

## 第一部分

---

# 心理测验的功用和起源



# 第一章 心理测验的性质和用途

心理测验是工具。为了获得测验能够提供的益处，我们必须记住这个重要的事实：任何工具都可能是有益也可能是有害的，这取决于怎样使用这种工具。测验的发展越来越快，并且测验正在越来越多的日常生活领域中起着重要作用。<sup>[1]</sup>但是这种发展也伴随着一些不现实的期望和误用。测验使用者需要了解怎样评价测验。正在使用的测验适用于特定目的的程度如何？关于参加测验的个体，测验能够提供什么类型的信息？怎样能够使测验结果并入作出决策的资料网？正是出于对这类问题的考虑，我们编著了本书。本书不是为测验专家而是为一般的心理学学生编著的。今天，不仅编制或实施测验的人员，而且在作出关于自己或他人的决策时将测验结果用作一种资料来源的任何人士，都需要了解一些测验的基本知识。

## 心理测验的用途和种类

传统上，心理测验的功用是测量个体之间或同一个体在不同情况下的反应之间的差异。促进心理测验发展的最早课题之一，便是鉴定智力落后个体。至今，检测智力落后仍然是某些类型的心理测验的一种重要应用。测验的临床使用则是检查具有

严重的情绪障碍和其他类型的行为问题的个体。同样，教育上的评定需要也大大推动了早期测验的发展。第一个智力测验即著名的比纳(Binet)测验正是如此。现在，学校是最大的测验用户。教育测验的多种用途包括：按能力将学生分类，使他们从不同类型的学校教学中得益；鉴定明显落后或明显超常的学生；中学生和大学生的教育和职业咨询；专业学校选拔学生。

工业人员的选拔和分类是心理测验的另一种主要应用。从装配线操作人员或文书到上层管理人员，几乎对于每一类工作，某种心理测验都有助于其招聘、工作分配、调动、提升、解聘等。诚然，在许多这类情境中，尤其关于高水平的工作，有效使用测验通常要求我们将测验用作为熟练面谈的附属工具，这样可以依据个体的其他背景资料适当解释测验分数。不过，测验仍然构成人事总计划的一个重要部分，与此密切相关的一种心理测验应用是军事人员的选拔和分类。从第一次世界大战开始到第二次世界大战期间，军事情境中使用心理测验的范围和种类迅速增加。后来，关于测验编制的研究，在军队的各个兵种中继续大规模地进行。

个体咨询中测验的使用，从教育和职业计划的狭义指导，逐渐扩大到包括个体生活的各个方面。情绪健康和有效的人际关系日益成为咨询的主要目的，使用测验来增进自我了解和个体发展也日益受到重视。在这种背景下，测验分数是给予个体的资料的一个组成部分，帮助个体自己作出决策。

显然，心理测验目前正被用来解决一系列实践问题。然而，我们不应该忽视心理测验在基础研究中也起着重要作用。例如，几乎在差异心理学的各种问题中，都需要将测验方法作为收集资料的一种手段。作为说明，可以提到的是关于个体差异的性质和程度、心理特质的结构、团体差异的测量、与行为差异有



关的生物学因素和文化因素的鉴定等研究。对于所有这类研究领域，并且对于许多其他研究领域，一种重要的先决条件就是使用编制良好的测验去准确地测量个体差异。同样，对个体一生的发展变化、不同教育方法的相对效果、心理治疗的效果、社区规划的影响、环境变量对行为的影响等种种问题进行研究，心理测验都提供了标准化的工具。

出于这些不同的目的而设计的多种类型的测验，在其他一些主要特征上也有所不同。包括实施测验的方法有所不同，例如，一个被试都由一位经过训练的主试进行测量的个别测验，大型团体的同时测验，或由计算机实施的测验。测验所包括的行为方面也有所不同。一些测验主要评定认知特质或能力。它们的范围从上大学的准备等广泛的能力倾向，到完成某种简单的手工操作所需要的高度具体的感觉运动技能。另一些测验测量情感变量或人格，例如情绪或动机特质、人际行为、兴趣、态度、价值观等。

面对多种多样的性质和目的，心理测验共同的区别性特征到底是什么？怎样将心理测验区别于收集资料的其他方法？回答在于测验编制和测验使用的某些基本特征之中。本章讨论的正是这些特征。

## 心理测验是什么？

行为样本 (behavior sample) 心理测验实质上是对行为样本的客观的和标准化的测量。就我们对小型但经过仔细选择的个体行为样本进行观察这一点而言，心理测验像任何其他科学中的检验一样。在这方面，心理学家使用的方法与生物化学家

通过分析一个或几个样本而检验病人血液或社区用水是相同的。如果心理学家希望测验儿童掌握词汇的程度、办事员的计算能力或飞行员的手眼协调能力,则他们就可以使用一组有代表性的词汇、算术问题或运动测验来检查被试的表现。显然,测验是否适当包括所考虑的行为,取决于样本中项目的数量和性质。假如算术测验只有 5 道题,或只包括乘法题,那就不能很好地测量个体的计算能力。同样,假如词汇测验全为棒球术语,也就不能可靠地估计儿童的总词汇量。

心理测验的诊断或预测价值,取决于它预测相对广泛和重要的行为领域的程度。对测验所直接包括的行为样本的测量,如果有过的话,那也很少是心理测验的目的。就儿童对特定的 50 个单词表的了解本身,对我们来说没有什么意义。同样,工作申请人在一组特定的 20 道算术题上的成绩,也没有多大的意义。不过,如果能够证实儿童对单词表的了解和他掌握的总词汇量之间,或者申请人在算术题上的分数和他在工作上的计算表现之间密切对应,那么测验就算符合它的目的了。

在这一点上应当指出,测验项目不必与测验所预测的行为非常相似,而只需要证实两者之间的经验性对应。测验样本与所预测的行为之间的相似性程度可以大不相同。在一个极端,测验可以与所预测的行为的一部分完全相合。一个例子是外语词汇测验,它检查学生已经学过的 50 个生词中的 20 个。另一个例子是驾驶员获得驾驶执照之前所参加的路考。工作培训之前所实施的许多职业能力倾向测验,是相似性程度较小的例子,工作上所完成任务与测验中所编入的任务之间只有适度的相似性。在另一个极端,我们看到的是罗夏 (Rorschach) 墨迹测验之类的投射人格测验。人们试图根据被试对墨迹的联想去预测他或她对他人、对情绪刺激、对日常生活中其他复杂情境会如何反

应。尽管表面上有差异,所有这些测验都由个体的行为样本所构成。每个测验必须在经验上证实被试在测验中与其他情境中的表现相对应而证明其价值。

在这一点上使用“诊断”一词或“预测”一词也表示微小区别。预测通常意味着时间估计,例如根据被试现在的测验表现去预测他们未来的工作表现。然而,在广义上,甚至对现在情况的诊断,例如智力落后或情绪障碍等,也意味着预测个体在不同于现在测验的情境中将做什么。把所有测验都看作是行为样本,根据行为样本能够预测有关的其他行为,这在逻辑上比较简单。于是,不同类型的测验都能够被表示为这种基本型式的变式。

一开始就应该考虑的另一点是能力 (capacity) 的概念。例如,完全可能设计一个测验,在个体开始学习法语之前去预测个体法语能够学得多么好。这样一个测验包括学习新语言所需要的各类行为样本,但本身不含有法语知识。这样就可以说,这个测验测量个体学习法语的“能力”或“潜力”(potentiality)。不过,在心理测验中使用这类术语应该谨慎。只有在现在的行为样本能够被用来预测将来其他行为的意义上,我们才能够说一个测验可以测量“能力”。心理测验的基本任务是测量行为。只有通过经验性试验,才能证实这类行为是否能够作为其他行为的有效指标。

标准化 (standardization) 回忆一下,在最初的定义中,心理测验被描述为标准化的测量。标准化是指测验的実施和评分中程序的一致性 (uniformity of procedure)。如果要使不同被试所获得的分数可加比较,显然,测验的条件对所有被试都必须是相同的。这一要求只是所有科学观察中需要控制条件的一种特殊应用而已。在测验的情境下,单一的自变量通常是正在接受测

验的个体。

为了确保测验条件的一致性,测验编制者为实施每一个新编的测验而提供详细的说明。编写说明是新测验标准化的一个重要部分。这种标准化包括所使用的确切材料、时限、口头指导语、例题、处理测验参加者的询问方法,以及测验情境的其他每一个细节。其他许多更细微的因素都会影响测验的成绩。因此,口头念出指导语或提出问题时,必须考虑到讲话的速度、音调、停顿、面部表情等。例如在一个检验荒谬的测验中,假如主试读到关键词时,微笑或停顿,也许就会泄露正确答案。本章后面一节论述测验实施的问题,将从主试的观点进一步讨论标准化的测验程序。

测验标准化的另一个重要步骤是建立常模 (norm)。心理测验没有预先确定的及格或不及格的标准;每个测验的成绩都是根据经验性资料加以评价。对于大多数目的来说,解释个体的测验分数时,是将它与其他人在同一测验上所获得的分数相比较。顾名思义,常模就是正常的或平均的成绩。因此,如果正常的 8 岁儿童在某个算术推理测验中答对 50 题中的 12 题,那么这个测验中 8 岁常模就为 12 分。这种分数称之为测验的原始分数。它可以是答对的项目数、完成一项任务所需要的时间、错误数目,或适合于测验内容的其他一些客观度量。只有按照合适的解释资料加以评价,这种原始分数才有意义。

在测验标准化的过程中,将测验实施于为之所设计的那类个体的一个大型的、有代表性的样本。这个样本称之为标准化样本,用于建立常模。这种常模不仅指出平均成绩,而且指出平均数上下不同程度离差的相对频数,因而能够评价不同程度的优劣。第三章将考虑表示这种常模的具体方法。所有这些使我们可以表示个体相对于常模或标准化样本的位置。

也应当指出，建立人格测验常模的方法，本质上与能力倾向测验的一样。能力倾向测验的常模未必是满分即没有错误的分数，人格测验的常模也未必是最合乎要求的即“理想”的成绩。在这两类测验中，常模均与典型或一般个体的行为表现相对应。例如，在支配—顺从测验中，常模处于一般个体所表现的支配或顺从程度的中间点。同样，在情绪调节问卷中，常模也会引发不受欢迎的即适应不良的反应。标准化样本的大部分“正常”个体也发生一些这类反应，适应不良反应的这个数目因此就代表常模。

难度的客观测量（objective measurement of difficulty）本次讨论开始所提出的心理测验的定义就表明，心理测验的特性是客观的和标准化的测量。心理测验到底在哪些方面是客观的呢？讨论标准化时已经提到心理测验的客观性的某些方面因此，如果测验的实施、评分、分数解释等与特定主试的主观判断无关，则它们都是客观的。理论上，不管主试是谁，任何一个测验参加者应该获得相同的分数。当然，情况也不完全如此，因为实际上尚不能达到完满的标准化和客观性。但是，客观性至少是测验编制的目标，并且在大多数测验中已经达到相当高的程度。

在其他一些主要方面，也能够适当描述心理测验的客观性。项目或整个测验的难度水平的确定要根据客观的、经验性的程序。比纳和西蒙（Simon）最初编制 1905 年智力量表时，将表的 30 个项目由易到难加以排列。项目难度是通过将 50 个正常儿童和一些智力落后儿童试测这些项目来确定的。有些项目答对的人数最多，根据这一事实，则作为最容易的项目；有些项目通过的人数相对较少，则作为较难的项目。根据这种程序，可以建立难度的经验性顺序。这个早期例子的特征是难度水平的客

观测量,这一点现在是心理测验编制的共同做法。

不仅测验项目的安排,而且测验项目的选择,都能够以测试样本中通过每个项目的人数比例为指导。因此,如果量表的易端或难端项目过多,则要删除一些项目。同样,如果难度范围的某些部分项目稀少,则要增加补充新项目。第七章将讨论项目分析的技术问题。

信度(reliability)这个测验好不好?这个测验真的有用吗?这些问题可能,有时也确实导致长时间的无效讨论。主观的意见、感觉和个人偏向有时使人们对某个测验能够达到什么目的提出过分的要求,有时又使人们顽固地拒绝测验。最终回答此类问题的惟一方法是通过经验性尝试。心理测验的客观评价主要是确定测验在规定情境中的信度和效度。

心理测量学中所使用的信度一词主要表示一致性。测验信度是使用相同的测验或测验的等值型式再次测验相同被试时,他们所得分数的一致性。如果一名儿童星期一获得 IQ 为 110,而星期五再测时 IQ 为 80,显然,两个分数都没有或没有多少可靠性。同样,一名被试在一组 50 个单词中认对 40 个,而在另一组假设为等值的单词中只答对 20 个,那么两个分数都不能作为其词汇理解的可靠指标。诚然,在两个例子里,两个分数中可能只有一个是错误的,但这只能通过进一步的再测来确定。根据已知数据,我们只能得出结论,两个分数都是不正确的。没有另外的资料,就不能确定是两个都不能还是有一个能适当估计个体的词汇能力。

发行心理测验以供全面使用之前,应该进行详尽的、客观的测验信度的检验。第四章将考虑不同类型的测验信度及其测量方法。比较相同的测验参加者在不同的时间、使用不同组的项目、不同的主试或评分者,或在任何其他有关的测验条件下所获

得的分数，都可以检验信度。必须说明信度的类型和确定信度的方法，因为同一个测验在这些不同的方面可以有所不同。同样应该报告检验信度所使用的被试的性质和人数。有了这类资料，测验使用者才能够预测对于他们期望测验的团体，测验是否同样可靠，或测验的可靠性可能增大还是减小

效度( validity) 毫无疑问，心理测验所要求回答的最重要的问题当是测验效度，即测验实际上测量到其意欲测量的东西的程度。效度直接检查测验实现其功用的程度。不管设计测验去测量什么东西，效度的确定通常需要独立的、外部的效标( criteria)。例如，如果使用一个医学能力倾向测验来选拔有希望的考生进入医学院，那么这些学生在医学院中最终成功就是一种效标。在建立这类测验效度的过程中，将它实施于医学院大型的学生团体，根据学业成绩、教师评定、完成训练的成功或失败等，最终获得每个学生在医学院表现的某种度量。这种综合度量便是一种效标，计算每一个学生最初的测验分数与它的相关。较高的相关或者说效度系数( validity coefficient)表示那些测验得高分的个体在医学院中相对成功，而那些测验得低分的个体在医学院中学得不好。较低的相关表示测验分数和效标度量之间没有多少对应，因此测验的效度较差。效度系数使我们能够确定根据测验分数预测效标成绩的程度。

同样,为其他目的所设计的测验,也能够以适当的效标来建立效度。例如，职业能力倾向测验，能够以新员工试用的工作成绩来建立效度。飞行员能力倾向成套测验，能够以飞行训练的成绩来建立效度。为一般的、多种不同用途所设计的测验，以一些独立获得的行为指标来建立效度。只有根据许多不同类型研究中逐渐积累的资料，才能建立这些测验的效度。

读者也许已经注意到测验效度概念中有一个明显的自相矛

盾之处。如果必须追踪测验参加者或以其他方法获得测验试图预测的东西的独立度量,那么为什么还需要测验呢?要解开这个谜,就要区分建立效度的团体和出于使用目的最终对之实施测验的各个团体。在测验准备使用之前,必须根据一个有代表性的被试样本建立测验效度。这些被试的分数本身不用于使用目的,而仅仅在检验测验的过程中起作用。如果使用这种方法证明测验为有效的,那么在缺乏效标度量的情况下,能够将测验使用于其他样本。

也许有人仍会提出,为了获得测验试图预测的资料,我们只需要等待任何一个群体的效标度量的形成,并加以利用。但是这种方法耗费大量的时间和精力,在大多数情况下都是不允许的。因此,我们要确定哪些应聘者会工作成功或哪些学生会顺利读完大学,就要招收所有被试(或其中一个随机样本),并且等待他们后来的发展。设计测验正是要把这种过程的极大耗费及其对个体有害的情绪影响减到最小限度。通过测验,我们能够在可测定的误差范围之内评定个体必备的技能、知识和其他有关特征的目前水平。测验的效度和信度越高,这种误差范围就越小。

第五章和第六章将讨论确定不同类型测验的效度时所碰到的特殊问题,以及所使用的具体效标和统计方法。不过,现在应该考虑一个进一步的问题。效度告诉我们的不止是测验实现其功用的程度。它实际上告诉我们测验测量什么东西。研究建立效度的资料,我们能够客观地确定测验测量什么东西。因此,按照我们对测验测量什么东西的了解程度去定义效度就更为准确。假如按照经验上所确定的各种关系有系统地命名测验,而建立测验效度正是通过这些关系,那么测验分数的解释无疑会更为清楚和明确。用‘学业评定测验’和‘人员分类测验’之类测



验名称来取代模糊的名称“智力测验”，我们就能够看到这种趋势。

## 为什么要控制心理测验的使用

“我可以要一份斯坦福一比纳测验空白卷吗？我的侄子想进某学校，下个星期参加斯坦福一比纳测验，我很想让他练习一下，他就能通过了。”

“为了改进学校的阅读课程，我们需要一个排除文化影响的智力测验来测量每个儿童的潜力。”

“昨晚我回答一本杂志上所发表的智力测验问题，IQ 得了 80 分，我认为心理测验是无聊的。”

“我的室友正在学习心理学。她让我做了一个人格测验，结果是我有神经症。自此，我心烦意乱，不能上学。”

“去年你出于研究目的让我们员工做了一个新编人格测验，我们想把分数归入他们的个人档案。”

这些说法都是假设的，但每种说法根据一件真实事情，心理学家很容易再列举下去这类说法说明，心理测验的潜在的误用或误解，会使测验没有价值，或使个体受到伤害。像任何科学仪器或精密工具一样，心理测验必须加以适当使用才会奏效。在马马虎虎或本意虽好但缺乏有关知识的使用者手中，心理测验会造成严重危害。控制心理测验使用的主要理由有两条：（1）保证由合格的主试来实施测验，并且适当使用测验分数；（2）防止大家都熟悉测验内容，这样测验就没有效度。

合格的主试 测验的选择、实施和评分，以及分数的解释是测验情境中三个主要的方面，而其中每一方面都显然需要合格

的主试。测验不可能像割草机那样,根据邮购目录加以选择。测验也不可能根据名称、作者或其他容易鉴别的标记加以评价。诚然,考虑运输测验材料的费用、体积和便利性,以及测验所需要的时间、评分的容易和快速之类因素,都不需要心理学训练。关于这些实际问题的资料,通常能够根据测验目录获得,并且在制订测验计划时应该加以考虑。然而,要使测验发挥功用,根据效度、信度、难度水平和常模之类特征来评价测验的技术优劣是至关重要的。只有如此,测验使用者才能确定测验对于特定目的的适合性,及其对于所计划实施测验的被试类型的适合性。

本章前面测验标准化的讨论,已经提到经过训练的主试的重要性。如果要使不同主试所获得的测验分数可作比较,或者要根据所发表的常模来评价被试的分数,那么就需要准确遵循指导语和彻底熟悉标准指导语,还必须严格控制测验条件。同样,不正确或不准确的评分会使测验分数没有价值。缺乏适当的检验方法,可能发生的评分误差就比一般所了解的大得多。

测验分数的适当解释,需要彻底理解测验、测验参加者和测验条件。只有参照建立特定测验的效度所使用的具体方法,才能客观地确定测验正在测量什么东西。关于信度、常模样本的性质等其他资料,同样也是重要的。关于被试的一些背景资料,在解释测验分数时也是必不可少的。不同的被试,由于截然不同的原由,也有可能获得相同的分数。根据这类分数所得出的结论,因此就大不相同。最后,还必须考虑可能已经影响某测验分数的一些特殊因素,例如,不同寻常的测验条件、测验参加者暂时的情绪状态或身体状态、测验参加者先前的测验经验等。

测验使用者的作用 80年代和90年代心理测验的一个重要发展,便是人们日益认识到测验使用者的关键作用(Anastasi, 1990b)。在这方面,任何人士,只要把测验分数用作实际决策的

一种资料来源,他就是测验使用者。测验使用者可以是也可以不是实施和评分测验的主试。测验使用者包括学校系统的教师、咨询人员和管理人员,以及工业或政府部门的人事管理人员等。大多数对测验的批评,不是针对测验的内在特征,而是针对不合格的使用者误用测验结果。一些误用的原因是人们对于现实生活问题,希望有解决的捷径、快速的回答和简单的常规解法工作负担过重产生的时间压力,会怂恿人们去依靠这类权宜之计。不过,测验误用的最常见原因,也许是测验知识的不足或欠缺(Eyde, Moreland, Robertson, Primoff & Most, 1988; Moreland, Eyde, Robertson, Primoff & Most, 1995; Tyler & Miller, 1986)。

全国专业机构的各特殊委员会与测验出版者联合工作,日益关注并防止测验误用。一个著名的例子是测验使用者资格工作小组(Test User Qualification Working Group, TUQWoG)所进行的项目(Eyde 等, 1988)。TUQWoG 的主要目标是制订一套经验性的、基于数据的不同类型测验使用者的必备资格,测验出版者将它编入测验购买者资格表。通过 5 年深入的全国性研究, TUQWoG 项目建立了给人深刻印象的数据库。一些出版者已经开始将这种结果用于购买者资格表。后来成立第二个工作小组,其目的是使用 TUQWoG 的数据库去制订测验使用者的准则和训练材料,即被称之为经修改的首字母缩略词 TUTWoG(T 表示训练)。它的第一个成果是一本防止常见测验误用的个案资料(Eyde 等 1993)。这些个案都根据种种环境下所实际观察到的具体的误用例子。它们是在项目的大范围调查中所提出的,后来由莫兰(Moreland)等作出总结(1995)。

测验内容的可靠与测验信息的交流 显而易见,假如某人记住色盲测验的正确反应,则这种测验就不再测验该被试的色视觉。在这种情况下,测验完全没有效度。为了防止故意伪造

分数，对测验内容显然必须加以限制。然而，在其他情况下，熟悉测验内容对测验分数的影响也许不是那么明显，或者说测验完全由于被试得到不宜的信息而失去效度。例如，一位学校教师给班级做一次十分类似于智力测验问题的特殊练习，“所以孩子们就会作好参加测验的准备”。这种态度只是通常准备学校考试的程序的继续。不过，这类具体的训练或辅导用到智力测验上，就可能提高测验分数而不影响测验试图取样的广泛的行为领域。在这种情况下，作为一种预测或诊断的工具，测验的效度就减小了。

保证具体的测验内容的可靠，不必也不应该妨碍与测验参加者、有关的专业人员和一般大众进行有效的测验信息交流。这种交流符合若干目的。第一，它有助于消除与测验有关的任何神秘感，因而有助于纠正关于测验用来测量什么东西和测验分数表示什么意思的普遍误解。一些主要的测验出版者所发送的书面出版物就是出于这种目的。第二，是关于编制和评价特定测验的技术程序，它们提供测验的信度、效度和其他心理测量特性的有关资料。每个测验的技术手册中一般都包括这类信息，任何感兴趣的人士都可得到。

测验信息交流的第三个目的是使测验参加者熟悉测验程序、消除焦虑、保证每个被试表现出最佳能力。出于这种目的，准备了若干解释性小册子，有些是通用性的，有些适合于大学入学考试委员会的学业评定测验(Scholastic Assessment Test)之类具体测验。本章后面一节将进一步讨论这些资料。十分重要的第四类交流是提供给测验参加者的关于他们自己所参加的测验成绩的反馈。心理学家相当注意在不同情况下传达这种信息的最有用的和最有意义的方法。第十七章和第十八章将讨论适当的方法。

测验信息的传播是至关重要的，其中所采用的方法良莠不齐。在联邦和州两级水平都采用政府控制，这种轻率的立法试图就是有害方法的一例（）

年代后期，加利福尼亚和纽约实际上制定了管理泄露测验信息的州法律，其中纽约州法律更为极端，它要求不加限制地泄露所有大规模高等教育机构招生测验的测题和答案。

因为这种泄露要求使得每次实施必须编制一种测验的新型式，所以它具有若干不利的影响。其中包括一年中可用于测验日期的减少、考生费用的增加，以及测验编制和使不同时间所实施的测验分数均等中所能够使用的质量控制程序的下降。值得注意的还有，没有多少测验参加者利用测验泄露的机会，并且这种泄露并没有显著提高另一型式上的再测成绩（Stricker, 1984）。促动人们提出测验泄露法所要达到的目标，能够通过加强测验信息交流的现有程序而更为有效地达到，并且没有不利的副作用。

## 测验实施

测验的基本原理是从测验情境所观察到的行为样本而概化到在其他非测验情境所表现出的行为。测验分数应该帮助我们预测患者在诊所之外的感觉和行为如何、考生在大学课程上的成绩如何、应聘者在工作上的表现如何。测验情境特有的任何影响均构成误差方差，并且降低测验效度。因此重要的是要鉴别可能限制或降低测验结果概化力的任何与测验有关的影响。

我们很可能要用整本书来专门讨论合乎需要的测验实施的程序。但是这种全面评述超出本书的范围。此外，在具体背景

之内获得这种技术更为切实可行，因为我们通常不会关心从婴儿测验到精神病患者的临床测验等一切形式的测验或军事人员的团体测验的实施。因此，现在的讨论主要涉及测验实施的一般原理而不是实施的具体问题。在萨特勒(Sattler, 1988, 第五章)关于儿童个别评定的综合讨论中,有一个测验实施的优秀例子。

**主试的预先准备** 良好测验程序的最为重要的要求是预先准备。测验中有时会有突发事件，因此必须作出特别努力，预见和防止突发事件。只有这样，才能保证测验程序的一致性。

预先准备可采取许多方式。在大多数个别测验中，必须牢记确切的指导语。甚至在对测验参加者宣读指导语的团体测验中，预先熟悉指导语可以防止读错和犹豫，而且使主试在测验期间举止更为自然、轻松。测验材料的准备是另一个重要的预备步骤。在个别测验尤其在操作测验的实施中，这种准备包括必需材料的实际安排，这便于随后的使用，以免东寻西找。材料一般应该放在靠近测验桌的地方，以便主试容易拿取材料而不会使测验参加者分心。如果使用复杂的仪器，必须经常进行定期的检查和校准。在团体测验中，所有的测验空白卷、答卷、特殊用笔，或其他所需要的材料，应该在测验日之前仔细点数、检查和安排。

彻底熟悉具体的测验程序是个别测验和团体测验中另一个重要的前提。对于个别测验，特定测验的实施的培训通常是必不可少的。这种培训可能需要几次演示和练习或一年以上的教学，这取决于测验的性质和所测被试的类型。对于团体测验，尤其是大规模测验，准备工作还包括事先对主试和监试作任务分工，以便他们完全了解各自的职责。一般来说，主试宣读指导语、掌握时间和主持每个测验场所的工作 监试分发和收集测验材料、保证测验参加者遵循指导语、在测验手册规定的范围内回

答他们的各个问题、防止作弊。

**测验条件** 标准化程序不仅适用于指导语、时间、材料和测验本身的其他方面，而且也适用于测验环境。应该注意选择合适的测验房间。该房间应该没有过度的噪音，以免被试分心，还应该提供适当的光线、通风、座位和测验参加者的工作空间，并且采取特殊的步骤，以防止测验中断。在门上贴出表示测验正在进行的记号是行之有效的办法，当然，其前提是所有人员都知道这种记号表示在任何情况下都不得入内。大型团体测验时，有必要锁上门或在门外布置助手，以便阻止迟到者进入。

重要的是要认识到测验条件可能影响分数的程度。甚至在测验情境十分次要的方面，也可能明显改变成绩。例如，在中学生团体测验中，使用书桌或使用带有一小桌面的椅子是重要的，使用书桌的团体往往获得较高分数（T. L. Kelley, 1943; Traxler & Hilbert, 1942）。也有证据表明，所使用的答卷的类型可以影响测验分数（F. O. Bell, Hoff & Hoyt, 1964）。由于设立独立的测验评分和数据处理机构，而该类机构提供他们自己的机器评分答卷，所以主试有时实施团体测验使用的答卷不同于标准化样本。如果缺乏经验性证据，就不能假设这些答卷等价。测验五年级以下的儿童时，使用任何分开的答卷，都会显著降低测验分数（Cashen & Ramseyer, 1969; Ramseyer & Cashen, 1971）。对处于这些年级水平的儿童，一般来说，最好让他们就在测验小册子中标出答案。

在任何年龄水平，相同测验的纸笔实施与计算机实施之间可能的差异更为显著。人们相当注意测验实施中这种差异对常模、信度、效度的影响，这些都涉及测验参加者群体和测验的性质。已经提出特殊的专业准则，帮助测验使用者评定在两类实施中所得出的测验分数的可比性（Butcher, 1987; Hoger & Green,

1985)。

已经表明许多其他更为次要的测验条件影响能力测验和人格测验的成绩。主试是陌生人还是测验参加者所熟悉的人员，这可以产生测验分数的显著差异 (Sacks, 1952; Tsudzuki, Hata & Kuze, 1957)。另一个研究表明，主试的一般习惯和行为，例如微笑、点头、作出‘好’或‘不错’之类评论等，对测验结果具有决定性的影响 (Wickes, 1956)。在要求被试编写符合给出图片的故事的投射测验中，房间内主试的存在往往抑制故事中包括强烈情绪的内容 (Bernstein, 1956)。在实施打字测验时，工作应聘人单独测验的打字速度显著快于两个或两个以上的团体测验 (Kirchner, 1966)。

很容易举出更多的例子。这有三重含意。第一，一丝不苟地遵循标准化程序。在测验手册中全面而清楚地叙述这种程序是测验编制者和出版者的责任。第二，记录任何不同寻常的测验条件，不管其多么不重要。第三，解释测验结果时考虑测验条件。通过个别测验全面评定某个被试时，有经验的主试出于特殊原因偶尔也会偏离标准化测验程序，以便得到另外的信息。在这种情况下，不能再根据测验常模来解释测验结果。这时，测验刺激仅仅用于定性探索，应该以处理其他日常行为观察或面谈资料的同一方式来处理被试反应。

**引入测验：和睦关系和测验参加者的倾向性** 在测验实施中，“和睦关系”指主试努力激起测验参加者对测验的兴趣、得到他们的合作、鼓励他们以适合测验目的的方式作答。能力测验的目的是要求对给出的任务全神贯注，尽最大努力取得好成绩。自陈人格问卷的目的是要求对于自己通常行为的问题，作出诚实坦率的回答。某些投射测验的目的是要求详尽报告由刺激所引起的联想，对内容不作删改或编辑。另一些测验可能有其他



的要求。但是在各种条件下，主试都要努力促使被试尽量全面而认真地遵循指导语

主试培训的内容包括建立和睦关系的技术以及直接有关测验实施的技术。像在其他测验程序中一样，建立和睦关系时，条件的一致性对于结果的可比性是必不可少的。如果每当儿童答对一个测验问题时，就给她所渴望的奖品，则她的成绩不能直接与常模或仅仅以标准的言语鼓励或表扬所激励的其他儿童的成绩相比较。解释成绩时，应该注意和考虑偏离某测验标准激励条件的情况。

在个别测验中较容易更为全面地建立和睦关系，在团体测验中也应该采取步骤去激励测验参加者，消除他们的焦虑。建立和睦关系的具体技术，随测验的性质与所测被试的年龄和其他特性而有所不同。测验学龄前儿童时，所考虑的特殊因素包括对陌生人的害羞、分心、不合作等。主试友好、愉快、轻松的态度有助于儿童消除疑虑。害羞胆小的儿童需要更多的准备时间来熟悉环境。由于这种原因，主试最好不要太着急，可以等到儿童愿意进行最初的交往时再正式开始。测验的时间应该简短；测验的任务应该丰富多彩，引起儿童兴趣。测验应该如同游戏般呈现给儿童，实施每种新任务之前应该激起儿童的好奇心。在这个年龄水平，由于可能的拒绝、不感兴趣和其他消极的表现，测验程序的某种灵活性是必不可少的。

像学龄前儿童一样，小学一至三年级的儿童也有许多同样的测验问题。游戏仍然是激起测验兴趣的最为有效的方式。通过激发竞争精神和取得好成绩的愿望，通常能够激励年龄稍大的学龄儿童。然而，如果测验教育背景差的儿童或不同文化的儿童，主试不能假设激励他们更好完成学业任务的程度与标准化样本的儿童相同。第九章、第十二章和第十八章将进一步研

究这个问题和关于测验不同背景被试的其他问题。

测验有情绪障碍的个体、囚犯或少年罪犯,可能遇到特殊的动机问题。尤其在公共机构中测验时,这类个体可能表现出一些不合作的态度,例如猜疑、不安全感、担忧、冷淡等。以往经验中的特殊情况,也可能对他们的测验成绩产生不利影响。例如,由于学校早期经历的失败和挫折,他们对学业怀有敌意和自卑感,而测验又类似学业任务。在这种情况下,有经验的主试作出特殊努力去建立和睦关系。无论如何,主试必须对这些特殊困难十分敏感,并且在解释测验成绩时加以考虑。

主试应该记住,无论测验学龄儿童或成人,每个测验都隐含着对被试自尊心的威胁。因此在开始时应该使被试消除疑虑。例如,最好解释一下,没有一个人能够做完或全部答对。否则,当被试做到较难的项目或在规定的时间内不能完成某个分测验时,他或她可能会感到自己很失败。

最好尽量消除被试对测验情境的惊奇感,因为出乎意料和不了解都可能产生焦虑。许多团体测验包括一个解释性开场白,由主试向团体宣读。更好的办法是预先给每个测验参加者提供材料。这种材料解释测验的性质和目的,提出怎样参加测验的一般建议,还包括一些例题。在许多大规模测验例如大学董事会举行的测验中,参加者通常可得到这种解释性小册子。

成人测验还会遇到另外的一些问题。与学生不同,成人不可能仅仅因为给他分配任务而去努力完成。因此向成人“推销”测验目的更为重要,虽然中学生和大学生也会对这种要求作出反应。使测验参加者相信获得有效的分数是为了他们自己的利益,这通常能够保证他们的合作。有效的分数正确表明他们能够做什么,而不是过高或过低估计他们的能力。大多数被试会明白,无效的测验分数可能造成不正确的决策,这意味着他们后

来的失败、浪费时间和挫折。这种方法不仅能够激励测验参加者在能力测验上尽力而为，而且在人格问卷上减少作假、鼓励坦率报告，因为被试认识到，要不然最终承受损失的是他们自己。录取被试学习一门课程但他们缺乏必要的知识和技能，或者给他们分配一项工作但他们不能完成或感到志趣不合，这决不是为了他们的最大利益。

## 主试和情境变量

人们定期发表关于主试和情境变量对测验分数的影响的综合评述(Lutey & Copeland, 1982; Masling, 1960; S. B. Sasason, 1954; Sattler, 1970, 1988; Sattler & Theye, 1967)。虽然已经证实这类变量对客观的团体测验有一些影响，但所获得的大多数资料或来自投射技术，或来自个别智力测验。比起界定明确的、认识充分的机能来，没有结构的、模糊的刺激，以及难度大和新颖的任务，更可能受到这些外部因素的影响。一般来说，儿童比成人易受主试和情境变量的影响。测验学龄前儿童时，主试的作用尤为关键。任何年龄的有情绪障碍和不安全感的个体也比顺应良好的个体易受这种条件的影响。

有人研究个别实施智力测验和投射技术的成绩与许多主试变量的关系，其中包括年龄、性别、种族、职业地位或社会经济地位、训练和经验、人格特征、外表等。虽然发现若干重要的关系，但研究结果往往是非结论性的或令人误解的，因为实验设计没有控制或分离不同的主试和被试的特征的影响，因此可能混淆两个或多个变量的影响。

在测验即将实施和实施期间，主试的行为会影响测验结果，

这一点得到较为明确的证实。例如,主试和被试之间的人际关系是‘热情’还是‘冷淡’,或主试的举止是刻板还是自然,加以控制的研究得出智力测验成绩有显著差异(Exner, 1966; Masling, 1959)的结论。此外,作为主试自己人格特征的一种机能,相同的主试特征或测验方式对不同的被试可能产生不同的影响,在这种意义上,主试和被试的特征之间存在着显著的相互作用。测验性质、测验目的、给予测验参加者的指导语等任务变量也会发生类似的相互作用。戴尔(Dyer, 1973)还提出另一些变量,他注意到测验实施者和测验参加者对测验的功用和目标的各种不同理解可能产生的影响。

主试自己的期望也可能无意中影响测验参加者的反应。这只是自我实现的预言的一个特例(Harris & Rosenthal, 1985; R. Rosenthal, 1966; R. Rosenthal & Rosnow, 1969)。使用罗夏测验所进行的一项实验证实这种影响(Masling, 1965)。主试是 14 名自愿的研究生,告诉其中的 7 人,有经验的主试引发的人类反应比动物多,而告诉另外 7 人,有经验的主试引发的动物反应比人类多。在这种条件下,两组主试从其被试中获得显著不同的动物对人类的反应比例。尽管主试或测验参加者都报告没有意识到任何影响,这种差异仍然发生。此外,所有测验时间的磁带录音也表明,没有主试方面言语影响的证据。主试的期望显然是通过测验参加者对主试微妙的姿势暗示和面部暗示作出反应而起作用。

除了主试之外,测验情境的其他方面也可能显著影响测验成绩。例如,军队新兵往往在入伍不久对陌生的、有压力的情境进行紧张的再顺应期间参加测验。在一项旨在检验适应这种情境对测验成绩影响的研究中,2724 名新兵在海军培训中心的第九天参加“海军分类成套测验”(Navy Classification Battery, L. V.

Gordon & Alf, 1960), 另外 2180 名新兵在常规时间即第三天参加测验, 把两组新兵所获得的分数相比较, 第九天组在成套测验的各分测验上的分数显著较高。

测验参加者临测验之前的活动也影响他们的成绩, 尤其当这类活动产生情绪干扰、疲劳或其他不利条件时。在一项对三、四年级学校儿童的研究中, 有证据表明, 儿童在“画人测验”(Draw-a-Man Test)上 IQ 受到他们先前的课堂活动的影响 (McCarthy, 1944) 在第一次测验时, 学生刚写完作文“我所遇到过的最高兴的一件事”。在第二次测验时, 他们也刚写完作文, 但这次是“我所遇到过的最讨厌的一件事”。第二次测验即令人沮丧的情绪体验之后的 IQ, 平均比第一次低 4 ~ 5 分。这些研究结果在后来一项研究中得到证实, 该研究特地设计来确定临测验之前的体验对“画人测验”的影响 (Reichenberg-Hackett, 1953)。在这项研究中, 儿童成功解决一道有趣难题之后奖赏玩具和糖果, 具有这种愉快体验的儿童测验分数的提高大于经历中性体验或没有什么愉快体验的儿童。戴维斯 (W. E. Davis, 1969a, 1969b) 对大学生的研究得出类似的结果。在言语理解测验上有过失败体验之后, 算术推理测验的成绩显著低于没有参加先前测验的控制组和在通常条件下参加标准言语理解测验的控制组。

若干研究关注测验分数反馈对个体随后测验成绩的影响。在一项对七年级学生的设计良好的某研究中, 布里奇曼 (Bridge-man, 1974) 发现, 获得‘成功’反馈的学生在随后类似测验上的成绩显著高于获得‘失败’反馈的学生——实际上他们开始做得一样好。这类动机反馈主要通过参加者在随后作业中为他们自己设立的目标而起作用, 因此是自我实现的预言的另一个例子。然而, 这种一般动机反馈不应该混淆于矫正反馈, 通过矫正反馈

告诉被试做错的具体项目，并且给予补救性辅导。在这种条件下，反馈更可能提高最初得低分被试的成绩。

本节所引用的例子证实各种各样与测验有关的变量可能影响测验分数。在大多数认真实施的测验中，出于实际目的，这类变量的影响可以忽略不计。然而，有经验的主试仍经常注意去发现这类变量的可能的作用，并且把它们的影响降到最低。如果环境不允许控制这些条件，则应该对从测验分数中所得出的结论加以限定。

## 从测验参加者角度看问题

**测验焦虑** 关于被试对测验情境的反应，最早研究之一是测验焦虑。毫无疑问，测验焦虑的显著性及其对成绩的明显不利的影响，激起了对这类反应的早期兴趣。在测验实施中增进和睦关系的许多做法，也是为了减少测验焦虑。使测验参加者消除对测验情境的惊奇和陌生感，使他们安心并且加以鼓励的做法，必然有助于降低焦虑。主试自己举止得当，和组织良好、实施顺利的测验，也有助于同一个目标。

有人研究小学生和大学生的测验焦虑的个体差异（Gaudry & Spielberger, 1974; Hagtvet & Johnsen, 1992; I. G. Sarason; Spielberger, 1972）。萨拉森（S. B. Sarason）及其同事在耶鲁大学进行过许多这类研究（Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960）。其第一步是编制问卷来评定被试参加测验的态度。例如，小学生问卷包括如下项目：

你在参加测验之前十分忧虑吗？

当老师说她想知道你学会了多少，你的心跳开始加快吗？

当你参加测验时，你经常想到做得不好吗？

学业成就和智力测验分数都与测验焦虑产生显著的负相关，这一研究结果最为重要。对大学生的研究也得出同样的相关(I. G. Sarason, 1961)。纵向研究同样表明焦虑水平变化与智力或成就测验成绩变化之间的相反关系(K. T. Hill & S. B. Sarason, 1966; S. B. Sarason, K. T. Hill & Zimbardo, 1964)。

当然，这种研究结果并不表示孰因孰果。很有可能，学生产生测验焦虑是因为他们在测验上做得很差，因而在先前的测验情境中经历失败和挫折。在智力测验高分者组内，焦虑水平和测验成绩之间的负相关不复存在，这一研究结果证实上述解释(Denny, 1966; Feldhusen & Klausmeier, 1962)。另一方面，也有证据表明，焦虑对测验成绩的不利影响至少造成某种关系。在一项研究中(Waite, Sarason, Lighthall & Davidson, 1958)，让智力测验分数相同的高焦虑儿童和低焦虑儿童再做一次学习作业，结果发现，虽然两组被试最初的学习测验得分相同，但低焦虑组的提高显著高于高焦虑组。

若干研究人员比较了在引起‘焦虑’状态和‘放松’状态条件下的测验成绩。例如，曼德勒和萨拉森(Mandler & Sarason, 1952)发现，告诉测验参加者我们期望每个人在规定的时间内完成测验之类的专注自我的指导语，对低焦虑被试的成绩具有有利影响，而对高焦虑被试的成绩则具有不利影响。其他研究同样发现测验条件与焦虑水平和成就动机这类个体特征之间的相互作用(Lawrence, 1962; Paul & Eriksen, 1964)。看来焦虑和测验成绩之间的关系可能是非线性的，即轻度的焦虑有利，而高度的焦虑则有害。引起适度焦虑的测验条件有利于通常低焦虑的个体，而在比较轻松的条件，通常高焦虑的个体则会做得较好。

毫无疑问，长期高度焦虑水平对学校学习和智力发展产生

不利影响。焦虑干扰信息的获得和提取 (Hagtvet & Johnsen, 1992)。然而,这种影响应该区别于本次讨论所关注的限于测验的影响。在什么程度上,测验焦虑使得个体的测验成绩不能代表他或她非测验情境下的惯常成绩?有人提出,因为准备升大学的高中生所经历的竞争压力,测验焦虑可能过度影响大学招生测验的成绩。在对这个问题的一项控制良好的深入研究中,弗伦奇(French, 1962)比较了高中生在定期实施的“学业能力倾向测验”的一个测验上的成绩与在另一次“放松”条件下所实施的一个平行测验上的成绩。后面一次测验的指导语说明,只是出于研究目的而实施测验,并且不会将分数送往任何大学。结果表明,标准条件下实施的成绩不比放松条件下实施的成绩差。此外,在这两种条件下,根据高中课程成绩确定的测验分数的同时效度没有显著差异。若干近期研究也对测验焦虑学生知道有关内容但参加测验时“愣住”的通常旧印象表示怀疑(参见 Culler & Holahan, 1980)。这项研究发现,在测验焦虑量表上得高分的学生,年级平均成绩低于测验焦虑量表上得低分的学生,学习习惯也往往较差。

关于测验焦虑的性质、测量和处理的研究,以越来越快的速度继续发展(I. G. Sarason, 1980; Spielberger, Anton & Bedell, 1976; Spielberger, Gonzalez & Fletcher, 1979; Spielberger, Gonzalez, Taylor, Algaze & Anton, 1978; G. S. Trton, 1980)。关于测验焦虑的性质,已经鉴别出两个重要的组份即情绪性和忧虑。情绪性组份包括紧张、心跳加快之类感觉和生理反应。忧虑即认知组份包括做不好的预料、担心失败的结果之类消极的自我定向想法。这些想法分散了对测验所要求的任务定向行为的注意力,因而使成绩降低。两个组份都可以通过若干测验焦虑问卷加以测量。虽然这类问卷在研究中得到广泛使用,但直到最近为止,只有通过



研究文献中的报告才可看到。斯皮尔伯格 (Spielberger) 及其同事所编制的“测验焦虑问卷”(Test Anxiety Inventory) 是已出版的测验的一例; 第十三章将叙述这个问卷。

人们作出相当努力来发展和评价处理测验焦虑的方法。其中包括若干行为治疗方法, 用于降低测验焦虑的情绪性组份(第十七章)。一般来说, 结果是积极的, 但因为评价研究方法的缺陷, 难以把提高归因于某种特定的方法 (G. S. Tryon, 1980)。事实上, 甚至在有治疗干预的控制组中, 以及在给予可信的假治疗的特殊控制组中, 从初测到再测, 测验焦虑的情绪性组份也往往减少。此外, 情绪性组份的减少对成绩水平没有或没有多少影响。

如果处理是针对自我定向的认知反应, 则较可能提高测验和课程作业的成绩。迄今为止的研究提出, 使用联合处理的方法, 能够获得最佳结果。这种方法能消除情绪性和忧虑, 并且改进学习技能。测验焦虑是一种复杂的现象, 有多种原因, 并且不同原因的相对作用随个体而有所不同。为了行之有效, 应该使处理方法适合个体的需要。也必须认识到, 测验焦虑只是降低个体学习效果的一般条件中的一个条件而已。

关于测验参加者对测验看法的综合研究 虽然测验焦虑是测验参加者行为的一个令人注意的重要方面, 但能够进行有益研究的方面还有许多。1993 年, 巴鲁克·内瓦和杰格 (Baruch Nervo & R. S. Jager) 编辑了一本书, 书中广泛收集了在教育、工业、临床、咨询等环境下被试对测验反应的现有资料。该书有 15 章, 作者均是在测验应用及其他方面公认的研究人员, 书中报告每一论题现有的各国出版物内容以及作者自己的研究成果。它试图根据资料来认真回答此前在报章杂志和政治、法律等文献中主要探讨的问题。这本书因此矫正现今关于测验的可能带有

偏见的、相冲突的各种意见。例如,第一章引用 10 个实施良好、范围广泛的人群样本对测验态度的民意调查,结果揭示了公众的观点与广泛听到的但没有代表性的代言人对测验问题的某些主张之间的差异。

各章的论题多种多样。有几章讨论设计和使用反馈问卷和团体面谈来评定参加测验的不同团体对某一测验的态度和对该测验测量什么东西的理解。有一章比较学生对论文式课堂测验和多项选择式课堂测验的意见,结果是他们强烈偏爱多项选择形式。一些作者论述工作应聘者对测验的公正性及其与工作相关性的反应。作为作者自己的研究结果,有几章提出改进测验实施和测验环境的方法。整体来说,各章都展现了一个有前途的探索领域,试图对目前测验中的一些社会和实践问题寻找解答,还提出了增强测验使用者和测验参加者之间相互理解的方法。

## 训练对测验成绩的影响

评价训练或练习对测验分数的影响时,一个根本问题是,提高是仅局限于测验所包括的具体项目,还是要扩大到测验所评价的广泛的行为领域(Aanstasi, 1981a, 1981b)。对这个问题的回答表示辅导和教育之间的差异。测验取样于有关行为的各个方面,显然,个体的测验成绩,应该反映他的正规的或非正规的、学校之内或之外的教育经历。这种广泛的影响决不会使测验没有效度,因为测验分数准确地描述个体在所考察的能力上的位置。当然,这是一种程度的差异。我们不能将影响分类为或狭窄的或广泛的,不过其范围确是大不相同,从只对一次测验的影响,

到对某种类型的所有项目成绩的影响，再到对个体在绝大多数活动中成绩的影响。然而，从有效测验的观点来看，作出一种切实可行的区分是可能的。因此可以说，只有当特定经验提高测验分数但没有明显影响测验所要测量的行为领域时，测验分数才没有效度可言。

**辅导** 辅导对测验分数的影响得到广泛研究。英国心理学家进行了若干早期研究，由于以前曾使用测验把 11 岁儿童分配到不同类型的中学，因此他们着重研究练习和辅导对这种测验的影响(Yates 等, 1953 ~ 1954)。可以预料，提高的程度取决于测验参加者的能力和以前的教育经验、测验的性质，以及所提供的辅导的数量和类型。比起受教育较多，并且对做好测验已经有所准备的个体来，受教育较少的个体更可能从特殊辅导中受益。同样显而易见，测验内容和辅导材料之间越是类似，测验分数的提高也越大。另一方面，教学越是局限于具体的测验内容，提高就越是不可能扩大到效标成绩。此外，由于严重的方法缺陷，许多辅导研究得出模棱两可的、不可解释的结果 (Anastasi, 1981a; Bond, 1989; Messick, 1980a)。这些缺陷中，主要是没有使用一个确实可与经过辅导的样组相比较的、未经辅导的控制组。例如，参加商业性辅导课程的学生均为志愿者，他们最初的能力、动机以及影响测验成绩的其他个体特征，往往不同于控制组。另外，在使用前测和后测的实验设计中，难以保证测验参加者做好两次测验的动机相同。如果一次测验以常规方式实施，另一次测验仅仅出于练习或研究的目的以特殊方式实施，情况尤为如此。

大学入学考试委员会已经对缺乏准备的商业性高考辅导课程的蔓延予以关注。为了澄清这个问题，该委员会进行了若干控制良好的实验，以便确定这类辅导对“学业能力倾向测验”(SAT)的影响，并且调查其他独立的研究者所进行的类似研究

的结果( Donlon, 1984; Messick, 1980a, 1981; Messick & Jungeblut, 1981)。这些研究包括多种多样的辅导方法,同时针对公立和私立中学的学生,也研究城市和乡村的少数民族学生的样本。所有这些研究的结论是,对类似于 SAT 项目的强化练习,比之学生在一年常规的中学教学之后再测 SAT,不可能产生显著的提高。

也应该指出,在测验编制方法上,大学董事会和研究生资格考试委员会之类机构研究新的项目类型是否易受辅导的影响( Evans & Pike, 1973 ; Powers, 1983 ; Powers & Swinton, 1984; Swinton & Powers, 1985)。如果通过短期针对性的练习或教学能够明显提高某些项目类型的成绩,则在施行的测验形式中便不再保留这种项目类型。一个明显的例子是有一类问题要求简单的顿悟性解答,即一旦得出答案,就能够直接用于所有类似的问题。将来再遇到这类问题时,它们便成了测验回忆而不是测验解决问题的技能的问题。另一个例子是某些复杂的项目类型,它们使用新颖的或陌生的材料,并且需要长时间的、复杂的教学( Powers, 1986)。

在狭义的传统意义上,进行辅导是为了发展在生活活动中可能没有多少用处的、适用范围高度有限的技能。同样,“教会测验”的练习往往针对测验所包括的知识和技能的特定样本,而不是测验所要评定的广泛的知识领域。所谓的“测验真实”法,或者说公开法,要求在一次实施之后完全公开测验型式,鼓励人们把重点放在适用范围有限的具体的测验技能上。最后,一些测验参加者经过辅导,其他测验参加者则没有,这往往在严格界定的参加测验的技能上产生个体差异,由此降低测验的诊断价值。

**测验的技巧** 测验的技巧即纯粹的参加测验的练习,对测

验成绩也有影响。使用同一测验的复本进行研究，第二次的分数往往较高。一些研究者报告，即时或 1 天至 3 年之后实施复本测验，平均数都显著提高（Donlon, 1984; Droege, 1966; Peel, 1951, 1952）。对于智力正常的或超常的小学生、中学生和大学生以及员工样本，都得出同样的结果。测验手册中应该提供使用平行形式再测时所预期的分数提高的分布资料，解释测验分数时也应该考虑到这种提高。

分数提高并不局限于复本。比起第一次参加测验的个体来，具有丰富的参加标准化测验经验的个体，在测验成绩上具有某种优势（Millman, Bishop & Ebel, 1965; Rodger, 1936）。这种优势，一部分是由于克服了最初的陌生感，以及形成了更多的自信和更好地参加测验的态度；一部分是由于许多测验所包括的内容和机能的类型中有一定数量的重叠。熟悉通常的项目类型，练习使用客观型答卷，也可能使成绩获得稍许提高。比较参加测验的经验大不相同的被试所获得的分数时，考虑测验技巧尤为重要。对于计算机实施的测验，应该注意测验参加者对这种测验实施形式的熟悉程度（Hofer & Green, 1985）。

要使被试的测验技巧相等，短期的适应和练习是十分有效的（Anastasi, 1981a; Wahlstrom & Boersman, 1968）。这种熟悉性训练减少参加测验经验的先前差异的影响。由于这些个体差异是测验情境所特有的，减少差异应该可以更为有效地评定测验所要测量的广泛的行为领域。例如，大学董事会发行一本书名为《参加 SAT I 推理测验》（*Taking the SAT I: Reasoning Test*）的小册子，将其分发给报名参加该测验的所有考生。小册子提出有效的参加测验行为的建议，说明和解释测验所包括的不同类型的项目，而且复制一份完整的测验，建议学生在标准时限条件下完成，并根据给出的答案评分。一本类似的测验小册子《参加 SAT

《学科测验》( *Taking the SAT II: Subject Test* ),则说明和解释不同学科测验的项目。

研究生资格考试( GRE )也提供熟悉测验的材料。分发给所有考生的《信息公告》( *Information Bulletin* ),包括对“一般测验”( *General Test* )的例题的解释和一套完整的以前实施的测验及其评分答案。以小册子形式,即《参加 GRE 一般测验的练习》( *Practicing to Take the GRE General Test* ),定期发行该测验的另外的测验型式,也发行提供若干学科领域的各种 GRE 测验的类似的练习小册子。

80年代和90年代出现的熟悉测验材料热,并不局限于印刷媒体,而且包括投影、幻灯、电影、盒式录像带和计算机软件。教育测验服务中心( *Educational Testing Service* )制作和发行其中许多材料。一些材料适用于具体测验,例如,《参加 SAT》( *Taking the SAT* )配有一部幻灯片。另一些材料则解释 SAT 分数和大学入学考试委员会成就测验的分数,还可得到帮助理解 SAT 分数的计算机程序。有人为准备参加 GRE 一般测验的学生而编制相对精致的准备测验的软件包。这个软件包通过相互作用的计算机程序,提供例题,一种模拟的规定时间的测验情境,解释答错的项目,分析测验参加者的长处和不足。

其他材料(印刷品、电影、多媒体包、计算机软件)适用于较一般的参加测验的目的,其范围从小学生到成人,如适合中学生个别或团体使用的一盘录像《依靠自己:准备参加标准化测验》( *On Your Own: Preparing for a Standardized Test*, 1987 )。小册子《如何参加测验 尽力而为》( *How to Take a Test: Doing Your Best* )是一本简明的综合指南( *Dobbin*, 1984 )。一些主要的商业性测验出版者和政府机构也制作适应测验的辅助材料。后者如适用于美国职业服务中心( *Employment Service* )《一般能力倾向成套

测验》(General Aptitude Test Battery, GATB)的一套辅助材料。

**广泛的认知技能的教学** 一些研究人员对提高测验成绩的相反方法作了探索。他们的目的是发展广泛适用的智力技能、工作习惯和解决问题的策略。在测验分数和大学课程之类效标成绩中,都应该表现出这种干预的影响。根据本节开头所作出的区分,这类计划是为了提供教育而不是辅导。其中一些研究者的工作对象是可教育的智力落后儿童和青少年(Babad & Budoff, 1974; Belmont & Butterfield, 1977; A. L. Brown, 1974; Budoff & Corman, 1974; Campione & Brown, 1979, 1987; Feuerstein, 1979, 1980; Feuerstein, Rand, Jensen, Kaniel & Tzuriel, 1987)。另一些研究者关注教育背景差的大学和专业学校的学生(Linden & Whimbey, 1990; Whimbey 1975, 1977, 1980)。

这些计划中所使用的许多训练方法,是为了发展有效地解决问题的行为,例如仔细分析问题,得出答案时考虑各种不同方法、有关细节和含义,深思熟虑而不是凭一时冲动提出或选择一个答案,使用高标准来评价自己的成绩。显然,这些策略不仅应该在测验上,而且在学业活动和许多其他的日常生活活动中,提高个体的智力机能。然而,值得注意的是,在训练中所使用的内容和环境的类型中,这种影响的迁移和概化又达到何种程度。迄今为止所报告的结果都是有前途的,但是这类计划仍然处于探索阶段,需要进行更多的研究,以便证实所获得的提高的广度和持久性。

**总评** 我们考虑了测验之前的三类训练,它们的目的大不相同。三类训练怎样影响作为评定工具的测验效度及其实际用处?第一种是辅导,对类似于测验的项目做大量强化练习。已经指出,编制良好的测验选择的是受这种辅导影响最少的项目类型。因为这种辅导可能提高测验成绩,但并没有相应提高效

标行为。它由此降低测验效度，使其不能有效测量它所评定的广泛的能力，也不能准确弄清个体是否具有在效标情境中取得成功所必需的知识和技能。

另一方面，适应测验的方法是为了消除或平衡以前参加测验经验的差异。像辅导的影响一样，这些差异也是影响测验分数本身的条件，但在所要评定的广泛的行为领域中未必反映出来。因此，适应测验的方法减少测验特有的因素的影响，应该使测验成为更为有效的工具。

最后，训练广泛适用的认知技能，如果行之有效的話，则应该提高受训者处理后来智力任务的能力。测验成绩中必然而且应该反映这种提高。因为测验分数和效标成绩都得到提高，这种训练使测验效度保持不变，却增加个体达到理想的目标的可能性。

## 测验信息的来源

心理测验处在迅速变化的状态。新测验不断涌现，旧测验又出修订本，另外的资料可能改进或改变对现存测验的分数的解释。变化的加速，加上大量的现行测验，使得任何一本教科书都不可能全面评说具体的测验。在论述咨询、临床实践、人员选拔、教育等领域的测验使用的书籍中，能够较深入地介绍各特殊领域的测验工具和问题。本书的有关章节注明这类书籍的出处。然而，为了与目前的发展同步，同测验打交道的任何人都需要熟悉测验信息的直接来源。

最重要的来源之一是奥斯卡·K·伯勒斯(Oscar K. Buros)创立并编辑的直至1978年的《心理测量年鉴》(*Mental Measurements*)



Yearbook, MMY)。自从 1985 年以来, MMY 由内布拉斯加大学伯勒斯心理测量研究所出版。这套年鉴几乎包括以英语出版的所有商业性心理测验、教育测验和职业测验。所介绍的纸笔测验尤为完整。每本年鉴包括规定期间所出版的测验, 因此它补充而不是取代前面的年鉴。这套年鉴的最早几本只是测验的文献目录。1938 年开始, 年鉴采取目前的形式, 包括一位或几位测验专家对测验的批评性评论, 以及每个测验所附的已出版的全部参考书目。也定期给出出版者、价格、型式、测验所适用的被试的年龄等常规信息。目前的计划是每二三年出版一本新的 MMY, 补充新出版的测验。

现在使用电子设备通过西尔弗普拉特信息中心 (Silver Platter Information), 可以得到 MMY 的测验条目和批评性评论。数据库从第九期 MMY 中的条目开始, 每 6 个月更新一次。伯勒斯研究所的另一份刊物是《已出版的测验》(Tests in Print), 现在是第四卷 (TIP-IV, 1994), 由墨菲、康诺利和伊姆珀勒 (L. L. Murphy, Conoley & Impara) 编辑。这份刊物累积介绍所有以英语出版的商业性测验, 以及事实信息和参考书目。《已出版的测验》的每个相继的版本, 也能够用作所有在它之前的 MMY 的一个索引。

已出版测验的信息的另一个主要来源是教育测验服务中心 (Educational Testing Service, ETS) 编制的《测验文献目录集》(Test Collection Bibliographies)。在各个具体的内容领域, 可以得到加注释的测验文献目录。介绍是综合性的, 包括各种类型的测验, 以及为身体残疾者这样的特殊群体和特定用处而设计的测验。每个测验条目提供事实信息, 包括作者、出版日期、出版者、目标总体、测验目的、所测量的变量等。人们能够以很少的费用, 从教育测验服务中心测验集分部 (Test Collection, ETS) 购买各个领域的测验文献目录。这是提供测验当前信息的若干 ETS 刊物

之一。

除了已出版的测验之外,还有书籍、杂志或未出版的报告中所描述的或复制的大量未出版的测验。这些测验主要是供研究人员使用的,各种摘要都对此有评述(例如,Goldman & Mitchell, 1995)。通过教育测验服务中心测验集分部发行的《微缩胶片测验》(*Tests in Microfiche*),可以得到未出版测验的目前信息。每年增加一组新的未出版的测验,也可获得它们的索引。合格的使用者可以购买各个测验。从美国心理学会科学理事会(Science Directorate of the American Psychological Association)可以得到清楚而简要的指南,用于查找已出版测验和未出版测验的信息(《查找信息》, *Finding Information*, 1995)。这种资料定期更新,任何索取该资料的人,都会收到一本最新版本。

对于测验使用者来说,各个现行测验的信息的最直接来源是测验出版者的目录和每个测验配有的手册。在最新的《心理测量年鉴》中,能够找到测验出版者及其地址的全面目录。向主要测验出版者索取,即可获得现行测验的目录。合格的使用者能够购买测验的手册和样本。

测验手册应该提供实施、评分、评价某测验所需要的基本信息。手册中应该包括全面而详细的指导语、评分答案、常模以及信度和效度的资料。此外,手册应该报告建立常模、信度和效度的样本的性质和人数,以及计算信度和效度指标所使用的方法。如果必要的信息太多,收在手册中不方便,则应该提到能够很容易找到这类信息的技术手册或其他印刷资料。换句话说,手册应该使测验使用者在按具体目的选择测验之前能够评价测验。还应该指出,一些测验手册尚未达到这个要求。但是大型的、专业的测验出版者正日益重视编制符合一定科学标准的手册。许多有见识的测验使用者坚信,将来一定会保持和提高这种标准。

评价心理测验的简明而全面的指南，可以参阅美国心理学会(APA)和其他两个与测验有关的学会即美国教育研究协会(American Educational Research Association, AERA)和全国教育测量委员会(National Council on Measurement in Education, NCME)合作制定的《教育和心理测验的标准》(*Standards for Educational and Psychological Testing*)。1954年《标准》首次出版，1966年、1974年和1985年，分别进行修订。三个参与学会正在进行进一步的广泛修订。

80年代开始出现制订《测验标准》<sup>[2]</sup>的需要，这种标准不仅关心测验的技术质量，而且关心测验对个体福利的影响(参阅图1-1)。最新修订的《测验标准》的性质表明这种关心依旧。图1-2包括1966年由三个学会组成的一个联合委员会所制订的《标准》中所提出的目录。显然，它关心测验的选择是否适合于对测验参加者经历的了解，也关心测验分数的解释和使用。这些关心在不断增强。应该指出，图1-2第二节标题是“测验的公平性”。测验使用者日益意识到，测验的使用不当会损害个体，减少她或他为社会作出的贡献。对误用测验的大众批评大大有助于测验使用者提高这种意识。这应该能够减少这种误用，而反过来这又增加了大众对测验使用的潜在好处的认识。

---

## 教育和心理测验的标准

### 第一部分 测验编制和评价的技术标准

1. 效度
2. 信度和测量误差
3. 测验的编制和修订
4. 制定量表、制定常模、分数可比性和等值
5. 测验出版 技术手册和使用者指南

### 第二部分 测验使用的专业标准

6. 测验使用的一般原则
7. 临床测验
8. 教育测验和学校中的心理测验
9. 咨询中的测验使用
10. 就业测验
11. 专业和职业的执照和证书
12. 方案评价

### 第三部分 特殊应用的标准

13. 测验少数民族语言的个体
14. 测验具有不利条件的个体

### 第四部分 实施程序的标准

15. 测验的实施、评分和报告
16. 保护测验参加者的权利

---

图 1-1 《教育和心理测验的标准》(AERA, APA, NCME, 1985)所包括的细目

---

## 教育和心理测验标准提出的目录

### 第一部分 测验的编制、评价和文件提供

1. 效度
2. 信度、测量误差和测验分数信息函数
- 3 测验的编制和修订
4. 制定量表、制定常模、标准和分数可比性
5. 测验的实施、评分和报告
6. 测验的文件

### 第二部分 测验的公平性

7. 公平和偏向
8. 保护测验参加者的权利
9. 测验第一语言为非英语的个体
10. 测验具有能力障碍的个体

### 第三部分 测验应用

11. 测验使用的一般原则
  12. 心理测验和评定
  13. 教育测验和评定
  14. 就业测验、执照和证书
  15. 方案评价和政府政策中的测验
- 

图 1-2 《教育和心理测验的标准》(AERA, APA, NCME)  
修订版所选择的细目(预备稿)

## 注 释

[1] 关于清楚说明心理测验的潜在作用以及现实生活中的例子,参阅达尔  
斯罗姆(Dahlstrom, 1993b)的有关论述。

[2] 为简便起见,我们在全书中采用通常做法,把它们称作“测验标准”。

## 第二章 现代测验的历史背景

概览心理测验的历史背景和起源，可以提供观察问题的视角，并且有助于理解当今的各种测验；<sup>[1]</sup>参照心理测验的前身，有助于了解当代心理测验的发展方向。同样，研究心理测验产生的背景，有助于了解当前测验的优点和不足之处。本章主要讨论测验运动的历史背景和早期发展。后面几章结合一些具体测验讨论各类测验的最新发展，例如能力测验（第八章至第十二章）、兴趣测验（第十四章）以及测验应用领域，例如教育、工业、临床、咨询（第十七章）等。

心理测验的历史根源无从考究，但有人反复提到在中国盛行了大约 2000 年的科举考试制度（Bowman, 1989）。在古希腊，测验附属于教育过程，使用测验来评定是否掌握体育技能和智育技能（Doyle, 1974）。在中世纪，欧洲的一些大学授予学位和荣誉时，采用正式考试。不过，为了确认使当代测验得以形成的主要发展，我们只需要追溯至 19 世纪。我们现在正是要讨论这些发展。

### 对智力落后者分类和训练的早期关注

19 世纪，社会上对智力落后者和精神病患者实行人道主

义。在此之前,通常这些个体受到冷落、嘲笑甚至迫害。随着社会日益关注于智力有问题的个体,这就需要某些统一的标准来对这些个体进行鉴定和分类。欧洲和美国都建立许多社会收容所来护理智力落后者,这就急需建立收容标准和客观的分类系统。首先需要区分精神病和智力落后。前者表现为情绪障碍,可能伴随也可能不伴随智力退化;后者主要特征是从出生或婴儿早期就存在的智力缺陷。1838年法国医生埃斯科罗(Esquirol)出版了两卷本著作,用100多页的篇幅论述现在所说的“智力落后”,第一次明确作出了这种区别。埃斯科罗还指出,从正常到“最严重的白痴”,智力落后有许多等级。为了建立某种系统来区分智力落后的不同等级,埃斯科罗尝试若干种方法后得出结论,个体使用语言的能力是衡量他或她智力水平的最可靠的标准。值得注意的是,目前智力落后的标准主要仍是语言,而且当今智力测验中言语内容的比例也很大。后面几章都会反复说明,言语能力在我们的智力概念中的重要地位。

另一位法国医生塞戈因(Seguin)的贡献也特别重要,他是训练智力落后个体的先驱。塞戈因摒弃智力落后不可治这一流行观念,多年来进行他称之为生理训练法的实验(1866/1907),并且在1837年建立第一所专门教育智力落后儿童的学校。1848年他移居美国,他的思想在这里受到普遍重视。智力落后个体教育机构所应用的许多感觉和肌肉训练方法,最先都由塞戈因提出。使用这种方法,能集中训练严重智力落后儿童的感觉辨别和运动控制。塞戈因此所提出的一些方法,最后成为操作智力测验或非言语智力测验的组成部分。一个例子是塞戈因的“形状板(Form Board),要求个体尽快把不同形状的积木插入相应的凹处。

在埃斯科罗和塞戈因的研究工作半个多世纪之后,法国心

理学家阿尔弗雷德·比纳( Alfred Binet)提出,不能适应正常学校学习的儿童,应该在退学之前被鉴别出来,如果认为他可以接受教育,则安排在特殊班级里(T.H.Wolf,1973)。比纳和儿童心理研究会(Society for the Psychological Study of the Child )的同事敦促法国公共教育部采取措施,改善智力落后儿童的学习条件。结果,成立了一个研究智力落后儿童的管理委员会,比纳是其中一员。这项任命是心理测验历史上的重大事件,后面还会谈到。

## 首批实验心理学家

一般来说,19世纪的早期实验心理学家并不关心个体差异的测量,当时他们的主要目的是对人类行为作出一般化的描述,研究的重点是行为中的共同性而不是差异性。个体差异或者被忽略或者被认为不利于研究结果的推广应用。因此,在相同的条件下所观测到的不同个体的不同反应,被认为是一种误差。这种误差即个体变异性的存在,使得推广是近似的而不是准确的。冯特(Wundt)1879年在德国莱比锡大学创建实验室,许多早期实验心理学家在这里接受训练。在这类实验室中,这种对待个体差异的态度具有普遍性。

实验心理学的缔造者在课题选择以及在研究工作的其他方面,都受到他们的生理学和物理学背景的影响。实验室中所研究的问题大多是视觉、听觉和其他感觉刺激的感受性以及简单的反应时。这种对感觉现象的重视又转过来影响到第一批心理测验的性质(参见后面几节)。

应当指出,19世纪实验心理学影响测验运动的另一个方面。早期心理实验需要在观测中严格控制实验条件。例如,在



反应时实验中，给予被试的指导语的措词可能明显加快，或者减慢被试的反应速度。再如，周围环境的颜色和亮度也会明显改变视觉刺激的外观。这些都充分表明在标准化条件下观测所有实验被试的重要性。这种测验过程的标准化最终成为心理测验的特性之一。

## 弗兰西斯·高尔顿的贡献

英国生物学家弗兰西斯·高尔顿(Francis Galton)是倡导测验运动的主要人物。在高尔顿多种不同的研究活动中，始终如一的是他对人类遗传的兴趣。高尔顿在研究遗传的过程中认识到，需要测量有遗传关系和没有遗传关系的个体的属性。只有这样，才能发现父母和子女之间、兄弟姐妹之间、堂表兄弟姐妹之间、双生子之间的确切相似程度。为此，高尔顿敦促许多教育机构保存对学生所进行的系统的人体测量记录。在 1884 年国际博览会上，他还设立了一个人体测量实验室 (anthropometric laboratory)，参观者付 3 个便士，就能测量某些身体属性，如测验视听觉敏锐度、肌肉力量、反应时及其他简单的感觉运动机能。博览会闭幕后，实验室迁至伦敦南肯辛顿博物院，在那里继续开办达 6 年之久。通过这种方法，高尔顿逐渐积累了关于简单心理过程中个体差异的第一批大规模的系统资料。

高尔顿亲自设计了许多简单测验，在他的人体实验室里进行实施，其中不少测验至今仍以原型或修订型继续使用。例如，用于视觉长度辨别的高尔顿棒、用于确定听觉最高音频的高尔顿笛、用于测量动觉辨别的刻度砝码系列等。高尔顿认为，感觉辨别测验可以用作测量个体智力的工具。在这一点上，他受到

洛克(Locke)理论的某些影响。因此,高尔顿写道:“我们关于外部世界仅有的信息,看来通过我们的感觉通道而获得。我们的感觉越敏锐,我们的判断和智力也就越是大有用武之地。”(1883,p.27)高尔顿也注意到,极度智力落后的个体往往在热、冷、痛等辨别能力上也有缺陷。这种观测结果使他进一步深信,感觉辨别能力“基本上是心智能力中的最高能力”(1883,p.29)。

高尔顿还是应用等级评定量表、问卷法以及自由联想技术的先驱,后来自由联想技术应用于多种目的。高尔顿的另一个贡献是他发展了分析个体差异资料的统计方法。高尔顿对先前数学家推导出的一些统计技术进行选择 and 修改,使得未经专门数学训练的研究人员也能使用这些技术来定量处理测验结果。因此,他极大地推动了应用统计程序来分析测验数据。这项工作后来由高尔顿的许多学生推进一步,最为突出的便是卡尔·皮尔逊(Karl Pearson)。<sup>[2]</sup>

## 卡特尔和早期“心理测验”

在心理测验的发展史上,另一位著名代表人物是美国心理学家詹姆斯·麦基恩·卡特尔(James McKeen Cattell)。他将刚刚建立的实验心理学和新兴的测验运动结合起来。卡特尔师从冯特,在莱比锡大学完成了关于反应时的博士论文,获博士学位。

1888年,卡特尔在剑桥大学讲课期间与高尔顿的接触,使他对测量个体差异的兴趣得到强化。返美之后,卡特尔积极从事于建立心理实验室和传播测验运动。

卡特尔在1890年撰写的一文中所使用的“心理测验”(mental test)一词是第一次出现于心理学文献中。该文描述了一套测

验,当时每年都用它来测试大学生,以确定他们的智力水平。这些个别实施的测验,包括测量肌肉力量、运动速度、疼痛感受性、视听敏感性、重量辨别力、反应时、记忆力等等。在选择测验时,卡特尔与高尔顿的观点基本相同,认为智力机能的测量可以通过感觉辨别力和反应时测验来获得。卡特尔偏爱这类测验是有一定道理的,因为在当时它已能够准确地测量简单的机能,而要客观测量复杂的机能还是毫无希望的事情。

在 19 世纪的最后十年编制了不少成套测验,卡特尔测验是其中的典型。这类测验实施于中小學生、大学生及各种类型的成年人。1893 年芝加哥举行哥伦比亚博览会,贾斯特罗 (Jastrow) 设立了一个展览室,邀请参观者参加感觉、运动、简单知觉过程等测验,并把他们的技能与常模相比较 (J. Peterson, 1926; Philippe, 1894)。有人试图评价这些早期测验,所得结果却并不令人满意。个体在不同测验上的成绩不相一致 (Sharp, 1898 ~ 1899; Wissler, 1901), 而且测验成绩与教师独立评定的智力水平 (T. L. Bolton, 1891 ~ 1892; J. A. Gilbert, 1894) 或者与学业成绩 (Wissler, 1901), 都没有多少关系。

当时欧洲心理学家编制的不少测验,往往包括较为复杂的机能。克雷佩林 (Kraepelin, 1895) 主要对精神病患者的临床检查感兴趣,他编制了一系列测验来测量他所认为的个体特性中的基本因素。该测验主要包括简单的算术运算,用来测量练习效果、记忆、对疲劳与分心的敏感性等。另一位德国心理学家艾宾浩斯 (Ebbinghaus, 1897) 则用算术计算、记忆广度、句子完成等测验施测于学校儿童。三个测验中最为复杂的要数句子完成,但也只有这个测验与学生的学业成绩十分吻合。

1895 年,比纳和亨利 (Henri) 在法国发表文章,批评大多数测验感觉方面内容太多,而且过分集中于简单的、特殊的能力。

他们进一步阐述道，测量较为复杂的机能，不必要求十分准确，因为这些机能有较大的个体差异。他们还提出各种大量的测验，包括记忆、想像、注意、理解、受暗示性、审美等。从这些测验中，我们可以看到最终导致编制著名的比纳智力量表的趋向。

## 比纳和智力测验的产生

比纳和他的同事多年来积极地、创造性地研究测量智力的方法。他们尝试了许多方法，甚至包括测量头盖骨、脸型、手相以及分析笔迹。然而，研究的结果使他们逐渐相信，只有用直接的方法——哪怕是粗糙的——才最有希望去测量复杂的智力机能。一个特殊的机会使得比纳的设想很快付诸实践。前面讲过，1904年，法国公共教育部任命一个委员会，专门研究智力落后儿童的教育方法，比纳也是委员之一。正是为了实现这个委员会的目标，比纳与西蒙(Simon)合作，编制了世界上第一个“比纳-西蒙量表”(1905)。

这个量表，即1905年量表，包括30个题目，由易到难排列。对50名3~11岁的正常儿童以及一些智力落后儿童和成人实施这些题目，通过实证确定其难度水平。所设计的题目包括各种机能，但特别强调判断、理解、推理三者，比纳认为它们是智力的基本组份。虽然也包括一些感知觉题目，但是该量表比起当时大多数测验，言语内容的比例要大得多。不过1905年量表还是一种初步的、尝试性的工具，尚无准确的客观方法来表示总分。

第二个量表即1908年量表，增加了题目数量，删除了1905年量表中不满意的题目。根据约300名3~13岁的正常儿童的

成绩,所有题目都按年龄水平分组。因此,80%~90%的3岁正常儿童通过的所有题目放入3岁水平组,同样,80%~90%的4岁正常儿童通过的所有题目放入4岁水平组,以此类推,直到13岁水平组。儿童在整个测验上的分数表示为智力水平(mental level),这相当于成绩同他或她一样好的正常儿童的年龄。在比纳量表的各种翻译本和改编本中,通常用“智力年龄”(mental age)替代“智力水平”。因为智力年龄这一概念易于理解,引入这一术语无疑有利于普及智力测验。<sup>[3]</sup>但是,比纳本人一直避免使用“智力年龄”一词,因为他觉得智力年龄的发展含义无法证实,他比较喜欢“智力水平”这一中性术语(T. H. Wolf, 1973)。

比纳—西蒙量表的第三次修订本发表于比纳逝世的1911年。这次修订没作什么重大改变,只对少量题目进行修改和重新安排。在若干年龄水平组增加了一些题目,量表扩展到成人水平组。

早在1908年修订本问世之前,比纳—西蒙量表就引起全世界心理学家的广泛注意。许多国家有了翻译本和修订本,美国也有几个版本。戈达德(H. H. Goddard)第一个在美国修订比纳量表,当时他是维恩兰德(Vineland)训练学校(专收智力落后儿童)的心理学家。戈达德修订本对于医学界接受智力测验很有影响(Zenderland, 1987)。当时急需标准化测量,以便对智力落后个体进行诊断和分类,它的产生恰逢这一良机。但是,作为一种测验工具,它不久就被斯坦福大学的推孟(L. M. Terman)及其同事所修订的斯坦福—比纳量表(Terman, 1916)抛在后面,后者适用范围更为广泛,心理测量特性也更为完善。该量表首次采用智力商数(intelligence quotient, IQ),即智力年龄和实足年龄的比值。该量表的最新修订本现在仍广为使用,这在第八章将作全面讨论。第一版库尔曼—比纳(Kuhlmann-Binet)修订本也值得

一提，他把量表的适用范围向下延伸到 3 个月年龄水平（Kuhlmann, 1912）。该量表可算是最早编制的学前儿童和婴儿智力测验。

## 团体测验

比纳测验及其所有修订本都是个别量表（individual scale），因为每次只能施测一个个体。这种量表中许多题目要求被试口头反应或操作材料。一些题目分别规定反应时间。由于这些原因和其他原因，这类测验不适合于团体实施。比纳一类测验的另一个特性是主试必须经过严格的训练。此类量表实质上是临床工具，适用于对个案作深入细致的研究。

像第一个比纳量表一样，团体测验（group testing）的编制也是出于实际需要。1917 年美国参加第一次世界大战，美国心理学会任命一个委员会研究心理学如何为战争服务。在罗伯特·耶基斯（Robert M. Yerkes）的领导下，该委员会认识到，需要根据一般智力水平，对 150 多万新兵进行快速分类。这类资料与多项管理决策有关，例如剔除不符合兵役要求的对象、分配不同工作、进入军官训练营等。正是在这种背景下，编制了第一个团体智力测验。在这项工作中，军事心理学家参考了所有可以得到的测验材料，尤其是阿瑟·奥蒂斯（Arthur S. Otis）本人编制的未经发表的团体智力测验——他把它交给了军队。该测验是奥蒂斯当推孟的研究生时设计的，其主要贡献是采用多项选择题和其他“客观性”题目。

军事心理学家最后编成的测验最终称之为“军队甲种测验”（Army Alpha）和“军队乙种测验”（Army Beta）。前者为一般常规

测验；后者为非言语测验，适用于文盲及不能参加英语测验的外籍新兵。两个测验都适用于大规模团体实施。

第一次世界大战结束不久，两个军队测验转为民用。不仅“军队甲种测验”和“军队乙种测验”本身出现许多修订本，而且它们也作为大多数团体智力测验的模型。测验运动取得巨大发展。不久设计了许多团体智力测验，适用于从学前儿童到大学生等各种年龄和类型的团体。大规模测验计划在先前是不可能的，现在却进行得如火如荼。因为团体测验设计为大规模测验工具，它们不仅可以同时测量大型团体，而且使用十分简单的指导语和施测程序，因而主试不需要多少训练。学校教师开始对班级学生实施智力测验，并在招生前对大学生进行常规测验。还有人开始深入研究特殊的成人团体，例如囚犯。不久，一般公众也有了 IQ 意识。

这种团体智力测验的应用，对测验技术的改进并无多大帮助。在匆忙收集测验分数，并从中作出实际结论时，人们常常忘记了测验在技术上仍然是很粗糙的。当测验不能符合本来就没有理由达到的期望时，人们又往往怀疑甚至敌视一切测验。因此，本世纪 20 年代不加区别地使用测验的这种狂热，既可以说促进了心理测验的发展，也可以说阻碍了心理测验的发展。

## 能力倾向测验

虽然最初设计智力测验来对各种机能取样，以便估计个体的一般智力水平，但是不久就发现智力测验的内容十分有限，不能代表各种重要的机能。事实上，大多数智力测验主要测量言语能力，在有限的程度上，也测量处理数字关系以及其他抽象符

号关系的能力。心理学家逐渐认识到“智力测验”有点用词不当，因为这种测验只测量了智力的某些方面。

诚然，智力测验所测的各种能力，在它们所适用的文化中十分重要。但是人们认识到，最好还是按照这些测验能够获得的信息类型，使用更为确切的名称。例如，有些测验在 20 年代也许被称为智力测验，但后来却被称为“学业能力倾向测验”(scholastic aptitude test)。之所以发生这种用词变化，是因为人们认识到，许多所谓的智力测验测量的是学业所要求和培养的各种能力。

甚至在第一次世界大战之前，心理学家就开始认识到，需要用特殊能力倾向测验来补充一般智力测验。编制这些特殊能力倾向测验(special aptitude test)，主要用于职业咨询以及工业和军事等部门人员的选拔和分类。使用最多的是机械、文书、音乐、美术等能力倾向测验。

对于 20 年代普遍地、不加区别地使用智力测验的批评，也说明了另一个值得注意的事实：个体在智力测验不同组份上的成绩往往大不相同，团体测验的情况尤为如此。在团体测验中，题目通常分为相对同质性内容的分测验。例如，某被试可能在言语分测验上得分相对较高，而在数字分测验上得分则相对较低，或出现相反的情况。在某种程度上，斯坦福-比纳之类测验的这种内部变异性也是明显的，例如，某被试可能在词汇题目上得低分，而在图片或几何图形题目上则得高分。

测验使用者，特别是临床心理学家，经常利用这种内部比较来深入了解被试的心理结构。因此，在评价每一个案时，不仅考虑 IQ 或其他总分，而且也考虑在题目组群或分测验上的成绩。然而，一般来说，这种做法并不可取，因为智力测验的设计，不是出于分析不同能力倾向的目的。要比较的各个分测验包括的题



目往往太少，不能稳定地、可靠地估计特殊能力。结果，假如在不同的时间或使用同一测验的不同型式对个体进行再测，所得出的分测验分数之间的差异也许会相反。如果要进行这种个体内部的比较，就需要设计揭示不同机能差异的专门测验。

测验的实际应用表明需要多重能力倾向测验，与此同时，研究特质结构的平行发展则逐渐为编制这种测验提供了方法。关于智力性质的统计研究一直在探索许多个体在各种不同测验上得分的相关。英国心理学家查尔斯·斯皮尔曼 (Charles Spearman, 1904, 1927) 在本世纪头十年开创了这项研究。后来，在美国心理学家凯利 (T. L. Kelley, 1928)、塞斯顿 (L. L. Thurstone, 1938, 1947b) 以及美国和英国其他研究人员的研究工作基础上发展起来的方法，被称为因素分析 (factor analysis)。

第十一章将全面讨论因素分析方法对测验编制的贡献。现在只需指出，使用这种程序所收集的数据表明，存在着若干相对独立的因素，即特质。传统的智力测验程度不同地体现其中的一些特质，例如言语理解特质和数字推理特质。而另一些特质，例如空间、知觉、机械能力倾向，更多地出现在各种特殊能力倾向测验而不是智力测验中。

因素分析的主要实际成果之一，便是编制出多重能力倾向成套测验 (multiple aptitude batteries)。设计这种成套测验来测量个体在每一特质上的位置。不提供总分或 IQ，而是分别得出各种特质的分数，例如言语理解、数字能力倾向、空间视觉、算术推理、知觉速度等。测验使用者多年来一直试图根据智力测验，进行个体内分析或区别性诊断，但所得结果是粗糙的，而且经常是错误的；而这类成套测验倒是一种合适的工具。多重能力倾向成套测验还把先前特殊能力倾向测验所获得的许多信息，编入一个综合的、系统的测验计划，因为它们包括智力测验中通常没

有包括的一些特质。

多重能力倾向成套测验代表了测验领域相对较新的发展。它们几乎都出现于 1945 年之后。在这一点上, 应该注意第二次世界大战期间军事心理学家的研究工作。军队所进行的许多测验研究都是根据因素分析, 目的是编制多重能力倾向成套测验。例如, 美国空军编制了特殊能力倾向成套测验, 用于选拔飞行员、投弹手、无线电操作人员、光学测距仪操作人员以及许多其他军事专业人员。第二次世界大战期间航空心理学计划有 19 卷, 其中关于美国空军编制成套测验的一项报告就达 9 卷之多 (《空军》, *Army Air Forces*, 1947 ~ 1948)。现在军队各兵种仍在研究能力倾向成套测验。同样也编制了一些民用多重能力倾向成套测验, 广泛应用于教育咨询和职业咨询以及人员的选拔和分类。第十章和第十七章将讨论其中的一部分。

本世纪 80 年代后期和 90 年代初期出现的最新发展, 将传统智力测验和多重能力倾向成套测验这两种先前不同的心理测量方法相结合 (Anastasi, 1994)。现在人们逐渐认识到, 运用测验手段, 能够在不同的广度水平上恰当地评价人类能力, 即从狭义的具体测验 (甚至各个项目) 的能力倾向, 到广度逐渐增加的特质水平, 再到传统 IQ 之类的总分。不同广度的水平最适用于不同的测验目的。因此, 最新编制的智力测验例如 “区别能力量表” (Differential Ability Scales), 或者早期测验的最新修订本, 例如第四版斯坦福-比纳量表 (第八章讨论这两个量表), 都将多重能力倾向的综合内容和适用于具体测验目的的灵活的多水平评分相结合。虽然这两个量表是个别实施的智力测验, 但是这种综合而灵活的编制和使用测验的方法, 也对团体实施的成套测验产生影响, 第十章讨论这一点。第十一章结合智力性质问题的当前进展, 讨论这种能力测验合并的理论基础和实践意义。

## 标准化成就测验

正当心理学家热衷于编制智力测验和能力倾向测验的同时,传统的学校考试方法也在进行技术改进(O. W. Caldwell & Courtis, 1923; Ebel & Damrin, 1960)。其中一个重要步骤是, 1845 年波士顿公立学校开始用书面考试代替主试对学生的口头提问。那时支持这项改革的理由是: 采用书面考试, 所有学生都处于相同情境, 且内容广泛、减少题目选择的偶然性、消除主试可能的偏爱。所有这些理由似曾相识, 后来使用这些理由来证明用客观性多项选择题代替论文题是正确的。

到了 20 世纪, 测量学校教学成就的第一批标准化测验开始出现。受桑代克(E. L. Thorndike)工作的影响, 这些测验利用心理实验室中所发展起来的测量原理。例如, 评定书写和书面作文的量表, 以及拼写、算术计算、算术推理等测验。1923 年发表第一版“斯坦福—比纳成就测验”(Stanford Achievement Test), 导致了第一个成套成就测验的产生, 其编制者为三位早期测验发展的领导人杜鲁门·凯利(Truman L. Kelley)、贾尔斯·鲁赫(Giles M. Ruch)、刘易斯·推孟。该套测验预示了现代测验的许多特征, 根据单一的常模团体作出评价, 可以比较不同学科的成绩。

同时, 许多证据表明, 教师之间评定论文测验缺乏一致性。到了 1930 年, 人们普遍认识到, 论文测验不仅花费主试和被试较多的时间, 而且所得结果的可靠性不如“新型”的客观性题目。<sup>[4]</sup>随着标准化成就测验中逐渐使用客观性题目, 日益强调题目设计, 以便测验对知识的理解 and 应用等教育目标。新型客观性测验十分适合于机器评分, 30 年代评分机器应运而生。

值得一提的是,与此同时,全州、全地区全国性的测验机构开始建立起来,其中最为著名的当属大学入学考试委员会(College Entrance Examination Board, CEEB)。该机构成立于 20 世纪初,以便减少大学新生入学考试的重复,现在它的测验程序以及参加的大学的数量和性质都发生了巨大变化,这种变化反映了测验和教育二者在这一时期的发展。1947 年,CEE B 同卡内基公司(Carnegie Corporation)和美国教育理事会(American Council on Education)的测验职责合并,成立教育测验服务中心(Educational Testing Service, ETS)。此后,ETS 负责为大学、职业学校、行政部门及其他部门编制许多测验。还应该提到美国高等学校测验方案(American College Testing Program),它始于 1959 年,以甄别没有包括在 CEEB 中的大学申请人。还有若干全国性测验,用于选拔获得奖学金的高才生。

成就测验不仅应用于教育部门,而且也用于工业部门和行政部门中的人员选拔。前面已经提到,大约公元前 150 年起,中国采用科举考试制度(Bowman, 1989)。18 世纪末 19 世纪初,现代的欧洲国家采用考试选拔政府雇员。1883 年,美国文官选拔委员会(Civil Service Commission)将竞争考试作为一项常规方法(Kavrucek, 1956)。1922 年,奥洛克(L. J. O'Rourke)被任命为新建的研究分会的主任,随后美国文官选拔委员会的考试中引进第一次世界大战期间及之前发展起来的测验编制技术。时下,这项工作由隶属于美国人事管理局(Office of Personnel Management)的一个大型高级技术研究机构负责。

随着越来越多受过心理测量训练的心理学家参与编制标准化成就测验,成就测验的技术逐渐接近于智力测验和能力倾向测验。编制和评价这三类测验的方法具有许多共同之处。让成就测验测量广泛的教育目标,而不是对事实细节的回忆,这种不

断的努力使得成就测验的内容更为接近于智力测验。现在，这两类测验之间的差异主要在于内容的特异性程度，以及在什么程度上测验要求事前有指定的教学课程。

## 人格评定

心理测验的另一领域涉及情感行为或非智力行为，我们将在第十三至十六章讨论这一问题。适用于这种目的的测验一般称之为人格测验，虽然许多心理学家喜欢在广义上使用“人格”一词来表示完整个体。因此人格包括智力特质和非智力特质，不过，在心理测验术语中，“人格测验”一词往往表示测量情绪状态、人际关系、动机、兴趣、态度等特性。

早期人格测验的先驱克雷佩林对精神病患者实施自由联想测验。在这个测验中，给予被试经过特殊选择的刺激词，要求他报告想到的第一个词。克雷佩林 (1892) 也使用这种技术去研究疲劳、饥饿、药物等的心理效应。他得出结论，所有这些情况都增加表层联想的频数。在 19 世纪最后 10 年，萨默 (Sommer, 1894) 认为，可以使用自由联想测验来区别不同形式的心理障碍。后来，自由联想技术应用于种种目的，至今仍被应用。还应该提到高尔顿、皮尔逊、卡特尔等在发展标准化问卷和评定量表技术方面的工作。虽然他们最初设计这些方法出于其他目的，但是最终由他人来编制成目前最常见的人格测验。

人格问卷的原型，即自陈问卷 (self-report inventory, 见第十三章)，是伍德沃斯 (Woodworth) 在第一次世界大战期间编制的“个人资料调查表” (Personal Data Sheet) (DuBios, 1970; Franz, 1919, pp. 171 ~ 176; L. R. Goldberg, 1971; Symonds, 1931, 第五章)。这个

测验是一种初步的甄别工具，用于鉴定不宜服役的严重精神病患者。问卷包括一些与常见心理病理学症状有关的问题，由被试自己回答。然后，计算所报告的症状数目得出总分。在第一次世界大战结束之前，“个性资料调查表”尚未编制完成。但是，战争一结束，就编制了民用型式，包括适用于儿童的特殊型式。而且“伍德沃斯个人资料调查表”后来成为大多数情绪适应问卷的原型。在一些问卷中，试图将情绪适应再分成具体类型，例如家庭适应、学校适应、职业适应等。另外一些测验则测量较狭小的行为领域或较独特的社会反应，例如人际接触中的支配—服从。后来也编制了定量测量态度和兴趣的测验（第十四章）。这些测验本质上也是根据问卷技术而编制。

人格测量的另一种方法是使用表演（performance）或情境测验（situational test，见第十六章）。在这种测验中，被试完成一项目的被掩盖的测验。大多数测验模拟日常生活情境。20年代末和30年代初，哈茨霍恩（Hartshorne）、梅（May）及他们的同事第一次大量应用这种技术来编制测验（1928, 1929, 1930）。这套测验根据学校儿童进行标准化，测量作弊、说谎、偷窃、合作、坚持等行为。众多测验都得到客观的量化分数。第二次世界大战期间，美国战略服务局（Office of Strategic Services, OSS）评定处编制了适用于成人水平的系列情境测验。这种测验测量相对复杂的、难以捉摸的社会行为和情绪行为，实施时需要十分复杂的设备和受过训练的人员。此外，对个体反应的解释比较主观。

投射技术（projective techniques，见第十五章）是研究人格的第三种方法，发展很快，在临床上尤其如此。在这种测验中，给予被试相对非结构性的任务，允许自由作答。这种方法的基本假设是，被试会把他特有的反应模式投射到这类任务中去。像表演和情境测验一样，投射测验或多或少掩盖测量目的，因而减

少被试故意制造令人满意的印象的可能性。前面提到的自由联想测验是最早的投射测验之一。投射技术也使用句子完成测验。投射技术常用的任务还有画图、排列玩具创设场景、即兴演戏、解释图片或墨迹等。

各类人格测验都存在着某种理论和实践上的问题。每种方法都有特殊的长处和不足。总的来说,人格测验在实际技术上落后于能力测验。但是,这种落后不能归因于努力不够。1950年以来,人格测量的研究占据很大比重,人们研究许多新型方法和技术改进问题。正是人格测量中所碰到的特殊困难,使得这个领域的研究进展不快。

当前人格测验研究中出现两种一致的重要趋向(参见 Anastasi, 1985b, 1992a, 1993; Digman, 1990; L. R. Goldberg, 1993; Simon, 1994) 第一,越来越多的证据表明,在任务完成和行为发展中,情感(“人格”)特质和认知(“能力”)特质相互影响。人们逐渐认识到,两类特质之间的传统区分是人为提出的,只是为了便于描述和测量行为的不同方面而已。第二,对人格的特质和构成的理论分析,支持将认知特质和情感特质重新统一为人类活动的综合模型,它包括各种类型的行为。这种综合模型与智力特质(第十一章)和情感特质(第十三章)的基础研究有关。

## 注 释

- [1] 关于早期心理测验产生的更为详细的叙述,参见古迪纳夫(F. L. Goodenough, 1949)和彼得森(J. Peterson, 1926)的有关论述。关于更一般化的背景,参见博林(Boring, 1950)以及墨菲和科瓦奇(G. Murphy 和 Kovach, 1972)的有关论述。关于心理测验历史较新的阐述,参见杜波依斯(DuBois, 1970)和麦克雷诺兹(McReynolds, 1975, 1986)的有关论述。关于个体差异研究的历史背景,参见阿纳斯塔西(Anastasi, 1965)的有关

论述。关于心理测验的发展趋势,参见阿纳斯塔西(Anastasi, 1993)的有关论述。

- [2] 关于早期主要统计概念的发展及其代表人物,参见考尔斯(Cowles, 1989)的有关论述。
- [3] 古德依纳夫(E. L. Goodenough, 1949, pp. 50 ~ 51)提到,1887年,即1908年比纳—西蒙量表发表之前21年,夏耶(S. E. Chaile)在《新奥尔良内外科杂志》(*New Orleans Medical and Surgical Journal*)上发表一套婴儿测验,按照通常通过各测验的年龄排列。由于杂志的发行量有限,或许也由于科学界准备不足,古德纳夫所提出的年龄量表这一概念的意义在当时未引起注意。比纳的量表受到同时期一些心理学家特别是布林和戴梅(Blin & Damaye)的影响,他们编制了一组口头问题,从中得到每个儿童的总分(T. H. Wolf, 1973)。
- [4] 第十七章教育测验一节中,将引用论文项目和客观性项目相对效果的研究。



## **第二部分**

---

### **技术原理和方法原理**



### 第三章 常模和测验分数的意义

由第三章至第七章组成的第二部分，介绍基本的概念和方法。这些概念和方法对于理解心理测验和合理解释测验结果，都是必要的。这 5 章分别讨论常模、信度、效度以及项目分析和测验设计。本章讨论常模的建立和使用，以及便于测验使用者解释测验分数的其他程序。缺乏另外的解释资料，任何心理测验上的原始分数(raw score)都是毫无意义的。某个体在一个数学推理测验中做对 15 题，或者在一个词汇测验中认识 34 个单词，抑或用 57 秒成功地拼装一个机械物体，这些说法都没有提供或提供很少有关该个体在这些机能中的任何一项上排位的信息。大家所熟悉的百分数分数，对于解释测验分数这个问题，也没有提供令人满意的解答。例如，在一个词汇测验上做对 65% 的分数，也许等价于在另一个测验上做对 30%，也许等价于在第三个测验上做对 80%。显然，构成每一测验的项目的难度水平将决定分数的意义。像所有原始分数一样，只有按照明确规定的、统一的参照框架，才能解释百分数分数。

心理测验上的分数，通常要参照常模(norm)来解释，常模表示标准化样本的测验分数。通过确定代表性样本的成员在测验中的实际行为，从而在经验上建立常模。然后，任何个体的原始分数，参照由标准化样本所得出的分数分布，以找出他或她在那个分布中处于何处：该分数恰好是标准化样本的平均分吗？它

稍稍低于平均分吗？或者，它接近分布的高端吗？

为了较为精确地弄清该个体参照标准化样本所具有的确切位置，就要把原始分数转换成某种相对测量。设计这些导出分数(derived score)有双重目的。首先，它们表明了该个体在常模样本中的相对位置，因此参照其他个体，可以评价他或她的成绩。其次，它们提供可资比较的度量，对个体在不同测验上的成绩，可以进行直接比较。例如，如果我们仅仅知道某女孩在词汇测验上的原始分数为 40，在算术推理测验上的原始分数为 22，显然，关于她在这两个测验上的相对成绩，我们仍一无所知。是她的词汇测验成绩好？还是算术测验成绩好？抑或两个测验成绩一样好？由于不同测验上的原始分数通常采用不同的单位，因此就不可能对这些分数进行直接比较。特定测验的难度水平也影响原始分数之间的这种比较。另一方面，导出分数具有相同的单位，并参照适用于不同测验的相同的或极为相似的常模样本，因此能够比较个体在许多不同机能上的相对成绩。

为了实现上述两个目标，有多种转换原始分数的方法。然而，导出分数基本上采用两种主要方法之一：(1)所达到的发展水平；(2)在某个特定团体内的相对位置。本章中有几节分别讨论这些分数类型和一些常用的变式，但首先必须讨论一些基本的统计概念，因为它们是建立和使用常模的基础。下面一节，只是为了阐明某些常用统计量的意义。仅仅出于这个目的而举出简化的计算例子，并非提供统计方法的训练。在实际应用这些技术中所采用的具体程序和计算详情，读者可参阅任何一本最近的心理统计课本（例如，D. C. Howell, 1997; Runyon & Haber, 1991; West, 1991）。现在人们日益认识到，不仅对于测验使用者，而且对于任何想要看懂心理学任一领域所发表的研究报告的人，都需要基本的统计方法知识(L. S. Aiken, West, Sechrest &

Reno, 1990; Anastasi, 1991; Lambert, 1991; S. T. Meier, 1993).

## 统计概念

统计方法的一个主要目的,就是组织和概括数量资料,以便于理解。1000 个测验分数的罗列,可能使人看不出什么东西,那种形式的分数可能没有多大的意义。整理一组无序的原始分数,第一步是把分数列为频数分布(frequency distribution)表,如表 3-1 所示。把分数按适宜的组距分组,把每一分数标记在适当的组中,这类分布就准备好了。当所有的分数都标入后,就计算标记,得出每一组距内的频数,即个案数。这些频数的总和等于团体中个案的总数  $N$ 。表 3-1 表明 1000 名大学生在代码学习测验上的分数。在该测验中,用一组人造单词即无意义音节来替代另一组。在两分钟尝试期间,给出正确的替代音节的数目为原始分数,全距从 8 分到 52 分。分数的组距定为 4 分,从分布上端的 52~55 分至下端的 8~11 分。频数栏表明,2 人得分在 8~11 之间,3 人得分在 12~15 之间,8 人得分在 16~19 之间,等等。

频数分布所提供的信息,也能够以分布曲线图来表示。图 3-1 以图形表明表 3-1 的资料。在基线即水平轴上,是分为组距的分数;在垂直轴上是频数,即处于每一组距内的个案数。图形以两种方式绘制,这两种方式都是常用的。在直方图中,每一组距上方垂直柱的高度,相当于分数在该组距内的人数。我们可以想到,每一个体站在另一个体肩上,就形成垂直柱。在频数多边图中,组距的中心和适当频率的交点,表明每一组距内的人数。然后用直线逐一连接各点。

表 3-1 1000 名大学生的代码学习测验分数的频数分布表

| 组 距      | 频 数 |
|----------|-----|
| 52 ~ 55  | 1   |
| 48 ~ 51  | 1   |
| 44 ~ 47  | 20  |
| 40 ~ 43  | 73  |
| 36 ~ 39  | 156 |
| 32 ~ 35  | 328 |
| 28 ~ 31  | 244 |
| 24 ~ 27  | 136 |
| 20 ~ 23  | 28  |
| 16 ~ 19  | 8   |
| 12 ~ 15  | 3   |
| 8 ~ 11   | 2   |
| N = 1000 |     |

(资料摘自安娜斯塔西, 1934, p. 34)

除了少数例外,图 3-1 所示的分布类似钟形的正态曲线(normal curve)。数学上完美的正态曲线见图 3-2。这类曲线具有重要的数学特性,它为多种统计分析提供基础。然而,出于现在的目的,只需注意若干特性。曲线实质上表明,在全距的中点个案数最多;接近两端时,个案数就逐渐减少。曲线两边对称,单峰在中心。人类特质,从身高体重到能力倾向和个性特征,大多数分布近似于正态曲线。一般来说,团体越大,分布就越类似于理论上的正态曲线。

按照集中趋势(central tendency)的某一度量也能够描述一组分数。这类度量提供一个最典型的即最有代表性的分数,来表示整个团体的成绩。这些度量中最为熟知的当属平均数

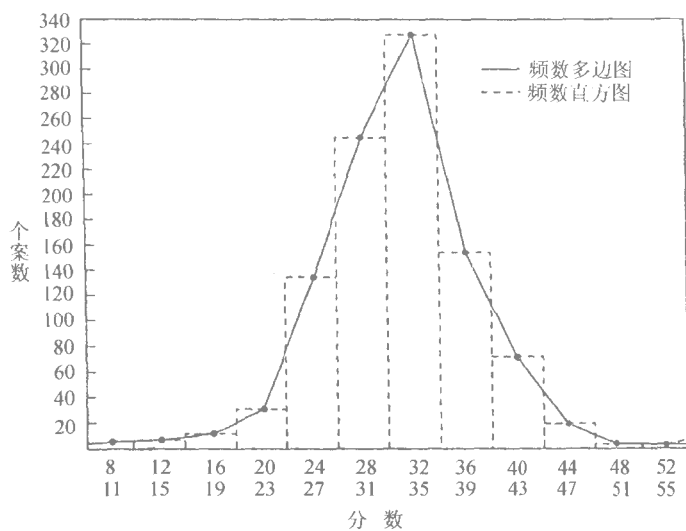


图 3-1 分布曲线：频数多边形和直方图（资料根据表 3-1）

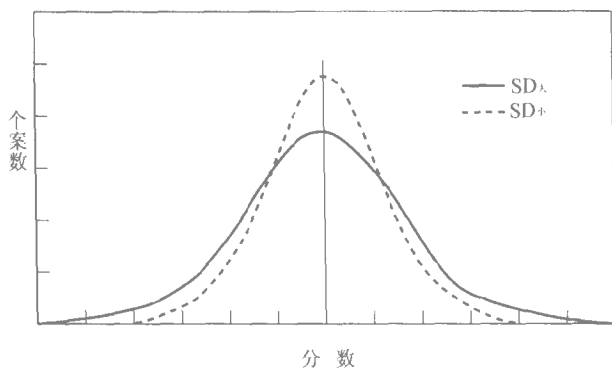


图 3-2 平均数相同、离散性不同的两个频数分布

(mean,  $M$ )。把所有分数相加,然后把总和除以个案数( $N$ ),就得出平均数。集中趋势的另一种度量是众数(mode),即频数最多的分数。在频数分布中,众数是具有最高频数的组距的中点。因此,在表 3-1 中,众数处于 32~35 的中间,为 33.5 分。应当指出,这个分数相当于图 3-1 中分布曲线的最高点。集中趋势的第三种度量是中位数(median),即所有分数按照大小排列时的最中间的分数。中位数是把分数分为两组的那个点,半数个案处于中位数以上,而另半数个案则处于中位数以下。

进一步描述一组测验分数,就要测量离散性(variability),即围绕集中趋势的个别差异的程度。报告离散性的最明白、最熟悉的方式,当属最高分和最低分之间的全距(range)。然而,全距是极其粗糙的、不稳定的,因为它仅仅由两个分数来决定。一个不同寻常的高分或低分,就会明显影响全距的大小。测量离散性的较为精确的方法,则根据每一个体的分数与团体平均数之间的差异。

在这点上,查看表 3-2 中的例子将对我们有所帮助。该例根据 10 个个案计算出所讨论的各种度量。选取如此小的团体,是为了简化演算,虽然在实践中,我们通常处理较大的团体。表 3-2 也用于介绍应当指出的某些标准的统计符号,以便将来参照。习惯上,大写  $X$  表示最初的原始分数,小写  $x$  表示每个分数与团体平均数的离差,符号  $\Sigma$  表示“总和”。可以看到,表 3-2 中第一列用于计算平均数和中位数,平均数为 40,中位数为 40.5,处于 40 和 41 的中间——5 个个案(50%)在中位数之上,5 个个案(50%)在中位数之下。在如此小的团体中,得出的众数就没有什么意义,因为个案没有群集在任何一个分数上。然而,在技术上,41 可代表众数,因为两个个体得到该分数,而所有其他分数只出现一次。



表 3-2 集中趋势和离散性的实例

|              | 分数 (X)         | 离差 ( $x = X - M$ ) | 离差平方 ( $x^2$ )   |
|--------------|----------------|--------------------|------------------|
| 50% 个案       | 48             | + 8                | 64               |
|              | 47             | + 7                | 49               |
|              | 43             | + 3                | 9                |
|              | 41             | + 1                | 1                |
|              | 41             | + 1                | 1                |
| 中位数 = 40.5 → | 40             | 0                  | 0                |
| 50% 个案       | 38             | - 2                | 4                |
|              | 36             | - 4                | 16               |
|              | 34             | - 6                | 36               |
|              | 32             | - 8                | 64               |
|              | $\sum X = 400$ |                    | $\sum x^2 = 244$ |

$$M = \frac{\sum X}{N} = \frac{400}{10} = 40$$

$$\text{方差} = \sigma^2 = \frac{\sum x^2}{N} = \frac{244}{10} = 24.40$$

$$\text{SD 或 } \sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{24.40} = 4.9$$

注：本表中给出的符号  $\sum$  和  $\sigma$  是同一希腊字母的大小写，读作“西格玛”。在许多统计学著作中，SD(或简化为 S)表示实得数据的样本的标准差，而  $\sigma$  则表示抽样总体的标准差的估计值。

第二列表明每个分数偏离平均数 40 之上或之下有多远。这些离差的总和总是等于零，因为平均数上下的正负离差必然平衡，即相互抵消（+ 20 - 20 = 0）。当然，假如我们忽视符号，求绝对离差的平均数，那么得出每一个体偏离团体平均数的平均值这个度量。这类“平均离差”虽然有描述性的价值，但由于人为地忽视符号而不适用于进一步的数学分析，因此在实践中并不应用。

对离散性更为有用的一种度量是标准差 (standard deviation, 用符号 SD 或  $\sigma$  表示), 就是把每个离差平方, 合理地消去负号。表 3-2 的最后一列采用这种程序。这列总和除以个案数<sup>[1]</sup>

( $\frac{\sum x^2}{N}$ ) 称之为方差 (variance) 或均方差 (mean square deviation)。

在区分不同因素对于测验成绩中个别差异的作用方面, 方差被证明为极其有用。然而, 出于现在的目的, 我们主要关注 SD, 它是方差的平方根, 如表 3-2 所示。该度量通常用于比较不同团体的离散性。例如, 图 3-2 中的两个分布, 平均数相同, 但离散性不同。较宽个别差异的分布, 比起较窄个别差异的分布, 得出较大的标准差。评价两个团体的相对成绩时, 我们应该比较标准差和平均数两者。如果两个团体的离散性不同, 这也许表明高分比例、低分比例或两者的团体差异, 而不管平均数差异。现在有一些综合程序, 把平均数差异和标准差差异的影响相结合 (例如, 参见 Feingold, 1995)。

根据常模来表示个体在不同测验上的分数, 标准差也为之提供了基础, 我们将在标准分数一节中说明这一点。应用于正

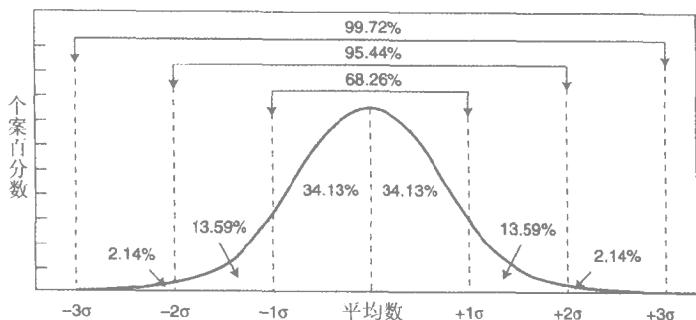


图 3-3 正态曲线中个案的百分数分布

态或近似正态分布曲线时,对标准差的解释尤为明确。在这种情况下,标准差和个案比率之间存在着确切的关系,如图 3-3 所示。在这条正态曲线的基线上,已经标出表示平均数上下 1、2、3 个标准差的距离。例如,在表 3-2 给出的例子中,平均数为 40 分;  $+1\sigma$  为 44.9 (即  $40 + 4.9$ );  $+2\sigma$  为 49.8 (即  $40 + 2 \times 4.9$ ) 等等。在正态曲线中处于平均数和  $+1\sigma$  之间的个案百分数为 34.13。由于曲线对称,个案的 34.13% 同样处于平均数和  $-1\sigma$  之间,所以平均数两边  $+1\sigma$  和  $-1\sigma$  之间,有 68.26% 的个案。几乎所有个案 (99.72%) 处于平均数  $\pm 3\sigma$  之内。这些关系在解释标准分数和百分等级中尤为重要,后面几节将讨论这一点。

## 发展常模

使分数具有意义的一种方法,是表明个体在正常发展线上处于怎样的水平。例如,在智力测验上,一个 8 岁儿童的成绩和一般的 10 岁儿童一样好,则可以说他具有 10 岁的智力年龄,而一个智力落后的成人的成绩也是如此,则同样可以说他的智力年龄为 10 岁。另一方面,我们可以说一个四年级儿童在阅读测验上达到六年级常模,而在算术测验上达到三年级常模。其他的发展系统则高度定性地描述具体机能的行为,例如感觉运动或概念形成。无论怎样表述,基于发展常模的分数,在心理测量上往往是粗糙的,不能适用于精确的统计处理。然而,它们十分适用于描述的目的,尤其对个体作深入细致的临床研究时,它们也适用于研究的目的。

智力年龄 (mental age) 我们在第二章指出,比纳—西蒙量

表的种种译本和改编本,普遍采用“智力年龄”一词,虽然比纳本人采用较为中性的“智力水平”一词。在比纳量表及其修订版(1986年之前)等年龄量表上,项目按年龄水平分组。例如,标准化样本中大多数 7 岁儿童所通过的那些项目,就放置在 7 岁水平,大多数 8 岁儿童所通过的那些项目,则放置在 8 岁水平,等等。儿童在测验上的分数,将相当于他成功完成的最高年龄水平。在实际测试中,个体的成绩会出现一定程度的分散(scatter),换句话说,被试没有通过低于他的智力年龄的一些测验项目,但通过了高于他的智力年龄的一些测验项目。由于这种原因,习惯上先计算基础年龄(basal age),即全部项目都通过的最高年龄。然后再加上各个较高年龄水平上通过的所有项目所代表的月龄。两者之和就是儿童在测验上的智力年龄。

智力年龄常模也适用于不划分年龄水平的测验。在这种情况下,首先确定儿童的原始分数。这类分数可以是整个测验上答对项目的总数,也可以是根据时间、错误数或这类度量的某种结合。标准化样本中每一年龄组的儿童所获得的平均原始分数,构成这类测验的年龄常模。例如,8 岁儿童的平均原始分数作为 8 岁常模。如果某个儿童的原始分数等于 8 岁组的平均原始分数,那么他在测验上的智力年龄就为 8 岁。参照年龄常模,能够以相似的方式转换这类测验上的所有原始分数。

应当指出,智力年龄的单位并非随着年龄保持不变,而往往随着年龄的增加而缩小。例如,4 岁时智力落后 1 年的儿童,在 12 岁时大约落后 3 年。3~4 岁智力成长 1 年,相当于 9~12 岁智力成长 3 年。由于智力发展在年龄小时较为快速,随着个体接近成熟年龄而逐渐减慢,因此智力年龄的单位随着年龄增加而相应缩小。如果我们想到以“身高年龄”来表示个体的身高,那么对这种关系也许就不难理解了。以英寸为单位的身高年龄

之间的差异,3岁和4岁就要大于10岁和11岁。由于智力年龄单位的逐渐缩小,5岁时超前或落后1年所表示的与常模的离差,就要大于10岁时超前或落后1年。

年级当量(grade equivalents)教育成就测验上的分数,经常按照年级当量来解释。这种做法是可以理解的,因为测验应用于学校环境之内。说某学生的成就分数相当于七年级的拼写、八年级的阅读和五年级的算术,这与传统的智力测验中使用智力年龄同样流行。

计算每一年级儿童所获得的平均原始分数,就得出年级常模。因此,如果标准化样本中四年级学生,在算术测验上答对问题的平均数是23,那么原始分数23就是年级当量4。中间的年级当量代表一个年级的各个分数,通常用内插法得出,虽然在学年之内的不同时间对学生进行测验,也能够直接得出中间的年级当量。因为学年包括10个月,能够用十进制小数表示连续的月份。例如,4.0表示四年级开始时(9月份测验)的平均成绩,4.5表示四年级中期(2月份测验)的平均成绩,等等。

尽管年级常模也很流行,但它仍存在着若干不足之处。首先,各个年级的教学科目多少有点不同。因此,年级常模仅仅适用于测验各年级都开设的共同学科。它们一般不适用于中学水平,中学的许多学科可能只学习一二年。甚至对于每一年级都教的学科,各个年级的教学重点也可能不同,因此,在一个特定的年级期间,一门学科的进度会比另一门学科快。换句话说,年级单位显然是不等距的,不同的学科领域内发生的不等性又没有规律。

年级常模也容易引起误解,除非测验使用者牢牢记得导出年级常模的方法。例如,假设一个四年级儿童在算术测验中获得年级当量6.9,这并非意味着她已掌握了六年级所教的算术

内容。毫无疑问，她主要由于在四年级算术中的优异成绩而获得该分数。我们当然不能假设，她具备学习七年级算术的必要条件。最后，年级常模往往被误以为是成绩标准。例如，一位六年级的教师可能假设，她班上的所有学生在成就测验上应该达到或接近六年级常模。使用年级常模时，这种误解当然不会令人惊奇。然而，任何一个年级之内的个别差异，使得成就测验分数的全距必然延伸若干年级。

顺序量表(ordinal scales)发展常模的另一种方法源于儿童心理学的研究。通过对婴幼儿的行为发展的经验观察，人们描述诸如运动、感觉辨别力、语言交流、概念形成等机能的随年龄发展的典型行为。一个早期例子由格塞尔(Gesell)及其同事在耶鲁大学的研究工作所提供(Ames, 1937; Gesell & Amatruda, 1947; Halverson, 1933; Knobloch & Pasamanick, 1974)。格塞尔发展顺序量表(Gesell Developmental Schedules)按月份显示了儿童在运动、适应性、语言、社会性等4种主要行为领域中所达到的大致发展水平。将儿童的行为和从四周到36个月的8个关键年龄的典型行为相比较，就得出这些水平。

格塞尔及其合作者强调早期行为发展的顺序模式。他们引用许多证据，以说明发展顺序的一致性和行为变化的顺序性。例如，在眼睛注视及手和手指运动中，儿童对放在他或她面前的小物体的反应，表现出一种特有的年龄顺序。先是使用整只手，尝试手掌抓握；然后是使用拇指和手掌；再次是使用拇指和食指钳子般夹持物体。在头几年的行走、爬楼梯以及大多数感觉运动发展中，同样观察到这类顺序模式。各发展阶段遵循不变的顺序，每一阶段以掌握先前阶段的必要的行为特征为前提，<sup>[2]</sup>在这种意义上，在这个框架之内所编制的量表是顺序量表。

本世纪60年代，出现了瑞士儿童心理学家让·皮亚杰(Jean

Piaget)的发展理论热(参见 Flavell, 1963; Ginsburg & Oppen, 1969; D. R. Green, Ford & Flamer, 1971)。皮亚杰的研究着重于从婴儿到十多岁少年的认知过程的发展。他关注具体的概念而不是一般的能力。这类概念或图式的一个例子便是物体永久性,当儿童从不同的角度看到物体或看不到物体时,他靠什么意识到物体的同一性和继续存在?另一个广泛研究的概念是守恒 (conservation),即认识到一种属性经过知觉外形变化而保持不变。例如,把等量的液体倒入不同形状的容器中,又如把相同长度的小棒进行不同的空间排列。

发展心理学家在研究中广泛使用皮亚杰作业,有些作业已被编入标准化量表,第九章将讨论这一点( Goldschmid & Bentler, 1968b; Pinard & Laurendeau, 1964; Uzgis & Hunt, 1975)。同皮亚杰的方法相一致,这些量表均为顺序量表,在概念发展中,一个阶段的到达要依据先前各阶段的完成。各种作业被用来揭示每一发展阶段的主要方面,只是到后来,它们才成为所收集的关于一般到达每一阶段的年龄之经验资料。在这方面,不同于编制年龄量表所采用的程序,后者首先根据连续年龄之间的区分来选取项目。虽然人们继续关注皮亚杰方法的贡献,但批评性的分析和经验评价同时强调它的建设性和局限性 ( Sugarman, 1987)。<sup>[3]</sup>

总之,顺序量表能用来鉴别儿童在具体行为机能的发展中所达到的阶段。虽然这类分数可以按照近似的年龄水平来报告,但对儿童的特征行为的定性描述是次要的。这类量表的顺序性指通过连续阶段的发展一致性。顺序量表一般提供儿童实际上能做什么的信息(例如,没有帮助能爬楼梯;把液体倒入不同形状的容器时,能认识到液体容量相同),就这点来说,它们具有本章后面一节所讨论的领域参照测验的重要特点。

## 团体内常模

现在，几乎所有的标准化测验都提供某种形式的团体内常模。这类常模是根据最可比较的标准化样本的成绩，来评价个体的成绩。例如，将一名儿童的原始分数与相同年龄或相同学校年级的儿童相比较。团体内分数具有统一的、定义明确的定量意义，能够适用于大多数类型的统计分析。

百分位数(percentiles) 百分位数分数表示标准化样本中低于某一特定原始分数的人数百分比。例如，在算术推理测验上，如果 28% 的个体答对 15 题以下，那么原始分数 15 就是第 28 百分位数( $P_{28}$ )。百分位数表示个体在标准化样本中的相对位置。也能够认为百分位数是 100 人团体中的等级，只是在评定等级时，习惯上从高分开始计算，团体中最好的个体获得等级 1 另一方面，我们在百分位数中则从低分开始计算，所以百分位数越小，个体的地位就越差。

第 50 百分位数( $P_{50}$ )就是中位数，我们已经讨论过，这是一种集中趋势的度量。50 以上的百分位数表示高于平均成绩；50 以下的百分位数则表示较差的成绩。第 25 百分位数和第 75 百分位数分别称为第 1 四分位点和第 3 四分位点( $Q_1$  和  $Q_3$ )，因为它们分别划出分布中最低和最高的  $1/4$ 。像中位数一样，它们也适用于描述分数分布及把该分布和其他分布相比较。

不要把百分位数混淆于熟悉的百分数分数。百分数分数是原始分数，表示答对项目的百分数；百分位数则是导出分数，表示人数的百分数。一个原始分数，低于标准化样本所获得的任何分数，百分等级为 0( $P_0$ )；一个原始分数，高于标准化样本所



获得的任何分数，百分等级为  $100(P_{100})$ 。然而，这些百分位数并不表示零分的原始分数和满分的原始分数。

百分位数分数具有几个优点。它们便于计算，易于理解，甚至对于未经专业训练的人士也是如此。此外，百分位数用途广泛。它们能够同样适用于成人和儿童，并适用于任何类型的测验，而不管它是测量能力倾向还是人格变量。

百分位数分数的主要缺点是单位明显不等，尤其在分布的两端。如果原始分数的分布近似正态曲线，大多数测验分数都如此，那么，靠近分布中部的原始分数，转换成百分位数时，其差异便被扩大；相反，靠近分布两端的原始分数，转换成百分位数时，其差异便被大大缩小。分数之间距离的这种失真，可以参见图 3-4。回忆一下，在正态曲线中，个案集中在中部，当接近两端时，个案就分散。结果，靠近中部的任一百分数的个案在底线上的距离，要短于靠近分布两端的相同百分数。在图 3-4 中，如果我们比较百分等级 40 和 50 之间的距离与百分等级 10 和 20 之间的距离，就很容易看到百分等级 (PR) 间距的这种差异。这两种距离与百分等级 10 和 1 距离之间的差异，则更为明显（数学上导出的正态曲线中，百分位数零处于无穷大，因此在图上看不到）。

如果我们考虑一下离开正态曲线平均数相等  $\sigma$  距离的百分等级，那么从相反方向也能看到同样的关系。在图 3-4 的正态曲线下给出了这些百分等级。因此，平均数和  $+1\sigma$  之间的百分位数差异是 34 (84 ~ 50)。 $+1\sigma$  和  $+2\sigma$  之间的百分位数差异只有 14 (98 ~ 84)。

显然，百分位数表示每一个体在标准化样本中的相对位置，而不是分数之间的差异量。然而，如果把百分位数标绘在算术概率纸上，它们也能够提供分数之间差异的正确的直观图。算

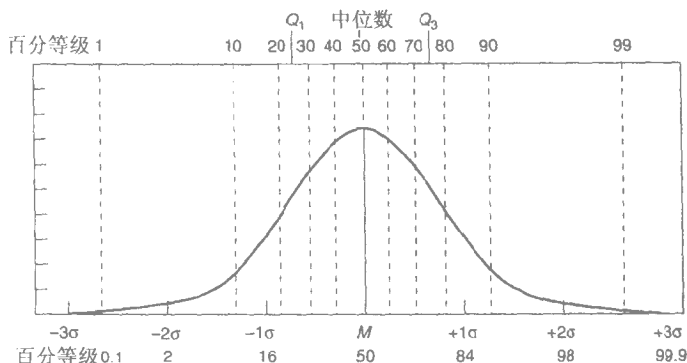


图 3-4 正态分布中的百分等级

术概率纸是一种剖面纸，纵线的间距和正态分布中百分位点一样（如图 3-4），而横线的间距相同，或纵横线反之亦然（如图 3-5）。这种正态百分位数图（normal percentile chart）适用于标绘不同个体在同一测验上或同一个体在不同测验上的分数。在两种情况下，都能正确地表示分数之间的实际差异。现在，许多能力倾向成套测验和成就成套测验都在分数剖面图上使用这种技术，分数剖面图表示个体在每个测验上的成绩。

标准分数（standard score）目前测验日益使用标准分数，从大多数观点来看，标准分数是一种最令人满意的导出分数。标准分数按照分布的标准差来表示个体与平均数的距离。

对最初的原始分数，进行线性转换或非线性转换，都可以得出标准分数。通过线性转换（linear transformation）得出标准分数时，它们保持最初原始分数的准确的数量关系，因为把每个原始分数减去一个常数，然后把结果除以另一个常数，就计算出标准

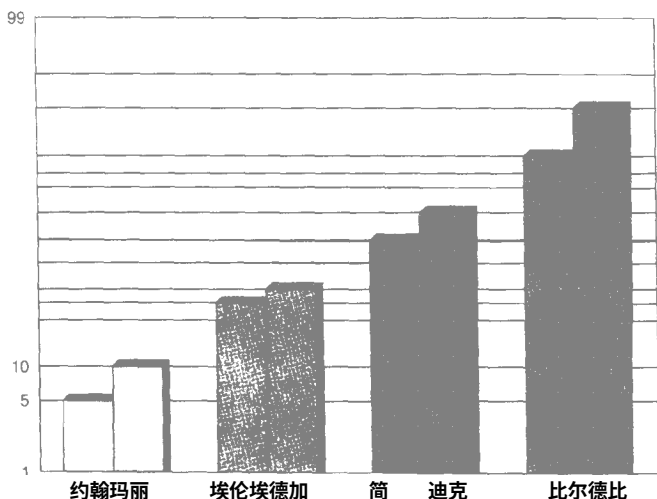


图 3-5 正态百分位数图。百分位数的间距相当于正态分布中的等距。比较一下约翰和玛丽之间以及埃伦和埃德加之间的分数差异 在两对之内,百分位数都相差 5 分。简和迪克百分位数相差 10 分, 比尔和德比百分位数也相差 10 分。

分数。通过这种线性转换所导出的标准分数之间差异的相对大小, 准确地相当于原始分数间差异的相对大小。这种标准分数分布的所有性质, 同最初原始分数分布的完全一样。由于这种原因, 使用最初的原始分数能够进行的任何计算, 使用线性标准分数同样能够进行, 所得结果没有任何差异。

线性导出的标准分数, 通常简称为‘标准分数’或‘ $z$  分数’。为了计算  $z$  分数, 我们首先算出个体原始分数和常模团体平均数之差, 然后把这个差除以常模团体的标准差。表 3-3 说明两个个体  $z$  分数的计算: 一个个体处于团体平均数之上一个标准差,

另一个体处于团体平均数之下 0.40 标准差,正好等于平均数的任何原始分数,其  $z$  分数为零。显然,采用这样一种程序,所有处于平均数之下个体的导出分数为负值。此外,因为大多数团体的全距不会超出平均数 3 个标准差,为了足以区分个体,这种标准分数必须至少报告一位小数。

表 3-3 标准分数的计算

| $z = \frac{X - M}{SD}$    |  | $M = 60$                  | $SD = 5$ |
|---------------------------|--|---------------------------|----------|
| 海伦的分数                     |  | 比尔的分数                     |          |
| $X_1 = 65$                |  | $X_2 = 58$                |          |
| $z_1 = \frac{65 - 60}{5}$ |  | $z_2 = \frac{58 - 60}{5}$ |          |
| $= +1.00$                 |  | $= -0.40$                 |          |

以上两种情况,即出现负值和小数,对于计算和报告两种目的来说,往往产生令人迷惑、使用不便的数字。由于这种原因,通常采用进一步的线性转换,把分数转换成方便的形式。例如,大学入学考试委员会的学业评定测验(Scholastic Assessment Tests, SAT)的分数,被转换成平均数为 500、标准差为 100 的标准分数。因此,在这种测验上,标准分数 -1 就是 400( $500 - 100 = 400$ )。同样,标准分数 +1.5 就是 650( $500 + 1.5 \times 100 = 650$ )。把最初的标准分数转换成新量表,只需要把标准分数乘以设定的标准差(100),然后从设定的平均数(500)中,加上或减去该乘积。可以任意选取其他方便的数值作为新的平均数和标准差。例如,韦克斯勒智力量表(Wechsler Intelligence Scales)各个分测验的分数,都转换成平均数为 10、标准差为 3 的分布。所有这

些测量都是线性转换的标准分数的实例。

回忆一下，把原始分数转换成导出量表的原因之一，就是使得不同测验上的分数可作比较。在前面一节所讨论的线性导出标准分数，只有从大致相同形式的分布中得出时，才是可作比较的。在这种条件下，一个分数，例如平均数之上 1 个标准差，表示个体在两个团体中处于同样的位置。在两个分布中，该分数大致超过同样百分数的个体，如果再知道分布的形式，就能确定这个百分数。然而，如果一个分布为明显偏态，另一个为正态，那么，标准分数 + 1.00 在一个团体中可能仅仅超过 50% 的个体，但在另一个团体中却超过 84% 的个体。

为了使不同分布形态的分数具有可比性，可以采用非线性转换，使分数符合指定类型的分布曲线。前面几节中所描述的智力年龄和百分位数，就是非线性转换，但它们又有已经讨论过的其他局限性。出于这个目的，通常采用正态曲线，虽然在某种情况下，另一种类型的分布也许更为合适。这种选择的主要原因之一，则是大多数原始分数分布最为近似正态曲线，而不是近似任何其他类型的曲线。此外，诸如身高、体重等身体度量，采用通过实际运算所导出的等距量表，一般也产生正态分布。正态曲线的另一个优点是，它具有许多有用的数学性质，这便于进一步的计算。

把原始分数分布转换成正态分布，由此所得的标准分数则称为正态化标准分数(normalized standard scores)。参照离开正态曲线的平均数不同标准差距离的个案百分数的表格，就能计算正态化标准分数。首先，计算标准化样本中处于每一原始分数及以上的人数百分数。然后，在正态曲线频数表上查到该百分数，就得出相应的正态化标准分数。正态化标准分数的形式同线性导出的标准分数一样，即平均数为 0，标准差为 1。因此，正

态化标准分数为零,表示个体处于正态曲线的平均数,超过团体的 50%; -1 分,表示他大致超过团体的 16%; +1 分,则表示他大致超过团体的 84%。这两个百分数分别相当于正态曲线的平均数上下 1 个标准差的距离,参照图 3-4 的底线能够看出这一点。

像线性导出的标准分数一样,正态化标准分数能够被转换成任何方便的形式。如果把正态化标准分数乘以 10,再加上或减去 50,它就被转换成 T 分数(T score),这是由 W. A. 麦柯尔(McCall, 1922)最先提出的一种分数。在该量表上,50 分表示平均数,60 分表示平均数之上 1 个标准差,等等。另一种众所周知的转换便是标准九分(stanine)量表,由美国空军在第二次世界大战期间提出。该量表是一位数字分数系统,平均数为 5,标准差近似为 2。<sup>[4]</sup>“标准九分”(“stanine”是“standard nine”的缩约词)根据分数范围从 1 到 9 而得名。分数限于一位数字,便于某些计算,因为每个分数在计算机上仅仅占用一列。

原始分数能够被方便地转换成标准九分,首先按大小顺序排列最初分数,然后根据表 3-4 中给出的正态曲线百分比来指定标准九分。例如,如果团体恰好由 100 个个体组成,4 个最低分数的个体得标准九分 1,其次 7 个得 2 分,再其次 12 个得 3 分,依次类推。当团体包括的个案多于或少于 100 时,首先计算相当于每个指定百分比的人数,然后,这些个案数就分别得到合适的标准九分。因此,200 个个案中,有 8 个得标准九分 1(200 的 4%为 8)。150 个个案中,有 6 个得标准九分 1(150 的 4%为 6)。对于包括 10~100 个个案的团体,巴特利特和埃德加顿(Bartlett & Edgerton, 1966)编制了表格,这样就能够直接把等级转换成标准九分。由于标准九分在理论和实践两方面的优点,因而得到广泛使用,尤其适用于能力倾向测验和成就测验。

表 3-4 转换标准九分时使用的正态曲线百分比

|       |   |   |    |    |    |    |    |   |   |
|-------|---|---|----|----|----|----|----|---|---|
| 百 分 比 | 4 | 7 | 12 | 17 | 20 | 17 | 12 | 7 | 4 |
| 标准九分  | 1 | 2 | 3  | 4  | 5  | 6  | 7  | 8 | 9 |

虽然对于大多数目的来说，正态化标准分数是令人最为满意的一种分数，然而，对所有分布进行例行正态化，在技术上存在着某些异议。只有当样本很大又有代表性，并且有理由认为造成偏态分布只是由于测验本身的缺陷，而不是由于样本的特性及影响所考虑的行为的其他因素，才能进行正态化转换。也应当指出，原始分数的最初分布近似正态时，线性导出的标准分数与正态化标准分数是十分相似的。在这种条件下，导出两种分数的方法大不相同，但是得出的分数几乎完全相同。显然，对于实际上已经是正态的分布进行正态化处理，不会产生多少变化或不会产生变化。每当可行时，为了获得原始分数的正态分布，一般总是希望适当调整测验项目的难度水平，而不是随后对明显非正态分布进行正态化。对于近似正态的原始分数分布，线性导出的标准分数同正态化标准分数具有同样的目的。

离差智商(deviation IQ) 为了把智力年龄 (MA) 分数转换成个体相对地位的统一指标，早期的智力测验采用比率 IQ (智力商数)。这种 IQ 只是智力年龄与实足年龄的比率，再乘以 100 消除小数 ( $IQ = 100 \times MA/CA$ )。显然，如果一名儿童的 MA 等于他或她的 CA，该儿童的 IQ 恰好为 100。因此 IQ 为 100 表示正常的或平均的成绩。100 以下的 IQ 表示“落后”，100 以上的 IQ 则表示“超前”。

传统的比率智商貌似符合逻辑、简单易行，但是不久就证明是靠不住的。一个重要的技术难点是，除非智商分布的标准差

随着年龄增长能够大致保持不变,否则不同年龄的智商就不可比较。例如,10岁时 IQ 为 115,与 12 岁时 IQ 为 125,可以表示同样程度的优秀,因为它们在各自的年龄分布上,都处于离平均数 1 个标准差的距离。在实际工作中,使编制的测验符合心理测量的要求即比率智商在所有年龄具有可比性,又是十分困难的。主要由于这种原因,比率智商大多被所谓的离差智商所取代,离差智商实际上是熟知的标准分数的另一种变式。离差智商是一种标准分数,平均数为 100,标准差接近于斯坦福一比纳智商分布的标准差。由于斯坦福一比纳比率智商(最后使用于 1937 年版本)的标准差在所有年龄并非一成不变,它在稍大于 16 的中位数值上下波动,因此,在报告新编测验的标准分数时,如果选择接近 16 的标准差,就能采用和斯坦福一比纳比率智商同样的方法,来解释所得分数。由于斯坦福一比纳智商使用多年,测验者和临床人员习惯于按照这类智商水平来对测验成绩进行解释和分类。他们知道,从个体的智商如 40、70、90、130 等等,就能预期什么。因此,一种导出量表相当于熟悉的斯坦福一比纳智商分布,在使用中就有某种优点。选择与斯坦福一比纳智商分布中数值相同的平均数和标准差,就能使分数单位相当。

还应当指出,使用“IQ”一词来表示这种标准分数,可能会令人误解。导出这种 IQ,与得出传统的比率 IQ 中所采用的方法不同,它们不是智力年龄与实足年龄的比率。这样做的道理,在于“IQ”一词家喻户晓,假如这种分数的标准差近似于先前所熟知的 IQ 的标准差,就能够像 IQ 那样进行解释。韦克斯勒智力量表最早按照离差 IQ 表示测验分数。在这些测验中,平均数为 100,标准差为 15。目前的一些团体智力测验和斯坦福一比纳量表的第三次修订版(1960),也使用离差 IQ。

随着日益广泛地使用离差 IQ,重要的是要记住,只有当不



同的测验采用相同或接近数值的标准差时,它们的离差 IQ 才是可以比较的。测验手册中都应该报告这个数值,测验使用者也应该认真注意。如果测验编制者在编制离差 IQ 量表时,选择不同数值的标准差,那么在该测验上任何 IQ 的意义,就会与其他测验完全不同。表 3-5 说明这些差异,该表表示标准差为 12 到 18 的正态分布中,获得不同 IQ 水平的个案百分数。在所出版测验的 IQ 量表中,实际上采用这些标准差数值。例如,表 3-5 表示,当 SD 为 16 时(例如斯坦福一比纳量表),IQ 为 70 划出分数最低的 3.1%,但它划出的百分数可以少至 0.7%(SD = 12)或者多至 5.1%(SD = 18)。传统上以 IQ70 作为临床鉴别明显智力落后的临界点。当然,同样的差异也存在于 IQ130 及以上,该 IQ 用于选择智力天才儿童接受特殊课程。智商分布范围在 90 和 110 之间的 IQ,一般称之为正常,包括总体的少至 42%

表 3-5 平均数均为 100 但标准差不同的正态分布中每一 IQ 组的个案百分数

| IQ 组      | 百分数频数   |         |         |         |
|-----------|---------|---------|---------|---------|
|           | SD = 12 | SD = 14 | SD = 16 | SD = 18 |
| 130 及以上   | 0.7     | 1.6     | 3.1     | 5.1     |
| 120 ~ 129 | 4.3     | 6.3     | 7.5     | 8.5     |
| 110 ~ 119 | 15.2    | 16.0    | 15.8    | 15.4    |
| 100 ~ 109 | 29.8    | 26.1    | 23.6    | 21.0    |
| 90 ~ 99   | 29.8    | 26.1    | 23.6    | 21.0    |
| 80 ~ 89   | 15.2    | 16.0    | 15.8    | 15.4    |
| 79 ~ 79   | 4.3     | 6.3     | 7.5     | 8.5     |
| 70 以下     | 0.7     | 1.6     | 3.1     | 5.1     |
| 总 计       | 100.0   | 100.0   | 100.0   | 100.0   |

或多至 59.6%，这取决于所用的测验。当然，在新编测验和早期测验的新版本中，测验出版者尽力采用统一的标准差 16。然而，在目前使用的测验中，标准差仍有很大差异，因而必须核对一下。

**团体内各类分数的相互关系** 在我们讨论导出分数时，读者可能已经注意到各种类型分数之间的相互关系。百分位数与正态化标准分数至少在图形上有类似之处。如果最初的原始分数分布近似于正态曲线，则线性标准分数与正态化标准分数也没有什么区别。最后，标准分数可转换成 IQ，反之亦然。关于最后一点，再次考察斯坦福—比纳之类测验上比率 IQ 的意义，这将会表明我们能够把这些早期比率 IQ 本身看作标准分数。如果我们知道，斯坦福—比纳比率 IQ 分布的平均数 100，标准差近似为 16，我们能够得出结论，IQ116 处于平均数之上一个标准差的距离，它表示标准分数 +1.00。同样，IQ132 相当于标准分数 +2.00，IQ76 相当于标准分数 -1.50，等等。此外，斯坦福—比纳比率 IQ116，近似相当于百分等级 84，因为正态曲线中 84% 的个案处于 +1.00SD 之下。

图 3-6 总结了本章至此所讨论的正态分布中各类分数之间存在的关系。这些分数包括  $z$  分数、大学入学考试委员会 (CEEB) 分数、韦克斯勒离差智商 ( $SD = 15$ )、T 分数、标准九分、百分位数等。任何测验上的比率智商，如果是正态分布，并且标准差为 15，都与离差智商量表相一致。任何其他正态分布智商，假如我们知道它的标准差，就能够把它加到图上去。例如，如果其标准差为 20，则 IQ120 相当于 +1SD，IQ80 相当于 -1SD，等等。

总而言之，报告分数的正确形式，主要取决于方便、熟悉、易于建立常模等。某种形式（包括离差智商）的标准分数，已经普

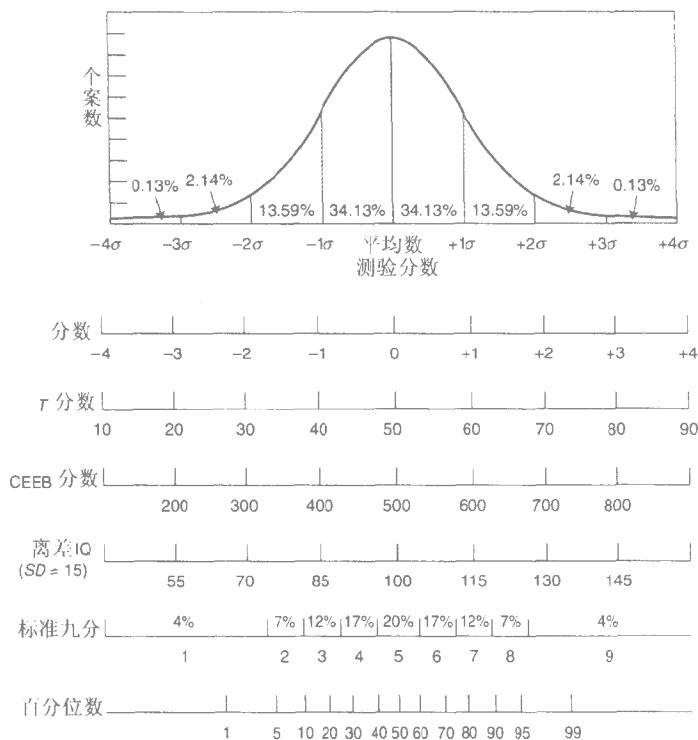


图 3-6 正态分布中不同类型测验分数之间的关系

遍取代其他类型的分数，因为它们具有测验编制和数据统计处理上的某些优点。然而，如果我们认真导出分数，又作出合适解释，则大多数类型的团体内导出分数基本上是类似的。符合某种统计条件时，我们能够方便地把其中的每一种分数转换成任何其他一种分数。

## 常模的相对性

**测验之间的比较** 报告智商或任何其他分数，总是应该同时报告得出该分数的测验名称。不可能抽象地解释测验分数，它们必须参照特定的测验。如果学校记录表明，比尔·琼斯获得 IQ94，特里·布朗获得 IQ110，没有进一步的信息，就不能以表面数值来认可 IQ。把他们在各自学校中所做的特定测验交换一下，这两个学生的位置也许会颠倒过来。

同样，由于测验常模缺乏可比性，也许会完全弄错个体在不同机能上的相对位置。让我们假设，对一个学生进行言语理解测验和空间能力倾向测验，以确定她在两个领域的相对位置。假如言语能力测验的标准化根据中学生的随机样本，而空间测验的标准化则根据参加选修课程的经过选择的学生团体，主试可能得出错误的结论，个体的言语能力比空间能力好得多，而事实上情况也许正相反。

另一个例子是同一个体在不同时间的测验成绩的纵向比较。如果一个学生的累加记录表明，四、五、六年级时的智商分别为 118、115、101 等，解释这些变化之前要问的第一个问题就是：“该学生这三次参加了什么测验？成绩明显下降也许只反映测验之间的差异而已。假如那样的话，即使三个测验都在一周之内实施，该儿童也会得到这些分数。

考虑同一个体在不同测验上所得分数之间的系统差异，有三个主要原因。第一，尽管测验名称类同，测验内容也许不同。所谓的智力测验便是这种混淆的许多例证。虽然用同一术语来描述测验，但其中一个测验也许仅仅包括言语内容，另一个测验

也许主要包括空间能力倾向,再一个测验也许按大约相等的比率,包括言语、数量、空间等内容。第二,量表单位也许是不可比较的。本章的前面解释过这一点,如果一个测验上 IQ 的标准差为 12,另一个测验上 IQ 的标准差为 18,那么,假如某个体在第一个测验上得到 IQ 112,多半会在第二个测验上得到 IQ 118。第三,不同测验在建立常模中所使用的标准化样本的组成有所不同。显然,与能力较高的团体相比较,或与能力较低的团体相比较,同一个体看来在后一种情境中会做得更好一些。

参照测验本身或测验手册,通常能够发现测验内容或量表单位缺乏可比性。然而,各个常模样本之间的差异,多半被忽视了。测验结果中许多不一致无法解释,也许原因就在于此。

常模样本 任何常模,不管怎样,都限于导出常模的特定的常模总体。测验使用者永远不应该忘记建立常模的方法。心理测验常模决不是绝对的、普遍的或永久的。它们只是表示组成标准化样本的个体的测验成绩。在选取这种样本时,通常尽力从测验所适用的总体中获得一个有代表性的部分。

在统计术语中,区分样本(sample)和总体(population)。样本指实际参加测验的个体组成的团体。总体指从中抽取样本的、范围更大而性质相同的团体。例如,假如我们希望建立城市公立学校 10 岁男孩这个总体的测验成绩常模,我们会测验一个经过仔细选择的样本,它包括 500 名 10 岁男孩,他们在美国若干城市的公立学校上学。要参照地理分布、社会经济水平、种族及其他有关的性质核查样本,以保证样本确是所规定总体的代表。

在测验常模的建立和应用中,应该十分注意标准化样本。显而易见,建立常模的样本应该很大,足可提供稳定的数值。从同一总体中同样选择另一个样本,得出的两个常模不应该有显著差异。如果常模的取样误差很大,在解释测验分数时,显然就

没有什么价值。

同样重要的必要条件是，样本应该是所考虑的总体的代表。可能使样本失去代表性的细微的选择因素，都应该认真加以研究。公共机构样本可以说明其中的一些选择因素。因为这种样本通常很大，十分适合于测验目的，所以，对于积累常模资料，不失为一个理想的对象。然而，这些样本的特殊的局限性，也应该认真加以分析。例如，对在校学生进行测验，由于能力较差学生逐渐退学，在后续年级中，就会增加优胜个案的选择。这种淘汰对不同的亚团体的影响也不同。例如，男孩的淘汰率高于女孩，较低社会经济水平学生的淘汰率又高于较高社会经济水平的学生。

在其他的公共机构样本中，例如囚犯、心理医院的病人，或教养院智力落后儿童，选择因素同样起作用。由于公共机构本身具有许多特殊因素，这类样本对于犯法者、心理障碍者或智力落后者的总体，都不具有代表性。例如，智力落后个体中，身体残疾者比起身体健康者，更有可能进入教养院。同样，严重智力落后个体的相对比率，公共机构样本比总体大得多。

与样本代表性问题密切有关的是，需要规定常模所能够推广的特殊总体。显然，保证样本代表性的一种方法是，按照现有样本的性质来限定总体。例如，如果规定总体只包括 14 岁学校儿童，而不是全体 14 岁儿童，那么学校样本就具有代表性。当然，在理论上，应该按照测验目的，事先规定合乎需要的总体，然后再组成合适的样本。然而，由于获得被试中的实际困难，可能达不到这种目标。在这种情况下，比较可取的做法是重新规定较小范围的总体，而不是报告理想总体的常模，这个总体是不能用标准化样本代表的。在实际工作中，没有什么测验根据一般假设的广泛总体进行标准化。没有一个测验能提供人类的常

模令人怀疑的是,对于“美国成人”<sup>34</sup>“10岁美国儿童”之类定义宽泛的总体,许多测验是否确实给出合适的常模。结果,不同的编制者所得到的样本,由于侧重于不同的方面,对于他们所宣称的总体,也许就失去代表性。因此,得出的常模也许是不可比较的。

解释测验分数时,测验使用者应该考虑各种具体条件,它们可能影响特定测验进行标准化中所使用的常模样本。这类条件包括特别的选择因素,以及收集常模资料时主要的社会条件 (Anastasi, 1985d)。

**全国参照常模** 解决常模缺乏可比性问题的一种办法是,使用另一个参照测验,以编出不同测验上分数的等价表。设计这类等价表是为了说明,测验 A 上的什么分数等价于测验 B 上的某一分数。采用同等百分位数法,便可做到这一点,即分数在假定的团体中具有相同的百分位数时,就认为它们是等价的。例如,如果同一个团体的第 80 百分位数,相当于测验 A 上 IQ115 也相当于测验 B 上 IQ120,那么就认为测验 A 的 IQ115 等价于测验 B 的 IQ120。一些测验出版者在有限的程度上采用这种方法,他们为自己的一些测验编制了等价表(例如, Lennon, 1966a)。

有人时常提出雄心勃勃的建议:依据单一的参照测验校准每个新测验,参照测验本身则实施于一个有高度代表性的全国常模样本 (Lennon, 1966b)。这种程序的一个例子,便是在美国教育总署的领导下,教育测验服务中心进行的参照测验研究 (Jaeger, 1973)。这项研究作出系统的努力,为 7 个使用最多的小学生阅读成就测验提供可资比较的、确有代表性的全国常模。通过严格控制的实验设计,测试 50 个州的约 30 万个四、五、六年级学生。参照测验由“城市成就测验”(Metropolitan Achieve-

ment Test) 的阅读分测验和词汇分测验所组成, 在这项研究的一个阶段建立该测验的新常模。在该研究的校准阶段, 每个学生参加 7 个成套测验中的两套测验的阅读理解分测验和词汇分测验, 每套测验依次和其他每一套测验组成一对。一些团体参加同一套测验的两个分测验的平行型式。为了控制测验顺序, 在另一些团体中, 所有的配对测验以相反顺序重复。对所有这些资料进行统计分析, 采用同等百分位数法编制分数等价表。也编制出解释分数的手册, 供学校系统和其他感兴趣的人员使用 (Loret, Seder, Bianchini & Vale, 1974)。

后来使用参照测验研究的校准阶段的资料, 编制成单一分数量表, 称为全国参照量表 (National Reference Scale, Rentz & Bashaw, 1977)。因而所编制的转换表, 准许在任何年级水平, 把 7 个测验的任何型式的分数, 转换成统一的、连续量表上的 3 点分数。编制该量表时, 使用拉希 (Rasch) 模型的项目分析和评分方法。拉希模型是最简单的项目分析模型之一, 这在本章后面一节中讨论, 在第七章中还将有较为全面的描述。

对于许多测验目的来说, 不同测验的分数可作比较是十分有用的。这些分数按统一的测量量表表示, 并参照单一的常模样本。然而, 应当指出, 分数可比性的程度和类型不尽相同。特定情境中所获得的可比性, 取决于测验内容的相似性及信度、难度等心理测量性质的相似性, 也取决于用于获得可比性的统计程序 (Angoff, 1984; Angoff & Cowell, 1986; P. W. Holland & Rubin, 1982) 不应该把几个测验描述为相等的或完全等价的, 除非它们确是相互可换。然而, 假如可作比较的分数使用得当, 又完全了解它们是如何导出的, 那么, 不同的类型和程度的可比性倒能够方便测验结果的解释。

具体常模 解决现有常模的非等价问题的另一种方法是



——对于大多数测验来说，也许是一种较为现实的方法——根据所规定的小范围的总体对测验进行标准化，以便适合每一测验的具体目的。在这种情况下，应该和常模一起，明确报告常模总体的范围。因此，可以说常模适用于“大型商务机构聘用的办公室人员”或“一年级工程专业学生”。对于许多测验目的，高度具体的常模是合乎需要的。甚至当有代表性的常模适用于所规定的大范围的总体时，分别报告亚团体常模，常常也是十分有益的。每当可以识别的亚团体在特定测验上得到显著不同的分数时，情况就是如此。年龄、年级、课程类型、性别、地理区域、城乡环境、社会经济水平及许多其他变量，都可以形成亚团体。测验的用途决定哪一种区分类型最为恰当，也决定是一般常模还是具体常模较为合适。

还应该提到地方常模，它常常由测验使用者自己在特定背景之内建立。导出地方常模所使用的团体，其规定的范围，甚至比上面所考虑的亚团体还要小。因此，雇主可以根据特定公司之内某类工作的申请人来建立常模。大学招生办公室可以根据它自己的学生总体来建立常模。或者一所小学可以根据它自己的分数分布来评价每个学生的成绩。对于许多测验目的，例如，预测后来的工作成绩或大学成就，比较儿童在不同学科的相对成就，或者测量个体随时间的进步，这些地方常模都比大范围的全国常模合适。

**固定参照组** 虽然计算大多数导出分数的方法，可以对测验成绩进行直接的常模解释，但也有一些值得注意的例外。一种非常模量表使用固定参照组，以便保证分数的可比性和连续性，而没有提供成绩的常模评价。使用这种量表时，常模解释则需要参照从合适总体中独立建立的常模。地方常模和其他具体常模常常用于这个目的。

