教育的标准化测验

王　权　邱学华

- 标准化测验的概况和特点
- 标准化测验的设计
- 标准化测验试卷的质量要求
- 测验分数的解释
- 成绩测验的编制
- 能力测验的编制
- 美国斯坦福标准化测验
- 日本小学数学标准化测验

河南教育出版社

教育的标准化测验

王 权　邱学华

河南教育出版社

教育的标准化测验

王 极 卓学华

责任编辑 侯耀宗

河南教育出版社出版

河南金水印刷厂印刷

河南省新华书店发行

787×1092毫米 32开本 7.5印张 146千字

1988年10月第1版 1988年10月第1次印刷

印数1—6120册

ISBN7—5347—0155—4／G·125

定价 1.65元

前　言

当前教学改革中，考试方法的科学化和标准化是急待研究的问题。标准化测验是近几十年国际上广为流行的考试方法，我国借鉴国外经验进行标准化测验的试验，已先在高考中试用，尔后中小学也开始研究和应用。

为了推进考试方法的改革，我们编写了这本书。这本书系统地介绍标准化测验的基本知识，从历史概况，标准化测验的设计，直到具体地如何编制学习成绩测验和能力测验。

编写过程中，我们尽可能应用国内的资料，但是，国内的标准化测验尚处于试验阶段，还没有完整的资料。我们主要参考了美国伊利诺斯大学格朗劳特(N·E·Gronlund)所著的《教学的测量和评价》(Measurement and Evaluation in Teaching,1981)及密执大学梅伦斯等(W·A·Mehrens)所著的《教育的标准化测验》

(Standardiged Test in Education, 1980)

附录资料由孙重恩同志协助翻译，上海市教科所和常州市教科所的同志曾给予不少帮助，在此一并表示感谢。

由于缺乏资料，加上水平有限，本书难免会有疏漏和各种缺陷。我们热诚欢迎大家批评指正。

编　者

1987年9月

目　录

第一章　标准化测验的概况和特点

第一节　标准化测验的历史概况

标准化测验起源于智力的研究。英国生物学家高尔顿(F·Galton)对人类的遗传问题有浓厚的兴趣。他发现双亲与儿女之间、兄弟与姐妹之间、孪生孩子之间在生理、心理品质上有很多相似性。1882年他在伦敦创立了人体测量实验室，在那里他测量了人类视觉和听觉的敏感性、肌肉的强度和反应速度以及其他的一些感觉运动功能，并试图从中发现个体的智力与感觉特征之间的关系。他的这一研究虽然没有成功，但是高尔顿利用个体的外部行为特征来推测个体智力水平的思想，引起了心理学家极大的关注和兴趣。

一、心理测验的历史

首先把高尔顿研究智力问题的思想和方法介绍到美国的是心理学家卡特尔（J·M·Cattell)。卡特尔曾在德国从事实验心理的研究，1890年他在英国的《心理》杂志上发表了“心理测验与测量”一文，指出：“心理学若不立根基于实验与测量上，决不能够有自然科学之准确性。”他在美国编制了多种测验（都属感觉能力和动作过程的测量），并在

1894年对哥伦比亚大学的学生进行了测量。1905年法国心理学家比奈(A·Binet)经过多年的研究，与医生西蒙(T·Simon)一起发表了“诊断异常儿童的智力之新方法”的论文。论文中的量表有30个测验项目。可以从多方面来观测个体的智力水平的表现，并按由易到难的顺序排列，以便于比较不同个体之间的差异。1908年比奈又对1905年的量表作了修订，改编成包括59个测验项目，都以其难度归入一定的年龄组别，分作3岁组到13岁组。儿童考试结果，完成到哪一年龄组的测验题，就以该年龄表示之，称作智力年龄。智力年龄的核算方法有详细的规定。智力测验的许多基本观念实际上就是由比奈首先具体化的。所以心理学家宾特纳(R·Pintner)说：“在心理学史上，假若我们称冯德为实验心理学的鼻祖，我们不得不称比奈为智力测量的鼻祖。”

