研究生教学用书

教育部研究生工作办公室推荐

现代教育与心理测量学原理

Principles of Modern Educational and Psychological Measurement

漆书青　戴海崎　丁树良　编著

高等教育出版社

研究生教学用书
教育部研究生工作办公室推荐

现代教育与心理测量学原理

Principles of Modern Educational and Psychological Measurement

漆书青　戴海崎　丁树良　编著

高等教育出版社

内容提要

本书是教育部研究生工作办公室推荐的研究生教学用书，是作者在其原作（江西教育出版社 1998 年版）的基础上修订而成的。本书全面系统地介绍了该学科的经典测验理论以及概括力理论、项目反应理论等现代测量理论，内容丰富，阐述深刻，较好地做到了基础性、先进性与实用性的结合。本书吸收了国外的相关研究成果，融合了编著者多年的教学与科研经验，不但具有一定理论水平，也是开展测验分析、建设题库和实际施测的参考工具。

本书适合教育测量与评价专业的研究生以及有一定数学基础的教育工作者使用参考。

图书在版编目(CIP)数据

现代教育与心理测量学原理/漆书青，戴海崎，
丁树良编著．—北京：高等教育出版社，2002.8

ISBN 7-04-010733-3

Ⅰ.现… Ⅱ.①漆…②戴…③丁… Ⅲ.教育心理学：心理测量学-研究生-教材 Ⅳ.G449

中国版本图书馆 CIP 数据核字(2002)第 050606 号

现代教育与心理测量学原理
漆书青　戴海崎　丁树良　编著

出版发行	高等教育出版社	购书热线	010-64054588
社　　址	北京市东城区沙滩后街 55 号	免费咨询	800-810-0598
邮政编码	100009	网　　址	http://www.hep.edu.cn
传　　真	010-64014048		http://www.hep.com.cn
经　　销	新华书店北京发行所		
排　　版	高等教育出版社照排中心		
印　　刷	北京铭成印刷有限公司印刷		
开　　本	787×960　1/16	版　　次	2002 年 8 月第 1 版
印　　张	20.25	印　　次	2002 年 8 月第 1 次印刷
字　　数	370 000	定　　价	28.10 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

序

心理与教育测量是了解和评价个体发展水平的重要手段。测量的实践必须要有理论的指导。心理与教育测量理论的发展经历了两个时期:20 世纪 50 年代之前只有经典测验理论起作用,称为经典测验理论阶段;50 年代至今,除经典测验理论外,还有项目反应理论、概括力理论等,可称为多种理论并存阶段。经典测验理论在测验发展中有着特殊的地位,它既是历史上的第一个测验理论,也是测验的最一般、最基本的理论,并且目前仍具有很强的生命力,应用极为广泛。现代测验理论大多是在经典测验理论的研究基础上,针对它在某个方面存在的问题发展起来的。如项目反应理论,就是为了克服经典测验理论中题目参数等指标的变异性发展起来的;概括力理论是针对经典测验理论的信度问题发展起来的。新的理论形成后,又可能衍生出更新的理论或模式,其中项目反应理论衍生能力最强。从项目反应理论的基本模式出发,已形成了许多新的模式,如多维模式、多变量模式、多等级模式、多成分模式等。测验理论的发展也促进了测验应用的发展,如标准参照测验的编制、题库建设、计算机化自适应测验、测验等值、测验偏差侦查等等。在目前这个多种理论并存阶段,我们应该看到各种理论都有其合理之处,同时也各有其局限性。虽然某一理论在某些方面比另一理论更具优越性,但却难以在各个方面均超越另一理论。所以,众多的理论将会长期并存,并且在相互促进的过程中一起发展、完善。

江西师范大学的漆书青、戴海崎等同志十多年来,努力开展心理与教育测量学方面的研究,取得了重要成绩。他们承接了大量的国家级和省级课题,都取得了良好的效果,在国内有很大影响。尤其是在项目反应理论方面的研究和应用达到领先水平,受到有关部门的重视。这本书是他们多年学习、研究和实践的结晶,内容丰富,科学性实用性都很强,书中既有各个重要理论的介绍和分析,又有作者的实践经验和示例,有利于各个层次的读者从中获益。值本书出版之际,谨作此小序以示对作者的祝贺,并愿他们今后在推动我国心理与教育测量工作的发展中取得更大的成绩。