按照固定参照组编制量表的最早的实例之一，便是大学入学考试委员会“学术能力倾向测验”(Scholastic Aptitude Test)<sup>[5]</sup>的分数量表(Donlon, 1984)。在 1926 年(该测验首次实施的时间)和 1941 年之间, SAT 分数根据常模量表来表示, 即根据每次施测中参加测验的考生的平均数和标准差。由于大学入学考试委员会会员大学的数目和种类的增加, 以及考生总体组成的变化, 这就得出结论: 应该保持量表的连续性。否则, 一个考生的分数就会取决于特定年份期间参加测验的团体的属性。要求量表具有连续性的一个更为重要的原因, 则是观察到: 由于选择因素的不同作用, 在一年某些时候参加 SAT 的学生, 成绩不如在其他时候参加 SAT 的学生。因此, 1941 年以后, 所有 SAT 分数, 都根据 1941 年参加测验的大约 11000 名考生的平均数和标准差来表示。这些考生组成固定参照组, 后来编制的测验型式都使用它。因此, 在 SAT 任何型式上的 500 分, 相当于 1941 年样本的平均数; 600 分处于平均数之上一个标准差, 等等。

为了能够把 SAT 任何型式上的原始分数, 都转换成这种固定参照组分数, 每一型式都包括一个简短的参照测验(即共同项目组)。因此, 每个新型式与一二个较早型式相联系, 较早型式又通过项目链依次与其他型式相联系, 一直追溯到 1941 年型式。把这些非常模 SAT 分数与合适的分数分布, 例如某特定的大学、某类型的大学、某地区等等分数分布相比较, 就能够解释 SAT 分数。在大学招生中, 这些具体常模比基于全体考生总体的年度常模有用。此外, 只有使用固定分数量表, 才能发现一段时间考生总体的变化。最近, SAT 量表根据 1990 年中学毕业、在三四年级时参加该测验的 100 多万学生, 重新定出平均数和标准差。1995 年 4 月 1 日以后, 则根据从 1990 年参照组导出的“新”量表报告参加 SAT 学生的分数。已经编制了说明书和工

具,以帮助测验使用者把从前的量表转换成单项的分数和总分,反之亦然(参见第十七章)。因此,对于具体的测验目的,准许对个体成绩作出全面的、多样的解释。<sup>[6]</sup>

根据固定参照组建立的量表,在一个方面类似于物理测量中使用的量表。在这一点上,安格夫(Angoff, 1962, pp. 32 ~ 33)认为:“几乎没有一个人知道,在测量高度或距离中所使用的英尺长度的最初的确切定义,即最初约定用哪位国王的脚作为标准;另一方面,没有一个人不知道,怎样按照该单位确定长度和距离。我们忽视英尺的最初的确切意义或由来,无论如何不会减少它的实用性。其理由是:英尺不会随时间而变化,我们自己对它也十分熟悉。自然,同样的考虑完全适用于英寸、英里、华氏温度等其他的测量单位。在心理测量领域,同样有理由说,量表的最初定义应该是无足轻重的。重要的倒是,量表保持不变——在多重型式测验程序中,通过严格的型式等同,做到了这一点——并且辅之以常模数据,以帮助解释和作出具体决定,而条件许可的时候,则要经常修订这种数据。”

**项目反应理论** 自从本世纪 70 年代以来,一种数学原理十分复杂的程序引起了广泛关注,它适用于分析测验项目的难度(Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991; Jaeger, 1977)。由于需要大量计算,只有随着高速计算机的日益使用,这种程序才切实可行。虽然这些方法的复杂性和具体数学程序有所不同,但最初一般都称之为潜在特质模型(latent trait models)。它们所使用的基本度量是,某种能力(所谓的潜在特质)的个体通过某种难度的项目的概率。然而,这并不意味着在物理的或生理的意义上存在着这类潜在特质或基本能力,它们也不引起行为。潜在特质是一种统计构想,数学上可从实际观察到的测验反应之间的关系中导出。对被试潜在特质的大致的初步

sonality Inventory, MMPI) 为例, 测验使用者可以得到关于测验参加者的人格倾向和情绪状态的计算机打印结果, 有数字分数, 还有解释性陈述。自己拥有计算机的测验使用者更多地购买计算机程序, 对于特定的测验, 例如韦克斯勒儿童智力量表修订版 (WISC - R) 和成人智力量表修订版 (WAIS - R), 这种程序不仅提供数字分数, 而且提供解释性报告。

在更为复杂的水平上, 交互式计算机系统 (interactive computer systems) 说明测验分数的个别化解释。在这种系统中, 个体使用反应站直接接触计算机, 实际上与计算机进行对话 (J. A. Harris, 1973; Holtzman, 1970; M. R. Katz, 1974; Super 等, 1970)。制订教育和职业计划及作出决定时, 都需要这种技术。在这种情况下, 通常把测验分数和学生或患者所提供的其他信息一起, 编入计算机数据库。实质上, 计算机把所有关于个体的信息与存储的关于教育计划和职业的数据相结合; 运用各种有关的事实和关系, 回答个体的问题, 并帮助他作出决定。这种交互式计算机系统的一个实例, 便是交互式指导信息系统 (System for Interactive Guidance Information, SIGI, 1974 ~ 1975)。该系统在学院和大学中使用十多年, 后来经更新和修订, 不仅适用于学生, 而且适用于准备进入或再进入职业市场以及考虑职业改变或发展的成人 (M. R. Katz, 1993; Norris, Schott, Shatkin & Bennett, 1986)。

**危险和准则** 计算机无疑为改进心理测验的各个方面开辟了广阔的道路, 但是, 计算机的某些应用也可能引起对测验分数的误用和误解 (Butcher, 1985a; J. J. Kramer & Mitchell, 1985; Matarazzo, 1983, 1986a, 1986b)。要努力防止这种危险, 就要十分注意制订用计算机进行测验的准则。《测验准则》(AERA, APA, NCME, 1985) 包括用计算机进行测验的若干准则。特别是关于在测验的各个方面使用计算机, 已经制订了一组较为全面的、详

细的准则(例如,参见 Butcher,1987, pp.413 ~ 431)。关于测验中使用计算机的全面评价,尤其关于计算机化分数的解释,参见莫兰德( Moreland,1985,1992)的有关著作。

计算机化测验关注的两个问题是,分数可比性和叙述性解释分数。采用计算机方式和传统的方式实施同一测验时,需要研究分数的可比性 (Mazzeo, Druesne, Raffeld, Checketts & Muhlstein,1991)。除非能够表明两种方式产生完全等同的测验型式,否则,同一组常模不能同时适用于两种方式;测验的信度和效度也会有所变化。尤其重要的是要查对分数可比性,因为不同的个体或团体使用计算机的经验,特别对于计算机化测验的经验,都可能截然不同。

提供叙述性解释分数报告的计算机服务迅速发展,这已经引起特别关注。两条基本原理是许多有关准则的基础。第一,必须提供充分的信息,使测验使用者能够评价在编制软件程序中所应用的解释系统的信度、效度和其他技术特性。如,解释性陈述是怎样从分数中导出的?该系统的理论基础和研究基础又是什么?陈述是从数量分析中导出,还是从专家的临床判断中导出?如果是后者的话,应该给出参与的专家资格的一些信息。

第二,解释性分数报告用于临床或咨询目的时,或者它们以其他方式作出关于个体的重要决定时,必须考虑其他可利用的关于测验参加者的资料来源。由于这种原因,只能由完全合格的专业人员来使用分数报告。应该把分数报告看作为专业人员的辅助工具,而不是他们的代用品。

sonality Inventory, MMPI) 为例, 测验使用者可以得到关于测验参加者的人格倾向和情绪状态的计算机打印结果, 有数字分数, 还有解释性陈述。自己拥有计算机的测验使用者更多地购买计算机程序, 对于特定的测验, 例如韦克斯勒儿童智力量表修订版 (WISC - R) 和成人智力量表修订版 (WAIS - R), 这种程序不仅提供数字分数, 而且提供解释性报告。

在更为复杂的水平上, 交互式计算机系统 (interactive computer systems) 说明测验分数的个别化解释。在这种系统中, 个体使用反应站直接接触计算机, 实际上与计算机进行对话 (J. A. Harris, 1973; Holtzman, 1970; M. R. Katz, 1974; Super 等, 1970)。制订教育和职业计划及作出决定时, 都需要这种技术。在这种情况下, 通常把测验分数和学生或患者所提供的其他信息一起, 编入计算机数据库。实质上, 计算机把所有关于个体的信息与存储的关于教育计划和职业的数据相结合; 运用各种有关的事实和关系, 回答个体的问题, 并帮助他作出决定。这种交互式计算机系统的一个实例, 便是交互式指导信息系统 (System for Interactive Guidance Information, SIGI, 1974 ~ 1975)。该系统在学院和大学中使用十多年, 后来经更新和修订, 不仅适用于学生, 而且适用于准备进入或再进入职业市场以及考虑职业改变或发展的成人 (M. R. Katz, 1993; Norris, Schott, Shatkin & Bennett, 1986)。

**危险和准则** 计算机无疑为改进心理测验的各个方面开辟了广阔的道路, 但是, 计算机的某些应用也可能引起对测验分数的误用和误解 (Butcher, 1985a; J. J. Kramer & Mitchell, 1985; Matarazzo, 1983, 1986a, 1986b)。要努力防止这种危险, 就要十分注意制订用计算机进行测验的准则。《测验准则》(AERA, APA, NCME, 1985) 包括用计算机进行测验的若干准则。特别是关于在测验的各个方面使用计算机, 已经制订了一组较为全面的、详

细的准则(例如,参见 Butcher,1987, pp.413 ~ 431)。关于测验中使用计算机的全面评价,尤其关于计算机化分数的解释,参见莫兰德( Moreland,1985,1992)的有关著作。

计算机化测验关注的两个问题是,分数可比性和叙述性解释分数。采用计算机方式和传统的方式实施同一测验时,需要研究分数的可比性 (Mazzeo, Druesne, Raffeld, Checketts & Muhlstein,1991)。除非能够表明两种方式产生完全等同的测验型式,否则,同一组常模不能同时适用于两种方式;测验的信度和效度也会有所变化。尤其重要的是要查对分数可比性,因为不同的个体或团体使用计算机的经验,特别对于计算机化测验的经验,都可能截然不同。

提供叙述性解释分数报告的计算机服务迅速发展,这已经引起特别关注。两条基本原理是许多有关准则的基础。第一,必须提供充分的信息,使测验使用者能够评价在编制软件程序中所应用的解释系统的信度、效度和其他技术特性。如,解释性陈述是怎样从分数中导出的?该系统的理论基础和研究基础又是什么?陈述是从数量分析中导出,还是从专家的临床判断中导出?如果是后者的话,应该给出参与的专家资格的一些信息。

第二,解释性分数报告用于临床或咨询目的时,或者它们以其他方式作出关于个体的重要决定时,必须考虑其他可利用的关于测验参加者的资料来源。由于这种原因,只能由完全合格的专业人员来使用分数报告。应该把分数报告看作为专业人员的辅助工具,而不是他们的代用品。

## 领域参照测验的解释

**性质和用途** 本世纪 70 年代,尤其在教育界广泛使用的一种测验方法,最初称之为“效标参照测验”。这个术语最先由格拉泽(Glaser, 1963)提出,但迄今为止还没有一个严格的、公认的定义。而且,另外若干不同的术语,例如内容参照、领域参照、目标参照等也同时流行。这些术语有时用作效标参照的同义词,有时又有稍微不同的内涵。此后,较为确切的描述性术语逐渐取代早期的“效标参照”这个名称。出于这种原因,本书从现在起使用“领域参照”(domain-referenced)一词。

典型的领域参照测验的解释参照系,是规定的内容领域而不是规定的人员总体。在这一方面,可以把它与常模参照测验相对照。在常模参照测验中,解释个体分数是把它与在同一测验上其他人的得分相比较。在领域参照测验中,报告测验参加者的成绩,则可以根据他所掌握的算术运算的具体类型、词汇量的估计大小、能够理解的阅读材料的难度水平(从连环画到文学经典)等,或者在外部效标上(教育的或职业的)达到指定成绩的可能性。

迄今为止,领域参照测验主要应用于一些教育改革。它们之中知名的是,计算机辅助教学系统、计算机管理教学系统,以及其他个别化的自定进度的教学系统。在所有这些系统中,测验和教学融为一体,在完成每一教学单元之前、期间和之后,都使用测验来检查必要的技能、诊断可能的学习困难,以及确定后续的教学程序(Nitko, 1989)。

从另一个角度来看,领域参照测验用于大规模的教育成绩



调查,例如‘全国教育发展评定’(National Assessment of Educational Progress)(E. G. Johnson, 1992; Messick, Beaton & Lord, 1983; F. B. Womer, 1970),也用于满足教育成绩责任制的需求。从再一个角度来看,达到最低要求的测验,例如取得驾驶员执照或飞行员执照,也是领域参照测验。另一种相关应用是测验职业水平,即评定是否掌握少量的明确规定的职业技能,例如军事专业的技能(Maier & Hirshfeld, 1978; Swezey & Pearlstein, 1975)。

最后,熟悉领域参照测验的概念,能够有助于教师改进在课堂上使用的自编的、传统的非正式测验。林和格朗伦德(Linn & Gronlund, 1995)出于这种目的,提供了详细的指导,并简单而全面地论述了领域参照测验。埃贝尔(Ebel, 1972)简洁而精辟地分析了领域参照测验的主要局限性;伯克(Berk, 1984a)则全面讨论了领域参照测验的编制和评价中的许多技术问题。

**内容意义** 领域参照测验(不管怎样给它下定义,也不管称之为该名称还是它的另一个同义词)的主要特征是,依据内容意义来解释测验成绩。显然,它强调测验参加者能够做什么,知道什么,而不是他们与别人相比较情况如何。编制这类测验的一项基本要求是,确定测验要评定的知识或技能的领域。要使这类测验上的分数具有交流意义,取样的内容领域必须被公认为是重要的。然后,必须把选出的领域细分为由操作术语所界定的小型单元。在教育上,这些单元就相当于在行为上所界定的教学目标,例如“三位数乘以两位数”或“找出拼错的单词,它在加上-ing时保留了最后的e”。在编制个别化教学的程序中,对于一门学校学科,这些目标可以多达几百个。教学目标确定以后,就编制项目,要在每个目标中都取样。这种程序既困难又费时。然而,假如对内容不加认真规定和控制,则无法解释领域参照测验的结果。一种实际的折中办法是,通过专家评判来界定

关键的概念、原理、方法或教学目标，然后，使用合适的测验项目，对所界定的每个重要领域进行全面取样。毫无疑问，需要评定的行为领域的专门化程度，随着测验的性质和目的而有所变化 (Popham, 1984; Roid, 1984)。

在严格的意义上，领域参照测验最适合于测验初级水平的基本技能（例如阅读和算术）。在这些领域，通常能够以顺序等级来排列教学目标，获得较初级的技能是获得较高水平技能的前提条件。<sup>[7]</sup>然而，在结构性较差的学科中，为高级水平知识确定高度具体的目标，或许既不切实际，也不合需要。在这些水平上，学习的内容和顺序多半十分灵活。

另一方面，领域参照测验在解释测验分数中强调内容意义，一般来说这对测验产生有益的影响。例如，解释智力测验分数就会得益于这种方法。按照具体的智力知识和技能来描述儿童的智力测验成绩，有助于消解传统上对 IQ 认识上的混乱和误解。然而，以这些一般术语来陈述时，领域参照方法相当于按照特定测验的实证效度，而不是按照模糊的基本性质来解释测验分数。当然，这类解释能够和常模参照分数相结合。

**掌握测验 (Mastery Testing)** 领域参照测验的第二个特征是测验掌握水平。本质上，这种测验得出一个全有或全无的分数，表明个体达到或没有达到预先制定的掌握水平。测验基本技能时，一般期望近乎完全的掌握（例如，答对项目为 80 ~ 85%）也可以使用三种级别，即掌握、没有掌握及另一个中间的、不确定的‘再审查’组。

一些教育家认为，在个别化教学中，如果提供适当的时间和教学方法，几乎每一个人都能完全掌握所制定的教学目标。因此个体差异表现在学习时间上，而不是像传统的教育测验那样，表现在最后成绩上 (Carroll, 1963, 1970; Cooley & Glaser, 1969,

Gagne, 1965)。由此推断,在掌握测验中,成绩的个体差异并不重要。因此,正如一般人设想的那样,在合适的训练之后,领域参照测验使成绩的个体差异缩至最小。掌握测验经常用于前面所提到的个别化教学程序。它也是所出版的适用于小学基本技能的领域参照测验的特征。

编制这类测验的两个重要问题是:(1)为了可靠地评定测验所包括的每个具体的教学目标,必须使用多少项目?(2)可靠地确定达到掌握水平,必须答对多少比例的项目?在许多早期领域参照测验中,凭主观判断来回答这两个问题。然而,在通过发展合适的统计技术以提供客观的经验回答方面,已经取得重大进展(Berk, 1984a, R. L. Ferguson & Novick, 1973; Hambleton, 1984a, 1989; Hambleton & Novick, 1973)。我们将举例说明这些工作的性质和范围。

关于项目数量和临界分数两个问题,能够合并为一个假设,在决策理论和序列分析的框架之内进行测验(Hambleton, 1984a; Wald, 1947)。我们特别希望检验这个假设:在测验项目所取样的内容领域或教学目标中,测验参加者已经达到所要求的掌握水平。序列分析包括,每次取得一个观察,每个观察之后决定或(1)接受假设,或(2)拒绝假设,或(3)继续观察。因此,得到可靠的结论所需要的观察数(此处为项目数)本身在测验过程中确定。不是预先提出一个固定的项目数,而是让被试继续测验,直到确定他掌握或没有掌握为止。这时,测验不再继续,学生转到下一个教学水平或回到没有掌握的水平进一步学习。使用本章前面所描述的计算机设备,这类序列决策程序是切实可行的,并且能够减少测验总时间而可靠地估计掌握水平。

一些研究人员一直在探索贝叶西(Bayesian)评定技术的使用,该技术把一些间接资料结合起来,十分适用于掌握测验所需

要的这类决策。由于要测验大量具体的教学目标，领域参照测验一般只为每个目标提供少量项目。为了弥补这种信息有限的不足，已经发展出各种程序，以便使用学生以前的成绩记录和其他学生的测验结果等间接资料（R. L. Ferguson & Novick, 1973; Hambleton, 1984a; Hambleton & Novick, 1973）。

当个别化程序不切实际时，通过分析各合适团体在某测验上教学前后的分数，能够在经验上确定临界分数。所选择的临界分数，最能鉴别接受和没有接受有关训练的团体（Panell & Laabs, 1979; L. A. Shepard, 1984）。在具体情境中也需要判断让不合格者‘通过’与让合格者‘失败’的相对严重程度。可以相应提高或降低临界分数，以便调整错误分类后果的严重性。

和常模参照测验的关系 除了基本技能，掌握测验就是不适用的或不充分的。在水平较高、结构性较低的学科中，成就是无止境的。在诸如理解、批判性思维、欣赏、创造力之类机能中，个体几乎可以无限制地长进。此外，内容范围涉及许多不同的方面，这取决于个体的能力、兴趣、目标以及当地的教学设施。在这种条件下，完全掌握就是不现实的，也是不必要的。因此，在这种情况下，一般使用常模参照测验来评定达到的程度。一些出版的测验编制得十分灵活，准许同时应用常模参照和领域参照。斯坦福阅读诊断测验和数学诊断测验便是一例。这种测验提供每一水平的合适常模，同时又准许对儿童达到详细的教学目标进行定性分析。

应当指出，领域参照测验并非像一些支持者认为的那样，是一种新型测验，或完全脱离常模参照测验。以字母等级或项目百分数等绝对术语来评价个体的测验成绩，当然比常模解释还要陈旧。按照内容意义较为确切地描述测验成绩的尝试，也先于采用“效标参照测验”一词（Ebel, 1962; J. C. Flanagan, 1962;

Nitko, 1984, pp. 14 ~ 16)。在早期作品量表中,也可以找到其他的例子,如把个体的作业样本和一组标准样本相比较,来评价书写、作文、绘画等的质量。而且,埃贝尔(1972)注意到,教育‘掌握’这一概念——在具体单元的全或无学习的意义上——在本世纪 20 至 30 年代相当流行,但后来却被抛弃了。

不管怎样表示分数,所有测验都内含常模参照系 (Angoff, 1974; Nitko, 1984)。主试的这种认识——预期人类有机体在特定的发展阶段或教学阶段能够做什么——影响所要测量的内容或技能的选择。这种选择预先假定具有其他个体在类似情境中已经做了什么的资料。此外,掌握测验并没有因为把统一的临界分数强加于能力连续体之上而消除个体差异。把个体的阅读理解水平描述为‘理解《纽约时报》内容的能力’,这仍然存在着大范围的个体理解程度的差异。运用临界点将成绩简单地分为两类,这不仅忽视了在两类之内仍然存在的个体差异,而且还抛弃了可能有用的资料。

## 最低资格和临界分数

**实际需要和缺陷** 领域参照测验中,掌握这一概念仅仅是作出决策中实际使用临界分数的一个例子。对于日常生活中多种目的,必须制定和执行最低资格。在许多情况下,例如颁发驾驶员执照、选拔飞行员、原子能工厂招聘员工等,出于安全考虑,需要设立作业的临界点。在教育中,通过课程或毕业也属这类情况,要求对个体作出全有或全无的分类 (Jaeger, 1989)。在临床和咨询实践中,关于治疗和行为建议的决策,也要求同样的判断。

坚决赞成使用临界分数的理由是，某种机能的作业要求一些关键性变量。这些变量的能力不足，不能由其他变量的优秀能力来补偿。在这种情况下，综合选拔成套测验上的高分，会掩盖关键性技能上的不足。在关键性技能上处于所要求的最低资格以下的个体将被淘汰，而不管他们的其他能力如何。例如，声纳设备操作员必须具有优秀的听觉鉴别力。在第二次世界大战期间，美国海军根据在听觉鉴别测验和机械理解测验上的联合分数来选择新兵，培训他们当声纳操作员。结果，在机械理解上优秀但恰好在关键的听觉技能上不足的一些大学男生，被分配去接受这种培训，后来却失败了。标准海军程序要求在第一项培训任务中的失败者，作为海军三等兵转到一般的海军勤务，他们由此失去潜在的专业化的工作。进一步分析情况后，出于这种选拔目的，及时改用临界甄别程序。然而，大多数同工作相关的变量与工作成绩的关系往往是线性的，所以测验分数越高，工作成绩也越好(Coward & Sackett, 1990)。在这种情况下，比起个体相对于临界点的位置，他的实际分数倒是较好的预测因素。

在许多实际决策中不可避免地会使用临界分数，因此必须意识到这类评价中的缺陷，并且使用各种方法减少判断误差。例如，应该努力减少单一测验分数的局限。条件可行的时候，临界分数应该是一组分数，而不是实施一次特定测验所得出的单一分数。此外，关于个体决策应该取决于每一个体的多种资料，其中的测验分数，要由过去和现在的其他有关作业数据来补充。如果测验上的临界点由一组评判员来制定，该组应该包括合适的专家代表，既有作业的相关领域的专家，又有测验编制和使用原理方面的专家。最为重要的是，如有可能，应该根据经验数据来制定或检验临界分数。这特别意味着，测验分数应该从各种团体中得出，他们在相关的效标行为上，例如在某类工作的实际

成绩上大不相同。当然,人们设计测验要预测的正是这种实际成绩,又设计临界分数来保证一种安全的、可接受的、合乎需要的最低资格。下面一节所讨论的期望表,清楚说明在人员选拔上制定临界分数的一种经验方法。

**期望表** 解释测验分数意义的方法之一,是根据个体的期望效标成绩,例如在训练课程中或在工作上的成绩来解释测验分数的。“效标”一词的这种用法遵照标准的心理测量实践,例如说到一个测验以某特定效标来建立效度(参见第一章)。严格地说,“效标参照测验”应该指对这类成绩的解释,而前面一节所讨论的其他方法则能够较为确切地描述为内容参照或领域参照。

期望表对于得到每一测验分数的个体,给出获得不同效标结果的概率。例如,如果一个学生在大学入学考试委员会的学术评定测验(SAT)上得 530 分,他在一所具体大学的一年级平均成绩处于 A、B、C、D、F 等类的可能性各是多少?考察预测分数(SAT)对效标成绩(一年级平均成绩)的双变量分布,就能获得这类信息。如果把这种双变量分布的每个单元的人数转换成百分比,结果就是期望表,如表 3-6 所示。该表的数据从 211 个学习数学课程的七年级学生中得出。预测分数是第一学期末所实施的“区别能力倾向测验”中的数推理测验,效标则是第二学期最后一次考试的课程成绩。测验分数和效标的相关是 0.60。

表 3-6 第一列表示测验分数,分为 4 组;第二列表示每一分数组的学生人数。表中每一行的其余项目,则表示每一测验分数组在课程结束时得到的每一成绩的学生人数百分比。因此,在数推理测验上得 30 分及以上的 22 个学生中,5%得 D 及以下,没有人得 C,36%得 B,59%得 A。另一方面,在数推理测验得 10 分以下的 14 个学生中,43%得 D 及以下,36%得 C,14%

得 B。异常的 7%得 A,只表示 1 个学生,这对于概化目的实质上是无用的资料。测验分数得 30 分及以下学生中的 5%得 D 及以下,也只表示 1 个学生,情况同样如此。然而,在所得数据的范围之内,表 3-6 的百分比估计一个学生得到某一效标成绩的概率。例如,如果一个新生得到 DAT 数推理 24 分(即在 20~29 分之内),我们会推断出,他得到课程成绩 A 的概率是 27%,得到成绩 B 的概率是 43%,等等。

表 3-6 期望表,它表示 211 个七年级学生 DAT 数推理  
测验与数学课程成绩之间的关系

| 测验分数   | 人数  | 得到每一成绩的百分比 |    |    |    |
|--------|-----|------------|----|----|----|
|        |     | D 及以下      | C  | B  | A  |
| 30 及以上 | 22  | 5          | 0  | 36 | 59 |
| 20—29  | 104 | 9          | 21 | 43 | 27 |
| 10—19  | 71  | 37         | 37 | 24 | 3  |
| 10 以下  | 14  | 43         | 36 | 14 | 7  |

(摘自《区别能力倾向测验技术手册》,第 5 版,p.152)

在许多实际情况下,能够把效标分成工作、学习课程或其他任务中的‘成功’和‘失败’两类。在这种情况下,能够绘制期望图,它表示相当于每个分数组的成功或失败的概率。图 3-7 便是这种期望图的一个例子。根据美国空军所编制的飞行员选拔成套测验,该期望图表示在成套测验上得到各个标准九分而不能完成初级飞行训练的实习飞行员的百分比。能够看到,标准九分为 1 分的学员中的 77%在训练课程中遭淘汰,而标准九分为 9 分的学员中只有 4%不能圆满完成训练。在这两个端点之



间,失败者的百分比随着标准九分的逐次提高而不断减少。根据这个期望图能够预测,例如,标准九分为 4 分的学员中,近似 40% 不能完成初级飞行训练,近似 60% 则能圆满完成。对于得到各个标准九分的个体,都能够作出关于成功和失败概率的类似陈述。因此,标准九分为 4 分的个体,完成初级飞行训练的概率是 60:40 即 3:2。能够看到,期望表和期望图除了提供测验分数的效标参照解释之外,也给出预测某一效标的测验效度的一般概念。由于这种原因,在第六章的结尾即关于公平使用测验的决策模式一节中,较为详细地讨论了制定临界分数的经验程序。该节也讨论在各种条件下,确定最理想的临界分数的数学程序。第十七章则讨论临界分数在主要的心理学实践领域中的具体应用。

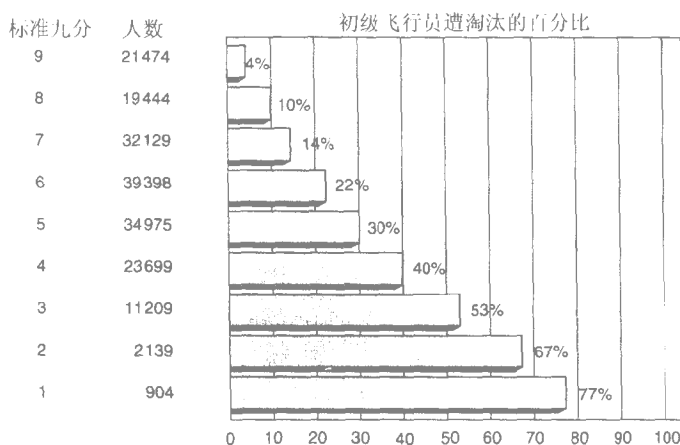


图 3-7 期望图, 它表示飞行员选拔成套测验的成绩与初级飞行训练遭淘汰之间的关系 (摘自 J.C. 弗拉纳根, 1947, p.58)。

## 注 释

- [1] 本章中所说明的计算法讨论描述统计 (descriptive statistics), 表示实际所测量的样本; 在推断统计 (inferential statistics) 中, 用  $N-1$  来代替, 以便从样本数据估计相应总体的值。样本越小, 样本和总体值之间的差异就越大。关于解释, 参见当前任何一本统计课本。
- [2] “顺序量表”一词的这种用法不同于统计学, 这里的顺序量表仅仅指个体的排列顺序, 而不知道他们之间的差异量。在统计的意义上, 顺序量表相对于等距量表。设计儿童发展的顺序量表实际上根据一种格特曼 (Guttman) 量表即单向模式, 在一种水平上的成功作业意味着在所有较低水平上的成功 (L. Guttman, 1944)。巴特和艾拉赛恩 (Bart & Airasian, 1974) 描述了包括非线性等级的格特曼分析的扩展, 尤其是关于皮亚杰量表。
- [3] 关于皮亚杰方法的进一步评价, 参见第九章。
- [4] 凯泽 (Kaiser, 1958) 修改了标准九分量表, 它的百分数稍有变化, 标准差恰好为 2, 因此便于定量处理。标准化九分量表的其他变式有 C 量表 (Guilford & Fruchter, 1978, pp. 484 ~ 487), 它由 11 个单位组成, 标准差也为 2 还有 10 个单位的标准十分量表, 5 个单位在平均数之上, 另 5 个单位在平均数之下 (Canfield, 1951)。
- [5] 20 世纪末, 为了反映测验分数的性质发生变化的趋向, 这项测验后来改名为《学术评定测验》(Scholastic Assessment Test) (特别参见第十二章关于测验操作经验史中个体差异的影响)。
- [6] 感谢大学入学考试委员会韦恩·卡马勒 (Wayne Camara) 的帮助, 使我们得到这里所报告的资料。
- [7] 理想上, 这类测验采用格特曼量表的单一性模型 (参见 Popham & Husek, 1969), 如同第九章所讨论的皮亚杰顺序量表。

## 第四章 信度

“信度”指在不同时间,使用同一测验,或者使用两个不同项目的等值测验,抑或在其他不同的测试条件下,对同一组被试实施两次测验所得分数的一致性。信度这一概念是计算各个分数的测量误差 (error of measurement) 的基础,据此我们能够预测各个分数受到无关因素或未知的偶然因素的影响而可能发生的波动范围。

信度概念用于分数一致性的若干方面。在最广泛的意义上,测验信度表示测验分数中个体差异可归因于所考虑的属性中“真实”差异的程度,以及可归因于偶然误差的程度。用专门术语来说,测验信度的度量,使得估计测验分数的总方差中误差方差 (error variance) 的比例成为可能。在通过改进方法而能够避免或校正的意义上,它们称不上“误差”。误差这一术语沿袭自早期心理学,当时人们的兴趣集中于发现普遍的行为规律,并且根据所假设的一成不变的基础特质来评定个体。今天,心理学家认识到变异性是所有行为的本质属性,并且试图对变异性的多种来源进行研究和分类。

分数信度的关键在于确定误差方差。对于一种目的来说,被认为是误差方差的因素,对于另一种目的来说,则可能属于真实方差,例如,如果我们有意测量心境的波动,那么,在快乐—沮丧测验上每天的分数变化与测验目的有关,因此便是分数的

实方差的组份。另一方面，如果我们设计测验来测量较持久的人格特征，同样的每天波动则属于误差方差了。

本质上，与测验目的无关的任何条件都影响误差方差。因此，当测验者控制测验环境、指导语、时限、和睦关系及其他类似因素，试图保持统一的测验条件时，他们正在减少误差方差，使得测验分数更为可靠。然而，尽管在最佳测验条件下，也没有一种测验是完美可靠的工具。因此，应该说明每一个测验的信度。在标准条件下，对类似于标准化样本的被试实施测验时，信度的这种度量就是测验的特性。因此，应该详细说明该样本的特性和所测量的信度类型。

当然，有多少种影响测验分数的条件，就有多少种测验信度，因为任何一种条件都可能与某种测验目的无关，因而属于误差方差。然而，在实际工作中所计算的信度类型相对不多。本章将讨论测量测验分数信度的主要方法和每种方法所确定的误差方差的来源。<sup>[1]</sup>

各种类型的信度都关注两组独立导出分数之间的一致性程度，所以它们都能够用相关系数（correlation coefficient）来表示。因而下一节将考虑相关系数的一些基本特性，以便讲清楚它们的使用并作出解释。在任何一本教育或心理统计的基础教材中，例如鲁尼恩和哈伯（Runyon & Haber, 1991）或豪厄尔（D. C. Howell, 1997）的著作中，都能够找到对相关的专门讨论，以及对计算程序的详细说明。

## 相关系数

**相关的含义** 本质上，相关系数( $r$ )表示两组分数之间的对

应程度或关系。因此,如果在变量 1 上得最高分的个体,在变量 2 上也得最高分,在变量 1 上得第二名的个体,在变量 2 上也得第二名,以此类推,直到这组中最差的个体,那么,变量 1 和变量 2 之间存在着完全相关。这类相关的数值为 1.00。

图 4-1 说明一个假设的完全正相关的例子。该图系一种相关散布图即双变量分布。图中每个计数符号表示一个个体在变量 1(水平轴)和变量 2(垂直轴)上的分数。应当指出,团体中所有 100 个个案,沿着从图的左下角到右上角的对角线而分布。这类分布表示完全正相关 (+1.00),因为它表示每一个体在两个变量上处于同一相对位置。双变量分数的分布越接近这条对角线,正相关就越高。

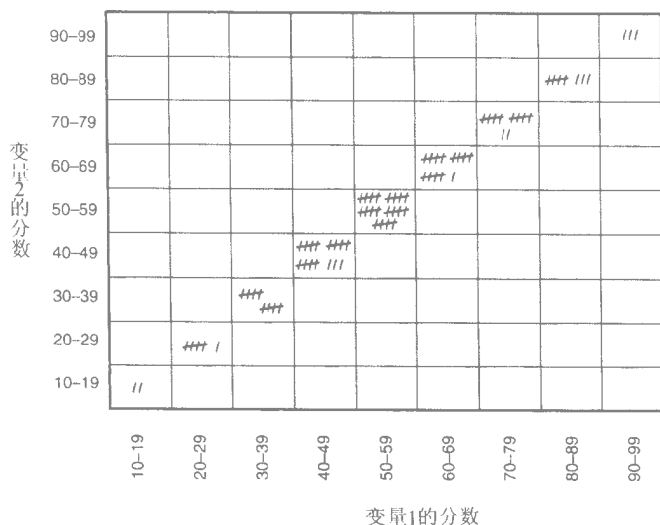


图 4-1 一个假设的相关为 +1.00 的双变量分布

图 4-2 说明一个完全负相关 ( $-1.00$ ) 的例子。在这种情况下, 从一个变量到另一个变量, 分数完全相反。在变量 1 上最好的个体, 在变量 2 上倒是最差的, 反之亦然。这种相反在整个分布上保持一致。应当指出, 在这个相关散布图中, 所有个体处于从左上角到右下角的对角线上。这条对角线和图 4-1 中的对角线方向相反。

87 零相关表示完全没有关系, 这样的情况偶尔也会发生。假如从盒子中随机抽出某一个体的名字, 以确定他在变量 1 上的位置, 对于变量 2 也重复这个过程, 结果就可能是零相关或接近零相关。在这种条件下, 知道某个体在变量 1 上的分数, 也不可能预测他或她在变量 2 上的相对位置。在变量 1 上得最高分的个体, 在变量 2 上可能得高分, 可能得低分, 也可能得平均分。

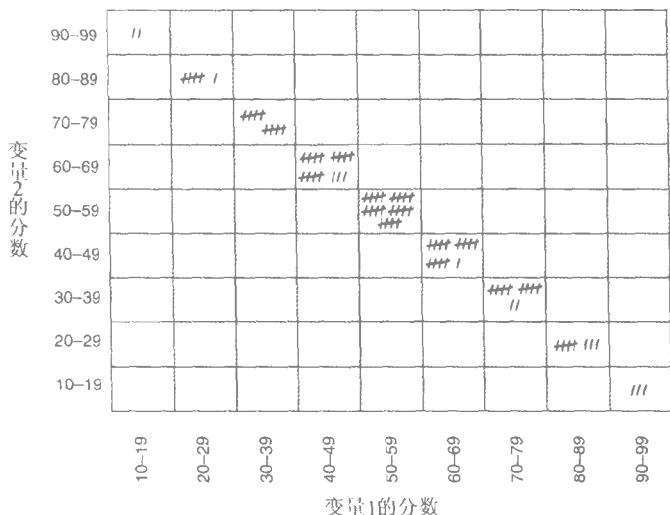


图 4-2 一个假设的相关为  $-1.00$  的双变量分布

一些个体在两个变量上的得分可能都在平均数之上,或都在平均数之下;另一些个体可能在一个变量上处于平均数之上,而在另一个变量上处于平均数之下;还有一些个体可能在一个变量上处于平均数之上,而在另一个变量上则为平均数,这些都纯属偶然。个体之间的关系没有规律可循

实际所得出的相关系数一般在这些极端之间,它们的数值大于零但小于 1.00。能力测量之间的相关几乎总是正的,虽然常常较低。当两个能力变量之间得出负相关时,通常是表示分数的方式所造成的。例如,如果计算时间分数和数量分数的相关,就可能得出负相关。因此,如果把每一个体在算术计算测验上的分数,记录为完成所有项目所需要的时间数,而在算术推理测验上的分数,则是正确解答的问题数,就可能得出负相关。在这种情况下,最差(即最慢)的个体会在第一个测验上得到数字上的最高分,而最好的个体则会在第二个测验上得到最高分。

相关系数可以用不同方法计算,这取决于数据的性质。最为普通的是皮尔逊积差相关系数(Pearson Product-Moment Correlation Coefficient)。这种相关系数不仅考虑个体在团体中的位置,而且考虑他在团体平均数之上或之下的离差大小。回忆一下,用标准分数表示每一个体的位置时,处于平均数之上的个体得到正标准分数,而处于平均数之下的个体则得到负标准分数。因此,在两个相关的变量上都较好的个体,会得到两个正标准分数;在两个相关的变量上都较差的个体,则会得到两个负标准分数。

现在,假如每一个体在两个变量上都处于平均数的同一边,那么,我们把每一个体在变量 1 上的标准分数,乘以他在变量 2 上的标准分数,所得到的乘积都是正的。皮尔逊相关系数只是这些乘积的平均数而已。在两个变量上的对应的标准分数符号相同,并且大小近似相等时,皮尔逊相关系数为最高的正值。在

表 4-1 皮尔逊积差相关系数的计算

| 学生  | 数学<br>X   | 阅读<br>Y   | x         | y         | x <sup>2</sup> | y <sup>2</sup> | xy       |
|-----|-----------|-----------|-----------|-----------|----------------|----------------|----------|
| 比尔  | 41        | 17        | +1        | -4        | 1              | 16             | -4       |
| 卡罗尔 | 38        | 28        | -2        | +7        | 4              | 49             | -14      |
| 杰弗里 | 48        | 22        | +8        | +1        | 64             | 1              | 8        |
| 安   | 32        | 16        | -8        | -5        | 64             | 25             | 40       |
| 鲍勃  | 34        | 18        | -6        | -3        | 36             | 9              | 18       |
| 简   | 36        | 15        | -4        | -6        | 16             | 36             | 24       |
| 埃伦  | 41        | 24        | +1        | +3        | 1              | 9              | 3        |
| 鲁思  | 43        | 20        | +3        | -1        | 9              | 1              | -3       |
| 迪克  | 47        | 23        | +7        | +2        | 49             | 4              | 14       |
| 玛丽  | <u>40</u> | <u>27</u> | <u>-0</u> | <u>+6</u> | <u>0</u>       | <u>36</u>      | <u>0</u> |
| Σ   | 400       | 210       | 0         | 0         | 244            | 186            | 86       |
| M   | 40        | 21        |           |           |                |                |          |

$$SD_x = \sqrt{\frac{244}{10}} = \sqrt{24.40} = 4.94 \quad SD_y = \sqrt{\frac{186}{10}} = \sqrt{18.60} = 4.31$$

$$r_{xy} = \frac{\sum xy}{(N)(SD_x)(SD_y)} = \frac{86}{(10)(4.94)(4.31)} = \frac{86}{212.91} = 0.40$$

一个变量上处于平均数之上的个体，在另一个变量上处于平均数之下时，对应的交叉乘积将是负的。如果交叉乘积的加和是负的，相关就是负的。当一些乘积为正，另一些乘积为负时，相关就接近于零。

在实际工作中，得出交叉乘积之前，不必把每个原始分数都转换成标准分数，因为把交叉乘积相加之后，能够一次性地进行这种转换，计算皮尔逊相关系数有许多捷径。表 4-1 中所说明的方法并不是最快的，但比起其他使用计算捷径的方法，它比较清楚地说明相关系数的含义。表 4-1 说明 10 个儿童的数学分数和阅读分数之间的皮尔逊相关  $r$  的计算。每个儿童的名字后



面,是他在数学测验( $X$ )和阅读测验( $Y$ )上的分数。在每列的下面给出 10 个分数的加和与平均数。第三列是每个数学分数与数学平均数的离差( $x$ );第四列是每个阅读分数与阅读平均数的离差( $y$ );后面两列是这些离差的平方。按照第三章所描述的方法,使用平方和来计算数学分数和阅读分数的标准差。不是把每个  $x$  和  $y$  除以对应的  $SD$  来得出标准分数,我们只是在最后一次性进行这种除法,如表 4-1 中相关公式所示。把  $x$  列和  $y$  列中对应的离差相乘,就得出最后一列中的交叉乘积 ( $xy$ )。为了计算相关( $r$ ),把这些交叉乘积加和,除以个案数( $N$ ),再除以两个标准差的乘积( $SD_xSD_y$ )。

统计显著性表 4-1 中所得出的 0.40 的相关,表示数学分数和阅读分数之间中等的正向关系。虽然这种关系不是十分密切,但在数学测验上做得好的那些儿童,往往也在阅读测验上做得好,反之亦然。如果仅仅关注这 10 个儿童的成绩,我们可以认为,这个相关合适地描述了该团体在两个变量之间存在的关系的程度。然而,在心理学研究中,我们通常关注的是,超出所测验的个体的特定样本而概化到他们所代表的更大的总体。例如,我们想知道,像我们所测验的同龄美国儿童中,数学测验和阅读测验是否相关。显然,实际所测验的 10 个案例是这个总体的一个极不合适的样本。同样大小的另一个可比较的样本,可能得出低得多或高得多的相关。

有一些统计程序,可以估计相关、平均数、标准差及其他团体测量的数值在样本之间可能的波动。然而,通常所提出的关于相关的问题,只是相关是否显著大于零。换句话说,如果总体的相关为零,与我们的样本中所得出的一样大小的相关,是由于取样误差造成的吗?当我们说相关的“显著性水平是 1%(0.01)”时,我们是指总体相关为零的可能性不会大于 1%。因此,我们得出

结论,这两个变量确实相关。显著性水平指从数据推出结论中我们愿意承担的错误风险。如果说相关的显著性水平为 0.05,错误的可能性就是 5%。大多数心理学研究采用 0.01 或 0.05 水平,虽然出于特殊原因,也可以采用其他的显著性水平。

表 4-1 中所得出的 0.40 的相关值,甚至没有达到 0.05 的显著性水平。我们可以预测到这一点,只使用 10 个案例,难以得出结论性的一般关系。使用这样大小的样本,达到 0.05 显著性水平的最小相关值为 0.63。低于那个数值的任何相关,都不能证明在从中抽取样本的总体中,两个变量是否相关。查阅统计课本中的相关显著性表,可以找到在 0.01 和 0.05 显著性水平对于不同大小团体的最低相关。然而,本书出于解释目的,仅仅需要了解一般概念。

多年来,显著性水平一直是评价相关的传统方法。然而,现在日益意识到这种方法的不足之处。说明信度系数(即相关)显著大于零,这对于理论或实践目的,都没有提供什么信息。尤其在小样本上得出时,甚至高相关也没有满足“显著性检验”。另一种替代方法正日益引起关注,它考虑所得出相关的实际大小,并且估计置信区间(confidence interval),总体相关可能处于其中一个指定的置信水平(例如,参见 Carver, 1993; J. Cohen, 1994; Hunter & Schmidt, 1990; Olkin & Finn, 1995; Schmidt, 1996; W. W. Tryon 1996)。这种用置信区间补充显著性水平——如果不是替代的话——的趋向,预示着在未来的相关系数分析中的重要变化。

**信度系数** 相关概念在分析心理测量数据中有多种用处。测量测验的信度便是相关系数的一种应用。图 4-3 给出用皮尔逊积差方法计算信度的一个例子。在这种情况下,104 个个体在词汇流畅测验<sup>[2]</sup>两个等值型式上的分数相关。在一种型式上,给测验参加者 5 分钟,尽可能多地写出以一个给定的字母开

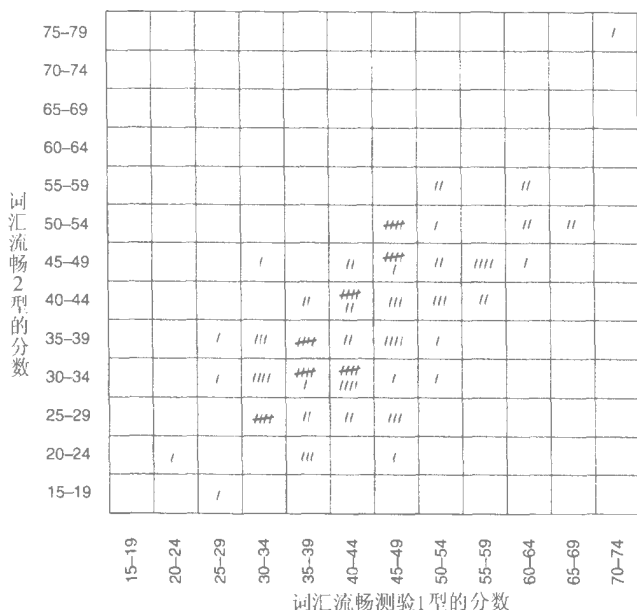


图 4-3 信度系数为 0.72(资料摘自安娜斯塔西和德雷克,1954)

头的单词。第二种型式完全相同,除了使用一个不同的字母之外。出于这种目的,测验编制者选择两个难度近似的字母。

在这个测验的两个型式中,写出的单词数之间的相关为 0.72。这个相关是高的,达到 0.01 显著性水平。使用 104 个案例,0.25 及以上的相关都达到这个显著性水平。然而,得出的相关略微低于信度系数所需要的相关,它们通常处于 0.80 ~ 0.90。图 4-3 中的相关散布图,表示高度正相关分数的典型双变量分布。应当指出,记数符号集中于从左下角到右上角的对角线附近;虽然存在着个别记数符号的某种散布,在这个方向上

的趋势则是明确的。下面一节将考虑使用相关系数来计算不同类型的测验信度。

## 信度的类型

再测信度 (test-retest reliability) 得出测验分数信度的最易理解的方法,是第二次重复实施完全相同的测验。同一组被试在两次测验上所得分数之间的相关,便是再测信度系数 ( $r_{tt}$ )。误差方差相当于两次测验成绩的随机波动。这种变化可能部分由于未加控制的测验条件而产生,例如天气剧烈变化、偶然的噪音及其他干扰,如铅笔头折断等。然而,在某种程度上,是由于测验参加者本身条件的变化而产生,例如疾病、疲劳、情绪紧张、忧虑、最近心情的愉快与否,等等。再测信度表示能够把一次测验上的分数概化到不同时间的程度;信度越高,分数受到测验环境或测验参加者条件中的日常随机变化的影响就越小。

在测验手册中报告再测信度时,总是应该说明再测间隔的时间。因为再测相关随着间隔时间的增长而逐渐减小,任何一个测验的再测信度系数都不是一个而是无限个。也需要说明再测样本的有关的中间经历,例如教育经历或工作经历、咨询、心理治疗,等等。

除了需要报告间隔时间之外,选择间隔时间还应该考虑什么?我们很容易引证,一些测验在几天或几周中具有高信度,但是,当间隔时间长达 10 年或 15 年时,测验分数就几乎完全失去一致性。例如,许多学前儿童智力测验,在学前期得出比较稳定的量数,但是要预测青少年期或成人智商,事实上就无能为力了。然而,在实际工作中,通常能够作出简单的区分。测验分数

的误差方差中,一般包括几小时到几个月间隔期产生的短期的、随机的波动。因此,检查这类测验信度时,要尽力使时间间隔短些。测验幼儿的时间间隔应该比成人短,因为对幼儿来说,一个月甚至更短的时间中,生长发展的变化是很大的。对于任何个体,再测的间隔时间一般不应该超过 6 个月。

较长时间间隔中发生的个体相对测验成绩的另外变化,可能是累积的、逐渐的,而不是完全随机的。此外,它们可能表示比测验成绩本身所包括的范围更大的行为的属性。因此,由于不同寻常的中间经历,人们的学术能力倾向、机械理解、艺术判断等一般水平在 10 年时间中可能发生明显的变化。由于环境特别是个体自己的家庭、学校、社区环境,或者由于疾病、情绪障碍之类原因,个体相对于其他同龄人的地位,可以明显上升或下降。

这类因素能够影响个体心理发展的程度,这是一个重要的课题。然而,这个问题不应该混同于特定测验的信度。例如,当测量斯坦福一比纳量表的信度时,我们计算再测分数相关的间隔一般不是 10 年,甚至也不是一年,而是几个星期。诚然,对这类测验也进行长期再测,但是,一般认为这种结果不是特定测验的信度,而是预测度,即从儿童期成绩预测成人智力。信度这一概念一般限于短期的、随机的变化,它表示测验本身的属性,而不是所测验的整个行为领域的属性。

应当指出,不同的行为机能在它们所表现的日常波动的程度上,本身就有所不同。例如,灵巧的手指运动的稳定性,无疑比言语理解容易受到个体条件的微小变化的影响。假如希望全面评价个体的惯常的手指稳定性,我们也许需要在数天时间里每天重复测验,而对于言语理解,一次测验就足够了。因此,我们必须分析测验目的和透彻了解测验所要预测的行为。

虽然再测方法看来简单易懂,但是应用于大多数心理测验

时，它也出现了种种困难。练习对于不同个体的再测分数，也许会产生不同数量的提高。此外，如果再测间隔时间太短，测验参加者可能回忆起许多他们前一次的回答。换句话说，同样型式的对错回答完全可能通过记忆而再次发生。因此，两次测验上的分数不是相互独立的，它们之间会出现假性高相关。测验本身的性质也可能由于重复而改变，推理问题的情况尤为如此。一旦测验参加者掌握了问题中所包括的原理或作出了答案，他或她在将来能够再次作出正确回答而不需要经过中间的步骤。只有不大受重复影响的测验，才适用于再测方法。一些感觉辨别和运动测验就属于此类。然而，对于大多数心理测验来说，使用完全相同的测验来再测，这不是得出信度系数的合适方法。

复本信度(alternate-form reliability)使用测验复本，可以避免再测信度中所碰到的一些困难。因此，能够在第一次使用一种型式，而在第二次使用另一种等值型式，对同一组被试进行测验。两种型式上得出的分数之间的相关，便是测验的信度系数。应当指出，这种信度系数既测量时间稳定性，又测量对不同的项目样本（即测验型式）的回答的一致性，因而把两种类型的信度结合起来。然而，两类信度对于大多数测验目的都是重要的，所以，复本信度为评价许多测验提供了有用的度量。

项目取样即内容取样这一概念，不仅是复本信度的基础，而且也是所讨论的其他信度的基础。因而应该仔细考虑这个概念。许多学生都有过参加学科考试的体会，也许有时他们感到碰上“好运气”，因为许多项目他们碰巧复习得最认真而有时恰恰相反，所考的大量项目他们没有复习过。这种熟悉的情境说明内容取样所造成的误差方差。本测验上的分数取决于特定的项目选择的程度究竟如何？假如另一位独立工作的研究人员，按照同样的规定编制另一份测验，那么被试在这两个测验上的

分数究竟会相差多大？

让我们假设,编制一份 40 个单词的词汇测验,测量一般的言语理解。现在又假设,编制第二份 40 个不同单词的测验,测验目的相同。两个测验的难度分布范围也相同。同一组被试在这两个测验上所得分数的差异,则说明这类误差方差。由于不同个体过去经验中的偶然因素,这两个测验上的相对难度,对于不同个体就有所不同。因此,被试 A 可能在第一个测验上有较多的生词,而被试 B 则可能在第二个测验上有较多的生词。如果这两名被试的词汇量(即‘真分数’)大致相同,B 还是会在第一个测验上超过 A,而 A 则会在第二个测验上超过 B。由于项目选择中的偶然差异,因此这两个被试在两个测验上的相对位置相反。

像再测信度一样,复本信度总是应该同时说明测验实施的时间间隔,也应该说明有关的干预情况。如果两个复本连续实施,得出的相关只表示跨型式的信度,而不表示跨时间的信度。这种情况下的误差方差指两组项目成绩中的波动,而不是时间上的波动。

当然,编制复本时应该注意,保证它们确实是平行测验。十分重要的是,一个测验的两个平行型式应该是符合相同规定、独立编制的测验。平行测验应该包括相同数目的项目,项目的表述型式应该相同,项目的内容类型应该相同,项目难度的水平和范围也应该相同。此外还应该检查指导语、时限、例题、型式以及测验的所有其他方面是否相同。

还应当指出,除了确定测验信度之外,出于其他原因,也需要平行的测验型式。在追踪研究或研究某种实验干预因素对测验成绩的影响方面,复本是十分有用的。使用复本也可以减少辅导或欺骗的可能性。

复本信度的应用比再测信度广泛得多,但是,它也有某些局限性。首先,如果所考虑的行为机能受到练习的影响很大,则使用复本只能减少但不能消除这种影响。当然,假如所有测验参加者由于重复而出现同样的提高,他们的分数之间的相关仍然不受影响,因为每个分数加上一个常量,不会改变相关系数。然而,情况多半是由于先前练习类似材料的程度、参加测验的动机以及其他因素,被试分数的提高量有所不同。在这种情况下,练习的影响就产生另一种方差,它往往减少两个测验之间的相关。如果练习的影响不大,则这种减少可以忽略不计。

另一个有关的问题是,测验的性质会由于重复而有所改变。例如,在某些推理问题中,大多数被试一旦作出答案,就能够很容易地解答属于同样原理的项目。在这种情况下,第二种型式中项目具体内容的改变,不足以消除来自第一种型式的迁移。最后,还应当指出,编制真正的等值测验时实际困难重重,所以,许多测验没有复本。由于所有这些原因,常常需要使用其他方法来估计测验信度。

分半信度(split-half reliability)使用一种测验型式,实施一次测验,也可能采用种种分半方法而得到信度量数,在这类方法中,把测验分成两个最为等值的一半,每一被试就得到两个分数。显然,分半信度提供内容取样一致性的量数。因为只实施一次测验,这种信度就不包括分数的时间稳定性。因为只需要对一种型式实施一次测验,有时也称之为内部一致性系数。

要算出分半信度,第一个问题是怎样把测验分成两个最为等值的一半。把一个测验分为两半有许多不同的方法。大多数测验的前面一半和后面一半不是等值的,由于项目的性质和难度水平有所不同,而且,准备状态、练习、疲劳、厌倦等因素的累积影响,从测验的开始到结束也有所变化。在大多数情况下,可



以得出测验的奇数项目和偶数项目的分数。如果项目最初大致按照难度顺序排列的话,这种分法就得到两个几乎等值的半测验分数。在采用这种奇偶分半模式时,应该注意到,一组项目是涉及同一个问题的,例如阅读测验中的某段文章或某个机械图。在这种情况下,应该把整组项目原封不动地分到这一半或另一半。假如把这组项目分到不同的半测验,两个半测验分数的相似性就会假性提高,因为对这个问题理解上的任何一个错误都会影响对两个半测验中项目的回答。<sup>[3]</sup>

一旦得到每一被试的两个半测验分数,就可以用通常的方法计算它们的相关。然而,应该指出,这种相关实际上只是半个测验的信度。例如,整个测验由 100 个项目组成,那么计算两组分数之间的相关,而每组分数只是根据 50 个项目。另一方面,在再测信度和复本信度中,每个分数都根据测验的全部项目。

在其他条件相等的情况下,测验越长,测验就越可靠。<sup>[4]</sup> 我们有理由设想,使用较大的行为样组,就能够得到较为合适和一致的测量。斯皮尔曼—布朗 (Spearman-Brown) 公式能够估计增长或缩短一个测验对其信度系数的影响。公式如下:

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}$$

在这里,  $r_{nn}$  表示估计的系数,  $r_{tt}$  表示实得的系数,  $n$  表示测验增长或缩短的倍数。因此,如果测验项目数从 25 增加到 100,  $n$  等于 4。如果测验项目数从 60 减少到 30,  $n$  等于 1/2。用分半方法确定信度时,广泛使用斯皮尔曼—布朗公式,许多测验手册用这种型式报告信度。当公式应用于分半信度时,它总是把测验长度增加一倍。在这种条件下,能够把公式简化如下:

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}$$

在这里,  $r_{hh}$  表示两个半测验的相关。

鲁隆(Rulon)提出了另一种得出分半信度的方法(1939)。这种方法只需要每一被试在两个半测验上分数之差的方差( $SD_d^2$ )和测验总分的方差( $SD_x^2$ )把这两个数值代入下列公式,就直接得出整个测验的信度:

$$r_{tt} = 1 - \frac{SD_d^2}{SD_x^2}$$

注意这个公式与误差方差定义的关系,这是令人感兴趣的。被试在两个半测验上分数之差就是无关方差即误差方差。这些差的方差除以总分的方差,就得出分数中误差方差的比例。作为一种规定的测验惯例,从 1.00 中减去这个误差方差,就得出“真实方差”的比例,它等于信度系数。

库德—理查逊(Kuder-Richardson)信度和  $\alpha$  系数 得出信度的第四种方法也对一种型式实施一次测验,它的基础是对所有测验项目的反应一致性。这种项目间一致性(interitem consistency)受到两种误差方差的影响:(1)内容取样(像复本信度和分半信度一样);(2)所取样的行为领域的异质性。行为领域的同质性越高,项目间一致性就越高。例如,如果一个测验只包括乘法项目,而另一个测验则包括加法、减法、乘法、除法等项目,那么,前者的项目间一致性也许就会高于后者。后者是一个较为异质性的测验,一位测验参加者也许在减法项目上的得分相对较好,另一位参加者也许在除法项目上的得分相对较好,等等。一个较为极端的例子是,一个测验由 40 个词汇项目组成,而另一个测验则包括 10 个词汇项目,10 个空间关系项目,10 个算术推理项目和 10 个知觉速度项目。在后一个测验中,被试在不同类型项目上的成绩之间,就没有多少关系甚至没有关系。

显然,从相对同质性的测验上得出的测验分数,其意义较为

明确。假设在上述高度异质性的 40 个项目的测验上,史密斯和琼斯都得 20 分。我们能够得出两人在这个测验上成绩相等的结论吗?根本不能。史密斯可能做对 10 个词汇项目和 10 个知觉速度项目,而算术推理项目和空间关系项目则一个也没有做对。相反,琼斯得到 20 分,可能做对 5 个知觉速度项目,5 个空间关系项目,10 个算术推理项目,而词汇项目则没有做对。

许多其他的组合显然也会得出相同的总分 20。通过这种不同的项目组合得出这个分数时,它会有迥然不同的意义。另一方面,在相对同质性的词汇测验上,20 分也许就表示,测验参加者大致做对前 20 个项目,如果项目按照难度上升顺序来排列的话。也许他没有做对两三个较易的项目,却做对了两三个第 20 题以后的较难的项目;但是,这类个体变化的程度与较为异质性的测验相比,则微乎其微。

在这一点上,极为重要的是要搞清楚,测验试图预测的效标本身究竟是相对同质的,还是相对异质的。虽然人们比较喜欢同质性测验,因为可以对它们的分数作出十分明确的解释,但是,单一的同质性测验显然不能恰当预测一个高度异质的效标。此外,在预测异质性效标时,测验项目的异质性不一定是误差方差。设计异质性测验来预测异质性效标,传统的智力测验便是一个很好的例子。然而,在这种情况下,最好编制若干个相对同质性的分测验,而每个分测验测量异质性效标的一个不同方面。因此,对测验分数的明确解释可以和合适的效标范围相结合。

得出项目间一致性的最为通常的方法由库德和理查逊所提出(1937)。像分半方法一样,得出项目间一致性也对一种型式实施一次测验。然而,这种方法不需要两个半测验的分数,它考虑每个项目上的成绩。最初文章中提出的一系列公式,其中下列公式应用最为广泛,一般称为“库德—理查逊公式 20”:

$$r_{tt} = \left( \frac{n}{n-1} \right) \frac{SD_t^2 - \sum pq}{SD_t^2}$$

在这个公式中,  $r_{tt}$  表示整个测验的信度系数,  $n$  表示测验的项目数目,  $SD_t$  表示测验总分的标准差。把每个项目上做对人数比例( $p$ )和做错人数比例( $q$ )列表,就可得出这个公式中唯一的新概念  $\sum pq$ 。计算每个项目的  $p$  和  $q$  的乘积,然后,把所有项目的这种乘积相加,就得出  $\sum pq$ 。在编制测验的过程中,为了得出每个项目的难度水平,按惯例通常记录  $p$ ,所以,这种确定信度的方法就没有多少额外的计算了。

数学上能够证明,库德—理查逊信度系数实际上是一个测验所有不同的分半方法所得出的分半系数的平均数 (Cronbach, 1951)。<sup>[5]</sup>另一方面,普通的分半信度根据两个等值的半测验得出。因此,除非测验项目高度同质,库德—理查逊信度系数将低于分半信度。试举一个极端的例子说明这种差异。假设我们编制一个 50 个项目的测验,项目有 25 种不同类型,如果项目 1 和项目 2 是词汇,项目 3 和项目 4 是算术推理,项目 5 和项目 6 是空间关系,等等。在理论上,这个测验上的分数十分一致,因而得出较高的分半信度系数。然而,这个测验的同质性很低,因为整个测验的 50 个项目之间成绩一致性不会很高。在这个测验中,我们会预期到库德—理查逊信度比分半信度低得多。事实上,库德—理查逊信度系数和分半信度系数之间的差异,可以作为测验异质性的一项初步指标。

库德—理查逊公式仅用于项目按照对错或有无等二分法记分的测验。然而,一些测验包括多重记分项目。例如,在人格问卷上,被试在一个项目上可以得到不同的数字分数,这取决于他选择‘通常’‘有时’‘偶尔’‘从不’等。对于这类测验,已经推导出称之为  $\alpha$  系数的公式 (Cronbach, 1951; Kaiser & Michael, 1975;

Novick & Lewis, 1967), 在这个公式中, 用项目分数的方差之和  $\sum (SD_i^2)$  代替了  $\sum pq$ 。具体做法是, 先算出每个项目上所有被试分数的方差, 然后把所有项目的方差相加。a 系数的完整公式如下:

$$r_{tt} = \left( \frac{n}{n-1} \right) \frac{SD_t^2 - \sum (SD_i^2)}{SD_t^2}$$

评分者信度 (scorer reliability) 显然, 不同类型的信度在误差方差中所包括的因素有所不同。有时误差方差包括时间的波动, 有时误差方差则指平行项目之间的差异, 而有时误差方差又包括项目间的不一致性。另一方面, 误差方差的度量中没有包括的因素主要是两类: (1) 其方差应该保留在分数中的那些因素, 因为它们是被考虑的真实差异的组份; (2) 实验中能够控制的那些无关因素。例如, 在令人分心的条件下实施测验或时限比手册规定的长些或短些, 这时通常不报告得出的测量误差。客观上能够消除测验情境中的用时误差 (timing errors) 和严重的注意分散。因此, 不必报告相当于“注意分散方差”或“用时方差”的特别的信度系数。

同样, 大多数测验提供高度标准化的实施和评分的程序, 以至于由这些因素造成的误差方差可以忽略不计。用于大规模测验和计算机评分的团体测验, 情况尤为如此。对于这类测验, 我们只需要仔细遵照和认真检查所规定的程序。另一方面, 对于深入细致的个别检查中所使用的临床测验, 就要考虑主试方差 (examiner variance)。通过特别的实验设计, 有可能把主试方差从测验参加者条件中的时间波动或使用测验复本所造成的方差中分离出来。

能够十分简单地检查出的一个误差方差是评分者方差。某种类型的测验, 特别是创造力测验和人格投射测验, 大多取决于评分者的判断。对于这类测验, 像需要通常的信度系数那样, 也

需要评分者信度的度量。由两名主试对测验卷样本独立评分，就能够得出评分者信度。每个测验参加者因而获得两个分数，然后用通常方法计算分数的相关，得出的相关系数便是评分者信度的度量。当研究中使用主观性评分测验时，通常计算评分者信度。在适当的情况下，测验手册也应该报告评分者信度。

总览表 4-2 和表 4-3 概括了本节所讨论的不同类型的信度系数。在表 4-2 中，获得每种信度所使用的方法，按照所需要的测验型式和测验次数进行分类。表 4-3 说明每种信度所处理的误差方差的来源。

表 4-2 测 量信度的方法及其与测验型式和测验次数的关系

| 所需测验次数 | 所需测验型式数       |          |
|--------|---------------|----------|
|        | 一             | 二        |
| 一      | 分半信度、库德—理查逊信度 | 复本信度(即时) |
| 二      | 再测信度          | 复本信度(延迟) |

表 4-3 信度系数与误差方差来源的关系

| 信度系数类型                | 误差方差       |
|-----------------------|------------|
| 再测信度                  | 时间取样       |
| 复本信度(即时)              | 内容取样       |
| 复本信度(延迟)              | 时间取样和内容取样  |
| 分半信度                  | 内容取样       |
| 库德—理查逊信度和 $\alpha$ 系数 | 内容取样和内容异质性 |
| 评分者信度                 | 评分者之间差异    |

按照不同的来源所造成的分数方差的百分比 (percentage of score variance)，可以直接解释任何信度系数。因此，信度系数 0.85 表示，测验分数方差的 85% 取决于所测特质的真实方差，而 15% 则取决于误差方差 (由所使用的具体方法进行操作性定义)。有统计经验的读者可以回忆一下，相关系数的平方表示共同方差的比率。测验分数中真实方差的比率，实际上就是一次测验上的分数和不受偶然误差影响的真实分数之间相关的平方。这种相关称之为信度指数，<sup>[6]</sup> 等于信度系数的平方根 ( $\sqrt{r_{tt}}$ )。把信度指数本身平方，结果就是最初的信度系数 ( $r_{tt}$ )，作为规定的测验惯例，可以把信度系数直接解释为真实方差的百分比。

表 4-4 一个假设测验的误差方差来源之分析

|             |                                |
|-------------|--------------------------------|
| 延迟的复本信度     | $1 - 0.70 = 0.30$ (时间取样加上内容取样) |
| 斯皮尔曼—布郎分半信度 | $1 - 0.80 = 0.20^*$ (内容取样)     |
| 两者之差        | $0.30 - 0.20 = 0.10^*$ (时间取样)  |
| 评分者信度       | $1 - 0.92 = 0.08^*$ (评分者之间差异)  |
| 总的测量误差方差*   | $0.20 + 0.10 + 0.08 = 0.38^*$  |
| 真实方差        | $1 - 0.38 = 0.62$              |

\* 误差方差

一些实验设计得出同一团体的一种以上的信度系数，这就可以把总分的方差分成不同的组份。让我们考虑下列假设的例子。创造力测验有 A 和 B 两种型式，间隔两个月，施测 100 个六年级儿童，得到的复本信度为 0.70。随便根据哪一种型式上的分数，都能够计算分半信度系数。<sup>[7]</sup> 用斯皮尔曼—布郎公式校正

| “真”方差:62%                        |  | 误差方差:38% |      |         |
|----------------------------------|--|----------|------|---------|
|                                  |  | 20%      | 10%  | 8%      |
| 跨时间的稳定性;跨型式的一致性;<br>不受评分者之间差异的影响 |  | 内容取样     | 时间取样 | 评分者之间差异 |

图 4-4 一个假设测验的分数方差的百分比分布

的分半信度系数为 0.80。最后,第二位评分者对 50 份随机样本进行再评分,从中得到评分者信度为 0.92。现在能够分析这三个信度系数,得出表 4-4 和图 4-4 所示的误差方差。应当指出,从内容和时间两种取样造成的误差方差中(复本再测信度),减去由内容取样单独造成的误差方差(分半信度),我们得出,由时间取样单独造成的误差方差为 0.10。把内容取样造成的误差方差(0.20)、时间取样造成的误差方差(0.10)、评分者之间差异造成的误差方差(0.08)相加,就得出总的误差方差 0.38,因而真实方差为 0.62。在图 4-4 中,以较为熟悉的百分比来图示各种比例。对方差来源的这种分类,是所谓的信度概化理论(generalizability theory of reliability)的本质。使用复杂的实验设计,可以同时评定分数方差的多种来源及其相互作用,请参见布伦南(Brennan,1984)、克龙巴赫等(Cronbach et al,1972)、费尔特和布伦南(Feldt & Brennan 1989)以及谢夫森和韦布(Shavelson & Webb,1991)等的专题论述。

## 速度测验的信度

在编制测验和解释测验分数时,区分速度测验和难度测验



是十分重要的。在纯速度测验 (speed test) 中, 个体差异完全取决于作业的速度。编制速度测验, 全部采用低难度的项目, 所有项目都在测验所适用的个体的能力水平之内。但是规定的时限很短, 所以没有人能够答完所有项目。在这种条件下, 每一被试的分数仅仅反映他作业的速度。另一方面, 纯难度测验 (power test) 的时限很长, 每个被试足可尝试所有项目。但是项目的难度逐渐增加, 并且测验还包括一些难到任何人都答不出的项目, 所以没有人能够获得满分。

应当指出, 设计速度测验或难度测验, 都要防止被试获得满分。其理由是, 满分的意义是不明确的, 因为不可能知道, 如果测验包括更多的项目或更难的项目, 个体的分数还会高出多少。为了使每一个体充分表现他能够完成什么, 测验必须在项目的数目上或在难度水平上提供合适的上限。这项规则的一个例外是掌握测验, 例如第三章所讨论的领域参照测验。这类测验的目的, 不是建立个体能够做什么的限度, 而是确定个体是否达到预先建立的作业水平。

在实际工作中, 速度测验和难度测验的区别只是程度上的不同, 大多数测验兼有难度和速度两者, 只是比例有所不同。我们需要知道每个测验的速度和难度的比例, 不仅为了知道测验测量什么, 而且为了选择合适的方法来评估测验的信度。施测一次测验的信度系数, 例如用奇偶分半或库德—理查逊方法得出的信度系数, 都不适用于速度测验。当测验分数的个体差异取决于作业的速度时, 用这些方法得出的信度系数会是假性高相关。举一个极端的例子有助于说明这一点。让我们假设, 一个 50 个项目的测验完全取决于速度, 所以分数的差异完全根据所完成的项目数, 而不是根据错误率。这样, 如果个体 A 得到 44 分, 显然他做对 22 个奇数项目和 22 个偶数项目。同样, 如果

个体 B 得到 34 分，则奇数项目和偶数项目各得 17 分。结果，除了在一些项目上偶然的粗心错误之外，奇偶项目分数就会完全相关，即  $+1.00$ 。然而，这种相关完全是假性相关，并不表示测验的信度。

考虑一下得出分半信度和库德-理查逊信度所采用的方法，这表明两种方法都根据被试的错误数(number of errors)的一致性。现在，如果测验分数的个体差异不是根据错误率而是根据速度，则信度的度量显然必须根据作业速度(speed of work)的一致性。一旦测验成绩取决于速度和难度的结合，施测一次测验的信度系数虽然低于  $1.00$ ，但是这个系数仍然是假性高相关。只要测验分数的个体差异明显受到速度的影响，就不能恰当解释施测一次测验的信度系数。

那么使用哪些替代方法来确定速度为主的测验的信度呢？如果可以使用再测方法，那么它就比较合适。同样，复本信度也适用于速度测验，也可以使用分半方法，但要按照时间而不是按照项目来进行分半。换句话说，两半分数必须根据分别规定时间的两个半测验。一种分半方法是，实施两个等值的半测验，分别规定时限。例如，可以把奇偶项目分别印制在两张纸上，每组项目的用时为全测验时限的一半。这种方法等于在同一时间实施两个等值测验。但是，每个测验的长度仅为全测验的一半，而测验参加者的分数通常根据全测验。由于这种原因，应该使用斯皮尔曼-布朗公式或其他合适的公式来得出全测验的信度。

如果分别实施两个半测验行不通的话，一种替代方法是把总时间四等分，算出每段时间的分数。这种方法简单易行，每当主试发出事先安排的信号，就要测验参加者在他们正在做的项目上打个记号。然后，把第一段和第四段时间内答对的项目数相加，得出一个半分数；把第二段和第三段时间内答对的项目数

相加,得出另一个半分数。四段时间如此相加,往往可以平衡练习、疲劳和其他因素的累积效应。当项目的难度水平不是逐渐增加时,这种方法尤其令人满意。

在什么情况下一个测验是速度测验?在什么条件下必须采用本节所讨论的特别方法?显然,单单使用时限并不表示一个测验是速度测验。如果所有测验参加者都在规定的时限之内完成,那么,作业的速度在分数高低中就不起作用。没有完成测验的被试的百分比,可以作为速度对难度的一项粗略指标。然而,甚至在没有一个人完成测验时,速度的作用也可以忽略不计。例如,某测验有 50 个项目,在规定的时限内,每个人恰好都完成其中 40 个项目,虽然无人再有时间去完成所有项目,但是在速度上,完全不存在个体差异。

当然,问题的关键是:测验分数中由速度造成的个体差异达到什么程度?如果使用专业术语来陈述的话,即我们要知道测验分数的总方差中速度方差的比率是多少。算出不同个体所完成的项目数方差,然后把它除以测验总分的方差( $SD_c^2/SD_t^2$ ),这就能够大致估计这个比率。在上一段举出的例子中,每一个体都完成 40 个项目,所以,在完成的项目数上没有个体差异,该分数的分子就为零( $SD_c^2 = 0$ )。整个指标因而在纯难度测验中就等于零。另一方面,如果测验总分方差( $SD_t^2$ )由个体的速度差异造成,两个方差就会相等,它们的比率就等于 1.00。现在已发展若干种较为精确的方法来确定这个比率,但是,对于它们的详细讨论,超出本书的范围。

试举一例,对第 1 版 11~17 岁基本心理能力的 SRA 测验的研究中所收集的数据(Anastasi & Drake, 1954),说明速度对施测一次测验的信度系数上的影响。在这项研究中,首先用通常的奇偶分半法计算每个测验的信度系数,列于表 4-5 第一行。然

表 4-5 11~17 岁基本心理能力的 4 个 SRA 测验的  
信度系数(第 1 版)

| 得出信度系数的方法        | 言语意义测验 | 推理测验 | 空间测验 | 数字测验 |
|------------------|--------|------|------|------|
| 施测一次测验的<br>奇偶分半法 | 0.94   | 0.96 | 0.90 | 0.92 |
| 分别规定时间的<br>分半法   | 0.90   | 0.87 | 0.75 | 0.83 |

(资料摘自安娜斯塔西和德雷克, 1954)

后, 计算分别规定时间的两个半测验的分数相关而得出信度系数, 列于表 4-5 第二行。计算速度指标表明, 言语意义测验基本上是难度测验, 而推理测验则稍偏于速度测验。空间测验和数字测验大致属于速度测验。应当指出, 在表 4-5 中, 采用适当方法计算时, 空间测验的信度为 0.75, 而奇偶分半系数则为 0.90, 出现假性高相关。同样, 言语意义测验相对来说不属速度测验, 用两种方法计算信度时, 其差异就甚小。

## 所测样本对信度系数的影响

**变异性** 影响信度系数大小的一个重要因素是所测量信度的团体的性质。首先, 任何相关系数都要受到团体中个体差异的分布范围的影响。如果某团体每个成员的拼写能力几乎相同, 那么在该团体中, 拼写与任何其他能力的相关就接近于零。在这样一个团体之内, 知道某个体的拼写分数, 显然不可能预测他或她在任何其他能力上的地位。

两个能力倾向测验例如言语理解测验和算术推理测验之间

的相关，是另一个常见的例子。如果实施这两个测验的样本高度同质，例如作为 300 个大学二年级学生的团体，则二者的相关必然很低。在这样一个经过选择的大学生样本之内，由于分数的分布范围的限制，任何个体的言语能力和算术推理能力之间都没有多大关系。另一方面，如果实施这两个测验的 300 个被试为异质性样本，包括从智力落后患者到大学毕业生，则二者之间无疑获得较高相关。智力落后被试在这两个测验上的得分必然低于大学毕业生。同样的关系也适用于在这种高度异质性样本之内的其他亚团体。

考察一下图 4-5 中给出的假设的相关散布图，它进一步说明团体之内的变异性即个体差异的分布范围对相关系数的影响。这个散布图表示整个异质性团体内的高度正相关，因为分数都集中在从左下角到右上角的对角线附近。现在，如果我们仅仅考虑处于图的右上部的小矩形之内的亚团体，显然，两个变量的相关接近于零。在两个变量上处于这个较小的分布范围之内的个体，代表一个高度同质性的团体，例如前面提到的大学二年级学生。

像所有相关系数一样，信度系数也受到所测样本的变异性的影响。因此，如果测验手册中所报告的信度系数是从四年级儿童到中学生的某团体中得出，我们就不能假设在八年级样本中的信度必然同样高。一旦样本的同质性高于标准化团体，而测验又要用于区别该样本之内的个体差异，则应该根据这样一个样本重新确定信度系数。团体的标准差增大或缩小时，估计预期的信度系数的公式，在初级统计课本中都可找到。然而，最好还是根据一个团体重新计算信度系数。如果所设计的测验需要包括较大的年龄或能力范围，则测验手册应该分别报告标准化样本之内各个相对同质性的亚团体的信度系数。

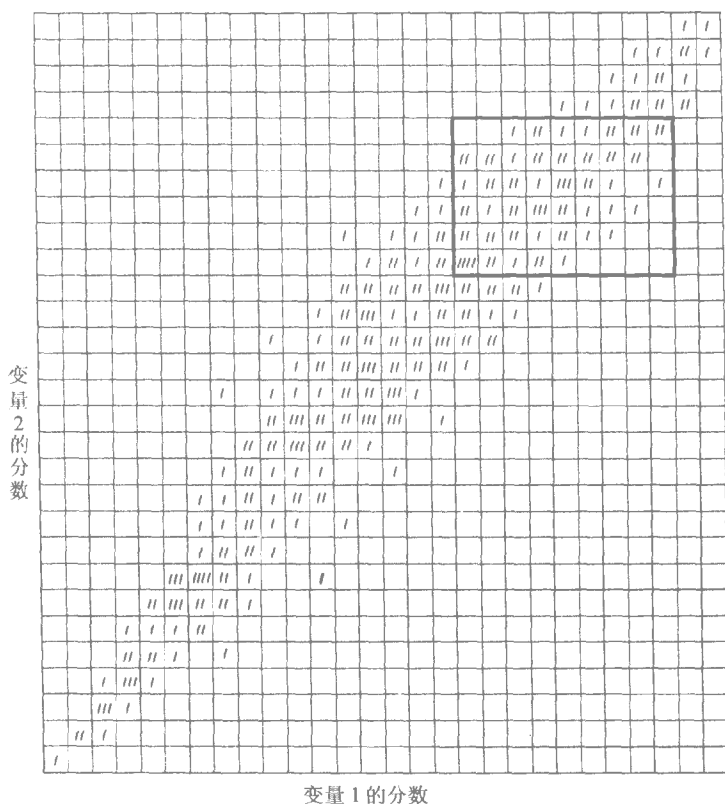


图 4-5 较小的分数分布范围对相关系数的影响

**能力水平** 信度系数不仅随着样本中个体差异分布范围而变化，而且在平均能力水平不同的团体之间也有所不同。此外，这种差异一般不能用统计公式来预测或估计，只能通过对年龄或能力水平不同的各种团体进行测验的客观试验来发现。单一

测验的信度会产生这类差异，或者由于在测验的不同难度水平上测量的是稍有不同的能力组合，或者由于测验的长度在不同的年龄水平上有所不同。即使项目数相同，但是高低两端可能没有编排适当难度水平的足够的项目，让个体充分表现他们能够做什么（上限效应或下限效应）。在其他测验中，年幼者团体和能力较低者团体的信度相对较低，因为他们的分数受到猜测的影响过大。

显而易见，每个信度系数都应该同时充分说明确定信度系数的团体的类型。应该特别注意样本的变异性和能力水平。所报告的信度系数仅仅适用于类似于计算信度系数的样本。测验编制中一种合乎需要、日益增多的做法是，按照年龄、性别、年级水平、职业等等，把标准化样本分成几个较为同质的亚团体，分别报告每个亚团体的信度系数。在这种条件下，信度系数多半适用于实际工作中所测样本。

## 测量标准误差

**解释个体分数** 测验的信度还可以用测量的标准误差 (standard error of measurement, SEM) 来表示，也称之为分数标准误差。标准误差这个度量尤其适合于解释个体分数。因此，对于许多测验目的来说，测量标准误差比信度系数更为有用。根据测验的信度系数，用下列公式能够容易地计算测量标准误差：

$$SEM = SD_t \sqrt{1 - r_{tt}}$$

在这里， $SD_t$  表示测验分数的标准差， $r_{tt}$  表示信度系数，两者都根据相同的团体计算。例如，如果某个智力测验离差 IQ 的标

准差 15,信度系数为 0.89,则该测验 IQ 的 SEM 为  $15 \sqrt{1-0.89}$   
 $= 15 \sqrt{0.11} = 15 \times 0.33 = 5.00$ 。

为了理解 SEM 如何解释个体分数,让我们假设,我们得到儿童珍妮特在上述测验上的 100 个 IQ 分数。由于本章所讨论的各种偶然误差,这些分数各不相同,它们形成围绕着珍妮特真分数的正态分布。作为规定的测验惯例,这 100 个分数分布的平均数称为“真分数”(true score),分布的标准差则称为 SEM。像任何标准差一样,这种标准差能够按照第三章所讨论的正态曲线频数来解释(参见图 3-3)。回忆一下,正态曲线中平均数  $\pm 1\sigma$  之间,大约有 68% 的个案。因此,我们得出结论,大约有 2:1(即 68:32)的概率,珍妮特在该测验上的 IQ 处于  $\pm 1\text{SEM}$ ,即在她的真分数两边各 5 分之间。我们可以预期,她的得分在 105 和 115 之间的次数大约为 2/3(68%)。

如果要作出更为可靠的预测,我们可以选择大于 2:1 的概率。查阅图 3-3(第三章),它表明  $\pm 3\sigma$  包括 99.7% 的个案。根据正态曲线频数表可以确定,平均数两边  $2.58\sigma$  的距离恰好包括 99% 的个案。因此,珍妮特的 IQ 处于她的真 IQ 两边各 2.58 个 SEM 即  $2.58 \times 5 = 13$  分之内的概率为 99:1。因此我们能够以 99% 的置信水平(错误的概率仅为 1%)陈述,珍妮特在任何一次测验上的 IQ,必然处于 97 和 123 之间( $110 - 13$  和  $110 + 13$ )。如果对珍妮特实施 100 个等值的测验,则她的 IQ 只有一次处于这个数值带之外。

当然,在实际工作中,我们并不知道真分数,而只知道在一次测验中的实得分数。在这种情况下,我们能够逆向应用上述推理。如果一个体的实得分数不可能偏离他的真分数 2.58 个 SEM 之上,我们就可以说他的真分数必然处于他的实得分数的 2.58 个 SEM 之内。虽然不能确定这种陈述对于某实得分数的



概率，但我们能够说这种陈述对于全体个案的 99% 必然正确。根据这种推理，格利克森(Gulliksen, 1950, pp. 17 ~ 20 )提出,应该按上述说明来使用测量的标准误差，以便为任何已知实得分数的个体估计真分数的合理范围。在心理测验中，习惯上正是按照这种“合理范围”来解释测量的标准误差，本书中也作如此解释。<sup>[8]</sup>

显然，测量的标准误差和信度系数是表示测验信度的两种不同方法。与信度系数不同，测量的标准误差不受所测团体的变异性的影响。测量的标准误差以个体分数来表示，无论是同质团体还是异质团体，所得标准误差保持不变。另一方面，测量的标准误差以分数单位来报告，在测验之间不能直接比较。当以算术问题、词汇测验中的单词等等来报告测量的标准误差时，就会产生单位可比性这一常见问题。因此，如果我们要比较不同测验的信度，信度系数比测量标准误差好。而解释个体分数，测量标准误差则较为合适。

然而，当能力水平变化很大时，不能假设信度系数和测量标准误差都保持不变。在同一测验的不同水平上计算测量标准误差时，上一节所讨论的信度系数的差异仍然存在。第三章所讨论的项目分析的 IRT 技术，可以全面解决这个问题。这种技术跨越较大的能力分布范围，把测验的测量准确度表示为能力水平的函数。这种做法得出一条仅仅取决于测验所包括项目的测验信息曲线(test information curve)，并能够在每一能力水平估计测量标准误差。第七章进一步讨论这种技术。

SEM(或测量准确度的其他指标)谨防过分重视单一的数字分数。SEM 的这种应用很重要，现在出版的越来越多的测验报告分数,不是使用单一的数字,而是使用个体真分数可能处于中的一个分数带。大学入学考试委员会在分发给中学和大学顾

问的材料中,以及在分发给测验参加者的 SAT 个体分数报告中,提供 SEM 数据极其使用说明。为了使学生了解测验分数的意义,教材中同样包括 SEM。SEM 资料也用于解释研究生资格考试的分数(GRE,1995-96 指南)。

**解释分数差异** 在评价两个分数之间的差异时,考虑测验信度和测量标准误差尤为重要。考虑到每个分数可能波动的分布范围,就会避免过分重视分数之间的微小差异。不论比较不同个体的测验分数,还是比较同一个体在不同能力上的分数,都要小心谨慎。同样,指导语或其他实验变量引起的分数变化,也需要按测量标准误差来解释。

关于测验分数的一个常见的问题是个体在不同领域的相对地位。多丽丝的言语能力比数字能力强吗?汤姆的机械能力倾向比言语能力倾向强吗?如果在能力倾向成套测验上,多丽丝言语测验的得分比数字测验高,汤姆机械分测验的得分比言语分测验高,则我们有多大把握,用成套测验的另一个型式进行再测,他们的得分仍然如此?换句话说,他们的分数差异仅仅是由所用的特定的言语、数字、机械等分测验中具体项目的偶然选择造成的吗?要适当地解释能力和人格特质的多重评分成套测验上的分数,这个问题尤为重要(Anastasi,1985a)。解释个体在这类成套测验上分数剖面图时要考虑种种问题,对它们的进一步讨论及其例子,可以参见第八章和第十章(能力测验)以及第十三章(人格测验)。

由于日益关注意解释分数剖面图,测验出版者正在研制各种报告型式,准许以测量标准误差来评价分数。一个例子是区别能力倾向测验所使用的个体报告型式,它包括图 4-6 所示的资料。在这种型式上,把成套测验的每个分测验上百分位数分数绘制成百分位数带,实得分数则处于每个数带的中心。每个百

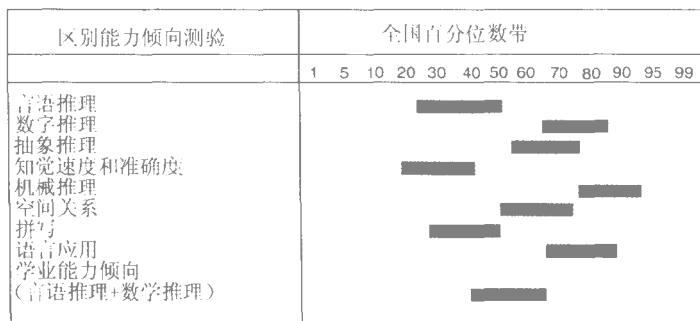


图 4-6 区别能力倾向测验分数剖面图，它说明百分位数带的使用

分位数带相当于实得分数两边各一个 SEM 的距离,因此,个体“真”分数处于这个带之内的概率大约为 2:1(0.68:0.32)。解释剖面图时,建议测验使用者不要过分重视百分位数带重叠的分数差异,特别是如果百分位数带一半以上长度相重叠的话。例如图 4-6 所示的剖面图中,言语推理和数字推理的分数差异也许反映能力水平的真正差异;数字推理和抽象推理的分数差异也许并非如此;抽象推理和机械推理的分数差异则是不确定的。

最好记住,两个分数之间差异的标准误差要大于其中任何一个分数的测量标准误差。这是由于这种差异受到两个分数中存在的偶然误差的影响。根据两个分数的测量标准误差,用下列公式,就能得出两个分数之间差异的标准误差:[9]

$$SE_{\text{diff.}} = \sqrt{(SEM_1)^2 + (SEM_2)^2}$$

在这里, $SE_{\text{diff.}}$ 表示两个分数之间差异的标准误差, $SEM_1$ 和 $SEM_2$ 分别表示两个分数的测量标准误差。用  $SD \sqrt{1 - r_{11}}$  代替

SEM<sub>1</sub>, 用  $SD \sqrt{1 - r_{22}}$  代替 SEM<sub>2</sub>, 我们可以直接以信度系数来重写公式如下:

$$SE_{\text{diff.}} = SD \sqrt{2 - r_{11} - r_{22}}$$

在这个替代公式中, SD 表示测验 1 和测验 2 使用的相同的标准差, 因为在比较分数之前, 两个测验的分数必须以相同的量表来表示。

我们以韦克斯勒成人智力量表一修订版 (WAIS - R) 上言语 IQ 和操作 IQ 为例来说明上述做法。两个分数的分半信度分别为 0.97 和 0.93。表示 WAIS - R 离差 IQ 的平均数为 100, SD 为 15。因此, 两个 IQ 分数之间的差异标准误差能够得出如下:

$$SE_{\text{diff.}} = 15 \sqrt{2 - 0.97 - 0.93} = 4.74$$

为了确定在 0.05 显著性水平上, 由偶然误差造成的分数差异为多大, 我们把差异标准误差 (4.74) 乘以 1.96, 结果为 9.29 分, 即近似 10 分。因此, 个体 WAIS - R 言语 IQ 和操作 IQ 之间差异应该至少为 10 分, 才能达到 0.05 显著性水平。<sup>[10]</sup>

## 掌握测验的信度和临界分数

回顾第三章, 领域参照测验通常(不是必然)以掌握而不是成就程度来评价作业。掌握测验的一种主要统计含义是, 减少个体分数的变异性。从理论上讲, 如果每个人继续训练直到掌握某种技能为止, 变异性就减少到零。在本章前面一节中, 我们看到, 包括信度系数在内的任何相关, 都会受到团体的变异性的影响。样本的变异性减小时, 相关系数也减小。显然, 在一组个体达到事先规定的掌握水平之后, 就不能使用通常的方法来估

计大多数领域参照测验的信度。在这种条件下,甚至一个稳定性和内在一致性都很高的测验,也会得出接近于零的相关系数。

估计信度中的这种明显的困难,起因于没有考虑到设计领域参照测验来测量什么东西。在实际工作中,这种测验本质上用于区别个体是否获得一项指定活动所要求的知识和技能。从获得驾驶员执照或分配专业,到在个别化教学程序中进到下一单元或学习某门课程,实施测验的具体目的大不相同。不过,在所有这类情境中,使用测验这个事实意味着预期个体之间作业的变异性。这种变异性主要反映有关机能上以往训练数量的个体差异。

为了估计领域参照测验的信度,特别设计了一打以上的不同方法(Berk, 1984b; Brennan, 1984; Subkoviak, 1984)。其中一些方法适用于简单的掌握与否的两分决策,即所有的分类误差都被认为同样严重而不管它们离开临界分数的距离。在这种情况下,可以使用平行型式进行再测,以便得出在两次测验上得到相同决策的个体的百分比。要进一步分析这些数据,还可以计算适当的一致性指数和显著性的值。另一些方法考虑在两次测验上的实得分数,并且提供一些指数来反映每一个体在某临界分数之上或之下的离差。选择哪种方法,应该考虑测验的性质和用途、临界分数的位置以及测验的其他心理测量特性。技术文献中广泛详尽地讨论了需要考虑的有关事宜(参见 Berk, 1984a; Feldt & Brennan, 1989)。

## 注 释

- [1] 研究分数信度的这种方法有时称之为信度的概化理论 (generalizability theory, 参见 Brennan, 1994; Crick & Brennan, 1982; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Feldt & Brennan, 1989; Shavelson & Webb, 1991)。然而,

这种名称不够具体,因为概化可应用于测验分数的各个方面,事实上,可应用于所有科学数据。一种较为确切的描述则是,这种信度方法的基础是把方差组份确定为有关的或无关的。

- [2] 适用于 11~17 岁的《基本心理能力》的 SRA 的分测验之一。阿纳斯塔西和德雷克(Anastasi & Drake, 1954)的一项研究中得出这批数据。
- [3] 现在有充分的证据证实这种设想,关于这类整合的项目分组即“小测验”(testlets)的统计处理,已经积累了相当多的研究(Sireci, Thissen & Wainer, 1991)。
- [4] 然而,增加测验长度,只是增加内容取样的一致性,而没有增加跨时间的稳定性(参见 Cureton, 1965; Cureton 等, 1973)。
- [5] 严格说来,只有当分半系数用鲁隆公式(根据两个半测验分数之差的方差)得出时,情况才是如此。当分半系数用两半相关和斯皮尔曼-布朗公式得出时,情况并非如此(Novick & Lewis, 1967)。
- [6] 根据两组不同的假设推导信度指数,参见格利克森(Gulliksen, 1950, 第二章和三章)的有关论述。
- [7] 为了较好地估计内部一致性系数,可以计算每种型式的分半相关,再用合适的统计程序(例如费希尔  $z$  转换)计算两个系数的平均数。
- [8] 有人提出其他的方法,把估计的“真”分数作为置信区间的中心(Dudek, 1979; Glutting, McDermott & Stanley, 1987)。如果信度系数较高,这种方法就没有多少影响;如果信度系数较低,真分数和置信区间的大小都是根据同一个难免有错的信度系数来计算。而且,最理想的方法随着测验分数的使用目的而有所不同(例如长期预测和当前作业评定)。
- [9] 不要把这个公式混淆于计算两个团体平均数之间差异标准误差的公式,当要比较的两个变量存在相关时,后一个公式包括一种相关概念。而两个变量中的测量标准误差则是随机误差,因此假设它们不相关。
- [10] 利用每个年龄组的信度和标准差,可以得到更为精确的估计值。测验手册中报告这样计算时,在 0.05 显著性水平上最小的言语-操作差异,分布范围从 8.83 分至 12.04 分。然而,大多数数值接近 10 分。

## 第五章 效度：基本概念

一个测验的效度涉及这个测验测量什么东西，以及它在什么程度上测量到所要测量的东西。测验效度告诉我们，从测验分数能够推出什么结论。在这一点上，我们应当谨防把测验名称当作测验测量了什么东西的一个指标。测验名称只不过是识别测验目的的一个简短而方便的标记。虽然现在日益使用较具体的、经验界定的测验名称，但大多数测验名称太笼统、太含糊，从中难以看出所包括的行为领域。只有检查建立测验效度中所使用的客观资料和经验操作，才能确定测验所测量的特质。此外，不能一般性地报告测验效度，不能抽象地说某个测验的效度“高”或“低”。建立测验效度，必须参照所考虑的测验的特定用途。

各种确定测验效度的方法，基本上都涉及测验成绩和独立观察到的被研究的行为特征的事实之间的关系。研究这种关系可以使用许多具体方法，它们的名称也各不相同。这些传统的方法强调效度的不同方面，尤其是不同的测验用途。随着测验应用的发展和扩大，效度概念也随之发生改变（Anastasi, 1986a; Messick, 1988, 1989）。

## 测验效度概念的发展

测验的最早用途之一，便是评定个体在指定的内容领域已经学会了什么东西。今天，学校中学科的期末考试，以及驾驶员执照或某职业资格考试，都体现了这种用途。这类测验通常称之为成就测验。而评价成就测验，一般是把测验内容与测验所要评定的内容领域相比较。这种描述性方法，还与建立测验的某种应用的效度有关，本章有一节将讨论这种方法。

测验发展到第二阶段时，重点则转到预测。现在或将来某个时候，不同的个体在一定的情境中会怎样反应？同一个体在不同的指定情境中又会怎样反应？所要预测的行为在情境中的表现逐渐被称为效标。出于这种目的，通常把测验效度报告为测验分数与该效标的直接的、独立的测量之间的相关系数。在教育、职业或特定治疗上，使用测验来选择或安排个体时，这种方法尤其合适。因此，对于一个机械能力倾向测验，其效标可以是成为机械师后的工作表现；对于一个学术能力倾向测验，其效标可以是大学的成绩；而对于一个神经过敏症测验，其效标可以是有关人员的评定或其他可得到的关于个体在各种生活情境中行为的资料。

测验在目前的阶段出现两种主要趋向：(1)加强理论趋向；(2)心理学理论与通过经验和实验的假设检验之间密切联系。这两种趋向在测验编制和效度建立中较为明显，总的来说，在心理学的其他领域也是如此（Anastasi, 1992a, 1992b, 1995）。这两种趋向的一种结果是，日益认识到结构在描述和理解人类行为中的重要性。结构是一种广泛的范畴，源自直接可观察的行为



变量共同具有的一般特性。结构又是一种理论实体，本身无法直接观察。对结构的兴趣导致引进最初所认为的第三种测验效度即结构效度（AERA, APA, NCME, 1985; APA, AERA, NCME, 1974; Cronbach & Meehl, 1955）。最后，就它说明测验测量什么东西而言，结构效度逐渐被认为是基本的、包括一切的效度概念。许多资料来源都有助于界定和理解一个测验所评价的结构，其中包括内容效度方法和预测效度方法。同时，这两种方法本身也提供有价值的资料，在评价测验的某种用途时卓有成效。因此，尽管这两种概念(和术语)可以并入综合的结构效度概念，它们还是保存下来了。

## 内容描述程序

**性质** 内容描述的效度分析程序实质上是要系统检查测验内容，从而确定测验是不是包括了所要测量的行为领域的代表性样本。一些测验测量个体掌握某种技能或学习某门课程所达到的程度，它们通常使用内容描述的效度分析程序。似乎仅仅检查测验内容，就足以建立测验在该目的上的内容效度。如果一个测验分别由乘法、拼写、簿记等项目组成，这个乘法测验、拼写测验或簿记测验似乎就是有效度的了。

然而，解决问题的办法并非像看起来的那般简单。一种困难是对项目总体进行合适的取样。必须系统地分析所要测验的行为领域，以便弄清楚测验项目是否以恰当的比例包括所有的主要内容。例如，一个测验很容易偏重于行为领域中比较适合于编制客观性项目的那部分内容。应该事先完整描述所考虑的行为，而不是在编制测验之后再作规定。例如，一个编制良好

的教育测验应该包括各种教学目标，而不仅仅是学科内容。因而内容必须是广义的，包括各种主要目标，例如应用原理解释数据以及事实知识。此外，内容效度取决于个体的测验反应与所考虑的行为领域的关系，而不是取决于项目内容的表面适当性。仅仅检查测验内容并不能揭示被试在测验中所实际使用的过程。

另一种应当注意的倾向是，谨防把测验所取样的行为领域过于概化。例如，一个多项选择题的拼写测验可以测量识别拼对和拼错单词的能力，但是不能假设这个测验也测量正确听写的 ability、作文中的拼错率以及其他方面的拼写能力（Ahlstrom, 1964; Noel & Harris, 1952）。另一个困难在于测验分数中可能包括无关因素。例如，一个测量数学能力或机械能力的测验，可能过度受到言语理解能力或完成简单日常工作的速度的影响。

**具体方法** 建立测验的内容效度，从一开始就要选择合适的项目。对于教育测验，在准备项目之前，应该全面地、系统地检查有关的课程大纲和教材，并且征求学科专家的意见。根据所收集的资料，为项目编写者起草测验细目表（test specifications），这种细目表应该说明要包括的内容领域或主题、要测验的教学目标或过程，以及每个主题和过程的相对权重。最后的细目表应该指出每个主题所要准备的每类项目的数目。例如，阅读能力测验可以包括理解上下文中的词汇，逐字逐句理解内容，根据某种资料作出正确推断等。这种测验也可以从不同来源的材料中取样，例如论文、诗歌、报纸上的文章或操作设备的说明书。数学测验则可以包括计算技能、解答文字题，以及把学过的过程应用于不熟悉的新情境。

在教育成就测验手册中讨论内容效度时，应该说明如何保证测验内容的合适性和代表性的程序。假如学科专家参与编制

测验的过程,则应该说明专家的数目和专业资格。假如他们作为项目分类评判员,则应该报告给他们的指导语以及评判员之间的一致性程度。由于课程和课程内容经过一段时间会发生变化,因此特别需要报告学科专家咨询的时间,也应该报告所研究的教材和课程大纲的数量、性质以及出版时间。

为了补充教育成就测验的内容效度分析,通常也采用一些经验方法。可以检查每个项目上的分数和总分随年级升高的进步程度。一般而言,应该保留从低年级到高年级这一阶段中,儿童通过项目的百分比增加最大的那些项目。在适宜的情况下,还可以采用其他的补充方法,包括分析测验中常见的错误和观察被试使用的解题方法。后者可通过个别测验进行,在学生解答每个问题时,要求他们‘出声思维’(think aloud)。注意一下多少学生没有做完测验,或者采用第四章所讨论的较为精确的方法来检查速度的影响。为了检查阅读指导语的能力对测验成绩可能产生的无关影响,可以计算该测验分数与阅读理解测验分数的相关。另一方面,如果测验本身要测量阅读理解能力,则提出一些问题而又没有与这些问题有关的阅读材料,这样可以说明被试仅仅根据先前的信息或其他无关线索能够回答多少问题(Scherich & Hanna, 1977)。

**应用** 特别当用已经说明的那些经验检查来确定内容效度时,内容效度分析就为评价成就测验提供了一种合适的方法。它能让我们回答教育成就测验的效度的两个基本问题:(1)该测验覆盖了规定的知识和技能的代表性样本吗?(2)测验成绩不受无关变量的影响吗?内容效度分析尤其适合于第三章所描述的领域参照测验。因为以内容意义来解释领域参照测验上的成绩,显然内容效度是有效使用这类测验的必要条件。然而,全面评价领域参照测验的有效性,还需要其他类型的效度证据(参见

Hambleton, 1984b)。

内容效度分析也适用于某些对员工进行选拔和分类的职业测验，第十七章将讨论这种测验。当测验内容是取自实际工作，或者是实际工作所需要的知识和技能时，这类效度证据才是合适的。在这种情况下，应该进行彻底的工作分析，以便证实工作活动和测验之间的相似性。舍恩菲尔特、舍恩菲尔特、阿克、珀尔森等( Schoenfeldt, Schoenfeldt, Acker & Perlson, 1976 )清楚而详尽地说明了编制工业阅读测验时怎样应用这些效度分析方法。研究人员同公司的任职人员和管理人员密切合作，按照题材和理解技能，调查了一个大制造公司中各种起点工作的阅读要求，然后编制与这些要求相称的测验项目。在编制联邦政府和州政府公务员测验时，这种方法得到了广泛的采用( Hardt, Eyde, Primoff & Tordy, 1981; Menne, McCarthy & Menne, 1976; Primoff & Eyde, 1988; Tordy, Eyde, Primoff & Hardt, 1976)。

另一方面，对于能力倾向测验和人格测验，内容效度分析通常是不合适的，事实上甚至可能被误解。虽然在开始编制任何测验时，显然都应该考虑内容的恰当性和代表性。但是能力倾向测验或人格测验的最终效度分析，需要用下面章节中所描述的方法在经验上加以证实。比起成就测验来，这些测验与所取样的行为领域的内在相似性要低。所以，能力倾向测验和人格测验的内容只能显示某些假设，而这种假设指导测验编制者去选择某种内容来测量规定的特质。为了建立测验效度，需要在经验上证实这类假设。

能力倾向测验和人格测验与成就测验不同，它们不是根据指定的教学课程或统一的先前经验来抽取测验内容。因此，在能力倾向测验和人格测验中，对相同的测验项目作出反应时，被试使用的作业方法或心理过程可能大不相同。同一个测验可能

测量不同个体的不同机能。在这种情况下,实际上不可能从检查测验内容来确定测验所测量的心理机能。例如,大学毕业生可能以言语或数学能力来解答某个问题,而机械工则以空间想象能力得出相同的解答。又如,某个测验用于中学一年级学生时是测量其算术推理能力,用于大学生时仅仅测量其计算速度的个体差异。

**表面效度** 内容效度不应该混淆于表面效度 (face validity)。在技术意义上,表面效度不能算是一种效度;它并不是指测验实际上所测量的东西,而只是指测验表面上看起来所要测量的东西。表面效度是指对于参加测验者、决定采用测验的管理人员,以及其他未经专门训练的旁观者来说,测验是否“看起来有效”。表面效度的问题主要涉及友好关系和公众关系。虽然在这点上“效度”一词的通常用法可能引起混乱,但是表面效度本身也是一种合乎需要的测验特征。例如,一些测验最初为儿童设计,而且使用于课堂环境,一旦首次扩大为成人使用,它们常常由于缺乏表面效度而遭到反对和批评。毫无疑问,如果测验内容看起来不相干、不合适,或者愚蠢、幼稚等,其结果必然是被试合作不好,而不管测验的实际效度如何。特别在成人测验中,一个测验有客观效度还是不够的,它还需要表面效度,以便在实际情境中有效地实施。表面效度也影响在立法和司法决定中对于一项测验的可接受性,以及一般公众对于测验的评价。

巴鲁克·内瓦 (Baruch Nevo) 及其同事研究了表面效度 (B. Nevo, 1985, 1992; B. Nevo & Sfez, 1985), 在这项创新的研究计划中,请测验参加者评定测验(参见第一章)。首先,他们注意到,关于表面效度的研究甚少,尽管表面效度可能对测验的一般态度起作用。然后,他们提出对表面效度进行定量评定,请测验参加者和其他对心理测量虽不精通却感兴趣的人士来评定一项测验对

于其用途的适合性；也可以使用相同的方法来评定每个测验项目或一套测验。1385 名以色列学生参加了包括 6 个测验的大学入学考试，分析他们在“被试反馈问卷”（Examinee Feedback Questionnaire）上的反应，从而报告解说性数据。结果表明，评定者之间一致性、再测信度，以及测验之间的区别性和计划主修不同专业的被试亚团体之间的区别性均较好。他们建议，在测验手册中通常应该报告关于表面效度的定性资料和定量资料。

修改测验项目措辞，使之在应用测验的特定情境中看起来相关或合理，常常能够改进表面效度。例如，如果编制一个用于机械工人的简单算术推理测验，应该以机械作用语来表述项目，而不要用“8 角 6 分钱能买多少橘子”或其他传统教科书中的问题。同样，用于海军人员的算术测验应该以海军用语来表述，这不会改变所测量的功能。当然，千万不要把表面效度看作为客观效度的替代物。不能假设，改进一个测验的表面效度，必然改进它的客观效度。也不能假设，为了增加测验的表面效度而修改测验时，测验的客观效度仍然不变。总是应该直接检查测验最后型式的客观效度。

## 效标预测程序

同时效度分析和预测效度分析 效标预测效度表明测验预测个体在指定活动中行为表现的有效性。用于得出测验分数效度的效标度量，可以在同测验分数大致相同的时间获得，也可以在相隔一定的时间之后获得。1985 年“测验标准”（testing standards）根据效标和测验之间的时间关系，区分出同时效度和预测效度。“预测”一词可以在广义上使用，表示以测验来预测任何

效标情境也可以在狭义上使用,表示相隔一段时间的预测。在“预测效度”这种说法中,正是在后一种意义上使用预测一词。用于人员的选拔和分类的测验,大多使用预测效度提供的资料,例如,招聘人员、大学和专业学校的招生选拔、分派军事人员参加职业培训等,在分类决策中都需要了解测验的预测效度。其他例子还包括使用测验来筛选在恶劣环境中可能产生情绪障碍的报考者,以及使用测验来辨认最可能得益于某特定治疗的精神病患者。

在一些情况下,同时效度分析(*concurrent validation*)只是作为预测效度分析的替代品。经过一段时间建立预测效度,或者事先选择一个适合于测验目的的样本,这往往都是行不通的。因此,作为折衷的解决办法,可以对已有效标资料的团体实施测验。例如,可以把大学生的测验分数与他们测验时的学科平均成绩相比较,或者把员工的测验分数与他们当前的工作成功相比较。

另一方面,在心理测验的某些应用中,同时效度本身也是最为合适的一种。预测效度和同时效度之间的逻辑区分,不是根据时间而是根据测验目标。测验的同时效度多用于诊断现在的状态,而不是预测将来的结果。这种差异可以用两种问法来说明:“史密斯有资格当优秀飞行员吗?”或者“史密斯具备成为优秀飞行员的必要条件吗?”第一个问题需要同时效度分析,第二个问题则需要预测效度分析。

因为在测验时总是可以得到同时效度分析的效标,我们可能会问,在这种情况下,测验还有什么功用。一般来说,这类测验提供的资料比效标更简单、更迅速、更经济。例如,如果效标是在医院连续两个星期的对病人的观察,而测验能够区分正常人和有障碍的、可疑的个案,那么测验就会明显减少这类详尽观

察所需要的人员数目。

**效标污染** 得出测验效度时，一种必要的预防措施是弄清楚测验分数本身不会影响个体的效标成绩。例如，如果一位大学教师或一位工厂管理人员知道某人在一个能力倾向测验上得分很低，这类消息可能会影响他对该学生或该员工的评定。又如，在作出学业成绩或职业等级评定时，对测验得高分的人可能有好处。显然，这类影响会完全是假性的，即人为提高测验分数和效标之间的相关。

测验效度分析中这种可能的误差来源称之为效标污染 (criterion contamination)，因为效标评定受到评定者知道测验分数的“污染”。为了防止这类误差的作用，绝对不能让参加效标评定的人员知道被试的测验分数。出于这种原因，“对测验进行检验”时使用的测验分数必须严格保密。有时很难使教师、雇主、军事官员或其他行业人事干部相信，这类预防措施是必不可少的。这类人员急切地希望利用所有可以得到的资料来作实际决策，他们可能没有认识到，在得出效标数据并检验效度之前，测验分数必须置之一边。

**效标度量** 一个测验有多少种具体用途，就可以根据多少种效标进行效度分析。评定任何情境中行为的任何方法，都为某种特定的目的提供一种效标度量。不过，测验手册中所报告的进行效度分析时应用的效标，可以分为几种常用类型。对智力测验进行效度分析时最常用的一种效标是学业成就 (academic achievement)。正是出于这种原因，往往把智力测验更确切地描述为学业能力倾向的测量。用作效标度量的具体指标包括在校成绩、成就测验分数、升级或毕业记录、某种荣誉或奖励，以及教师对学生的“智力”评定等。至于在学业环境之内作出的智力评定多半带上浓厚的个体学业成绩的色彩，可以使用学业成就效



标将它们适当分类。

种种学业成就指标，提供从小学低年级到大学和研究生院的所有教育水平的效标数据。虽然它们主要用于一般智力测验的效度分析，但也可以作为某些多重能力倾向测验和人格测验的效标。例如，用某个测验来选拔大学生，其效度的常用效标就是大学一年级学生的平均成绩。这种度量是大学一年级期间学习的各门课程的平均成绩，每门成绩的权重则是它得到的学分数。

经常用于校外成人的学业成就效标的一种变式，是个体已经完成的教育量。一般认为，智力较高的个体受教育的时间较长，而智力较低的个体则较早退学。这条效标的基本假设是，教育阶梯起着一种不断选拔的作用，每上一个梯级，就淘汰一些没有继续学习能力的个体。例如，比起小学毕业生，大学毕业生是一个经过高度选拔的团体。这无疑是对的，但是，教育量和学业能力倾向之间的关系决不是完全正相关。特别对高等教育水平，经济因素、社会因素、动机因素和其他非智力因素都会影响个体继续受教育的可能性。而且，使用这类同时效度分析，难以分清因果关系。在什么程度上，智力测验分数上所得出的差异只是不同的教育量的结果？又在什么程度上，测验能够预测后来教育进展中的个体差异？只有像预测效度那样，在得出效标之前实施测验，才能回答这些问题。

在编制特殊能力倾向测验时，常用的效标是特殊训练课程的成绩。例如，机械能力倾向测验可以根据工艺课程的期终成绩进行效度分析。诸如打字或簿记之类的各种商学院课程，为这些领域的能力倾向测验提供效标。同样，在音乐或美术学院的成绩，也可以用来对音乐或美术能力倾向测验进行效度分析。一些专业能力倾向测验的效标，可以是法学院、医学院、牙科工程学院以及其他专门学院的成绩。至于专门化的测验即适用于

具体测验课程的测验，训练记录是常用的效标数据。举一个著名的例子，空军飞行员选拔测验的效度分析是根据基本飞行训练的成绩。在其他军事专业和一些工业测验的效度分析中，通常也以训练课程的成绩作为测验的效标。

训练成绩作为效标，其具体指标可以是完成训练后所实施的成就测验的成绩、正式评定的等级、指导教师的评定、成功完成训练或者被中途淘汰等。多重能力倾向成套测验的效标，常常是中学或大学里特定课程的成绩，以此可以分别确定它们的预测效度。例如，言语理解测验上的分数可以同英语课程的成绩相比较，空间视觉测验上的分数可以同几何课程的成绩相比较，等等。

一般来说，在用训练成绩记录作为效标度量时，区别中间效标( intermediate criteria )和最后效标( ultimate criteria )是十分有用的。例如，空军飞行员测验或医学能力倾向测验，最后效标应该分别是战斗成绩和开业医生的最终成就。显然，获得这种效标数据需要很长时间。而且，在实际工作中是否能获得真正的最后效标，也是令人怀疑的。最后，即使可以得到这种最后效标，它也常会受到许多不可控制的因素的影响而没有多大价值。例如，很难评价在不同地区和不同专业的从业医生的相对成功的程度。由于这个原因，中间效标，即某一训练阶段的成绩记录，经常被用作效标度量。

对于许多测验目的来说，最令人满意的效标度量是实际工作表现( job performance ) 的追踪记录。这种效标在某种程度上用来对一般智力测验和人格测验进行效度分析，在很大程度上也用来对特殊能力测验进行效度分析。在对用于具体工作的专门化测验进行效度分析时，它也是一种常用的效标。所讨论的‘工作’在水平和种类上大不相同，包括商业工作、工业工作、专业工

作、军事工作等。虽然大多数工作表现也不一定代表最后效标，但对于许多测验目的来说，它们至少是良好的中间效标。在这一点上，实际工作表现比训练成绩更为可取。另一方面，工作表现测量做不到像训练期间那样尽可能保持条件一致性。而且，由于它通常要作较长时间的追踪，工作表现的效标上可以得到的参加者的人数可能会减少。因为名称相同的工作的性质，在不同组织中有所差别，所以测验手册以工作效标来报告效度时，不仅应该报告所使用的具体效标度量，而且应该报告员工从事的工作的职责。

采用对照组(contrasted groups)方法进行效度分析，一般包括一个合成效标，它反映日常生活中累积的、不可控制的选择影响。这种效标最终是根据某特定团体之中的幸存者与从中被淘汰者。例如，把音乐学院或工程学院学生的分数分别与未经选择的中学生或大学生的分数相比较，就可以得出音乐能力倾向测验或机械能力倾向测验的效度。当然，选取对照组，可以根据任何效标，例如学校成绩、评定或工作表现，只要选择效标度量分布的两端就行。不过，现行类别中包括的对照组，是由于日常生活的多项要求的作用而逐渐成为具有特点的团体。因此所考虑的效标比先前讨论的效标更复杂，而且难以明确界定。

在人格测验效度分析中，也经常使用对照组方法。因此，在进行一个社交特质测验的效度分析时，可以把推销员或行政官员的测验分数与办事员或工程技术人员的测验分数相比较。这样一种做法的基本假设是，在许多社交特质上，从事售货或行政工作之类职业的个体，整体上优于从事工程或办公室工作之类职业的个体。同样，参加许多课外活动的大学生可以与在大学学习的同一期间不参加任何活动的大学生相比较。编制兴趣测验，例如斯特朗职业兴趣测验(Strong Vocational Interest Blank,

SVIB),经常采用各种职业对照组进行效度分析,编制态度量表也是如此。态度量表有时也采用其他对照组进行效度分析,例如政治团体、宗教团体、地理团体,以及众所周知在某些问题上代表不同观点的其他特别团体。

领域参照测验的经验效度分析,除了采用通常的内容效度分析程序之外,也采用对照组方法的若干改变形式(Hambleton, 1984b)出于这种目的,可以比较某种教学量不同的各组的测验分数。如果应用掌握分数,则可以作一个  $2 \times 2$  分析,把教学前团体的通过或失败的分数比例同教学后通过或失败的分数比例相比较(Panell & Laabs, 1979)。对某一年级教授测验所评定的特定的概念或技能,然后对低一年级和高一年级的学生实施测验,也可以进行同样的比较。如果在若干不同的教学周期之后可以得到测验分数,那就可以算出实际成绩与教学量之间的相关。

在编制某些人格测验时,精神病诊断(psychiatric diagnosis)既是选择项目的基础,也是测验效度的证据。假如精神病诊断是根据长期的观察和详尽的个案史,而不是仓促的会谈或检查,它可以成为令人满意的效标。在后一种情况下,就没有理由认为,在表示个体情绪状态方面,精神病诊断比测验分数本身更优越。这类精神病诊断不能作为效标度量,而只是一种连它自己的效度也有待确定的指标或预测。

在讨论其他的效标类型时,我们已经提到由学校教师、专业课程的指导教师、工作单位的管理人员等作出的某些评定(ratings)。另外还有军事单位的官员的评定,学校咨询人员对学生的评定,由同事、同学、同一俱乐部成员以及其他团体会员出的评定。前面所讨论的评定,只是获得学业成绩、专业训练的成绩、工作成功等效标资料的一种辅助方法。现在我们考虑把

评定作为效标度量的核心。在这种情况下,评定本身就界定效标。而且,这类评定不局限于评价具体成就,还包括观察者关于心理测验所要测量的某种特质而作出的个人判断。因此,可以根据支配性、机械独创性、创造力、领导能力、诚实等特性,来评定效度样本的被试。

评定几乎可以用于每一种测验的效度分析,特别适用于为 人格测验提供效标,因为人格测验的客观性效标很难找到。对于独特的社交特质,情况尤其如此。在这种情况下,根据个人接触作出的评定,也许是逻辑上最为正确的效标。虽然评定常会发生判断误差,但是在严格控制条件下得到的评定,仍然是效标数据的一种有价值的来源。第十六章将讨论改进评定精确性和减少常见误差的各种方法。

最后,一个新测验同先前已有的测验的相关,也经常作为效度的证据。当新测验是目前已有测验的简化型式时,完全可以把后者作为一种效标的度量。因此,纸笔测验可以用先前已经建立效度的、较复杂且费时的操作测验作为效标。同样,团体测验可以用个别测验作为效标。例如,“斯坦福一比纳量表”(Stanford-Binet)反复作为各种团体智力测验效度分析的效标。在这种情况下,新测验充其量也不过是与先前测验相近似而已。应当指出,只有新测验比先前测验更简单、更省时,用后者作为效标才是合适的。

在 20 世纪 80 年代和 90 年代,测验编制中的一个重要发展是注重于效标分析(criterion analysis)。在传统的测验研究中,通常正是忽略了测验编制的这一方面。在几年时间里,不时有人主张进行系统的效标研究,但实际上做得甚少(L. R. James, 1973; Tenopir, 1986)。甚至在精心设计的测验计划中,虽然把仔细的工作分析作为测验编制的一条准则,但工作分析的结果,对

于后来测验效度分析中所使用的效标度量没有多少影响。效标通常被认为是‘理所当然的’;人们一而再、再而三地把基于评定或成就记录的工作表现这个单一综合指标作为效标。

现在人们普遍认识到,首先确定某种工作表现的主要结构,然后选择或编制测验,而该测验的分数要评定所需要的结构,这样就能最有成效地研究测验的效度(J. P. Campbell, 1990a; J. P. Campbell, McHenry & Wise, 1990; L. V. Jones & Appelbaum, 1989; Messick, 1995)。把详尽的效标研究作为编制成套测验的第一步,一个著名的例子便是“美国军队选拔和分类计划”(U. S. Army's Selection & Classification Project),通常称之为“计划 A”(J. P. Campbell, 1990b)。因为它与在工业和组织中使用测验有关,所以第十七章将较为全面地讨论这个大规模的七年计划。

效度的概化(validity generalization) 效标预测效度时常用于局部效度的研究,即评价一个测验适用于一个具体计划的有效性。例如,某公司希望评价某个选拔应聘人员的测验,或者某大学希望确定某个学业能力倾向测验在多大程度上能够预测学生的课程成绩,此时就采用这种方法。效标预测效度最好被描述为一个测验适用于指定目的时的实际效度。

在工业测验的效度研究中,最初计算标准化能力倾向测验和大体类似的工作成绩之间的相关,发现效度系数大不相同(Ghiselli, 1959, 1966)。当效标是学校中不同课程的成绩时,同样可以观察到效度系数的差异(G. K. Bennett, Seashore & Wesman, 1984)。这些研究结果导致人们普遍对测验效度可以概化到不同情境抱悲观态度。直到 70 年代中期,一般还认为“情境具体性”(situational specificity) 是标准化测验用于人员选拔的严重的限制因素。然而,施米特(Schmit)、亨特(Hunter)及其同事对这个问题进行了复杂的统计分析,结果表明,得出的效度系数

之间的大量差异都是一种统计计算的产物,即由于样本太小,效标不可靠和受聘人员样本分布范围有限造成的。<sup>[1]</sup>

适用于测验效度分析的工业样本一般都很小,以致不能稳定地评估预测工具和效标之间的相关。由于相同的原因,得出的相关系数可能太低,达不到所采用样本的统计显著性,因此也不能提供测验效度的证据。据估计,在工业测验效度研究中使用的大约一半的效度样本,包括的个案数不超过 40 ~ 50 (Schmidt, Hunter & Urry, 1976)。对于如此小的样本,使用效标预测效度在技术上也是行不通的。

施米特、亨特及其同事从各行各业抽取大量样本,运用他们新发展的技术来处理样本数据,研究结果表明,言语测验、数字测验和推理能力倾向测验的效度能够概化到不同行业的程度,比先前认识到的要大得多。他们也证明,在早期工业测验效度研究中普遍发现的效度系数的差异,与所预期的偶然差异不相上下。即使各种工作之间特定的工作机能看来大不相同,情况也是如此。显然,各行各业上的成功表现,在很大程度上取决于各种认知技能的一种共同内核。在他们的研究中所采用的测验,主要包含传统智力测验和学术能力倾向测验所取样的内容和技能。由此可见,这组认知技能和知识,能够广泛预测先进的技术社会所需要的种种学术活动和职业活动中的表现。考虑一下在二三个一般认知测验上的分数,最好再测量一下适合于特定工作的具体的技能,通常就能作出更为准确的选拔决策 (Hartigan & Wigdor, 1989; L. L. Wise, McHenry & Campbell, 1990; Zeidner & Johnson, 1991)。

元分析(meta-analysis) 在研究效度概化时使用的统计程序,本质上是一种综合不同研究结果的方法。可以把过去研究和现在研究的资料相结合,把在不同地方进行的研究资料相结

合,也可以把各种已发表论文的研究资料相结合。虽然基本程序在 70 年代开始称之为元分析,并被引入心理学研究 (Glass, 1976; Schmidt & Hunter, 1977),但它已经使用了几十年,尤其在其他科学中 (Hartigan & Wigdor, 1989, 第六章)。元分析取代传统的文献研究,在心理学中受到越来越多的关注 (Lipsey & Wilson, 1993; Schmidt, 1992)。这种研究方法一般报告在统计上有显著意义的研究数字,例如,实验组和控制组的平均数的差异,或者测验分数和其他变量的相关。使用传统的研究方法,则经常会失去有价值的、正面的研究结果,因为在一次研究中使用的样本太少,就不会产生显著差异。

在元分析中,把发表的若干研究的结果相结合,并且在可能的条件下,根据每个研究中有关的方法特点和本质特点,对研究结果进行加权,这样就可以揭示重要的、正面的研究结果。元分析的另一个优点是可以计算影响大小 (effect sizes)。对于理论和实际两种目的来说,比起简单地说明它远大于零,估计出其差异或相关的大小则更为有用。

在 80~90 年代的十几年中,元分析研究几乎在心理学的每一个领域都迅速发展。它在职业选拔和分类研究中的应用,也许引起最为广泛的关注 (参见第十七章)。现在,人们对元分析的兴趣日益增加,研究方法也正在不断改进之中。虽然关于具体的统计技术还有一些争议,但是各种方法的主要研究结果并无明显分歧。<sup>[2]</sup>

## 结构鉴别程序

现行的“测验标准”(Testing Standards)的第一版,即 1954 年



“心理测验和诊断技术的方法推荐”(Technical Recommendations for Psychological Tests and Diagnostic Techniques, APA, 1954)正式将“结构效度”一词引入心理测量词汇( APP, 1954)。第一次对结构效度的详尽说明,出现在次年克龙巴赫和米尔( Cronbach & Meehl)的一篇论文中( 1955)。随后对结构效度的讨论——现在仍方兴未艾——使结构效度的含义更为明确,也为结构效度的使用提供了系统的理论基础。结构效度关注心理学理论在测验编制中的作用,也关注提出假设的需要,这种假设在效度分析中能够被证明成立或不成立。结构效度也促使人们去寻找收集效度资料的新方法。虽然研究结构效度中所使用的若干技术早已熟悉,但结构效度的有效领域已经扩展到认可多种程序。

测验的结构效度是指测验能够测量到理论上的结构或特质的程度。这类结构有学业能力倾向、机械问题理解、言语流畅、步行速度、神经敏感症、焦虑等等。每个结构的提出,都是为了理解和组织观察到的反应一致性,从行为度量之间所建立的相互关系中推导出结构。结构效度需要逐渐积累各种来源的资料。任何阐明所考虑的特质的性质的资料,以及影响该特质发展和表现的条件,都是结构效度的合适证据。下面几节讨论有助于结构鉴别的若干具体技术。

**发展变化** 在一些传统的智力测验中,效度分析的一种主要效标是年龄差异。例如,斯坦福—比纳测验和大多数学前测验,便是通过考察实际年龄来确定测验分数是否随着年龄的增长而增加。因为一般认为各种能力在儿童期随年龄而增长,因此如果测验有效,测验分数也应该随年龄而增加。由比纳首创的智力年龄量表的概念,就是假设‘智力’随年龄而增加,至少到成年期是如此。

当然,年龄差异效标不适合那些没有表现出明确的、一致的

年龄变化的机能。例如，人格测验中很少使用年龄差异。而且，应当指出，即使在使用年龄差异时，它也只是效度的一个必要条件而不是一个充分条件。因此，如果测验分数没有随着年龄而增加，这个结果也许表明测验不是能力样本的有效测量。另一方面，如果证明测验测量了随着年龄而增加的某种东西，这并不能准确界定测验所包括的领域。身高或体重的度量也随年龄有规律地增加，但是它们显然不能称为智力测验。

最后还应当强调的一点是对年龄效标的解释。采用年龄效标的测验，只有在该测验标准化的环境下，才测量随年龄而增长的行为特性。由于不同的文化培养和促进不同的行为特性的发展，因而不能假设年龄效标是普遍适用的。像所有其他效标一样，年龄效标也受到这种效标的特定文化背景的限制。

发展分析也是第三章和第九章所讨论的皮亚杰量表的结构效度分析的基础。这种量表的基本假设就是发展的顺序型式 (sequential patterning)，即到达概念发展中的前一阶段是获得后一阶段概念技能的先决条件，因此这些量表在内容上具有内在的层次。因而顺序量表的结构效度应该包括相继阶段的顺序不变性的经验资料。这就要检查在所测验的某概念例如守恒性或客体永久性的发展中，不同水平儿童的表现：掌握某一水平概念的儿童，也掌握较低水平的概念吗？

与其他测验的相关 一个新测验与类似的旧测验之间的相关，有时可以用来证明新测验与其他冠以相同名称的测验例如“智力测验”或“机械能力倾向测验”；大体测量相同的行为领域。与效标预测效度中得出的相关不同，这些相关应该适度，而不是太高。如果新测验与已有测验的相关太高，而又没有另外的优点例如简单或易于实施，那么就没有必要编制这个新测验。

与其他测验的相关还有另一种用途，就是表示新测验相对

不受某些无关因素的影响。例如，特殊能力倾向测验或人格测验与一般智力测验或学业能力倾向测验之间，不应该有高相关。同样，阅读理解不应该明显影响这种测验的成绩。因此，与一般智力测验、阅读测验或言语理解测验的相关，有时作为效度的间接证据或反证来报告。在这种情况下，高相关会使测验受到怀疑。然而，低相关本身也不会保证效度。应当指出，与其他测验相关的这种用途，类似于内容描述程序所讨论的辅助技术之一。

**因素分析** 因素分析是确定心理特质的一种方法，因而特别适用于结构效度。因素分析本质上是分析行为资料相互关系的一种精确的统计技术。例如，如果给 300 名被试施测 20 个测验，第一步是计算每一测验与其他各个测验的相关系数。检查得出的 190 个相关系数表，可以发现测验的某些组群，这表明存在某些共同特质。因此，如果词汇测验、同义词测验、反义词测验、句子完成测验等彼此之间具有高相关，而与所有其他测验具有低相关，我们则可以暂时推断存在一种言语理解因素。因为对相关表作这种检查分析是十分困难的，又是不确定的，因而已经发展了更精确的统计技术来找出共同因素，它们可以解释所得出的相关。第八章将进一步讨论因素分析的若干技术，及其在研究智力性质中的用途。

在因素分析中，用于描述每个被试表现的变量或类型的数目，从最初的测验数目减少到几种因素或共同特质。在上述例子中，5~6 个因素就足以解释 20 个测验之间的彼此相关。因此可以用 5~6 个因素的分數而不是最初的 20 个分數来描述每个被试。因素分析的主要目的是，把类型的数目从最初众多的测验变量减少到若干种共同因素即特质，从而简化对行为的描述。

在因素被确定之后，就可以用来描述测验的因素组成。因此可以用决定测验分数的主要因素，加上每个因素的权重即负

荷以及测验和每个因素的相关，来表示每个测验的特性。这种相关有时也称为测验的因素效度（factorial validity）。因此，如果言语理解因素与词汇测验的相关为 0.66，那么这个词汇测验测量言语理解特质的因素效度为 0.66。应当指出，因素效度实际是该测验与一组测验所共有的特性或行为的其他指标的相关。当然，这组经过分析的变量可以包括测验资料和非测验资料，因此可以利用评定和其他的效标度量，同别的测验一起，探究一个特定测验的因素组成，并定义它所测量的共同特质。

**内部一致性** 某些测验尤其是人格测验，在出版的说明书中写道，用内部一致性方法得出测验效度。这种方法的基本特征是，效标恰恰是在自身测验上的总分。有时也采用对照组方法的一种改变形式，即根据测验总分来选取高分组和低分组，然后比较在每个测验上高分组和低分组的成绩。如果高分组的“通过”（或标准反应）率，没有明显大于低分组，则认为项目无效，应淘汰或修改。相关方法也可以应用于这种目的。例如，可以计算每个项目上“通过或失败”与测验总分的二列相关，只保留与测验显著相关的那些项目。用这种方法选择测验项目，我们就说测验具有内部一致性，因为在每个项目上得高分的被试，在整个测验上也得高分，反之亦然。

内部一致性效标的另一种用法是分测验分数与总分的相关。例如，许多智力测验由分别实施的分测验（例如词汇、算术、填图等）组成，分测验的分数联合起来就得出测验总分。在编制这种测验时，常常计算每个分测验上分数与总分的相关，删除与总分相关太低的分测验。剩余的分测验与总分的相关就作为整个测验内部一致性的证据。

显而易见，内部一致性相关，无论根据项目还是根据分测验，本质上都是对同质性的度量。因为它有助于描述测验所取

样的行为领域或特质，测验的同质性程度就同测验的结构效度有某种关系。然而，内部一致性资料对于测验效度的作用十分有限。在缺少测验本身的外部资料时，关于测验到底测量什么，我们知之甚少

**幅合效度和区别效度** 在认真分析结构效度时，坎贝尔(D. T. Campbell, 1960)指出，为了证实结构效度，我们必须表明，一个测验不仅与理论上应该相关的那些变量具有高相关，而且与理论上应该区别的那些变量具有低相关。在较早的一篇文章中，坎贝尔和菲斯克(Fiske, 1959)把前者称为幅合效度(convergent validation)，把后者称为区别效度(discriminant validation)。例如，一个数量推理测验与后来数学课成绩的相关，就是幅合效度；而同一个测验与阅读理解测验分数的较低的、不显著的相关，则是区别效度，因为在设计测量数量推理的测验时，阅读能力是一种无关变量。

回忆一下，在内容效度分析的辅助方法和预防方法中，我们讨论过与无关变量低相关这一要求。区别效度分析也特别适用于人格测验的效度分析，因为无关变量会以多种方式影响人格测验的分数。

坎贝尔和菲斯克(1959)提出一种适合于幅合效度分析和区别效度分析双重方法的系统实验设计，他们称之为多重特质—多重方法矩阵(multitrait-multimethod matrix)。此种方法必须采用两种或两种以上的方法来测量两种或两种以上的特质。现以坎贝尔和菲斯克所假设的例子来加以说明。表5-1表明用三种方法测量三种特质时所得分数之间所有可能的相关。这三种特质是三种人格特性，即(A)支配性(B)社交性、(C)成就动机。三种方法则是(1)自陈问卷、(2)投射技术、(3)同伴评定。因此， $A_1$ 表示自陈问卷上的支配性分数， $A_2$ 表示投射测验上的支配性

分数,  $C_3$  表示同伴评定的成就动机分数。

表 5-1 中列出的假设相关包括信度系数(主对角线上圆括号里的数字)和效度系数(三条较短的对角线上的黑体数字)。这些效度系数是用不同方法测量相同特质所得分数的相关;对同一种特质的每种度量因此都用其他的独立度量来检查,如同在熟悉的效度分析方法中所做的那样。这张表也包括用相同方

表 5-1 假设的多重特质—多重方法矩阵

|      |                | 方法 1           |                |                | 方法 2           |                |                | 方法 3           |                |                |
|------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 物质   |                | A <sub>1</sub> | B <sub>1</sub> | C <sub>1</sub> | A <sub>2</sub> | B <sub>2</sub> | C <sub>2</sub> | A <sub>3</sub> | B <sub>3</sub> | C <sub>3</sub> |
| 方法 1 | A <sub>1</sub> | (.89)          |                |                |                |                |                |                |                |                |
|      | B <sub>1</sub> | .51            | (.89)          |                |                |                |                |                |                |                |
|      | C <sub>1</sub> | .38            | .37            | (.76)          |                |                |                |                |                |                |
| 方法 2 | A <sub>2</sub> | .57            | .22            | .09            | ( .93 )        |                |                |                |                |                |
|      | B <sub>2</sub> | .22            | .57            | .10            | .68            | ( .94 )        |                |                |                |                |
|      | C <sub>2</sub> | .11            | .11            | .46            | .59            | .58            | (.84)          |                |                |                |
| 方法 3 | A <sub>3</sub> | .56            | .22            | .11            | .67            | .42            | .33            | ( .94 )        |                |                |
|      | B <sub>3</sub> | .23            | .58            | .12            | .43            | .66            | .34            | .67            | ( .92 )        |                |
|      | C <sub>3</sub> | .11            | .11            | .45            | .34            | .32            | .58            | .58            | .60            | (.85)          |

注 字母 A、B、C 表示特质,下标 1、2、3 表示方法。对角线上三组黑体数字表示效度系数(相同特质—不同方法)主对角线上圆括号里的数字表示信度系数(相同特质—相同方法)。实线三角形内的数字表示不同特质—相同方法的相关;虚线三角形内的数字表示不同特质—不同方法的相关(摘自坎贝尔和菲斯克,1959,p.82)。

法测量不同特质所得的相关系数(实线三角形内的数字)和用不同方法测量不同特质所得的相关系数(虚线三角形内的数字)。对于令人满意的结构效度来说,效度系数应该显著高于用不同方法测量不同特质的相关系数;它们也应该高于用相同方法测量不同特质的相关系数。例如,自陈问卷上支配性分数与投射测验上支配性分数的相关,应该高于自陈问卷上支配性分数与社交性的相关。后者表示共同方法的方差,如果很高的话,则可能表明被试在这个问卷上的分数,过分受到某种共同的无关因素的影响,例如理解问题的能力或希望自己在所有特质上都获得社会赞同。

**实验干预** 通过实验来研究经过选择的变量对测验分数的影响,这是提供结构效度资料的另一种方法。例如,在个别化教学程序中使用一个测验,现在要检查该测验的效度。一种方法是比较前测和后测的分数,这种测验的基本原理要求,在前测即有关教学之前实施的测验上得低分,而在后测上则得高分。这种关系也可用来检查测验中的每个项目。理论上,绝大多数被试在前测时不应该通过某个项目,而在后测时则应该通过这个项目。对于这种测验的目的来说,在两次测验中都没有被通过的项目太难了,而在两次测验中都被通过的项目则太容易了。如果相当多的被试在前测时通过某个项目,但是在后测时却没有通过这个项目,那么显然这个项目或这种教学抑或两者都有点不当。

对被试实施一个焦虑倾向测验,随后让被试置身于一种唤起焦虑的情境,例如让被试在干扰和紧张条件下参加考试。然后计算最初的焦虑测验分数与考试期间和考试之后焦虑表现的生理指标和其他指标之间的相关。要评价关于焦虑测验的一种不同的假设,可以在唤起焦虑体验前后分别实施测验,再看看在

后测时测验分数是否显著提高。这个实验的研究结果果真如此，那就表明测验分数反映被试当前的焦虑水平。采用同样的方法，可以设计实验来检验某测验所测特质的其他假设。

结构方程建模除了用测验分数和效标成绩来鉴定结构之外，测验效度分析中的一种重要进展是，考虑结构之间的关系和一个结构影响效标成绩的方式（J. P. Campbell, 1990a; Messick, 1989; Schmidt, Hunter & Outerbridge, 1986）。例如，一个体对特定领域的兴趣可以通过不断学习有关的实际知识，获得必不可少的技能，或由于动机的激发而尽最大努力进行必要的活动，从而影响工作成绩。研究一种经过鉴定的结构或人格特质怎样导致好的或差的成绩，这大大有助于理解为什么一个测验在某种情境中具有高效度或低效度。一种称为结构方程建模（structural equation modeling, 1994）的统计方法促进了这种分析，在 80 年代和 90 年代越来越多地使用这种模型，给人留下深刻印象。《结构方程建模》杂志的创刊（1994）就是一例。这种方法同路径分析的若干变式关系密切，两者常常非正式地称之为“因果建模”（causal modeling）。<sup>[3]</sup>

具体来说，设计结构方程建模要解决什么问题？它又是怎样提出来的？人们很快认识到，初级统计中的相关不表示因果关系。一个熟悉的例子是假性的年龄相关。年龄从 6 岁到 14 岁的学校儿童，他们的身高与算术能力之间的相关可能很高；但是我们不能得出结论：其中一个变量影响另一个变量。相关主要是年龄造成的，而年龄又与学校教育量有关系。在努力研究因果关系中，研究人员在 60 年代和 70 年代开始使用跨时间滞差的实验设计（cross-lagged experimental designs, D. T. Campbell & Stanley, 1966; Cook & Campbell, 1976, pp. 284 ~ 293）。例如，为了分析学生的数学态度与数学成绩之间相互的因果影响，可以在



两个时间点测量数学态度和数学成绩；然后分别计算时间 1 的数学态度和时间 2 的数学成绩之间，以及时间 1 的数学成绩和时间 2 的数学态度之间的跨时间滞差的相关。两个相关的大小就表示每个方向上影响的相对强度。在几年时间里，这看来是评定两个变量相互影响的一种有前途的方法。

然而，不久以后，逻辑分析和统计分析都表明，使用跨时间滞差相关存在着严重缺点。虽然，基本的跨时间滞差设计是优秀的，但是使用简单的、顺序抵消的相关可能歪曲结果，并且得出关于因果关系的错误结论(Rogosa, 1980)。这种方法中的误差来源于没有考虑到：第一，最初的变量之间的相关和随后的变量之间的相关 第二,变量的信度及其跨时间的稳定性 第三,没有测量到的变量可能产生的作用，例如上述经典例子中的年龄和学校教育。结构方程建模提供了解决这类困难的方法。之所以如此，本质上是由于使用回归方程，根据跨时间滞差模型或其他因果模型中的自变量来预测因变量。在这种方法中，使用偏相关来得出回归系数，由此考虑到变量之间的所有相关，以及测量误差和取样误差，至少认识到，还存在另外的、没有测量到的因果变量(Bentler, 1988; L. R. James 等, 1982; Loehlin, 1992; Rogosa, 1979)。

结构方程建模的第一步是设计一个有待检验的、假设的因果关系模型。必须指出，建立该模型的基础是，完全熟悉所研究的变量和情境的现有知识。假设的因果关系应该具有成熟的理论基础。解答一组线性回归联立方程式而对模型进行实际检验。<sup>[4]</sup>在因果模型中，方程式通常比未知数多，这准许解答出若干不同的模型。把每个模型与最初的、通过实证得出的相关矩阵相比较，以便获得较好的拟合。然而，若干因果模型可以同样较好地拟合资料 (MacCallum, Wegener, Uchino & Fabrigar, 1993)。

这些统计上等值的模型可能表示不同的因果路径，因而对经验上观察到的效果作出不同的解释。然后根据对情境的认识，研究人员必须对不同模型的合理性和实际意义作出评价。

结构方程建模的另一特点是，计算结构之间典型的因果关系，而不是计算孤立地测量的变量之间的因果关系。例如，要评价一个学生的数学态度，可以采用若干指标，诸如兴趣、目标定向、数学能力倾向的自我概念，以及其他有关的情感变量等度量。这些指标之间的共同方差就界定该学生数学态度的一种结构，它本身与随后的数学成绩有关。这样使用结构，各个指标的误差方差和特定方差相互抵消，因而提供更为稳定和可靠的估计值。

关于结构方程建模，以及个别修改和程序上的改进，目前有若干种方法(例如,参见 Andeerson & Gerbing, 1988; bEntler, 1990, Bollen & Long, 1993; Breckler, 1990; Cole, Maxwell, Arvey & Salas, 1993; James, 1980, Mulaik 等, 1989)。虽然结构方程建模仍处于发展状态,但这是理论、实验、统计三结合的一种有前途的方法。它已经广泛应用于发展心理学、人格和社会心理学、工业心理学、教育心理学等领域的种种问题(例如, Graves & Powell, 1988; L. A. James & L. R. James, 1989; MacCallum & Browne, 1993; McCordle, 1989; Parkerson, Lomax, Schiller & Walberg, 1984; Shavelson & Bolus, 1982)。还有人致力于方法的一致和简化(例如, Joreskog & Sorbom, 1993)。

认知心理学的贡献在 70 年代,实验心理学和心理测量学相互融合,这对于理解智力测验和其他广义的能力倾向测验所评估的结构,开始产生重要的作用( Ronning, Glover, Conoley & Witt, 1987; R. E. Snow & Lohman, 1989)。早在 50 年代,认知心理学家就开始运用信息加工的概念来研究人类的问题解决。一些

研究人员设计出计算机程序来执行问题解决过程,由此模拟人类思维。所编制的程序可以模拟不同技能水平个体的表现,运用这类程序也有可能预测不同反应所发生的错误的数目和类型,以及不同反应所需要的时间。在设计程序时,研究人员通常从任务分析开始,包括内省自己解决问题的方法,要求被试出声思维,或者运用更为精确的观察方法。把解决同一个问题时计算机的表现同儿童和成人相比较,或者同专家和新手相比较,研究人员就能检验关于完成任务时人们做了些什么的假设。这些方法中研究的作业通常包括传统的难题、逻辑问题、象棋比赛、代数文字题、物理问题、医学实践中的诊断问题等(Chi, Glaser & Farr, 1988; J. H. Larkin, McDermott, Simon & Simon, 1980a, 1980b; Newell & Simon, 1972, Simon, 1976)。

这些研究所确定的变量包括过程(程序技能)和陈述性知识(事实和信息)。认知模型详细说明完成任务的智力过程,组织这些过程的方式、相关知识的储存以及这种知识如何在记忆中表征并在需要时被提取出来。也越来越注意所谓的执行过程或元认知,它是指个体控制他自己在完成任务时对过程、表征和策略的选择。在 70 年代,一些认知心理学家开始应用这些任务分析和计算机模拟技术来探索智力测验到底测量什么东西。不同的研究人员从不同的角度来探讨这个问题(参见 Resnick, 1976; Sternberg, 1981, 1984, 1985b)。研究结果不断推进测验的编制和使用。

恩布里森(Embretson)的工作十分清楚地说明了认知研究对于结构效度分析的含义(1983, 1986, 1995a)。恩布里森认为传统的结构效度分析方法过于局限,他提出测验效度分析的两个主要方面:(1)结构表征;(2)一般规律研究的范围。传统的结构效度分析正是着重于第二方面即一般规律研究的范围。这涉及到在其他变量的“一般规律研究网”之内测验成绩的关系。研究这

种关系,一般通过测验分数与其他度量如效标成绩和其他现实生活资料的相关。

另一方面,结构表征的目的,则是识别出完成测验项目提出的任务所需要的具体的信息加工组份和知识储存。在实验中可以应用任务分解 (task decomposition)<sup>[5]</sup>来进行这类分析。通常可能的做法有操纵任务复杂性、呈现部分任务,或提供改变任务要求的暗示。已经设计出一些数学模型来测量不同的反应组份对项目成绩的作用。另一种广泛使用的认知任务分析的方法是口语记录分析(protocol analysis)(Ericsson, 1987; Ericsson & Simon, 1994; van Someren, Barnard & Sandberg, 1994)。在这种方法中,要求被试在完成一项任务或解决一个问题时“出声思维”。任务的范围从对两个给出的数字做心算乘法,回忆一个过去事件的细节,或找出一个设备功能障碍的原因,到回答一个智力测验上连续的项目。这种方法的副产品是可能发现,经验背景有所不同的被试,对于同一测验项目可以进行截然不同的认知过程。

关于认知心理学至今对结构效度的作用,我们能够得出什么结论呢?信息加工方法尚处于探索阶段,一些启发式概念有助于指导进一步研究。认知心理学显然着重于反应过程,而传统的心理测量学研究则着重于思维的最后结果。按照具体的认知过程来分析测验作业,当然会加强和拓宽我们对测验测量什么东西的理解。而且,在基本的组份水平上分析个体的作业,最后应该有可能指出每一被试的长处和短处,由此增强测验的诊断功用(Embretson, 1987, 1994; Estes, 1974; Pellegrino & Glaser, 1979; Sternberg & Weil, 1980)。这又转过来促进我们去编制适合个体需要的训练程序。总之,第一,从应用研究和实践的观点来看,心理测量方法和认知方法之间的关系是互补(complementary)的。这时,每种方法对于不同的评定目的都有其长处。

二，从基础研究和理论的观点来看，它们的关系则可以表示为交互(reciprocal)。这时，每种方法都有助于澄清和丰富另一种方法；两者相结合，可以使我們更好地理解智力行为。

## 总结和整合

各种效度分析方法的比较“这个测验的效度如何？”我们考虑了这个问题的几种问法。一个测验由 50 道各类算术问题组成，让我们依次应用各种方法于这个测验，以说明不同的效度分析方法的不同特点。表 5-2 说明可以适用于该测验的 4 种方法，以及适合于每种方法的效度分析类型。这个例子说明效度分析方法的选择取决于测验分数的用途。同一个测验，适用于不同的目的时，就应该使用不同的效度分析方法。如果一个成就测验用于预测后来的高等教育水平的成绩，例如大学招生，那么对这个测验的评价就需要以后来的大学成绩为效标，而不是按照测验的内容效度。

表 5-2 一个算术测验适用于不同目的的效度分析

| 测验目的            | 解说性的问题                 | 效度的证据  |
|-----------------|------------------------|--------|
| 小学算术成就测验        | 迪克在过去学会了多少？            | 内容描述   |
| 预测中学数学成绩的能力倾向测验 | 珍妮在将来会学得好吗？            | 时间效标预测 |
| 诊断学习无能的技术       | 比尔的成绩说明具体的学习无能吗？       | 同时效标预测 |
| 数量推理的度量         | 海伦的分数与她的推理能力其他指标的关系如何？ | 结构鉴定   |

结构效度的含义表 5-2 中给出的例子着重于不同类型的效度分析方法之间的差异。然而,对这些方法的进一步考虑则表明,内容关联效度、效标关联效度和结构关联效度并不是性质不同的或逻辑上并列的类型。相反,结构关联效度是一个综合的概念,它包括其他两类效度。本章前面所讨论的、适用于内容分析和测量效标关系的各种具体技术,也可以列在结构鉴定之下。例如,机械能力倾向测验同工艺课程成绩和各种工作成绩的相关,有助于我们理解测验所测量的结构。把成功员工和不成功员工的对照组的分数相比较,则会进一步证实结构鉴定。

测验手册通常采用各种实际效标的效度,帮助有可能使用测验的人士理解测验测量什么东西。虽然测验使用者也许不会直接预测测验所采用的某一具体的效标,但是检查一下这类效标,则能够建立测验所取样的行为领域的概念。如果沿着这条思路再进一步,我们就能够看到,所有的测验使用和对测验分数的所有解释都意味着结构效度,这一点正被人们日益认识到(J. P. Campbell, 1990a; Guion, 1991; Messick, 1980b, 1988, 1989, Tenopir, 1986)。因为使用测验的条件难得与收集效度资料的条件完全相同,某种程度的结果概化就是不可避免的。测验分数的解释意义总是根据各种结构,而结构的广度或概化程度在各种行为领域、人群总体、背景等方面又大不相同。

梅西克(Messick, 1980b, 1989)令人信服地证明“效度”一词,就它指出一个测验的解释意义而言,应该用结构效度替代。他坚持认为,传统上与“效度”一词有关的其他方法,应该用描述较具体的名称来称呼。因此,内容效度可以称之为“内容恰当”和“内容范围”,分别表示领域细目表和领域代表性。效标关联效度可以称之为“预测效用”和“诊断效用”,相当于预测效度和同时效度。这些描述较清楚的名称无疑有助于更好地理解各种

方法实际上做了些什么。然而,不同类型的效度分析方法的不同之处,有助于找出适合于特定目的的测验。因此,测验手册中应该十分清晰地报告这种差异。

从另一个角度,即使目前的实际应用着重于内容描述(例如教育评定)和效标预测(例如职业选拔),使用广度合适的结构也比具体的测验成绩的度量来得有效。效标研究越来越清楚地表明,效标度量和测验分数若与结构相配,它们就能更有效地得到表现。而且,人们现在认为,对结构之间的因果关系的研究,例如结构方程模型,十分有助于理解为什么要使用测验和怎样使用测验。<sup>[6]</sup>

**测验编制过程中的效度分析** 现在人们日益认识到,编制一个有效度的测验需要多种方法,而这些方法在测验编制的不同阶段将被相继使用(Anastasi, 1986a; Guion, 1991; Jackson, 1970, 1973; N. G. Peterson 等, 1990)。因此效度从一开始就融入测验,而不是限于测验编制的最后几个阶段,如传统的效标效度分析那样。效度分析过程首先根据心理学理论、先前研究,以及对有关行为领域的系统的观察和分析,提出详细的特质或结构的定义。然后准备符合这种结构定义的测验项目。接着进行经验的项目分析,从最初的项目库中选择最有效度的项目,然后可以进行其他合适的内部分析,包括对项目组或分测验的统计分析。最后阶段根据外部现实生活的效标通过统计分析,对不同分数和分数的解释性组合进行效度分析。

在编制测验和使用测验的过程中所收集到的信息,几乎都同测验效度有关。当然,内在一致性和再测信度的资料有助于定义结构的同质性及其时间稳定性。常模可以提供另外的结构说明,年龄、性别或其他人口统计变量都影响个体经验史,继而影响个体的测验成绩,常模为依据这些变量所分类的亚团体分

别提供标准化资料时,情况尤为如此。而且,测验发表供使用之后,通过逐渐积累临床观察资料,以及通过特别的研究项目,可以继续阐明和丰富测验分数的解释意义。<sup>[7]</sup>

