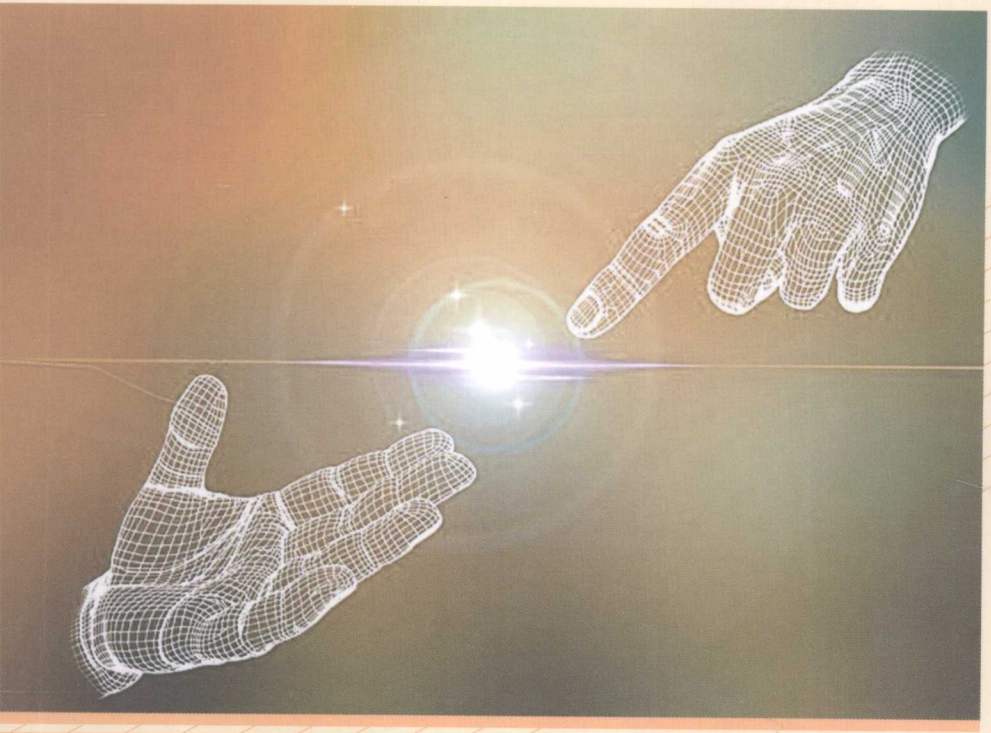
21世纪教育科学系列教材

Psychological and Educational Measurement

心理与教育测量

顾海根 主编



北京大学出版社
PEKING UNIVERSITY PRESS

21世纪教育科学系列教材

心理与教育测量

主　编　顾海根
副主编　沐守宽　刘世宏

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

心理与教育测量/顾海根主编. —北京：北京大学出版社，2008.1
(21世纪教育科学系列教材)
ISBN 978-7-301-12204-4

Ⅰ. 心… Ⅱ. 顾… Ⅲ. ①心理测量学—高等学校—教材②教育测验—高等学校—教材 Ⅳ. B841.7 G449

中国版本图书馆CIP数据核字(2007)第083834号

书　　名：心理与教育测量
著作责任者：顾海根　主编
责任编辑：王　艳　刘　维
标准书号：ISBN 978-7-301-12204-4/G·2089
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn　　电子信箱：zyl@pup.pku.edu.cn
电　　话：邮购部62752015　发行部62750672　编辑部62767346　出版部62754962
印　刷　者：北京飞达印刷有限责任公司
经　销　者：新华书店
730毫米×980毫米　16开本　16印张　280千字
2008年1月第1版　2008年1月第1次印刷
定　　价：28.00元

作者简介

顾海根，男，上海师范大学应用心理学系教授、博士生导师。主要从事心理测量的教学与研究，曾出版《学校心理测量学》、《学校心理辅导系列教材》、《爱国情感教育心理学研究》、《心理差异与教育》、《人员测评》等10部著作，发表论文60多篇。曾获国家级教学成果二等奖、上海市教学成果一等奖、上海市哲学社会科学研究成果著作类二等奖、上海市教育科研成果一等奖等。

