描述。有一些章节在试题编写方面非常详尽,使人深受启发,如《测试研发指南》(*Handbook of Test Development*)第12、13、14章(Downing, 2006; Welch, 2006;以及Sireci & Zenisky, 2006),《教育测量》(*Educational Measurement*, Ferrara & DeMauro, 2006;及Schmeiser & Welch, 2006)第9、16章。还有一些书整篇都是针对试题编写的,如《高级思维评价的试题编写》(*Writing Test Items to Evaluate Higher Order Thinking*, Haladyna, 1997)以及《单项选择试题的开发及验证》(*Developing and Validating Multiple-Choice Test Items*, Haladyna, 2004)。这些教材都是深入研究试题编写技术的强有力的工具。

1. 单项选择题的编写准则

第一个以研究为基础的关于单项选择题的编写准则分类是由哈拉代那和唐宁(Haladyna and Downing, 1989a)提出的,并且带有实证证据总结(1989b)。2002年这一分类法得到了进一步修正,补充了更多的实证证据以及对这些证据的元分析研究(Haladyna, Downing, & Rodriguez, 2002)。大

多数单项选择型试题编写准则基于逻辑推理和命题经验,几乎没有以实证证据为基础的。其编写准则涉及以下四个方面:内容、形式与风格(formatting and style)、题干(the stem)、选项(the options)。

内容是试题编写中最重要的方面,学科命题专家是编写成功试题的领路人。试题编写必须仔细,以便能将重要相关的内容及认知技能包含进去。这些试题编写准则大多都是以命题专家的逻辑论证和经验以及考生的反应为基础的,已有的文献中除了一些对试题明晰度及用词恰当性等泛泛的研究外,有关这些试题的编写准则没有什么实证性研究。有关试题内容方面的准则包括:

(1)每道题都应该反映考试说明(双向细目表、命题蓝图)中明确的、特定的内容以及某|具体的心理行为。

(2)每道题都须建立在重要的学习内容基础上,避免测试不重要的内容。

(3)用新的材料去测试高层次学习。试题中要避免使用教材中的语句,那些在课本中或课堂上使用的语句在命题时要进行修改,以免考生仅凭记忆作答。

(4)确保测试中每道题的内容独立。

(5)在编写单项选择题过程中,须避免过于具体或过于宽泛的内容。

(6)避免经验性的试题(opinion-based items)。

(7)避免脑筋急转弯性质的试题(trick items)。

(8)确保题目语言对应试群体而言是简单的。

形式和风格是建立在良好的试题编写经验基础之上。有一些实证证据为大多数题型的一般性应用提供了支持(Haladyna, Downing & Rodriguez, 2002)。同时,还有一些实证数据表明,某些形式,如复杂单项选择题形式(the complex MC format)增大了试题难度,而且与测试构念(construct)毫无干系。在下面的准则中,准则(9)和(13)已经过大量的证据支持,尤其需要重视。

(9)使用传统单项选择题、二选|题、判断正误题、多项判断正误题(multiple true-false, MTF)、搭配题、依据上下文回答题以及试题组形式,但是避免使用复杂单项选择题型(K型)。

(10)试题应纵向排列,而非横向排列。

(11)对试题进行修订和验证。

(12)试题语法、标点、字母大小写以及拼写正确无误。

(13)尽量减少每道题的阅读量。

编写题干是另一个缺乏实证证据的领域。实证研究结果表明,试题中应尽量避免使用否定式的题干,这一准则尤其适用于题干编写,它是试题整体编写风格的延续。尽管阿巴蒂(Abedi)与其他人的研究结果为这些证据提供了支持,但是他们进行的研究并非是有意设计用来检测某一具体试题编写方法的效度。关于题干编写的准则有:

(14)确保题干指向性明确。

(15)要在题干,而非选项中体现主题思想。

(16)避免语言拖沓冗长(过度繁琐)。

(17)题干使用肯定语句,尽量避免否定语句,如"不是"或"除……"这样的语句。否定式语句须谨慎使用,始终确保否定词须大写或黑体标出。

选项编写是研究中关注最多的部分。我们注意到,这里列举的 14条编写准则中只有 5条经过实证研究(18, 24-27条)。研究文献中最注重的一条编写准则就是选择题中的选项数目。罗德里格兹(Rodriguez, 2005)就这一论题对过去 80年间的研究进行过一次综合分析,他得出的结论是,在大多数情况下

