基于认知理论的测验编制技术—项目生成*

周　骏　戴海琦　徐淑媛

(江西师范大学 教育学院,南昌 330027)

摘　要:传统测验重视统计技术,不重视测量结构的心理学意义,使得它的功能局限于筛选,不能提供更多的信息。由于认知心理学理论、心理计量学以及计算机技术的发展,使得基于认知理论指导下的项目生成技术逐渐成熟。该文从项目生成的概念、方法,及研究的意义和难点等方面对项目生成技术作了一个简要述评,以期推进我国认知心理学与心理测量学相结合的研究工作。

关键词:项目生成;认知理论;心理测量理论

中图分类号:B841.2　　文献标识码:A　　文章编号:1003 - 5184(2007)04 - 0064 - 05

1　引言

当代认知科学的迅速发展对心理与教育测量的发展提出了更高的要求,同时也提供了发展的基础。这就是发展新一代测量理论、探索测量与认知相结合的方法,从只测量'结果'到既测量'结果'又测量"过程",变只能测量笼统的'心理特质'为测量各种实质性的'心理认知成分",这是一项心理与教育测量发展史上具有革命性意义的工作。项目生成是以认知理论为指导的一种测验编制方法,它与传统测验编制方法有本质的不同。

2　传统测验编制方法的缺陷

Carroll 和 Maxwell 在 Annual Review of Psychology 中指出自从 25 年前首位学者就测验的发展进行回顾以来,此领域少有变化,但这并不等于传统测验的发展已臻成熟[4]。由于心理学理论和测验分析技术的限制,使得早期的测验编制者不得不忽略那些早已存在的问题。在心理学界尚以行为主义马首是瞻的时候,实验心理学家主要是通过外在刺激与反应行为之间的关系,推断被试答题时的心理过程。

测验的开发是从项目开始的。传统上,项目的规格属性通常是含糊的,也就是说,测验开发者编制出的项目通常只包含概括性的内容或描述的很含糊的处理水平,一旦项目被编制出来,只能通过考察被试在项目的作答结果来保证答案的可辨别性,以及用多种质量标准去检测项目内容。

另外,传统心理计量取向,虽然重视抽象的心理特质(如:智力、动机等),并且使用相关、因素分析等统计技术来避开心理特质量尺不明的窘境,但其效度问题无法从根本上得以验证,受到人们较多的质疑与批评。依据传统心理计量的做法,主要是依据 Cronbach 和 Meehl 对测验结构效度的定义,即,通过测验分数与外部变量的相关来表明所设计测验的结构效度。如果检验结果不支持所要测量的测验结构,测验开发者要么重新定义结构要么重新设计测验。就测验编制而言,在效度研究之前测验编制者大部分的精力主要用于测验的开发和项目的编制上,因此,测验开发者更倾向于重新定义结构,而不是重新设计测验。

测验的开发者利用所编制的测验与其他效标测验间的相关来支持该测验的结构效度,这种做法并不能说明被试解题时的特定技能、知识结构和心理过程等。之所以如此,最主要的原因还是在于传统的心理计量取向强调以统计技术从测验的结果来推论被试的能力构成,并没有建立被试实际答题过程的理论基础,也就没有办法建立有效的心理计量模型来说明测验刺激特征与个人特质间的关系。传统的测验编制方法,由于缺乏心理学实质理论作基础,其效度无法从根本上得以验证,受到人们较多的质疑与批评。

2　认知理论在测验编制中的作用

近年来,认知心理学理论的发展和心理计量技术的革新对测验发展的方向影响巨大。认知心理学理论在心理和教育测量的设计、评估中的作用越来越普遍。认知心理学不再将人类的内心世界视为"黑箱",也不再将被试作答的过程抛开。认知心理学中的认知成分分析以及 Vygotsky 的认知发展理论对测验的发展有关键性的影响。

Sternberg 指出未来测验发展的方向应结合认知

* 基金项目:高等学校博士学科点专项科研基金项目(20050414001),全国教育考试'十一五'科研规划课题(2006JKS3057)。

心理学理论、心理计量学理论以及教学，使测验的发展具有认知心理学的基础，测验的结果能提供有关对被试的诊断信息，测验的分数能反映出答题的心理过程等[16]。有了对智能活动的认知分析结果后，引入合适的测量模型对这类分析给予量的刻画，使认知分析结果能获得测量数据的实际验证，就能对测量作更有效的解释。这就是认知心理学与现代测验理论的结合，是新一代测验理论的核心思想。将认知心理学与测量的结合，可以提高所编制测验的结构效度[1,8]。该取向的基本假设是：测验编制的目的是希望通过特别设计的项目特征来引发符合测量目标的特质或是认知过程；在测验编制的过程中，认知心理学理论或是相关研究可以对哪种项目特征会引起被试哪种认知过程提供理论基础，并赋予被试在项目反应(item response)实质上的意义；而心理计量方法能将项目反应的实质意义与个别差异相联系，并可验证被试在答题时的认知过程是否如认知理论的假设，用来验证该测验的结构效度。