张厚粲

1998 年 11 月

目　录

第三篇 应用技术原理

绪言　测量过程和心理计量学

测量，就是按照一定规则给研究对象在一定性质的数字系统（尺度）上指定值，目的就在于正确地认识和对待客体对象。客观事物经测量后在数字（即量，从而也反映出质）上就会显出差异，人们把握了事物的个别差异性，就有可能更好地来对待它们。心理和教育测量，也正是要给受测者的心理特性指定值，以便有针对性地正确教育、使用、矫治和发展他们。

心理和教育测量古已有之。我国的科举考试取士，也是一种测量，要把人分出高下，但这还不是科学的现代测量。现代的科学测量，随着大生产与科技的发展，先是在物理量的测量中大规模地发展，后来经过一个时期的心理—物理量的研究，才逐渐在心理与教育测量领域发展起来。当代，心理与教育测量已经科学化、现代化，在教育、经济和社会生活中正显示出重要的作用。

心理—物理量的研究，是一种实验心理学研究。19 世纪中后期韦伯（E. H. Weber）和费希纳（G. H. Fechner）的工作以及冯特（W. Wundt）的心理实验室就对之做出过重大贡献。研究结果说明了一般人各种绝对感觉阈限和差别感觉阈限取值有多大，以及各种感觉领域中心理量与物理量对应的函数关系如何（总体说是非线性的）。这就不但显示了人的心理特性（比如感觉）的确是可测量的，而且把物理测量中严格控制误差的方法，如施测条件与操作手续的标准化等带进到心理测量过程中来了，强调了客观性原则。但是，心理—物理量的研究以发现人类普遍存在的共同规律为主，受测者的个别差异反被当作讨厌的误差而受忽视与被设法消除。

人的心理的个别差异是客观存在的事实。受洛克（J. locke）经验哲学的影响和心理—物理学研究的启发，19 世纪末不少人对简单心理过程个别差异的测量作了认真研究。高尔顿（F. Galton）是一位杰出代表。他认为，“外在世界的任何信息欲传至个人，唯一的途径是经过我们的感官，因此感官的区辨力愈强，我们的判断力与智力所能运作的范围愈大”。所以，感官的区辨能力“整体上说，在智能最高的人的身上也是最灵敏的。”[①]卡特尔（J. M. Kattell）也如此，企图通过对视、听、痛觉的敏度，重量的区辨力，反应时等简单心理过程的测量来推论人的聪

① Anastasi A 著，[台] 黄安邦译.《心理测量》（第六版）第 8 页. 台北：五南图书出版公司，1987

明程度。然而比纳(A.Binet)不同,经过长期研究,他深信高级的复杂的智力功能是可以测量的,尤其强调判断、理解与推理能力的测量。同时,他又认为,在测量复杂功能时,不必要求太高的精确性,因为在这些功能上个别差异比简单功能上要大得多。

比纳在心理测量技术上的划时代贡献,就是提出了试测样本资料项目分析及其基础上的常模这一概念。没有常模这一概念从方法学上作革新与突破,现代科学的心理测量就不可能诞生。在20世纪初年的比西智力量表中,比纳把同一年龄的儿童划分为同一被试组,然后,找出每一年龄组中有80%~90%正常儿童能够通过的试题,用这种难度的试题来代表该年龄的"心智水平"(Mental level)。以后,经过推孟(L.M.Terman)等人的修改,就变成了"心理年龄"以及发展出了"智商"的概念。虽然比纳的"心智水平"更多的是属于发展常模的范畴,而后来韦氏量表和现在斯比量表上的常模都是组内常模,但采用代表性样组的实测资料作通过率意义上的试题难度分析,这样来求取常模的基本做法,却始终未变。编制智力量表求取常模的做法,后来也就推广到人格测验常模的求取上去了,这也一直沿袭至今。第一次世界大战时,在继续开发个别测验的同时,开始团体测验的研究。奥梯斯(A.S.Otis)提出了"多择一"的选择题题型。这就使被试的作答反应高度结构形式化,能有力控制阅卷者的评分误差,为机器阅卷和施测开辟了可能,在题型格式上进行了革新,更加突出了心理和教育测量的"客观性"。如今,在教育测验和人格测验中,都大量使用选择型试题。

