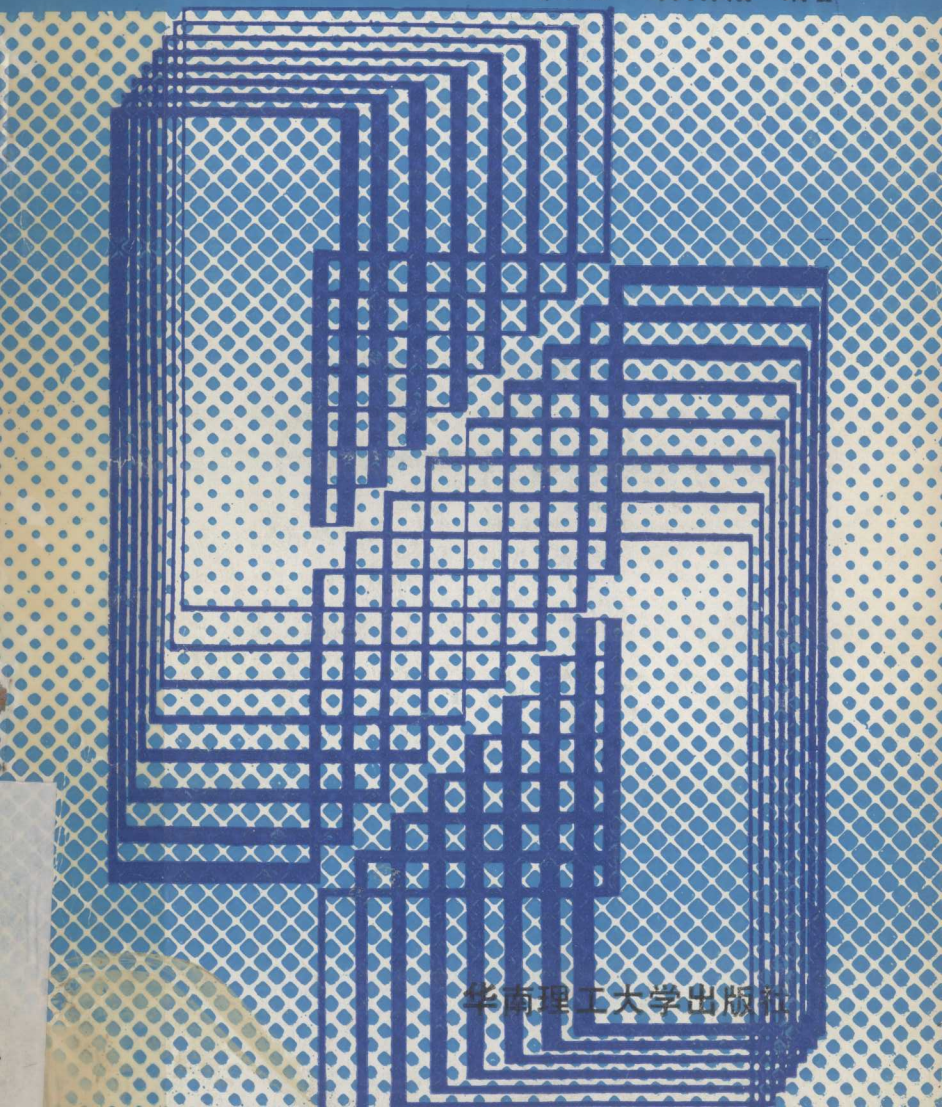
·现代教育测评丛书·

现代教育测量理论

邢最智　司徒伟成　编著



华南理工大学出版社

现代教育测评丛书

现代教育测量理论

邢最智　司徒伟成　编著

·现代教育测评丛书·

现代教育测量理论

邢最智　司徒伟成　编著

华南理工大学出版社出版发行
（广州　五山）
华南理工大学印刷厂印刷

开本787×1092　1/32　7.4印张　171千字
1989年12月第1版　1989年12月第1次印刷
印数：1—2 000
ISBN 7—5623—0160—3/G·22
定价：3.80元

序　言

教育的测量与评价是一个极其重要的问题。因为只有科学地测量与评价教育、教学工作，才能从根本上提高教学效果和达到教育的目的。过去有的教育测评工作由于在测量标准和评价依据等方面有较大的主观随意性，所以并没有起到上述的作用。目前这种状况已引起人们的注意，教育测评越来越受到人们的重视。