目前，认知心理学已经应用到测验开发的各个阶段，如定义结构、选择项目类型、为诊断性的分数解释提供基础、为自动计分定义原则以及为项目生成算法提供结构等。尽管认知心理学在当前的测验编制中很重要，但是项目还是没有完全由认知理论来设计。

在项目生成中，认知心理学原理是为每个指定项目刺激特征的关键，这样结构效度能被扩展到项目水平上，因此研究者能够利用认知复杂程度的来源为被评估者开发特定难度水平的项目[12]。

3　基于认知理论的测验编制方法——项目生成

3.1　项目生成的概念

什么是项目生成(Item Generation)？简单地说，项目生成是以认知理论和心理计量学理论为基础，在能力或成就测验领域，依据一套完善的项目设计原则，由计算机自动编制项目的一种技术。当被试参加测验时，所使用的项目都是计算机适时生成的。

3.2　早期项目生成技术

项目生成，长期以来都受测验开发者所关注。二十世纪七十年代，Wells Hively 和他的同事构建了一个可以生成大量数学题目的系统[15]。facet 理论假设：可以通过确定规则，从而将具体属性映射到项目之中。因此，早期的项目生成研究假设：项目由固定成份(项目框架)和可变成份(替换它们即可生成更多项目)组成[13]。如：

小刚在超市买了一些物品，他买了 3 斤青椒，每斤 1.25 元，4 斤西红柿，每斤 1.75 元，和 2 斤洋葱，每斤 1.58 元，小刚共花费多少钱？

只要在一定范围内替换数量、价钱，就能从这种项目形式中生成大量项目。如果所替换的数字限制在一定范围之内，就保证了问题的难度相等。计算的对象也可被替换，同时也假设这样做不会影响难度。

从上面的描述来看，项目生成似乎很简单，只要知道了某个类型的项目中哪些是可变成份和固定成份，或者说，知道某个题型的生成规则，就能完成该类题型的项目生成工作。实际上，仅仅知道替换规则是远远不够的，很有必要将认知理论应用于所研究的项目类型中，以设计出源于坚实的心理学理论的项目生成规则。

在 Hively 等人的研究中，他们意识到在生成项目时需要同时控制项目的同质性和难度，但当时对难度的考虑还相对比较少，特别是没有意识到从心理计量学方面生成项目需要认知理论的指导[3]。

3.3　基于认知理论的项目生成技术

在利用项目生成技术编制测验时，引入认知理论指导测验编制一个较好的理由是能获得项目水平上的结构效度，即对测验的解释基于每个项目的属性，因为对项目而言影响项目难度的具体认知来源都已知[5]。在项目水平获得结构效度，有利于组织有效的测验，有利于生成诊断性较强的项目。一般使用的认知理论是 Sternberg 的认知成分分析理论，该方法是分析被试在解题过程中所需的认知成分，以确定被试在解题时所需特定技能、知识结构，以及被试解题所应用的策略。

目前，使用认知理论来进行项目生成的方法有两种：Embretson 提出的认知设计系统方法(Cognitive Design System Approach，简记 CDSA)和 Bejar 提出的项目模型(Item modeling)方法。

◇　认知设计系统方法(Cognitive Design System Approach)

为了将认知理论应用于测验编制中，Embretson 定义了认知设计系统方法，并将其作为开发测验的框架，在这个框架中，可以使用特定的认知模型来描述项目[6]。认知设计系统方法的开发是为了强调能力和成就测验中认知理论的作用，为了强调测验开发中的认知理论，认知设计系统方法包括概念框架和程序框架两部分。

在概念框架中，Embretson 认为 Cronbach 和 Meehl 在 1955 年提出的结构效度概念并不能说明认知理论在测验开发中的作用，他们的概念强调建立一个经验网络(empirical networks)去确定一个测验测量了什么，在测验被开发出来以后再进行结构效度