测验的个体结果和社会结果 一些心理测量学家建议,测验效度概念中应该包括另外一个的特点,即测验对个体和对社会的结果。拓宽效度概念的一个著名代表人物是梅西克(1980b,1988,1989,1995),他特别重视使用具体测验的意料之外的结果,因为测验可能不利于某些个体,例如某个少数民族或具有不同经验背景的其他团体的成员。全国科学研究协会(National Research Council)召集的一个专家委员会非常全面地研究了这种情形,他们的报告详细分析了在评价职业应聘人员中平衡不同的目标和价值观的种种问题(Hartigan & Wigdor,1989,尤其参见第十三和十四章)。

测验使用的道德含义和社会含义当然需要引起广泛关注。第十八章将介绍这个问题,第六章则将联系“测验偏向”问题来论述技术方面的问题。然而,其他心理测量学家(例如, Cole & Moss,1989)注意到,把效度概念中的这些问题合并在一起,看来并不是解决这些问题的最为有效的方法。单单使用经验资料和统计分析尚不能回答这些问题。在统计处理中也不应该取消所期望的价值观。这些问题是一种独立的目标,需要进行明确的陈述和讨论;使用特定测验时,除了考虑严格的经验效度和统计效度之外,也需要考虑这些问题。关于平衡相冲突的目标,使用适合于处理价值系统的方法则可作出明智的决策(Mullen & Roth,1991;Zeichmeister & Johnson,1992)。<sup>[8]</sup>这类方法需要人类判断、系统的争论、妥协方案、解决冲突等,在这些方法中,不同价值系统的拥护者都得到充分的表现。使用特定测验时,把确定效度的经验的、统计上的方法,与社会和道德结果的评价相结



合,只会混淆和模糊问题。

考虑这个困难而重要的问题的一个结论是,重申第一章已经指出的测验使用者的主要作用。当需要价值判断时,尤其在个别测验中,测验使用者负有一种更重要的责任。在选择合适的测验,以及在解释测验分数时,测验使用者能够控制测验结果。不仅从科学的观点,而且从道德的观点来看,测验使用者的广泛的价值取向和社会敏感性,十分有助于恰当使用测验。<sup>[9]</sup>

## 注 释

- [1] 这项研究工作是许多论文和专论中所报告的一项连续研究计划的组成部分。和本课题关系最大的论著包括(而不是限于)皮尔曼、施米特和亨特(Pearman, Schmidt & Hunter, 1980) 施米特、加斯特-罗森伯格和亨特(Schmidt, Gast-Rosenberg & Hunter, 1980); 施米特和亨特(1977) 施米特、亨特和皮尔曼(1981); 以及施米特、亨特、皮尔曼和沙恩(Schmidt, Hunter, Pearlman & Shane, 1979)。
- [2] 关于最新的应用、对程序的详细解释、评价等, 参见哈蒂根和威格多(Hartigan & Wigdor, 1989) 赫奇斯(Hedges, 1988) 亨特和施米特(1990); 詹姆斯、德马里、穆莱克和拉德(L. R. James, Demaree, Mulaik & Ladd, 1922) 琼斯和阿普尔鲍姆(L. V. Jones & Applebaum, 1989); 罗森塔尔(R. Rosenthal, 1991) 施米特(1992) 施米特等(1993) 以及施米特、奥尼斯和亨特(Schmidt, Ones & Hunter (1992) 的有关论述。关于统计程序的简单介绍, 参见沃尔夫(F. M. Wolf, 1986) 的有关论述。关于在行为研究中应用元分析的一种主要观点, 参见库克等(Cook et al. 1992); 库珀和赫奇斯(Cooper & Hedges, 1994); 哈斯尔布莱德和赫奇斯(Hasselblad & Hedges, 1995); 以及瓦赫特和斯特拉夫(Wachter & Straf, 1990) 的有关论述
- [3] 关于某一事件的最终的或完全的因果关系, 为了避免其哲学含义, 心理学家比较喜欢中性的表述, 例如 A 决定或影响 B 这一说法。然而, 有时使用“因果”这一术语来表示所有这些关系而假设其局限性是众

- 所周知的(例如,参见 L. R. James, Mulaik & Brett, 1982, 第一章;P. A. White, 1990)。
- [4] 关于详细方法的介绍,参见博伦(Bollen, 1989)和莱林(Loehlin, 1992)的有关论述。使用计算机软件包例如 LISREL(Hayduk, 1988; Joreskog & Sorbom, 1986, 1989)和 EQS(Bentler, 1985),能够进行实际的计算。
- [5] 详情参见巴特菲尔德、尼尔森、坦根和理查森(Butterfield, Nielsen, Tangen & Richardson, 1985),恩布里森(Embretson, 1985b)和斯腾伯格(Sternberg, 1977, 1980)的有关论述。
- [6] 应用这些较复杂的测验效度分析方法的一个例子,参见金和金(L. A. King & D. W. King, 1990)的有关论述。
- [7] 关于成功应用测验效度分析的这种综合模式,参见埃利奥特(Elliott, 1990b, 第九章)的有关论述。
- [8] 关于一种较广泛的观点和另外的参考书目,也可参见阿克斯(Arkes, 1993)的有关论述。
- [9] 附带指出一下,研究心理学的一种新途径是“漫无边际心理学”(discursive psychology),即通过日常生活中个体之间的谈话和传统的实验方法来研究问题(例如,参见 Harre & Stearns, 1995; J. Smith, Harre & Van Langenhove, 1995)。

## 第六章 效度：测量和解释

第五章讨论了效度概念和效度分析资料的来源；本章讨论效度的定量表述及其解释。测验使用者在两个阶段都关心效度。第一阶段，他们考虑一个测验对于他们目的的适合性。他们检查测验手册或其他发表的报告中所报告的效度资料。通过这种信息，他们得到测验实际上测量什么心理机能的暂时概念，并判断这种机能是否适合于他们所提出的测验用途。实际上，当测验使用者依靠所发表的效度分析资料时，他们是在处理结构效度，而不管收集资料中所使用的具体方法。我们在第五章看到，不能假设所发表的研究中使用的效标同测验使用者要预测的效标完全相同，就好比两个不同的公司里相同名称的职位难得完全相同，或者不同的大学里所教授的一年级学生英语课程可能完全不同一样。因此，选择测验时需要效度有某种程度的概化。

因为测验需要的多样性和从测验分数作出论断的多样性，一些测验使用者希望以地方效标来检查某一选出的测验的效度。虽然所发表的资料有力地表明某一测验在特定情境中具有高效度，但是在技术上可行的时候，最好进行直接的证实。测验使用者评价效度的第二阶段是以具体的地方效标来确定效度。本章要讨论的技术主要是分析测验使用者自己获得的效度资料。然而，大多数技术也有助于理解和解释测验手册中所报告

的效度资料。

## 效度系数和估计误差

**关系的测量** 效度系数是测验分数和效标度量之间的相关。因为效度系数提供了测验效度的一种数量指标，所以在可以得到每种效标数据时，测验手册中通常使用效度系数来报告测验对每种效标的效度。计算效度系数中所使用的数据，也能够用预期表或预期图的形式来表示，第三章说明过这一点。事实上，这类表和图提供一种方便的方式来表明效度系数对于所测个体意味着什么。回忆一下，预期图给出在测验上获得某一分数的个体达到规定的效标作业水平的概率。例如，如果我们知道一个学生在“区别能力倾向测验”(DAT)的数学推理测验上的分数，使用表 3-6，我们就能够查出他在七年级数学课中得到某一等级的可能性。同一批数据得出效度系数为 0.60。当测验变量和效标变量都是连续变量时，可以使用熟悉的皮尔逊(Pearson)积差相关系数。当数据用不同的形式来表示，如使用通过与失败两分效标时(如图 3-7)，则可以计算其他类型的相关系数。计算这些不同类型相关的具体做法，在任何普通统计课本中都能找到。

**影响效度系数的因素** 同信度的情况一样，效度系数根据样本团体得出，因而必须详细说明样本团体的性质。同一个测验实施于性质不同的被试时，可以测量不同的机能。这些性质包括年龄、性别、教育水平、职业，或其他任何有关的特性。例如，具有不同经验背景的被试，可能采用不同的作业方法来解答同一个测验问题。结果，一个测验预测某一效标时，在一个总体

中具有高效度，而在另一总体中则效度很低甚至没有效度。或者，测验可能有效地测量两个总体的不同机能。适用于多种总体的测验，应该在技术手册中报告关于总体间概括化的合适数据。而且，在测验分数上变化很大的一个总体之内，效度系数在分数分布的不同部分明显不同，因而应该检查各个合适的亚团体中的效度系数(R. Lee & Foley, 1986)。

同信度度量一样，样本异质性的问题也与效度测量有关，因为信度和效度通常都以相关系数来报告。回忆一下，在其他条件相等的情况下，分数分布范围越大，相关也越高。解释测验手册中给出的效度系数时，应该记住这一点。

许多效度分析样本中遇到的一个特殊困难由预先选拔而产生。例如，研究一个新编职业选拔测验的效度，可能对新雇员的一个样本进行施测而最终得到职业成绩的效标度量。然而，这类雇员可能代表所有职业申请人员的一个经过选拔的优胜者样本。因此，这样一个样本的测验分数和效标度量的范围在分布的较低端就会缩小。这种预先选拔的影响由此降低效度系数。在以后的测验使用中，出于选拔目的对所有申请人员进行施测时，可以期望效度会有所提高。

由于选拔标准的变化，效度系数也会随时间而变化。试举一例，比较经过 30 年间隔所得出的耶鲁大学学生的效度系数(Burnham, 1965)。根据大学入学委员会测验和中学成绩记录作出一种预测效标，计算这种效标同大学一年级平均成绩之间的相关，结果发现这种相关在 30 年后从 0.71 降低至 0.52。对双变量分布的检查清楚地揭示出这种下降的原因。因为入学标准的提高，后面的班级在预测指标和效标成绩上的同质性都高于 30 年前的班级。因此现在这个团体的相关系数较低，虽然预测个体成绩的准确性没有发生什么变化。换句话说，观察到的

关系数的下降,并不表示预测指标的效度不如 30 年之前。假如忽视样本同质性的差异,就会得出错误的结论。

为了恰当地解释效度系数,也应该注意测验和效标之间的关系型式。计算皮尔逊相关系数是假设两个变量是线性的,并且在整个分布范围内是均匀的。对测验分数和职业效绩的关系的研究表明,这些条件一般来说可以满足 (Crowd & Sackett, 1990; Hawk, 1970)。但是,特殊的情况可以改变这种关系,测验使用者应该警惕这种可能性。例如,某项职业可能要求最低的阅读理解水平,使工作人员能够阅读说明书、标记等。然而,一旦超过这个最低水平,阅读能力的进一步提高与职业成功的程度就没有什么关系。这个例子中测验分数与职业成绩之间就是一种非线性关系。检查一下阅读理解分数对效标度量的双变量分布即散布图,可以看出在到达最低阅读能力之前,职业成绩一直随阅读能力的提高而升高;但到达该点后,职业成绩则趋于稳定,不再随阅读能力的提高而提高。因此,散布点围绕着一曲线而不是一条直线。

在另一些情况下,最优拟合线可能是一条直线,但是个体的散布点在上面偏离这条直线的范围要大于下面。假设在一个学业能力倾向测验上的成绩是成功学习某门课程的必要而非充分的条件,换句话说,测验低分的学生在该课程上的成绩必然很差,但是测验高分的学生中,有些在课程上的成绩会很好,有些则由于学习动机水平低下、缺乏学习兴趣,或其他不利因素而在课程上的成绩会很差。在这种情况下,测验高分学生的效标成绩的变异性必然大于测验低分学生。双变量分布中的这种因素称之为异质变异性。皮尔逊相关假设双变量分布的整个范围的同质变异性或均等的变异性。在现在这个例子中,双变量分布将是扇形——上面宽下面窄。对双变量分布本身的检查,通常

会很好地说明测验和效标之间关系的性质。预期表和预期图也能正确揭示在各个不同水平上测验的相对有效性。

**效度系数的大小** 效度系数应该为多大？对这个问题不可能作出一般的回答，因为解释效度系数必须考虑一些伴随的情况。获得的相关系数当然应该很高，达到某种可接受的统计显著性水平，例如第四章所讨论的 0.01 或 0.05 水平。换句话说，得出测验效度的结论之前，我们应该有理由确信，获得的效度系数不可能由相关为零的总体通过取样的偶然波动而产生。

然而，确定测验分数和效标之间的显著相关之后，我们需要根据测验的用途来评价相关的大小。如果我们希望预测个体的精确的效标分数，例如学生在大学里将获得的年级平均成绩，对效度系数的解释不妨根据估计的标准误差 (standard error of estimate,  $SE_{est}$ )，这类似于讨论信度时的测量误差。回忆一下，测量的误差表示个体分数中预期的误差大小，这是测验信度不完美的结果。同样，估计的标准误差表示预测的个体效标分数中预期的误差大小，这是测验效度不完美的结果。

估计的标准误差用下列公式计算：

$$SE_{est} = SD_y \sqrt{1 - r_{xy}^2}$$

公式中， $r_{xy}^2$  表示效度系数的平方， $SD_y$  表示效标分数的标准差。应当指出，假如效度是完美的 ( $r_{xy} = 1.00$ )，则估计的标准误差就是零。另一方面，如果一个测验的效度为零，则估计的标准误差就和效标分布的标准差一样大 ( $SE_{est} = SD_y \sqrt{1 - 0} = SD_y$ )，在这种情况下，预测与猜测无异；预测误差的变动范围和效标分数的整个分布一样大。介于这两者之间的是相应于具有各种不同效度的测验的估计标准误差。

关于  $SE_{est}$  的公式表明， $\sqrt{1 - r_{xy}^2}$  这一式子用来表示相对于

由纯猜测造成的误差而言的误差大小。换句话说，如果  $\sqrt{1 - r_{xy}^2} = 1.00$ ，则估计的误差就同我们所猜测的个体的效标分数一样大，因此由测验带来的预测提高为零。如果效度系数为 0.80，那么  $\sqrt{1 - r_{xy}^2} = 0.60$ ，误差为偶然误差的 60%。换一种说法，使用这个测验使我们能够预测个体的效标成绩，而误差的大小比猜测误差小 40%。

因此看起来似乎是，即使效度为 0.80——这是异常高的效度了——但是预测分数的误差还是相当大的。假如心理测量的基本功能是预测每一个体在效标分布中的精确位置，则前景就是十分令人沮丧的。一旦根据估计的误差来检查测验，大多数测验看来就都没有多大效率。然而，在大多数测验情况下，没有必要去预测个体的具体效标成绩，而应该确定哪些个体会超过效标上某一最低成绩标准或临界点。玛丽·格林从医学院毕业的可能性是多大？汤姆·希金斯通过微积分课程的可能性是多大？贝弗莉·布鲁斯成为宇航员的可能性又有多大？哪些应聘人员可能成为令人满意的办事员、保险代理人、机床操作人员等？此类信息不仅对团体选拔有用，而且对个体职业计划也有用。例如，对于一个学生来说，知道自己很有可能通过法学院的全部课程，这是十分有用的，即使我们不能准确估计他的平均成绩会是 74 分还是 81 分。

如果测验与效标有任何显著性相关，不管如何低，也可以明显提高预测效率。在某种情况下，甚至低到 0.20 ~ 0.30 的效度，也可以证明在选拔计划中使用测验是正确的。对于许多测验目的来说，根据估计的误差来评价测验是不现实的。必须考虑用其他方法来评价测验的作用，这类方法考虑根据分数作出决策的类型。下面一节将说明其中的一些方法。





取 45 个最有希望的人 (录取率 = 0.45)。在这种情况下, 垂直粗线右边的 45 人被录取。可以看到, 有 7 人工作失败, 即处于水平粗线之下的“错误接受”(false acceptance), 而有 38 人工作成功。因此, 工作成功百分率现在是 84 而不是 60 (即  $38 \div 45 = 0.84$ )。这种增加可归因于使用测验作为甄别工具。应当指出, 预测的效标分数中的误差并不影响决策, 因此可以忽略不计。只有越过临界线并由此把个体放入错误的类型的那些预测误差, 才会减少测验的选拔有效性。

为了全面评价测验作为甄别工具的有效性, 还应该检查图 6-1 中另一种个案类型。这个类型是错误拒绝 (false rejection), 包括 22 人, 他们的分数虽低于测验的临界点, 但高于效标的临界点。根据这些数据, 我们可以估计, 如果测验用作甄别工具, 并且采用现在的临界点, 那么全体应聘者样本中, 22% 潜在的工作成功者将被淘汰。确定测验的临界点时, 应该注意错误拒绝的百分率, 以及经过选拔的团体之内成功者和失败者的百分率。在某种情况下, 临界点应该定得很高, 以尽量排除一些可能的失败者。当不合格的员工会给工作带来重大损失时, 情况就是如此, 例如飞行员。在另一种情况下, 冒招进较多失败者的危险而尽可能多地招收合格的个体, 这才是重要的。这时, 选择较低的临界分数, 就能减少错误拒绝的人数。正常情况下, 确定临界分数位置的其他因素包括应聘人员供应情况、招聘岗位的数目, 以及招聘工作的紧急性或时间性等。<sup>[1]</sup>

在许多人事决策中, 录取率取决于情境的实际要求。例如, 由于招聘岗位与应聘人员的供求关系, 在一种情况下, 可能录取应聘者中测验分数最高的 40%, 而在另一种情况下, 则可能录取 75%。当外部没有规定录取率时, 测验上的临界分数可以定在对效标团体之间作出最大区别的那一点上。比较两个效标团

体的测验分数的分布,就可以大致定出临界点。现在也提出了确定最佳临界分数的更为精确的方法(Darlington & Stauffer, 1966; I. Guttman & Raju 1965; Jaeger, 1989; Livingston & Zieky, 1982; Martin & Raju, 1992; Rorer, Hoffman & Hsieh, 1966)。这些方法有可能考虑有关的其他参数,例如错误拒绝对错误接受的相对严重性。但是,至于是否把这类评价引入这个过程,还需要山人来判断。

使用决策理论的术语,图 6-1 中给出的例子说明一种简单的策略(simple strategy),即决定接受哪些申请者、拒绝哪些申请者的计划。用更一般的说法,策略就是利用信息而对各个个体作出一种决策的技术。在本例中,策略便是接受获得最高测验分数的 45 名个体。成功者的百分率从 60 提高到 84,由此可以估计使用测验而得到的净收益。

沃尔德(Wald, 1950)提出的统计决策理论,主要是关于工业产品的检验和质量控制中所需要的各种决策。克龙巴赫和格莱塞(Gleser, 1965)系统地阐述统计决策理论对于编制和解释心理测验的含义。决策理论本质上试图把作出决策的过程纳入数学模型,所以我们可以使用可得到的信息来作出指定条件下的最有效的决策。决策理论的一些基本概念有助于重新阐述和澄清某些测验问题。在统计决策理论正式提出之前,其中一些观点就被引入测验,后来又被认为符合测验框架。

**结果的预测** 心理测验决策理论的先驱泰勒和拉塞尔(Taylor & Russell)提出了泰勒—拉塞尔预期表(1939),可以确定录取准确度中可归因于使用测验的净收益。该表所需要的资料包括测验的效度系数、必须接受的申请者比率(录取率 selection ratio),以及不使用测验而录取的申请者的成功率(基础率, base rate)。这三个因素的任何一个人的变化,都会影响测验的预测效率。

表 6-1 基础率为 0.60 时, 通过采用某种效度的测验和某种  
录取率而预期的‘成功’率

| 录取率  |      |      |      |      |      |      |      |     |     |     |     |
|------|------|------|------|------|------|------|------|-----|-----|-----|-----|
| 效度   | .05  | .10  | .20  | .30  | .40  | .50  | .60  | .70 | .80 | .90 | .95 |
| .00  | .60  | .60  | .60  | .60  | .60  | .60  | .60  | .60 | .60 | .60 | .60 |
| .05  | .64  | .63  | .63  | .62  | .62  | .62  | .61  | .61 | .61 | .60 | .60 |
| .10  | .68  | .67  | .65  | .64  | .64  | .63  | .63  | .62 | .61 | .61 | .60 |
| .15  | .71  | .70  | .68  | .67  | .66  | .65  | .64  | .63 | .62 | .61 | .61 |
| .20  | .75  | .73  | .71  | .69  | .67  | .66  | .65  | .64 | .63 | .62 | .61 |
| .25  | .78  | .76  | .73  | .71  | .69  | .68  | .66  | .65 | .63 | .62 | .61 |
| .30  | .82  | .79  | .76  | .73  | .71  | .69  | .68  | .66 | .64 | .62 | .61 |
| .35  | .85  | .82  | .78  | .75  | .73  | .71  | .69  | .67 | .65 | .63 | .62 |
| .40  | .88  | .85  | .81  | .78  | .75  | .73  | .70  | .68 | .66 | .63 | .62 |
| .45  | .90  | .87  | .83  | .80  | .77  | .74  | .72  | .69 | .66 | .64 | .62 |
| .50  | .93  | .90  | .86  | .82  | .79  | .76  | .73  | .70 | .67 | .64 | .62 |
| .55  | .95  | .92  | .88  | .84  | .81  | .78  | .75  | .71 | .68 | .64 | .62 |
| .60  | .96  | .94  | .90  | .87  | .83  | .80  | .76  | .73 | .69 | .65 | .63 |
| .65  | .98  | .96  | .92  | .89  | .85  | .82  | .78  | .74 | .70 | .65 | .63 |
| .70  | .99  | .97  | .94  | .91  | .87  | .84  | .80  | .75 | .71 | .66 | .63 |
| .75  | .99  | .99  | .96  | .93  | .90  | .86  | .81  | .77 | .71 | .66 | .63 |
| .80  | 1.00 | .99  | .98  | .95  | .92  | .88  | .83  | .78 | .72 | .66 | .63 |
| .85  | 1.00 | 1.00 | .99  | .97  | .95  | .91  | .86  | .80 | .73 | .66 | .63 |
| .90  | 1.00 | 1.00 | 1.00 | .99  | .97  | .94  | .88  | .82 | .74 | .67 | .63 |
| .95  | 1.00 | 1.00 | 1.00 | 1.00 | .99  | .97  | .92  | .84 | .75 | .67 | .63 |
| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | .86 | .75 | .67 | .63 |

注：完整的预期表可参见 H.C. 泰勒和拉塞尔(1939) 麦考密克和伊尔根(1980)(摘自 H.C. 泰勒和拉塞尔(1939, p.576)。

为了说明起见,我们复制了泰勒—拉塞尔预期表之一,即表 6-1。当基础率即在使用测验之前所录取的申请者的成功百分率为 60 时,可以使用该表。泰勒和拉塞尔编制的其他预期表适用于其他基础率。表中最上端一行列出不同数值的录取率,左边第一列为测验效度。表中各个数字表示使用测验之后所录取的个体的成功率。因此,表中任何一个数字同 0.60 的差数,就表明归因于测验的成功录取率的增加。

显而易见,如果录取率为 100%,即如果所有的申请者都必须被录取时,任何一个测验都不能改进录取过程,而不论其效度如何。查阅表 6-1 表明,当必须录取 95% 的申请者时,即使测验具有完美效度( $r = 1.00$ ),成功者比率的增加也仅为 3%(从 0.60 到 0.63)。另一方面,当只需要录取 5% 的申请者时,效度系数仅为 0.30 的测验,也能把所录取的申请者的成功百分率从 60 提高到 82。从 60 到 82 的增加,表示测验的增值效度(incremental validity, Sechrest, 1963),即归因于测验的预测效度的提高。它表示测验对于录取在效标成绩上符合最低标准的申请者时所起的作用。当然,应用泰勒—拉塞尔预期表时,应该根据同估计先前成功者百分率所使用的相同类型的团体而计算测验效度。换句话说,对测验作用的评价不是根据偶然成功,除非先前是随机录取申请者的——而这是一种最不可能的情况。如果申请者的录取是根据先前工作史、推荐信、面谈等,对测验作用的评价则应该根据测验使先前的这些录取方法改进了多少。

使用测验得到的增值效度不仅取决于录取率,而且也取决于基础率。在先前举出的工作录取的情况下,基础率表示使用测验之前而录取的成功者的比率。表 6-1 表明基础率为 0.60 时的预测效果。关于其他的基础率,我们需要查阅所引用的参考文献中(H. C. Taylor & Russell, 1939) 其他合适的预期表。让我

们举一个例子,测验效度为 0.60,录取率为 40%。在这种条件下,如果我们一开始的基础率为 50%,测验的增值效度的作用是多少?如果我们从较极端的基础率 10%和 90%开始,这种作用又是多少?查阅一下适用于这些基础率的泰勒—拉塞尔预期表,它表明在第一种情况下,成功者的百分率从 50 提高到 75,在第二种情况下从 10 提高到 21,在第三种情况下从 90 提高到 99。由此可见,当基础率为 50 时,归因于使用测验的成功者百分率的提高为 25 而当基础率处于近两端时,这种提高只有 11 和 9。

临床心理学特别关注极端基础率的含义,这时基础率表示在所测验的总体中诊断为病理学状态的频数 (Buchwald, 1965; Cureston, 1957a; Meehl & Rosen, 1955; J. S. Wiggins, 1973/1988)。例如,如果某诊所收治总体的 5% 为器质性脑损伤,那么 5% 就是该总体中脑损伤的基础率。虽然使用任何效度的测验都会提高预测或诊断准确度,而当基础率接近 50% 时,提高则为最大。但是,对于罕见的病理学状态所发现的极端的基础率,这种提高便是微不足道的。在这种情况下,考虑到测验的实施和评分的费用,使用测验就没有什么道理了。在临床情况下,这种费用包括专业人员的时间,要不然他们可以把这些时间用于治疗另外的病例 (Buchwald, 1965)。错误肯定的数目,或正常个体被错误地诊断为患者,这当然会增加临床情境中的这种总费用。

某种罕见的病理学状态十分严重而需要紧急诊断时,在连续决策的早期阶段可以使用中等效度的测验。例如,首先使用一个中等效度的、容易实施的测验来筛选所有个体。如果临界分数定得很高(高分),那就几乎没有错误否定但有许多错误肯定,即正常人被诊断为患者,但对测验作出肯定诊断的所有个体进行深入细致的个别检查,还是能够检测出后一种人。例如,使用已有的工具对所有个体进行深入细致的个别检查行不通时,

这种解决方法就是合适的。

**效度和生产率的关系** 在许多实际情境中,评估选拔测验的效果,不是根据超过最低成绩的被试的百分率,而是根据录用被试的总的生产率。如何比较根据测验而录用的员工与不用测验而录用的全体申请者二者之间的工作能力,即效标成就的实际水平呢?在泰勒和拉塞尔的研究之后,一些研究者也致力于这一问题的研究。布罗格登(Brogden, 1946b)第一个提出,预期的产量增长与测验的效度成正比。因此,使用效度为 0.50 的测验所产生的增长,就是一个效度完美的测验所预期增长的 50%。

在表 6-2 中,<sup>[2]</sup>我们可以容易地看出测验效度与预期的效标成就增长之间的关系。效标分数是平均数为零、标准差为 1.00 的标准分数,该表列出某种效度的测验和某种录取率下所录用员工的预期平均效标分数。其中,效度为零的这一列给出基础的产品平均数,它相当于不使用测验而录用的申请者的成绩。使用一个效度为零的测验,等于根本没有使用测验。为了说明如何使用该表,让我们假设使用一个效度为 0.50 的测验,录用分数最高的 20% 的申请者(录取率 = 0.20)。参照表 6-2 可以看出,该样本的平均效标成绩高于预期的没有经过测验的样本的基础平均数 0.70 个标准差。使用一个完美的测验(效度系数 = 1.00),录取率仍为 20%,则录用的申请者的平均效标成绩将是 1.40,恰好是效度为 0.50 的测验的 2 倍。如果把表 6-2 之内任何一行中其他的平均效标成绩相比较,可以看到同样的正向线性关系。例如,采用 60% 的录取率,效度为 0.25 测验产生的平均效标成绩为 0.16,而效度为 0.50 测验产生的平均效标成绩则为 0.32。同样,效度增加一倍,产量也增加一倍。

施米特(Schmidt)及其同事进一步分析了生产率与测验效度

表 6-2 录取被试的平均标准效标成绩与测验效度和录取率的关系

| 录取率 | 效度系数 |     |     |     |     |     |     |     |     |     |      |      |      |      |      |      |      |      |      |      |      |
|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|------|------|------|------|------|------|------|
|     | .00  | .05 | .10 | .15 | .20 | .25 | .30 | .35 | .40 | .45 | .50  | .55  | .60  | .65  | .70  | .75  | .80  | .85  | .90  | .95  | 1.00 |
| .05 | .00  | .10 | .21 | .31 | .42 | .52 | .62 | .73 | .83 | .94 | 1.04 | 1.14 | 1.25 | 1.35 | 1.46 | 1.56 | 1.66 | 1.77 | 1.87 | 1.98 | 2.08 |
| .10 | .00  | .09 | .18 | .26 | .35 | .44 | .53 | .62 | .70 | .79 | .88  | .97  | 1.05 | 1.14 | 1.23 | 1.32 | 1.41 | 1.49 | 1.58 | 1.67 | 1.76 |
| .15 | .00  | .08 | .15 | .23 | .31 | .39 | .46 | .54 | .62 | .70 | .77  | .85  | .93  | 1.01 | 1.08 | 1.16 | 1.24 | 1.32 | 1.39 | 1.47 | 1.55 |
| .20 | .00  | .07 | .14 | .21 | .28 | .35 | .42 | .49 | .56 | .63 | .70  | .77  | .84  | .91  | .98  | 1.05 | 1.12 | 1.19 | 1.26 | 1.33 | 1.40 |
| .25 | .00  | .06 | .13 | .19 | .25 | .32 | .38 | .44 | .51 | .57 | .63  | .70  | .76  | .82  | .89  | .95  | 1.01 | 1.08 | 1.14 | 1.20 | 1.27 |
| .30 | .00  | .06 | .12 | .17 | .23 | .29 | .35 | .40 | .46 | .52 | .58  | .64  | .69  | .75  | .81  | .87  | .92  | .98  | 1.04 | 1.10 | 1.16 |
| .35 | .00  | .05 | .11 | .16 | .21 | .26 | .32 | .37 | .42 | .48 | .53  | .58  | .63  | .69  | .74  | .79  | .84  | .90  | .95  | 1.00 | 1.06 |
| .40 | .00  | .05 | .10 | .15 | .19 | .24 | .29 | .34 | .39 | .44 | .48  | .53  | .58  | .63  | .68  | .73  | .77  | .82  | .87  | .92  | .97  |
| .45 | .00  | .04 | .09 | .13 | .18 | .22 | .26 | .31 | .35 | .40 | .44  | .48  | .53  | .57  | .62  | .66  | .70  | .75  | .79  | .84  | .88  |
| .50 | .00  | .04 | .08 | .12 | .16 | .20 | .24 | .28 | .32 | .36 | .40  | .44  | .48  | .52  | .56  | .60  | .64  | .68  | .72  | .76  | .80  |
| .55 | .00  | .04 | .07 | .11 | .14 | .18 | .22 | .25 | .29 | .32 | .36  | .40  | .43  | .47  | .50  | .54  | .58  | .61  | .65  | .68  | .72  |
| .60 | .00  | .03 | .06 | .10 | .13 | .16 | .19 | .23 | .26 | .29 | .32  | .35  | .39  | .42  | .45  | .48  | .52  | .55  | .58  | .61  | .64  |
| .65 | .00  | .03 | .06 | .09 | .11 | .14 | .17 | .20 | .23 | .26 | .28  | .31  | .34  | .37  | .40  | .43  | .46  | .48  | .51  | .54  | .57  |
| .70 | .00  | .02 | .05 | .07 | .10 | .12 | .15 | .17 | .20 | .22 | .25  | .27  | .30  | .32  | .35  | .37  | .40  | .42  | .45  | .47  | .50  |
| .75 | .00  | .02 | .04 | .06 | .08 | .11 | .13 | .15 | .17 | .19 | .21  | .23  | .25  | .27  | .30  | .32  | .33  | .36  | .38  | .40  | .42  |
| .80 | .00  | .02 | .04 | .05 | .07 | .09 | .11 | .12 | .14 | .16 | .18  | .19  | .21  | .22  | .25  | .26  | .28  | .30  | .32  | .33  | .35  |
| .85 | .00  | .01 | .03 | .04 | .05 | .07 | .08 | .10 | .11 | .12 | .14  | .15  | .16  | .18  | .19  | .20  | .22  | .23  | .25  | .26  | .27  |
| .90 | .00  | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 | .10  | .11  | .12  | .13  | .14  | .15  | .16  | .17  | .18  | .19  | .20  |
| .95 | .00  | .01 | .01 | .02 | .02 | .03 | .03 | .04 | .04 | .05 | .05  | .06  | .07  | .07  | .08  | .08  | .09  | .09  | .10  | .10  | .11  |

(摘自布朗和吉塞利, 1953, p. 342)



的关系(Schmidt, Hunter, McKenzie & Muldrow, 1979)。以联邦政府中计算机程序员的工作为例, 这些研究者估计出一年来录取新员工中使用计算机能力倾向测验(效度 = 0.76)所产生的生产率增长的美元值。他们把决策理论技术应用于从美国人事管理处得到的数据, 由此作出估计。对于从 0.05 到 0.80 的九种录取率, 以及从零(随机录取)到 0.50 的先前录取方法的五种效度系数, 他们计算出预期的增长。

研究结果表明, 在所有这些条件下, 使用测验均产生可观的生产率的增长。把使用测验同随机录取相比较, 美元增长的范围从录取率为 0.05 时的 9720 万, 到录取率为 0.80 时的 1650 万。与先前 0.50 的录取效度相比较, 相应的美元增长的范围从 3330 万到 560 万。这些增长贯穿新近录用员工的预期使用期, 他们在联邦政府中从事计算机程序员平均将近 10 年。也应当指出, 这种估计所根据的假设是, 录取从高分申请者往下进行, 直到达到规定的录取率。换句话说, 这种程序采用最合适的录取过程。

使用人口普查数据来评定美国总人口中被聘为计算机程序员的个体数目, 研究者还作出在全国水平上使用某一测验的相应的效果估计值。在后来范围更大的研究中, 亨特(Hunter)和施米特(1981)探索把同样的统计技术应用于跨所有行业的全国劳动力的可能性。这些初步估计值被公认为是粗糙的、试验性的, 关于这类分析的其他方法都产生较低的估计值(Burke & Frederick, 1984; 美国劳工部 1983b; Weekley, Frank, O'Connor & Peters, 1985) 然而, 得出的研究结果有力地表明, 把人们分派到各种工作的有效方法, 对全国生产率起着重大促进作用。劳动生产率的性质以及影响生产率的个体因素和组织因素, 正受到越来越多的关注。这是测验效度中效标研究的一个发展中的领域,

它正取得显著的理论进展和方法论进展 (J. P. Campbell, Campbell & Associates, 1988 ; Hunter, Schmidt & Judiesch, 1990; Raju, Burke & Normand, 1990)。

**决策理论中的效用概念** 按照测验在特定情境下的有效性来评价测验,这是决策理论的特征。这种评价不仅考虑预测特定效标中测验的效度,而且也考虑一些其他参数,包括基础率和录取率。另外一个重要参数是预期结果的相对效用 (utility),即判断每一种结果的利弊。现在尚不能按照统一的效用量表来确定各种结果的价值,这是应用决策理论的一大障碍。在工业决策中,经常以美元为单位来确定不同的结果。但即使在这种情况下,关于良好意愿、公共关系、员工的精神等的某些结果,也很难用金钱来评定。教育决策必须考虑教学目标、社会价值和其他相对无形的因素。在个体决策咨询中,则必须考虑个体的偏爱和价值系统。但是,有人反复指出,决策理论并没有把价值问题引入决策过程,而只是提出这一问题而已。各种决策中都会碰到价值系统,但是迄今为止我们尚不能清楚地认识并系统地处理价值系统。

上面所引用的施米特、亨特及其同事的生产率研究,说明了决策模型中价值确定方法的进展。虽然那个研究关注货物和员工工作的美元价值,但是其中所发展的技术,也适用于测量其他价值。按照人类判断的量化,同样的方法也可以用于任意的数字量表,假如清晰地定义并且前后一贯地应用量表的话。应当指出,决策模型所需要的估计值仅仅是关于不同结果的相对价值,而不是绝对价值。对人员决策中效用的一种综合技术处理,参见布德罗(Boudreau, 1991)的有关论述。<sup>[3]</sup>

在选择决策理论时,目的是要使预期的所有结果的总效用最大化。参照图 6-2 所示的简单决策策略,它有助于阐明这种

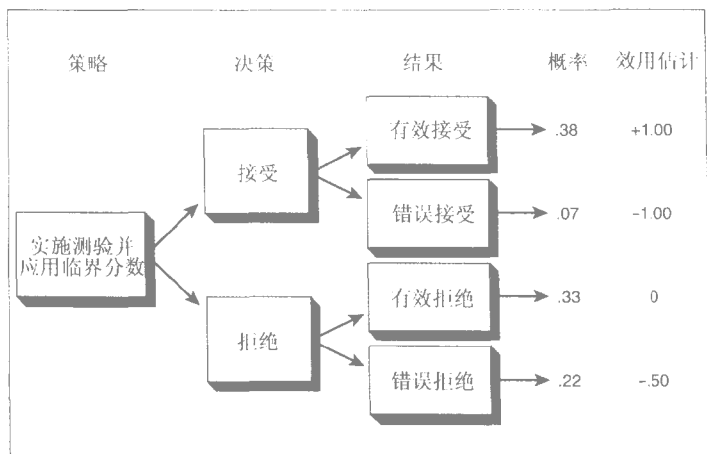


图 6-2 一种简单的决策策略

方法。该图说明图 6-1 中的决策策略，对一组申请者实施一次测验，然后根据测验上的临界分数而作出接受或拒绝某个申请者的决策。共有四种可能的结果，即有效接受、错误接受、有效拒绝、错误拒绝。从图 6-1 的四个分区中的人数，我们可以得出每一种结果的概率，现将它们列于图 6-2 中。

我们还需要另一类数据，即表示在一个共同量表上的不同结果的效用。图 6-2 的最后一列给出这些假设的数值，无论用什么样的判断方法来得到。每种结果的概率乘以这种结果的效用，然后把 4 个乘积相加，再减去相当于测验费用的数值，就得出预期的策略的总效用。这最后一项强调，如果某个测验项目不多、费用不高，稍经训练的人员就能实施，而且适用于团体实施，虽然效度较低但也可能被采用。某个测验需要训练有素的

主试或昂贵的设备,那就必须要有较高的效度,才有被采用的价值。在图 6-2 所示的假设例子中,测验费用在效用量表上被估计为 0.10,这个决策策略的总的预期效用(EU)为:

$$EU = 0.38 \times 1.00 + 0.07 \times (-1.00) + 0.33 \times 0 + 0.22 \times (-0.50) - 0.10 = 0.10$$

然后可以把这个 EU 与不同临界点、不同的测验(效度不同、费用不同)或成套测验,以及不同决策策略<sup>[4]</sup>所得出的其他各种 EU 相比较。

**序列决策和适应性处理** 在一些情况下,如果采用较为复杂的决策策略,考虑较多的参数,则可以增加测验的有效性。试举两例,用以说明这种可能性。第一种情况,使用测验来作出序列决策而不是最终决策。图 6-1 和图 6-2 中所示的简单决策策略,不论接受还是拒绝,所有决策都是最终决策。另一方面,图 6-3 说明两个阶段的序列决策。测验 A 是一个简短而且容易实施的甄别测验。根据该测验上的成绩,将被试分为三类,即明确接受者、明确拒绝者,以及介于这二者之间的‘不确定’者,需要用深入细致的技术例如测验 B 对这类被试作进一步的检查。根据第二阶段的测验,再把这类被试分为接受者和拒绝者。

另一种策略适合于心理障碍的诊断,虽然只采用两种类别,但要经过初步甄别测验归入阳性者(即可能的病理患者)的所有个案,作进一步的测验。这就是本节中前面所述的策略,使用测验来诊断基础率很低的病理状态。

也应当指出,许多人事决策实际上也是序列决策,虽然看起来并非如此。由于预测错误而被录用的不够格的员工,通常在一段试用期之后被解聘;不及格的学生在若干阶段被大学淘汰。在这种情况下,不录取的决策才是最终决策。确实,错误的录取决策在后来再作纠正,按一些价值系统来说代价昂贵,但是它们

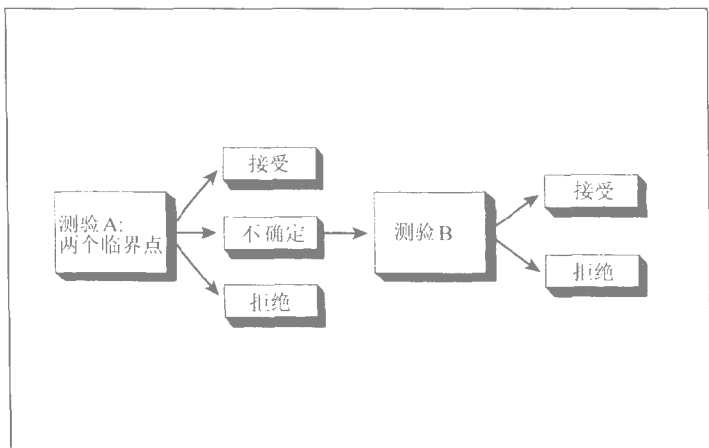


图 6-3 一种序列的决策策略

的代价通常毕竟小于最终的错误决策。

改变心理测量有效性的第二种情况是，针对个体特性，采用不同的处理和适应性处理(adapting treatments)。例如，对不同能力倾向水平的员工，采用不同的训练方法，或者为具有某种教育障碍的学生，开设补偿教学课程。在这种条件下，随后对各个被试作出决策时，应该考虑最初的测验分数和不同处理的相互作用的数据。采用适应性处理，可以大幅度提高成功率。因为把个体分配到不同的处理组，本质上是分类问题而不是录取问题，后面分类决策一节中会详细讨论所需要的方法。

决策理论的概念和基本原理有助于以具体的测验目的来评价心理测验，上述例子说明其中的一些方法。决策理论本质上集中注意于因素的复杂性，这些因素决定某个测验在某种特定

情境下能够起的作用。效度系数本身尚不能说明是否应该使用某个测验，因为在评价测验对总决策过程的功效的影响时，效度系数只是要考虑的各种因素之一。<sup>[5]</sup>

**调节变量** 对于某种效标而言，测验效度在被试特性不尽相同的亚团体之间有所不同。经典的心理计量模型假设，预测误差是测验的特性而不是被试的特性，并且假设预测误差在被试之间是随机分布的。决策理论所引入的灵活方法，促进人们去探索包括被试和测验相互作用的预测模型。这种相互作用意味着，同一个测验预测某些类型或某些子集的被试的效度，比另一些类型或子集的被试好。例如，某个测验预测男子效标成绩的效度比女子好，或者预测社会经济水平较低申请者的效度，比社会经济水平较高的申请者好。在这些例子中，性别和社会经济水平被称为调节变量(moderator variables)，因为它们调节测验的效度(Saunders, 1956)。

兴趣和动机也起着调节变量的功用。因此，如果申请者对工作没有什么兴趣，他们也许会做得较差，而不管他们在有关的能力倾向测验上的分数如何。在这类被试中，能力倾向测验分数和工作表现之间的相关就会很低。另一方面，对于有兴趣和动机的被试，能力倾向测验分数和工作成功之间的相关就会很高。本世纪 50 年代和 60 年代广泛研究各种可能的调节变量，吉塞利(Ghiselli, 1956, 1960, 1963, 1968) 的系列研究着重于预测工作表现；另一些研究者检验有关人格变量作用的假设，尤其是对大学成绩的预测(N. Frederiksen & Gilbert, 1960; N. Frederiksen & Melville, 1954; Grooms & Endler, 1960; L. J. Stricker, 1966)。

一个相对一致的研究结果是预测学业成绩时的性别差异。对于多种资料的数百个相关系数的研究都报告，能力倾向测验分数和学业成绩之间的相关，女性高于男性(Gross, Faggen & Mc-

Carthy, 1974; Schmidt, Mellon & Bylenga, 1978; Seashore, 1962)。在中学和大学都出现这种倾向,但是在大学水平这种倾向更为明显。虽然这种数据没有说明预测学业成就中性别差异的原因,但是根据其他已知的性别差异对它进行推测,这是令人感兴趣的。如果一般来说女生往往遵守和接受学校情境的价值观和标准,则她们的学业成就也许主要取决于能力。另一方面,如果男生往往致力于感兴趣的活动(校内或校外),则兴趣差异会成为学业成就的另一种变量,这就难以从能力倾向测验分数预测学业成就。不过,应当指出,这些效度系数中的性别差异虽然十分一致,但一般来说都很小。此外,在后来的研究中都表明性别差异往往较小,这个研究结果反映 60 年代后期和 70 年代女性态度的变化。

一般来说,早期关于调节变量作用的预期并没有实现 (Abrahams & Alf, 1972; Pinder, 1973; Zedeck, 1971)。方法分析揭示出许多不足之处。新样本的交叉效度分析通常没有证实最初的研究结果。企图使用调节变量显著改进通过其他方法能够获得的预测,这看来不大可能。根据现在的知识,在缺乏明确证据的情况下,没有一种变量能够被假设为有调节效度。然而,调节变量的概念仍有启发价值,它有助于理解个体行为,例如在临床个案研究中,以及提出应该使用合适方法进行研究的新假设。事实上,80 年代和 90 年代调节变量再次引起人们的兴趣。虽然使用早期方法不能检测某些调节变量,但是,现在通过复杂的统计分析已能够作出鉴定 (Morris, Sherman & Mansfield, 1986; E. F. Stone & Hollenbeck, 1989)。

## 不同测验的信息组合

预测实际效标，往往需要几个而不是一个测验。大多数效标都是比较复杂的，效标度量取决于若干不同的特质。用来测量这类效标的测验因此必须是高度异质性的。但是我们已经指出，主要测量单一特质的相对同质性的测验是比较令人满意的，因为它得出较为明确的分数（第五章）。因此，通常最好使用由若干相对同质性的测验组成的成套测验，其中每个测验包括效标的一个不同的方面，而不是使用由许多不同类型的项目凑合成的单一测验。

使用若干经过专门选择的测验来预测单一的效标，它们被称为成套测验（test battery）。使用成套测验的主要问题是，如何组合不同测验的分数而对每一个体作出决策。主要有两种方法，即多元回归方程和剖面分析。对每一个案实施测验进行深入研究时，例如临床诊断、咨询或评价高水平的行政干部，主试的通常做法是利用测验分数而不作进一步的统计分析。准备个案报告和提出建议时，主试依赖于判断、以往经验和理论基础来解释分数类型和组合不同测验的分数。

**多元回归方程** 多元回归方程根据个体在成套测验上的各个分数来预测他或她的效标分数。以下列回归方程为例，根据学生在言语测验(V)、数字测验(N)和推理测验(R)的分数,应用回归技术来预测他在中学数学课的成绩：

$$\text{数学成绩} = 0.21V + 0.21N + 0.32R + 1.35$$

在这个例子中，测验分数和效标分数都表示为标准九分。在回归方程中也可以使用其他分数量表。在上述方程中，学生



在三个测验上的标准九分乘上方程式中给出的权重。这些乘积的加和，再加上一个常数，就可以预测学生在数学课的标准九分位置。

假设贝蒂·琼斯得到如下分数： $V = 6, N = 4, R = 8$ ，那么，其预测的数学成绩如下：

$$\text{数学成绩} = 0.21 \times 6 + 0.21 \times 4 + 0.32 \times 8 + 1.35 = 6.01$$

贝蒂的预测标准九分近似为 6。回忆一下(第三章)，标准九分以 5 为平均数，因此可以预期贝蒂数学课的成绩稍高于平均成绩。她在推理测验上得分很高 ( $R = 8$ )，在言语测验上也高于平均数 ( $V = 6$ )，这就弥补了她在计算速度和准确性上的不足 ( $N = 4$ )。

计算回归方程的具体技术，请参阅心理统计学教材（例如，D. C. Howell, 1997; Runyon & Haber, 1991）。回归方程实质上是根据每个测验与效标的相关，以及各测验之间的相关。显然，那些与效标有较高相关的测验的权重较大。但是，考虑每个测验与成套测验中其他测验的相关同样重要。彼此高度相关的测验是不必要的重复，因为它们在很大程度上包括效标的相同方面。即使这样两个测验都与效标高度相关，使用这两个测验也不会明显增加整套测验的效度。在这种情况下，其中一个测验的作用和两个测验大体相同，因此在成套测验中只需要保留一个测验。

不过，即使严重重复的测验被删除之后，成套测验中所保留的测验彼此相关也会有所不同。为了达到最大的预测值，比起同其他功用部分重复的测验来，那些对整套测验起着独一无二作用的测验应该获得较大的权重。在计算多元回归方程时，每个测验的权重与它和效标的相关成正比，与它和其他测验的相关成反比。因此，最大的权重应该分配给具有最高效度，同时又

与其他测验最少重复的测验。

要得出整套测验的效度，可以计算效标与成套测验之间的多重相关( $R$ )。对于预测所讨论的效标而言，每个测验都获得最佳权重时，该相关就表示从某成套测验能够获得的最大预测值。通过回归方程所确定的权重最为合适。

应当指出，最佳权重只是对于得出权重的特定样本而言。由于得出权重时所使用的相关系数的偶然误差，回归权重随样本而有所不同。因此，应该计算所预测的效标分数与新样本中实际的效标分数的相关，从而得出成套测验的交叉效度。当回归方程应用于第二个样本时，可以使用公式来估计所预期的多重相关的减少量，但是，在可能的情况下，最好使用经验证实。得出回归权重的样本越大，这种减少就越小。<sup>[6]</sup>

在某些情况下，在回归方程中包括一个与效标零相关但与成套测验中另一个测验高度相关的测验，这可以改进成套测验的预测效度。当与效标不相关的测验作为抑制变量 (suppressor variable) 即消除或抑制其他测验的无关方差时，就出现这种奇妙的情况 (Conger & Jackson, 1972)。例如，阅读理解也许与数学或机械能力倾向测验分数高度相关，因为做题需要理解复杂的书面指导语。如果阅读理解与要预测的职业行为无关，那么测验所需要的阅读理解就会带来误差方差而降低测验的预测效度。实施阅读理解测验，并且将其分数包括在回归方差之中，那就会删除这种误差方差而提高成套测验的效度，抑制变量在回归方程中为负的权重。因此，个体的阅读理解分数越高，从他的数学或机械能力测验分数中就减去越多。然而，在任何情况下，修改测验从而消除无关方差的直接方法，都要优于通过抑制变量而消除无关变量的间接统计方法。在修改测验行不通的情况下，才应该考虑去研究抑制变量。在这种情况下，总是应该以新

样本来检查抑制变量的效果。

**剖面分析和临界分数** 除了在临床评定中应用个别化剖面分析之外,我们能够按照多重临界分数来评定使用人员选拔成套测验所获得的测验分数型式。简单地说,这种方法就是对每一个测验都建立一个最低的临界分数。严格运用这种方法时,在任何一个合适的测验上处于最低分数之下的每一位被试,都将被拒绝。为某种职业选择合适的测验和建立临界分数时,单单考虑测验效度通常是不够的。假如仅仅考虑产生显著效度系数的测验,就会忽视所有在职员工都擅长的一种或几种基本能力。因此也需要考虑总体来说员工都擅长的那些能力倾向,即使超过某种最低水平的个体差异与职业成功的程度无关。而且,在一些职业中,员工们也许在关键特质上高度同质,其个体差异的范围甚小,不可能得出测验分数和效标之间的显著相关。

美国职业服务中心(United States Employment Service)编制了“一般能力倾向成套测验”(General Aptitude Test Battery, GATB),<sup>[7]</sup>各州职业服务处将它用于职业咨询和分配计划(美国劳工部,1970),此例说明多重决断方法的这种用途。该成套测验有9个能力倾向分数,根据效标相关,也根据该职业员上的平均数和标准差以及定性的工作分析,从中选择每种职业所要考虑的那些分数。

赞成使用多重决断方法而不是回归方程的最主要理由是分数补偿的可能性。换句话说,被试成套测验的总分中,由于某一个测验上的高分,会掩盖另一种技能的严重不足。如果被录取的申请者 in 特定职业的某种关键技能上存在缺陷,那么他将面对工作失败的结果。但是,鉴定出某特定职业所要求的一种或几种关键技能,并且在相应的测验上应用临界分数,那就能够避免这种情况。对于大多数测验来说,通常最好保留实际分数,因

为被试的测验分数越高，他的工作表现也越好。在大多数职业中，预测分数和效标成绩之间存在着线性关系。还应当指出，使用 GATB 的广泛研究有力证实了这种线性关系（Coward & Sackett, 1990; Hartigan & Wigdor, 1989; Hawk, 1970）。在这种条件下，比起接受所有超过最低临界分数的被试来，根据实际的测验分数的大小去选择被试，倒会获得较好的工作成绩。

## 分类决策中测验的应用

**分类的性质** 心理测验可以用于选拔、安置或分类等多种目的。在选拔时，每一位个体或被接受或被拒绝。例如，大学是否录取某学生、单位是否录用某职业应聘者，或军队是否接受某新兵参加军官培训等，都属于选拔决策。采用序列决策进行选拔时，前面阶段通常称之为“筛选”（screening），而深入细致的最后阶段才称之为“选拔”（selection）。“筛选”也可以用来表示任何快速的、粗糙的选拔过程，即使后面没有进一步的选拔程序。

安置和分类不同于选拔，因为没有一位被试被拒绝或被淘汰。所有个体都被分派到合适的岗位，以便产生最大的效益。在安置时，根据单一的分数进行分配。这种单一的分数也许来自单一的测验，例如数学成就测验。如果实施成套测验，则应用单一回归方程计算出一个合成分数。安置决策的例子包括：根据成就测验分数将大学新生分配到不同的数学班级；将应聘者分派到各种办公室工作，其要求的技能和责任性水平各不相同；出于治疗目的，将精神病患者鉴定为“重度障碍”或“轻度障碍”，等等。显然，所有这些决策都只使用一种效标，而且这种安置取决于个体在单一预测量表上的分数。

另一方面,分类总是包括两种或多种效标。例如,在军事部门,分类是一个重大问题,因为必须把人员库的每一个体分派到他或她最能发挥作用的军事专业。工业部门同样需要分类决策,把新员工分派去参加不同的职业培训。其他例子还有:关于大学生选修课程(理科、文科等等)和专业的咨询。咨询基本上是根据分类,因为被试被告知他在不同学业计划或职业中成功的可能性。临床诊断同样也是一个分类问题,每种诊断的主要目的便是作出最适于治疗的决策。

安置可以使用一个或多个预测测验,但是分类则需要多重预测测验,其效度分别由每一效标所确定。分类成套测验需要为每一效标建立不同的回归方程,其中有些测验在各个回归方程中都有权重,虽然其数值有所不同;有些测验也许只被包括在一个或两个回归方程之中,而对于其他效标的权重则为零或很小。因此,从整套测验中使用一些测验,这些测验的组合,以及具体的权重,都随特定的效标而变化。一个早期例子是美国空军所编制的分类成套测验,用于分派人员参加不同的训练。该成套测验包括纸笔测验和仪器测验,提供飞行员、领航员、投弹手及其他空勤专业人员的标准九分。根据不同的回归方程估计出个体的各种效标分数,这就能够预测他或她是否更可能成为合格的飞行员而不是领航员,等等。一个近期的、使用更为广泛的例子是计划 A,即美国空军选拔和分类计划(U. S. Army Selection and Classification Project, J. P. Campbell, 1990b)。

**区分效度** 在评价用于分类的成套测验时,主要考虑其对各个效标的区分效度。这种成套测验的目的是预测每一个被试在两种或多种职业、训练计划或其他效标上表现的差异。组成成套测验的各个分测验,对于各种效标应该得出大不相同的效度系数。例如,在两个效标的分类问题中,理想的测验应该与其

中一个效标高度相关而与另一个效标零相关（最好为负相关）。一般智力测验相对来说不适合于分类目的，因为它们同样准确地预测许多领域的成功。因此，智力测验与所要区分的各种效标的相关十分接近。在智力测验上得分高的被试被认为随便分配在哪个岗位上都会成功，这就不能预测他在哪个岗位上会做得更好一些。在用于分类的成套测验中，我们需要其中一些测验能较好地预测效标 A 而预测效标 B 则弱些，另一些测验能较好地预测效标 B 而预测效标 A 则弱些。

已经发展了若干统计方法来选取分测验，以便使得用于分类的成套测验具有最大的区分效度（Brogden, 1946a, 1951, 1954; Horst, 1954; Mollenkopf, 1950b; Zeidner & Johnson, 1991）。不过，当效标的数目多于两个时，这个问题就十分复杂了。实际上采用种种经验方法可以接近理想的目标。坎贝尔（J. P. Campbell, 1990a, pp. 715 ~ 721）详尽分析了分类问题的复杂性。

**多重鉴别功能** 处理分类决策的另一种方法是使用多重鉴别功能(French, 1966)。实质上,这是一种数学方法,用于确定某个体在整套测验上的分数在多大程度上近似于一般群体在某一职业、课程、精神病症状或其他类别上的分数。然后把被试分派到他或她最接近的特定团体中去。虽然回归方程可以预测各个被试在每一领域的成功程度,但是多重鉴别功能则把同一类别的所有被试都视为是相同状况的。团体成员是这种方法所使用的惟一效标资料。得不到效标分数而只能确定团体成员时,就可以使用鉴别功能。例如,对不同职业的被试实施某些测验,虽然没有测量各个被试在每一领域之内的工作成功的程度,但也可以建立这些测验的效度。

效标与一个或多个预测测验之间存在非线性关系时,同样可以使用鉴别功能。例如,某种职业的某些人格特质有其最佳

范围，在所讨论的特质上得分过高或过低的被试因而都不是理想的人选。有理由预期，售货员的社会支配性适度高分最可能成功。但是该分数太高或太低时，成功的可能性反而会下降。使用多重鉴别功能，我们就可以录取分数处于最佳范围之内的个体。另一方面，使用回归方程，支配性分数越高，所预测的结果也越好。当然，假如预测测验和效标之间为负相关，则回归方程会预测低分者做得较好。但是，中等的测验分数不可能直接获得最大预测分数。虽然在许多情况下，两种方法都会作出同样的选择，但在某些情况下，使用回归方程和多重鉴别功能，将对被试作出不同的分类对于大多数心理测验目的来说，回归方程是一种更为有效的方法。然而，在某些情况下，多重鉴别功能倒更适合于得出所需要的信息。

**人力资源的充分利用** 比起使用单一的一般测验或从单一的回归方程所得出的合成分数能够作出的预测来，使用成套测验对多个效标作出不同的预测则更能充分利用现有人才资源。在本章泰勒—拉塞尔预期表和其他地方，我们看到，一个测验在录取某职业申请者时的有效性取决于录取率。在分类决策时，我们采用较低的录取率，因而能够把更为合格的被试分派到每项工作中去。如果有 100 个申请者，两项不同的工作各需要 10 人，当每项工作各自使用预测测验时，则每项工作的录取率各为 10%。假如使用单一的预测测验（例如一般智力测验）来录取两项工作的申请者，录取率就是 20% 了，因为我们只能录取测验分数最高的 20 个申请者。

甚至当两项工作的预测测验高度相关时，使得某些申请者两项工作都合格，使用两个预测测验也是获益匪浅。以表 6-3 为例说明这种情况，表中所示为两项工作使用两种策略所录取者的效标标准分数的平均数。一种是选拔策略（单一预测测

验)，另一种是包括两个不同的预测测验的分类策略，而每个预测测验以各自的工作效标建立效度。假如不经选拔而随机分派被试，则该量表的标准分数平均数就为零。这就相当于每项工作的录取率都是 50%，结果申请者 100% 被录用。应当指出，即使在这种条件下，使用两个预测测验也会改进工作成绩，如表最后一行所示。使用两个互不相关的预测测验，工作成绩平均数为 0.31(近似高于随机值  $3/1$  个标准差)。随着两个预测测验之间相关的增高，所录用者的工作有效性逐渐降低；但是相关高达 0.80 时，也仍然比随机录取好。当然，使用较低的录取率，我们能够录取更为合格的人员。不过，在表 6-3 中能够看到，对于每一种录取率，由分类策略所录取的人员的工作成绩，都要优于由选拔策略所录取的人员。

试举一个实例说明分类策略的长处，即美国陆军使用“能力倾向领域成套测验”(Aptitude Areas) 分数把人员分配到各军事

表 6-3 使用选拔策略或分类策略所录取者在两项工作  
效标上标准分数的平均数

| 每项工作的录取率 | 选拔：单一预测测验 | 分类：两个预测测验的相关 |      |      |      |     |
|----------|-----------|--------------|------|------|------|-----|
|          |           | 0            | .20  | .40  | .60  | .80 |
| 5%       | .88       | 1.03         | 1.02 | 1.01 | 1.00 | .96 |
| 10       | .70       | .87          | .86  | .84  | .82  | .79 |
| 20       | .48       | .68          | .67  | .65  | .62  | .59 |
| 30       | .32       | .55          | .53  | .50  | .46  | .43 |
| 40       | .18       | .42          | .41  | .37  | .34  | .29 |
| 50       | .00       | .31          | .28  | .25  | .22  | .17 |

(摘自布罗格登, 1951, p.182)



专业(Maier & Fuchs, 1973)。在这项研究中,每个能力倾向领域代表一组需要相同的能力倾向、知识和兴趣的军事工作。该成套测验有 13 个分测验,每个能力倾向领域由其中 3~5 个分测验构成。图 6-4 表示使用能力倾向领域分数与使用综合的筛选测验即“陆军资格测验”(Armed Forces Qualification Test, AFQT),对 7500 名应征入伍者的比较研究结果。应当指出,只有 56% 的人员达到或超过 AFQT 的第 50 百分位数,而 80% 的人员在各自最佳能力倾向领域上达到或超过标准分数的平均数 100。因此,如果根据每项工作所需要的能力倾向,把个体分配到各专业工作领域,则绝大部分人员的工作表现能够达到或超过整个样本的平均数。几乎每个人都在平均数之上,这看来是

16

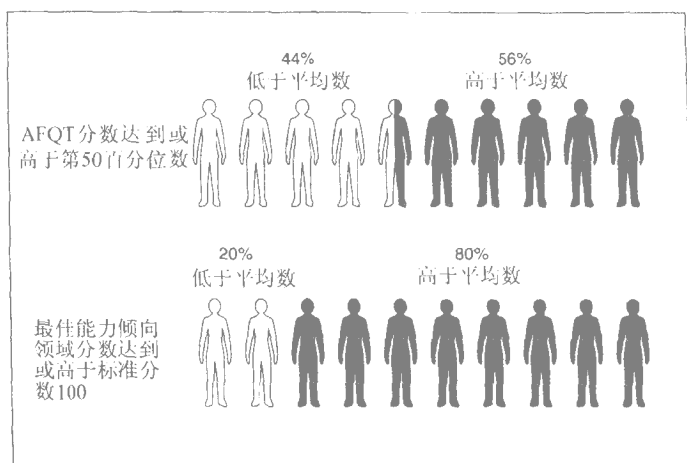


图 6-4 7500名应征入伍者样本 AFQT 分数及陆军分类成套测验最佳能力倾向领域分数高于平均数的百分比(资料摘自美国行为和社会科学研究院军队研究所)

不可能的事情，但是利用每个人在某种能力倾向上胜过别人这个事实却能够办到。

费尔德曼和布拉顿(Feldman & Bratton, 1972)研究超常儿童时使用截然不同的总体也得出相同的观点。出于研究目的，他们使用 19 种度量来评价小学五年级两个班级的 49 名学生。所有这些度量以往曾被用来选拔超常学生学习特殊课程，例如团体智力测验和教育成就成套测验的综合分数、各能力倾向测验与阅读和数学之类各学业领域测验、创造性思维测验、音乐和美术的成绩，以及教师们对每个班级中最‘聪明’和最‘有创造性’学生的提名。如果分别使用每个效标来选拔得分最高的前五名学生，那么团体中 92% 的学生都被选中。因此，如果我们使用多变量效标，那么几乎某团体的所有成员都有超过他人之处。

## 测验偏向的统计分析

**问题** 如果我们要使用测验去预测某种未来情境中的结果，例如某申请者在大学里或在某项工作上的表现，那么我们需要测验对特定的效标具有高度的预测效度。在编制所谓的文化公平测验(culture-fair test)时，通常忽略了这种要求(第九和十二章进一步讨论这个问题)。为了在这类测验中仅仅包括不同的文化或亚文化共同的机能，我们可以选取对于我们希望预测的效标没有什么关系的内容。一种较好的解决方法是，选取和效标有关的内容，然后研究对于测验目的而言，测验有效性中可能存在的总体差异。效度系数、回归权重、临界分数等都可以随测验参加者经验背景的差异而有所变化。因此，对于有理由预期会有这类影响的各亚团体，我们应该检查其中的这些数值。这

些可能的亚团体差异被认为是前一节所讨论的调节变量的一种特例。回忆一下，探索显著的、稳定的调节变量证明是令人失望的。本节我们将讨论把这类分析具体应用于美国少数民族总体。

应当指出，如果测验和效标成绩有内在关联性，那么测验分数的预测性在各文化团体之间不大可能有所变化。如果使用一个言语测验去预测非言语工作的成绩，由于某个文化团体具有以往经验的传统联系也许会得出意外的效度。然而，在另一个具有不同经验背景的团体中，该测验就没有效度可言了。另一方面，如果一个测验直接从效标行为中取样，或它测量本质的、必要的技能，那么它的效度就可能在不同团体中都保持不变。

自从 60 年代中期以来，关于测验分数的预测意义中可能存在的种族差异，进行了大量的研究。<sup>[8]</sup> 至今所进行的绝大部分研究涉及美国黑人，虽然一些研究也包括其他少数民族。所研究的各种问题一般归类于测验偏向 (test bias)。在这种情况下，我们是在严格的统计意义上使用“偏向”一词，表示恒定误差即系统误差是相对于偶然误差而言的。在同样的意义上，我们讲到有偏向的样本是相对于随机样本而言的。测验偏向的两个主要问题涉及效度系数(斜率偏向)以及测验和效标上团体平均数之间的关系(截距偏向)。下面两节将讨论这些问题。

**斜率偏向** 为了便于理解测验偏向的各个专门问题，让我们首先考虑第四章所讨论的相关散布图或双变量分布（尤其是图 4-3）。出于现在的目的，横轴(X)表示测验分数，而纵轴(Y)则表示效标分数，例如大学年级平均成绩或某项工作成绩的指标。回忆一下，计数符号表明每一个体在测验和效标上的位置，它们表示两个变量之间相关的方向和大小。通过这些计数符号的最优拟合线称之为回归线，其方程便是回归方程。在本例中，

回归方程只有一个预测测验。虽然本章前面所讨论的多元回归方程具有若干个预测测验,但是其原理是相同的。

当测验分数和效标分数都以标准分数 ( $SD = 1.00$ ) 来表示时,回归线的斜率 (slope) 就等于相关系数。因此,如果一个测验在两个团体中得出显著不同的效度系数,那么这种差异就称之为斜率偏向。这类团体差异通常称之为“区分效度”(differential validity)。一些研究人员也使用“单一团体效度”(single-group validity) 一词来表示某测验的效度系数在一个团体中达到统计显著性,而在另一个团体中则没有达到。

图 6-5 用图解说明几种双变量分布的回归线。<sup>[9]</sup>椭圆表示每个样本的计数符号落入的区域。例 1 所表示的两个团体的双变量分布,其预测测验的平均数不同,但是预测测验和效标之间的回归线相同。在这种情况下,不存在测验偏向,因为任何一个给定的测验分数 ( $X$ ) 在两个团体中都对应于同一个效标分数 ( $Y$ )。例 2 表明斜率偏向,少数民族团体的效度系数较低。

在区分效度的研究中,一个共同的困难在于,少数民族样本的个案数通常远远少于主体民族样本。在这些情况下,同一个效度系数在主体民族样本中具有统计显著性,而在少数民族样本中则没有显著性(所谓的单一团体效度)。例如,对于 100 个个案而言,相关系数为 0.27,就在 0.01 水平上达到显著性而对于 30 个个案而言,同一个相关系数,即使在 0.05 水平上,也远远没有达到显著性所要求的最小值。由于这种原因,在这类区分效度研究中使用的合适方法,是评价两个效度系数之间的差异,而不是分别检验每个效度系数的显著性 (Humphreys, 1973)。假如使用后一种方法的话,人们就很容易“证明”,某测验对于白人有效度而对于黑人则没有效度。出于这种目的,我们所需要的只不过是很大的白人样本和很小的黑人样本而已!

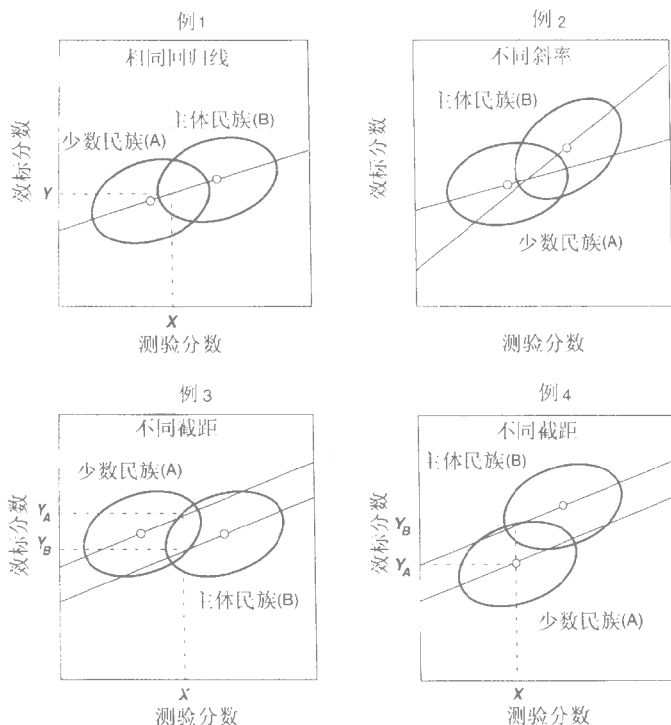


图 6-5 预测效标分数时的斜率偏向和截距偏向。椭圆表示根据效标成绩绘制测验分数点的位置时，每一团体的成员所落入的区域(例 1、例 2 和例 4 摘自戈登, 1953, P.3)。

19 个发表的研究报告了黑人职业样本和白人职业样本的效度系数，对这些研究的结果进行复杂的统计分析，使得人们十分怀疑早期研究所得出的结论 (Schmidt, Berner & Hunter, 1973)。考虑一下每个研究所得出的效度系数和样本的大小，研究人

员证明,黑人和白人之间所得出的效度系数的差异,与随机预期没有什么两样。后来,一项包括 39 个更为全面的研究分析证实了这个结论(Hunter, Schmidt & Hunter, 1979)。关于少数民族和非少数民族职业申请者的区分效度的课题,持续讨论了十多年。一些研究人员注意到,由于方法的局限性,研究结果就得出不出什么结论。然而,应当指出,对各个工业样本(J. T. Campbell, Crooke, Mahoney & Rock, 1973)和军队人员(Maier & Fuchs, 1973)进行设计良好的、大规模的研究,并没有得出区分效度的证据。一般来说,方法较为成熟的研究不大可能得出区分效度的证据。

许多关于黑人和白人大学生的研究也得出类似的结果(Breland, 1979)。大学入学考试委员会“学业能力倾向测验”(Scholastic Aptitude Test)和其他大学招生测验的效度系数,一般来说黑人学生和白人学生一样高,甚至更高一些。当黑人和白人样本进入相同的大学时,以及当他们分别进入不同的大学时,都得出这种关系。在另一个不同的水平上,米切尔(B. C. Mitchell, 1967)研究了两个教育准备测验对小学一年级儿童学期终成就测验分数的效度。在黑人和白人被试儿童的大样本中,两个种族团体的总分效度和分测验的效度都十分相似,虽然黑人儿童的稍高一些。总之,对可得到的各种研究所进行的全面调查和批评分析并不支持如下假设:能力测验预测黑人的职业或教育成绩的效度不如白人(Hunter, Schmidt & Rauschenberger, 1984; Linn, 1978)。

虽然关于其他少数民族的比较研究甚少,但对美籍西班牙人进行的教育测验和职业测验,都得出相似的结果(Breland, 1979; Duran, 1983, 1989; Pennock-Roman, 1990; Schmidt, Pearlman & Hunter, 1980)。然而,解释美籍西班牙人的测验分数是十分复杂的,由于不同程度地使用两种语言,也由于家庭中态度和和文化

两种变量不仅影响测验分数，而且也影响学业成就和职业成就。在这种条件下，我们不会预期预测效度受到影响。杜兰(Duran, 1983)对所发表的关于使用测验招收大学的研究进行了设计良好的调查。他注意到，关于美籍西班牙人的这些问题，修改测验并不是一种有价值的方法；相反，需要直接研究和解决这些问题。不过，解释测验分数，还是应该全面理解各种情况下的背景调节变量。而且，关于美籍西班牙人的研究结果的概化，应该考虑亚团体之间可能的差异，例如波多黎各人、美籍墨西哥人，等等。

**截距偏向** 即使一个测验在两个团体中得出相同的效度系数，它还是有可能存在截距偏向。回归线的截距指回归线与纵轴的交点。如果一个测验过低预测或过高预测某特定团体的效标成绩，那么它就存在截距偏向。让我们再看一下图 6-5 中的例 1，即少数民族样本和大民族样本具有相同的回归线。在这种条件下，既没有斜率偏向，也没有截距偏向。虽然两个样本的测验分数的平均数大不相同，但是它们的效标成绩也存在着相应的差异。另一方面，在例 3 中，两个样本的回归线的斜率相同，但是截距不同。在这种情况下，少数民族样本(A)的截距大于主体民族样本(B)；换句话说，少数民族回归线与 Y 轴的交点高于主体民族回归线与 Y 轴的交点。虽然每个样本之内所计算出的效度系数相同，但是任何一个测验分数(X)在两个样本中对应于不同的效标分数，如点  $Y_A$  和点  $Y_B$  所示。相同的测验分数对于两个团体因而具有不同的预测意义。

一些心理学家关注测验对于少数民族团体成员可能存在的不平等，他们设想例 3 所示的情景。应当指出，在这种情况下，虽然主体民族的测验平均数较高，但主体民族和少数民族的效标成绩则不分上下。假如根据为主体民族团体所建立的测验临

界分数来选拔所有的申请者，就会对少数民族申请者抱歧视态度。在这种情况下，对于两个团体都使用主体民族回归线，就会过低预期少数民族团体成员的效标成绩。如果测验方差的大部分与效标成绩互不相关，而且在所测量的机能上主体民族优于少数民族，那么就可能发生这种情况。全面的工作分析，加上令人满意的效度，会对选取这类测验起到一种防护作用。

截距偏向问题与一般所称的“测验公平”(test fairness)关系最密切。虽然有时在广义上交替使用‘测验公平’和‘测验偏向’这两个术语（包括对文化少数民族所实施的测验），但是把测验公平（或不公平）当做截距偏向已经成为惯例。《雇员选拔程序的统一指导纲要》(Uniform Guidelines on Employee Selection Procedures, 1978)中就采用这种用法。在“公平”(14B)一节中明确陈述如下：

如果某种族或性别团体的成员在选拔测验上的得分系统地低于另一团体成员，而且工作成绩度量中的差异并没有反映上述分数差异，那么使用该选拔测验就使得得分较低团体的成员不公平地失去了工作的机会。

然而，对于实际使用测验的经验研究表明，并没有显著的截距偏向，或往往有相反方向的轻微倾向，如图 6-5 中例 4 所示。此处主体民族样本 (B) 的截距大于少数民族样本 (A)。在这种情况下，如果两个团体都使用主体民族回归线和临界分数，就会过高预期少数民族团体成员的效标成绩，因而歧视倒有利于少数民族团体。在预测大学成绩 (Breland, 1979; Duran, 1983; Zeidner, 1987)、法学院成绩 (Linn, 1975)、陆军和空军训练课程的成绩 (M. A. Gordon, 1953; Maier & Fuchs, 1973; C. W. Shore & Marion,



1972),以及种种工业效标(参见 Hunter 等,1984)时,都得出这类结果。

数学上已经证明,如果两个团体在一个或多个与测验和效标都为正相关的附加变量(additional variable)上有所不同,那么就会发生图 6-5 中的例 4(Linn & Werts,1971;Reilly,1973)。微小的过高预测则是一次考虑一个预测测验的统计产物。关于从法学院和办公室人员到幼儿园儿童的若干总体的经验研究证实:成套测验中加上较多的预测测验,这种过高预测就会减少(参见 Hunter 等,1984)。

应当指出,令人感兴趣的是,在按照教育或社会经济水平来分类的团体之间进行比较研究,也得出相同的结果。“陆军分类成套测验”往往过低预测大学毕业生在军事专业课程中的成绩,同时过高预测中学辍学者的成绩(Maier,1972)。同样,各种学业能力倾向测验过低预测其父亲为专业人员的学生的成绩,而又往往过高预测来自较低职业水平家庭的学生的成绩(Flewer,1965)在所有这些研究中,对高分样本和低分样本的比较研究表明,它们没有显著的截距差异,或存在有利于测验得低分的样本的轻微偏向。

**公平使用测验的决策模型** 研究的重点逐渐开始从评价测验偏向,转移到对文化上的少数人群体公平使用测验而设计选拔策略。如果某种选拔策略采用图 6-5 中所示的回归(参见 Cleary)模型,那么将只根据所预测的效标分数来选取被试(大学招生、职业招聘,等等)。这种策略使得总效标成绩达到最大,而不管选拔过程的其他目标。按照这种策略,在选拔中公平使用测验就是指仅仅根据每一个体的效标成绩的最佳估计值。

有人也提出若干其他的决策模型,其效果是从低分团体中选取较大部分的被试。这种效果与一般称之为“肯定行为”(af-

firmative action)的目标相一致,或在选拔过程中减少“不利影响”(adverse impact)。最初采用这些不同的模型时,它们似乎使用与回归模型大不相同的程序。<sup>[10]</sup>但是后来表明,它们都能够被表示为一种综合模型的变式(Darlington, 1971; Gross & Su, 1975; Petersen & Novick, 1976)。各种模型之间的差异,能够按照每种模型固有的价值判断加以解释。本章前面部分(参见图6-2)说明过决策策略中价值观的作用。回忆一下,确定每种结果的相对效用,需要判断每种结果的利弊程度。计算决策策略的总预期效用(EU)时,我们使用这种价值判断以及每种结果的概率。

对公平使用测验的决策理论的分析表明,所提出的各种模型在公平的定义上有所不同,因为它们使用不同的价值观,接受和拒绝少数民族和非少数民族总体之内潜在的成功者和失败者。所预期的有效模型使得潜在的社会价值观成为显明的。这种方法要求公开陈述各种效用,这不是通过统计手段所能达到的,而是要求公开的讨论和不断的相互接近,以协调各种相互冲突的目标(N. S. Cole & Moss, 1989; Darlington, 1976; Messick, 1989)。所要协调的目标包括:为所有个体提供平等的机会、使成功率或生产率达到最大、增加劳动力的人口统计组合和代表性(至少在某些职业),以及优待由于以往不平等而处于不利地位的团体。

最后,应当强调一下,对测验分数、临界分数和预测公式加以统计校正。然而,这作为校正社会不平等的一种方法,其实没有多少价值。使用统计方法(即分别建立各亚团体或种族的常模)来掩盖分数差异,<sup>[11]</sup>虽然能将个体分配到与之技能或知识不相匹配的职业或教育计划中去,但这样可能反而对他们不利。结果往往是他们的成绩很差,这不仅对个体的自我概念和工作态度产生消极影响,而且强化一种成见,即来自某种文化或种族

的个体都是成绩较差者。本章前面部分使用其他方法提出了一些更有建设性的方案。一种就是多重能力倾向测验和分类策略,这可以最为全面地利用不同文化背景所培养的多样化的能力倾向类型。广泛考虑有关的人格特质、动机和态度,也有利于预测工作或学习成绩。另一种方法是采用适应性处理,例如个别化训练课程。为了最大限度符合这类个别化课程,至关重要的是测验尽可能全面和准确地反映个体目前在必不可少的能力上的发展水平。综合决策模型提供了一种框架,可以把多种多样的方法和价值系统组合起来,并且评定每种解决方案的总效果。

#### 注 释

- [1] 第三章介绍临界分数时,从不同的角度也得出相似的观点。
- [2] 内勒和夏因(Naylor & Shine) 编制的表包括更多的录取率和效度系数的数值(1965)。
- [3] 关于效用评定的其他观点,参见卡肖和莫里斯 (Cascio & Morris, 1990); 梅西克(Messick, 1989, pp. 78 ~ 81); 萨德克、坎贝尔、迪法齐奥、舒尔茨和怀特(Sadacca, Campbel, Difazio, Schultz & White, 1990) 的有关论述。
- [4] 说明所有计算步骤的若干决策策略的例子,可以参见威金斯(J. S. Wiggins, 1973/1988, 第六章)的有关论述。
- [5] 关于较全面地讨论决策理论对于测验使用的含义,参见威金斯 (1973/1988, 第六章)的有关论述,在较专业水平上的讨论,则参见克龙巴赫和格莱塞(Cronbach & Gleser, 1965)的有关论述。
- [6] 在某些条件下,单位权重或其他权重可能都优于回归权重。关于不同权重研究的小结,参见邓尼特和博尔曼 (Dunnette & Boman, 1979) 的有关论述。
- [7] 第十七章进一步讨论这个广泛使用的成套测验,及其工业和管理上的测验使用。

- 8] 从大量的文献中,我们只能引用一些有代表性的研究。关于这个问题的综述以及各方面的分析,不妨参见科尔和莫斯(N. S. Cole & Moss, 1989);亨特、施米特和劳斯琴伯格(Hunter, Schmidt & Rauschenberger, 1977)雷诺兹和布朗(C. R. Reynolds & Brown, 1984)的有关论述。
- [9] 图 6-5 所示的这种测验偏向分析,已经被称为“克利里模型”,因为在一个被广泛引用的、关于少数民族学生在大学入学考试委员会“学业能力倾向测验”分数的研究中,克利里(Cleary, 1968)使用了这种模型。这个数学程序,由格利克森和威尔克斯(Gulliken & Wilks, 1950)提出汉弗莱斯(Humphreys, 1952)建议将其应用于少数民族和性别的比较研究。图 6-5 的图形摘自戈登(M. A. Gordon, 1953)在汉弗莱斯的指导下在美国空军进行的一项研究。
- [10] 公平使用测验所用的各种决策模型的文献很多,其中许多是十分专业的。关于不同模型的主要特点和含义的简单概括,参见邦德(Bond, 1981);邓尼特和博尔曼(Dunnette & Borman, 1979, pp. 497 ~ 500);格罗斯和苏(Gross & Su, 1975, pp. 350 ~ 351)雷诺兹(1982)的有关论述。关于较为全面的解释,参见亨特和施米特(1976)以及亨特等(1977)的有关论述。
- [11] 例如,参见布朗(D. C. Brown, 1994)、戈特弗里森(L. S. Gottfredson, 1994)、萨基特和威尔克(Sackett & Wilk, 1994)的有关论述。

## 第七章 项目分析

像测验编制的其他阶段的知识一样，熟悉项目分析的基本概念和技术，能够帮助测验使用者评价所出版的各种测验。此外，项目分析尤其适合于编制非正式的地方性测验，例如教师自编的适合课堂使用的测验和考试。有效地编写项目的一些一般准则，以及较为简单的项目分析的统计技术，都能够大幅度地改进课堂测验。即使对于小团体，也值得使用。

我们能够按照其内容和形式，对项目进行定性分析；也能够按照其统计属性进行定量分析。定性分析包括考虑第五章所讨论的内容效度，以及按照编写项目的有效程序来评价项目。定量分析则主要包括项目难度和项目鉴别度的测量。任何测验的信度和效度最终都取决于测验项目的特性。通过项目分析，能够预先使得测验具有较高的信度和效度。通过选择、替代或修改项目，能够改进测验。

项目分析有可能缩短测验，同时又提高测验的信度和效度。在其他条件相等的情况下，较长测验的信度和效度都高于较短测验。第四章曾经讨论过增长或缩短测验对信度系数的影响，也介绍了估计这种影响的斯皮尔曼—布朗（Spearman-Brown）公式。如果删除的项目与保留的项目等值，或者给测验加上等值的新项目，这种所估计的信度变化才会发生。删除或加上等值效度的项目，才会发生类似的效度变化。对信度或效度变化的

所有这类估计，都涉及通过随机选择项目而不作项目分析来增长或缩短测验。如果删除令人最不满意的项而缩短测验，那么较短测验的信度和效度倒可能高于最初的较长测验。

## 项目难度

通过百分率 对于大多数测验目的，项目难度被定义为答对项目人数的百分率（或比例）。项目越容易，所通过的百分率就越大。一般认为，标准化样本中 70% 对一个词给出正确的定义（ $p = 0.70$ ）那该词比只有 15% 答对的词（ $p = 0.15$ ）容易。习惯上按照难度来排列项目，这样测验参加者先做相对容易的项目，后做难度逐渐增加的项目。这种安排使得测验参加者有信心完成测验，也可以使他们不在超出其能力的项目上浪费过多的时间，以致忽视他们本来能够正确完成的较容易的项目。

在测验编制的过程中，测量项目难度的一个主要原因是选取难度水平合适的项目。大多数标准化能力测验旨在尽可能准确地评定每一个体在特定能力上达到的水平。出于这个目的，如果没有一个人通过某个项目，那么它就是测验中不必要的东西。人人都通过的项目同样如此。这两种类型的项目都没有提供关于个体差异的信息。由于这类项目没有影响测验分数的变异性，它们对测验的信度和效度就不起什么作用。项目的难度越接近于 1.00 或 0，它提供的关于测验参加者的区分信息也越少。相反，难度水平越接近于 0.50，项目能够作出的区分就越大。假设 100 个被试中，50 个通过某个项目，另 50 个没有通过（ $p = 0.50$ ），这个项目最能使我们作出每个通过项目者和每个没有通过项目者之间的区分。我们因此具有  $50 \times 50 = 2500$  对比

较,即区分信息为 2500 比特。70%的被试所通过的项目提供的信息为  $70 \times 30 = 2100$  比特;90%的被试所通过的项目提供的信息为  $90 \times 10 = 900$  比特;100%的被试所通过的项目提供的信息为  $100 \times 0 = 0$  比特。对于通过人数不到 50% 的较难项目,同样的关系也成立。

为了得到最大区分,看来人们应该选取所有项目都为 0.50 的测验难度水平。但是,由于一个测验之内的各个项目往往彼此相关,决策就是一件复杂的事情。测验的同质性越高,项目彼此相关也越高。试举一个极端的例子,假如所有项目彼此都是完全正相关,并且所有项目都是 0.50 难度水平,100 人中同样的 50 人就会通过每个项目。结果,一半测验参加者得满分,另一半则得零分。由于项目彼此相关,最好选取难度水平适度分开的项目,但是它们的平均难度为 0.50,而且,项目彼此相关(或项目与总分的相关)越高,项目难度的分布也应该越广。

在选取合适的项目难度时,还要考虑多项选择题的猜测概率。考虑到一定比例的测验参加者靠猜测而选择正确选项,理想的正确率要定得比自由反应项目的高一些。例如,对于 5 个选项的多项选择题,平均正确率应该近似为 0.69(Lord, 1952)。

**间距表** 通过一个项目的人数百分率是按照顺序量表来表示项目难度的;换句话说,它正确指出项目的等级顺序即相对难度。例如项目 1、2、3 的通过率分别为 30%、20%、10%,我们能够作出结论,在三个项目中,项目 1 最容易,而项目 3 最难。但是我们不能推论,项目 1 和项目 2 之间的差异,等于项目 2 和项目 3 之间的差异。相等的百分比差异只是在矩形分布中对应于相等的难度差异,在这种分布中各个案在整个范围内均匀分布。这个问题类似于百分位分数中碰到的问题,百分位分数也是根据各个案百分比。回忆一下第三章,百分位分数不代表相

等的单位,从分布的中部到两端,单位大小有所不同(图 3-4)。

如果我们假设任一项目所测量的特质呈正态分布,则可以参照正态曲线频数表,把项目的难度水平按单位相等的间距量表加以表示。例如在第三章我们看到,在正态分布中,大约 34% 的个案处于平均数之上或之下一个标准差的距离之间(图 3-3)。据此,我们可以讨论图 7-1,它表示某个项目的难度水平为通过率 84%。因为在分布中分数较高的个体通过项目而分数较低的个体没有通过项目,这个 84% 就包括分数较高的一半,加上分数较低一半中的 34% ( $50 + 34 = 84$ )。因此,该项目处于平均数之下一个标准差,如图 7-1 所示。通过率为 16% 的项目处于平均数之上一个标准差,因为在这点之上,有 16% 的个案 ( $50 - 34 = 16$ )。通过率恰为 50% 的项目处于平均数,因此在这个量表上的值为零。较难的项目为正值,较容易的项目为负

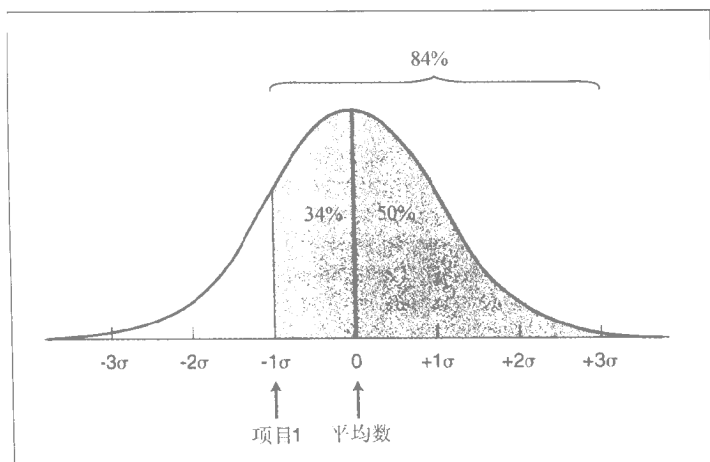


图 7-1 项目通过率和使用正态曲线单位的项目难度的关系



值。参照正态曲线频数表，我们可以得出对应于任何通过率的难度值，在任何一本普通统计教科书和适用本书的学习指南中，都附有正态曲线频数表。

**瑟斯顿绝对量表法** 以百分率或正态曲线单位来表示的项目难度指数，都限于得出这些指数的样本所涉及的能力范围。但是，出于某些目的，需要测量适用于不同能力水平样本的项目难度。例如，在教育成就测验中，最好能够在一个统一的量表上，比较一名儿童在几个连续年级上的分数。但是，对一个团体施测一些项目来排列适合于所有年级的项目，这显然是行不通的，因为几乎对于该团体中的每个人来说，有的项目太难了，有的项目太容易了。

再举一个例子，大规模测验计划需要多种等值型式，以便在不同时间实施，例如大学入学程序。第三章讨论过这个问题，因为它影响到对“学业评定测验”(Scholastic Assessment Test)之类测验的总分的解释。在那种情况下所陈述的解决办法是，使用一个固定的标准参照样本来界定量表的单位和原点，然后把所有后来的分数都转换成那个量表。这种转换需要一组参照项目即联系项目，对任何一对团体实施的测验中都包括这组项目。这组项目构成一个小型测验，因为它们代表整个测验的内容和型式。可以对不同的成对团体实施不同的联系项目。通过一连串的小型测验，每一种新的型式和一二种较早的型式相联系，它们又和其他的型式相联系，直至返回最初的参照组。

我们能够使用大体相同的方法，来测量一个统一量表上各个项目的难度，而这个量表适用于任何数量的彼此联系的团体。这种方法称之为绝对量表法，由瑟斯顿提出(Thurstone, 1925, 1947a)，并在测验编制中被广泛应用(例如，Donlon, 1984)。这种方法实际上包括两个步骤。第一步，我们在每个样本之内分别

得出每个项目的量表值，把每个项目的通过率转换成正态曲线的标准差距离即  $z$  分数值。第二步，我们选取其中一个团体为标准化团体或参照团体，把所有这些量表值都转换成该团体的对应值。可以选取任何一个团体为参照团体，例如所测的第一个团体、年龄最小的团体、在分布范围中间的团体，或某种其他方便的团体。所需要做的事情就是对两个或两个以上团体实施一组共同的参照项目，并且在每个团体之内排列这组项目。

相同的一些项目在两个(或两个以上)团体的量表值可以确定团体之间的关系，并且可以把所有的项目难度值从一个团体转换到另一个团体。图 7-2 用图解说明这种关系，它表示同一个项目在两个相邻团体 A 和 B 中的标准差距离( $z$ )。同一个项目(i), 团体 B 的通过率大于团体 A。项目(i)至团体 B 平均数的距离( $z_B$ )因此小于至团体 A 平均数的距离( $z_A$ )。所有共同项目

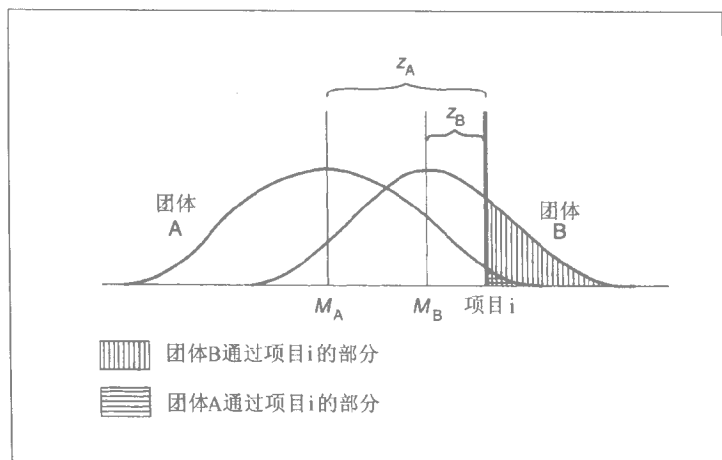


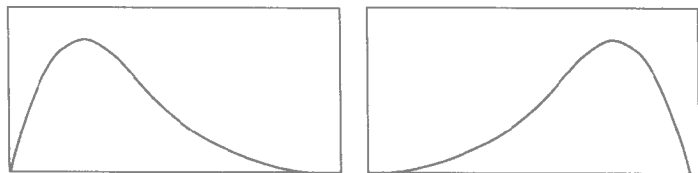
图 7-2 正态曲线两个  $z$  值,表示同一项目在团体 A 和 B 中的相对难度

的团体 A 和团体 B 的对应值,是转换公式的基础,据此我们能够把实施于团体 B 的所有项目都转换成团体 A 的难度值,反之亦然。一种简单的近似法是,对照团体 B 的  $z$  值,标绘团体 A 的  $z$  值的各点,并通过这些点画一条直线。然后利用这条直线,可以找出实施于团体 B 的所有其他项目的  $z_A$  值。

使用成对的相邻的重叠团体,可以把相同的转换方法扩大到任何数量的团体。例如,某个测验适用于一年级到八年级,可以把八年级量表转换成七年级量表,七年级量表转换成六年级量表,依次类推,直至一年级量表。两个相邻的年级团体通常十分相似,能够合用测验的大部分项目而达到联系的目的。但是,任何一个年级与相邻的较高年级和较低年级则合用不同的部分。

**测验分数的分布** 整个测验的难度当然直接取决于构成该测验的各个项目的难度。根据总分的分布,可以全面检查整个测验对于它所适用的总体的难度。如果标准化样本是这种总体的一个有代表性的样本,那么一般可以预期分数大致呈正态分布曲线。

让我们假设,所获得的分布曲线不是正态的,而是明显偏态的,如图 7-3 的 A 和 B 所示。在 A 分布中,分数集中在低端,



A. 分数集中在分布的较低端      B. 分数集中在分布的较高端

图 7-3 偏态分布曲线

这表明测验对于所研究的团体来说太难，缺少足够数量的容易项目对难度范围较低端进行适当的区分。结果是，本来会正态地分布在相当大范围的被试，在这个测验上得零分或接近零分。因此获得一个在量表低端的高峰。图 7-4 说明这种人为的分数集中，图中一个正态分布团体却产生一个在某特殊测验上的偏态分布。图 7-3 的 B 则说明一种相反的偏态，即分数集中在高端，这种结果表明测验难度不够。把一个适合于一般总体的测验，施测于经过选择的大学生或研究生样本，通常会产生这种偏态分布，一些学生几乎得满分。使用这样一个测验，不可能测量样本中能力较高学生之间的个体差异。假如这个测验包括一些较难的项目，一些个体的得分无疑会脱颖而出。

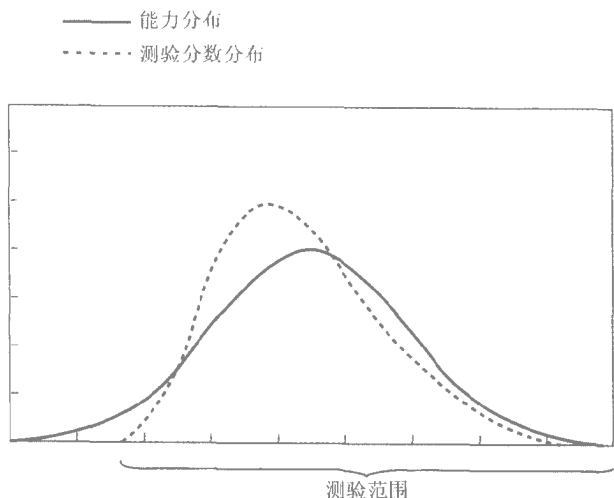


图 7-4 测验太难造成的偏态

如果标准化样本在一个测验上产生明显的非正态分布，通常要对测验的难度水平进行修正，直到近似于正态曲线。根据所呈现的偏态的类型，增加较易的或较难的项目、删除或修改其他项目、改变项目在量表中的位置，或者修改某些反应的得分权重。这类调整一直继续到分布至少大体呈正态为止。在这种条件下，大多数被试最可能得到的分数，通常是答对 50% 左右的项目对于不熟悉心理测验编制方法的人士来说，50% 的得分似乎极低。有时他们正是据此提出反对意见，说主试设定的通过测验的标准太低；或者得出推论，认为所测团体特别差。按照编制心理测验所采用的方法来看，这两种结论当然都是毫无意义的。这类测验精心编制、特意修改，正是为了产生答对的平均数大致为 50%。惟有如此，使用测验才能获得各种能力水平个体之间的最大区别。使用答对项目大约为 50% 的平均数，最有可能产生正态分布，个体分数在两端之间分布很宽。<sup>[1]</sup>

**项目难度和测验目的的关系** 一般来说，标准化心理测验的目的是得出各种水平个体之间的最大区别。我们对项目难度的讨论至此都是关于这类测验的。但是，在编制适用于特殊目的的测验时，选取合适的项目难度，以及测验分数分布的最适型式，都取决于所要求的鉴别度的类型。因此，用于筛选目的的测验，应该采用难度值最接近于所要求的选择率的项目。例如，要挑选分数较高的 20% 的个案，最好的项目便是集中在通过率为 20% 左右的项目（或考虑到猜测可以略高一点）。由于在筛选测验中，在接受组或拒绝组之内不需要作出区别，所以如果项目都集中在临界分数附近，则能最为有效地使用测验。例如，如果某个测验用于筛选大学生总体中奖学金申请者，则项目应该比大学生总体测验的一般水平难得多。同样，如果挑选学习困难者参加补偿训练课程，那么比一般水平容易得多的项目才是合乎

需要的。

按照特殊的测验目标选取项目难度水平的另一个例子便是掌握测验。回忆一下（第三章），掌握测验通常和领域参照测验相联系。如果测验的目的是确定个体是否已掌握技能的基本要素，或是否已获得必要的知识，以便进入到学习计划的下一步，那么项目的通过率最好应该为  $0.80 \sim 0.90$ 。在这种条件下，我们可以预期，大多数考生几乎答对所有项目。因此，非常容易的项目（甚至通过率为  $100\%$  的项目），在一般的标准化测验中由于没有鉴别度而被删除的项目，倒正好是掌握测验要包括的项目。同样，为了确定学生是否已经获得要教授的技能而在一个学习单元之前实施的预测中，每个项目的通过率就会很低。在这种情况下，通过率很低甚至为零的项目，也不应该被删除，因为它们揭示尚待学习什么东西。

从这些例子中显然可以看到，合适的项目难度水平取决于测验的目的。虽然在大多数情境中，集中在中间难度（ $0.50$ ）附近的项目，提供关于每一个体作业水平的最大信息，但是，如果不知道测验分数会被如何使用，则不能按常规作出关于项目难度的决策。

## 项目鉴别度

**效标的选取** 项目鉴别度是指一个项目在测验所要测量的行为上，正确区分各个测验参加者的程度。如果用效标关联效度来评价整个测验，则可以根据各个项目和同一个外部效标的关系对项目本身进行评价和选择。尤其在编制某些人格测验或兴趣测验时，多采用这种方法。第十三章和十四章将讨论这一

点。选取传记式问卷所包括的项目时,一般也采用这种方法,这种问卷一般收集异质性的个体背景资料。在这类测验中,没有先前的基础来评定某种应答的对错,或给它指定分数权重,只能和作出那种应答的个体的效标状况相比较。以不同的效标类型诸如各种职业或精神病综合征对个体进行分类,从最初的项目库中,保留那些最能区分各类个体的项目。通常,一门学业课程、一项训练计划或一类工作中的成功和失败构成各效标团体。

在第三章所叙述的领域参照掌握测验中,各类个体在有关的机能上接受不同的教育量,比较他们的项目成绩,就能评价项目(Panell & Laabs, 1979; L. A. Shepard, 1984)。通常是比较教学之前和教学之后在某个项目上正确作答人数的比例。因为这种测验用于确定个体是否达到规定的掌握水平,在一次测验上总成绩的个体差异并不重要。因此,内部的项目分析毫无意义,倒是需要外部的效标,例如有关的教学量。

在其他类型的成就测验中,例如在许多能力倾向测验中,通常是根据测验本身的总分来研究项目鉴别度。<sup>[2]</sup>对于教育成就测验来说,一般得不到外部效标。对于能力倾向测验来说,对结构效度分析的日益重视使得总分成为项目选择的合适效标。在测验编制的最初阶段,总分就是对所研究的能力、特质或结构的一种初步的、近似的测量。

让我们进一步考察根据一个外部效标和根据测验总分来选取项目的不同含义。前者往往使得根据外部效标的测验效度最大,而后者则使得测验的内部一致性即同质性最大。在某种条件下,两种方法可能得出相反的结果。根据外部效度所选取的项目,可能正是根据内部一致性所淘汰的项目。让我们假设,某个学业能力倾向测验的初步型式由 100 道算术题和 50 道词汇题组成。为了从这个最初的项目库中选取项目而采用内部一致

性方法，就要计算每个项目成绩和 150 个项目总分之间一致性的某种指标。显而易见，这样一种指标往往是算术项目高于词汇项目，因为总分中算术项目分数是词汇项目的两倍。如果希望在这个测验的最后型式保留 75 个“最佳”项目，则很可能这些项目中的大多数都会是算术问题。但是，按照学业成就的外部效标，词汇项目的预测效度倒可能高于算术项目。如果正是如此，项目分析将会降低而不是提高测验效度。

淘汰与总分低相关的项目，是使测验趋于同质的一种方法。使用这种方法，将保留彼此相关平均数最高的项目。只有当最初的项目库测量单一特质，并且在要评定的效标或结构中出现该特质时，这种选取项目的方法才会增加测验效度。但是，某些类型的测验，测量的是某种复杂的效标所要求的特质的组合。在这种情况下，使测验同质也许会减少其效标内容，因而降低效度。

为使效标关联测验效度最大而选取项目，可以比作为使成套测验产生最大效度而选取各个测验。回忆一下（第六章），对成套测验效度最起作用的测验，是和效标具有最高相关而又和成套测验中其他测验具有最低相关的那些测验。如果应用这条原理来选取项目，这就意味着最令人满意的是具有最高外部效度而又具有最低内部一致性系数的那些项目。因此，一个项目同外部效标高相关但又同总分相对低相关，另一个项目同效标和测验总分都是高相关，前者就比后者更可取，因为前者可能测量了测验的其余部分没有充分测量到的效标的某一方面。

看来，可以采用成套测验中选取测验的相同方法来选取项目。因此，每个项目可以和外部效标相关，也和其他每个项目相关。然后，使用回归方程，对用这种方法得出的最好的项目进行加权。但是，这种做法实际上行不通，理论上也站不住。不仅计



算工作过多,而且项目之间相关常发生较大的取样波动。另外,得出的回归权重也不稳定,不能为选取项目提供令人满意的基础。然而,一个更为严重的缺点是,如此编制的测验的内容太异质,无法对测验分数进行有意义的解释。

外部效度和内部一致性是编制测验的两个希望达到的目标两者的相对重要性随测验的性质和目的而变化。对于大多数测验目的来说,一个令人满意的折中方法是,把相对同质性的项目分到独立的分测验中,每个分测验覆盖外部效标的一个不同的方面。因此,覆盖宽度是通过多种分测验(每个分测验产生一个相对明确的分数),而不是通过单一测验中各个项目的异质性来获得的。采用这种方法,内部一致性指标较低的项目不是被淘汰而是被分开。结果,每个分测验或项目组之内,内部一致性相当高。

**项目鉴别度的统计指数** 由于项目的回答一般记录为正确或错误,项目鉴别度的测量通常涉及一个两分变量(项目)和一个连续变量(效标)。在某种情况下,效标也可以是两分变量,如大学毕业对大学没有毕业,或某种工作上的成功对失败。另外,一个连续效标出于分析目的也可以作为两分变量。

在测验编制中,已经提出和使用了 50 多种不同的项目鉴别度指数。其中的一种差异是对两分变量或连续变量的适用性。而且,在适用于两分变量的指数中,一些指数假设某种基本特质是连续的正态分布,对它的两分是人为提出的;另一些指数则假设一个真正的两分变量。另一种差异是项目难度和鉴别度的关系。某些测量鉴别度的指标独立于项目难度之外。在另一些指数中,接近 0.50 难度水平的项目的鉴别度值,要大于难度在两端的项目。

尽管在方法和假设上有所不同,但大多数项目鉴别度指数

所得的结果十分相似(Oosterhof, 1976)。虽然各种指数的数值不尽相同, 但根据不同的鉴别度指数而保留哪些项目和淘汰哪些项目, 则基本上是相同的。事实上, 一般来说, 样本之间的项目鉴别度数据的变化要大于不同方法之间的。

**两端组的使用** 在项目分析中, 一种通常做法是比较某个项目在两个对照效标组中的通过率。如果效标用连续量表测量, 如学科成绩、工作评定、产品记录、测验总分等, 则可以从分布的两端选取高分效标组(U)和低分效标组(L)。显而易见, 高低分两组越是极端, 鉴别度就越明显。但是, 由于所利用的个案人数太少, 使用非常极端的高低分两组, 例如高低分各 10%, 又会减小所得结果的信度。在正态分布中, 这两种条件达到平衡的理想点是高低分各 27%(T. L. Kelley, 1939)。如果分数分布比正态曲线平坦, 理想的百分数可以稍大于 27%而接近 33%(Cureton, 1957b)。对于小样本, 如一个常规教学班, 项目统计的取样误差太大, 只能获得大致的结果。因此在这种条件下, 我们不必太关注两个对照组中精确的个案百分数。从 25%到 33%之间的任何方便的数字, 都是令人满意的。

对十编制标准化测验中所采用的正态分布的大样本, 习惯上采用效标分布的高低分各 27%。便于计算项目鉴别度指标的许多表格和计算方法, 都是基于使用“27%规则”的假设。随着高速计算机的日益普及, 各种节省劳力的、简便项目分析的方法, 将逐渐被更精确的复杂方法所取代。由于计算机的便利, 最好分析整个样本的结果数据, 而不是使用高低分两个极端组。

**小样本的简易分析** 因为进行项目分析通常都使用小样本, 例如参加一次课堂测验的学生, 所以我们首先考虑一种适合于这种情境的简易方法。让我们假设, 一个班级有 60 个学生, 我们选取 20 个(33%)测验高分的学生和 20 个低分的学生。我

们现在有 3 组试卷，可以称之为高分组(U)、中间组(M)和低分组(L) 首先，我们需要计算 3 个组中学生答对每个项目的人数。这很容易完成，如果我们在一栏中列出项目号，再准备其他 3 栏，分别标为 U、M 和 L，我们拿起每个学生的试卷时，只要在他答对的项目旁边，标上一个符号。首先——记录 U 组中 20 份试卷，然后——记录 M 组中 20 份试卷，最后——记录 L 组中 20 份试卷。现在我们可以统计计算符号，并记录每组的总数，如表 7-1 所示。为了说明起见，前面 7 个项目已经写上。把 U 组中答对人数减去 L 组中答对人数，就能得出每个项目的鉴别度值的大致指数。表 7-1 最后一栏列出(U-L)的差。使用相同的数据，把 3 个效标组中(U+M+L)答对每个项目的人数相加，就能得出对项目难度的测量。

表 7-1 简易的项目分析方法：每个效标组中答对的人数

| 项目 | U<br>(20) | M<br>(20) | L<br>(20) | 难度<br>(U + M + L) | 鉴别度<br>(U - L)   |
|----|-----------|-----------|-----------|-------------------|------------------|
| 1  | 15        | 9         | 7         | 31                | 8                |
| 2  | 20        | 20        | 16        | 56 <sup>a</sup>   | 4                |
| 3  | 19        | 18        | 9         | 46                | 10               |
| 4  | 10        | 11        | 16        | 37                | - 6 <sup>a</sup> |
| 5  | 11        | 13        | 11        | 35                | 0 <sup>a</sup>   |
| 6  | 16        | 14        | 9         | 39                | 7                |
| 7  | 5         | 0         | 0         | 5 <sup>a</sup>    | 5                |
| ⋮  |           |           |           |                   |                  |
| 75 |           |           |           |                   |                  |

<sup>a</sup> 选出要讨论的项目

对表 7-1 的检查表明,所选出的 4 个有问题的项目需要进一步的考虑或课堂讨论。项目 2 和项目 7 之所以被选出,是因为一个项目太容易,60 个学生中有 56 人通过;另一个项目则太难,只有 5 人通过。项目 4 和项目 5,虽然难度水平是令人满意的,但鉴别度值分别为负值和零。我们也要考虑该栏中 ( $U-L$ ) 的差为很小正值的项目,如果比较的团体大小大致如此,那么这个值大致上为 3 或 3 以下。对于较大的团体,我们可以预期,在一个没有鉴别度的项目中,偶尔也会产生较大的差值。

在教师自编测验中,项目分析的目的,是找出测验或教学的不足之处。和班级学生一起讨论有问题的项目,通常足以诊断问题所在。如果某项目的措辞不妥,则在随后的测验中加以修改或删除。然而讨论也许表明,某项目是令人满意的,但是测验要点没有被完全理解。在这种情况下,可以检查和阐明项目的主题。为了缩小问题所在的范围,至少对所选出要讨论的一些项目进行补充分析,如图表 7-2 所示,这往往有所帮助。这张表格列出  $U$  组和  $L$  组回答特定项目时选择每个选项的人数。

虽然表 7-2 中包括项目 2,但在表中列出每个错误选项的频数,我们并不能了解更多的情况,因为  $L$  组中仅有 4 人而  $U$  组中没有一人去选择错误答案。然而,和学生一起讨论这个项目,却有助于确定是否总体上太容易以至于没有什么内在价值;是否项目编制中的一些不足之处泄露了正确答案;或是否这本是一个很好的项目,它涉及的要点恰好是教授过的并已牢牢记住。在第一种情况下,也许要淘汰这个项目;在第二种情况下,则对其加以修改;在第三种情况下,则原封不动地保留这个项目。

关于项目 4 的数据表明,第三个选项出乎意料之外,9 个高分学生宁愿选它而不愿选正确选项。这点很容易解决,只要问问那些学生,让他们解释一下为什么选择该选项。在项目 5 中,

表 7-2 个别项目的反应分析

| 项目 | 组别  | 反应选项     |           |   |           |   |
|----|-----|----------|-----------|---|-----------|---|
|    |     | 1        | 2         | 3 | 4         | 5 |
| 2  | 高分组 | 0        | 0         | 0 | <b>20</b> | 0 |
|    | 低分组 | 2        | 0         | 1 | <b>16</b> | 1 |
| 4  | 高分组 | 0        | <b>10</b> | 9 | 0         | 1 |
|    | 低分组 | 2        | <b>16</b> | 2 | 0         | 0 |
| 5  | 高分组 | 2        | 3         | 3 | <b>11</b> | 2 |
|    | 低分组 | 1        | 3         | 3 | <b>11</b> | 2 |
| 7  | 高分组 | <b>5</b> | 3         | 5 | 4         | 3 |
|    | 低分组 | <b>0</b> | 5         | 8 | 3         | 4 |
| ⋮  | ⋮   |          |           |   |           |   |
|    | ⋮   |          |           |   |           |   |

注 黑体为正确选项

不足之处看来在于题干或正确选项的措辞，因为做错这题的学生均匀地分布在 4 个错误选项上。项目 7 太难了，U 组中 15 人和 L 组中所有人都做错了。较多的人选择错误选项 3，这表明该选项表面上的吸引力，尤其是对容易被迷惑的 L 组。同样，L 组中没有一个人选择正确反应（选项 1），这表明该选项的措辞，表面上或对水平低的人来说，似乎是错误的。良好的测验项目当然需要具有这两个特点。课堂讨论也许表明，项目 7 是一个很好的项目，它涉及的要点班上没有几个人实际上学会了。

**鉴别度指数** 如果把 U 效标组和 L 效标组中通过每个项目的人数表示为百分率，则这两个百分率之差就是一种项目鉴别度指数，可以独立于得出这种指数的特定样本的大小而对它

加以解释。心理测量文献中多次论述这种指数(例如,参见 Ebel, 1979; A. P. Johnson, 1951; Oosterhof, 1976), 分别称之为  $U-L$ 、 $ULI$ 、 $ULD$ , 或简称为  $D$ 。虽然它很简单, 但同其他各种较复杂的项目鉴别度的测量十分一致(Engelhart, 1965; Oosterhof, 1976)。参照先前表 7-1 中所报告的数据, 可以说明  $D$  的计算。首先, 把  $U$  组和  $L$  组中通过每个项目的人数转换成百分率。两个百分率之差就是鉴别度指数(Index of Discrimination,  $D$ ), 如表 7-3 所示。这个指数的取值为  $+1.00$  到  $-1.00$ 。如果  $U$  组所有人而  $L$  组没有一个人通过某个项目, 则  $D = 1.00$ 。相反, 如果  $L$  组所有人而  $D$  组没有一个人通过某个项目, 则  $D = -1.00$ 。如果两个组通过某个项目的百分率相等, 则  $D = 0$ 。

表 7-3 鉴别度指数的计算

| 项目 | 通过百分率 |     | 鉴别度指数( $D$ ) |
|----|-------|-----|--------------|
|    | 高分组   | 低分组 |              |
| 1  | 75    | 35  | 40           |
| 2  | 100   | 80  | 20           |
| 3  | 95    | 45  | 50           |
| 4  | 50    | 80  | -30          |
| 5  | 55    | 55  | 0            |
| 6  | 80    | 45  | 35           |
| 7  | 25    | 0   | 25           |

注 数据取自表 7-1

同项目鉴别度的其他指数一样,  $D$  值和项目难度不是彼此独立的, 但是中等难度水平时,  $D$  值可以最大。表 7-4 说明具有不同答对百分率的各个项目的可能的最大  $D$  值。如果总样

本通过某个项目为 100% 或 0% ,则 U 组和 L 组的通过百分率没有差异,因此 D 为零。在另一个极端,如果 50% 通过某个项目,则可能是 U 组所有人而 L 组没有人通过,因此得出 D 为 100 ( $100 - 0 = 100$ )。如果 70% 通过某个项目,则 D 的最大取值说明如下:  $(U)50/50 = 100\%$  ;  $(L)20/50 = 40\%$  ;  $D = 100 - 40 = 60$ 。回忆一下,对于大多数测验目的来说,接近于 50% 难度水平的各个项目是更为可取的。因此,偏向这种难度水平的项目鉴别度指数往往适合于项目选择。

表 7-4 最大 D 值和项目难度的关系

| 通过项目的百分率 | 最大 D 值 |
|----------|--------|
| 100      | 0      |
| 90       | 20     |
| 70       | 60     |
| 50       | 100    |
| 30       | 60     |
| 10       | 20     |
| 0        | 0      |

**phi 系数** 项目鉴别度的许多指数,都以相关系数来报告项目和效标之间的关系。其中之一便是 phi 系数( $\phi$ )。 $\phi$  的计算是根据四格表,即根据某个项目在 U 效标组和 L 效标组的通过率和未通过率。像所有的相关系数一样,它的取值为 +1.00 到 -1.00。 $\phi$  系数假设项目反应和效标变量都是真正的两分变量。因此,严格说来, $\phi$  系数仅适用于得出它的两分条件,而不能概化到项目和效标所测的各特质之间的基本关系。像 D 指数一样, $\phi$  也偏向于中等难度水平——即对于最接近 50—50 的

两分变量，它得出可能的最高相关。

通过  $\varphi$  同  $\chi^2$  和  $\varphi$  同正态曲线比值的的关系，可以容易地计算  $\varphi$  系数的显著性水平。应用正态曲线比值，我们使用下面的公式可以确定达到 0.05 或 0.01 统计显著性水平时  $\varphi$  的最小值：

$$\varphi_{0.05} = \frac{1.96}{\sqrt{N}} \quad \varphi_{0.01} = \frac{2.58}{\sqrt{N}}$$

上述公式中， $N$  表示两个效标组合起来的总人数。因此，假如  $U$  组 50 人， $L$  组 50 人，那么  $N$  为 100，而达到 0.05 显著性水平的最小  $\varphi$  为  $1.96 \div 100 = 0.196$ 。 $\varphi$  值达到或超过 0.196 的任何项目，因此在 0.05 显著性水平上和效标相关。

**二列相关** 作为项目鉴别度常用度量的最后一个指数，我们可以考虑二列相关 ( $r_{bis}$ )，它在两个主要方面不同于  $\varphi$  系数。  
第一， $r_{bis}$  假设作为二分的项目反应和效标变量的基础特质都是连续的正态分布。第二，它产生独立于项目难度的项目和效标关系的度量。

二列相关的计算要利用通过项目者的平均效标分数和未通过项目者的平均效标分数、整个效标组的标准差，以及通过项目的人数百分率和未通过项目的人数百分率等。大多数统计课本中都给出计算  $r_{bis}$  的公式(例如，Guilford & Fruchter, 1978, pp. 304 ~ 306)。通过一个简单的公式，使用  $r_{bis}$  公式中本身已有的条目，可以计算  $r_{bis}$  的标准误差。还应该补充一句，如果使用计算机，则可以直接得出二列相关及其标准误差。

## 项目反应理论

项目一测验回归      在项目 — 测验回归 (item-test regression)



曲线图中，可以同时表示项目难度和项目鉴别度。为便于说明，让我们讨论一个假设的 12 个项目的测验，它要求自由反应类型的简短回答，例如个别实施的智力量表中的词汇测验。表 7-5 给出在每一个总分水平上，被试对两个项目的答对率。这些数据绘制成图 7-5。

表 7-5 说明项目一测验回归的假设数据

| 总分 | 答对率  |       |
|----|------|-------|
|    | 项目 7 | 项目 11 |
| 12 | 1.00 | .95   |
| 11 | .82  | .62   |
| 10 | .87  | .53   |
| 9  | .70  | .16   |
| 8  | .49  | .05   |
| 7  | .23  | .00   |
| 6  | .10  | .00   |
| 5  | .06  | .00   |
| 4  | .03  | .00   |
| 3  | .00  | .00   |
| 2  | .00  | .00   |
| 1  | .00  | .00   |

可以把每个项目的难度水平定为它的 50% 阈限，这是心理物理学中确定感觉阈限的惯例。在图 7-5 中就是这样做的。从两条项目曲线与 50% 线的交点，作水平轴的垂线，这就得出相应的总分。由此可见，总分约为 8 分的被试通过项目 7 的机会是 50%，总分约为 10 分的被试通过项目 11 的机会也是 50%。

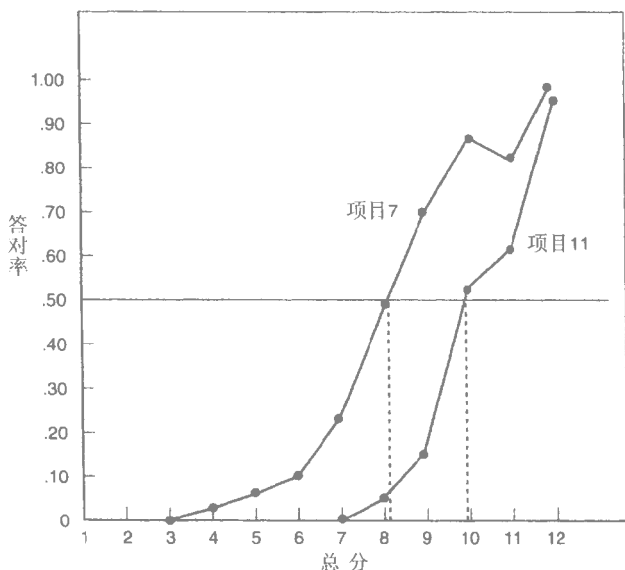


图 7-5 项目 7 和项目 11 的项目一测验回归 (数据取自表 7-5)

曲线的斜率表示每个项目的鉴别力：曲线越陡，项目成绩与总分的相关就越高，鉴别度指数也越大。凭观察，项目 7 和项目 11 的鉴别力看来大致相似。

检查一下图 7-5 中所示的项目一总分回归曲线，使我们能够直观地看出一个项目的功能如何。这类图不仅把项目难度和项目鉴别度的信息相结合，而且完整地描述项目成绩和总分之间的关系。例如，项目 7 曲线上可见一个逆转，总分为 10 的被试通过该项目的比率反而比总分为 11 的被试大。假如这些数据根据小样本，这种逆转就可以忽略不计；但是它说明，对项目

数据进行这种分析能够发现这类信息。

然而，出于数学处理以及对项目进行精确的评价和选择之目的，这类曲线图还是粗糙的、十分有限的。它是发展高度复杂的项目分析的起点，70 年代和 80 年代初期开始关注复杂的项目分析。复杂的项目分析日益盛行的一个原因，无疑是高速计算机的迅速普及。没有高速计算机，复杂的项目分析所需要的计算工作就难以进行。对于若干项目分析的模型，都准备了计算机程序，这样，实际中应用这种精细的方法才切实可行。下面几节将叙述这种方法的主要特征。

项目反应理论 (item response theory, IRT) 的基本特征<sup>[3]</sup> 现在所讨论的数学方法——项目反应理论，也可称之为潜在特质 (latent trait) 理论和项目特征曲线 (item characteristic curve, ICC) 理论。这种方法的基本特征是，项目成绩和被试的“潜在特质”的估计量有关，后者用希腊字母  $\theta$  (theta) 表示。在这种情况下，“潜在特质”表示一种统计构想，而并不表示一种独立存在的心理学或生理学的实体。在认知测验中，一般认为潜在特质是测验所测量的能力。测验总分经常作为这种能力的初始估计。

项目特征曲线的绘制是根据数学上导出的函数，而不是根据项目——测验回归曲线中所使用的经验数据。不同 IRT 模型根据各自不同的假设而使用不同的数学函数。一些模型使用正态拱形函数 (即累积正态分布) 另一些模型使用逻辑斯谛函数，即利用对数关系的一些数学上很方便的性质。一般来说，用不同模型所获得的结果基本相似，假如它们的假设在特定情境下符合的话。图 7-6 表示 3 个假设项目的项目特征曲线。水平轴为能力量表 ( $\theta$ )，根据特定样本中测验总分和可得到的关于测验反应的其他信息来加以估计。纵轴为  $P_i(\theta)$ ，表示对项目  $i$  的正确反应的概率，它是个体在能力量表 ( $\theta$ ) 上的位置的函数。这

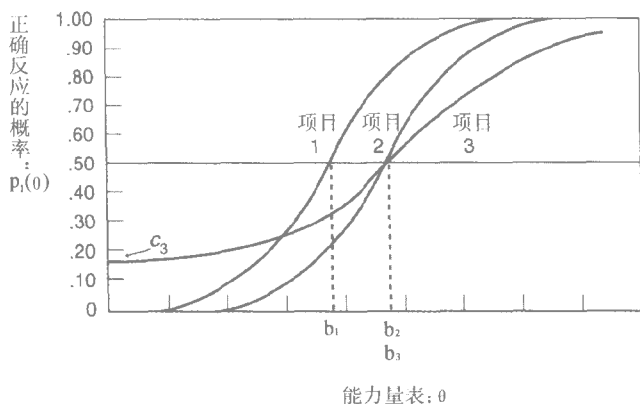


表 7-6 3个项目的假设的项目特征曲线

个概率的得出，是根据不同能力水平上通过该项目的人数比率。

在完整的三参数模型中，每一条项目特征曲线由 3 个参数来描述，而参数在数学上根据经验数据得出。项目鉴别度参数 ( $a_i$ ) 表示曲线的斜率。为了增加  $P_i(\theta)$ ，人们必须在能力量表 ( $\theta$ ) 上前进一段距离，而项目鉴别度参数则与这种距离成相反关系。 $a_i$  值越大，曲线越陡。图 7-6 中，项目 1 和项目 2 具有相同的  $a_i$ ，即鉴别度值，而项目 3 的  $a_i$  值较小，其曲线上升较慢。项目难度参数 ( $b$ ) 表示正确反应概率  $P_i(\theta)$  为 0.05 时在能力轴上的位置。可以看到，项目 2 和项目 3 具有相同的  $b_i$ ，而项目 1 则较为容易，对于 0.05 的正确反应概率，它需要的能力较低。用于处理多项选择题的 IRT 模型通常还包括第三个参数，即所谓的猜测参数 ( $c_i$ )<sup>[4]</sup>，它表示偶然得到的正确反应的概率。在多项选择题中，即使能力水平很低的被试，其得出正确反应的概率也会

大于零。图 7-6 中项目 3 说明了这一点，其渐近线低端远在零之上。

对项目参数和能力的估计一般都使用重复的或逐次近似的方法来计算；重复近似直到数值稳定为止。除了得出数学上精确的项目难度和项目鉴别度的指数之外，IRT 技术还具有若干其他的长处。这种方法的一个重要特征是，通过项目信息函数 (item information functions) 来处理信度和测量误差。计算出每个项目的项目信息函数，这样就为在测验编制中选择项目提供了可靠的依据。项目信息函数考虑到项目的所有参数，并且表明在不同能力水平上项目的测量功效。

IRT 最为公认的贡献是，其结果具有样本自由的性质，技术上称之为项目参数不变性 (invariance of item parameters)。在不同能力的各个团体中计算项目参数时，项目参数应该不变，这是 IRT 理论的一个基本概念。这不仅意味着，对于不同的团体可以使用一种统一的测量量表；也意味着，对于不同的团体或个体，可以使用适合于其能力水平的一组不同的项目进行测验，并且其分数可以直接比较。每一被试的测验分数，不仅根据答对项目的数量，而且根据答对项目的预先定出的难度水平。

当测验许多不同的样本时，一种方法便是使用根据随机大样本已经预先校准的大型项目库即项目银行。当能力分布范围很大时，例如成就测验系列涉及学校中许多年级，就有必要使用一些共同项目 (称之为参照项目、联系项目、校准项目等)，以便解决各个团体之间的差距。一旦项目库里的项目被校准，那么这些项目的任何子集都可以用来施测任何团体或个体，并且所得分数之间可作比较。

其他的 IRT 模型 在前面一节中，我们讨论了三参数模型。当猜测对测验成绩的影响可以忽略不计时，删去机遇反应参

( $c_i$ ) 的二参数模型也是合适的。一参数模型只考虑一组项目的难度( $b_i$ )，它是由拉希提出的 (Rasch, 1966; 也参见 Andersen, 1983)，后又得到一些研究人员的大力支持和改进 (例如 Wright, 1977; Wright & Stone, 1979)。这一模型假设，猜测和项目鉴别度差异均可以忽略不计。在编制测验时，拉希模型的支持者通常要淘汰那些没有满足这一假设的项目。也有人认为，IRT 模型在统计意义上十分‘强大’(robust)，这意味着可以适度违背各种假设而不致歪曲测验结果。当然，这是一个需经验验证的问题。

至此，所讨论的各种模型都假设测验的单维性 (unidimensionality)，即假设项目反应归因于某单一特质。一般来说，如果测验成绩取决于单一的主要特质，那么，即使其他特质对成绩有些许的影响，单维性假设也可以予以适当的满足。现在也设计出适用于多维测验的更为一般的模型，只是计算过程更为麻烦。还提出了一些其他变式来处理等级（不是对错或两分反应）项目反应 (Samejima, 1969)，或分析多项选择题中不同的反应选项 (Bock, 1972)。

**目前状态** IRT各种模型的相对长处仍有很大争议。需要对数学上导出的值作更多的检验，不仅使用人为的数据和计算机模拟，而且也使用日常生活数据。项目参数不变性尤其需要在各种现实情境中进行广泛研究。例如，当被试具有不同的经验背景或同一被试在不同的学习阶段，他们所完成的相同项目可能涉及不同的能力组合。从另一个角度来看，虽然已经编制出适用于 IRT 数据的多种计算机程序 (例如，参见 Hambleton, 1989, pp. 171 ~ 172)，但这些程序由于再评价、修订、取代等而处于不断的变化之中。<sup>[5]</sup>

虽然也在提出多种理论和方法，但在实际的测验编制中更多地使用 IRT 方法。IRT，技术正迅速体现在新近编制的测验

中,以及由测验出版商发行的、被人们广泛使用的成套测验的修订版中。其中包括第八章中讨论的“加利福尼亚成就测验”(California Achievement Tests)、“基本技能综合测验”(Comprehensive Tests of Basic Skills)和“区别能力量表”(Differential Ability Scales)。IRT尤其适合于某些新近出现的测验,例如第十章中讨论的计算机化适应测验(CAT)。在这类测验中,每个测验参加者对一组不同的项目作出反应,但是所有项目都使用同一个量表来记分(Wainer 等,1990)。IRT 的一项重要应用则是发展“军队服务职业能力倾向成套测验”(Armed Services Vocational Aptitude Battery)CAT 版本的长期规划(Wiskoff & Schratz,1989)。

## 快速测验的项目分析

不管速度是否和正在测量的机能相干,从快速测验中计算出的项目指数也许会让人产生误解。除了所有被试或几乎所有被试都有时间去做各个项目之外,从速度测验中得出的项目指数将反映项目在测验中的位置,而不是项目内在的难度或鉴别度。测验中排在后面的项目,总样本中的通过率相对较少,因为只有少数被试有时间做到这些项目。不管项目如何容易,如果在速度测验中排在后面,则它看起来就是很难的。如果某项目在快速测验中排在最后的话,即使它只是询问姓名,被试的通过率也会很低。

同样,对于并非由所有测验参加者都做到的那些项目来说,项目鉴别度指数往往被过高估计。因为水平较高的被试往往做得快,他们在速度测验中较有可能做到排在较后面的项目。因此,如果某项目在速度测验中排在后面的话,不管项目本身的性

质如何，总会得出项目和效标之间的某种相关。

为了避免其中的一些困难，我们把每个项目的分析限于做到该项目的那些被试。然而，这不是一个完全令人满意的解决办法，除非没有做到该项目的人数甚少。这种办法中使用的被试人数大为减少，因此使得较后面的项目的分析结果十分不可靠。而且，对较后面的项目进行分析时所使用的被试，可能构成一个经过选择的样本，因此不能同较前面的项目所使用的较大样本进行比较。我们已经指出，做得较快的被试往往也是水平较高的，因而较后面的项目是按照一个水平较高的被试样本来进行分析的。这类选择因素的一种影响便是降低较后面项目的显见难度水平，因为在经过选择的水平较高样本中的通过率要大于整个样本。应当指出，这种误差同按照整个样本来计算通过率时的误差正好相反。在那种情况下，显见的项目难度被假性提高。

上述方法对项目鉴别度的影响不太明显，不过也是存在的。例如，已经观察到，一些得低分的测验参加者往往匆忙完成测验，他们几乎随机做项目而努力在准许的时间之内完成所有项目。这种倾向在得高分的测验参加者中就较少出现。结果，对后面的项目进行分析所依据的样本中，可能包括一些很差的被试，他们在项目上的成绩同随机无异。但也包括许多水平很高且做得快的被试，他们可能答对项目。在这样一个样本中，项目一效标相关可能大于较有代表性的样本。另一方面，缺少这类随机被试，对较后面的项目进行分析所依据的样本，则包括较小的能力范围。在这种条件下，较后面的项目的鉴别度指数往往小于依据未经选择的整个样本所得出的数值。

速度对项目难度和项目鉴别度指数的预期影响得到经验证实，不论使用整个样本(Wesman, 1949)，还是只使用做到该项目



的那些被试(Mollenkopf, 1950a)来计算项目统计量。在后一个研究中,对中学生各可比较团体实施两种型式的言语测验和两种型式的数学测验。在两种型式中,每种型式包括的项目和另一种型式相同,只是在一种型式中放在前面的项目在另一种型式中则放在后面。每种型式分别以短促的时限(速度条件)和充裕的时限(难度条件)进行实施。因此可能在型式和时限条件之间进行相互比较。研究结果清楚表明,在速度测验中一个项目的位置影响其难度和鉴别度的指数。当同一个项目在快速测验中放在较后面时,做到该项目的被试的通过率较大,并且项目一效标相关也较高。

快速测验的项目分析中碰到的困难,基本上类似于第四章所讨论的关于速度测验信度的困难。为了解决这些困难,已经提出种种解决方法,有经验方法,也有统计方法。一种经验解决方法是,对于进行项目分析所依据的样本,使用较长的时限实施测验。假如速度本身不是测验所要测量的能力的一个重要方面,这种解决方法还是令人满意的。然而,除了具体测验所提出的技术问题之外,最好记住,使用快速测验所得出的项目分析数据是可疑的,需要加以仔细研究。

## 交叉效度分析

**交叉效度分析的含义** 计算测验效度所依据的被试样本,必须不同于选择项目所依据的被试样本。这种独立确定整个测验的效度称之为交叉效度分析(cross-validation)。如果计算效度系数和选择项目都依据同一个样本,那么这种效度系数就会利用那个特定样本之内的随机取样误差,结果就是假性的高。事

实上,在这种情况下,甚至当测验在预测特定的效标中根本没有效度时,也会得出较高的效度系数。

让我们假设,一个样本为 100 名医学院学生,从中选取医学院成绩最高 30 名和最低 30 名作为相对照的效标团体。现在,如果以一些实际上与医学院成功毫不相干的特质来比较这两个团体,无疑会发现某种随机差异。比如,在高分效标团体内可能有许多私立学校毕业生和红头发学生。假如我们给每一个体评分时,采用私立学校毕业生记一分,红头发也记一分,那么这类分数的平均数,当然是高分效标团体高于低分效标团体。然而,这并非是预测因子效度的证据,因为这种证实过程根据一种循环论据。这两个预测因子的选择首先是根据这个特定样本的随机特性。另外,在产生总分的平均数差异时,同样的随机差异也在起作用。然而,当测验另一个样本时,私立学校毕业生和红头发的出现率中的随机差异就可能消失或相反,分数的效度因而不复存在。

一个经验的例子 需要交叉效度分析的一个经典例子,是使用罗夏墨迹测验(Rorschach Inkblot test)进行的一项早期研究(Kurtz, 1948)。为了确定罗夏测验是否有助于选拔人寿保险公司的营业经理,对 80 名这类经理实施该测验。这些经理是从 8 家人寿保险公司所聘用的数百人中精心选出来的,被各自公司认为十分满意的 42 人为高效标组,而被认为不满意的 38 人则为低效标组。80 份测验记录由一名罗夏测验专家加以研究,他选择了一组 32 个迹象或反应特性,它们在一个效标组中出现的次数比另一个效标组多。在高效标组中出现较多的迹象,如果有,记 1 分 如果没有,则记 0 分。在低效标组中出现较多的迹象,如果有,记 -1 分 如果没有,则记 0 分。因为每种类型有 16 个迹象,理论上总分范围为 -16 到 +16。

把基于这 32 个迹象的记分答案再次应用于最初的 80 人团体,其中 79 人被正确分类为高效标组或低效标组。测验分数和效标之间的相关因此接近于 1.00。但是,另一个比较样本有 41 名经理,其中 21 人为高效标组,20 人为低效标组。当依据这第二个样本对测验进行交叉效度分析时,效度系数就下降到微不足道的 0.02 了。显而易见,在第一个样本中所得出的记分答案对于选拔这类人员毫无效度可言。

一个具有随机数据的例子 丘顿 (Cureton, 1950) 的一项经典研究清楚表明,使用单一的样本来选择项目和证实测验效度,在纯随机条件下,也会得出完全假性的效度系数。要预测的效标是注册心理学课程的 29 名学生每人的平均成绩。效标成绩分为两种:B 及以上和 B 以下。“项目”由 85 个标签组成,在一个面上写上数字,从 1 到 85。为了获得每个学生的测验分数,把 85 个标签装入一个容器中摇动,然后落在桌上。对落下时数字一面朝上的所有标签加以记录,这表示在学生成绩中包括那个特定项目。将 85 个标签掷 29 次,因而得到每个学生的完整记录,这说明在每个学生的成绩中是否包括各个项目或反应。由于在产生这些随机分数中所使用的方法,丘顿把该测验戏称为“乙种心理运动投射测验”(B-Projective Psychokinesis Test)。

然后,以每个学生的平均成绩为效标进行项目分析。由此从 85 个“项目”中选出 24 个。其中 9 个项目在平均成绩 B 及以上的学生中间出现较多,每个权重为 +1 另 15 个项目在平均成绩 B 以下的学生中间出现较多,每个权重为 -1。这些项目得分的加和,构成每个学生的总分。尽管知道这些“测验分数”的得出是随机的,但在最初 29 名学生的团体中,测验分数和成绩效标的相关为 0.82。这个研究结果类似于前面引用的研究中使用罗夏测验分数所获得的结果。在两个例子中,测验分数和效

标之间表面上的对应，是由于在选择项目和确定测验总分效度时利用相同的随机差异造成的。

**影响效度减小的因素** 交叉效度分析中效度系数的减小量，部分取决于最初项目库的大小和所保留的项目的比率。当最初的项目数量很大而所保留的比率又很小时，利用随机差异的机会就多了，因此得出假性高效度系数。影响交叉效度分析中效度系数的减小量的另一个因素是样本的大小。由于最初样本中假性高效度是取样误差的累积所造成的，较小的样本（它产生较大的取样误差）就出现较大的效度减小。

如果项目的选取是根据先前提出的假设，而这种假设源自心理学理论或以往使用效标的经验，交叉效度分析中效度的减小就是最少的。例如，假如某个假设要求成功学生中“是”的回答应该较多。一旦不成功学生作出明显较多“是”的回答，那么该项目则不予保留。相反，盲目的经验方法则是：拼凑一组杂七杂八的问题，不大考虑它们同效标行为的相干，然后保留同效标产生显著的正相关或负相关的所有项目。我们可以预期，后面一种情况下的效度减小远大于前面一种情况。米切尔和克利默斯基(Mitchell & Klimoski, 1986) 提供的一个设计良好的实例说明，当根据基本原理或经验策略来选取项目时实际产生的效度减少的差异。总之，如果样本很小，最初的项目库很大，所保留的项目比率很小，项目的汇编没有考虑先前提出的基本原理，那么交叉效度分析中测验效度的减小就是最大的。

## 项目的不同功能

统计学的方法“项目偏向”的分析日益受到关注，这是研

究少数民族团体的测验偏向的一个方面。这类分析基本上关注各个测验项目对于不同的文化或经验背景的团体的相对难度。以心理测量的术语来说,项目分析的这个领域称之为项目的不同功能(differential item functioning, DIF)。它试图鉴别的项目是,不同文化团体中相同能力的个体具有不同的项目成功概率。出于这个目的,相同的能力意味着测验所要评定的结构相同,或测验所要预测的效标行为相同。人们提出了许多方法来鉴别这类具有不同功能的项目,包括统计学的方法和判断的方法(Berk, 1982; Camilli & Shepard, 1994; Hambleton & Rogers, 1989; P. W. Holland & Thayer, 1988; P. W. Holland & Wainer, 1993; Osterlind, 1983; C. R. Reynolds & Brown, 1984)。

一个重要问题是,项目难度的人口统计上的(或其他经验的)团体差异与整个测验上成绩水平的平均差异密切相关。结果,按照测验总分具有良好鉴别度值的项目,看来倒可能是“有偏向的”项目而遭到淘汰。可以使用若干方法来控制这种总分差异。随着计算机的日益普及,最有希望的方法之一基于项目反应理论(IRT)。如果可以使用大型样本,这些方法尤为合适。本章前面部分说过,每一项目的项目特征曲线(ICC)表示正确反应的概率和测验的能力量表的关系(图 7-6)。比较一下两个团体中同一项目的 ICC,我们能够鉴别相对于在统一的量表上所表示的各团体的测验总成绩而言,具有实质性的不同功能的那些项目。图 7-7 中用两个项目说明这种比较。可以看到,对于项目 1,团体 A 和团体 B 的 ICC 截然不同,而对于项目 2,两条 ICC 则十分相似。对于每个项目来说,可以使用两条 ICC 之间的区域来鉴别所测能力的范围,其中有 DIF 的证据。无论使用什么方法,一旦鉴别出 DIF 项目,下一步则是探索差异的性质和来源。对这个问题的回答决定对项目的处理。出于这个目的,

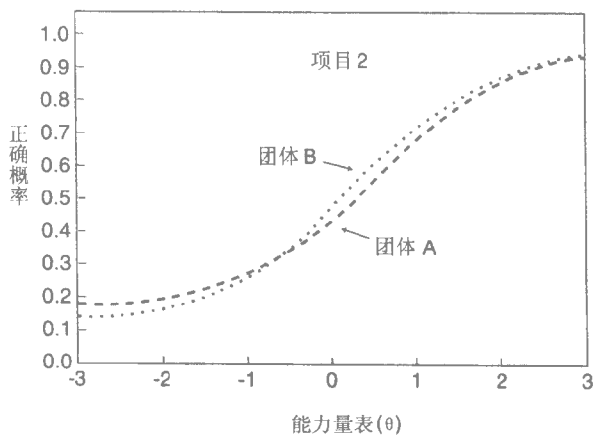
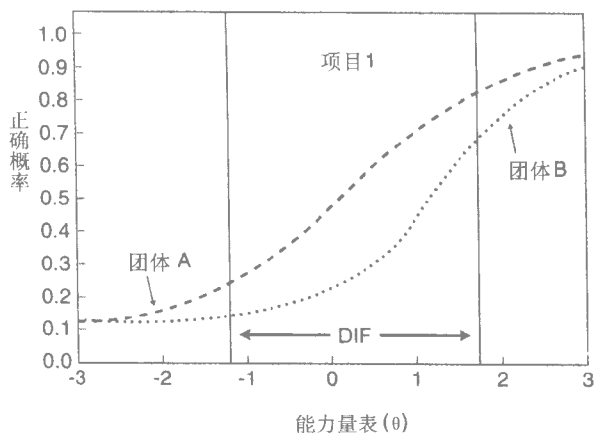


图 7-7 两个项目的项目特征曲线(ICC),说明项目具有不同大小功能(DIF)(摘自帕希利,1992)。

可能需要判断的方法，也许要和进一步的统计方法相结合。

**判断的方法** 对于各种目的来说，不存在单一的项目偏向分析的“最佳方法”。因为不同的方法提供不同类型的信息，所以最好联合使用各种方法。合适的联合取决于所预期的测验用处，也取决于从分数中所作出的推论的类型。一般来说，最佳联合包括一些统计学的方法和一些判断的方法。

当恰当应用判断的方法时，它可以提供其他方法得不到的有用信息(Scheuneman, 1982; Tittle, 1982)。在测验编制的最初阶段和最后阶段，在统计分析之前和之后，判断分析尤为合适。测验编制伊始，通常采用判断分析，以便找出可能冒犯少数民族或降低少数民族身份的内容，或者在职业角色和其他社会角色上存在文化偏见的内容。出于这个目的，大多数测验出版商经常由他们自己的工作人员或有代表性的外界顾问进行初步的项目检查(Berk, 1982，第九章)。这类检查也有助于鉴别因文化局限性而对特定的测验参加者总体来说不熟悉的测验内容。但是，应当指出，在预测项目对于不同总体的相对难度或相对鉴别力方面，判断的检查一般来说尚未证明是成功的(Plake, 1980; Sandoval & Miille, 1980; Scheuneman, 1982)。出于这个目的，需要对经验结果作统计分析。

另一方面，不能假设由统计方法所鉴别出的所有不正常项目都是有偏向的。需要通过第二次不同类型的判断检查再来解释结果。在这个阶段，对项目作检查是为了找出不正常项目的可能的来源。统计结果通常没有揭示不正常项目的共同特征或明显的原因；每个项目都需要个别考虑。一些统计结果可能只是反映由特定方法而引起的统计现象。在另一些情况下，不正常的成绩是由对测验解释具有不同涵义的多种因素造成的。对这类不正常项目作出合适评价，要求熟悉测验的内容领域和各

个测验参加者总体的经验差异。

不正常项目的一种可能原因是,它没有测量不同群体中的相同结构。例如,某个比喻语,对一个群体来说,可能测量的是言语推理,但对另一个群体来说,测量的也许是词汇知识了,因为它包含的一个关键词汇,是这个群体许多成员都不熟悉的。同样,某数学问题可能是测量的是一个群体的数学能力,而对另一个群体来说是测量其理解复杂的言语陈述的能力。在这两个例子里,缺乏的知识同整个测验所测量的结构并不相干。但是,假设我们发现,包括小数的数学项目对于某群体的成员相对较难,这种差异就同数学能力的结构有关。因此,在这种意义上,这类结果就不是有偏向的项目。

当统计上鉴别出不正常的项目时,通过进一步的统计方法,例如分析多项选择题中被选中的错误选项,可以部分阐明不正常的来源。这些补充分析,再结合判断的检查,应该提供如何适当处理项目的基础。不正常的项目,可以淘汰,可以全盘修改,也可以只修改不足之处。测验指导语,可以扩充,也可以阐明。在按照测验说明再次考虑之后,也可以保留最初的项目。项目分析甚至可以建议,对测验说明本身进行再评价。结果或是改变测验说明,或是阐明从测验分数中能够作出的特定推论。

误用 DIF 的一个著名案件 有一个案件主要的依据是“测验偏向”评价中的项目分析,这也许是法院所受理的第一个此类案件,因而得到了大力宣扬。这个案件被称为“戈尔登·鲁尔案”(Golden Rule case),因为它涉及戈尔登·鲁尔保险公司招聘员工时使用执照考试。公司名称中有两个词,但其常用的意义完全不同,这可能引起普遍混乱。案子的最终解决是比较某项目的纯通过率的团体差异,而不是力求使各团体在测验所评定的能力的任何度量中相等,也不考虑适用于测验预期目的的项目效



度。显然，这种决策完全违反项目具有不同功能的概念，并且可能删除了最能预测工作成绩的项目。

因为其广泛的误解及其对职业上和教育上使用测验可能产生的影响，在批评性评价中，从不同的角度来讨论“戈尔登·鲁尔”判决中的谬误(例如, Lim & Drasgow, 1990), 包括美国心理学会(American Psychological Association, APA)的一份正式声明。<sup>[6]</sup>这个判决也是 APA 年会专题讨论会的一个议题，其中大多数论文后来发表于专刊《教育和心理测量：问题和实践》(*Educational and Psychological Measurement : Issues and Practices*, Bond, 1987; Faggen, 1987; Linn & Drasgow, 1987) 上。讨论这个著名的案件旨在强调, 试图通过表面性的、不完全的指征来评定‘测验偏向’会产生潜在的实际危害。

## 项目编制的探索

80年代和90年代计算机利用的迅速发展，加上认知心理学的进展，促进广泛研究项目编制的新方法。传统上项目编写与其说是一门科学，不如说是一门艺术。甚至在最好的条件下，项目编写者接到的指示也只是规定了项目型式和内容范围。通常的做法仍然是依靠项目的经验预测去评定其难度水平和鉴别力。在预测之前，仅仅分析一下刺激的物理或语义特性，有没有办法去预测这些项目的统计量？更好一点，能不能编制出具有所要求的难度和鉴别度值的项目？能不能系统操纵刺激特性而预先确定测验项目的认知要求？在不断深化的研究中，通过实验和数学两种方法，这些问题正在得到进一步的探究 (Bejar, 1985, 1991; Carroll, 1987; Embretson, 1985a, 1985b, 1991, 1994,

1995; Freedle, 1990).

