

略论心理和教育测量理论的发展历程^①

漆书青

(江西师范大学 教育科学学院,江西 南昌 330027)

摘要:从19世纪末科学的心理和教育测量诞生到现在,其测量理论不断发展和丰富。具体内容与形态有:常模理论、真分数信度理论、实证效度理论、概括力理论、项目反应理论等。测量理论跟心理学实质理论、计量模型与认知模型的相互结合,是新世纪测量理论发展的重要方向。测量理论研究存在一个由客观性科学原则的简单移用,到对被测现象能动性、与测量活动主体间的自觉把握,这样一个方法论上逐步探索转变的过程。

关键词:心理量测量;真分数理论;概括力理论;项目反应理论;计量模型;心理学实质理论

中图分类号:B84 G449 **文献标识码:**A **文章编号:**1000-579(2001)01-0094-06

测量理论是测量活动内在关系的抽象理性和逻辑形式化的反映;就其内容说,包括测量对象的辨别、导引与确定,测量规则选用与体现测量规则的测量工具制作使用,以及测量结果分析处理的理论。就物理测量来说,测量理论中的前两方面,跟物理学本身的实质理论直接融合在一起。无论力学还是热、声、光、电磁学的测量,对象的导出与确定,工具的制作和使用,都是力学以及其他物理学本身内容的一部分。由于物理测量中,直接与间接测量、基本与导出测量,其中各有关物理量的联系都是确定性的函数关系,因而干扰测量过程导致误差的因素,均属外部因素。所以,测量结果分析中的计量理论,主要就是误差分析。正因如此,测量正确有效性分析,主要就是有关物理量关系即物理学实质理论的分析,以及测量工具是否正确体现这种物理量关系的分析,包括如何实际确定测量参照点和取定测量单位的分析,均未超出物理学理论本身的范围。只有由外部因素造成的误差,是要在施测过程中认真加以控制的,而它们又是随机出现的。这样,分析误差的计量理论就不能再采用确定性关系的函数模型,而应采用概率模型了。随着人类物理测量实际活动的深入开展,到18世纪,人们已经明确认识到,测量误差服从正态分布,或称高斯分布;提高测量精度的工作,就要遵循误差理论的指导来进行。误差理论是科学的测量理论,尤其是测量计量理论的一项重大成就。

心理物理量的测量是19世纪才开展起来的。由于绝对和差别感觉阈限这种心理物理量,跟外部物理刺激量间,也存在着相当确定的函数关系;虽然这属于非线性的对数函数或幂函数关系,但终究跟物理量的测量十分类似,因而,其测量理论中除对象引出和工具制作等有特殊内容外,测量计量理论的中心仍属误差理论。

测量高级心理过程的心理和教育测量,这时心理量的测量,情况就大不相同。内部心理量与外部动作、语言和情感表达的关系,认知变量与人格、动机变量,以及认知领域和人格领域内部变量之间的关系,均不再存在确定性函数关系了,都是主体能动性的表现。这样,要寻求其

① 收稿日期:2000-10-13

作者简介:漆书青(1937-),男,江西萍乡人。江西师范大学教育科学研究所所长,教授。

间的定量关系,就不能不采用概率模型这样一类非确定性函数关系模型。也就是说,心理量测量理论的各种部分,普遍要采用概率模型。

心理量的辨明、导引与确定,当然也要依靠心理学实质理论。然而,由于心理量的能动与复杂性和心理科学本身发展水平的限制,测量对象的导出,测量工具的编制,常常只能借助于经验,而很难按照严格形式化的科学理论来进行。其测量过程,大体说来,无非是采用口头语言、书面文字以及动作操作等方式。学业成绩考试、智力测验、人格问卷、社会态度测查等等的编制,无不体现出这种浓烈的经验性。从19世纪末高级心理量测量的科学研究工作开展至今,这方面的改进极其有限。试题与项目的陈陈相因现象,一直非常突出。比较可予称道的,就是出现了选择题这种客观题型。