前　言

我长期担任我系心理学本科生和研究生基础课《心理与教育测量》的教学工作，但该课程一直没有一本合适的教材。我虽曾于1999年出版过《学校心理测量学》一书，但该书已出版多年，有些内容已陈旧，且该书在理论深度与具体应用方面还嫌不足。于是我想重新写一本适用面更广的教材。2003年我申请到我校研究生教材资助经费，就开始着手《心理与教育测量》的撰写。

心理与教育测量是一门理论性与应用性都极强的学科。在理论方面，心理与教育测量已形成经典测量理论、概化理论和项目反应理论三大理论体系，每一理论都有相应专著出版。在应用方面，心理与教育测量有各种心理与教育量表的使用方法与编制技术、题库建设与计算机自适应测验的编制方法、测验的等值技术、项目功能差异的侦查技术等。在本科阶段，心理与教育测量课程一般偏重于心理与教育测量的基本概念和各种心理与教育量表的使用方法，而对心理与教育测量的理论讨论不多，特别是对现代测量理论一般很少介绍。对各种心理与教育量表的编制方法、测验等值方法以及题库建设方法也很少涉足。鉴于此，本书较系统介绍了心理与教育测量的基本原理和具体的应用方法。全书共分11章，前面7章主要介绍经典测量理论、概化理论和项目反应理论，后面4章主要介绍题库建设、计算机自适应测验、心理测验的编制、测验等值与项目功能差异侦查等具体方法。

在论述测量理论，特别是现代测量理论时，会涉及一些多元统计和高等数学的知识，如项目反应理论中的参数估计、多维模型等，读者和教师可以根据实际情况做些删减，对这部分章节的删减不会影响对其他章节的理解与掌握。

本书是我与我的研究团队共同努力的成果。先由我写出全书的三级提纲，再由下列各位负责撰写初稿：第一章：刘世宏；第二章：潘文；第三章：吴杲；第四章：沐守宽；第五章：毛媚；第六章：顾海根；第七章：王立君；第八章：顾海根；第九章：艾平；第十章：顾海根；第十一章：王金吉。初稿完成后再由我修改

和统稿。沐守宽、刘世宏、赵必华和徐雷参加了部分统稿和校对工作。

本书适合心理学、教育学、管理学等专业的研究生作教材使用，也适合教育硕士、考试院和其他专业人员培训之用。对心理学、教育学等优秀本科生和对心理与教育测量感兴趣的人士而言，本书也可作为自学参考书。

本书在撰写过程中参阅了国内外许多文献，本书的出版也得到我校研究生教材建设经费的资助。同时，北京大学出版社教育出版中心姚成龙副主任对本书的出版给予了大力支持，本书责任编辑王艳、刘维为本书作了仔细的审校，在此一并表示衷心感谢！

由于本人水平有限，书中错误、缺点难免，恳请专家与广大读者指正。

顾海根

2007 年 5 月 24 日

于上海师范大学应用心理学系

目　录

第一章　经典测量理论概述

经典测量理论（classical test theory，简称 CTT）形成于 19 世纪末，经过几十年的探讨与摸索，在 20 世纪 50 年代就形成了一套相当完整的理论体系，对测验的编制提出了一系列具体实用的统计分析方法，这些方法在实际测量工作中产生了巨大影响，至今仍然在广泛使用。

比奈-西蒙的第一个心理测验问世，标志着心理计量学的诞生。自第一本心理测验理论的著作——美国心理学家桑代克（R. L. Thorndike）1904 年发表的《心理与社会测量理论导论》问世后，随后经诸多学者（如：Cronbach，1951；Guilford，1954；Gullikson，1987；Guttman，1944；Lord & Novick，1968；Richardson，1936；Terman，1916；Thurstone，1929；Tucker，1946）不断对他的理论进行扩充和严格化，他们的研究与阐述归纳形成了经典测量理论。

经典测量理论主要是以真实分数模型（true score model）为基础（Gullikson，1987；Lord & Novick，1968），围绕被试对试题的应答结果（观测分数）和被试所具有的真实的心理特质（真分数）之间存在的误差进行分析，发展并形成了包括信度、效度、区分度、等值等概念在内的比较完整的心理与教育测量理论体系，故经典测量理论又称为真分数理论。经典测量理论借用普通物理测量的基本假设，采用的计算公式简单明了、浅显易懂，能够对测试结果做出合理的解释，并且可操作性强，便于在实际测验情境（尤其是小规模资料）中实施，能满足人们对将测试作为一种选拔工具的需要，在测验实际工作中有着较强影响力，并发挥着重要的指导作用。它适用于大多数的心理与教育测量资料，以及社会科学资料的分析，是目前应用最广泛的心理与教育测量理论。