比奈的智力量表对心理和教育研究的影响是极其深刻的。在心理学方面，智力测验开始成为心理研究和心理诊断不可少的工具；在教育方面也引起了教育家对于个别差异的注意。为了适应个别差异，各种教育措施的研究也引起了重视，所以比奈的量表一经发表立即被介绍到英、美、德、意各国，以后又传到日本、中国、瑞典、土耳其等国。

1910年戈达德(H·H·Goddard)在美国首先把法文比奈量表翻译成英文，但是很快就发现法、美两国的不同文化背景使翻译的量表在美国不适用。于是斯坦福大学心理学家特曼(L·M·Terman)教授主持修正比奈量表。特曼花了五年的时间，按严格的手续修订，在1916年发表了“比奈——斯坦

福修正量表”。这个量表共有90个测验项目，其中54个是比奈量表原有的，36个是重新编制的。修订本对实施测验的步骤和计分方法作了详细的规定。此外，特曼的一个重要贡献是引入了智商的概念。在比奈的量表中分数是以智力年龄表示的。但是智力年龄仅表示一个儿童的智力的绝对数值，不能表明其智力水平的高低程度。所以特曼认为智力年龄必须与实足年龄结合起来考虑，才可以说明其高低程度。他提出比值

$$IQ=\frac{\text{智力年龄}}{\text{实足年龄}}\times 100$$

可以表明智力水平的高低程度，并称之为智商。乘以100是为了化去小数点，便于使用。例如一个12岁的儿童只能完成10岁组的测验题，则其智商 $IQ=\frac{10}{12}\times 100=0.831\times 100=83$。反之，10岁的儿童完成12岁组的测验题，则智商

$$IQ=\frac{12}{10}\times 100=120$$

特曼的1916年量表经过多年的使用，发现其内容重在测量语言方面的能力，缺乏足够难度的项目来评价较高的智力水平，而且仅有一种型式的测验，再测时只能重复使用。所以，1937年特曼与他的同事梅里尔(M·Merrill)又发表了比奈量表的第二个斯坦福修订本，分作L和M两种形式。1960年，L和M型两份测验中的最优项目又重新组合成一份测验。象前面两个修订本一样，它依然受到智力测验使用者的赞

赏，是著名的智力测验量表。

比奈式量表的主要局限性是测验只能个别进行，它是一种个别测验的量表。第一次世界大战期间，美国军方急需进行大规模的军事人材的选拔。当时奥蒂斯（A·Otis）和特曼已编有团体智力测验，军队的心理学家就以此为根据编成了第一个正式的团体智力测验，这就是著名的“陆军智力测验”。在战争以后的二十多年中，陆军智力测验一直是编制团体测验的样例。第二次世界大战期间，虽然智力测验得到广泛使用，并且出现了许多特殊测验，但很少有独特的或是创新的方法出现。当时的许多机械化部队，迫切需要选拔具备特殊能力的军人，于是成套的特殊能力测验陆续编制出来。所以，在这期间用以评价能力的各个分量的成套测验得到了迅速的发展。

第二次世界大战以来，智力测验在进一步探讨和处理智力的定义问题以及智力与种族、社会等级差异之间的关系的研究中发挥了极其重要的作用。

1958年，魏克斯勒（Wechsler）为成人设计了一套智力测验。在魏克斯勒的智力测验中，智商 IQ 只代表被试在同年龄组分数分布中的一个相对位置，称为离差智商。它是目前流行的一种成人智力测验。

二、教育测验的历史

教育测验主要是成绩测验，教育测验与心理测验所要测量的内容虽然不一样，但是所依据的基本原理和方法都是相同的。教育测验的发展虽在很大程度上受心理测验的影响，

但是教育的标准测量观念由来已久。1864年英国格林威治医学院教师费希(G·Fisher) 搜集了许多学生成绩样本，编成“量表集”一书。该书备有各科的学生作业样本，并对每一样本作业评有一个分数，以示优劣。当评定某生的成绩时，就将该生的作业与量表集中的各样本进行比较，找出与其相同优劣的样本，于是书中的样本分数即是该生的分数。费希的方法虽然还有不少主观性，但评定时已有标准可循，不失为一大进步。可惜他的工作当时没有被引起重视。

