心理测量学：心理学皇冠上的数学明珠*

文 / 余嘉元
南京师范大学心理学院 南京 210097

【摘 要】心理测量学是采用定量方法对心理特质进行研究的学科，是在社会需求的强劲推动下产生和发展的。经典测验理论、项目反应理论和概化理论是它的基础，在和计算机科学、认知科学的结合中产生的计算机化自适应测验和认知诊断测验正在得到应用，计算智能也开始成为研究的重要手段。它在教育、医学、管理、工业、军事等领域都有广阔的应用前景，在建模、参数估计、等值、项目功能差异、标准设置等方面还要加强研究，心理测量的应用也需要有良好的法制环境，以及具有高素质的人才队伍。

【关键词】心理测量学，经典测验理论，项目反应理论，概化理论，计算机化自适应测验，认知诊断理论

DOI：10.3969/j.issn.1000-3045.2012.Z1.021

翻开人类文明发展的绚丽画卷，处处闪烁着数学的光辉。恩格斯说，数学在一门科学中应用的程度，标志着这门科学的成熟程度。自从冯特（W. Wundt）于1879年在莱比锡大学建立了第一个心理学实验室以来，人们采用定量方法研究个体差异的兴趣就一直没有停止过。诗人赞美数学是科学皇冠上的明珠，心理学家孜孜不倦地把这颗明珠镶嵌到自己的皇冠上，在这个艰辛的过程中诞生了心理测量学。

1 来到人间，只为苍生

在人类文明初期就有识人用人的需求，它引发了古代哲人对心理特质定量描述的思考。我国春秋时期，孔子就将人区分为中人、中人以上和中人以下，古希腊则把人的气质分为多血质、胆汁质、黏液质、抑郁质4种类型，这是人类最早的研究探索。

科学家是最富有探索性的人群，冯特、高尔顿（F. Galton）、卡特尔（J. M. Cattell）的杰出工作为心理测量学诞生写下了浓墨重彩的篇章[1]。冯特在实验心理学研究中，发现了人的个别差异，并给出了定量的描述，这直接导致了心理测量的开展。英国剑桥大学教授高尔顿建立了人类学测量实验室，把统计学方法运用到心理测量数据的分析中。美国宾夕法尼亚大学教授卡特尔综合

* 修改稿收到日期：2012年12月23日

了冯特和高尔顿的学说，对于个别差异进行了深入的研究，指出“心理学若不立根于实验和测量上，决不能有自然科学的准确[2]。”这些心理学家的研究比孔子迟了2 000多年，是什么让远隔万水千山的中国人记住了他们的名字呢？重要的原因是他们采用了实证和测量的方法，把数学融合进了自己的研究工作，运用定量的方法对人的心理特质进行了描写。

社会需求是科学发展的动力，1904年，法国教育部组织了一个委员会，研究公立学校中低能儿童班级的管理问题，该委员会成员比纳（A. Binet）和他的助手西蒙（T. Simon）经过精心研究，于1905年提出了世界上第一个科学的心理量表，称为比纳-西蒙智力量表，使得心理测量摆脱了对颅相、面相、手相的分析，步入了运用科学量表进行测量的新时代[3]。美国著名心理学家波林（E. G. Boring）指出，在心理测量领域，“19世纪80年代是高尔顿的10年，90年代是卡特尔的10年，20世纪头10年是比纳的10年[4]。”这是对心理测量学诞生时期各位代表人物所做贡献的最好总结。

在社会强劲需求的推动下，涌现出了许多著名的智力测验，例如推孟（L. M. Terman）教授修订的斯坦福-比纳量表，从此智商（Intelligence Quotient，IQ）一词风靡全球。韦克斯勒智力量表、瑞文推理测验的推出，都是智力测量领域的大事。在教育测量方面，桑代克（E. L. Thorndike）把统计理论引入了心理和教育测量，为测验的编制奠定了理论基础。心理测量的方法也进入了人格领域，1917年第一个现代意义上的人格问卷伍德沃斯个人资料调查表发表，接着明尼苏达多相人格调查表（MMPI）、加利福尼亚心理调查表（CPI）、卡特尔16种人格因素问卷（16PF）、艾森克人格问卷（EPQ）、罗夏墨迹测验等相继问世。