3个选项已经足够, 虽然这不一定是最佳数目。

(18)提供的有效选项越多越好, 但是研究显示, 3 个选项就已经足够。

(19)确保只有| 个选项是正确答案。

(20)根据选项数目调整正确答案的位置。

(21)按照逻辑程序或数字顺序排列选项。

(22)选项须独立, 选项间不得相互包含。

(23)选项内容和语法结构须同质。

(24)选项长度大致相当。

(25)慎用“以上都不对”(*none-of-the-above*)这样的选项。

(26)避免使用“以上都对”(*all-of-the-above*)这样的选项。

(27)选项尽量用肯定句, 避免否定句。

(28)避免提供正确答案的线索, 如:

a 特定限定词, 包括“常常”, “从不”, “完全”以及“绝对”。

b 语音联想意义, 与题干中词汇相同或相近的选项。

c 避免考生通过语法的不| 致性找到正确选项。

d 显而易见正确的选项。

e 某两 个或三 个选项给予受试者暗示, 令其找到正确答案的选项

f 可笑, 荒谬的选项。

(29)让所有的干扰项都看上去有道理。

(30)用学生特别容易犯错的答案做干扰项。

(31)如果与教师以及学习环境相吻合, 题中可以加入幽默元素。

2 建构反应题的编写准则

有关建构反应题编写准则的研究起步较晚, 研究成果甚少, 测试专家们对于编写建构反应题的重点也缺乏统一认识。奥斯特林德和莫兹 (Osterlind and Merz) 1994年提出了建构反应题分类法, 这一分类法很大程度上是建立在认知心理学基础上, 其包括三个方面: (1)使用的推理能力类型, 包括①事实回忆 (factual recall), ②阐释性推理 (interpretive reasoning), ③分析性推理 (analytical reasoning), ④预见性推理 (predictive reasoning); (2)采用的认知连续性 (cognitive continuum)性质, 包括①集合思维 (convergent thinking), ②发散思维 (divergent thinking); (3)获得的答案种类, 包括①开放性答案形式 (open product), ②封闭性答案形式 (closed product), 共有 16种组合。前两个方面强调的是认知过程, 第三个方面强调的是可能获得的答案的种类。封闭答案形式指的是答案的选择性很少, 计分点相对较少; 而开放性答案形式可以有多种答案, 评分根据更加复杂的标准, 允许出现创意性以及意料之外的答案。

大部分测试服务公司都制定了建构反应题的编写准则, 作为其命题人员的工作指南。例如, ETS开发了几个包括建构反应题的大型考试 (如 NAEP及 AP测验)。这些考试项目促进了对建构反应型试题的大量研究, 但研究结果很少公布。此外, ETS还有专门的命题文件, 如《建构反应型试题及其他表现性评价的编写准则》(*Guidelines for Constructed-Response and Other Peformance Assessments*; Baldwin, Fowles, & Livingston, 2005), 它提供了许多很好的命题准则。这些命题准则都是一般性的, 可作为制定具体测试项目命题方案的基础, 包括:

(1)确保对评估起决定作用的人的知识和技能得到评估, 同时其能代表这 个受评估群体在人口、民族及文化等方面的多样性。

(2)在早期阶段公开评估的相关信息,

使需要了解以及希望了解情况的人可以对信息加以评论。

(3)向即将参加评估的人提供信息,对为何进行评估,评估如何进行,应试人的答案如何进行评分等信息加以解释。

四、试题的效度验证(Item Validation)

在测量与考试中,试题的效度常常定义为证据对测验分数解释或应用的支持程度。目前,不同领域对试题的效度有不同的定义。但是在教育测试方面,大多数人都认同《教育及心理测试标准》(the *Standards for Educational and Psychological Testing*, AERA, APA, NCME, 1999)一书中的定义,即"试题的效度指证据和理论对测试分数解释的支持程度"(AERA, APA, NCME, 1999)。《教育及心理测试标准》将试题的效度验证描述为为实现某些目标而收集证据的过程,这些证据包括测试构念,测试内容,答题过程(response processes),内部结构(internal structure),与其他变量关系(relations to other variables)以及预期结果和非预期结果(intended and unintended consequences)等。