通过认知心理学所发展的任务分解技术,能够探索测验刺激的认知要求。使用这种方法,能够研究项目的不同特性与作业的速度和错误之间的关系。对于空间项目,已经进行了若干这类研究(Embretson, 1994; Pellegrino, Mumaw & Shute, 1985)。例如,某空间类推测验中所呈现的刺激能够被分解为:(1)复杂性,即必须被识别的独立元素的数目(例如,形状、大小、位置);(2)转换,即要评价的一对刺激之内,刺激变化的方法的数目。在某几种空间想象问题中,要求测验参加者选出物体的各部分,把它们拼配成一个给出的整体,各部分可能只是被分开,或被移动,或被转动,或把这些方法结合起来加以变动。

另一些研究则是关于言语刺激的语义特性。例如,在言语推理测验中,能够按照已知的逻辑原理来编制项目(Colberg, 1985; Colberg, Nester & Trattner, 1985; Scheuneman, Geritz & Embretson, 1991; K. Sheehan & Mislevy, 1989; Shye, 1988)。这种方法能够保证,只有一个反应选项是真正正确的,并且在项目例题中,按某个预先确定的比例呈现不同的逻辑关系。这种方法也能够操纵项目的逻辑复杂性,然后能够经验地研究它和难度水平的关系。一些研究人员进行实验,编制字母系列去测验归纳推理(Butterfield 等, 1985)。首先为系统编制这类字母系列而提出一组详细规则,然后形成关于试图理解某个系列时个体要做些什么的假设。通过对完成项目难度的经验研究来检验假设。

恩布里森(Embretson, 1994)彻底分析并更新了项目编制的过程。这个过程首先界定要评定的结构,进而提出测验的认知模型。这种认知模型的详细特征为项目编制提供具体说明。项目

的经验效度分析在于确定在实践应用中这些项目实际上符合认知模型的程度。“空间学习能力测验”(Spatial Learning Ability

Test) 的编制说明了完整的程序, 该测验不仅测量最初的空间能力, 而且测量标准化指导之后的可变性。

关于根据刺激的物理和语义特性预测项目难度的研究, 不仅促进项目编写者去编制有效的测验, 而且导致使用计算机来编制项目。诚然, 能够把详细的项目说明编入计算机程序(例如, 参见 Butterfield 等, 1985; Embretson, 1994)。毫无疑问, 这些测验编制新方法的潜在优点给人深刻印象。然而, 我们必须谨防对任何一种方法期望过高。例如, 情况往往可能是, 某个测验全面而有效地测量某些可明确辨认的认知结构, 但是对于某种重要的实践用途则没有高度的预测效度。出于这种原因, 就有必要既考虑结构效度分析, 恩布里森(1983)称之为结构描述, 也考虑一般规律广度(nomothetic span)。任务分解提供了结构描述的信息; 一般规律广度则要求研究测验分数与其他外部变量网的关系, 其中包括效标测量。其次, 不要过度概化; 因为在任何一个论题领域或专门知识领域的出色表现, 都需要有关的内容知识。过程往往与内容相联系, 缺乏合适的内容, 就不能有效地评价过程。

最后需指出, 本节所引述的新方法被恰当应用时, 对系统地、有控制地编制测验项目有很大的帮助。此外, 通过鉴别测验所测量的结构, 这些方法有助于我们理解为什么特定的测验可以预测效标情境中的表现的原因。另一个益处同测验的诊断用途有关, 我们能够借助于它, 将被试与特定的认知过程相联系。这些都是美好的目标, 但实际贯彻仍然有待于对尚未解决的问题的大量研究(例如, 参见 Wainer, 1993a)。目前对项目编制正在进行许多研究, 以便确定各被试在解答特定的项目时所使用的认知过程(Willson, 1994)。分析各被试的错误类型, 为此目的提供了有价值的线索(Kulikowich & Alexander, 1994)。

## 注 释

- [1] 实际上, 正态曲线中, 量表两端的鉴别度比中部的要好。在量表的所有点上都相等的鉴别度则需要均匀分布。但是, 如果随后要对分数进行统计分析, 正态曲线自有其长处, 因为目前的许多统计技术都假设分布近似为正态。由于这种原因和其他原因, 未来一段时间里, 大多数适合于一般使用的测验可能会继续采用正态曲线型式。
- [2] 测验及其项目中共同的具体方差和误差方差会稍稍增大项目一测验的相关, 可以利用公式来校正这种部分一整体的影响 (Guilford & Fruchter, 1978, pp. 165 ~ 167)。
- [3] 关于 IRT 方法学及其应用的总览, 参见汉布尔顿等 (Hambleton et al., 1991) 的有关论述。较为技术性的评述和批评性的评价, 参见汉布尔顿 (1989)、德拉斯高和休林 (Drasgow & Hulin, 1990) 的有关论述。关于心理测量学中引用 IRT, 参见洛德 (Lord, 1980)、韦斯 (D. J. Weiss, 1983)、韦斯和戴维森 (Davidson, 1981) 的有关论述。
- [4] 一些研究人员建议,  $c_i$  只应该称之为渐近线低端或 ICC 的机遇参数, 因为三参数模型认为  $c_i$  独立于能力, 而猜测也是能力的一种机能。
- [5] 当前一个著名的例子是 ASCAL 程序, 适用于二参数或三参数的逻辑斯蒂 IRT 校准, 由评价系统公司 (Assessment Systems Corporation, 地址参见附录 B) 销售。另一个例子是新近发展的概化线性项目反应理论 (GLIRT), 从中可以导出不同的 IRT 模型, 它也可以适用于不同的项目型式 (Mellenbergh, 1994)。
- [6] 由美国心理学会心理测验和评定委员会提出, 美国心理学会各有关委员会和代表委员会都同意这份声明。

## **第三部分**

## **能力测验**



## 第八章 个别测验

在第二部分，我们讨论了心理测验的主要原理，现在我们准备应用这些原理来评价各种具体测验。我们已经知道，关于每个测验要问些什么问题，以及到哪里去寻找答案，测验手册和《心理测量年鉴》是主要资料，从中可以查到所引用的任何测验的有关信息。<sup>[1]</sup>

本书其余部分的讨论有双重目的。一个目的是有机会看一下测验原理在各种测验中的应用，另一个目的是使读者熟悉每个主要领域中的一些著名测验。我们并不想对任一领域之内可得到的测验作全面评述，这种评述超出本书范围。而且，这种评述在发表之前就可能过时，因为新的测验或修订的测验迅速出现。由于这些原因，讨论将集中于每一类型中一些有代表性的测验。之所以选择这些测验，或因为其得到广泛使用，或因为它们标志着测验方法中的重要发展。我们在第三部分讨论能力测验，在第四部分讨论人格测验，在第五部分讨论在各种情境中的测验应用。除非另外说明，书中所讨论的测验的所有资料，均取自测验出版者提供的测验手册或补充技术资料。读者自己要考察特定的测验，可以查阅用于本书的《研究指南》中所提出的测验评论概要（Urbina, 1997）。为此，测验标准（AERA, APA, NCME, 1985）提供了更为详细的准则。

本章和下一章所要讨论的一类测验，即所谓传统的“智力测

验”，都直接源于最初的比纳量表。这类测验适用于种种情境，以相对广义的效标来建立效度(参见 L.R. Aiken, 1996)。它们的特征是提供单一的概括分数，例如传统的 IQ，作为被试成绩一般水平的一个指标。它们通常也产生分测验或分测验组的分数，来评定狭义的能力倾向。由于那么多智力测验的效标是学业成就度量，因此，它们经常被称之为学术能力倾向测验或学业智力测验。智力测验常被用作初步的甄别工具，第二步再用特殊能力倾向测验进行施测。在正常青少年或成人的教育和职业咨询、人员选拔以及类似目的的测验中，这种做法尤为盛行。一般智力测验的另一个常被采用之处是在临床测验中，尤其是在智力落后个体的鉴别和分类中。对于临床目的，一般采用个别施测的测验，本章所讨论的斯坦福一比纳量表和韦克斯勒量表是使用最广的个别(与团体相对照)智力测验。斯坦福一比纳量表是本书要讨论的第一个测验，所以对它的讨论比全书其他测验更全面。这样做是为了在开始时就说明在评价一个测验时要考虑的信息的类型。然而，应当指出，不要把对全书中特定测验的讨论看做测验评论，如同《心理测量年鉴》之类资料中提供的那样。出于本书的目的，通常注意某一特定测验的特殊贡献，或区别于其他测验的特点。<sup>[2]</sup>

## 斯坦福一比纳智力量表

■表的发展 最初的比纳-西蒙量表，1905 年、1908 年和 1911 年发表于法国，这在第二章中已作过简单描述。回忆一下，在美国出现的早期比纳测验的若干英译本和改编本中，最负盛名的是斯坦福一比纳量表。<sup>[3]</sup>由斯坦福大学的推孟及其同事



编制的比纳—西蒙量表的第一个斯坦福修订本，出版于 1916 年。该修订本作了许多修改和补充，实际上是一个新测验。1/3 以上是新项目，一些原有项目作了修改，重新划归到不同的年龄水平或被删除。根据大约 1000 名儿童和 400 名成人的美国样本，整个量表重新标准化。编写了施测和评分的详细指导语，在心理测验中首次采用 IQ。1937 年出版的第 2 版斯坦福修订本，由两个相等的型式 L 型和 M 型组成（Terman & Merrill, 1937）。在这个修订本中，量表的样本再次扩充，根据美国人口的新样本完全重新标准化。尽管作了认真努力以便获得有代表性人口的样本，但是，3184 名被试样本的社会经济水平略高于美国总体状况，它包括过多的城市儿童，而且只包括本地出生的白人。

1960 年出版的第 3 版修订本，从 1937 年两种型式中选出最佳的项目合并成单一的型式 L-M（Terman & Merrill, 1960）。在准备 1960 年版斯坦福—比纳量表时，作者面临心理测验的两难困境。一方面，经常需要修订本，以便得益于测验编制的技术发展和先前使用测验的经验，以及被更新的测验内容。最后一点考虑对于常识项目和图片材料项目尤为重要，家庭用具、车辆以及其他常见物品式样的变化，都会影响到这类项目。使用过时的测验内容可能严重损害和被试的友好关系，并可能改变项目的难度水平。另一方面，修订可能使许多积累的数据不适用于新的型式。广泛使用多年的测验获得了一大批解释材料，在需要修订时应该认真考虑到这点。出于这一原因，斯坦福—比纳量表的作者想把较早的两种型式浓缩为一种，因此选择了在过时和不连续性这两种缺陷之间的中间道路。这样，虽然丧失了平行型式，但不至于为达到目的而付出太大的代价。到了 1960 年，已不像在 1937 年那样需要替代型式，那时得不到其他编制良好的个别智力量表。1960 年版斯坦福—比纳量表没有对常

模量表重新标准化。采用新样本,只是为了找出项目难度经过一段时间之后所发生的变化。结果,1960年L-M型式上的智龄和IQ仍然以1937年标准化样本来表示。

下一个阶段是1972年对L-M型式的重新标准化(Terman & Merrill, 1973, Pt. 4)。这一次,测验内容实际上没有改变,但是,常模是从1971~1972学年期间所测的大约2100名被试的新样本中导出的。和1937年常模相比较,1972年常模基于更有代表性的样本,并且是最新的,因此反映了一段时间的文化变化对测验成绩的影响。令人感兴趣的是,后一种常模表明在所有年龄中测验成绩的某种提高,尤其在学前年龄阶段提高很多,IQ平均提高大约10分。测验作者把这种提高归因于其他文化变化中的大众传媒对幼儿的影响,以及父母的文化教育水平的提高。在15岁及以上,测验成绩的提高虽较少,但是显而易见,这点如同作者所提出的,可能与70年代继续接受中学教育的学生的比例大于30年代有关。根据横向和纵向两种比较,R.L.桑代克(1977)进一步探究了这些标准化样本的变化,还提出了其他起作用的因素,包括为了促进学前儿童智力发展的特别电视节目的采用。

在其他一般智力水平测验中,也发现从30年代或40年代到70年代测验常模的提高(Flynn, 1984, 1987)。从测验使用者的观点来看,这种发现的一种重要含义是,先后用较早的型式和较近的型式所测验的个体或团体,会出现能力下降,因为他们的成绩根据后来型式中的较高标准来评价。测验使用者应该意识到解释分数中这种可能的假象。

**第4版斯坦福-比纳量表(SB-IV)总述** 这个编制良好的量表的当前版本是经过最大规模修订后的产物(Delaney & Hopkins, 1987; Thorndike, Hagen & Sattler, 1986a, 1986b)。作为个

别施测的临床工具,该修订本既保留了早期版本的主要优点,又反映了智力机能的理论概念和测验编制的方法学两方面的发展。通过保留早期型式中的许多项目类型,使得该修订本部分保持了与早期版本的连续性。更为重要的是保留了适应性测验方法,每一个体只需做难度适合于其表现出的作业水平的那些项目即可。

同时,内容范围大大拓宽,不像早期型式中大多以言语为主,使得测验更有代表性,包括数量作业、空间作业和短时记忆作业。而且在尽可能大的年龄范围内采用每种项目类型,由此能够进行跨年龄水平的、更精密的比较评定。当前量表适用的年龄为 2 岁到成人。

**施测和评分** 施测斯坦福一比纳量表所需要的标准材料如图 8-1 中所示。主要包括 4 本印有卡片的小册子,用于翻页呈现测验项目 各种测验物体,例如积木、形状板、多种颜色和形状不同的小珠子,以及一张画着一个同一性别不同种族的娃娃的大图片;一本记录小册子,主试用来记录被试反应及其评分;一本关于量表的施测和评分的指南。

与大多数个别智力测验相同,斯坦福一比纳量表需要训练有素的主试,施测、评分和解释结果都需要特殊的训练和经验。顺利的施测需要实践和对量表相当的熟悉。主试的犹豫和笨拙可能损害与被试的和睦关系,尤其对于年幼的测验参加者。用词方面微小的、漫不经心的变化可能改变项目的难度。另一个复杂情况是,施测项目时,必须对项目评分,因为随后的测验进程取决于个体在先前所施测的项目上的成绩。

几十年来,临床心理学家已经把斯坦福一比纳量表和类似的个别量表不仅作为标准化的测验,而且作为临床面谈的工具。这个特点使这些量表难以施测,但也创造了主试和被试之间交

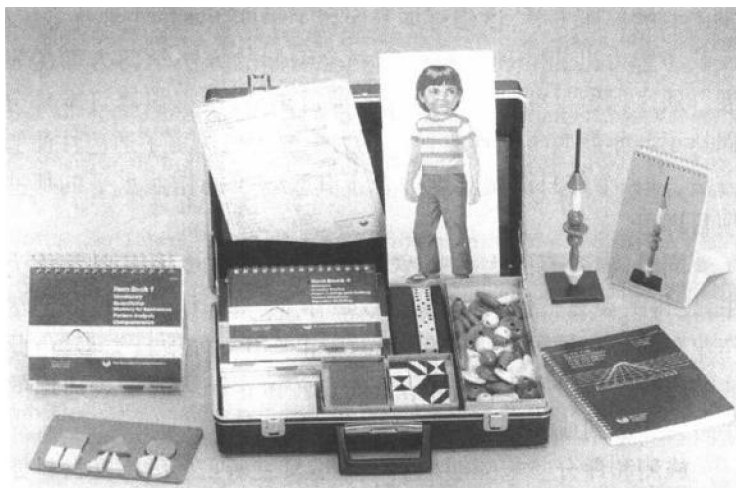


图 8-1 施测第4版斯坦福—比纳量表所用材料

谈的机会，并且为有经验的临床心理学家提供其他的线索。斯坦福—比纳量表和本章中所讨论的其他测验，使主试有可能观察被试的工作方法、解决问题的方法以及作业的其他性质。主试也有机会去判断某种情绪特性和动机特性，例如集中精力的能力、活动水平、自信、持久性等。当然，在施测个别量表的过程中所作的任何定性观察，应该被明确地认为是定性观察，而不应以客观测验分数的相同方式来加以解释。这种定性观察的价值，主要取决于主试的技能、经验和心理学知识，同时还取决于主试对这类观察固有的危险和局限性的意识。

同早期版本以年龄分组不同，第4版斯坦福—比纳量表表现在每种类型的项目都按难度升序放置在各个测验中。有15个

测验，用来代表 4 个主要的认知领域 言语推理、抽象／视觉推理、数量推理和短时记忆（参见图 8-2）。虽然出于评分目的，15 个测验在这 4 个标题下分组，但它们是以混合的顺序来施测的，以保持测验参加者的兴趣和注意。6 个测验的难度范围跨

| 年龄             | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18+ |
|----------------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|-----|
| <b>言语推理</b>    |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 词汇 .....       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 理解 .....       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 挑错 .....       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 言语关系 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| <b>数量推理</b>    |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 数量 .....       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 数字系列 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 列出等式 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| <b>抽象/空间推理</b> |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 图型分析 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 临摹 .....       |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 矩阵推理 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 折纸和剪纸 .....    |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| <b>短时记忆</b>    |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 小珠记忆 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 句子记忆 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 数字记忆 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |
| 物体记忆 .....     |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |     |

图 8-2 第4版斯坦福-比纳量表 15 个测验的年龄跨度。注意白色区域：对于年龄范围有限制的 9 个测验，因为例行测验成绩不同寻常的高或低，标准化样本中处于这些年龄之外的一些成员，也参加了测验。他们的成绩被用于估计常模表中整个年龄样本的成绩，但对这些估计值应特别谨慎。详情参见《指南》（桑代克等，1986a, p.7）和《技术手册》（桑代克等，1986b, p.30）（摘自第 4 版斯坦福-比纳量表，施测和评分指南, p.7）。

越整个量表的年龄范围,由于任务的性质,其余 9 个测验或开始于较高水平,或结束于较低水平,这可以在图 8-2 中看到。

第 4 版斯坦福一比纳量表的施测过程为两个阶段。在第一阶段,主试给出词汇测验,这是一个例行测验,用来选择所有其余测验的进入水平(entry level)。在词汇测验上从哪个项目开始,只取决于个体的年龄。对于所有其他测验,从词汇测验和年龄的结合图上,可以找到进入水平。在测验的第二阶段,主试必须按照个体的实际成绩,确定每个测验的基础水平(basal level)和上限水平(ceiling level)。在两个连续的水平上,通过 4 个项目,这就达到基础水平。如果在进入水平上没有发生这种情况,那么测验就向下级水平继续进行,直到达到基础水平为止。在两个连续的水平上,4 个项目中的 3 个项目或全部 4 个项目均没有通过,这就达到上限水平。对于个体来说,那个特定的测验就不再继续进行。

每个项目施测时,主试在记录小册子上记录分数。记入所施测的最高项目的项目号,从中减去尝试的项目中没有通过的项目总数,这就得出每个测验的原始分数。11 个测验还有例题,这只是为了熟悉测验,不记分数。对于大多数测验,每个项目只有一个正确答案,主试在项目书和记录小册子上可以找到。根据规定的标准,各个项目或通过或没有通过。5 个测验要求自由反应,因此需要使用指南中所提供的扩展评分指导。<sup>[4]</sup>一些模糊反应需要主试进一步询问,在评分指南中也有说明。

虽然整个量表包括 15 个测验,但没有一个个体参加所有这些测验,因为一些测验仅仅适用于有限的年龄范围。一般来说,完整的成套测验包括 8~13 个测验,这取决于测验参加者的年龄及其在例行测验上的成绩。整套测验的施测时间假设为 30~90 分钟,但经验不足的主试可能费时更多一些。量表通常可

以一次测完,在各测验之间可以有几分钟的休息。为适用于某种目的,指南中提出了某些简化的成套测验,这种成套测验费时较少,但集中于最适用于特定测验目的的那些测验。简化成套测验包括由 6 个测验组成的一般目的成套测验和由 4 个测验组成的快速筛选成套测验,这两个成套测验在 4 个认知领域的每个领域中,至少包含 1 个测验。还提出评定天才学生的 3 个成套测验,分别适用于 3 个年龄水平。另有 3 个成套测验用于学校中有学习问题的学生,也分别适用于 3 个年龄水平。在所有这些简化成套测验中,都提出了进入水平以及施测和评分的标准程序。《主试手册》(Delaney & Hopkins, 1987)说明了有关对于不同类型的被试施测第 4 版斯坦福—比纳量表和评分的许多程序问题。

**标准化和常模** 标准化样本的组成为 5000 余名 2~23 岁的个体,测验遍及 47 个州(包括阿拉斯加和夏威夷)和哥伦比亚特区,严格按照 1980 年美国人口普查中地理区域、社区大小、种族、性别的比例进行分层取样。另外,按照父母的职业水平和教育水平来检查社会经济地位。结果表明,在较高水平上代表过多,在较低水平上代表不足。计算常模表中分值时,通过频数的不同权重,对这种差异进行调整。因此,来自较高社会经济背景的每一个体计数为不足一个案,而来自较低社会经济背景的每一个体计数为一个案以上。

使用常模表,把 15 个测验的每个测验上的原始分数,转换为标准年龄分数(SAS)。<sup>[5]</sup>这是常态化标准分数,在每一年龄组之内,平均数为 50,标准差为 8。其中,2~5 岁,常模表的年龄间隔为 4 个月;6~10 岁,常模表的年龄间隔为 6 个月;11~17 岁,常模表的年龄间隔为 1 岁;18~23 岁,则为单一的常模表。记录小册子提供图形,可以绘制测验参加者在所施测的每个测验

上的 SAS 成绩的剖面图。

对于 4 个认知领域的每个领域,以及对于整个量表的综合成绩,都可以得出标准年龄分数。根据施测于每一个体的测验的标准年龄分数,参照合适的常模表,可以得出领域的标准年龄分数和综合的标准年龄分数。这些标准年龄分数也是常态化标准分数,但平均数为 100,标准差为 16。因此,表示标准年龄分数的单位,同较早版本的斯坦福一比纳量表上的离差智商相同。然而,术语“IQ”的使用现在已被完全放弃。为适用于特殊目的,主试可以得出任何想要的两个或两个以上领域结合(“部分综合”)的标准年龄分数。例如,言语和数量推理的结合大致相当于“学术能力倾向”;特别与评定学业成就或学习准备有关系。

**信度** 由于第 4 版斯坦福一比纳量表没有复本,只能通过内在一致性或再测方法来评价信度。大多数分析使用库德—理查逊技术,这一技术适用于从整个标准化样本中得出的数据。可以预期,在整个成套测验上的综合分数在所有年龄水平都将产生最高信度,范围从 0.95 至 0.99。每个认知领域的信度也是高的,虽然这种信度随着每个领域所包括的测验数目而变化,其范围为 0.80 ~ 0.97。对于各个测验,大多数信度处于 0.80 ~ 0.90,除了物体记忆之外,这是一个简短的 14 个项目的测验,它的信度范围从 0.66 至 0.78。一般说来,在较大年龄水平的所有效度往往稍高于较小年龄水平。

间隔 2 ~ 8 个月之后,对 57 名 5 岁学前儿童和 55 名 8 岁学龄儿童进行再测,得出再测信度上的补充数据。一般说来,综合分数的稳定性是好的,两个样组的信度系数分别为 0.91 和 0.90。言语推理领域分数的信度为 0.80,而其他领域分数和各个测验的波动则相当大。这类结果难以解释,由于有限的范围可能对一些测验有影响,以及明显的练习效果,因此,个体之间



大不相同。

除了信度系数之外,‘指南’和‘技术手册’给出每个测验、领域分数和综合分数在每个年龄水平之内的测量标准误差(SEM)。评价个体分数,以及在剖面分析中解释分数间差异,都需要这类测量标准误差。总的综合标准年龄分数(平均数为100,标准差为16)的测量标准误差约为2~3分。例如,如果我们取近似的中间值2.5,这意味着有1/2的可能,个体的‘真’综合分数与其实得分相差2.5分或以下有95%的可能;‘真’综合分数的变化不超过5分( $2.5 \times 1.96 = 4.90$ )。

斯坦福一比纳量表的《主试手册》(Delaney & Hopkins, 1987)提出一种解释框架,试图形成和交叉证实从成套测验所收集的定量数据和定性数据中得出的假设。定量分析所采用的模型,由戴维斯(F. B. Davis, 1959)最先提出,考夫曼(Kaufman, 1979, 1994)及其他人应用于韦克斯勒量表。实质上它在综合分数和4个领域分数(参见图8-2)之间作比较,按照测量标准误差来评价差异的统计显著性,也对实得差异的频数与标准化样本的基础率数据作比较。另外,还对个体在综合和领域分数上的平均成绩与各个测验上的分数作比较,从而可以系统地评定个体内部在由每个测验所要求的各种能力上的长处和不足。手册提供了进行这些剖面分析所需要的信息,以及4个完整的应用实例,这对于新任主试和有经验的主试,应该都有很大价值。

效度 和测验效度的当前概念相一致,第4版斯坦福一比纳量表的编制采用多种方法来确定和界定所包括的合适结构。对所能得到的关于智力的性质和测量的研究文献的考虑,指导了最初对结构的选择(R. L. Thorndike 等, 1986b, chap. 1)。关于斯坦福—比纳量表本身的早期版本及其出现的长处和不足的经验的,进一步指导了测验编制的计划和决定。例如,把项目类型分

成可靠的分测验，替代传统的临床做法即通过主观的项目分组进行非正式的模式分析。

对第 4 版斯坦福—比纳量表中所要评定的结构进行最初选取和初步定义之后，对项目进行鉴定，同时发展新的项目来适应新的结构定义。对项目库进行大量的、使用高级统计程序的项目分析，包括对项目偏向的判断评定和统计评定(R. L. Thorndike 等, 1986b, 第二章)。从若干初步试验和现场试验中得出量表的最后版本, 对标准化样本进行施测, 又进一步研究了 3 种主要类型的效度数据: (1) 分数的相关和分数的因素分析; (2) 和其他智力测验的相关; (3) 先前所鉴定的各特殊团体的成绩比较 (Thorndike 等, 1986b 第六章)。

首先, 对于整个标准化样本, 在每一年龄水平之内, 计算所有的测验分数、领域分数及综合分数之间的相关。在证实性因素分析中使用所有年龄相关的中数。这种分析的主要目的是检验假设: 一种一般因素可解释不同领域的测验之间的相关, 而各种群因素则解释每一领域之内的残余相关。在 3 个年龄组(2~6 岁、7~11 岁、12 至 18~23 岁)内, 都用相关的中数来进行类似的因素分析。

所有因素分析的结果表明, 所有测验在一般因素上负荷都很大, 由此证明使用一个总的综合分数是正确的。对于 4 个领域中的 3 个领域, 群因素解释领域之内相当比例的残余共同方差。抽象/视觉推理领域是例外, 4 个测验之间得出高度的特异性。我们可以推测, 在这个领域没有得出群因素的明显证据, 可能与学校课程的累积效果有关, 关于言语内容和数量内容都有高度组织, 而关于空间知觉内容则没有。有助于发展空间知觉能力的个体日常经验, 没有像学业经验那样被系统地组织为“课程”或内容领域。因此, 它们不大可能建立跨个体的共同的关系

模式(Anastasi, 1970, 1986b)。

对测验手册中所报告的因素分析结果的考察,以及使用第4版斯坦福—比纳量表标准化数据的其他研究人员所独立进行的考察,都支持用一个综合分数作为对一般智能的度量(R. M. Thorndike, 1990)。然而,关于范围较小因素的性质和数量,研究人员意见不同(也参见 McCallum, 1990)。由于第4版斯坦福—比纳量表在不同年龄由不同的测验组所构成,因素分析的原始数据(即测验分数之间的相关)也因此而不同,这种事实使得情况更为复杂。这就造成在不同的年龄水平出现的因素的类型和数量不同——在2和4之间。在不同的研究中应用多种因素分析方法,解决了这种混乱。然而,一般说来,第4版斯坦福—比纳量表编制中所假设的四因素模型,更接近于日益增长的年龄的实际情况,当使用证实性因素分析而非探索性因素分析时,情况尤为如此。

效度数据的第二个来源系一组研究,这组研究所使用的团体参加过第4版斯坦福—比纳量表和其他一项智力测验,包括斯坦福—比纳量表本身的L-M型式。<sup>[6]</sup>团体包括由学校称为“非特殊的”、编入常规班级的学生,以及学习特别课程的三种“特殊”团体即天才学生、学习障碍学生、智力落后学生。在非特殊样本中,早期斯坦福—比纳量表(L-M型)上的离差智商与第4版上综合分数的相关为0.81,与言语推理领域的相关为次高(0.76),与抽象/视觉推理领域的相关为最低(0.56),根据两个版本的内容的相似和差异,可以预期到这一点。对于所有团体,第4版斯坦福—比纳量表的综合分数和领域分数与其他智力测验上的总分或部分总分的相关,一般来说符合所检验的关于结构的假设,同时,对第4版上特定分数与其他智力测验上特定分数之间所得出的所有相关进行考察,这有助于进一步理解

斯坦福一比纳量表现在所测量的结构。

使用特殊样本的第三组专门研究表明,第4版斯坦福一比纳量表能够合适地鉴定天才的、学习障碍的、智力落后的等测验参加者的成绩水平。天才者样本在综合分数和4个领域分数中每个领域分数上获得的平均数显著高于标准化样本。学习障碍学生和智力落后学生的综合分数和领域分数的平均数都显著低于标准化样本,智力落后被试的平均数显著低于学习障碍被试。应当指出,在关于特殊团体的所有研究中,参加者先前是根据测验或根据其他的成绩指标来鉴定的,但第4版斯坦福一比纳量表本身决没有被包括在这类鉴定中。

后来通过对第4版斯坦福一比纳量表效度研究的评论(Laurent, Swerdlik & Ryburn, 1992)得出结论,看来该量表至少和当前可得到的其他量表一样可以准确地测量一般智能。该量表与成就度量高相关,该量表能够区别智力落后个体、天才个体和神经损伤的个体。评论者认为第4版斯坦福一比纳量表是评价天才儿童的合适工具,因为测验的年龄范围提供了很高的上限;另一方面,他们认为,测验缺乏下限项目,这种项目难度很低,足可诊断年龄最小的测验参加者中的智力落后。

为了进一步解释第4版斯坦福一比纳量表上种种测验分数和分数组合的意义,我们正在做有关的研究,并已有可喜的成果。另外,从萨特(Satter, 1988)、格卢汀和卡普兰( Glutting & Kaplan, 1990)、坎普豪斯(Kamphaus, 1993)等处可得到关于该量表使用的全面指导。斯坦福一比纳量表现在的版本,在量表的编制方面取得真正的进展。第4版让主试可以灵活机动地评定适合于具体测验目的的各种能力。该量表与当前关于智力性质理论和研究的结果也相一致(参见第十一章)。

## 韦克斯勒量表

戴维·韦克斯勒编制的智力量表包括 3 个量表的若干连续的版本,一个为成人量表,一个为学龄儿童量表,一个为学前儿童量表。韦氏量表除了用于测量一般智力之外,有人还把它作为精神病诊断中潜在的辅助工具。韦克斯勒和其他临床心理学家最初观察到,大脑损伤、精神衰退和情绪困难对某些智力机能的影响要大于其他机能。他们认为,分析个体在不同分测验上的相对成绩,可以揭示具体的精神障碍。第十七章中分析了对韦克斯勒量表进行这类剖面分析中的问题和结果,这是临床使用测验的一个实例。

由韦克斯勒量表所激起的兴趣,以及量表使用的范围,由数千种关于该量表的出版物所证实。除了《心理测量年鉴》中通常的测验评述之外,关于韦克斯勒量表的研究,各种杂志定期地进行评述(Guertin, Frank & Rabin, 1956; Guertin, Ladd, Frank, Rabin, & Hiester, 1966; Guertin, Ladd, Frank, Rabin & Hiester, 1971; Guertin, Rabin, Frank & Ladd, 1962; T. D. Hill, Reddon & Jackson, 1985; Littell, 1960; Rabin & Guertin, 1951; I. L. Zimmerman & Woo—Sam, 1972), 好几本专著也作了概括(例如, Forster & Matarazzo, 1990; Gyurke, 1991; Kamphaus, 1993; Kaufman, 1979, 1990, 1994; Sattler, 1988, 1992)。

**韦克斯勒智力量表的前身和发展** 韦克斯勒量表的最初形式,称为韦克斯勒—贝勒维智力量表(Wechsler-Bellevue Intelligence Scale), 1939 年出版。编制量表的主要目的之一是提供适合于成人的智力测验。首次提出这个量表时,韦克斯勒(1939)

指出,先前可得到的智力测验主要是为学龄儿童设计的,加上一些相同类型的、较难的项目后,改为成人使用。成人对这类测验的内容往往没有什么兴趣,除非测验项目具有某种最低的表面效度,否则难以适当地与成人测验参加者建立和睦关系。许多智力测验项目,主要涉及学龄儿童日常活动,对于大多数成人,显然缺乏表面效度。

大多数测验中过度强调速度,这也往往不利于年龄较大的个体。同样,韦克斯勒认为,相对而言,言语的常规使用在传统的智力测验中的权重不当。他同样注意到智龄常模不能适用于成人,并指出,个别智力测验的标准化样本中,先前几乎没有包括成人。

正是为了符合这种种目的,编制了最初的韦克斯勒-贝勒维智力量表。在形式和内容上,该量表为后来所有的韦克斯勒智力量表建立了一种基本的型式,后来的每个量表依次对原有的量表作精心的改进。1949年编制韦克斯勒儿童智力量表(Wechsler Intelligence Scale for Children, WISC),是韦克斯勒-贝勒维量表向小年龄的延伸(Seashore, Wesman & Doppelt, 1950)。许多项目直接取自成人测验,将同样类型的较易的项目加到每个分测验上。1955年韦克斯勒-贝勒维量表本身被韦克斯勒成人智力量表(Wechsler Adult Intelligence Scale, WAIS)所取代,该量表改进了早期量表在常模样本的大小和代表性及分测验的信度方面的技术缺陷。1967年出版韦克斯勒系列中适用年龄最小的量表,即韦克斯勒学龄前及小学生智力量表(Wechsler Preschool and Primary Scale of Intelligence, WPPSI),最初的适用年龄为4~6岁半,这是适用年龄为5~15岁的WISC向小年龄的延伸。

WISC的编制有点自相矛盾,因为韦克斯勒着手其最初事业

的部分原因是需要一个成人量表,但它又不仅仅是可得到的儿童量表向大年龄的延伸。第一版 WISC 事实上遭到批评,由于它的内容并非完全儿童化。修订版(WISC - R)出版于 1974 年,年龄范围为 6~16 岁,特别致力于替换或修改成人化的项目,以便使内容较接近儿童期的共同经验。例如,在算术分测验中,“雪茄烟”改为“棒糖”。其他改变包括删除对于各种特定儿童团体熟悉程度不同的项目,以及在分测验的图片内容中使用较多女性和黑人。为了增加信度,增加了若干分测验的长度。对施测和评分程序也作了改进。

**量表的描述** 到现在,3 个韦克斯勒量表中的每一个都经过了一次或多次修订。当前的各种版本,以戴维·韦克斯勒的姓名来出版,甚至在他 1981 年逝世以后也是如此,它们是韦克斯勒成人智力量表—修订版(WAIS - R——Wechsler, 1981),适用的年龄范围为 16~74 岁;韦克斯勒儿童智力量表—第 3 版(WISC - III——Wechsler, 1991),适用的年龄范围为 6~16 岁 11 个月的儿童;韦克斯勒学龄前及小学生智力量表—修订版(WPPSI - R——Wechsler, 1988),现在的年龄适用范围为 3~7 岁 3 个月。自 1992 年以来一直在修订中的第 3 版 WAIS,预期 1997 年出版。

WAIS - R、WISC - III 和 WPPSI - R 具有许多共同特点,它们的基本结构分为言语量表和操作量表,每个量表最少由 5 个分测验(最多由 7 个)组成,分别得出离差智商。个体在 10 个常规施测的分测验上(WAIS - R 为 11 个)的分数合并为一个全量表智商,其平均数为 100,标准差为 15,言语智商和操作智商的平均数和标准差也是如此。WAIS - R、WISC - III 和 WPPSI - R 共使用 17 个不同类型的分测验,其中 8 个分测验(5 个言语分测验和 3 个操作分测验)对于 3 个量表都是共同的。施测 3 个量表

时,言语分测验和操作分测验交替进行,进行的顺序均为事先确定,每个量表有所不同。

在 3 个量表中,所施测的第一个言语分测验都是常识分测验,尽力避免专业化的知识,以便与被试建立一种和睦的关系。第一个项目都很容易,绝大多数被试都可通过,除非他们智力落后或有现实取向问题。在这种情况下,主试可以马上决定不再测验。WAIS-R 和 WISC-III 版本的常识分测验中的问题,包括美国大多数个体都有机会学习的基本知识,例如,“12 月之前是哪一个月?”或“马克·吐温是谁?”WPPSI-R 中也有类似的问题,只是难度水平更低一些。实际上,头几个项目以图片形式呈现,仅仅要求指出正确反应是什么。例如,呈现一张画着几个家用物品的图片,问儿童哪一个用于清洁工作。算术分测验是另一个言语测验,在韦克斯勒各量表中,项目难度的分布范围很大。最简单的 WPPSI-R 算术项目,要求在图片卡上一组物体中,指出一个表明一种数量概念的物体(例如,“最少”或“较多”)。较复杂的项目包括数数或解决算术问题,其中最难的项目要求完全理解分数。

韦克斯勒量表的操作分测验一般要求操作各种物体,例如拼板和积木,或者阅览所印制的材料,例如图片或符号。操作分测验对测验参加者都规定时限,在大多数情况下,给被试加速度奖励分。与之相对照,在言语量表中,只有一个分测验(算术)是速度测验。图画补缺是 3 个韦克斯勒量表中都有的一个操作分测验;它要求被试从画着常见的物体或情境的图片中,找出哪个重要的部分缺少了。用于小年龄的项目要求基本的视觉检查——例如,呈现的图片上,动物少了一条腿。较难的项目还要求某种演绎推理、具体知识,或兼而有之,以便找出所缺少的东西。

图 8-3 说明两个相对容易的图画补缺项目,它们类似于韦克斯



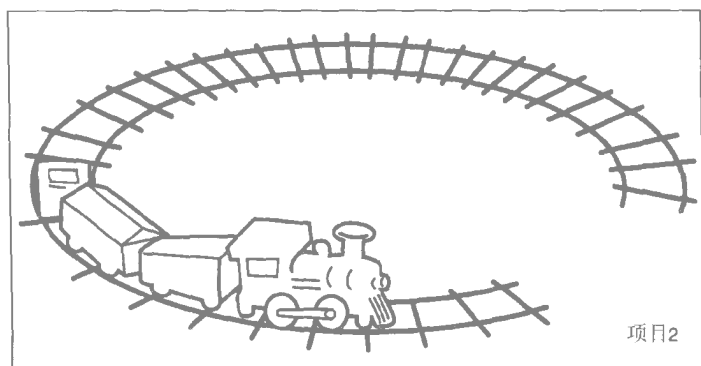
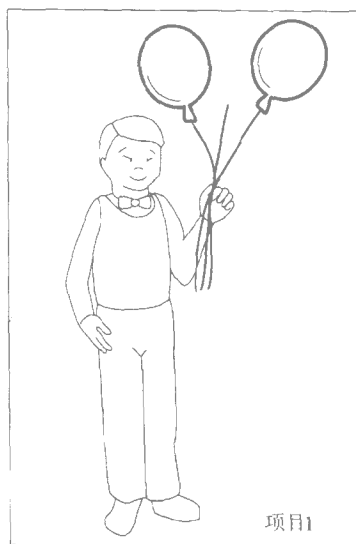


图 8-3 两个图画补缺项目，它们类似于韦克斯勒学龄前和小  
学生智力量表修订版所用的项目。

勒量表中所使用的项目。

**简化量表** 自从最初的韦克斯勒-贝勒维量表发表以来,提出了众多韦克斯勒量表的简化量表或简式。缩短这些量表的目的是,大量减少施测的时间,以获得一个估计的全量表智商,并能根据所发表的常模来进行评价。编制这种简式的一种方法是,仅仅省去一些分测验而按比例分配分数。减少分测验之内的项目数,也可以形成简化量表。

几种分测验的组合与全量表智商的相关为 0.90 以上,这鼓励人们出于快速筛选目的而编制和使用简化量表。人们进行广泛的研究,以便在预测言语智商、操作智商和全量表智商方面,找出 2、3、4、5 个分测验的最佳组合 (Matarazzo, 1972; McCusker, 1994; Sattler, 1988, 1992)。其中的许多研究使用标准化数据,但一些研究根据特殊团体例如精神病患者和智力落后个体来进行。

在汇编和检验韦克斯勒量表的简式中,看来花费了过多的精力。然而,有人对在得出简化量表中所使用的实际程序的质量如何提出了问题 (Silverstein, 1990)。例如,最初的全量表常模适用于简式量表上按比例分配的总分,这一假设并非总是正确的。而且,在施测个别量表中可能进行的许多重要的定性观察,一旦使用简化量表则无法进行。因此,使用这类简化版本也许是不可取的,而作为粗略的筛选工具,当属例外。

**常模和评分** 最新的韦克斯勒量表的标准化样本是精心选取的,以保证他们的代表性。每个量表的常模样本约为 2000 名,男女人数相等,分布在各个合适的年龄组中。被试的选择尽可能符合进行标准化时美国人口调查得出的最新报告,考虑的变量为地理区域、种族、职业水平和教育水平。至于儿童被试,则使用父母的职业水平。对于后来的每个版本,在分层取样

划中所使用的变量都稍有改变,以便使样本更具有包容性。例如 WISC-III 种族类型包括 4 个团体(白人、黑人、西班牙人、其他),而较早的 WISC-R 仅使用两类(白人或非白人)。而且,与较早的量表不同,WISC-III 样本明确包括一定数量、在学校环境中接受特殊学习的学生,例如学习不能儿童和接受天才课程的儿童。

韦克斯勒量表是当前应用最为广泛的个体智力测验,从而引发了一些试图拓宽其用途的研究。例如,作为对梅约(Mayo)诊疗院老年美国人的系列常模研究之一,对 56~97 岁的 222 名个体的一个样本进行施测,以便提供超出标准化样本中最大年龄组的 WAIS-R 常模数据(Ivink 等,1992)。另一项研究则根据 130 名 75 岁以上的个体得出 WAIS-R 的常模(Ryan, Paolo & Brungardt, 1990)。

韦克斯勒量表的每个分测验上的原始分数,都被转换成平均数为 10、标准差为 3 的标准分数。所有分测验所评定的分数因而可以作比较的单位来表示。然后,把适当的言语分测验分数、操作分测验分数及全量表分测验分数分别相加,并转换成平均数为 100、标准差为 15 的标准分数,称之为‘离差智商’(deviation IQ)。WISC-III 也提供另外 4 个基于因素的指标分数,即言语理解(VCI)、知觉组织(POI)、不分心(FDI)、加工速度(PSI)。这些指标分数的构成,十分类似于由许多独立的研究人员从较早的 WISC-R 中推导出的典型的因素构成。新的指标分数则根据 2 个或 4 个分测验的组合,其平均数为 100、标准差为 15。3 个韦克斯勒量表均提供了按照适当的年龄常模评价个体在每个分测验和一组分测验上成绩所需要的资料。

信度 韦克斯勒量表提供了每个年龄组在每个分测验分数<sup>[7]</sup>、指标分数及智商的分半信度系数。对于所有的韦克斯勒

量表,全量表智商的分半信度范围从 0.90 至 0.98,言语智商的分半信度系数范围从 0.86 至 0.97,操作智商的分半信度系数范围从 0.85 至 0.94。WISC-III 基于因素的指标分数的分半信度系数范围从 0.80 至 0.95。可以推测,分测验的信度较低。3 个韦克斯勒量表的分测验的信度系数范围从 0.52 至 0.96,其中绝大多数在 0.70 以上。在评价同一个体在分测验分数之间的差异显著性时,如在剖面分析中,考虑分测验信度就尤为重要(J. H. Kramer, 1990, 1993; Sattler, 1988, 1992)。韦克斯勒量表手册也报告所有韦克斯勒分数的测验标准误差。言语智商的测量标准误差范围从 2.50 分至 4.98 分,操作智商的测量标准误差范围从 3.67 分至 4.97 分,全量表智商的测量标准误差都低于 4.00 分。由此我们可以得出结论,例如,大致有 2:1 的可能性,个体的真正的全量表智商处于其实得的全量表智商的 4 分之内。

每一版本的韦克斯勒量表分数的再测信度(即稳定性)的数据,均被详尽收集。成人的稳定性系数往往高于儿童。再测研究总是表明,经过 12 天至 9 周的间隔,从首测至再测,各种智商都增加 2~13 分,全量表智商一般增加 5~7 分。这种所预期的练习效果虽然不大,但在短时间之后再测个体时,也应该考虑到。

在随时间不断改进的系列中, WPPSI-R 手册和 WISC-III 手册都是最新的。值得注意的一个特点是,对于在评分中要求作出大量判断的分测验给出的评分者信度数据。这种数据表明,由经过训练的、具有足够经验的主试来对这些分测验进行评分,其信度是可接受的。这两个量表的手册中另一项革新是,关于个体内分数差异,提供了相当广泛的数据。手册的附表表明达到统计显著性所需要的差异量,另外,手册也列出标准化样本内得出的差异频数。在临床使用这些测验时,这类资料尤有价

值(参见第十七章)。

**效度** 在连续的几本韦克斯勒手册中所作的改进,最多的当属效度领域。1981 年发表 WAIS-R,手册本身没有关于量表的效度数据,只有该量表与较早韦克斯勒量表的两个相关研究。该量表的效度资料甚至不及 WPPSI 和 WISC-R,后者至少还有本身量表与其他非韦克斯勒智力量表例如斯坦福一比纳量表的相关资料。然而,可得到的关于各种韦克斯勒量表效度的独立研究的结果<sup>[8]</sup>,多少可以补偿以往手册有限的效度资料。韦克斯勒手册中不强调效度资料,部分是由于韦克斯勒深信,其量表中作业包含多种具体能力,足可保证有效地评定一般智力(Wechsler,1958;Zachary,1990)。

韦克斯勒对效度的处理实质上反映一种内容描述倾向,虽然它具有一种结构鉴定的方法的含义,却没有什么证实性的数据。例如,韦克斯勒量表与其他的智力总体度量例如斯坦福一比纳量表的相关均为 0.80 左右。另外,多年来独立的研究人员所进行的因素分析研究,也得出相当一致的结果。在所有年龄组中,几乎总是得出一个言语因素和一个知觉组织或非言语因素。一般来说,在大龄儿童量表和成人量表中也出现另外的记忆因素和/或注意因素。各种团体在所预期的方向上显示出差异,这在某种程度上证明在职业的和教育的选拔决策中使用韦克斯勒量表是正确的。

相比之下,关于最新韦克斯勒量表的各种类型的效度数据则是十分丰富的。WPPSI-R 手册和 WISC-III 手册分别用 21 页和 38 页篇幅来讨论效度,而 WAIS-R 手册中关于效度仅有 2 页。分测验之间的相关和对分数的因素分析,都提供了结构关联效度数据。标准化样本的各个年龄组的言语量表和操作量表之间的平均相关,WAIS-R 为 0.74,WISC-III 为 0.66,WPPSI-

R 为 0.59。这些统计数据表明,存在一种广泛的一般因素,对各量表的大多数因素分析也证实这点。

对 WAIS-R 标准化样本的 9 个年龄组的因素分析研究,屡次表明三因素模型最好地解释了 11 个分测验之间的相关。这些因素为言语理解、知觉组织、记忆/不分心等,它们似乎能概括不同类型的样本(Leckliter 等,1986;Waller & Waldman,1990)。手册和其他论著中所论述的、用 WPPSI-R 标准化数据所进行的同类分析则提出二因素模型,这与分测验的结构分为言语量表和操作量表相一致(Blaha & Wallbrown,1991;LoBello & Gulgoz,1991;B.J.Stone,Gridley & Gyurke,1991)。另一方面,手册中所描述的一些探索分析和证实分析,从一开始就使用 WISC-III 数据,看来这种数据倒最适合于四因素模型,即言语理解、知觉组织、不分心、加工速度等。这 4 种因素已被编入 WISC-III 的标准评分。

WPPSI-R 手册和 WISC-III 册也提供了许多研究的效度数据,这些研究计算这两个量表与其他个别施测的测验的相关,虽然使用方便的小型样本。在 WISC-III 的手册中,还引用了该量表与团体施测的成就量表及与学校成绩的一些相关。此外,一系列研究使用特殊团体,例如天才儿童、智力落后儿童、学习无能儿童及其他类型儿童,研究了 WISC-III 和 WPPSI-R 的诊断功用和效标预测功用。

关于韦克斯勒量表的最后评论 3 个韦克斯勒量表的连续版本,反映了韦克斯勒量表数十年发展中测验编制的经验和水平的提高。与其他个别施测的测验相比较,韦克斯勒量表的主要长处在于标准化样本的规模和代表性,尤其对于成人和学龄前总体,外加测验编制程序的技术质量。WISC-III 手册中对信度和效度的处理尤其值得称道。目前,韦克斯勒量表的盛

行又使该量表的研究基础得到不断的拓宽。韦克斯勒量表的使用者也可得到许多辅助材料,例如计算机辅助解释程序、训练主试的手册(例如, Fantuzzo, Blakey & Gorsuch, 1989)、解释指南等(例如, Kaufman, 1994; Nicholson & Alcorn, 1994; Whitworth & Sutton, 1993)。然而,一些批评家注意到,按照目前对于评定工具和干预策略之间的联系的要求,即便是韦克斯勒量表的最新的、改进最多的版本,不久也可能过时(Shaw; Swerdlik & Laurent, 1993; Stenberg, 1993)。在这方面,所有韦克斯勒量表的最大缺点在于缺乏理论基础,这使得它难以找到一致的解释基础。另外,量表的组成似乎假设,各个年龄水平的分测验所要求的能力领域是相同的,因为测验材料和测验任务都具有表面上的相似性。鉴于我们现在所知道的关于一生中智力性质的发展变化的知识,这种假设可能是不正确的(参见第十一章)。

## 考夫曼量表

考夫曼量表是个别施测的临床工具,适用于多种功能,在这方面已经编制了诸如斯坦福一比纳量表和韦克斯勒氏量表之类测验,传统上也使用这类测验(Kaufman & Kaufman, 1983a, 1983b, 1990, 1993)。考夫曼量表编制于 20 世纪 80 年代和 90 年代早期,体现出测验编制中的最新进展。考夫曼儿童评定成套测验(Kaufman Assessment Battery for Children, K-ABC——Kaufman & Kaufman, 1983a, 1983b)以及考夫曼青少年和成人智力测验(Kaufman Adolescent & Adult Intelligence, KAIT——Kaufman & Kaufman, 1993)尤其体现了作者的意图——两位作者都参与编制 WISC-R——改变以往智力量表的非理论化观点。他们试

图编制一套量表,这套量表以发展中的智力理论为依据,又包括发展上适当的作业,并为多种评定情境提供有用的资料。

### 考夫曼儿童评定成套测验

性质和编制 K-ABC 的设计从形成要评定的结构着手。和认知心理学的总倾向相一致,重点放在信息加工上。他们选取特定的方法来区分同时加工 ( Simultaneous Processing )与系列加工 ( Sequential Processing ),前者由 7 个分测验所代表,后者则由 3 个分测验所代表 ( J. P. Das. 1984; J. P. Das, Kirby & Jarman, 1975 . 1979; J. P. Das & Molloy, 1975 ; Kaufman & Kaufman. 1983b ,第二章 ; Luria, 1966) 。同时加工分测验要求被试能够从总体上观察空间的和视知觉的内容,并对这种内容进行综合和组织。系列加工分测验要求被试进行系列的或时间的排列,他们使用言语的、数字的、视知觉的内容以及短时记忆。在联合的心理加工量表 ( Mental Processing Scale ) 中所出现的一些作业,类似于神经心理检查所使用的作业 ( 第十七章讨论这一点),选取这种作业也出于这个原因。

该成套测验还包括一个成就量表,由 6 个分测验组成。虽然成就分测验评定阅读、算术、词汇知识、一般常识等方面的能力,但设计这个分测验显然不是测量学校中所教授的实际知识。它们更像传统的能力倾向测验或智力测验中所包括的作业,而不像传统的教育成就测验中所包括的测验。例如,在算术测验中,儿童阅读一组某家庭参观动物园的图片,然后用每张图片中的物体来计数和进行简单的数字运算。在阅读理解中,儿童阅读句子,然后表演每句中所描述的动作。

K-ABC 根据 2000 名儿童 ( 年龄为 2 岁半 ~ 12 岁半 ) 的全国样本进行标准化。为了编制关于种族和父母教育的社会文化



常模,还对另外的黑人儿童和白人儿童的样组进行测验,这种常模可以作为补充的解释辅助工具。而且,K-ABC的设计和组织实施试图适用于测验特殊团体的需要,例如残疾儿童和不同文化和语言的少数民族儿童,以及帮助诊断学习不能儿童(Kaufman, Kaufman & Harrison, 1990)。成套测验得出4个综合分数系列加工、同时加工、心理加工组合(前两个加工相联合)、成就等分数。每个综合分数都是以平均数为100、标准差为15的标准分数。

一般评价 K-ABC 具有许多技术和实践上的积极特点。<sup>[9]</sup>一种流行倾向是用一个单一的评价数字,例如 IQ 来给儿童一个标记,有了 K-ABC,使用《解释手册》(Kaufman & Kaufman, 1983b)第五章和第六章所描述的多种分数、剖面分析及诊断解释,无疑会制止这种流行倾向。该手册第六章提供一则从假设提出到假设检验的优秀实例,这是临床诊断方法的本质。而且,作者努力消除某些流行的错误概念,他们在《解释手册》的第一章中(Kaufman & Kaufman, 1983b, pp. 20 ~ 24)明确陈述:“K-ABC 不是‘测量固有的或不可改变的能力’;看来所有认知作业均测量个体学会的东西”。他们告诫,像其他任何测验一样, K-ABC 不是‘完全的成套测验’;而应该由符合个体需要的其他工具加以补充。

尽管作者有言在先,但使用“成就测验”这个术语也许是不恰当的选择,因为关于能力倾向测验与成就测验之间的关系,有一种流行的错误概念。当一个测验与具体的教学内容有着密切的联系,同时又假设测验参加者学过这种内容,该测验就能明确地称之为成就测验。在 K-ABC 中所说的‘成就’测验,情况并非如此,作者尽力使测验与课堂上所获得的具体知识无关。事实上,在所发展的各种能力的连续统一体之内,这些测验与能力倾向一端的距离,比之距成就一端近得多,这个结论由分测验的

相关所证实。当一个术语获得了附加的意义,带有不是原来想要表达的含意,又导致了流行的错误概念时,它就没有什么理由再使用了。

K-ABC 的理论框架是编制项目和规定作业的指南,对这种理论框架作明确陈述则是一项可喜的革新,符合测验编制原理。然而,K-ABC 发表 10 多年后,仍然存在着所选择的特定的理论倾向是否最适合于测验目的这个问题。同时加工与系列加工之间的区分,是理解 K-ABC 成绩的基础,这种区分尤其受到挑战。据认为,如此命名的两组分测验,也可以表示为言语推理测验和非言语推理测验(J. P. Das, 1984; Goetz & Hall, 1984; A. R. Jensen, 1984; Keith, 1985; Keith & Dunbar, 1984)。

另一方面,关于 K-ABC 的诸多研究则认为,根据分数的预测效度及测量‘一般智力’的程度,K-ABC 的综合分数类同于 WISC-R 的分数(Kamphaus, 1990)。由于 K-ABC 不大依赖言语技能,对于英语水平有限的儿童或听力残疾儿童,它倒是测量的一种选择。在《K-ABC 的临床应用和研究应用》(Kamphaus & Reynolds, 1987, 第八章)一书中,可以找到对这种相对新型的工具之长处和局限性的全面陈述。

### 考夫曼青少年和成人智力量表

性质和编制 KAIT(Kaufman & Kaufman, 1993)适用于测量年龄 11~85 岁及以上者的智力,它试图将霍恩和卡特尔(Horn & Cattell, 1966)所表述的流体智力和晶体智力的理论与其他理论家所提出的关于成人智力的各种观点(Golden, 1981; Luria, 1980; Piaget, 1972)相结合。

这套测验由一个晶体量表(Crystallized Scale)和一个流体量表(Fluid Scale)所组成,前者测量学校教育和文化适应中获得的

概念，后者则测量解决新问题的能力。核心成套测验由每个量表中各取 3 个分测验所组成，再加上一个从 4 个指定的分测验中任取的一个分测验就组成扩展的成套测验，我们也可以把它用于被怀疑为神经损伤的个体。KAIT 还包括一个简短的心理状况测验，用来评定认知损伤严重、不能参加完整的成套测验的被试的注意和定向。

一般评价 在技术质量方面，KAIT 看来符合心理测量标准及当代任何一个主要智力量表。常模取样是合适的，手册中所报告的信度和效度数据看来也有前途。这套测验相对容易施测。此外，手册中包括的资料在测验的施测和评分时十分有用（例如，某被试以英语之外的语言作答时怎么办）。

然而，KAIT 真正有别于其他成人智力量表的东西，倒是作者编制和试验最初的项目库中 2500 多个项目时的一丝不苟。所设计的项目适用成人测验参加者。它们往往要求皮亚杰形式运算思维中典型的问题解决的类型，以及卢利亚（Luria, 1980）和高尔顿（Golden, 1981）所提出的、成人思维特有的有计划的评价机能。结果，大多数项目都是不同寻常且趣味横生。许多分测验甚至在标题上也如同游戏——例如著名人像（Famous Faces）、神秘代码（Mystery Codes）、双重意义（Double Meanings）等。其他分测验也以新颖作业为特色，例如画谜学习（Rebus Learning）。在该分测验中，被试学习与特定的画谜（图画）有关的单词，然后“读出”由画谜所构成的词组或句子。图 8-4 说明一个例题。同任何新型量表一样，对 KAIT 的决定性的检验，在于它的吸引力是否足可引起大量的研究和应用，以便建立丰富的、持久的基础数据库。

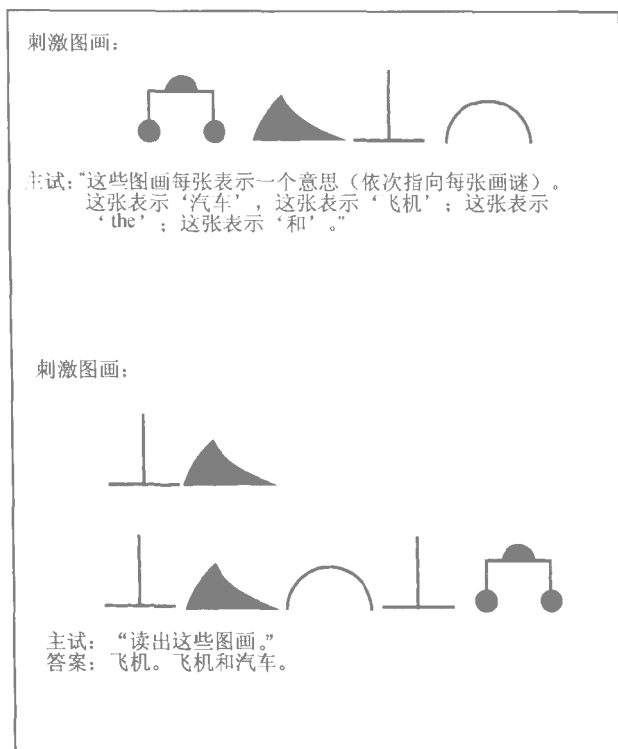


图 8-4 考夫曼青少年和成人智力量表画谜学习分测验  
 的例题(摘自考夫曼和考夫曼,1993,P.5)

### 考夫曼简短智力测验

考夫曼简短智力测验 (Kaufman Brief Intelligence, K - BIT, Kaufman & Kaufman, 1990) 是一种评估智力机能水平的快速筛选工具。虽然该测验是个别施测的，但它十分简单，只需一个办事

员就能施行。K-BIT 适用年龄范围为 4~90 岁。K-BIT 兼用 KAIT 的部分常模,大约使用 2000 个 KAIT 标准化样本中的 20%。

K-BIT 不是 K-ABC 或 KAIT 的缩短的版本。它的组成为一个言语分测验和一个非言语分测验,前者包括 45 个表达性词汇项目和 37 个定义项目,后者包括 48 个矩阵。K-BIT 所得出的 3 个分数(言语、非言语、综合),同其他考夫曼量表一样,按照离差智商单位来表示。K-BIT 分测验的长度,使得它的信度系数高于其他量表的简式。然而,关于它的分数与完整量表分数之间的相关这个关键问题,看来 K-BIT 并不比其他智力测验的简式有什么改进。

## 区别能力量表

埃利奥特(C. D. Elliott, 1990a, 1990b)编制的区别能力量表(Differential Ability Scales, DAS)是英国能力量表(British Ability Scales, BAS)的修订和扩展,后者 70 年代发表于英国(Elliott, Murray & Pearson, 1979)。DAS 与斯坦福一比纳量表和韦克斯勒量表的当前版本具有同样的目的,即按照一般能力水平对个体进行分类,并得出个体智能的长处和不足的剖面图。然而,在程序和技术质量上,DAS 与众不同,体现出其他量表所没有的心理测量的理论和实践上的进展。在这点上,值得注意的是,DAS 手册的序言宣称“智力”和“IQ”这种术语不是 DAS 的用语(Elliott, 1990a, p. vi),相反,量表的组织、评分、解释等都强调实际上所评定的确切界定的行为。如此明确的陈述首次出现在一般能力量表的手册中,这应该有助于消除使用这些术语中普遍的成见和

误解。

**描述** DAS 主要以可靠的信度测量各种具体的能力,以便有助于实现个体评定的较复杂的目标,即作出区别诊断和处理计划。在理论和经验原则的指导下,选择各种作业组成成套测验。其理论基础是折衷主义。DAS 使用一种层次方法来研究各种心理能力,这种方法能适应不同水平的概括性,又能提供广泛的信息基础,从中得出关于各个被试的假设。这种结构与关于认知能力发展的经验研究结果十分一致。以往的量表不管是否合适都把数据强加于理论模型而 DAS 则不一样,它仅仅保留经验和理论原则都支持的那些组份。

如图 8-5 所示,DAS 由 20 个分测验组成,分为 3 个主要组份:(1)核心分测验;(2)诊断分测验;(3)成就测验。这些测验和分测验的标题,描述了它们所包括的各种作业,其方法主要是自我解释。12 个核心分测验和 5 个诊断分测验构成认知成套测验(cognitive battery),它本身分为学龄前水平和学龄水平。在学龄前水平,4 个核心分测验适用年龄为 2 岁 6 个月至 3 岁 5 个月,6 个核心分测验适用年龄为 3 岁 6 个月至 5 岁 11 个月。在学龄水平(6 岁至 17 岁 11 个月),有 6 个核心分测验。对于每个年龄水平,使用核心分测验的总和得出一般概念能力(General Conceptual Ability, GCA)分数,这是 DAS 的总体概括的分数。DAS 成套测验称为‘核心分测验’的各个测验,在成套测验的一般因素即  $g$  上的负荷较高。另一方面,诊断分测验与  $g$  的相关则较低,没有形成一组因素;这就意味着它们测量相对独立的各种能力。根据年龄水平,使用 2~5 个诊断测验。最后,提供 3 个成就测验,一般用于 6 岁以上的被试。

在图 8-5 中能够看到,在所规定的正常年龄水平之外,也可以施测并恰当解释 DAS 的每个主要组份中的一些分测验。

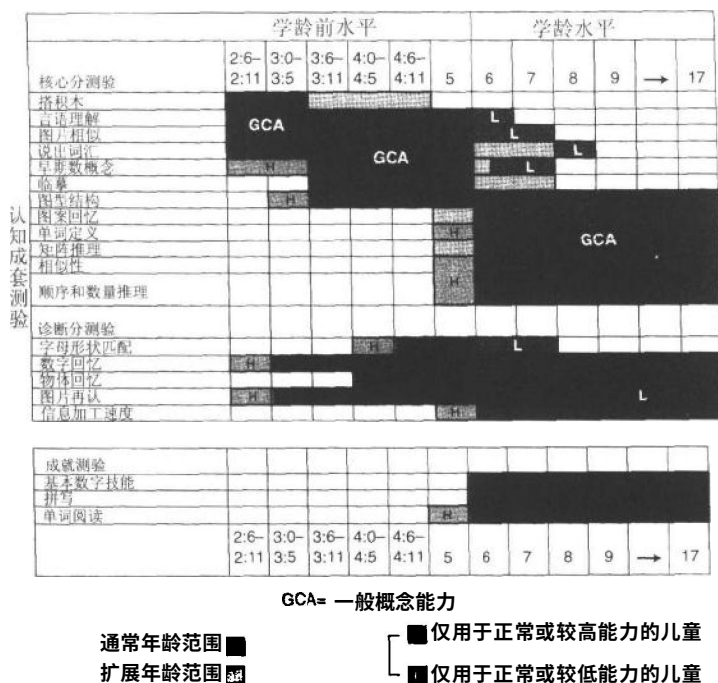


图 8-5 区别能力量表的结构 (摘自埃利奥特, 1990b, p. 4)

称为“扩展的年龄范围”(extended age range)和“水平之外”(out of level)的分测验上的成绩, 可以同合适的年龄样本的常模相比较, 收集年龄样本是 DAS 标准化过程中的一部分。扩展的年龄范围中的各种分测验, 当它们的内容和评定目的有关时, 可以用作某个被试的附加诊断测量。例如, 搭积木分测验也能够用于年龄为 3 岁 6 个月至 4 岁 11 个月的儿童, 以便在核心成套测验之外获得那个年龄上更多的关于知觉技能和精细运动技能的

信息。另一方面,称为水平之外测验的分测验(图 8-5 中标有“H”或“L”),仅仅适用于正常至较高(“H”)或正常至较低(“L”)能力水平的个体。这个特点的益处在于,准许主试以前所未有的精确程度来评定相对于他们的年龄而言成绩处于异常的高水平或低水平的特殊个体。

**编制量表和建立常模** DAS 在观念和技术上成功的一个主要原因是,它从原有量表即 BAS 的编制所积累的研究和思考中获益匪浅。BAS 的设计、编制及标准化历时约 20 年,在这期间,心理测量的理论和实践取得了重大进展。因此,BAS 是一种具有许多最新特点的新型工具,它还反映了本世纪 60 年代至 80 年代所取得的知识和经验。

DAS 的标准化,在样本规模和取样认真两方面都是范例。该样本由 3475 名被试组成,对于个别施测的测验来说,这是一个不同寻常的、相当大的数字。该样本设计来代表的目标总体为在数据收集期间(1987~1989),居住在美国的、专门机构(指教养院、疯人院等)之外的 2 岁 6 个月至 17 岁 11 个月的所有精通英语的个体。分层取样主要根据年龄、性别、种族、父母教育及地理区域。目标总体的统计数字不像通常那样,只根据一次人口统计的一个变量,而是根据各个变量的结合。例如,DAS 样本中居住在东北部的白人女性的父母教育的分布,接近美国东北部总体中这类女性的相应分布。使用从人口统计局得到的、当时最新的原始数据,以计算出目标统计数字。虽然标准化样本代表了总体的种族构成(分类为黑人、西班牙人、白人及其他人),但收集了约 300~600 名黑人和西班牙人的附加数据,仅仅用作偏向分析。常模样本中并没有排除特殊教育班级中的学生,例如轻度损伤儿童或天才儿童,这就意味着它不仅是一个“正常的”样本,而且包括所有学龄儿童总体。



DAS 使用项目反应理论 (IRT) 中的一元参数模型,<sup>[10]</sup>以便校准每一项目的难度值结果,能够用最适合于他们能力水平的项目(即适应测验)对个体进行测验。个体的分数根据所完成的项目的数量和难度水平两者得出。这些数据被置于一个共同的、非常模性的量表上,利用该量表,把每一分测验上的原始分数转换成能力分数。良好的拟合性这个统计量根据所预测的项目反应和所观察的测验反应之间的对应性,也可以使用它来找出和淘汰不符合模型的项目;这样做的结果是得到较具同质性的项目集。

在 DAS 的实际施测中,完成适应测验策略是使用所谓的根据年龄的起始点(starting points)、根据从起始点到决策点上成绩的决策点(decision points),以及根据每个具体分测验的不同停测规则(alternative stopping rules)。这些点之间所包括的项目集凭经验确定,以便得到信度和测验长度之间的最有效的平衡。DAS 的适应测验策略的主要优点是灵活性,它让主试选择适合于每一被试的项目子集。甚至当实施不同的项目子集时,人们仍然能够获得基于共同的项目难度量表的能力估计值,这个事实具有另外的优点,可以比较不同个体的分测验能力分数,以及同一个体在不同情况下的分测验能力分数。这个特点使得 DAS 及按照同样方式编制的其他工具,特别适合于使用纵向法或交叉顺序设计的发展研究。

一旦获得认知成套测验的分测验上的能力分数,就可以把它们转换成平均数为 50、标准差为 10 的正态化标准分数(T 分数)以及百分量表。每个年龄组都提供这两类分数。成就测验使用平均数为 100、标准差为 15 的标准分数而不是 T 分数,使用年级百分量表而不是年龄百分量表。对于所有 DAS 分测验,都能够获得年龄当量分数。对于各个成就测验,也能够获得年级

当量分数。这些当量表示被试的能力分数相当于所在年龄（或年级）分数的中位数。由于 DAS 认知测量和成就测量的编制及常模建立同时进行，通过所有这些分数的转换有可能进行常模比较，这使得使用者能够提出适合于对各个问题进行精确研究的种种议题。

DAS 核心分测验分数相加，就导出在不同年龄水平上可利用的适当的合成分数。所有合成分数都表示为平均数为 100、标准差为 15 的标准分数。如图 8-6 所示，对于最小年龄水平的儿童（2 岁 6 个月至 3 岁 5 个月），惟一可得到的合成分数是

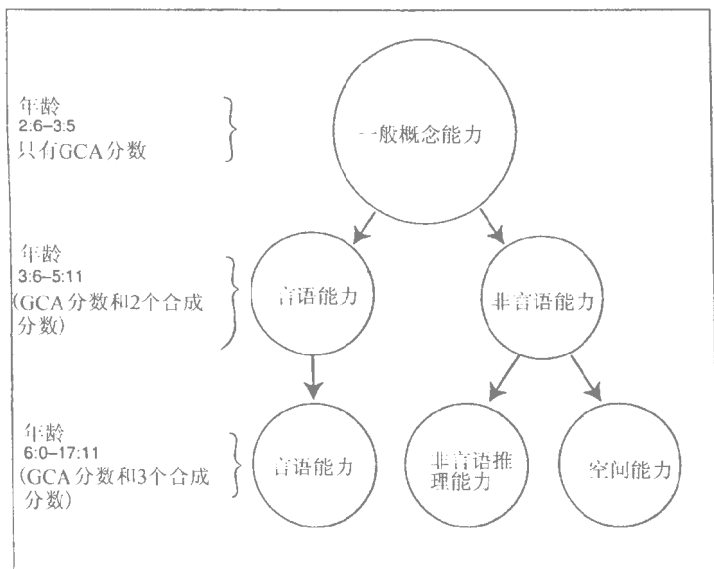


图 8-6 区别能力量表认知成套测验的合成分数  
(摘自埃利奥特, 1990b, P.21)

GCA 对于年龄在 3 岁 6 个月至 5 岁 11 个月之间的儿童,除了 GCA 之外,DAS 还提供 2 个合成分数(言语能力和非言语能力);对于所有学龄被试(6 岁 0 个月至 17 岁 11 个月),则还能够得到 3 个合成分数(言语能力、非言语推理能力、空间能力等)。通过在不同年龄原始成绩和 GCA 分数之间的外推关系,还可以把 GCA 常模的使用扩展到远低于标准常模的成绩水平。这种规定便于评定在常规样本中不具代表性的严重智力落后个体。

**信度和效度** 同其他智力测验相比,DAS 的信度指标较好。两次测验间隔 2~7 周,GCA 和合成分数的再测信度系数为 0.79~0.94。分测验的可资比较的再测信度估计值为 0.38~0.94,其中绝大多数在 0.60~0.90 之间。评分中需要大量判断的开放式应答分测验<sup>[11]</sup>的评分者信度,多为 0.95 左右。

DAS 的内在一致性系数通过应用项目反应理论 (IRT)来评定。这种程序可以计算出对应分测验上每个可能分数的精确的信度和测量误差。这些结果在整个能力范围内变化很大,这证实了众所周知的倾向,即处于分数分布范围两端的个体的信度系数要低于靠近中部的个体。DAS 在所有适当年龄水平上的内在一致性信度估计值,对于分测验为 0.66~0.95,对于合成分数为 0.86~0.94,对于 GCA 分数为 0.89~0.96。DAS 手册也报告了基于 RIT 的内在一致性系数和传统的内在一致性系数之间的一些比较。在这种情况下,两种方法所得出的信度统计数字十分一致。

从内部和外部两种观点来探索 DAS 的效度。在内部效度方面,进行探索性因素分析和证实性因素分析,得出 DAS 分数的结构(参见图 8-6)。两类分析的结果十分相似,从整体来看,这为各种能力随着年龄增长而逐渐分化这个研究结论,提供了进一步的证据 (Anastasi, 1970)。对于 2 岁 6 个月至 3 岁 5 个

月的儿童,得出 GCA 分数要使用 4 个核心分测验,包括这些分测验的单因素模型(最符合那个年龄组儿童的数据)。对于 3 岁 6 个月至 5 岁 11 个月之间的儿童,则需要二因素模型。在这个年龄范围内,非言语能力因素主要由图形构建和临摹分测验上的较高负荷所确定,而言语能力因素则主要由言语理解和词汇命名分测验所确定。在 6 岁 0 个月至 17 岁 11 个月,使用言语能力、非言语推理能力、空间能力等组成的三因素模型,最能说明核心分测验之间的相关。GCA 或结合分数中均不包括 DAS 的 5 个诊断分测验。这些分测验一般由需要记忆和信息加工速度的各种作业所组成。诊断分测验在一般因素(即“g”)上的负荷很小,而在具体变量上的负荷很大,这就使得它们十分适合于鉴定个体的长处和不足。

手册中报告了关于 DAS 外部效度的大量数据。这类数据的主要来源是:(1)DAS 合成分数和分测验分数与斯坦福-比纳量表、韦克斯勒量表等其他多种能力成套测验上分数之间的相关;(2)DAS 分测验分数与皮博迪图片词汇测验修订版(PPVT-R)、伍德科克阅读掌握测验修订版(WRMT-R)等其他具体认知能力测验和学业成就测验之间的相关,以及与学校年级的相关;(3)对诸如天才儿童、学习不能儿童、心理落后儿童等特殊总体的 DAS 分数剖面图的研究。所有这些数据来源,一般来说都证实 DAS 的层次结构,以及它的合成分数、分测验分数和成就测验分数与相似度量的可比性。对于鉴定学习不能儿童亚团体,诊断分测验的效度看来是有价值的,但需要进一步的研究。

一般评价 其他人(Aylward, 1992; Reinehr, 1992)认为,DAS 施测和评分程序的复杂性,可能妨碍它在应用环境下的传播和使用。另外,DAS 是一种相对新型的、未经临床检验的工具,尚需证实它的实用价值。再一个局限性是,对于 DAS 所包括的最

小年龄组和最大年龄组(即 2 岁 6 个月和 17 岁 11 个月)中的被试,项目难度的下限水平和上限水平也许不足。然而,DAS 是同类量表一种“达到目前工艺水平”的工具,它给使用者提供的各种可能的用途和优点,是其他量表所无法比拟的。它的层次结构,它所包括的能力的多样性及评定多种能力的信度,让使用者具有前所未有的灵活性。具体说来,主试可以从范围广泛的作业中,选取最适合于评价目的和测验中的被试的需要的那些作业。DAS 的另一个显著特点是高质量的《施测和评分手册》和《介绍和技术手册》(Elliott, 1990a, 1990b)。尤其后者,内容十分丰富,信息量大且精确、清晰、简要。它对未来的使用者应该特别有帮助,尤其是对于那些希望更好地熟悉智力及智力测量领域中理论和实践进展的使用者,因为 DAS 正是充分体现了这种进展。

## 达斯—纳格里里认知评定系统

达斯—纳格里里认知评定系统 (Das-Naglieri Cognitive Assessment System, CAS) 发表于本世纪 90 年代后期,这是个别评定认知机能的另一个重要的新型工具。该量表的编制历时十余年,它所根据的是作者提出的 PASS 智力模型 (J. P. Das, Naglieri & Kirby, 1994; Naglieri & Das, 1990, 1997a, 1997b), 这种模型又根植于俄罗斯神经心理学家鲁利亚 (Á. R. Luria) 提出的认知和大脑组织的理论。

CAS 的各种作业旨在测量基本的认知机能,这种机能与学习有关,但又被假设独立于教育。这种机能包括计划、注意、同时加工和系列加工。该系统使用言语测验和非言语测验,而通

过视觉和听觉通道来呈现测验。计划测验要求评定被试在完成作业中所使用的策略。CAS 适用于 5 岁 0 个月至 17 岁 11 个月的个体, 尤其把评定和干预相联系。由于它的成熟的理论基础和实验基础, 以及它所经过的认真的、大规模的标准化, 许多测验使用者都热切期待 CAS 的完成。事实上, 已经有人对它的试用本发表评论(Lambert, 1990; Telzrow, 1990)。根据可得到的关于 CAS 效度的初步证据, 看来该测验会成为评定认知状况的一个重要的、革新的工具。

## 注 释

- [1] 10 卷本《测验评论》丛书(Keyser & Sweetland, 1984 ~ 1994)是另一个有用的信息源,它提供数百个测验的评论。
- [2] 参见弗拉纳根、金沙夫特和哈里森(Flanagan, Genshaft & Harrison, 1997)所编的《当代智力评定》卷,它对本书第 3 部分所讨论的许多主题都作了很好的评述。
- [3] 关于比纳-西蒙量表以及斯坦福-比纳量表的发展、使用和临床解释的详尽评述,参见萨特勒(Sattler, 1982, 1988)的有关论述。
- [4] 谈到的 5 个测验是词汇、解释、挑错、临摹和言语关系。
- [5] 常模表参见桑代克等(Thorndike et al. 1986a, pp. 183 ~ 188)的有关论述。一些 SAS 数值,其根据不到 100 名个案,用统计方法来估计整个年龄组,在常模表中用阴影表示。当个体在例行测验上的成绩相对于其年龄而言,不同寻常的高或低时,就会出现这类分数(Thorndike 等, 1986b, pp. 29 ~ 30)。
- [6] 其他测验为 WISC - R、WAIS - R、WPPSI 和 K - ABC,本章后面将分别讨论。
- [7] 不宜适用分半信度的分测验除外(即数符号、译码、动物房、符号搜寻和数字广度)。
- [8] 参见迪安(Dean, 1977, 1979, 1980) 古特金和雷诺兹(Cutkin & Reynolds, 1981) 霍伦贝克和考夫曼(G. P. Hollenbeck & Kaufman, 1973) 卡恩斯和

布朗 (Karnes & Brown, 1980) 考夫曼 (1975); 考夫曼和霍伦贝格 (1974); 莱克利特、马特拉泽和西尔弗斯坦 (Leckliter, Matarazzo & Silverstein, 1986) 西尔弗斯坦 (1982a, 1982b) 以及沃拉和沃尔德曼 (Waller & Waldman, 1990) 的有关论述

- [9] 关于评论和批评讨论, 主要参见米勒 (T. L. Miller, 1984) 的有关论述也参见阿纳斯塔西 (Anastasi, 1984a, 1985c) 科夫曼 (Coffman, 1985) 坎弗斯 (Kamphaus, 1990); 克兰、斯奈德和卡斯特利诺斯 (Kline, Snyder & Castellanos, 1996) 以及佩奇 (Page, 1985) 的有关论述。

- [10] 关于解释, 参见第七章。

- [11] 讨论中的分测验为单词定义、相似性、临摹、图案回忆等。

## 第九章 特殊总体的测验

本章所讨论的测验包括个别量表和团体量表。编制这类测验，主要适用于特殊个体，而传统的工具，例如上一章所讨论的个别量表或下一章所要讨论的典型的团体测验则不能恰当地测验这种个体。历史上，本章所讨论的这类测验，称为操作测验、非语言测验及非言语测验。

从整体来看，操作测验(performance tests)系对物体进行操作，最低限度地使用纸和笔。非语言测验(nonlanguage tests)就主试或被试而言，都不需要语言。这类测验的指导语能够通过演示、手势、哑剧等给出，而不使用口头语言或书面语言。军队乙种测验(Army Examination Beta)是非语言团体测验的原型，编制该测验，是为了在第一次世界大战期间测验操外语的新兵和文盲新兵(Yerkes, 1921)。后来该测验的修订本转为民用。对于大多数测验目的，在测验施测中并不需要排除所有语言，因为测验参加者通常懂得一点常用语言。此外，一般来说，可以把简短的、简单的指导语翻译一下，或以两种语言相继给出，这样不会明显改变测验的性质或难度。但是，这些测验都不需要测验参加者本身使用书面语言或口头语言。

与之相关的是非言语测验(nonverbal tests)，也许称之为非阅读测验更为合适。适用于小学和学龄前儿童的大多数测验属于此类，适用于任何年龄水平的文盲和没有阅读能力者的测验



也是如此。这些测验不需要阅读或书写，而主试则大量使用口头的指导语和交流。此外，通过使用附有口头指导语的图片项目，它们经常测量言语理解，例如词汇再认及理解句子和短语。同非语言测验不一样，它们不适合于听力残疾的个体或不讲英语的个体。

虽然操作测验、非语言测验及非言语测验之间的传统区分，有助于理解不同测验的目的，但是由于越来越多地编制了涉及这数类测验的成套测验，这些区分变得有点模糊起来。韦克斯勒量表中言语测验和操作测验相结合，就是一个典型的实例。

本章中测验的分类，不是按照内容或施测程序，而是参照主要用途。根据这一标准，可以将这些测验分为 4 个主要类型 适用于婴儿和学龄前水平的测验；适用于综合评定智力落后个体的测验；适用于有各种感觉残疾和动作残疾个体的测验；适用于跨文化或亚文化的测验。然而，这种分类必须保持灵活性，因为其中若干测验被证明不止在一种情境中 useful。对于最初设计为跨文化测验而现在多用于临床检查的那些量表，情况尤为如此。<sup>[1]</sup>

最后，虽然本章中所包括的一些测验设计为团体测验，但经常用于个别施测。一些测验广泛应用于临床测验情境，用来补充常规的智力测验，由此较全面地描述个体的智力机能。有几个测验可以与个别测验联系起来进行定性观察，但需要相当的临床经验，才有可能对测验成绩作出详尽解释。从整体来看，它们更接近于第八章所讨论的个别测验而不是第十章所讨论的团体测验。

## 婴儿和学龄前儿童测验

为婴儿和学龄前儿童设计的所有测验都需要个别施测。使用一些为低年级儿童编制的测验,以小团体形式,可以对一些幼儿园儿童进行测验。一般来说,儿童要达到入学年龄之后,才能实施团体测验。大多数适用于6岁以下儿童的测验,或是操作测验或是口头测验。一些测验包括初步的纸笔操作。

习惯上,把生命的头5年再分为婴儿期和学龄前期,前者从出生到约18个月,后者从18个月到60个月。从施测的观点来看,应当指出,对婴儿进行测验时,婴儿必须躺下、扶着或抱在膝上,对此本章后面会加以说明。给出测验指导语时,主试不使用什么语言,虽然有关资料表明,儿童已有自己的语言发展。许多测验针对感觉运动发展,婴儿的这种能力表现为抬头、翻身、拿取和抓握物体,以及眼睛追随运动的物体。另一方面,学龄前儿童会走路,或坐在桌边,用手操作测验物体,以及用言语进行交流。在学龄前水平,儿童也往往把主试作为一个人而对他作出回答。而对于婴儿,主试则主要作为提供刺激物体的工具。学龄前测验是一种高度人际交流的过程——这个特点增加了测验情境所提出的机遇和困难。

对年幼儿童进行适当的心理检查,需要考虑一系列广泛的行为,这包括社会特质和情绪特质、运动、语言及其他认知能力。此外,人们日益认识到,在评定儿童时需要考虑儿童环境的性质(Vazquez Nuttall, Romero & Kalesnik, 1992)。本章所讨论的一些量表反映这种生态学倾向。本节所讨论的一些典型量表为婴儿和幼儿所设计,表现出方法的多样化。韦克斯勒学龄前和小学

生智力量表也属于这种类型,但被包括在第八章中,以便保持韦克斯勒系列量表的连续性。在评定学龄前儿童时,也可以使用第八章中所讨论的斯坦福—比纳量表、考夫曼儿童成套评定测验以及区别能力量表,因为它们也包括 2~6 岁阶段。

**婴儿和学龄前儿童测验的历史背景** 耶鲁大学阿诺德·格塞尔(Arnold Gesell)及其同事进行的一系列纵向研究,是了解正常的婴儿和学龄前儿童发展的最早的系统研究之一(Ames, 1989)。这些研究共历时 40 年,结果编成格塞尔发展顺序量表(Gessel Developmental Schedules),首次发表时(Gesell 等, 1940),它代表一种开拓性的尝试,为评定幼儿行为发展而提供一种系统的、基于经验的方法。通过直接观察儿童对标准玩具和其他刺激物体所作出的反应,并且补充由父母或主要监护人所提供的信息,从而获得量表的大多数资料。多年来,其他研究人员对格塞尔量表进行了修订和更新,从事研究和实践的心理学家和儿科医生广泛使用格塞尔量表。至今一些人仍然使用该量表,用于补充医学检查,尤其适用于鉴定婴儿的神经缺陷及器质性行为异常。<sup>[2]</sup>虽然在大多数临床环境下,新近编制的、在心理测量上也更高级的量表取代了格塞尔量表,但是为婴儿水平所设计的大多数其他发展量表,承袭了格塞尔及其同事所开创的项目和程序。

从 60 年代至 90 年代的几十年里,出现了婴儿和学龄前儿童测验热。其中最初起作用的一种因素是,关于智力落后儿童的教育课程快速发展;另一种因素是关于处在文化劣势环境中儿童的学前补偿教育课程的广泛发展。关于婴儿和学龄前儿童中各种生理和心理残疾的早期鉴定和治疗,近来颁布了一系列法令,其中的一些法律(例如, P. L. 99 ~ 457) 是“残疾儿童教育法案”(P. L. 94 ~ 142)的修正案或补充,本章后面将详细讨论该法

案。总而言之,为了满足实践的迫切需要,新的测验和书刊以及关于新的评定方法的研究,如同雨后春笋般地涌现。[3]

### 儿童早期发展的标准化测验

贝利婴儿发展量表(Bayley Scales of Infant Development) 图 9-1 所说明的贝利婴儿发展量表是一个编制十分出色的、用于最小年龄水平的测验,现在它已有了第 2 版(Bayley - II——Bayley, 1993)。该量表从格塞尔量表及其他婴儿和学龄前测验中吸收了一些项目,最初发表于 1969 年。它体现了贝利及其同事多年研究的最终成果,其中包括对伯克利发展研究(Berkeley Growth Study)的纵向研究。

贝利量表第 2 版为评定年龄从 1 个月至 3 岁半之间儿童发展状况而提供 3 个互补的工具:心理量表、运动量表和行为评定量表。心理量表测量感知觉敏度、记忆、学习、问题解决、发音、初步的言语交流、初步的抽象思维等机能。运动量表测量坐、站、走、爬楼梯等大运动能力以及双手和手指的操作技能;也包括评定感觉和知觉—动作统合的项目。在婴儿水平,运动和操作的发展,在儿童与环境的相互作用中起重要作用,因此在儿童的智能发展过程中也起着重要作用。行为评定量表设计来评定个性发展的各个方面,例如情绪行为和社会行为、注意广度和唤醒、持久性、目标定向等。对于每个项目和所评定的具体行为,它都采用五级评分制。主试在施测其他两部分之后再来完成行为评定量表。它根据从儿童监护人处得到的信息,也根据主试自己的印象。

在测验编制程序的技术质量上,贝利量表在婴儿测验中是杰出的。常模的建立所根据的 1700 名儿童,从 1 个月至 42 个月,分为 17 个年龄组,每组包括男女各 50 人。为了代表美国人

口总体状况,按照种族、主要地理区域、父母教育等来选取标准化样本。样本只考虑正常儿童,即怀孕 36~42 周后出生,没有明显的药物并发症,没有因心理问题、生理问题或行为问题接受过治疗的那些儿童。心理量表和运动量表分别得出发展指标,以平均数为 100、标准差为 15 的正态化标准分数来表示。在

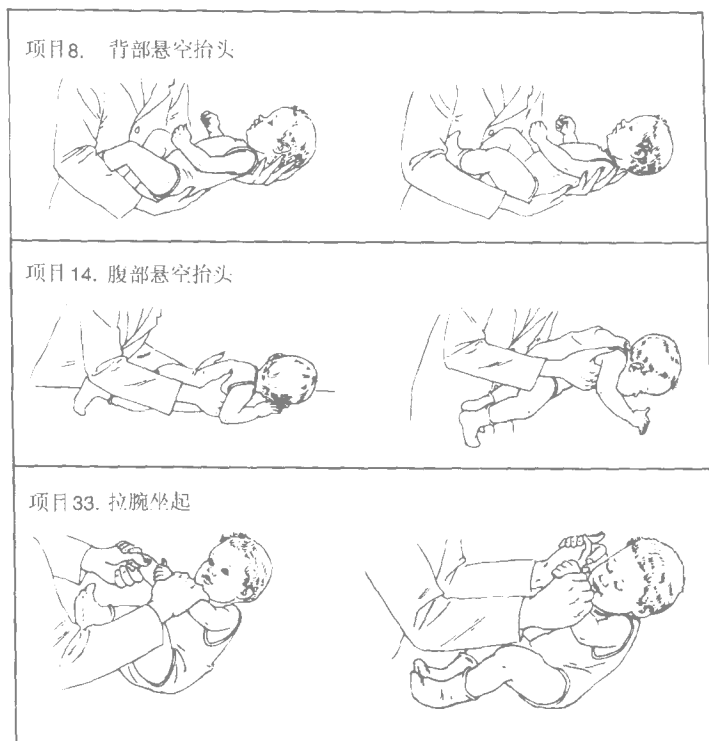


图 9-1 贝利婴儿发展量表第2版列举的婴儿测验

(摘自贝利,1993,pp.143,145 和 150)

儿童自己的年龄组内可以找到这些指标,从 1 个月至 36 个月,每隔 1 个月为一个年龄组;此后每隔 3 个月为一个年龄组。行为评定量表则得出百分等级分数,依次分类为“不理想”、“有问题的”、“在正常范围内”。最近通过对若干学龄前儿童评定量表进行评价,证明贝利量表第 2 版是符合大多数技术标准的两个测验之一(D. P. Flanagan & Alfonso, 1995)。<sup>[4]</sup>

贝利指出,像所有婴儿测验一样,她的量表应该主要用于评定当时的发展状况,而不是预测以后的能力水平。婴儿期的能力发展很容易受到许多干预因素的影响,所以作出长期预测就没有什么价值,第十二章将讨论这一点。<sup>[5]</sup>

自从贝利量表发表以来,婴儿评定的焦点,从评价正常婴儿转变到评价有可能会发展迟缓的儿童或实际上发展已经迟缓的儿童。虽然临床使用贝利量表第 2 版尚需要进一步的研究,但现有量表已经包括了一些项目,它们是用来最大限度地鉴别正常样本和临床样本,并提供了各种特殊的临床总体的资料,因此,它应该有助于早期发现感觉缺陷和神经缺陷、情绪障碍及环境障碍。此外,艾尔沃德(Aylward, 1995)使用取自贝利量表第 2 版和其他神经测验中的 11 和 13 个项目,编制成贝利婴儿神经发展筛选量表(Bayley Infant Neurodevelopmental Screener, BINS),该量表用来快速评定 3 个月至 24 个月的婴儿。

**麦卡锡儿童能力量表** 麦卡锡儿童能力量表(McCarthy Scales of Children's Ability, MSCA——McCarthy, 1972)是编制出色的学龄前水平量表,适用于 2 岁半至 8 岁半之间的儿童。它由 18 个分测验组成,让主试有多种机会观察儿童解决种种问题和刺激的方法。分测验分为 6 个部分重叠的量表:言语、知觉—操作、数量、一般认知、记忆、动作等。一般认知量表的分数,根据成套测验 18 个测验中的 15 个,最接近于传统的智力发展的全

面测量。这种一般认知指标(GCI)是一种正态化的标准分数,以同传统的 IQ 一样的单位来报告(平均数为 100,标准差为 16),每隔 3 个月为一年龄组。在 MSCA 的编制中,故意回避 IQ 这一术语,因为它包含许多令人误解的涵义。GCI 作为测验当时儿童机能的一项指标,并不意味着它的不变性或原因。在另外 5 个量表上的分数都根据同一个年龄组,它们的平均数为 50,标准差为 10。

麦卡锡量表是惟一专门为评定幼儿认知而设计的量表,自从它发表以来的 20 年里,关于使用该量表的研究非常之多。在许多研究中,瓦伦西亚(Valencia)使用麦卡锡量表评定少数民族儿童的详细报告(1990)、布拉肯(Bracken)总结的使用麦卡锡量表评定幼儿的大量效度资料(1991a)等尤其有价值关于麦卡锡量表的临床使用,考夫曼和考夫曼(Kaufman & Kaufman, 1977)编制的手册仍然是一本必不可少的工具书。众多评述中的主要观点是,尽管有一些不足之处,麦卡锡量表仍然是有效的,也是有用的。该量表的心理测量特性适用于大多数目标总体,尤其是中间范围的目标总体。

### 皮亚杰量表

依照让·皮亚杰发展理论编制的量表,虽然主要适用于学龄前水平之后,但至今在儿童早期也有重要应用。所有这类量表都采用实验型式,没有什么量表是商品化的。虽然大多数量表只是用于编制者自己的研究项目,但其中一些量表也适用于其他研究人员。皮亚杰量表对儿童心理测验的主要贡献是,其理论框架关注思维过程的发展顺序,其评定程序具有灵活性和定性解释的特征。

第三章中讨论过皮亚杰量表的一些特点,尤其是关于成绩

上,

sensorimotor stage)

preoperational stage)

concrete operational stage)

formal opera-



6 种水平。为了提高量表的敏感性,尤季里斯和亨特把反应分为 6 种以上水平,在不同的量表上水平数目在 7~14 之间各不相同。该测验包括 6 个量表,名称如下:

1. 物体永久性(object permanence)——儿童眼睛追踪一个物体:把一个物体藏起来,隐藏程度不断增加,然后儿童去搜寻这个物体。这就表明儿童形成了物体独立存在的概念。

2. 达到所期望的环境目的之工具的发展(development of means)——为了得到物体,儿童用自己的双手和其他物体,例如绳子、小棒、支撑物等。

3. 模仿(imitation)——包括姿势的模仿和声音的模仿。

4. 操作性的因果关系(operational causality)——眼睛观察自己双手、从他人处得到所期望的行为、发动机械玩具,这些反应说明儿童认识和适应客观的因果关系。

5. 物体的空间关系(object relations in space)——儿童协调视听图式,从而对物体进行空间定位,理解容积、平衡、重量等关系。

6. 与物体有关的图式的发展(development of schemata for relating to objects)——儿童通过看、摸、操作、丢下、扔开等动作,以及通过适合于特定物体的社会驱动的图式(例如“开”玩具车、搭积木、戴珠子、给物体起名)对物体作出反应。

虽然没有提供常模,但作者对 84 名伊利诺斯大学毕业生和员工的婴儿进行施测,收集到关于量表心理测量特性的资料。报告了 48 小时之后的再测一致性和观察者一致性。一般来说,测验在两个方面看来都是令人十分满意的。根据同样的 84 名儿童的分数,计算出每个量表的顺序性指标,所报告的这些指标也是令人十分满意的。<sup>[7]</sup>

虽然尤季里斯和亨特所编制的顺序量表本来只是临时性使用的,但现在广泛运用于各种研究目的。<sup>[8]</sup>量表最初设计来测量

具体的环境条件对婴儿发展的进程和速度的影响。关于在不同条件下长大的婴儿的研究,以及关于参加干预课程的婴儿的研究,表明这些环境变量对儿童达到发展量表中不同阶段的平均年龄的显著影响。量表作者在所编辑的专著中,评述了这些研究及使用顺序量表测量正常婴儿和特殊婴儿认知发展的其他研究(Uzgiris & Hunt,1987)。量表所探索的发展顺序,主要适用于婴儿与无生命物的相互作用,各种相互作用又反过来被认为是发展交往行为和其他适应技能的先兆(Dunst & Gallagher,1983; Kahn,1987)。

要讨论的皮亚杰量表的第二个实例是守恒概念评定成套测验(Concept Assessment Kit - Conservation, CAK),同其他心理测验一样,这是由一家测验发行公司所发行的一个测验。它适用于4~7岁的儿童,测量一个众所周知的皮亚杰概念。所谓守恒(conservation)指儿童认识到,物体的形状、位置、形式或其他属性发生变化时,物体的重量、体积、数量等特性仍然不变。编制者(Goldschmid & Bentler,1968b)强调,守恒是儿童思维由前运算阶段向具体运算阶段转变的一种标志,皮亚杰大致把它放在7岁或8岁。

整个测验的程序实际上相同。给儿童呈现两个相同的物体,然后,主试将其中一个物体作某些变化,再问儿童二者是相同还是不同。回答以后,再要求儿童作出解释。每个项目,等效的正确判断记1分,可接受的解释另记1分。例如,主试拿出两个普通玻璃杯,里面装着等量的水(连续的量)或谷粒(不连续的量)。然后,将所装物质倒入一个浅碟子或几个小玻璃杯中。又如,主试给儿童看两个相同的橡皮泥球,然后将其中一个压平成饼状,问球和饼是否一样重。

测验有3种型式。A型和B型是平行的,各包括6项守恒

作业 二维空间、数、实物、连续的量、不连续的量、重量等。这两个型式上分数之间的相关为 0.95。C 型包括两项不同的作业即面积和长度,它与 A 型和 B 型的相关分别为 0.76 和 0.74。

建立常模的标准化样本由 560 名 4~8 岁的男儿童组成,从加利福尼亚的洛杉矶地区的小学、日托中心及幼儿早期教育中心选取。由于每一年龄组中个案人数很少,也由于样本代表性不足,因此必须认为这些常模是试验性的。每一年龄组的平均分数随着年龄而有系统地上升,在 6~8 岁之间则急剧上升,根据皮亚杰理论可以预期到这一点。

编制者进行种种统计分析来评定:评分者信度、库德—理查逊信度、复本信度、再测信度、等级性即顺序性,以及因素组成(也参见 Goldschmid & Bentler, 1968a)。虽然样本很小,但总的说来,研究结果表明,信度是令人满意的;而且令人满意地证明了顺序性及全部作业中存在着守恒这个广泛的共同因素。

在 7 个国家进行的比较研究说明:测验广泛适用于各种不同的文化,信度高,年龄趋向大致相似(Goldschmid 等, 1973)。然而,研究发现,在获得概念的平均年龄上,存在着文化和亚文化之间的差异——即各条年龄曲线平移 1~2 岁(也参见 Figurelli & Keller, 1972; Wasik & Wasik, 1971)。研究也发现,训练显著提高守恒作业的分数(也参见 Goldschmid, 1968; B. J. Zimmerman & Rosenthal, 1974a, 1974b)。手册提出了测验结构效度的数据,坎贝尔和雷米(F. A. Campbell & Ramey, 1990)最近进行的一项研究也是如此。

**对皮亚杰方法的评价** 关于皮亚杰认知发展方法的理论基础和经验基础,都仍然存有争论(例如,参见 Inhelder, de Caprona & Cornu-Wells, 1987; Liben, 1983; Sugarman, 1987)。关于训练效果的涵义、跨文化差异对解释皮亚杰发展阶段的影响,都是有疑问

的。在确定通过顺序量表的各阶段中所碰到的一个主要障碍，是皮亚杰学派研究人员所称的“滞差”（*decalage*），<sup>[9]</sup>即在所预期的顺序中的不一致性。越来越多的资料使人们怀疑智力发展中隐含的连续性和规律性。和某个体成绩对应的阶段往往随作业而变化，不仅当作业解答中需要不同的过程时，而且当同一过程应用于不同的内容时都是如此（Dasen, 1977; Goodnow, 1976; Horn, 1976; J. McV. Hunt, 1976）。

还应当指出，有证据表明：皮亚杰量表同标准化的智力测验高度相关（Gottfried & Brody, 1975; Kaufman, 1971; M. E. Sexton, 1987）。皮亚杰量表和团体智力测验，同一年级儿童的学业成就的相关大致一样高（Kaufman & Kaufman, 1972）。使用不同量表的独立的研究人员都确认了这类重叠（Humphreys, Rich & Davey, 1985）。这些研究结果所表明的是，尽管方法上有显著差异，但是皮亚杰量表、标准化的智力测验、学业成就测量等具有许多共同的东西。此外，它们对于全面评价儿童，各自起着独特的、有价值的作用。皮亚杰量表施测较为困难，花费时间也更多，但是，一旦同常模参照测量和领域参照测量相结合，它们却能够全面描述一个儿童能做什么以及怎样做的（D. Sexton 等, 1990）。

目前关于幼儿心理活动的研究进展迅速。经验的研究结果修正和拓宽了早期皮亚杰概念（参见 Butterworth, Harris, Leslie & Wellman, 1991; Whiten, 1991）。事实上，现在存在着一些统称为“新皮亚杰的”（*neo-Piagetian*）新方法，它们提出各种认知发展的问题，但其观点在不同程度上都受到皮亚杰理论及信息加工理论的影响（Beilin & Pufall, 1992; Demetriou, 1988）。在评定领域，一些新皮亚杰研究人员将种种动力方法结合起来，并以一种形式化的方式使用间接学习，试图评价同个体先前的知识基础关系最少的心理能力（Pascual-Leone & Ijaz, 1991）。这些技术仍然

是试验性的，其目的在于适用于小至二三岁的儿童，以及各种文化、社会、语言背景的个体。

### 目前婴儿和幼儿评定中的趋势

历史上，智力测验的效度主要与年龄差别的效标以及同学业成绩的相关相联系。测量婴儿的适当的发展，几乎毫无例外地把他们在贝利等发展量表中所包括的广泛作业上的结果与同龄常模相比较。然而，最近社会上致力于早期鉴定和矫正缺陷，则要求用来评定婴儿认知机能的量表具有预测力。因此，尽管试图追踪智能在各个年龄水平的发展变化困难重重，人们还是作出新的努力，试图编制在实际应用中具有足够的预测价值的量表和程序。

一个更令人感兴趣的新方法是对信息加工技能的测量，例如费根婴儿智力测验（Fagan Test of Infant Intelligence, Fagan, 1992; Fagan & Detterman, 1992）。这种方法根据所确认的婴儿偏爱新奇事物的研究结果，反过来又使得研究婴儿抽取和保持信息的能力成为可能。费根测验旨在区分正常儿童和认知缺陷儿童，评定 3~12 个月婴儿对新奇事物的有选择的视觉注意。刺激是人脸图片，“评分”根据注视相对于熟悉刺激而言的新奇图片的时间量。图 9-2 说明便携式费根测验，研究发现，预测后来智力测验上的成绩，费根测验等同于或优于标准婴儿智力测量。费根测验分数与 3 岁 IQ 的相关范围大约为 0.45~0.60。费根测验仍在发展中，在一些问题上也受到批评（例如，参见 Benasich & Bejar, 1992; Goodman, 1990）。毫无疑问，关于在预测认知缺陷包括心理落后的临床使用方面，需要积累更多的资料。然而，它所代表的方法具有坚实的经验基础，并且同第十二章中所讨论的婴儿智力性质的研究结果十分一致。

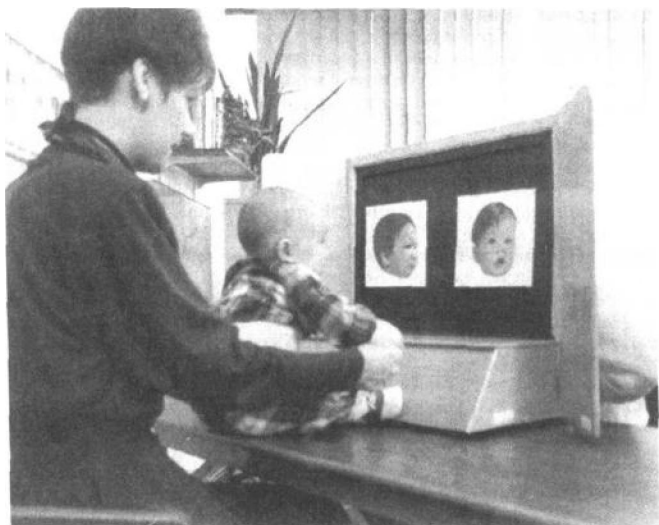


图 9-2 便携式费根婴儿智力测验 (摘自费根和德特曼, 1992, p. 189)

现在人们日益认识到,为了使干预有效果,对幼儿的评定必须是全面的、准确的、有效的。依靠单一的综合分数,不论是称之为 IQ 还是发展指标,都不适合于大多数实践目的。这类分数可以用来把儿童分为各种类型,但量表使用者得不到关于儿童的长处和不足的资料。对婴儿和儿童以及大年龄为个体,进行个别化评定,都需要使用多种方法,以及诸如语言、运动、社会技能等有关领域的多种资料。为了达到这些要求,正在编制新的评定系统,其中涉及各门学科的专业人员和方法。

婴幼儿发展评定 (Infant-Toddler Developmental Assessment, IDA) 就是体现这种趋向的一种实例。它是根据一组多学科幼儿专家完成的研究工作而设计的 (Provence, Erikson, Vater &

Palmeri, 1995a, 1995b, 1995c)。IDA 实际上是一个框架,指导人们鉴别出生至 3 岁中有可能发展迟缓的儿童。IDA 所提供的材料,有利于收集、记录、解释和综合从一项计划的引入直到计划发展的资料。评定程序包括父母参与每一阶段的活动、婴幼儿健康检查、根据观察以及同父母和其他监护人的面谈作出发展评定。IDA 中最类似于传统的婴儿发展测量的部分,是“普罗文斯出生至 3 岁发展剖面图”(Provence Birth to Three Developmental Profile)。虽然剖面图在评定 8 个发展领域和顺序时,施测和评分都采用标准程序,但其分数则是以年龄为参照的、按照文献确证的发展里程碑,而不是通过标准分数、百分等级或其他组内比较。在这点上,编著剖面图的儿科专家萨利·普罗文斯(Sally Provence)遵循了盖塞尔在他的发展时间表中所建立的传统。像任何其他量表一样,IDA 的效果显然是同使用 IDA 的专业人员的训练和经验相联系的,并且需要通过研究和临床使用来作进一步评价。然而,人们设计 IDA 和其他像它一样的评定系统,是为了应付对于过分依靠智力测验的做法所提出的批评(参见 Goodman, 1990),如果使用得当,会证明 IDA 具有巨大的实践价值。应当指出,使用 IDA,并不排除在需要的情况下,出于测量一个儿童在同龄伙伴中的相对地位等目的而施测传统的认知机能测量或任何其他量表

应当指出,除了更全面地综合评定这种趋向之外,另两种倾向对测验幼儿及测验本章所讨论的大多数其他特殊总体,也有重要影响。第一种倾向是生态学观点在评定过程中的影响,这种观点要求考虑儿童环境的各个方面。<sup>[10]</sup>第二种倾向是日益强调评定和干预之间的联系,诊断不能脱离治疗,临床工作中早就认识到这一点。在早期干预和教育的情况下,现在认为必不可少的一种需要是,根据个体能做什么和不能做什么的独特的剖

面图，为教师规定指导原则（Bagnato & Neisworth, 1991; Witt, Elliott, Gresham & Krammer, 1988）。

## 对智力落后者的综合评定

在美国，随着“1975 年各种残疾儿童教育法”（P. L. 94 ~ 142）——现在经修订称为“残疾人教育法”（IDEA）——和有关法律<sup>[11]</sup>的通过，对心理残疾或生理残疾儿童的测验获得了令人瞩目的进展。履行这种法规需要 4 个基本步骤：（1）必须通过初步的筛选工具来鉴定各种残疾儿童；（2）由一组专家对如此鉴定出的儿童进行评价，确定每个儿童的教育需要；（3）学校必须制定个别化教育计划来满足这些需要；（4）在计划进程中，定期对每个儿童进行再评价。本书的若干部分，包括第八章、第十七章以及本章的本节和后两节，讨论在符合法律要求的计划中，使用合适的测验的问题（也参见 Jacobson & Mulick, 1996）。

美国智力落后协会（AAMR, 1992）在关于定义和分类的手册中写道：“智力落后指现存机能中有严重的限制。它的特性是智力机能明显低于平均水平，在下列适应技能领域中，同时存在着与两个或更多领域有关的限制：交流、自我照顾、日常生活、社交技能、社会习惯、自我指导、健康和安全、机能学习、闲暇、工作等。”（p.1）手册还规定，这种状况在 18 岁之前有所表现。这个定义与先前的定义并没有实质上的不同（Grossman, 1983, p. 11）。然而，有关的分类系统变化显著，以适应这样一个观点：智力落后不是一种特质，而是一种源于个体能力限制与环境要求之间相互作用的残疾。先前的分类系统根据智力测验成绩来规定从轻度到极度的智力落后水平。现在的分类系统则对个体在 4 个



方面(智力机能和适应技能、心理／情绪因素、生理健康／病因学因素、环境因素)所需要的支持强度进行分类。表 9-1 列出支持强度的定义及实例。<sup>[2]</sup>

表 9-1 支持强度的定义和实例

**间断性**

支持是一种“个体所需要的基础”。其特点是插曲性，个体并非总是需要支持，或者说在一生的转折期（例如，失业或急性疾病）需要短期支持。提供间断性支持时，强度或高或低。

**有限性**

支持强度的特点是，时间上的连贯性，虽时间有限但不间断，比起更强的支持，所需要的工作人员较少，费用也较低（例如，时间有限的职业培训或从学校到成人社会的转折期的支持）。

**广泛性**

支持的特点是，至少在某种环境（例如，工厂或家庭）中的经常性（例如每天）和没有时间限制（例如，长期支持和长期家庭生活支持）。

**弥漫性**

支持的特点是，经久不变的高强度；各种环境中都提供支持；潜在的生活支持性。比起时间有限性支持或广泛性支持，弥漫性支持一般需要较多的工作人员，更有侵入性。

摘自(AAMR, 1992, p. 26)

现在的定义重申：智能限制是智力落后的必要条件，但不是充分条件。如果存在智力落后，智能限制一定对适应或应付技能有影响。用于划分智力落后的智力测验成绩的阈限水平是，在以平均数为 100、标准差为 15 的量表上的 70~75 分，即大约低于平均数 2 个或 2 个以上标准差。考虑到测量误差，也强调“智力落后”和“正常”之间没有明显的界线。取消各种落后水平以及日益强

调适应技能和环境因素,其目的在于强调个体的独特的长处和不足,以及发展的可能性。对 AAMR 的修订本有些争议。一些批评指责新的规定不精确,不能作出可靠的评定,会增加学校中需要特殊服务的人数比例(Gresham, MacMillan & Siperstein, 1995; MacMillan, Gresham & Siperstein, 1993; Matson, 1995)。也有人反驳说,这些反对缺乏经验基础(Resis, 1995)。总之,采用不同的标准,就可能使得进行评价的方式以及评价本身的结果发生变化。然而,全面评价修订本的效果,也许为时过早。

除了第八章中讨论过的那些个别智力测验之外,评定智力落后个体的程序一般还包括测量日常生活情境中的适应行为。<sup>[13]</sup>设计来评定适应行为的典型量表是“瓦因兰德社会成熟量表(Vineland Social Maturity Scale),由瓦因兰德训练学校校长埃德加·多尔(Edgar Doll, 1935/1965)发表于 30 年代。作为观察智力落后患者之间的差异的结果,多尔创造了一种标准化记录形式,设计来评定个体在寻求他或她的实际需要和担负日常生活中的任务两方面的发展水平。最新的修订本是瓦因兰德适应行为量表(Vineland Adaptive Behavior Scales, VABS——P. L. Harrison, 1985; Sparrow, Balla & Cicchetti, 1984a, 1984b),有 3 种版本,或独立使用或联合使用。两个量表是“面谈版本”(Interview Editions),通过与一位家长或其他监护人的半结构性面谈来获得资料。其中一个 297 个项目的“调查表格”(Survey Form),它同较早的瓦因兰德量表最为相似。另一个是 577 个项目的“扩展表格”(Expanded Form),它也为准备个别化的教育计划或治疗计划提供系统的基础。两种版本都适用于从出生至 18 岁,再到机能低下的成人。第 3 个版本是“课堂版本”(Classroom Edition),包括 244 个项目的问卷,由课堂教师填写,它适用于 3~12 岁的儿童。‘课堂版本’和‘调查表格’分数之间的相关从 0.31 至 0.54,

这说明两者不应该被互换使用。

所有‘瓦因兰德量表’强调儿童通常、习惯上做什么,而不是他能做什么。项目分类为 4 个主要适应领域,表 9-2 说明这 4 个领域及其子领域以及有关行为的简短描述。两个“面谈版本”也包括一组 32 个选做项目,它涉及可能妨碍个体机能的适应不良或不合需要的行为。所有版本提供向家长报告结果的设计良好的表格。

表 9-2 瓦因兰德适应行为量表的内容

| 领域和子领域                         | 描 述                                                |
|--------------------------------|----------------------------------------------------|
| 交 流<br>接受<br>表达<br>书面          | 个体理解什么<br>个体说什么<br>个体读什么,写什么                       |
| 日常生活技能<br>个人<br>家庭<br>社区       | 个体怎样进食、穿衣及搞个人卫生<br>个体做哪些家务<br>个体怎样支配时间、钱钞、电话及职业技能  |
| 社会化<br>人际关系<br>游戏和闲暇时间<br>应付技能 | 个体怎样同他人交往<br>个体怎样游戏,怎样度过闲暇时间<br>个体怎样表现出对他人的责任心和敏感性 |
| 动作技能<br>大动作技能<br>精细动作技能        | 个体怎样使用手臂和腿进行运动和协调<br>个体怎样使用手和手指来操纵物体               |
| 适应行为综合                         | 上面列出的 4 个领域的综合                                     |
| 适应不良行为*                        | 可能妨碍适应性机能的不合需要的行为                                  |

\* 《调查表》和《扩展表》中所包括的选做项目,《课堂版本》中则没有(摘自斯帕罗、巴拉和西切提,1984a,p.3)。

两个“面谈版本”的标准化样本采用全国性取样,共 3000 名被试,年龄分布从出生至 18 岁 11 个月,按照 1980 年美国人口调查,根据性别、种族团体、社区大小、地域、家长教育水平等分层取样。也建立特殊团体的补充常模,包括住宿和不住宿的智力落后成人的样本,以及住院的情绪障碍、视觉受损、听觉受损等儿童样本。“课堂版本”的标准化样本约为 3000 名儿童,年龄分布从 3 岁至 12 岁 11 个月,从 38 个州的学校中抽取,根据其他版本中所采用的同样基础分层取样。

对于 4 个领域分数以及所有 3 种形式上的适应行为合成,提供平均数为 100、标准差为 15 的标准分数。给出 5 个不同置信水平(从 68% 至 99%)的误差范围(根据 SEM)。也可得到百分等级、标准九分、年龄当量、适应水平(即定性描述的类型)等。对于子领域分数,可以把结果表示为适应水平和年级当量;对于适应不良行为,只能把结果报告为定性(适应不良)水平。可以使用补充常模,在每个适当的特殊团体内,找到百分等级和适应水平。可以利用软件,把原始分数转换成导出分数,以及进行剖面图分析。

各种版本的领域分数和合成分数的内部一致性信度系数的中位数大部分高达 0.90。可以理解的是,子领域的信度较低一些,并随年龄水平和内容领域而大不相同。然而,子领域信度的中位数大部分在 0.70 ~ 0.80。对相同的被试施测两个不同的“面谈量表”,再测信度和评分者信度的数据分别表明,短时间间隔的高稳定性和令人满意的一致性。

三种形式的手册中所概括的几类数据说明结构效度。最初就阐述适应行为结构,用以指导编制和选择项目,在某种程度上,这就使量表具有效度。对标准化样本的分析中,以及由独立的研究人员进行的研究中,都提供了经验效度数据。其中包括

领域和子领域成绩的发展趋向；领域和子领域分数的因素分析；对补充常模中所包括的智力落后个体和残疾个体样本的分数剖面图的比较；以及同其他适应行为调查表和 WISC - R、K - ABC、“皮博迪图片词汇测验”等能力测验的相关

一般来说，“瓦因兰德量表”的编制和评价中所使用的程序，技术质量是高的，手册也全面、清楚地阐述了这一点。它们反映了较早版本发表以来所取得的进展。但是，量表的实际效用取决于手册中和所积累的文献资料中所阐述的心理测量学特性的全面知识。<sup>[14]</sup>

以评定结果为基础，为智力落后者设计和选择合适的训练计划，这引起广泛关注。这种关注反过来又导致去编制越来越多的测量适应行为的量表。<sup>[15]</sup>一个实例是 AAMR 编制的“适应行为量表”(Adaptive Behavior Scales, ABS)，其一般目的和“瓦因兰德量表”一样。“AAMR 适应行为量表——住宿和社区，第 2 版”(ABS - RC:2 - Nihira, Leland & Lambert, 1993) 的标准化样本为居住在社区或寄宿环境中的 4000 多名残疾成人，它对 18 个领域进行评分，其中 10 个为应付技能，8 个为社会行为，包括各类适应不良模式。另一方面，“AAMR 适应行为量表——学校，第 2 版”(ABS - S:2 - Lambert, Nihira & Leland, 1993) 的常模样本为智力落后和非落后学生，年龄分布为 3 ~ 18 岁。

在智力落后个体中需要评定的另一个领域是动作发展（婴儿量表中也检查这一点），用于这种目的的原型是“奥西里斯基动作水平测验”(Oseretsky Tests of Motor Proficiency)，这一测验最初是在 1923 年出版于俄罗斯。在测验运动受损、轻微脑机能障碍或学习失能等儿童时，尤其在实施个别训练计划时都可使用奥西里斯基测验。奥西里斯基量表的目前版本是“布鲁宁克斯—奥西里斯基动作水平测验”(Bruininks-Oseretsky Test of Motor

Proficiency, Bruininks, 1978)。这套测验由 46 个项目组成,分为 8 个分测验,需要 45 ~ 60 分钟。得出 3 个分数:大动作合成分数,测量肩膀、躯干、腿等大肌肉的运动;精细动作合成分数,测量手指、双手、前臂等小肌肉的运动;成套测验总分数。还有一个 14 个项目的简短版本,需要 15 ~ 20 分钟,提供一般动作水平的单一指标。可以用基于年龄的标准分数、百分等级、标准九分等表示成绩。也可得到每个分测验的年龄当量。这套测验的标准化样本为 765 名儿童,年龄为 4 岁半 ~ 14 岁半,代表美国总体。合成分数间隔 7 ~ 12 天的再测信度多数为 0.80。研究者以若干种方法来研究效度,包括项目的因素分析、年龄差别、智力落后和学习无能儿童与非智力落后和非学习无能儿童的比较。

对评定智力落后的主要挑战之一是如何区分智力落后和发展迟缓,尤其在婴幼儿期。不仅婴儿期的认知评定的可靠性不如其他年龄,我们已经知道这一点,而且其认知机能受损,可能是其他多种因素导致的(Hodapp, Burack & Zigler, 1990)。负面影响儿童智能成绩和适应技能的主要因素,是感觉受损和运动受损,以及不利的家庭环境。本章的其余部分将讨论有关这两种因素的问题,当然,这两种因素可以单独存在也可以同时并存。

## 对身体残疾个体的测验

测验残疾个体所提出的问题在几十年里受到关注,本世纪 70 年代以来所制定的法律正促进人们努力解决这些问题。前面所引用的残疾人教育法提出要为各种身体残疾者提供合适的教育。首先通过“1973 年改善残疾人生活法案”第 504 款,近来通过“1990 年美国残疾人法案”(ADA - P. L. 101 ~ 336),<sup>[16]</sup>在更

广泛的水平上,把用于其他少数人群的一般公民权条款扩大到身体残疾个体。这些法律禁止歧视的领域为:(1)就业;(2)利用体育设施;(3)学前教育、小学教育和中学教育;(4)中学后教育;(5)健康、福利和社会服务 ADA 加强早期托管,并把这种托管扩大到私立部门中的各组织。

为了从一开始就提供合适的教育,对身体残疾儿童的早期测验尤为重要。这类方法有助于预防累积的学习无能,否则可能扩大残疾对智能发展的影响。<sup>[17]</sup>在任何年龄,测验身体残疾者都会出现关于测验施测和合理解释测验成绩等特殊问题。迄今为止,处理这类测验的主要方法包括:(1)修改测验手段、时限、现有的测验内容等;(2)个别化的临床评定,把测验分数同被试经历、面谈以及教师等日常生活观察者所提供的判断等其他资料相结合(AERA, APA, NCME, 1985,第十三章;Bailey & Wolery, 1989; Barnett, 1983; Culbertson & Willis, 1993; Eyde, Nester, Heaton & Nelson, 1994; Scarpati, 1991; Sherman & Robinson, 1982)。

单独建立特定的身体残疾者的常模,或专为这类团体设计测验,这种努力往往因为可得到的个案太少而止步不前。这种局限性尤其表现在发生率低的多种残疾,以及在经过特殊选择的样本例如研究生和职业学校招生的情况下使用的测验。然而,人们仍在继续研究不同身体残疾者参加标准版或特殊改编版的各种测验上的成绩。

教育测验服务中心(Educational Testing Service)所进行的是最著名的系列研究之一,他们把标准版和非标准版的大学董事会 SAT 和 GRE 一般测验用于 4 类残疾申请人员,即听力残疾、视力残疾、学习无能、肢体残疾等(Willingham 等,1988)。所研究的心理测量特性包括信度、项目区分功能、因素结构以及关于成绩水平和预测力的其他效度指标;也研究测验内容、时限和材

料。总之,研究结果表明,在分数的意义等许多方面,程序改编本可以比得上标准测验。然而,从分数或以前的成绩预测学业成绩,对于残疾个体的准确性就不如其他人。关于测验改编本的某些项目功能及因素结构,也是有疑问的(R. E. Bennett, Rock, & Novatkoski, 1989; Rock, Bennett & Jirele, 1988; Willingham, 1988)。此外,研究发现,非标准版中的时限相对宽松,这个研究结果支持在使用这些版本所获得的分数上打个标记这一引起争论的做法。因此最为重要的是,需要根据经验设定可比较的时限(例如,参见 Wainer, 1993a, pp. 9 ~ 10)。

关于测验残疾个体,还有许多尚未解决的心理测量问题和道德问题。虽然需要更多的研究,但也必须认识到,一些问题也许是无法解决的,因为它们源于一个事实:关于能力、能力障碍、个人性格等的类型和程度,每一个个体都具有一种独特的结构。但是,出于实际的目的,现在比以前更为清醒地认识到残疾个体对测验的需要,以及改进如何对他们进行测验的指导。此外,测试设备中的新进展,例如模拟言语的声音和计算机控制的其他电子设备,有可能提供一大批有希望的测验新方法,其中包括许多在这个领域中有价值的测验方法(参见 Educational Testing Service, 1992; Wilson, 1991)。

下面几节考虑特殊的测验问题和程序,它们有关三类主要的身体残疾,即听觉受损、视觉受损和运动受损。

听觉受损<sup>[18]</sup> 由于听觉受损儿童语言发展普遍落后,言语测验通常对他们不利,即使视觉呈现言语内容也是如此。幼儿耳聋的年龄越小,这种不利条件就越严重。幸运的是,评定听觉机能的最新进展,使我们有可能在出生的头几个月里,准确诊断听觉丧失,并开始治疗程序(Shah & Boyden, 1991)。

某些最早编制的操作测验,例如“平特一佩特森操作量表”



(Pintner-Paterson Performance Scale) 和“阿瑟操作量表”(Arthur Performance Scale), 主要目的就是测验聋儿。韦克斯勒量表的特殊改编本也经常用来测验耳聋被试。如果把口头问题打印在卡片上, 就可以施测大多数言语测验。有人制定了交流操作测验指导语的种种程序(参见 Sattler, 1988, 1992)。事实上, 在美国, WISC—R 操作量表是使用最广的听觉受损儿童智力测验。然而使用这类标准测验程序的改编本, 人们不能假设信度、效度和常模保持不变。由于韦克斯勒量表广泛使用于听觉受损个体, 关于在此类情况下的心理测量特性, 就有大量的文献资料(参见 Braden, 1985; Maller & 1993; Sullivan & Schulte, 1992)。一般来说, 这些研究表明, 操作量表用于听觉受损儿童和听觉正常儿童, 其因素结构以及预测效度和结构效度都十分相似。

至此所提到的测验都根据听觉正常的个体进行标准化。许多研究人员得出结论, 听觉受损个体的操作水平与听力正常的个体相似时, 不需要为前者单独建立常模, 韦克斯勒操作量表就是如此。同时, 根据聋儿所获得的常模, 在一些有关聋儿教育发展的情境中也是有用的。为了满足这种需要, 有人努力建立现行测验的特殊常模, 例如 WISC—R 操作量表用于聋儿的标准化(R. J. Anderson & Sisco, 1977)。

在更基础的水平上, 编制了“希斯基—内布拉斯加学习能力倾向测验”(Hiskey-Nebraska Test of Learning Aptitude, Hiskey, 1966) 根据聋儿和重听儿童进行标准化。这是一个个别测验, 适用年龄为 3~7 岁。由于向聋儿难以传达速度概念, 该测验排除了速度, 也试图使智力机能的取样范围比大多数操作测验大一些。测验中使用手势和实际练习来表达指导语, 也使用有内在兴趣的项目来建立和睦关系。项目的选取特别考虑聋儿的局限性; 项目的最后选择主要根据年龄差异的效标。分别根据

1079 名聋儿和 1074 名听力正常的儿童得出常模；手册中提供了测验两类儿童的平行指导语，手册中也讨论了测验聋儿时所需要的练习。现在“希斯基—内布拉斯加测验”的常模过时了，但它具有合适的信度和效度的证据，仍然被认为是用于听力残疾儿童的最佳测验之一(Sullivan & Burley, 1990)。

在过去的 50 年里，有关耳聋造成的智能结果方面的认识进展是令人惊讶的。布雷登(Braden, 1994)讨论了这段历史中的许多情况，综合考察了 200 多项关于聋人的研究，这包括 17000 余名研究被试。这项工作描述了这方面文献中许多令人感兴趣的研究结果，其中包括父母耳聋的聋儿在操作测验中的得分高于非残疾儿童的常模。虽然这些研究结果的原因尚未完全明了，但毫无疑问，耳聋这个变量比先前所认识到的要复杂得多。而且，失聪的开始和发现的年龄、程度、病因学，以及交往模式、教育状况、父母听力状况和其他残疾的存在，这种因素的相互作用，对听觉受损个体的不同的认知作业也起着相当的作用。

视觉受损<sup>[19]</sup> 测验盲人所碰到的问题与测验聋人大不相同。口头测验最适合于盲人，而操作测验则是最不可能应用的。除了主试常用的口头表达之外，也使用其他的测验技术，例如磁带录音。一些测验，例如“大学入学考试委员会学业评定测验”(College Board Scholastic Assessment Test, SAT)，也使用大型开本或布莱叶盲文。然而，后者的应用性多少有限，因为同油墨相比，用布莱叶盲文所印刷的材料体积越大，阅读盲文的速度就越慢，同时不能流畅阅读布莱叶盲文的盲人数量很多。可以用布莱叶盲文或在键盘上记录测验参加者的反应。对于是非题、多项选择题及其他选择反应的项目，也可使用特制的凹凸答卷纸或卡。当然，在许多个别施测的测验中，能够使用口头反应或手势反应。

一般智力测验被改变用于盲人的最早实例当属比纳量表。根据 1916 年斯坦福—比纳量表,编制了测验盲人的第一个海耶斯—比纳(Hayes-Binet)版。1942 年,根据 1937 年斯坦福—比纳量表,编制了“临时性海耶斯—比纳量表”(Interim Hayes-Binet)<sup>[20]</sup> (Hayes, 1942, 1943)。最新的改编本——与斯坦福—比纳 L-M 型相类似——则是“帕金斯—比纳盲人智力测验”(Perkins-Binet Tests of Intelligence for the Blind)。这个测验根据弱视儿童和盲童分别进行标准化、并有相应的两种型式 (C. J. Davis, 1980)。

韦克斯勒量表也被改编用于盲人测验参加者。这些改编本实质上在于使用言语测验而省去操作测验。一些不适合盲人的项目也被替换。一般来说,对弱视儿童或盲童的研究说明,这种条件对儿童的认知发展产生负面影响,甚至在言语领域也是如此,因为这种条件限制了儿童经验的范围和多样性。视觉受损儿童的韦克斯勒量表剖面图表明,不同研究揭示一种类似的模式。研究结果指出,对于视觉受损儿童和视力正常儿童,测验作业的因素构成不尽相同。虽然不能认为总 IQ 精确测量视觉受损儿童的全部认知机能,但对于见多识广的主试,韦克斯勒量表能够提供有用的诊断信息,了解这些儿童的长处和不足 (Groen-veld & Jan, 1992)。

几乎没有什么量表是专门编制来用于视觉受损个体的。其中最为著名的也许是“盲人学习能力倾向测验”(Blind Learning Aptitude Test, BLAT, Newlan, 1979)。BLAT 是一种个别施测的测验,收编的项目由其他测验例如瑞文渐进推理测验和其他非言语项目改编,呈现时采用凹凸版本。重点放在学习过程而不是过去学习的结果,后者可能不利于盲童。现在其常模资料虽然过时了,但与通常可得到的那些为特殊总体所设计的量表相比较仍然具有一定价值。信度和效度的资料不足,需要进一步的

研究。然而,在评价小学年龄的盲童时,与言语测验一起,BLAT是一个有用的测验。

一个较新的测验仍在编制阶段,一组荷兰研究人员试图综合测量视觉受损儿童智力。“视力受损儿童智力测验”(Intelligence Test for Visually Impaired Children,ITVIC;Dekker,Drenth,Zaal & Koole,1990),把触觉形式的积木图案之类作业编入一个成套测验,使其包括若干非言语分测验和言语分测验。<sup>[21]</sup>需要对ITVIC的一系列问题作进一步的研究,但是,初步的研究结果表明,量表效度是有价值的(Dekker,1993;Dekker,Drenth & Zaal,1991;Dekker & Koole,1992)。

如同本章中所讨论的其他各种特殊条件,视觉受损有一系列等级,又往往和其他问题结合在一起。因此,究竟使用标准测验或使用标准测验改编本,抑或使用专为盲人设计的测验,这取决于评定的目标和所研究个体的独特性。一般来说,测验使用者应当记住,对测验作修改,例如视觉图形的触觉呈现或延长时限,我们就不能假设,该测验测量与当初版本同样的结构。

**运动受损<sup>[22]</sup>** 运动受损个体通常能够接受听觉和视觉的刺激,但是他们可能具有严重的运动障碍,不能作出口头或书面反应。试图使用形状板或其他操作材料,同样碰到困难。在限定时间或陌生的环境下作业,往往增加运动受损者的运动障碍。他们容易疲劳,有必要安排短时测验。

最严重的运动受损者为大脑瘫痪者。对运动受损个案的检查通常使用普通的智力测验,例如斯坦福一比纳量表。在这类研究中,最严重的运动受损者因为不能测验通常被排除在外。对测验程序通常也作非正式的调整,使测验适合于儿童的反应能力。当然,这两种做法都是权宜之计。

令人较为满意的方法是,编制出甚至适合于最严重运动受

损个体的测验工具。出于这种目的,现在有一些专门设计的测验或现行测验的改编本,虽然它们的常模资料和效度资料往往不足。下一节所要讨论的若干测验,最初是用于跨文化测验,也证明适用于运动残疾个体。现在编制了“利特国际操作量表”(Leiter International Performance Scale)和“波蒂厄斯迷津”(Porteus Mazes)的改编本,适用于测验大脑瘫痪的儿童(Allen & Collens, 1955; Arnold, 1951)。在这两个改编测验中,主试操作测验材料,而测验参加者只需以适当的头部运动作答。瑞文渐进推理测验也证明是用于这种目的的一种有用工具。由于该测验没有时限,也由于作答可以口头、书面、或手指或点头,看来它尤其适合于运动受损个体。该测验作答灵活简单,有各种形式,包括一系列难度,提供相当高的测验上限。据报告,在研究大脑瘫痪个体和其他运动受损者中,已成功地使用了这个测验(参见 Capitani, Sala & Marchitti, 1994)。

另一类准许使用简单的手指反应的测验则是图片词汇测验。这些测验快速测量“使用”词汇,这就使它们尤其适用于不能完全发声的个体(例如大脑瘫痪者)和聋人。由于图片词汇测验容易施测,大约 15 分钟就能完成,在不能施测综合的个别测验的情况下,图片词汇测验也是一种十分有用的快速筛选工具。

“皮博迪图片词汇测验”(Peabody Picture Vocabulary Test)是一个典型的图片词汇测验。它的现行版本(PPVT - R; Dunn & Dunn, 1981)由 175 张卡片组成,每张包括 4 幅图形。呈现每张卡片时,主试口头提供一个刺激词汇,要求测验参加者指出或以某种其他方式表示最能说明刺激词汇意义的图形。整个测验适用的范围从学前到成人水平,每一个体只需做适合于他自己作业水平的项目,这由规定的成功点和失败点所决定。原始分数能够转换成标准分数(平均数为 100、标准差为 15)、百分等级、

标准九分等。这些导出分数绘制在一张图上,置信区间为 $\pm 1\text{SEM}$ (测量标准误差)也提供年龄当量分数。PPVT-R 没有规定时限,但一般需要 10~20 分钟。它有两种平行型式,分别使用一组不同的卡片和不同的刺激词汇。

PPVT-R 的全国标准化样本,包括 4200 名 2 岁半~18 岁的儿童和青年,以及 828 名 19~40 岁的成人。测验的心理测量特性十分令人满意(关于评述,参见 McCallum, 1985; Wiig, 1985)。内部一致性、复本、再测等信度系数分布从中等至较高。PPVT-R 的许多效度证据,建立在对 PPVT 的坚实的研究基础之上,PPVT-R 与 PPVT 的相关中位数为 0.70。对 300 多项使用 PPVT 的研究所作的全面考察得出结论,PPVT 与其他词汇测验为高相关,与言语智力和学业能力倾向测验为中等相关,与教育成就测验上的成绩为有价值的相关。在各种总体中得出的相关十分相似,这包括经济条件差的团体,以及残疾和智力落后者的样本。PPVT 上的分数部分反映了被试的文化同化和接触标准美国英语的程度。

使用 PPVT-R 本身的研究表明,它也与其他言语理解测验高相关(参见 Elliott, 1990b, p. 235)。一项对 PPVT-R 的研究特别令人感兴趣,它使用词汇获得顺序的结构方程模型,证实了量表的结构效度(L. T. Miller & Lee, 1993)。第 3 版皮博迪图片词汇测验“PPVT-III”,与一个表达词汇测验一起制定常模,将于 1997 年出版。

图片分类测验中也采用类似的测验施测程序,例如‘哥伦比亚心理成熟量表’(Columbia Mental Maturity Scale, CMMS; Burge-meister, Blum & Lorge, 1972)。最初编制该量表是用来测量大脑瘫痪的儿童,由 92 个项目构成,每个项目为一组 3 或 4 或 5 张彩色图形,印制在一张大卡片上。要求被试找出与其他图形不

相属的那张图形,用指点或点头表示他的选择(参见图 9-3)。CMMS 的标准化样本由 2600 名 3 岁 6 个月至 9 岁 11 个月的儿童组成,代表 1960 年美国同龄总体。分半信度和再测信度为 0.84~0.91。一组 52 名学前和一年级儿童中,与斯坦福一比纳 IQ 的相关为 0.67。一、二年级样本中,与成就测验分数的相关大多数高于 0.40 而低于 0.60。CMMS 早期和现行测验型式对于种种残疾团体的适用性和效度的研究已积累了广泛资料(参见已出版的第 2、3、4 版测验)。然而,由于常模过时,加上所评定的能力范围狭小,CMMS 的适用性十分有限。

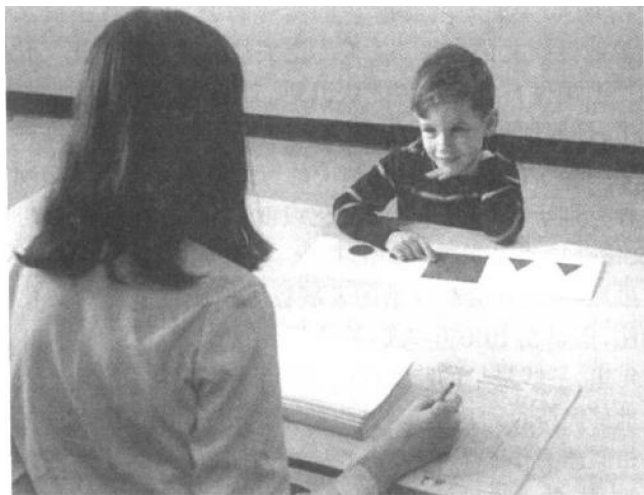


图 9-3 主试对儿童施测哥伦比亚心理成熟量表(摘自哥伦比亚心理成熟量表 施测和解释的指南,伯奇迈斯特等,1972,p. 11)

## 多重文化测验

**问题** 自本世纪 50 年代以来,测验文化背景十分不同的被试日益受到关注。为了最大限度地利用世界上许多新兴发展中国家人力资源,人们需要测验。这些国家中教育设施迅速发展,招生及个体咨询都需要进行测验。随着社会日益工业化,对测验的需求也大为增加,以便有助于人员的选拔和安置,尤其在信息加工、机械、专门职业等领域。

在美国,多重文化测验<sup>[23]</sup>的实际问题,主要与主流文化之内的少数民族文化有关。人们广泛关注各种测验对于文化劣势个体的适用性。应该清楚地认识到,文化劣势是一个相对概念。客观上,任何两种文化或亚文化之间只有文化差异。每种文化都强化与其价值观和要求相适应的行为发展。当个体必须适应于一种有别于养育他或她的文化或亚文化,并在其中参与竞争时,文化差异就可能成为文化劣势,当然,也可能是文化优势。

虽然,最近社会和政治的发展极大地激发了对跨文化测验的关注,但早在 1910 年人们就认识到这个问题了。在 20 世纪头十年里,一些最早的跨文化测验的编制,是为了测验迁入美国的大批移民(Knox, 1914)。另一些则源于对相对孤立的文化团体的相似能力作基本研究。这些文化往往同编制了大多数心理测验的<sup>[24]</sup>、技术先进的西方社会很少或没有接触。

传统上,跨文化测验试图排除各种文化不同的一个或多个参数。其中一个众所周知的参数是语言。如果被测验的各文化团体使用不同的语言,就得编制主试和测验参加者都不需要语言的测验。当教育背景大不相同或主要为文盲时,则排除阅读。



这些测验中并不排除口头语言，因为它们被设计来适用于讲一种共同语言的个体。文化或亚文化有所不同的另一个参数是速度。不仅日常生活的速度，而且匆忙行事的动机和对快速操作的价值观，在各国文化之间、在一国之内的各少数民族之间，以及在城乡亚文化之间都大不相同（参见 Klineberg, 1928; R. R. Knapp, 1960; M. Womer, 1972）。因此，跨文化测验往往——虽然并非总是——准许较长的时限，且对较快的作业不加分而试图排除速度的影响。

各文化不同的另一个参数是测验内容。例如，若干非语言测验和非阅读测验，需要某种文化特有的常识项目。因此，它们要求测验参加者理解诸如小提琴、邮票、手枪、小刀、电话、钢琴、镜子之类物体的功能。在某种文化中长大的个体也许缺少答对这类项目的经验背景。最初编制典型的“超文化”（culture-free）测验，主要正是为了控制这类文化参数。下面简短讨论一下设计来消除一个或多个上述参数的典型测验，然后我们将分析跨文化测验的不同方法。

典型的传统量表<sup>[25]</sup> 在努力编制适用于跨文化的测验时，心理测量学家使用种种程序，本节对其中一部分加以讨论。“利特国际操作量表——修订版”（Roid & Miller, 1997）是一个个别施测的操作测验，最初发表于 1940 年，是通过对夏威夷不同民族团体的数年试用后正式编制的。后来，波蒂厄斯（Porteus）将该量表应用于若干非洲团体，其他研究人员则把它应用于另外一些国家的团体。1948 年发表的修订本是根据对第二次世界大战期间美国的儿童、中学生及军队新兵的进一步测验。1997 年版本则根据 2000 多名 2~20 岁的正常的和不正常的个体样本。利特量表与众不同的一个特点是几乎完全消除言语指导语，这一特点自此也被其他量表所采用。每个测验开始时都是

一项十分容易的作业，其类型与整个测验中要碰到的相同。它属个别施测且没有时限，对作业的理解也作为测验的一部分。所有图形刺激材料都呈现在画架上，测验参加者的反应是选择最合适的图片，并放置在反应盘上，如图 9-4 所示。

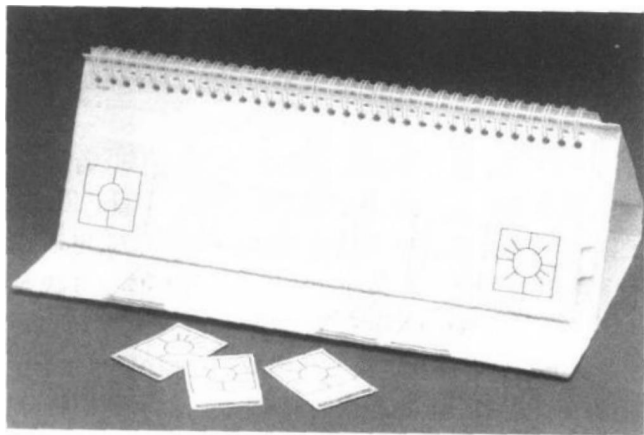


图 9-4 利特国际操作量表修订版所用典型材料。  
所列举的部分完成的作业属于排顺序测验，要求被试从 6 张反应图片中选出 5 张，并正确放置在反应盘上。

利特量表包括一系列类似于言语量表中的机能。在它的目前型式中，所包括的内容明显扩展，现在包括 4 个领域，即推理、想象、注意和记忆。在推理和想象领域，在各种年龄中所包括的作业为图案类同、形式完成、匹配、排顺序（如图 9-4 所示）、分类、折纸、图形旋转等。注意和记忆领域测验包括持久注意和分配注意度量，以及多种瞬时记忆和延迟记忆作业。可以预期，修订版利特量表是相当新的，心理测量特性比早期版本更先进。

例如，使用项目反应理论来校准难度水平，不再以比例智商来表示分数。此外，有代表性的当代常模和广泛的内容，也应该提高量表的效用。新的手册包括各类信度和效度的资料。

“瑞文渐进推理测验”(Raven's Progressive Matrices, RPM)主要用来测量斯皮尔曼的  $g$  因素即一般智力(J. Raven, 1983; Raven, Raven & Court, 1995)。和斯皮尔曼对  $g$  的理论分析相一致，该测验主要测量抽象项目之间的关系推理。各项目都由一组矩阵图案组成，即图案元素按行和列排列，每个项目中都缺少一部分。任务是从给出的选项中选出缺少的图案。较容易的项目需要准确鉴别力；较难的项目则涉及图形的类推、排列和替换，以及其他的逻辑关系。图 9-5 列举“标准渐进推理测验”(Standard Progressive Matrices) 中的两个典型的项目。实施测验通常没有时限，既可以个别施测，也可以团体施测，只需要十分简单的口头指导语。

RPM 有三种形式，难度水平各不相同。“标准渐进推理测

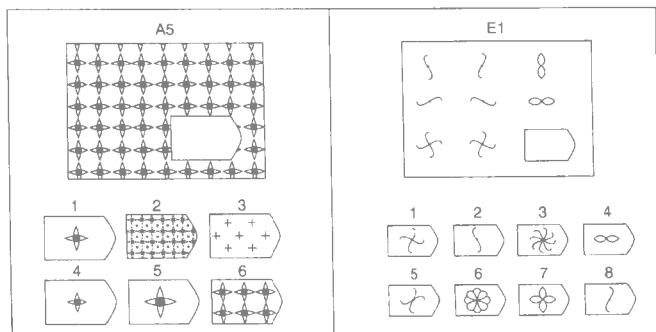


图 9-5 标准渐进推理测验(SPM)的典型项目。一个较易的项目(A<sub>5</sub>)和一个较难的项目(E<sub>1</sub>)

验”(SPM, 1996 年版)适用于年龄 6~80 岁的正常个体。较容易的型式为“彩色渐进推理测验”(Coloured Progressive Matrices, CPM, 1990 年版),适用于年幼儿童及由于种种原因而不能用 SPM 合适测验的特殊团体。CPM 的常模适用于 5 岁半~11 岁的儿童,以及智力落后成人和 60~89 岁非智力落后个体。第三种形式为“高级渐进推理测验”(Advanced Progressive Matrices, APM, 1994 年版),适用于正常智力之上的青少年和成人。

RPM 所有水平的手册分几部分,可以单本购买,也可以任意选购几本,几本手册可以合订成一册。第一本手册为总述,最新版为 1995 年版其他手册出版日期为 1990 年至 1996 年,分别为 RPM 三种水平的特定手册。与 RPM 同时使用的两个标准化词汇测验的手册,也是瑞文测验包的一部分。最后一本手册总结了在各个国家和特殊团体上得出的信度、效度和补充常模的研究(Court & Raven, 1995),还包括使用英国标准化资料的若干补充研究和在北美、以色列、德国所收集的常模资料,以及 2000 多项研究所注释的文献目录。<sup>[26]</sup>

虽然发表了大量关于 RPM 的研究,但这些研究是分散的、不同一的,而且目的各异。测验编制者建议,可能的使用者要找出与他们自己的兴趣最有关系的那些研究和总体;但是他们告诫,各种研究在方法、样本大小和技术质量上都各不相同。

一般来说,年龄大致相同的成人和大年龄儿童团体中的再测信度大约为 0.70~0.90。然而,在较低分数的范围,信度就大大低于这些数值。内在一致性系数大多为 0.08~0.90。与言语智力测验和操作智力测验的相关为 0.40~0.75,与操作测验的相关往往高于言语测验。对智力落后个体以及对不同的职业和教育的团体的研究,表明有相当的同时效度对学业效标的预测效度系数稍稍低于通常的言语智力测验。一些因素分析指

出, 渐进推理测验在多数智力测量共同的因素(被许多心理学家鉴定为斯皮尔曼的  $g$ ) 上负荷高, 但是空间能力倾向、归纳推理、知觉精确性和其他群因素也影响成绩。

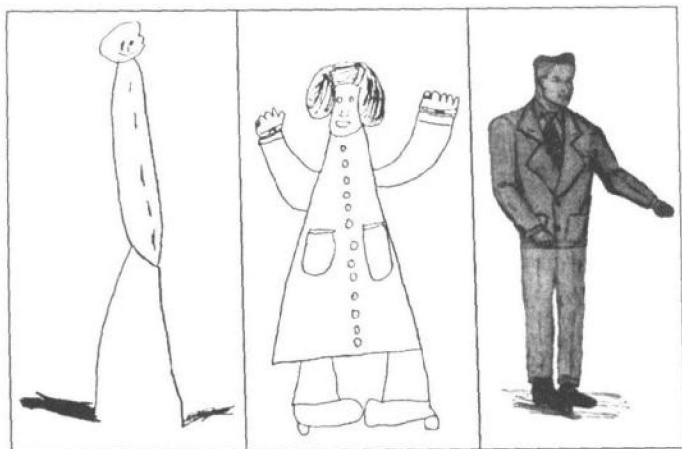
“古迪纳夫画男人测验”(Goodenough Draw-a-Man Test) 说明非言语测验的另一种不同的方法, 其间只要求测验参加者“画一个男人, 尽可能画得最好”。该测验最初标准化在 1926 年, 一直使用到 1963 年仍未作修改。称为“古迪纳夫—哈里斯画人测验”(Goodenough-Harris Drawing Test) 的扩展和修订本于 1963 年发表(D. B. Harris, 1963)。同当初的测验一样, 修订本强调儿童观察的精确性和概念思维的发展, 而不是绘画技能。记分包括身体各个部分、服饰的细节、身体各个部分的比例、透视以及类似特点。根据年龄区别、与测验总分的关系以及与团体智力测验分数的关系, 合计选取 73 个记分项目。

在修订本量表中, 也要求被试画一个女人和他自己。“女人量表”的评分方式类似于“男人量表”。“自我量表”编制为人格投射测验, 虽然从这种应用中得到的结果不容乐观。<sup>[27]</sup>每个量表上的得分都转换成平均数为 100、标准差为 15 的标准分数。图 9-6 列举 3 幅图画以及对应的原始分数和标准分数, 绘画儿童的年龄分别为 5 岁 8 个月、8 岁 8 个月和 12 岁 11 个月。人们使用种种程序重复研究了“古迪纳夫—哈里斯画人测验”的信度, 再测相关以及分半信度和评分者信度的数据都不错, 学校中美术训练对分数的影响显然可以忽略不计(J. A. Dunn, 1967; D. B. Harris, 1963)。

除了量表编制中所收集的项目分析资料之外, 与其他智力测验的相关提供了测验的结构效度的资料。这些相关大不相同, 但大多数在 0.50 以上。幼儿园儿童的“画男人测验”与数字能力倾向的相关高于四年级儿童, 幼儿园儿童的“画男人测验”

与知觉的速度和准确性的相关则低于四年级儿童 (D. B. Harris, 1963)。这些研究结果说明, 在不同年龄, 测验也许测量稍有不同的机能。关于不同的文化和种族团体的大量研究都使用这两种版本。这类研究表明, 该测验上的成绩对文化背景差异的依赖比当初假设的要大。例如, 丹尼斯(Dennis, 1966)分析了在 40 种截然不同的文化团体中使用该测验所得到的比较资料, 并发现, 团体平均分数看来与每种文化之内表现艺术的经验总量的关系最为密切。

一项设计良好的、使用古迪纳夫-哈里斯画人测验对墨西哥和美国儿童的比较研究, 再次揭示经验背景中的文化差异



男人 原始分数 7

实足年龄 5~8

标准分数 73

女人 原始分数 31

实足年龄 8~8

标准分数 103

男人 原始分数 66

实足年龄 12~11

标准分数 134

图 9-6 古迪纳夫-哈里斯画人测验实得的样图

(Laosa, Swartz & Diaz-Guerrero, 1974)。最近在伊朗进行的对该测验的一项大规模研究中, 6 ~ 13 岁儿童的平均分数都稍低于美国常模, 但分数显示出良好的年龄区别以及与社会经济状况和学业成就度量的正相关 (Mehryar, Tashakkori, Yousefi & Khajavi, 1987)。还应当指出一点, 关于“古迪纳夫—哈里斯画人测验”的这些研究结果, 也代表了最初设计为“超文化”(culture-free)或“文化公平”(culture-fair)的各种测验所获得的结果 (Samuda, 1975, 第六章)。

现在有“画男人测验”的新版本, 称为“画人定量评分系统”(Draw A Person: A Quantitative Scoring System, DAP; Naglieri, 1988)。打算更新和改进古迪纳夫—哈里斯版本的技术质量。DAP 提供更新、更详尽的常模, 但施测程序多少有点不同, 评分系统也经过修订, 试图减少“古迪纳夫—哈里斯画人测验”中的模棱两可。此外, 它包括黑人和西班牙儿童样本的常模资料。然而, 尽管有所改进, 有人仍批评该版本的范围相对有限, 以及没有证据表明所修订的评分系统有优点 (Cosden, 1992)。

关于本节所讨论的量表, 应当重申最后一点。虽然当初编制为跨文化测验, 但其中一些测验倒主要为临床心理学家和咨询心理学家所应用, 他们用来补充从斯坦福—比纳量表和韦克斯勒量表之类工具上所获得的资料, 在测验各种残疾个体中, 又作为一种资料来源。这种认识导致了新一代此类工具的编制。其中之一是“非言语智力测验”(Test of Nonverbal Intelligence), 已经出第 2 版 (TONI-2; L. Brown, Sherbenou & Johnsen, 1990), 它的内容和适用范围都类似于 RPM (关于 TONI-2 的评述, 参见 K. R. Murphy, 1992 & Watson, 1992)。这种类型的其他主要新量表目前正在标准化, 将于本世纪 90 年代后期出版。<sup>[28]</sup>

跨文化测验的方法 编制适用于在不同文化或亚文化中长

大的个体的测验,在理论上,我们能够找出三种方法,尽管在实践上,这三种方法也许兼有某些特点。第一类方法是,选择对多种不同文化共同的项目,所得测验则以多种不同文化中的当地效标来建立效度。这是编制文化公平测验的基本方法,虽然在不同文化中所重复的测验效度往往被统统忽略或实施不当。然而,没有这一步,我们就不能确信,测验相对不受文化限制元素的影响。而且,不可能设计出单一的测验而完全满足多种文化的要求。

然而,关于一些重大问题的基础研究则需要多重文化评定技术。这些问题之一是单一文化之内得出的心理学原理和结构的普遍性(Anastasi, 1958, 第 18 章;Berry 等, 1992; Irvine, 1983; Irvine & Carroll, 1980)。另一个问题是环境条件在发展个体行为差异中的作用——在截然不同的文化所提供的广泛的环境变化范围之内,能够更为有效地研究这个问题。这类研究所要求的量表,至少能够在不同文化中相对类似的条件下施测。在适当的实验设计中以及在研究人员对所研究的文化或亚文化完全熟悉的情况下,才能防止对使用这类量表所获得的结果作出不正确的解释。所要求的事情就是鉴定某种文化之内特定的经验变量,它们可能与那种文化特有的行为发展中社会性显著差异有关(J. W. Berry, 1983; Brislin, 1993; Segall, 1983; Whiting, 1976)。在发表的“跨文化焦虑系列”中,能够发现人格测验领域中努力进行这类研究的一个著名实例。这个系列全部致力于测量跨文化的焦虑,在阐述焦虑结构及全世界不同地域的人怎样体验焦虑方面,特别富有成效(参见 Spielberger & Diaz-Guerrero, 1990)。

第二种方法是,在一种文化之内编制测验,把它施测于不同文化背景的个体。在这种情况下,我们应该避免一种错误,即把单一文化背景之内所编制的测验当做测量‘智力’或其他结构的



通用标准。我们也不应该假设,在这种测验上的一个低分,对于测验一种文化中的成员和另一种文化中的成员,具有同样的因果解释。这种方法所能够确定的是,团体之间的文化距离,以及个体的文化适应程度和他对于特定文化的教育活动和职业活动的准备。一些研究人员试图突出一个事实:养育个体的文化环境影响他所获得的认知技能和知识。早期实例包括根据澳大利亚土著居民进行标准化的“脚印识别测验”(Porteus, 1931)和根据印第安儿童进行标准化的“画马测验”(DuBois, 1939)。在两种情况下,编制测验所根据的文化团体在与其他团体的比较中都占据优势。

第三种方法是,在每种文化之内编制不同的测验(或对现行测验作重大修改),以当地效标建立效度,且仅仅使用于适当的文化之内。在特定的文化之内,编制工业人员和军事人员的测验正是这种方法的具体运用。由美国国际发展局(United States Agency for International Development)发起,在非洲、亚洲和拉丁美洲若干发展中国家所进行的测验发展计划,提供了这样的实例(Schwarz & Krug, 1972)。在这种情况下,以测验所要预测的具体教育效标和职业效标来建立效度,且按照当地常模来评价成绩。每个测验仅仅与编制该测验所在地的文化相适应,不试图进行跨文化比较。然而,如果所要预测的效标是技术性的,那倒可能需要“西方型的智力”,测验也将反映特定文化发展的方向,而不是自身盛行的文化特性。此外,正如对全面使用测验的最新一次调查所指出当前的现实是,整体来看,使用最多的——至少用于儿童和青年——是在美国和欧洲所编制的那些测验。事实上,在参加调查的各类国家中,最不发达的国家也许最不同于美国和西方国家,却也是最依赖外国测验技术的国家(Hu & Oakland, 1991; Oakland & Hu, 1992)。<sup>[29]</sup>

在诸如美国、以色列、荷兰之类多元化社会之内,关于少数民族文化的心理测验,积累了大量文献资料(参见 Bleichrodt & Drenth, 1991; Duran, 1989; Figueroa, 1990; Hessel & Hamers, 1993; Samuda, Kong, Cummins, Lewis & Pascual-Leone, 1991; Zeidner, 1988)。在本书中,分别在相应的章节中对此展开讨论。第十八章讨论对少数民族文化使用测验中的社会道德问题和责任。第六章和第七章讨论了测验偏向和项目一团体相互作用的心理测量技术问题。本章讨论为跨文化能力测验所编制的量表。第十二章讨论对跨文化测验结果的解释以及当前趋向等问题。

今天,多重文化测验不再编制特殊的测验,倒是越来越重视测验过程中主试的作用。主试的基本职责是:(1)获得被试文化背景的资料;(2)选择最适合于测验目的的测验;(3)给特定个体有效地呈现和实施测验;(4)解释测验结果时,要根据个体的经验背景,以及所要评价的被试的各种具体条件(职业、教育、社会等)。第十二章将进一步讨论主试作用的这些特征。

**环境的评定** 虽然讨论传统的跨文化测验具有历史价值,由此能加深理解现行测验的起源和性质,但世界上各种文化之间的接触日益加快,从根本上改变了对这类测验的需要。在特定的文化之内,出于具体的目的,例如教育、工业、咨询等,会编制或改编越来越多行之有效的测验。追求一个通用的人类智力测验被认为是徒劳无益的,因为人们越来越意识到个体所处的环境和经历对其完成智力作业所具有的重要作用。这种情况导致人们愈来愈重视对个体所处环境的评价。<sup>[30]</sup>

传统的环境评定的方法,依靠某些社会经济水平的总体的综合指标。社会学家使用复杂的程序来鉴定个体作为社会阶层成员的资格(Warner, Meeker & Eells, 1949)。然而,简便易行的指标证明为同样有效,它与复杂方法所得结果十分一致。事实上,