心理测量学的诞生，就是为社会服务的，它向着教育、医学、军事、工业等各个领域浩浩荡荡地进军，经过100多年的勤奋努力，终于筑成了目前这座心理测量学的大厦。在这座高耸如云的建筑中，每个窗户都有圆溜溜的脑袋在向你呼唤：你要知道自己孩子的智力吗？你要了解朋友的性格吗？你要招聘到优秀的员工吗？你要明白自己最适合从事什么职业吗？请到我这里来吧！这一浪高一浪的吆喝声，也许会让你感到迷惑：仅仅凭着回答的这几十道题目，就能把复杂的心理特质测量出来吗？如果我们对大厦地基进行仔细探测，就可以找到问题的答案。

2 心理测量学大厦的基石

心理测量学家是兼有科学家和工程师素质的人才，他们试图对心理特质进行定量描述的同时，兼顾研制心理测量的工具。为了解答测量数据是否精确可信、是否真正测到了想测的特质、如何编制测量的工具、怎样解释测验的分数等问题，他们建立了心理测量学的基本理论。

2.1 经典测验理论

首先出现的是经典测验理论（Classical Test Theory，CTT），也称为真分数理论，这是英国心理学家斯皮尔曼（C. Spearman）提出并经过洛德（F. M. Lord）、诺维克（M. R. Novick）等多位学者重新陈述和精心构建的理论[5]。它的数学模型是：X = T + E，其中X是测验分数，T是真分数，E是随机误差。该理论提出了重测信度、复本信度、同质性信度、评分者信度等多种估计测量精度的方法，以及内容效度、构念效度和效标关联效度等一系列估计测量有效性的方法。它采用难度、区分度作为分析题目质量的指标，运用常模对测验分数进行解释和比较，它成为心理测量学大厦的第一块基石。

在CTT得到广泛应用的同时，人们也发现了它的缺陷，首先是测验分数依赖于题目的难度，因此在高难度测验中得到低分的考生就可能受到“虎妈狼爸”的严惩。同时题目难度是依赖于考生水平的，如果是根据弱智考生的作答分数计算题目难度，那么每道题目都会难于上青天。为了解决这些问题，项目反应理论（Item Response Theory, IRT）在心理学家的探索中诞生了。

2.2 项目反应理论

人的心理特质是潜在的、无法直接观察到的,心理学家是通过被试对于题目(又称“项目”)的作答来推测其心理特质的,因此就必须探索具有某种水平心理特质的人是如何对某个题目做出反应的,这就是项目反应理论的由来。经过洛德[6]、汉布尔顿(Hambleton)[7]等人的持续努力,建立了该理论的数学模型,其中最常用的是三参数逻辑斯谛模型:

$$P_i(\theta)=c_i+\frac{(1-c_i)}{1+\exp\left[-1.7a_i(\theta-b_i)\right]}$$

该模型表示能力为 θ 的被试,对于区分度为 a_i、难度为 b_i、猜测参数为 c_i 的第 i 道题目的正确作答概率为 $P_i(\theta)$。在这个公式中,被试的能力、题目的 3 个参数和正确作答的概率都是未知的,测量学家的工作就是要根据被试对于一组题目的作答情况(通常称为“反应矩阵”)把这些未知量估计出来。如果有 N 个被试,对 n 道题目进行了作答,那么需要估计的参数就有 N+3n 个,这显然是一个非常困难的问题。幸运的是,心理学家是一支高智商的队伍,很快就提出了运用极大似然法[8]、贝叶斯方法[9]、马尔可夫链蒙特卡罗方法(MCMC)[10]等参数估计技术。在 IRT 的发展过程中,还涌现出了多级记分模型[11]、多维项目反应模型[12]、展开式模型[13]等,这些模型组成了项目反应理论丰富多彩的大家庭。