任何情况下,效度验证都是一个不间断的过程。同时,效度证据最重要的来源就是直接推理以及针对测试结果提出的主张。在试卷设计、实施、试卷分析以及分数报告的各个阶段收集效度证据都很重要。在试题开发过程中,可以收集多种形式的效度证据,以支持某一具体试题的使用。唐宁和哈拉代那(Downing与Haladyna)1997年提出了一个模型,用来搜集关于试题质量的定性的效度证据。

1. 试题效度证据模型:定性证据(A Model of Item Validity Evidence: Qualitative Evidence)

证据类型	活动	需要的证据
内容定义	完成角色描述,任务/工作分析;完成实践分析	记录选择试题内容的方法
测试规范	创建测试框架/测验蓝图	记录测验内容和测验蓝图之间的系统联系
试题编写人员培训	开发培训材料以及方法;培训试题编写人员	记录命题的方法、准则、命题材料及样题
遵照试题编写准则	所采用的标准试题编写规则	遵守命题规则的证据,以及审核试题采取的程序文件
认知行为	用来对试题进行分类的认知分类系统	记录所采用的系统以及基本原理;对使用系统的研究进行报告
审核试题内容	试题内容专家对试题进行审核评判	内容专家审核程序中试题内容专家的证书记录
试题编辑	审核试题并进行专业编辑	编辑人员证书及经验;编辑及风格的准则,记录编辑/审核周期
偏差或敏感性审核	偏差/敏感性审核政策及实施过程	偏差/敏感性审核记录;政策基本原理;审核人员证书
试题预测	预测/试点测验;试题性能数据;与被试面谈	被试预测数据记录;被试及试题的特征
答案的审核确定	参考答案的正确性经过内容专家组确认	答案确认政策及流程;记录审核答案的结果
测验安全计划	制定确保测验安全的政策及程序	详细列举确保试题安全的方法及流程指南的副本

2 创新试题形式与技术改进

计算机化测试为各种新题型出现创造了机会,新的试题形式不断开发涌现。有些人认为,这些新题型使更多的参试者能参加考

试。还有人认为,新的题型能更好地对目标测试构念进行测量,即提高了效度。2006年,斯里茨和泽尼斯基(Sireci&Zenisky)归纳了13种基于计算机测试的题型,这些题型有助于提高目标测试构念的代表性,减少了与构念不相关的因素干扰。增强对目标测试构念的测量能力对保证试题的效度是十分重要的,以下对新出现的部分题型进行介绍。

拓展型单项选择题(ExtendedMC Items)通常在段落阅读题中出现,段落中的每一个句子都可以作为具体问题的答案。这种考试针对一个阅读段落可以设计一系列的问题。例如,要考查一个段落的主题思想,答案就是将这个段落中的某个句子凸显出来。这种题型的特别之处在于选项和答案都是从阅读段落中选出来的句子,不同于传统的单项选择题中的选项要重新设计、编写。

其他的新题型包括概念联系题(dragging and connecting concepts)以及信息挑选分类题(sorting and ordering information)。计算机环境下允许其他新型的回答方式,包括改正有语法或数学表述错误的句子,完成句子或方程式,制作或完成图形模型、几何图形或数据趋势等。计算机提供了广泛的可能性。

基于计算机测试的新题型已经在许多中学后及职业考试中得以验证。GRE和TOEFL考试的创新促进了很多有影响的研究的出现。1999年,本内特、莫里、夸德特以及洛克等(Bennett,Morley,Quard&Rock)对GRE考试中使用图形模型(graphical modeling)测试考生数学推理能力的试题进行了研究,这一类型的试题要求考生在回答时使用图解表示(graphical representations)。研究结果提供了与GRE定量部分总分(GRE quantitative total score)及相关变量具有中等程度相关的分数,这些分数具有很高的信度。例如,考生可以在网格上(grid)描点(plot points),然后利用工具将这些点连接起来。尽管考生们认为这些图形题能更好地反映学生在研究生院的学习成功潜力,但是他们还是更倾向于传统的单项选择题(在对比单项选择题型和建构反应题的过程中,这种观点很常见)。