本章着重讨论测量误差与真分数、经典的信度和效度理论以及误差来源等方面的内容。

第一节　测量误差与真分数

一、测量误差

（一）误差理论的发展进程

误差理论的起源最早可以追溯到 18 世纪，二百余年的发展历程可以划分为经

典误差理论的萌芽期、成熟期和现代误差理论的形成发展期。经典误差理论的萌芽期：1794 年，德国数学家高斯(C. F. Gauss)首次提出最小二乘法原理，并于 1809 年在其著作《天体沿圆锥截面围绕太阳运动的理论》中发表。同一时期，法国数学家勒让德(A. M. Legendre)也于 1805 年在其著作《决定彗星轨道的新方法》中应用最小二乘法原理处理观察结果。这为测量数据处理奠定了理论基础。经典误差理论的成熟期：20 世纪前后，苏联科学家切比雪夫（П. Л. Чеοышев)、利柯夫(М. Ф. Маников)等对误差理论进行了系统研究，取得了许多成果，其中最著名的是马利柯夫在 1949 年出版的《计量学基础》(Основы Метрологии)，全面系统地介绍了误差理论，成为误差理论的科学总结。经典误差理论以统计学理论为基础，以静态测量误差为研究对象，以服从正态分布为主的随机误差估计和数据处理理论为特征，用测量误差来表征测量结果的可靠程度。真值的相对性、理论性导致测量误差无法确定。

(二) 误差的概念

误差是由于各种因素造成的测量偏差。是否定义为误差变量取决于研究的需要：某一研究目的成为误差变量的因素，很可能在另一研究目的中成为需要研究的因素。例如，当在研究心情波动的一系列测验中，被试每日兴奋一沮丧的测试得分与我们的研究目的有关，并且是测验内容的一部分。但是，当以测量个性特征稳定性为研究目的时，同样的每日心情波动得分便成为误差变量。

简而言之，与测量目的无关的任何因素都是误差变量。测量误差由与测量目的无关的因素引起的，而且是不准确或不一致的测量结果。可用图 1-1 加以说明。

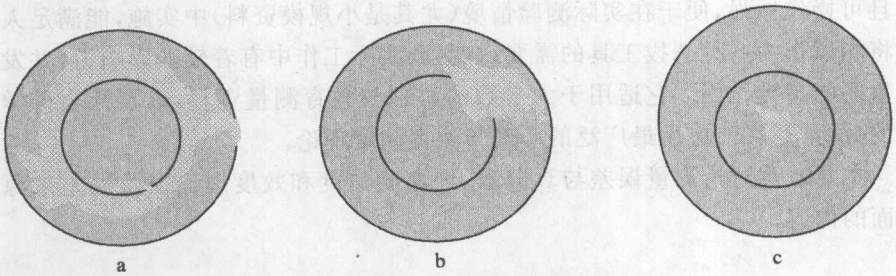


图 1-1　打靶图

从图 1-1 中可看出，图 a 弹着点随机地散落在靶心的四周，既无准确性，又无一致性；图 b 弹着点集中但远离靶心，说明一致性好，准确性差；图 c 弹着点集中于靶心，说明一致性好，准确性也好。此打靶图反映了误差的两种形式：随机误差和系统误差。

（三）误差的种类

一般心理统计资料中，存在三种误差：取样误差、随机误差和系统误差。

1. 取样误差

取样误差又称抽样误差，是由抽样变动引起的误差。取样误差与取样方法及样本容量有关。如果采用完全随机的取样方法，那么平均数的标准误 $SE_{\bar{X}}=S/\sqrt{N}$（这里 S 为样本标准差，作为总体标准差的估计值，N 为样本容量。）我们知道，样本均数与总体均数存在差异，样本之间平均数也存在差异；同时由于抽样误差与样本规模的平方根成反比，在编制一个正式的心理量表时，如果样本容量 N 足够大，算出来的 $SE_{\bar{X}}$ 很小，取样误差可以忽略不计；此外，抽样误差代表样本均数与总体均数的离差，与测量的好坏无必然关系，所以，研究信度或效度时，可以忽略取样误差。