物理量一般都能在等距乃至比例尺度上指定值,借助测量工具来进行比较与加和操作,在线性的等距或比例量尺上指定值,逻辑清楚明白,是不困难的。所以,物理量测量中,量度化(量表化)理论实不复杂,并未构成测量理论的显著部分。心理量的测量情况大不相同。首先,基础测量中,每个项目的结果,要量化都要经过主体(如阅卷者,或态度调查中的被试)作出判别、意义解释,甚至是一定的分析推论,才能在顺序尺度而很少能在等距尺度上指定值。其次,导出测量值,如整个测验的总分,能否由项目得分累加确定,也是很成问题的。因为,各项目的难度(或强度)并不相等,各分值间未必存在可交互替换性。因而,心理量的量表化(量度化)理论,确应成为测量理论的关键部分。19和20世纪之交,当时心理和教育测量的先驱们,如比纳和桑代克等,面对这一困境,就先把基础与导出测量间的关系置于不顾,而简单地把测验中各个分价值直视为可交互替换的,只要总分高就认定为所测特性“含量多”。这样,就把量表化(量度化)的重点转移到测验总分的处理上来,并采用概率统计观点,提出了常模理论。常模理论就是心理量测量中第一个量表化(量度化)理论。通过常模,尤其是组内常模,就将所测对象转化到线性的等距量尺上,或说给其在线性的等距量尺上指定值了(当然,早期的比纳年龄常模,还只能说是在线性的顺序量尺上指定值)。这里,测量参照点就是常模组的平均数,而单位就是常模组的标准差,并且具有等单位性。常模理论是测量理论中超出物理量与心理物理量测量已有成就的、人类科学理论的一大成就。它用常模组内相对地位量数来刻划所测被试的心理量,使其实现了量表化,为心理量计量提供了一个科学武器,影响深远。这样做解决量表化问题的策略确属独创,但其逻辑基础却不能说毫无瑕疵。

承接物理量和心理物理量测量的研究传统,20世纪前半叶发展起了高度理论形式化的心理量测量的误差理论,即真分数理论(CTT)的信度理论,并同样相应地构成了20世纪前期与中期的心理量测量理论的主导部分。结果,通常所说的心理测量的经典理论,意即指真分数理论,主要指其信度理论。跟物理测量类似,真分数理论的经典理论假定:①所测的量(即真分数值)是不变的(这当然偏离了心理量的本质);②测量误差跟所测心理量彼此独立(这也缺乏足够的事实根据),取值呈正态分布;③观察分数等于真分数加误差分数之和。意即无限多次反复测量,观察分数的平均数就等于所测心理量的真值。所以,真分数理论可用下列数学模型表述:

$$X = T + E \quad (1)$$

其实,这一线性数学模型是一种测量操作过程模型,是从物理测量活动中经验地概括得到的,但却被当作心理量测量的经典模型。而真分数理论的信度概念,除了要遵循这一并不很适合心理量测量实际的线性模型,即要以这个模型代表的三个假设为前提来导出,还要遵循所谓

的“平行测验”观念(当然,严格的平行测验实际上是编制不出来的)。即还假定,在大量平行测验施测于被试时其真值保持不变,并且这样得到的误差分数方差会等于任一平行测验测量大量被试后所得到的误差分数方差(这种假设实际上也无法验证)。在以上所有这些假设成立的前提下,真分数理论导出了如下结论:“测验信度系数等于真分数方差对误差分数方差的比。”即

$$r_{xx} = \frac{\sigma_T^2}{\sigma_x^2} \quad (2)$$

另外还有相应的测量标准误与信度指数(即观察分数跟真分数的相关系数)的求取公式。这样一套理论,虽然并不严谨,颇显粗糙,但却也较实用。从诞生起到今天,确也为控制测验误差、改进心理量的测量技术作出了贡献。真分数理论的测量误差,是就某一被试组来说的,各种水平被试同等看待。这就抹煞了测验难度跟被试水平相适应时,其误差要小于难度跟被试水平相距很远时的误差这种重要事实。另外,根据这种信度理论的要求,就要选取内部相关一致性高,而且难度只适合于中等水平被试的测验项目来构成测验,这实际上也就限制了多种功能类型测验的编制。