从家庭主要供养者的职业中,就能够得出社会经济水平的合理的近似值。已经编制了若干粗略的量表,对父母的职业水平分类;一些量表把父母的职业资料和教育水平相结合,例如广泛使用的“社会地位的二因素指标”。这个指标首先由霍林斯黑德(Hollingshead,1957)描述,在几个资料中都能够找到(例如,Bon-jean,Hill & McLemore,1967;Hopkins & Stanley,1981)。也有人研制了记录职业资料 and 得出职业水平指标的更客观的程序(Duncan,1961;Stricker,1985)。

传统的总体指标的主要局限是,它们按照好坏或高低这个单一连续体对环境进行分类。然而,环境所强化的特定行为有所不同,因而对具体的个体特性的影响也有所不同(参见 McAndrew,1993)。因此,对于发展运动技能、学业成就、创造力、社会适应等,最理想的环境也许大不相同。在罗杰·巴克(Roger Barker)关于生态心理学的开拓性著作修订版中,能够找到关于对行为环境作描述和分类的一种经验方法的有价值的指导性意见(Schoggen,1989)。

跨文化测验强调父母教养和家庭环境在儿童的智力发展中所起的作用(参见 M. H. Bornstein,1991)。现在人们也认识到,这类环境差异并非仅仅存在于一些可清楚鉴别的文化或种族总体,从而可以显著地影响任何个体的心理发展。而且,需要根据环境所鼓励的特定行为,更为具体地鉴别环境。现在人们日益关注更为准确地评价不同的家庭及其环境对心理的影响。

现在有多种量表和技术,用来评价家庭及其环境(Bradley & Brisby,1993;Paget,1991)。一个著名的、广泛使用的家庭环境调查表为“环境测量的家庭观察”(Home Observation for Measurement of the Environment,HOME;B. M. Caldwell & Bradley,1984)。这个量表集中于家庭环境中促进认知发展的刺激和父母行为的类型

(Bradley & Caldwell, 1984; B. M. Caldwell & Bradley, 1978; J. H. Stevens & Bakeman, 1985)。HOME调查量表目前有 3 个版本,分别适用于出生至 3 岁、3 岁至 6 岁、6 岁至 10 岁。HOME 提供几个量表上的分数,评价诸如提供合适的游戏材料、刺激多样性、鼓励社会性成熟及刺激学业行为之类变量(关于评述,参见 Boehm, 1985)。婴儿的家庭社会经济地位(SES)的指标与儿童早期的智力成绩的相关,等于或优于 HOME 分数与儿童早期的智力成绩的相关。然而,在某种情况下,SES 和 HOME 分数相结合,能够提高智力预测性(参见 D. L. Johnson 等,1993)。而且,HOME 和类似的量表所评价的变量,能够为出于许多其他目的评价儿童而提供独特的、有价值的资料。

## 注 释

- [1] 关于本章中所列举的许多类型的测验,进一步的资料、评价及参考书目,参见萨特勒(Sattler, 1988,第十二、十四和十五章)的有关论述。
- [2] 关于最初格塞尔量表的修订版的最新手册,参见诺布洛奇、史蒂文斯和马隆(Knobloch, Stevens & Malone, 1980)的有关论述。还有一些其他测验在标题中也使用格塞尔这个冠名,但其中没有一个测验包括婴儿期(参见 TIP-IV 现有测验一览表及第 9 版 MMY 对其中一些测验的评述)。
- [3] 关于学龄前儿童评定的简明扼要的历史,可以参见凯利和瑟贝克(M. F. Kelley & Surbeck, 1991)的有关论述。关于婴儿和学龄前测验的其他有关资料,参见艾尔沃德(Aylward, 1994)、布拉肯(Bracken, 1991b)、卡伯特森和威尔斯(Culbertson & Wills, 1993)、坎弗斯(Kamphaus, 1993)、雷诺兹和坎弗斯(C. R. Reynolds & Kamphaus, 1990a),以及瓦兹奎兹-纽塔尔、罗梅罗和卡尔斯尼克(Vazquez Nutall, Romero & Kalesnik, 1992)的有关论述。
- [4] 另一个测验是伍德科克-约翰逊(Woodcock-Johnson)心理教育成套测

验—修订版：认知能力测验 (Woodcock & Johnson, 1989, 1990)。

- [5] 关于婴儿智力测验的用途和局限性的讨论，参见古德曼 (Goodman, 1990) 的有关论述。关于最初贝利量表的心理测量学特性的一系列论文，参见罗维—科利尔和利普西特 (Rovee-Collier & Lipsitt, 1992) 的有关论述。
- [6] “schemata”是“schema”的复数，这个词在皮亚杰著作中经常碰到，它实际上表示一种框架，而个体使感觉资料符合于这种框架。
- [7] 解释所报告的顺序性指标时，应该记住一个事实：测量顺序性的程序及把难度次序分析应用于皮亚杰量表，都是引起争论的问题（参见 F. H. Hooper, 1973; A. C. Rosenthal, 1985）。
- [8] 量表在临床评定中的潜在价值也被公认；认识这种潜力的一个重要步骤，则是发表了专为临床和教育上使用量表而设计的手册和评分表 (Dunst, 1980)。
- [9] 照字面意义“不符合”理论上所预期的反应模式。
- [10] 本章的后面一节讨论对环境的评定。
- [11] 主要是 1986 年和 1990 年分别通过的 P. L. 99 ~ 457 和 P. L. 101 ~ 476。关于测验和评定儿童的联邦立法和法庭案例的涵义的讨论，参见艾尔斯、戴和罗塔特里 (Ayers, Day & Rotatori, 1990)，德默斯、菲奥雷洛和兰格 (DeMers, Fiorelle, & Langer, 1992)，凯利和梅尔顿 (M. P. Kelly & Melton, 1993)，以及萨特勒 (1998, pp. 767 ~ 784) 的有关论述。
- [12] 美国精神病学协会目前的《心理障碍的诊断和统计手册》(DSM-IV—1994)，仍然根据 IQ 水平规定智力落后的 4 种严重程度，即轻度、中度、重度、极度。
- [13] 应当指出，所有主要的个别智力量表应用于智力落后个体时，效果并不是同样理想 (参见 Spruill, 1991)。
- [14] 例如，参见米德尔顿、基恩和布朗 (Middleton, Keene & Brown, 1990)，波恩和巴尼特 (Poth & Barnett, 1988)，拉乔和马辛盖尔 (Raggio & Massingale, 1990)，沙茨和汗旦—艾伦 (Schatz & Hamdan-Allen, 1995)，以及西尔弗斯坦 (Silverstein, 1986) 的有关论述。关于瓦因兰德量表三种形式的进一步描述和独立评价，参见坎贝尔 (I. A. Campbell, 1985) 和雷诺兹

(1986)的有关论述。

- [15] 关于对其中许多量表的评述, 参见福克斯和迈耶( Fox & Meyer, 1990), 诺夫( Knoff, 1992)以及萨特勒( 1988 第十五章) 的有关论述。
- [16] 关于 ADA 的心理测验涵义的评论, 参见内斯特( Nester, 1994) 的有关论述。美国心理学会评价、测量及统计分会的一份声明中, 全面讨论这项法律所提出的心理测量和评定问题, 文章发表于该分会的通讯杂志 1993 年 1 月号《评分》。
- [17] 关于残疾幼儿的发展模式及其评定程序的进一步的资料, 参见瓦尔斯和希恩( Wachs & Sheehan, 1988) 的有关论述。
- [18] 关于评定听力残疾儿童的问题和讨论的评述, 参见布拉德利-约翰逊和埃文斯( Bradley-Johnson & Evans, 1991) , 马伦 Y. Mullen, 1992) 以及沙利文和伯利( Sullivan & Burley, 1990) 的有关论述。
- [19] 关于评定视力残疾儿童的问题和程序的评述, 参见布拉德利-约翰逊( 1994) , 菲韦尔( Fewell, 1991) , 穆尔和麦克劳克林( M. S. Moore & McLaughlin, 1992) 以及奥兰斯基( Orlansky, 1988) 的有关论述。
- [20] 最初称之为临时性版本, 因为量表标准化是试验性的, 现在该版本在文献中已被冠以这个名称。
- [21] 触觉智力量表类似于本世纪 50 年代和 60 年代早期根据盲人成人所编制和建立常模的测验( Shurrager & Shurrager, 1964) 。它由 6 个操作分测验组成, 仿照韦克斯勒-贝勒维量表, 即数字符号、物体拼配、积木图案、物体补缺、形状板、小珠子算术。
- [22] 一些测验在评定幼儿大动作机能中十分有用, 对它们的评述, 参见威廉斯( H. G. Williams, 1991) 的有关论述。鲁宾逊和菲伯( C. Robinson & Fieber, 1988) 描述了一种强调过程的方法, 它使用皮亚杰感觉运动和前运算作业来评定幼儿。
- [23] “多重文化测验”也被广泛地称为跨文化测验或超文化测验。
- [24] 关于早期测验的例子, 参见安娜斯塔西( Anastasi, 1954, 第十章) 的有关论述。
- [25] 对本节和前几节所讨论的若干非言语测验以及其他这类测验的评述, 参见纳格利里和普鲁伊特( Naglieri & Prewett, 1990) 的有关论述。

- 
- [26] 在 RPM 出版商登录的地址,从考特(J. H. Court)处的硬盘复制和软盘中,都可以获得后一种资料。
- [27] 第十五章中讨论使用画人的其他投射方法。其中有伊丽沙白·科皮茨(Elizabeth Koppitz)对这项技术的研究工作,包括认知和情绪两方面。
- [28] 例如布拉肯和麦克勒姆(B. A. Bracken & R. S. McCallum)编制的“通用非言语智力测验”。
- [29] 认识到这种现实,国际测验委员会精心制订了一组改编教育和心理测验的有关准则(Hambleton, 1994, 1996; Van de Vijver & Hambleton, 1996)。盖辛格(Geisinger, 1994)提出了许多类似的问题。
- [30] 第十二章进一步讨论这一点。

## 第十章 团体测验

斯坦福一比纳量表和韦克斯勒量表之类的个别测验，主要应用于临床，而团体测验则主要应用于教育、政府、工业、军事等部门。回忆一下，第一次世界大战期间，美国军队编制和使用“军队甲种测验”和“军队乙种测验”，团体测验就此开始。前者是言语测验，适用于一般的筛选和安置的目的。后者是非言语测验，适用于操外语的被试和文盲被试，显然使用甲种量表不能合适地对他们进行测验。后来编制了大量民用团体测验，都遵循这两个量表所建立的模式。

两个最初军队测验的民用型修订本，持续使用了几十年。在军事部门，后来编制了“陆军资格测验”(Armed Forces Qualification Test, AFQT)，这是一个初步的筛选工具，随后使用多重能力倾向分类成套测验来分配专业。AFQT 根据相同数量的词汇、算术、空间关系、机械能力等项目，提供单一的分数。后来又编制了“军队服务职业能力倾向成套测验”(Armed Services Vocational Aptitude Battery, ASVAB)，适用于各个军事部门，这是一个综合性选拔和分类成套测验。ASVAB 的一些分测验构成共同的资格测验。对人员分类时，每个部门则选择和同时使用几个分测验，以符合具体专业的需要。

我们在本章中，将首先讨论团体测验和个别测验之间的主要差异；其次讨论适合个体的团体测验的新方法，及计算机在测



验程序中的应用；随后我们介绍一些内容范围广泛的通用团体测验；最后，我们将探讨团体测验以及第八章所讨论的个别测验中测验编制和使用方面的当前趋向。这种趋向是，当初设计为对单一综合能力（例如智力、学业能力倾向）的总测量同多重能力倾向成套测验相融合。越来越多的能力测验正在被改编，以便于灵活使用，单个的量表因而能够产生不同范围的分數，从一般到具体，以符合多种不同的测验目的和情境。

## 团体测验与个别测验

**测验设计的典型差异** 团体测验在项目的形成和安排上必然不同于个别测验。虽然可以使用自由反应的开放式问题——早期团体测验也曾经使用过——今天典型的团体测验则使用多项选择题(multiple-choice items)。这种变化显然是评分的统一性和客观性所需要的。传统的个别测验和团体测验之间的另一个主要差异是控制项目的难度。在个别施测的测验中，主试遵守始测规则以及下限和上限规则，以保证用适合于其能力水平的项目来测量每个测验参加者。在团体测验中，相似内容的项目按难度升序安排在分别限定时限的分测验之内。这种项目组织使每个测验参加者都有机会去尝试每种类型的项目，例如词汇、算术、空间等，以便他在可能浪费大量时间去尝试较难的项目之前，完成每种类型中较易的项目。

然而，使用分开的分测验所碰到的一个实际困难是，经验不足的或不够细心的被试也许在时间安排上犯错误。比起整个测验的单一的长时限，这类错误在若干短时限中更可能发生，相对来说也更为严重。为了调和使用单一的时限与让所有测验参加

者尝试难度水平逐渐增加的各类项目，一些测验使用螺旋混合形式(spiral-omnibus format)。采用这种形式的最早测验之一是“奥蒂斯心理能力自我测验”(Otis Self-Administering Tests of Mental Ability)，正如名称所暗示，它力图把主试的作用减到最低限度。在螺旋混合形式测验中，首先呈现每种类型中最容易的项目，其次是第二容易的项目，依此类推。现说明如下：

- |                                 |     |
|---------------------------------|-----|
| 1. 恨的反义词是：                      | 答案  |
| (1) 敌人 (2) 害怕 (3) 爱 (4) 朋友      |     |
| (5) 高兴 .....                    | ( ) |
| 2. 3支铅笔要花2角5分，7角5分能买多少支铅笔？..... | ( ) |
| 3. 鸟并非总是有：                      |     |
| (1) 翅膀 (2) 眼睛 (3) 脚 (4) 巢       |     |
| (5) 喙 .....                     | ( ) |
| 4. 荣誉的反义词是：                     |     |
| (1) 光荣 (2) 耻辱 (3) 胆怯 (4) 害怕     |     |
| (5) 失败 .....                    | ( ) |

为了避免重复每个项目的指导语，以及减少测验参加者所需要的指导语的变换次数，一些测验采用5~10个项目组的螺旋混合安排而不是单一项目的螺旋混合安排。

**团体测验的优点** 团体测验主要设计为进行大规模测验的工具。与个别测验相比较，团体测验既有优点，又有缺点。其优点表现在：第一，团体测验能够对许多个体同时进行测验，只要他们在测验场地能够自如地通过话筒作答。团体测验技术的发展，使得大规模测验计划可能实现。仅仅使用书面的项目和在

测验小册子或答卷上抑或在计算机上都能够记录的简单反应，不再需要主试和被试之间一对一的关系。

团体测验有利于大规模测验的第二点是，大大简化了主试的作用。例如，施测斯坦福一比纳量表需要大量的培训和经验，与之相对照，大多数团体测验仅仅需要对测验参加者念出简单的指导语和准确地控制时间。当然，也需要进行一些初步培训，因为毫无经验的主试可能无意中改变标准化程序而影响测验结果。然而，由于主试的作用减小到最低限度，团体测验能够比个别测验提供更为一致的条件。使用磁带录音的指导语，以及计算机施测，为大规模测验中标准化程序和消除主试变量，提供了进一步的机遇。一般来说，团体测验的评分较为客观，一名办事员就能完成。现在大多数团体测验能够由计算机评分。

团体测验的再一个特点是，所建立的常模比个别测验要好。由于团体测验收集资料相对容易和快速，在标准化过程中往往测验大型的、有代表性的样本。在大多数最近进行标准化的团体测验中，常模样本的人数通常在 10 万 ~ 20 万之间，与之相对照，大多数认真编制的个别智力量表进行标准化时，所费力积累的个案为 1000 ~ 8000。

**团体测验的缺点** 虽然团体测验具有若干令人满意的特点，在目前测验中几乎是必不可少的，但也应当指出它们的局限性。在团体测验中，主试没有什么机会与被试建立和睦关系，获得其合作并维持其兴趣。与个别测验相比，在团体测验中不易发现被试当时的状态，例如疾病、疲劳、忧虑、焦虑等，这也可能影响测验成绩。一般来说，不习惯测验的个体在团体测验上的障碍比在个别测验上多。也有证据表明，情绪障碍儿童个别测验上的成绩比团体测验上的好（Bower, 1969; Willis, 1970）。

团体测验遭到批评的另一个缺点是，对被试反应所强加的

限制。批评尤其针对多项选择题以及诸如类推、类同、分类等标准项目类型 (Hoffman, 1962; LaFave, 1966)。其中一些论点具有独创性且引起争议。一个论点是, 这类项目可能使卓越的、有创造性思维的被试处于不利地位, 因为他无法在答案中展示其不同凡响的一面。应当附带指出, 根据项目分析和效度资料, 如果发生这种情况, 也只是偶发事件。而且, 如果在一名个体的测验中, 在一二个项目上发生这种情况, 那对被试的总分几乎没有什么明显影响。一些批评强调应该分析错误以及询问个体为什么选择某特定答案的理由, 如同典型的皮亚杰方法那样 (Sigel, 1963)。无疑的, 在团体测验中, 主试没有多少甚至根本没有机会去直接观察被试的反应或确定非典型行为的原因。由于上述原因及其他原因, 一旦作出关于个体的重要决策, 就需要对团体测验作补充, 或对可疑个案进行个别测验, 或从别处获得另外资料。

传统的团体测验的再一个局限性是缺乏灵活性, 得按顺序对每个被试测验所有项目。假如每个被试只做适合于其能力水平的项目, 就可以更为有效地利用测验时间。而且, 这种做法一方面可避免因为做太容易的项目而产生乏味感, 另一方面可避免因为尝试超出个体有能力的项目而产生挫折和焦虑。个别测验一般让主试根据测验参加者的先前反应来选择项目。当测验的设计针对广泛的能力范围时, 个别测验与团体测验之间的这种差异就尤为重要。

## 适应性测验和计算机施测

**适合个体的测验** 人们正在探索若干技术, 努力使个别测验和团体测验的一些优点相结合。至今的主要兴趣集中在如何

使项目内容适应各个测验参加者的反应特征。在关于这个主题迅速增加的文献中,这种方法被不同地称之为适应测验、顺序测验、分支测验、特制测验、个别化测验、程序测验、动态测验、依反应而定的测验等。虽然有可能设计出体现这类适应程序的纸笔团体测验(Cleary, Linn & Rock, 1968; Lord 1971a),但是这些技术本身最适宜于计算机化的测验实施。

适应性测验能够采用多种程序模型(De Witt & Weiss, 1974; Larkin & Weiss, 1974; Weiss, 1974; Weiss & Betz, 1973)。图 10-1 说明一个包括两阶段测验的简单例子。在这个假设的测验中,所有被试参加 10 个项目的常规测验,其项目包括广泛的难度范围。根据被试在常规测验上的成绩,让他或她做另外 3 个测验中的一个,每个测验均有 20 个项目,但难度水平不同。因此,虽然整个测验由 70 个项目构成,但每个被试只做 30 个项目。角

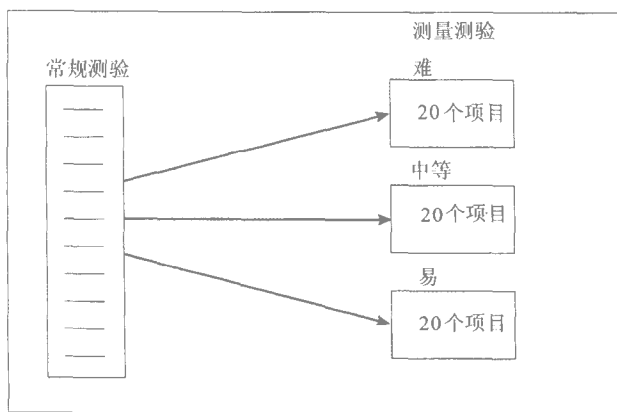


图 10-1 两阶段适应性测验,有三种测量水平。每个被试做常规测验和一个测量测验

锥形测验说明另一种不同的安排,如图 10-2 所示。在这种情况下,所有被试先做一个中等难度的项目。如果个体做对该项目,他朝上做下一个较难的项目;如果做错了,他朝下做下一个较易的项目。每个项目反应之后都重复这种程序,直到个体作出 10 个反应为止。这个例子说明一个 10 阶段的测验,测验共有 55 个项目,但每个被试做其中 10 个项目。粗线表示某被试所走的路线,在上面用 + (对) 或 - (错) 列出他的反应。

计算机化适应性测验(CAT) 图 10-1 和图 10-2 说明两种适应性测验模型,其中若干变式都采用纸笔和计算机两种形式尝试过。然而,较为复杂的模型并不使用预先建立的、固定的

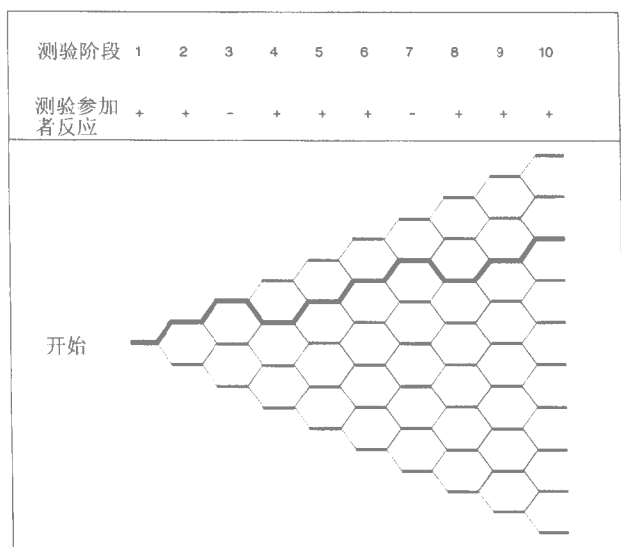


图 10-2 角锥形测验模型。粗线表示测验参加者的路线,在上面列出他的项目反应。

项目顺序模式，只有使用计算机化适应性测验才行得通（Embretson, 1992; B. F. Green, 1983; Wainer 等, 1990）。这些 CAT 方法利用第七章所描述的项目反应理论（IRT）技术来建立项目库、测验个体、评分个体作业等。对于库中的每一个项目，都有通过该项目的 50 对 50 的可能性所需要的能力估计值。这种能力估计值便是个体通过该项目所获得的分数。它反映了有关该项目的难度水平、鉴别度值和猜对的可能性。还有项目信息函数，它表示测量精确性。项目信息函数的总和为测验信息函数，它与传统的测量标准误差（SEM）具有同样功能。每个项目反应之后，计算机根据到此为止的个体总的反应情况，来选择下一个项目。用新增的项目继续测验，直到测验信息函数达到预先建立的标准为止。因此，所有个体都被测量到相同的精确度水平。

个体的分数不是根据答对的项目数，而是根据那些项目的难度水平和其他心理测量学特性。从对应于所通过的每个项目的的能力估计值中得出测验总分。每当加上新项目时，就再次调整 and 精制这种能力估计值，直到达到预先决定的测量精确度为止。对于用该项目库所测验的所有个体，这类分数都是可作比较的，而不管每一个体所做的特定的一组项目。使用评价系统公司（Assessment Systems Corporation）<sup>[1]</sup>所分发的若干现有的计算机程序，例如微软 CAT，有利于编制 CAT 量表的实用程序。

一般来说，使用各种方法的研究表明，个别化适应性测验能够达到常规测验同样的信度和效度，但使用的项目却少得多，测验时间也少。对于测验所包括的能力范围中高低两端的个体，它的测量也更为精确（Lord, 1970, 1971a, 1971b, 1971c, Weiss, 1982）。也有相当多的研究表明，编得好的测验的 CAT 形式和其纸笔形式之间的相关，大致与大多数测验的信度系数一般高。这类研究结果表明，两种形式实质上能够测量同样的结构（Mead

& Drasgow, 1993)。同时,某些测验情境不适用于 CAT,例如按照临界分数分类个体的筛选测验和速度测验 (Wainer, 1993b)。现在尤为关注制订评价 CAT 工具的技术指标 (Green, Bock, Humphreys, Linn & Reckase, 1984)。

适应性测验尤其适用于第三章中所讨论的个别化教学程序,其间学生按自己的进度进行,因而需要难度水平大不相同的项目。计算机化测验会使得个体的反应一旦提供足够的信息,就可以确定他或她已掌握教学内容,测验也随之停止。目前正在积极研究和发展在若干环境中计算机化适应性测验的其他用途。教育测验服务中心和大学入学考试委员会联合编制的一个计算机化适应性测验说明了一种用途,把入学新生分配到合适的英语和数学专业 (Smittle, 1990; Ward, Kline & Flaughner, 1986)。由于个别化项目选择,该测验测验时间短,且立即评分。因此,在注册时就能够施测,而且可以立即把学生安排到最适合于其现在成绩水平的课程或小组。

CAT 的另一种主要应用在于工业、政府、军事等部门的大规模的人员选拔和分类程序。CAT 特别适用于这种目的,至少有 3 条理由:(1) 要测验的申请人源源不断,因而不可能在任何一个时间和地点进行大型团体测验;(2) 要包括的能力水平范围很大;(3) 测验安全性较大,每一申请人可从计算机所储存的大型题库中得到一组不同的项目。关于“军队服务职业能力倾向成套测验”(Armed Services Vocational Aptitude Battery, ASVAB) CAT 版的解释性研究,进行了一段时间 (McBride & Martin, 1983; Moreno, Wetzel, McBride & Weiss, 1984; Wiskoff & Schratz, 1989, 1989)。现有主要团体测验的 CAT 版目前正在陆续编制当中,包括本章最后一节要讨论的“区别能力倾向测验”(Differential Aptitude Tests)。<sup>[2]</sup>对于许多实际应用,以及关于个体差异的性质



和根源的重大研究，CAT 无疑具有种种优点。恩布里森（Embretson, 1992）清晰而有效地描述了未来测验中 CAT 的前景。

## 多重水平成套测验

**总述** 与大多数个别量表和计算机化适应性测验不一样，传统的团体测验对所有测验参加者呈现同样的项目，而不管他们各自的反应。由于这种原因，任何一个团体测验必须包括相对有限的难度范围，以适合于它所设计的特定的年龄、年级或能力水平。为了提供在较大范围里可作比较的智力发展的度量，现在编制了重叠的多重水平成套测验。因此，任何一个个体仅仅用适合于他或她的水平来进行测验，在后来的年份里对同一个体再次进行其他水平的测验，或对不同年龄的团体进行比较评价。连续的重叠的成套测验，为处于各自年龄或年级分布的两端的个体，提供了适当的下限和上限。当然，应当指出，多重水平成套测验所提供的项目难度和测验参加者能力之间的匹配，至多也是近似的。而且，与 CAT 中所采用的个别化程序不同，这种匹配根据事先对测验参加者的了解，例如年龄或学校年级等，而不是根据那时他们自己的测验反应。

多重水平成套测验尤其适用于学校，因为需要若干年份里的分数具有可比性。由于这种原因，一般根据年级来描述水平。关于所包括的内容或智力机能，大多数成套测验都提供适当程度的连续性。各水平的分数按照统一单位的量表来表示。现在人们日益使用项目反应理论（IRT，第七章所讨论）的程序，来获得整个范围内分数的连续性和可比性。在测验标准化的过程中，各学生组做互有重叠的测验水平，以提供必需的联系资料。

在不同水平所测验的常模样本，也比单独标准化测验的更为等值。各个水平通常包括 1~3 个年级。另一方面，使用某个多重水平成套测验所能够统一测验的范围，可以扩展到从幼儿园到大学。

大多数成套测验产生一个总标准分数，相当于个别测验中传统的“IQ”。一些成套测验提供若干种常模，包括百分等级、标准九分、年级当量、标准分数等。除了总分之外，大多数成套测验也产生言语分数和数量分数，或语言分数和非语言分数。这种分类符合一种研究结果，即个体在言语分测验和其他类型分测验上的分数也许大不相同，在较高水平尤为如此。

成套测验的名称也令人关注。诸如“智力”“一般能力”“心理能力”“心理成熟”“学业潜力”“学业能力”等术语被用来称呼本质上同类的测验。在心理测量学家词汇中，这些术语实质上是同义词，可以互换。值得一提的是，在大多数最近编制或修订的成套测验中，“智力”这个术语已被较为具体的名称所取代。这种变化反映出人们日益认识到，“智力”这个术语获得了太多的附加意义，这可能导致对测验分数的误解。多重水平成套测验从重要的智力技能中取样，这被认为是学校作业的必要条件。它们的基本功用是评价个体在学校教育过程的每个阶段的学习准备。

代表性的成套测验表 10-1 所列出的三个成套测验，能够说明当前多重水平能力成套测验的性质和范围。之所以选择这些成套测验，是由于它们是近期的最新修订本、高质量的测验编制程序，以及标准化样本的规模和代表性。这些成套测验的另一个值得注意的特点是，每个成套测验与一二个多重水平教育成就成套测验同时进行同一年级的标准化，第十七章将讨论后一类测验。对同一标准化样本实施两类测验，就有可能建立

表 10-1 代表性的多重水平成套测验

| 成套测验                    | 年级范围       | 水平数    | 联合建立常模的测验                                                  |
|-------------------------|------------|--------|------------------------------------------------------------|
| 奥蒂斯—莱农学校能力测验(OLSAT,第7版) | 幼儿园至十二     | 7      | 斯坦福成就系列测验(第9版)                                             |
| 认知能力测验(CogAT-5型)        | 幼儿园至三、三至十二 | 2<br>8 | 依阿华基本技能测验(幼儿园至九年级)<br>成就和水平测验(九至十二年级)<br>依阿华教育发展测验(九至十二年级) |
| 认知技能测验(第2版,TCS/2)       | 二至十二*      | 6      | 加利福尼亚成就测验(第5版)<br>基本技能综合测验(第4版)                            |

\* 也有单独的认知技能初级测验(PTCS),以一组不同的测验适用于幼儿园和一年级水平。

两组分数之间的对应。结果就能够联合使用两类测验,以便较全面地探索学生的教育发展及其影响条件。

目前已使用合适程序广泛研究了这些成套测验的信度和效度。在每一一年级水平之内,都计算出总分和二三个不同内容成套测验上分数的库德—理查逊信度,大部分为0.90。再测相关也很高,这表明稳定性令人满意。和学校年级的相关以及和成就测验的相关,则表明良好的预测效度。各部分分数的相关及因素分析,揭示出每个成套测验中有一种广泛的一般因素。

不同水平上的典型的测验内容 适用于团体测验的最小年龄,是幼儿园和一年级水平。在学前年龄则需要个别测验,以便建立和维持和睦关系,以及实施适合于这类儿童的口头项目和操作项目。然而,到了五六岁,就有可能对不超过10~15名儿

童的小型团体实施书面测验。在这类测验中，主试仍然必须对测试儿童予以个别的关注，以保证他们照令行事，查看测验小册子是否翻到应翻的页码，以及监督其他的程序细节。如果需要测验较大的团体，则可以使用一二个主试助理。

初级水平的团体测验一般包括幼儿园和小学一二年级或三年级。在这类测验中，给每个儿童提供一本小册子，上面印着构成测验项目的图片和图解。所有指导语都是口头念出，通常也作演示。实际练习通常是，儿童试做一二道例题，主试或助理检查他们的反应，以保证他们恰当地理解指导语。儿童用一支颜色笔或软度铅笔在测验小册子上标上他们的反应。大多数测验仅仅要求儿童在一组图片中标出正确的一张。一些测验也要求简单的动作协调，例如画一条线把两点连起来。显然，初级水平的测验不要求测验参加者阅读或书写。

大多数多重水平能力成套测验包括适用于初级水平 (primary level) 的测验。图 10-3 说明这种水平上所使用的测验项目的类型。该图表示 OLSAT 的水平 A 的一些例题，适用于幼儿园。认识到儿童早期智力快速成长，OLSAT 的最新版本为幼儿园和一至三年级分别提供 4 种水平(A、B、C、D)。这表示该成套测验比较早版本及其他多重水平成套测验更加细化。在水平 A 所有指导语由主试口头念出。儿童的反应是，在所选的答案图片下涂满圆圈，如图 10-3 所示。

图中说明水平 A 所包括的 10 个项目中的 4 例。整个测验大约需要 75 分钟。分两个时间段施测，每段先做 15 或 20 分钟，中间休息 5 分钟。也可在几天之前做一个练习测验，使用类似的项目和指导语。图 10-3 所示的例题相对简单，以使儿童熟悉在本身测验中会碰到的项目。每个项目都有详尽的口头指导语及对项目内容的清楚描述。图 10-3 中所给出的解释则是

图片分类:标出和其他图片不相属的那张图片。


☐

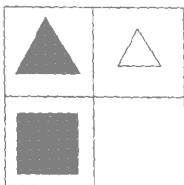
☐

☐

☐

☐

图形推理:在应该放在空方格的那张图形下面标上记号。

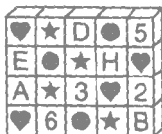

☐

☐

☐

☐

照令行事:标出正好在心形下面的那个数字。



2

☐

3

☐

5

☐

6

☐

图片系列:在应该放在空方格的那张图片下面标上记号。


☐

☐

☐

☐

图 10-3 奥蒂斯-莱农学校能力测验所用的一些例题

(OLSAT,第 7 版,水平 A)

高度浓缩的形式。也有一些微小的格式差异,以利于幼儿理解和保持注意。例如,页和行的确定,不仅用数字,而且用熟悉物体的小图,例如杯子、鞋子、剪刀等也给每个儿童一张书签,以

留意在合适的行上做作业。

适用于小学水平 (elementary school level) 的各种测验,从三四年级向上延伸,在内容和一般设计上具有许多共同之处。由于预先假设在这些水平有一定的读写能力,测验主要为言语内容;大多数测验也包括算术问题或其他数字测验。此外,一些成套测验提供非阅读测验,设计来评定同样的抽象推理能力,适用于操外语的、有阅读障碍或其他教育障碍的儿童。

图 10-4 说明适用于小学水平的项目类型。这些是认知能力测验(CogAT) 中级水平所使用的典型项目。如表 10-1 所示,CogAT 包括从幼儿园至三年级的 2 种水平,以及从三年级至十二年级的 8 种水平。每种水平印成单本小册子。参加不同水平的被试做一组不同的项目。如此设计测验,大多数被试所测验的项目难度中等而鉴别度则最为有效。

从水平 A 至水平 H 包括相同的 9 个分测验,分为如下 3 组成套测验:

言语成套测验——词语分类、句子完成、词语类推

数量成套测验——数量关系、数字系列、等式建立

非言语成套测验——图形分类、图形类推、图形分析

这些分测验不使用文字或数字,而使用几何或图形元素;项目与学校正规教学的联系相对较少。每个分测验,先做练习,作详细解释。在实际施测之前,也可做练习测验。图 10-4 说明 9 个分测验中 6 个的典型项目,指导语经浓缩,措词也稍作改变。这些项目的难度水平,大致相当于四至六年级所做的项目。手册建议,在 3 段测验时间里,每个儿童都要完成 3 个成套测验。对于大多数儿童,使用非言语成套测验预测学业成就,不如使用言语成套测验和数量成套测验的好。然而,对 3 个成套测验上的成绩作比较,可以提供关于特殊能力或特殊障碍的有用信息。

词语分类:想一想,黑体单词怎样类同。然后,在下面一行找出一个与它们相属的单词。

和蔼的      友好的      有益的

A. 有能力的    B. 积极的    C. 慷慨的    D. 美丽的    E. 强壮的

词语类推:想一想,第一对两个单词有什么关系。然后,找出一个单词,它和第二对另一单词形成同样的关系。

船 → 港口:卡车 →

A. 驾驶    B. 公路    C. 汽车库    D. 汽油    E. 货物

数字系列:想一想,下面一行数字排列的规律。然后,选出应该出现的下一个数字。

3    2    1    3    2    1 →

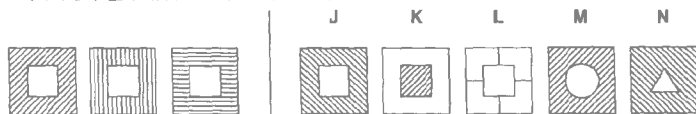
A. 0      B. 1      C. 2      D. 3      E. 4

等式建立:上面一行的数字和运算符号可以连成一个式子,得出一个答数。在下面一行数字中标出这个答数。

2    4    8    -    -

J. 0      K. 2      L. 4      M. 6      N. 10

图形分类:前面 3 个图形在某个方面是相同的。在右面图形中找出一个图形,它和前面 3 个图形相同。



图形分析:想一想,前面两个图形互相有什么关系。然后,在右面图形中找出一个图形,它和第二对的另一图形形成同样的关系。

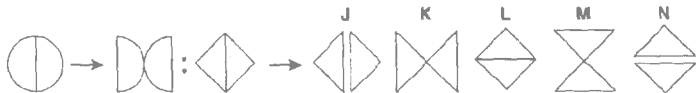


图 10-4 认知能力测验的一些项目类型。回答标在另外的答卷上。正确答案为:C,C,D,K,J,K(摘自 CogAT-5 型,水平 A-H 的练习测验)。

多重水平成套测验的较高水平适用于中学生 (high school students), 除了难度之外, 该测验基本上与适用于小学生的并无二致。出于种种目的, 这些水平也适用于测验一般未经选择的成人团体。认知技能测验 (TCS/2) 最高水平的项目, 可以说明该水平的测验内容。该成套测验的每个水平包括 4 个测验:

序列——理解和应用在一组或一串图形、字母或数字中的规律或原理。

类推——找出一对图片中的关系, 使第二对图片形成同样的关系 图片包括景色、人物、动物、物体、图形符号等。

言语推理——使用多种项目类型来测验, 这包括找出物体或概念的基本元素, 根据共性来对物体分类, 推断一组单词的关系, 根据短文得出逻辑结论等。

记忆——呈现一组人造单词(无意义音节)的定义, 以供学习。大约 25 分钟之后进行回忆, 在这期间实施其他测验。

还有一个练习测验, 在实际测验之前一两天施测。图 10-5 给出 4 种项目类型中的 3 种实例。在该成套测验中, 从四年级至十二年级使用相同项目类型的序列、类推、言语推理等测验; 所有这些水平包括相同的例题。较高的两个水平相当于中学年级, 根据经验上所建立的较高的项目难度作出区分。

TCS/2 成套测验的一个显著特点是包括一个记忆测验。言语、非言语、记忆等测验分别提供分数。通过初步的因素分析确定这些能力倾向领域, 指导编制和选择测验项目。在标准化过程中, 通过项目反应理论(IRT 参见第七章)的技术, 编制跨水平的量表。出于这种目的, 各学生样本参加包括相邻两个水平的项目的相联系的测验(《TCS/2 技术报告》, 1993, pp. 113 ~ 114)。在测验编制和评分中, TCS/2 成套测验使用 IRT 尤为出色。因此, 获得的分数有可能不仅反映所通过的项目数, 而且反映所通



## 序列

想一想,每组的型式。然后,在下面一行中选出一个项目,它应该放在空白处。

1.



A

B

C

D

2.

A C E G I \_

H J K L

A

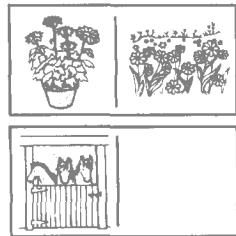
B

C

D

## 类推

想一想,上面长方形框中两张图片有什么关系。然后,在右面图片中找出一张图片,它和第二对的另一张图片形成同样的关系。



F

G

H

J

## 言语推理

1. 看一看下面划线的单词即字母表。下面哪一个单词提到某事物,而它一定是字母表的一部分?

字母表

A 单词

B 字母

C 数字

D 句子

2. 根据上面两句中给出的信息,下面哪一句一定是对的?

大钟是英国的一只钟。

朱迪参观过大钟。

A 人们经常参观大钟。

B 英国的许多钟都是很大的。

C 大钟以某人命名。

D 朱迪去过英国。

图 10-5 认知技能测验所用的一些项目类型。回答标在另外的答卷上(摘自 TCS/2,水平 2—6 的练习测验)。

过的各个项目的难度水平。

对多重能力倾向的认识 如同本章头几段所述,有一种明显的趋向,即在测量全面的一般能力和测量相对独立的单独能力倾向之间建立联系。这种联系来自最初引起争议的、暂时互不相容的能力测验的两种方法。这种趋向与个别测验所发生的情况极为相似,第八章讨论过这一点。在目前情况下,最初的多重水平成套测验作为个别智力测验的团体版本,尽管它们通常具有比较狭小的目标,即评定学业能力倾向或提高教育水平的准备性。人们逐渐明白,可用比较具体的分数,对单一的全面分数作有用的补充。本节所讨论的 3 个成套测验中都有这种进展。

有人指出,在 OLSAT(1996 版)中,总分本身限于一组“言语—教育”能力倾向,成套测验并没有试图评价一般智力的“实践—机械”部分,而且为在言语分数和非言语分数中找出更明确的补充分数作了准备。然而,这种区分主要集中在测验项目在作答时需要使用语言与否。这种区分因而主要针对把测验使用于英语知识有限的学生。然而,在广泛的言语类型和非言语类型之内,也有可能获得 5 个测验组的标准九分(在年龄或年级之内)。这些测验包括言语理解、言语推理、图片推理、图形推理、数量推理等。手册提出,比较个体在这些测验上的相对成绩,有助于识别学生的长处和不足(第 7 版 OLSAT 技术手册,1997)。

CogAT(5 型,1993) 不仅提供总合成分的常模,而且提供言语分数、数量分数、“非言语”(即空间)推理分数的常模。此外,个体分数报告包括这 3 个领域分数及总分的直方图。手册特别关注直方剖面图对于预测学业成就的价值。关于分数解释的说明,再次提到考虑个体分数剖面图的实际意义(参见 Riverside 2000,1994,p.44),因而对多重能力倾向的认识又前进了一步。

认知技能测验(TCS/2, 1992)编制者更为明确地认识到,多重能力倾向资料在评定学业成绩中的价值。该测验不仅有成套测验上总分的常模,而且有4个分测验中的每个分数及“非言语”总分(序列分测验和类推分测验)的常模。此外,编制成套测验本身就是评定由因素分析研究所找出的3个广泛的认知特质,即言语推理、非言语推理、记忆。因而人们特别认识到需要多重能力倾向测验,下一节将讨论这一问题。

## 测量多重能力倾向

传统的智力测验,不管个别测验还是团体测验,主要设计来对个体认知发展的一般水平,进行单一的、总体的测量,例如IQ。然而,实践和理论两方面的发展,不久就使人们关注在早期智力测验所代表的松散组合之内某些不同的能力倾向。这种发展导致单独测验的编制,以测量一些广泛应用的能力倾向。同时,这种发展也导致精确定义和更全面理解智力测验本身所测量的东西。

若干事件促使人们不断关注测量不同的能力倾向。首先,人们日益认识到智力测验成绩中个体内的变化。试图大致比较个体在不同的分测验或项目组上的相对位置,比编制多重能力倾向成套测验早了好多年。然而,智力测验的设计不适合于这种目的。分测验或项目组往往很不可靠,不能证明个体内比较是有道理的。而且,编制智力测验时,一般选择项目或分测验,以提供单一的、内部一致的度量。因而在这样一种选择中,努力使个体内变化减至最小,而不是增至最大。一般来说,与量表的其余部分相关甚低的分测验或项目被排除在外。但是,如果强

调不同的能力,那倒也许正是要保留这些部分。限于大多数智力测验的编制方法,这种测验上的作业分为言语和非言语,或语言和数量两大类,而不可能再分为更多的类。

人们逐渐认识到,所谓的一般智力测验的一般性,事实上比当初所假设的要差,这进一步促进了编制多重能力倾向成套测验。显而易见,许多这类测验主要测量言语理解。除了在某些操作量表和非语言量表中之外,通常没有测量诸如机械能力之类领域。由于认识到智力测验的这种局限性,心理学家开始限定‘智力’这个术语。有人建议区分‘学业’智力和‘实践’智力。有人提出‘抽象’智力‘机械’智力和‘社会’智力。有人编制‘特殊能力倾向’测验,以补充智力测验。但是,严谨的分析表明,我们能够说智力测验本身是测量言语能力和数量能力等特殊能力倾向的某种组合。

心理学家不断参与职业咨询,以及工业和军事等人员的选拔和分类,也促进了区别能力倾向测验的编制。早期编制的办公、机械及其他职业领域中的专门化测验,就反映了这类影响。编制适用于医学院、法学院、工学院、牙科学院和其他专业学院申请人入学选拔的成套测验,同样反映了多年来的进展。而且,若干区别能力倾向成套测验,例如军队和美国职业服务中心所编制的一些测验,则是职业选择或分类工作的直接结果。

最后,通过因素分析<sup>[3]</sup>技术对特质结构的研究,为编制多重能力倾向成套测验提供了理论基础。通过这类因素分析研究,能够更为系统地鉴定、分类和定义在“智力”之内松散组合的不同能力,因而能够选择测验,使每个测验最好地测量因素分析所鉴别的一个因素或特质。

区别能力倾向测验 使用最广泛的多重能力倾向成套测验之一便是“区别能力倾向测验”(Differential Aptitude Tests, DAT)。

DAT 最初出版于 1947 年,定期进行修订(第 5 版,C 型,1992)。该成套测验主要适用于八至十二年级学生的教育和职业咨询。第 5 版 DAT 有两种水平 水平 1 主要适用于七至九年级学生以及完成这些年限学业的成人;水平 2 适用于十至十二年级学生以及完成九年以上学业但中学未毕业的成人。

DAT 由下列 8 个测验组成 言语推理、数推理、抽象推理、知觉速度和准确性、机械推理、空间关系、拼写、语言应用等。图 10-6 说明其中 4 个测验的例题。有一个练习测验,包括全部 8 个领域,被试可以事先使用。还有一个特殊型式,即“人事和职业评定区别能力倾向测验”(Differential Aptitude Tests for Personnel and Career Assessment),8 个测验中的每个都经缩简,单独印在小册子上。使用该型式,可以选择与具体职业有关的特定测验,可以单独使用。<sup>[4]</sup>

像大多数现行主要测验一样,DAT 能够完全以计算机化版本施测。现在正在探索的一种最新发展则是 1987 年以来所使用的计算机化适应性测验(CAT)型(DAT 适应型)。如同在所有适应性测验中,被试只做适合于其作业水平的那些项目。这种 CAT 版本所使用的项目选自较早的 DAT(V 型),再用项目反应理论的简化的一元参数模型,即拉希模型(参见第七章)进行项目分析。

测验出版人员和独立的研究人员多年来都在广泛收集 DAT 效度资料,后者在不同的职业环境中或在研究项目中应用该成套测验。其中大多数资料是关于学业和职业两种课程中中学成就的预测效度。许多效度系数都很高,甚至在测验和效标资料间隔长达 3 年的情况下也是如此。关于区别预测,结果就并不怎么鼓舞人心了。虽然一般来说,言语测验和英语课程的相关与数字测验和数学课程的相关度都较高,但却表明有一种广泛

## 言语推理

选择一对正确的单词填空。其中第一个单词填入句子开头的空格，而第二个单词填入句子结尾。

……对鳍 好比 鸟对……

- A. 水—羽毛      C. 鱼—翅膀  
B. 鲨鱼—巢      D. 鳍形肢—飞  
E. 鱼—天空

正确答案是 C

## 数推理

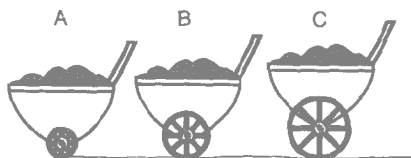
在这个加法题中，哪个数字应该代替 R？

$$\begin{array}{r} 7R \quad \quad F \quad \quad G \quad \quad H \quad \quad J \quad \quad K \\ + R \\ \hline 88 \quad \quad 9 \quad \quad 6 \quad \quad 4 \quad \quad 3 \quad \quad \text{都不是} \end{array}$$

正确答案是 F

## 机械推理

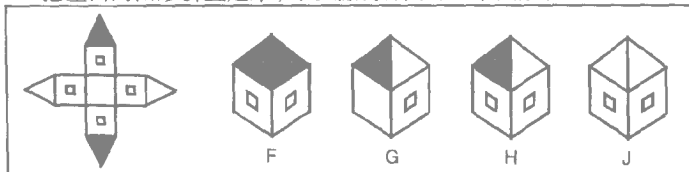
推哪辆车容易通过松软的沙滩？



正确答案是 C

## 空间关系

把左面的图形折叠起来，可以制成后面哪一个图形？



正确答案是 H

图 10-6 第5版区别能力倾向测验的例题

的一般因素,它构成各门学业课程的基础。例如,言语推理和大多数课程都有高相关。主要由于这种原因,所以把 VR + NR 分数作为学业能力倾向的一种指标。它是言语推理分测验和数推理分测验上分数之和,这个指标和学业成就总效标的相关为 0.70 ~ 0.80。有这种指标的常模,该指标是 DAT 剖面图通常包括的分数之一(参见图 4-6)。还有一个部分成套测验(Partial Battery),只包括 VR 分测验和 NR 分测验,只有当需要学业能力倾向的一般指标时,才能够使用该成套测验。

从另一个角度来看,越来越多的证据表明,传统的“一般智力”测验或“学业能力倾向”测验——不管个别测验还是团体测验——根据种种广泛的教育和职业效标而得出相当高的效度系数(L. S. Gottfredson, 1986a; Guion & Gibson, 1988; Pearlman 等 1980; Schmidt, Hunter, Pearlman & Shane, 1979)。这类测验实质上包括 DAT 的 VR + NR 分数所评定的同一群认知知识和认知技能。因而能够看到,智力测验和多重能力倾向测验之间的原先差距正从两端开始缩小。DAT 之类测验更为强调使用 and 解释广泛的分数,例如 VR + NR。同时,最初的一般测验也更为强调使用和解释各个分测验的分数以及剖面图分析。关于 DAT 的评述,参见哈特鲁普(Hattrup, 1995)和施密特(N. Schmitt, 1995)的有关论述。

**多维能力倾向成套测验** 最近编制的“多维能力倾向成套测验”(Multidimensional Aptitude Battery, MAB),不失为能力测验的一种新型式。该测验最初出版于 1984 年,后来对施测程序、常模、手册等都作了修订(Jackson, 1994b)。MAB 属团体测验,旨在评定如同韦克斯勒成人智力量表一修订版(WAIT-R, 参见第八章)一样的能力倾向。它包括言语量表<sup>[5]</sup>和操作量表,各由 5 个分测验组成,得出 V、P 和全量表的离差 IQ。MAB 适用于青少

年和成人,但不适用于智力落后或智力缺陷个体,他们的智力状况会妨碍他们理解或执行测验指导语。

10 个分测验的名称与相应的 WAIS - R 相同(其中一个例外),现列出如下:

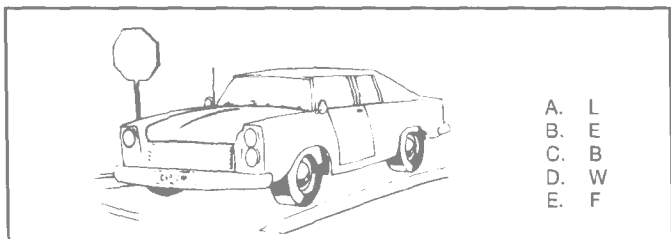
| 言 语 | 操作   |
|-----|------|
| 常识  | 数字符号 |
| 理解  | 图画补缺 |
| 算术  | 空间关系 |
| 类同  | 图片排列 |
| 词汇  | 物体拼配 |

MAB 的空间关系代替了 WAIS - R 的积木图案分测验。设计纸笔项目来测量个别施测的 WAIS - R 所包括的同样的机能,这表现出相当的独创性;对于操作测验尤为如此。图 10 - 7 说明图画补缺分测验和空间关系分测验的各一道简单的例题。在每个分测验的所有项目上,被试的任务同这些项目中所说明的一样。在图画补缺分测验中,被试必须想到缺少部分的名称,然后在给出的选项中找出该名称的首字母。在空间关系分测验中,把左面的图形在纸上转动一个位置,只能得到右面 5 个图形中的一个;而选项中给出的其他图形则需要把左面图形翻转。

一本小册子包括 5 个言语分测验,另一本小册子包括 5 个非言语“操作”分测验。每本小册子一开始是练习题,说明 5 个分测验的其中 3 个所使用的项目类型;每个分测验一开始还有 1~3 道例题。手册中给出每个分测验的一般指导语和具体指导语;指导语可以口头读出,也可由录音机或计算机呈现。在单独的答卷上或在计算机上记录反应,用套板手工评分或由计算机评分。另有一种选择,MAB 现行版本之一由当地测验使用者完全进行计算机化施测和评分。

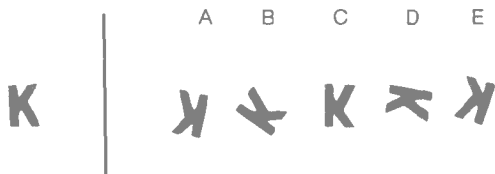


图画补缺:选择一个字母,该字母打头的单词说明图中缺少的部分。



答案是 Light(灯),所以应该选 A

空间关系:在竖线右面选择一幅同左面一样的图形。该图形逆转后,看上去就同左面图形一样了;而其他图形则必须翻转才行。



正确答案是 A,所以应该选 A。BCDE 等必须翻转

图 10-7 多维能力倾向成套测验 (MAB)两个操作测验的例题

参照常模表,能够把 10 个分测验的原始分数分别转换成等值的、统一的量表分数(平均数为 50、标准差为 10 的标准分数)。从 16 岁至 74 岁分为 9 个年龄组,使用言语量表、操作量表和全量表上的这些量表分数之和,可以查出每个年龄组的离差 IQ (平均数为 100,标准差为 15)。也有单独的常模表,可以查出每个年龄组的等值的量表分数;使用这些量表分数,能够绘制合适的年龄剖面图。使用单一的成套测验而不是一组不同测验的分数,其优点是成套测验根据同一标准化样本建立所有常模,因而

可以直接比较。

整体来说,编制 MAB 所采用的心理测量程序具有高度的技术质量;每一步骤都由十多年的广泛研究所证实。<sup>[6]</sup> MAB 另一个值得注意的特点则是它和 WAIS - R 的经验联系。首先,在 145 名青少年和成人的异质性样本中, MAB 每个分测验分数以及言语总分、操作总分、全量表总分等,与 WAIS - R 相应分数的相关很高。事实上,除了 2 个分测验之外(数字符号和空间关系/积木图案)这些相关与 WAIS 和 WAIS - R 分数之间的相关一般高甚至更高一些。应当指出,尽管 MAB 所有项目都是新编的,但仍然获得这些相关;两个量表之间没有共同的项目。160 名 16 ~ 35 岁的被试参加了 MAB 和 WAIS - R 两个测验,根据这个异质性样本,建立起两个测验分数的线性方程。学生样本和精神病患者样本的初步方程表明,这类校准可推广到不同的测验参加者样本。这种校准程序提供了一种有希望的方法,即根据从大型的、有代表性的标准化样本中得出的资料建立一种统一的量表,然后根据这种量表来表示各种新编测验上的分数。这种程序实际上是实现第三章所讨论的全国参照常模这一目标的一个步骤。

从另一个角度来看,这个成套测验清楚地体现了目前对于层次分数的趋向。在 10 个分测验水平,在较广泛的言语和操作水平,以及整套测验的总分,它都提供可充分解释的分数。测验使用者因而具有高度的灵活性,选择合适的分数水平以适合特定的测验目的——看来这可能是 21 世纪能力测验的一个特点。

#### 注 释

[1] 参见夸恩、珀克、桑德尔和沃尔夫(Quan, Park, Sandahl & Wolfe, 1984)以及韦斯和韦尔(Weiss & Vale, 1987)的有关论述。

- [2] DAT-适应版(1987)相当于纸笔 DAT-V 型(1981)。
- [3] 第十一章将讨论这个问题。
- [4] 职业兴趣问卷(Career Interest Inventory)和第 5 版 DAT 联合进行标准化。该问卷适用于教育和职业咨询,能够把它和 DAT 一起施测、评分和解释。
- [5] WAIS-R 数字广度分测验在 MAB 中没有平行型式。该分测验难以用纸笔形式呈现;再则,它和韦克斯勒全量表分数的相关最低。
- [6] 参见雷诺兹(S. B. Reynolds, 1989)和西尔弗斯坦(Silverstein, 1989)的有关论述。

## 第十一章 智力的性质

所有心理测验都是设计来测量行为的。因此，选择合适的测验和解释测验的结果，都需要有关人类行为的知识。不仅测验编制者，而且测验使用者，都需要熟悉有关的行为研究。在本章和下一章中，我们要考察一些积累起来的关于行为的知识，智力技能测验就是要评定这些行为。心理研究对于理解下列问题能够起到什么作用：(1) 认知能力测验即‘智力测验’所测量的行为；(2) 这类行为中个体差异的来源；(3) 在后来某一时间和其他情境中这类行为的可预测性。

首先，应当指出，不仅一般大众，而且诸如生物学、哲学、教育等不同学科的人员(参见 Sternberg, 1990)，以及从事不同专业领域或持有不同理论倾向的心理学家(例如，H. Gardner, 1983, 1993; Sternberg, 1985a, 1989 也参见 Brody, 1992; Lubinski & Benbow, 1995; Messick, 1992; H. Rowe, 1991) 都在不同的意义上使用未加限定的“智力”一词。早在 1921 年，《教育心理学杂志》主编，邀请 17 位著名研究人员探讨智力的定义和概念 (1921)，结果众说纷纭。65 年之后又进行了一次相似的调查 (Sternberg & Detterman, 1986)。阅读一下这些论著，具有相当的理论意义，也为展开热烈的讨论和在相互冲突的观点中间寻求某些可能的共同点，提供了一种基础。不过，出于现在的目的，我们将目标限于狭小的范围：关于传统的智力测验所评定的、并且用 IQ 所表

示的人类智力的特定部分，我们应该了解什么？现在的讨论正是在这种限定的意义上，使用“智力”一词（参见 Anastasi, 1983c）的。

## IQ 的涵义

对于一般大众来说，IQ 不是表示某个特定测验上某种特定类型的分数，而往往是智力的一种速记名称。<sup>[1]</sup>这种用法十分流行，所以就不能忽略或只是悲叹，这是一种大众的误解而已。诚然，当我们考虑某个 IQ 的数值时，总是应该指明得出这个 IQ 的测验。不同的智力测验都得出 IQ，但是事实上，它们在内容和影响测验分数解释的其他方面，却有所不同。在前面几章所讨论的一些例子中，共同使用“智力测验”名称的测验之间的一些差异是显而易见的。但是，仍然需要再讨论一下由 IQ 所表示的“智力”一词的一般涵义。

首先，应该把所测验的智力看作为是一种描述的概念，而不是一种解释的概念。IQ 表示个体在某个时间相对于年龄常模而言的能力水平。没有一种智力测验能够说明个体表现的原因。把一个测验上或日常生活中不好的表现归因于“不好的智力”，这是同义重复，决不会有助于我们了解个体的障碍。事实上，它倒可能延误我们在个人历史中努力探索障碍的实际原因。

智力测验，以及任何其他类型的测验，不应该被用于给个体贴上标签，而是应该有助于对他们的了解。在多种资料中，从各个心理学家的著作到全国委员会的正式报告（Hobbs, 1975a, 1975b；全国委员会……，1990），多年来一直强调这一点。一本广为宣传的书（Hermstein & Murray, 1994）——书名是《钟形曲

线》(The Bell Curve)——使得关于智力测验成绩中种族差异和性别差异的种种框框和误解继续存在,并且给本来已经复杂的问题增添了混乱和争论。美国心理学会智力工作小组(American Psychological Association Task Force on Intelligence)在一份报告中客观而真实地论述了有关论点(Neisser等,1996)。美国心理学会1995年大会上,也有一个讨论会专门阐明有关问题(Steele, Chair, 1995年8月)。为了使人们达到最大的机能水平,我们需要从其时他们所在之处开始;我们需要评定他们的长处和短处并相应地作出计划。如果一个阅读测验表明,某儿童阅读落后,我们不是把他称为阅读落后儿童就完事;我们也不是实施一个非言语测验而隐藏阅读障碍。相反,我们全力去教儿童阅读。此外,当代测验的一个重要目的,便是有助于自我理解和个人发展。测验所提供的信息,正日益被用于帮助被试制定教育计划和职业计划,以及作出关于他们自己生活的决策。人们注重将测验结果告知被试的有效方法,这表明人们正日益认识到测验的这种应用。

要记住的第二个要点是,智力不是一种单一的能力,而是由若干机能构成。智力这个词通常用来表示特定的文化之内生存和发展所需要的各种能力的组合(Anastasi, 1986c)。由此推断,这个合成物内所包括的各种具体能力及其相对权重,都随时间和地点而有所变化。在不同的文化中,以及在同一文化之内不同的历史时期,达到成功应有的条件有所不同。在个体的一生中,从婴儿期到成人期,也能够看到智力构成的变化。在其价值被个体的文化或亚文化所强调的那些机能上,个体的相对能力往往随着年龄而上升;而在其价值不被强调的那些机能上,个体的相对能力则往往随着年龄而下降。

适用于学龄儿童或成人的典型智力测验主要测量言语能

力；在较小的程度上，它们也测量处理数字和其他抽象符号的能力。这些都是学校学习中的主要能力，因此能够认为大多数智力测验是测量学业能力倾向或学业智力。IQ 既反映先前的教育成就，也预测以后的教育成绩。因为教育系统中所教授的各种机能，在技术先进的现代文化中是十分重要的，所以学业智力测验的分数，也能够有效预测个体在这种文化中的许多职业和日常生活中的表现。

当然，应当指出，许多重要的心理机能，智力测验从来没有做过测量。机械、运动、音乐、美术等能力倾向都是显而易见的例子。动机、情绪、态度等变量对于各个领域的成就来说，都是重要的决定因素。事实上，一些心理学家在智力定义中包括人格成分(例如，H. Gardner, 1983)。同样，创造力研究也识别出与创造性生产有关的认知和人格两种变量。当然，所有这些意味着，作出个体决策和组织决策，都应该根据合情合理收集的多种有关资料。单单根据测验，尤其是单单根据一二次测验就作出决策，这显然是误用测验。决策必须由人来作出，测验只是作出决策中所使用的资料的一种来源，测验本身不是作出决策的工具。

我们关于智力测验测量什么东西的许多信息，来自于预测教育成就和职业成就时对测验效度的实践研究。在较为理论的水平上，70 年代后期对于智力测验所测量的智力结构兴趣大增(Humphreys, 1979; Resnick, 1976; Sternberg & Detterman, 1979)。这种兴趣方兴未艾，涉及心理学的不同领域，并且由不同的理论观点，通过各种各样的方法途径继续下去。在连续出版的系列丛书中(Dettermann, 1985 ~ 1993; Sternberg, 1982 ~ 1989)，以及在关于这个主题的一本完整的大全中(《人类智力大全》，*Encyclopedia of Human Intelligence*, 1994)，都反映出这一点。

努力了解智力测验测量什么东西，这不仅包括结构效度的标准统计方法，例如因素分析，而且包括将信息加工技术应用于智力测验中所呈现的任务（参见第五章）。后者强调测验参加者作出测验问题答案的基本过程，而不是仅仅考虑答案的正确性。这类分析应该大大有助于测验的诊断效用，以及发展满足个体具体需要的训练程序。

## 遗传力和可变性<sup>[2]</sup>

许多混乱和争议起因于把遗传力估计值应用于测验分数。一个著名的例子是詹森（Jensen, 1969）的一篇掀起轩然大波、引发许多激烈争论的论文，其中一些争论持续到 90 年代。虽然这场争论包括若干方面，论题也十分复杂，但争论的一个主要的实质原因同解释遗传力估计值有关。尤其是遗传力指数，它表示在现存条件下，某个总体中某种特质的总方差中，由基因或遗传因素造成的方差比例。例如，美国城市中学生中斯坦福一比纳（Stanford-Binet）IQ 的遗传力为 0.70，这句话表示这些分数中所得出的方差的 70% 可归因于遗传差异，30% 可归因于环境差异。

遗传力指数可用不同的公式来计算（参见 Jensen, 1969; Loehlin, Lindzey & Spuhler, 1975），但是其基本数据都是对所考虑的特质中亲属之间相似性的量数。一种常用的方法是比较同卵（完全相同的）双生子和异卵（兄弟姐妹般的）双生子的智力测验相关。也使用比较一起抚养的同卵双生子的相关和在领养家庭中分开抚养的同卵双生子的相关。

解释遗传力估计值时应该注意几点。首先，由于未加评定



的环境因素的作用，亲属相似性的经验数据容易失真。例如，显而易见，同卵双生子共同的环境比异卵双生子更为相似（Anastasi, 1985, pp. 287 ~ 288; Koch, 1966），一起抚养的兄弟姐妹的环境在心理上也许大不相同（Daniels & Plomin, 1985）。另一个困难是，分开抚养的双生子不是像在理想的实验中那样随机分到不同的领养家庭；众所周知，领养家庭的安排是经过选择的，要考虑儿童和领养家庭的特点。因此，每对双生子的领养家庭环境可能具有相似性，足以说明测验分数之间的某种相关。也有数据表明，关于遗传力的双生子数据，也许不能随便地概化到总体，因为双生子容易遭受胎儿期损伤而导致严重的智力落后。一个样本中包括这类严重落后个案，可以大大增加双生子智力测验分数的相关（Nichols & Broman, 1974）。

除了有疑问的数据之外，遗传力指数还有其他内在的局限（参见 Anastasi, 1971; Hebb, 1970）。应当指出，在先前所引用论文的前面部分，詹森（1969, pp. 33 ~ 46）清楚地列举了这些局限。第一，遗传力这一概念适用于总体，而不适用于个体。例如，试图确定一名特定儿童的智力落后的原因时，遗传力指数就无能为力了。不管总体中遗传力指数的大小，该儿童智力落后的原因或是一种基因有缺陷，例如苯丙酮尿症（PKU），或是胎儿期大脑损伤，或是极端的经验剥夺。

第二，遗传力指数涉及其时得出它的总体。遗传条件或环境条件的任何变化，都会改变遗传力指数。试举一例，近亲繁殖的增加，例如在孤岛上，会减少可归因于遗传的方差，因而降低遗传力指数；另一方面，增加环境的同质性，则会减少可归因于环境的方差，因而提高遗传力指数。此外，一个总体之内所得出的遗传力指数不适用于分析两个总体的测验成绩的差异，例如不同的种族团体之间的测验成绩的差异。

第三,遗传力不是表示一种特质的可变性的程度。即使在某个总体中,一种特质的遗传力指数为 100%,但不能因此说环境对于那种特质的作用就是无关紧要的。一个极端的例子也许有助于说明这一点。假定在一个假设的成人社区里,每个人的饮食完全相同,所有人都得到同样数量的相同食物。在这个总体中,食物对于健康和体格条件的总方差的作用就为零,因为食物方差不能说明健康和体格方面的个体差异。但是,假如突然切断食物供给,这个社区中的入仍会死于饥饿。相反,改进饮食质量,结果就会普遍提高社区的健康水平。

不管各种总体中所得出的 IQ 遗传力指数的大小如何,一个明确证实的经验事实是: IQ 不是固定不变的;通过环境干预, IQ 可以改变。下一章将和纵向研究一起,讨论得出这个结论的一些证据。在识别 IQ 的加速环境和减速环境的特点方面,已经取得一些进展。儿童生活中偶然发生的环境变化和有计划的环境干预,都可以造成 IQ 的上升和下降。家庭结构的重大变化,家庭收入水平的急剧上升或下降,或被在一个领养家庭收养,都可以产生 IQ 的显著的提高或降低。

全世界都对发展智力的系统计划感兴趣,这种兴趣在 70 年代后期和 80 年代迅速增长,并且一直继续到现在。从 1980 年到 1987 年,出版《人类智力国际通讯》(*Human Intelligence International Newsletter*)就说明了这一点。这份通讯有一个国际编辑委员会,每季度提供一份关于认知研究及其教育应用的报告;另一个重要例子是由委内瑞拉政府所制定和系统实施的十年计划。这项计划包括从婴儿期到老年期“思维技能”发展的许多具体规划,它推动其他一些国家也开始实行相似的规划(Collins & Mangieri, 1992; Greenwald, 1982, 1984; Hermstein, Nickerson, Sanchez & Swets, 1986; Nickerson, 1988; Spitz, 1986; Sternberg, 1986)。

有人研究有计划的干预在婴儿和学龄前阶段的效果,第十二章将讨论这一问题。不过,也应当指出,越来越多的证据表明这类干预在人生的后面阶段的有效性。虽然关于学龄儿童的各种计划之规模,不如学龄前儿童的那般大,但它们得出了令人振奋的结果 (Bloom, 1976; Brown & Campione, 1986; Campione & Brown, 1987; Jacobs & Vandeventer, 1971; Olton & Crutchfield, 1969; Resnick & Glaser, 1976)。一些研究人员把注意力集中于更大的年龄阶段,研究大学生和专业学校的学生;他们报告学业成绩和学业能力倾向测验成绩两者都有显著提高 (Bloom & Broder, 1950; Whimbey, 1975, 1977, 1980)。关于老年人的研究,证实了训练干预之后学习和迁移的效果 (Willis, Blieszner & Baltes, 1981)。还有一些研究人员关注可教育的智力落后儿童和青少年,发现他们也有显著的提高 (Babad & Budoff, 1974; Budoff & Corman, 1974; Feuerstein, 1980; Feuerstein 等, 1987; Hamilton & Budoff, 1974; Rand Tannenbaum & Feuerstein, 1979)。<sup>[3]</sup>

这些计划训练可广泛适用于认知技能、解决问题的策略,以及高效的学习习惯。尤其强调自我监控或自我批评,这是取得有效作业的一个条件 (Flavell, 1979; Owings, Petersen, Bransford, Morris & Stein, 1980; Whimbey, 1975)。评价自己的作业,并且认识到自己理解什么和不理解什么,这是改进作业的重要的第一步。不成功的学习者,往往不能区分真正的理解与不准确的或表面上的理解。研究表明,学习困难儿童在自我批评和监控他们自己的认知方面,尤为不足 (Kotsonis & Patterson, 1980)。

在这些智力发展计划中所教授的各类认知技能的其他例子,第一章曾经引用过。在第一章中,将广泛适用的智力技能中的这类训练与仅仅限于对测验项目的辅导相对照。在那一方面可以看到,在评价智力发展计划中要研究的一个关键问题是,超

出训练的内容和场合，训练效果的迁移或概化的程度。另一个有关的问题是训练效果的持久性。

知识也是智力的一个组成部分，并且有助于个体准备学习较为高级的材料，进一步的考虑便是大年龄儿童或成人积累这种知识所需要的时间。研究表明，除了在十分初级的水平之外，解决问题的图式和概念是领域的或特殊的，换句话说，它们与特定的问题领域有关。解决问题的技能与个体在具体的知识领域之内所积累的、经过组织的内容贮存密切相关（Bransford, Sherwood, Vye & Rieser, 1986; Brown & Campione, 1986; Glaser, 1984; Larkin, McDermott, Simon & Simon, 1980a; Neimark, 1987; Resnick & Neches, 1984; Richardson, Angle, Hasher, Logie & Stoltus, 1996）。虽然大年龄个体具有高效的学习技术，能够比他或她儿童时更快建立这种知识贮存，但期望在短期的训练计划期间发生这件事则是不现实的。个体的年龄越大，要填补的知识空白也越大，不能认识到这一点，就可能对所有这类训练计划的功效感到失望，使自己缺乏信心。

## 动机和智力

把测验分为各种类型,这符合习惯,也提供方便,但是应该认识到,这类区分只是表面性的。解释测验分数时,不能脱离人格和能力倾向。个体在能力倾向测验上的成绩,以及他或她在学校里、在工作上或在其他任何情况下的表现,都会受到他或她的成就驱力、持久性、价值系统、是否有情绪障碍问题,以及传统上归类为‘人格’的其他特征之影响。

现在人们日益认识到学生的动机在学校学习中的作用

(Bloom, 1976, 第四章; Budoff, 1987; Feuerstein 等, 1987; J. G. Nichols, 1979; Renninger, Hidi & Krapp, 1992; R. E. Snow, 1989)。个体的兴趣、态度、学习者的自我概念等,都影响到他或她对学习任务的接受、学好的愿望、对教师的注意和用于这种任务的时间。研究表明,个体的这些反应都与教育成就显著相关 (Baron, 1982; Dreger, 1968; J. McV. Hunt, 1981)。

在较为基础的水平上,一种越来越多的共识是,研究能力倾向不能再脱离情感变量(Anastasi, 1985b, 1994; Izard, Kagan, & Zajonc, 1989; Kanfer, Ackerman, & Cudeck, 1989, 第 4 部分; Moore & Isen, 1990; Saklofske & Zeidner, 1995; Salovey & Sluyter, 1997; R. E. Snow, 1992; Spaulding, 1994; Sternberg & Ruzgis, 1994)。

暂时的情感状态对个体当前表现的影响也得到证实。人格特质对个体智力发展的方向和程度的累积影响更为重要。有关的数据来自若干类型的研究,包括长期的纵向研究 (Eichom, Clausen, Haan, Honzik & Mussen, 1981), 以及最近将结构方程模型应用于因果关系分析(Shavelson & Bolus, 1982)。这类研究证实,将关于动机和态度的信息与能力倾向测验的分数相结合,能够大大提高对个体后来智力发展的预测。

动机和其他情感变量对能力倾向发展起作用的途径之一,是通过相对于其他引起注意的活动而言,个体用于某一特定活动的时间量。根据对成就动机的约 25 年的研究,阿特金森及其同事(J. W. Atkinson, 1974; Atkinson, O' Malley & Lens, 1976)提出了一种表示能力、动机和环境变量的综合图式。这种图式中的一个关键概念是用于任务的时间 (time-on-task), 即个体用于一种活动的时间,例如学习或执行一种与工作有关的机能。动机既影响完成一种任务的效率,也影响相对于其他活动而言,用于这种任务的时间。作业的水平取决于个体的有关能力和他或她

将那些能力应用于当前任务的效率。最终成就或产物反映作业水平和花在工作上的时间的联合影响。

阿特金森图式的另一个重要组份是，任务成绩对个体自己的认知发展和动机发展的持久的、累积的影响。这一步表示对个体自己特质的反馈回路。其影响应该表现在该个体将来在测验上和现实生活中的成绩。对中学生的纵向资料进行计算机模拟和经验分析，都表明阿特金森图式的预测价值（Atkinson, 1974; Atkinson 等, 1976; Lens, Atkinson & Yip, 1979）。

注意控制可以加大用于任务的纯时间的影响。个体注意什么东西，注意的集中有多深，注意持续多久等，都对个体的认知发展起作用。注意的选择性导致有选择的学习，并且这种选择在面临相同情境的各个体之间有所不同。此外，这种有选择的学习可以影响不同能力倾向的相对发展，因而有助于形成各种各样的特质型式（Anastasi, 1970, 1983a, 1986b）。注意控制的若干方面，基本上能加强用于有关活动的�间的影响，因而增强它对能力倾向发展的影响。

人格和智力之间的关系是双向的。不仅人格特征影响智力发展，而且智力水平也可以影响人格发展。普兰特和米尼姆（Plant & Minium, 1967）在一项研究中所提供的启发性数据证实了这种关系。他们找到关于准备升大学青年的 5 个纵向研究，利用其中所收集的数据，按照智力测验分数在每个样本中各选取高低 25%。然后，根据已经对一个或多个样本所实施的人格测验，将这些对照组相比较。人格测验包括测量态度、价值观、动机、人际关系和其他非认知特质。这个分析的结果表明，能力倾向测验高分组较低分组具有经历更多‘心理上积极的’人格变化的强烈倾向。

个体在发展和使用能力倾向中所达到的成功，必然会影响

个体的情绪适应、人际关系、自我概念等。在自我概念中,我们能够最为清楚地看到能力倾向和人格特质的相互影响。儿童在学校里、在操场上或在其他情境中的成就,有助于形成他或她的自我概念;而这种自我概念在任何阶段都会影响儿童在螺旋型发展中后来的表现。在这一点上,自我概念起着一种个人自我实现预言的作用。

近年来,人们对研究婴儿发展中的情感因素的兴趣日益增强。若干研究将婴儿在人格变量上的行为所作的评定和使用韦克斯勒儿童智力量表修订版(WISC-R)以及斯坦福一比纳量表(Stanford-Binet)等工具对儿童后来的认知发展所作的评定作了一个比较,发现二者之间呈显著相关(Bims & Golden, 1972; R. B. McCall, 1976; Palisin, 1986; Yarrow & Redersen, 1976)。一般来说,在测验情境中表现出积极情感、兴趣活跃和易起反应的婴儿,可能在认知发展中学得更多,提高也更快。显然这得益于他们早期经验的积累。他们也可能在后来的学业活动中反应积极,这种活动涉及目标定向任务中与成人的相互作用。另一个优点源于这类婴儿行为对成人看护者的社会行为的影响,它转过来又会增加儿童的学习机会(Haviland, 1976; Wilson & Matheny, 1983)。

更具体地说,关于婴儿掌握环境的动机的研究,表明它与以后的智力测量有一些有价值的关系。婴儿掌握环境的行为包括观察、探索、操纵他或她环境。这种动机理所当然地应该是认知发展的一种基本的作用因素,实验文献也日益表明事实的确如此(Hmcir, Speller & West, 1985; White, 1978; Yarrow 等, 1984; Yarrow 等, 1983)。事实上,一些研究结果表明,儿童掌握环境动机的早期指标较能力的早期测量,也许能更好预测后来的智力。

关于婴儿的研究,导向情感发展和认知发展研究之间的良好关

系。它有助于在任何年龄水平解释测验结果时，更为综合地利用情感资料和认知资料。

## 智力的因素分析

由于对人类智力的性质和组成感兴趣，人们才进行鉴别心理特质的心理学研究。<sup>[4]</sup>这类研究始于计算一个被试样本在多个能力测验上所得分数的彼此相关，然后对相关表进行进一步的数学分析，以便鉴别各测验之间的共同因素或特质。目前适用于这种目的的各种方法，总称为因素分析(factor analysis)。

**因素矩阵** 因素分析的主要目的是减少变量或维度的数目，从而简化对数据的描述。因此如果我们发现，5个因素足以说明一套20个测验中的共同方差，则出于大多数目的，我们能够用5个分数来代替最初的20个分数而不会牺牲重要的信息。通常的做法是，从最初的测验中保留对每个因素提供最佳测量的那些测验。

因素分析的各种技术首先都列出一张一组测验之间的全部相关表。这种表称之为相关矩阵(correlation matrix)。每一种因素分析最后都得出一个因素矩阵，就是一张表明在每个测验中每个因素的权重即负荷的表格。一个假设的只包括两个因素的因素矩阵见表11-1最上面一行列出这两个因素，对应的各行中给出它们在10个测验中的权重。

现在已经推导出若干不同的方法，可以把一组变量分析为几个共同因素。早在1901年，皮尔逊(Pearson)就指出了这类分析的方法。斯皮尔曼(Spearman, 1904, 1927)是现代因素分析的先驱。美国的凯利(T. L. Kelley, 1935)和瑟斯顿(Thurstone, 1947b)



表 11-1 一个假设的因素矩阵

| 测 验        | 因素 I | 因素 II |
|------------|------|-------|
| 1. 词汇      | .74  | .54   |
| 2. 类推      | .64  | .39   |
| 3. 句子完成    | .68  | .43   |
| 4. 句子排列    | .32  | .23   |
| 5. 阅读理解    | .70  | .50   |
| 6. 加法      | .22  | -.51  |
| 7. 乘法      | .40  | -.50  |
| 8. 算术问题    | .52  | -.48  |
| 9. 方程关系    | .43  | -.37  |
| 10. 数字系列完成 | .32  | -.25  |

与英国的伯特(Burt, 1941) 为发展这种方法作出了许多贡献。其他许多人也提出不同的方法加以改进。高速计算机的普及, 使得人们采用更为精致和复杂的技术。虽然在最初的假设中有所不同, 但其中大多数方法得出相似的结果。关于因素分析的具体方法的介绍, 读者可以参阅科姆里和李 (Comrey & Lee, 1992) 或洛林(Loehlin, 1992) 所编的教科书。关于基本的概念和方法的简单综述, 则参阅金和米勒(Kim & Mueller, 1978a, 1978b) 或克兰(P. Kline, 1993) 的著作。

讨论因素分析的数学基础或计算方法, 这超出本书的范围。不过, 理解因素分析的结果, 并不一定限于掌握这种专门方法的那些人员。即使不知道怎样计算因素负荷, 也有可能明白怎样利用一个因素矩阵去鉴别和理解各个因素。然而, 熟悉一些其他的概念和术语, 有助于有效阅读因素分析的研究报告。

参照坐标轴 习惯上使用几何方法把因素表示为参照坐标

轴,据此能够标出每个测验的位置。图 11-1 是对这种方法的说明。在该图中,参照相当于 I 轴和 II 轴的两个因素,逐一标出表 11-1 中 10 个测验的位置。因此,沿 I 轴移动 0.74 个距离而沿 II 轴移动 0.54 个距离,就标出表示测验 1 的点。使用表 11-1 中给出的权重,以相同的方法标出相当于其余 9 个测验的各点。虽然在因素 I 上的所有权重都是正的,但应当指出,在因素 II 上,一些权重是正的,另一些权重是负的。在图 11-1 中也能够看到这一点,测验 1 至测验 5 集中在图的一个部分,而测验 6 至测验 10 则集中在图的另一个部分。

在这一点上应当指出,参照坐标轴的位置不是由数据固定的。最初的相关表仅仅决定各个测验彼此之间的位置(即图 11-1 中的各点)。使用任何位置的参照坐标轴,都能够标出相同

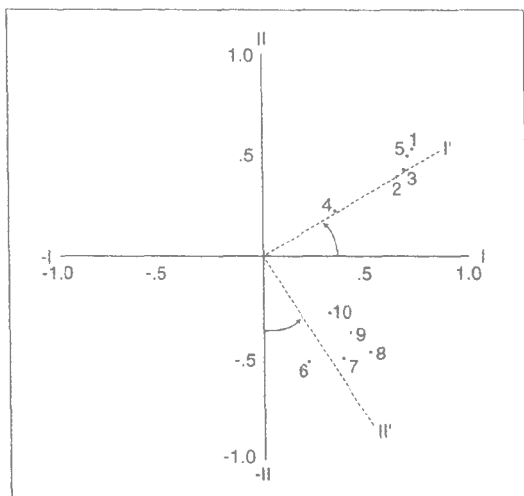


图 11-1 一个假设的因素模式,说明两组因素在 10 个测验上的权重

的各点。由于这种原因,因素分析者通常将坐标轴旋转,直到获得最满意的、最容易解释的形式。这是一种合理的方法,有点类似于从芝加哥而不是从格林威治去测量经度。

将图 11-1 中的参照坐标轴旋转到用虚线所表示的  $I'$  和  $II'$  的位置。<sup>[5]</sup> 根据瑟斯顿的正流形 (positive manifold) 和简单结构 (simple structure) 这两条准则,进行这种旋转。前者要求将坐标轴旋转到这样一个位置:删除所有较大的负权重。大多数心理学家认为,负的因素负荷逻辑上不适用于能力倾向测验,因为负的负荷意味着,个体在特定因素中的评定越高,他或她在测验上的成绩就越差。简单结构这一准则实质上意味着,每一测验在尽可能少的因素上具有负荷。<sup>[6]</sup> 这两条准则旨在得出能够最容易、最明确加以解释的诸因素。如果一个测验在单一因素上具有高负荷,而在其他任何因素上都没有什么负荷,则我们检查一下测验内容,就能够认识到这种因素的性质。相反,如果测验在 6 个因素上都有中度至低度的负荷,它就不能告诉我们多少关于其中任何一个因素的性质。

在图 11-1 中能够看到,在经过旋转的坐标轴上,所有的言语测验(测验 1 至测验 5)都处于或非常接近于  $I'$  轴。同样,所有的数字测验(测验 6 至测验 10)都集中于  $II'$  轴附近。表 11-2 给出参照经过旋转的坐标轴所测量的新的因素负荷。读者准备一把纸“尺”,标度单位和图 11-1 的相同,便很容易证实这些因素负荷。使用这把尺子,能够参照经过旋转的坐标轴来测量距离。表 11-2 的因素负荷中没有负值,除了可归因于取样误差的微不足道的数值之外。所有的言语测验在因素  $I'$  上都具有高负荷,在因素  $II'$  上实际上没有负荷。另一方面,所有的数字测验在因素  $II'$  上都具有高负荷,在因素  $I'$  上则具有较低的、可忽略的负荷。两种因素的鉴别和命名,以及每个测验的因素构成

的描述,由于使用参照坐标轴的旋转而变得简化了。在实际工作中,因素的数目往往多于 2,这种情况使得几何表示和统计分析复杂一些,但是基本方法没有改变。

表 11-2 经过旋转的因素矩阵

| 测 验        | 因素 I' | 因素 II' |
|------------|-------|--------|
| 1. 词汇      | .91   | -.06   |
| 2. 类推      | .75   | .02    |
| 3. 句子完成    | .80   | .00    |
| 4. 句子排列    | .39   | -.02   |
| 5. 阅读理解    | .86   | -.04   |
| 6. 加法      | -.09  | .55    |
| 7. 乘法      | .07   | .64    |
| 8. 算术问题    | .18   | .68    |
| 9. 方程关系    | .16   | .54    |
| 10. 数字系列完成 | .13   | .38    |

(资料根据图 11-1)

一些因素分析者使用一种理论模型作为坐标轴旋转的指南。一些因素分析者还考虑不变性,即在独立的但可作比较的各研究中确定相同的因素。目前,因素分析愈来愈多地被用于证实研究而不是探索研究。关于不同变量在完成任务中的作用,理论上提出了一种模型(参见 Loehlin, 1992),在评价这种模型时,往往把因素分析和结构方程模型相结合(参见第五章)。

**因素的解释** 一旦计算出经过旋转的因素矩阵,我们便能够进行因素的命名和解释,这一步要求心理学洞察力而不是统计训练。为了认识某个特定因素的性质,我们只要检查一下在

该因素上具有高负荷的各测验，并且尝试发现它们所具有的共同的心理过程。在某个因素上具有高负荷的测验越多，我们就越能清楚确定该因素的性质。例如，在表 11-2 中，显而易见的，因素 I' 为言语因素，因素 II' 为数字因素。

表 11-2 中所给出的因素负荷也表示每个测验与因素的相关。<sup>[7]</sup>回忆一下，这个相关就是测验的因素效度(第五章)。根据表 11-2，我们能够说，词汇测验测量言语因素时的因素效度为 0.91，加法测验测量数字因素时的因素效度为 0.55，等等。显然，前面 5 个测验测量数字因素时没有多少效度，后面 5 个测验测量言语因素时实际上也没有效度。

**测验的因素构成** 因素分析的基本原理之一是，测验的总方差等于共同因素（与其他测验共有）的方差和独特因素（该测验独有）的方差之和，再加上误差方差。我们已经碰到测验分数分析中的误差方差（第四章）。例如，如果某个测验的信度系数为 0.83，我们得出结论，该测验上分数方差的 17% 为误差方差（ $1.00 - 0.83 = 0.17$ ）。通过因素分析，我们能够进一步划分对测验分数起作用的方差的来源。

让我们关注表 11-3 中所列出的两个假设的测验。表中给出每个测验在言语(V)因素、数字(N)因素和推理(R)因素中的因素负荷，也给出测验的信度系数。因为因素负荷也表示测验和因素之间的相关，因素负荷的平方就告诉我们测验和该因素之间的共同方差的比例。在表的最后部分，将每一因素负荷平方，以便说明该因素对测验分数总方差的贡献比例。我们因此得出，算术推理测验方差的 16% 归因于言语因素，30% 归因于数字因素，36% 归因于推理因素。只要从总方差中减去信度系数，就得出最后一列的误差方差（ $1.00 - 0.90 = 0.10$ ）。所剩下的部分就表示该测验的特殊性，即进行因素分析时，该测验与其

表 11-3 测验分数的方差来源

| 测验      | 共同因素负荷 |     |     |     | 信度系数 | 贡献比例 |     |     |      |      |
|---------|--------|-----|-----|-----|------|------|-----|-----|------|------|
|         | V      | N   | R   |     |      | V    | N   | R   | 独特方差 | 误差方差 |
| 1. 算术推理 | .40    | .55 | .60 | .90 |      | .16  | .30 | .36 | .08  | .10  |
| 2. 乘法   | .10    | .70 | .30 | .85 |      | .01  | .49 | .09 | .26  | .15  |

他任何测验都不共有的其“真”方差的那部分。关于算术推理测验,我们得出:

$$0.16 + 0.30 + 0.36 + 0.10 = 0.92$$

$$1.00 - 0.92 = 0.08$$

对于表 11-3 中两个测验的方差来源,图 11-2 进行图形分解。

任何个体在这两个测验上的成绩,都取决于他或她所具有的每一种有关能力即因素的总量,也取决于这些因素在特定测验中的相对权重。因此,假如我们得到个体在 V、N 和 R 因素中以相同单位所表示的分数,把它乘上对应的因素负荷,我们就把

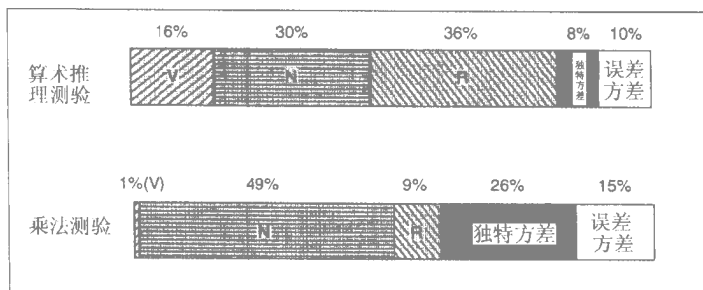


图 11-2 两个假设测验的共同方差、独特方差和误差方差的百分比(资料根据表 11-3)

每一分数加权。这些乘积之和便是个体测验分数的估计值。测验的独特因素和误差因素的贡献越小，这个估计值就越准确。

在表 11-3 给出的例子中，如果某个体 V 上的评定很高，则他在算术推理测验上的得益远大于乘法测验。事实上，这种得益高达 4 倍，因为 V 因素在算术推理测验中的权重是乘法测验的 4 倍（0.40 对 0.10）。3 个共同因素中，N 对乘法测验的影响最大（负荷为 0.70），R 对算术推理测验的影响最大（负荷为 0.60）。

**因素负荷和相关** 因素分析的第二条基本原理是关于变量之间的相关和因素负荷之间的关系。任何两个变量之间的相关等于它们共同因素负荷的乘积之和。因为每个变量的独特因素和误差因素都是独有的，它们对变量之间的相关不起作用。任何两个变量之间的相关仅仅取决于这两个变量共有的因素。两个变量中共同因素的权重越大，变量之间的相关也越高。把两个测验中 3 个共同因素的负荷彼此相乘，然后把乘积相加，就能够得出表 11-3 中所给出的两个测验之间的相关，如下面所示：

$$r_{12} = 0.40 \times 0.10 + 0.55 \times 0.70 + 0.60 \times 0.30 = 0.60$$

**斜交坐标轴和二级因素图** 图 11-1 中所使用的坐标轴称之为直角坐标轴(orthogonal axes)，因为它们互成直角。有时，使用斜交坐标轴(oblique axes)能够较好地拟合一组测验的位置。在这种情况下，各因素本身相关。一些研究人员认为，总是应该使用直角的即非相关的因素，因为它们更简单、更清楚地描述特质的关系。另一些研究人员则认为，当斜交坐标轴与数据拟合得更好时，就应该加以使用，因为最有意义的范畴不一定是非相关的。身高和体重就是一个明显的例子。众所周知，虽然身高和体重是高相关，但测量体格时，它们证明是有用的范畴。

我们能够对测验之间的彼此相关进行统计分析；当各因素

本身相关时,我们也能够对因素之间的彼此相关进行同样的统计分析。换句话说,我们能够‘将因素再分解因素’(factorize the factors),并且得出二级因素(second-order factor)。关于能力倾向变量和人格变量的一些研究都遵循这一过程。某些研究能力倾向测验的研究人员得出一种单一的二级一般因素(single second-order general factor)。一般来说,美国因素分析者首先通过群因素来解释尽可能多的共同方差,然后认为一般因素就是二级因素,如果数据证明它是正确的话。另一方面,英国心理学家通常从一般因素开始,他们把共同方差的大部分归因于一般因素,然后研究群因素,解释剩余的相关。这些程序上的差异反映理论观点的差异,下面一节就要讨论这一问题。

## 特质组织的理论

半个多世纪以来,为了努力理解各种各样人类活动中所包括的能力的性质和组织,因素分析的统计技术得到广泛应用。不过,这种技术仍然与认知能力即‘智力’研究的联系最为密切,因素分析也来源于此。最近,有人对所有收集到的关于能力的因素分析研究作了一番全面评述,这是对这类研究的一次重要总结(Carroll, 1993)。此番总结包括 70 多年的研究,远远超出文献综述;它也包括对原始研究中的 450 多套数据的再分析,以及在历史框架之内,对不同的智力理论模型所作的叙述和评价。我们在本节中仅仅考虑一些众所周知的智力理论,之所以选择这些理论,是因为它们对测验编制和使用的影响。

二因素论(Two-Factor Theory)建立在测验分数的统计分析基础之上的第一种特质组织理论,是由英国心理学家查尔斯·斯



皮尔曼 (Charles Spearman, 1904, 1927) 所提出的二因素论。在最初的阐述中, 这种理论认为, 所有智力活动都具有一种单一的共同因素, 称之为一般因素即  $g$  因素。此外, 这种理论假设许多特殊因素即  $s$  因素, 每一种特殊因素严格地为某一活动所特有。

任何两种机能之间的正相关因此归因于  $g$  因素。两种机能中“包含的”  $g$  因素越多, 它们之间的相关就越高。另一方面, 特殊因素的存在, 往往降低机能之间的相关。

虽然这种理论假设两类因素, 即一般因素和特殊因素, 但只有单一的  $g$  因素才解释相关。因此, 与其他的特质关系理论不同, 它能够被更为确切地描述为单因素论, 虽然最初的名称仍保留着。图 11-3 说明按照这种理论解释测验之间相关的基础。在该图中, 测验 1 和测验 2 彼此高相关, 因为每个测验包含的  $g$  因素很多, 如阴影区域所示。每个测验的白色区域表示特殊因素方差和误差方差。测验 3 和其他两个测验之一都是低相关, 因为它包含的  $g$  因素很少。

根据二因素论必然得出, 心理测验的目的应该是测量每一个体的  $g$  因素的量。如果所有能力之中都贯穿着这种因素, 则它为从一种情境预测另一种情境中个体的表现提供了惟一的基础。测量特殊因素则是无用的, 因为根据定义, 每一种特殊因素仅仅在单一活动中起作用。因此, 斯皮尔曼提出, 应该用包含大量  $g$  因素的单一测验来代替智力测验中所见到的一组异质项目。他建议, 处理抽象关系的测验也许是对  $g$  因素的最佳测量, 因此它能够适用于这种目的。为测量  $g$  因素而编制的测验例子包括‘瑞文渐进推理测验’(Raven's Progressive Matrices)和‘卡特文化公平智力测验’(Cattell's Culture Fair Intelligence Test)。

斯皮尔曼从一开始就认识到, 二因素论必然有一定的适用范围。当相比较的智力活动相似时, 除了可归因于  $g$  因素的相

关之外,还可以产生某种程度的相关。因此,除了一般因素和特殊因素之外,可能还有另一类中间因素,既不像  $g$  因素那样一般,也不像  $s$  因素那样特殊。这样一类因素称之为群因素(group factor)。它是一群智力活动而不是所有智力活动共同具有的因素。在对其理论的早期阐述中,斯皮尔曼承认,可能存在十分狭窄的、可忽略不计的群因素。根据他的几个学生后来的研究,他认为还应包括一些广泛得多的群因素,例如算术、机械、言语等能力。

**多因素论(Multiple-Factor Theories)** 美国关于特质组织的主要观点建立在早期因素分析研究的基础之上,这种观点提出若干中等广度的群因素,每个群因素在不同的测验上具有不同的权重。例如,言语因素在词汇测验中的权重较大,在言语类推测测验中的权重较小,在算术推理测验中的权重则更小了。图 11-4 按照多因素论模型来说明 5 个测验之间的彼此相关。测

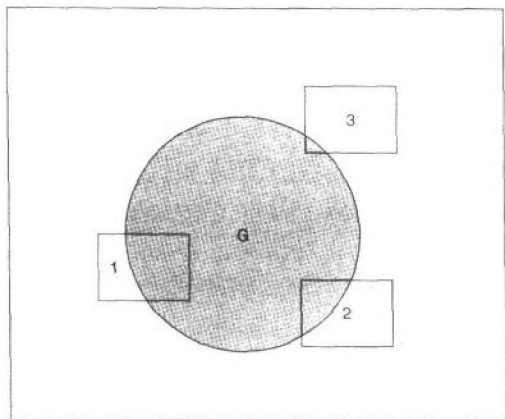


图 11-3 二因素论的相关模型

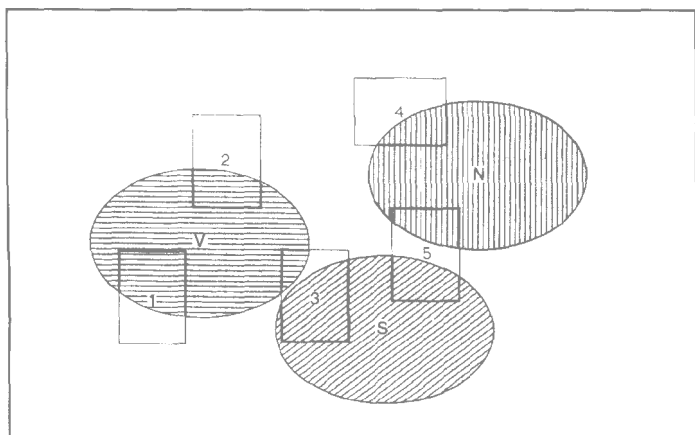


图 11-4 多因素论的相关模型

验 1、测验 2 和测验 3 这三者之间的彼此相关，是它们的共同负荷言语因素 (V) 造成的。同样，测验 3 和测验 5 之间的相关是空间因素 (S) 造成的，测验 4 和测验 5 之间的相关是数字因素 (N) 所造成的。测验 3 和测验 5 的因素构成较为复杂一些，其中每个测验在一个以上因素上具有较大负荷：测验 3 中具有 V 因素和 S 因素，测验 5 中具有 N 因素和 S 因素。根据前面一节中所讨论的因素分析的第二条基本原理，我们还可以讨论一下测验彼此相关的相对大小。例如，测验 3 和测验 5 的相关高于测验 3 和测验 2 的相关，因为测验 3 和测验 5 中 S 因素的权重(斜纹区域)大于测验 2 和测验 3 中 V 因素的权重(横纹区域)。

凯利《人类智力的十字路口》(Crossroads in the Mind of Man, 1928)一书的出版，为许多探求特定的群因素的研究铺平了道

路。凯利所提出的主要因素是：空间关系的操作、数字能力、言语材料能力、记忆、速度等。后来的一些研究人员使用前面一节中所讨论的现代因素分析模型，对凯利的 5 种群因素加以修改和扩充。

多因素论的主要代表人物之一是瑟斯顿。他根据自己及其学生的广泛研究，提出了约 12 个群因素，他称之为‘基本心理能力’(primary mental abilities)。在瑟斯顿的研究以及其他独立研究人员的研究中得到证实最多的群因素如下：

V. 言语理解(verbal comprehension) 阅读理解、言语类推、句子排列、言语推理、谚语配对之类测验中的主要因素。使用词汇测验测量该因素最为适当。

W. 语词流畅(word fluency) 字谜游戏、押韵、列举某种类型的单词(例如，男孩的名字、以 T 开头的单词)之类测验中所得出的因素。

N. 数字(number)：简单算术运算的速度和准确性。

S. 空间(space)：该因素可能表示两种不同的因素。一种包括知觉固定的空间关系或几何关系；另一种是操作性想象，必须想象经过变化的位置或变换(McGee, 1979; Portegal, 1982)。

M. 联想记忆(associative memory)：主要指在配对联想中要求机械记忆的测验中找到的因素。有证据表明，该因素反映利用记忆支撑物的程度(Christal, 1958)。这种证据反对所有记忆测验中存在一种广泛的因素。一些研究人员提出时间顺序记忆和空间位置记忆等其他范围有限的记忆因素。

P. 知觉速度(perceptual speed)：迅速而准确地掌握视觉细节、相同性和不同性。该因素可能与凯利和其他早期研究人员所鉴定的速度因素相同。这是知觉任务中后来所鉴定的若干因素之一(Thurstone, 1944)。

I(或 R). 归纳推理(或一般推理)(induction or general reasoning): 该因素的鉴定最为模糊。瑟斯顿最初提出归纳艺术因素和演绎因素。后者由三段论推理测验加以测量最为适当,前者则最好由要求被试找出某种规则的测验,如数字系列完成测验加以测量。不过,关于演绎因素的证据比归纳因素少得多。此外,其他研究人员提出一般推理因素,由算术推理测验加以测量最为适当。

应当指出,一般因素、群因素、特殊因素之间的区分没有初看时那么重要。如果成套测验中测验的数目或种类很少,则单一的一般因素可以说明测验之间的所有相关。但是,如果同一批测验被编入由一组异质性测验构成的一个大型成套测验,则最初的一般因素也许成为一种群因素,若干测验而不是所有测验具有这种共同因素。同样,某种因素在最初的成套测验中仅由一个测验所代表,但在大型成套测验中由若干测验共同具有。这样一种因素在最初的成套测验中被鉴定为特殊因素,但在综合成套测验中则成为群因素。根据这种观点,对于各特殊领域进行深入的因素研究而得出许多因素来取代每个领域最初所鉴定的一二种基本心理能力,这就不足为奇了。在研究言语、知觉、记忆、推理等测验时,情况正是如此。

因素研究已经得出的因素倍增令人迷惑。迄今为止,不同的研究人员所报告的认知因素的数目远远超过 100。通过不同研究人员所报告的因素交叉鉴定而得出因素种类的量,往往被给予不同的名称(Ekstrom, French & Harman, 1979; French, 1951; Harman, 1975)。当相比较的各研究共同具有若干测验时,就能够完成这类交叉鉴定。为了使这一过程尽可能方便,一组因素分析人员编制了一套“参照测验”来测量其时所鉴定的基本能力倾向因素。教育测验服务中心(Educational Testing Service)所发

行的这套测验(Ekstrom, French, Harman & Dermem, 1976; ETS kit, 1976), 使打算进行因素分析研究的不同的研究人员比较容易在他们的成套测验中包括一些共同测验。

显而易见,即使在作出简化和调整的努力之后,因素的数目仍然是很多的。人类行为本来就是多样、复杂的,期望用一打左右的因素去描述人类行为是不现实的。不过,出于具体目的,我们能够选取性质和广度均为适当的因素。例如,假如我们选拔申请人从事一项困难的、高度专业化的机械工作,我们也许会测量相当狭窄的、与工作要求相配的知觉因素和空间因素。另一方面,假如我们选拔大学生,最为有关的则是一些广泛的因素,例如言语理解、数字能力、一般推理等。

**智力结构模型** 一些因素分析者通过将特质组成某种系统图式,从而试图简化对特质关系的描述。根据 20 多年的因素分析研究,吉尔福德(Guilford, 1967, 1988; Guilford & Hoepfner, 1971)提出一种箱状模型,称为智力结构(Structure of Intellect, SI)模型,如图 11-5 所示。该模型<sup>[8]</sup>按 3 个维度把智力特质分类:

**操作(operation)**——指被试所做的事情。操作包括认知、记忆登录、记忆保持、发散作业(在创造活动中很突出)、集中作业和评价。

**内容(content)**——指完成操作所根据的材料或信息的性质。内容包括视觉、听觉、符号(例如字母、数字)、语义(例如单词)和行为(关于他人的行为、态度、需要等的信息)。

**产物(product)**——指被试处理信息的形式。产物分类为单元、类别、关系、系统、转换和蕴涵。

由于这种分类包括  $6 \times 5 \times 6$  个类型,所以模型中有 180 个单元。每个单元至少包括一种因素即能力,一些单元可能包括一种以上的因素。按照这三个维度描述每种因素。对 SI 模型

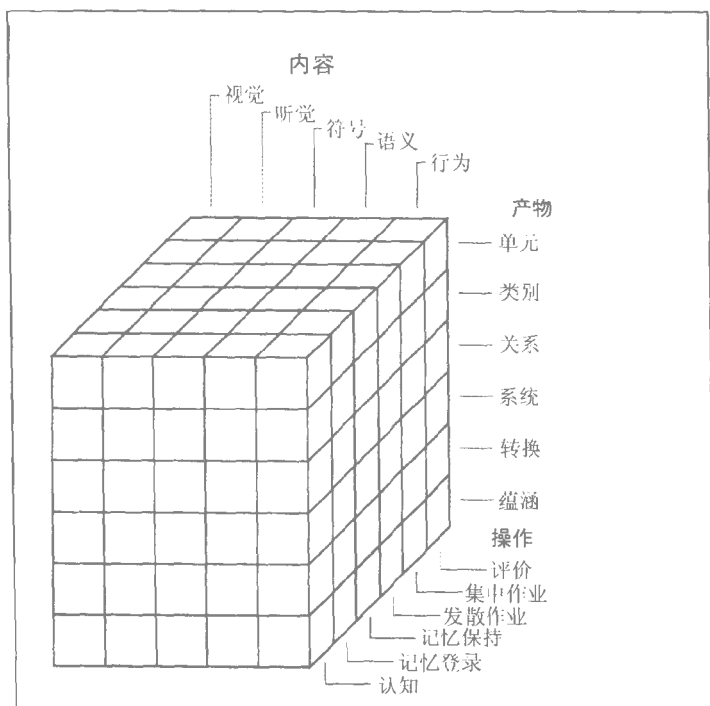


图 11-5 智力结构三维模型(根据《教育和心理测量》(吉尔福德, 1988 年修订, p.3)。

的 20 年协作研究计划——能力倾向研究计划 (Aptitudes Research Project) 一经完成, 吉尔福德及其同事就鉴定了 98 种所预期的因素 (Guilford & Hoepfner, 1971)。在吉尔福德和霍夫纳 (Hoepfner) 的著作中 (1971), 按字母顺序列出并简单描述在这 20 年计划期间所编制的许多测验。

虽然根据 SI 模型至少直接设计了一个测验,即智力结构—学习能力测验(Structure of Intellect Learning Abilities Test, Meeker, Meeker & Roid, 1985),但模型本身对一般测验的编制和应用没有多少影响。<sup>[9]</sup>应当记住,与各种特质组织的模型一样,SI 模型提供一种图式来表示所得出的变量之间的相关。因为在坐标轴的旋转中所使用的方法的关系,所以经验上证实 SI 模型,并不排除其他模型(Carroll, 1972; Horn & Knapp, 1973)。对相同数据进行不同的旋转可同样拟合其他模型。事实上,对吉尔福德原始因素分析的数据进行再分析,结果发现,其他模型能比 SI 模型更好地拟合数据,并且与理论解释和实践解释更相一致(Bachelor 1989; Carroll, 1993)。

另一方面,SI 计划的一种有利的间接影响,是将注意集中到因素鉴定中操作和内容之间的区分。这种区分有利于阐明通过因素分析所鉴定的因素和通过认知心理学所研究的过程,以及两者之间的关系。本章最后一节在论述特质的性质和发展时,将更为详细地讨论这一点。SI 模型的另一种间接影响,便是区分集中思维和发散思维。发散思维这一概念,其含义为非典型行为,已在创造力分析中广泛采用。不过,试图编制与内容领域无关的发散思维测验,一般来说被证明是不成功的。发散思维和创造力以及它们所适用的领域或课题,诸如一门特定的科学(例如生物学或物理学),或者一种特定的艺术形式(例如音乐或雕刻),有着特殊的关系。此外,创造性的观念或产物在其特定的文化之内必须是有意义的或有用的。如果只有单纯的发散性而没有质量上的提高,则不能构成创造力。关于创造力本身这一广泛课题的研究,在表现范围和形式方面,都进展迅速。90 年代出版的、艾尔伯特(R. S. Albert, 1991 ~ 1994)编辑的创造力研究丛书,正说明人们对这一领域的兴趣在不断地增加。这



套丛书中所包括的各本书涵盖了这个课题的广泛内容，从创造力的各种表现到培养以及发展创造行为的方法。

层次理论( Hierarchical Theories ) 一些英国心理学家，例如伯特( Burt, 1949)和 弗农( Vernon, 1960)，以及美国心理学家汉弗莱斯( Humphreys, 1962) 都提出了另一种因素组织图式。图 11-6 的图解说明弗农对这种系统的应用。弗农把斯皮尔曼的  $g$  因素放在最高层次。第二层次是两个大群因素，相应于言语—教育(v:ed) 能力倾向和实践—机械(k:m) 能力倾向。这两个大因素群可以再作划分。例如，言语—教育因素分为言语、数量等次因素。同样，实践—机械因素分为机械—信息、空间能力、心理运动能力等次因素。通过进一步分析，还能鉴别出范围更小的次因素，例如言语作业的因素。在最低层次，则是特殊因素。在后来精心设计的模型中，在第三层次，尤其关于教育成就和职业成就，弗农( 1969) 包括了因素之间某些较为复杂的相互联系和交叉作用。例如，科学技术能力与空间能力和机械信息能力均相联系；数学能力与空间能力和数量能力均相联系，并且通过归纳因素更为直接地与  $g$  因素相联系。

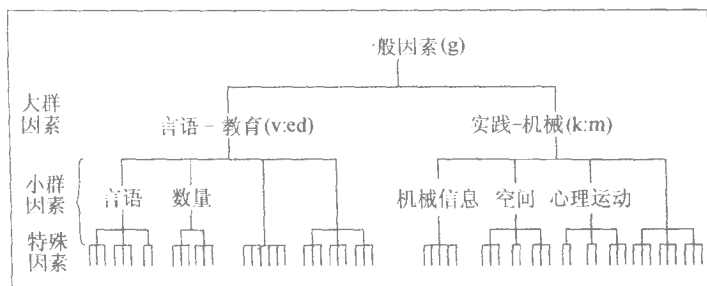


图 11-6 能力的层次结构模型 (根据 P.E. 弗农, 1960, p. 22)

汉弗莱斯(1962,1970)也提出一种层次模型,作为应付因素激增的一种方法。不过,他不把因素的任何单一的层次看成是基本的,而是建议每个测验编制者或使用者都应该选择最适合于自己目的的那个层次。此外,汉弗莱斯认识到,根据内容、过程等方面,可以把单一的测验分为一个以上的层次。他提出,为了测量任何一个方面,应该编制对于其他各个方面来说是异质的测验。例如,如果我们对个体解决类推问题的能力感兴趣,则应该使用一个包括言语类推、数量类推、图形类推和空间类推的测验。同样,如果我们希望测量言语能力,则应该使用各种各样的项目类型,例如词汇、类推、系列完成等。这种方法与吉尔福德所使用的方法不同,吉尔福德为三维模型中每个同质单元而探索各自独立的因素(和测验)。不过,在后来的研究中,吉尔福德(1981)使用一种部分层次组织去鉴定最初的智力结构三维模型中一些因素之间的较高级因素。

出于理论和实践两种目的,智力层次模型正越来越广泛地为人们所接受(参见 Anastasi, 1992a, 1994; Carroll, 1993; Gustafson, 1984, 1989; Lubinski & Dawis, 1992)。作为一种特质关系的理论模型,它调解单一的一般因素(斯皮尔曼的  $g$  因素)与多因素模式的分歧。在方法论上已经证明,多因素解法和层次解法在数学上是等价的,并且可以相互变换(Harman, 1976, 第十五章; Schmid & Leiman, 1957)。斜交坐标轴解法(相关因素)使人们得出层次模型,我们能够将斜交坐标轴解法变换为直角坐标轴解法(不相关因素)。在直角坐标轴解法中,二级因素作为不同宽度的一级因素出现。范围广大的因素比范围狭小的因素在更多的变量上具有负荷。

从实践的观点来看,根据层次模型所编制的测验的主要优点是,这类测验把综合的能力倾向内容范围与使用的灵活性相

结合。为了适应不同的测验目的，测验使用者可以选择单一的成套测验的总分数，或者选择测量狭义因素的各组测验中的一个或几个分数。在某种条件下，各个分测验的分数有助于鉴别专门技能的长处或不足。

第二章曾讨论过，作为传统的智力测验和多项能力倾向成套测验相统一的必然性发展，层次模型对能力测验的编制和使用的影响正在不断增强。讨论个别测验的第八章和讨论团体测验的第十章中都有这类层次方法的例子。综合的内容范围与灵活的使用相结合的最明显的例子有：用于个别测验的“区别能力量表”(Differential Ability Scales, Elliott, 1990b)和用于团体测验的“多维能力倾向成套测验”(Multidimensional Aptitude Battery, Jackson, 1994b)。体现这类测验中各评分水平灵活性的一个值得注意的特点是，为在所选择的水平上解释分数而提供各种常模。没有这类各自水平的常模，就不能在不同层次水平上对测验参加者的表现进行适当评价。

## 特质的性质和发展

如果我们认识到，通过因素分析所鉴定的特质仅仅表示行为测量之间的相关，那么，对于不同的研究人员得出不同的特质组织模型，就不会感到困惑了。特质不是基础的实体或偶然因素，而是描述的范畴。因此可以想象，不同的分类原则可以应用于相同的数据材料。英国的汤姆森(Thomson, 1948)、伯特(1941, 1944)和弗农(1960)，以及美国的特赖恩(R. C. Tryon, 1935)，在他们的早期论著中，都明确地把因素这一概念作为描述的范畴。所有这些作者都关注多种多样的行为元素，认为它们可以通过

遗传联系或经验联系而组成群集。

**经验史** 现在人们日益认识到个体的经验史在智力发展和群因素形成中的作用( Anastasi, 1986b; Greeno, 1989)。不仅不同能力的行为表现水平, 而且行为表现组成不同特质的方式, 都受到经验背景的影响。研究发现, 因素模型的差异与不同的文化或亚文化、社会经济水平、学校课程类型等相联系( 参见 Anastasi, 1970, 1983a, 1986b, 1994; Vernon, 1969)。因素模型经过一段时间, 有所变化也与此有关, 其中包括长期变化——这也许反映日常生活经验的累积影响——和短期变化——这由练习和实验控制的其他学习经验所造成( Baltes, Cirmelius, Spiro, Nesselroade & Willis, 1980; Birren, Cunningham & Yamamoto, 1983; Fleishman, 1972; Fleishman & Munford, 1989; Khan, 1970, 1972; Reinert, 1970)。对动物的研究也得出启发性的证据, 在实验中控制早期经验而形成各种因素( Whimbey & Denenberg, 1966)。

同一个客观任务的因素构成在不同经验背景的个体之间有所不同。可以发现, 这些个体差异的一种原因在于使用不同的方法去执行同一个任务。例如, 言语能力高度发展的个体往往使用言语中介去解决机械问题或空间问题; 而机械经验为主的个体往往使用知觉方法或空间方法去解决同一个问题。弗伦奇(French, 1965) 提供了有关的证据, 他发现, 按照解决问题的典型风格将被试分组, 则相同测验的因素构成在各组之间有所不同。在弗雷德里克森(C. H. Frederiksen, 1969) 关于记忆单词时使用认知策略的一项研究中, 也能够发现启发性的证据。在学习过程中, 个体可能改变策略的选择, 因而改变他们所完成的任务的因素构成。研究证明, 在教学过程中, 经过较长时间会发生能力倾向要求的相似变化( R. B. Burns, 1980)。

**特质形成的机制** 学习定势和训练迁移这两个熟悉概念,

是因素出现的一种机制( Carroll, 1966; G. A. Ferguson, 1954, 1956; Simon, 1990; Whiteman, 1964)。当呈现同样类型的新问题时, 学习定势的建立使得个体能够高效地学习。在哈洛 ( Harlow, 1949, 1960) 对猴子的经典实验中, 那些已经解决要求区分某些形状 (例如三角形和圆形) 的问题之后的猴子比没有这类经验的猴子能更快学会区分其他形状。这些猴子已经建立起区分形状的学习定势, 知道碰到一个新问题时应该如何面对。猴子因此 “已经学会怎样去学习” 这类问题。

同样, 通过阅读和算术计算等正规学习所发展的许多技能, 也能应用于后来多种多样的学习情境。高效而系统的解决问题的技术, 同样能够被用于解决新问题。在许多不同任务的完成中, 都会反映个体获得这些技能的程度差异; 对这些任务进行因素分析时, 这些广泛适用的技能就作为群因素出现。迁移效果的广度, 或者说某种技能适用的任务的多样性, 决定所得出的群因素的广度。

特质形成的另一个主要来源是学习经验的接近 (contiguity) 即共同发生。例如, 处于教育先进的文化中的个体, 可能发展一种广泛的言语—教育因素, 它贯穿于学校中所学习的所有活动。一种较狭窄的数量能力倾向因素, 可能由同一位教师在同一个教室里教授所有算术加工所造成。因此, 在算术课期间失去信心、心生反感或厌烦的儿童往往在学习所有算术加工中都落后; 而在算术课上感到兴奋和高兴的儿童往往对算术课期间所教授的一切都学得很好, 并且往往较快地发展那些促进他或她后来数量学习的能力倾向。

不管因素是怎样发展的, 通过因素分析所鉴定的能力即因素都是描述的范畴, 它们反映种种情境中行为表现的变化着的相互关系。这些因素不是静态的实体, 它们本身是个体累积经

验的产物。经验的相互关系在个体之间和团体之间都有所不同,因而可以预期他们之间的不同的因素模型。通过正规教育、职业培训、其他继续活动等,个体的经验有所变化,因而可以分化出新的特质,或者先前存在的特质也可以并入范围较大的特质之中。

**因素分析和认知任务分析** 最近发展起来的认知心理学的信息加工程序,不仅对于理解智力测验测量什么东西(参见第三章),而且对于理解因素的形成和发展,都作出了重大贡献。口语记录分析要求个体在解决一个问题或完成一项智力任务时“出声思维”,这为分析人类思维提供了一种有价值的方法(Ericsson & Simon, 1993)。不过,随着对认知过程的研究的继续,越来越多的证据表明思维过程的领域特殊性(domain specificity)。除了最基础的水平之外,加工技能对于所加工的内容的类型往往是特定的;个体掌握内容知识和为迅速提取而组织内容知识时,他获得了加工技能。

另一方面,通过因素分析所鉴定的共同特质,主要涉及言语、数量、空间等内容领域。那些由过程所定义的因素没有什么共同性,如联想记忆、一般推理、发散思维,如果通过专门编制的测验评定这种因素,则发现它们本身具有领域特殊性。一个测验是根据因素分析还是根据认知分析而编制,其基本区分是过程与应用过程的领域之间的区分。有关领域可以指所加工的内容(content,例如语言、数学、机械等),或指操作内容的背景(context),这由文化、社会、地理、职业或其他环境的类型所鉴定。

强调领域特殊性的出版物明显增多,<sup>[10]</sup>这反映了认知心理学的发展,从早期努力鉴定人类智力的过程,到探索各个认知过程的领域特殊性的迅速进步。同样值得注意的是,为评定因素分析所得出的内容明确的特质而编制的各种测验可以良好地预

测日常生活表现。这类因素很可能间接地适合于某个内容领域的一组加工技能。因此,某一个体在言语能力倾向测验上获得高分,他不仅在言语知识的范围和组织方面,而且在使用言语领域所需要的专门加工技能方面,均表现出色。

**一般智力** 对智力的因素分析研究不仅关注智力测验测量什么东西,而且关注智力的性质,而不论智力表现在什么地方。如果根据这种广义的观点来考虑,则人类智力由认知技能和知识联合组成,而这些技能和知识为个体在其中发挥作用的经验背景所要求、所培养、所奖赏(Anastasi,1986c)。领域特殊性在这种概括化的智力结构中,比它在口语记录分析或因素分析所鉴定的范围狭小的过程中更为基本。当个体在学习特定的文化背景或特殊的职业等单一领域之内的实际知识和加工技能时,智力的许多方面得到了发展。

早期的研究者努力地鉴定一种单一的、普遍适用的智力,这就产生了传统的“智力测验”。不久人们认识到这类测验是测量学业智力或学业能力倾向的,这是这类测验实际评定的特定智力。此后是“实践智力”(practical intelligence)这一术语的提出,这是传统的测验所没有包括的一种不同的智力(参见 Anastasi, 1986b; Neisser, 1976; Sternberg & Wagner, 1986)。不过,实践智力不是一种单一的智力,而是多重智力,适用于不同的实践领域(Lave, 1988; Rogoff & Lave, 1984)。智力不局限于学业智力和实践智力这两类智力,而是一个多面性的结构。

传统的因素分析的一个不足之处是,对于要进行因素分析的变量的选择,给予不适当的关注(Anastasi, 1988a)。显然,这类分析所鉴定的因素是从被选变量之间的彼此相关中得出的。多数智力研究始于测验的一组分数,而这种测验是设计来测量学校所教授的作业的。因此,所得出的因素表示学业智力。在

教育先进的文化中，这类因素被证明可以良好地预测学校成就；它们也和多种工作上的表现具有适度的高相关，而这些工作和学校教育本身是有关联的。然而，如果我们要评定广义的智力，则需要首先研究人们在现实生活中、文化上有价值的各种活动中的表现。虽然在测验研究中容易得到学校学生被试，但显然也可使用各种职业的成人。研究工作可以组织得既对参加者有利，又产生控制良好的科学资料。在工业和农业环境中有效地进行这类合作研究的例子越来越多（例如，Fleishman & Reilly, 1992b; Lubinski & Dawis, 1992; Whyte, 1991）。

认知心理学中所发展的一些方法本身突出了智力的领域特殊性。研究认知过程的一种方法是通过各对照组某项特定活动中的专家和新手（如棋手、打字员、计算机程序员）的口语记录进行分析。已经多次证实，个体的表现水平和认知技能的应用严格限于特定的领域。这种认知心理学方法的一个特殊优点是，可以对现实生活环境中的种种活动进行研究。这类活动的范围从服务员对餐馆定菜的记忆到医学诊断和法院判决（Chi 等，1988; Ericsson & Smiss, 1991）。

领域特殊性概念对心理学研究的若干课题的影响越来越大。例如，目前对天才的定义中明显反映出这一点。天才学生的鉴别，已经从规定的 IQ 水平（Terman 等，1925）转变到社会所需要的许多领域之一的超常才能（Csikszentmihalyi, Rathunde & Whalen, 1993; Feldman & Bratton, 1972; Horowitz & O'Brien, Subotnik & Arnold, 1994）。同样，早期的努力是设计创造力或发散思维的一般过程测验，现在显然已遇到领域特殊性的问题（Baer, 1993; Runco, 1991, 1994; Subotnik & Arnold, 1994）。目前，关于如何使领域与创造力测验相结合的研究已经取得一些进展。看来，认知心理学与心理测量的因素分析相统一，不仅能够增进对



智力机能的<sup>理解</sup>，而且使两类研究更接近日常生活领域。

## 注 释

- [1] 首次使用 IQ(智力商数)时,它确实表示一种分数——即智力年龄对实足年龄的比率(参见第三章)。
- [2] 关于行为发展中遗传和环境的作用,这是一个广泛的研究领域。本节所讨论的问题,只是其中一个很小的部分。对这个课题的全面论述,包括对遗传力指数的批评性评价,参见布劳恩、霍尔和杜林(Brauth, Hall & Dooling, 1991); 布朗芬布伦纳和赛西(Bronfenbrenner & Ceci, 1994),霍罗威茨(Horowitz, 1994),普洛明和麦克利恩(Plomin & McClean, 1993),以及普洛明和里德(Plomin & Reade, 1991)的有关论述
- [3] 关于对富伊尔斯坦(Feuerstein)方法的批评性评价,参见安娜斯塔西(Anastasi, 1980)、布莱格(Blagg, 1991)的有关论述。
- [4] 关于历史观点,参见阿纳斯塔西(1984b)的有关论述。
- [5] 读者也许感到,经过旋转的 II' 轴应该标为 -II', 以便相当于未经旋转的 -II。不过,轴的哪极标为正,哪极标为负,这本是一件任意的事情。在现在这个例子中,将经过旋转的 II' 轴“倒映”,以便删除负权重。
- [6] 这条准则要求各个测验在一些因素上的负荷与零没有显著不同。现在使用得出因素负荷的标准误差的统计方法,可以在经验上检验这一要求(Cudeck & O'Dell, 1994)。
- [7] 只有当应用坐标轴的直角旋转(orthogonal rotation)时,情况才是如此。采用斜交旋转(oblique rotation),本章后面一节会解释这一概念,因素负荷和因素相关彼此具有一种简单的关系,通过合适的计算方法,就能够根据其中的一个得出另一个。
- [8] 这是吉尔福德后来研究的结果,该模型与早期模型(Guilford, 1967)有所不同,它把图形内容面分为视觉成分和听觉成分,把记忆操作面分为记忆登录和记忆保持(Guilford, 1988)。
- [9] 有关的参考书目,参见卡罗尔(Carroll, 1993)的有关论述。
- [10] 例如,参见格里诺(Greene, 1989)、亨特(E. Hunt, 1987)、施奈德和韦纳

特(Schneider & Weinert1990)、西蒙(Simon1990)，以及斯滕伯格和弗伦斯奇(Sternberg & Frensch,1991)的有关论述。

## 第十二章 能力测验中的心理学问题

任何科学活动的扩展及逐渐复杂化不可避免的结果是科学工作者的兴趣和作用越来越特殊化。这种特殊化很清楚地表现在心理测验和当代心理学主流之间的关系中（Anastasi, 1967, 1991）。心理测量学的专家们把测验编制技术的质量提到了极高的水平。然而，他们提供技术优良的工具，却不太注意确保测验使用者具有恰当使用这些工具所必需的心理学知识，其结果是常常不参照相关行为的研究结果，仍对测验成绩作出过时的解释。心理测验同心理学其他领域部分地分离，以及对测验结果的错误解释及其误用，可部分地说明公众为何对心理测量不满。这种不满始于 50 年代，到 70 年代时表现得愈来愈强烈，直至现在仍然在各种不同领域内弥漫。本章讨论的主题是：说明心理研究的发现有助于有效使用能力测验的方式，帮助纠正对于“IQ”和一些相似测量分数的流行的错误概念。

### 儿童智力的追踪研究

理解“智力”构念的一种重要途径是对同一些个体进行长时间的追踪研究。虽然这种研究可能有助于某些测验的长时预测效度分析，但它对智力的本质和智力测验分数的意义有着更为

广泛的含义。当智力被认为主要是遗传潜力的展现时,就会认为每一个体的“IQ”在其一生中几乎保持稳定。在重复测验中,任何观察到的变化可归结于测量工具的缺点,如,信度不足或所测的机能选择不当。然而,随着对智力本质研究的深入,人们已逐渐认识到智力本身既复杂又多变化。在以下部分,我们将讨论智力追踪研究具有代表性的发现,探究造成所测能力既稳定又不稳定的条件。

**智力测验成绩的稳定性** 已经积累的大量资料显示,在整个小学、中学和大学阶段智力测验的成绩是十分稳定的(参见 Anastasi, 1973, pp. 232 ~ 238; Bornstein & Krasnegor, 1989; McCall, Appelbaum & Hogarty, 1973)。例如,在一项早期的对相对未经筛选的瑞士人的研究中,胡森(Husen, 1951)发现:613名三年级的学校男生的测验分数和10年后同一批被试在军队入伍测验中所获分数之间的相关系数是0.72;在后来一项瑞士人的研究中,哈恩奎斯特(Harnqvist, 1968)报告在4500名年轻男性中13岁和18岁之间的测验分数相关是0.78。甚至学前测验和后来的测验之间也有显著的高相关;在菲尔斯研究所(Sontag, Baker & Nelson, 1958)进行的一项140名儿童追踪研究中,3岁和4岁时获得的斯坦福-比纳分数的相关是.83。后续测验和3岁时的相关随着重测时间间隔的增长而变小,但到12岁时,它们之间仍然有0.46的相关。与斯坦福-比纳测验特别有关的是后来由布拉德维(Bradway)、汤普森(Thompson)和卡温斯(Carvens, 1958)进行的研究。研究对象是1937年斯坦福-比纳标准化测验样本的一部分,最初测验时是2~5岁半的儿童,研究发现:起初的IQ和10年后的重测相关是0.65,和25年后的重测相关则是0.59;10年后(平均年龄14岁)的重测和25年后(平均年龄29岁)的重测相关是0.85。

正如我们所预想的一样，测验之间间隔越短相关就越高；此外，当测验之间间隔恒定，重测的相关倾向是儿童年龄越大相关就越高。年龄和重测间隔对重测相关的影响是相当有规则的，本身具有较高的可预测性( R. L. Thorndike, 1933, 1940) 。智力测验分数的稳定性随年龄而增加的一种解释是“智力发展的累积性质”。个体每一年龄的智力技能和知识包括他或她所有早期的技能和知识加上新获得的增加部分。即使每年的增加部分相互没有任何关系，成绩水平上的前后一致性也会增加，这是因为，以前获得的东西随年龄增长在整个技能和知识中所占的比例也在增加。

不同年龄的技能和知识的这种重叠可以部分地说明个体发展中测验分数稳定性增加的原因。但另外两个条件也值得考虑：第一个是环境的稳定性，代表了大多数人发展的特点。在发展某一阶段具有的有利条件或不利条件在重测的间隔时间内一直保留着；儿童在其成长过程中，处在同一家庭中，具有同样的社会经济水平和文化环境，一般不可能任意地从智力刺激良好的环境转向智力刺激迟钝的环境。同时，我们必须注意到同一家庭子女的心理环境远非完全相同。越来越多仔细收集到的资料显示：在一起抚养的子女的经历可能具有多方面的差异( Boer & Dunn, 1992; Dunn & Plomin, 1990, Hetherington, Reiss & Plomin, 1993)。有年幼或年长的兄弟姐妹，其本身就是一种不同于独生子女的具有重要意义的不同的心理经验。父母育儿实践的性質以及父母对儿童行为的反应对不同时间出生的子女也可能是不同的。此外，影响家庭生活的重要事件，如离婚、家庭收入的大变动，从农村转到城市环境，可能对不同发展阶段的儿童产生不同的影响。除了这些可能的差异以外，还有不同子女可能拥有家庭外累积的不同经验，因而，尽管家庭环境的连续性可能有

利于某个儿童测验成绩的稳定，但并不必然导致子女之间的相似。

第二个有利于个体智力测验成绩稳定性的条件是先前学习技能对后继学习的作用。个体不仅保存先前学习的结果，而且其中的许多为后继学习提供了工具，因此，儿童在任何一个时候获得的智力技能和知识越多，对他随后的学习帮助越大。教育中学习准备的概念就是这个一般原理的表述。学习的序列性质也隐含在前面讨论过的皮亚杰的心理发展研究途径以及各种个体化的教育计划中。

同一原理的应用是为教育不良的学前儿童而准备的“早期教育项目”(Project Head Start)和其他补偿性计划的基础(Stanley 1972, 1973; Zigler & Valentine 1980)。因为来自不良背景的儿童，缺少一些在学校中有效学习的重要先决条件，随着年级的升高，他们在学业成绩上可能会越来越落后。必须认识到：学习的先决条件，不仅包括诸如获得语言和数量概念这样的智力技能，还包括态度、兴趣、动机、问题解决的策略、对挫折的反应、自我概念和其他人格特征。补偿教育计划的目的是提供学习必须的前提条件，以便使儿童能得益于以后的学校教育。当然，这些计划希望这样做能破坏 IQ 分数的“稳定”，否则，它将会停留在较低的水平。

**智力测验成绩的非稳定性** 测验分数的相关研究提供了可应用于群体预测的统计数据。基于以上的理由，从统计学意义上看，分数倾向于十分稳定。但另一方面，对个体的研究可能会发现测验分数向上或向下较大变动的现象。急剧的上升或下降可能是儿童主要生活环境变化的结果。家庭结构或家庭条件的激烈变化，如被寄养家庭收养、长时间的严重疾病、治疗或补偿计划都是可能改变儿童以后智力发展的生活事件的具体例子；

即使是保留在同一环境中的儿童，在重测时测验分数也可能表现出较大的上升或下降。当然这些变化意味着儿童以一种高于或低于测验赖以标准化的正常人群的速度在发展。一般来说，随着年龄的增长，教育不良环境中的儿童的重测分数倾向于降低，而教育良好环境中的则倾向于上升。具体说明这些环境因素和儿童自身特征的研究，在当代的研究中逐渐受到了重视（参见 Carroll 1993 . pp.669 ~ 674; Detterman & Sternberg, 1982）。

智力测验中个体分数变化幅度的资料最初由“加利福尼亚指导研究”（Honzik, Macfarlane & Allen, 1948, California Guidance Study）提供。这一研究对 222 名被试的重测资料进行分析，发现个体 IQ 变动幅度可大至 50 点，从 6 岁到 18 岁重测的相关一般是高的，59% 的儿童的变化是 15 点或 15 点以上，37% 的是 20 点或 20 点以上，9% 的是 30 点或 30 点以上。实质上这些变动大多数不是偶然或反复无常的。相反，在连续的几年里，儿童表现出一种前后一致向上或向下的变动趋势，而这些变动跟环境的特征有关。在加利福尼亚研究中，仔细地调查了家庭条件和亲子关系，表明了 IQ 向上或向下的较大变动与抚育儿童的文化环境和情绪氛围有关。参加者年满 30 岁时的进一步跟踪研究，仍发现测验分数和家庭环境之间的显著相关和年仅 21 个月时的评定一样（Honzik, 1967）。像其他反映父母对儿童一般的物质关心的变量一样，父母对教育成绩的关注是后继测验成绩的一种重要相关因素。

有些研究者更为明确地集中在与智力发展加速和减速有关的人格特征研究上。费尔斯研究所的一项详尽追踪研究包括 140 名儿童，年龄从婴儿早期到青春期再到青春期之后（Kagan & Freeman, 1963; Kagan, Sontag, Baker & Nelson, 1958; Sontag 等, 1958）在这些儿童中，从各种人格和环境的度量方面，比较了

4岁半到6岁之间 IQ 提高最多和降低最多的儿童，也同样比较了6到10岁之间 IQ 有最大的幅度变动的儿童。在学前期，对父母的情绪依赖是与 IQ 降低有关的主要因素；而在学龄期间，IQ 的增加主要与高成就动机、竞争内驱力和对自然的好奇心相关联。一些获得的资料也显示出父母态度和其儿童抚育实践在这些特质发展中的作用。

对同一批被试，后来的一项分析扩展到了17岁，主要集中在随时间而发生的分数变动的模式上（McCall 等，1973）。这一研究对表现出不同模式的儿童进行比较，并把它和通过定期的家庭访问来评定儿童抚育实践联系起来。一项典型的发现是：在学前期，分数表现上升趋势儿童的父母表现出了“一种鼓励性和奖赏性的氛围，但其策略具有某种结构和强制性”（McCall 等，1973, p. 54）与分数上升相连的一种主要条件是，促进性的努力或“父母有意地在各种还没有成为重要的心智和动作技能方面训练儿童的程度”（p. 52）。

关于与智力测验分数增加或减小有关因素的研究，说明了决定一般智力发展的条件。它也显示了：如果个体情绪和动机特征方面的度量，以及其环境方面的度量能跟最初的测验分数结合起来，那么对个体以后智力状况的预测可以得到改善。从另一角度来看，这类研究的发现指明了通向这种干预计划的途径。这种干预计划可以有效地改变智力发展的进程，使其沿着所希望的方向发展。

## 儿童早期的智力

年龄两端的智力评定提出了一些特殊的理论和解释性问



题,其中之一是应该测量的机能是什么,包括婴儿和学前儿童智力的组成是怎样的?老年人的智力组成又是怎样的?第二个问题不是完全独立于第一个的。不像学龄儿童、婴儿和学前儿童还没有以学校课程代表的标准化系列经验;在编制小学、中学、大学水平的测验过程中,测验编制者有大量的共同的经验材料用于起草测验项目;另一方面,入学前,尽管在儿童的抚育实践中有某些文化一致性,但儿童的经验是远非标准化的,在这种条件下的测验编制和测验结果的解释都要困难得多。在某种程度上,老年人的测验中也面临同样的难题,他们的学校教育在许多年前已经完成,此后从事各种各样的活动。在本节和下一节里,我们将分别考察儿童早期和成年测验中这些问题的一些含义。

**婴儿和学前测验的预测效度** 追踪研究得出的结论是:学前测验——尤其是2岁以后施测——对以后的智力测验成绩具有中等的效度,而婴儿测验实际上没有效度可言(Bayley, 1970; Lewis, 1973; McCall, Hogarty & Hurlbart, 1972)。麦克卡尔(McCall)和其同事(1972)将8项研究结果的报告结合起来计算出:出生后30个月期间的测验和3~18岁获得的IQ具有中等程度相关。在他们的发现中有几种明显的趋势:第一,出生第一年内施测的测验有较少或不具有长期的预测价值;第二,婴儿测验对预测学前年龄(3~4岁)的IQ具有一定的效度,但当儿童达到学龄年龄以后这种相关就急剧下降;第三,18个月以后,预测效度是中等的和稳定的,大多数是在0.40和0.50之间,当预测是从这一年龄作出的,不管从3岁到18岁的任何时候,相关似乎具有同样的大小。

婴儿测验缺乏长时的预测效度,需要在考虑其他相关发现的情况下进一步加以评价。第一,预测可以通过重复测验来考虑发展趋势而得到改善。第二,几个调查研究发现,婴儿测验在

不正常的、有临床问题的人群中比在正常的人群中具有高得多的预测效度 对原先 IQ 分数低于 80 的儿童以及已发现或怀疑具有神经异常的人群来说, 已经报道其具有 0.60~0.70 的显著的效度系数(Ireton, Thwing & Gravem, 1970; knobloch & Pasaman-ick, 1963, 1966; Werner, Honzik & Smith, 1968)。在因遗传或环境原因引起的机体病变而导致的发展缺陷的诊断中, 婴幼儿测验作为辅助工具似乎是最有效的。

在没有机体病变的情况下, 儿童以后的发展主要由他养育的环境所决定, 不能期望测验来预测。事实上, 对以后的智力水平来说, 父母的教育及家庭环境中其他更为特殊的特性是比婴儿测验更佳的预测指标; 还有, 18 个月以后, 如果测验分数能跟家庭的社会经济状况指标结合起来, 其预测效果会得到明显的改善 (Bayley, 1955; McCall 等, 1972; Pinneau, 1961; Werner 等, 1968) 也有人曾经提出婴儿的个体差异可能相对较小, 具有短暂性, 因为在这个早期阶段的正常发展基本上是种族共同的 (R. B. McCall, 1981)。在其后的年月里, 个体的差异增大, 在不同的年龄间变得更加稳定, 与遗传和环境因素都具有更高的相关 (Plomin, De Fries & Fulker, 1988) 此外, 我们还必须注意到, 在 90 年代, 关于幼儿认知行为预测价值的研究活动逐渐增加, 并得到了一些有希望的结果 (Colombo, 1993)。

**儿童早期智力的性质** 当我们考察了关于儿童早期智力性质的研究, 就可以更全面地理解婴儿智力测验的效度和早期行为测量的意义。这些研究结果不支持婴儿期发展稳定和单一性智慧能力的概念 (Lewis, 1973, 1976; McCall 等, 1972)。甚至在只有 3 个月间隔的测验中得到的相关也是微不足道的, 2 岁和其后在同一或不同量表上的成绩的相关经常不显著。此外, 在同一年龄施测的不同量表间的结果也没有多少相关。这些结果是

通过标准化的工具，如贝利幼儿发展量表和皮亚杰类型的顺序量表得到的（Gottfried & Brody, 1975; King & Seegmiller, 1973; Lewis, 1976, Lewis & McGurk, 1972）。

几个研究者得出结论，认为婴幼儿智力测验虽然缺少预测效度，但是它是儿童当时智力发展的指标（Boyley, 1970; Stott & Ball, 1965; Thomas, 1970）。根据这一观点，婴儿测验和以后成绩间的相关很小的主要理由是智力的性质和成分随着年龄而变化。婴儿期的智力和学龄期有质的不同，它包含了不同的能力组合。

在一系列深入细致的研究中，麦克卡尔和其同事探讨了出生后两年内间隔 6 个月的婴儿智力的性质变化（R. B. McCall, 1976; McCall, Eichorn & Hogarty, 1977; McCall 等, 1972）。通过对每 6 个月水平内不同技能之间的相关，以及跨年龄的同一些技能和不同技能间的相关的统计分析，这些研究者试图找出婴儿行为以后发展的前兆。这一研究得出的一个结论是：不同年龄中起支配作用的行为表现出质的变化，它是有序的、合理的转变。在连续几个 6 个月年龄水平上，对幼儿人格塞尔测验项目上的反应分别进行因素分析时发现：每一水平上的第一因素的分在不同的年龄间的相关显著；但这些第一因素的行为组成有着年龄差别。用另一句话来说就是，虽然当通过与年龄相宜的行为来评定每一年龄时，某一年龄的能力能预测以后年龄的能力，但不同年龄水平的心理能力的具体表现是不同的。

引入“发展转换”（developmental transformations）的概念，用来描述和年龄相适应的智慧能力表现的变化。能力行为的这种质的变化的进一步证据是由雅洛（Yarrow）和其同事关于幼儿对其周围环境掌握的研究提供的（Messer 等, 1986; Morgan & Harmon, 1984; Yarrow 等, 1983. Yarrow 等, 1984; Yarrow & Messer, 1983）。

这一研究结果显示：引发这种行为的任务，和所引出的行为如视觉注视、操作和问题解决中的坚持性这两方面都有暂时的进步。例如，幼儿首先发现他或她能影响周围环境，如丢下一块积木看它落下并听到它碰到地板的声响，或摆动铃铛使它发出声音；其后对环境的掌握，通过更复杂目标定向的活动表现出来，如利用迂回或“手段—目的”的关系获得某一玩具。通过这些与具体年龄相适应行为的确定，儿童早期智力评定的构想效度和预测效度都能够得到更有效的研究。知识内容在智力加工和认知策略运作中的作用也是重要的（Reese, 1987），这是整个认知心理学研究中逐步认识到的一个事实。

**干预计划的涵义** 儿童早期教育的各种干预计划表现出来的效果依赖于特定计划的性质（R. C. Collins, 1993; Haskins, 1989; Zigler & Muenchow, 1992; Zigler & Styfco, 1993）。这些计划主要是为了提高不良背景中儿童的学习准备性，其方法和结果具有很大的差异。大多数项目是失败的，原因是对其实施或评估没有适当的计划；还有几种项目使得儿童学习成绩有了实质性的改善，但这种改善经常是有限的、短时的。与大多数只想通过含糊选择的方法提高 IQ 分数的计划相反，有几项高质量的项目，清楚地确定了要改进的具体智力技能，选择了恰当的训练程序，在这种情形下进行的跟踪计划在相关技能上表现出了显著的持久的改善。有一些关注将干预计划放在更广泛情境中，包括家庭所需要的健康照顾和社会服务。父母的参与对家庭中学前计划的完成，对确保正规计划终止后的连续性具有尤其重要的意义（Jaynes & Wlodkowski, 1990）。

用跟踪计划来评定干预的性质和持续时间，也需要加以密切注意。这种项目结果有效性的评价，要求有相当复杂的方法（Collins & Horn, 1991; Willett & Sayer, 1994）。除了实验设计的因

素外,与评估工具的心理测量学特性有关的统计方法问题,也可能产生虚假的积极或消极的结果(Bejar,1980)。实验组和控制组在测验项目难度或区分度上的差异,或同一组被试测验前后成绩上的差异,对干预计划的成功和失败可能会得出错误的结论。在项目反应理论(IRT,第七章)和适应被试特点技术(CAT,第十章)基础上编制的记分的测验的使用,可避免其中的一些难题。最近,人们对设计周密的干预计划产生了兴趣,它是以比较有效的早期干预计划为模型的(R. C. Collins,1993; Consortium, 1983; Haskins, 1989; Whimbey, 1990; Zigler & Styco, 1993)。此外,新计划的发展可以从儿童智力研究快速积累的资料中得到益处(参见 Horowitz & O'Brien, 1989)。一项特别有希望的长期研究项目,集中在父母对 1~2 岁儿童的行为对其以后智力表现的影响上(Hart & Risley, 1995)。这项研究最初的资料已经强有力地支持父母与儿童接触的性质及程度和儿童的智力发展具有紧密关系的观点。

## 成年智力测验中的问题

随年龄而发生的衰减 韦克斯勒量表用来测量成人智力的一种显著特征是:使用递减常模来计算离差智商(第八章)。韦氏测验(WAIS和 WAIS-R)几个分测验上的原始分数首先转化为均数为 10、标准差为 3 的标准分数。这些量表分数是按固定的参照组来表述的,它由标准化样本中的 500 名年龄从 20 岁到 34 岁的被试组成。为了寻求合适的年龄组中的离差智商,我们使用了 11 个分测验量表的总分。然而,如果我们直接检测量表总分,可以将单一的连续量表与不同年龄组的成绩做一比较。

图 12-1 显示了全国标准化样本中不同年龄水平上这些量表总分的平均分数。

从图 12-1 可以看出,分数在 20 岁和 34 岁时达到高峰 其后,在所有年长组中稳定下降。离差智商是把一个个体量表总分与其年龄组常模相参照而查找出来的。所以,如果被试的成绩表现出和常模样本一样随年龄下降的趋势,那么他的 IQ 应该是保持不变的。韦克斯勒这一方法的最初原理是个体所测试的能力在 30 岁以上随年龄而下降是正常的。

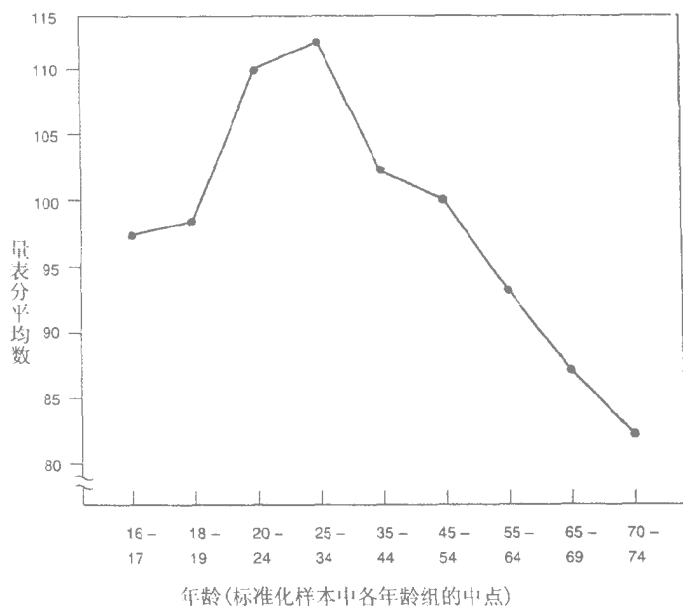


图 12-1 不同年龄组 WAIS-R 全量表量表分的递减 (数据引自韦克斯勒 1981, p.26.)