2. 随机误差

随机误差是使用测量工具进行心理测量时，由与测量目的无关的、偶然变异引起的误差，又称观察误差、偶然误差。如图 1-1(a)，弹着点随机地散落在靶心的四周，无规律性可言。在做测试时，被试随便回答，或遗漏一定题目或有测试外因素时，会引起被试临时反应，产生测量误差，使得几次测量结果既不准确又不一致。随机误差是由种种随机原因造成的，这种误差不易控制，它的方向和大小的变化完全是随机的，多次测量可能产生不一致的结果。随机误差是有缺点的测量造成的，是引起测量结果不一致的原因，因而与信度相关；随机误差还影响结果的准确性，因此也与效度相关。随机误差与效度、信度都有关，且无法消除，因此随机误差引起的测量误差是不可忽略的。

3. 系统误差

系统误差是由与测量目的无关的变异引起的一种恒定而有规律的效应，误差集中向着一个方向分布，数量稳定地存在于每一次测量中，如图 1-1(b)所示，几乎所有的弹着点都落在靶心的一侧，射击的偏差具有一致性和系统性。就心理测验而言，一方面，任何心理测验只有被应用于它所适用的目标群体时，才能显示出它的效能。由于人的心理在不同文化、教育和社会环境中有很大差异，如果把一个测验应用于目标群体之外的个人或团体，那么使用其原有的常模来做评定就会出现偏差；此外，编制心理测验时，研究者可能会存在偏见，如对遗传、环境对智力发展的影响存在不同的理解，选题不当，种族歧视等，都会在命题过程中产生误差，这些误差一般都是影响考试结果的恒定因素，就属于系统误差；另一方面，被试对所有问题可能有选“是”或“否”的倾向；所有这些情况对多次测验的影响是一致的，测量值虽然一致，但不准确。系统误差影响了测量值的准确性，与效度有关，但由于测量结果并不引起不一致性，它的效应在信度上不被察

觉。系统误差同样不可忽略。

由于研究信度或效度时，可以忽略取样误差，一些书提及误差常指两类：系统误差和随机误差。在物理测量中，通常把对于随机误差的控制程度叫做精密度，把对系统误差的控制程度叫做正确度，把对系统误差与随机误差的综合控制程度叫做精确度。在心理测量中，把测试中偶然因素所引起的随机误差的控制程度叫做信度，把系统误差和随机误差的综合控制程度叫做效度。测量误差中绝大多数是随机误差，这是由于导致系统误差的原因常为人们特别关注，因而容易避免，而随机误差则是由原因不明的因素造成的，因而难以发现。当然，某些造成随机误差的因素也可由未知转为已知。从这个角度来看，被试在测量工具上所测特性的真实值，并不是直接得到的观测值，我们得到的测量分数实际上是由两部分合成的，一部分是“真分数”，另一部分是由测量误差造成的“误差分数”，这两个分数的和就是测量得出的“总分”。

二、真分数

（一）实测分数、真分数与测量误差

心理测量和物理测量一样，必须做出尽可能准确的估计。但心理测量和物理测量相比，有着显著的不同。首先心理测量是通过行为表现进行的间接测量，受多方面因素的影响，误差不可避免。受测者知道自己是处于被观察的状态中，因而会产生一些不自然的反应，这势必影响测量结果的真实性。把对某一对象实际进行测量时直接获得的值称为实测分数，或实测值或观测分数（X）；而被试在某种程度上所具有的稳定的心理特质被称为真分数，也称真值（T）；实测分数与真分数的差被称为测量误差（E）。实测分数、真分数与测量误差三者的关系可用下列等式表示：

$$X=T+E \tag{1-1}$$

公式(1-1)的含义是将任何一个测验成绩都看做是真分数和测量误差之和，这是经典测量理论的基本思想。对特定被试 a 来说，a 的真正水平即真分数是一个确定值，是常数，用 T_{ga}（g 为某一具体测量）表示，与 T_{ga} 相对应的实测值（观察值）为 X_{ga}，测量误差为 E_{ga}，则 $X_{ga}=T_{ga}+E_{ga}$。