按照上述这种测量操作过程模型的逻辑,信度就是反复施测时所得测验分数的一致性。然而,反复施测所得分数虽然前后一致,仍然难以保证测验所测到的东西就是应该测到的东西。这跟物理测量不同。在物理量的测量中,直接与间接测量间,基本与导出测量间,有关物理量的关系是由严格逻辑甚至是数学形式化理论所保证了,是可以直接作客观检验的。而心理量的测量所测的是主体的具有能动性质的量,心理量间的关系是非确定性函数关系,测验编制具有很强的经验性。于是,测量工作的正确有效性,就难以从心理量内部关系的检验来得到保证,而要从系统外,将测验所得结果去跟其他效标测量值作比较来予以检查。这样在信度理论之外,也就有必要发展起效度理论来。历史上首先发展起的效度理论,是统计效度,即并存效度与预测效度理论。但是并存效度逻辑上有循环论证的毛病。其效标测量(如教师评价)若高度可靠,所考察的测验(如入学考试)实际上就没有编制的必要了;而效标测量若不可靠,正因此而需要另行编制新测验,这种新编测验有效性的证据又难以找到了。这种逻辑上循环的毛病十分显著。预测效度的效标资料,则要在测验实施后,常常是按测验结果决策处理后(如根据入学考试成绩选录学生一学年后)很久才能得到。这样效度验证对测验编制工作的实行与改进来说,实在是过分拖后的“迟到客”了。所以,统计效度并不理想。其次是内容效度理论,但其经验性成分过强,在计量理论中难于取得重要地位。后来在智力测验相关分析研究的推动下,发展起了结构效度理论。到20世纪中后期,不少人(如美国的安娜斯塔西)认为这是测验效度的最根本的概念形式。它在真分数与误差分数观念的基础上,使用多元统计方法,经过20世纪中后期的数十年的发展,现已构建起了探索性因素分析与实证性因素分析等模型,对心理量测量所测特质内部结构的探讨,确实作出了不小贡献。

心理量测量中,施测者与受试者都是具有主观能动性的主体,是主体间的交往关系,而不是物理量测量中的主客体作用关系。在同一测量情境关系条件下,视测量目标为何者的不同,施受双方地位可能对换变化。也就是说,有时学生是测量目标而阅卷教师是作为测量条件的测量侧面,而有时阅卷教师是目标而学生是侧面,地位可能变换。测量目标的差异带来的是真分数方差,而测量侧面的差异带来的是误差方差。真分数与误差也就不是固定不变的东西了。这跟物理量测量中的情况大不一样。那里带来误差的外部变量,是绝对不会跟测量目标

作对换转化的;并且它们的作用影响,可当作微弱的随机变量来看待。而在心理量的测量中,无论是学生差异,阅卷教师差异,以及试题性能差异,对测量结果取值的总方差的影响都是不容忽视的。于是,20世纪中后期人们就运用方差分量模型,在真分数理论基础上,对分析出真分数方差后的误差分数方差,再作进一步的方差分量分析。把原来笼统的误差方差明晰化,为测量误差的分析与控制开辟了深入发展的新方向,提出了概括力理论(GT)。概括力理论主要是一种信度理论,它用更准确有效的概括力系数(即G系数)来取代真分数理论的信度系数,深化和丰富了信度的内涵。并且,还在G研究的基础上提出了要作D研究(即测验设计研究),并开发出了D研究的策略与技术,为误差的预先主动控制开拓了道路,很值得认真注意。另外,它是在测量目标与各测量侧面综合考察的思想构架下来讨论问题的,所以也对测验效度即测验正确有效性研究作出了自己的贡献。