2.3 概化理论

概化理论(Generability Theory, GT)也是为了克服经典测验理论(CTT)的缺点而发展起来的。CTT 把测验分数划分为真分数和误差分数,这种貌似简单的方法使得人们不能判断误差究竟是何种原因造成的,也就无法针对性地寻找减少误差的措施。GT 采用了方差分析的方法,把造成误差的各种来源都进行了考虑和分析,并提出了绝对误差和相对误差的概念和计算方法。

CTT 的信度理论建立在严格平行测验的强假设基础上,也就是要求两个平行测验的实测分数必须具有相同的平均数和方差,这对于实际工作者来说,是勉为其难的事情。GT 在这方面就比较有人情味,它的分析计算建立在随机平行测验基础上,即随机取自同一题库的长度相同的测验,这让大多数实际工作者松了口气。

在 CTT 中通常用多种信度来描述同一个测验的精确度,而这些信度之间又缺乏内在的关系。GT 则采用了概化系数、可靠性指标、信噪比等指标来描述测验的精确程度,而这些指标具有内在逻辑关系[14]。基于概化理论,人们可以对诸如作文、面试等多个评委主观打分的测量结果进行深入分析,并找到减少误差的方法。

3 与新时代同步,和高科技结缘

心理测量学是一门充满灵气的学科,它深知要持续汲取高科技的雨露,才能永葆青春的活力。自 20 世纪后期以来,计算机科学、认知科学、计算智能等迅速发展,心理测量学尽情地吸允这些新兴科学技术的营养,使自己成长得更加枝繁叶茂。

3.1 计算机化自适应测验

在“国家形象宣传片”中有姚明和丁俊晖站在一起的镜头,设想让他们穿上完全相同的服装,那一定会使观众忍俊不禁。可是在我们的生活中,人们总是习惯地认为,让所有的考生都使用相同试卷是最公正的,其实并非如此,考生的能力水平有高有低,对于同一张试卷,优秀考生的水平不能充分发挥,而后进考生则是依靠猜测来答题,对他们来说,这是一张不公正的试卷。

心理学家机灵地把项目反应理论和计算机科学相结合,提出了计算机化自适应测

验[15]，其核心是由计算机根据考生的能力水平自动选择测试题目，并最终对考生能力进行估计。心理学家设计了多种试题选择方法，包括最大全局信息量、最大加权信息量、分层选题、全贝叶斯准则等策略[16,17]，这样每个考生所面对的是最适合其水平的题目，考生能力可以得到充分发挥，考试时间将会大大缩短，同时也保证了题库的安全性。

3.2 认知诊断理论

认知科学是21世纪的前沿科学，认知诊断理论是心理测量学和认知科学相结合的产物。传统的测验只能提供一个分数，但实际上，得分相同的考生未必是完全相同的，例如两位数学得分相同的中学生，可能其中一位的代数能力较强，而另一位则是几何能力较强。如果只看其数学测验的总分，就无法将他们区分开来。

认知科学的蓬勃发展为心理学家提供了天赐良机，他们果断地将心理测量学和认知科学联姻，诞生了崭新的认知诊断理论[18]。从线性逻辑斯蒂模型、多成分潜在特质模型、规则空间模型、属性层次模型[19]到统一模型、总分模型、NIDA模型、贝叶斯网络模型、DINA模型[20]等数十种认知诊断模型相继问世。这些研究成果能够为学生提供诊断性报告，使得他们摆脱题海战术，提高学习效率。

3.3 计算智能的应用

计算智能是以生物进化的观点认识和模拟智能，它的主要方法有人工神经网络、遗传算法、模拟退火、蚁群算法、粒子群算法等，心理学家始终对该学科的发展保持敏锐的关注。

人工神经网络具有很强的自学习能力，已经被应用于各个领域的模式识别，认知诊断的实质就是对学生进行模式识别，Almond等人在2007年就提出将神经网络方法运用于认知诊断[21]。由于神经网络能够很好地对各个变量间的非线性关系进行拟合，它通常被运用于测验效度检验，如，心理学家采用该方法帮助军队进行人员选拔[22]。在远程教育中，它还和心理测量的项目反应理论结合起来，用于设计个性化的e-learning系统[23]。