工作实践必然推动理论研究的发展。20世纪20年代以来,心理与教育测量理论有迅速的发展,现在已取得巨大的成就,并建立起了心理计量学这一专门的学科分支。测量理论必须剖析与概括测量过程。心理计量学现在一般均认为,心理与教育测量既然是按一定规则给心理特性在一定的数字系统(尺度)上指定值,那么其过程就必然要包含如下阶段和问题:

第一,测什么和如何测?前已述及,不但简单的心理过程要测,高级的复杂的心理过程也要测;不但人的认识方面的特性要测,人的情感与人格方面的特性也要测;不但作为教育活动直接后果的知识与技能成绩要测,作为人的心理的基本素质与基本特点的因素也要测。总之,心理和教育测量要测人的各种心理特性。然而,问题是如何来测。测量工作应该如何进行取决于所测对象本身的性质,这是不言而喻的事。由于心理现象的非实体性、动力性等性质,就使心理与教育测量跟物理测量比,更加复杂,并具有显著的间接性。测量工作所直接面对的,只能是作为心理过程活动与心理特性作用结果的外显行为。许多这种行为,还采取言语表达的方式来显现。这当然也是人类所特有的优点。比如,人所具有的知识、心智技能、社会态度和情感体验等,都可借助语言来表达从而成为外显可测量的东西。这样,心理测量所直接测到的就是人的心理的外显行为,许多

还是语言表达的行为。结果，不但教育成就测验，而且人格和智力测验中，言语方式的测验特别是纸笔测验就成了一种主要形式。但外显行为与内部过程和潜在特性，不能简单等同。相同的认知试题的正确解答，可以通过不同的认识策略来实现。所以，按简单的 S—R 模式只管结果不管内部过程是远远不够的。因此，测量过程的第一步就是要确定测量对象，进行测验设计。这就要对心理现象的实质作深入的探究，并以此为据来设计测验刺激结构和开发试题，建构测验情境关系，编制测验和制作测量工具。

第二，在什么性质的量尺上如何来指定值？对这些值如何作加工处理？有四种不同性质（水平）的测量量尺（量表），即称名性质的、顺序性质的、等距性质的和比率性质的。称名量表上的值只有类别代号的意义，不能进行数学运算。有人把在称名量表上指定值只叫作分类，而不认为是测量。顺序量表上的值有可比性，可求中位数等，但不能加减求平均。等距量表上的值才是有可加性，有相对零点，容许作线性变换，这时谈测量参照点与单位才有实质意义。心理特性一般认为可在顺序量表上取值。然而比西量表通过项目分析确定了各年龄组的代表性试题，答对一题相当于一年的几个月，实际上是想在等距量表上取值。比率量表上的值具有可比、可加、可除性，能作各种数学运算，是测量水平最高的量表。各种心理特性能在什么性质的量尺上取值，决定于心理特性自身的性质，而不决定于人们的愿望，要通过对其作深入的研究来解决。要指出的一点是，不少分布形式对数据的性质提出了特定的要求，比如正态分布就要求数据至少是等距量表上的值。因此，指定一个分布形式等效于采用一个特定性质的量表。经过研究确定了心理特性能在某一测量水平的量表上指定值，在给被试具体赋值时，还要认真研究所测心理特性的具体状况，提出指定数字的规则。很明显，这种规则与前面所要求开发设计的试题、编制的测验、建构的情境关系是分不开的。它们在一起构成了一个统一的反映心理特性实质的规则体系。测量就要按这一系统规则为所测对象指定值。

第三，所得测值可靠吗？测量误差控制得如何？心理与教育测量的对象不同于物理测量的对象，具有动力学性质。人的心理具有主观能动性，可以学习、迁移，并形成与改变态度。所以，心理测量不像物理测量那样，可以无限制地对同一对象反复施测。虽然如此，仍可把所测心理特性看成具有某种稳定性，视为是人的“潜在特质”，内隐于人的“心理结构”。事实上，对人的同一“特质”与“结构”，还是可以在一定程度上作“重测”的。通过反复重测，人们就可发现测值是否一致、可靠，误差是否得到较好控制。为分析和控制误差，有必要采用数学模型。在物理世界，人们可大量采用确定性模型。在心理科学中，人们只好主要采用概率模型。因为，除了在感觉阈限中物理量与心理量的对应关系外，大家几乎再也找不到心理变量间这种确定性函数关系。采用数学理论与方法，特别是采