教育测评工程，一般可分为教学测评和教育测评两大部分。教学测评主要指具体的学科测评，即根据数学、物理、化学、生物、语文、外语、历史、地理、政治各科的特点，研究各科学业成就测量与评价的理论和方法；而教育测评则综合研究与教育有关的测量与评价的一般原理和方法。《现代教育测评系列丛书》正是研究上述问题的系列成果，暂包括《初中测试丛书》、《高中学业测评丛书》和《现代教育测评丛书》三辑。

《现代教育测评丛书》的特点是：一、各论注重吸收当今国内外教育测量与评价的一系列专著和参考文献的成果，使各书具有现代特色，反映现代趋势；二、各书注重与其他学科的相互渗透，特别注重用数学及统计方法研究教育测量和评价问题，使各书有显著的数量化的特色；三、各书在内容上注意理论联系实际，有实用性，对测量和评价方法均辅之以实例，方便学以致用，同时重视计算机在教育测评中的应用。

总的来说，《现代教育测评系列丛书》有其实践和理论意义，它既提供了中学各科学业成就测量与评价的方法，同

时具有一定的理论意义，为研究教育测评问题提供了理论依据。它的出版在国内还属首次，它将对现代教育的测量与评价起开拓作用。当然，该项测评系列研究才开始不久，难免存在不足之处，如实验的取样不够广泛，研究的结果还不够完善等。进一步的工作还有待大家共同努力。借《丛书》出版之际恳切希望国内外教育专家，广大教师和读者对该项研究工作提出批评和改进的意见，使我国的现代教育测评工作能逐步完善，实现国家培养四化人才的目标；同时期望《丛书》能为其他测评的研究指出一条路子和提供参考。

中国教育统计与测量研究会（筹）理事长

叶佩华

1987年6月于广州

目　录

第一章　测验理论的基本问题……（1）
第一节　什么是测验理论……（1）
第二节　标定方法……（4）
第三节　测验编制过程……（5）
第二章　信度……（18）
第一节　经典真实分数模型……（18）
第二节　信度估计方法……（24）
第三节　影响信度系数的因素……（35）
第四节　真实分数的回归估计……（38）
第五节　差分数的信度……（38）
第三章　拓广理论导言……（41）
第一节　G-研究和D-研究……（42）
第二节　单面设计下的拓广系数……（44）
第三节　单面嵌套G-研究……（57）
第四节　固定单面情形的拓广系数……（60）
第四章　准则参照测验的信度系数……（62）
第一节　范围分数估计的信度……（64）
第二节　掌握程度分类的信度……（67）
第五章　效度与验效……（79）
第一节　内容验效……（81）
第二节　准则—关联验效……（87）
第三节　造念验效……（92）
第四节　验效研究中的其它多元统计方法
……（99）

第六章　选拔测验中的偏见……（117）
第一节　基本术语与概念……（117）
第二节　偏见的心理测量学定义……（121）
第三节　公平选拔方法……（123）
第四节　选拔的决策论方法……（127）
第七章　测验构造中的题目分析技术……（131）
第一节　经典题目分析技术……（131）
第二节　IRT简介……（141）
第三节　IRT的应用……（161）
第四节　题目偏见的检验……（169）
第八章　测验记分及其解释……（177）
第一节　记分方法及对猜题的修正
……（177）
第二节　准则参照测验中标准的设置
……（183）
第三节　常模参照测验中的记分与解释
……（188）
第四节　来自不同测验的分数等值问题
……（208）
后　记……（225）

第一章　测验理论的基本问题

第一节　什么是测验理论

一、心理造念

许多学科都有自身的测量问题和测量理论，但不管具体的测量问题如何，各学科的测量均有一个共同点：所要测量的东西并非研究对象本身，而是对象的某方面特征。例如，化学家要测量某物质的分子结构，物理学家要测量物体的应力强度，生物学家要测量某细菌的存活时间等等，都是对现象或对象的特定属性进行测量的。类似地，教育心理学家并不是测量学生本身，而是学生的某种心理上的属性，诸如记忆能力发展、社会成熟程度、数学运算能力、智力、创造性等，我们将这些能表征一个人在工作、学校、社会和家庭中的行为的属性，叫做**心理属性或心理特质**。然而对心理特质的测量比较困难，因为一个人的心理属性仅仅是一种思维产物或造念（Construct），这是人们为了描述和解释人类行为而形成一种假设的概念。这种造念的存在是不能完全得到证实的，因此也就不象人的身高、体重那样可以直接测量到的。

对人的心理造念的测量只能通过对表征该造念的行为的观测而获得。也就是说，为了测量一个造念，必须在理论的造念和观测到的行为之间建立某些对应法则，从而获得该造念的一个适当的表征。又由于一个人的行为表现不能真正全部被观测到，我们只能观测到一个行为表现的样本，于是可以简单说：测验是从一特定范围中获取行为的样本的标准方