遗传算法作为一种高效的全局并行搜索优化算法，适合于处理多目标优化问题。在心理测量中存在很多优化问题，如项目反应模型的参数估计，就是要寻找一组项目参数和被试能力值，使得它们代入模型后所得到的反应矩阵和被试的实际作答情况最为接近，有学者将遗传算法运用于项目反应模型的参数估计，取得了很好的结果[24]。测验试卷的组卷是多目标优化问题，它要满足测量误差最小、试卷的内容最符合预先设计要求等目标，因此遗传算法也是智能组卷的好方法[25]。

在心理学的历史上，心理测量学就是这样与新时代同步，永远保持着青春的活力；和高科技结缘，时刻放射着灿烂的光辉。

4 广阔的前景，严峻的挑战

几十年来，心理测量理论从CTT的线性模型到IRT的非线性模型，从IRT的单维模型到多维模型，从只考虑单个方差的CTT模型到专注方差分解GT模型，从只有一个总分的测验模型到对被试进行模式识别的认知诊断模型，从使用统计方法到计算智能的应用，心理测量学从来没有停止过前进的步伐。

4.1 教育考试

教育是使用测量手段最多的领域，其中最主要的就是考试。我国对于教育考试一贯高度重视，无论是在考试的科学性还是安全性方面都做了大量卓有成效的工作。但是，从测量理论的角度来看，还有许多课题值得深入研究。

首先是项目反应理论、概化理论、认知诊断理论等测量理论如何与我国的各种考试相结合。高考牵动着亿万人民的神经，究竟如何将现代的测量理论运用到高考中，使之更加科学、更加公平？由于我国高考的内容多、题型多，测验的数据是否能够拟合于现有的测量模型？是否还需要研发新的模型？

对于一个使用区域宽广的考试，应该重视对它的项目功能差异（Differential of Item Function, DIF）进行研究[26]。任何一个测验都会受到各种无

关因素的影响,而这些因素对于考生全域中的各子总体的影响是不同的,这样就会形成项目功能差异。我国幅员辽阔、民族众多、城乡差异明显,这些都可能使得高考试题中出现项目功能差异,然而我们对高考中DIF的研究进行了多少呢?这是否会影响考试的公平性呢?

在教育考试中,还存在测验的统一性和人才需求多样性的矛盾。在教育改革中已经涌现出了多种测评方法,其中有许多采用了主观评价的方法,那么概化理论是否可以运用到这些方法中?

4.2 资格证书考试

我国设立的各种资格证书考试对于规范人才培养和用人制度虽然起到了重要作用,但也存在着某些亟需解决的问题。

首先是合格分数线的制定,也称为"标准设置"。这是心理学家长期关注的问题,不合格的医生会让病人血染手术台,不称职的律师会在法庭里上演闹剧,心理测量专家研发了多种划定分数线的方法[27],包括Nedelsky法、Angoff法、Ebel法、Bookmark法等多种方法,每种方法都有各自的优缺点,我国的资格证书考试是否对这些方法进行过深入研究?采用何种方法来划定分数线?其科学性如何?标准误差是多少?

其次是测验的等值,测验如同一把尺子,对于同一种资格证书考试,每年的试题都不同的,因此要把不同年份的测验分数转换到同一把尺子上,这就是测验的等值,它是保证测验公平性的重要手段。心理学家提出了多种测验等值的方法,包括平均数等值、线性等值、等百分位等值、动差方法等值、特征曲线等值,同时还给出了等值误差的计算方法[28]。在我国的各种资格证书考试中,采用何种等值方法?等值的误差是多少?能否做到每年都是用相同的尺子去度量考生?