3.提高应试者的可参与性

教育测试的一个新重点是要提高测验的可参与性。可参与性指不同能力的学生(包括身有残疾的学生)可以无须适应,即能按照与测试构念相适应的测试方式展现其最佳的能力程度。关于这方面内容,目前已有很多的资源,包括最近的杂志特刊,如《皮博迪教育期刊》(*Peabody Journal of Education*, Volume 84, Number 4, 2009)及《教育应用测试》(*Applied Measurement in Education*, Volume 23, Number 2, 2010)等。提高测试参与性最常用的做法就是贯彻"共用性设计"原则(the Principles of Universal Design)。使用这一原则的中心思想是消除那些妨碍测试者展现其才能的障碍,而实现这一目标的方法就是为所有考生,包括那些有残疾或认知缺陷的考生设计出具有最大参与性的测试,并提高重要知识、技能和能力的衡量水平。

4.试题编写准则:共用性设计(Item Writing Principles: Universal Design)

在许多测试及评估体系,包括教育系统的测试及评估中,试题编写的准则都是围绕一套称之为"共用性设计"的原则而制定的。"共用性设计"这个词来源于建筑工程领域,指的是所有人都可以使用,范围最大化,无需更改或调整的设计。这意味着共用性设计的有效性很强,因为考生的参与可以让他们展现其最佳表现,提高施测者解读分数的能力,得到有效的学业表现推论。"共用性设计中心"(the Center for Universal Design, 1997)认为,共用性设计结构包括以下7个主要原则:

(1)公平使用:这种设计对不同能力的使用者来说都是有用的。

(2)弹性使用:这种设计涵盖了广泛的个人喜好和能力。

(3)简易及直觉使用:不论使用者的经验、知识、语言能力或集中力如何,这种设计的使用都很容易了解。

(4)明显的信息:不论周围状况或者使用者感官能力如何,这种设计有效地针对使用者传达了必要的信息。

(5)容许错误:这种设计将危险及以为或不经意的动作所导致的不利后果降至最低。

(6)省力:这种设计可以有效、舒适及不费力的使用。

(7)适当的尺寸及空间供使用:不论使用者体型、姿势或移动性如何,这种设计提供了适当的大小及空间供操作及使用。

这些原则之间相互联系,可以将其概述为两个总的设计目标:第一,去除所有对提高使用性毫无裨益的元素或部分;第二,尽量让所有人可以参加。

共用性设计原则已经扩展到了各种领域,包括课程、教学(National Center on Universal Design for Learning, 2011)及心理教育测试领域。利用这一原则,可以对最广泛的考生群体设计相关的评估及测试工具,而不用再专门针对某一特定的人群设计试题。美国国家教育成就中心(The National Center for Educational Outcomes; Thompson & Thurlow, 2002)修正了原有的共用性设计原则,总结出以下评估要点:

(1)包含整个评估群体。

(2)精确定义的构念。

(3)可理解、无偏见的试题。

(4)可随时进行改编。

(5)简单、清楚、直观的说明和流程。

(6)可读性强,容易理解。

(7)最大的清晰程度。

在测试方面,共用性设计的目标有两个,第一,消除所有对测试构念有干扰的选项特征,即消除那些与构念无关的变量;第二,在构念的各个水平上,对所有考生无障碍。这两个目标要体现在有关考试目的说明中。考试开发和试题命制的准则首先要把对考试目的说明作为第一也是最重要的一步(Downing & Haladyna, 2007)。要得到高质量的测试(由高质量试题构成的测试),必须在试题编写过程中完全恪守测试目的。测试目的之所以重要,是因为其界定了测试的构念以及测试的考生群体范围。例如,美国明尼苏达州教育局(the Minnesota Department of Education)对其州测试的目的作出如下表述:“衡量明尼苏达州学生在明尼苏达 12年制教育学术标准下的学业成绩。”(明尼苏达州教育局, 2009)。尽管这一说明很简短,但它描述了考生群体(明尼苏达州学生)以及测试构念或测试内容(明尼苏达学术标准中有进一步界定)。从考试目的说明角度来讲,共用性设计要求测试开发人员做到以下两点:第一,精确界定测试构念,从而很容易地排除无关试题;第二,清楚地界定考生群体,并尽量扩大考生范围,确保考试构念与所有考生相关。共用性设计的价值特别在后一点中体现,即把有各种本领、才能、背景的考生都包含进来。很多情况下,考试与评估都会对某一群体的学生存在障碍,或是持有偏见。例如,语言障碍与学习障碍虽被看做与考试构念不相关,但往往会影响考生回答试题的能力。共用性设计鼓励的是那种有最广泛参与性的高质量的评估测试。