不论真分数理论还是概括力理论,都是在随机抽样的大理论框架下来展开自己的推论分析的。所以,不但要求抽取的被试是代表性样组,而且对试题(项目)的抽取也提出了要求。不过在试题的抽取问题上,真分数理论要求应是严格平行的测验,提出的是强平行条件;而概括力理论却不要求严格平行,只要是来自测查相同特质的同一题库的试题组就行,是弱平行条件。但两种理论都是以整个测验的性能为分析对象,并未去逐个考察作为测验基础的试题(项目)跟整个测验的关系,所以都属测量操作过程的随机抽样理论范畴。

20世纪中后期,逐步发展起了项目反应理论(IRT),又称潜在特质理论。它不再以整个测验为考察对象,而是径直以项目(测验试题)为考察对象,并在此基础上,再去分析作为测验基础的项目跟由它所构成的整个测验的关系。大家知道项目施测后,被试在其上可观察到的反应作答表现,理应由被试内部不可直接观察的心理量即潜在特质,以及用来测查这种潜在特质的项目性能如项目难度或强度等,这两个方面的因素所决定。对项目作答反应过程作出这样的分析,无论从逻辑还是实际经验的角度看,显然都是合理的。项目反应理论,就正是体现这种逻辑的项目作答的反应过程模型理论。按照项目反应理论的观点,被试在项目上答对答错的反应表现,跟其内部被测量的潜在特质的关系,仍然不是确定性函数关系而是一种由概率模型描述的关系。若潜在特质为 θ ,项目难度(或强度)为 b ,则被试在项目上的答对概率 $P(\theta)$ 就应该是潜在特质跟项目难度的差 $(\theta - b)$ 的函数(注意,这里是概率函数而非确定性关系函数)。亦即 θ 大于 b 时答对概率大, θ 小于 b 时答对概率小(也就是答错可能性大), θ 等于 b 时答对答错的概率相等。然而 $(\theta - b)$ 可以在正负无穷大的区间 $(-\infty, \infty)$ 取值,而答对(或答错)概率取值范围只能是在零与一之间 $(0, 1)$,经理论分析与经验观察(如考察题分在总分上的回归)可以发现,答对概率跟被试特质水平间的关系形成一条单调上升的点对称的“S”型曲线,它可用如下的项目特性函数来表述:

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}} \quad (3)$$

$$P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}} \quad (4)$$

前一式是当项目特性只须用难度 b 来刻划时的项目特性函数式,即单参数模型(Rasch模型)的项目特性函数式;后一式是当项目特性须用难度 b 、区分度 a 、猜测参数 c 来刻划时的项目特性函数式,即三参数模型(Logistic函数模型)的项目特性函数式。根据上述项目特性函数,采用最大似然估计,以及边际最大似然估计/EM方法,就可从被试的项目作答反应资料出发,直接将潜在特质水平标刻到线性等距尺度上,而不必再经由题分累加为测验总分,借助于总

分的常模资料,才迂回曲折地转化到等距尺度上去。前已说明,后面这种迂回道路的逻辑基础是很可疑的。正因为项目反应理论实现了在测验项目基础上直接将被试水平量表化(量度化)的突破,完成了比纳以来许多测量学家的心愿,因而它跟常模理论、真分数理论以及概括力理论都不同,不再同属于随机抽样理论范畴,而属于量表化理论范畴。