用统计数学的理论与方法，来分析、处理心理和教育测量问题的学科分支，叫心理计量学。心理计量学在19世纪末开始酝酿发展，其第一个理论体系——真分数测验理论在20世纪50年代臻于成熟。这是一个经典真分数理论体系，它以测验信度理论为核心内容，认为观察分数是真分数与误差分数的和。心理和教育测验后所得测值是否可靠，就看观察分数跟真分数的相关如何。观察分数与真分数相关的平方，就等于信度系数的值。信度系数可通过"重测"或平行形式测验结果的分析等方法来估出。当相关系数趋近于1.00从而其平方根亦即观察分数与真分数的相关也趋近于1.00时，观察分数就可视为真分数的线性变换值，测验所得结果就很可靠了。历史上，人们在开始提出常模这一概念时，并不把观察分数视为不可靠的，而重点在找出一个较科学的办法来解释实测分数的意义。随着实践与认识的深入，人们才日益要求更好地通过控制误差来提高测验质量，因而发展起有关信度的真分数理论。在50年代以后，在随机抽样理论框架内，又发展起了概括力理论；另外，在随机抽样理论框架外，还发展起了项目反应理论。这些都不属经典测验理论范畴，而被称为现代测验理论。现代心理计量学的发展，越来越不再局限于只研究测验信度问题了。

第四，测验测到的是否真的就是本来打算要测的东西？心理和教育测量的对象，应该是某种特定的"潜在心理特质"或"内隐心理结构"。所编测量工具实际施测后所测到的东西，是否真是这特定的"特质"与"结构"，自然是测量工作中最具根本性的问题，在测量过程中既是据以开始又是最后归结所在的问题。由于心理测量对象的间接性，这个问题的考察最为复杂。由此发展起了多种测验效度验证的理论与方法。历史上人们开始时着重于考察所得测量结果跟测验外部效标的相关，利用实证资料分析测验的统计意义上的预测能力。但是，后来又认识到单作统计学分析是很不够的。心理计量学不能只重"计量"而忽视"心理"。心理计量学要和实质心理学有机地统一起来。如果老把"特质"、"结构"当作一个纯"统计学构造"而不认真充实其心理学意义，出现"缺失心理学意义"的现象，是会严重阻碍心理与教育测量的进一步发展的。要深入研究心理现象的本质、结构、机制和功能，以此为指导，更多地采用实验设计的技术来编制测验；还要深入分析被试接受测验的具体心理过程，看到前一部分测验作答过程对后一部分作答的迁移与影响；要把心理学的过程模型跟心理计量学模型结合起来，定性定量地统一进行分析；不能只对作答反应结果资料作探索性因素分析，而且可运用结构方程分析模式作验证性因素分析，主动进行有关心理结构的假设检验。总之，既要大力强化现代科学心理学，又要大力强化现代数学的理论与方法在心理和教育测量中的应用，切实保证提高测量的正确性与有效性，以适应当前社会与科技发展的需要。

美国测量学专家苏恩(Suen)曾把测量过程总结成三个阶段。他写道："实质

上，在测验情境中，我们首先设计量表化规则把一组反应变成数字观察分数。其次，我们推论观察分数充分可信地反映了真分数。最后，我们推论真分数确实真正地反映了(心理)结构"。他认为，这就是心理计量学关于测量过程的推论递进阶段。并作图解如图 0－1：①