4.3 工业和组织心理测量

在国际上,心理测量还被广泛地应用于工业和组织心理学领域[29],人员的招聘、选拔和考核,员工满意度、工作负荷、组织氛围、自我效能感、团体凝聚力的测量都是成熟企业的常规工作。

在民用产品开发方面,国际上通常采用"消费者驱动的产品开发"[30],新产品的设计决不是技术人员拍脑袋的产物,而是首先测量消费者对于产品功能和外形的需求,然后根据测量结果进行产品设计,否则等产品问世后再去做推销工作,就事倍功半了。

由此可见,一个优秀的企业对于内部的员工和外部的消费者都需要进行心理测量,我们的企业做到了吗?

5 共同呵护这颗美丽的明珠

心理测量学是我们身边的明珠,人们爱她疼她。爱她,是因为她毫无半点私心,来到人间就是为了苍生;爱她,是因为她绝不说半句假话,她是建立在CTT、IRT和GT科学基础上的;爱她,是因为她永远那样坦诚,非但向你提供测量的数据,还告诉你估计的误差;爱她,是因为她和时代同步,与科技结合,不断朝气蓬勃向前进;爱她,是因为她有广阔的前景,社会各领域都在期待她的加盟。疼她,是因为在前进的道路上,她还面临着巨大的挑战,多维模型的参数估计、测验等值、项目功能差异、标准设置等都有许多困难需要克服。对于这样的学科,我们要爱她疼她,更要竭尽全力支持她。

5.1 建立良好的法制环境

测验是一把尺子,各类考试机构挥舞着这把尺子,给考生们贴上合格和不合格、录用和不录用的标签。对于这关系到广大人民群众切身利益的事情,我们是否应该制定一些法规?中国心理学会公布过《心理测验管理条例》和《心理测验工作者的道德准

则》，但这仅仅是学术团体的文件。政府是否也应该有所作为？是不是可以要求考试机构把他们如何设计和制造这些尺子、如何保证尺子质量的信息向社会公示？国家是否应该制定相应的法律来规范这些尺子的研制工作？政府是否应该有专门机构来监督各种测验的编制和施测？

5.2 培养高质量的人才

作为一门学科，心理测量学工作者不仅需要掌握心理学的知识，还需有深厚的数学和计算机功底，这就需要在学科规划、教学计划制定方面站得更高、看得更远。要培养优秀的学生，首先要有潜心做学问的老师，可以调查一下，一年365天中，我们的老师有多少日子是全身心地投入到测量理论的钻研和计算机程序的编制中呢？如果这些专业工作者都不能潜心于学问，哪里还能培养出高质量的学生？

5.3 加强科研力度

心理测量学是通过对人的外显行为的分析，做出对其内隐心理特质的定量描写，这个任务非常艰巨！然而研究手段很少，除了测验还有哪些是真正有战斗力的？只有加强科研力度，才能让这颗明珠更加闪亮！

心理测量学，当我第一次和她相见时，就被她的无穷魅力所吸引。她让各种内隐的心理特质用数字的形式呈现在我的面前，她把心理变量间的复杂关系转化为清晰的数学公式，她让数学的明珠在心理学的皇冠上闪烁光芒。我真心希望每天跟她一起去迎接前进道路上的挑战：建模、参数估计、等值、项目功能差异、标准设置、认知诊断等等。心理测量学向我描绘着美好的前景：教育、医学、工业、军事等各个领域都可以大显身手，让我永远充满青春的激情。

致谢 感谢南京师范大学钱锦昕、沙如雪、张潇，丹麦哥本哈根大学梅竹等同志在文献搜集、资料整理等方面给予的帮助和大力支持。

参考文献

1 Anastasi A. Psychological Testing (7th ed.). New York: Macmillan, 2009.

2 Cattell J. Mental tests and measurement. Mind, 1890, 15: 373-381.

3 Weimer W B. The history of psychology and its retrieval from historiography: The problematic nature of history. Social Studies of Science, 1974, 4: 235-259.