项目反应理论具有许多特点和优点。首先,它在估计被试特质水平时,不必要求抽取被试的代表性样组,具有被试水平参数不依赖于测验项目组,以及项目特性参数不依赖于所测被试组的参数不变性优点。其次,跟真分数理论中被试水平是项目组上的通过率,项目难度是被试组上的通过率,二者定义在不同的量尺上有别,项目反应理论中被试水平 θ 跟项目难度 b 定义在同一量尺上,因而可直接比较,更确切而真实地反映了 θ 与 b 的本质。第三,项目反应理论定义了项目信息函数与测验信息函数,正视了测量误差的大小跟项目性能是否与被试水平相适应有关这一客观事实,从而开辟了另一种新的测验信度观,并为误差的预先分析与控制提供了强而有效的方法与技术。第四,提供了计算机化自适应测验的测验实施的崭新策略,不但有利于提高测验的信度、效度与效率,而且为心理和教育测量技术跟现代信息技术的结合,开辟了广阔前景。第五,在常模参照特别是标准参照测验的编制,测验等值,题库建设,试卷的自动生成,项目功能差异分析,乃至效度验证等方面,都从计量学角度提出了自己的新观点与新技术。总之,它推动了心理量测量理论的全面改进,故被人誉为有别于真分数理论这类经典(古典)理论的现代测量理论。

然而在实际上,直到今天为止,项目反应理论提出的项目特性函数,在技术上较为成熟的仍然是适于 $(1, 0)$ 计分的单维模型。当然,单维的等级计分模型,如塞姆吉玛的等级计分模型、单参数的以及拓广的分部评分模型等等,在技术上也有了长足进展。现在国外流行的 MUTHLOG 和 PARSCALE 等参数估计与测验分析系统,就都是这方面突出的技术成果。这样,不但测验中不同难度项目所得分数缺乏交互替代性问题,而且同一项目不同等级间难度有别、同一项目不同加工步骤上难度有别等等带来的困难,均在一定程度上获得解决。只是这类模型在测验实际工作中的应用还相当有限。就我国情况说,更属起步阶段。至于多维模型,无论国内外当前均仅在作理论探讨,真正大规模应用,尚需时日。

20 世纪中期以来,认知心理学勃兴,而且统计分析与实验设计的方法技术也有很大进步,再加上计算机技术的广泛应用,对人的心理特性结构与心理加工过程的探索,取得大批突出成果,许多过去的“黑箱”问题正在逐步“灰化”,渐趋明朗。无论是真分数理论还是项目反应理论,把心理量或“潜在特质”视作单纯统计结构而不具实质心理学内容的态度,自然就显得落伍而不合时日了。克隆巴赫很早就批评只搞相关分析而跟实验心理设计取隔膜态度的心理计量学家,要求把相关分析传统跟实验设计传统二者结合起来。到七、八十年代,美国的不少心理计量学家更明确提出,测量学要跟实质心理理论、计量模型要与认知模型两相结合,创建出新一代测验理论(Test Theory for a New Generation of Tests)的口号。实际上,项目反应理论所提出的量表化(量度化)理论概率模型,是一个宽广的构架,不仅可以在试题(项目)这个层次上将反应资料量表化(量度化),而且可以在作答项目时的加工成分层次上、加工过程步骤层次上,将反应资料量表化(量度化),从而为这种结合提出了必要性和给出了实际可能性。当前,测量理论跟心理实质理论的结合,已经形成势头良好的研究潮流。斯腾伯格、戴斯、考夫曼、埃姆布里逊等人都已就此作出出色贡献。这种结合不但会对量表化(量度化)理论,而且会对多年来困扰测量学界的效度理论的发展,带来强有力的推动。测验编制工作包括测验目标导出、测验项

目开发与性能预控、测验结构掌握控制等工作,长期依赖经验的局面,或许会出现重大改观!

从19世纪末科学的心理和教育测量诞生至今,其测量理论不断发展丰富,具体内容与形态择其大端依次有:常模理论、真分数信度理论、实证效度理论、概括力理论、项目反应理论等。当然,测验编制从来就离不开心理学实质理论支撑,计量分析时由于采用概率模型而非确定性函数模型,所以,测量计量理论一度曾出现过统计概念缺失心理学意义的问题。在20和21世纪之交,测量模型与心理学实质理论相结合要求的提出,正是自觉克服这种弱点的科学意识的觉醒与光大。相信在新的世纪中,现代测量理论必将取得飞速的进步,在心理和教育科学的百花园中开出鲜艳的奇葩。

(责任编辑:宋 毅)