然而在解释图 12-1 所表明的年龄下降趋势时,我们必须考虑测验标准化中采用样本的一个重要特性。因为每一个标准化的样本就是常模样本(normative sample),它应该反映每一年龄水平现有人群的特征(Anastasi, 1956)。几十年来,一般人群的教育水平都一直在提高,任一时间年长组受到的教育要比年轻组少。这种教育的差异清晰地反映在韦克斯勒(1953~1954 施测的)量表和其修订版(1976~1980 施测)的标准化样本中。在这两种样本组中,学校教育最多年限是在 20~34 岁的水平,在年长组里教育水平呈一致性下降的趋势。尽管韦克斯勒量表修订版的样本比原量表的样本受到的教育要多得多,但在其修订版中,学校教育年限随年龄而减少的趋势和早期的标准化样本中的趋势一样。还有,在韦克斯勒修订版中,量表分数的年龄减少趋势和早期标准化样本中的趋势极为相似。

如果一项测验的标准化样本真实地代表了常模建立时整个国家的人口总体,那么其受教育数量的年龄差异是不可避免的。尽管如此,教育差异使得对实测的分数减少的解释变得复杂化。标准化样本中的年长组,在测验中的成绩不佳,不是因为他们变老,而是因为他们比年轻组受到的教育较少。

从韦克斯勒量表标准化样本中所获得的结果,是传统的成人智力的横向研究结果的代表。不同年龄个体在同一时间进行测验的横向比较可能表现出了一种明显的年龄衰减趋势,因为文化上的变化和老龄化的效果交织在一起。受正规教育的数量只是年龄组间许多不同的变量之一。在过去的半个世纪里,在人类社会发生的其他一些文化变化,使得 20 岁人的经验背景十分不同于 70 岁的。像收音机、电视机和互联网等交流媒介以及交通设施的变化大大增加了发展中个体获得信息的范围;营养和医疗护理的改善也可能间接影响行为发展。

以同一批被试,从 5 年到 40 年的重复测验为基础的追踪研究发现了相反趋势:随年龄增长,分数趋向于增长。某些追踪研究是用智力优良组进行的,如大学研究生或开始时因为较高的 IQ 分数而选取出来的个体 ( Bayley & Oden, 1955; R. B. Burns, 1966; D. P. Campell, 1965; Wisbet, 1957; Owens, 1953, 1966) 。鉴于这种理由,有些作者争辩说:这些发现只能局限于具有较高智力或教育水平的人群,不能应用于一般的人群。然而,另一些对中等程度样本 ( Charles & James, 1964; Eisdorfer, 1963; Tuddenham, Blumenkrantz & Wilkin, 1968 ) 及公共机构以外的智力落后成人 ( Baller, Charles & Miller, 1967 ; Bell & Zubek, 1960; Charles, 1953 ) 的追踪研究也获得了相似的结果。

不管是横向研究还是追踪研究都不能单独对观察到的年龄变化提供结论性的解释。一方面,教育水平的年龄差异,在横向研究的成绩中,会产生虚假的随年龄递减现象;另一方面,随着个体的年龄增长,他们本身也接触到了文化的变化,这有可能改善他们在智力测验上的成绩。对每种程序中存在的方法上的难题以及它们所要求的实验设计,已经出版的几种报告进行了极好的分析。<sup>[1]</sup>基本上,所需要的是几种程序的组合,如横向序列设计 ( K. W. Schaie, 1965, 1994, Shock 等, 1984 ) 这一设计把传统的横向和追踪测验的数据与“时间延滞比较”(time-lag comparisons) 结合起来。后者要求在不同的时期对同龄群体 ( the same-age cohorts )<sup>[2]</sup>进行测量,如 1940 年测量的 20 岁组和 1970 年测量的 20 岁组进行比较。

有几项研究提供了至少能对起作用的因素进行部分分析的数据。欧文斯 ( Owens, 1966 ) 在对依阿华州立大学的学生进行 40 年的重测时和坎贝尔 ( D. P. Campbell, 1965 ) 在对明尼苏达大学的学生进行 25 年的重测时,分别测量当时的一年级大学生,从



而能对两组同样年龄但时间相隔为 25 ~ 40 年的被试测试成绩,对一组被试在同一时间间隔前后的成绩进行多重比较。在这两项研究中,相隔 40 年的组比他们自己早期的成绩有了进步,与相隔 25 年的组相比,则处于同等水平。这种发现表明:是文化的变化和其他因素的影响而不是年龄本身,在比较有限的实验设计中产生分数的上升和下降。库斯勒 (Kausler, 1994) 的一次综合调查表明,对老年人学习的研究兴趣正在明显地逐步增加。这本书必定能帮助我们去掉学习老龄效应的刻板印象。尽管这本书较多地建立在科学发现的基础上,但它也提供了许多与日常生活相联系的经验。

**西雅图追踪研究** 一项使用横向序列设计技术、计划得很好的长时间研究是西雅图追踪研究 (The Seattle Longitudinal Study, SLS, K. W. Schaie, 1994; Schaie & Hertzog, 1986)。该研究开始于 1956 年,研究者从一项大约包括 18000 名成员、费用先付的医疗计划的人群中,分层随机抽取 500 名被试,施测了一套能力测验。<sup>[3]</sup>这个群体被认为能在相当程度上代表大都市地区的人口统计特征。样本从 21 岁到 70 岁,每间隔 5 岁分别包括 25 名男性和 25 名女性。在 SLS 的后继阶段,测验施测经过了 6 个周期(从 1956 年到 1991 年),每一周期中包括原始样本存活者,以及每期引进的新样本。

SLS 的基本实验设计包括:同一个体的重复追踪重测,同一时间测量的不同年龄群体的横向比较(如 1977 年测量的 30 岁组和 50 岁组),不同时间测量的特定年龄群体的比较(如 1963 年 30 岁组和 1984 年的 30 岁组)。通过对不同种类的比较数据进行恰当的统计分析,有可能确定与年龄、特定社会文化变化、个体实践及经验等因素有关的成绩变化。如图 12-2 所示,结果表明大多数机能的年龄衰减现象开始较晚,其趋势比传统横

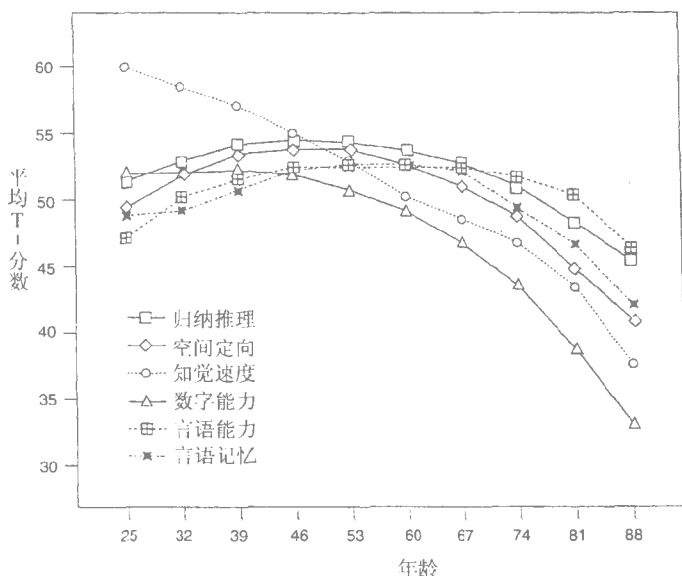


图 12-2 各能力构成的平均分数的追踪值 (引自 K.W. 谢伊, 1994, p. 308。其中的数据是根据个体 7 年的重测积累起来的)。

向比较研究所显示的要平缓一些。

除了对长期争论的年龄对成绩影响问题进行综合的系统探讨以外, SLS 还研究了其他相关问题。例如, 它发现年龄的变化随评定的智力机能的不同而不同, 如, 言语能力、数字能力和知觉速度 (见图 12-2)。因此对智力的整体测量, 如 IQ 分数, 会使能力的上升和下降变得含混不清。有相当多的研究还关注引起这些变化的原因, 尤其是年龄较老的人身上发现的成绩下降的原因。与成绩下降相关的主要变量有: 健康状况不良、特殊疾

病、不活动、在某些机能上缺少连续练习、低动机水平等人格因素和灵活性的减少等。这些发现的一项结果是：形成了一些干预程序，努力阻止或逆转随年龄而发生的能力减弱。当前 SLS 的研究，尤其是朝着形成该种补偿性计划而努力的（参见 K. W. Schaie, 1994; Schaie & Hertzog, 1986）。

**个体差异和年龄** 一项基本的发现是：年龄衰减比以前假设的幅度要小，出现的时间要晚；除此以外，当前的研究总体上揭示了在所有年龄水平上能力都有很大的个体差异。不管是关于年龄衰减还是关于同辈群体差异的任何概括，都必须认识到所有情况下都能发现的个体差异性，才能成立。任何一个年龄水平内的个体差异比年龄水平间的平均差异要大得多，其结果是：不同年龄个体获得的分数分布广泛地交织在一起。这意味着，可以发现大量年长者的成绩水平等于年轻人的成绩；此外，年长者中成绩最好者超过年轻组中成绩最差者。这种交织状况不局限在相近的年龄水平上，当比较两端组的被试时，成绩范围仍然交织在一起；一些 80 岁的老人会比 20 岁的年轻人做得更好。

然而与现在讨论的主题更有关系的是各个体由于年老而发生的变化有所不同。例如，在 50 岁和 60 岁之间的成人测试成绩，一些个体可能表现出了下降变化，一些则没有明显的变化，甚至有一些还呈增加趋势；其变化的数量，不管是下降的或上升的，在个体间将有很大的差异。此外，对年长个体的详细研究扩展到 60 多岁、70 多岁、80 多岁时，显示了智力功能和个体的健康状况存在的关系比和实际年龄的关系更为紧密（Birren 等，1983; Polmore, 1970; Schaie & Gribbin, 1975）。其他起作用的因素包括具有各种智力刺激的适宜环境和灵活生活方式的维持（K. W. Schaie, 1994, Schaie & Hertzog, 1986）。

**成年人智力的性质** 对人的整个一生,传统智力测验主要定向在学龄儿童和大学生。在这些水平上,测验编制者利用那些已经组织在学习课程内的许多共同经验。大多数智力测验测量个体获得学校所传授的智力技能的程度;反过来它们可以预测在教育阶梯中学生为下一水平所做准备的程度。包括韦克斯勒量表和西雅图追踪研究中使用的基本心理能力测验在内的成人测验,都主要依靠这种可确定的共同经验。随着个体年龄的增长,他或她自身的正规教育的经验已经消退,这些共同的经验可能逐步变得不适合用来评价其智力机能。成年人的职业比学龄期的学校教育更为多样,成年期的累积经验可能刺激不同的人发展不同的能力。

因为智力测验与学术能力紧密相连,我们会毫不奇怪地发现:甚至那些最早关于成年人的研究,也已经表明那些继续接受长时间教育的个体的分数随年龄有较大的增长(D. P. Campell, 1965; Harnqvist, 1968; Husen, 1951; Lorge, 1945; Owens, 1953)。同样,如果个体的职业在其内容上具有较多的“学术性”,要求使用言语和数字能力,他们在智力测验分数上可能维持其水平或有所进步;而那些从事机械活动或人际关系职业的人则可能会有下降的现象。威廉斯(M. Williams, 1960)报道的一些具有启发性的数据支持这种假设。威廉斯在一系列的言语和非语言测验中比较了年龄范围从 65 岁到 90 多岁的 100 名被试的成绩,发现个体的职业和他在这两种类型作业上的相对成绩具有显著的对应性。对成年人的追踪研究也发现整个 IQ 的变化和某些生平问卷的项目具有明显的关系(Charles & James, 1964, Owens, 1996)。

人们无时不在促进适合其特性要求的技能的发展。在人的整个一生中,从婴幼儿到学龄儿童,到不同职业的成人,再到退

体的 70 ~ 80 岁的老人, 这些要求各不相同 ( Baltes, Reese & Lipsitt, 1980)。戴明和普莱西( Demming & Pressey, 1957)较早地显示了这一事实在智力测验中的涵义。这些研究者开始时通过非正式的方式, 调查成人阅读材料和其报告的日常活动及问题, 对典型的成人机能进行任务分析; 在这一基础上, 他们准备了适合年长者的初步的 20 种测验。这些测验强调实践知识、判断和社会知觉。这些测验中的三种测验的结果以及同时对不同年龄样本实施的标准言语和非言语测验结果显示: 年长者在新的测验中优于年轻人, 而在传统测验中恰好相反。所有这些类型的研究表明: 成年期智力测验分数随年龄增长是上升还是下降, 主要依赖于个体在这些年内所拥有的经验, 依赖于这些经验和测验所包括的功能之间的关系。从广义的水平来看, 所有测验结果只有在具体情境的框架中才能更好地理解。包括对职业和其他有社会意义活动领域在内的多种不同场合情境下行为的任务分析, 应该既能改善测验的结构又能清楚地说明测验分数的意义 (Anastasi, 1986)。值得注意的是, 在 80 年代和 90 年代, 对成人、尤其是处在生命阶段上端的那些人<sup>[4]</sup>的研究的显著增加。毕生发展心理学是一个正在逐步发展的研究领域 ( Rutter & Rutter, 1993)。

## 智力测验成绩的人群变化

**上升分数** 经过一段较长的时间后, 某一人群的智力测验成绩会发生什么变化? 这是我们在几个方面已经碰到的问题。在前一节里我们看到, 这几十年来随着成年人群教育水平的提高, 智力测验成绩的平均数也相应上升。这种变化的结果是 常

模样本中的年长成员平均完成的教育比年轻成员要少，他们在测验上的得分要比年轻成员低。相似的现象也发生在儿童测验的标准化样本中。在以后标准化的斯坦福—比纳测验和韦克斯勒量表中，后来标准化样本的成绩明显要比前面样本的好；如果以修订版来施测，任何一个儿童都比他在早期版本中得到的 IQ 分数要低，原因很简单，因为其成绩是以较高的常模来评估的。后来样本中，施测儿童的父母教育水平较高，则是智力测验分数上升的原因之一。

这种类型的比较研究可以称为人群追踪研究。在心理学研究中，追踪研究通常包括经过一段时间的间隔后对同一些个体进行重复施测。然而在人群的追踪研究中，人群是在不同期间取样的；其比较是在不同时间出生的、但在同一年龄施测的群体之间进行的。<sup>[5]</sup>在 20 世纪前 50 年内所进行的几个大规模的研究揭示了标准化智力测验测量的人群中智力有上升趋势（Anastasi, 1985d, pp. 26 ~ 130）。显然，几十年来随着受教育者的增加，较高的教育水平和其他文化变化，使所有年龄人群的平均智力呈现了一种稳定的上升趋势。

在这些比较研究中采用了各种程序。其中一种是经过一段时间后施测同一测验，如 1932 和 1947 年进行的对 11 岁苏格兰儿童的研究（苏格兰议会，1949）。另一程序是对有代表性的样本实施两种测验，目的是确立这两套测验分数间的对应关系，而后能把一种测验的成绩转换为另一种测验的成绩。在一战和二战时，美国军队中士兵的成绩比较曾使用这一程序，对士兵分别施测军队甲种测验和军队一般分类测验（Tuddenham, 1948）。第三种技术上更合理的方法，其基础是通过使用参照项目（anchor items）建立一种绝对的与样本无关的分数量表，就如大学董事会所采用的测验一样。项目反应理论的应用（第七章）代表了这一

方法的进一步完善。

**下降分数** 随着时间的变化,特定人群智力测验分数是上升、下降或保持稳定依赖于许多条件。所经历的年代和随之而发生的文化变迁,显然是主要因素;测试个体的年龄也能带来一些差异。例如,人群教育水平的提高将直接影响成年人的测验成绩,但它只能间接影响儿童的操作,因为所比较样本中的儿童在测试时受到同样程度的教育。另一需要特别考虑的重要事实是,当测试挑选出亚群体时,不同时期选择程度的变化。例如,1960 年升入高中学习的人口比率超过 1910 年的;实际上,1910 年的高中生和 1960 年的相比,代表了他们那时代的总人群中经过高度选择的样本。测验分数上升和下降之间的明显不一致性可能源于采用测验本身的性质、测验的特定的亚群体(如准备上大学的高中高年级学生、一般成年人群、小学儿童)、覆盖的具体时期(Flynn,1983,1987)等因素。

智力测验中影响人群分数下降或上升的因素的多样性和复杂性的一个良好说明是:对大学入学考试的学习能力倾向测验(SAT 上众所周知的分数下降现象的研究分析 Donlon,1984, pp. 188 ~ 191; Wirtz,1977)。从 1963 到 1977 年,平均 SAT 言语分数由 478 下降到 429,平均数学分数由 502 下降到 470。在解释这种 14 年来分数稳定下降现象的努力中,一个特地指定的专门小组实施了由来自不同领域专家进行的 38 项研究,研究了一组重要的因果假设。

这一小组得到的主要结论是:14 年中,前半部分和后半部分的因果模式是不同的。在前 7 年里,分数的下降主要来自于参加 SAT 测试的组群的组成成分的变化;因为这一期间进入大学的高中毕业生的比例持续增加,在测验所测量的认识技能方面,取样逐步变得具有较少的选择性。然而在后 7 年里,进大学

的人口变得稳定，取样选择只能说明分数下降中少得多的部分；对于这一期间的解释不得不主要从一般的家庭、学校和社会条件中寻找。这一小组观察到，获得的数据不能决定各种不同文化变化对分数下降的相对作用。然而，在许多影响因素中可能重要的因素是：学术标准的降低，年级学生数膨胀和自动升级，家庭作业布置的减少，学校缺勤率的增加，忽视技能和知识的掌握，过度地观看电视节目，以及在学生生活中引起注意的社会剧变。<sup>[6]</sup>

一项后来的分析研究 (Turnbull, 1985) 认为，后 7 年内 SAT 分数下降，可能部分是前 7 年内组成成分变化的延迟的间接结果。因为较大一部分准备不足的学生逗留在高中（并申请入大学），学校教育条件的许多变化可以看成是学校对学生有较大差异的适应性反应。这种反应倾向于降低学校计划的要求水平，这可以从年级学生数膨胀、职业和一般选修课程的激增、课本简单化和家庭作业布置的减少等现象来得到说明。所以学生人群的变化导致教育计划的变化，反过来又导致了分数的下降。这一假设和 70 年代高分数较少的发现一致，在班级里名次高的学生的分数继续下降，而那些名次低的学生的分数稳定不变或有所提高。然而，在 20 世纪 70 年代晚期和 80 年代，在学校学业要求水平方面有所转变，这在 SAT 的分数有所上升的趋势中明显地反映出来。

总述 在评估人群智力分数变化的努力中碰到的方法问题，由一项对 14 个国家智力测验成绩提高的调查研究而变得更加明显。这一研究结果显示存在着许多的不一致性和不相容性。它们的解释也是如此难以捉摸，以至于使得这一项研究的作者没有得出任何结论，而只有一种消极性的建议，即智力测验不能真实地衡量智力！测量人群变化的努力仍在进行当中，较



好的程序仍未完全发展起来。对这些相关程序问题的恰当的处理方法可以在由柯林思和霍恩(1991)编辑的关于这一主题的一次会议报告中找到。

目前,对于人群测验分数上升和下降的研究结果的恰当理解,要求有几种相关的信息。第一,对施测的各种具体测验必须加以鉴别,要特别注意它们所覆盖的具体过程和内容(如知觉速度、记忆、语言理解、空间定向),以及它们常模的来源。第二,必须报告施测和重测的日期。第三,相关的数据资料必须包括样本抽取的人群以及重测时取样的任何变化,如起初成绩较好或较差者在后继样本中的丢失。第四,应该有影响所研究人群的任何重要的文化变迁的信息,这些信息可以是关于教育的量和质的,或者是可能影响文化间接触的传媒或交通的发展,或在数量或方向上可能改变智力进步的个体经验中的任何其他事件。

使得对不同人群分数变化研究产生混乱的因素中,还包括对‘智力’这一术语的普遍误用(尤其是 IQ)。有些人将智力视为代表一种统一的可视为同一的机体特性(参见 Flynn, 1987),另一些人则将智力解释为在特定的情境中有效地运作和发展所要求的各种能力的结合(见第一章)。因此对个体测验成绩或不同时间不同地点人群平均成绩的适当解释,当然要求上面阐述的主要条件的有关知识。人们越来越认识到在测定变化时会遇到许多技术问题,以及在不同情境下为不同目的而评定变化将有许多不同的方法(尤其见 Gottman, 1995 Gottman, 1995)。

## 文化多样性

本书的不同部分已经从不同的角度考察了不同文化背景中

的人们的测验使用问题。第十八章是关于测试中的社会和伦理考虑，特别是关于多民族文化中的少数民族群体。关于测验偏见和项目偏见的技术问题在第六章和第七章进行了分析；第九章我们讨论了最初设计的各种跨文化应用的典型测验。本节我们将提出关于文化在行为中的作用的一些基本理论观点，特别是关于智力测验分数的使用 and 解释。

**文化心理学的领域** 20 世纪的最后 30 年，有关文化心理学的研究及其著作显著增多 (Berman, 1990; Irvine & Berry, 1988)，甚至有些国际会议主要就是关于这一主题的 (参见 Brislin, 1993; Cronbach & Drenth, 1972; Manoleas, 1995)。这一领域主要关注的是：在不同文化背景中养育和生活的群体之间的行为差异。这些背景，可以狭小到邻里之间或一个村庄，也可以广大到一个国家或一块大陆。此外，一些广义的文化 (如美国的西班牙文化) 有其亚文化，如墨西哥的、古巴的、波多黎各的，或者中美洲的、南美洲的。在理解个体的行为中，它们有很大的差异，需要加以认定 (参见 Geisinger, 1992; Marin & Marin 1991)。

文化在人类行为中的作用，可以看成是领域特殊 (domain-specific) 的，与第十一章讨论的认知心理学所认为的相似。认知心理学家的研究开始于对如学习、记忆、问题解决和情绪等基本心理过程的分析，不久，他们发现这些过程在行为表现中是领域特殊的，例如，当一个人在下棋、解决数学问题或写文章时，其记忆、问题解决、推理过程或许是有极大差异的。

一开始，文化心理学研究那些与研究自身文化有明显不同的、具有广泛多样性、以前不熟悉的文化中的行为，继而迅速发展到了系统地探究不同文化背景下养育的个体的不同经验史。当前文化心理学这一领域，实质上代表了对所有人类行为文化特殊性的认识。不同文化群体中的成员，基本的心理过程可能产

生差异较大的行为表现、态度、自我概念和世界观 (L. L. Adler & Gielen, 1994 ; Berry, Poortinga, Segall & Dasen, 1992 ; Diaz-Guerrero, 1990; Shweder & Sullivan, 1993)。文化的作用正在被逐步地认识并整合到了心理学的所有领域：一方面是毕生发展、社会行为、情绪、思维的研究和理论。<sup>[7]</sup>另一方面是工业组织、临床、咨询心理学的实践。<sup>[8]</sup>

逐步觉察到文化因素在所有心理学领域的作用在美国心理学会 (APA) 的 1994 年年会中得到了证明。除了一整天关于“评估和干预中文化的敏感性”的继续教育研讨会以外, 年会的名家演讲都致力于这样一个一般性主题——“跨文化心理学的国际视角”。在一系列演讲中, 6 个国际知名心理学家讨论了这一主题的不同方面。这些演讲本身也是关于“文化差异性: 美国的未来”这一主题分会议的一部分, 它分布在为期 5 天的整个 APA 年会中。跨文化取向的心理学快速发展的进一步证据是新学术期刊《文化和心理学》(Culture and Psychology, 1995) 的建立。

文化差异和文化障碍在 20 世纪第一个 25 年内, 心理学家开始发展跨文化测验的工具, 他们希望它至少在理论上可能测量独立于文化经验影响的“遗传性智力潜力”。认为个体的行为覆盖着一种文化的表层饰面, 其渗透力 (penetration) 成为当时被称为“与文化无关”的测验的对象; 以后的遗传学和心理学的的发展表明这一概念是错误性的。我们现在认识到在机体发展的所有阶段遗传和环境的联合作用, 其影响在所产生的行为结果中是不可分离地交织在一起的。对人类来说, 文化几乎渗透到环境的所有领域。由于个体的所有行为都受生活于其中的文化环境的影响, 也由于心理测验只测量行为的样本, 所以文化的影响会而且应该反映在测验成绩中; 试图设计一种不受文化影响的测验是毫无益处的。后来, 目标成为建构一种以不同文化中

所共有经验为前提的测验。因此,诸如‘文化—共同’‘文化—公平’‘跨文化’的术语,代替了早期的与文化无关的标记。

尽管如此,没有一个测验能够被普遍应用或对所有文化都同样公平。有着各种各样的文化公平测验,就像文化差异有着许多参数一样。体现文化公平的,在一种情境中可能是非阅读测验;在另一种情境中可能是非语言测验;第三种情境中可能是操作测验;在第四种情境中则可能是言语测验的翻译改写。各种可用的跨文化测验是不能互换的,但在不同类型的跨文化比较中却是有用的。而且,任何测验对一种文化以上的群体,尤其是这些文化极不相似时,它不可能是同样公平的。一方面,我们在测验操作中努力减少文化的差异;另一方面,跨文化测验不可能完全排除这种差异。每种测验倾向有利于来自形成测验的文化中的群体。譬如只用纸笔或呈现抽象的任务而没有即时实践意义的测验,会有利于一些文化群体,而不利另一些文化群体。情绪和动机因素可能影响测验成绩。在许多随文化不同而不同的相关因素中,可能提到的是测验内容的内在兴趣、与测验者的关系、做好测验的驱力、超过其他人的愿望、过去单独地或合作地解决问题的习惯。

当个体离开他原先生活的文化或亚文化,在另一种文化中的奋斗、竞争或取得成功时,文化差异就变成了文化障碍。然而从更广阔的视野上看,正是不同文化间的接触和交流促进了文明的进步。文化孤立或许能使个体感到自在、舒适,却导致了社会的停滞。

另一个相关的概念是文化剥夺。尽管这一术语在许多不同的意义上使用,但弗恩斯坦 (Feuerstein, 1980, 1991; Feuerstein & Feuerstein, 1991) 赋予这一概念以特定的含义,并且使它成为他的认知训练计划中的重点。他认为文化剥夺是由缺少中介学习

经验而产生的一种“认知可调节性的减少”状态。文化积累起来的知识的代代传递是一种特殊的人类行为，在这一过程中，父母或其他照顾者担任“起中介作用的人”的角色，选择和组织儿童所面对的各种刺激。弗恩斯坦认为这种中介学习是儿童认知发展所必需的，因为它有利于儿童确定学习态度、方向或其他有利于后来学习的行为模式。不管何种理由，没有经历这种中介学习的儿童，缺少高水平认知功能的前提条件；相反，那些在他们自己文化中已有中介学习经验的个体，发展了继续调节所必需的技能 and 习惯，经过一段时间不长的转换期后，他们便能够适应新文化的要求。

从不同的观点来看，文化刻板印象(cultural stereotypes)的存在可能直接影响个体的测验成绩(Steele, Spencer & Aronson, 1995) 一项长期的研究计划发现，个体对现存刻板印象的认识可以通过分心、自我概念、减少努力和降低成功期望等来影响一些被试的测验态度和动机。这种反应被称为刻板印象脆弱性，可以发现它在性别及种族比较中影响测验分数。有效的肯定行计划需要的不仅仅是具体的补偿性训练。至少在一部分个案中，要求特别注意如语言、数学或空间等具体领域中的成绩期望，防止那些预期有正常或较好成绩的个体所经受的失败。

跨文化测验中的语言 大多数传统的跨文化测验，使用非言语内容，希望获得比较近似于文化公平的度量，表示与言语智力测验所测量的同一些智力机能（见第九章）。作为这一方法基础的两种假设是值得质疑的：第一，不能假设非言语测验和言语测验能测量同一些机能，不管它们看起来是如何相似；某种空间类比测验，不只是某种言语类比测验的非言语版本。一些早期的非言语测验，如军队乙种测验在空间视觉和知觉能力上有较大的负荷，而这与言语和数字能力没有什么关系。甚至如端文

渐进矩阵和其他非言语测验,其目的是要引发推理和抽象概念化,因素分析的结果揭示了在测验分数的变异中非言语因素起了很大的作用(R. S. Das, 1963)。这些发现得到认知心理学最近研究的证实 认知心理学的研究,不断地表明了思维过程的内容特殊性(content specificity) 问题解决的策略和技能,在对特定内容领域的反应中和特定情形中得到发展。

逐渐积累的大量资料,从另一个角度表明,非言语测验可能比言语测验有更多的文化负荷。在许多国家,对各种不同文化组群的研究发现,操作和其他非言语测验比言语测验具有更大的群体差异(Irvine, 1969a, 1969b; Jensen, 1968; Ortner, 1963, 1972; Trimble, Lonner & Boucher, 1983, Vernon, 1969)。也有证据表明,图形测验比言语和数字测验更容易受训练效果的影响(Irvine, 1983)。使用图片呈现本身可能不适合那种不习惯于用图画表现的文化。对某一客体两维的重现不是与其原型完全一样的复制品,它只是呈现了某些线索,这些线索作为过去经验的结果导致对这一客体的知觉;如果这些线索是高度凝缩的,如一张头像代表整个人,或者是缺少必需的去经验,就不可能产生正确的知觉。现在大量的经验研究资料表明了不同文化中个体图片知觉的显著差异(R. J. Miller, 1973; Segall, Campbell & Herskovits, 1966)。

从另一角度来看,非言语测验、空间知觉测验,常常需要相对抽象的思维加工和分析性认知风格,这是西方文化中产阶级的特征(J. W. Berry, 1972; R. A. Cohen, 1969) 其他文化背景中抚养的人可能不习惯这种解决问题的方法。各种文化对于概括化,对于在各种不同经验中寻找共同点的价值是不同的。在一些文化中,行为经常与具体领域和具体情境相联系。对某一问题的反应,可能依赖于谁提出问题和它涉及什么类型的内容

(Cole & Bruner, 1971; Goodnow, 1976; Neisse, 1976, 1979).

当然，在说同一种语言的不同文化中施测言语测验不存在程序上的难题。当语言不同需要对测验进行翻译时，就产生了常模可比性和分数等值性的问题。应该注意到，简单的翻译往往是不够的，一般需要进行一些改编和修订；有许多项目内容可能在一种文化中比在另一种文化中更为熟悉，甚至一些细微的差别都可能显著影响测验成绩。例如一种语言中不同词之间，词的相对长度或声音的相似性，可能改变两种语言的阅读困难程度(Duran, 1989; Marin & Marin, 1991)。由于测验的翻译本在许多方面不同于其原版，因此不能认为它们之间有可比性，《测验标准》(AERA, APA, NCME, 1983, 第十三章)清楚地说明了将测验使用于任何人群时，其翻译版本的信度、效度和常模都必须单独建立。

应该牢记，影响测验反应的文化因素，也可能对测验样本的更为广泛的行为领域产生影响。例如，在一个说英语的文化里，英语不熟练，可能不仅妨碍儿童的智力测验，而且不利于其学校作业、与同伴交往及其游戏活动；还可能因此妨碍儿童以后的智力和情绪发展。在成年人身上，它可能严重限制其工作绩效、人际关系和其他有意义的活动。还可举出许多相似的文化差异。有些是认知方面的差异，如阅读不能或解决抽象问题的无效策略；有些是态度或动机方面的，如对智力活动缺少兴趣、对权威人士的敌意、成就动机低或不良的自我概念。以上的这些状况可以通过语言指导、机能性读写算训练、个人咨询和心理治疗等方法得到改善。所有这些都可能既影响儿童和成人的测验成绩也影响其日常活动。

测验成绩和日常活动中语言的重要性，导致了在本土语言和外来语言方面的语言熟练性测验的发展。在美国，大多数可

用的测验是关于英语和西班牙语的,测验的清单以及其研究、发展及评价的参考资料,可在杜兰(Duran)的著作中(1989, pp. 574 ~ 577)找到。关于双语者的一般讨论见德格洛特和贝里(de Groot & Barry, 1993)的著作。

**测验情境** 当今世界快速扩展的文化间接触,增加了测验施测于不同文化个体的可能性。每一个施测者都可以期望测验一个或多个来自不同于其自身文化的个体。因此测验施测者的训练应包括一种或多种不同文化的相关知识,并特别注意那些可能影响个体行为发展的文化差异,尤其是这些差异对被试测验反应的可能影响。不同测验行为的一般来源有:个体自我概念、世界观、自我暴露程度、单独或合作地解决问题习惯等方面的变量。

正如本章前面部分说明的那样,从测验中去掉那些由于个体文化背景而认为特别困难的部分,只能减弱预测效度,却不能使被试去注意为了在可能的情境中进行有效操作而需要加强的方面。相应的,传统的方法正在被一种新的方法所代替,即把重点转移到测验情境中的施测者行为。

80年代和90年代出版了许多关于测验其他文化成员的施测者训练和适当行为的手册、指导纲要、学术论文及其他文献(Atkinson, Morten & Sue, 1993; Myers, Woyhlford, Guzman & Echemendia, 1991; Stricker 等, 1990)。其中有些注重于学生的施测(Samuda, Kong, Cummins, Lewis & Pascual-Leone, 1991),有些把重点放在儿童身上(Miller-Jones, 1989; Rogoff & Morelli, 1989),还有一些着重于某一特定的文化人群或其中的分支(Dana, 1984; Diaz-Guerrero & Szalay, 1991),但大部分是关于跨文化测验中的一般问题。这些出版物,从美国心理学会出版的非常一般性的、简要阐述的指导纲要,如1991年的简明手册和1993年重印在



《美国心理学家》(American Psychologist) 上的指导大纲, 到达纳(Dana)1993 年出版的论述详尽的著作。任何准备实施测验的人员, 应该仔细参考这两类一般性的资料。另外, 还需注意的是有关双语者的多样性、问题和优点的综合讨论, 这在德格洛特和贝里(de Groot & Barry, 1993)的著作中可以找到。

实质上, 跨文化测验中施测者的特殊作用包括: 第一, 与被试在一起的开始阶段, 获得有关文化认同、同化的程度和类型、最初文化的特征等方面可能影响个体测验成绩的充分信息。第二, 施测者自身的行为需适应被试的需要。在这方面, 施测者必须考虑如何介绍测验, 如何解释测验目的, 以及如何激励被试进行适当的操作。而且, 施测者自身要和被试建立良好的人际关系。测验分数的解释和使用, 应该明显地考虑到文化因素, 还应该考虑反馈的性质和反馈的对象。

理查德·达纳(Richard Dana, 1993) 编制了一套问卷, 以便帮助施测者从被试身上获得需要的信息。其中一部分是普遍适用的, 另一部分则是为特定文化(如土生土长的美国人、亚裔美国人)或其中的特定亚文化而编制的。作为一种结束性的评论, 必须还要补充的是: 至少要获得有关被试经历的一些相关信息, 以便恰当地理解和使用测验分数, 要注意不论被试是什么人, 不管其文化差异如何, 测试本身都应是一种令人满意的活动。

## 注 释

- [1] 见巴尔特斯(Baltes, 1968)、波特威尼克(Botwinik, 1984, 第二十和第二十一章)、布斯(Buss, 1973)、内塞尔洛得和李塞(Nesselroade & Reese, 1973)、内塞尔洛得和凡爱(Nesselroade & Von Eye, 1985)、K. W. 斯切爾(K. W. Schaie, 1973, 1988a)、斯切爾和荷兹格(Schaie & Hertzog, 1986)等人的著作

- [2] 这一情境下使用的群体是指同一年龄的一组人群,也就是说,他们出生于同一年或其他特定的时期。
- [3] 起初,测验包括瑟斯顿(Thurstone)通过对一般认可的智力的某些方面进行因素分析而编制的“基本心理能力成套测验”(Primary Mental Abilities Battery)(见第十一章);后来的版本是为这一项目而编制的谢伊-瑟斯顿(Schaie-Thurstone)“成人心理能力测验”(K. W. Schaie, 1988);最后的分数是整体结构的而不是单个测验的,目的是为了提供较为稳定和一般性的结果。
- [4] 见下列诸人的著作:本森和斯切尔(Bengtson & Schaie, 1989),比勒恩和本森(Birren & Bengtson, 1988),比勒恩和斯切尔(Birren & Schaie, 1991),卡莱克和萨尔桑瑟(Craik & Salthouse, 1992),费斯克和奇里伯格(Fiske & Chiriboga, 1990),哈里普和帕内克(Haplip & Panek, 1993),考斯勒(Kausler, 1994),纳迪恩(Nadien, 1989),索得雷格(Sonderegger, 1992),斯库尔兹和艾温(Schulz & Ewen, 1993),威里斯和斯切尔(Willis & Schaie, 1986)。
- [5] 这种一般程序的具体使用,可以在 K. W. 斯切尔(1965)以前引用的横向序列研究设计中所体现的时间滞后比较中见到。
- [6] 尽管大多数全面调查是针对 SAT 的,但在其他大学入学测验中如美国大学测验(American College Test, ACT)计划,在中学和小学水平测验中,也有分数下降的相似现象出现。
- [7] 见下列诸人的著作:戈曼里和布罗兹恩斯凯(Gornly & Brodzinsky, 1993),柯塔亚马和马库斯(Kitayama & Marcus, 1994),米斯基和罗格弗(Mistry & Rogoff, 1985),鲁格特、勒斯特和布纳兹尔顿(Nugent, Lester & Brazelton, 1990),罗格弗和查瓦加(Rogoff & Chavajay, 1995),史密斯和邦得(Smith & Bond, 1993),托平、克罗威尔和柯巴亚史(Topping, Crowell & Kobayashi, 1989)。
- [8] 见下列诸人的著作:弗雷里奇、雷贝克和萨弗谢克(Freilich, Raybeck & Savishinky, 1991),培得森(Pedersen, 1987),培得森和艾维(Pedersen & Ivey, 1993),特安迪斯、顿内特和霍格(Triandis, Dunnette & Hough, 1994)。

## **第四部分**

---

### **人 格 测 验**



## 第十三章 自陈人格调查表

虽然“人格”一词有时是从广义上使用的，但在传统心理测量术语中，“人格测验”是测量情绪、动机、人际关系和态度特征的工具，以区别于能力。<sup>[1]</sup>在后面的4章中，我们将考察各种各样的人格测验，就当前的目的而言，对可资利用的工具根据从个体那里获得资料的方法加以分类。这一章涉及自陈调查表，第十四章我们将讨论测量兴趣和态度的技术。这两章涉及的测验主要是纸笔式的、适合团体进行的自陈调查表，虽然它们中的大多数当然也可以采用个别评定的方式。人格特征评定的投射技术的使用在第十五章讨论，第十六章我们将对若干人格评定方法——其中有些还处于实验阶段——作一评述。

可资利用的人格测验达数百种，特别大量的是人格调查表和投射技术。本书中我们主要关心的是已经被用于评定人格的方法类型，作为举例说明我们将简要地描述每一种类型中最广为人知的少数几个测验。已经有了几本关于通过测验和其他技术进行人格评定的书籍。关于这一主题更详细的资料，读者可参考有关专著(如 Angleitne & Wiggins, 1986; Butcher, 1995; Lanyon & Goodstein, 1997; and J. S. Wiggins, 1973/1988)。

在人格调查表的发展中，在项目的编制、收集、选择和分组方面已经产生了几种方法。目前使用的主要方法是基于内容关联、经验准则确定、因素分析和人格理论的，每一种方法将在以

后章节中逐一讨论和说明。然而应当指出的是，它们不是二择一的或相互排斥的技术。理论上所有这些方法都能在某一人格调查表的编制中结合起来。事实上，目前大多数调查表使用两种或两种以上方法。

虽然某些人格测验被用作团体筛查工具，但大多数人格测验主要应用于临床和咨询。因为这一原因，接下来的 4 章须与第十七章中提及的那些特定情境问题一起讨论。就它们目前的发展状况而言，大多数人格测验应该或者被看作是个体评定中的辅助手段，或者被看作是研究的工具。

## 内容关联程序

自陈人格调查表的原型是伍德沃斯个人资料调查表 (Woodworth Personal Data Sheet)。该调查表在第一次世界大战期间产生并使用 (参见第二章)，主要的目的是对精神病人的访谈标准化，并采用团体测试的方式。因此，伍德沃斯通过精神病学文献和精神病医生的会议两条途径收集了有关神经症和前神经症一般症状的信息。调查表的问题最初就是参照这些症状而形成的。问题涉及这样一些偏离行为，如变态恐惧或恐怖、强迫、噩梦和其他睡眠障碍、过度疲劳和其他身心症状、幻觉以及运动障碍，如痉挛和震颤。在最后的选项选择中，伍德沃斯运用了某些统计检验，这将在下一部分讨论。不过，正如这些项目的来源和一般把某些行为当作适应不良的行为所表明的，显然这一调查表的编制和使用主要强调项目的内容关联性。现代人格调查表的编制继承了伍德沃斯的传统，这在下列事实中是很明显的，即当新的项目产生时，项目作者的说明一般还是基于对所评定

的行为领域的内容分析。

自陈人格调查表编制的内容关联方法的现代例子是症状检查表 90 项—修订版 (Symptom Checklist - 90 - Revised, SCL - 90 - R - Derogatis, 1994), SCL - 90 - R 旨在筛选心理问题和心理病理症状,包括简要描述的 90 项症状(如食欲不振、软弱乏力或头晕目眩),要求受测者用五点量表指出在过去 7 天中他们在每一问题上的苦恼程度。SCL - 90 - R 提供了男女成人和青少年非病人及精神病住院病人和出院病人的各种常模。然而,有些常模不具有充分的代表性,例如,精神病住院病人常模组主要是社会经济地位较低的,而青少年组主要是中产阶级并几乎都是白人样本。

SCL - 90 的项目类似于那些早期的检查表,不仅项目的选择基于内容关联和临床效用,而且有些项目通过中介的量表,如霍普金斯症状检查表(Hopkins Symptom Checklist)和康奈尔医学索引(Cornell Medical Index)(Derogatis & Lazarus, 1994),可追溯到伍德沃斯个人资料调查表。项目由心理病理学的九个维度组成,称之为躯体化(somatization)、抑郁(depression)、焦虑(anxiety)、敌意(hostility)、精神病性(psychoticism)、人际敏感性(interpersonal sensitivity)、恐怖性焦虑(phobic anxiety)、偏执(paranoid ideation)及强迫(obsessive-compulsive)症状。对这些量表的因素分析表明,它们彼此相关,因此在鉴别诊断方面不太有用。不过,从检查表得到的总体指标(global indices)被证明是心理病理症状及其严重程度的可靠指标(Payne, 1985)。SCL - 90 和一些相关的工具,如简明症状问卷(Brief Symptom Inventory)在评定通过治疗过程的变化和研究各种治疗的效果方面,作为成套测验的一部分使用是最合适的。

人格调查表编制的内容关联方法的主要优点在于方法的简

单和直接。尽管这些特征使它们有可能成为相对简明和经济的工具，但它们的透明度也较其他方法为被测者提供更多有意识地操纵结果的机会。基于内容的工具一般缺乏防止和监测反应偏向的特征，这在本章的后面加以讨论(Bornstein, Rossner, Hill & Stepanian, 1994)。由于这个原因，惟一地依靠它们作为决策的基础是不提倡的。

## 经验准则确定

**基本方法** 经验准则确定(empirical criterion keying)指的是根据某些外部准则而形成的一种记分的基准。这一程序涉及选择保留的项目和每一反应的分数权重的安排。在先前提到的武德沃斯个人资料调查表的编制中，应用于最后项目选择的统计检验提出了准则确定的方法。基于这一程序的原则是在常模样本中以这种频率发生的行为特征不能作为异常的指标。如果这一问卷中 25%或更多的常模样本以否定的方向对一个项目作出回答，则这个项目不被保留。对照不同组的方法也被用于项目选择。只有那些在先前诊断为精神神经症组的出现频率至少两倍于常模组的症状被保留。

然而,尽管使用了这些经验检查的方法,内容关联方法基本上仍然是依赖于对问卷项目的一种原原本本的或真实的解释。对每一问题的反应被看作是该问题所描述的特定疑难问题、信念或行为的实际存在或不存在的一个指标。而在经验准则确定中,反应被作为与之关联的准则行为的特征或症状看待。米尔(Meehl, 1945)早期在关于这一方法的描述中写道:



……把人格调查表的言语型式作为一种“自我评定”或自我描述，其价值需要假定被测者观察自我的正确性，这不是最有效的。宁可把对测验项目的反应作为言语行为中真正感兴趣的一部分，是一种认识，其价值比认识这个项目表面上所要调查的事实材料更大。因此，如果一个疑病症患者(hypochondriac)说他“头很疼”，重要的事实是他这样说了(p.9)。

一份自陈调查表无疑是一系列标准化的言语刺激，当遵循准则确定程序时，由这些刺激诱发的反应按照由经验确定的行为相互关联加以评分。对待它们就像对其他心理测验反应一样。问卷反应可能与个人对现实的知觉相应，这并不改变这个情况，它只是提供了一种假设去解释由经验确定的某些项目的效度。

### 明尼苏达多相人格调查表

人格测验建构方面经验准则确定的突出例子是明尼苏达多相人格调查表(The Minnesota Multiphasic Personality Inventories, MMPI)。最近,MMPI 已经再版,并重新建构为两个独立的版本,即 MMPI - 2 (Butcher, Dahlstorm, Graham, Tellegen & Kaemmer, 1989)和 MMPI - 青少年版(MMPI - A - Butcher 等,1992)。尽管有了这些新版本,对它们的讨论仍然不能不参照最初的 MMPI,不能不提到它在人格评定历史中的作用。虽然对它进行透彻的描述超过了本书的范围,但必须指出,在差不多半个世纪中,MMPI是使用最广泛和研究得最透彻的人格测验。[2]

在许多方面,MMPI 作为一种工具,是它自身成功的牺牲品。一位临床心理学家(Starke R. Hathaway)和一位神经精神病

学家(J. Charnley Mckinley)于 30 年代认为,它最初于 40 年代通过一系列论文发表,成为精神病诊断过程中的一种辅助工具。<sup>[3]</sup>从那以后,它在检测心理病理和鉴别那时还相当粗糙的疾病种类方面的效果导致它被用于日益扩大的目的,超越了它最初的目标。到了 60 年代,MMPI 稳定而牢固地作为主导的人格测验,用于正常被试的咨询、就业、医学、军事和法律等方面,其使用次数和用于精神病人的相同,甚至更多。到了 80 年代,MMPI 文献累计数千种,它的 13 个分量表在各种不同的人群中使用,基于它的项目编制了数百种特殊量表,量表和剖面图之间有各不相同的经验相关。然而,按照心理病理学、人格理论以及测验建构方面正视图的进步来看,它在概念和心理测量学问题上出现了更多的麻烦。而且,已经证据清楚地表明,基础狭窄和过时的测验常模对当前的被试者是不适合的,因此确定异常所依赖的基础是不可靠的(Colligan, Osborne, Swenson & Offord, 1983, 1989)。事实上,最初的标准化样本已经变得与非常模参照组有些相似,分量表就是根据非常模参照组来确定的。随后收集的关于剖面图代码的许多资料为常模解释提供了基础。

因此,负责 MMPI 重新标准化的委员会,在试图汇集与人格评定有关,尤其是包含在 MMPI 基本结构中 with 心理病理学评定有关的丰富的解释性资料时,面临着使工具现代化的艰难任务。为了保持连续性,委员会有选择地保留了大量的项目——所有最初的临床和效度量表以及很多补充量表——连同它们内在的弱点。主要的变化是:调查表常模的完全重新制定,8 个原始临床量表 and 所有内容量表 T 分数的建立,修订和删除过时的或引起反对的项目,增加新的项目 制定一些新的效度、补充量表 and 内容量表,以及把调查表分为两个版本,以适合不同的年龄组。

明尼苏达多相人格调查表—2 (The Minnesota Multiphasic

Personality Inventory - 2) MMPI - 2 包括 567 个肯定的陈述项目。参加测验者只需回答“是”或“否”。开始的 370 个项目除了编辑的变化和重新安排次序外, 实际与 MMPI 一致, 它们提供了对原先 10 个“临床”量表和 3 个“效度”量表评分所需的所有反应。其余的 197 个项目(其中 107 个是新的) 是对全部 104 个新的、修订的和保留的效度量表、内容量表和补充量表及分量表进行记分所需要的。这些量表和分量表组成了完整的调查表。项目的内容范围很广, 包括一般的健康、情感、神经和动作症状, 性、政治和社会态度, 教育、职业、家庭、婚姻问题, 以及很多众所周知的神经病的或精神病的行为表现, 如强迫状态、妄想、幻觉、多疑、恐怖(phobias) 及性施虐(sadistic) 和受虐(masochistic) 的趋向。达尔斯琼姆(Dahlstrom, 1993a) 已经编写了一份补充手册, 提供了比较 MMPI - 2 和 MMPI 原版的项目的所有必要的信息。有些作为例子的项目, 在目前测验中的序号如下<sup>[4]</sup>。

我的睡眠是断断续续的, 容易被惊醒(39)

我相信有人在暗算我(138)

我对性感到担心(166)

无聊的时候, 我就会惹事以求刺激(169)

大多数人内心不喜欢帮助其他人(286)

MMPI - 2 提供了 10 个基本的“临床量表”的分数, 它们和原先 MMPI 的一样, 如下所示:

1. Hs: 疑病症(Hypochondriasis)
2. D: 抑郁症(Depression)
3. Hy: 歇斯底里(Hysteria)

4. Pd : 精神病态偏离( Psychopathic deviate)
5. Mf : 男子气—女子气( Masculinity-femininity)
6. Pa 妄想狂( Paranoia)
7. Pt : 精神衰弱( Psychasthenia)
8. Sc : 精神分裂症( Schizophrenia)
9. Ma : 轻躁狂( Mania)
10. Si : 社会内向性格( Social introversion)

这些量表中的 8 个量表产生于 40 年代,通过项目的准则确定在小临床样本和正常控制组之间进行区分,前者每个样本近 50 人,代表那时传统的精神病诊断,常模控制组有 724 位明尼苏达大学医院病人的来访者和亲戚( Hathaway & McKinley, 1940, 1943)。男子气—女子气量表( Masculinity - femininity scale )最初打算区分同性恋和异性恋的男子,产生于男性士兵和女性空中服务小姐在项目赞成频率上的差异。在这一量表上的得分表明一个人的兴趣和态度与对他或她的性别组的刻板印象的匹配程度。后来加上社会内倾量表( Social introversion scale )来自两组不同的大学生的反应,这些学生是根据一个内倾—外倾测验的极端分数进行选择的。

在完整地保留 MMPI 的基本量表时(除了 9 个删除的项目、一些编辑上的变化和项目的重新排序), MMPI-2 编制者力求保留与剖面图代码的解释有关的大量临床上有用的信息。这种代码解释基于这些量表的得分形式( Graham, 1993; Greene, 1991),然而,也完整地保留了隐含在这些量表中的心理病理学的陈旧观点以及不成熟地、有缺陷地应用对照标准组记分经验方法导致的结果。另外的问题,如基本量表的多维性和重叠性,也被保留了下来( Helmes & Reddon, 1993)。

原始的 MMPI 的最显著的特点是 3 个所谓的效度量表的使用,它们也被保留在 MMPI-2 中。<sup>[5]</sup>从技术意义上来说,这些量表与效度无关。实际上,它们是对粗心、误解、装病以及对特殊反应的操作和参加测验者态度的检查。效度分数包括:

撒谎分数(L):它基于一组测验题,这些题目使回答者表现出赞成的样子,但不可能以赞成的方式真实地回答(如:我不喜欢我认识的每一个人)。

不寻常分数(F):由 60 个项目(来自于原始量表的 64 个项目)的答案决定,MMPI 标准化组群中不到 10%。尽管这些项目代表着不合要求的行为,但它们也不属于特定的异常形式,因此任何人都不能实际表现出这些症状或绝大部分症状。一个高的 F 分数可能表明存在记分错误、回答中的粗心、十足的怪僻、精神病过程、故意装病等现象。

校正分数(K):使用专门挑选出来的项目的又一组合,这一分数提供相当微妙的测试态度的度量,一个高的 K 分数可能表明了防御性或试图“装好”;一个低的 K 分数表明非常诚实和自我批评或故意“装坏”。

前面两个分数(L、F)通常用于对测试记录的全面评价。如果其中一个分数超过一个特定值,这个记录就被认为无效。另一方面,K 分数是想作为抑制变量起作用,它用来计算一个校正因素和某些临床量表分数相加,获得经过调整的总分。使用 K 分数的有效性值得怀疑,因此,量表得分可以有也可以没有这种校正。尽管一个不寻常的高的 K 分数本身使记录可疑,需要进一步详细研究,但中等高度的 K 量表分数事实上可能反映了自我的强度和积极的调整。通过个人的生活经历和生活环境来评

定分数高低尤其重要。

在 MMPI-2 的 21 个补充量表中有三个新的“效度”指标,能有助于评定参加测验者回答调查表的仔细和诚实程度。它们是 F 反向量表(Fb)、变量反应不一致性量表(VRIN)和真实反应不一致性量表(TRIN)。Fb 量表基本上是调查表后半部分原始 F 量表项目的扩展,而 VRIN 和 TRIN 是两份新量表,由相似或相反意义的成对项目组成,目的是检测不一致或矛盾的回答。

MMPI-2 的基本剖面图(图 13-1)包括 13 个来自原始版本的效度和临床量表,还有 15 个内容量表、27 个内容成分量表、21 个补充量表和 28 个哈利-林戈斯(Harrin-Lingoes)分量表的单独的剖面图。<sup>[6]</sup>这些量表和分量表有些是新的,有些是从最初的量表中保留下来的,然而,所有这些量表的记分都使用 MMPI-2 年龄在 16~84 岁之间的 2600 名成人常模样本。这一样本远比最初的明尼苏达常模组更能代表当前的美国人口状况,他们来自 7 个不同的州。该样本正是试图根据重要的人口统计变量,包括性别、年龄、种族,反映美国人口状况(Dahlstrom & Jellagen, 1993)。不过,这个样本的代表性受到质疑,主要是因为与 1980 年人口调查数据相比,它的职业水平和教育程度较高,不能很好地代表西班牙裔和亚裔美国人(Duckworth, 1991)。

MMPI 重新制定常模的主要结果是临床剖面图的分数高度下降,这个受到预期的变化除了是由于最初样本所特有的一些具体因素和发展 MMPI 所使用的方法之外,还可能受到代际差异的影响(D. S. Nichols, 1992)。无论如何,在研究具有临床意义的量表正视图时所需要的临界 T 分数现在是 65,或大约高于平均数 1.5 个标准差,而不是 70。MMPI-2 中的另一个富有新意的变革是在 10 个临床量表中的 8 个和所有的内容量表中使用统一的 T 分数——这与线性导出的或标准化的不同,它使所



面图形式和代码的不同,因此,将来自大量 MMPI 剖面图的解释文献和临床经验的结果概括到 MMPI - 2 的可行性产生了许多争议(Chojnacki & Walsh, 1992; Morrison, Edwards & Weissman, 1994; Tellegen & Ben - Porath, 1993)。资料表明对于能清楚区分量表分数的明确的剖面图, MMPI 和 MMPI - 2 代码类型之间的一致性似乎与其中一个版本重复施测所获得的结果之间的一致性一样好( Archer, 1992b; Graham, 1993)。不管怎样, MMPI - 2 手册具有这样的信息,它使得使用者能根据一个版本的反应对两个版本所产生的分数进行比较。虽然这个设想还是有问题的(参见 Ben - Porath & Tellegen, 1995),但是,在两个版本的转换时期作为使用者所采用的一种在经验上有道理的方法还是被一些人所认可(参见 Humphrey & Dahlstrom, 1995)。

明尼苏达多相人格调查表一青少年版(MMPI - A) MMPI - A 是 MMPI 特别用于青少年的新模式。它体现了 MMPI 和 MMPI - 2 的主要特征,包括 13 个基本量表,但减少了测试题目,只有 478 项,以适应年轻的参加测验者。其中的新项目 and 量表覆盖了与他们特别有关的领域,如学校、家庭问题,最重要的是提供了适合于年龄的常模。MMPI - A 使用年龄在 14 ~ 18 岁之间的 1620 名当代青少年为常模样本,同时收集了 713 名同年龄段的临床样本,用于比较和效度研究。

除了与 MMPI - 2 共有基本的临床和效度量表外, MMPI - A 有自己的效度量表(F1 和 F2),有些内容量表和补充量表及分量表是 MMPI - A 特有的,有些是这两个工具共同的。虽然大量研究,包括由达尔斯琼姆等(1922)和马克斯、西曼和海勒(Marks, Seeman & Haller, 1974)发表的常模和转换表都支持 MMPI 用于青少年,但这些研究不一定适用于 MMPI - A。它是一种新的工具,而不只是一个修订本。因此,它的使用不得不通过研究和解



释材料的积累来决定，而这些研究从它发表时就开始了( Archer, 1992a; Butcher & Williams, 1992 ; Williams, Butcher, Ben-Porath & Graham, 1992)。

对明尼苏达多相人格调查表的结论性评价 尽管 MMPI 的最初量表被认为是不成熟地运用经验主义的典型产物，并不断有传说说它的地位岌岌可危，但它还是存活下来了。事实上，虽然它的最初目的是帮助精神病分类的进程，因而它的发展中所遵循的程序使它不适用于正常个体的人格评定，但 MMPI 除了广泛地用于精神病病人之外，还广泛地被应用于正常人。它和它的修订本已把一些附加的程序和解释策略与最初经验式获得的调查表框架结合起来。这些特征包括根据内容把项目分组而编制的许多量表( Butcher, Grahem, Williams & Ben-Porath, 1990 )和编制补充量表时因素分析的使用 ( Welsh, 1956 )，这些特征在过去已经提到过，但值得进一步注意。

探索 MMPI 解释这一复杂任务的新方法继续向前发展。最新的方法之一是使用结构概要 ( structural summaries )，它使我们在明尼苏达调查表中各种相互有关的多重性量表时，具有内部一致性和简易性。这一基于量表、分量表和项目分析的方法，目的在于减少为解释结果所必需的维度，它们存在于任意分类的各种量表中，如“补充量表”“内容量表”“哈利—林戈斯量表”中。最突出的维度用来把各种类别组织成一个“结构概要”型式，这类似于埃克斯纳( Exner )使用罗夏测验时发展起来的方法(见第十五章)，这种 MMPI 解释的方法在发展的早期阶段，需要进一步的研究和确认。不过，这种方法已被运用于 MMPI - 2 和 MMPI - A，前者使用在内容分析基础上获得的结构维度，后者使用通过因素分析获得的维度，看来它们是很有前途的 ( Archer & Krishnamurthy, 1994; Archer, Krishnamurthy & Jacobson,

1994; Nichols & Greene, 1995)。

与大部分其他测验一样,MMPI-2 和 MMPI-A 飞速发展的两个方面是施测、记分和解释的程序计算机化,以及把测量工具翻译成多种语言。传统的适用于 MMPI 的计算机施测和记分,也可用于新的调查表,几个自动的解释性工具也一样。另外,MMPI-2 的计算机化改编版本已发表,并有良好的结果(Roper, Ben-Porath & Butcher, 1991, 1995)。

第一次把最初的 MMPI 翻译出来花费了几乎十年的时间,有关 MMPI-2 跨文化改编本的工作甚至在它公开发表之前就开始了。在它出现的头三年里,MMPI-2 有 15 个翻译计划已经完成或在进行中。6 个西班牙文翻译版或改编版,包括 2 个用于美国的西班牙文版本已编写好。MMPI-2 的翻译和国际改编手册也可资利用(Butcher, 1996)。那些使用翻译本和改编本的人所收集的数据中一个值得注意的结果是,现在的跨文化常模样本得分与 MMPI-2 标准化常模比先前的这些样本与 MMPI 常模更接近。

为了在基本方面不变的情况下改进人格调查表的经典版本,MMPI 重新标准化委员会给它自己设立了两个困难的、在很大程度上来说是矛盾的目标,时间将作出判断:委员会所作出的决定究竟是会把明尼苏达调查表的优越性延伸到下个世纪,还是使得 MMPI 被相似的新一代测量工具所赶上,如杰克森的基本人格调查表(将在这一章的后面部分讨论)或莱斯里·摩莱(Leslie Morey, 1991)编制的人格评定调查表(PAI)——它利用复杂的序列策略,即把逻辑的和经验的方法结合起来,以保证它的量表具有心理测量学的良好性。在此期间,有关 MMPI-2 和 MMPI-A 的书籍和文章发表的速度和范围似乎一直没有衰减(Butcher, 1990; Butcher Graham & Ben-Porath, 1995; Keller & Butch-

er, 1991; Pope, Butcher & Seelen, 1993).

### 加利福尼亚心理调查表

MMPI 在出现多年后, 已作为发展其他广泛使用的调查表的基础。最明显的例子就是加利福尼亚心理调查表 (California Psychological Inventory, CPI)。CPI 采用了 MMPI 中将近一半的项目, 但它是专门用于正常成人的, 在它最近的修订本 CPI 第 3 版) 中包括 434 个项目, 回答‘是’和‘否’, 并产生 20 个量表分数 (Cough & Bradly, 1996)。3 个是用来评定参加测试者的态度的“效度”量表, 这些量表分别称之为安宁 (Well-being, Wb) 以要求常模‘装坏’的反应为基础 好印象 (Good impression, Gi) 以要求常模‘装好’的反应为基础 社会性 (Communality, Cm), 以易为大家接受回答的频次计算为基础。其余 17 个量表提供了支配性、社交能力、自我接受、责任性、社会化、自我控制、适应的成就、独立的成就、同情心和独立性等人格维度的分数。最后两个量表是 1987 年修订版中新增加的。

17 个量表中有 13 个项目的选择以对照组的反应为基础。这些组是根据课程成绩、社会阶层成员、参加课外活动和等级评定等标准划分的。等级评定通过同伴提名获得, 它已被认为是评定人际特质的有效技术 (见第十六章)。对于其余的 4 个量表, 项目最初主观地加以分组, 然后对其内部的一致性进行检查。所有大样本量表的交叉效度分析产生了显著的组间差异, 尽管各个不同的标准组的重叠性是相当大的, 标准相关通常也是比较低的。

和 MMPI-2 一样, 所有 CPI 分数都根据平均数 50 和标准差 10 的标准分量表报告。目前, 这个量表的常模样本是从 CPI 的档案中抽出女性 3000 人, 男性 3000 人, 以便在年龄、社会经济

地位及地理位置方面代表美国的一般人口。常模有男女分开的,也有合在一起的。此外,提供了许多特殊人群的每个量表分数的平均数和标准差。

CPI最早作为‘开放系统’发表于1956年,在需要时有些成分可以除去,有些成分可添加进去(Gough, 1987, p. 1),因此,通过修订,调查表从最初的480项减少到1987年版的462项再到最近的434项——删除了那些引起反感和按照1990美国残疾人条款而引起法律问题的项目(P.L. 101-336),尤其是人事选拔中会引起反感或受到法律质疑的项目。高夫和勃莱德利(Gough & Bradley)利用档案中13000多个被试在所有CPI项目上的数据,力图在内容长度不变的条件下保持基本量表的信度和效度。他们根据与量表成员标准的相关程度,用功能相当的项目来代替那些删除的项目。

对CPI的研究得到大量的信息,这些信息有助于根据单个量表正视图和两个或多个量表上分数的结构或类型对剖面图进行分析,在一定程度上,类似于MMPI的代码类型解释传统(McAllister, 1996)。跨文化研究表明,CPI在探索种族之间的人格差异方面是有用的(见Dana, 1993; Davis, Hoffman & Nalson, 1990)。除了基本量表外,几种其他量表也通过不同途径形成了,例如管理潜能量表(Managerial Potential Scale)和工作定向量表(Work Orientation Scale)(Gough, 1984, 1985)。这些量表备有应用指导手册以帮助专业使用者,特别是有关CPI在人员选拔和职业发展方面的使用(Meyer & Daris, 1992)。

自从1987的修订版以来,CPI还包括了一个三维类型的模型,用来区分通过因素和项目分析确定的3个结构或向量量表上的高分者和低分者。结构量表测量内在一外在、规范接受一规范拒绝和自我实现的较高层次维度。这些量表的前两个量表

分用来把回答者分成四种人格类型( $\alpha$ 、 $\beta$ 、 $\gamma$ 、 $\delta$ ),而第三个向量量表分目的是测量与一个人的类型相关的积极潜能的整合或实现水平。这个类型的模型对调查表的组织使用者来说显然是有吸引力的,但是由于用来获得该模型的程序描述含糊而受到了批评(Engelhard, 1992)。此外,它具有概念的和心理测量学的弱点,这一弱点将一个或多个连续维度任意地分成两部分,并以此为依据对人进行分类这种方法所固有的。<sup>[7]</sup>

### 儿童人格调查表

儿童人格调查表(Personality Inventory for Children, PIC)虽然没有利用 MMPI 项目或资料,但它是用与 MMPI 和 CPI 相同的基本方法学来建构的(Wirt & Lachar, 1981; Wirt, Lachar, Klinedinst & Seat, 1991)。明尼苏达大学的一个研究小组受到 MMPI 的原理的和临床使用的影响,花了近 20 年时间,为年龄在 3~16 岁之间的儿童和青少年设计了 PIC。PIC 和 MMPI 的主要区别在于获得信息的方法不同,是一否的问卷项目不是由孩子而是由了解孩子的成人(通常是母亲)回答。这一方法和在儿童诊所的一般实践是一致的,即以与父母访谈,作为孩子目前的问题和个案史的信息来源。事实上,这个调查表提供了收集这些信息和根据常模和诊断资料进行解释的一种系统的方法。

最初的 PIC 一共有 600 个项目,分为 3 个“效度量表”、1 个一般筛选量表和 12 个临床量表。效度量表包括撒谎量表,包括孩子似乎好得离奇的项目;频次量表,由很少赞同的项目组成;防御量表,用来评定父母对于孩子行为的辩护。筛选量表也叫调节量表,用来鉴别孩子是否需要心理评估。12 个临床量表旨在评定孩子的认知发展和学业成就、几种已经确定的情绪和人际问题(如抑郁、焦虑、退缩和多动),以及家庭的心理气氛。

在现行的 PIC 修订版(PIC-R)中,项目已重新排序,并减至 420 项;施测题册把项目分成逐步加长的三个部分。第一部分(项目 1~131)产生撒谎量表分和 4 个新的广泛的因素量表分;第二部分(项目 132~280)在第一部分上加上其他常规量表的缩减本和一部分关键项目 第三部分(项目 281~420)加上其余部分的项目,这些项目是对 16 个原始量表分、4 个广泛的因素量表分和整套关键项目的全面补充。

原始的 16 个 PIC 量表中的 7 个是在标准组和控制组反应频率的实证性比较基础上发展起来的,在此期间通过一个重复的程序增加项目,直到获得最佳的量表效度。其他 9 个量表基本上遵循内容效度分析程序,依赖这一程序,项目最初的选择根据判断者的提名或量表关联性的等级评定。然而甚至在这些情况中,量表中项目反应的内部一致性的评定和项目的因素分析也有助于量表的构想效度。

在拉泽和格道斯基(Lachar & Gdowski, 1979)的专题论文中,以系统的、全面的效度研究为基础,提供了最初的 PIC 广泛的解释性资料;为修订版编写的补充手册对因素量表也做了同样的工作,并提供了关于简化量表的心理测量资料(Lachar, 1982)。另外,拉泽和他的助手以 PIC 剖面图为基础,用聚类分析对异质的儿童样本进行分类,并研究了与不同剖面图类型相关联的具有诊断意义的显著特征。根据这些不断的研究,发展了剖面图分类的序列规则和计算剖面图相似性指数的方法(Gdowski, Lachar & Kline, 1985; Kline, Lachar, & Gdowski, 1992)。这种解释的方法也用于 MMPI 和 CPI,它是从测验编制阶段到解释阶段的经验准则确定传统的扩展和深入(参见 Kline, Lachar & Boersma, 1993)。

值得注意的是, PIC-R 不是自陈调查表,而是行为观察调

查表。因此,它与第十七章阐述的临床心理学取向的行为评定是一致的。然而父母报告确实有局限性,这些局限性已被 PIC 作者普遍认识和注意到。正如他们所指出的,反应可能部分地反映了父母的动机、态度和个人或文化标准。因此,可以预期在不同观察者如父亲和母亲、在父母的报告和儿童的自我报告之间存在着一些不一致的地方,事实上也已经发现了这些不一致性。处理这一无法避免的问题的一个办法是对特定的反应倾向进行评定。这种反应倾向可能通过“效度量表”而歪曲资料,第二种选择是收集和比较多个观察者的报告,还有第三个途径是收集和比较自我报告和观察者报告的资料。

最后,青年人格调查表(Personality Inventory for Youth, PIY)作为与 PIC-R 类似的自陈测量也已产生。尽管 PIY 在理论上可单独使用,但用来自父母和儿童相结合的剖面图能为全面评定提供基础(Lacher & Gruber, 1995a, 1995b)。PIY 的作者使用 PIC-R 的前面 280 个项目作为其起点,大多数项目经过小小的改变,把第三人称变为第一人称(如“我的孩子经常带朋友们回家”变为“我经常带朋友们回家”),一些项目必须经过更多的修订以符合它们的意图,另一些项目由于不适合用于年龄较小(在 9~18 岁之间)被试的自我报告而不得不删掉,少数项目是全新的。PIY 的最后版本包括 270 个项目,组成 9 个不重复的临床量表、24 个无重叠的分量表和 4 个“效度量表”。前面 80 个项目可作为简化形式用于筛选。PIY 的编制和标准化采用受常规教育的 2337 名学生样本和 1178 个儿童和青少年的临床样本(Lacher & Geuber, 1993)。

关于 PIC-R 和 PIY 还有大量工作要做,50 年代后期到 60 年代早期收集的 PIC 的常模应该更新,有关学前年龄段的研究需要扩充(Knoff, 1989)。至于 PIY,它自己单独使用和它与 PIC

- R 一起使用的临床效用还待确定。然而，这两个工具都有良好的实证基础并专为儿童和青少年使用。

## 测验发展中的因素分析

为了实现对人格特征的系统分类，一些心理学家求助于因素分析技术。这一技术已经结合认知能力的组织问题讨论过了，它适用于通过寻找行为发生的一致范型，来减少解释行为现象的范畴数目。我们可以回忆，主观成分可能参与因素的确定，因为这一过程依赖于每个因素上有最高负荷的度量或项目的考察（见第十一章）。因此，确定用不同测量方法的独立研究得到的因素中的共同因素是困难的，它是特质的名称和数目方面差异的来源。另外，因素分析可以各种方式用于人格特质的研究。其中，两种主要的研究传统已共存了数十年，近年来两者在某种程度上已联合起来。

传统方法之一，集中于人格问卷数据的使用，以吉尔福德（Guilford）和他的合作者们的一系列研究为代表（见 Guilford, 1959,第十六章；Guilford & Zimmerman, 1956）。这些研究者计算来自许多人格调查表的单个项目之间的相关，而不是现有调查表总分的相关。作为这一研究的副产品，编制了 3 个人格调查表，并最后组合成吉尔福德-齐默尔曼气质调查表（Guilford-Zimmerman Temperament Survey）。这一调查表产生 10 个独立的特质分数，每一分数基于 30 个不同的项目，这些特质包括：克制性、支配性、情绪稳定性和友好性。

另一个早期应用于发展人格调查表的因素方法，有时称为“词汇”的传统，存在于卡特尔 40 年代开始的工作中（John, An-



gleitner & Ostendorf, 1998)。为了获得对人格的全面描述, 卡特尔开始在词典中(由 Allport 和 Odbert 编辑, 1936)或精神病学和心理学文献中收集所有人格特质的名称, 差不多有 1800 个条目, 先通过合并明显的同义词加以缩减, 然后将减少了的特质条目让不同的成人组进行关联性评定。这些评定和自我报告资料的相关及因素分析确定了卡特尔称之为“人格的基本根源特质”, 这一名称似乎暗示着比先前的研究证明的结果更具普遍性和稳定性。卡特尔研究途径的特征是不把因素分析看作是减少数据的技术, 而是看作发现潜在的具有因果关系特质的方法( Cattell, 1979)。

**16 人格因素问卷(The Sixteen Personality Factor Questionnaire, 16PF)** 卡特尔和他的合作者以他们的因素研究为基础, 编制了一些人格调查表, 其中最著名的是 16 人格因素问卷, 目前已是第 5 版( Cattell, Cattell & Cattell, 1993; Conn & Rieke, 1994; Russell & Karol, 1994)。16PF 最早发表于 1949 年, 是为 16 岁和 16 岁以上的人编制的, 并产生 16 个特质分数, 如敢为性( Social Boldness)、支配性( Dominance)、警惕性( Vigilance)、情绪稳定性( Emotional Stability)和规则意识( Rule Consciousness)。在 16PF 的不同版本中用相同的字母确定的 16 个因素, 多年来不断精练并已重新命名, 卡特尔最早用来标明特质的深奥的术语大部分已经删除。例如, 现在称为敢为性的这一维度早先用“ Threctia”和“Parmia”表示, 两端分别为害羞和大胆。16PF 的第二级因素, 起先是在 4 和 8 个数目之间, 现在称为“综合的因素”, 并限定为 5 个。这符合后面讨论的普遍的五因素模型。

16PF 的第 5 版只有一种形式和 185 个项目, 大部分内容是根据它们的内容和心理测量学特性从先前的问卷中挑选出来的。16PF 已经根据 2500 个个体样本重新标准化, 这个样本大

致代表 1990 年美国人口调查的情况,包括性别、种族、年龄分布及受教育程度等各方面。16PF 的独特特点之一是调查表的末端在“问题解决问卷”标题下接着呈现的 15 个项目。这些项目组成推理量表,旨在作为心理能力的快速测量。此外,问卷现在还有 3 个反应风格的指标,期望用来评定默认、随机反应和不现实地表现自己是否具有社会需要的特质。

第 5 版 16PF 的 16 个主要因素量表的内部一致性和重测信度,要比早期那些版本好。而且,这一版本的技术手册比早期的手册包含多得多的效度信息,但明显存在于早期版本中的 16 个量表缺乏因素独立性的问题似乎仍然存在。别的研究者用卡特尔最初的变量都不能重复 16 个左右的因素,上述困难就更为突出。相反,采用作为卡特尔系统基础的数据进行的大多数研究只发现 4~7 个因素(Digman, 1990; L. R. Goldberg, 1993)。重复卡特尔的主要因素的第一个尝试是由费斯克(D. W. Fiske, 1949)做的,其结果产生了 5 个因素,这工作现在作为下面将讨论的当代模型的最早版本而被广泛引用。

“五因素模型”及其原理 目前关于人格评定的文章越来越多地注意所谓的五因素模型(Five-Factor Model, FFM),这代表了来自各种因素分析传统的人格研究者之间不同寻常的一致性(Costa & Widiger, 1994; Digman, 1990; McCrae & John, 1992; Wiggins & Pincus, 1992)。同时,模型表示的特定方法导致了相当多的批评和争论(Block, 1995; Carlson, 1992; Goldberg, 1993; Kroger & Wood, 1993; Loevinger, 1994)。<sup>[8]</sup>实质上,五因素模型试图采用分析的层次范型,以便简化大量有关个体情感行为的可资利用的资料,因此,在评定个体和预测他们在特定环境中的行为方面提供了更易处理的信息。这些因素是描述性的而不是解释性的,而且它们不比获得它们的个别项目或具体测验更为基本。

五因素模型与第十一章讨论过的通过对能力测验的因素分析所获得的层次结构相似。虽然在过去的 40 年中,一些研究者都找到了 5 个因素——解释大量人格描述资料之间的大部分相关的必要数目——但是,为了特定的目的,因素的适当数目实际上可能更好是  $5 \pm 2$ , 每个因素所给的名称也是有争议的(参见 Digman, 1990, p. 423; Loevinger, 1994; Paunonen, 1993)。由于因素反映了调查表、量表、反应型式及获得因素的样本的不同,这一点是可以预期的。

有些混淆和误解来自报告五因素模型程序的方法。有人认为通过一些新方法,人格的基本因素已被“发现”。事实上,这些因素和人格测验和评定量表资料的因素分析结果的第二层次水平是一致的。值得注意的是,产生中等广度因素的第二层次已被证明在认知和情感评定两方面都是最可操作和可重复的。因为它们集中或组合了来自行为描述较低水平和定义较狭窄的特质的度量,所以这些因素产生更可靠的分数——且产生的效度也相应较高。如果一个成套测验由层次模型建构,而且如果常模能反映不同水平,如第八章描述的诸如区分能力量表之类的认知成套测验,那么该成套测验就能最大限度地灵活适用于具体的目的。例如,在鉴定出最能说明一个个体的特征的中间水平因素后,能通过考察个人在较低水平的较窄、较详细的测验中的得分进行补充分析。

与五因素模型最密切相关的两个研究者已编制了一种适应其模型的测验。在它的现行版本中,修订的 NEO 人格调查表<sup>[9]</sup> (NEO PI - R, Costa & McCrae, 1992b)提供了 5 个主要的人格维度或范畴的分数和 30 个用来鉴别每一范畴的附加的特质或界面的分数。考斯塔和马克雷 (Costa & McCrae) 在命名可能适用于任一层次水平的量表或成分时避免使用“因素”一词, 5 个主要

范畴是神经质(Neuroticism, N)、外倾性(Extraversion, E)、经验的开放性(Openness to Experience, O)、一致性(Agreeableness, A)、自觉性(Conscientiousness, C)。它们各自的方面列在表 13-1 中。

表 13-1 修订的 NEO 人格调查表(NEO PI-R)的范畴和界面

| 神经质       | 外倾性      | 经验的开放性 |
|-----------|----------|--------|
| 焦虑(N1)    | 热情(E1)   | 幻想(O1) |
| 愤怒的敌意(N2) | 合群性(E2)  | 美学(O2) |
| 抑郁(N3)    | 武断性(E3)  | 情感(O3) |
| 自我意识(N4)  | 能动(E4)   | 行动(O4) |
| 冲动(N5)    | 寻找刺激(E5) | 观念(O5) |
| 脆弱(N6)    | 积极情绪(E6) | 价值(O6) |
| 一致性       |          | 自觉性    |
| 信任(A1)    |          | 能力(C1) |
| 坦率(A2)    |          | 规则(C2) |
| 利他主义(A3)  |          | 尽职(C3) |
| 顺从(A4)    |          | 成就(C4) |
| 谦虚(A5)    |          | 自律(C5) |
| 敏感(A6)    |          | 谨慎(C6) |

(摘自 Costa & McCrae, 1992b, p.2, 心理评定资料公司, 复印得到允许)

NEO PI-R 量表是在 15 年的研究后发展起来的, 它开始是对正常成人样本中老年人的纵向研究, 后来扩展到临床、就业和大学样本。虽然它是作为“正常人格特质”的测量而设计的, 但考斯塔和马克雷打算把它应用于临床和其他应用情境及研究。在方法学中, 引入 NEOPI-R 的新方法是自我报告形式(S型)和观察者报告形式的两个版本(R型—男性和 R型—女性), 后者包含了以第三人称描述的与 S型相同的 240 个项目, R型有可能获得与自我评定相同的领域中来自同伴、配偶和其他人的独

立评定。就 NEO PI - R 来说,这是特别重要的,因为调查表假定参加测试者是诚实和合作的,而且没有包含检查反应真实性的量表,成年男女的常模可用两种形式,S 型有大学年龄的男女常模。

马克雷和考斯塔提出的五因素(或‘大五’)模型,虽然没有普遍地把它作为探求人格特质的一个有用的框架,但这种看法已被广泛接受。当然,因素分析取向的研究者之间在他们提出的中间水平的因素数目和定义两方面仍存在差异(L. R. Goldberg, 1993; Zuckerman, Kuhlman, Joireman, Teta & Kraft, 1993)。不过,在它的各种版本中,这一模型已刮起了一股研究活动的旋风,目的在于对因素的交叉鉴别和各种观点的整合,如正常的和病理的人格的重要方面的问题(参见 Hofstee, de Raad & Goldberg, 1992)。在测验发展和现有量表的精制方面有关的工作也在快速进行着(Cost & McCrae, 1994, 1995; Costa & Widiger, 1994; Harkness, McNulty, Ben-Porath, 1995; Hogan & Hogan, 1992)。

在评价这一活动的结果时,应切记因素分析仅仅提供了一种技术,把项目分成相对同质的组和独立的类,这种分组促进了根据经验效标的效度研究,使得对特定效标的预测分数进行更有效的组合,并有助于构念的界定,同质性和因素纯度是测验编制中理想的目标,但它们不能代替经验的效度分析和良好的理论基础。

## 测验发展中的人格理论

人格理论通常起源于临床情境,随后对它们进行的实验验证的数量在各理论系统之间有很大的差异。不管这种客观证实

的程度如何, 一些人格测验已在一个或另一个人格理论框架中构成。临床上形成的假设在投射技术发展中显得尤为突出, 这将在十五章中讨论, 而测验建构的这种方法在自陈调查表中较少出现, 只有几个突出的例子。

米隆临床多轴调查表(Millon Clinical Multiaxial Inventory, MCMI) 虽然最初发表于 1977 年的米隆临床多轴调查表 III (Millon Clinical Multiaxial Inventory - III, MCMI - III, Millon, Millon, Davis, 1994) 基于相同的目的, 在几个方面追随 MMPI 传统, 但它引进了重要的方法学变革。事实上, 它的发展是谨慎进行的, 既考虑了对 MMPI 的批评, 又利用了心理病理学诊断和测验编制方面的进展。

MCMI - III 是基于米隆关于人格功能作用和精神病理学的生物心理学观点的(Millon, 1969, 1981, 1990; Millon 等, 1996)。他的理论包括了一个人格类型的矩阵。这一矩阵来自两个维度, 即强化源(如孤立的、不一致的、依赖的、独立的及矛盾的)和应付行为范型(如主动或被动)的结合。米隆的人格类型理论是美国精神病协会编辑的《心理障碍诊断和统计手册(DSM - III, 1980)》的轴 II 人格障碍分类的最初形式的概念基础之一, 现在已发展到第 4 版(DSM - IV, 1994)。反过来, MCMI - III 临床量表与 DSM - IV 的分类系统一致, 虽然不完全相同。将调查表与 DSMs 框架的发展变化尽可能紧密联结的努力, 实际上促进了对 MCMI 的不断修订。

MCMI - III 包含了 175 个简短的自我描述的句子, 要求作出“是”或“否”的回答。分数剖面图包括 24 个临床量表, 每个量表基于 12 ~ 24 个重叠项目, 这些项目常常在 3 个不同量表中出现, 虽然权重不同。满足“中心”量表的所有效度标准的项目给予权重 2, 而辅助项目得到权重 1。正如表 13 - 2 中看到的, 临床

量表分为 4 个主要的类型,称之为临床人格型式、严重的人格病理学、临床症状和严重的症状。这些类型中的前两个包含用来评定持久的 DSM 轴 II 不同程度人格障碍类型的量表,其他两个包含了 DSM 的轴 I 症状中的一些内容,还有三个校正指标和一个用来检测不正规的回答形式和测试者偏见的效度检查表。最初 MCMI 只能由计算机记分,最近除了记分服务系统和用来产生剖面图和解释报告的软件外,也可以手工记分,当然十分费力,因为需要进行分数转换和几种调整。

表 13-2 米隆临床多轴调查表 - III

|            |         |      |
|------------|---------|------|
| 临床人格型式     | 临床症状    |      |
| 精神分裂症样的    | 焦虑      |      |
| 回避的        | 躯体形式    |      |
| 抑郁的        | 双极:躁狂   |      |
| 依赖的        | 情绪不良    |      |
| 做作的        | 酒精依赖    |      |
| 自恋的        | 毒品依赖    |      |
| 反社会的       | 创伤后应激障碍 |      |
| 攻击的(施虐的)   |         |      |
| 冲动的        | 严重的症状   |      |
| 被动攻击的(受虐的) | 思维障碍    |      |
| 自我挫败       | 重度抑郁    |      |
|            | 妄想障碍    |      |
| 严重的人格病理学   | 调节指标    | 效度指标 |
| 精神分裂类型的    | 泄露      |      |
| 临界         | 满意      |      |
| 妄想狂        | 降低      |      |

(摘自 Millon 等,1994,p.2,由 Dicandeien, Inc.1994 年出版,引用得到允许)

MCMI 中引进的最有意义的一个革新叫作基本比率 (BR) 的标准分数的使用。它替代了常模化, 参照被测特征的一般比率。MCMI 量表的 BR 界定分的设置反映了量表评定的特定条件下的精神病人人群中计算出来的基本比率数据, 这样可加强鉴别性诊断。因为不同的临床人群和情境的一般比率各不相同, 一些 MCMI - II 量表的 BR 分可以根据情境、慢性程度和焦虑、抑郁的得分以及一定的反应型式进行调整。

MCMI 的项目发展采用多界面方法, 这是近来人格调查表的编制和效度分析方面的实践性特点。在这一点上, MCMI 在方法学上与在本章的几个部分里已提到的都有一些共同之处。这一程序包括三个主要步骤: (1) 理论的一实体的 (如写出和选择适合临床上恰当构念的项目); (2) 内部的一结构的 (如项目一量表的相关和认可频率); (3) 外部的一效标的 (如参照组与诊断组的区别和新样本的交叉效度分析)。

用于 MCMI - II 及其前身的项目分析的参照组由不加区分的精神病人组成, 而不是正常样本。主要由于常模和转换分数的产生仅仅使用临床样本, 因此, 作者明确地指出“MCMI - II 不是一个用于正常人群和其他目的的一般人格评定工具, 它只是用于诊断筛选或临床评定”(Millon 等, 1994, p.5), 而且量表的界定分和剖面图解释适用于那些显示出存在中等严重程度的心理病理的人, 而不适用于那些接近正常或问题非常严重的人。一些用正常被试的 MCMI 研究已发现他们的得分高于精神病人, 但处于次临床水平, 而另一些研究显示一些正常人某些量表得分处于病理水平 (如有意做作和自恋)。这些结果证实了用此量表对正常人施测是不可取的, 并表明在中等水平正视图上, 这些量表甚至可以测量健康人格特征 (Wetzler, 1990)。

MCMI 的主要目的之一就是对鉴别诊断有所帮助, 但由于



同一个体能同时共存几种精神病状况而使这项工作变得复杂化。由于 MCMI 的初版受到称赞,因而经历了相对频繁修改和精心完善的过程,但这也使得评定它在多大程度上符合其目的变得困难。然而,积累了大量有关早期版本的研究经验,大量有助于使用者的相关出版物正在出现,研究结果证实了各量表的项目重叠能降低它们的辨别力,尤其是当测试者处于焦虑和抑郁状态时;有关临床症状量表的诊断效果还需要更多的数据。不过,MCMI 仍是诊断人格障碍和评定它们的治疗效果的一个潜在的有价值的工具(Choca, Shanley & Van Denburg, 1992; Craig, 1993; Goncalves, Woodward & Millon, 1994; Retzlaff, 1995; C. R. Reynolds, 1992a)。<sup>[10]</sup>

最近,米隆又编制了两个扩展其人格和心理病学评定方法的新工具。一个是米隆青少年临床调查表(MACI, Millon, Millon & Davis, 1993),试图作为临床环境中评定 13~19 岁青少年的可供选择的工具。它从米隆青少年人格调查表发展而来(MAPI, Millon, Green & Meagher, 1982),后者是一个老的工具,最初用于临床评定、职业咨询和学业指导,它用量表来评定青少年基本的人格类型、关心的事物和行为倾向。<sup>[11]</sup>另一个是米隆人格类型指标(MIPS, Millon, 1994),用来测量在各种咨询条件下工作、家庭、社会问题等方面寻求帮助的正常成年人的性格。MIPS 以成人和大学生为样本来标准化,它把米隆的人格理论中的原理与弗洛伊德和荣格的理论结合在一起。

埃德华个人偏好表(Edwards Personal Preference Schedule, EPPS)在促进测验发展的人格理论中,最富有成果之一的是由默里(Murray)和他的助手在哈佛心理诊所(1938)提出的明显需要系统(manifest need system)。旨在评定这种需要的强度的第一批调查表之一是埃德华个人偏好表(EPPS, Edwards, 1959)。

从默里所列的需要中取出 15 种开始,埃德华编写了其内容适合这些需要的一套项目,包含成就需要(尽力完成困难的事情)、遵从需要(按别人希望的做)、表现需要(成为注意的中心)、支配需要(影响别人,被认为是领导)、养育需要(帮助有麻烦的人)。

调查表由 210 对陈述句子组成,其中 15 个量表中每个量表的项目与其他 14 个量表的项目配对。<sup>[12]</sup>在每一对中,让被试选一个更能说明自己特点的陈述。记住这点是很重要的,由于这是被迫选择,所以 EPPS 产生的是自比分数,即每一个需要的强度不是以绝对的方式表达,而是相对于该个体的其他需要的强度。在自比记分中参照框架是个体而不是常模样本。因为所有分量表分数的总和对所有被试是固定的,所以,如果某个人一个分量表的得分移上了一点,则另一分量表的得分必定移下一点。在这种情况下,EPPS 中两个得分相同的个体可能在他们需要的绝对强度上有显著差异。虽然 EPPS 提供把得分转化成百分数的常模,但因为分数的自比性质,转换的合理性可能会受到质疑。参照的自比框架对个体内部的比较是最合适的,如兴趣和其他爱好的评定中需要的比较,但常模参照数据对个体间的比较也是必要的,例如用于评定异常的比较 (Fedorak & Coles, 1979) 然而,参照的两种框架的结合与前后一致地采用自比或常模的方法相比,会使对分数的解释含混不清,缺少意义。

虽然在 EPPS 手册中缺少效度资料,但大量独立的效度分析研究已经发表。然而,这些研究的结果很难解释,原因是大部分研究没有考虑数据的自比性质。用自比分数各个量表的相关的平均值是负的,所有量表与外部变量的相关的平均值趋于零 (Hicks, 1970)。鉴于这些人为的限制,自比分数不能用通常的相关方法进行恰当分析,因此公布的效度分析研究产生了相互抵触和无说服力的结果并不奇怪。尽管 EPPS 有其简易性和值得

注意的一些特征，但需要在修订版中排除与项目形式和分数解释有关的技术弱点。

**人格研究表和其他杰克森调查表** 人格研究表 (Personality Research Form, PRF) 反映了测验建构方面的许多技术进步，包括可以使用高速计算机之前根本不可能的一些项目选择技术。PRF 是杰克森 (Douglas N. Jackson) 的人格测验方法的一个例子，它开始于对所测构想作明确的、详细的描述。这些描述形成了项目文字和特质定义的基础。这些特质在效度分析研究中由评判者进行评定 (Jackson, 1970, 1989b)。

PRF 有 5 种不同的选择，包括了两套平行形式 (A、B 和 AA、BB)，分别有 300 个和 440 个项目。较长的表有 22 个量表分，每个量表有 20 个项目，包括两个效度分数：不寻常 (Infrequency) 和符合期望 (Desirability)；较短的表只有 15 个 20 项的量表。后来采用更复杂的项目分析技术发展了增加版 (E 表)，由长表中最好的 352 个项目组成，共有 22 个量表分，每个量表只有 16 个项目。现在最常用的 E 表，较其他表的词汇水平要低。不寻常分数是根据被试很少可能选择的反应数，作为粗心、没有理解指导语和其他无目的的反应的指标，如“每晚我试着至少睡着一会儿”；我所有的衣服和鞋子都是自己做的”。虽然符合期望的偏向预先通过项目编写和选择程序大大地减少了，但还是提供了符合期望的量表。使用手册正确地显示量表上特别的高分或低分，可能不仅表明被试不一般的态度（如故意制造好印象与装病），而且可以表明他们自身的重要的人格特征（如高自尊或高度的传统社会化与低自尊）。

和一些别的人格测量工具一样，PRF 以默里的人格理论为出发点。利用过去 30 多年中积累的广泛的研究和理论文献，杰克森表述了行为定向的和 20 个相互排斥的人格构念或特质定

义。在所有这些定义中,12个与 EPPS 中的名称一样,手册对 20 个双极人格量表中的每一个都提供了高分者的说明和一组界定特质的形容词。两个量表定义的例子如表 13-3 所示。

通过仔细控制的程序,每个量表还有 100 多个项目的题库,然后,根据与量表总分二列相关高、与其他特质量表得分和符合期望量表得分相关低,为每一个量表选择 20 个项目,排除了认可比率极端的项目。通过一种特别编制的计算机程序,项目按照与它们自身的量表的二列相关和认可频率被安排到平行的模式中。用于 E 表编制中的程序包括计算出每个项目的效能指标。根据从各种统计的项目参数中获得的权重,这个指标使每个项目根据它的效能排列在量表中。

表 13-3 取自人格研究表的量表定义例子

| 量 表  | 高分者特点                                                                   | 界定特质的形容词                                                                      |
|------|-------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| 认知结构 | 不喜欢信息的模糊性或不确定性;想要完整回答所有的问题;希望基于确定的知识而不是凭猜测或概率进行决策。                      | 精确,严格,确定,寻求确实性,细致的,完美主义的,清晰的,显见的,准确的,严密的,朴实的,避免模糊性的,明确的,刻板的,需要结构的。            |
| 感觉能力 | 注意气味、声音、光线、味道和事物感觉的方式;记住这些感觉,并相信它们是生活的重要部分;对多种形式的经验敏感;保持一种基本乐观的或审美的生活观。 | 审美的,喜欢身体感觉的,观察力敏锐的,世俗的,觉醒的,注意环境的,富于感情的,敏感的,感官的,经验开放的,有理解力的,注意的,有辨别能力的,对印象敏感的。 |

(取自杰克森,1989b,pp.6~7. Douglas N. Jackson 出版,引用得到允许)

PRF 的构想效度很大程度依赖于每个量表的项目发展和选择所遵循的程序。随后的因素分析结果使项目分成 20 个量表的做法得到了进一步巩固。与加州心理调查表、吉尔福德—齐默尔曼气质调查表及 NEO-PI 等其他相似调查表的相关为特质的确定提供了又一支持。虽然用来测量默里需要的各种工具——如 PRF、EPPS、TAT(第十五章)和 ACL(第十六章)之间的相关研究产生了不一致的结果,但与别的工具相比,似乎更支持 PRF 的构想效度;只要仔细关心一下 PRF 的发展过程中特质的表述就不会对此感到惊讶了 (Cocta & McCrae, 1988; D. W. Fiske, 1973; Rezmovic & Rezmovic, 1980)。与来自大量的同伴评定和自我评定的数据相比,PRF 的经验效度资料也产生了有希望的结果。值得注意的是,PRF 已证明可应用于几种文化,甚至是非西方文化 (Jackson, Guthrie, Astilla & Elwood, 1983; Paunonen, Jackson, Trzebinski & Forsterling, 1992)。总之,正如麦克莱纳 (MacLennan) 汇编的详细注释的文献目录证实的, PRF 是一个非常优秀的研究工具,但仍需有更多的信息来决定它在实际情境中的有效性。<sup>[13]</sup>

杰克森人格调查表一修订版(JPI-R)是继 PRF 之后通过相似但更精制的量表编制程序发展起来的,其实践定向性更强 (Jackson, 1976, 1994a)。特质量表是从人格心理学和社会心理学的文献中选出来的,部分是因为它们与正常被试在各种情境中的行为预测有关。15 个量表所涵盖的特质是焦虑、合作、责任心、社会敏锐力、忍耐。效度资料的收集不仅采用多特质—多方法矩阵,以求与同伴评定和自我评定的相关,而且通过对特定群体的研究,这些群体在真实生活情境中的相关行为资料是可以获得的。在最近的调查表修订版中,大学生的常模被更新了,还形成了蓝领和白领阶层的常模,支持它在学业指导和就业情境

中的应用。此外,在量表上还做了许多技术修改和微小变动。现在量表分为五个高层次群,这样与前面讨论过的五因素模型(FFM)的分类基本上一致。特别要指出的是,JPI-R的外倾的、可靠的、分析的丛分别与 FFM 的外倾性、诚实、经验的开放性维度很相似。另外,JPI-R 的情绪丛似乎是 FFM 中的神经质和一致性的结合。然而 JPI-R 的机会主义丛,取自它的社会敏锐力和风险承受量表,与 FFM 没有直接的联系,因此被看成是一个独特的因素。

至于心理病理学的评定,杰克森应用与 PRF 和 JPI 相同的严格标准以编制基本人格调查表(BPI, Jackson, 1989a)。BPI 的发展大约历时 15 年,它使用在内容、心理测量纯度和可应用范围方面更有价值的量表,其目的在于恢复与 MMPI 相联系的诊断效能。尽管 BPI 需要更有代表性的常模,尤其是成人的,但它已显示出有希望用于青少年犯罪领域的临床应用(Holden & Jackson, 1992 评价见 Urbina, 1995)。

在这个有关自陈调查表的概述中,读者可能已注意到各种不同方法相互渗透的倾向正在加强,杰克森调查表和过去 12 年中发展的其他新的调查表尤其如此。除了项目的经验准则确定,其他所有的策略都用上了。此外,甚至是明尼苏达调查表,作为经验准则确定的突出例子,现在所含量表也是以内容和因素分析为基础的。一些证据表明由本章提到过的以四种主要方法中的任何一种方法编制的人格调查量表都是有效的,至少在聚合和预测效度方面(Bunsch, 1986)。然而,以内容和理论为基础的方法与以经验准则确定的方法相比,更容易编制,使用效力更高,更可能显示出区分效度,反之,经验准则确定方法与其余的方法相容性最小。而且,充分一致的是:(1)调查表的发展工作应该从所要测量的特质或构念的明确定义开始;(2)由坎贝尔

(Campbell) 和费斯克在他们的多特质—多方法矩阵中(见第十五章)概括的组织结构,为人格调查表的构想效度的检验提供了一个最佳的策略(Angletiner, John & Lohr, 1986; Hogan & Nicholson, 1988; Ozer & Reise, 1994)。

## 参加测试的态度和反应偏向

**伪装和符合社会期望** 自陈调查表特别具有故意虚假作答的可能性。尽管指导语要求不要这样,但调查表的大部分项目还是有一个答案比别的答案被认为更令社会满意和更为社会接受。在这种测验里,回答者可能受‘装好’的动机驱使,选择造成好印象的答案,如申请工作或寻求教育机构的入学资格时就是如此。在其他情况下,回答者可能受‘装坏’的动机驱使,这样使他们表现得更有心理困扰,这种情况可能会发生在对犯人作测试的时候。

在人格调查表中,回答者能成功地进行掩饰的证据很多(参见 Jacobs & Barron, 1968; Radcliffe, 1966; Stricker, 1969; J. S. Wiggins, 1966)。一个普通的班级实验,要求不同小组担任规定的各种角色,如要求一部分人作为快乐的、适应很好的大学生来回答每个问题,而要求另一部分人按适应不良的方式来回答问题,最后一部分人按自己的行为真实地回答问题;或者同一个人做两次测试,第一次的指导语是以规定的方式作虚假回答,第二次是在一般情况下进行。这些研究的结果清楚显示,在测试中,容易故意造成令人满意的印象。有趣的是,甚至可以成功地进行一个特殊职业目标的虚假回答(Wesman, 1952)。现有的研究证实,特别是在自陈调查表中,在真实和控制的实验条件下,项目的

“表面效度”<sup>[14]</sup>会增强伪装的可能性；回答者越能从项目的文字中识别出要测量的特质，就越常作出令人满意的回答(Bornstein, 等, 1994)。

在自陈调查表中选择令社会满意反应的倾向并不一定说明回答者是故意欺骗。第一个对社会满意变量进行研究的人是 A. L. 埃德华(1957), 他把它基本上看成是一种‘装饰门面’的外表效应或倾向, 对此, 回答者基本上是无意识的。这种倾向可能表明对一个人的性格、自我欺骗或不愿意大胆面对一个人等局限性缺少洞察力。其他研究者(Crowne & Marlowe, 1964; N. Frederiksen, 1965)已经提出证据表明, 社会满意反应定势的强度, 与个体自我保护、避免批评、社会遵从和社会赞同等较一般的需要有关。另一方面, 在自我描述中选择不利项目的个体可能是受关心、同情、遇到个人问题时接受帮助的需要的驱使。例如, 寻求心理治疗的人们, 在调查表中很可能表现得比真实的他们更加适应不良。

而且, 不能假定基础的研究不受反应定势的影响。例如在关于态度变化的研究方面, 研究者已经表明结果可能受到这样一些条件的影响, 如参与者对实验者期望什么的觉察、保护自己形象的欲望, 以及使实验者愉快或失败的希望(Silverman & Shulman, 1970)。这些反应定势中毋庸置疑的差异可以部分地解释为什么重复实验时不能得到相同的结果。

一些研究者(Paulhus, 1984, 1986; Paulhus & Reid, 1991)强调印象处理<sup>[15]</sup>和自我欺骗的概念之间的差异, 作为对按社会满意或不按社会满意方向回答的解释。印象处理是指有意识的掩饰, 为的是产生回答者期望的特定效果, 它被看作是对自我报告资料的污染。它本身就是必须评定, 并且必须尽可能减少或遏制的东西。另一方面, 自我欺骗包括被试相信确实是正确的正



面的反应偏向，它更为复杂。自我欺骗与别的一些关于自我的概念相联系，如自我形象和自尊，也与心理分析的防御机制的概念相联系。这样就其本身来说，它作为良好适应指标和其他独立效标的预测，是非常值得研究的变量。例如，研究者已经注意到，一些自陈量表的所得结果表明防御否认者具有“幻觉的心理健康”，他们力求维持自己是适应的信念(参见 Shedler, Mayman & Manis, 1993)。因此，自我欺骗和适应之间的关系似乎既不简单又不直接。项目的某些语言特征与产生偏向性反应的反应者变量之间的相互作用使情况变得更加复杂(Helfrich, 1986)。

已有几种方法试图解决人格调查表中伪装和有关的反应定势的问题。相对“微妙”或中性的项目建构，可能会减少一些调查表中这些因素的作用，然而，正如杰克森(1971)所指出的，这些项目通常对所考察的维度来说具有低效度。理想上而且事实上在某些情况下，测验指导语和友好关系的建立应该鼓励被试如实回答，如果他们确信这样做对他们是有好处的。然而，这种方法在某些情境中是无效的，而且对个体没有意识到的令社会满意的反应定势可能也没有多少效果。

评定令社会满意反应和印象处理其他形式的另一个方法是构造特殊量表——此量表或者能包含在调查表里或者能在成套测验中为了该目的自行施测。最早的这种量表之一是埃德华(1957)的社会满意(SD)量表。该量表是根据各评判者关于项目的最高或最低社会满意值的一致性，通过项目选择发展而来的。其他如 CPI 的好印象量表和维新(J. S. Wiggings, 1959)的 SD 量表，以被试在“装好”条件下和通常情况下的认可频率的差异为基础选择项目。第三种方法的例子是 MMPI 的说谎量表，它包含这样表述的项目，只有显示出非现实偏向的人才会按令社会满意的方向回答。其他专门用于检测装病、粗心和随机回答的

测量,如 MMPI 的 F 量表已经设计出来,<sup>[16]</sup>——还有另一种不用来检测而是用来预防掩饰的方法——是迫选项目的使用。

**迫选技术** 实质上,迫选技术要求答题者在两个同样可接受但效度不一样的描述词或短语中进行选择,这种配对短语可以是两个都令人满意或都不令人满意。迫选项目也可以是三项、四项或五项,在这种情况下,答题者必须确定哪一个是最符合或最接近自己特点的。还有另一种变化的情况是要求在同一特征的两个相反反应之间选一个,作为单一量表记分。虽然这种项目形式很少用于人格调查表,但它具有产生常模而不是自比分数的优点,因此在不同量表之间的相互关系方面没有人为的限制。一个例子是由梅耶斯——布里格斯(Myers-Briggs)提供类型指标,这将在第十六章中讨论。

采用迫选技术来控制社会满意这一因素要求有关于每个反应选择的两种类型的信息,即社会满意或“偏好指标”和它的效度或“辨别力指标”。后者可以根据调查表所要预测的特定效标确定,如学业成就和特定种类的工作上的成功;或者可以根据项目的因素负荷或项目与不同特质的理论关联性来确定。社会满意可以通过有代表性的一组人对项目中这个变量的评定,或通过确定该项目在自我描述中认可的频率而查明。研究显示出选择频率和社会满意的相关性很高(Edwards, 1957)。换句话说,一个群体自我描述的平均值与符合期望的人格描述的平均值十分接近。

虽然迫选项目会减少社会满意的影响,但不能认为会完全排除这些影响。EPPS 项目以自由选择形式呈现,所得分数与相同测试在迫选形式下所得分数的相关性相当高(Lanyon, 1966)。而且,项目的社会满意判断不是对一切目的都固定不变,而是在不同情况下有所不同。因此,其项目在一般社会满意方面都有

相同的迫选测验,就职申请者、入专业学校学习的候选人以及其他有具体定向的群体还可以进行伪装。从另一角度来看,当项目是以一般社会满意的组群平均判断为基础进行配对时,对个体来说仍可发现它们是不相当的(N. Wiggins,1966)。

总之,迫选技术在控制伪装或令社会满意的反应定势方面没有证明具有预期的效果。同时,迫选项目形式,特别是产生自比分数时,引起了别的技术困难,并排除了关于个体特征的绝对强度的信息,这些特征在某些测试情境中可能最为重要。

**反应定势和反应风格** 根据社会满意选择反应的趋向,只是自陈调查表回答中已经确定的几种反应定势中的一种(Lanyon & Goodstein,1982,pp. 158 ~ 169),虽然关于反应定势在人格调查表中的作用的文献主要始于 50 年代,但反应定势在能力和人格测验中的影响,早有研究者观察到了(Block,1965,第二章)。早期引起注意的反应定势之一是默认或回答‘真’或‘是’的倾向。默认作为一个连续变量,在量表的一端是一致的“赞同者”;另一端是一致的‘否认者’(Couch & Keniston,1960)。就人格调查表的编制来说,这种反应定势意味着在任何特质量表中“是”和‘真’回答是正面记分的项目,应与‘否’和‘假’回答是正面记分的项目的数量相等。这种平衡可以通过项目的适当选择或文字加工来实现。PRF 是这样做的,现在大多数新的调查表也正在这样做。<sup>[17]</sup>

另一个反应定势是偏离,或是给出不寻常的或不一般的回答,伯格(Berg,1967)提出这一假设,并证明了它在一个特别编制的要求表示对几何图形的偏好的测验中的作用。量表是由几乎全部被试可能都以一个方向回答的项目组成,如杰克森的 PRF 的不寻常量表,用来鉴别这种偏离反应形式。然而,杰克森自己也指出这些量表与外在准则之间缺少概念性相关,因此它

们就产生了问题，特别是在像应聘这样的情境中。在这些情境中，问卷项目的相关性是非常重要的。因为这一点，JPI的不寻常量表在修订时被去掉了(Jackson, 1994a)。在评定等级量表中采用极端选择(如七点量表上的1和7)的倾向也已经被确定为一种可能的反应偏向(Paulhus, 1991)。

关于反应定势，如令社会满意、默认和偏离的研究，已经经历了几个阶段。开始时，反应定势被认为是与测量得分无关或误差变异的来源，应从测验分数中加以排除。后来，逐渐认识到就它们自身来说是值得测量的广泛和持久的人格特征的指标(Jackson & Messick, 1958, 1962; J. S. Wiggins, 1962)。在这一阶段，它们被普遍描述为反应风格，经验数据的精细构造是围绕着它们而建立起来的。最后，这些数据受到了多方面的挑战(Block, 1965; Heilbrun, 1964; Rorer, 1965)。如布洛克(Block)提出强有力的证据支持对两个主要因素的内容取向解释，这两个因素一般可用来说明MMPI量表的大部分共同变异，反应定势和反应风格的拥护者把它们解释为社会满意和默认。

人格评定中有关反应定势和内容一风格的争论尚未完全解决(Edwards, 1990; Hogan & Nicholson, 1988; Jackson & Paunonen, 1980)。<sup>[18]</sup>大多数测验编制者和研究者似乎都同意人格调查表分数可能是自我欺骗、印象处理和理想的自我描述的混合体，而每一个组成部分的权重随个体和场合的变化而变化。然而有些人认为，试图通过特殊的量表和项目以提高自我报告资料的可靠性的做法可能会起反作用，因为它们可能会降低量表的效度，特别是对与病理样本不同的正常样本来说。这些作者赞同在诱发病人的合作和解释分数方面运用临床技能，以及在有理由怀疑分数被严重歪曲时要参考来自了解情况的信息提供者的评定。

大多数其他人员,特别是心理病理评定工作者,继续使用所谓的‘效度’量表,认为它们可能也反映了人格类型和特征。事实上,一些技术较先进的最新评定心理病理的工具,如杰克森的BPI和默雷(Morey)的PAI,利用平衡的项目记分方法和特殊量表来检测使调查无效的反应定势;现在还有各种各样新的量表,如MMPI-2和MMPI-A的VRIN和TRIN量表,它们利用特别挑选出的在内容上相似或相反的成对项目来检测不一致的或自相矛盾的回答。VRIN和TRIN量表与初版MMPI的格林尼(Greene,1978)的粗心量表相似,由于其制定方法的关系,它们不可能因为人格特质变异而变得混乱(Ozer & Reise,1994)。

不管怎样,有关反应定势和风格的争论刺激了广泛的研究,并已经产生了数百种论著。和许多科学的争论一样,它的最后效果已深化了我们对方法论问题的理解,因此促进了人格调查表的编制,以及它们在研究和应用情境中的使用。

## 特质、状态、个人和情境

**个人和情境的相互作用** 关于心理特质的一般性与行为的情境特殊性的长期争论在60年代末和70年代达到高峰。60年代关注的焦点在于狭义的“重要行为”和远离广义的特质。在能力领域,这一关注体现在个别化指导方案与领域参照测验(第三章)和学习无能的诊断和治疗方面(第十七章)在人格领域,对测验中行为特殊性最强大的推动力来自作为行为矫正和行为治疗的基础的社会学习和社会认知理论(Bandura,1969,1986;Golfried & Kent,1972;Mischel,1968,1969,1973)。对此的批评集中于早期把特质——包括认知的和非认知的——看作是固定、不

变、潜在的因果实体的观点，这类批评在几位心理学家的早期研究和著作中已预见到（见第十一章）。虽然极少数心理学家确实为这种极端的特质观点争辩过，但在情境特殊性争论的全盛时期，很难找到自认为是“特质理论家”的人（Jackson & Paunonen, 1980）。

情境特殊性更多是人格特质所具有的特征，而不是能力所具有的特征。例如，一个人在办公室里表现十分友善、开朗，但在社会集会时表现害羞和沉默，或一个考试作弊的学生，在钱的问题上却可能表现诚实。米歇尔和皮特森（Mischel & D. Peterson, 1968）收集的大量研究证据表明，个体在许多非智力方面的确表现出相当的情境特殊性，诸如攻击行为、社会遵从、依赖性、坚定性、诚实性以及权威的态度。认知的功能比非认知的功能具有更高的跨情境一致性，部分解释可能是个体在智慧领域比在人格领域具有更多的标准化反应经历（Anastasi, 1958, 第十一章；1970, 1983a）。例如，正规学校课程在言语和数学方面发展了广泛可用的认知技能，另一方面，人格的发展则处于非常不一致的条件下。何况在人格领域，相同的反应在一种情境中可以导致正强化而在另一情境中得到负强化的社会后果，这样个体就会在不同的背景中以完全不同的方式作出反应。这种不同个体和不同情境与不同文化中经历的差异也导致人格测验项目比认知测验项目有更大的模棱两可性。因此，对人格调查表里某一问题（它本身可以作为一种“情境”）的相同回答，对不同个体可能具有不同的意义。

应该注意到，特质—情境问题也与熟悉的遗传环境问题有关（D. C. Rowe, 1987）。遗传的影响可能通过相对持久的个体特质得到表现，不过它可能包含一种对情境要求适应性的特质。环境的影响可能对情境变异（或特殊性）也对特质的稳定性起作

用,因为个体环境本身可能在各个时间和情境中表现出一致性。适当的实验设计通过纵向和跨情境的重复测量,应该促进我们对遗传—环境和特质—情境两个问题的理解。

关于个体与情境相互作用的理论探讨和研究无疑丰富了我们决定个体行为的许多条件的认识,对形成成熟的研究设计也起作用。同时,相反观点的追随者逐渐形成共识,认为行为的特质和情境解释可以共存,事实上,行为是由特质和情境变量的相互作用决定的。从 70 年代后期到 80 年代后期出版的关于这一问题的一些论著中,<sup>[19]</sup>这种和谐关系特别明显。其中有几点值得注意,在测量为可靠时,即总结熟悉被评定者的多个观察者所做的重复观察,行为表现出相当明显的时间稳定性。对个案和情境的随机样本的研究发现,个体差异比情境差异对总的行为变异作用更大,个体和情境之间的相互作用与个体差异的作用相似或略大。为了鉴别广义的人格特质,我们必须利用可观察的维度和与之相关的行为,测量多种情境中的个体,并聚集各种结果(Epstein,1980;Kenrick & Funder,1988)。尽管争论的双方对这些问题的许多方面达到了一定的共识,并且相互有所了解,但关于个体—环境仍有许多问题有待解决并不断产生新的问题。

**个体** 每个个体在不同情境中的行为特殊性的程度各不相同。个体不同程度地改变其行为以满足每一情境的要求。在这方面,中等的 inconsistency 表明有效的和适应的灵活性,而过分一致表示适应不良的刻板性。而且,各人在其中表现出行为一致性的情境各不相同。这种情境之间的一致性受个体对情境的知觉和分类的影响。情境的这种分类反过来又依赖于个体的目标、动机和情感以及她或他对相似情境的以往经验。<sup>[20]</sup>

一致性方面的个体差异对心理学家具有重要意义,其原因

是很明显的，这种差异若能精确地评定，它们能作为行为预测的中间变量。另外，从心理测量观点出发，一致性的个体内和个体间差异在调节所有心理测量的效度和信度方面被认为是决定性的。因此，研究者正在从多个方面努力寻求评定这些差异的方法，是不足为奇的。一种方法是根据等级评定，人们在不同的特质维度上评定自己，分数来自每个人等级评定之间的变异。量表项目之间低的变异代表评定特质 (traitedness) 的质量，事实上，它与较高的效度系数有关 (Ameloang & Borkenau, 1986; Baumeister & Tice, 1988)；另一种方法以兰宁 (Lanning, 1991)“可等级性” (scalability) 的观点为代表，它是指在自陈评定时人们保持不同情境中行为项目的标准次序的程度。

根据人格问卷项目反应的过程，把个体—情境相互作用的概念和社会满意联系起来，这一令人感兴趣的提议是杰克森的反应的‘阈限理论’所提出的(参见 Helmes & Jackson, 1989; Jackson, 1986b)。这一模型是以调查表项目是真实世界中行为的缩影这一前提为基础的，用项目反应理论方法把人格项目内容加以度量(见第七章)。杰克森提出，正如能用许多个体对单一项目的反应得出项目特征曲线一样，单个被试对许多项目的反应同样也能形成被试特征曲线。这一曲线以被试符合期望维度的突出程度、她或他根据社会满意对项目作出积极反应的阈限和愿望以及项目本身是否符合期望为基础，能预测一个人对项目表示赞同的概率。

**情境** 情境在它对行为的限制方面也不同。我们可完全肯定地预测，读者在图书室能保持安静，驾驶员看见红灯会停车。相似地——无论人们的特质结构如何——人们去海滨浴场大多是为了游泳，去图书馆是为了看书。当然，某些人可能会花时间在海滨看书，某些人也可能会花许多时间在图书馆里做游泳的



白日梦。较好地理解情境对行为的限制的一个方法，是研究各种行为情境的特征。罗格·巴克(Roger Barker)关于生态心理学的最近研究提供了一种有希望的工具，用于行为情境的分类和对情境各方面的描述(Schoggen, 1989)。

跨文化差异可以看作是一个特殊的、更普遍的情境变化的例子。正因为如此，它们提供了研究个人行为的一致性和非一致性的独特机会，所以，第九章和第十二章中讨论的能力测验方面的跨文化测试的方法也可用于研究别的心理特质。

在一种文化中发展的人格测验应用于其他文化中时，显示出较大的差异。这种文化和亚文化差异的任何解释，需要有每一组群的主要环境条件的特殊知识，这点在美国国内不同文化群体的心理病理评定中越来越被认识到 (Malgady, Rogler & Costantino, 1987; Paniagua, 1994)。例如，像 MMPI 这种测试中的组间差异只是反映对个别项目和指导语理解的差异。有关令社会满意的行为的文化差异同样也可能影响测验分数，在一些组群中抑郁量表的高分可能来自自我贬低和谦虚的强烈传统，另一些组群中高分可能说明真正的情绪问题，这些问题起源于童年教养实践、社会角色的冲突、少数人群体的挫折感和其他众多的文化差异。

跨文化的人格评定问题比测验的可转换性 (transportability) 问题更广泛，更确切地说，它包括应用于行为的概念系统的可转换性，诸如特质和特质等级 (Guthrie, Jackson, Astilla & Elwood, 1983) 而且，正如早期在能力评定领域发现的那样，现在人们越来越认识到一些重要人格维度并不具有普遍性。因此，除了不断努力去改编和翻译传统的英美的工具以供不同的人群使用外，一些用于评定本土文化和亚文化群体的特殊领域的人格测验正在编制中 (参见 Dana, 1993; Lonner & Berry, 1986)。然而，一

些特殊项目和调查表不可能在不同的文化间转换,有理由相信等级特质模型在综合跨文化的研究结果中可能有用。假如用适用于每一文化的工具测量这一文化中具有重要意义的行为,以此为起点,那么确定一些高水平的具有跨文化的普遍性或至少具有较广的可概括性的心理结构是可能的 (Anastasi, 1992c; Diaz-Guerrero & Diaz-Loving, 1990)。

**特质和情境** 特质和情境在行为分类中并不是不相容的方面,这可由评定测验焦虑的自陈调查表来说明 (I. G. Sarason, 1980)。斯皮尔伯格和他的助手 (Spielberger 等, 1980) 编制的测验焦虑调查表 (Test Anxiety Inventory, TAI)<sup>[21]</sup> 就是一个例子。这一工具主要是特质测量。然而,特质是按照一组规定的情境来界定的,这种情境集中于测验和考试,高测验焦虑的人倾向于把评价性情境知觉为对个人的威胁。调查表包括 20 个描述在测验或考试之前、之中或之后的反应的陈述句。要求回答者通过标记他们体验每一反应的频度 (几乎从不、有时、经常、几乎总是),表明他们一般是如何感受的。典型的例子包括“对重要的考试我变得麻木”和“每当考试时,我都有不适和不安的感觉”。TAI 对在测验情境中的焦虑倾向给出一个总分,并把通过因素分析确定的两个主要成分的分量表分数称之为担心 (worry) 和激动 (emotionality)。在这种情境中,担心被定义为“对失败的结果的认知忧虑”,而激动被定义为“由评价紧张而引起的自主神经系统的反应”(Spielberger 等, 1980, p. 1)。

测验焦虑剖面图更进一步提供了对情境的详细说明 (Oetting & Deffenbacher, 1980)。<sup>[22]</sup> 在这一工具上,回答者评定他们在各个项目上的反应,这些项目包含了焦虑感和思维干扰感觉。获得两类焦虑分数,每类有 6 个测验情境,要求回答者想象自己处于这些情境之中,从“多项选择测验”和“不宣布的突然袭击的

小测验’到‘作一次演说’。

也许因为测验焦虑的普遍性和在学业情境中研究它较容易，所以在美国和别的国家中，关于这一主题的研究和理论化的热情一直没有降低(Hagtvet 和 Johnsen, 1992)。不管怎样，这一构念为特质和情境概念如何用于行为分类——特别是在人格领域里——提供了良好的模式。特质构念可有不同程度的广义与狭义的定义，并可能与特定的情境类型有关。

**特质和状态** 由人格测验来评定的行为领域概念化的另一途径涉及特质和状态的区分。这种区分在斯皮尔伯格和他的合作者(Spielberger, 1985; Spielberger 等, 1983)编制的状态—特质焦虑调查表(STAI)中得到极其清楚的说明。在这一工具的编制中，状态焦虑(S—焦虑)被定义为一种短暂的主体感到紧张和忧虑的情绪状态，这一状态的强度和波动随时间变化。状态焦虑通过 20 个简短的描述句来测量，个体回答他或她当时的感受(如‘我感到平静’；‘我极度紧张不安’)，答案用个体的感受强度表示(完全不、有点、中度、非常)。

**特质焦虑(T—焦虑)** 指的是相对稳定的焦虑倾向，即个体以高状态—焦虑强度对感到威胁的情境作出反应的倾向。要求回答者对 20 个陈述中的每一个标记出现频率(几乎从不、有时、经常、几乎总是)以表明他们的一般感受，如“我对事情放心不下”和“我是个沉着的人”。高特质焦虑的个体比低特质焦虑的个体倾向于更经常表现出状态焦虑的加强，因为他们把更多的情境作为威胁性的或危险的加以反应。他们特别敏感于对他们的自尊形成威胁的人际情境，如绩效评定或失败的经历。然而，在一定的语境中，状态焦虑是否增加取决于个体基于她或他过去的经历把情境知觉为威胁性的或危险的程度，STAI 及其儿童版，即儿童状态—特质焦虑调查表(STAIC)已被译成 43 种语言，

并已有超过 6000 份的研究文献(Spielbergen, 1989; Spielberger & Sydeman, 1994)。

状态—特质的区分被斯皮尔伯格和他的同事们应用于另一个随后发展起来的调查表,即状态—特质愤怒表现调查表(STAXI, Spielberger, Johnson, Russell, Crane, Jacobs & Worden, 1985)。STAXI 包含 44 个项目,代表愤怒经验和愤怒表现领域。愤怒经验领域由与 STAI 的量表类似的两个量表评定,称之为状态愤怒(S—愤怒)和特质愤怒(T—愤怒),T 愤怒有两个分量表 愤怒气质和愤怒反应,各有 4 个项目。表现愤怒的频率可用 3 个八项目的量表来获取:愤怒内向(Anger-in)、愤怒外向(Anger-out)和愤怒控制(Anger-Control)。比斯金(Biskin, 1992)和雷茨拉夫(Retzlaff, 1992)关于该测验的评述可见于第 11 版 MMY。

## 人格调查表的现状

人格调查表的编制和使用为所有心理测试中所遇到的共同问题以外的特殊困难所困扰。人格测量中的印象处理问题远比能力倾向测试要严重得多,人格测验中测量的行为也比能力测验中所测内容更易随时间而变化。后一事实使测量信度的确定变得复杂,因为测验成绩中随机短暂的波动,很可能与全面的系统行为变化混淆起来。即使只是较短的时间间隔,也不能假设测验反应中的变化只限于测验本身,而不具有非测验行为方面的特征。与此有关的一个问题是,非认知领域中反应的情境特殊性要比认知领域中的大。

面对人格评定的复杂性和寻求对这些持久问题的创新的解

决方法的研究在 90 年代已经得到复苏。这一时期以重要的理论和方法学上的进步为特征。<sup>[23]</sup>对人格测量的早期批评无疑起着有益的作用，并部分地刺激了心理测量学这一领域随后的发展。然而，我们必须防止这样的危险，即在消除错误想法的热情中可能丢失良好的和有用的概念。例如，完全抛弃诊断性人格测验和特质概念的偶然建议，表明对这两个概念的界定过于狭窄。诊断不一定意味着使用传统的精神病分类方法或应用医学的“疾病”模式给个人加上一个标签。诊断性测试应该用作描述和理解个体，确定她或她的问题，并作出恰当的行动决策的一种辅助方式。同样，如果我们要在科学上或从其他方面了解行为，必须对行为进行必要的分类。而特质就是指这些行为类别，最佳的类别广度将随着评定的具体目的不同而变化。第十一章中提到的层次特质模型能同样好地用于非认知行为的描述。在某些情况下，相对宽广的特质最好；而在另一些情况下，需要评定狭窄的特别界定的行为。

## 注 释

- [1] 关于目前人格理论和研究的方法更详细的背景材料，见艾肯 (L. R. Aiken, 1993)、伯格 (Burger, 1993) 和麦迪 (Maddi, 1989)。
- [2] 关于初版 MMPI 的相当简洁的说明，见本书的早些版本 (如 Anastasi, 1988b)，对这一工具较广泛的论述，可见达尔斯琼 (Dahlstrom)、威尔希 (Welsh) 和达尔斯琼 (Dahlstrom, 1972, 1975) 的经典著作。
- [3] 最初的文章重印于达尔斯琼和达尔斯琼 (Dahlstrom & Dahlstrom, 1980) 的著作中。
- [4] 来自明尼苏达人格调查表 - 2，版权归明尼苏达大学董事会，版权保留，重印需经允许。
- [5] 也有一个不确定 (?) 分数代表双重标记或遗漏的项目，如果这类计算超过 30 项，则这一测试记录就被认为极其值得怀疑或很可能无效。

- [6] 从 1996 年起, MMPI-2 出版者不再提供 10 个威纳—哈蒙 (Wiener-Harmon) 细微—明显分量表
- [7] 见第 1 章梅耶斯—布里格斯 (Myers-Briggs) 类型指标的讨论。
- [8] 还可以看美国心理学家 ( *American Psychologist*, 1993, pp. 1289 ~ 1304 ) 上相同的评论部分中克劳格 (Kroger) 和武德 (Wood) 之后的论文。
- [9] NEO 三个字母代表神经质 (Neuroticism)、外倾性 (Extraversion) 和经验的开放性 (Openness to Experience), 但这些词本身并不用于调查表的全称。
- [10] 不适应性和适应性人格表 (SNAP) 是一个新的有趣的测量, 也用来评定人格异常。与 MCMI 相反, SNAP 使用通过因素分析获得的量表 (参见 Clark, McEwen, Collard & Hickok, 1993)。
- [11] 现在 MAPI 只被推荐用于正常青少年人格的非临床评定。
- [12] 这种项目型式代表了 EPPS 的重要特征, 将在下一部分作为迫选技术的例子加以进一步讨论。
- [13] 这一工具已在第 7 版、第 8 版、第 10 版心理测验年鉴中得到全面评价
- [14] 见第五章。
- [15] “操纵”和“欺骗他人”有时代替“印象处理”使用。
- [16] 对评定装病的量表的深入评论见伯里 (Berry)、威特 (Wetter) 和贝尔 (Baer, 1995)。
- [17] 海尔默斯 (Helmes) 和莱登 (Reddon, 1993) 也已经指出, 如果量表是双极的和它的项目记分是不平衡的, 那么由低分揭示的信息量大大减少。
- [18] 还见于美国心理学家 ( *American Psychologist*, 1990, pp. 289 ~ 295 ) 中同一评论部分埃德华 (Edwards) 的其他文章。
- [19] 见于安娜斯塔西 (Anastasi, 1983a) 的回顾。
- [20] 行为一致性这一概念来自阿尔波特和凯利 (Allport, 1973 & G. A. Kelly, 1963) 表述的人格评定的早期独特方法。
- [21] 在被试的表格上标为“测验态度问卷”, 在第 9 版 MMPI 和 TIP-IV 中也这样列出。
- [22] 测验焦虑剖面图不再印刷, 然而, 写信给欧基纳·奥汀 (Eugene Oetting), (Department of Psychology, Colorado State University, Fort Collins, CO

80523}), 可获准复制用于研究目的。

- [23] 勃劳顿(Broughton) 将原型概念和多维量表技术应用于人格评定的工作, 说明了这一领域某些最引人入胜和快速发展的方法学上的进展(Broughton, Boyes & Mitchell, 1993)。在理论方面, 这一环状模型已经提供了人际行为领域内各种传统的整合框架(参见 Hofstee 等, 1992; J. S. Wiggins, 1996; Wiggins & Pincus, 1992)。另外, 人格和智力领域之间的多重关系正在不断讨论和探索(见第十六章)。

## 第十四章 兴趣与态度测量

一个人兴趣和态度的性质和强度是体现其人格的一个重要方面。这些特征在很大程度上影响一个人在教育和职业方面所取得的成就，影响他的人际关系、从休闲活动中享受到的乐趣，以及日常生活的其他主要方面。尽管测验一般是针对某个变量的测量而设计的，但现有工具不能刻板地划分成彼此分离的兴趣类或态度类，而是相互重叠和交叉的。因此，一个旨在评定对探究性、艺术性或传统性职业的兴趣的相对强度的问卷，也可以用于测定一个人对纯自然科学、纯艺术、实际任务及其他方面的态度。

价值显然也和生活中的选择相关，对价值的探讨也总是和兴趣、态度、爱好相联系的。人格和社会心理学家们对价值问题已做了大量的研究，并且这些研究仍继续在进行，其中包括就价值跨文化普遍性问题开展的有趣而广泛的调查（S. H. Schwartz, 1992, 1994; S. H. Schwartz & Sagiv, 1995; Super & Sverko, 1995）。然而，近年来完全以价值评定为目标的、标准化的、商业上可利用的测试工具的发展相对滞后。这和价值测量所特有的一些问题有关，诸如从价值范畴中系统抽样的困难性，确定合适的抽象概括水平等。此外，业已证明一些早期被广泛使用的价值测量和这一领域最终形成的价值概念并不相符（Braithwaite & Scott, 1991）。<sup>[1]</sup>许多正规的价值评定目前已合并到了用来促进择业决



策和评定工作态度与动机的测量工具之中，<sup>[2]</sup>而一些“单独的”测验，如“生活价值调查”(Brown & Grace, 1996)和“价值量表”(Nevill & Super, 1989)仍在出版或修订。

对兴趣研究的最强烈的原动力来自于教育和职业的咨询，这将在第十七章中进行探讨。其次，职业选择和职业分类也推动了这一领域的测量的发展。从雇员和雇主双方的观点来看，考虑个人的兴趣是很重要的。对观点和态度的评定原本主要是社会心理学的问题，例如，对不同团体的态度显然会影响团体之间的关系。同样，对各种各样的问题、机构、活动的公众意见的测量和预测，社会心理学家和商务人员、从政者和其他应用领域的从业人员都非常关注。对观点和态度的测量在消费者研究和雇员关系领域也已取得了显著的进展。

这一章所考察的工具都是自陈调查表，这些调查表是采纳了第十三章中所述的一种或几种方法编制而成的。然而，应该注意的是，这一领域的研究和所有人格构念的测量一样，正在不断探索其他的方法，这将在第十五章和第十六章进行讨论。

## 兴趣调查表 现状

绝大部分兴趣调查表<sup>[3]</sup>的设计是为了评定个人对不同领域工作的兴趣。有些兴趣调查表也提供对课程或其他研究领域的兴趣分析，而这通常又和择业有关。尽管从 50 年代以来，用于咨询中的各种测试在总量上并没有大的起伏，但相对能力测试而言，兴趣测试的使用增长较快 (Zytowski & Waman, 1982)。从最近设计或修订的调查表中可以看到职业咨询方面的一些重要变化，其中之一就是越来越强调自我探索。越来越多的工具给

个人提供各种机会，去研究详尽的测试结果，并将测试结果与职业信息以及其他有关个人条件和经验的资料相联系。职业决策的培训也日益备受关注。本章将讨论的职业兴趣调查表应该以第十七章中的职业评定为背景加以考虑。

其次的相关变化是关于兴趣测定的目标。现在，越来越强调拓宽个人的职业选择面。事实上，已采用“探索效度”(exploration validity)这一术语来指兴趣调查表在有助于职业探索的行为方面的效果(参看如 Randhal, Hansen & Haverkamp, 1993)。与第十七章中列举的较为综合的职业定向方案一样，兴趣调查表现已用于使个人了解什么职业是适合自己的。

第三个重要的变化是它自身与拓宽职业选择面有关，关系到兴趣调查表中性别公平的问题。一般而言，兴趣调查表是将个人所表达的兴趣与从事不同职业的人群的兴趣进行比较。这种比较或是对各项回答进行评分，或是对广泛的兴趣领域测验成绩进行解释，或两者兼而有之。这种方法是一种评价个人兴趣的主观的经验性的程序，但它倾向于使现存的职业组群差异存在下去。在某些职业中，如果男性和女性的比率存在巨大的差异，例如在工程界和护理界，在某种程度上这种差异会影响用兴趣调查表获得的男性和女性结果的解释。为此，人们对减少兴趣调查表中性别偏见的方法进行了大量的讨论和广泛的研究(Tittle & Zytowski, 1978; Zytowski & Borgen, 1983)。现在，已经制定出评定职业调查表中性别偏见和性别公平性的一套规则，并已得到广泛传播。<sup>[4]</sup> 近来，不仅已经修改和编制的几乎每一种调查表都体现出这些规则的影响，而且还制定了一些条款用来处理性别公平性的问题。采取的步骤之一是在调查项目的用词上消除性别偏见。此外，其他常用的解决办法，包括根据性别角色社会化中的典型差异来平衡项目内容，并提供各性别组在调

查表上的最适当的常模。这些测量改善了为职业顾问和兴趣调查表的其他使用者所使用的技术资料的质量。然而,由于在兴趣模式和职业群体成员的性别差异问题上的不确定性,还存在一些涉及有关解释和使用兴趣调查表资料的社会政策问题,这些问题在女性的职业评定中需慎重考虑(Hackett & Lonborg, 1994)。

最近出现了一个新的情况,主要是对为残疾的和环境不利的群体提供职业指导的法律的反应(L. S. Gottfredson, 1986b; Reed, Rotatori & Day, 1990; Szymula, 1990)。<sup>[5]</sup>在职业评定方面残疾者的特殊要求已促进了一些图片兴趣调查表的发展,它们使用图画或视听刺激,如幻灯片和录像(Elksnin & Elksnin, 1993; Kapes 等, 1994, pp. 307 ~ 345)。这些工具的最初版本,回避了传统纸笔式兴趣调查表的阅读要求,从心理测量学的角度来说,这是不适当的,特别在效度方面。然而,图片兴趣调查表代表了测验刺激的一种创新方式,随着技术的改进和为特殊人群提供各种测量的需要的扩展而进一步发展。传统工具的改进,如第九章中所描述的关于特殊人群的能力测验的改进,当然也能用于兴趣评定。1990年美国残疾人法令(P. L. 101 ~ 336)在某种程度上激发了这一领域的研究兴趣和工作热情(Bruyere & O'Keefe, 1994)。

## 斯特朗兴趣调查表(SII)

**SII 的初版与发展** 这个兴趣调查表的最新版本出版于1994年,但是已有很长的历史。斯特朗(E. K. Strong)于1919 ~ 1920年参加卡耐基科技学院举行的关于兴趣测量的研究生讨

论班时首先提出了编制此调查表的一般观点 (D. P. Campbell, 1971,第十一章;Fiyer, 1931, 第三章)。当斯特朗职业兴趣表 (Strong Vocational Interest Blank, SVIB)在 1921 年第一次发表时,引入了测量职业兴趣的两个主要程序。首先,这些项目涉及回答者对于日常生活中经常碰到的各种具体活动、对象或不同类型的人的喜欢或不喜欢。其次,这些回答被经验地对应于不同的职业。因此,这些兴趣调查就是第一批采用项目的标准分类的测验。在这以后,像 MMPI 和 CPI(第十三章)这些人格调查也都遵循这种方法。人们发现从事不同职业的人具有一些共同的兴趣特征,使他们区别于其他职业的人。这些兴趣方面的差异不仅扩展到直接与他们工作活动有关的事物,而且包括学校科目、癖好、运动、个人喜欢的游戏方式或书的类型、社会关系和日常生活的许多其他方面。因此,编写一份调查表,以探求个体对熟悉事物的兴趣,并且由此来判断他或她的兴趣与那些成功地从事某种特定职业的人的相似程度是可行的。

从 20 世纪 70 年代开始,人们在 SII 的后继版本方面引入并实施了广泛的革新 (D. P. Campbell, 1974; D. P. Campbell & Hansen, 1981; Hansen & D. P. Campbell, 1985; Harmon, Hansen, Borgen & Hammer, 1994)。主要的变化包括:(1)提出了一个理论框架来指导对分数的组织和解释;(2)对先前的男女兴趣调查表进行了合并,根据新的男女样本,对所有的职业量表重新制定了常模;(3)用于那些不要求大学学历的职业/技术职业的量表,在数量上有了很大的增加,这是一个在以前的版本中没有充分体现的范畴。

**The SII - Form T317** :一般描述 现在的斯特朗兴趣调查表包括 317 个项目,分为 8 个部分,在前 5 个部分里,回答者用 L “喜欢”、I “中性”、D “不喜欢”来表示他或她对事物的态度。在这

5个部分里的项目分别属于以下一些范畴：职业、学校科目、活动(例如演讲、修钟、为慈善事业筹款)、休闲活动及与各种不同类型的人接触(例如很老的人、军官、处于危险中的人)。另外两部分，要求回答者从成对活动(如和物打交道还是和人打交道)与从工作领域的四要素(即思想、数据、事物和人)的所有可能配对中选择自己的爱好。调查表的最后一部分要求回答者用“是”“否”或“不确定”标记一系列自我描述的陈述。

SII 兴趣调查表只能通过电脑记分，在出版商指定的记分中心或者使用从各地出版商那里得到的软件进行。图 14-1 表示了最新的 SII 剖面图的第一页，提供了做测验的人在调查表的主要量表上的最高得分的总体印象或概况。<sup>[6]</sup>该表将得分分为三种水平，它们的广度各不相同。最广泛最全面的是 6 种一般的职业主题分数，下一部分包括 25 种基本的兴趣量表，而最具体的层面提供了 211 种可利用的职业量表。除这些以外，斯特朗(Strong)的表 T317 发展了 4 种新的个人风格量表，评定在工作类型、学习环境、领导风格和风险承担/冒险精神方面的爱好。这个剖面图也提供了一套施测指标，包括回答的总数(以检测过多的遗漏)、罕见的或不寻常的回答数目，及调查表 8 个部分中每一部分三种回答选择各自的百分率。这些指标可以用来检查粗心和特殊的反应定向。

SII 的职业兴趣分类渊源于霍兰德(John Holland, 1996, 1985/1992)建立的理论模型，并得到霍兰德和其他独立的研究者广泛研究的支持。由霍兰德的模型所确定的基本职业主题被称为现实的(R)、研究的(I)、艺术的(A)、社会的(S)、进取的(E)和传统的(C)。<sup>[7]</sup>每一种职业主题不仅描述了一种人，而且描述了一种他觉得最适宜的工作环境。每一种工作环境出现在相应的那类人中并在他们中占主导地位。据霍兰德的观点，人并不是严格

# 斯特朗兴趣调查表

总体报告: 咨客 1

身份:  
年龄: 20  
性别: 男

测试日期:  
记分日期:

总体印象: 咨客 1 的结果概况

1 页 / 6 页

R I

C A 基本职业主题

E S

一般的职业主题描述了对六大领域的兴趣, 包括工作、休闲活动、各类人和工作情境。你在每一领域的兴趣都以等级表示, 注意每一主题有一代码, 由该主题名称的第一个字母代表。

你可以用打印在你的结果下面的主题代码确定你可能觉得感兴趣的学校科目、非全日制工作、大学专业、休闲活动或职业。如何使用你的主题代码的建议见这一简介的最后面。

| 主题<br>代码 | 主题  | 很少                       | 一点                                  | 中<br>等                              | 高                        | 很高                       | 典型兴趣    |
|----------|-----|--------------------------|-------------------------------------|-------------------------------------|--------------------------|--------------------------|---------|
| I        | 研究的 | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 研究、分析   |
| R        | 现实的 | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 建造、修理   |
| C        | 传统的 | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 计算、处理数据 |
| A        | 艺术的 | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 创造或欣赏艺术 |
| S        | 社会的 | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/> | <input type="checkbox"/> | 帮助、指导   |
| E        | 进取的 | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/> | <input type="checkbox"/> | 销售、管理   |

你的主题代码是 IRC—。

你可以用包含这些字母的任何组合的代码探索职业。

R I

C A 基本兴趣量表

E S

基本的兴趣量表测量你在 25 个具体领域或活动方面的兴趣, 只有 5 个你最感兴趣的领域列于下面的评定等级中, 你在 25 个基本兴趣量表上的结果在第 2 页上。

每一量表的左边是一个字母表示该活动与六个基本职业主题中的哪一个关系最为密切, 这些代码能帮助你确定其他你可能感兴趣的活动的。

| 主题<br>代码 | 基本<br>兴趣 | 很少                       | 一点                       | 中<br>等                              | 较<br>高                              | 很<br>高                              | 典型<br>活动   |
|----------|----------|--------------------------|--------------------------|-------------------------------------|-------------------------------------|-------------------------------------|------------|
| R        | 体育运动     | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | 游戏或观看运动    |
| I        | 数学       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | 数字或统计工作    |
| I        | 医学       | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | 医药或生物工作    |
| R        | 机械活动     | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | 使用工具或仪器的工作 |
| A        | 应用艺术     | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/>            | 生产或欣赏视觉艺术品 |

R I

C A 职业量表

E S

职业量表测量你的兴趣与喜欢从事这些职业的人的兴趣在多大程度上相似。只有你的兴趣与那些人最相似的 10 个量表列于下面的评定等级中, 你在所有 221 个职业量表上的结果在第 3、4、5 页上。

续 表

每一量表左边的字母表示最能描述从事该职业的人的兴趣,你可以用这些字母发现另外你可能感兴趣的职业。在这份资料的第 6 页上审阅你的结果,回头看第 5 页末尾,在你最感兴趣的主题中找出其他职业。

| 主题<br>代码 | 职业     | 很<br>不<br>相<br>似         | 不<br>相<br>似              | 中<br>等<br>相<br>似         | 相<br>似                              | 很<br>相<br>似                         |
|----------|--------|--------------------------|--------------------------|--------------------------|-------------------------------------|-------------------------------------|
| RI       | 工程师    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> |
| RI       | 音响师    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input checked="" type="checkbox"/> |
| IR       | 牙科医生   | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| R        | 管道工    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| CI       | 保险统计员  | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| IR       | 配镜师    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| CE       | 会计师    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| IR       | 化学家    | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| R        | 自动化机械工 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |
| RIC      | 电工     | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/>            |

个人性格量表  
测量你的工作风格、学习环境、领导风格和风险承担/冒险精神方面的水平这一信息,帮助你确定你觉得满意的工作环境、教育情境和活动类型。你在这四个量表上的结果在第6页上。

CPP 咨询心理学家出版公司 3803 Bayshore Road, Palo Alto, CA 94303

图 14-1 斯特朗兴趣调查表结果概况。剖面图的第 1 页提供了一个 20 岁男性大学生获得的分数的“简单印象”[引自斯特朗兴趣调查表:应用和技术指导, L. W. 哈蒙(L. W. Harmon), 乔-伊达·汉森(Jo-Ida C. Hansen), 弗雷德·鲍根(Fred H. Borgen)和艾伦·L·哈默(Allen L. Hammer)]。

地分为这 6 种主要类型,更确切地说,它们的特点是与其中一种或几种的相似程度。这种按相似程度建立的类型与类型组的联合,就提供了模式或代码的多重性来描述个体差异的多样性。

图 14-2 把这 6 种职业主题放在霍兰德的 6 边模型的 6 个角,为了简洁起见,通常用每一种主题的第一个字母来表示。这样,按这些主题在六边形上的顺序就得到了 R-I-A-S-E-C 这个缩略词,于是,人们就用这个词来表示这个模型。要指出的是这个六边形中相关最高的是边长上处于相邻位置的职业主题,相关最低的是处在对角线上的职业主题。例如,现实的量表与研究的量表相关为 0.53,但与艺术的量表相关只有 0.04,与

社会的量表相关只有 0.06。类似地,在这个六边形上的这些职业量表大部分遵守这个规则,例如,工程师的代码是 RI,处在现实的和研究的之间,银行家的代码是 CE 处在传统的和进取的之间。通常,一种职业在一种职业主题上得分高,那么在相对的那种职业主题上得分就会低(如艺术的和传统的)。当一种职业量表与对角线上的几个职业主题出现明显的相关时,则这种职业通常包括不同性质的亚群,它们具有公认的不同职业功能。

25 种基本的兴趣量表分为 6 个基本的职业主题,这些量表包括若干类密切相关的项目。基本兴趣量表在内容上与职业量表相比更为同质,因此有助于理解为什么一个人在特定的职业量表上得分高。

各种职业量表构成了最初 SVIB 的主体,现在已按照一般职业主题加以分类。在该调查表的后续研究中,已增加了新的量表,旧的量表现在已用新的标准样本加以更新。T317 包括 211 个职业量表,其中 83% 的常模根据 80 年代和 90 年代获得的样本制定。除了 7 个量表之外,所有量表都能确定足够的回答者的位置,以建立女性常模和男性常模(即每一种性别有 102 个量表)。迄今为止,5 个量表只有女性常模,2 个量表只有男性常模。

编制每一个职业量表所采用的大部分样本包括 200 或更多的人,但是实际人数从 60 到 1187 不等。就 1994 年的版本来说,人数在 55000 以上,其中将近 4 万人符合用于量表编制的规定要求。职业标准组主要包括年龄在 25~60 岁之间的人,他们从事该职业至少 3 年,对工作表示满意,并正在履行该职业人群的典型职责。还有两个一般参考样本,包括 9484 个男性和 9467 个女性,所有这些人的测试都在 90 年代进行。GRSs 由 98 种不同的职业组成,其中 90 种职业由每一种职业可得到的人群中随



机选取 200 人代表,8 种职业由只有 92 人和 195 人之间的样本代表。<sup>[8]</sup>斯特朗 GRSs 中的个体受教育水平远远高于美国一般人群的水平,事实上,他们中近 80% 的人具有学士、硕士或更高的学位(Harmon 等,1994, p. 110)。大学毕业生的这一比例超过了一般人群中的相应数字,甚至超过了专业和白领的比例,这些工作是斯特朗 GRSs 中列举的职业的绝大部分。因此,和它以前的版本一样,斯特朗的最新版本可能因它的样本缺乏代表性而受到批评(Worthen, 1995)。

SII 的项目通过两种方法放入不同的量表。那些在基本职业主题和基本兴趣量表中的项目根据内容和人们回答方式的相

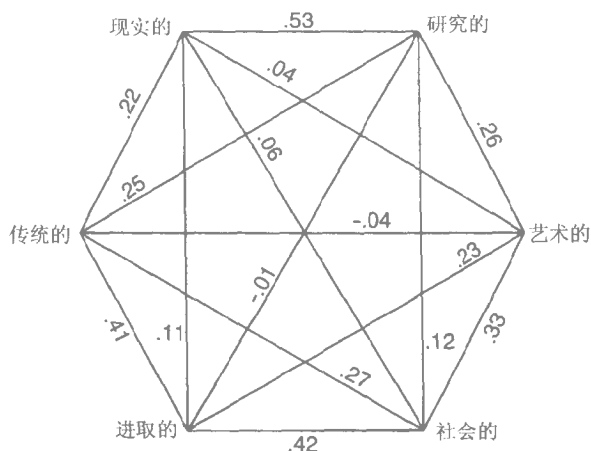


图 14-2 霍兰德的六边形基本职业主题模型。对每一对主题来说,相关系数基于由 9484 男性和 9467 女性组成的一般参照样本(翻印征得出版商同意,引自斯特朗兴趣调查表:应用和技术指导, p.51)。

似性,通过因素分析,分为同质的几个群组。另一方面,每一个职业量表的项目根据每一性别的职业标准样本和 GRSs 之间项目反应的百分比的差异进行选择 and 加权。例如,在女性会计师量表中, $a+1$  权重表示该回答在女性会计师中较一般女性中出现的多,而  $-1$  权重表示该回答在女性会计师中较一般女性中出现的少,那些不能明显区分女性会计师和 GRS 中女性的回答不出现在女性会计师量表中,不管会计师们是否经常选择这些回答。每一种职业量表上原始总分就是回答权重加减的代数和。

**记分和解释** 所有斯特朗调查表的得分用平均数为 50、标准差为 10 的标准分数报告。对基本职业主题和基本兴趣量表来说,计算标准分数的常模样本是男女混合在一起的一般参照样本 ( $N=18951$ )。然而,剖面图上分数的解释和图示都基于同一性别常模,也可以通过参考剖面图上呈现的图形,与其他性别常模比较。在职业量表上,每一位回答者实际上得到两个标准分数——一个来自女性职业样本,另一个来自男性职业样本。尽管主要是同性别的比较,但这些报告程序为咨询者和答卷者进行跨性别的比较提供了必要的资料,以便对个体的反应范型进行更充分和更有效的解释。

《SII 应用和技术指导》( *The SII Applications and Technical Guide*, Harmon 等,1994) 质量上乘,并包含大量的信息,可以帮助咨询人员解释结果、与咨客讨论这些结果的可能含义。例如,产生内部一致结果的剖面图倾向于具有较高的预测性。另一方面,如基本兴趣量表和职业量表之间某些不一致,有助于我们了解回答者所表示的爱好的性质和来源。从另一个角度来看,通过与美国劳动部(1991)颁布的《职业名称词典》( *Dictionary of Occupational Titles*, DOT ) 和其他类似的参考资料( Gottfredson & Hol-

land, 1989; Harmon 等, 1994) 中列出的职业之间已经建立的联系, 调查表获得的结果可以推广到其他相关的职业。《应用和技术指导》也有几章专门讲述将该调查表如何用于特殊的人群, 如各种各样的文化群体和残疾人。此外, 还有一些其他的出版物用来帮助咨客和咨询人员理解和使用斯特朗调查表的结果。

**心理测量学的评价** 对斯特朗调查表已经开展了连续的研究, 产生了大量有关它的信度和效度的资料 (D. P. Camoell, 1971, 1977; Hansen & Campell, 1985; Harmon 等, 1994)。就职业量表来说, 对 191 个受雇的成人间隔 3~6 个月的重测信度中数为 0.90 就基本兴趣量表来说, 相应的信度为 0.86 基本职业主题的相应信度为 0.89。斯特朗先前版本的职业量表的长期稳定性也是高的, 在这 20 年间相关性下降的主要是 60 年代和 70 年代年龄在 25 岁以下的人和 80 年代年龄在 25 岁以上的人。

同时效度资料由不同的职业样本之间及职业样本和参照样本之间的差异程度提供。对包括在 1994 年修订版本中的 211 个样本来说, 中数交叠为 36%, 这一点反映了职业样本和 GRSs 之间的平均分距离在两个标准差之下。预测效度已经在先前版本的几个样本中经过长时间间隔后得到检验。有证据表明最先的职业剖面图和最后从事的职业之间具有明显的一致性。一个具体的例子是跟踪了 40 年的心理学家样本, 他们的职业生涯显示了他们最初的剖面图的曲线和诸如工作变化及从教学或研究转向行政或应用性工作的频率之间具有某些有意义的关系 (Vinitsky, 1973)。另一个研究发现了在 9 个西方国家测试的心理学家样本之间剖面图的跨文化高度相似性 (Lonner & Adams, 1972) 尽管斯特朗调查表多年来积累了广泛的研究基础, 它们作为职业选择与职业稳定的指标也具有明显的可靠性, 但为了直接研究最新版本的预测效度, 有必要用它进行新的研究。

关于构想效度、职业量表与基本职业主题之间的关系,及图 14-2 中描述的各主题之间的相互关系,都具有特别的意义。霍兰德(1966,1985/1992)提出的 R—I—A—S—E—C 模型在职业兴趣研究方面具有巨大的启发价值,斯特朗调查表和其他使用相同格式的兴趣测量一起,反过来又成为这一研究传统的有机组成部分。使用斯特朗调查表和其他调查表的研究结果,一般来说与该模型的预测很一致。

SII 的结构和组织使它能通过不断的修订得到一些完善和扩展。斯特朗调查表不仅是兴趣测量领域的先锋,而且是美国咨询心理学家中使用最广泛的工具(Watkins, Campbell & Nieberding, 1994)。其他少数几个工具的使用也差不多同样悠久和普遍。然而,自从 60 年代以来,这一领域在新的测量工具方面经历了迅速的发展。这一发展部分地反映了对职业探索日益增长的注意以及对兴趣在这种探索中的作用的认可。

第十七章中讨论的职业探索的全面计划一般包括职业兴趣的测量,它与多重能力倾向测验分数和职业信息结合使用。例如,用职业兴趣调查表(Career Interest Inventory, CII)得到的兴趣资料,能和标准化的区别能力倾向测验一起使用(Psychological Corporation, 1991a)。类似地,职业能力倾向调查表和兴趣表-2(OASIS-2)通过提供 6 个能力倾向和 12 个兴趣因素的分數,对高中生的职业探索和职业发展具有辅助性作用。在职业计划领域的另一个变化可以哈林顿—奥西的职业决策系统修订版(Harington-O'Shea Career Decision-Making System-Revised, CDM-R, Harington—O'Shea, 1993)为例加以说明,它试图将兴趣、价值和能力的自陈资料与职业信息整合在一起。坎贝尔兴趣和技能调查(Campbell Interest & Skill Survey, CISSbCampbell, Hyne & Nilsen, 1992)也测量自陈的兴趣和技能,并以与斯特朗调查表相

似的方式组成,坎贝尔——CISS 的作者专注于斯特朗调查表有相当一段时间。增加技能方面的资料可以把兴趣量表上和能量量表上的高低分类型进行比较,这反过来扩展了该调查表为职业探索和决策提供的基础。CII、OASIS - 2、CDM - R 和 CISS 的评述可在卡佩斯(Kapes)等(1994)的著作中找到。

## 兴趣调查表：评论和某些启示

在近来许多可利用的兴趣调查表中,我们挑选出 4 种单独地加以介绍,因为其中每一种都在它的理论定向、方法论或适用人群类型方面说明了一些值得注意的特征。我们并不试图详细地描述或评价这些调查表,所有这些测量工具都在最近的《心理测量年鉴》中加以评论,其中既有目前的版本也有先前的版本。

杰克森职业兴趣调查表 ( Jackson Vocational Interest Survey, JVIS) 我们之所以选择 JVIS(Jackson, 1977)给以特别的注意,首先,因为它是复杂的测验编制程序的例证 其次,因为在几个方面它的方法与 SH 中所遵循的方法有很大差别。斯特朗和杰克森调查表的初版日期相隔 50 年,斯特朗调查表在项目选择和常模解释两方面都集中于具体的职业,而杰克森调查表在项目发展和记分系统两方面都使用宽泛的兴趣领域。斯特朗职业量表代表了经验准则确定和准则关联效度分析的极端例子,而杰克森调查表列举了发展每一阶段的构想效度;在斯特朗调查表中,绝大多数项目由回答者独立标记喜欢、中性或不喜欢,而在杰克森调查表中所有的项目都是迫选型的。

正如第十三章所描述的人格研究表和杰克森人格调查表, JVIS 编制的第一步是确定测量的构想和维度。根据已发表的工

作心理学研究和职业兴趣项目的因素分析与理性分类，选出的维度有两种类型。一类根据工作角色界定，另一类根据工作风格定义。工作角色是指一个人在该职业中做什么，某些角色与特定的职业或职业类型密切相关，如工程师、律师或小学教师，另外一些角色，如人际关系处理和专业指导与很多职业有关。工作风格不是指与职业有关的活动，而是指对工作环境或情境的爱好，在这些环境或情境中，期望会出现某类行为。工作风格维度一般与价值直接或间接相关，工作风格的例子包括计划性、独立性和支配性的领导。

调查表的编制经历几个阶段，包括项目的逐次试验和统计分析，以适合每一个工作角色和工作风格的详细规定。从超过 3000 个项目的初表开始，最初只分别表示喜欢一不喜欢的反应，程序包括为每一个量表准备的各部分项目的因素分析。反映在一个大的一般因素中的反应偏向，在进一步的项目分析之前在统计上加以消除。选择与自身量表的总因子分相关高而与其他量表相关低的项目。代表不同工作角色或工作风格，而且当它们分别呈现时显示出相似的认可频率的成对项目，组合为迫选形式。<sup>[9]</sup>

JVIS 的最终形式包含 34 个基本兴趣量表，覆盖 26 种工作角色和 8 种工作风格。该调查表同样适用于两种性别，虽然也有分别用于男性组和女性组的百分位常模。常模来自美国和加拿大的大学生和中学生大样本，34 个 JVIS 基本兴趣量表上的任何一个高分都表明其对特定工作领域中人们所做的事情和该工作情境所期望的行为方式感兴趣。

JVIS 可以很快地、很方便地对 34 个量表进行手工记分，但已有的基于计算机的记分方式利用更新的常模，并在简要报告或新版的扩展的陈述报告中介绍了另外一些得分分析。后者提

供了个别化的描述和解释材料及大量信息以帮助寻找职业,例如电脑上的报告包括来自 34 种基本兴趣量表的因素分析。这些分数仿效霍兰德的 6 个主题,包括 10 个一般兴趣主题:善于表达的、逻辑的、好奇的、实践的、坚持己见的、社会化的、助人的、传统的、进取的和交际的。它还提供学术满意感的度量、回答的一致性和不寻常性指标以及不能记分的回答计算方法。其他的分数分析涉及回答者的剖面图作为一个整体与从 17 类专业大学生和 32 类职业人群中获得的众数剖面图进行比较。职业剖面图来自 JVIS 和斯特朗调查表的结合使用,通过统计分析,在两者之间建立了联系。这种联系提供了在解释 JVIS 的得分时利用斯特朗调查表的职业数据库的一种方法。〔10〕

一些评论者认为,JVIS 项目的措词对许多高中学生来说可能太复杂了(D. T. Brown, 1989; J. W. Shepard, 1989)。调查项目手册最近对措词作了些许改变,但是这些改变是更新而不是简化所用术语。收入 90 年代中期常模的修订版手册也即将出版。同时,关于 JVIS 应用于职业咨询和职业指导的手册(Verhoeve, 1993, Jackson, 1995)也可以从出版商那儿得到。

库德职业兴趣调查表(The Kuder Occupational Interest Survey, KOIS) 库德(Frederic Kuder)编制的兴趣调查表的使用和斯特朗调查系列的历史几乎一样悠久。最早的版本是《库德爱好档案——职业版》,它的兴趣测量方法与斯特朗的主要有两个方面的不同 第一,库德采用 3 项迫选方法,回答者必须指出 3 种活动中哪一种最喜欢,哪一种最不喜欢;第二,获得的分数并不是针对具体的职业,而是针对 10 种广泛的兴趣领域:户外的、机械的、计算的、科学的、追求的、艺术的、文学的、音乐的、社会服务的和办事的。每一量表的项目都是根据内容效度进行阐述和暂时分组的,最后的项目选择基于内部一致性和与其他量表

的低相关性。库德一般兴趣调查 (KGIS) 后来是作为《库德爱好档案——职业的》的修订版和向下延伸而发展起来的。这一形式是为六至十二年级设计的, 使用较简单的语言和较容易的词汇, 只需要六年级的阅读水平。关于 KGIS 的评论 (M. Pope, 1995; D. Thompson, 1995) 可在第 12 版 MMY 找到。

更新版本的库德职业兴趣调查表 (KOIS) 和斯特朗兴趣调查表一样, 提供了关于具体职业群的分數 (Kuder, 1966; Kuder & Diamond, 1979; Kuder & Zytowski, 1991)。然而, 和斯特朗不同的是, KOIS 没有采用一般的参照组, 相反, 回答者在每一职业量表上的得分用他或她的兴趣类型与特定职业群的兴趣类型之间的相关来表示。<sup>[11]</sup> 这种兴趣调查表可以在现场记分或通过出版商用光学扫描记分, 还可以在计算机上施测和记分。得分目前可用于 109 种具体的职业群和 40 种大学专业。虽然, 这些量表只是基于男性或只是基于女性或是基于两者发展起来的, 然而, 所有量表上的得分都向男性和女性回答者报告。这种调查表覆盖的职业水平范围很广, 从美容师、卡车司机到药剂师、律师等。

通过对 3000 人 (代表该调查表覆盖的职业和大学专业领域的 30 个核心组, 每组 100 人) 的得分的深入细致的统计分析, 库德证明用该调查表采用的记分系统较通过使用一般参照组得到的职业量表能更好地区分职业群。关于其他职业量表的研究还在继续 (Zytowski, 1992)。