4 Boring E G. A History of Experimental Psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1950.

5 Lord F R, Novik M R. Statistical Theories of Mental Test Scores. Mass: Addison-Wesley, 1968.

6 Lord F M. Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

7 Hambleton R K, Swaminathan H. Item Response Theory: Principles and Applications. Boston, MA: Kluwer-Nijhoff, 1985.

8 Bock R D. Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 1981, 46(4): 443-459.

9 Baker F B, Kim S H. Item Response Theory: Parameter Estimation Techniques (2nd ed.). New York: Marcel Dekker, 2004.

10 Kim J, Bolt D M. Estimating item response theory models using Markov Chain Monte Carlo methods. Educational Measurement: Issues and Practice, Winter，2007, 38-50.

11 Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969, 34: 386-415.

12 Reckase M D. Multidimensional Item Response Theory. In: C. R. Rao (Ed.), Handbook of Statistics (Volume 26). Elsevier B.V. 2009.

13 Andrich D, Luo G. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Applied Psychological Measurement. 1993, 17: 253-276.

14 Brennan R L. Generalizability theory. New York: Springer-Verlag, 2001.

15 van der Linden W J, Pashley P J. Item selection and ability estimation. In W. J. van der Linden & C. A. Glas (Eds.), Elements of Adaptive Testing. New York, NY: Springer, 2010.

16 Barrada J R, Olea J, Ponsoda V, Abad F J. A method for the comparison of item selection rules in computerized adaptive testing. Applied Psychological Measurement, 2010, 34, 438-452.

17 Chang H H, Qian J H, Ying Z L. A-Stratified multistage computerized adaptive testing with b blocking. Applied Psychological Measurement, 2001, 25: 333-341.

18 DiBello L W, Stout W. Editors' Introduction and overview: IRT-based cognitive diagnostic models and related methods. Journal of Educational Measurement. 2007, 44: 285-291.

19 Gierl M J. Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. Journal of Educational Measurement, 2007, 44: 325-340.

20 de la Torre J. DINA model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics, 2009, 34(1):115-130.

21 Almond R G, DiBello L, Moulder B et al. Modeling diagnostic assessments with Bayesian networks. Journal of Educational Measurement, 2007, 44: 341-359.

22 Arendasy M, Sommer M, Hergovich A. Statistical judgment formation in personnel selection: A study in military aviation psychology. Military Psychology, 2007, 19 (2): 119-136.

23 Baylari A, Montazer G A. Design a personalized e-learning system based on item response theory and artificial neural network approach. Expert Systems with Applications, 2009, 36(4): 8 013-8 021.

24 Li F, Jing F. Estimation of multidimensional item response theory models in person parameter base on genetic algorithm. Proceedings 2010 International Conference on Anti-Counterfeiting, Security and Identification, 2010, 207-210.

25 Yong O Y, Luo H F. Design of personalized test paper generating system of educational telenet based on genetic algorithm. Proceedings of 2009 4th International Conference on Computer Science and Education, 2009, 170-173.

26 Kim S H, Cohen A S, Alagoz C et al. DIF detection and effect size measures for polytomously scored items .Journal of Educational Measurement, 2007, 44 (2): 93-116.

27 Lin J. The bookmark procedure for setting cut-scores and finalizing performance standards: strengths and weaknesses. The Alberta Journal of Educational Research, 2006, 52(1): 36-52.

28 Kolen MJ. Brennan R L. Test Equating, Scaling, and Linking: Methods and Practices. New York, NY: Springer, 2004.

29 Budworth M, Latham G P. New directions in industrial-organizational psychology. Canadian Journal of Behavioral Science, 2009, 41(4): 193-194.

30 Sandmeier P, Morrison P D, Gassmann O. Integrating customers in product innovation: Lessons from industrial development contractors and in-house contractors in rapidly changing customer markets. Creativity and Innovation Management, 2010, 19(2): 89-106.

Psychometrics: Embedding Mathematical Pearl in Psychology Crown

Yu Jiayuan

(School of Psychology, Nanjing Normal University 210097 Nanjing)

Abstract Psychometrics is a subject which studies psychological trait with quantitative method. It has emerged and developed under the powerful promotion of social demand. Classical test theory, item response theory and generability theory are its foundation. Computerized adaptive test (CAT) is the combination of psychometrics and computer science. Cognitive diagnosis theory (CDT) is the integration of psychometrics and

(转至 140 页)