（二）真分数理论的基本假设

真分数理论提出以下基本假设，这是真分数理论的逻辑前提。

假设一：在讨论范围内，真分数具有某种程度的稳定性，即真分数不变，是常数。

心理学的真分数理论借鉴了物理测量的经验，虽然心理测量和物理测量有着显著的不同，在反复测量中被试的心理特质是否真能保持不变是值得怀疑的，

但心理特质可以看做保持相对稳定的，否则无法研究。因此我们假定被试的心理特质在所讨论范围、所经历的时间过程中保持恒定不变。至于真分数具体是什么，人格或是智力或其他心理特质则都可以。

假设二：测量误差的期望值为0，即

$$E=0 \tag{1-2}$$

由于测量误差 E 是指随机误差，是由原因不明的因素引起的，对测量结果来说影响有正有负。一个人的实测分数可能大于真分数，也可能小于真分数，总是围绕真分数上下波动。当重复测量次数足够多，测量误差的正负值就会相互抵消，测量误差的平均数就会为零。也就是说测量误差是服从平均数为零的正态分布。

假设三：测量误差与真分数相互独立，真分数与测量误差的相关为0，即

$$\rho_{TE}=0 \tag{1-3}$$

真分数 T 与测量误差 E 是两个相互独立的变量，它们之间没有相关。如果测量误差与真分数存在相关，那么它也可以部分地反映出一组被试的不同水平，就不成为误差。这是测量误差的随机性所决定的。

假设四：不同测量误差之间的相关为0。

测量误差之间、测量误差与被测特质外其他变量间也彼此相互独立。用公式表达为

$$\rho_{E1E2}=0 \tag{1-4}$$

测量误差是随机出现的，意味着每一次测量所产生的误差是独立的随机变量，两次测量之间没有必然的联系，不存在统计意义上的相关，这也被称为独立性假设。

在这些基本假设的基础上，经典测量理论得以发展(Gulliken,1950；Lord & Novick,1968)，并且可推出用途广泛的真分数理论的另一个等式：实测分数的方差等于真分数方差和测量误差方差之和，即

$$\sigma_X^2=\sigma_T^2+\sigma_E^2 \tag{1-5}$$

推论如下：$X=T+E$，

$$\sigma_X^2=\sigma_T^2+\sigma_E^2+2\rho_{TE}\sigma_E\sigma_T$$

其中 σ_E 和 σ_T 分别是测量误差和真分数的标准差，根据假设三，$\rho_{TE}=0$，因此公式(1-5)成立。

假设五——平行性假设：以相同的程度测量同一心理特质的测验称为平行测验。其内容和形式既可能完全相同(如同一测验重复施测)，也有可能存在某

注：按统计学规定，总体相关系数用 ρ，样本相关系数用 r；总体标准差用 σ，样本标准差用 s。本书将按上述规定使用这些符号。

种差异，但要求能以相同的程度测量同一心理特质。用统计学来表达，就是用两个测验来测量同一心理特质，二者的测量误差的方差相等，就是平行测验。即有两个测验 X 和 X'，若

$$X=T+E,\ X'=T+E',\ 且\ \sigma^2(E)=\sigma^2(E')$$

则测验 X 与 X' 是平行测验。

真分数理论假设测量误差是完全随机的，若每个平行测验的被试数量多，平行测验的个数也多，平行测验又假设每个平行测验能以相同的程度测量同一心理特质，假设平行测验 k 个，被试 n 个，平行测验 1 的测量误差平均数记为 μ_{E1}，平行测验 2 的测量误差平均数记为 μ_{E2}，平行测验 k 的测量误差平均数记为 μ_{Ek}，k 个平行测验在被试 a 上的测量误差平均数记为 μ_{Ea}，k 个平行测验在被试 b 上的测量误差平均数记为 μ_{Eb}，k 个平行测验在被试 n 上的测量误差平均数记为 μ_{En}，可以做出如下推论：

(1) 每个平行测验的平均数应相等，等于同一心理特质，即真分数：

$$\mu_1=\mu_2=\cdots=\mu_n=T \tag{1-6}$$

(2) 各平行测验的测量误差的平均数相等，且等于 0，即

$$\mu_{E_1}=\mu_{E_2}=\cdots=\mu_{E_k}=\mu_{E_a}=\mu_{E_b}=\cdots=\mu_{E_n}=0 \tag{1-7}$$