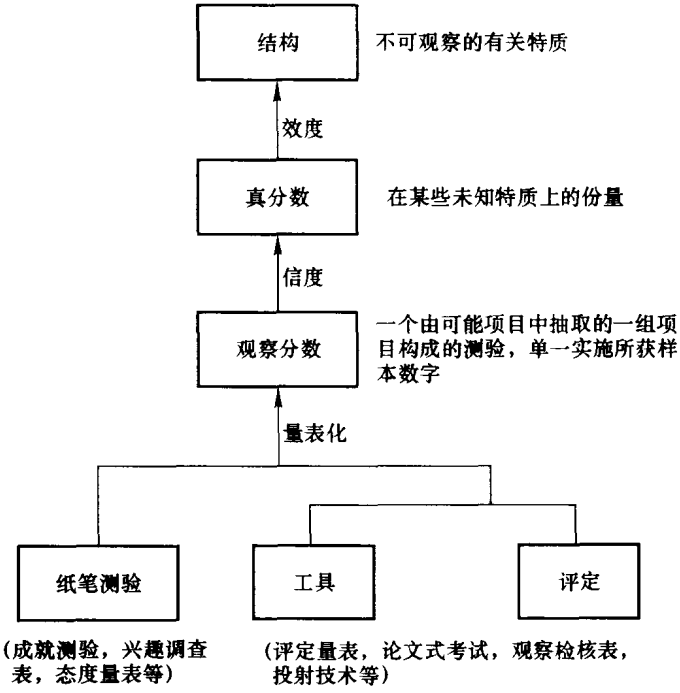

图 0－1　测量过程的阶段划分

我们的看法跟他基本一致，只是将他的按规则指定值的阶段进一步细化成测验设计与量表化两步，然后才是信度和效度的验证。这样，或许才能将测验编制、项目分析、分数合成与解释(含常模的建立)，乃至参数估计等都更好地包括进去。

本书就是打算论述心理与教育测量过程的理论原理，而且力图着重介绍先进的现代心理计量学原理的。按照逻辑和历史顺序，共分成了三篇。先从传统的项目分析与分数常模开始。第一篇是随机抽样理论，包括经典真分数理论与现代的概括力理论。概括力理论不再像真分数理论那样，把测量误差单纯看成混沌一团的东西。它采取统计学中的方差分量模型，具体分析实际的测验情境

① Suen H K. *Principles of Test Theories*, pp. 5－6. Lawrence Erlbaum Associates, Inc, 1990

关系，根据不同的确定测量目标与侧面的做法，针对性地考察多种信度与效度。第二篇是项目反应理论。真分数理论与概括力理论都要求抽取总体的有代表性的随机样本，项目反应理论却不要求这样，样本的结构可以跟总体分布结构不一样，只要在分布全距的各个区间都抽取到必要数量个体就行。所以，项目反应理论，不属随机抽样理论框架范围。它主要是一种量表化模型理论，其突出优点有三：① 参数不变性；② 被试能力与项目难度定义在同一量表上；③ 提出了信息函数概念，从崭新角度来处理信度问题。本篇以单维的适合二值评分资料的模型为范例，论述了项目反应理论的基本原理；同时，还介绍了多种其他类型的模型。任何理论原理都应该用来解决实际问题。所以，第三篇是应用技术原理，探讨了现代测量原理在等值、测量偏差（项目功能差异）、题库、自适应测验等问题上的应用。最后，由效度验证问题转入心理计量学与实质心理学的有机结合这样一个当前测量发展的主导潮流之中，从而结束全书。这样，也许能更好地引起读者深思吧。

第一篇　随机抽样理论

第一章　传统的项目分析与分数常模

作项目分析和建立分数常模，是实现测验标准化的关键步骤。传统的项目分析，就是从应予测验的被试总体中抽取代表性样组，然后在这样的被试样组上施测项目，并根据实测资料对项目的测验性能，如难度和区分度等作统计分析，求出项目的测验性能指标。经过项目分析，就可以筛选出合乎要求的项目来。用这样一批符合要求的项目组成测验（或分测验），再在前述代表性被试样组上施测，就可求取到测验（或分测验）分数的常模，从而得到标准化测验解释分数意义的参照量表。这就是常模参照测验编制的基本程序。历史上第一个标准化常模参照测验是比纳—西蒙智力量表，比纳和西蒙也是最早按上述科学程序进行传统的项目分析与求取测验常模的人。他们的工作，促进了心理计量学的建立和发展。由于这一切都是针对代表性被试样组来进行的，因而其观点与方法属于随机抽样理论范畴。