测试构念的相关与不相关(即目标一的焦点)对于编写高质量的试题至关重要。有关测试构念相关性的讨论源于经典测量理

论,该理论认为测试分数可以分成两部分,一是真分数(与测试构念相关),二是误差分数(与测试构念无关)。一道好的试题要把真分数最大化,同时将误差分数最小化。共用性原则有助于命题人员在试题编写过程中确定如何实现测试构念相关性的最大化。

此外,还应把更多的注意力放在其他的替代性测试上,特别是针对那些有中等或严重认知缺陷的学生(他们通常都无法参加一般的教育测试项目),要重视为他们提供相应的替代性测验。《皮博迪教育期刊》(Volume 85, 2009)最近就介绍了目前有关替代性测试研究的最新情况。不少研究对以下方面的内容进行了重点调查,如识别适宜参与的考生群体、考生参与的水平、适应情况、与一般教育课程的联系以及总体表现等。凯特勒、艾略特及白窦(Kettler, Elliott, Beddow, 2009)展示了他们在开发一个有理论指引,并有实证支撑的工具方面所做的工作,这一工具,即"测试可及性及调整清单"(The Test Accessibility and Modification Inventory, TAMI,可以登录 http://peabody.vanderbilt.edu/tami.xml),其可以对测试作出调整从而达到扩展测试的可参与性的目的。该工具以下列要素为指导:共用性设计原则、测试可参与性、认知荷载理论、测试公平、测试适应度以及试题编写研究等。

TAMI提供了一个评分系统,根据阅读材料或其他试题刺激(item stimuli)、题干、图像、答案选项、页码及试题分布、公平度以及计算机化测验等几个方面,来评估测试的可参与性。这一评级系统以一系列评估准则为指导,这些评估准则包括对上述各项的总体可参与性的评级,还包括对试题调整提出建议的可能性,以增加评估的"无障碍性"。TAMI还提供了几个标准的调整样例作为指导,试题调整的目的是为了确保试题编写准则有效实施,同时消除与测试构念不相干的因素,最大限度提高学生的可参与性。TAMI还推荐使用认知实验室(cognitive labs)来指引试题及测验的调整。

TAMI已经在美国几个州得以应用。凯特勒、罗德里格兹等(Kettler, Rodriguez),以及艾略特、凯特勒等(Elliott, Kettler, 2010)已经公布了使用TAMI调整州测试的多州联合研究项目结果。这些研究人员发现,调整保持了分数的可靠性,改善了学生表现(分数增加)。而有资格参加替换性测试的学生(如有中等认知缺陷的学生)的表现比参加其他测试的要好。罗德里格兹(Rodriguez, 2009)提交了一份有关代替换性测试面临的心理测量挑战的详细分析报告。这种测试中常见的调整之一就是把选项减少到3个,将最无效的干扰选项删除。在数学及阅读题中,保留下来的干扰选项更具区分度。这一系列的研究中最重要的一个原理就是试题调整往往涉及多方面的调整。由于调整是多方面的,要孤立地讨论某种单一的调整产生的影响是非常困难的。另外,试题编写应把考生可参与的最大化作为最基本的目标,而要实现这一目标,TAMI可以作为一个有效的指南。

总结

有关试题编写的研究非常有限,但是试题的质量却很重要。所有基于测试分数作出的决定,包括分配、升级、入学、颁发证书或资格证等,都使得测试分数的质量成为重中之重,这当然和测试效度密不可分。因此,要提高解释、使用测试分数的效度,我们需要确保试题的高质量。

以上对普通单项选择题及建构反应题题型模式进行了回顾,同时评论了单项选择题

及建构反应题的一般以及基于证据的试题编写准则。此外,文本还探讨了与试题效度相关的重要问题,包括试题编写过程中的定性有效证据、新型题型以及增强测试参与度的方法。所有这些都是测验效度的重要方面,其作为测量的基本构成要素,都旨在改善测试的质量。

参考文献:

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999), *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

[2] Attali, Y., & Burstein, J. (2006), Automated scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment*, 4(3), 1-30.