KOIS 目前提供了职业分数和 10 个广泛的、同质的基本兴趣分数, 称为职业兴趣评估 (Vocational Interest Estimates, VIE)。VIE 从简短的量表中获得百分位分数, 相当于早期的库德爱好档案中的 10 个兴趣领域分数。它们也能通过某些量表的直接对应和两个或三个库德量表的平均百分位数与其他量表的对应 (如库德量表中艺术的、文学的和音乐的平均百分位数对霍兰德



量表艺术的)转换为霍兰德的 R—I—A—S—E—C 主题代码。评论家一般在指出这一测量工具缺乏预测效度资料的同时称赞 KOIS 的技术特点,不能说明迫选形式对分数的影响也受到批评(见 Herr,1989; Tenopyr,1989)。

职业评定调查表—职业版 ( Career Assessment Inventory - The Vocational Version, CAI - VV)<sup>[12]</sup> CAI ( Johansson, 1984)第 1 版发表于 1975 年,形式接近斯特朗调查表。然而,与大多数兴趣调查表不同,它特别为寻找不要求四年大学学历或高级职业训练的职业的人设计,集中于技能性的行业、技术工作和服务业。目前可用的职业量表例子有空军机械工、牙科卫生人员、自助食堂工作人员、计算机程序员及注册护士。 305 个调查项目按 3 个内容类别分组 活动、学校科目和职业。每一个项目提供 5 种选择,从“非常喜欢”到“非常不喜欢”。CAI 只要求六年级的阅读水平,也可用于阅读技能差的成人。和斯特朗量表一样,CAI 提供 3 类主要量表的分数,包括 6 个霍兰德一般主题量表,22 个同质的基本兴趣领域量表和 91 个职业量表,还包括施测指标和 4 个非职业量表。

虽然产生 CAI 的程序近似于斯特朗量表所使用的程序,但该量表所有的数据收集和统计分析都独立进行,因此,除了一般主题量表,每一类别的量表是 CAI 特有的。关于 CAI - VV 的评论见基欧(Kehoe,1992)和瓦斯(Vacc,1992)的论著。

自我定向的探索(Self-Directed Search, SDS) 职业兴趣评定的另一种方法可由自我定向探索加以说明。这一测量工具是由霍兰德发展的,他的一般职业主题(本章前面已讨论过)的异质模型已吸引了广泛的注意,并已结合到目前的几个调查表中(Holland, 1985/1992; Holland, Fritzsche & Powell, 1994; Holland & Gottfredson, 1976; Holland, Powell, Fritzsche, 1994)。

正如它的标题所隐含的,SDS 为一种自我施测、自我记分和自我解释的职业咨询工具,虽然围绕兴趣组织,这种方法也要求对能力和胜任感进行自我评定。个人填写自我评定表,对回答进行记分并计算 6 个与霍兰德模型的主题(现实的、研究的、艺术的、社会的、进取的和传统的)相对应的总分。3 个最高的总分被用来形成 3 个字母的代码。<sup>[13]</sup> 使用一本附带的小册子《职业探测》,在 1335 种职业中找出其代码与被试的总分相似的那几种职业。挑选出这些职业,以便代表差不多所有的美国工人,使用者也可以使用转换代码来考察《职业名称词典》中的所有职业。它还提供另外的说明、程序和信息资源以促进个人的职业决策。

SDS 被广泛地用于各种情境中,并由作者和独立的研究者进行了大量的研究。从它出版以来的这些年中,已经过几次修订,以简化程序并减少职业决策中的性别偏见。SDS 以其简洁明了、自己做的特点以及在扩展个人的职业选择中的作用而受到青睐。除了调查表的正规形式(R 型),SDS 还有 3 个版本:(1)E 型(容易型),用于阅读水平有限的个人;(2)CP 型(职业策划型),为职业转换中的成年人设计;(3)职业探索型,以初中和高中学生为对象。

关于 SDS 的心理测量学特征,总分的信度指标一般来说令人满意。基本的 6 个主题构想效度分析,主要依赖于导致对主题进行系统阐述的研究和后来大部分起支持作用的鉴别因素分析研究(参见 Oosterveld, 1994),SDS 的同时效度和预测效度随样本组成的年龄、性别、教育水平及类型分布而波动(Holland, Fritzsche, Powell, 1994)。对 SDS 的批评主要针对它的记分和解释程序(M. H. Daniels, 1989; Manuele—Adkins, 1989)。不过,评论者的一致观点是该测量工具提供了探索职业的一种简单、价

廉和相对正确的方式关于 SDS 的参考文献及一些和它一起使用的提供信息的小册子可从出版商那里得到。

霍兰德(1966,1985/1992)明显地和那些将职业偏好当作生活方式加以选择的人结成同盟。这种选择反映了个人的自我概念和主要的人格特征。霍兰德的每一个职业主题相应于人格特征的一个“类型”或一群人格特征。某一个人可以根据一个或多个占优势的类型加以描述。这些主题也与各典型环境相对应,不同职业的情境特点可根据这些典型环境加以描述,环境不仅包括物理特征和工作要求,而且包括一起工作的人的类型(合作者、上司、顾客、咨客、学生)。根据霍兰德的观点,个人追求与他们的人格类型相一致的环境,而且这种一致性促进工作满意感、工作稳定性的提高及成就的取得。

从不同的观点看,霍兰德评定职业兴趣的一般方法与职业决策心理学的某些发展相一致。舒伯(Super, 1953, 1957, 1990)已经反复强调职业选择是自我概念的具体落实。关于各职业人群之间的人格差异有一个相当长的研究传统(参见 Borgen, 1986; Costa, McCrae & Holland, 1984; Osipow, 1973 第六章; Pietrife-sa & Splete, 1975, 第四章; Super & Bohn, 1970, 第五章)。职业的选择经常反映一个人的基本情感需要,而且职业适应是一般生活适应的重要组成部分(Tait, Padgett & Baldwin, 1989),因此,职业兴趣的评定——以及进一步识别职业群,某个个体的兴趣和态度与这些群体的兴趣和态度具有许多共同之处——成为理解不同人格的焦点之一。<sup>[14]</sup>

## 某些重要的趋势

**调查表的发展和使用** 当前兴趣测量中最明显的变化是职业心理学中两大理论观点的融合、编制调查表的不同方法的融合和进行解释时经验数据资料的交叉利用。越来越多的工具既提供同质的、广泛的兴趣量表的分数,又提供特定职业量表的分数。霍兰德模型的 6 个职业主题体现在近来发展和修订的日益增多的调查表中。

从不同角度来看,可用于大量人群的经验性职业资料间的联系正在建立起来。这一点可以从利用斯特朗职业资料来解释杰克森职业兴趣调查的结果中加以说明,也可以从许多当前的调查表,如霍兰德的 SDS 和 SII 及《职业名称词典》(美国劳动就业和培训部,1991)所提供的资料间的联系中得以阐明。这些发展加强和提高了任何单个工具的有用性。这些结果应该是有益的,只要这些联系是通过良好的心理测量学程序建立的,并采取步骤以防止在分数解释上可能的过于泛化。

新近发展和近来修订的兴趣调查表的另一个特点是它们所覆盖的职业范围的扩大。开始时,兴趣调查表集中在专业性职业,涉及低于大学或专业学校教育要求的职业。虽然某些早期的努力曾试图扩大所包含的范围(如 Clark, 1961),但这样的调查表未能得到广泛的使用。相反,我们现在发现斯特朗的最新版提供了大量的专业的/技术的职业量表,进入这些职业不求大学学历。而且,某些更新编制的工具,如职业评定调查表(Career Assessment Inventory, Johansson, 1984)职业版和职业方向调查表(Career Directions Inventory, Jackson, 1986)主要是为那些文

化程度不超过中等或技术学校的人设计的。这些发展至少部分反映了对所有职业水平有效的职业选择的重要性的认识,以及所有工作中兴趣对工作成功的体验和工作满意的体验起重要作用的认识正在不断提高。

还有一个明显的趋势,其意义已超越了兴趣调查表而扩展到其他种类的测验上。关于兴趣调查表的未来的一个引起争论的看法是,霍兰德认为这种调查表是干预技术。调查表对测验参加者具有哪些影响?例如,它可以支持和加强一个人现有的职业精神;对另一个人,它可以激起他对工作范围的广泛探索,注意到迄今尚未考虑的方面;对第三个人,它能增进自我理解。这多种可能的影响不仅可以在个人的分数解释方面,而且在兴趣调查表的实施和编制方面反映出来。这也许代表了计算机化适应性测验的另一种可能的应用。

在 90 年代,职业选择和职业发展心理学由于应用来自认知心理学领域的范式而得以丰富(参见 Peterson, Sampson & Reid, 1991)。把职业决策看作是一个问题解决活动,在人的一生中可能重复出现,这强调了有必要为个体提供实现最佳解决方案所需要的知识和信息加工技能。许多新的工具包含在职业探索的全面计划之内,这些计划是舒伯关于职业发展的观点和霍兰德职业选择观点的逻辑延伸。现在已越来越清楚的是,职业行为领域为个体积极地 and 直接地利用心理学见解和工具提供了空前的机会(Borgen, 1991)。

职业模型(Models of Occupations)这一章中所讨论的大多数职业测量,或者起源于霍兰德的职业主题模型(如 CAI 和 SDS)或者将这个模型或多或少地结合进他们的方法中(如 SH 和 KOIS)。事实上,在过去 20 年中,不仅在美国,而且在世界其他地方,霍兰德的理论假说在刺激职业心理学领域的研究方面

起着卓越的作用(见 Borgen, 1991; Lokan & Taylor, 1986)。在对模型进行非常全面的检验时, 特雷西和朗茨 (Tracey & Rounds, 1993)对各种工具的 R—I—A—S—E—C 量表进行结构性的元分析。这一研究使用了 104 个美国样本, 发现了支持该模型的证据。然而, 由相同的作者随后进行的对国际的和美国少数民族样本所作的类似的元分析并没有产生这种良好的拟合性 (Rounds & tracey, 1996)。另一个跨国家和跨民族的研究产生了混合的结果 (Fouad & Dancer, 1992; hansen, 1987; Khan, Alvi, Shaukat & Hussain, 1990; Swanson, 1992)。关于 R—I—A—S—E—C 模型的跨文化研究表明, 它的各个方面可能并不同等地适用于不同的文化。

无论如何, R—I—A—S—E—C 六边形的简单性和过去 20 年来它在研究方面绝对优势的地位导致大量的研究者提出, 现在可能到了这样的时刻, 为了加强其效用, 应该超越它, 或者在它里面加入更多的成分。<sup>[15]</sup>一些人要求通过发展新的理论结构来说明职业的其他方面, 如能力和对强化物的偏好, 而另一些人希望增加研究的变量和维度 (见 Dawis, 1992; Prediger & Vansickle, 1992; R. H. Schwartz, 1992)。霍兰德本人已指出研究应针对兴趣调查表的使用和解释, 以及它们对参加测试的人的影响, 而不是将六边形反复地应用于不同的样本 (Holland & Gottfredson, 1992)。

对霍兰德模型进行的最大胆的修正和扩展之一由特雷西和朗茨 (1996) 完成。他们应用了新的和非常复杂的数据分析技术, 包括多维量表法。他们的职业兴趣三维表现使用了霍兰德的分类, 两个维度可能是基础,<sup>[16]</sup> 它是一个扩展的球状模型的出发点, 这个模型能根据测试者的需要容纳各种数量的兴趣类型, 还有一个是威望维度。近期的《职业行为杂志》(Journal of

*Vocational Behavior*) 专门介绍了这一新的模型,并发表了各种研究者对它的评论。对特雷西和朗茨的工作的反应一般是相当积极的,并表明这一理论很可能成为职业理论和测量领域的整合与进一步发展的催化剂(见 Borgen & Donnay, 1996; Gonzalez, 1996; G. D. Gottfredson, 1996; Hansen, 1996; Harmon, 1996; Prediger, 1996)。

对基本兴趣的性质、数量及其组织的探索,类似于为了在能力和人格领域(第十一章和第十三章已讨论过)中确定基本的因素所作的工作。<sup>[17]</sup>在所有的这三种情况下,通过数据分析所揭示的范畴是随着所使用的特定变量和样本而变化的。而且,这些范畴是描述性的而不是解释性的,它们的主要价值在于其简化收集和利用评价和预测行为的信息的能力。

## 意见调查和态度量表

**工具的性质** 态度经常被定义为对特定类型的刺激,如一个国家的或种族的群体、一种习惯或一个机构的赞成或不赞成的反应趋势。显然,态度是隐含的,不能被直接观察而必须从外部行为(言语的和非言语的)进行推测。用较客观的术语,可认为态度的概念意味着关于某类刺激的反应一致性。在实践中,“态度”一词往往与社会刺激和情绪性反应相联系,也常常涉及价值判断。

意见有时不同于态度,但提出的各种区分既不是一致的,在逻辑上也没有充分的理由。这两个术语经常互换着使用,这里的讨论也是如此。然而,有关评定方法,传统上意见调查不同于态度量表。意见调查的特点是关于具体问题的回答,它们不必

相互有关,对这些问题的回答相互独立而不是组合成一个总分。例如,一份雇员意见调查表可以包括关于工作日程、报酬、优惠、公司自助食堂及与上司的关系等问题,因为它们与雇员关系的改进有内在联系,每个问题的回答都分别统计列表,以便确定雇员满意或不满意的根源。<sup>[18]</sup>

另一方面,态度量表一般有一个总分,表示个体对公司、人群、政策或其他刺激类别的态度的取向和强度。在态度量表的编制中,不同的问题旨在测量单一的态度或单维的变量,为了实现这一目标常遵循一些客观的程序。例如,一份雇员态度量表产生的单一分数表明个人对工作满意的程度或对公司的整个态度。

**态度量表的主要类型** 在所有的态度量表中,回答者表明他们对一系列有关态度对象的陈述的同意或不同意。为了达到项目的单维性和同质性、量表单位的等距以及量表之间分数的可比性,研究者设计了专门的方法。编制态度量表中涉及的技术问题已经引起了广泛的注意,而且在方法论上取得了显著的理论的和统计的进展。讨论专门的量表技术已经超出了本书的范围,目前这已形成了统计方法中一个不断发展的领域 ( Jones & Koehly, 1993 ; Mueller, 1986 ; Ostrom, Bond, Krosnick, Sedikides, 1994 ; Procter, 1993 ; Reckase, 1990 ; Young, 1984 ) 。不过,我们可以简要地考察心理测验文献中会遇到的态度量表编制的三种主要方法,这些方法由瑟斯顿(Thurstone)、格特曼(Guttman)和利克特(Likert)量表代表。

瑟斯顿将心理物理学方法应用于判断材料的量化,代表了态度量表编制方面一个重要的里程碑 ( Thurstone, 1959 ; Thurstone & Chave, 1929 ) 。运用这些程序,瑟斯顿和他的合作者编写了约 20 个量表,用于测量对战争、死刑、宗教、爱国主义、书刊审查制



度及许多其他机构、活动、问题及民族或种族群体的态度。瑟斯顿量表首先收集大量的陈述，它们表示对所研究主题的各种态度。要求大量的评判员各自将陈述分成不同的组（通常是 11 组）以表示不同的赞同程度。评判员不表示他们自己的态度，只是将陈述分类。评判员对每个陈述所作判断的中数位置就是该陈述的量表值。判断的变异作为意义不明确的指标，因为不同的评判员把陈述归在不同的类别。项目的选择是为了使变异尽可能的小、有较广的量表值范围，以及在十一点量表范围中间隔近似相等。在最后的态度量表中，陈述以随机顺序呈现，不表明它们的量表值。回答者的分数是他或她认可的所有陈述的量表值中数。

格特曼量表最初是作为决定一系列陈述是否是一种单维的技术而产生的(L. Guttman, 1944, 1947)。在格特曼看来,如果同意一个特定态度的某种陈述的回答者也同意该态度的较温和陈述,则是一个完美的量表。换句话说,这样的态度量表项目能按强度的连续体或接受的难度排列。因此,每个人在量表上的位置将完全决定她或他的回答。如果我们知道一个人会接受的最极端陈述,我们就可以重复她或他的所有回答。实际上,由于每一个回答中测量的误差,这种重复不可能完全达到,它只能在一定的限度内近似。格特曼量表中的基本程序是确定一系列项目,根据回答者的赞同程度排列。不符合这一要求的项目被剔除。一个人在格特曼量表上的分数通过考察她或他所赞同的项目范型来得到。我们可以回忆起,同样的顺序性概念或行为的一致性进步是第三章和第九章讨论的皮亚杰量表的基础。

由于瑟斯顿量表的编制要求相当精致的程序,而格特曼量表的条件在实际中难于满足,利克特(1932)发展了一种量表,它易于编制而有同样满意的信度。利克特型量表以一系列陈述开

始,每一个陈述表达一种或明确赞成或明确不赞成的态度,项目根据测验编制过程中施测对象的回答而进行选择。虽然也采用可利用的外部准则,但项目选择的主要基础是内部一致性。利克特量表要求对每一个陈述进行等级回答,回答通常根据下列5种类型表达:很同意(SA)、同意(A)、不能确定(U)、不同意(D)、很不同意(SD)。为了量表的记分,回答选择从赞成到不赞成记为5、4、3、2、1。例如,对一个喜欢的陈述“很同意”得分为“5”,对一个不赞成的陈述“很不同意”也一样。项目分数的总和代表个人的总分,它必须根据经验建立的常模来解释。

大多数态度量表为特定的研究目的而发展起来。有些是为研究雇员的态度和士气的,其他一些用于评定教育和训练项目的结果。态度量表有助于评价为改变态度而设计的不同的教学程序,或者可以用来测量学生对文学、艺术、不同的种族和文化群体或伴随某一个教育计划而产生的社会和经济问题的态度。态度测量最广泛的应用之一是社会心理学研究。实际上每一本社会心理学教科书都包括态度及其测量部分。通过态度测量进行研究的许多问题中可能会提到态度的群体差异、态度在群体内部关系中的作用、与态度发展相关的背景因素、态度之间的关系(包括因素分析和其他多元分析方法)、态度的趋势和短暂的转化,以及通过插入的经历而发生的态度的实验性改变。<sup>[19]</sup>已发表的态度量表相对很少,虽然大多数在研究文献中有充分的描述。为各种目的而编制的早期态度量表收集在肖和赖特(M. E. Shaw & Wright, 1967)的论著中。有关较近期发展的某些态度和价值的测量,诸如异化和反常、自尊以及控制点,可见于《人格和社会心理态度的测量》(*Measures of Personality and Social Psychological Attitudes*, Robinson, Shaver & Wrightsman, 1991)一书中。

与性别有关的变量和度量的说明 在该书的前面几版中,

已经在某种程度上包括了涉及性别角色和有关概念的测量部分,如男子气、女子气和双重性,而目前的版本没有包括这些内容。虽然旨在评定这些现象的研究和工具不断增多(见 Lenney, 1991),但是这一领域的许多研究者一致认为它还处于概念混乱的状态之中。从与性别相关变量的个体差异的较广阔角度来看,贝茨(Betz, 1995)极好地论述了与这些变量有关的知识的当前状态。她认为缺乏理论框架以及实质性的概念定义已经导致与性别有关现象的研究进展受到限制。关于这一点,应该补充的是,过去 20 年在美国和世界其他地方性别文化观方面史无前例的变化也是有目共睹的。无论如何,正如贝茨所指出的,与性别有关的差异的完整的解释性概念,如性别角色或男子气—女子气,基本上没有获得实验研究结果的支持。根据她的看法,为了进一步的发展,有必要对构念的概念化和定义进行更多的研究工作,这些构念可置于一个有意义的和一致的理论框架中。

## 控 制 点

被描述为“控制点”的构念首先出现在罗特(Rotter, 1966)的一篇论文中。在这篇论文中,罗特发表了他编制的一种量表,用于评定个体对强化的内部—外部控制的一般性期望(I—E 量表)。这一工具是在社会学习理论的背景中编制的,在解释它的用途时,罗特写道:“伴随某些行为的强化效果……不是一个简单的铭记过程,而是依赖于个人是否知觉到他自己的行为 and 奖励之间的因果关系。”(1966, p. 1)内部控制指的是把事件知觉为和个人的行为或个人相对持久的特征相联系的。另一方面,外部控制指的是伴随个体某些行为的积极或消极强化被知觉为并

不完全是和他或她的行为相联系的，而是机遇、命运或运气的结果；或者可能被知觉为在强大的外力控制之下或由于个体周围的力量复杂性而不可预测。

I—E 量表是一种迫选自陈调查表，表 14-1 有两个作为例子的项目，完整的项目单和标准的实施说明书一起，可以在罗特的论文中看到(1966)。与 I—E 量表有关的大量信息出现在最初发表的论文中，包括来自一所大学的几百名男女大学生的百分位常模，以及主要由大学生组成的数十个样本的平均数和标准差。许多其他群体的资料也随后通过一些独立的研究项目而累积起来，关于构想效度的大量资料也已具备。因素分析最初表明，单一的一般因素可以解释大多数反应变异，然而，其他因素分析随后表明，这种构念可以被细分为几个不同的因素，如相信有一个艰难的世界、一个不公平的世界、一个不可预测的世界和一个政治上反应迟钝的世界等 (B. E. Collins, 1974)。后来的几项类似的研究也发现了多因素结构。

表 14-1 来自内部—外部(I—E)量表的两个项目说明

- 
- (a)从长远来看，人们能从世人中获得他们应有的尊敬。
- (b)不幸的是，不论他多努力，个人的价值常常被忽略。
- 
- (a)成为一个成功的人，最重要的是努力工作，运气与它几乎无关或完全无关。
- (b)得到一份好的工作，主要在于把握机遇。
- 

指导语：“这是一个问卷调查，以发现在我们的社会中，某些重要事件影响不同人的方式或途径……就你而言，请你从每对陈述中选择你认为正确的一条(而且只能一条)。”

---

(来自 Rotter, 1966, p. 11. 同意引用)

直到 20 世纪 70 年代中后期,已经确定由 I—E 量表获取的这类控制期望在帮助预测某些行为中起着重要的作用。然而,这一领域的研究者也认识到,为了最大限度地达到预测的准确性,对控制期望的测量有必要针对研究的特定人群和行为领域。从那时以来,已经产生了若干不同的控制点量表。它们有些适用于不同的人群,包括学龄和学前儿童(见 Connell, 1985; Herzberger, Linney, Seidman & Rappaport, 1979; Nowicki & Duke, 1983),另一些其目的在于评定与特定领域有关的控制信念,如婚姻满意感或心理健康状态(D. J. Hill & Bale, 1980; P. C. Miller, Lefcourt & Ware, 1983),还有一些包含关于控制的不同领域(如成就和参与)或不同范围,如个人效能、人际控制及社会政治控制的因果信念(Lefcourt, von Baeyer, Ware & Cox, 1979; Paulhus, 1983)。几个可资利用的量表的复制和基本心理测验数据,可以在勒夫克特(Lefcourt, 1991)的著作中找到。

评定控制点构念的工具的数量和种类证实了它具有持久的生命力,在心理学文献 1984 ~ 1995 年 PsycINFO 数据库中这一主题下的条目超过 5000 条,关于控制期望在有关健康的实践中的作用的研究特别多,另外,控制点一般是动机的一个重要方面,且与人格研究的其他重要领域密切相关,包括归因理论、学习无助以及自我效能感等(Skinner, 1995)。

## 注 释

- [1] 例如,第一批价值调查表之一的价值研究(Alport, Vernon & Lindzey, 1960),主要包括与爱好、信念、兴趣有关的项目,今天看来并不属于价值范畴。现在它已不再印刷。
- [2] 主要或惟一与价值有关的许多工具的描述见安娜斯塔西(Anastasi, 1988b, pp. 560 ~ 583),布雷思韦特和斯科特(Braithwaite & Scott 1991),以

- 及戴维斯(Dawis, 1991, pp. 845 ~ 850)。
- [3] 兴趣调查表和其他职业评定工具的全面指导手册, 现在已是第 3 版, 由美国职业发展协会 (Kapes, Mastie & Whitfield, 1994) 出版。对这一领域重要的理论和方法问题的总结和一些主要工具的例子, 见鲍根 (Borger, 1986) 和汉森 (Hansen, 1990) 的论著。
  - [4] 作为国家教育研究所一个研究的一部分而制定, 塔特尔 (Tittle) 和齐特沃斯克 (Zytowski) 的著作中有复制 (1978, pp. 151 ~ 153)。
  - [5] 这种法律的两个具体例子是 1984 年卡尔·D·伯卡 (Carl. D. Perkins) 的职业教育法令 (P. L. 98 ~ 524) 和 1990 年的卡尔·D·伯肯的职业和应用技术教育法令修正案 (P. L. 101 ~ 392)。
  - [6] 完整的 SH 轮廓图已经重新设计, 现在有 6 页, 每一页的最前面呈现参加测试的个体在所有量表上的得分, 每一页的后面有关于分数的意义的一般信息和如何进行职业探索的建议。除了这份轮廓图, 还可以获得更全面和定性的描述报告, 解释回答者的得分。
  - [7] 这些主题和价值研究 (Allport 等, 1960) 所评定的价值态度在很大程度上相互重叠, 价值研究本身又由斯伯莱格 (Spranger) 在人的类型中提出 (1928)。
  - [8] 用于 SH 1994 年版的所有职业样本的完整目录, 见哈蒙 (Harmon) 等 (1994)。
  - [9] 然而, 与其他迫选测量工具相反, JVIS 的项目配对以产生的分数不是自比性的这种方式完成。
  - [10] 所用程序的完整说明, 见杰克逊 (Jackson, 1977, 第四章) 及杰克逊和威廉姆 (Jackson & Williams, 1975)。
  - [11] 采用的相关是克莱曼斯 (Clemans, 1958) 提出的  $\lambda$  (lambda) 相关系数, 这基本上是一个对不同效标组的同质性差异进行调整的点二列相关, 二分变量包括个体答案纸上做标记与未做标记的回答, 连续的变量是效标组中标记各个回答的人数的比率。
  - [12] CAI 目前有两个版本可用, 职业版 (Vocational Version, VV) 和提高版 (Enhanced Version, EV)。这一部分讨论的只是 VV, EV 虽然在结构上与其非常相似, 但却是完全独立的工具 (Johansson, 1986), 适用于较大

数量和范围的职业，包括许多要求中学后教育的职业。

- [13] 虽然 SDS 的设计是自我记分的，但手册推荐几种监督和检查分数的方法。一项对随机选取的不同年龄的 107 人进行 SDS 目前版本测试的研究表明，7.5% 的人得到的编码包含了不正确的字母或转换（Holland, Powell & Fritzsche, 1994, p. 16）。
- [14] 沃勒（Waller）、利肯（Lykken）和特利根（Tellegen）（1995）报告了关于职业兴趣、休闲时间兴趣及人格特质之间的相互关系的有关结果，用的是明尼苏达双生子登记处的大样本。他们的研究提出人格特质明显是个人职业适合性的中介，三组特点可以作为各不相同的领域加以概括
- [15] 《职业行为杂志》的一个特刊（1992），是这一部分引用的几个参考文献的来源，这一特刊全部都是讨论霍兰德的理论的。
- [16] 这些双极的维度是人／事和数据／思想，由伯里迪格（Prediger, 1982）确定，并被广泛使用。
- [17] 关于兴趣维度的确定和分类以及它们相互关系的说明的历史，见郎兹（Rounds, 1995）的范围广且具有启发性的评论和展望。
- [18] 由费克（Fink, 1995）出版的多卷系列丛书提供了专门针对调查表的详细计划、编写、实施和分析的全面指导
- [19] 涉及态度的社会心理学分支的精辟总结可见艾格利（Eagly）和蔡肯（Chaiken, 1993）。关于态度和态度变化的文献的简要回顾，见奥尔逊（Olson）和赞娜（Zanna, 1993）。

## 第十五章 投射技术

可资利用的投射技术(projective techniques)多而杂。本章讨论这些技术的几个主要种类以及一些人们熟知的例子。除了某些具体技术的特别之处,对个别工具不作评论性的评价。相反,有专门的部分对投射技术作总结性评价,重点强调共同的方法论问题。投射技术的理论和实践之间有难以理解的不一致,当把它们作为心理测量工具加以评价时,其表现多数不佳,但这并不影响它们继续广泛地被应用于临床(Bellak, 1992; Lubin, Larsen & Zalewski, 1993; Watkins, 1991)。这种不一致的性质和意义将在最后部分讨论。

有关投射技术的文献浩如烟海,关于每一个投射工具的参考文献就有 6000 项之多,如想对已有的投射技术有一个更广泛的了解,读者可参考克洛普弗(Klopfer)和道尔比(Taulbee, 1976)拉宾(Rabin, 1981, 1986)、雷诺德(Reynolds)和坎普豪斯(Kamphaus, 1990b, 第三至八章)的论著。心理测量年鉴囊括了对目前大多数投射工具的评论。《人格评定杂志》<sup>[1]</sup>(*Journal of Personality Assessment*)一方面刊用各类测验的文章,同时还刊出大量有关投射技术使用的材料。



## 投射技术的性质

投射技术一个突出的特点是它们指定一个相对非结构性的任务，即这个任务允许作出几乎无限可能的反应。为了使个人展开自由想象，只提供简短的、一般性的指导语；出于同样的原因，测验刺激物是不确定的或模棱两可的。这里的基本假设是，个体觉察、解释测验材料，或者对眼前情形进行‘建构’的方式会反映出他或她心理机能的基本方面。换言之，它期望测验材料作为一种屏幕，被试把他们特有的思想过程、需要、焦虑和矛盾“投射在它上面”。

一般来说，投射工具是经过掩饰的测验方法，使被试难以觉察对他们的反应所做的心理学解释。投射技术的另一个特征是，它是对人格进行评价的综合方法。它主要描述人格整体，而不是测量个别特质。最后，投射技术的拥护者通常认为，它们在揭示人格中隐蔽的、潜在的和无意识的方面特别有效。此外，有人争辩说，测验越是非结构性的，对这种隐蔽内容越敏感。它来自这样一个假设，刺激越是非结构性的或模棱两可的，激起被试防御反应的可能性越小。

投射法源于临床，至今仍然主要是临床医生的工具。有些由针对精神病人的治疗方法（比如艺术疗法）演化而来。在它们的理论框架中，大多数投射技术反映了传统和现代精神分析的观念。也有一些零星的尝试，试图将投射技术建立在刺激反应理论和人格知觉理论基础之上（参见 Lindzey, 1961/1977）。当然，应当指出，没有必要根据理论倾向或特定的历史渊源对某一具体技术进行评价。一个方法证明在实际应用或研究中有效，

并不是因为当初引用它时是如何论证的。

## 墨迹技术

罗夏墨迹技术( The Rorschach ) 最流行的投射技术之一是使用罗夏墨迹 ( Rorschach inkblots, Aronow & Reznikoff, 1983; Aronow, Reznikoff & Moreland, 1994; Erdberg & Exner, 1984; Exner, 1993)。这一技术是由瑞士精神病医生海曼·罗夏( Hermann Rorschach, 1921/1942 ) 创造的。最早的描述见于 1921 年。虽然早就有研究想象和其他心理机能的心理学家使用一系列标准化的墨迹, 但罗夏第一个用它对整体人格进行诊断性研究。

创造此项技术时, 罗夏用大量墨迹进行了实验, 施用于不同的精神病群体。作为临床观察的结果, 能够将不同的精神病症状区分开来的各种反应特征逐渐汇总成为一个记分系统。通过对智力落后和正常人群, 对艺术家、学者及其他特殊人群的补充测验, 使记分方法进一步细化。因此, 罗夏的方法代表了对标准记分的早期非正式的、相对主观的应用。

由于罗夏于 1922 年早逝, 正在进展中的工作由他的同事和学生继续。在随后几十年中, 罗夏技术在欧洲和美国广泛使用。但是, 由于没有人进行统一的系统化, “罗夏墨迹测验” 的实施、记分、解释方法增多并演化成为几种方式或系统。<sup>[2]</sup> 实际上, 到 20 世纪 60 年代, 说它是一个单一的标准化测验已不确切。对不同的系统和使用者, 只有 10 张原始刺激卡片和一些基本的解释性假设共同来自最初的罗夏测验。

每一张罗夏卡片印有类似图 15-1 所示两边对称的墨迹。5 张墨迹只做成黑、灰色, 有 2 张另外添加了少许鲜红色; 剩余

的 3 张由几种淡而柔和的色彩组成。实施罗夏测验时，典型的做法是向被试出示每一个墨迹，一次一张，要他讲出墨迹可能代表什么。除了记录对每张卡片的言语反应，主试通常还记录反应时、反应持续时间、抓取卡片的位置、自言自语、情绪反应以及被试在测试期间其他偶然的行爲。在呈现 10 张卡片之后的某一时间，多数主试都对被试个人就每一墨迹进行联想的部分和方面进行系统提问。在提问期间，被试有机会对先前的反应澄清或细化。

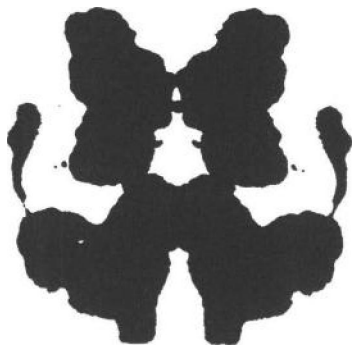


图 15-1 罗夏技术使用的一幅墨迹

盛行于 30 ~ 60 年代的不同的罗夏系统,主要区别在于记分方法以及相应的解释不同。在根本上,罗夏测验的解释主要关心反应的内容或者它们形式上的特点,例如位置、决定因素、形式的特质,以及反应的各种量的概括。虽然不同罗夏测验系统对卡片的反应进行记分和解释的细节有巨大差异,但是各种记分的基本类别却是相同的。位置指的是每次引发被试联想的墨迹的相应部位。他或她是使用整个墨迹、常见的细节、不常见的

细节、空白处,还是利用其中几个部分的组合 反应的决定因素包括形式、颜色、阴影和运动。墨迹本身当然没有运动,但被试把墨迹理解为代表运动的物体,就计入这一类。这些类别之内又有进一步区分。例如,人的运动、动物运动、抽象的运动或无生命物的运动,都分别记分。与此类似,阴影可能被理解为代表深度、质地、模糊不清的云状物,或照片上的非彩色。反应形式的特质和水平,指所作出的反应与使用的位置相匹配的准确性、反应的新颖性,或二者兼顾。另外,有些系统还对反应的认知复杂程度和知觉的其他特点分别记分。

虽然记分通常采用某几个种类,但不同的罗夏系统对内容的处理各不相同。主要种类包括人像、人像细节(即人像的某些部分)、动物形象和动物形象的细节。其他更宽泛的种类还包括艺术品、植物、地图、云、血、X 射线、衣服、性印象和风景。根据一般人各种反应的相对频率,通常可以找到一个普遍分数。对每张卡片的某些反应记普遍分,是因为这些反应是常见的。另外,大多数系统还记录被试在完成罗夏测验过程中的异常言语。这种言语对检测某些严重的病态心理特别有用。

一般而言,对罗夏反应作进一步分析的根据,除了不同种类的反应比例和相互关系,还有不同种类反应的相对频率。通常对罗夏反应作定性解释的例子包括“整体”反应与理性思维的联想,颜色反应与情绪的联想;“人的运动”反应与想象和幻想的联想。通常使用罗夏测验时,着重点放在对个体作最后总的描述,临床医生在描述中将原始记录中各部分的结果进行整合,同时考虑不同分数和指标之间的相互关系。在实际应用中,这些描述还使用其他测验、谈话、病史记录等外部资料信息。

埃克斯纳综合系统到 20 世纪 60 年代,作为心理测量工具的罗夏测验已经声名狼藉。这一方法本身固有的缺陷使研究

者们进退两难,比如,除了记分系统繁杂以外,还有总的反应次数变化不定,主试效应产生的影响,以及不同分数间的相互依赖。这种情形使罗夏技术的信度和效度的研究变成零打碎敲的工作,方法漏洞百出,结果令人失望。虽然很多临床心理学家仍然经常使用罗夏测验,但他们大都承认自己并不完全遵循某一个系统。相反,他们倾向于以自己喜欢的的方式使用罗夏测验资料,从对资料进行印象性的定性说明,到视实际情况而或多或少地与某个或几个系统保持一致。

在美国使用的 5 个主要罗夏系统的巨大差异由约翰·埃克斯纳(John E. Exner, Jr)进行了说明。他曾经与萨缪尔·贝克(Samuel Beck)和布鲁诺·克洛普弗(Bruno Klopfer)一起工作,后两者对罗夏测验所作的系统化工作差别最大。<sup>[3]</sup>虽然埃克斯纳对罗夏测验的临床使用和研究文献作了广泛研究,但是他的兴趣在于是否能够把这一方法所具有的在实证上站得住脚而且有用的特征精炼成一个单独的系统。在过去的 1/4 世纪中,他作出了最雄心勃勃和富有成效的努力,为罗夏测验奠定了良好的心理测量学基础(Exner, 1974, 1991, 1993, 1995; Exner & Weiner, 1995)。首先,埃克斯纳研究出一个综合的罗夏测验系统,它综合了从 5 个主要方法中挑选出的因素。在这个综合系统中,标准化的实施、记分、解释方法是在各种实际应用中通过实证比较后挑选出来的。它的重点在于结构变量而非内容变量。实际上,根据埃克斯纳的意见,对反应进行记分的目的是获得结构性概括。它是系统的核心,并且是大多数解释性假设的基础。每个反应根据几个不同的记分种类进行编码,包括位置、决定因素、形式的特性、内容、有组织的活动、普遍性等。将编码后的反应列表,统计代码的频率。利用这些因素计算比例和百分比,以及组成结构性概括的各种指标。参照整个罗夏测验记录,根据

不同复杂水平上的变量得出解释性的陈述。

有些假设只涉及简单的频率，比如在多大程度上使用某个决定因素(如阴影)另一些假设则涉及两种或两种以上变量同时出现的情况，例如人和动物内容的数量。最复杂水平的分析包括几个变量的群和由实证研究得到的临界分数。这些变量分成几组指标(例如精神分裂指标、抑郁指标和应付缺失指标)，并假设它们可能反映已经发生的某些精神失调。<sup>[4]</sup>

埃克斯纳及其同事使用这个经过 20 年发展完善起来的统一系统，收集了大量心理测量资料，除了各种精神病样本以外，还包括成人、儿童、青少年的常模。根据几天到三年等不同时距的再测信度研究，说明多数记分的变量在时间上相当稳定。综合系统的记分指导有详尽的描述，使得经过训练的主试之间的评分有较高的一致性。实际上，埃克斯纳工作的主要贡献之一就是提供了一个统一的罗夏测验系统，使不同研究者的研究结果有了可比性。<sup>[5]</sup>因此，综合系统成为最经常讲授的一种罗夏测验的记分和解释方法，并被证明对提高罗夏测验研究的统计学效力极具价值(见 Acklin, McDowell & Orndoff, 1992)。

尽管埃克斯纳系统在方法上有了显著的改善，但罗夏测验的几个重要问题仍然存在。其中首要和最复杂的是效度问题。关于这个问题的文献众多，但研究结果相互矛盾。罗夏测验和 MMPI 一样，被用作各种目的，其中很多远远超出原作者的意图。这种使用上的多样化更增加了效度研究的复杂性。一般而言，元分析研究已经表明，罗夏测验的聚敛效度指标与 MMPI 得到的指标相似 (Atkinson, Quarerington, Alp & Cyr, 1986; K. C. H. Parker, Hanson & Hunsley, 1988)。另外，埃克斯纳个人对罗夏测验广博的研究，使许多构念通过他的系统进行效度评价时得到强有力的支持，而且可以有效地用来描述人格机能的某些方面。

然而,当对目前的复杂情况进行诊断,或者对未来行为进行预期时,研究结果就很复杂了(Exner, 1996; Weiner, 1994a; Wood, Nezowski & Stejskal, 1996a, 1996b)。

对罗夏测验分数进行解释时,使问题复杂化的一个主要因素是反应总数——称为反应产生力(respons productivity)或者 R。当不同个体或群体的 R 产生巨大差异时,也可能随之产生其他记分类别的差异。因此,某些类别的差异可能仅仅是 R 变化的结果。反应产生率似乎与其他诸如智力水平和教育程度之类的变量相关,这一事实可能更突出了罗夏测验分数这一内在的特点。<sup>[6]</sup>一般而言,罗夏测验的研究者对反应产生力变化带来的问题的重要性和范围看法不一。例如,迈耶(G. J. Meyer, 1992, 1993)要求进一步研究 R 的心理学意义以及对这一变量进行控制或校正的利与弊。其他一些人则认为它的影响早已证明在大多数情况下并不重要(Exner, 1997; Weiner, 1995b)。还有一些人认为 R 的问题需要根据对资料的使用情况、根据研究时资料分布在多大程度上偏离常态来作不同处理(Kinder, 1992; Lipgar, 1992)。

对埃克斯纳综合系统也并非没有批评。很多罗夏测验的使用者反对埃克斯纳方法的非理论化特点和对内容资料利用不足。他们认为后者严重降低了它的临床使用价值。批评还指出,这个系统除了偶有模糊不清和相互矛盾,还过于复杂。再者,埃克斯纳自己的研究还有样本容量小、变量多、缺乏交叉效度分析,无法让公众仔细审查等缺点。<sup>[7]</sup>即使如此,埃克斯纳系统的有效性以及他和同事积累的研究资料给作为心理测量工具

的罗夏测验注入了新的生命力。

其他方法 尽管埃克斯纳综合系统被广泛接受,同时仍有许多与他并存的其他罗夏测验方法。实际上,很多人从他们自

己以及其他观点出发，认为埃克斯纳的工作使人们恢复了对罗夏测验的兴趣。一个可供选择的带有更强临床倾向的方法由阿若诺(Aronow)及其同事作了描述(Aronow & Reznikoff, 1976, 1983; Aronow 等, 1994, 1995)。这个方法实质上把罗夏测验当做一种标准化临床谈话，对人的知觉操作取样。它的重点在于对内容的解释，而不是结构变量或反应的知觉决定因素。但是，人们认为，把现有的内容量表和记分系统作为心理测量工具做个别诊断不可靠。相反，这些作者建议把临床上严格使用罗夏测验作为提高独特地理解个案的手段。他们观察到，因为这种方法在心理治疗过程中有用而受到大多数有经验的临床医生的偏爱。他们的解释首先根据反应内容，再辅之以言语或非言语行为。根据已有的研究和临床经验，阿若诺和同事们编写了一套指导原则，以便更有效和更可靠地作出独特的解释。例如，他们认为，那些非同一般的反应和与墨迹的刺激特点迥异的反应对个案很重要。与此类似，作者们告诫说，不要使用僵化的符号解释系统，也不要赋予某类内容固定的意义或给罗夏墨迹一成不变的引发性特点。相反，他们为研究反应的意义提供了一些保守的、与一般心理动力学原则相一致的方法，同时还利用个体自己的经历。

目前另一个可供选择的罗夏测验方法由保罗·莱纳尔(Paul Lerner)所倡导。这一工作与埃克斯纳非理论化态度形成鲜明对比，因为它从20世纪70年代以来就扎根于现代心理动力学理论。埃克斯纳主张罗夏测验是一个测验，投射很少起作用，而莱纳尔却认为它基本上是一种达到个体内心世界的投射方法。<sup>[8]</sup>莱纳尔的阐述为罗夏测验的临床使用，为研究客体表征、防御策略的评定和其他现代心理动力学理论的核心概念提供了指导。

罗夏测验的一个具体的临床应用在一致性罗夏测验(the



consensus Rorschach) 中得到说明 (Aronow 等, 1994, 第十三章; Blanchard, 1968; Cutter & Farerow, 1970)。在这个改编本中, 呈现的墨迹由一对夫妻, 或者其他家庭成员、同事、青少年团伙的成员或其他自然群体共同作出解释。通过讨论协商, 参与者必须就一套共同的反应达成一致。这个技术被用来作为研究人际关系和多种其他社会行为的基础, 并取得了明显的成功。

罗夏测验被确切地描述为“垂而不死”的测验 (Peterson, 1994, p. 396)。它已屡次被预言将寿终正寝, 因为像所有那些最常用的心理测验一样, 罗夏测验也成为最滥用的测验之一。目前, 这一技术经历着研究活动和理论建设的巨大复苏。虽然实施罗夏测验的方法仍然不同, 但所有类型的罗夏测验的使用者似乎都确信, 它在研究人格机能作用的知觉、认知和情感方面有特殊的价值。有人满足于简单地把罗夏测验视为一种获得从不同观点都可利用的资料的一种方法, 而另一些人还力图将罗夏测验的不同理论途径, 以及实证与理论的传统结合起来, 形成一个整合的系统 (见 Acklin, 1905; Blatt, 1990; P. M. Lerner, 1994; Meloy & Singer, 1991; Weiner, 1994b; Willock, 1992)。

霍兹曼墨迹技术 (Holtzman Inkblot Technique HIT) 甚至在埃克斯纳开始研究他的综合罗夏系统之前, 韦恩·H·霍兹曼 (Wayne H. Holtzman) 就做了一系列努力使墨迹技术成为具有心理测量取向的技术。由于霍兹曼墨迹技术, 是继罗夏墨迹技术之后产生的, 它的设计就避免了早期工具的主要技术缺陷 (Holtzman, 1961 . 1986; Holtzman, Thorpe, Swartz & Herron, 1961)。然而, 刺激材料和方法有了足够大的变化, 因此应把 HIT 看作一种不同的测验, 并独立于罗夏测验加以评价。霍兹曼技术提供了两套平行的卡片, 各 45 张, 墨迹是从大量预备材料中选出来的, 选择根据是实验得到的标准, 以便使它们的效能最大化。每

张卡片只得到一个反应。有非彩色和彩色两种卡片，有几张卡片是明显不对称的。

HIT 的实施和记分经过了很好的标准化，从一开始测验就有清晰的描述。分数来自 22 个反应变量，包括很多与罗夏测验中相对应的变量，如焦虑和敌意。对每一个变量，都有正常儿童、成人样本和很多种异常群体的常模百分位数（E. F. Hill, 1972; Holtzman, 1975）。评分者信度似乎高得令人满意。虽然大多数结果令人鼓舞，但分半信度、复本信度和再测信度的研究表明反应变量之间有差异。一种使用幻灯片的团体测验在大多数变量上的得分与个别施测的得分接近（Holtzman, Moseley, Teinehr, & Abbott, 1963; Swartz & Holtzman, 1963）。HIT25 是霍兹曼近来研究出的较小的版本（1988），由 HIT 卷的前 25 张卡片组成，每张卡片有两种反应，人们正在制定它的常模（Swartz, 1992）。

目前积累了大量关于 HIT 效度的资料，大多数结果很有前途（Gamble, 1972; Holtzman, 1975, 1986, 1988; Leichsenring, 1991; Sacchi & Richaud de Minzi, 1989; Swartz, 1973）。效度研究的途径很多，包括发展趋势的研究、跨文化比较研究、与其他测验的相关、与人格特征的行为指标的相关、正常被试和精神病人的对照组比较研究。E. F. 希尔（E. f. Hill, 1972）的手册专门指导 HIT 的临床使用。

HIT 显然比罗夏测验有一些心理测量学方面的优势。采用平行卷不仅可以测量再测信度，而且可以做充分的跟踪研究。限制对每张卡片只做一个反应，使反应产生力保持恒定，避免了罗夏记分的很多缺陷。但是，应当指出，反应长度（词数）仍未得到控制，并和罗夏测验一样，它与 HIT 中的几个分数的相关显著（Megargee, 1966）。尽管 HIT 有自己的优势，然而，与罗夏测验相

比,HIT 的资料仍然缺乏,需要更多的资料确定各种分数的诊断意义以及这一技术所评价的人格变量的构想效度(有关评论见 Cundick, 1985 ; Dush, 1985)

## 图片技术

主题统觉测验(Thematic Apperception Test)与墨迹技术相比,主题统觉测验(TAT)提供了更结构化的刺激并要求更复杂和经过意义组织的言语反应主试对反应的解释通常是根据内容分析而不是根据质的特点。TAT 最先由哈佛心理诊所的亨利·墨里(Henry Murray)及其同事研究出来,它不仅在临床实践和研究中广泛应用,还用来作为开发很多其他工具的模型(J. W. Atkinson, 1958 ; Bellak, 1993 ; Dana, 1996b ; R. Harrison, 1965 ; Holmstrom, Silber & Karp, 1990 ; Kloper & Taulbee, 1976, pp. 554 ~ 558 ; Obrzut & Boliek, 1986)。

TAT 材料由 19 张模棱两可的卡片组成,卡片是黑白色的,另有一张是空白卡片。<sup>[9]</sup> 要求被试根据每张图片编一个故事,说出是什么导致图片所示的事件,描述此时正在发生什么和图片中人物的所感、所思,并给出最终结局。对空白卡片,要求被试想象出一幅在卡片上的画并进行描述,讲一个关于它的故事。墨里在测验手册中按最初方法描绘所要求的测验时间为两个阶段各 1 小时,每一阶段完成 10 张卡片。第二阶段故意选择那些更加不常见、更有戏剧性和更古怪的卡片,并伴有指导语促使被试发挥自由想象有部分重合的 4 套卡片各有 20 张,分别适用于男童、女童、14 岁以上男性和女性。大多数临床医生使用经过特别挑选和删减的一套卡片,很少让一个被试做 10 张以上。

在对 TAT 故事进行解释的最初方法中(Murray 等, 1943), 主试首先确定谁是“主人公”, 即假设被试当作自己的那个男性或女性。对故事内容的分析主要参照墨里的“需求”和“压力”项目, 所列出的某些需求在第十三章和爱德华个人偏好测量技术联系起来进行描述。例如成熟、归属和攻击。压力指来自环境中的促进或干扰需要得到满足的力量。遭到攻击或批评、受到关爱、被安慰、身体面临危险等都是压力的事例。评定某一需求或压力对个体的重要性或强度时, 要特别注意这一需要或压力在不同故事中的强度、持续时间和发生频率, 以及和某一图片联系的独特性。假设认为, 不寻常的材料, 即对各张图片的不同寻常的反应, 可能对个体更重要。

关于每张卡片最经常诱发出的反应的特点, 包括对每一张卡片以什么方式理解、发展出的主题、给人物什么角色、表达情绪的基调、反应速度、故事长度等, 已经发表了相当多的常模资料(J. W. Atkinson, 1958; W. E. Henry, 1956; Murstein, 1972)。虽然这些常模资料为解释个人的反应提供了一般框架, 但大多临床医生主要依靠通过自己对测验的经验建立起来的“主观常模”和通过其他手段得到的有关被测者的知识。很多量化记分的方案和评定量表已经研制出来, 且有较好的评分者信度。但是, 因为它们相当费时, 所以这种记分方法在临床实践中很少使用。虽然 TAT 在临床情境中一般作为个别口头测验, 但它也可以在团体测验时书面回答。

TAT 已广泛应用于人格研究。不幸的是, TAT 实施和记分方法, 甚至刺激材料的多样性和 TAT 的规程一起, 已经波及到研究领域和临床实践(Keiser & Prather, 1990)。这种多样化使得对‘TAT 测验’这一独特心理测验的心理测量学特点的研究非常困难, 因为这缺乏最基本的统一性要求。另外, 已有大量实验资

料表明,诸如饥饿、睡眠剥夺、社会挫折等情况显著影响 TAT 反应(J. W. Atkinson, 1958)。TAT 虽然支持投射假设,但这种对暂时状态的敏感性可能使反应的意义复杂化。TAT 反应的内部一致性问题也受到关注(J. W. Atkinson, 1981; Entwisle, 1972),而且应该对故事长度或生产力进行某种控制——这是 TAT 和罗夏测验的共同问题(J. W. Atkinson & Raynor, 1974, 第三章)。

但是,一般来说,主题统觉技术,特别是 TAT 的价值是毫无疑问的。最近的研究证实了各种版本的 TAT 测验的临床使用价值,比如用做评定心理病态的程度和防御机制的使用等传统用途,以及评定问题解决技能之类的新用途(Cramer & Blatt, 1990; Hibbard 等, 1994; Ronan, Colavito, & Hammontree, 1993; Ronan, Date, & Weisbrod, 1995)。TAT 最有前途的应用之一是新研究出的客体关系临床评定量表(Alvarado, 1994; Barends, Westen, Leigh, Silbert, & Byers, 1990; Freedendfeld, Ornduff & Kelsey, 1995; Westen, 1991; Westen, Lohr, Silk, Gold, & Kerber 1990)。TAT 的用途并不囿于对反应作主题分析, TAT 故事的结构和内容形式特征都可以卓有成效地应用于个体或团体研究(见 Cramer, 1996; McGrew & Teglassi, 1990; Teglassi, 1993)。

**TAT 测验的改编本与相关的测验** 已经编制了 TAT 的很多改编本以适应多种特殊用途。它们与原版的 TAT 有程度不同的相似。但是, TAT 的修订版与用 TAT 的一般方法编制的新测验之间有何区别并不明确。有几种版本的 TAT 用来调查对劳工问题、少数人团体、权威等问题的态度(D. T. Campbell, 1950; R. Harrison, 1965)。其他一些改编本也已经编制出来用于职业咨询、行政评价和种类繁多的研究项目。还有一些问卷编制出来用于特殊的人群,包括学前儿童、小学儿童、残疾儿童、青少年和各种民族和种族群体(R. Harrison, 1965)。

有些 TAT 的改编本集中对某一种需要或驱力——性或攻击——进行深入细致的测量。特别重要的是对成就需要（n-Ach）的广泛研究。麦克莱伦（McClelland）、阿特金森（Atkinson）及其同事从事这方面的工作达 30 多年之久（J. E. Atkinson, 1958; J. W. Atkinson & Feather, 1966; J. W. Atkinson & Raynor, 1974; McClelland, 1985; McClelland, Atkinson, Clark & Lowell, 1953/1976）。测量 n-Ach 要使用 4 幅图画, 其中两幅来自 TAT。详细的方案业已研制出来, 以便根据所表达出的 n-Ach 对最后形成的故事进行记分。这项技术已广泛应用于成就动机的研究项目。研究的问题包括从基本的动机理论（J. W. Atkinson & Feather, 1966）到 n-Ach 的社会根源及其结果, 以及它在社会兴衰中的作用（McClelland, 1961/1976）。元分析研究将 TAT 和对 n-Ach 问卷的测量进行比较, 表明两种方法均有效, 尽管测量目的和测量成就动机的侧面均不相同（Spangler, 1992）。

查尔斯·斯密斯（Charles Smith）和约翰·阿特金森（John W. Atkinson）、戴维·C·麦克莱伦（David C. McClelland）、约瑟夫·弗洛夫（Joseph Veroff）共同制订了对言语材料进行内容分析的计分系统大纲（1992）。其中有继承长期研究传统的记分系统（例如成就、归属和权力动机的记分系统），还有许多关于政治思想和应对能力等主题的记分系统。除考虑取样、记分和分析言语材料等方法外，还讨论了理论问题。虽然上述很多系统都是在 TAT 基础上进行的修改，反映了墨里的观点，但它们也代表了几种其他的理论观点。此外，这些系统用来分析思想样本的表现而非象征内容——这与‘投射’相反——同时，它们用于研究而非临床。

据说, 最初的 TAT 可用于小至 4 岁的儿童,<sup>[10]</sup>但是儿童统觉测验 (CAT) 专门用来测 3 ~ 10 岁儿童 (Bellak, 1993)。儿童很

容易投射到动物图画而不是人的图画,根据这一假设,CAT 卡片图片中的人换成了动物。图画中的各种动物被描绘在典型的人类环境中,画成连环漫画和小人书中特有的人形。这些卡片用来引发喂养和其他口唇活动、同胞竞争、亲子关系、攻击、大小便训练和其他儿童期经历事件的想象。CAT 的作者用人的图画对测验进行修订(CAT-H),用于较大儿童,特别是那些心理年龄超过 10 岁的儿童(Bellak & Hurvich,1966)。作者们坚持说,测验的人类试卷和动物试卷都可以更为有效,这取决于儿童的年龄和个性特征(有关的评论见 Hatt,1985;Shaffer,1985)。

一个近期研究出来的罗伯特儿童统觉测验(Roberts Apperception Test for Children, RATC),比其他同类技术更符合测验的编制和评价的心理测量学标准(McArthur & Toberts,1982;也见 Sines,1985)。RATC 提供了两套有重叠的卡片,各 16 张,一套用于男孩,一套用于女孩有一套黑人儿童图片的补充测验,但还未标准化图片的选择是为了诱发熟悉的人际关系情境,包括儿童与成人和其他儿童的关系(图 15-2)。故事的记分依据一系列量表,量表包括被送到诊所就诊的儿童的各种问题。清晰明确的指导使得对反应的记分相当客观;常模是根据 200 个名教提名、适应良好的儿童的反应制定的。这些反应与 200 名儿童里辅导诊所的儿童的反应进行比较,提供了手册中引用的效度资料显然,这个工具表现了一种认真的努力,试图将投射技术的灵活性与标准化测验的实施、记分和测验的评价方法结合起来。对各种用途的 RATC 效度的研究继续显示出有利的结果(Palomares, Crowley, Worchedl, Olsin & Rae, 1991)。另外,格伦·E·罗伯特(Glen E. Roberts, 1994)已经制订出对 RATC 的记分、解释进行详尽指导的临床使用手册。

TEMAS 是西班牙语的“主题”一词,也是“给我讲一个故事”

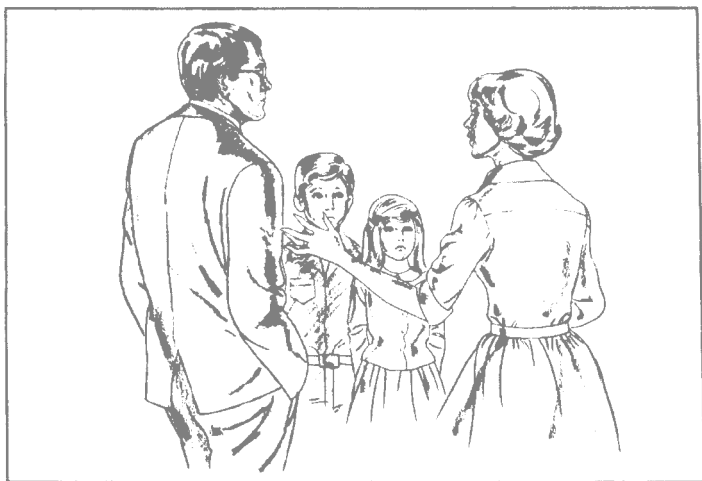


图 15-2 罗伯特儿童统觉测验使用的图画之一  
(西部心理服务公司, 1982 年版权, 经同意复印)

(“Tell-Me-A-Story”) 巧妙的首字母缩略词。它是专门用来评定 5~8 岁儿童的认知、情感和人格特征的工具 (Costantino, Malgady & Togler, 1988)。TEMAS 使用两套平行的彩色刺激卡片, 一套用于少数民族儿童, 一套用于白人儿童。刺激材料经过精心设计以促进言语反应, 激发那些在相互矛盾的目标之间作出选择的故事, 例如立即满足对延迟满足。那套少数民族儿童的卡片描绘的人物特征和深色皮肤使人想起黑人或拉美人血统。虽然 TEMAS 由于适合非洲和拉美裔的美国儿童而被称赞为是对原有 TAT 卡片的显著改进, 但是这个工具的心理测量学特性, 特别是再测信度和内部一致性却一再受到质疑 (有关评论见 Dana, 1993, 第八章; Lang, 1992; Ritzler, 1993a)。



用于老年人的类似的主题统觉测验也已研究出来,包括老年人统觉测验(Wolk & Wolk, 1971)和年长者统觉测验(Bellak, 1993; Bellak & Bellak, 1973) 两者都使用了成套卡片,画有一个或几个年长者,表现老年人可能关心的问题,例如孤独、家庭困难、无助。这两个工具因为发表时不够成熟和使用的卡片倾向于使人对老年人形成不利的刻板印象而受到批评(J. P. Schaie, 1978; K. W. Schaie, 1978)。此外,它们在测验老年人时都没有表现出超越 TAT 的优点(Fitzgerold, Pasewark & Fleisher, 1974; Foote & Kahn, 1979)。老年人统觉测验已不再版。

**罗森茨威格图片一挫折研究 (the Rosenzweig Picture-Frustration Study)** 我们已经考虑过 TAT 及其相关技术,使用图片激发自由想象和详尽的言语反应。相比之下,本部分描述的罗森茨威格图片一挫折研究涉及范围比较有限,要求的反应相对简单。这个工具有用于 14 岁以上的成人(Rosenzweig, 1950, 1978a, 1978b),用于 12~18 岁的青少年(Rosenzweig, 1970, 1976b, 1981a)和用于 4~13 岁的儿童几种形式。根据作者的挫折和攻击理论,P-F 研究呈现一系列卡通画,画中一个人使另一个人受挫,或者使他的注意指向令人沮丧的情形。取自儿童测验的两幅卡通画如图 15-3 所示。在空白处,被试写下受挫人会怎样反应。

P-F 研究参照攻击的类型和方向对反应分类。攻击的类型包括障碍—控制,强调引起挫折的物体自我—防御,重点在于受挫者的自我保护;需要—坚持,重点在于建设性地解决引起挫折的问题。攻击方向的分类有:外向攻击,即指向外界环境;内向攻击,即指向自己内部;不攻击,即试图掩饰或逃避挫折情境。测验记分时每一种反应的百分比都拿来与常模中相应的百分比作比较。也可以得出群体一致性评定,它表明个体反应与

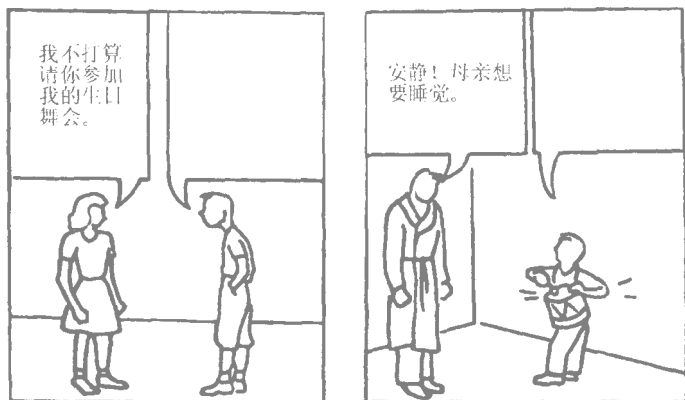


图 15-3 罗森茨威格图片——挫折测验儿童卷中的典型题目  
(沙尔·罗森茨威格, 1976 年版权, 经同意复印)

标准样本众位数反应相一致的反应倾向。

由于 P-F 研究涉及范围有限, 结构性强, 记分方法相对客观, 所以它比多数其他投射技术更适合于统计分析。从一开始, 测验的作者和其他研究者就有系统地努力收集常模资料并做信度和效度检验, 以致 50 多年来, 关于 P-F 研究积累了相当多的成果。这些研究文献涉及这个工具的心理测量学特性和临床诊断、发展变化、性别差异、文化差异、幽默和攻击的关系等 (Grayhill, 1990, 1993 ; Nevo & Nevo, 1983 ; Rosenzweig, 1976a, 1978b, 1978c; Rosenzweig & Adelman, 1977 ; 有关评论见 Viglione, 1985; Wagner, 1985)

## 言语技术

迄今为止,虽然讨论的所有投射工具都要求言语反应,但某些投射技术全部是言语的,刺激材料和反应都只用言语。这些言语技术(verbal techniques)<sup>1</sup>中的一部分可以口头或书面的形式实施,但都适合团体书面施测。当然,团体书面施测的先决条件是被试有最起码的阅读水平和完全熟悉测验的语言。因此,这些要求使此种技术不能用于幼儿、文盲和不会英语的人。

在投射技术蜂拥而至之前的半个多世纪中,有一项技术是词汇联想测验,这一技术起初称做“自由联想测验”,最早由高尔顿(Galton, 1879)作系统描述。随后冯特(Wundt)和 J. M. 卡特尔(J. M. Cattell)将其引入心理学实验室,经改编用于多种用途。测验方法包括简单地呈现一系列无关的词,要求个人对每个词作出反应,说出想到的第一个词。最初的心理测验人员和早期的实验心理学家都把此种联想测验视为探索思维过程的工具。

尽管其他精神病学家——例如克雷佩林(Kraepelin)——早先也研究过这种技术,但词汇联想方法的临床应用主要受到精神分析运动的推动。在精神分析学家当中,荣格(Jung)对词汇联想测验系统发展的贡献最引人注目。荣格(1990)选择刺激词代表普通“情结”;根据反应时、反应内容和情绪紧张的身体表现对反应进行分析。30多年后,梅宁格诊所(Menninger Clinic)的拉帕波特(Rappaport)与同事(1946/1968)研究出类似的联想技术。根据作者的意见,测验有双重目的:帮助检测思维过程的损害;指出重要的冲突领域。作者还提到将词汇联想技术用作“测谎仪”。<sup>[1]</sup>这种应用同样始于荣格,随后在实验室和应用条件下

受到广泛研究(Burt, 1931; Lindsley, 1955)。词汇联想用来检测撒谎或犯罪的基本原理与用来发现情绪冲突的原理很相似。

词汇联想测验的一个不同方法由肯特(Kent)和洛沙诺夫(Rosanoff, 1910)早期的工作得到说明。肯特-洛沙诺夫自由联想测验(Kent-Rosanoff Free Association Test)主要是一个精神病筛选工具,它使用完全客观记分。刺激词是100个普通中性词,选择这些词是因为这些词倾向于激起一般人同样的联想。例如,对桌子这个词,多数人反应“椅子”对黑暗,他们说“光明”。他们设计了一套频率表——每个刺激词一个表——表明在1000个正常成人样本中每一种反应的次数。量表记分时,根据每个被试反应的频率值可得到“共同性指标”(index of commonality)。精神病人与正常个体的比较显示出,精神病人得到的共同性指标低于常人。

但是,由于逐渐认识到反应频率也随年龄、社会经济和教育水平、地区和文化背景、创造性及其他因素而有很大差异,词汇联想技术在诊断上的使用减少了。因此,要恰当地解释结果,就要求随着词汇使用的变化收集亚群体的常模资料,并定期更新。另外,促进这些技术发展的传统精神分析的观念也不那么流行了(Rabin & Zlotogorski, 1981)。无论如何,肯特-洛沙诺夫测验作为标准的实验室工具保住了自己的地位。另外还收集了几个国家的常模,而且此技术已被广泛应用于言语行为和人格的研究(Goldfarb & Halpern, 1984; Isaacs & Chen, 1990; Jenkins & Russell, 1960; Palermo & Jenkins, 1963; Postman & Keppel, 1979; Van der Made-Van Bekkum, 1971)。

另一言语投射技术——句子完成——已广泛应用于研究和临床实践(P. A. Goldberg, 1965; Haak, 1990; D. H. Hart, 1986; Lah, 1989)。在反应长度、结构数量和其他方面,句子完成测验居于

词汇联想和主题统觉测验技术之间一般来说，开放词或句干所允许完成的句子，其变化几乎是无穷的。例如：我的抱负……；女人……；令我担心的是……；我的母亲……。句干的表述方式通常是为了引发与所研究的人格侧面有关的反应。句子完成技术的灵活性是它在临床和研究上的优势之一。尽管如此，已出版了一些标准化的问卷以供更一般的使用。

一个广泛应用的例子是罗特不完全句子填空（Rotter Incomplete Sentences Blank, RISB, Rotter & Rafferty, 1950），由 40 个句干组成，给被试的说明写着“完成这些句子以表达你真实的感情。努力完成每一个句子，一定要写完整的句子”。

根据每一个完成的句子所表明适应或适应不良的程度在七点量表上进行评定使用手册给出了与每一个评定等级相对应的句子，作为例句加以说明。借助这些反应的样本，才有可能使计分相当客观。个人得到的评定分数的总和提供一个总的适应分数，供筛选之用。反应内容还可作临床分析，以便得到更具体的诊断线索。新修订的 RISB 手册包括了增补的常模信息和 20 世纪 50 年代以来所作研究的评论（Rotter, Lah, & Rafferty, 1992）。

其他很多句子完成测验已经编制出来，用于评价不同的目标人群，供各种研究和心理诊断使用（几种传统工具的描述见 D. H. Hart, 1986; Lah, 1989, Ravin & Zlotogorski, 1981）。<sup>[12]</sup>最近这一领域令人感兴趣的新成员包括：用来在残疾检查中检测诈病、预测管理效率和评定防御机制等构念的工具，它们在人格评定中可能有重要意义（Carson & Gilliard, 1993; N. L. Johnson & Gold, 1995; Timmons, Lanyon, Almer & Curran, 1993）。

## 自传体记忆

言语投射技术领域里最新和最有前途的发展之一是对人格评定中使用自传体记忆(autobiographical memories)重新感兴趣。分析记忆,特别是早期记忆,以理解后期生活中经常发生或难以解决的冲突,当然是弗洛伊德时代以来心理动力学的心理治疗中常用的方法。<sup>[13]</sup>另外,弗洛伊德最初的追随者之一、后来建立了自己的个体心理学学派的阿尔弗雷德·阿德勒(Alfred Adler)认为,正是最初的记忆掌握着理解个人“生活风格”的钥匙。结果,阿德勒流派的心理学家自 20 世纪 30 年代后把自传体记忆用作临床工具,偶尔也用于研究(参见 Hafner, Fakouri & Labrentz, 1982; Slavik, 1991)。其他理论家也承认自传体记忆——通常被视为构建或投射而非真实的历史陈述——在人格发展中的核心作用。但一般而言,这一表面上的人格信息资料的来源经过 20 世纪早期的短暂关注之后,直到最近才被系统、持续地研究。

20 世纪 80 年代早期以来,作为心理学认知观点影响的结果,人们重新关注自传体记忆,特别关注它在人格组织中的特殊作用(Bruhn & Last, 1982; Ross, 1991; Rubin, 1986; Singer & Salovey, 1993)。布鲁恩(A. R. Bruhn)在这个领域做了出色的工作(Bruhn, 1984, 1985, 1990a, 1990b)。对以前的弗洛伊德学派、阿德勒学派和自我心理学家用来解释早期记忆的模型进行回顾之后,他提出了更系统的使用它们的方式,还为它们提供了一个新的理论框架。在布鲁恩的认知-知觉理论中,自传体记忆对理解人格极其重要。因此,发展一种收集和解释自传体记忆的标准化方法是布鲁恩优先考虑的事情之一。早期记忆方法(EMP,

Bruhn, 1989, 1992a, 1992b) 是一个自测的纸笔工具。它不仅仅在童年期, 而是从整个人生历程选取 21 个自传体记忆做样本, 第一部分要求 6 个一般的或“自发的”记忆, 主要由时间框架来界定, 即 5 个最早的记忆和一个特别重要的、终生难忘的记忆; 第二部分由 15 个具体的或“有指导”的记忆组成, 考察可能在临床上有关的各种事件或领域(例如, 造成精神创伤的记忆, 第一次受惩罚的记忆, 或最幸福的记忆)。<sup>[14]</sup>除了对每一个记忆的陈述, EMP 还包括各种各样的探测, 涉及记忆的清晰度、情绪色彩、重要性以及其他各种因素。布鲁恩认为对具体事件的早期记忆是一些故事或隐喻, 反映出人从生活经历中有意识地学习或直观感受到的东西。他还相信这些故事常常不准确或被扭曲, 不过他坚持认为, 它们的真实性对临床使用并不重要, 因为像其他投射结果一样, 早期记忆的价值在于它们能够揭示目前的忧虑、态度、信念和情感状态。

虽然布鲁恩和同事们设计并修订了一个早期记忆综合记分系统(Comprehensive Early Memories Scoring System, CEMSS—R, Last & Bruhn, 1991), 但布鲁恩对自传体记忆记分和解释的方法还是相当灵活的。<sup>[15]</sup>他把 EMS 视为复杂的心理现象, 对它进行解释可能需要不同的理论模型, 因此需要不同的记分系统。实际上, 他提倡根据对标准群体早期记忆的一些方面的实证观察, 来制定符合规格要求的系统或“专业化”系统, 以作出明确的预测。关于用来预测少年犯罪和暴力倾向的记分系统, 布鲁恩和他的同事已经获得了很有前途的资料(Davidow & Bruhn, 1990; Tobey & Bruhn, 1992)。

EMP 技术显然仍处在发展之中, 还没有进行标准化。虽然布鲁恩与同事设计的几种记分类别获得了相当水平的评分者一致性, 但是其他几种信度的研究证据还为数不多。无疑的要获

得 EMP 的信度或其他心理测量资料会很困难。像其他投射技术一样，自传体记忆的分类和量化的结果是使可能对理解某一研究对象有独特价值的信息丧失。但是，这个方法有可能成为人格评定的有用工具，特别是在心理治疗领域内（见 Ritzler, 1993b）。此外，在临床上，自传体回忆的系统化样本可能比其他在目的和使用方式上都相类似的言语材料——例如梦的报告、自由谈话或故事——更加重要。<sup>[16]</sup>

## 表演技术

有一类庞大而散乱的投射技术由很多比较自由的自我表现组成。所有这些技术的特点是除了用作诊断方法外，还用作治疗方法。据说，通过这些活动提供的自我表现机会，个体不仅展示出她或他的情绪问题，而且使问题得到缓解。此类方法中最常用的是各种绘画和游戏技术，包括戏剧化地使用玩具。毫不奇怪，这些方法中的绝大多数专门用来评价儿童，尽管很多情况下也可用于成人。

绘画技术(Drawing Techniques) 虽然几乎每一种艺术媒介、技术和材料都曾被研究过，以探求人格评定中的诊断线索，但人们特别关注人像绘画。<sup>[17]</sup> 一个众所周知的早期例子是马乔弗画人测验 (Machover Draw-a-Person Test, D - A - P, Machover, 1949) 在这个测验里，给个体提供纸和笔并告诉他“画一个人”。完成第一幅画后，要求他或她画一个与第一个人物性别不同的人。被试绘画的同时，主试记录他或她的评论、画不同部位的顺序和其他步骤上的细节。绘画之后通常提出一系列问题，使他或她说出所画人物的年龄、受教育程度、职业和有关其他事



实。询问还可能包括要求被试对画中的每一个人讲一个故事。

马乔弗尔对 D-A-P 的解释实质上是定性分析,充满着根据单个指标进行的广泛概括,例如“不成比例的大脑袋常常出自患器质性脑病变的个体之手”。最初发表的测验报告虽然参考了“临床情境下的成千上万幅绘画”,也引用了一些实例作说明,但是没有系统的材料。另外,其他研究者随后的实证性研究普遍不支持马乔弗尔的诊断性解释(见 Klopfer & Taulbee, 1976, pp 558 ~ 561)。

使用儿童青少年人像绘画(HFDs)的另一个方法——建立在更坚实的实证基础上的方法——是由考比兹(Koppitz, 1968, 1984)设计的。由于考比兹一直相信 HFDs 在评定儿童中的临床效用,因此她根据 1856 名 5 ~ 12 岁公立学校儿童的绘画研究出两个客观的记分系统,并使之标准化。其中一个系统主要建立在古德纳夫-哈里斯绘画测验(Goodenough-Harris Drawing Test, 见第九章)和考比兹自己的临床经验基础之上,把 HFDs 用作心理成熟的发展测验。第二个系统起源于马乔弗尔和其他一些人的工作,是对儿童人际态度和忧虑的投射测验。它由 30 个“情绪指标”组成,用于区分有情绪问题和没有情绪问题的儿童所画的图画。这些指标很少发生在样本中的正常儿童身上。与发展指标不一样,它们被认为与年龄和成熟水平无关。它们包括:

- (1) 表示特点的迹象,例如透明度和脸部的明暗阴影;
- (2) 特殊特征,例如极小的脑袋或怪诞的体形;
- (3) 省略应有的部分,例如没有脖子或眼睛。

HFDs 某些集中的特征,例如“情绪指标”的异常或总数,似乎能将有问题和适应良好的儿童区别开来(D. T. Marsh, Linberg & Smeltzer, 1991; Naglieri & Pfeiffer, 1992; Yama, 1990)。但是,考比兹和其他研究者都警告说,不要用单一指标或“迹象”来诊

断。关于 HFDs 比较一致的意见好像是它只能对儿童情绪适应水平提供非常一般化的了解。而且,就临床应用而言,大多数专家都赞同,绘画只能用来产生假设,对它们的解释必须在关于个体的其他信息背景中进行 (M. V. Cox, 1993; Knoff, 1993; Tharinger & Stark, 1990)。

尽管有这些警告和限制,人像绘画的流行不仅没有减弱,而且,人们又设计出另外几个绘画任务。使用最广的一个是房子一树一人 (H-T-P) 技术,正像其名称所表明的那样,它要求被试分别完成一座房子、一棵树和一个人的图画 (Buck, 1948, 1992)。绘画本身的特点、特征与绘画任务之后广泛的询问通常都用作冲突和担忧等一般领域提出假设的来源。作为临床工具似乎有独特价值的一项较新技术是动态家庭绘画 (Kinetic Family Drawing, KFD, R. C. Burns, 1982; R. C. Burns & Kaufman, 1970, 1972)。这个测验要求儿童画出他们家庭中的每一个人,包括他们自己“在做什么”。KFD 产生了大量的研究。在最近对这些文献的回顾中,汉德勒和哈贝内希特 (Handler & Habenicht, 1994) 指出,尽管对这个技术所做的研究有方法问题,但仍有很多富有前景的研究结果,以保证用更高级的分析方法,例如多元回归,作进一步的探索。例如, (KFI) 中人物之间的距离和交往的程度是绘画中最具有心理学意义的特征。

更富有想象的任务在继续被设计出来,例如,合作绘画技术。根据字面意思,它要求全家或一对夫妻合作完成一幅绘画,同时有一个或多个治疗专家在面前,密切观察所有参与者的行为 (G. Smith, 1991)。这个互动式技术源于 KFD,主要用于家庭治疗。<sup>[18]</sup>

游戏技术和玩具测验 (Play Techniques and Toy Tests)<sup>[19]</sup>  
各种游戏技术和玩具测验,包括诸如玩偶、玩具娃娃、微型物体

画等物体,都被投射测验广泛采用。这些材料源于儿童游戏治疗,后被改编用于成人和儿童的诊断测验。之所以选择这些材料是因为它们的联想值。对成人和儿童诊断时最常用的物件,有代表男女成人和儿童的玩具娃娃、动物、家具盥洗室和厨房用具及其他家庭用具。戏剧测验可见图 15-4,由一套标准化的此类器材组成,并附有手册。它于 20 世纪 60 年代在瑞士发表,最近在美国也可见到(Staabs,1991)。人们期望用这些物件的游戏揭示同胞对抗、恐惧、攻击、冲突等,还揭示儿童对他或她的家庭的态度。主试记录儿童的言语、情绪表现和其他外显行为,还要记录儿童选择什么物件,对它做什么。

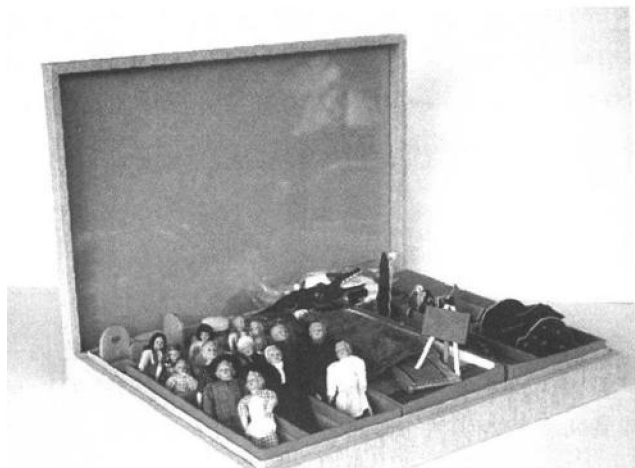


图 15-4 戏剧测验的标准化材料  
(霍格雷非和休柏出版公司,1994 年版权,经同意复印)

对儿童，主试只是提供一组自由游戏的玩具。对成人，与材料同时呈现的还有指导语，要求完成高度非结构化的任务。当然，这些指导语也可用于儿童。通常，任务带有戏剧化特点，就像小型舞台上人物的安排。例如，戏剧测验的材料放在一个扁平手提箱里，有一个盖子可被用作各种人物和附属品的“舞台”（见图 15-4）。

用于儿童诊断和评定的游戏技术已被编入夏弗（Schaefer, Gitlin & Sandgrund, 1991）主编的综合手册里。除了木偶技术这样的投射工具，这个测验还描述了评定具体问题——从孤独症到歇斯底里——和发展性评价，例如能力、权力动机以及婴儿气质等种类繁多的游戏量表。用于游戏治疗和评定亲子交往、同伴交往的量表也包括其中。像夏弗及其同事所承认的那样，他们介绍的很多技术仍处在初级发展阶段。但是，他们收集的各种各样的方法包括了一些巧妙的测验，并提供了高度结构化的观察方法。它们很多是临床工具，而且很适用于研究。

## 对投射技术的评价

显然，投射技术本身存在很大差异。似乎一些比另一些更有希望，因为它们有更有利的实证研究结果，理论取向更加合理，或者两者兼备。关于某些技术，例如罗夏测验，虽然它们的解释仍不确定，但已积累了大量的资料。至于另一些技术，人们所知甚少，因为它们产生不久，或者客观验证受到这些工具自身的性质或它们的倡导者态度的阻碍。

一个有关的观察表明，投射技术和标准化测验并不像第一眼看上去那样有巨大的或根本性的差异。有人已经令人信服地

证明 投射技术与自陈问卷的差异, 不仅在于它们的心理测量学特点, 还在于呈现给被试任务的性质和结果解释的方式上, 而这些都是程度上的差异而不是本质的差异 (Levy, 1963)。各个单独的工具属于一个连续体; 虽然两个极端的差异容易辨别, 但是在中间部分, 一些共同的特征也是显而易见的

要对每一个投射工具作出单独评价, 并试图对有关的大量文献进行概括, 需要单独一本书的篇幅。本章只对表现出独特特征——不管在性质上是赞同的还是不赞同的——工具, 才插入批判性的评论。但是, 有些一般性问题可能在某种程度上针对大多数投射技术提出。为方便起见, 这些问题可以总结的形式予以考虑。

融洽性和适用性 (rapport and applicability) 大多数投射技术为临床医生和病人最初接触时打破僵局提供有效的工具。测验任务本身通常有趣而引人入胜。这有助于把个体的注意力从他或她自身引开, 从此减轻窘迫感和防御性。它对被试的名声极少或不构成威胁, 因为任何反应都是“对的”。

某些投射技术可能特别适用于幼儿、文盲和有语言困难或言语缺陷的人。非言语媒介手段容易应用于所有这些群体。而且从前两类人那里可以获得对图片和其他非言语材料的口头反应。对所有这些受言语局限的群体, 投射技术可以帮助被试与主试交流。这些技术还可以帮助个人澄清以前从没用语言表达过的个人行为的某些方面。

伪装 (faking) 一般而言, 投射工具不如自陈问卷容易伪装。投射技术的目的通常是隐蔽的。<sup>[20]</sup> 即使某人具有心理学经验, 熟悉某个工具的一般性质, 例如罗夏测验或 TAT 测验, 他或她仍然不可能对反应记分和解释的复杂方式作出预见。最近一系列研究比较了对依赖性的客观测量和投射测量, 正如结果表

明的那样，测验的表面效度和它所伪装的可能性似乎成反比关系( Bornstein, Rossner, Hill & Stepanian, 1994)。此外，被试很快就会专注于测验任务，因而不大可能采取人际交往中惯常的伪装和克制。

另一方面，不能假定投射测验完全避免了伪装。使用罗夏测验、TAT 测验和其他投射工具的一些实验表明，当指导被试改变他们的反应以创造好印象或不好的印象时，或者给他们一些陈述，暗示某种类型的反应合乎要求时，的确产生显著的差异(Masling, 1960)。有相当多的实验证明，对投射测验的反应事实上能成功地向‘伪装好’和‘伪装坏’两个方向改变，虽然后者更容易做到。几种测验工具，包括罗夏测验、TAT、罗森茨威格 P-F 测验和句子完成测验，都得到了这样的结果(Albert, Fox & Kahn, 1980; Exner, 1991; Kaplan & Eron, 1965; Meltzoff, 1951; Netter & Viglione, 1994; Perry & Kinder, 1990; Schwartz, Cohen & Pavlik, 1964)。有经验的主试对伪装的迹象很警觉，包括个别反应的性质和反应模式以及与被试其他资料的不一致。

主试和情境变量 很显然，多数投射技术的实施和记分都不够标准化，或者在临床实践中没按标准化方式使用。有证据表明，甚至言语指导的措辞和被试—主试关系的细微差异就能够明显地改变这些测验的成绩(Baughman, 1951; Exner, 1993; Hamilton & Tobertson, 1966; Herron, 1964; Klinger, 1966; Klopfer & Taulbee, 1976)。甚至使用一模一样的指导语时，有的主试可能让人感到令人鼓舞、让人放心，而其他则令人感到威胁，这归因于他们的举止和外貌。这种差异可能影响反应产生力、防御性、刻板模式、想象力和其他表现方面的基本特点。鉴于这些研究结果，实施和测验条件问题比在其他心理测验中更重要。

同样重要的是计分解释缺乏客观性。甚至当研究出客观的

记分系统并根据它记分时，最后对原始资料的评价和整合，通常仍然依靠主试的技巧和临床经验。最麻烦的是，对分数的解释常常是对主试的投射，就像测验刺激对被试的投射，换言之，投射测验反应的最终解释，可能更多地揭示主试的理论倾向、他偏爱的假设和个人特点，而不是被试人格的动力特征。

**常模** 很多投射工具另一个共同的、引人注目的缺陷涉及常模资料。这种资料可能完全缺乏或严重不足，或者常模群体描述不清楚。由于缺乏足够的、客观的常模，临床医生依赖她或他的“一般临床经验”解释投射测验的成绩。但是，这样的参考框架容易被记忆歪曲，它们反映了临床医生的理论倾向、先入之见和其他特点。此外，每个临床医生接触的人也可能范围有限，在教育程度、社会经济水平、性别比率、年龄分布或其他有关的特征等方面不具有代表性。至少在一个方面临床医生的经验几乎肯定会产生误导，因为他或她主要与有问题或病态的个体打交道。因此，临床医生可能对测验中正常人有代表性的反应缺乏足够的第一手了解。埃克斯纳收集的罗夏测验常模就代表了一种对这方面的某些不足进行修正的努力。

对投射测验成绩的解释常常涉及主观或客观的亚群体常模。这种常模可能导致错误的解释，除非亚群体的其他方面相同。例如，如果精神分裂症患者与产生常模的正常人教育程度也不同，观察到的精神病患者和正常人表现的不同可能是不同的教育导致的，而不是精神分裂的结果。将不同的精神病症状相比较，会发现系统误差或恒定误差也在起作用。例如，有证据表明，临床医生在某些种族群体或年轻人中倾向于过多地诊断出精神分裂症，而不是躁狂抑郁症。与此相类似，据报告，转换性障碍在社会经济地位较低的个体中更普遍（美国精神病学会，1994）。

信度<sup>[21]</sup> 鉴于投射测验记分方法的特殊性和常模资料的不足,评分者信度成为考虑的重点。<sup>[22]</sup>对投射技术,恰当地测得评分者信度不仅应该包括比较客观的前期记分,而且应该包括最后的综合解释阶段。例如,仅仅表明主试掌握同一个罗夏测验记分系统,以及他们对反应整体、异常细节或者颜色反应等特点的记分高度一致是不够的。关于像罗夏测验这样的工具,这些原始的量数不能像通常的心理学测验一样根据常模表直接得到解释。对测验进行解释的评分者信度,涉及到不同的主试在多大程度上根据同一个记录认为某个被试具有相同的人格特点。关于投射测验,此类评分者信度的研究很少。有的研究者揭示出,相当有资格的测验使用者所作的解释也有显著的分歧。这种结果中主要的含糊不清源于不知道解释者技能的作用。评分者信度高低都不能直接概括到其他评分者,因为在不同的研究中评分者各不相同。实际上,适用于罗夏之类测验计算机解释系统广为流传的主要原因之一,是它们的结果在解释水平上是一致的。

在投射测验领域内,试图测量其他几种测验信度的结果甚至更差。计算出的内部一致性系数通常很低。在诸如罗夏测验、TAT 测验和罗森茨威格 P-F 研究之类的测验中,有人认为不同的卡片或项目没有可比性,因而不能用来计算分半信度。实际上,这种工具中的不同项目测量不同的变量。此外,对前后相继题目的反应趋势常常被认为对解释结果很重要。阿特金森(1981;J. W. Atkinson & Birch, 1978, pp. 370 ~ 374)通过计算机模拟证实,当内部一致性很低时(例如 0.70),用类似 TAT 测验的方法,总分仍可能获得很高的构想效度(例如 0.90)。他观察到,个体对连续卡片的反应并非独立的,而是连续的活动,它反映了不同行为倾向相对强度的起伏。活动倾向一旦表现出来就



降低了强度。被试在描述上所花时间的比例，例如，被试在对不同的卡片作出反应时，描述由成就动机驱动的活动所花时间的比例，是随下述这些因素而变化的：对连续卡片反应的累积效应、不同的成就诱因和一张张卡片中其他相互竞争的动机。考虑到反对投射测验中使用内部一致性信度的各种理由，一个解决办法是编制实际上可以比较的平行本，就像霍兹曼墨迹技术所做的那样。

再测信度也有特殊的问题。时距长，可能会发生人格的真正变化，测验应该检测到这些变化。时距短，再测显示出的可能只是对初测反应的回忆。当为了确定同一主题是否再出现，研究者告诉被试在再测中要写出不同的 TAT 故事时，多数变量两次测量的相关显著。另外一点很值得注意，从投射技术得到的很多分数所依据的反应样本很不充分。至于罗夏测验，举例来说，在某个人的原始答案中属于动物运动、人的运动、阴影、颜色、异常细节之类的反应可能很少，以至于产生的指标极不可靠。在这种情况下，大量的偶然变化都在预料之中。根据这种不可靠的度量计算出的比率和百分比，比个别的度量本身更不稳定(Cronbach, 1949, pp. 411 ~ 422)。

**效度** 对任何测验，最根本的问题是效度问题。很多投射测验的效度研究涉及同时效标关联效度。其中大多数研究将不同组的成绩作对比，例如职业群体和诊断群体。但是，正像在讨论常模时指出的那样，这些群体在其他方面，例如年龄或教育，也有差异。其他对同时效度的研究实质上使用了匹配技术，将同一个人来自测验记录的人格描述与病史、精神病临床谈话或长期行为记录等描述或资料作比较。有几个研究以特殊训练、工作业绩和对心理治疗的反应方面的成功为效标，对预测效度作了研究。一个日益增长的趋势是，通过检验每个测验的使用

和解释的具体假设，来研究投射工具的构想效度。

由于实验控制或统计分析方法上的缺陷，或两者都有缺陷，已发表的投射技术的效度研究大多数都没有确定结果。有些方法上的缺陷可能导致产生子虚乌有的效度。效标或测验资料的污染就是一个例子。因此，效标评定者可能早已对被试的测验成绩有所了解。与此类似，主试可能从测验实施过程中与被试的谈话，或者从个案历史中，以及其他非测验原始资料中获得关于被试特点的线索。对效度研究中的后一类污染习惯上的控制方法是采用盲分析法，由那些与被试没有接触并且不了解她或他的记分者，而不是参与施测的记分者，来解释测验记录。但是，临床医生争辩说，盲分析是解释投射测验反应的非自然的方式，与临床实践中使用这些工具的方式不相符。

虚假效度资料另一个一般来源是未进行交叉效度分析(Kinslinger, 1966)。由于从大多数投射测验可以得到大量潜在的迹象或可以记分的成分，因此很容易只是由于偶然而发现一系列这种迹象，它们在标准群体之间有显著差异。然而，当应用于新样本时，这种记分方法的效度将降到零。

一种更微妙的误差源于刻板印象的准确性程度。某些描述性陈述，例如在罗夏原始记录中可能出现的陈述，也可以广泛地应用于一般人、年轻人、住院病人或者某一研究选取的任何一类人。<sup>[23]</sup>效标和测验资料在这种陈述上的一致会产生效度的虚假印象。有必要对这个误差进行某种检查，例如在同一类别内对一个被试的测验评定和对另一个被试的效标评定的一致性度量。这个度量会表明在某种研究条件下由刻板印象的准确性带来的虚假一致的程度（见 L. H. Silverman, 1959）。

还有另一个常见的误差来源，产生于确定诊断迹象时对临床经验的依赖，查普曼(Chapman, 1967)称之为“错觉效度分析”

(illusory validation)。此现象可以部分解释为什么一些工具和诊断系统实证研究效度结果主要是否定的，但仍在继续使用。在验证这一现象的一系列经典实验中，查普曼和查普曼（1967）呈现给大学生一系列与马乔弗画人测验（D-A-P）类似的人像绘画。结果表明，参加实验者根据预先建立起来的流行的刻板印象作出反应，即使在实验的“经验学习”期间呈现给他们的材料不支持这些联想也是如此。例如，他们把不规则的眼睛与多疑联系起来；大脑袋与智力联系起来；宽肩膀与男子汉气概联系起来。不仅这些解释与参加实验者“学习”过的经验无关，而且在其他实验研究中，即使在精心训练、旨在建立相反联想的情况下，这些刻板文化联想也表现出对变化的抗拒。换言之，即使面对互为矛盾的观察资料，人们仍坚持先前的期望。

错觉效度分析是迷信残存机制的一个具体例子。我们倾向于注意和回忆期望的东西；倾向于忽略和遗忘与期望相反的内容。在强烈执著于某一特定诊断系统的临床医生的临床观察过程中，这个机制可能在实际上干扰发现和使用有效的诊断迹象。查普曼和查普曼最初用 D-A-P 的研究证实了罗夏测验和不完全句子填空测验所作的类似研究（Chapman & Chapman, 1969; Golding & Rorer, 1972; Starr & Karkin, 1969）。

另一方面应该注意，实验设计的不当可能有相反的低估诊断工具效度的结果。例如，人们都认识到，传统的精神病类别，像精神分裂症或人格分裂，是对病人实际表现出的障碍的粗略分类。因此，如果这种诊断类别用作检查人格测验效度的惟一标准，则否定的结果不能令人信服。同理，不能预测职业效标，可能只是因为主试忽略了所研究的工作要求的特质。当采用这种效标时，有可能投射测验对它所要测的人格特质是有效的，但是这些特质与在所选择的效标情境中的成功无关。

越来越多的测验使用者强调人格评价中的全面性和综合性原则，例如使用结构模式( configurad patterns )和情境变量。其中很多情境变量在使投射测验中对单个指标、单独分数或诊断“迹象”的效度分析中起到重要作用。相关不显著可能是由于没有考虑人格变量关系的复杂模式，这可以由临床医生认为最有用的许多投射技术存在大量矛盾的结果加以说明。例如，像 TAT 所揭示的那样，假设中的幻想攻击和外显攻击行为之间的关系，就不是简单的关系。高幻想攻击可能与高或低的攻击行为都有关系，这取决于其他并存的人格特征，例如对惩罚的焦虑或恐惧水平( R. Harrison, 1965; Massen & Naylor, 1954)。

与预期相一致，TAT 故事中表现出的攻击和随机样本个体的外显行为缺乏显著相关。因为它们之间的相关对某些人是正的，对其他一些人则是负的。但是，很明显的是，缺乏相关也与这一假设相一致：此测验不能检测攻击倾向。当然，需要更多使用复杂实验设计的研究，以便对每个假设适用的情况进行分析。

**投射假设** 关于投射技术的传统假设是，个人对呈现给他或她的模棱两可的刺激的反应，反映出重要的和相对持久的人格特点。很清楚，投射测验反应能够也的确反映了个体的反应风格和持久的特质，但大量不断增多的研究指出，很多其他因素也影响那些反应。正如再测信度所测出的那样，经常观察到随时间发生的变化。这表明有大量随机误差的作用。几个实验研究表明，诸如饥饿、睡眠剥夺、药物、焦虑、挫折等因素对投射测验反应有影响，这为投射测验反应易受暂时状态的影响提供了更直接的证据。同样还发现，指导语、主试特点和被试对测验情境的理解不同，反应有显著差异。能力因素——特别是言语能力——显然影响大多数投射测验的分数。根据这些研究结果，只有当主试广泛地了解作出反应的环境、被试的能力和实验

景，才能对测验反应作出有意义的解释。

使用非结构性和模棱两可刺激的优势从另一个角度受到质疑 (Epstein, 1966)。这种刺激不仅对被试模棱两可，对主试也是如此；因此，它们增加了对被试反应解释的不确定性。另一方面，使用有结构的刺激，就能够选择与它要测的人格特点有关的刺激，改变刺激的性质以便对特定的人格维度进行充分的考察。这种方法可能比非结构性刺激这种漫无目标的方法更清楚地解释测验成绩。也有证据反对这个普遍的假设，即刺激越没有结构，就越可能引起投射和触及“深层”人格 (Kloper & Taulbee, 1976; Murstein, 1963)。实际上，模棱两可与投射的关系似乎是非线性的，中等程度的模棱两可效果最佳。

认为 TAT 之类的投射技术引起的幻想可揭示隐蔽的动机倾向，这个假设也遭到异议。例如，有一项对 TAT 幻想和相关的外显行为的纵向研究持续了 20 年，结果表明，青春期活动对成人 TAT 的想象性描述的预测比青春期 TAT 想象成人活动的预测好得多 (McClelland, 1966; Skolnick, 1966)。在一个实例中，显示出提高自己的社会地位和能力的倾向的个体在成人时获得较高的成就需要分数；但青春期获得较高成就需要分数的人却没有在后来表现出上述能力和倾向。这类研究结果使投射技术传统的基本原理的关系发生逆转。如果我们不把 TAT 反应看作动机的投射表现，而是看作受到以前行为影响的、个人思想的样本，这些研究结果就可以得到解释。如果取得过较多成就和发展过程中经常接触到追求成就的榜样，则更倾向在模棱两可的图画中知觉到更多成就主题。

总之，很多研究倾向于对投射假设的许多方面提出怀疑。有充分证据表明，个体对非结构性和模棱两可测验刺激的反应可以有許多其他解释。

作为心理测量工具的投射技术 当很多投射工具根据测验标准评价时，就发现明显的不足。在前面关于实施测验和记分方法的标准化、常模不足、信度、效度几个部分概括的资料中，这一点显而易见。发表的研究没有证明 TAT 和 D-A-P 之类的技术有任何效度，这类研究的累积给人深刻的印象，即使扣除了很多研究方法上的不当也是如此。但是经过几十年结果相互矛盾的研究之后，投射技术的使用并没有实质的改变，实际上，这些矛盾还可能正在迅速增加。用一位评论家的话说，“仍然有热情的临床医生和怀疑一切的统计学家”（Adcock, 1965, p. 533）。

如果我们认识到，除了几个例外，投射技术并不是真正的测验，就可以理解这个明显的矛盾。几个显著的例外有霍兹曼墨迹技术、埃克斯纳综合系统、某些 TAT 的改编本、句子完成测验和罗森茨威格 P-F 研究。在本章未讨论的其余投射技术中无疑也可以找到另外几个准测验的例子。然而，即使是这些工具，除了需要界定明确的群体的常模资料，还需要根据多效度资料，以明确测验分数所测构念的性质。因此，尽管这些工具多数比其他投射技术更符合测验标准，它们中的大部分还是不能作为日常使用的工具帮助人们作出决策或对个体进行预测。

作为临床工具的投射技术 大多数投射工具不是被看作或评价为心理测量工具，或者严格意义上的测验，而是更多地被视为临床工具。因此，它们可以作为有经验的临床医生临床诊断时定性分析的辅助手段。它们作为临床工具的价值与临床医生的技巧成正比，因此，进行评价不能脱离使用它们的临床医生。试图根据通常的心理测量学方法评价它们是不合适的。由于同样的原因，使用量化分数的精细记分系统不仅浪费时间而且引起误导。这种记分方法使结果蒙上客观的外衣，而且可能产生不必要的印象，以为某技术可以作为测验使用。对投射技术进

行临床解释，较之对它们进行量化记分，把它们当做客观的心理测量工具进行解释，更有其特殊的价值。

克伦巴赫和格莱塞 (Cronbach & Glesser, 1965) 借用信息论的概念把谈话和投射技术描述为“宽波段”方法。带宽，即覆盖范围的广阔是以降低信息的真实和可靠性为代价的。客观的心理测量学测验一般产生的信息范围狭窄，可靠性高。相比之下，投射和谈话技术提供的信息范围广泛得多，可靠性低。此外，任何一个投射技术提供的资料种类都因人而异。例如，一个人的 TAT 反应可能告诉我们很多关于攻击行为的信息，关于创造性和成就需要的信息则可能没有或很少；另一个人的记录可使我们对她或他的创造性程度和成就需要的强度进行全面评价，却极少揭示其攻击性。各人提供的信息种类缺乏一致性有助于对投射测验反应进行某一人群中某一特质分析时发现的低效度作出解释。

注意到类似的不一致是临床医生解释个体记录的一个特点，这是很有趣的。因此，在一个早期对 TAT 效度的研究中，亨利和法利 (Henry & Farley, 1959, p. 22) 断定：

使用 TAT 的解释没有惟一的正确方法。不同评分者之间项目一致性很低，但是，每个评分者作出足够‘正确’的判断，产生显著一致的数字。各个评分者可能沿着不同的路线对测验报告作出实质上相同的解释；或者各个评分者在使用 TAT 对不同领域……或对不同被试进行预测的能力可能有个别差异。

使用投射和谈话资料对个别病人作出诊断的临床判断的性质，越来越受到心理学家的注意(见第十七章)。在这个过程中，

对资料进行组织所依据的构念或范畴，是通过考察每个案例资料独特的结合，经过归纳建立起来的。临床医生的特殊作用在于，根据不可能做成统计表格或方程的那些独特或罕见的事件的结合作出预测。临床医生创造新的构念适用于个体，使她或她能在从未遇见过的事件的组合中进行预测。在预测过程中，临床医生也可以考虑类似的事件对不同个体的重要性。假如这种临床预测不是作为最终结论被接受，而是与后来的询问、测验反应、治疗的反应，或者病人其他行为的信息进行不断的对照检验，它也是有帮助的。由于谈话和投射技术的性质，因此不能根据得到的单个资料或分数作出判断。这些技术最适合在连续判断中使用，以提出进一步研究的线索或随后对个体进行验证的假设。

## 注 释

- [1] 此刊物几次易名，最初由布鲁诺·克洛普弗(Bruno Klopfer)1936年创刊时名为《罗夏测验研究通讯》(*Rorschach Research Exchange newsletter*)；1950~1963年称为《投射技术杂志》(*Journal of Projective Techniques*)。
- [2] 罗夏测验及其演化的较为完整的历史，见埃克斯纳(Exner, 1969, 1993)。
- [3] 其他三个系统是玛格丽特·赫兹(Marguerite Hertz)、齐各蒙特·皮奥特洛维奇(Zygmunt Piotrowski)与戴维·拉帕波特(David Papaport)和罗易·夏弗(Roy Schafer)的系统。
- [4] 综合系统利用计算机的记分和解释辅助方法已经研究出来，它只要求对反应进行初级编码。
- [5] 但是，怎样最好地确定罗夏测验的评分者一致性这个问题仍未得到解决。对这个问题的讨论和罗夏测验资料评分者一致性的三种不同估计方法的比较，见麦克道尔和艾克林(McDowell & Acklin, 1996)。
- [6] 来自罗夏测验对装病敏感性的文献——与R有关的一个有趣发现是，当要求被试在罗夏测验中假装精神病时，反应产生力就下降(G. G.



Perry & Kinder, 1990)。

- [ 7 ] 对埃克斯纳工作的这种或其他的批评, 见阿若诺、雷兹尼克夫和墨兰 (Aromow, Reznikoff, & Moreland, 1995), 克雷格 (Kleiger, 1992), P. M. 莱纳尔和 W. 佩里 (P. Mlemer & W. Perry, 1993), 维格里奥尼 (Viglione, 1989), 文森特和哈曼 (Vincent & Harman, 1991), 武德、尼兹沃斯基和斯特基思克尔 (Wood, Nezorski & Stejskal 1996a, 1996b)。
- [ 8 ] 这一明显矛盾的原因在于这些作者关于投射性质和被试要完成的罗夏测验任务的看法有分歧。
- [ 9 ] 关于 TAT 渊源的历史陈述, 见 W. G. 摩根 (W. G. Morgan, 1995)。
- [ 10 ] 有关 TAT 临床使用和其他用于儿童青少年的讲故事技术的具体资料, 见特梅·勤西 (Teglass, 1993) 沃切尔和杜普雷 (Worchel & Dupree, 1990)。
- [ 11 ] 词汇联想技术已不再用于测谎。关于“测谎器”或测谎仪使用方法的近期应用的有关讨论见十七章。
- [ 12 ] 在研究中广泛应用的一个工具是华盛顿大学句子完成测验 (Washington University Sentence Completion Test), 它专门用来评定自我概念发展阶段, 在十六章与自我概念的其他测量一起讨论。
- [ 13 ] 人格评定中使用自传体记忆的简单历史, 见布鲁恩 (Bruhn, 1995a)。
- [ 14 ] EMP 指导语还引发“不适当的性经验”和“身体或情感虐待”的回忆; 但是, 被试可能会贴上没有此类经验的标签。这种情况以及 EMP 的纸笔形式, 降低了获得这种涉及心理卫生专业人员和一般公众关心的“暗示”性记忆的可能性。
- [ 15 ] 布鲁恩极力推荐的一种解释技术使用 EM 的摘要——或者简明的概括, 这是 TAT 和有关工具常用的方法。
- [ 16 ] 关于早期记忆的临床价值, 一些有趣的逸事性的例子, 包括尼克松 (Nixon) 和弗洛伊德 (Freud) 的自传作品的自传记忆的比较, 见布鲁恩 (Bruhn, 1995b)。
- [ 17 ] 把人像绘画用作测量认知机能的非言语测量, 在第九章中讨论。
- [ 18 ] 对使用投射绘画的进一步讨论和这部分提到的几个技术的评论, 见卡明斯 (Cummings, 1986), 哈默 (Hammer, 1986), 诺夫 (Knoff, 1990)。

- 
- [19] 从强调精神分析理论和发展前景的观点，对投射游戏的方法和解释的一般讨论，见克拉尔( Krall, 1986)。
- [20] 被试不知道使用这些技术进行评定的性质和目的，这些技术的隐蔽性是一个伦理问题(见十八章)。
- [21] 关于主题统觉测量信度问题的详尽讨论，特别关于动机变量的测定，见史密斯( C. P. Smith, 1992)。
- [22] 意识到实际记分时保证充分客观的重要性，因此自 1991 年起，《人格评定》杂志( *Personality Assessment* ) 要求报告罗夏测验研究的论文作者提供证据，证明所有主要记分种类的两个或更多评分者之间的一致性至少达到 80%。
- [23] 这种普遍适用的陈述的使用是十七章引用的“ Barnum effect ”的一个例子( Dunnette, 1957; Meehl, 1956)。对这一效应的广泛研究见克洛普弗( Klopfer, 1983, pp. 510 ~ 514)。

## 第十六章 其他评定技术

前几章概述的自陈调查表和投射技术,代表了最有名、应用最广泛的有关人格鉴定的工具。然而,还有很多其他手段也正在为此目的而被探究。在这多种方法中也会出现一些最终将拓展新方向的技术。这一章涉及的方法基本上还处于研究阶段,尽管有些在实际应用情境如咨询和组织心理学工作情境中被用作辅助的评定工具。多种多样的方法是由各种具体的技术所表现的,其中一些难以分类,因为它们评定的构念包括能力和人格两个领域。三种主要的类别包括认知风格和人格类型的测量、情境测验以及旨在评定自我概念和个体构念的技术。对这个考察加以进一步的展望,在人格评定方面非测验技术的应用正引起一些关注,包括自然观察、访谈、等级评定,以及对生活史资料的分析。

### 风格和类型的测量

虽然在本书和其他书中,能力测验和人格测验是分开讨论的,但实际上,测验的使用,尤其是对测验分数作解释时,认知和情感两方面是不能截然分开的。<sup>[1]</sup>组成心理测验的行为样本是个体全部行为的一个横切面,因此,它们同时包括个体各方面的

信息。比如,在第十一章,我们回顾了日益一致的证据,表明能力倾向调查不能脱离情绪变量,因为能力测验的成绩显然与成就驱力、持久性、价值观等等均有关系。同样,在第十三章,我们讨论了这么一个观点,个体在跨情境中的行为一致性是各不相同的,这表现在他们怎样觉察情境并把情境加以分类,而这又取决于他们以前的经验和学习。而且,在第三和第四部分的所有章节中我们讨论的一些方法,它们的主要目的在于评定能力或人格,也分别包括情感或认知的作用。

有多种方法可以处理人类行为中内在的多种因素。比如,我们能只检查如焦虑这样的特质和问题解决能力的相关(参见 Zeidner, 1995),也可采用多元技术,如因素分析、多维量表法,来分离一系列行为数据的不同成分(Jones & Sabers, 1992),还可以设计容纳几个变量并以图形来表示它们之间关系的结构图式。由弗农(Vernon)发展的能力组织的层次模式(见第十一章)、霍兰德(Holland)的职业主题的六角形描述(见第十四章)和人际周边模式(参见 Hofstee 等, 1992),是这类在单一行为领域内分类的例子。<sup>[2]</sup>另一种方法是利用类别,这些类别本身是复杂的,包括认知和情感两种成分。情绪能力(Salovey & Mayer, 1990; Mayer & Salovey, 1993)这一构念在最近引起了广泛的注意(参见 Goleman, 1995)。下面两部分讨论的认知风格和人格类型,也说明了这类方法。它们代表了为掌握人类行为模式或构念的本质区别所作的尝试。

**认知风格** 认知风格实质上是指一个人的知觉、记忆、思维、解决问题的偏好和典型的方式(Messick 等, 1976)。它们被认为是广泛的行为风格特性,跨越能力和人格,也表现在许多行为和其他中介环节上。已经积累了关于各种认知风格和相关概念,如学习和思维风格的许多研究文献(Brocklinsky, 1982; Furn-

ham, 1995; Globerson & Zelniker, 1989; Goldstein & Blackman, 1978a, 1978b; Grigorenko & Sternberg, 1995; Jonassen & Grabowski, 1993; Kogan, 1976; Messer, 1976; Sternberg, 1994; Witkin & Goodenough, 1981)。

认知风格分化的主要起源之一可以说是在知觉功能领域。许多实验文献已经证实个体的态度、动机、情绪等性格特征与知觉或认知任务的成绩之间的显著关系。同时也应该认识到,许多投射技术,如著名的罗夏测验,可以基本上被看作是知觉测验(参见 Blatt, 1990)。

在知觉因素分析确定的因素中,被证实的人格研究中特别有效的是闭合的灵活性 (flexibility of closure) (Pemberton, 1952; Thurstone, 1944)。这一因素的一般测验要求识别在分散注意、迷惑人的细节中的图形。图 16-1 说明一个测验的两个项目在这个项目上具有高负荷〔戈特沙尔特 (Gottschaldt) 图形〕。早期的几个研究得出了具有启发性的数据,它们表明知觉因素和人格

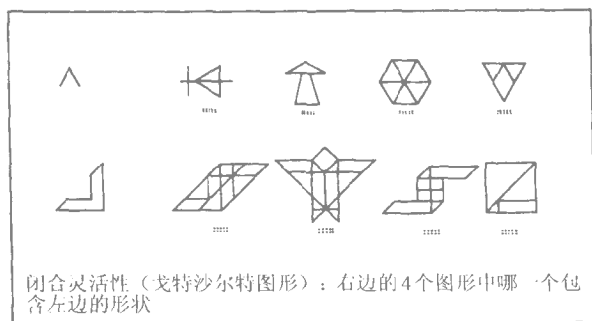


图 16-1 说明运用于人格评定的知觉任务的项目  
(取自瑟斯顿, 1950, p. 7)

特征的可能关系。其中一个调查发现,闭合灵活性优秀的人在有关特征描述上,如社会性退让、独立于其他人的观点、分析的、对理论和科学问题感兴趣以及不喜欢系统化和常规等方面,有高的自我评定(Pemberton, 1952)。在整个中间年份中,研究者已经在认知和非认知行为研究中使用了戈特沙尔特图形的修订本。

从另一角度来看,威特金和他的助手(Witkin 等, 1954/1972)在一项知觉空间定向的长期研究中,确定了抵御相互矛盾的情境线索的破坏性影响的能力是一个重要变量。通过利用可以单独移动的棒体和框架、一把倾斜的椅子以及一间倾斜的房子的各种测验,这些研究者揭示了个体在“场依存性”上的巨大差异,以及他们的垂直知觉受周围视觉环境影响的程度。奇偶分半和重测信度很高,许多不同空间定向测验间的相关都达到显著水平,因此,大量累积起来的数据都表明,场依存性具有相对稳定和一致的特征,具有一定的普遍性。

更有趣的是这些定向测验和包含图形测验(与图 16-1 中的戈特沙尔特图形相似)之间的显著相关,包含图形测验被认为是在纯视觉、纸笔式情境中场依存性的一种测量。随着更多的研究积累,场依存性和场独立性被看成是整体的或清晰的认知风格,或心理分化这样一个较为广泛的人格维度中的知觉成分(Withkin, Dyk, Faterson, Goodenough & Karp, 1962/1974)。有证据表明这些认知风格从孩提时代到成年早期表现出相当的稳定性,并与一些人格变量相关,比如领导能力(Weissenberg & Gruenfeld, 1966)、社会遵从(Witkin 等, 1974)以及其他等等(参见Jonassen & Grabowski, 1993, 第七章)。

场依存性研究的宽广范围和多样性给人以深刻印象,它包括了人际关系(Witkin & Goodenough, 1977)、学习和记忆(D. R. Goodenough, 1976)、数学成就(Vaidya & Chansky, 1980)、在大学和

研究生院学习领域的选择 (Raskin, 1988)、跨文化差异 (J. W. Berry, 1976) 和对工作环境的偏好 (Wooten, Bamer, & Silver, 1994) 等等。对许多研究进行的概括得到了一些令人感兴趣的结果, 其中一个发现场独立者倾向于积极地用“参与者”的方式去学习, 而场依存者倾向于用“旁观者”的方式。最近一个比较多重选择测验和基于操作的评定结果的研究发现, 后者更有利于场独立的学生 (Lu & Suen 1995)。另一方面, 在人际环境中, 场依存者在与他人相处时有一定的优势, 他们比场独立者倾向于更注意社会线索, 对他人的行为作出更多的反应, 情绪也更加外露。似乎场独立—依存连续体两端的任何一端都不是必然或者说总是有利或不利的, 两个方向的偏离值取决于特殊环境的要求。

许多这种研究都采用包含图形测验, 相对来讲这类测验比较容易施测。已出版的测验形式对成人、儿童和学前儿童水平以及团体施测都适用 (Coates, 1972; Witkin, Oltman, Raskin, & Karp, 1971)。团体包含图形测验中的两个示范项目如图 16-2 所示。复杂图形应该是儿童和学前儿童可以辨认的、熟悉的对象, 测验单独实施, 最初的成人测验形式也是单独实施的。

虽然关于这一主题有一系列连续不断的研究 (参见 Bertini, Pizzamiglio & Wapner, 1985), 但许多研究仍然只是澄清在场独立性研究中发现的相互矛盾之处及与各种各样行为变量的关系。其中一个问题与术语有关, 反过来, 这也反映出怎样看待场独立性上的差异。一些研究者根据它与  $g$  因素度量的相关总结出, 场独立构念是一种认知能力, 其他人则把它归类为认知“控制”, 把它放在一个中介的位置, 介于认知能力和认知风格之间 (McKenna 1984; Jonasson & Grabowski, 1993)。对概括的一个相关的障碍是研究缺乏一贯性, 无论是参加者还是方法学。例如, 研

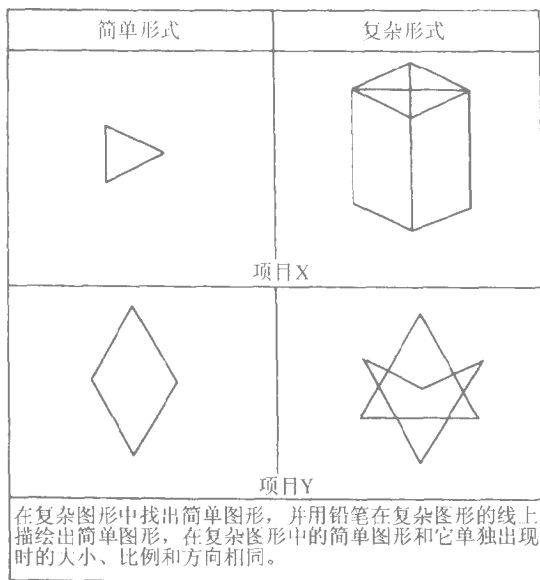


图 16-2 团体包含测验中的两个示范项目

(咨询心理出版社,1972 年版权,经同意复制)

研究表明纸笔式方法的结果，如团体包含图形测验在很大程度上受一般智力的影响，而操作测验如手提式棒框测验（Portable Rod-and-Frame）的分数，完全反映了场独立性的特点与价值无关的风格变量（Arthur & Day,1991）。

虽然对它们的研究可能有概念上和方法学上的障碍，但认知风格有着很大的理论和实践意义。<sup>[3]</sup>因为它们位于能力和人格的交叉点上，这些风格可以影响或调节情绪和智力领域中的行为。因此，有关它们的性质和特征问题，比如在什么程度上它们是固定的还是灵活的，显得相当重要（参见 Niaz,1987）。进一



步说,在学习是首要目标的努力中,人们越来越认识到如果对风格因素也加以考虑,教学和评价的功效也许会有相当大的提高(Furnham, 1995; Jonassen & Grabowsk, 1993; Lu & Suen, 1995; Sternberg, 1998, 1994b; Zelniker, 1989)。

**人格类型** 同认知风格一样,人格类型也指人的构念,被用来解释个体内和个体间思想、知觉和行为偏好方式的相似和差异。人格类型基本上是根据两个或两个以上特征或品质的构念加以确定的类别。至于人类行为的解释,类型学有一漫长的历史,一直可追溯到古希腊的体液理论——血、黑胆汁、黄胆汁、粘液——气质类型和它们有关,分别命名为多血质、抑郁质、胆汁质、粘液质。类型学系统通常得到大众的关注,因为它们为理解并解释一个人和其他人的行为提供相对简单、似乎坚实的基础。这些系统在可以用来解释类型的维度数目和种类上各不相同。

在心理学界,过去几年中提出了许多不同的类型。如谢尔顿(Sheldon)划分人格类型的依据是体格(Sheldon & Sterens, 1942/1970),主要以思辨为基础,由于缺乏确实的证据而被屏弃。其他如在第十三章中已加以描述的 MMPI 代码类型,来自于经验观察,并在不断地发展之中,至今仍在使用(Graham, 1993)大部分类型系统包括单一领域的差异如职业或者气质。少数包括一个以上的领域,其中米勒(Alan Miller)的研究综合了目前存在的涉及认知、情感、动机领域的各种类型。米勒(1991a, 1991b)以评论和分析在这三个领域内的大量类型为基础而把多个系统进行概念上的联结,被证明具有相当的启发性价值。

**梅耶斯—布里格斯类型指标** Myers-Briggs Type Indicator) 最悠久的类型分类之一是由荣格(Jung, 1921/1971)提出的,它已经作为梅耶斯—布里格斯类型指标(MBTI, Myers, 1962; Myers

& McCauley, 1985) 的基础, 而后者是评定正常个体人格的广泛使用的工具。<sup>[4]</sup> MBTI 应用了荣格著名的外倾和内倾态度的二分法( E 和 I) 和他对知觉的两种对立方式的分类( 感觉对直觉, S vs. N) 以及判断的相反方法的分类( 思维对情感, T vs. F)。它也包括对外部世界的偏好倾向( 判断或知觉, J 或 P), 这一点是荣格所没有涉及的。这 4 个可能独立维度上的分数导致 16 种可能的“类型公式”, 它用一个字母组合表示四个维度中每一维度的倾向。如组合‘INTP’表示‘内倾、直觉、思维、知觉’型。<sup>[5]</sup> 每一倾向也有一个数字的分数, 用来表示倾向的强度, 根据每一维度两极间分数的不同来计算。

MBTI 的结果与其他人格调查不同, 主要是为作答者使用, 并用非判断方式来呈现。解释 MBTI 结果的两个最基本的前提是: (1) 所有类型都是有价值和需要的, 各有特定的长处和弱点; (2) 个体在他们所偏好的机能、过程、态度方面显得更有技能。这些特征使得 MBTI 很流行, 应用于各种不同目的, 包括职业指导、咨询、团队选择和发展。已经证明 MBTI 分数造成了心理测量学的困难( 评论见 Decito, 1985; J. S. Wiggins, 1989)。评论者尤其反对梅耶斯—布里格斯的二分迫选, 在那里, 同一剖面图类型的字母根据分数的差异确定, 而这些分数可能在大小上有很大不同。然而, 利用这一工具的大量研究继续在进行。除了对 MBTI 的效度外, 许多研究已经探索了其他可供选择的记分方法( 参见 Davis, Grove & Knowles, 1990; Girelli & Stake, 1993; Harvey & Murry, 1994), 而且, 梅耶斯—布里格斯的变式和用于不同方面的改编本也已被编制出来。其中一个有价值的变式是新近发展的学生风格问卷( Student Styles Questionnaire, Oakland, Glutting & Horton, 1996), 主要用来评价 8 ~ 17 岁学生学习、工作和关系的风格。

尽管梅耶斯—布里格斯和其他一些类型在历史上都很流行,但大部分研究人格的心理学家把它看作是由不同成分、特性组成的精神实体。在此领域,人们之间有着量的差异,而不是一种基于质的差异的分类而且,在科学心理学领域,已经不再愿意把类型当作解释性的概念。这其中有三个主要原因:(1)许多人格理论强调理解和尊重个体独特性;(2)在类型和定型间有隐含的联系,而后者有其内在的危险性;(3)缺少定量分析方法用来鉴定和分析类别数据

然而,一系列的最近发展表明,对类型观念的一些传统的反对意见已受到挑战。例如米尔(Paul Meehl)认为,用经验方法来探索可能存在于非病理性的人格个体差异领域内自然发生而不是任意划分的种类,现在应该是时候了。他和他的助手建议用新的分类方法来鉴定这些种类(Meehl, 1992, 1995; Meehl & Golden, 1982; Meehl & Yonce, 1994)。<sup>[6]</sup>同样,在回顾了从单维的双极特质到多维的周边复杂特性的人格几何形特质模式以后,达尔斯特姆(Dahlstrom, 1995)总结道:“这些模式对于组织人格观念的任务和研究是不合适的。根据他的观点,人格观念和机能的内在的构造性质需要“一个分类架图式(schema of pigeon-holes),以使用可靠和准确的方式来概括我们彼此之间如何相异与相似”(p. 14)。最后,甚至对定型往往是不正确的和破坏性的这样一个已经牢固树立的观点,近来也提出了质疑(Lee, Jussim & McCauley, 1995)。

这些以及其他的动态预示着人们重新愿意把多维分类——如类型和风格——用作研究和解释行为个体差异的理论武器的一部分,也预示着对此会有一个比较成熟的研究途径。然而,尽管这些观念可能是那么有用和受欢迎,但它们仍将无疑地与强调个体和他们行为的独特方面的研究方向共存,而不是代

替后者。在对人们作评价时，把类型这个概念具体化并把它们简单地或过分广泛地当作解释性概念来使用的危险必须牢记在心。

## 情境测验

虽然“情境测验”这一术语在二战期间和战后很流行，但符合这一种类的测验先于那时已得到发展。实际上，情境测验就是把测验者放在类似或模拟“真实生活”的标准情境中。这些测验和应用在编制职业成就测验中的工作样本技术，以及学习成就的绩效评定有一些基本的相似之处，这两者将在第十七章中描述。然而，在目前的测验中，取出的效标行为经常变化且比较复杂，而且，成绩主要按照情绪、人际关系、态度和其他人格变量进行评价，而不是按照能力和知识进行评价。

性格教育调查测验 (Test of the Character Education Inquiry, CEI) 最早的情境测验——虽然在那时并不是这样称呼的——是由哈茨霍恩 (Hartshorne)、梅 (May) 和他们的助手为性格教育调查 (CEI) 编制的测验 (1928, 1929, 1930)。这些测验主要用作研究工具，广泛应用在儿童性格的性质和发展的研究项目中。然而，这些技术也适用于其他测验目的，有一些一直就这么使用。

通常 CEI 技术利用学生日常活动中熟悉的、自然的情境。这些测验以常规的班级测验形式进行，作为学生家庭作业的一部分，也可在体育运动竞赛活动或团体比赛中加以施测。而且，学生们没有意识到他们正在被测验，除非日常的学校考试包括在测验程序中。同时，所有的哈茨霍恩—梅测验是经过精心标

准化的工具，因此得到客观量化的分数

CEI 测验以新的和创造性的方式设计，用来测量这样一些行为特征，如诚实、自我控制和利他主义。最大量的是有关诚实和涉及使儿童相信在不受监视的情形下可以撒谎的情境。例如，在循环迷津（circles puzzle）中，要求儿童闭着眼睛在 10 个不规则排列的小圆圈上做一个标记。在排除偷看条件下的控制测验表明，在总共 3 次试验中正确标记不可能超过 13 分，因此，超过 13 分被作为偷看的证据。

大部分 CEI 测验被证实有很好的区分力，分数有很大的个体差异，信度也显得相当令人满意。然而，儿童的反应带有相当的情境特殊性。每一类中不同测验间的相关显得很低（例如诚实测验或坚持性测验中的各种测验）。当我们考虑到不同情境中儿童的兴趣、价值和动机的作用时，这一特殊性也是可以理解的。例如，一个在学习上有很强动机的儿童，就不一定关注运动比赛和团队比赛中的成就。这些动机的区别会在这些不同背景下进行的诚实测验的行为中反映出来，不论是在他们的发现还是他们的解释中。CEI 研究者强调情境特殊性已有 40 年，然而，后来对 CEI 数据进行的重新分析提示，一个更适宜的模式既包括诚实的一般因素，也包括情境成分，而且还有一些道德特性的一般因素的证据（Burton, 1963; Rushton, 1984）。<sup>[7]</sup>

评价中心的情境测验和角色扮演技术 情境测验是在二战期间由美国战略服务局（OSS）所引入的评价中心项目中的一个重要组成部分。评价中心技术主要包括几天“生活于其中”的阶段，在此期间，候选人接受用各种方式进行的观察和测验。其主要程序表现在挑选执行重大海外任务的军事人员（Murray & Mackinnon, 1946; OSS Assessment Staff, 1948）。相似的程序后来也出现在加利福尼亚大学的人格评定研究所的研究中，以及某些

用于军队和非军队人员的大规模评价方案中。<sup>[8]</sup>后来,评价中心技术和一些情境测验,被一些工业大企业采用,用来评价高层行政人员。

由 OSS 发展出来的一种测验叫情境压力测验,用来对个体在压力、挫折或情绪分裂条件下的行为进行取样调查。例如,被试分配到一项任务,与另两个“帮助者”一起完成,他们其实是障碍和不合作者。另一类型的情境测验,用无领导团体作为评价团队工作、善于应变、首创精神和领导能力这样一些特征的工具。在这些测验中,分派的任务需要一组被测者的合作努力,他们中没有任何一人被认为是领导或给了特殊的责任。这种 OSS 方案的例子包括小河情境——用最快速度和安全方式把人员和设备转移到小河对面;以及过墙情境,在那里人员和物资必须越过一个被虚构大峡谷分开的两堵墙运送过去。

这一技术的一个变式是无领导团体讨论 (the Leaderless Group Discussion, LGD)。由于只要求很少的设备和时间, LGD 已经广泛应用来选拔军队官员、公务员和行政人员、工业经理、管理培训人员、售货培训人员、教师和社会工作者等。基本上要求一个群体在特定时间谈论一个主题,测验者观察和评定每一人员的表现,但并不参与讨论。虽然经常在非正式和非标准化条件下使用,但关于 LGD 仍有很多研究。运用这一技术的研究表明,尤其是当评定者受到很好的训练时,它可以当作一个很好的工具来预测工作成绩。该工作要求口头交流、口头问题解决,以及被同伴所接受 (Bass, 1954; Greenwood & McNamara, 1967; Guilford, 1959; Thornton & Zorich, 1980)。

一些情境测验运用角色扮演或即兴表演激发行为者的兴趣。实际上,前面引用的无领导团体讨论和其他一些情境测验,可以看作是角色扮演的变式和精致化。虽然,角色扮演是应用

在 OSS 评价方案中的一个技术,但它有更早的历史和更广的应用。有关角色扮演或即兴表演技术的历史、原理和变化的全面调查可以在麦克雷诺兹和德沃格 (McReynolds & De Voge, 1978) 的著作中找到。这一技术明确要求个体扮演一个角色,或者外显地(有或没有其他人)或者用语言报告他或她要做(说)些什么。情境可以很真实,像在舞台上,或用录像带、录音带,或书面描述。<sup>[9]</sup>

虽然即兴演出技术的许多应用都是非正式的,并且根据特定情境和当地条件而变化,但仍然盛行不衰,其中一个很重要的应用是人员的职业评定,尤其是当人际行为在工作功效中很重要时(Stricker, 1982; Stricker & Rock, 1990)。一个特殊的例子是咨询员效果的评价。在这种情境中,对一个“受过训练的顾客”进行咨询。这个“顾客”其实是由工作人员或同学扮演的。当人们提出一个事先经过选择的标准化问题时,对未来的咨询员进行观察或录像(参见 Connor, 1994, pp. 72 ~ 75; Kelz, 1966; Neufeldt, Iversen & Juntunen, 1995; A. Williams, 1995)。当使用录像带时,除了受到评定者和同伴的评定之外,受训者还可以监控和评价自己的表现。<sup>[10]</sup>

已经反复证实评价中心技术具有很好的多种效标的预见性(参见 Coulton & Feild, 1995; Howard & Bray, 1988; Ritchie, 1994; Tziner, Ronen & Hachohen, 1993)。它们已经广泛地应用在选择标准很复杂的各个领域,如法律执行工作(J. L. Coleman, 1987; Moore & Unsinger, 1987),甚至人们已经把这些技术加以改编,使之用于聋人工作申请者身上(Berkay, 1993)。因为它们的应用范围很广,我们不能概括出评价中心技术的效度,一个中心与另一个中心的结果不同,这取决于所应用的特殊程序、效标的本质、评价者的特性。一般地讲,当方法最适当时,如在研究中使用几

种方法，包括同伴的评价，并集中于相关的和直接可观察的行为维度时，效度系数最高（Gaugler, Rosenthal, Thornton & Bentson, 1987; Shore, Shore & Thornton, 1992; Thornton & Byham, 1982）。尽管有了广泛的研究，有关评价中心技术的许多问题仍未解决。也许最令人头痛的问题是，有些研究者不能表明采用多种方法评定特定行为时的辐合效度。<sup>[11]</sup>

## 自我概念和人的构念

20 世纪 80、90 年代，人们重新对自我概念和相关构念显示出浓厚的兴趣（Byrne, 1996; Harter, 1990; Hattie, 1992; Markus & Wurf, 1987; Oosterwegel & Oppenheimer, 1993; Wylie, 1989）。<sup>[12]</sup> 目前几个有关人格评定的方案都关注个体看待自己和他人的方式。这些技术反映出现象心理学的影响，现象心理学关注事情是怎样被个体感知的，因此个体的自我描述本身是最重要的，而不被看作是替代其他行为观察的第二位的東西。兴趣也集中在个体表现出来的自我接受程度。

这一节所讨论的程序都集中于个体对自我和他人的知觉。尽管一些工具是已出版的测验，但许多是为具体研究项目而编制的，在已出版的项目报告中可以找到。有些工具很有意思，主要是因为它们与特定的人格理论有关，或与正在进行研究的领域有关，其余是可广泛应用于研究各种问题的技术。

华盛顿大学完成句子测验（Washington University Sentence Completion Test, WUSCT）也许有人会说，自我概念测验实质上与第十三章中已讨论的自陈调查表没有区别。更确切地说，自陈调查表实际上是自我概念的度量。根据自我概念化思想解



释调查反应构成了卢文格 (Loevinger) 系统阐述的人格发展理论观点的基础 (1966a, 1966b, 1976, 1987, 1993; Loevinger & Ossorio, 1958)。卢文格结合自己和其他人研究的不同结果, 提出了一个人格特质, 她把这个特质定义为把自我概念化的能力, 或者同自我和冲动“假设有距离”的能力。按照卢文格的观点, 人格调查表中这一特质的表现形式已经用这样一些术语加以描述, 如表面、参加测验的防御性、反应定势、符合社会期望、默认和个人风格。同其他许多心理学家一样, 卢文格认为这些参加测验的态度不是应排除的测验工具的错误, 而是人格调查中有效方差的主要来源。

根据来自多方面的数据, 卢文格认为形成自我概念的能力随年龄、智力、教育和社会经济地位的提高而提高。在最初——以婴儿为例——个体没有自我概念化的能力, 随着能力的提高孩子逐渐形成定型的、习俗的、为社会所接受的自我概念, 卢文格认为这是青春期的典型特征。随着走向成熟, 个体的进步已超越了定型的概念, 而到达分化的和真实的自我概念。这时, 个体充分意识到他们的特质, 并且接受自己是什么样的。

自我概念形成的这一特性被称为是自我发展或自我水平。卢文格和她的助手用华盛顿完成句子测验测量的就是这一特性 (Loevinger, 1985, 1987; Loevinger & Wessler, 1970; Loevinger, Wessler & Redmore, 1970)。作者的理论框架假设了自我发展的如下 9 个水平: 前社会、冲动、自我保持、遵从、自我知觉、认真、个人主义、自主、整合。除了言语技能出现之前的第一个水平, 都可通过 WUSCT 加以评定。每完成一句句子, 就给予一个自我水平的等级, 全部测验的综合分数根据这些值来计算。WUSCT 是以妇女和女性青少年的研究为基础的, 后来经修改也用于成年男子和男孩这个测验经过修改, 现在有适用于男性和女性的可比

较形式(Loevinger, 1985; Novy, 1992),新的记分手册提供了男性和女性的数据资料(Hy & Loevinger, 1996)。

对 WUSCT 不断的研究已经提供证据支持它的信度 (Novy & Francis, 1992; Weiss, Zillberg & Genevro, 1989) 和作为测量不同样本自我发展的构念效度 (Bushe & Gibbs, 1990; Novy, Gaa, Frankiewicz, Liberman & Amerikaner, 1992; Westenberg & Block, 1993)。这一工具特别富有成果的应用是自我发展性别差异的研究(Cohn, 1991)。与其他完成句子测验和允许开放的言语反应技术一样, WUSCT 存在的几个问题之一就是它们的分数与言语流畅性和词汇水平有关,或在有些情况下受它们影响(参见 Vaillant & McCullough, 1987; Westenberg & Block, 1993)。虽然这种可能性需要在设计 WUSCT 的研究中和该测验的其他应用中加以考虑,但这与其理论基础相一致,并不会使结果无效。

自尊问卷及相关测量 (Self-Esteem Inventories and Related Measures) 在更广范围内的研究中,自我概念的构念与被称作自尊(Baumeister, 1993; Bednar & Peterson, 1995) 的和自我效能感(Bandura, 1982, 1995; Maddux, 1995; Schwarzer, 1992) 的相关构念结合在一起。这一研究主要与个体的自我评价对她或他行为的影响有关。事实上自尊一般被描述为自我概念的评价成分。在长期积累的基础上,这些自我评价也许会影响认知、情感等特质的发展。尤其是大家都公认,自尊是应对能力和幸福感觉等这样一些重要心理变量的决定性因素。

乍一看,自尊的构念容易误解为很简单。它常被假设为对自己的一般评价态度,从极端肯定到极端否定,它是稳定的和完全主观的。用于研究和应用目的的自尊测量,在传统上都是以这些假设为基础的。事实上,布莱斯科维奇和托马卡(Blascovich & Tomaka, 1991)在全面回顾自尊和自我概念的测量后,确定罗

森伯格(Rosenberg, 1965)的自尊量表(SES)——一个有10个项目、具有表面效度的自陈问卷——为这类测量中使用最频繁的工具。

简单明了的综合性的自尊测量,如SES,一直都显得比较稳定,而且可以有效地用作为某些目的服务的一般自我尊重的判断手段。然而,许多研究者发现,如果前者被看成是一个有层次的、多维的构念,并得到相应的评定,那么,自我概念和行为的关系统能够更加清晰地加以说明(Fleming & Courtney, 1984; Marsh, Byrne & Shavelson, 1992; Marsh & Shavelson, 1985; Shavelson & Bolus, 1982; Shavelson, Hubner & Stanton, 1976; Uguroglu & Walberg, 1979)。在某些情况下,使用单一的、综合的自尊测量也许得到前后不一致的结果,或不能揭示与其他变量的显著相关;而一个较狭义的构念,如学术的自我概念,会得到一致的和有意义的结果。对个人价值系统中学术成就并不很高的人来说尤其如此。在这些情况中,倾向于反映反应者自我价值系统的综合性自我概念高,可能与学术成就或智力功能没有显著相关。而且,研究表明,基于自陈的“体验到的自尊”(experienced self-esteem)度和根据他人报告的“呈现出的自尊”(presented self-esteem)度并不总是相关的(Demo, 1985)。

作为这些发现的结果,近年来的研究已经从自尊的单维概念转移到它的一些特殊方面,而且发展得很快。有关儿童和青年的学业的自我概念和学术成就关系的调查特别多。在这一特殊领域,采用复杂方法的研究——如能分析因果关系方向的结构方程建模和纵向设计——支持这样的观点,学业的自我概念与研究的特定领域有关。更重要的是,它们表明,这些概念与以后的学术成就有关,能预测和影响以后的学术成就(Fortier, Vallerand & Guay, 1995; House, 1995; Lyon & MacDonald, 1990; H.