客观测验的实际发明者是赖斯(J·M·Rice)，赖斯是个训练有素的医生，在德国研究教育学，他对心理物理学方法应用于教育评价的问题有浓厚的兴趣。赖斯的最大成绩是首先编制了一个拼法测验，测试了将近三万个儿童。当时19世纪90年代没有自动的数据分析工具，甚至还没有台式计算机，所以全部计算均用手工操作。赖斯工作的更重要意义是发展了教育测量的一些基本观念。他认为要客观地评定学生的成绩，就必须在相同条件下，用相同的测验进行考查，在比较中才能作出客观的结论。

1904年美国心理学家桑代克(E·L·Thorndike)的“心理与社会测量导论”一书出版，系统地介绍了统计方法以及编制的测验基本原理。1909年桑代克在美国科学促进会的波士顿会议上首先发表了他的书法量表，这个量表是语文学科中第一种用科学方法编制的教育测量工具。1908年斯通(C·W·Stone)编制了一个算术推理测验，是最早的一种标准测验。初期的教育测验几乎全都限于小学的各科测验。1918年以后，

中学各科测验逐渐发展，一直扩大到高等教育。

第一次世界大战期间，由于美国军方利用智力测验筛选人材取得成功，大大地促进了客观性的成绩测验的发展。1923年，斯坦福成绩测验诞生，这是学校客观性成绩测验发展的里程碑。斯坦福成绩测验是一非单科性测验，它是当时美国公立学校主要学科的一个测验集，有几千名儿童作为参照组，用以解释测验分数。它的编制方法为以后的其他成绩测验提供了一个极有价值的蓝图。事实上它的一些基本方法一直沿用到现在。

三、我国的标准化测验历史

1918年我国小学教学法先辈俞子夷首先根据桑代克的书法测验仿编了小学语文毛笔书法量表，可算是我国最早的标准化测验。1922年“中华教育改进社”聘请美国教育测量专家麦柯尔来华指导编制测验，自此测验运动在中国开始。

在智力测验方面，1922年陆志韦主持修订了比奈——西蒙智力量表，陆氏修订比奈量表取样1400人，男女各半，年龄从3岁至30岁，程度从幼儿园到小学，有少数中学生。共包括65个测验项目。

廖世承参照美国“国家智力测验”编制了“廖氏团体智力测验”，分甲、乙两个量表，每个量表包括五种测验，适用于小学三年级至初中二年级。量表甲包含下列五种测验：(1)算术理解题，共15题，由浅入深；(2)填字，共21句，亦由浅入深；(3)理解的选择，共23题；(4)同——异测验，共50题；(5)形——数测验，共140个图形。量表乙的五个测验

是：(1)算术测验，共32题；(2)常识测验，共35题；(3)字汇测验；(4)比喻测验，共32题；(5)校对，共50题。

此外主要的还有刘廷芳的“刘氏中学智慧测验”，陈鹤琴的“陈氏图形智力测验”等。

在教育测验方面当时编制了较多数量的各科测验。在语文学科方面有默读、默字、识字、作文、文法和书法等量表，其中以陈鹤琴的贡献为最大，他所编制的测验几乎涉及了以上语文学科的各个领域。我国最早研究常用词汇的也得首推陈鹤琴，他结合编制测验调查各种语体文中的字汇，从554478个字的材料中分析得单字4261个，再统计各个单字在554478字的材料中出现的次数。他的研究成果对编写教科书和测验都有极其重要的价值。

在算术学科方面，俞子夷等人编就了四则测验、应用题测验等。此外尚有英文、史地、自然常识科的测验，无论从数量或质量方面看，成绩都是不小的。麦柯尔也曾表示说，有些测验并不亚于西方的测验。

由于旧中国政治腐败，教育行政部门的领导多为官僚，他们并不关心教育的发展。测验虽在研究人员中搞得轰轰烈烈，但没有能够逐步地普及到实际教育工作者，理论与应用严重脱节。

解放以后，由于教条主义地学习苏联，对测验和统计采取完全否定的态度，以致使我们对测验的学习、研究和应用中断了三十多年。四人帮粉碎以后，随着四化建设的发展，准确地选拔人材，科学地管理学校都急切地需要运用标准化