[3] Baldwin, D., Fowles, M. & Livingston, S. (2005), *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

[4] Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (1990), *Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning* (ETS RR-99-21), Princeton, NJ: Educational Testing Service.

[5] Center for Universal Design (1997), *The Principles of Universal Design*, Version 2.0. Raleigh, NC: North Carolina State University.

[6] Downing, S. M. (2006), Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301), Mahwah, NJ: Lawrence Erlbaum.

[7] Downing, S. M., & Haladyna, T. M. (1997), *Test item development: Validity evidence from quality assurance procedures*. Applied Measurement in Education, 10(1), 61-82.

[8] DuBois, P. H. (1970), *A history of psychological testing*. Boston, MA: Allyn & Bacon.

[9] Ebel, R. L. (1951). *Writing the test item. In E. F. Lindquist (Ed.)*, Educational measurement (1st ed., pp. 185-249). Washington DC: American Council on Education.

[10] Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M., Bolt, D., & Roach, A. T. (2010), *Effects of using modified items to test students with persistent academic difficulties*. Exceptional Children, 76(4), 475-495.

[11] Ferrara, S., & DeMauro, G. E. (2006), *Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.)*, Educational Measurement (4th ed., pp. 324-), Westport, CT: Praeger Publishers.

[12] Haladyna, T. M. (1997), *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.

[13] Haladyna, T. M. (2004), *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

[14] Haladyna, T. M., & Downing, S. M. (1989a), *A taxonomy of multiple-choice item-writing rules*. Applied Measurement in Education, 1, 37-50.

[15] Haladyna, T. M., & Downing, S. M. (1989b), *The validity of a taxonomy of multiple-choice item-writing rules*. Applied Measurement in Education, 1, 51-78.

[16] Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002), *A review of multiple-choice item-writing guidelines for classroom assessment*. Applied Measurement in Education, 15(3), 309-334.

[17] Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009), *Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know*. Peabody Journal of Education, 84, 529-551.

[18] Kettler, R. J., Rodriguez, M. R., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (in press), *Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm*. Applied Measurement in Education.

[19] Minnesota Department of Education (2009), Minnesota Comprehensive Assessments Series II (MCA-II): Test Specifications for Reading. Roseville, MN: Author. Retrieved online at http://education.state.mn.us/mdeprod/groups/Assessment/documents/Report/006367.pdf.

[ ] National Center on Universal Design for Learning (2011), *Universal design for learning guidelines, version 2.0*, Wakefield, MA: CAST.

[20] Osterlind, S. J., & Merz, W. R. (1994), *Building a taxonomy for constructed-response test items*. Educational Assessment, 2(2), 133-147.

(下转第 84页)

Analysis on the Controlling of Imperial Examination on Social System

Feng Jianmin

Institute of Education, Xiamen University, Xiamen, Fujian, 361005

Abstract Imperial Examination was founded in the Sui-tang dynasty, perfected in the Song-yuan dynasty, prospered in the Ming-ching dynasty and abolished in the end of Qing dynasty. It had lasted for more than one thousand years and dominated the center of political activities and social activities of ancient China. Imperial Examination, which is an examination system aiming to select nation-governing talents, has long histories, perfect setups and stable patterns. The system, with powerful social functions, touched every corner of the society and had a controlling impact on feudal political system, educational system and etiquette & custom system.

Key words Imperial Examination, Social System, Controlling

(上接第 94页)

[21] Rodriguez, M. C. (2005), *Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research*, Educational Measurement: Issues and Practice, 24(2), 3-13.

[22] Rodriguez, M. C. (2009), *Psychometric considerations for alternate assessments based on modified academic achievement standards*, Peabody Journal of Education, 84, 595-602.

[23] Schmeiser, C. B., & Welch, C. J. (2006), Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 324-), Westport, CT: Praeger Publishers.

[24] Shanedling, J., Van Heest, A., Rodriguez, M. C., Putnam, M., Agel, J. (2010), *Validation of an online assessment of orthopedic surgery residents' cognitive skills and preparedness for carpal tunnel release surgery*, Journal of Graduate Medical Education, 2(3), 435-441.

[25] Sireci, S. G., & Zenisky, A. L. (2006), Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum.

[26] Thompson, S., & Thurlow, M. (2002), *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

[27] Welch, C. (2006), Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303-327), Mahwah, NJ: Lawrence Erlbaum.