法。

二、心理造念的测量问题

从下面三个例子可以看出造念的测量在现实生活中是需要的。

例1　某校实验室需要招收实验员若干名，于是想对应聘者的技能进行测量。

例2　政府决策人想了解某项教育决策的社会心理反应，则需要对公民的政策态度进行测量。

例3　学校老师想要构造一个测验来测量学生对于微积分的掌握程度。

以上出现在例子中的造念，诸如技能、态度、成就水平，虽然互不相同，但都说明造念测量的必要性，也看出测量的复杂性。这主要表现在以下五个方面：

1。造念的定义十分难以精确化。我们是以体现这些造念的行为基础进行间接测量的。有可能不同的心理学家对同一个造念采用很不相同的行为来定义该造念。例如上面例子中，如何从学生表现出来的行为上判断他们对微积分知识掌握的程度呢？至少有三种方法测量：

方法1　拟定一组微积分计算题让学生去完成；

方法2　要求学生叙述微积分计算公式；

方法3　要求学生说明微积分与积分之间的关系。

这样一来，不同的测量方法导致了关于学生对微积分知识的掌握水平的不同定义。

2。行为表现的样本的局限性：我们无法对学生在解答微积分题上的全部行为表现进行观测，也就是说，测量只是以部分行为表现的样本为基础的。于是如何测量到最本质的有代表性的行为，便是测量理论的首要问题。

3．测量误差问题：对造念的测量容易受到众多误差来源的影响，诸如疲劳、厌倦、猜测、遗忘、粗心、错误评分、甚至题型等。于是如何去估计测量中的误差是一个重要的问题。

4．标定问题：测量量表上并没有一个定义好的单位。例如一个学生在一组题上没有做出回答，是否能说明他对该种知识一无所知？学生小张做对2题，小李做对4题，小王做对6题，是否就可以判定小王与小李之间的能力差异等于小李与小张之间的能力差异呢？从而量表的性质、标定单位、分数的解释便成为人们关心的问题。

5．对造念的测量必须表明与其它造念和可观察到的现象之间的联系。为此，Lord和Novick(1968年)指出：作为心理测量的造念必须从两个方面来定义：第一，必须借助于可观测到的行为来定义，这类定义说明测量是如何取得的；第二，必须从该造念与其它造念的逻辑关系或数学关系来定义。这是我们解释所得到的测量结果的基础。

可以说，整个测量理论的发展都是围绕着这五个基本方面的问题展开的。于是，测验理论的基本内容可以概括为二个方面：（1）如何估计这些问题对所获得的测量而产生的影响的程度；（2）构造各种方法来克服和减少这些问题的影响。

目前，世界教育与心理测量学界一般认为，R.L Thorndike在1904年发表的著作《心理测量与社会测量理论导论》是测验理论中的第一本著作，在以后的60多年间，人们不断对他的理论进行扩充和严格化，建立了有关能力、成就、个性和兴趣测量方面的基础，我们称此时的理论为经典测验理论。不过，在近20多年来，测验理论有了飞速的发

展，其涉及的内容和方法大大超出了经典测验理论的范围。尤其是电子计算机的应用，使我们可以建立数学上更为复杂的，统计效率更大的模型来处理测量分数。当然，对教育计划的评估和对学生或顾客各方面特质的测量的需求越来越高，也是促使测量理论发展的原因之一。

第二节 标定方法

任一种测量都要将观测结果量化，才有利于解释与应用。对造念的测量也不例外，需要制定有意义的测量单位，观测结果到量表值之间的对应法则，这就是标定问题，例如，古埃及人规定一个测量长度的单位叫“丘比特”（Cubit），此单位表示从肘到手中指顶端间的前臂长，从而凡被测物体含有此单位的数目就是它的长度。类似地，人们建立了各种教育和心理的测量量表来系统地对行为表现的样本进行搜集和赋值。为了得到教育或心理评估的一个量表，编制者应该预见由测量某造念而获得的资料进行标定的可能性如何？也就是要了解：（1）该造念的观测是否可用变动的量的形式来表现，如可以，则建立一个所谓心理连续谱将其量化；（2）在此连续谱系上具备实数系中的哪些基本性质（即序、原点、大小）。