的研究,所以认知理论对测验的影响被最小化了,经验性的结果不能为项目设计提供反馈。

Embretson 将结构效度分为结构表征(construct representation)及合法区间(nomothetic span)两个方面[5,9]。结构表征涉及解题所需的加工过程、策略和知识结构。在测验开发过程中利用认知心理学进行的相关研究属于结构表征范畴。项目属性(如项目难度)可以通过对代表认知加工过程的项目刺激特征建立数学模型来加以表征,依据所建立的数学模型不仅可以用来了解项目所测结构的性质,同时也能对项目的心理计量属性做出恰当的预测[11]。

结构效度的另一方面,即合法区间,关注的是测验分数与外部其他测量之间的相关。Embretson 特别指出由结构表征研究生成的假设系统应该指导合法区间的研究。

在结构效度中将结构表征区分出来目的在于帮助强调测验开发中认知理论的重要性。首先,结构效度可以在项目水平上进行评估,即,刺激特征影响加工过程,加工过程确定项目的结构表征;其次,认知理论能够在测验开发中发挥作用,因为结构表征依赖项目刺激特征,设计的项目能够反映认知复杂度的来源;最后,项目生成的原则能够建立在对效度影响已知的项目刺激特征之上。

认知设计系统的程序框架不仅详细说明了如何开发项目性能的加工模型,而且用项目加工过程来表述测验效度。认知设计系统的程序框架可以概括为如下一些步骤:确定测量的整体目标、确认项目的设计特征、建立与问题解决相联系的认知模型、决定想要操纵的项目内容特征及其复杂程度、生成设计规格相符的项目、将认知模型转换为数学模型、施测并评估项目的认知和心理属性、依据项目的参数估计被试能力等。这些步骤所表述的是项目生成的一个完整过程,项目的持续改善需要反复完成这些步骤。即使在已经生成项目之后,如果发现描述项目特征的最初的认知模型不完整或不足以让人理解,就应该回到认知模型的建立阶段,修改过重新建立认知模型。

通过这种方法设计的项目具有如下特点:1)项目的意义与解决问题的认知模型相连;2)每个项目的特点能用内容特征及其复杂程度来明确界定;3)项目的难度能由内容特征及其复杂程度来控制。

◇ 项目模型(Item modeling)

Bejar 描述了一个项目生成方法——生成反应模型方法(Generative Response Model)[2,3]。生成反应模型是利用被试对预先设计好的许多种项目模型中具体项目做出反应所需要的心理加工知识来估计出反应模型的项目参数,并据此编制测验的方法。简单来说,生成反应模型是以一个具有适当的心理计量品质的项目作为模板,通过替换其中一个或多个属性生成新项目,并且可以不用对新项目进行试验,就可以通过这些项目去估计被试能力,因为研究者假设项目模型的性能会延续到每个新项目上。

2003 年 Bejar 在其研究报告中,将生成反应模型方法改为项目模型法或称项目建模法(Item modeling),并明确定义项目模型法是一种生成测验的方法,利用这种方法能提高测验的效度。在这个系统中,项目表现为种种"项目模型"(item models),这种项目模型是指具有良好心理计量属性的项目。项目模型被认为是一种构建同质项目的方法,同质项目是指项目中包含可比较的内容,并且从心理计量的角度来说,这些可比较的内容允许在一定范围内被替换。因此,采用这种方法生成项目进而编制测验,可以看作是由结构指导的,由于项目模型是建立在对项目的认知分析基础上,当然测验的效度也就提高了,这种说法和 Embretson 是一致的。

项目模型方法与传统人工项目编制方法比较有许多优势,对于人工项目编制者而言,无论项目之间多么相似,他们都是将每个项目看作独立的实体。而对项目模型方法来说,当项目模型被明确表达和校准之后,项目模型方法可以使生成项目的许多细节自动化,另外,研究者还能通过被试在项目上的反馈来修正和扩展项目模型。

在使用项目模型方法生成项目时,除了控制项目的同质性以外,还需要控制项目的难度。在生成方法中 Bejar 认为至少有两种方法来控制难度:强理论(strong theory)和弱理论(weak theory)。无论使用哪种理论来控制生成项目的难度,都需要和心理计量的反应模型联系起来[11]。因此,在判断用项目模型生成项目时,不仅要检查心理计量模型与数据之间的拟合情况,还要检查观测数据与每个项目模型得出的预测值之间的理论上的拟合情况。除此之外,还要检查由模型产生的项目对特定的评估目标而言是否是可接受的,因为使用项目模型生成项目是在一定范围内变化刺激特征,如果刺激特征的变化超过了被评估目标解决问题的能力,就认为该项目模型所产生的项目是不被接受的。