第一节　项目难度

难度的定义　最简单最通用的难度指数，建立在通过率的基础上。一种办法是直接定义为被试样组上的通过率，或者说是得分率、答对率；一般用符号 p 代表。很明显，这样定义的难度指数取值越大，项目的难度反而越小，指数值跟其含义恰好相反，其实可说是“项目易度”。另一种办法是定义为被试样组上的未通过率，或者说是失分率、答错率；一般用符号 q 代表。这种定义可说直接描述了“困难程度”的本来涵义。然而，由于被试样组上通过率（得分率）p 跟未通过率（失分率）q 满足 $p+q=1$，所以这两种方法在数学上是等价的。又因为未通过率（失分率）q 一般要根据通过率（得分率）p 来求取，在工作中反而显得不

够方便，故难度指数常取第一种办法来定义。

若试题为选择题，或其他以“全或无方式”记分的试题，答对为 1 分，答错为零分，或者说采用 1,0 方式记分，项目的通过率就等于答对人数对被试总人数的比。这时，项目难度指数可按下式求取

$$p=\frac{r}{n}$$

$$q=1-p=1-\frac{r}{n} \tag{1.1}$$

这里，r 为被试样组中的答对人数，n 为被试样组容量。若试题采取全对评给 k 分，有错酌情扣分，全错评给零分的方式记分，或简言之以 k,0 方式记分，项目通过率就等于该项目上的平均得分值跟满分值的比。这时，项目难度指数可按下式求出

$$p=\frac{\sum x}{nk}=\frac{\overline{x}}{k}$$

$$q=1-p \tag{1.2}$$

这里，$\overline{x}$ 为项目上的平均得分值，x 为各被试在该项目上的得分。

下面，是包含了不同记分方式的几个项目难度指数计算的例子。

表 1-1 难度指数计算实例

得分/被试 / 项目		A	B	C	D	E	F	G	H	I	J	满分 (k)	得分率 (p)	失分率 (q)	记分方式
第一题		3	2.5	3	1.5	2	0	1.5	1	2	0.5	3	0.57	0.43	k,0
第二题	①	1	1	0.5	1	0	1	0.5	0	1	0	1	0.60	0.40	k,0
	②	1	0.5	1	1	1	0	0	1	1	0.5	1	0.70	0.30	k,0
	③	1	0	1	1	1	1	0	1	1	1	1	0.80	0.20	1,0
	④	1	1	0	0	1	0	1	0	1	0	1	0.50	0.50	1,0

表达成特质水平量表上的值 建立在通过率基础上的难度指数值，无论是 p 值(成功百分比)还是 q 值(失败百分比)，都不是等单位量度。一般可以假定，当对一个未经筛选、数量极大的被试团体施测时，被试的特质水平取值呈正态分布。测试项目在该团体上的通过率(成功百分比)，可视为正态分布曲线下，横轴(即特质水平量表)上从右端 $+\infty$ 开始，到某一特质水平值点(一般用符号 Z 代表)为止的区间，跟正态曲线所夹面积，对曲线下总面积的比；亦即特质水平强于某一特定值的被试，占该团体总人数的比。如图 1-1 所示。

若有项目 1,2,3,4，其通过率分别为 .50,.45,.10,.05，即 $p_1-p_2=.50-$

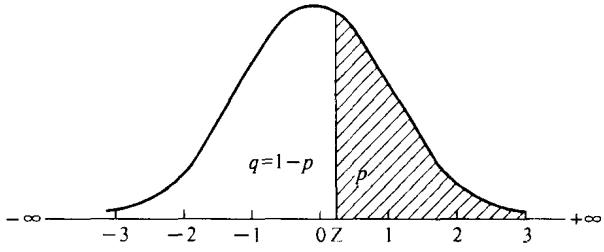

图 1－1　特质水平量表上的难度值

.45＝.05，$p_3 - p_4$ ＝.10－.05＝.05，两组项目通过率差值相等，即正态曲线下对应面积差相等。查正态分布表可知，它们对应的特质量表（即横轴）上的取值分别为 0.0000，0.1257，1.2816，1.6443。于是 $Z_1 - Z_2 = 0.0000 - 0.1257 = -0.1257$，$Z_3 - Z_4 = 1.2816 - 1.6443 = -0.3627$，对应差值显著不等。这可图解如图 1－2：