(3) 各平行测验的测量误差方差相等，各被试在平行测验上的测量误差方差也相等，且两种测量误差方差相等：

$$\sigma^2_{E_1}=\sigma^2_{E_2}=\cdots=\sigma^2_{E_k}=\sigma^2_{E_a}=\sigma^2_{E_b}=\cdots=\sigma^2_{E_n} \tag{1-8}$$

所有平行测验的测量误差的方差及所有被试测量误差的方差都相等，因此测量误差的方差就用一个统一符号 σ_E^2 来代表，这是在独立性假设和平行性假设的前提下，测量误差的一个重要性质。由于测量误差是在一定条件下某一测验的特性，而不是个人的分数，所需要的是对于一组被试的测量误差。

在以上基本假设的前提条件下，经典测量理论从数学模型上引入了测验的信度、效度的概念和定义。信度和效度是一个测验最重要的两个质量指标，代表测验工具的整体特性。

根据上面的假设，可以将真分数定义为：一个被试在某一测量中无限多次测量的均值或数学期望，即

$$T=\lim_{k\to\infty}\frac{\sum_{i=1}^{k}X_i}{k} \tag{1-9}$$

这是因为

$$\sum_{i=1}^{k}\frac{X_i}{k}=\left(\sum_{i=1}^{k}\frac{T_i+E_i}{k}\right)$$

$$\lim_{k\to\infty}\frac{\sum_{i=1}^{k}E_i}{k}\to 0$$

所以

$$\sum_{i=1}^{k}\frac{X_i}{k}=\sum_{i=1}^{k}\frac{T_i}{k}$$

而

$$T_1=T_2=T_3=\cdots=T_k=T$$

因此，

$$T=\lim_{k\to\infty}\frac{\sum_{i=1}^{k}X_i}{k}$$

显然，上述公式采用的是概率模型。这是构成经典真分数理论的重要概念。对实测分数求平均值，当 k 趋于无穷大，实测分数的平均数近似于真分数。也可用一组被试一次或两次测验来代替同一个人反复地施测。当然，对一组被试来说，所测的心理特质的真分数在不同被试间会有不同，但对未筛选的大被试团体，基本心理特质真分数可视为服从正态分布。实测分数是真分数的线性函数。

三、经典测量理论的优点与局限性

在20世纪大部分年代里，心理与教育测量的理论与实践一直建立在真分数模型基础上。真分数模型是以弱假设为基础的，因此这些假设容易验证并易于被绝大多数测验数据资料所满足。经典测量理论在此基础上建立了一整套理论及统计分析方法。经典测量理论浅显易懂，便于在实际测验情境(尤其是小规模资料)实施，是目前测量学界使用与流传最广的理论依据。在测验实际工作中有着强大的影响力，发挥着重要的指导作用。

经典测量理论的真分数线性模型为 $X=T+E$，同时 $X'=T+E'$(X 和 X' 为平行测验)，$\delta^2(E)=\delta^2(E')$，对误差的分析是粗糙的、笼统的，由于理论体系存在先天不足，经典测量理论有以下局限性：

(1) 该模型最突出的弱点就是把所有的测量误差都归为一类，而没能区分测验情景中的各类测量误差，在测量误差 E 中包括了类似评定者、测题、测验环境等影响测量目标的各种因素，也没有说明这些测量误差究竟来自哪些误差源，及各自产生的误差的大小。

(2) 该模型的第二个弱点是样本依赖性。经典测量理论的统计分析方法得到的各项指标，如难度、区分度和信度等，依赖于它们所来自的特定的被试样本。这些指标会因接受测验的被试样本的不同而不同，因此，同一份试卷很难获得一致的难度、区分度或信度。

(3) 在真分数模型中，问题的核心是实测分数 X(以及真分数 T)并不位于等距量表上，因而无法比较两组测验的得分。

(4) 真分数模型已经指出测量误差的存在,以一个相同的测量标准误作为每位被试的测量误差,显然这种作法的适当性受到怀疑。

(5) 在测验结果的精确程度上,是以测验信度和测量误差的方差来表示的,忽略了单个被试在项目上的得分。在此基础上给出的被试的能力分数是很笼统的。比如说,在瑞文标准推理测验中,最后给出的被试的智力分数是一个被试所属年龄组中的百分等级分数,也即是给出了被试在其相同年龄组中的一个相对地位。但是这样的解释并不能真正揭示被试的潜在特质水平以及被试的一些具体认知特征。