强理论使用认知心理学原理,在编制测验的模型中或生成项目的模型中控制难度。在前一种情况中,是先设计出能生成同质项目的每个模型,使用认知心理学原理在模型之间建立难度变量,然后估计

每个模型的参数,如:Embretson 在关于矩阵完成任务的研究使用的就是这种方法[7,11];在后一种情况中,认知心理学原理能够为每一个项目模型生成难度差异较大的项目。例如,Bejar 使用心理学理论指导心理旋转方面的项目生成,用初始图形和目标图形之间的角度差异来预测过程难度[2]。

一般而言,强理论适用领域比较狭窄,只有在那些有心理学理论支撑并能够进行深入认知成分分析的研究领域,具有很好的前景。在其他更广阔的研究领域,弱理论比较适用。弱理论的方法是从标刻一组不同的项目开始,这组项目包含了不同的难度和内容,然后,将每个项目作为进一步生成项目的基础模型。项目标刻包括为估计出每一个项目模型的项目参数(如:难度、区分度、猜测度等),然后,将这些估计值分别赋给每一个模型。这样,一个项目模型所生成的项目具有相同的项目参数估计值。

项目模型方法在实践中明显的一个优势是:一旦校准了项目模型,就能生成许多提前校准的项目。与此同时也存在一个问题,那就是项目模型生成的项目可能存在相似性,并且这种相似性可能被学生以多种方法发现从而提高测验分数。解决这个问题的方法有两个:一是建立生成不相似项目同时又能保持项目参数同质性的项目模型;再有就是在测验中建立数量较多的项目模型,在施测过程中控制每个模型生成项目的数量。

在项目生成研究中,两种方法异曲同工,无论使用 Embretson 的认知设计系统方法,还是 Bejar 的项目模型方法,都是基于认知理论的指导。许多研究表明认知心理学对提高测验的结构效度有很大的潜力。1986 年,Hornke 等研究者利用认知理论为指导,利用了认知操作中的分类系统,将心理特质理论、项目建构与项目分析结合,编制了一个抽象推理项目的题库,分类系统可以很好的预测 616 个生成项目的难度[14];1996 年,Embretson 基于认知设计系统方法编制了一套抽象推理测验[10]。

4　项目生成研究的意义

计算机化自适应测验是当前测验领域中的前沿技术,在这种测验中,计算机根据被试的能力为其逐个选择具有最佳信息的项目。但是,计算机化自适应测验需要一个大型多样题库,以便准确、有效地测量被试。传统的项目编写过程无法满足需要,手工编写项目速度慢,并且项目不合格率比较高。很大一部分项目要么不能符合标准,要么在实际预测当中不能获得充分的心理测量学特征;再者,测验编制者经常需要特定难度水平的项目,而项目编写者也无法做到这点;还有,在项目编写中,将刺激内容与心理计量属性相关联的信息很少。

如果将计算机化自适应测验与项目生成技术合起来称为自适应项目生成,这是一种新的测验形式。自适应测验利用标准化的心理计量属性,用计算机算法交互式地选择项目。如果被试完成一个项目,就选更难的;如果未完成,则选更易的。这种自适应的项目选择使得短而有效的测量成为可能,最佳项目选择是通过人工智能来实现的测量方法。自适应项目生成比自适应测验又前进了一步,由于自适应项目生成的测验编制是建立在对项目类型的认知分析的基础上,因此,在实施测验时可以为被试生成新的具有最佳信息的项目,计算机根据被试的上一次反应模式生成新项目。另外,在认知测验中,测验所用的测验材料已经为越来越多的人所了解,测验的安全问题日益受到重视,因为测试被试所使用的项目,是根据被试的能力适时生成的,所以自适应项目生成可以避免这类问题。

项目生成可以作为开发认知诊断测验的基础,测量技术发展到 21 世纪,人们对测验结果的渴望已经不能用只给出一个总分值来满足,人们更期待出现针对不同个体的、能够揭示其内在特质的诊断性测验。项目生成技术依赖于对所测领域认知加工过程与项目结构的细致分析,确定影响人们行为的刺激特征与其他因素,利用项目生成方法编制的测验,可以为被试提供有效且个性化的基于项目认知过程和刺激特征层面的诊断分析;换而言之,因为通过项目刺激特征的权重刻画每个项目认知复杂性的具体来源,因此能在项目刺激特征层次上对被试进行更细微的评价。