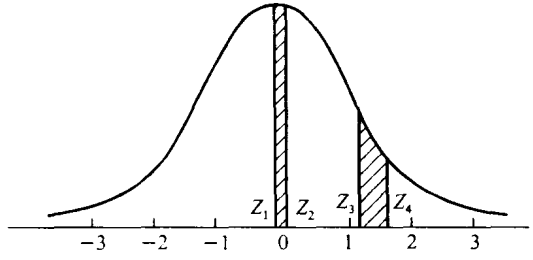

图 1－2　百分比难度变化引起的 Z 值变化图解

图 1－2 说明，项目的百分比难度作了相等变化（上图中阴影部分面积相等），对应的特质能力（水平轴上的取值）变化却并不相同。因此，百分比难度指数不是等单位量度，是顺序量表上的取值。假使要让难度指数成为一种等单位量度，成为间距量表上的取值，就不能直接用项目通过率，而要用这种通过率在正态曲线下所代表的面积在横轴上的对应值，亦即表达成特质水平上的值 Z。然而，正态分布中 Z 值有负值，故宜进一步作如下变换

$$\Delta = 4Z + 13 \tag{1.3}$$

表达成 Δ 值的难度指数，取值一般在 1 和 25 之间，并且是等单位量度，具有可加性，故便于作进一步的理论分析计算。表 1－1 中各题的难度指数，转换成 Δ 值可列成下表：

表 1-2 按表 1 数据求 Δ 值

项 目		p	Z	Δ
第一题		0.57	-0.1764	12.2945
第二题	①	0.60	-0.2533	11.9868
	②	0.70	-0.5244	10.9024
	③	0.80	-0.8416	9.6335
	④	0.50	0.0000	13.0000

恰当难度与恰当难度分布 一个测验常由许多项目组成,它们的难度不能都一样大。那么,什么是项目的恰当难度,以及整个测验的恰当难度分布呢?这个问题相当复杂,概括地说,它取决于所测特质的性质,测验的目的,项目的格式类型,以及项目间的相关性。

就标准参照性测验来说,人们常常力求全部达到标准,希望有百分之百的通过率。若果能如此,不但每个项目都有高成功百分比难度指数,而且各项目间难度指数差异也不大,难度分布围绕一高值点形成窄全距分布。

但是,许多心理和教育测验,其目的都是要把考生加以区分,甚或是典型的常模参照测验。正因为目的在于区分被试,突出个别差异的认定,所以就希望最后被试的测验总分能最大限度地"拉开距离",并使测验在分数分布的每个点上都有强的鉴别力。这时,各项目难度以多大为宜,测验项目难度分布以宽一点为好还是窄一点为好呢?

对自由反应型试题来说(猜测成功影响很小,可忽略不计),就上述目的而言,最适项目难度指数值应为 0.50。可以设想,假定 100 名被试接受测验,对错被试各为 50 名,这 100 名考生就可划分成彼此不同的 2500 个对子($50 \times 50 = 2500$)。若难度指数为 0.60,这种对子数就是 2400(即 $60 \times 40 = 2400$)。若为 0.70,就是 2,100,……若极难(0.00)和极易(1.00),这 100 名被试分数就全都一样(0 分和满分),测验就不能把任何两个被试区分开来。所以,中等难度是自由反应型项目的恰当难度;项目过难与过易都无助于把被试区分开来。

以上仅考虑单个项目,测验中所有项目是否难度都应为 0.50 呢?若测验完全同质,内部各项目间相关近于 1.00,或者被试组非常同质,这时难度分布应宽些为好,可在从 $\Delta = 3.7(p = 0.99)$ 到 $\Delta = 22.3(p = 0.01)$ 的区间内作均等分布。若测验完全异质,项目间相关近于 0.00,则项目难度都应为 $\Delta = 13(p = 0.50)$。事实上,测验的同质程度不会如此极端。当一个测验项目与总分间的平均二列相关为 0.60 时,就将视为非常同质。这时,难度可有一个由 $\Delta = 9(p = 0.84)$ 到 $\Delta = 17(p = 0.16)$ 的宽全距分布。当一个测验项目与总分间的平均二列相关为 0.40 时,就可视为相当异质。这时,需要有一个由 $\Delta = 12(p = 0.60)$ 到 $\Delta = 14(p$