(6) 结果应用的局限性。经典测量理论的测验信度是建立在严格平行测验假设基础上的,即两测验是以相同的程度测量同一心理特质。然而,这一理论假设在实际的测验情景中却难以满足,我们常常无法保证不同测量间得分的平均值和方差都相等,也没有一个统一的标准来判断究竟在多大程度上才是"平行测验"。同时经典测量理论指导下的测验还要求测量条件完全标准化,从施测指导语到测验记分都有严格而明确的规定,对于非复本但功能相同的测验所测得的分数间,无法提供有意义的比较,从而使测量目标变得狭小(例如只能对一种非常严格条件下的被试能力进行测量,不够灵活,不便于推广和实际操作),这样就不能对测验进行有效地改进。

(7) 信度估计的不精确性。经典测量理论对信度的假设是建立在平行测验的假设上,但是这种假设往往不存在于实际测验情境里。因为很难找到两个测验的测量误差完全相等。我们不可能要求每位被试接受同一份测验无数次,而每次测量间都彼此独立,因此平行测验的理论假设是很难满足的。在平行测验条件不满足的情况下,估计的各种信度可能有较大误差。

(8) 经典测量理论忽视被试的试题反应组型,认为原始得分相同的被试,其能力必定一样。其实不然,即使原始得分相同的被试,其反应组型亦不见得会完全一致,因此,其能力估计值相应会有所不同。

(9) 能力量表与难度量表的不一致性。在经典测量测验理论中,能力量表与难度量表没有定义在同一个参照系上,这样就找不到验证某个项目是否匹配某种能力水平被试的计量方法,这使得选题带有一定盲目性。被试能力的估计依赖于他完成测验各项目的情况。不管他做错什么题目都影响对他能力的估计,因此,他必须仔细,否则不可能得高分。对被试能力水平的估计精度也无法调节。

上述缺点限制了经典测量理论的进一步应用。经典测量理论致力于估计真分数在观察分数中所占的比例,这种方法不管估计的值多大,都是一种情况下的值,如果测量情境发生变化,真分数所占的比例也必定发生变化。鉴于经典测量理

论存在的不足，测量的理论和实践领域都呼唤一个全新的测量理论。20 世纪 60 年代后，认知心理学的崛起，将实验法与测验法结合，产生了信息加工测验；电子计算机的应用，建立了更为复杂、统计效率更高的模型来处理测量分数。正是在这样的理论背景之下，20 世纪 60 年代在 Cronbach 等学者的研究下（Cronbach，Gleser，& Rajaratnam，1963；Cronbach，Gleser，Nanda，& Rajaratnam，1972），概化理论（Generalizability Theory，简称 GT）应运而生，为测量理论界开拓出一片新天地。项目反应理论（Item Response Theory，简称 IRT）则是从另外一个角度来分析每一个项目的项目特征曲线（Item Characteristic Curve，简称 ICC）和项目信息函数（Item Information Function，简称 IIF）。现代测量理论提出了不同于经典测量理论的框架，使自身获得了超越经典测量理论的许多优良性质。当代测验理论虽然严谨，但理论高深难懂，仅适用于大样本测验资料的分析。所以，不同测验理论各有所长，在应用上也各有其限制。经典测量理论在当前与以后的一段时间内还有用武之地。

第二节 经典的信度理论

一、信度

（一）信度的概念

信度（Reliability）也称可靠性，是指测量结果的一致性程度。如用一质量天平称 50 g 的东西，其结果不会随着时间而变化，后一次称的结果与前一次的一样，就说这天平衡量质量是可靠的或可信的。信度是自然科学研究中的重要概念。信度的重要性可以用下面这个测验来说明：表 1-1 是一个关于不可靠的词汇测验的结果，这是三个学生星期一和星期三的词汇测验成绩（用答对百分率表示）。

表 1-1 不可靠的词汇测验①

学生	星期一	星期三
Judy	49%	94%
Kevin	86%	38%
Scott	52%	38%

表 1-1 中可看出，在星期一和星期三，三名学生的词汇测验成绩发生了不可思议的变化：星期一 Kevin 成绩最好，Scott 又比 Judy 好一点；星期三 Judy 最好，Kevin 和 Scott 成绩一样。显然这个词汇测验不能反映这三个学生学习词汇

① 转引自 Murphy K. R. & Davidshlfer C. O. Psychological Testing，1988，p. 22.