5　项目生成研究的难点

在国外项目生成技术已经在一些能力或学业成就测验中取得了成功,在国内相信其前景也会十分乐观;但国内至今未有研究者涉足此领域,概括起来项目生成研究的难点有这么一些:首先,由于项目生成研究的初期耗时且费用大。在能力或学业成就测验领域中,研究者都要根据自身的研究领域为每一项目类型设计认知模型,这种认知模型各自建立在独立研究的基础上;再有认知心理学提供的可操作的认知理论有限,研究者要提出自己的认知模型有一定的难度;最后,认知心理学和测量学两学科的沟通与结合不是很好。复杂的数学模型令认知心理学家望而却步;认知分析工作需要有丰富的认知心理学的知识、理论和实践经验才能胜任,对于测量学家来讲,认知心理学的那套话语还很陌生,不仅要接受

而且要娴熟运用更需较长时期的学习。在这种情况下,两条研究路线沿各自的方向前进,要真正实现“认知和测量相结合”,还需真正有志于此方向的研究者融两家之长,付出艰苦卓绝而富有成效的创造性的劳动。

参考文献

1 Alderton D L, Larson G E. Dimensionality of Raven's Advanced Progressive Matrices items. Educational and Psychological Measurement, 1990, 50: 877 – 900.

2 Bejar I I. A generative analysis of a three – dimensional spatial task. Applied Psychological Measurement, 1990, 14(3): 237 – 245.

3 Bejar I I. Generative response modeling: Leveraging the computer as a test delivery medium (ETS Research Report). Princeton, NJ: Educational Testing Service, 1996.

4 Carroll J B. Maxwell S. Individual difference in ability. Annual Review of Psychology, 1979, 30: 603 – 640.

5 Embretson S E. Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 1983, 93: 179 – 197.

6 Embretson S E. Processes in abstract reasoning: A test of the Carpenter, Just, Shell theory. Unpublished manuscript, University of Kansas, Lawrence, 1993a.

7 Embretson S E. Psychometric models for learning and cognitive processes. In: N. Frederiksen, R. J. Mislevy I I Bejar, Eds. Test theory for a new generation of tests. Hillsdale. NJ: Lawrence Erlbaum, 1993b. 125 – 150.

8 Embretson S E. Applications of cognitive design systems to test development. In: C. R. Reynolds, Ed. Cognitive assessment: A multidisciplinary perspective. New York, NY: Plenium Press, 1994. 107 – 135.

9 Embretson S E. Developments toward a cognitive design system for psychological tests. In: D. Lubinsky, R. Dawes, Eds. Assessing individual differences in human behavior: New concepts, methods, and findings. Palo Alto, CA: Consulting Psychologist Press, 1995.

10 Embretson S E. Cognitive design principles and the successful performer: A study on spatial ability. Journal of Educational Measurement, 1996, 33: 29 – 39.

11 Embretson S E. Generating items during testing: psychometric issues and models. Psychometrika, 1999, 64(4): 407 – 433.

12 Embretson S E. Improving construct validity with cognitive psychology principles. Journal of Educational Measurement, 2001, 38: 343 – 368.

13 Guttman L. Integration of test design and analysis. Proceedings of the 1969 invitational conference on testing problems. Princeton, NJ: Educational Testing Service, 1969.

14 HornkeL F, Habon M W. Rule – based item bank construction and evaluation within the linear logistic framework. Applied Psychological Measurement, 1986, 10: 369 – 380.

15 Hively W, Patterson H L, Page S. A “universe – defined” system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5: 275 – 290.

16 Sternberg R J. Cognitive theory and psychometrica. In: R. K. Hambleton, J. N Zaal, Eds. Advances in educational and psychological testing: theory and applications. Boston: Kluwer, 1991. 367 – 394.

Test Construction Technology Based on Cognitive Theory: Item Generation

Zhou Jun Dai Haiqi Xu Shuyuan

(Education College, Jiangxi Normal University, Nanchang 330027)

Abstract: Because the traditional test pays attention to statistical techniques, rather than psychological meaning of the structure, its function is restricted to screening and it can't provide more information. With the development of cognitive psychology theory, psychometrics and computer technology, item generation technology, which is based on cognitive theory, had become mature gradually. The paper makes a brief review on item generation technology from concepts, methods, research significances and difficulties etc, in order to push forward the research on combining cognitive psychology with psychometrics.

Key words: item generation; cognitive theory; psychological measurement theory