Torgerson（1958年）提出三种标定方法：

1．对象中心法——此标定法是设法将个体（考生）标定到心理连续谱系的不同位置上去。此法广泛用于许多能力、成就甚至情感领域的测量上。例如，一位临床心理学家要构造一量表来测量抑郁感而编写出20个陈述句子，对每一句子指定用明确式与非明确式两个方式来回答。且规定明确式选答为1分，非明确式选答为0分，总分为考生在各题上得

分的总和。从而将获得高分者标定到连续谱系的正端，而低分者靠近负端。这种方法的目的仅在于对个体给予固定的标定，并不关心考生在回答的题目上可能产生的差异。即使要求考生对每个题目的某些层次作出选答，例如从“很同意”——“很不同意”，也仍属于对象中心法。总分的计算只是简单地假定对每题的可能回答指定一分数值，然后将各题分数加权相加，不过多数情况是采用等权相加，这样得到分数值的数值尺度具有“序”和“等度量单位”的性质。

2．刺激中心法——此法是将“刺激”（或“题目”）标定到心理连续谱上的不同位置上去。此法最早由德国知觉心理学家应用于实验室中，他们的研究兴趣是：身体对物理刺激（如光、音调、重力等）的反应及这些刺激之间的关系。并测出能为人体感觉到的刺激的最小变异量。例如：测量对光的知觉，可给出二个光源，要求考生说出哪个较强，从而产生一个办法对光感进行标定。比方说让一个或一组考生重复地去比较两个光源，当两个刺激等频率地被选出时，就说这两个刺激在知觉尺度上的标定值相等。当一个刺激的选择频率超过另一个的75%时，则说这两个刺激在心理尺度上的差异具有“恰好显著差异”（JND）。之所以选75%为准则，因为它位于偶然性与完全精确性之间的一半位置上。

3．应答中心法——这是由考生的应答资料，根据正确回答的题目的程度将他们标定到心理连续谱系上去，同时又根据考生若要正确回答题目所应具有的特质的程度将题目标定，所以这个方法兼有上面两种方法的特点。

第三节　测验编制过程

我们以对象中心测量为例，说明测验编制的全过程。即

我们欲将个体就其在某一特定的造念上的行为表现，将他们标定在一个数量化的连续谱上，一般过程分为十步。

第一步，说明测验分数的基本作用和目的

例如我国高考的分数是为录取高等学校新生而提供信息的，某公司招聘新雇员的测验分数应提供有关考生是否具有某特定专长和能力的信息；用于识别低能学生某特定弱点的诊断测验应提供对个人作出诊断决策的信息。显然，这些测验目的与作用不相同，因而题目的范围、数目、难度等也不一样。

第二步　确定代表该造念的行为

把对欲测量的心理造念的测量转化为特定的一组题目的过程是十分隐性的。人们需要确定说明该造念的行为，即将这些行为概念化，再编制一套能够测到这些行为的题目。可惜常常会导致忽略了重要行为，其中包括了一些只是编造者头脑中才有的东西。为了避免主观定义，提炼和证实可欲测量的造念的内涵，测验编制者可以求助下面方法：

1．内容分析

提出一些有关该造念的有启发的问题让研究对象去回答，然后将他们的回答按专题分为类别，于是出现次数占优越地位的那些专题便是欲测造念的主要成分。

2．评论分析

以人们最经常研究的和评论的行为来定义欲测的造念。测验编制者可以综合选择某些专家的工作来说明行为类别。

3．临界行为

给出一系列行为，它们可以表明欲测造念的连续谱上的极端值，例如可以请公司主管人描述一个情况，在该情景下雇员会最有成效（或最无成效）地工作，这样可获得一系列

临界行为，利用它们对工作成绩排等次。

4．直接观察

编制者通过直接观察发现一些行为。

5．专家判断

从对该造念有第一手经验的“专家”那里获取行为资料，例如要对医院里的护士工作成绩进行评估，可以调查一批护士长，了解一个好护士的行为应包括哪些成绩类型。

6．教学目标

由熟悉某学科的专家（如多年的教师）根据教学目标和学生的表现确定能说明学生行为的教学目标，这些教学目标将告诉编题者题目应测的特定内容是什么以及考生去完成时的工作表现如何。

第三步　准备一组有关测验题目编制的说明，确定在第二步所选定的每一类行为上的测题的比重。

有两类作用不同的测验需要区分，一是“常模参照”测验，其作用是在所给定的造念上区分考生，将个别考生的分数与其它考生的分数相比较获得测验分数的意义；二是“准则参照”测验，其作用在于测量考生在所给定领域上的绝对水平。例如，为了证实考生是否已达到某学科上的最低能力水平的测验。前者测验编制者应该组织那些在考生组上表现出一定分数差异的题目；后者测验编制者开始应有一组指导目标，确定考生应达到的成就范围。凡测题分数能在此范围上对考生进行推断的题目之全体叫“题目域”。自然测验的编制就是在此域上取样的。因为我们不可能编制出所有题目来。目前广为采用“题目说明”法来取样。题目说明法的内容包括：题目的内容来源，问题或刺激的描述，正确答案的特征，以及在多项选择题型中错误回答的特点等。如果拟题