测验的测评手段，现在标准化测验已经开始引起教育行政部门和广大教育工作者的重视，首先在高考中应用，尔后在中小学各科的考试中试验应用。并且正开始着手研究和编制适应当前教育改革、有社会主义特色的标准化教育测验。

第二节　教育测量的基本特点和过程

教育是培养人材的专业，教育的产品是人材，任何产品合格与否都得运用一定的测量工具经过测定才能作出结论，教育也不应例外。但是对人的德、智、体，尤其是对"智"和"德"方面的测量与其他工、农业中的产品测量有着本质的差别。在工农业生产中，被测定的对象往往是一些实物的物理属性，我们可以使用尺、秤等测量工具直接与被测对象进行比较，从而获得某一数量结果，这种测量是一种直接测量。然而在教育测量（包括心理测量）中，我们所要测量的对象大多是学生对基础知识的理解程度、技能的熟练程度、记忆、想象和推理等能力的发展水平等等，这些学业成就，心理品质都是一些看不见、摸不着的不能感觉到的行为属性，我们不可能运用某种测量工具直接与这些行为属性进行比较，我们只能从行为表现来推测属性的品质，所以进行教育测量必须设法使被测属性以一定的行为表现出来，然后根据行为表现来推测属性的品质，也就是被测属性的品质是间接确定的，这种测量称为间接测量。

由于教育测量是一种间接测量，这就大大地增加了测量

工作的困难。这些困难之处正是教育测量不同于其他物理测量的特点，我们认清了这些特点，就会理解怎样才能正确地编制和使用教育测验，使教育测量在开始普及时就能沿着健康的道路发展。

一、教育测量的基本特点

1.被测对象的不明确性

教育测量的对象主要是人的行为属性或心理品质。行为属性与其他物理测量的物理属性，如长度、重量、温度、气味等是完全不同的。由于它不能为我们直接感知，所以教育测量面临的第一个问题就是我们所要测量的对象究竟是什么？比如我们要测量学生的数学能力，那末数学能力究竟是什么？各人可能会有不同的理解，没有统一的认识。再如测量人的智力，那末智力究竟是什么？从本世纪开始，世界上第一流的心理学家进行了大量的研究，但对智力的定义至今仍然没有统一的认识。数学能力的问题也有类似情况。在被测属性还没有统一的认识的情况下，要对其作出一个客观的并为大家所能接受的测定，当然是极其困难的。所以被测对象的不明确性是教育测量的第一个特点。当然并非所有的教育测量全部是如此地困难。例如测量的对象若是学生的某些学业成就，如学生学会了什么？做得怎样？则测量的对象就比较容易确定。

2.测定方法的间接性

如果对测量的属性有了统一的认识，继之而来的问题是如何来测定它们？假若被测属性不能为我们直接感知，则我

们有可能直接观测到的只能是它的行为表现。例如要测量学生的推理能力的发展水平，则就必须研究不同发展水平的推理能力会在怎样的不同行为中显示出来，或者说学生完成怎样的不同作业，则就可推测其不同的发展水平。所以进行教育测量，必须制订一套能使被测属性与行为表现之间建立一种对应关系的法则来，凭借这种对应关系，我们就可以透过"现象"，看到其"本质"。因此测定方法的间接性是教育测量的第二个特点。

3.数量表征的相对性

进行任何测量都必须使用测量单位，有了单位，被测对象才能用数量来表示。单位问题实质上也是个等值概念的问题。例如桌子的长是120厘米，即该桌子的长度有120个等长的距离。在教育测量中，一般总是以学生完成某一作业任务的行为，例如解释一个词或完成一道数学题，或进行一次推理活动来给相应于该行为的属性计分，从而得到一个数量结果。假如给学生甲解释6个词汇为6分，给学生乙解释3个词汇为3分。那末有什么理由认为解释这个词与解释那个词的行为是等值呢？有什么理由可以认为甲的行为的价值是乙的行为的价值的2倍呢？这种等值行为的规定在很大程度上是人为的，所以只能做到一定程度的合理，因此教育测量中的量值只能达到一定程度的准确，或者说数量表征的相对性是教育测量的第三个特点。

教育测量的这三个特点要求测验的编制者在编制测验时必须周密考虑、谨慎行事。教师明白了这三个特点，并能够