者都能按照题目域的说明进行，便可以产生大量的“平行”题目，即具有“相互可交换”性，即使是不同地点，不同时间所拟定的题目也相对等价的。在制作大型题库时，这是一个很好的方法。

下面给出一个例子示范。

子技能：表明小数乘法的能力

题目说明：小数乘法的运算

刺激特点：

1. 题目要求含有两位小数、分数或混合式的乘法。
2. 题目以语句形式书写，必须有“求积”或“相乘”的说明。
3. 因子中应有一个为含三个非零数字，其余的含有3个数字，但恰好有二个为非零数字。
4. 每个因子的小数点后至少有一个非零数字。
5. 小数点后不要多于四个非零数字。
6. 分步的步骤数至少二个。
7. 组成每个因子的数字不要重复使用二次以上。

应答特点：

1. 格式

所有数值答案的选择或按递增次序或按递减次序排出。

2. 每题设置四个备择答案

（1）一个正确的答案；

（2）一个陪衬答案是在运算中分步中间发生的错误结果；

（3）一个陪衬答案反映了在运算中排列式发生的错误；

（4）一个陪衬答案或反映丢失小数点或反映小数点位置上的错误。

题目算式法和题目形式法是另一种表示题目域的方法。前者用于数值内容的题目，后者用于文字内容的题目。

下面是题目算式法的例子。

要求学生会计算两个正整数之间的差，算式为

“当$a>b$，求$a-b=$_____”

按此算式，只要将a、b代入合乎条件的具体数值，便可拟定出大量的题目来。上例是用一个算式说明，也可用几个算式来说明题目的。

题目形式法是写出题目的固定完整结构来，但其中含有一个或多个可变换部分。当然可变部分应是题目的关键或重要部分。那么通过不同的替换便可拟定出一组题目来。

下面是题目形式法的例子。

当一株柠檬树表现出A时，这可能会是_____症状。

1．缺乏营养；

2．除莠剂损害；

3．冻伤；

4．病毒感染；

5．细菌感染。

此题的可变换部分（A）可插入该树的某个典型的病理学症状，这里当然是有关叶与树皮的症状（如叶黄、皮裂等症），从而要求考生选择正确的答案填上。

当测验题目的说明或行为类别的说明给后，编制者应进一步考虑各部分题目在程度上的均衡性问题，即行为的不同类别上的题目依照其重要性与复杂性按什么比例出现的问题。如不解决这个问题，那么有的出题者可能凭主观过分强调这个目标，而另一些出题者又强调其它的目标，这一来，编制出来的测验将会很不相同了。特别在成绩测验中，至少要考虑到基本内容与认识过程这两个方面的题目。例如平面几何某一教学单元的目标有二个：

1．圆的定义及有关术语（半径、直径、圆心角、周长、面积）；

2．计算圆周、角度、面积等。

显然，目标1属于基本内容的回忆，而目标2既需要1的知识，又需要对各概念及其关系的认识。可见仅说明题目复盖的范围而不说明它所需要认知的水平也是不行的。

许多人建议采用层次系统法对认知行为分类，最著名的算是Bloom（1956年）提出的层次分类法。即：

1．知识——回忆与教学过程中的具有相类似形式的实际材料；

2．理解——对概念进行变换、解释、或外推到与原来教授时稍有不同的形式上去；

3．应用——使用学过的原理或它的推广结论去解决新问题；

4．分析——通过认识基本要素，它们之间的关系或将原理重新组织、系统化的过程，将问题剖析为那些基本要素，寻找出解决问题的逻辑路经；

5．综合——通过对原来整体结构的认识，将分立的要素进行结合或要求将几个原理相继结合使用地去解答一个问题；

6．评价——使用内部准则（自我产生的）或外部准则，就精确性、逻辑性或学术性方面对某方面问题作出评论。

人们采用测验说明表（又有人叫“双向细目表”）来获得测验的均衡。此表考虑到测量的内容与认知水平两个方面，而构成一张双向权数分布表。

下页上表就是对教师进行资格考试的一张测验说明表。

表中每格中数字是测验编制者对该格的考试内容及认识