W. Marsh, 1990a, 1990b).

自我概念的最新测量方法,利用了在这一领域累积起来的广泛的理论和经验性文献。例如,学生自我概念量表(SSCS, Gresham, Elliott & Evans-Fernandez 1993),是一个商业上可行的测量,它以班杜拉(Bandura, 1982, 1986)的自我功效理论为起点,同时吸收其他理论和研究成果。它评定自我概念的三个主要方面,分别为学术的、社会的和自我形象的。在每一方面,反应者不仅要指出他们能做条目所呈现的事的自信程度,而且要指出条目对于他们有多重要,以及他们对抱一定态度或去做某事会带来某些结果有多少信心。SSCS有分别为小学和中学水平男女学生使用的常模参照分量表和合成分数。

一个类似的尽管在商业上并不有效的测验是以中学生和大学生为被试编制成的个人和学业自我概念问卷(Personal and Academic Self-Concept Inventory, PAsCI, Fleming & Whalen, 1990)。这一测量试图把夏弗尔森(Shavelson)和他的同事们发展起来的自我概念的等级和多因素模型操作化,并加以进一步研究。

PAsCI是50年代最初用于评定不适当感觉的实验量表的第4次修订本,是最初量表的扩展。<sup>[13]</sup>目前的版本包括一个综合的自尊量表和六个附加因素量表,其中两个用于处理自我概念的社会方面(社会接受和社会焦虑),两个属于身体方面(身体外表和体育能力),两个属于学业方面(数学能力和言语能力)。当然,SSCS和PAsCI都受反应偏向的影响,所有的自陈测量都如此,这在第十三章已讨论过,而且两者都需进一步证明它们在发展过程中的有效性。然而,根据构念分化和内容规范的观点,它们又是自尊概念化和测量方面取得进步的代表。

形容词核查表(the Adjective Check List, ACL)现在已经有几个专门用于评定自我概念的广泛定向的技术。一个广泛应

用并在商业上可行的工具是形容词核查表。它最初是为人格评定研究所(IPAR)的研究目的而编制的。这一工具提供按字母顺序排列的 300 个形容词,从“心不在焉”(absent-minded)到“滑稽的”(zany)(Gough, 1960; Gough & Heilbrun, 1983),要求被测者标出他们认为描写自己的形容词。

在目前的形式中,ACL 分成 37 个量表,其中 4 个评定反应定势。一组主要的量表最初建立在理性或内容的基础上,把各个形容词分配到 EPPS(见第十三章)所包括的默里的 15 个需要中。另外 9 个“主题量表”,主要是通过项目的经验准则确定而编制起来的,以测量在人际行为中被认为是重要的各种特质。剩下的两组量表是为了适应专门的人格理论即伯尔尼(Berne, 1961, 1966)的人际关系分析理论(transactional analysis theory)和威尔士(Welsh, 1975b)的创造性和智力理论而设计的。所有 37 个量表应用在每一量表发展的某一阶段的经验性效度分析数据的主要来源,是直接观察 IPAR 的评价中心项目的参与者和特质的评定。根据 IPAR 评定和其他一些辅助性研究,手册给出了在每一个量表上得高分者和得低分者的人格描述。

作为一个研究工具,ACL 已经应用于来自精神病理学、职业选择、创造力、政治和经济行为,乃至病人对畸齿校正术和对隐形眼镜的反应等各种各样的问题。它也应用于根据自传和已出版著作对历史人物所作的评价(Welsh, 1975a)和对无生命物体特征如城市和汽车的描述中。最近,ACL 已被用于中年妇女发展变化的研究(Helson & Wink, 1992; Wink & Helson, 1993; York & John, 1992)和自恋调查(Wink, 1991, 1992),以及其他研究(有关评论见 Teeter, 1985; Zarske 1985)。<sup>[14]</sup>

**Q 分类** 另一个适合于研究自我概念的特殊技术是 Q 分类。它首先由斯蒂文森(Stephenson, 1953)提出,用于进行称之为

Q 方法的研究 (参见 Kerlinger, 1986, 第三十二章; McKeown & Thomas, 1988)。在 Q 分类技术中, 给反应者一系列包含一些陈述或特质名称的卡片, 要求他们按最能说明自己的特征到最不能说明自己的特征的顺序, 把它们分成几类。这些条目可能来自标准化清单, 但更多的时候, 它们是为符合个别的案例而设计的。为了保证评级分布的一致性, 采用了一个硬性的常态 (forced normal) 分布, 要求反应者把规定数量的卡片放在每一堆中。这样的分布可用于关于某一个常态曲线表的任何规模的项目样本。必须注意, 像第十三章中提到的迫选技术一样, Q 分类得到自比 (ipsative) 指标而不是常模数据, 换句话说, 反应者告诉我们, 他认为哪个是他的强项和哪个是他的弱项, 但这并不表明他认为与其他人或外在的常模相比他有多强。

Q 分类已经被用于研究各类心理问题 (Bem & Funder, 1978; Block, 1961/1978; Kogan & Block, 1991; Ozer, 1993; Rogers & Diamond, 1954)。在个体人格的细致研究中, 反应者经常被要求在不同的参考构架下重新对相同的条目进行分类。例如, 把条目应用于自己和其他人, 如父亲、母亲、丈夫或妻子, 借此进行分类。或者把它们应用于自己所处的不同场合, 如工作、家庭或社会情境而加以分类。Q 分类也可根据个体认为他们实际上如何 (现实的自我)、他们认为别人如何看他们 (社会的自我)、他们想成为怎样的人 (理想的自我) 来进行。为了观察变化, Q 分类还能在心理治疗的不同阶段连续进行, 尤其是某些以患者为中心的治疗人员所采用的治疗程序中。随着治疗的推进, 自我概念变得较为良好, 也更接近个体理想的自我概念 (Rogers & Diamond, 1954, 第四章)。<sup>[15]</sup>

语义区分 (the Semantic Differential) 这一技术最初是由奥斯古德和他的助手 (Osgood, Suci & Tannenbaum, 1957) 作为意

义心理学的研究工具而发展起来的,随后它能用于人格评定的可能性很快被人们认识到。语义区分是一种标准化的和定量的程序,用来测量个体某一概念的理解。每一概念要在七点量表上加以评定等级,如图 16-3 所示。每一概念应用一系列的双极形容词量表,一般包括 15 或 15 个以上。由奥古德发展的最初 50 个量表的相关和因素分析,揭示出三个主要因素:(1)评价,在好—坏、有价值—无价值的、清洁—肮脏的这些量表上具有高负荷;(2)力量,从强—弱、大—小、重—轻这些量表中得出;(3)积极性,从主动—被动、快—慢、敏锐—迟钝这些量表中得出。而评价因素是最显著的,能说明总方差中的最大部分

| 父 亲  |                                                              |
|------|--------------------------------------------------------------|
| 好的   | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 坏的   |
| 清洁的  | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 肮脏的  |
| 冷酷的  | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 友好的  |
| 慢的   | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 快的   |
| 有价值的 | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 无价值的 |
| 紧张的  | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 放松的  |
| 强壮的  | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 瘦弱的  |
| 大的   | _____ : _____ : _____ : _____ : _____ : _____ : _____ : 小的   |

图 16-3 语义区分技术说明。评定‘父亲’这一概念,反应者在每量表适当的线段上做 | 记号,通常采用大得多的量表

语义区分量表反应的分析有多种方法对于定量处理,每一量表的评级都可给予一个数值,从 1 到 7 或从 -3 到 +3。对一个个体或一个群体来说,任何两个概念的相似性,可以根据它

们在所有量表中的位置而测得。由某一个体评定的所有概念的内涵可以通过计算每一概念三个主要因素的分数而得到了解。因此,在从 -3 到 +3 的量表上,对于‘我兄弟’这一概念也许会有这样的评定 价值 -2 力量 0.1 积极性 2.7。

被评定的概念可选择与正在被研究的问题相对应。可要求反应者评定人物,如他们自己、他们的家庭成员、朋友、雇员、教师或公众人物,不同种族和不同文化群体的成员,从事不同职业的人 活动,如学习和户外活动 抽象概念,如迷惑、憎恶、恶心、和平、爱;商品名称或商标名称;录音机或电视节目。语义区分已应用于许多不同的领域,在各种问题的研究中,如临床诊断和治疗、职业选择、文化差异和顾客对产品与商标名称的反应 (Sinder & Osgood),而且,这一技术本身的进一步精致化工作也在不断地进行(参见 Cogliser & Schriesheim, 1994),有关语义区分量表的参考文献目录超过 2000 种。

**角色构念项目测验 (Role Construct Repertory Test)** 专门用作临床实践的辅助技术,角色构念项目测验 (Rep)是由凯利 (G. A. Kelly, 1955, 1963, 1970)编制的。这一测验的发展与凯利的人格理论密切相关。这一理论的一个基本命题是,个体用来感知事物的概念或构念,影响她或他的行为。在心理治疗过程中,经常需要建立一些新构念,去除一些旧构念,从而得到一定的进步。

Rep 测验旨在帮助临床人员鉴别患者有关人的一些重要构念。虽然测验可以以多种方式实施,包括团体和个体的方式,但它总是用这种或那种方法对刺激进行分类。Rep 测验得到的数据可排列成一个矩阵或栅格,可以评定构念的相互关系。比较简单的一种变式可以作为例子说明其基本特点。<sup>[16]</sup>首先给反应者一份角色名称清单,要求根据她或他的经验来提出一个人。



这个人符合该角色名称典型的角色包括：你父亲、你妻子或目前的女友、你喜欢的一位老师、最近和你有密切关系但不喜欢你的人。主试接着选择 3 个被提出的人，并问：“在哪此重要方面其中两个人比较相像而与第三人不同？”这一程序随 3 个名称的不同组合而重复出现，其中一些名称会重复出现在不同的组合中。一旦有关构念的原始数据以这样的方式产生并被组织成栅格形式，它们的含义就可加以分析。虽然 Rep 测验得到许多定性的数据，但也发展出许多定量的程序，以确定诱发出的构念对这个个体的重要程度。这些程序从简单的描述统计到高度复杂的构念分析，软件可用于帮助口头记录的内容和构念分析（参见 Winter, 1992）。

Rep 测验的多种修正形式，已经应用于有关问题的大量研究。这些问题有人格理论、社会认知、教育、交往以及心理治疗和评定。由被试对熟人的分类得到的指标中，有一种称之为认知复杂性。这一指标是以个体所应用的不同构念的数目为基础的，被认为是认知风格的一个度量。认知复杂性的较高程度意味着个体应用较多的维度，因此以更分化的认知系统来组织或表征周围环境（Bieri, 1971; Bieri 等, 1966; Goldstein & Blackman, 1978a, pp. 483 - 487; 1978b）。然而，认知复杂性本身是一个多因素构念，它与其他变量的相关程度在很大程度上与它怎样被概念化和测量有关（参见 Goldsmith & Nugent, 1984）。

经过一段相对沉寂的时期后，20 世纪 80 年代人们重新又对凯利的个人构念理论显示出极大的兴趣，自此，在这一框架下的研究也以越来越快的速度进行着（Bannister, 1985; Burr & Butt 1992; Epting & Landfield, 1985; Fransella & Thomas, 1988; G. J. Neimeyer & R. A. Neimeyer, 1990; R. A. Neimeyer & G. J. Neimeyer, 1992）。凯利理论还被用来作为建构主义的支柱之一，<sup>[17]</sup>因

为它非常适合于研究个体产生和组织意义的独特方式，所以，Rep 测验本身在近年已得到更为广泛的应用。研究者现在使用更为复杂的统计技术如分层聚类分析和多维量表法，来分析项目栅格数据(参见 Merluzzi, 1991; Ogilvie & Ashmore, 1991)。该技术有如此多的变式，因此它的有效性和心理测验的特征不能被概括出来。

**环境感和社会气氛感 ( Perceived Environment and Social Climate)** 环境这一概念以多种方式进入心理学领域。<sup>[18]</sup>我们已经考虑了观察者对环境的评定在测验特殊人群中的特殊作用(第九章)，以及文化和情境变量对认知和情感特性表现的影响(第九、十一、十二、十三章)从现象学的观点来看，通过在每一环境中人们所报告的知觉而得到的环境和社会气氛的评定，大有助于理解个体和群体

许多方法可用于描述和评价环境的物理和社会方面(有关这些方面的几个评论见 Walsh & Betz, 1995, 第十一章)。斯坦福大学的摩斯(R. H. Moos, 1974, 1993a, 1993b, 1993c, 1994a; Moos & Spinrad, 1984)所编制的 10 个社会气氛量表，是其中经常被使用的通用工具。社会气氛量表可用于下列情境中：医院和社区的精神病治疗计划、保护性照料和矫正机构、军事场所、大学生的寝室、中学的教室、工作环境、家庭 还有较一般的群体环境量表用于工作定向的、社会的和相互支持的群体。大部分测验包括 90 或 100 个是非题，在这其中反应者描述她或他对给定环境的知觉。最初的项目编写与理论上挑选出的环境维度一致，例如，投入、自治、秩序，最后的项目是根据它们区分情境的能力和它们在分量表内的内部一致性而经验性地选择出来的。10 个量表中全都经过一次或多次修订，使之与当前情况相适应(Moos, 1994b; Moos & Moos, 1994; Trickett & Moos, 1995)。社会气

气氛量表可以一种不同形式加以实施,分别测量(1)现实环境的知觉,(2)理想环境的知觉,(3)对某种未知环境的期望

每一个环境方面根据 7~10 个分量表分数来描述,分量表评价环境的不同维度,某些维度在不同的分量表中重复出现事实上,尽管社会气氛量表包括多种多样的方面,但每一方面中的分量表都可分成同样的三类,包括(1)关系维度(投入、支持、同伴内聚力),(2)个人成长维度(自治、任务取向、竞争),(3)系统维持和改变维度(秩序和组织、明确性、革新)。还值得注意的是,社会气氛量表用于复杂和异质机构内的相对较小的单元,如一个教室而不是整个学校,一个治疗方案而不是整个医院,一个大学住宿中心而不是整个大学。因此,这些量表得到的数据,比对整个组织的混合性评定而得到的数据更易解释,更为明确。虽然社会气氛量表的使用指南指出,它们可用于临床评定、组织咨询、方案评价,但一些评论者认为这些量表更适合于环境知觉的决定因素和后果的研究(Allison, 1995; Loyd, 1995; R. O. Mueller, 1995; Saudargas, 1989; Sheehan, 1995; C. R. Smith, 1989)。

## 观察者的报告

到现在为止,我们讨论过的测验为人格评定所遵循的各种方法提供了充分的证据然而,对大部分测验而言,说它们是适用于研究目的的实验技术或熟练的临床人员结合个体其他信息解释人的人格特征时的有用工具是最恰当的了。很明显,人格评定不能完全依赖于标准化测验。为了追踪或增补由测验分数提供的线索,为了评定无适当的测验可用的特质,为了发展人格测验和获得使人格测验有效的标准资料,其他方面的信息资源

也是需要的。

不论是在诊所、咨询中心、教室、人事办公室,还是要求进行个体评定的任何其他场合,直接观察人的行为在人格鉴定方面起着重要的作用。要把这样的行为观察放在适当的视角中考虑问题,我们必须记住所有的测验本身只是小样本行为的评定。这些行为样本是在标准化的条件下观察到的。这样的标准化程序尽管有明显的优点,但通过在自然情境下的观察获得更广泛的行为样本的资料也有其长处,我们必须把两者加以平衡。举个极端的例子,假设我们有一个人从出生到 30 岁的详细记录,我们就可能对他的行为作出预测,而且比通过任何测验来预测更准。这种有关个人生活所有细节和周围环境的记录是很难进行的,但要是我们有的话,我们就可以根据 30 年的行为样本来预测,而不是通过一二个小时的测验来预测。

这一节所涉及的所有技术中,个体在自然情境下相对长期的所作所为,通过一个或几个观察者为中介来传播。为了提高观察的准确性及可交流性,有许多工作要做。

**自然观察** 儿童心理学家最广泛运用的是在自然情境下对自发行为的直接观察技术,尤其是对学前儿童。虽然,这种方法对任何年龄的人都适用,但因为年龄越小的个体其行为受观察者出现影响的可能性越少,或个体发展了的社会掩饰性使得对行为的解释复杂化的可能性也越少。这些观察技术也适用于教室,特别是观察者是老师或是很容易适应正常学校情境的其他人。这种评定技术主要可见于学校、家庭、儿童照料中心、诊所、医院或其他场合所实行的行为矫正计划中 (Hartmann & Wood, 1990; Kent & Foster, 1977; Lalli & Goh, 1993)。在社会心理学 (Webb, Campbell, Schwartz, Secrest & Grove, 1981) 和跨文化研究 (Bochner, 1986) 中也已运用了巧妙的自然观察技术。

自然观察涉及的方法多种多样 (Adler & Adler, 1994; Jones, Reid & Patterson, 1975; Sattler, 1988, 第十七章), 可以是全面的高度的方法, 如日记法, 也可以是限制在较窄范围较短时间内的高度受控制的观察, 如时间取样法。后者涉及到一个短期观察时间的分配问题, 这主要视观察性质和目的而定, 可以是短至不到 1 分钟的观察, 也可以是连续几个小时的观察, 但最普通的观察时间是 5 分钟或更少一点。观察可以集中在一天, 也可分散在几个月内。可以观察在特定时间内出现的所有行为, 但更为常用的还是限定一个特定的行为, 如语言、运动、人际行为或攻击。所要观察的内容的核查表是一种有用的辅助工具, 其他辅助工具有观察的时间表、记录表、编码系统和其他记录工具 (W. W. Tryon, 1985 第七章和第八章)。记录可以通过录音、胶片或录像带进行,<sup>[19]</sup> 手提式电脑也可用于收集和分析观察数据 (Kratichwill, Doll & Dickson, 1991, pp. 137 ~ 141; Repp & Felce, 1990)。

我们必须指出, 自然观察与前面所讨论的情境测验有很多共同之处, 而它们的不同之处在于两个方面。其一, 自然观察没有对刺激情境进行控制; 其二, 至少在大多数观察方法中, 观察的行为样本更为广泛。关于自然观察技术的研究兴趣正在增强, 特别是它们的生态效度和它们在评定随时间而变化的行为方面的有效性 (参见 Barkley, 1991; Kaminer, Feinstein & Seifer, 1995)。

**访谈** 我们应该提到历史悠久的访谈技术, 它也是提供信息的来源。访谈已在临床心理学、咨询、人事心理学和教育上用于多种目的。许多地方都已对访谈的方法、运用、效果及关于访谈过程的研究有所讨论。<sup>[20]</sup> 访谈有多种形式, 从高度结构化(与呈现口头实施的问卷没有多少区别)、涵盖一些预先确定范围的有结构的或指导性访谈, 到非指导性和深度访谈, 其中访谈者只

是设定一个活动舞台，鼓励被访谈者尽可能地自由谈论。结构式访谈现在已被普遍用于临床和精神病诊断领域的研究目的。这些工具是标准化的，除了诊断分类还提供定量的分数，因此，它们像所有的测验一样，用信度和效度的心理测量学标准来评定关于各种结构访谈的评论，见霍金和泽曼（Hodges & Zeman, 1993）、坎普汶斯和弗利克（Kamphans & Frick, 1996, 第十二章）、罗杰斯（Rogers, 1995）的论著。

访谈主要提供两种信息，首先，它们提供对访谈期间表现出来的有限行为样本进行直接观察的机会，如可以记录个体遇到陌生人时的言语、姿态、态度；第二，这也许是访谈更重要的作用，即可以引发生活史资料，个体过去做过的事是他或她未来会做些什么的预兆，尤其是根据相伴环境和个体自己对行动的评论来解释这些往事时。访谈不仅应关注什么已在被试身上发生，还应关心她或他对这些事件的知觉和目前对它们的评价。

作为访谈者，访谈需要收集和解释资料的技术。访谈可能会由于没有引发重要信息或对资料作出了不恰当的或不正确的解释，而导致错误的决定。一个成功的访谈者的关键素质是，具有鉴别被访谈者行为和他或她所报告事实的线索的敏感性。这些线索会导致对或支持或否定原先假设的其他事实的进一步探究。

**等级评定** 虽然在很多领域以及为了不同的目的都用到等级评定，但这里所讲的评定是评定者在对个人日常生活进行累加和非控制的观察基础上对个体的评定。这种评定不同于随意、非正式的积累材料的自然观察，它不仅简单地记录观察结果，还要进行解释和判断。既不同于自然观察，又不同于访谈，这种评定通常有较长的观察过程，信息是在较真实的环境中收集的。等级评定大量应用于教育或工业环境下对个人的评定，

从而为测验的效度和众多的研究目的分析收集效标资料从 70 年代起,掀起了一股研究等级评定技术的高潮,其重点在于全面而系统的研究,定义和程序的充分标准化,以促进各研究结果的比较(Borman, 1991; Landy & Farr, 1983; Ozer & Reise, 1994; pp. 370 ~ 371; Saal, Downey & Lahey, 1980; Sulsky & Balzer, 1988)

在提高等级评定的准确性方面还可以作许多改进。共同的困难来自于特质名称、量表单位或两者皆有的模糊性。为了解决这个问题,每一种特质都要用特定的名称,等级评定也要以所有的评定者都能一致理解的方式来表述。比起那些可以让不同的评定者产生歧义的数字或一般的描述性的形容词来,用精心表述的行为锚(behavioral anchors)来确定特质的程度可能更为明确(Dickinson & Zellinger, 1980)。也有证据表明,不同的量表结构的相对准确性随工作性质和所完成的行为功能的不同而变化(Borman, 1979; J. M. Feldman, 1986)。

影响到等级评定效度的一个因素是评定者与被评定者的接触程度(Freeberg, 1969; Landy & Farr, 1980; Paulhus & Bruce, 1992; Wiggins & Pincus, 1992, pp. 493 ~ 496)。仅仅认识很长时间是不够的,评定者应该有机会在行为可以得到表现的情境中观察被评定者。比方说,如果一个雇员始终没有机会进行决策,那么上司就难以对她或他的这种能力进行评价。在很多评定情境下,如果评定者没有机会观察到某个体的特殊特质,最好要有一个空间打上记号以代替评定。

等级评定也容易产生一些误差。<sup>[21]</sup>一个著名的例子就是光环效应(halo effect),对它的传统定义是评定者的倾向性不适当地受到了一个好的或差的特征的影响,它又影响到对个体其他特征的判断。光环效应的存在通常是从对不同行为维度的等级评定之间的相关推论出来的。虽然光环效应还是被一些研究者

按传统的观点来看待,并把它作为单一的构念,但是这个概念的其他说明方式和其他操作性定义已经出现 (Balzer & Sulsky, 1992; Kozlowski, Kirsch & Chao, 1986; Murphy & Anhalt, 1992; Nathan, 1986).<sup>[22]</sup>

另外一个会影响等级评定数据的误差是趋中误差 (error of central tendency) 即评分都位于量表中段,避免极端分数。还有一个误差是宽大误差效应 (leniency error),指的是很多评定者不愿意打否定分数。前者的分数都集中在量表中部,后者都集中于量表的上端。两种误差都缩小了分数分布,使等级评定不具有辨别力。消除误差的一个方法是进行排序或其他的优点排列法 (order-of-merit procedures),用这些方法可以迫使在个体之间作出区别,从而使等级评定的信息得到最大的利用。当然,这些技术只能应用于组内的比较,而不能应用于由不同评定者评定的不同组间的直接比较。

对等级评定误差的大量研究发现,光环效应或趋中效应和其他较为直接的评定准确性指标之间的关系十分复杂,并且常常和直觉相反。一些进行多元分析研究并对评定误差的文献进行评论的研究者断定,存在于行为评定中的概念和方法上的问题使得我们不能用这些错误的度量作为等级评定准确性的指标 (Balzer & Sulsky, 1992; Borman, 1991; Murphy & Anhalt, 1992; Murphy & Balzer, 1989)。

不过,通过训练评定者可以改进评定过程。不同场合下的研究都证明了训练在提高信度和效度以及减少普遍的判断错误方面的有效性 (Bernardin & Buckley, 1981; McIntyre, Smith & Hassett, 1984; Pulakos, 1986; Stamoulis & Hauentein, 1993; Sulsky & Day, 1992, 1994)。然而应该指出,评定者训练方案中包含了许多不同类型的训练,它们的效果在种类、数量和持续时间上都不同。



训练可以是给评定者提供如何评定行为的统一参照标准,也可以帮助分析一般的评定错误和将其影响降至最低程度,或提高观测的技巧等等。对于特殊的评定情境和目的,一种或几种训练的结合是最合适的。然而,在大部分情境下提高评定观察技能可以产生很好的效果。

临床评定通常需要从了解被评定者行为模式的那些人那里收集资料。等级评定量表为收集这类信息提供了一种有效的方式,并且非常适用于对儿童和青春期的评定。很多为父母和教师运用的标准化量表已在近年内出版,现在可以购到。<sup>[23]</sup>一个特别全面的例子是儿童行为评定系统 (Behavior Assessment System for Children, BASC, Reynolds & Kamphaus, 1992)。这个系统除了教师评定量表和父母评定量表外,还有自陈和观察者资料。它们分为从学前期到青春期的三种不同形式 (关于评论见 R. B. Kline, 1994)。

**提名技术** 一种在同伴之间互相评定的特别有用的方式是提名技术。这个技术最初是为了调查群体构念而在社会测量法 (J. L. Moreno, 1953) 中发展起来的,可以应用于任何群体中,例如班级、俱乐部、工厂、军事单位,只要群体中成员相处的时间很长,足以彼此熟识。要求每个人从群体其他成员中选择愿意与之共同学习、工作、进餐、游玩,或者进行其他活动的一个或几个同伴。所选同伴数目可由被试自己确定,选择方法或者是按一定的顺序选择(比方说第一名、第二名、第三名),或者每一项活动只选一名同伴。

当这种技术被用于个人评定时,根据某个成员被别人提名的次数,可以判断他是群体的潜在领导人物(获得很多提名)还是群体中的孤立者(很少或没有被提名)。另外,还可以对个体评定计算更加准确的指标,最简单的一种是数出某个成员在某

项活动中被提名的次数，以此作为对她或他的评定分数。提名技术可以应用于感兴趣的任何行为，例如，可以要求被试指出谁的独到见解最多、能依靠谁把工作做好、谁的体育最棒。他们有时不仅要指出与给出的描述最相符合的成员，还要指出与描述最不相符合的成员，在这种情况下，肯定提名记为 +1，否定提名记为 -1。<sup>[24]</sup> 同伴评定也可以通过其他方式，比如说排序或评定等级来获得，但提名技术被证明是最成功的，应用得也是最经常的。

同伴评定在各种各样的群体中已经成为最可信赖的评定技术之一，比如说部队人员、工厂管理人员、和平部队的志愿者、中小学生和大学生等群体（Cole & White, 1993; Gresham & Little, 1993; Hughes, 1990; Kamphaus & Frick, 1996，第十章；Kane & Lawler, 1978; J. S. Wiggins, 1973/1988, pp. 356 ~ 363）。根据人际关系的各种实际效标来检验这些评定，通常发现它们具有良好的同时效度和预测效度。如果我们考虑同伴评定的特征，这些结果是可以理解的。首先，参与评定的人员很多，包括所有的群体成员；其次，个体的同伴有观察到个体典型行为的非常有利的条件，对于某些人际关系的特征，同伴会得出比老师、管理人员或其他外部观察者更正确的判断；第三，可能最重要的是，群体成员的意见无论是正确的还是错误的，都会影响到他们的行为，因此也会部分决定这个个体与群体之间后来的相互作用的性质。其他的类似群体也会以相似的方式对待这个个体。因此社会测量评定和工作样本一样具有内容效度。

**核查表和 Q 分类** 任何一种自陈工具，如第十三章和第十四章讨论的人格和兴趣调查表，也可用于对他人进行描述。<sup>[25]</sup> 用于评定自我概念的测量特别适用于这个目的，如在 IPAR 研究计划中（Gough & Heilbrun, 1983），形容词核查表（ACL）就被广

泛用于获得观察者的评价。在一个 2~3 天的评定期内,受过训练的心理学家密切地观察被试,通过在表中恰当的形容词上做记号,记录下自己的评价。

Q 分类也被广泛用于获得观察者的评价布洛克(Block, 1961/1978)最初提出了“加利福尼亚 Q 分类卡片”(California Q-Sort Deck),为受过职业训练的观察者进行综合性的人格评定提供标准语言。以后,这套卡片发行更为广泛,并经过修订以适应于专业及非专业使用者。另一套适用于较小年龄段的“加利福尼亚儿童 Q-定向”(California Child Q-Set)也已编成(关于评论见 Heibrun, 1985)所有版本的材料都由 100 条陈述构成,按 9 点加以分类。分类是根据每条陈述对某个个体的显著性——即它们对说明被评对象的独特的和基本的特征的重要性而进行的。因此,Q 分类所特有的自比参照框架仍保留着,个体不和外在的常模标准作比较。

Q 分类项目的这种一致性有利于交流,也保证了不同观察者结果之间的可比性。标准 Q 分类也可用于其他许多研究目的(参见 Caspi 等,1992;Reise & Oliver, 1994;Wink, 1992)。另一个应用是把这个技术应用于个人评定,布洛克(1961/1978)提供了三个“限定性 Q 分类”的例子,表现了对一个理想的正常人和两个精神病综合征患者的一致性评价,对某一个体的 Q 分类可以和这些评价进行比较。可编制类似用于任何种类的人的限定性 Q 分类(defining Q sorts)。

## 个人传记资料

本章开头在讨论访谈技术时,曾提到生活史资料的重要性。

关于一个人的过去行为及经历的信息不仅对试图理解人格和认知发展模式的理论家，而且对试图评定个体并预见其行为的应用心理学家，都很有意义。因为个体在过去某个特定情境中的反应方式能为该个体在将来相似情境中如何反应提供信息。

生活史资料可通过几种方法获得，其中访谈和问卷可能是最普遍的，特别是在临床和咨询心理学中。日记和自传也可以为心理传记作者和其他对个人生活研究感兴趣的人提供大量信息(参见 J. S. Wiggins & Pincus, 1992, pp. 487 ~ 493)。<sup>[26]</sup> 进行纵向研究的研究者，在不同时间重复观察、测量被试，不仅收集而且建立了一个人的生活史资料(Funder, Parke, Tomlinson-Keasey & Widaman, 1993)。

然而，最具结构性的收集和使用个人生活史资料的方法是运用传记调查表或量表（现在称为个人传记资料测量）。它是专为在工业和教育环境中预测绩效而制定的。和第十三章、第十四章探讨的人格与兴趣调查表一样，自传调查表是一个标准化的自陈工具，对项目的反应从两个或两个以上的选项中加以选择来作出回答，而不是由反应者来建构。虽然大多数问题是属于相对客观的和容易验证的事实，但当它们被记分和用于评定或预测时，自传量表和结构访谈一样，仍须满足和其他测验一样的信度、效度的心理测量学标准。典型的项目涉及所受教育数量、性质、工作经历、特殊技能、爱好，以及娱乐活动的数量和性质。当反应者被问及在学校最喜欢或最不喜欢的课程，或在他们的工作经历中最喜欢和最不喜欢什么时，通常可以看出个体对先前经历的反应。<sup>[27]</sup>

历史上，自传量表的项目根据效标定项来选择和加权，和第十三、十四章中讨论过的 MMPI 和斯特朗这些调查表的编制一样。产生的调查表根据在新样本中的相同效标进行交叉效度分

析。当遵循这些程序时，个人传记资料调查表被证明对各种各样情境中的行为有一致的良好预测作用。已经根据各种效标，如人寿保险代理商出售的保险总额、银行职员的工作营业额、研究科学家的成果、中学生的艺术创造力以及海军人员潜水训练中的成绩，编制了一些调查表。当用于预测从非熟练工人和蓝领工人到高层管理人员的工作绩效时，这样的测量被证明是有效的（Anastasi, 1979, pp. 79 ~ 80; Owens, 1983）。同时，许多为特定工作编制的自传调查表的局限性是它们不可推广或不能一般化，它们的可应用性局限于编制它们时所采用的特定的场合和效标（Baird, 1985; Hunter & Hunter, 1984）。

自传测量一直是广泛研究的课题。研究者进行了大量理性的和因素分析方法的探索，以发展个人传记资料量表（Hough & Paullin, 1994; Schoenfeldt & Mendoza, 1994）。与项目的经验选择和效标确定相反的是，理性的方法通常先通过工作分析和已出版的实验的和理论的两方面文献资料的研究确定主要的有关构念。然后对初步的项目进行因素分析，从这些项目中挑选出最后的项目以产生因素量表。编制个人传记资料量表的各种方法的相对优点是一个不断争论的主题，虽然有一点是明确的，即每一种方法都有优点和缺点。理想的方法是构念效度分析策略，它合并尽可能多种多样的成分。

为了使产生的工具更一般化和可推广，除了传统的调查表编制方法，新的个人传记资料项目的产生、选择以及效标确定方法正在尝试（参见 Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990; Russell, Mattson, Devlin & Atwater, 1990）。欧文斯（William A. Owens）率先涉足一个最具创造性的研究领域，利用聚类技术确定个体的亚群，他们具有共同的生活经验史形态（Mumford, Stokes & Owens, 1990; Mumford & Stokes, 1992; Owens & Schoen-

feldt, 1979)。这样的研究结果可应用于多重效标的预测。更重要的是,通过这些方法确定的亚群的研究可以导致对个人生活发展型式的了解,从而达到这样一个水平,即将个人的和普遍的研究途径的成分结合在一起 (Hein & Wesley, 1994)。

经过了多年仔细的控制研究后,今天,个人传记资料测量已成为教育、工业、政府和其他场合最可靠、最有效的评定和选拔手段之一。而且,个人传记资料的研究历史,以不同寻常的清晰方式,强调了基础科学和应用科学彼此之间的相互依赖。有趣的是,为解决人事选拔中的实用问题所进行的研究,其方法和结果,对用来理解个体一生行为模式发展的理论框架的形成作出了贡献。

不过,个人传记资料测量的实施不是无懈可击的,在实践的、政治的,以及技术的方面均存在着问题。这些问题主要是关于(1)合法性,例如侵犯隐私、平等的雇佣机会等问题 (Sharf, 1994);(2)个人传记资料测量和回顾报告容易伪装和有其他不正确的地方 (Henry, Moffitt, Caspi, Langley, Silva, 1994; Lautenschlager, 1994; Trent & Laurence, 1993)。关于传记工具发展和使用的所有方面的大量有用的研究已概括在《个人传记资料手册》(*Bildata Handbook*)一书中(Stokes, Mumford & Owens, 1994)。

## 注 释

- [1] 把普通心理学的理论和研究以及实践中这些方面整合起来的需要,人们已有了越来越清楚的认识,参见最近由萨克劳费斯克和蔡德纳(1995)、斯登伯格和鲁茨格斯(1994)编的两本关于人格和智力相互作用问题的著作。
- [2] 第十四章提到的由特雷西和朗茨(1996)提出的职业兴趣的球形表现是比较复杂的包含多维度的图式的一个例子。

- [3] 应该指出, 认知风格的复杂性各不相同, 有些风格只有比较简单的差异, 如沉思和冲动, 再比如认知速度, 只是界定为个体对问题情境反应的速度, 特别是涉及高度不确定的情境时的反应速度 (Kagan, 1965; S. B. Messer, 1976)。另外一些认知风格, 如斯登伯格格的思维风格 (1988, 1994b) 包括了比较复杂的素质模式
- [4] 在第十三章关于人格自陈调查表中迫选技术的使用简单地提到了 MBTI
- [5] 根据 MBTI 手册, 这样的人是 (1) 安静和沉默寡言的; (2) 喜欢用逻辑和分析的方法解决问题; (3) 主要对观念感兴趣; (4) 具有明确界定的爱好 (Myers & McCaulley)
- [6] 虽然没有什么从一个情境到另一个情境的转移性的证据, 但通过对资料的聚类分析而产生的亚群——本章后面将讨论——从类型分类的观点来看, 也提供了一些有意义的可能性
- [7] 对哈茨霍恩和梅的研究在关于美德的教学和性格的争论发展中的作用进行了有趣的分析
- [8] 对这些开始于 OSS 项目的里程碑式的评定研究的解释, 可见于威根斯 (1973/1988, 第十一章)。
- [9] 多媒体和交互的计算机技术要求以全新的方式将真实的刺激情境和反应选择呈现给参加测验者。对发展多媒体测验的可能性和问题, 以及测量消除冲突技能的交互录像工具的讨论, 见德拉斯高、奥尔逊——布坎南和墨伯格 (1996) 与奥尔逊——布坎南、德拉斯高、墨伯格、米德和克娜 (1996)。
- [10] 在临床心理学中, 角色扮演已经在几个不同的理论取向中经历了广泛和系统的发展, 特别是在行为矫正程序、家庭治疗以及婚姻咨询中
- [11] 对评价中心技术的简洁而丰富的评价, 见兰迪、桑克斯特和苛勒 (1994, pp. 277 ~ 278), 与斯密特、奥尼斯和亨特 (1992, pp. 635 ~ 637)。
- [12] 对 17 ~ 20 世纪关于自我和自我相关过程的哲学和心理学理论的回顾, 可见勒温的论著 (1992)。
- [13] 布拉斯科维基和托马卡 (1999) 提供了评价这一工具的概况, 并对其他

许多自尊测量进行了评价。夏弗逊的模型是这一领域最有影响的模型之一,还被用于 SSCS 的概念化。关于这一模型的资料,见马什等(1992)、马奇和夏弗逊(1985)、夏弗逊和布勒斯(1982),以及夏弗逊等(1976)

- [14] 1980年前 ACL 的全部文献目录可以从测验出版商那里获得 (Gough & Heilbrun, 1980)。
- [15] 当然,所有这些程序也可为以前讨论过的形容词核查表或评定自我概念的任何其他技术所遵循。此外, Q 分类和核查表可用于本章后面将讨论的观察者的评价。
- [16] 对其他程序变式的说明,见班尼斯特和梅尔(1968)、比尔(1985)、莱兰菲尔德和埃普汀(1987)、内墨耶(1989)和威特(1992)。
- [17] 建构主义是心理学内一场活跃的和相当有争论的运动,它把人类看成是主动的(proactive)的作用者,它的主要努力是了解经验的意义和要求对知识有多重观点——超越实证论(Gergen, 1985; Mahoney, 1991; C. J. Neimeyer, 1993; R. A. Neimeyer & Mahoney, 1995)。
- [18] 事实上,在 30 年的活动以后,环境和行为的研究是一个多学科领域,不仅在美国而且在全世界得到迅速发展(参见 Bonnes & Secchiaroli, 1995; Croat, 1995; McAndrew, 1993; Stokols, 1995; Stokols & Altman, 1987)。
- [19] 一个主要基于观察资料的评定系统的不同寻常的全面的例子是,古特曼(1994, 1996)为分析婚姻关系过程设计的系统。该系统包括像评价中心方法学一样的元素,是通过长期的研究计划发展和测试的。这一研究计划目的在于分离出预测婚姻最后结果的因素。
- [20] 关于访谈的广泛的文献包括有关研究的定期的考察(Elder, Kacmar, Ferris, 1989; Graves, 1993; Groth-Marnat, 1990, 第三章 Landdy 等, 1994; McDaniel, Whetzel, Schmidt & Maurer, 1994)和访谈指导及建议,以改善访谈方法,特别是在临床实践(Bierman, 1990; Lukas, 1993; Morrison, 1995; Rogers, 1995; Shea, 1988)和人事选拔中(Fear & Chairol, 1990; Webster, 1982)。教授和学习适用于大多数访谈场合的基本技能的模型可见于高登(1992)。
- [21] 关于等级评定错误的大多数研究报告发表在《应用心理学杂志》



- ( *Journal of Applied Psychology* )。
- [22] 一些其他的概念，至少部分地，把光环错误看作不仅是评定者而且是被评定者，甚至是特定的评定环境的作用。
- [23] 对这类工具的讨论和对它们中几个评论可见于卡姆费斯和弗里克(1996)、皮塞迪尼(1993)、威特、赫弗和法伊弗(1990)。
- [24] 当获得否定的提名时，测验者应该特别仔细地防止程序的任何潜在有害效应。关于儿童和青少年同伴参照评定的伦理学讨论，见格瑞斯姆和利特尔(1993, pp. 174 ~ 175)，卡姆费斯和弗里克(1996, pp. 201 ~ 203)。
- [25] 儿童人格调查表(PIC)和青年人格调查表(PIY)——在第十三章阐述——事实上是平行的工具：(1) 自陈形式(PIY)和(2) 被观察行为的调查表形式(PIC)。
- [26] 自传体记忆作为一种投射工具在第十五章中讨论。
- [27] 这一领域的全面陈述和个人自传体资料项目的特征可见于梅尔(1991)。



## **第五部分**

### **测 验 的 应 用**



## 第十七章 当代测验使用的主要领域

心理测验的使用有着众多不同的目的，其应用领域也在不断地扩大。在详细回顾各种类型测验的代表性样例后，我们现在将考虑与其应用有关的问题。在本章中，我们将考察三个主要的领域，即教育、职业和临床／咨询。在这些领域中测验有着多种功能。在下一章，也就是最后一章里，我们将讨论与所有领域的测验实践有关的伦理和社会问题。

### 教育测验

几乎每类可利用的测验都在学校中得到应用。智力、特殊能力倾向、多种能力倾向和人格测验都能在教育咨询人员和学校心理学工作者的资料库中找到。教师和教育管理者常常不得不依据几种不同测验的结果而行动。然而，某些种类的测验专门是为教育领域的应用而发展起来的。<sup>[1]</sup>本节关注的就是这一类测验，它们包括在某些特定的教育场合为预测和分类而设计的各种工具和许多不同的教育成就测验。

**成就测验：性质和使用** 成就测验在绝对数目上超过所有其他类型的标准化测验，其目的在于测量某一具体的教学或训练的效果。习惯上，我们经常把成就测验和能力倾向测验相对

比,后者包括一般智力测验、多种能力倾向成套测验和特殊能力倾向测验。从某一方面来看,成就测验和能力倾向测验的区别首先表现为相关的先行经验一致性程度的不同。成就测验测量相对标准化的一组经验的效果,如初级法语、三角或计算机编程课,而能力倾向测验成绩反映了日常生活中多重经验的累积影响。我们或许可以这样说:能力倾向测验衡量的是相对没有控制的和不知道条件下的学习效果,而成就测验衡量的是部分控制和已知条件下出现的学习效果。

它们的第二个区别在于各自的使用目的不同。能力倾向测验是为了预测后续的成绩,它被用来估计个体将从某一特殊的课程收益的程度,或预测个体在新的情境中成绩的质量;相反,成就测验一般用来表示完成某种训练后个体状态的最终评价,成就测验中强调的是个体在当时能够做什么。

然而必须认识到,能力倾向测验和成就测验的区别不是固定的。有些能力倾向测验可能在相当程度上依靠特定连贯的前学习,有些成就测验则覆盖着相对广泛、未经标准化的教育经验。同样,某一成就测验可以用来预测将来的学习,因而它具有同能力倾向测验一样的功能,例如,学习医学课程以前的成就测验可以用来预测个体在医学院中的学习成绩。

为了努力避免‘能力倾向’和‘成就’术语带来的过多意义,现在越来越倾向于代之以用中性词“能力”(ability)来衡量认知行为。<sup>[2]</sup>任何认知测验,不管其是否属于传统型的,都提供了个体在测验时知道些什么的样本,测量个体所获得的一种或多种能力的发展水平;但没有一种测验能揭示个体是如何或为何达到该水平的。为了回答后一个问题,必须研究其他相伴出现的变量,尤其是应该了解个体的经验背景。从这个意义上来看,每一测验分数都有需要了解的过去,这样才能恰当地理解获得该

分数的个体；同样的测验分数也预示着未来，因为它使我们能够预测个体在其他非测验情境和将来的某段时间会做些什么。

已形成能力的测验，不管是一般智力测验、多种能力倾向成套测验，还是成就测验，都能按照它们所需的经验背景的特殊性在连续体上加以排列。图 17-1 说明了这一连续体。在它的一端是“课程定向”的成就测验，包括技能技巧或事实性知识测验，俄语词汇或电视机维修之类的测验就属于这种类型。紧接其后是“广泛定向”的成就测验，在今天一般用于评定是否达到主要和长期的教育目标；这些测验着重于科学原理的理解和应用、文学的解释、艺术的欣赏等。还有一些测验具有更广的定向，在于测量影响个体各种不同活动中表现的基本认知技能，如阅读理解、数学计算、逻辑推理等。显然，这里的大多数成就测验与传统的智力和能力倾向测验微妙地融合了。言语的认知成套测验传统上称之为智力测验，与最广义的成就测验相似其后是非语言和操作测验，经常不要求个体阅读或写作在连续体的“一般性”一端是跨文化测验，应用于具有广泛不同经验背景的个体。

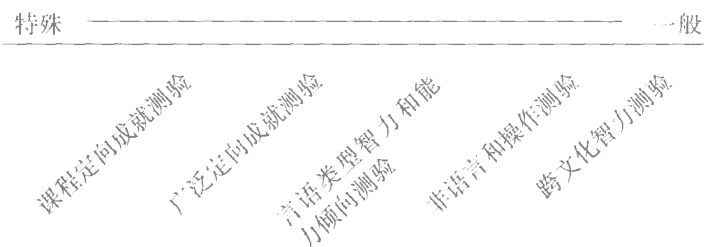


图 17-1 已形成能力的测验：经验特殊性的连续体

把一些工具称为能力倾向测验，另一些则称为成就测验，这在一定程度上导致了测验结果的误用。一个普遍的例子是：在实践中经常把成就测验分数低于其学业能力倾向或智力测验分数的儿童确定为“低成就者”(underachiever)。事实上，个体内的这种测验分数的差别反映了这样一个事实，即没有两种测验（或其他成绩预测工具，如课程分数）是完全相关的。高成就或低成就问题，可以更确切地称之为从第一个测验到第二个测验中出现的过高预期与过低预期问题。个体案例中预测错误的原因有：测验工具的不可靠性，测验覆盖内容的差别性，两种测验中态度和动机因素带来的不同效果，还有诸如补救教学或长期疾病等干扰经验的影响(R. L. Thorndike, 1963)。

人们很早就认识到了成就测验能在教育过程中起许多作用。在年级的确定或其他已获得能力的评定中，作为辅助工具的标准化成就测验具有客观性、一致性和有效性的优点；如果它得到恰当的建构，还具有其他优点，如内容涵盖的合理性，能减少无关和偶然因素在评分过程中的作用。成就测验也形成了补偿教育计划的一个重要特征，它既能用于确定特殊教育无能的儿童，又能用来衡量补偿作业中所取得的进步。

对于所有类型的学习者来说，周期性施测、建构良好、选择恰当的成就测验都有助于学习。这种测验能发现过去学习中的弱点，给后续学习提供方向，激励学习者。“知道结果”的激励作用，已不断地被许多不同学习情境、不同年龄和不同教育程度的学习者的心理学实验证实。这种自我检查的效果一般由于及时而得到加强。

从另外一个角度来看，成就测验提供了一种适应个体需要的教学手段。不管其处在何种阶段，当教育能够满足学习者的需要时，学习就会变得富有成效。有效教学所需的第一步是知



道个体已经能够做什么,和对一门学科他已经知道了什么;根据学年伊始所实施的测验,教育者能够采取建设性的步骤,弥补测验结果所发现的知识上的主要漏洞。成就测验在教育过程中的作用的其他例子,还可以在第三章所讨论的“领域参照测验”(Domain-referenced testing)和“适应个体差异指导系统”(the individually tailored instructional systems)中发现。

最后,成就测验可以作为辅助工具用于教学评价和改进以及教育目标的表述中。教育测验能提供事实上正在教的基本内容和技能是否合适的信息。通过对这些问题的关注和提供具体的事实,成就测验能引起对教育目标的分析,鼓励对教学内容和方法进行批评性的检测。<sup>[4]</sup>由于公众对责任的要求上升,在过去的一二十年的时间里,在所有水平的教育机构中,测验的数目达到了空前的程度。大量的测验由州政府、地方学区和联邦政府控制(B. Gifford, 1989b; Hart & Battaglia, 1993; National Council On Education Standards and Testing, 1992)。全国教育发展评定,俗称为“全国报告卡”,是现行由政府控制的最好测验项目之一(参见 Alexander & James, 1987; Gentile, Martin-Rehrmann, & Kennedy, 1995; E. G. Johnson, 1992; Messick, Beaton & Lord, 1983; NAEP, 1985; F. B. Womer, 1970)。出于标准化测验的流行和可能与其有关的范围广泛的经济效益的考虑,测验本身受到了仔细的审查和尖锐的批评。用于教育发展评定的方法曾经是测验专家所独有的职责范围,因此已经成为政治上争论的焦点,受到立法者、企业家和民众的关注(R. E. Bennett & Ward, 1993; Court & McInerney, 1993; Gifford & O'Connor, 1992; S. P. Robinson, 1993; G. P. Wiggins, 1993),对于由外部因素控制的与学校测验和评定项目有关的问题和发展趋势,请看林和格朗伦德(Linn & Gronlund, 1995, 第十八章)的论文。

**编制和选择** 历史上,传统的学校考试由一系列需要口头和书面回答的问题组成,两种回答都是由应试者构成并表述反应。“问答题”广泛用来指可以自由反应的问题,不仅包括需要有较长回答的问题,而且还包括要求应试者简单作答的问题或数学问题;而“客观问题”则用来指那些从每问题可供选择项目中选取正确答案的测题。要求应试者选取一种反应的项目具有好几种类型,如‘对—错问题’和‘匹配问题’;但到目前为止,研究最彻底、受到批评最多的测题类型是多项选择题。

多项选择形式的批评者认为:它促进的是孤立事实的机械记忆和学习,而不是发展解决问题的技能和对概念的理解;而且还有在许多教育和政治机构中,许多无知的人把多项选择题等同于标准化的测验,同时轻视两种评定方法的原则。<sup>[5]</sup>可笑的是,用来描述教育发展的同样一些标准化测验方案常常被认为促使它们所揭露的教育缺陷的形成 (Courts & McInerney, 1993; H. Gardner, 1992; Resnick & Resnick, 1992)。不幸的是,关于过度和不恰当使用标准化测验的批评在某些案例中得到了充分证实。无论如何,所有的教育家都提出了用测验推动课程和测验及课程两者都急需改革的要求,并且在最近的 20 年内这种要求愈来愈强烈。教育变革倡导者认为,需要对课程目标和教学方法和测量工具进行精细的检测。他们认为所有这些都是紧紧联系在一起的。

因为教育改革在哲学上、政治上和应用方面的问题超出了本书的范围,我们只局限于讨论某些可供选择的评定方法。这些方法常常用不同的标题描述,诸如“基于操作的评定”(performance-based assessment)、“可靠的评定”(authentic assessment)和“直接评定”(direct assessment)(参见 E. L. Baker, O'Neil & Linn, 1993; Linn & Gronlund, 1995)。虽然它们强调的重点是不同的,

但它们共同一个中心特征是明显偏向于任务,像前面的“问答题”一样,要求应试者自己产生答案这些项目现在描述为“建构反应或开放式任务”(constructed-response or open-ended tasks),它们与“选择反应任务”(selected-response tasks)相反,后者要求应试者从所呈现的可供选择的反应中选取一种反应,如多项选择项目和 其他称之为客观性类型问题。建构反应项目包括简单的完成作业(如填空)、问题解决、论文写作以及各种各样表现或操作,如使用仪器、发表演说、修理机器等<sup>[6]</sup>。

档案评定(portfolio assessment)提供另一套可供采用的办法。这类评定的主要目的是:使教育评定的过程尽可能有意义和现实。尽管这一术语应用于广泛的评价程序中,但是一个档案经常由累积的记录组成,在相当的一段时间内收集学生在特定领域中的作业样本,如写作或任何其他能说明学生进步的作业(Camp, 1993; Gitoner, 1993; D. P. Wolf, 1993)。档案评定的方法表现出广泛的灵活性,能或多或少地正式实施,使学生和教师之间具有不同程度的合作(这方面的例子见卡尔森(Karlsen)1992年的著作)。

即使从以上的简单综述中,读者仍会得到这样的印象:学习和学生作业的评定方法正在受到极大的关注。这种关注不仅扩展到不同的项目分别测量什么和它们是否能很好地进行测量。这些问题上,而且扩展到了测验项目的其他心理层面。例如泽纳(Zeidner, 1993)调查了学生对项目形式的态度,发现他们喜欢多项选择项目,而不是问答题;刘和宋(Lu & Suen, 1995)的研究表明“基于操作”的评定有利于场独立性的学生而不是场依存性的学生(见第十六章)。另一些研究者研究了测验焦虑和项目类型的关系,发现建构反应测验的分数似乎比选择反应测验的分数更多地受到焦虑的影响(Crocker & Schmitt, 1987)。对可能影响

应试者的动机,影响他们的敏感性和成绩的变量,如评定程序的目的,以及其他几种可能影响建构反应和多项选择测验解释的因素之讨论可以在斯诺(R. E. Snow, 1993)的著作中找到。

与此同时,使用于学校场合的“基于操作的任务”的严格的心理测量属性的经验的研究文献得到了不断积累。<sup>[7]</sup>随着研究项目类型的不同,其研究的进度和结果的方向有较大的差异。有相当数量的研究探讨了“建构反应任务”中记分程序的可信度,像我们在第十六章讨论的“情境测验”一样,它经常包括了等级评定(E. L. Baker 等, 1993; Linn & Gronlund, 1995, 第十章)。当记分规则清晰、完善和评分者受到适当的训练时,评分者之间的信度一般是十分满意的。在另一方面,跨课题和跨任务的测验结果的可概括性一般是低的,这就表示建构反应项目具有较高程度的任务特殊性。与选择反应项目相比,这些项目一般更复杂,允许反应的范围也更广;考虑到这一点,这些发现就一点也不奇怪了。

至于测验的效度问题,研究的基础仍然相当有限,至少对于那些最少限制、最具创新性的“建构任务作业”来说是如此。其中一个首要问题是建构反应项目和选择反应项目测量相同特质或技能的程度。尽管这一问题的证据仍然是有限的,但乔特劳伯(Traub, 1993)对这一问题的综述表明,其相同程度随领域不同而变化例如,当不同项目形式用来测量阅读理解、数学知识时,它们倾向于得出相同的结果,但在写作领域中似乎项目类型对测验分数有显著的影响。

同时,多项选择项目测验仍然广为流行。事实上,不同项目形式被引入教育测验及对多项选择项目的严肃批评,激起了改善多项选择项目测验的努力。关于如何发展、检查和评价多项选择项目的指导性意见已随手可得,对这种形式具体方面的研

究（如选择项最适宜的数目）仍在进行中（Trevizan, Sax & Michael, 1991, 1994）此外，选择反应任务类型的一些变化和改进也正在探究（参见 Linn & Gronlund, 1995, 第八章；Sax, 1991；Sireci, Thissen & Wainer, 1991；Wainer & Kiely, 1987；Wainer & Lewandowski, 1990）。

建构反应和选择反应项目的直接比较工作也已进行（如 Lukhele, Thissen & Wainer, 1994）。当比较的标准是经济、效率和预测效度时，这种比较许多时候有利于选择反应项目，尤其相对于传统问答题更是如此（Anastasi, 1988b, pp. 416 ~ 418；R. E. Bennett, 1993）。然而必须注意到当评定的目标和方法都处在一种不断变动状态时，评价和比较不同评定形式的问题决不是简单的。重要的是要记住，项目形式只是影响一种评定方法公正、准确和整体质量的许多相互作用变量中的一种。评定的目的、内容范围以及所评定的个体特征都需要考虑（E. L. Baker 等, 1993；R. E. Bennett, 1993；Dwyer, 1993；Mislevy, 1993）。例如，失败对应试者以后动机的不同效应可能比其他标准更为重要，尤其是对具有缺陷或其他不利因素的学生更是如此。然而，值得注意的是，没有任何理由认为基于操作的评定会减少在多项选择标准测验中存在于某些少数民族和白人测验分数上的差距。事实上有些研究表明这种分数的距离在建构反应测验中比在选择反应测验中可能更大（Hartle, & Bataglia, 1993）。当前对具有语言和文化差异学生的教育评定问题在凯萨林、阿雷纳、豪厄尔、鲁拉以及辛和贝克（Cancell, Arena, K. W. Howell & Ruecla, 1996；Shinn & Baker, 1996）的论述中得到了进一步的讨论。

### 教育测验的类型

在教育测验服务处（Educational Testing Service）的 1990 年年

度报告中,其理事会预测,在今后的 10 年内,教育测验将发生的变化比以前 50 年内发生的都要大(ETS, 1990)。现在看来这一预测是正确的,在今后的 10 年内它很可能被不断地强调;各种测验正在进行重要的修订,新的评价工具正以较快的速度被设计出来。所以,下面的简单回顾主要集中在教育领域中传统使用的工具类型和每种类型正在进行的一些变革,而不对各个测验进行细致的描述。

**一般学业成就成套测验** 有几套测验工具可以用来测量大多数课程领域的一般教育成就 (general education achievement)。该类测验可用于从低年级到成人的各种水平,但其主要应用范围是小学。这些测验工具一般提供个体在分测验或主要学科领域的分数剖面图;与独立建构成就测验 (independently constructed achievement test) 相比,它们的一个优点是可以进行横向或纵向的比较或同时进行这两种比较,因而可以用同一的常模样本来评价个体在不同学科领域或教育技能测验中的相对位置。个体从一个年级向另一个年级的进步可用单一的分数量表来报告。测验使用者必须检查某一测验是否已经标准化,因此能进行两种或其中一种的比较

尽管一些一般成就成套测验专为小学或中学设计,但大部分这类测验范围已包含这两种水平,偶尔还扩展到了大学一年级。少数测验是单为某一年级所用的,但大部分有几个相互重叠的测验分别使用于不同的水平;有些测验形成了相互配合的系列,对从十一年级到十二年级的学生进行可加比较的测量,如依阿华基本技能测验 (Iowa Tests of Basic Skill)、学习成就和效率测验 (the Tests of Achievement and Proficiency)、依阿华教育发展水平测验 (the Iowa Tests of Educational Development) 就是这种结合;另一个是斯坦福成就测验系列 (Stanford Achievement Test Series)

有些成就成套测验的一个值得注意的特征是，它具有与学业智力测验或学业能力测验的同一常模，主要例子是和第十<sup>早</sup>章<sup>早</sup>阐述过的一种多重水平测验配合的成就成套测验，即斯坦福成就测验系列与奥蒂斯-莱农学习能力测验（Otis-Lennon School Ability Test）相配合；依阿华测验系列，学习成绩和效率测验与认知能力测验（the Cognitive Abilities Test）相配合 加利福尼亚成就测验和基本技能综合测验（the California Achievement Tests and Comprehensive Tests of Basic Skill）与认知技能测验相配合（见表10-1）。在这些情况下使用同一标准化样本，可以对学生在两种类型测验上的得分进行直接比较通常，这两种测验相关较高，个体会获得很相似的分数；如果学生在某一种测验中的分数明显高于另一种，这就需要探讨导致这种差异的可能原因。学业成就成套测验主要测量个体在学校基础课程中所学到的东西，认知技能测验测量个体在学校内外学到的广泛的知识能力。任何显著的成绩差别可能反映了某种特殊能力或能力缺陷的影响或动机、兴趣、态度等非认知因素的影响。个体的经验背景常常提供了导致测验成绩出现异常差异的条件线索。

各种学业成就成套测验在其编制程序的技术水平上具有显著差别。然而，作为一类测验来看，它们都符合测验编制的高标准，尤其是常模样本的大小和代表性、信度和内容效度。对项目编写的详细具体的说明经常先要进行彻底的项目分析，包括项目反应理论方法的应用一些特殊的程序常常用来避免性别和种族偏见。所有这类测验一般都包括对阅读、语言和数学方面基本技能的评定，以及自然学科和社会学科中知识内容量的评定。有些测验还包括一些用来测量运用各种信息源进行学习的能力或技能。此外，为使测量更为灵活和提供更多的信息，几个主要标准化成就成套测验的出版者现在提供了更为多样的测验

项目和选项。现在正在使用的是开放式项目和范围更广的选择反应项目，旨在测量在较为有意义的情境中高级思维技能和心向。出版者通过使测验的内容适应地方课程的要求，使用项目内容、形式以及记分选择的各种不同的混合，从而增加了其灵活性。他们在测验和教学材料之间也提供了更多的联系桥梁。<sup>[8]</sup>

**基本技能的最低能力测验** 在过去的 20 年里,许多高中毕业生在阅读、写作和计算技能方面表现出来的低能力水平受到了越来越多的关注,导致了对基本技能测验的广泛需要,并且把这种测验作为证明获得最低能力的工具和是否颁发高中文凭的基础。这一要求引起了激烈的争论,大多数的反对意见集中在最低能力测验的误用和对其结果的错误解释以及可能导致教育的僵化和官僚主义的控制。<sup>[9]</sup>大多数州已确定了最低能力测验的政策,而它们的政策和程序在测验实施的时间、年级水平、测验结果的特殊使用、测验的编制或选择中地方自主性的性质及程度等方面有着较大的差异。而用来作出是否颁发毕业文凭的测验必须能反映其所覆盖的具体课程内容。正因为如此,恰当的测验现在由学校、学校系统或州机构自己编制,但它们是在测验出版者的帮助下进行的;测验出版者能提供符合地方目标的受过技术训练的人员、大量的项目库和报告等服务。该种测验也许包括前面讨论过的标准化成就成套工具的某些组成部分,它们或许是为某一地方而制定的。

近年来,确定基本技能掌握的兴趣点已扩展到了成人。<sup>[10]</sup>大量辍学率的累积效应、高中毕业生的低能力水平以及非英语国家移民的大量增加,引起了对美国劳动力在世界市场上竞争力的关注。1992 年,在教育部的赞助下,由 ETS 进行的全国成人读写算能力的调查结果加重了这种担忧。这个调查发现约有一半的美国人处在最低的二级水平(一共有五级)(Kirsch, Jurge-



blut, Jenkins & Kolstand, 1993)。

专为成人设计的基本技能的测验，一般和成人的教育类别、在刑事机构中进行的教育计划和职业训练计划联系起来编制。其中一个突出的实例是成人基础教育测验 (Tests of Adult Basic Education, TABE, Forms 7 & 8, 1994)。TABE 是一种包括阅读、语言、应用数学在内的五种不同内容领域，具有五级难度水平的测验，其结果是“常模参照分数”(norm-referenced scores)和可用于诊断的基于能力的信息；除了正规的测验形式外，TABE 还有特殊的版本适用于某个工作环境，也有用于测量说西班牙语成人的基本语言技能的西班牙语修订本(TABE Español)。

教师自编的课堂测验 数量最多的测验无疑是教师自编的用于他们自己课堂的测验，它们包含具体学科内容或学科的多个部分内容。众所周知，在同一学科和同一名称下各种课程的内容有很大差异，尤其是在高中或高中以上水平更是如此。在这种条件下，没有一种外部标准化测验是可以满足需要的。然而，地方性的课堂测验可以采用专业测验编制者的技术和积累的经验而得到实质性的改善。课堂测验的编制可以分为三个主要步骤：(1) 计划测验，(2) 项目编写，(3) 项目分析。在第七章已经描述了适合小团体使用的项目分析简单技术，下面的段落是其他两个步骤的简单阐述。<sup>[11]</sup>

直接埋头于测验项目编写的编制者可能会编制一些偏向性的测验；如果没有事先的计划，一些领域将会过多而另一些领域却不曾涉及。有些课题一般比另一些课题容易准备客观性的项目。需要简单事实回忆的项目的准备要比批评性评价的、不同事实整合的、在新情境中应用原理的项目容易设计；因而没有仔细计划而建构的项目测验，可能承载了一些相对临时的、不重要的材料。对选择反应测验普遍的批评，是针对一些编制不良的

测验过分强调机械记忆和琐事细节。

为了防止项目覆盖面的偶然不平衡性,在编写任何项目前应该拟订测验的详细说明。对于课堂测验来说,其说明须以教学目标大纲和其覆盖的内容为基础,应该反映关于这些方面相对重要性的判断,而这是通过分配给每个课题和目标的项目数来实现的。<sup>[12]</sup>测验的编制者必须确定适合于特定材料最适宜的项目形式。在本章的前一节,特地提到的客观及开放式项目使用在大规模、标准化的测验程序时的相对优点,在课堂测验编制中也需要加以考虑。最后,有效地编写项目的许多实践准则,已经在多年的项目准备的经验和反应评价经验的基础上有了系统的表述;参考这些总结性的建议有利于任何准备编制课堂测验的人的工作(Ebel,1979 第四至第九章;Haladyna,1994,第四至第六章;Linn & Gronlund,第六至第九章;Millman & Green,1989)。

**大学水平测验** 一些测验和测验方案是为大学生的入学、编班和咨询发展起来的。其中一个杰出的测验是大学董事会的学业评定测验方案(Scholastic Assessment Tests program, SAT)。它现在由两部分组成,即 SAT-I 和 SAT-II。前者是推理测验,用来代替学业能力倾向测验中的言语和数学部分;后者是学科测验,用来代替 SAT 前版本中的成就测验。<sup>[13]</sup> SAT-I 主要由测量言语和数学能力的多项选择题组成,作为高中生的成绩等级和其他信息的补充,用来评定学生对完成大学水平学习的准备程度,而 SAT-II 主要用来评价学生具体科目领域的知识,如文学、化学和世界历史,可以用于招生和编班。

SAT 在其内容、形式和名称上都发生了变化。如 SAT-I 的言语部分逐步强调了批判性的阅读和推理,同样,在数学部分现在要求提出而不是选择某些答案,并且在保持当前标准的同时,在整个数学部分允许使用计算器。SAT-II 测验也发生了变化,

现在通过使用听的成分和写的样本包括较为直接的技能评定。

1995年4月,SAT测验不再按照1941年以来的固定参照组的方法来报告分数(见第三章),SAT的分数量表进行了“重新聚中”(recentered),从90年代起采用新的参照组,因而平均成绩再一次用500左右的分数来表示。<sup>[14]</sup>“重新聚中”使得SAT的分数更精确可靠,尤其是分数处在量表的最高端和最低端时;还有量表分数的解释也变得容易起来,如现在的言语和数学分数可以不参照百分等级位进行直接的比较,因为两者都进行了重新组合而且,因为500分是200分和800分之间的中位数,所以“直觉”(intuitive)的均数将和实际的均数相同。大学董事会所采用的测验为了保持两个量表之间的连续性,颁布了表格和其他工具从而简化了从原始分数到“重新聚中”量表分数的转换。SAT测验修订本的信度、难度、完成的比率的信息,在专为咨询员、入学办公室工作人员而准备的使用手册中,在其他ETS和大学董事会的研究报告和出版物中都能找到。<sup>[15]</sup>一项比较传统的SAT和典型的SAT-I版本研究发现,新的版本比以前的版本能较好地预测大学一年级学生的年级平均成绩(Hale, Bridgeman, Lewis, Pollack & Wang, 1992)。修订后SAT测验效度的其他证据将包含在90年代后期即将出版的技术性补充读物中。

另一个开始于1959年的全国性方案是美国大学测验方案(America College Testing Program, ACT, 1995~1996)。起初主要用在州立大学系统中,后来得到了快速的发展,现在也用于全国的许多学院中。现行的ACT包括4个分测验,英语、数学、阅读和理科推理。这个测验反映创立者林奎斯特(E. F. Lindquist)的观点,提供了一套大学作业的样本。它与传统的能力倾向和成就测验有部分重叠,着重于在大学阶段有令人满意的成绩所需的基本智力技能。ACT评定的非认知组成部分是:高中课程/成绩

信息问卷,ACT 兴趣问卷,一份学生简介材料,询问学生的志向、计划和才能以及其他的背景信息。ACT 在历史上一直未达到 SAT 设立的技术标准,然而其效度的资料可以和应用于相似情境的其他工具相比美。

应该注意到,诸如 SAT 和 ACT 的测验并不试图代替中学的成绩来预测大学的成就。中学的成绩和大部分测验都能预测大学的成就,或者中学的成绩的预测性稍优于大部分测验;然而当测验分数和中学的成绩相结合时,它们对大学成绩的预测性便得到了加强。这种加强的部分原因是一种统一的、客观的测验能校正不同中学评分标准的差异。而且,这种测验不易受到个人偏见或其他因素的影响。这些偏见和因素可能参与了课程成绩的确定。

等同于大学考试的一些特殊成就测验的使用也在增加。在某一领域进行额外准备的中学生,可以参加大学董事会施测的高级编班项目(the College Board's Advanced Placement Program, APP)中的测验,在一门或几门科目中获得优势,从而取得入学资格。由大学董事会实施的大学水平考试大纲(College Level Examination Program, CLEP)也正在发展中,这一项目的总目的在于:促进通过考试而承认大学学分的进程;提供一种全国性的评定系统,用来评定那些通过自学或其他非正规教育而达到大学教育水平的个体。另一个相似的测验系列是由美国大学测验项目(the American College Testing Program, ACT)实施的“ACT 熟练性考试计划”(the ACT Proficiency Examination Program),它虽然包括了一些学术科目,如解剖学、生理学、变态心理学,但主要覆盖的是职业领域,如护理和财会等。

**研究生院入学测验** 入学测验的使用扩展到了研究生院和专业性学校。为此目的而设计的大部分测验是把一般智力和学

业成就测验结合起来。一个众所周知的实例是美国研究生资格考试(Graduate Record Examination, GRE)。这一测验产生于 1916 年,是卡耐基教学促进基金会和 4 所大学的研究生院的一项联合计划 现在得到了很大的发展。它是在研究生入学考试委员会的指导下,由教育测验服务中心负责实施。在进入研究生院之前,学生在全世界 100 多个国家指定的中心参加考试。这一考试的结果帮助大学作出录取和编班决策、选择奖学金、补助金和特殊职务接受者的人选。GRE 包括一般测验和不同的专业领域的学科测验。<sup>[16]</sup> 现行的一般测验分别得出 言语、数量和分析能力的分数;学科测验有 16 个领域,包括生物学、计算机科学、法语、数学、音乐、政治科学、心理学等等。GRE 的心理测量特征在最近出版的使用指导中作了说明 (GRE 1995 ~ 1996 Guide) 一般说来,学科测验分数比一般测验分数或大学本科二年级平均成绩,更能预测一年级研究生二年级平均成绩;但三者的结合能提供更高的预测效度。在不同的领域中它们的多重相关系数为 .45 ~ .60。

1992 年 10 月, GRE 项目对传统形式的一般测验开始实施计算机化版本; 1993 年 11 月, 引入了适应计算机的一般测验。虽然在 GRE 一般测验的计算机化的过程中, 一开始就碰到了可靠性风险等难题, 但计算机化的优点是, GRE 测验委员会在 1999 年可以抛弃纸笔测验。一般测验也正在进行修改, 以便包括写作、数学推理测验和一些建构反应的问题 (Update on the NewGRE, 1995)。

**诊断和预测测验** 与前面讨论的一般成就成套测验和其他成就测验不同, 本节讨论的测验, 旨在分析个体在具体的学科范围内的优点和弱点, 用来揭示其困难产生的原因。大部分的诊断工具是个别施测的, 因而常常认为是临床测验工具; 然而, 有

些是我们前面讨论过的几个主要成就测验的单独组成部分，并企图用于团体施测。大多数发行的诊断性团体测验，能测量阅读、数学和语言技能，并能提供常模和内容参照信息。这种方法的例证是斯坦福数学诊断测验（Stanford Diagnostic Mathematics Test）、斯坦福阅读诊断测验（Stanford Diagnostic Reading Test），加利福尼亚阅读和数学诊断测验（California Reading and Mathematics Test）。这两种测验系列的出版者也引入了单独的工具用于评定和诊断写作技能。斯坦福写作评定项目和 CTB 写作评定系统都采用各种写作形式的直接样本（如描述或记叙写作），并提供了几种记分的选择。

与所有诊断测验的使用有关，值得特别强调的一点是：学习无能的诊断和随后的补偿教育计划，是一个受过训练专家的分内职责。没有任何一套诊断测验能够完全达到这一目的。严重的学习不能的诊断和治疗，需要进行细致的临床个案研究，最好是跨学科的研究，包括感觉能力和动作技能发展、个体的医疗和健康状况、完成教育的历史这样一些补充信息、家庭和家族的背景材料，并彻底调查可能的情绪障碍。虽然调查和团体诊断测验可以用来确定需要进一步注意的个体，但学习不能的诊断和治疗需要特殊的技术；其中的一些方法在第九章已经讨论过，本章将进一步讨论和临床测验相关的问题。

某些旨在用于教育情境的测验，实质上也是预测的工具。因为它们起着能力倾向测验而不是成就测验的作用。同时，它们在内容上常常像成就测验，因为它们所预测的经常是特定学习课程中的表现。这种方法的典型例子是欧利恩斯—汉纳的代数诊断测验（Orleans-Hanna Algebra Prognosis Test, Hanna, Sonnenschein & Lenke, 1983）。在这一测验中，给学生上一组简短的代数“课”，然后立即测验他们学到了什么。这一测验由作业样本

组成,学 生以后的课程学习可以从他们在作业样本学习中的行为表现来预测。一个不同寻常的、实验性的诊断测验的实例是美|国人事管理处( U.S. Office of Personnel Management )和国防部设计的人工语言测验。这一测验用来预测个体学习一种新语言的能力( Diane, Brogan & McCauley, 1991 )

自 80 年代以来,另一种完全个别化定向的评定方法受到了越来越多的关注。实质上,这种方法遵循测验—教学—测验的程序,一般被描述为动力性或受指导的评定,它是和补救性教育联系在一起的在个体与具有三种角色功能即测验者、指导者、临床工作者的某个专业人员的一对一的关系中,通过观察个体的学习情况来评价他或她的学习潜力这一方法的主要倡导者是福伊尔廷( Feuerstein, 1979 )。布朗、坎皮尼、韦伯和麦基利( A. L. Brown, Campione, Webber & McGilly, 1997 )以及李兹( Lidz, 1987, 1997 ) 讨论了其他几种相关的方法 因为它具有显著的临床特征,在本章的后一部分将更充分地讨论动力评定问题

基于课程的测量代表了另一种目的在于把评定和干预联系在一起的技术( Deno, 1992; L. S. Fuchs, 1993; Fuchs & Deno, 1991; Shinn, 1989 )。有些基于课程的评定方法可以相当不正规,但也有一些包括了学生在诸如阅读、拼写、算术计算等基本能力方面成绩的标准化测验。不同模式的比较请看辛、罗斯菲尔德和克努森( Sin, Rosenfield & Knutson, 1989 ) 的论著。和传统心理测量工具的推断和常模参照性方法相反,这些技术的共同特征是较强的行为定向,与组成初等教育普通课程的一些任务有直接关系。基于课程的评定主要用于特殊教育。

早期儿童的教育评定 在过去的 20 年里,出版了许多用于测量年幼儿童教育发展的新工具。这一活动的数量和性质受到几种因素的影响(见第九章和第十二章)。早期认知发展的研

究、学前教育计划的迅速发展、对文化上的不利条件影响儿童接受学校教育能力这一问题的广泛关注起了主要的作用。<sup>[17]</sup>有些测验主要用于测量早期儿童的教育结果，因而它具有成就测验的功能；还有些测验是预测性的工具，用来评定儿童接受一年级教育的准备状态。然而，这两种类型的测验常常微妙地融合在一起，每一种通常为其中的一个目的服务。

入学准备实质上是指个体所获得的能使他最大程度受益于学校教育所需的技能、知识、态度、动机和其他适宜的行为特质。这些先决条件就是麦克冯特和可克 (J. McV. Hunt & Kirk) 所谓的儿童“入门技能”(entry skill)，即儿童处理一年级面临的“教—学”情境所必需的技能。学习准备包括通过成熟和以前学习而达到的身体和感觉运动发展最低水平。现在逐渐强调知识和技能的层次性发展，简单概念的获得为任何年龄的儿童学习较为复杂的概念作好准备。

学习准备测验一般在入学时进行。它们和小学低年级的智力测验有许多共同之处，不过更强调对学习阅读起重要作用能力能力。现在也注意了数学思维的先决条件和学习写字所需的运动控制能力。其中常常包含的技能是：视觉和听觉的辨别力、运动控制力、听觉理解、词汇、量的概念及其他一般知识信息。一个广泛使用的学习准备测验是大都市准备测验第 6 版 (Metropolitan Readiness Tests, MRT6; MRT 早期版本的回顾见马布利 (Mabry, 1995) 和斯托纳 (Stoner, 1995) 的著作)。对入学准备的一种不同的考察途径着重于儿童对普通关系概念的理解，如波恩的基本概念测验修订版 (the Boehm Test of Basic Concepts - Revised) 和布莱克恩的基本概念量表 (the Bracken Basic Concepts Scale) (综述见 Fitzmaurice & Witt, 1989; Linn, 1989; Turco, 1989; Ysseldyke, 1989)。对入学准备概念的讨论及其评定中的问题可



在格瑞德勒(Gredler, 1992) 的著作中找到。

总评 不管在美国还是世界各地,教育领域都是一个热点。教育测验正反映了这一点。在可预见的未来中,其变化可能会持续和加剧。<sup>[18]</sup>许多观察家同意,需要进一步把测评和教学结合起来,这样,教育领域中的各方面才能更好地相互补充,以有利于学习者(H. Gardner, 1992; Nitko, 1989)。<sup>[19]</sup>大多数人也认识到,测验和教育的方法总是经常随着理论和经验研究的进步而不断改善;没有单一的一种测评工具或教学方法是完美无缺的,每一种都有其局限性;而且不同的学生需要使用不同的方法。所以探索改进方法的工作必须而且也将进行下去。

## 职业测验

心理测验一般作为辅助工具用于职业决策中,包括个人咨询和机构人事选拔及分类。在本节中我们将讨论由负责人事选拔、分配和评价机构所认定的个体职业资格的测验应用问题。

在商业和工业企业中,在联邦、州、地方各级政府内,在军队不同的部门里,各级组织几乎把每一种能得到的测验都用于人事决策。多重能力倾向测验、特殊能力倾向测验,如同第十六章描述的情境测验那样,常常是为职业测验的目的而开发出来的。人格问卷(第十三章)和生平资料(第十六章)也得到了进一步的运用。兰迪(Landy, 1994)等人对人事选拔和分类中使用的测验和其他工具进行了简要的概述;对这一主题综合的处理方法,能够在最近由拉姆齐、沃克和哈里斯(Rumsy, Walker & Harris, 1994)沙米特和波曼等(Schmitt & Borman, 1993),柯勒、法尔和斯密斯(Schuler, Farr & Smith, 1993)编写的三本书中找到。在由

邓内特和霍夫 (Dunnette & Hough, 1990 ~ 1992)<sup>[20]</sup> 主编的手册中,有几章详细地讨论了工业和组织中使用测验的主要方面。工业和组织心理学学会 (SIOP, 1987) 准备和采纳了一系列关于人事选拔程序的确认和使用的原则。为了在人员选拔、发展和评价中能更好地使用所有的人事选拔程序,这些原则和标准化测验紧密相连。此外,现行的《测验标准》(Testing Standards, AERA, APA, NCME, 1985) 和它们的修订本(见第一章)都有一章阐述就业情境中的测验问题。在《测验标准》的两种版本中,职业测验的另一种主要应用是给那些具有资格从事某一行业和职业的人,颁发执照和资格证书。<sup>[21]</sup>

正如我们对教育测验的处理方法一样,本节我们将集中对在其他章节已讨论过的应用广泛的测验工具之外的特地为职业目的而设计的那些测验。我们也将简单考察就业情境中测验集合及有效化过程中所涉及的一些程序和问题。

### 就业测验的效度分析

从雇主和雇员的立场来看,把个体放在合适的能胜任的岗位上的重要性是显而易见的。有效的安置也意味着某种与工作无关的特性,不管其好坏,均不应影响选拔决策。如果,机械能力测验比某种工作需要更高的阅读理解水平,那么使用这种测验将不会导致这一岗位人员的有效安置。测验效度必须根据测验的特定用途加以确定这一简单的心理测量学事实早已为大家熟知。随着社会对文化和教育条件处于不利地位的少数民族工作安置问题的广泛关注(见第十八章),这一需要变得愈加紧迫。一种无效的或包括了与工作无关要素的测验,会不加考虑地把能满足这一工作要求的少数民族不公平地排除在外。

组织和社会关心的另一个相关问题,来自于工作效率和有

效的选拔工具效度之间所呈现的关系。我们在第六章阐述了用于评价这种关系和结果的程序。估计的生产效率得失与人事选拔程序效度的提高和降低有实质性的联系。在有许多雇员的组织中，如政府机构，这种得失的累积效应是十分巨大的，值得密切关注。

几十年来，人事心理学中流行的观点是：选拔测验须按照工作绩效的效标进行全量表效度分析。这种效标——预测效度分析的具体程序在第五章和第六章已作过讨论；然而，在大部分的情境中，对整个量表的纵向效度的研究是不现实的。即使拥有大量的雇员样本，具备相当理想的条件，某些实际上的局限性还是十分明显的(参见 Anastasi, 1972; J. T. Campell, Crook, Mahoney, & lock, 1973)。鉴于全量表地方性效标——预测效度分析实施中的实际困难，一系列替代性程序引起了关注。

**绩效评定的总体程序** 人事选拔中使用的一种评定程序是把所有的工作情境尽可能看成相似的。然而这种相似永远不可能是完全的。见习职位最类似于真实的工作情境，但是即使在这种条件下，工作期限的短期性和了解职务的见习性可能会以许多方式影响工作行为。工作样本 (job sample) 代表了另一种接近实际工作绩效的企图；在工作样本中，任务实际上是要完成的工作的一部分，但是任务和工作条件对所有求职者都是一致的。有些工作样本测验是专为某些特定的工作而编制的。工作样本的代表性及任务与工作条件的相似性，需要慎重地加以考虑。熟悉的实例有驾驶测验，办公技能的标准化测验，如打字、各种行业机器的操作。

有些测验采用情境模拟的方法来重现工作中的操作功能。情境模拟微妙地和工作样本融合在一起；它的例子可以从细微的穿孔操作到为火车司机和飞机驾驶员设置的模拟情境。在由

NASA 管理的美国空间计划和许多军事领域里，情境模拟已经应用于训练和测验中。

还有主要用来评价管理或行政人员的评定中心技术（见第十六章）（Bray, 1982; Finkle, 1983; Moses, 1985; Thornton & Byhan, 1982）。这种方法的一个显著特点是，它包含了情境测验。如在许多领域中用来测量行政人员的“篮子技术”（In-basket）（N. Frederiksen, 1962, 1966; Shapira & Dunbar, 1980）。模拟管理者办公桌上可以找到的、熟悉的“文件篮”，这一测验提供了一系列仔细准备好的来信、备忘录、报告、需签发的文件和一些相似的项目。在进行这一测验之前，应试者有机会研究背景材料和与假定工作有关的信息。在测验中，其任务是要求应试者像真实的工作情境一样处理篮子中的所有事项；所有的活动必须进行笔录，可以包括书信、备忘录、决议、计划、指示、获得或发送出去的信息、会议记要和任何其他类型的记录。其他有些评定中心技术可能运用角色扮演、团体问题解决和职业游戏。它们一个共同的特点是使用多种评定者或同事评定的方法。许多评价的特性是属于动机、人际技能和其他人格变量方面的。

虽然它们与工作相关的证据至少部分依赖于工作的相似性，但这些总体绩效评定方法也参照了一系列的标准来进行单独的和各种结合式的评价（参见 Landy 等, 1994; Schmidt, Ones & Hunter, 1992）。

**工作分析和工作要素方法** 内容效度分析在人事选拔测验中的应用逐渐引起人们的兴趣。在其所有的形式中，这一效度分析依靠全面的、系统的工作分析（McMormick, 1979）。为了使工作分析有效，必须确定区分不同工作的具体要求。可同样应用于大多数工作上的含糊的一般性描述，对于这一目的是没有用的。为了获得全面的工作活动情况，工作分析者可以依靠多

种渠道的信息已出版的训练和操作手册、操作记录,尤其是某个问题的专家,如某一领域的管理者、指导者、有经验的工作人员,常常是咨询的对象。

有效的工作分析也需集中于那些能显著区分好的和差的工作人员的行为表现方面赫尔(Hull, 1928)在其经典的著作《能力倾向测验》中强调了工作绩效区分的重要性。这点后来在弗拉纳根(J. C. Flanagan, 1949, 1954)的著作中得到了再一次的强调,他提出了关键事件技术。实质上这一技术要求,对代表满意或非满意工作人员特征的工作行为的具体实例进行事实性描述。

对关键工作要素的集中注意,导致了工作要素方法的发展,它用于编制测验和显示其内容效度(McMormick, 1979, 1983; McMormick, Jeanneret & Mecham, 1972; Primoff, 1975; Primoff & Eyde, 1988)。普里莫夫(Primoff)和他的合作者在美国人事管理处(前美国民政服务处)完善了并广泛使用这一方法。实质上,工作要素是描述关键工作要求的单元。虽然工作要素方法的各种变式在其具体程序上有着差异,但都按特定行为要求对工作活动进行描述,从而直接形成了测验项目。反过来,具体行为的阐述可以按比较广泛的范畴或构念加以分组,如计算的精确性、手臂的灵活性、视觉的辨别率,或压力下的工作能力。旨在按照广泛的行为构念对工作绩效进行一般分类的研究在逐步增加(Fleishman, 1975; Quaintance, 1984; Reilly, 1992b)。

工作分析的方法有助于加速测验在各种表面上不同的工种中的有效应用。这一点由弗莱希曼的工作分析调查(the Fleishman Job Analysis Survey, F-JAS)和工作关键系统(the Work Keys System)所证明。F-JAS是一种工作分析工具,从执行某一工作所需的知识、技能和能力方面来描述工作。F-JAS带有72种量表,其中有52种包括了精心确定的认知、心理运动、身体和感

知领域的能力,大多数是和现存的测验有关 (Fleishman & Mumford, 1991; Fleishman & Reilly, 1992a, 1992b)。其余的 20 种量表是关于人际/社会 and 知识/技能领域,现在仍处在发展之中。另一方面,最近由美国大学测验计划 (the America College Testing Program, ACT, 1995; Scruggs, 1994)设计的工作关键系统,着重于在一段适当的时间内能教会的一套少得多的的一般工作技能,如找出信息来源。然而,在这一框架内该系统提供了相互配合的一组工具用于:(1) 工作分析和工作描述 (job profiling);(2) 技能水平评定;(3)对个体、教育者和雇主作出反馈;(4) 在训练或课程教育的实施中提供指导性的帮助。

工作分析是工业心理学中最古老、最有活力的方法之一。由于计算机技术的发展使得工作资料的收集和分析愈加经济,<sup>[22]</sup> 工作分析在就业测验效度分析中的应用继续拓展。此外,许多通过细致的工作分析获得的资料还可以派作其他用处,如在市场上或工作设计中,确定某一工作的价值 (参见 Campell, 1994; I. L. Glodstein, Zedeck & Schneider, 1993)。

**绩效预测 局部效标——预测效度分析**(local criterion-prediction validation) 实施中存在的实际困难 (见第五、第六章) 导致了对这一问题研究的相对不足。许多组织要求对将来的工作绩效进行预测,以便能作出正确的人员选拔和安排决策。<sup>[23]</sup> 为了达到以上目的,综合效度分析和效度一般化技术逐步变成可供选择的方法。这两种方法都能在缺乏局部效度分析的情况下对某一工作测验的效度进行估计。随着经验资料的积累,两种方法都应成为所评定的结构的性质提供集中的证据 (J. P. Campell, 1990a)。

综合效度分析的概念以工作要素方法为基础,有可能确定许多不同工作的共同的技能、知识和其他操作要求。综合效度定义为“在特定的情境中,从对工作要素进行系统的分析而对效

度作出推论,确定这些要素的测验效度,以及把要素效度组合成一个整体(Balma, 1959, p. 395) ”实质上,这一技术包括三个步骤:(1)细致的工作分析,确定工作要素及其在特定工作中的相对权重;(2)对每一测验进行分析和经验研究,以确定其在多大程度上能测量执行每一工作要素的熟练性;(3)从这些要素在工作和测验中的权重,综合地找出有关某一工作的每个测验的效度。普里莫夫(Primoff, 1959; Primoff & Eyde, 1988)发展了一种计算这种效度的统计程序,被称之为工作相关系数(J-coefficient),这一程序实质上是第六章讨论的多重回归方程方法的适当改变。霍伦贝克和怀特曼(J. P. Hollenbeck & Whitemer, 1988)、莫斯霍德和阿维(Mosshodder & Arvey, 1984)描述了综合效度分析的其他方法

在第五章已描述的、最初由施米特和亨特(Schmidt & Hunter, 1977)发展起来的效度一般化的程序(validity generalization procedures),为人事选拔测验的效度分析提供了另一种途径。这种方法基本上能通过元分析技术,把以前的效度研究结果应用于新的情境(Schmidt, Hunter, Pearlman & Hirsh, 1985)。事实上,这种通过元分析而得到的结果概括化的实际程度引起了一些人的质疑批评家强调,不同工作情境和评定参数的方法问题都存在着差异反过来,这又导致元分析方法的完善,加强了其被接受和应用的程度虽然仍然存在一些争论的观点,但它还有较大的发展空间。效度一般化的方法十分有利于职业测验的理论、研究和实践,这一点是不可否认的(参见, L. R. James 等, 1992; Landy 等, 1994; Schmidt 等, 1993)。

**工作绩效的标准** 人事选拔和分类领域中一些很有前途的工作,来自于对标准的重新关注。我们可以回想起,从第五章开始就有大量的指数可以看成是衡量标准,这主要看对标准是如

何定义的。然而,在各个效度分析的研究中,不管预测过程的目的如何,对工作绩效单一的方便的度量,一般都用来代表特定的标准。直至现在,尽管几十年来反复要求慎重考虑标准问题(参见 L. R. James, 1973; Tenopry, 1986; Wallace, 1965),但对这一关键问题所做甚少然而,在过去的几年里,有几项研究是针对绩效概念的清晰化及更好地理解其决定成分的 (Borman, 1991; Campell, McCloy, Oppler & Sager, 1993; B. F. Green & Wigdor, 1991; Schmidt & Hunter, 1992)。

另一种有望起重要的启发式作用的工作绩效新模型是坎贝尔 (J. P. Campell) 和其同事与美国军队人员选拔和分类计划 (the U. S. Army Selection and Classification Project, Project A. J. P. Campell 1990a, 1990b, 1994; McHenry & Wise, 1990) 一起发展起来的多重因素理论。坎贝尔模型考虑到了工作绩效的多重层面的性质,并对这一主题下的各种要素进行分离。最初,这一模型对工作评价中工人可控方面如工作操作中的行为本身和不可控方面如工作操作的结果(效果)、其相对成本(生产率)、组织对这些方面所赋予的价值(功用)进行了基本的区分。对于工作操作本身,这个理论假设任何工作包含了多种操作成分(任务),每个操作成分的决定因素是由工人的技能、知识和动机要素的各种结合组成。还有,工作操作的每种决定因素、某些或多或少能说明的先行影响因素(如训练、强化和个体特征),可以通过影响个体的知识、技能和动机水平来间接影响操作。此外工作操作的决定成分之间是相互作用的,对操作产生必然的影响。

虽然工作绩效的多重因素理论仍然处在不断发展中,但它和工作行为评价中大多数概念和方法的进展是相容的 (Borman, 1991)。目前,这一模型确定了工作操作中 8 种一般因素,包括了努力的坚持性、个人的修养、领导特性、特殊工作任务的熟练



度和其他种类任务的熟练度。这些因素被假设有较大的包容性,包含了《职业名称词典》中所描述的所有工种所需的主要要素此外,这一模型除说明影响个体操作的先行条件外,特地说明了工作操作中三类具有决定意义的个体差异,即陈述性知识、程序性知识和技能、动机。这种综合的、清晰阐述的理论结构,必须在广泛的工作绩效结构的研究中证明其可应用性

### 测验在职业中的使用

正如前面讨论所提示的那样,不能离开特定的目的、情境和在一定条件下所涉及的人员,来考虑人事决策中测验使用的恰当性。<sup>[24]</sup>我们也必须注意到,为了讨论的方便,可以对测验按其类型归类,但在实践中,知识、能力、技能和人格特质之间的分界线不总是很清晰的。因此像某一些人 (Liubinski & Dawis, 1992) 所建议的一样,把工作行为看成是由反应能力决定的或许会更有效。此外,虽然一个测验的效度经常是单独评述的,但各种测验很少就其本身加以应用。大多数的基于测验的人事决策都采用一种或多种测量相结合的方法和一些其他种类的评定工具,如面试或背景资料。<sup>[25]</sup>在记住所有的这些告诫的条件下,我们进而考虑职业情境中测验的使用问题。

**学术智力的作用** 智力是一个范围较广的术语,有着许多定义。毫无疑问,在不同的文化背景中,在不同的历史时代里,在不同的生活阶段,智力的组成成分各不相同(见第十一、十二章)。传统的智力测验,虽然只包括一组范围有限的、可鉴别的认知技能和知识,但证明了其对个体在现代技术社会所要求的学术和职业活动中的表现具有广泛的预测作用;因为它主要涉及这种社会中正规学校教育过程中发展起来的知识和技能。这类能力经常被描述为学术智力或学术能力,其内容主要包括

语理解、数量推理和抽象思维的其他方面能力。

众所周知，学术智力测验中的成绩和所受教育的总量有着显著性的相关。似乎可以确定，教育的要求包括了求职者在重要的认知能力和知识群方面所需的条件。然而，这一解决办法在应用中存在许多困难，受教育的总量是个体认知发展状况一种间接的指标，两者的相关十分不理想。仅仅接受正规学校教育不能保证每一个体同样学到所教的内容；而且一般通过学校教育而发展起来的知识和技能，也可以通过其他途径而获得。因而，对个体比较公正的做法是，测量他或她的知识和认知技能，而不是根据正规教育总量接受或拒绝求职者。

在商业性出版的工具中，几种简短的学术智力测验是为工业领域的使用而特地开发出来的。温德利克人事测验（Wonderlic Personnel Test, Inc., 1992）是其中的一个实例。它开始作为一种早期团体智力测验（the Otis Self-Administering Tests of Mental Ability）的修订版，有 50 个项目，费时 12 分钟，它包括言语、数字、空间等不同项目类型，以一种螺旋形汇编的形式呈现，并产生单一的分數。它具有多种形式，在几十年的使用过程中，积累了各种不同职业组的大量常模，其训练和工作成功的预测值也得到了很好的证明（概述见 Belcher, 1992; Schmidt, 1985; Schoenfeldt, 1985）。

由于效度一般化的研究而重新引起了关于一般学术智力测验在人事选拔领域中潜在用途的兴趣（参见 Hunter, 1984）。研究发现，言语和数量推理测验对各种不同的工作具有一些预测效度，这是特别有意义的。此外，对需要更多决策和信息加工的工作来说，其预测效度也有所上升。然而，虽然一般认知能力测验，对工作绩效，尤其是复杂工作的绩效预测可以起很大的作用，但如能和其他变量的评定结合起来，其预测精确性将得到加

强,其中主要是特定工作所需的专门技能和知识(包括心理运动技能和隐含的或程序性知识),以及诸如气质和态度等非认知变量(参见 Ackerman, 1992; J. P. Campell, 1990b; Carroll, 1992; Kanfer, Ackerman, Murtha & Goff, 1995; Sternberg, Wagner, Williams & Horrvath, 1995)。个中的有许多在军队和某些政府机关中使用人事分类测验的大规模研究项目中得到了证实。

特殊项目的能力倾向成套测验<sup>[26]</sup> 一般能力倾向成套测验(the General Aptitude Test Battery, GATB)是由美国职业服务中心开发的,供政府职业服务机构的职业咨询师使用(美国劳工部,1970)。一些非盈利组织如高中、大学、监狱,可以获得该工具。现在 GATB 包括 12 个测验,其中的 4 个要求使用简单的仪器,其余 8 个是纸笔测验,完成整个测验大约需要 2 个半小时。它产生 9 种因素和 3 种主要组成成分的分數,所有的因素和成分见表 17-1。

表 17-1 一般能力倾向测验(GATB)的因素和成分

| 因 素            |                |                  |
|----------------|----------------|------------------|
| G. 一般学习能力      | S. 空间能力倾向      | K. 动作协调性         |
| V. 语言能力倾向      | P. 形状知觉        | F. 手指灵活性         |
| N. 数字能力倾向      | Q. 文书知觉        | M. 操作灵活性         |
| 成 分            |                |                  |
| 认知 = G + V + N | 知觉 = S + P + Q | 心理运动 = K + F + M |

GATB 分數的使用可以通过两种不同的方法来进行。第一种是对相对同质的一组工作所要求的最重要的能力倾向,使用多重临界分數 (multiple-cutoff scores)。和这一方法一起使用的

一个途径是,美国劳工部 70 年代开发的职业能力倾向模式 (the Occupational Aptitude Pattern, OAP) 结构 (美国劳工部, 1970, 1980)。OAP 包括了 60 多种工作门类,覆盖了数千种职业。对于每一组职业计算了表示有关能力倾向高、中、低三种资格水平的临界分数,可用于个体适宜职业的咨询。<sup>[27]</sup>通过使用效度一般化(VG)技术对 500 项前 USES 效度研究资料进行了分析,得到了 GATB 的第二种使用方法。这种方法大家称之为效度一般——一般能力倾向测验(VG-GATB)。它是在对每一门类职业所有工种的分数进行恰当综合的基础上,使用效度数值(美国劳工部,1983a, 1983b, 1983d)。预测以来源于原始因素分数的认知、知觉和心理运动三种成分的测量为基础。在这三种成分中,认知成分对于大多数的工作具有最高的效度系数,而心理运动成分的预测性随工作复杂性的降低而增强 (Hunter & Hunter, 1984)。

通过 USES 工具,积累了大量有关 GATB 的资料,其中许多证明了具有不同寻常的信度和很高的预测效度(综述见 B. Bolton, 1994; Kirman & Geisinger, 1986)。然而,在 20 世纪 80 年代,为了保证白人、黑人和西班牙人中的求职者参照标准的可比性,制定亚组常模的工作导致了关于雇佣前测验公正性的激烈的政治争论(Hartigan & Wigdor, 1989)。这一争论最后导致了 1991 年公民权法令(the Civil Rights Act of 1991, P.L. 102 ~ 166)的通过,禁止亚组常模的制定。这一立法影响了该工具的使用,使其前途变得不可捉摸(L. S. Gottfredson, 1994; Wigdor & Sackett, 1993;也见第十八章)。尽管如此,一项包括发展 GATB 两种新形式和一种计算机化的实验性版本在内的研究计划仍在进行当中。

另一种主要的人事选拔和分类工具是为所有的美国军队服

务机构使用而联合开发的军队服务职业能力倾向成套测验 ( the Armed Service Vocational Aptitude Battery, ASVAB, Bayriff & Fuchs, 1970)。这一测验实施的对象是对军队职业感兴趣的高中学生和那些 中请进入军队的人 ASVAB 现行的形式 包括表 17 - 2<sup>28</sup> 列出的 10 个分测验。陆军资格测验 (the Armed Force Qualification Test, AFQT) 是一种组合的分数, 被所有的军队机构使用, 作为甄别可能成为新兵的一般训练能力的衡量指标。此外, 每一机构可以把几个分测验结合起来, 形成适合其人事选拔和分类需要的组合, 例如军队作战部的测验组合由 AR + CS + AS + MC 成分的分测验组成 ASVAB 的标准分数, 以从美国青年有代表性的样本所获得的常模为基础 (美国国防部, 1982)。这一测验的因素结构经过相当详尽的检测, 其研究结果表明: 一般因素说明了整个 ASVAB 变异的 60%, 4 组因素得到重复表现 (J. R. Welsh, Watson & Ree, 1990) 在分测验中, 有最高负荷的四种因素是: (1) 语言 (WK 和 PC); (2) 速度 (NO 和 CS); (3) 数量 (AR 和 MK); (4) 技术 (AS, MC 和 EI)。

表 17-2 军队服务职业能力倾向测验 (ASVAB)

| ASVAB 分测验              |                        |
|------------------------|------------------------|
| 一般科学 (CS)              | 算术推理 (AR) <sup>a</sup> |
| 词的知识 (WK) <sup>a</sup> | 数学知识 (MK) <sup>a</sup> |
| 段落理解 (PC) <sup>a</sup> | 机械理解 (MC)              |
| 电子信息 (EI)              | 汽车和业务信息 (AS)           |
| 编码速度 (CS) <sup>b</sup> | 数字运算 (NO) <sup>b</sup> |

<sup>a</sup> AFQT 组合

<sup>b</sup> 速度组合

ASVAB 分测验及其组合的效度,按照各种不同的工作和教育绩效标准进行了研究。正如想象的那样,效度系数随所采用标准的类型和数量的不同而有很大的差异。一般说来,对于“能做”(can-do)标准(如一般的士兵和技术熟练的士兵)的效度要高于“想做”(will-do)(如努力、领导、个人意志)标准的效度。正如预期一样,前者是通过工作知识和动手操作的度量来评定的,而后者是通过上级、同事和本人的评定等级来评定的 (McHenry, Hough, Toquam, Hanson & Ashworth, 1990; T. L. Russell, 1994)。

作为联合—军种工作绩效测量/入伍标准 (the Joint-Service Job Performance Measurement/Enlistment Standards(JPM))项目的一部分(Wigdor & Green, 1991a, 1991b),1980年美国国防部开始对ASVAB进行最为仔细的检测。这一庞大项目旨在为各军队工作录取水平编制一个强有力的测量工具,这样就能为所有的志愿兵招募确定一种富有意义且有成效的入伍标准。JPM项目第一阶段确定:ASVAB是一种具有较高精确度的预测工具,是一种基于绩效的工作熟练度的指标。<sup>[29]</sup>它也为不同工作中的ASVAB组合分数的区分效度提供了支持。然而,黑人与非少数民族测验者平均分数的差异幅度,在AFQT测验和工作知识的纸笔测验上的差异幅度大大超过手工操作样本(the hands-on sample)测验上的。这表明ASVAB中某些测验可能过高地估计了工作绩效中实际的团体差异。如果有确凿的证据,这种情况将与GATB的一些发现相似(Hartogan & Wigdor, 1989)。此外,不同标准下ASVAB的效度系数足以说明对另外预测指标的研究是很有价值的。JPM项目第二个阶段是在成本和绩效水平上,探究能增强人事选拔和分类决策的整体效用性的入伍标准模式。

美国军队人事选拔和分类项目(Project A),包括了ASVAB

研究和军队工作绩效新的预测工具研究中另一些重要部分。作为军方本身对 JPM 项目颁布的广泛反应而产生的项目 A, 是有史以来规模最大、费用最高、争论也颇多的人事选拔研究项目 (Schmidt 等, 1992) 除了它对工作绩效理论建构的贡献外 (见本章前部分的讨论), 项目 A 的研究包括发展和评定了许多超越 ASVAB 传统认知功能的新型预测工具。这些工具, 既包括计算机实施的知觉和心理运动能力测验, 也包括用于评价人格、气质、背景变量和兴趣等特地建构起来的问卷 (McHenry 等, 1990; N. G. Peterson 等, 1990) 而且, 大规模的纵向性项目 A, 也使得研究随时间而产生的效度变化提供了空前的可能性 (J. P. Campell, 1990b)。<sup>[30]</sup>

**特殊能力倾向测验** 在多种能力倾向测验发展之前, 许多人就认识到智力测验在其所覆盖的能力上是有局限性的。不久, 人们就努力想使特殊能力倾向测验包含更具体、更实际的能力, 如机械能力倾向, 来弥补这一缺陷。职业选择和咨询的需要也刺激了空间、文书、音乐和艺术能力倾向测验的发展 视觉、听觉、肌肉操作和动作灵活性的测验还广泛应用于工业和军队领域的人事选拔和分类中。<sup>[31]</sup>

关于特殊能力倾向概念需要加以说明的是: 特殊能力倾向这一术语, 最初是将测试重点放在一般智力上时出现的; 在描述某一个体时, 机械、音乐和其他的特殊能力倾向被认为是对“智商”的补充。然而随着因素分析的出现, 人们逐渐认识到智力本身是由一系列相对独立的能力倾向组成, 如言语理解、数字推理、空间视觉等。此外, 一些传统的特殊能力倾向, 如机械和文书等, 现在已整合在一些多种能力倾向成套测验之中。

那么特殊能力倾向测验的作用是什么? 第一, 有某些领域, 如视觉、听觉、运动灵活性和艺术天赋等, 很少包括在多种能力

倾向成套的测验中。需要这些领域测验的情境是很特殊的，很难把它包含在标准的测验工具中。然而特殊能力倾向测验也被应用于多种能力倾向成套测验所覆盖的领域，如文书和机械能力倾向。在某几种测验中，学业智力测验和一些特地挑选出来的其他相关能力倾向的测验结合起来。这种做法的一个理由是，它们可以在某些广泛应用的特殊倾向测验所得到的常模和效度分析资料中找到。另一种理由则是这一程序的灵活性，它不仅表现在有关能力倾向的选择上，而且表现在能据某一特殊的目的来充分测量每种能力倾向上。

许多设计旨在测量速度、协调和其他心理运动技能的测验，大多数是关于操作的灵活性，也有一些是关于完成特殊工作所要求的腿部或脚部运动。有一些测验测量的是运动能力和知觉、空间或机械能力倾向的结合。这些测验主要应用在工业和军队的人员选拔中。虽然有些纸笔测验改编本是为团体施测而设计的，但心理运动能力测验具有仪器测验的特征。一项针对几种简单操作技能的克洛福德身体部分灵巧性测验 (the Crawford Small Parts Dexterity Test, Crawford & Crawford, 1981) 已经发布,如图 17-2。在这一测验的第一部分，应试者用镊子把小栓

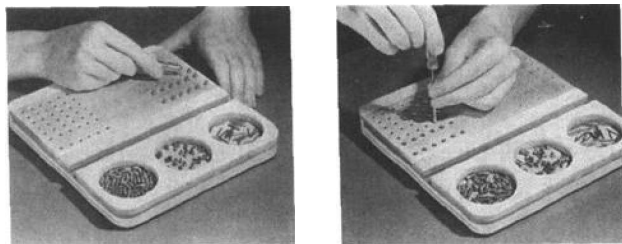


图 17-2 克洛福德身体部分灵巧性测验



插入最适合的小孔中，然后再给小栓戴上小栓母；在第二部分，把一些小螺钉放在可以穿过的孔中，用螺丝起子把它拧紧。其分数是完成每一部分所需的时间。

对于心理运动能力测验的整体效用能说些什么呢？在评价这类测验中值得注意的最重要之点是，它是一种高度特殊化的动作机能测验。大量动作能力测验的相关和因素分析没有发现智力机能测验中发现的群因素（Fleishman, 1975; Fleishman & Quaintance, 1984，见第十二章）。在考虑心理运动能力测验效度时，我们需要区分与试图预测的特定标准操作十分类似的复杂运动能力测验和能普遍使用的简单运动能力测验。前者是专门编制的测验（现在一般是计算机化的），再现了特定标准所要求的几个运动能力倾向组合，具有较好的效度，例如一些空军测验显示它能提高飞行操作的预测性（参见 R. H. Cox, 1989; Kantor & Carretta, 1988）。然而，这种测验的使用对大多数目的不适用，因为大量的测验必须适合各种不同标准。至于商业中可得到的运动能力测验，它们的测量是十分简单的，对大部分效标而言，其效度也是不高的。因此，这种测验最好作为某种选拔工具的一部分，而不是单一的预测工具。

机械能力倾向测验(Mechanical Aptitude Tests)包括了不同的机能。心理运动因素进入该类的一些测验中，或者因为在完成测验中需要有快速操作，或者因为一些旨在测量运动灵巧性的分测验也包含在纸笔测验中。在许多该类测验中知觉和空间能力倾向测验起着重要作用。最后，在许多机械能力倾向测验中，机械推理和机械信息使用能力倾向占主导地位。

认识到属于机械能力倾向的机能的多样性是十分重要的，因为每一种机能与其他变量的相关是不同的。例如，机械信息测验比抽象的空间或知觉测验更多地依赖于机械对象的过去经

验同样,各机能之间的性别差异可能是相反的。在操作灵巧性和知觉辨别测验中,女性表现良好;在抽象空间能力的测验中,明显可以发现两性之间均数差异非常微小,男性略占优势;在机械推理或信息测验中,男性具有明显的优势(Anastasi, 1981c; Hedges & Nowell, 1995)

所有多种能力倾向成套测验中都包括的一种能力倾向是空间能力倾向,这就是 DAT 空间关系测验(the Space Relations Test of the DTA)所测量的能力(见第十章)。已经发现它在许多操作和非语言的一般智力测验中具有较高的负荷。单独测量这一能力的最好测验之一是修订的明尼苏达纸笔测验(Minnesota Paper Form Board Test, Liker & Quasha, 1995)。机械能力倾向测验的另一种主要类型是关于机械信息、机械推理或机械理解等测验。虽然它们要求熟悉一些普遍的工具和机械关系,但其假设是:在现代的工业社会里,机械知识都是通过日常的经验获得的。这一领域的一些早期测验,要求应试者利用一些给定的部件组合常见的机械物体。纸笔团体测验现在被广泛应用于一般的测验目的。这种测验类型的一个众所周知的例子是本纳特机械理解测验(the Bennett Mechanical Comprehension Test, G. K. Bennett, 1994)。这一测验采用对应试者提出简短问题的图片,如图 17-3 所示,强调对各种日常情境中机械原理应用的理解。

文书能力倾向(clerical aptitudes)测验的共同特征是强调知觉速度和准确度。其典型的例证是明尼苏达文书测验(the Minnesota Clerical Test, MCT, Andrew, Paterson & Longstaff, 1979)。它由两个分别规定时限的分测验组成,即数目比较和名称比较。第一个分测验给应试者 200 对数字,每一数字的位数从 3~12 不等,如果两组的数字相同就在它们之间打一个记号。第二个分测验的操作任务和第一个相似,只是专有名称取代了数字。

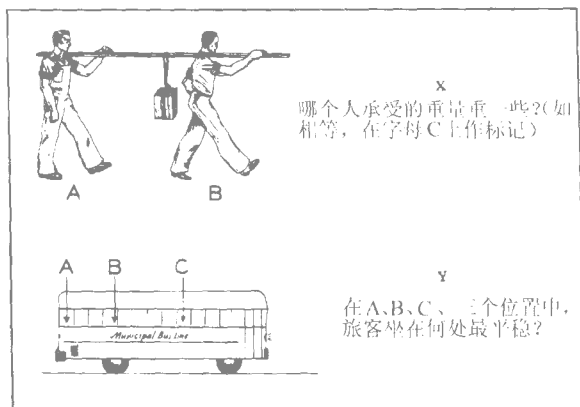


图 17-3 样例引自本纳特机械理解测验 (在答卷纸上作答)

像 MCT 这样相对同质性的测验,测量的只是文书工作的一个方面。文书的工作包括多种机能。此外,工作职责数量和组合,随工作类型和水平的不同而有巨大的差异。虽然职业活动有着这样的差异,但文书工作的分析一般显示,大部分的时间花费在任务所要求的知觉细节的速度和准确度上。的确,除文书外,许多类型的工作需要知觉速度和准确度。巡视员、质量检查员、包装员以及工厂的工人显然都需要这种能力,虽然这一类工作已被电子检测装置所代替。

有几种文书能力倾向测验把知觉速度和准确度与文书工作中所需的其他能力结合起来。在这些工具中使用像排序、分类、编码这样一些活动的工作一样本测验 (job-sample test);此外,一些测验包括了语言和数字能力作为一般智力测验的替代。其他一些文书能力倾向测验包括了办公技能,如职业词汇、职业信息、拼写和语言使用等。有些文书测验归为成就或工作样本

测验比较恰当,因为所测的技能是经特殊培训后获得的。这一类测验的例子有速记技能和打字技能的测验,还有像科学研究会(Science Research Associate,SRA,1990)的 CRT 技能测验中的数据登录和检索技能测验。

随着办公室工作使用计算机的快速增长,导致了几种计算机相关能力倾向测验的出版发行。其中许多用于咨询或筛选有潜力的受训人员,如计算机知识测验、计算机科学测验、计算机程序员能力倾向测验(这些工具的综述见 Mahurin,1992;Marco,1992;Schafer,1992) 这些测验及时地把 60 年代至 80 年代早期形成的心理测验应用于人事评定领域,但随着技术的快速变化,其中的一些测验已经过时。在教学和能力测验中,已逐步出现一些使用特殊软件程序<sup>[32]</sup>(如 dBASE,WordPerfect & Lotus1-2-3)的新工具。

近年来,对个体内和个体间机能的认知维度重要性的认识所取得的进展,可能刺激旨在评定工作情境中智力的社会 and 情绪层面评定的标准化客观工具的发展(参见 H. Gardner,1983;Salovey & Mayer,1990) 在此之前,这种工具在实际上是不存在的,因为工作情境中的个体内和个体间的机能传统上是通过人格测验或通过面谈和评定中心技术而进行测量的。团队工作—KSA 测验(The Teamwork - KSA Test) 是一种最近出版的旨在预测团队有效性的纸笔测验。它是该类新型工具中的第一批工具之一。通过总结组织内工作团队的各种资料,这种测验试图测量应试者关于人际和自我管理方面的知识、技能和能力。多项选择项目呈现假设的问题,涉及冲突的解决、交流、协同式的问题解决、目标设定、计划和其他一些自我管理技能(M. J. Stevens & Campell,1994)。对于这种或其他类似工具的进一步研究和实验定会紧随其后。

## 工作场所的人格测验

在 20 世纪 80 年代中期,伯纳迪恩和鲍纳斯 (Bernardin & Bownas, 1985) 指出,在工业情境中广为使用的人格评价技术(包括如笔迹分析在内的一些非科学的方法),在学术领域里却几乎被忽视了 20 年之久。此后,研究方法和理论的发展,重新激活了这一领域的大量研究。从研究方法上来看,元分析技术的应用和因果模型的使用引起了对影响工作绩效的非认知特性的研究(见第五章)。元分析技术主要用来研究不同情境中人格构念的效度和功效。路径分析和结构方程建模技术(见第五章)用来研究不同指标之间的关系,寻找导致不同绩效水平的原因和条件。如能力水平和工作自主性等一些关键变量对人格与绩效的关系的调节程度也在研究当中。总之,这些研究的目的不仅仅局限于预测,更是为了解决定工作绩效变化的结构和过程 (Barriek & Mount, 1991, 1993; Hough, Eaton, Dunntte, Kamp & McCloy, 1990; Schmidt & Hunter, 1992; Tett, Jackson & Rothstein, 1991)。

过去 10 年或 20 年里对方法的精细研究,为不同领域职业决策过程中使用精心编制的人格测验提供了相当大的支持。一个仍需进一步研究的问题是:确定与特定工作或某类工作的绩效相关最大的人格维度,如一些工作要求较高程度的社交性,而另一些工作则恰好相反;甚至,像认真这一性格特征,乍一看对所有的工作都是需要的,但进一步研究发现,它包含了对某些工作努力(如创造性追求)具有负作用的特性,至少在某些方面是如此。基十对这些问题的考虑,现在的注意力开始放在确定工作所需的气质和人际交往要求上 (R. Hogan, Hogan & Roberts, 1996; Landy 等, 1994)。

最近,大量的研究活动都在集中探讨各种形式的人格结构五因素模型在工作绩效标准预测中的作用(见第十三章和 R. 豪根 1990 年的著作)。五因素模型在人事选拔研究中的流行不是偶然的,因为  $5(\pm 2)$  种因素是从大量的人格描述资料中提炼出来的,代表了一种用来描述大家认可的正常人格特征的模型。该模型归纳的每一个方面都可以广泛地应用于各种日常行为中,包括工作职责的绩效;例如其情绪稳定性维度,对于在有压力条件下要求快速决策的工作如法律强制、飞机驾驶、卡车驾驶、急救等,是非常重要的另一方面,愉快的性格是任何需要广泛人际交往的工作不可缺少的品质。认真是经过多次元分析研究而得到一种无所不在的、有实际价值的工作绩效预测因素,这一点也不奇怪(Barrick & Mount, 1991, 1993; Schmidt & Hunter, 1992)。即使如此,对于认真这一构念的界定,或它的重要性问题的探讨远未达成共识(参见 Loevinger, 1994; Tett 等, 1991)。

**诚实测验** 1988 年通过的雇员测谎保护法案 [the Employee Polygraph Protection Act of 1988 (P. L. 100 ~ 347)], 禁止在政府机构等一些特殊的情境以外的人事筛选中使用测谎仪, [34] 此后,雇佣决策中诚实测验的应用问题变得显著起来。这一立法导致可信的纸笔测验的激增,它们被广泛地称为“诚实测验”。这些测验主要询问求职者对偷窃和其他不正当行为的态度,以及其历史上是否有过这种行为。这种精细检测通过心理学或心理学以外的方法迅速进行。[35] 在一个对诚实测验最为全面的研究中,奥尼斯(Ones)、维斯韦思瓦拉(Viswesvaran)和施米特(1993)对 665 个效度系数进行了元分析;他们的研究估计诚实测验对工作绩效的等级评定预测有 .41 的平均操作效度。研究表明,这些测验在预测工作中的破坏行为也有价值。

虽然根据某些参数,诚实测验的使用获得相当程度的支持,但它们在雇佣前选拔决策的应用中仍存在问题。当前关注的问题集中在测验可能容易受到指导和伪装的影响以及其预测某种特殊形式的不良行为(如偷窃)的相对无效性(参见 Alliger, Lilienfeld & Mitchell, 1996; Camara & Schneider, 1995)。

**领导能力** 有效领导者的选拔是人事决策领域中另一个有意义的需求。领导能力是一种在工作情境中最需追求的品质之一,因为它包含说服别人朝同一目的工作的能力。有效的领导能力在增强组织的功能方面具有强有力的影响,而无效的领导能力可能带来灾难。因而,许多类型的认知和人格测验、评定中心技术、面试、情境模拟、个人资料问卷等工具都应用在高层管理者的选拔中。选拔高级管理者的评定过程往往包括了多种方法,要经历几个小时或几天。这些方法与临床测验非常相似,其花费是相当昂贵的。

显然,用来精心选拔从日常工作到领导岗位的个体的资料与这些决策可能产生的影响相比,在很大程度上是相对贫乏的;在这种情况下一种可能的结果是管理无能的程度非常高(R. Hogan, Curohy & Hogan, 1994)。虽然在应用心理学中对领导行为的研究兴趣已有较长的历史,积累了许多文献资料,但在确定和阐明与有效领导有关的因素的经验研究方面仍有许多工作要做。在组织中的领导研究领域,领导能力的测量及理论和方法上最新进展的进一步信息,请看巴思(Bass, 1990)、克拉克(Clark, 1990)、尤克尔和范弗利特(Yukl & Van Fleet, 1992)的著作。

**工具** 除了把一些精细的评定方法用于高级管理者和一些专业人员(如情报人员、宇航员等)的选拔过程外,大部分参加人格测验的求职者只参加十三章讨论的自陈式问卷测验。一些旨

在检测心理障碍的问卷,如 MMPI, 虽然仍在筛选一些敏感职业时被使用,但因担心其对个人隐私的侵犯和其结果的误用,一般还是主要使用一些旨在评定正常被试的人格测验(见第十八章)。事实上,几种最初为正常个体设计的多维自陈问卷(如 CPI)最近进行了修订,使其更适用于工作情境中。同时,一些新的测量工具,如霍根的人格问卷(HPI-R., 1992),首先是为职业情境下的使用而设计的。其他一些测验则是为更小范围的应用而编制的,如 PDI 顾客服务问卷和就业问卷(the PDI Customer Service Inventory and the PDI Employment Inventory),该工具旨在从求职者中筛选出可靠的雇员(参见 Paajanen, Hansen & Mclellan, 1993)。

总评 技术发展正以大大超越工业革命所产生变化的速度改变着当今工作的性质。这种快速的变化对工业心理学家提出了巨大的挑战,也为新技术发展引起的方法进步提供了巨大机遇。<sup>[36]</sup>当研究者对于一些长期存在的问题有可能进行大规模的研究并可能取得一些明确的答案时,这些问题本身已经就开始发生了变化。同时,计算机技术对人事测验的影响,虽然才初见端倪,但已清晰可辨(Bartam, 1993; Burke, 1993; Schoenfeldt & Mendoza, 1991)。更令人激动的是某些新技术在测验中的应用,如相互作用的多媒体测验,尽管仍处在发展的初步阶段(Desmarais, Masi, Olson, Barbera & Dyer, 1994; Drasgow 等, 1996)。此外,尽管还有更多前景还没有成为现实,却是可以预见的。总之,心理学原理和方法在提高人力资源功效方面的应用,现在似乎比以前任何时候具有大得多的发展空间(Bray 等, 1991; Gatewood & field, 1993)。



## 测验在临床和咨询心理学中的使用

临床和咨询心理学家采用各种各样的测验，包括前面已讨论的大多数类型的测验。对理论定向和工作场所具有差异的临床和咨询心理学家使用测验的情况进行定期的调查发现 (Archer, Maruish, Imhof & Piotrowski, 1991; Lubin, Larsen & Matarazzo, 1984; Piotrowski & Keller, 1992; Watkins 等, 1994; Watkins, Campell, Nieberding & Hallmark, 1995)，使用频率最高的是像韦克斯勒量表 (Wechsler scales) 那样的个体智力测验和像 DAT 见第十章) 那样的多种能力倾向成套测验。从第十三章到第十六章讨论的许多人格测验是临床和咨询工作者的保留项目，具有重要的地位 (Butcher & Rouse, 1996)。一些诊断性的测验适用于学习不能案例和其他与学校有关的问题。此外，临床和咨询工作者使用许多简单的问卷和评定量表，对其实践中所碰到的多种问题进行快速的评定；许多该类测验在菲希尔和科科伦 (J. Fischer & Corcoran) 1994 主编的一本资料集中都有描述。

本节的重点是讨论在个体评定中要求使用多种测验及其他工具，而不是单独使用某种工具的一些领域。这种类型的心理评定由来自不同领域的心理学家进行，他们大部分受过临床和咨询心理学及其分支学科的训练。临床心理学家评定的传统目的是心理健康情况的诊断、预测和治疗 (Butcher, 1995; Hersen, Kazdin & Bellack, 1991; Maruish, 1994) 而咨询心理学家的最初专业局限于职业评定和指导 (S. D. Brown & Lent, 1992; Drummond, 1996; Gelso & Fretz, 1992)。应用心理评定的其他一些主要领域如学校心理学和司法心理学，<sup>[37]</sup>是由于其专业人员的工作场所

而得到确定的(Shapiro, 1991; Vance, 1993; Weiner & Hess, 1987)。另外一些使用心理评定技术的专业是由他们所研究的问题类型来确定的,如健康心理学(N. Adler & Matthew, 1994; Streiner & Norman, 1995; S. E. Taylor, 1990)、研究大脑和行为之间关系的神经心理学(Benton, 1994; Maruish & Moses, 1997)。还有一些是由他们病人或所服务对象的种类确定的,例如,儿童、夫妻或家庭(Conoley & Werth, 1995; Fruzzetti & Jacobson, 1992; Kamphaus & Frick, 1996)。许多实际工作者的工作领域更为专门,如儿科、神经心理学或医学康复,所有这些领域和分支领域都在继续不断地发展。此外,一些专业的界限变得模糊,在某些情形下甚至消失。

这种互相重叠的一个显著的事例是,临床和咨询心理学之间良好关系的发展(参见 Anastasi, 1979, 1990a; May, 1990)。<sup>[38]</sup>人们日益发现咨询心理学家活跃在一些私人业务、诊所和医院等场所中。同时,一些临床心理学家正在将他们的技能应用于职业评定和人事选拔(Lowman, 1989, 1991, 1993)。正在发生的一种显而易见的变化是:计划出版的新测验标准,在题为“心理测试和评定”的一章中提出在临床、咨询和学校等领域可使用测验进行诊断和个体评价,而 1985 年的标准则分别处理了临床测验和咨询测验的使用问题(见第一章)。

**心理评定** 那么心理评定在不同专业领域、不同场所、不同问题范围和不同人群中的实践应用带来了什么?它的一个主要特征是:运用多种资料集中对一个或多个个体(如某对夫妻或某个家庭)进行细致研究。通过建立和维持一种良好关系,临床工作者能引出其他方法不能得到的有关病人生活资料的相关事实。个人生活历史对了解个体和预见其后续行为提供了特别坚固的基础。而且,临床工作者在人际情境中可以作为刺激物,从

而有利于事实的发现过程；因此，临床面谈起着情境测验或模拟的作用，能提供具有或多或少控制条件下病人人际行为的样本。

通过观察、面谈和考察个案历史得到的信息和测验分数结合在一起，提供了一种综合的个体形象( Beutler & Berren, 1995)。临床工作者已经建立了某些防止测验分数过度泛化的防御机制，至少可以部分说明为什么一些在心理测量学特性上有弱点或未经检验的测验仍在继续使用。只要这些工具主要是为熟练的临床工作者提供一些提示，引起他们的重视，它们的保留便还是有价值的。当然也存在着危险，那些相对无经验、过分热诚的临床工作者不留意某种工具的局限性，可能对测验的分数过分相信。

心理评定另一明确的特征是其目的：一般来说，它是为诊断、职业选择、治疗建议、教育计划、儿童监护人的确定、处罚和许多对某人或多人有实际意义的诸如此类的事情进行有根据的决策提供帮助。决策可以借助相关行为资料的收集、分析、整合和思考的报告来进行。心理评定过程中的中心环节是对具体的个案连续不断地提出假设，检验假设。每项信息，不管它是个案历史记录中的事件，或是病人的评论，还是某一测验的分数，都提出了关于个体的一个假设，它在收集到其他事实材料时会得到证实或否定。在这一方面，必须牢记任何单一的资料源，不管它似乎是多么的可靠，有时会提供错误的信息。<sup>[39]</sup>

从清楚阐述的主要目的到结果的交流等等与心理评定相关的所有活动，基本上是一种以其所面对的特定问题和人群的知识为基础而进行的专业判断。而且，个别测验和其他评定工具的应用不仅需要专门的技能，还需要慎重考虑它们的特征以及评定的目的和情境( Drummond, 1996; C. T. Fisher, 1985; Groth-Marnat, 1990; G. Goldstein & Hersen, 1990; Hodd & Johnson, 1997; Mal-

oney & Ward, 1976; Tallent, 1992; Walsh & Betz, 1995)。在这一方面,临床方法和客观的、标准化的测验不同和那些通过回归方程、多重临界和其他的机械公式来处理数据的统计程序不同(见第六章,又见 Dawes, Faust & Meehl, 1993; L. R. Goldberg, 1991; Kleinmuntz, 1990; Wedding & Faust, 1989)。

强调需要考虑个人生活环境的生态学观点,像其对发展心理学和一些相关的领域一样,对评定工作有重大影响(参见, Moen, Elder & Luscher, 1995)。同样,对文化在所有行为——包括那些导致人们从心理健康专业人员那里寻找帮助的问题——中的作用的认识的提高,引起了人们对关于文化适应力的评定实践及其指导问题的浓厚兴趣(见,如美国心理学会 1994 年的报告, pp. xxiv 和 843 ~ 849; Dana, 1993, 1996; Prediger, 1993; Suzuki 等, 1996)。

**个体评定中的智力测验** 类似韦克斯勒量表和斯坦福一比纳测验(见第八章)等测验基本上是个别实施的临床工具;一个机灵的受过训练的临床工作者,在测验施测所要求的时间内和应试者主动接触,除了 IQ 或其他一些分数外,他或她几乎不可能得到有关个体的更多信息;如果某一测验是由某一技师施测的情况也是如此,倘若保存着应试者反应的一份完整记录。

临床工作者除了使用智力测验来评定个体的一般智力机能水平外,通常还探究测验分数的模式或剖面图以了解应试者的优点和弱点。剖面分析为大脑损伤的诊断和对智力产生不同影响的各种形式的心理病理因素的诊断提供了有用的资料。韦克斯勒量表特别适合于剖面图分析,因为所有分测验的分数可以直接用可比较的标准分表述。一开始,韦克斯勒就描述了其量表的一些诊断性使用。此后,部分临床工作者介绍了一些其他技术,把剖面图分析技术和其他工具结合起来使用(Delaney &

Hopkins, 1987; Elliot, 1990b; Kaufman, 1990, 1994; Matarazzo, 1972; Sattler, 1988, 1992) 大多数的剖面图分析采用三种主要程序的各种变式: 第一种, 对个体各种分数分布量或变异程度的评价, 包括言语和操作 IQ 的差异、全部分测验的分数分布、个别分测验的量表分和分测验的各种组合的均数的比较 (如言语或速度分测验)。第二种, 分析个体剖面图中显著的特征, 其分析是依据正常组群内该种特征出现多少的基线频率资料来进行的。第三种以分数模式为基础, 它与诸如阿泽梅斯 (Alzheimers) 疾病, 学习不能或焦虑状况这样一些临床症状有关。韦克斯勒和其他一些研究者描述了分测验分数高低模式和说明这些及其他一些心理障碍特征的分测验组合模式 (参见 Kaufman, 1990; Matarazzo, 1972)

几十年来对韦克斯勒量表各种形式的模式分析研究, 并没有为其诊断价值提供多少统计上的支持。<sup>[40]</sup>事实上, 对这一方法的批评几乎遍及所有方面而不是局限于某一点 (F. C. Goldstein & Levin, 1985; Kavale & Forness, 1984; Macmann & Barnett, 1994a, 1994b; MacDermott, Fantuzzo & Glutting, 1990)。尽管如此, 从韦克斯勒量表在临床实践领域中的继续流行来看, 从大量旨在促进和改善分数模式的分析并使之系统化的研究文献来看, 它仍然是解释智力测验数据的优先方法。

单纯从质的水平来看, 成绩的不规则性也可能提示了进一步研究的途径。有意义的提示可以来自于测验反应的形式, 也可来自于测验反应的内容。反应奇异、过分推敲或过度的自我中心可能显示了人格障碍。对错误和正确反应的质的分析, 可能为解决问题的方法、概念的形成或认知风格提供有用的线索。不正常的测验反应内容是另一种提示信息。在个体智力测验的实施中, 得到质的资料的另一个来源是测验情境中应试者的一

般行为反应,如动作、言语、情绪反应、对测验实施者的态度和对测验材料和测验环境的态度。因为其特殊性,它一个规则是,质的线索不能用适宜于测量群体趋势的数量方法来证实。尽管如此,因为认识到行为观察所起的重要作用,所以出现了一些用来把记录和解释测验过程中出现的行为加以系统化和量化的工具(参见 Guide to the Assessment of Session Behavior for the WISC - III and WIAT, Glutting & Oakland, 1992)。

对复杂的临床使用智力测验的清晰说明(把心理测量数据和质的观察结合起来)是由康福曼(S. Kanferman)提供的。在他的关于‘智力’和‘智力测验’的论著中,详细说明了临床工作者如何把测验分数的统计信息和有关人类发展、人格理论和心理学研究的其他领域的知识结合起来。康福曼强调考虑影响分测验成绩的技能 and 外部条件的重要性,强调需要一些解释分数模式的补充信息(可以从其他测验、个案历史、测验行为的临床观察得到)。测验分数和其他数据材料产生了关于个体的假设,它可以在获得更多信息情况下更加清楚地得到检验。康福曼方法的一个重要特征是要求对测验成绩作个体化的解释,而不是统一应用任何一类模式的分析;同一分数模式对不同的个体可能有不同的解释。

毫无疑问,康福曼描述的基本方法对智力测验的临床使用是有重要贡献的(参见 Roecker, 1995)。甚至它的批评者也承认,它已成为智力测验教学中所选定的方法,可用来帮助许多解释智力测验软件的开发(McDermott, Fantuzzo, Glutting, Watkins & Baggaley, 1992)。<sup>[41]</sup>由萨特(Satter, 1988, 1992)提供的指导手册是个别智力测验的心理测量和临床使用相结合的极好例证。然而必须认识到应用这些方法要求很有经验的临床工作者,他们熟知心理学的几个领域,并有充分的时间。而且,一些可利用的计

计算机化的解释程序,在某种程度上能促进这些方法的应用,然而也容易使性急或知识稍少的测验者过分解释程序提供的数据

同时,使用智力测验剖面图技术的新程序正在发展,其中有部分把心理测量和临床方法的一些要素结合起来,一个正在进行研究的有趣革新是“核心剖面图法”(core profile approach)。麦克德莫特(McDermott)、格鲁廷(Glutting)和他们的同事(Glutting, McDermott, Prifitera & McGrath, 1994, 1995; McDermott, Glutting, Jones & Noonan, 1989)以及唐德斯(Donders, 1996)把各种类型的聚类分析用于韦克斯勒量表的标准数据。这些分析的目的是产生基本剖面图类型,它有助于测验结果的分类及其临床意义假设的检验。另一个补充性的研究路线采用多维度的量表来确定人群中典型能力剖面图(Davison, Gasser & Ding, 1996)。这一研究的目的是把观察到的个体剖面图和从某一个测验工具中得到的典型剖面图之间相对应的程度进行量化。虽然这些研究十分有趣,但到现在为止,它们仍处在探究阶段,不具有临床应用价值(参见 Ryan & Bohac, 1994)。

### 神经心理评定

**脑损伤诊断的方法问题** 有关脑损伤的行为后果的认识大约开始于 20 世纪 20 年代早期戈尔茨坦(K. Goldstein)和其合作者的一本著作(Goldstein & Scheerer, 1941)。该书通过对一战时期受到脑损伤士兵的广泛观察,形成了对与脑损伤相连的智力损伤的经典性描述;其中的主要症状是抽象思维能力的减弱和对那些可能破坏正常知觉的无关刺激的反应倾向。

对儿童脑损伤的广泛兴趣开始于 20 世纪 30 年代后期和 40 年代斯特劳斯(A. Strauss)和其合作者的研究(Strauss & Lehtinen, 1947; H. Werner & Strauss, 1941, 1943)。这些研究者选取了

一组智力迟缓的儿童，他们的个案史表明由于出生前、出生时或出生后很短的时间内的创伤或感染而导致了脑损伤。对这些儿童的行为描述是戈尔茨坦提出的成人综合症状的扩展和精炼。它描绘那些广泛被认为是脑损伤儿童特征的智力和情绪障碍的独特模式，其中包括特殊的概念及知觉障碍和相对较高的言语能力，以及多动、注意力分散、攻击等症状。多年来，脑损伤儿童的研究和实践都被‘机体’(organicity)的单维度概念所支配。这一观点导致了对有机体本身诊断测验的探索并企图设计一些适合于作为一个整体的脑损伤儿童的补救性或教育计划。

自50年代以来，心理学家越来越认识到脑损伤可能会导致各种不同的行为模式。这一认识刺激了临床神经心理学的发展。该领域的目的是在脑损伤患者的诊断与康复中应用脑一行为关系的理论。不存在所有脑损伤案例都共同具有的某种症状或症状群。事实上，脑损伤对两个个体可能会造成相反的行为模式。这些发现与机体病理学自身的多样性相一致。脑一行为关系分析中的一个重要进展是拉弗·瑞坦(Ralph Reitan)及其同事在印第安纳州立大学医疗中心所作的研究(参见 Matarazzo, 1972,第十三章;Reitan, 1955, 1966)。这些研究表明,左半脑损伤倾向与韦克斯勒量表测验中言语智商分数低于操作智商分数( $V < P$ )的现象相关连，而在右半脑损伤及其弥漫性脑损伤小组中则相反的模式( $V > P$ )占主导地位。

神经心理的研究一直努力在寻找其他变量和脑病理学的行为后果之间复杂的相互作用(参见 Kolb & Whishaw, 1990)。有许多证据显示‘年龄’影响脑损伤所产生的行为症状。行为后果也取决于损伤前学习的总量和智力发展状况。例如，对学龄前儿童的研究显示，这一年龄水平上的脑损伤对智力功能的影响比其后年龄水平的大。



发现‘长期性’(chronicity)也影响测验成绩并和年龄效果相互作用。有资料表明,损伤以来经过的时间数量可能不仅与生理变化有关,而且和通过学习或补偿性治疗的行为恢复程度有关最后,必须注意在有些情况下,智力损伤可能是脑损伤的间接结果。在个体的整个发展过程中,机体和经验因素相互作用。例如,包含在脑损伤儿童典型症状中的一些行为障碍可能是机体因素导致的智力缺陷儿童所经受的挫折和人际问题的间接后果。这些行为继续存在与否,可能取决于儿童生活环境中父母、教师和其他重要的人的态度及理解程度。

有丰富的证据表明:脑损伤涉及各种各样的机体障碍,相应地具有各种行为表现。可以预料脑损伤患者的测验成绩随脑损伤的原因、程度、损伤点、损伤时个体的年龄、个体行为评定时的年龄、病理状况持续的时间、个体受到干预的程度和类型的不同而变化。希望脑疾病患者具有一致的行为表现是极不符合实际情形的。

从另一角度来看,同一种智力或行为障碍,并且在测验操作中具有同样的诊断性表现,可能是机体的、情绪的或多种病因的结果,一个明显的例证是永久性遗忘。各种类型的记忆障碍,可能是由机体原因或情绪抑制失调等引起的多种痴呆症中的一种症状。而且,机体记忆障碍一开始往往伴随着抑郁,表现出综合的状况。尤其是老年记忆问题的评定和对痴呆症及抑郁症的区分,是临床神经心理学中常常涉及的问题(Butter, Delis & Lucas, 1995; Poon, 1986; Reeves & Wedding, 1994; Storandt & VandenBos, 1994)经验因素在某些个案中可能和脑损伤无关,而在另一些个案中则或多或少直接和脑损伤有关,使得诊断更加复杂。因而对操作测验中某些具体诊断性症状的解释,需要有关个体的经验背景和个人历史的附加信息。例如,为了测定认知缺陷的

大小及智力机能恢复的程度,所需要的重要信息是个体生病前的能力水平(参见 Matarazzo, 1990),教育的程度经常作为病前机能的一种粗略指标,但以历史材料和损伤后测验成绩为基础的其他一些评定方法也已经设计出来(Vanderploeg, 1994b)。

总之,神经心理的评定实践是一项要求较高的临床任务,其应用需要有关正常人和病理个体的认知、人格、神经和一般生理机能的知识,因此,这一领域的研究文献和旨在指导学生和开业者的实践指导手册继续以巨大速度在增长。关于研究和临床信息的广泛的概述包括在波勒和格拉福曼(Boller & Grafman)主编的多卷本《神经心理学手册》(*Handbook of Neuropsychology*, 1988 ~ 1995)中。关于神经心理评定的一般实践及其干预的参考书籍最近由亚当斯、帕佐斯、戈尔伯福森和尼佐(Adams, Parsons, Gulberfson & Nizon, 1996)高尔顿、祖玛和埃普兹(Golden, Ziuner & Epierz, 1992),莱扎德(Lezak, 1995),陶兹、伯恩和格兰达斯(Touyz, Byrne & Gilandas, 1994),范德普卢和怀特(Vanderploeg, 1994a & White, 1992)提供。关于法学神经心理学(Valeiukas, 1995)、接触职业的毒害神经物质后的神经心理评定(Agnew & Masten, 1994)以及说西班牙语人的神经心理评定(Ardila, Rossecli, Puente, 1994),这些更为专门的专题论文继续以比较快的速率出现。

**神经心理的评定工具** 大量的测验是为评定神经心理损伤而专门设计的(参见 Lezak, 1995; Spreen & Strauss, 1991)<sup>[42]</sup>。这类测验经常作为机体或脑损伤的指标。这些工具评定的主要是那些认为最容易受病变过程影响的机能,如对空间关系知觉和新学材料的记忆。本德视觉运动格式塔测验(The Bender Visual Motor Gestalt Test)——一般称之为本德一格式塔测验(BGT—Bender, 1938; Canter, 1996; Heaton, Baade & Johnson 1978; koppitz,

1964, 1975; Pascal & Suttell, 1951)——和本顿视觉保持测验第 5 版(the Benton Visual Retention Test – Fifth Edition. BVRT, Sivan, 1991)是这类测验的具体例证,几十年来一直作为筛选测验使用。然而,因为机体脑功能失调及其伴发的行为障碍的多样性,任何一种测验都不足以检测一般的脑损伤,像 BGT 和 BVRT 等单个测验甚至不太适合于区别性诊断。

临床神经心理学家经常把可得到的测验结合使用,以评估不同的技能和缺陷,这被称为“灵活的成套测验”(flexible battery)方法(参见 Bauer, 1994; Goodglass, 1986)。这一方法的优点是根据个案所表现的问题而提供适合的测验组合;但它仍有若干局限性,在测验中可能不必要地重复测量了一些机能,而忽视了另一些重要的方面。挑选适合每一个案的测验给临床专家的评定和判断带来了较重负担。此外,一些单独编制的测验可能难以与常模和分数量表相比较。关于不同测验之间相关性的经验研究资料也可能不足,其结果是很难用分数模式来解释测验结果。

基于以上的这些原因,人们已作出一系列的努力来汇编综合性的标准化成套测验,以测量所有重要的神经心理技能。这种工具具有几种功能:它能成功地检测脑损伤,能帮助证实和确定脑损伤的区域,能区分与大脑皮层病变相连的各种特定症状,能揭示行为缺陷的具体类型,有助于恢复性训练的计划。综合性神经心理成套测验的两个主要例子是:霍尔斯特德-瑞坦(Halstead-Reitan)神经心理成套测验(HRB-Beitan & Wolfson, 1993)和鲁利亚-耐布拉斯卡(Luria-Nebraska)神经心理成套测验(LNNB – Golden, Purisch & Hammeke, 1985)。这两套工具都有适用于儿童的版本,它们有着共同的目的,但在几个重要的方面有区别:HRB 的形成要早,是由瑞坦从霍尔斯特德(1947)的工作

中发展起来,在施测的数量和选择方面为测验者提供了一些灵活性(综述见 Dean, 1985; M. J. Meier, 1985)。<sup>[43]</sup>而 LNNB 整合了鲁利亚的一些理论和诊断程序(Christensen, 1975; Luria, 1973, 1980),在内容、材料施测和记分上的标准化程序较高,施测的时间比 HRB 要短得多(综述见 J. H. Snow, 1992; VanGorp, 1992)。

通过脑电图和神经成像技术(Neuroimaging techniques)如磁共振成像(Magnetic Rezonance Imaging, MRI)和正电子发射型计算机断层(Positron Emission Tomography, PET),对脑损伤进行直接评定的新近发展,对基础的和临床的神经心理学带来了深远影响。尽管可利用的技术在不断改善,仍然没有一种诊断技术是 100% 可靠的。在大多数案例中,神经心理学家和神经学家以及其他一些专业人员一起工作,从不同的方面获得综合信息。在临床领域,精心标准化的行为测验和其他方法一道在评定和康复计划和监控中起着重要作用。还有在基础科学的领域里,神经心理和神经成像技术的结合,在一定程度上为进一步理解“脑—行为”关系提供了极大的希望(参见 Gur & Gur, 1991, 1994)。

### 特殊学习失能的鉴别

自 70 年代以来,掀起了对学习失能(Learning Disabilities, LD)的诊断和补偿研究的热潮。教育家越来越清楚地意识到在儿童中这类缺陷发生的频率是很高的,甚至在大学生和成人身上也是如此(参见 Gregg, Hoy & Gay, 1996; Kravets & Wax 1992; Wang, Reynolds & Walberg, 1991),虽然由于这一术语应用的随意性导致了一些个体被错误地归于此类。应用于这些案例的各种术语反映了方法的变迁和医学、教育学和心理学对 LD 取向的差别(参见美国精神病协会, 1980, 1994)。按照联邦政府为各种学习失能儿童的教育颁布的“94 ~ 142 公民法”(Public Law 94 ~

142) (见第九章) 确定的定义, 特殊学习失能被描述为:

理解或使用口头或书面语言中的一种或多种基本的心理过程的障碍, 它可能表现为听、想、说、读、写、拼写或数学计算能力等缺陷。这一术语包含了诸如知觉缺陷、脑损伤、脑功能轻微失调、诵读困难、发展性失语症等情形。这一术语不包括那些主要是因为视觉、听觉或运动缺陷, 智力发展迟缓, 情绪失调, 环境、文化或经济状况不良而造成的有学习问题的儿童 [联邦注册簿 (*Federal Register*), 1977, p. 65083]。

“94 ~ 142 公民法”进一步具体说明了学习失能的诊断只能应用于以下儿童: (1) 智慧能力与各种交流和数学技能上达到的水平有严重的脱节; (2) 即使提供了适宜的教育, 其能力水平仍然无法达到与他们年龄的相应的程度。<sup>[44]</sup>根据这一定义, 只有在排除其他几种引起儿童教育或心理发展困难的可能原因后, 才可使用学习失能的名称。多年来, 人们逐步认识到异质的学习失能群体是由几个亚群组成的, 它可按症状群加以区分 (Rourke, 1990; Feagans, short & Meltzer, 1991; Geary, 1993; S. R. Hooper & Willis, 1989; Pennigton, 1991; Shankweiler 等, 1995; H. L. Swanson & Keogh, 1990)。尽管如此, 学习失能专家们的理论定向仍有较大的差别, 这些差别反映在测验工具和补偿性计划中。

一般来说, 学习失能的儿童表现出正常或高于正常水平的智力, 而在学习一种或多种基本教育技能上 (大多数情况是阅读) 具有明显困难。然而, 须注意到特殊的学习失能可以出现在任何智力水平上, 即使弱智的学习失能儿童也不符合学习失能的法定定义。学习失能儿童也表现出有联系的一些行为症状的

各种组合；其主要症状是：信息知觉和编码困难，不同感觉器输入信息的不良整合，感觉运动协调性的破坏。语言发展障碍经常出现在学习失能儿童身上。像某些情绪和动机症状一样，记忆、注意力控制和概念技能方面的能力缺失是很常见的。特别是攻击和其他情绪及人际等问题的形成发展，经常是由儿童学习失能而引起的学业失败和挫折感的直接结果。在儿童行为评价中必须牢记：在儿童早期（3 岁以内）表现出的许多具体困难是正常的，但如果延续至年龄较大时，则表示功能失调。所以，一种质的常模（如果没有量的）参照性发展框架是很有必要的。

**评定技术** 不论理论取向如何，人们一般同意对学习失能的鉴别要求使用广泛的各种类型的测验和补充性观察性程序，诊断问题中至少有三种特征导致了这种状况：（1）与这一条件相连的行为障碍的多样性；（2）症状组合的个体差异性；（3）对于每一案例失能的性质和程度需要非常具体的信息。

对学习失能儿童的评定往往是专家组通力合作的过程。课程教师可以实施团体测验，可以采用其他筛选或众多的工具；普通的成就成套测验也是为此服务的，尤其是那些特地为小学和学龄前水平儿童设计、可以对具体优点和弱点进行“效标一参照”，即内容一参照分析的测验。在本章教育测验部分引述的几种工具是适合这一目的的。

个别实施的、范围较广的成就测验对学习失能的评定尤其有用。尽管辅助性的质的观察和分数解释可由临床工作者更有效地实施，但这些测验也可由教师施测。考夫曼教育成就测验（The Kaufman Test of Educational Achievement, K-TEA, Kaufman & Kaufman, 1985）就是该类工具中的一种。图 17-4 显示了使用普遍流行的画架型工具（easel-type kit）的该测验的施测情形。有几种成套测验包括了一些分测验，是测量能力倾向和认知能力



图 17-4 考夫曼教育成就测验(K-TEA)的施测。图中显示了画架工具的使用,活页封面上的卡形成了站立的架子,能同时把每个测验项目呈现给应试者,把相应的指导语呈现给主试(经美国指导服务处同意转载)。

的,另一些分测验是测量学生成绩的。最具有综合性的是伍德库克—约翰(Woodcock-Johnson)心理教育成套测验修订版(Psycho-Educational Battery, WJ-R, McGrew, 1994; McGrew, Werder & woodcock, 1991)。像斯坦福—比纳-K-ABC 和威克斯勒量表等测验,不仅提供了一种整体的指标(如 IQ)帮助区分智力迟缓和特殊学习失能,对于具体能力缺失也提供了许多质的信息(Kauf-man, 1990, 1994; Sattler, 1988, 1992) 例如,这些测验可以揭示知觉和视觉形状回忆、运动障碍、短时记忆局限、失能处理抽象概念和许多种类语言障碍方面的可能缺陷。韦克斯勒量表的出版者也编制了威克斯勒个人成就测验(The Wechsler Individual

Achievement Test, WIAT心理测验公司, 1992)。它是一种与智力量表相关的、想把它们应用于学习失能的评定中的综合性成套测验。

虽然, 现存的测验对于学习失能评定实施是有帮助的, 但许多研究者不断要求使用一些新的更有启发性的方法来对学习失能儿童进行诊断和评定。这些要求主要集中在需要一种清晰的理论原理来指导评定实践, 需要更深地理解每一个案的特殊过程, 以便提高评定和补偿性工作的成效(参见 Das, Naglieri & Kirby, 1994; R. B. Kline, Snyder & Castellanos, 1996; C. R. Reynolds, 1992b)。

**动力评定** 动力评定这一术语包括基本上是有意识偏离标准化或统一测验的各种临床方法, 其目的是引发一些关于个体的额外的质的信息。尽管熟练的临床工作者对这类方法的采用已有一段时间, 但其流行始于 20 世纪 70 年代(Lidz, 1981, 1987, 1991, 1995)。不仅在特殊学习失能的个案中, 而且在其他感觉学习困难的儿童, 如那些中度的智力迟钝者的诊断中, 它一直作为一种补充性的信息源。这一方法在天才(尤其是那些经济状况不良的儿童)的评定过程中的应用, 也得到一些尝试性的探究(参见 Bolig & Day, 1993)。

一种名为“测量限度”(testing the limits)的方法是这类定性的适应性测验方法中最早的一种。在这种方法中, 附加的线索可由测试者提供; 符合要求的操作需要的线索越多, 学习失能的程度就越大。在“测量限度”中使用的标准化程序的变化与在测验身体残疾个体时引入某些具体改编本相似, 同样的解释上的告诫和局限也适用于所产生的行为上(见第九章)。

最近发展了一种名为“学习潜力评定”(learning-potential assessment, Babad & Budoff, 1974; Campione & Brown, 1979; Feuerstein



1979; Glutting & McDermott, 1990; Hamers, Sijtsma & Ruijsenaars, 1993, Lidz, 1991) 的方法“潜力”这一词可能带有一种无需证明的涵义,即这种能力总是存在的,需要的只是显露而已。然而,事实上,这些方法包含了一种‘测验—教授—测验’的形式,通过各种方式教会学生完成他一开始不能执行的任务。表面上看,这一技术和一些预测性教育测验相似,后者给被试一个任务样本,它需要学习,这种学习和被试在某一课程学习中所遇到的学习是同类的。前者的应用至少在两方面和预测测验不同:(1) 给被试的指导和具体建议;(2) 任务常常包含范围更广的学习或问题解决技能。

福伊尔斯坦(Feuerstein)和其他人所首先探讨的动力评定技术从几个方面提出了发展的可能性。通过把评定和指导联结起来,他们把研究放在学业能力倾向可改变性和最佳的补偿性计划发展的基础上。此外,他们为合格的临床工作者提供了一种评定工具。与标准化的智力测验相比,它更能说明认知操作和对补偿干预的反应。尽管有这些优点,动力评定技术仍然受到批评,其中一个问题是关于他们的“可输送性”(transportability)或它能有效地被不同的临床工作者使用的程度;另一个问题是,关于这些广泛的作业(经常是与瑞文推理测验或韦克斯勒操作量表项目类似的非言语推理问题)所带来的补偿效果,是否能概括到真实的学校行为中去。而且,动力评定技术的大多数提倡者想抛弃对那些典型的‘静止’认知测量(如智力测验)分数的依赖;但如果不使用数字,如何使用其他方式来说明变化问题,这已经证明是十分困难的(A. L. Brown 等, 1992; R. E. Snow, 1990)。尽管如此,对动力评定技术的潜在使用价值的探讨仍在快速发展。最近对一些令人鼓舞的研究工作的报道包括:数学概念理解中不同指导策略作用的阐明,和多数字乘法问题操作的计算

机化动力评定系统的发展 ( Gerber, Semmel & Semmel, 1994; Jitendra, Kameenui & Carnine, 1994) 。

从另一角度来看, 埃布莱逊( Embretson, 1987, 1990, 1992 )提出的变化多维潜在特质模型(multimensional latent trait model), 防止了传统测验所固有的对行为的变化和学习进行测量时所存在的几乎是不可克服的许多技术难题 (参见 Gronbach & Furby, 1970)。埃布莱逊的方法使用项目反应理论 (IRT, 见第七章)和计算机化测验(CAT, 见第十章), 用接近每个个体阈限水平的项目 (在此水平上学习的效果最好) 来回避这些难题; 阈限水平上的项目还可以产生每一个体最可靠的成绩数值。而且, 通过应用认知心理学发展起来的任务分解程序, 埃布莱逊能够系统地改变完成每一测验项目的任务所要求的具体认知过程。例如: 它可以通过使用只要完全解决问题中的某些步骤就能完成的亚任务, 或呈现解决过程某一步或几步所需要的线索或具体知识, 或提供影响个体任务操作特定方面的指导语来做到这一点。显然埃布莱逊正在进行的工作, 代表了心理测量学上重要的进步。它把诸如临床心理学、认知研究、数学统计和计算机技术等各种不同领域的发展结合起来( Embretson, 1993, 1995)。

### 行为评定

集于‘行为改变’这一般概念下的各种技术, 表现了一些主要学习原理在行为变化的实际处理中的直接应用。这些技术基本上包含了条件作用原理在所需行为的获得及增强、和不需要行为的消退中的应用。行为治疗逐渐扩延, 以致包括了各种各样的心理问题, 包括了日益扩大的干预技术并考虑到认知、情感和外显的动作反应( Bandura, 1969, 1986; Hersen 等, 1991; Lazarus, 1981)。

评定技术 在行为治疗的早期应用中,对评定的注意甚少。然而,从 70 年代中期开始,人们逐渐认识到评定方法的重要性 (Barrios,1988 ;A.S. Beuack & Hersen,1988 ; Haynes,1991 ;Mash & Terdal,1980;Nelson & Hayes,1986;O' Brein & Haynes,1993; Ollendick & Hersen,1993)。<sup>45</sup> 在行为治疗中评定程序的主要功能可以归纳为三项:第一,评定技术通过对相关行为进行机能分析,帮助确定特定个体的问题。实际上,这种分析包括了对治疗目标(如恐惧或强迫观念的克服)的充分说明,对引发目标行为的刺激,目标行为实现的情境,以及对特定反应的性质、幅度和频率的描述。评定程序能指导行为治疗家的第二个方面是选择恰当的治疗措施。第三,需要评定治疗带来的行为变化,这种评定须包括监控变化的技术和终止性的测验;前者可以对治疗效果进行评定和在需要时引入另一种方法,后者确定是否达到满意状态和计划下一步所需追踪的程序。

在考虑具体的评定方法时,我们必须注意到 第一,同一方法常常能提供与以上三种功能相关的信息;第二,方法的选择依赖于问题的实质、患者的特征(包括患者必须面对的环境)和用于诊断的设施;第三,对大多数案例,几种评定方法的结合使用是较理想的。

可用的评定方法可分为三种类型 患者自我报告、行为直接观察和生理测量。虽然并不是每个中心都能提供生理测量所需的设备,但生理测量在某些评定中能提供补充的客观资料,如焦虑、性兴奋障碍和睡眠障碍 (sturgis & Gramling 1988)。皮肤电位、肌肉活动、眼电活动、心血管和大脑机能的测量等都是其中的具体例子。

目标行为的直接观察可由父母、教师、机构的人事部或特殊的观察者在自然条件下实行;观察可以使用的辅助工具如检查

项目表、评定等级量表、每日安排。这种观察具有几种弱点 (Barrios, 1993 也见第十六章),因而在临床中常采用情境模拟方法现在正在逐步尝试使用机械的感受器或监视器,以确保在实际和人为的情境中对行为进行连续的客观记录 (W. W. Tryon, 1985, 1991)。

患者自我报告由许多不同的技术组成,包括治疗家的临床面谈、目标行为和患者所保持的相关状态的自我监测记录、各种书面的检查项目及问卷。一些标准化自陈问卷的原始版或改编版,主要用来筛选、鉴定和监测目标行为。其中最简单、使用最广泛的是修订的贝克抑郁问卷 (Beck Depression Inventory, BDI, Beck & Steer, 1993)。这是一种具有 21 个项目,通过自评来评定抑郁严重程度的工具。<sup>[46]</sup>一种最新的标准化测量工具是酒精使用调查问卷 (the Alcohol Use Inventory, Horn, Wanberg & Foster, 1990) 它具有 228 个自陈项目,以在概念和心理测量上都很精致的方式来评定酒精的使用。它包括了不同概括水平的多种量表 (综述见 Drummond, 1995; McNeely, 1995)。许多为特定的研究项目或治疗计划的使用而准备的其他工具,尽管其中的一部分没有公开发行,但它们在一些杂志论文或著作中常常有充分的描述和转载 (参见 A. S. Bellack & Hersen, 1988)。

一些由多种报告者完成的客观评定量表组成的测量工具最近正在编制中。其中最全面的是塞西尔·雷诺兹 (Ceceil Reynolds) 和兰迪·坎普汉斯 (Randy Kamphans) 的儿童行为评定系统 (Behavior Assessment System for Children, BASC, 1992), 这在第十六章已经提到过。这一系统包括父母和教师评定的行为评定量表和课堂行为直接观察记录及编码的表格。此外, BASC 提供了儿童自陈问卷和一种结构性面谈程序,通过后者,父母能提供儿童的发展历史。社会技能评定系统 (the Social Skill Rating Sys-

tem, SSRS, Gresham & Elliot, 1990) 也提供由父母、教师、学生本人完成的表格, 来评定教育和家庭情境中学生的积极行为和问题行为。据一些评述者的观点, SSRS 一种特别有用的特征是它提供了一个组成部分, 把评定的结果和干预策略的计划联结起来(概述见 Benes, 1995; Furlong & Karno, 1995)。

### 职业评定

职业评定的实践是根据个体的能力、兴趣、目标、价值和气质, 以及特定职业的要求, 帮助个体辨别其最恰当的职业选择。生活中很少有像职业这样重要的领域了, 这不仅是因为大多数人大量的时间是花在工作上, 还因为工作提供了许多获得内外奖赏的机遇(Super & Sverko, 1995)。此外, 当代就业性质和条件快速变化的形势, 促使更多的人在一生中不止一次(常常几次)地考虑他们的职业选择。因而, 职业选择和职业发展的理论以一种较快的速度发展也就毫不奇怪了。在这一领域, 20 世纪 90 年代形成的主要新理论的数量就等于或超过 50 年代以来的总和。在 50 年代唐纳德·萨佩(Donald Super, 1953)和约翰·霍兰德(John Holland, 1959)首先在这一领域作出了有意义的贡献。<sup>[47]</sup>

从测验的角度来看, 我们已经考虑了那些直接应用于个体职业咨询的工具, 如第十四章讨论的兴趣问卷, 第十章描述的多重能力倾向成套测验和本章的职业测验部分的工具。选择一种职业常常意味着选择了一种生活方式, 它有一套特有的价值。因为兴趣问卷调查实质上是评定个体的价值系统, 在有效的职业计划中, 它们变得越来越重要。在这一节, 我们讨论两种更为专门的、特地用于职业咨询的工具, 即职业探究综合纲要(comprehensive programs for career exploration)和职业成熟测量(measures of career maturity)。卡佩斯、马思迪和怀特菲尔德(Kapes,

Mastie & Whitfield, 1994)编写了一套对所有类型职业评定工具都必不可少的指导手册,现在已是它的第 3 版。在其中有更多类型工具的描述和评论。

**职业探究综合计划** 有几种多重能力倾向成套测验已经整合到职业指导系统中,具体实例是第十章描述的区分能力倾向测验(DAT)。它可以和职业兴趣问卷(CII, Psychological Corporation, 1991a, 1991b)结合使用。在职业指导中,为了有利于比较这两种测验的结果,这两种工具是一起发展的。

另一个例子是美国职业服务处编制的计划,在本章职业测验部分讨论了其中的一般能力倾向成套测验(GATB)。其中源于 USES 的职业咨询项目的最有使用价值的工具是职业探究全面指南(*Complete Guide for Occupational Exploration*, CGOE, Farr, 1992)和职业探究增强型指南(*Enhanced Guide for Occupational Exploration*, EGOE, Maze & Mayall, 1995)。其目的是供咨询者、学生和求职者使用。这些指南按照主要兴趣领域、能力模式和成功绩效的其他要求对几千种职业进行了分组;个体能根据这些指导手册对职业进行初步探索,以确定自己对哪些职业组有浓厚的兴趣,然后再查阅它们所要求的训练和技能。CGOE 列举了《职业名称词典》(*Dictionary of Occupational Title*, 美国劳工部, 1991)中的 12741 种职业,而 EGOE 只列举了 2800 种,这覆盖 95% 的劳动力,但它对每种职业提供更多的信息。

职业咨询中另一种最新的方法是提供一种方法以整合来自于各种信息源的信息,形成一种综合性的探究纲要。这些信息可能包括各种测验的分数(每种有自身的常模和解释数据)、生平资料(包括教育和工作经历)和个体所表现出来的兴趣、偏好和价值观。几种可利用的工具,在不同程度上对这种方法作了说明,如哈林顿—奥西(Harrington-O'Shea)的职业决策系统修订

版(Career Decision-Making System-Revised, 1993 ) 和美国大学生测验职业计划纲要( ACT, 1994)。

这种综合性职业探究计划的一个杰出例子是第三章引用的相互指导信息系统的修订版 ( the System for Interactive Guidance Information, SIGI - PLUS)。通过使用交互作用程序, SIGI - PLUS 使个体能和计算机进行双向交流——提出和回答问题, 提供数据和征求信息。这一项目包括一个关于职业特性和要求的资料库, 并为组合其他的地方性数据作了准备。SIGI - PLUS 最初使用于专科和大学学生, 后经改进也适用于不同生活阶段希望变换职业或进入职业市场的成人。该计划的设计是: 通过考察一些相关的事实, 有计划地向有效决策前进, 从而对个体进行指导 ( M. R. Katz, 1993, Norris, Schott, Shatkin, & Bennett, 1986)。然而即使是这样一个设计良好的计算机实现的系统, 也不具有为个体完成整个职业决策的功能, 可能需要受过训练的咨询者的干预来鼓励个体考虑不同职业发展阶段的具体需要和个人特质 (参见 Tiedeman, 1994)。

**职业成熟评定** 另一类特地为职业咨询编制的工具是关于个体职业成熟水平的。这一概念起源于萨佩和其同事进行的一项关于职业发展的长期研究项目 (Super 等, 1957; Super & Overstreet, 1960)。职业成熟指个体掌握(mastery)和其年龄水平相适应的职业任务及处理这些任务的有效程度。这一研究本身由对约 100 名九年级水平的男孩长达 20 年的追踪调查构成, 其结果显示, 初中水平职业发展的主要任务是为职业选择作准备。另一项研究运用横向和纵向方法研究了职业发展, 提供了有助于解决这一问题的有益数据 (参见 Crites, 1969; Gribbons & Lohnes, 1982; Super, 1980, 1985)。最近研究的一项主要发现表明, 所有类型的职业行为可以出现在人一生中的大部分阶段, 尽

管每一人生阶段有某种起支配作用的职业行为。成人生活中职业变换问题逐渐受到了关注,职业的变换可能源于个体自身生活和职业性质两方面的各种因素(Kummerow, 1991; Lowman, 1991, 1993, Pickman, 1994; Walsh & Dsipow, 1993)。

职业发展研究的一个副产品是职业成熟标准测验的编制(Kapes 等, 1994, pp. 241 ~ 272),如萨佩编制的职业发展问卷,用来评定对职业决策的准备状况并着重于个体在职业定向中需要帮助的方面(A. S. Thompson & Lindeman, 1981, 1984)。另一种相关的工具是职业信念问卷(Krumoltz, 1991),是作为职业咨询的辅助工具而发展起来的,其具体目的是确定那些可能阻碍个体达到职业目标的信念。

### 临床判断

临床工作者在评定患者所做的工作可以认为是对人的认知或人际知觉——通过这一过程一个人开始认识理解另一个人——的一个特例(Kouglanski, 1989)。但临床情境中的判断正确性与日常社会环境中有几个重要方面的区别。关于临床情境中数据的加工、综合和临床工作者解释作用已有较多的论述。对临床判断过程的研究清楚地说明了判断过程中一些可能性的错误信息,如文化刻板印象的影响和对错误的预测原理的依赖;后者的事例包括忽视基线率或回归效应,以及假设相关较高的预测有较好的效度(参见 L. R. Goldberg, 1991)。

对于同一批事实,如测验分数或‘生活—历史’资料,临床判断是否比回归方程或其他经验研究所得到的公式的常规应用能对后续行为提供更为准确的预测?这一问题既具有实践意义又具有理论意义,因为某一统计策略或算法一旦发展起来,就可以由工作人员或计算机执行,能使临床工作者有时间从事其他工



作。在一本题为《临床和统计预测》(*Clinical Versus Statistical Prediction*)的经典著作中,米尔(Meehl, 1954)讨论了临床判断和调查研究过程,比较了两种类型的预测性。米尔的讨论表明,除了一个值得讨论的特例以外,统计程序的常规应用至少能产生与临床分析一样甚至更多的正确预测。米尔著作的出版和后来对其结论的重新肯定引起了至今仍然很活跃的激烈争论(参见Anastasi, 1988b, pp. 511 ~ 515; Dawes, Faust & Meehs, 1993; Kleinmuntz, 1990)。

当临床判断和统计程序用于同一批资料时,尽管统计程序具有明显的优势,但必须记住临床评定过程具有某些优点,这是很重要的。例如临床方法的一个主要贡献是:某些领域的资料不能通过满意的测验来获得时,可以通过细心面谈和行为观察来获得。临床方法还比统计更适合于对一些罕见的特殊事件进行处理,它的发生频率太低,不允许应用统计策略。

总之,在我们现在的知识状况下,最有效的方法是临床和统计方法的结合(Matarazzo, 1990)。临床工作者必须使用所有的客观测验数据和适合于具体情境的统计策略,而只能通过临床法获得的事实和推论补充这些信息。对实际结果所作出的临床预测效度,必须尽可能地对其进行系统的研究。<sup>[48]</sup>还需有更多的由不同临床工作者及同一临床工作者在不同的时候对同一个体所作预测的一致性的研究资料。只要有可能,临床预测所依据的过程和线索应明确记载在临床记录中。这种实践不仅有助于研究和训练,而且能促进对可靠的资料和有辩护力的解释的信任。最后,“作为工具的临床工作者”(clinician as instrument)是个重要的概念,这在正在进行的有关“调节临床判断正确性”的特征研究中得到了说明。通过系统的调查研究,斯彭格勒(Spengler)和他的同事得出结论:认知复杂性水平较低的心理学

家比水平较高的心理学家更容易形成偏向性临床判断(Spengler & Strohmmer, 1994; Walker & Spengler, 1995)。

**评定报告：最后的综合** 在第十八章，我们将讨论测验结果交流中一些广泛的问题，特别是关于其伦理和社会含义问题。对于临床工作者来说，这种交流经常包括书面测验或个案报告的准备，随后常常是和患者、父母、教师或其他有关专家进行讨论或协商。即使在一些不需要书面报告的情况下，准备一份记录作为将来的参考也是一个较好的主意。报告的准备过程也有助于组织和澄清临床工作者自己对于个案的思考，使他的解释具有洞察力。报告写作代表了临床工作者综合功能的最后阶段，报告的内容必须吸收临床工作者可利用的所有原始资料（测验的和非测验的）。

有几本书为报告的写作提供了指导。<sup>[49]</sup>虽然在这里我们不准备重复这些书中所列举的建议，但必须强调其中一些要点：第一，没有一个适用于所有报告的标准的形式或纲要，其内容和风格必须是实际上也确实随评定的目的、实施的情况、报告提交的对象、临床工作者的理论取向及专业背景而变化的。使报告适合收取人的需要、兴趣和背景，这是十分重要的。例如，给律师的报告需要十分不同于给心理治疗家的；然而，对于以上两种人，临床工作者都必须从他所收集的大量资料中选择有关的内容来回答一开始就存在的问题。

报告应该集中在每一个体具有区分意义的特征上（高位点和低位点），而不是那些个体位置与平均值相近的特质。对报告有效性的检测就得看它究竟是某一个体独特的，还是同样适合其他人。适合大多数人的一般性的、程式化的虚假报告的准备是件相对容易的工作。大量的研究显示，这些报告很容易被大多数人视为“十分正确”的自我描述（Goodyear, 1990; Klopfer,

1983; Snyder & Larson, 1972; Tallent, 1992, pp. 236 ~ 238)。这种虚假效应称为“巴纳姆效应”(Barnum effect)。菲尼亚斯·T·巴纳姆(Phineas T. Barnum)是一个著名的戏剧主持人,以一句“每分钟都有一个傻瓜出生”的评论而获得声望。相信这些能应用于一般情况的人格描述是算命者和其他骗子的最好手段。

尽管在某些案例中测验记录和其他详细数据可能分别附在报告之后<sup>[50]</sup>,但报告的重点在于解释和结论。一些特殊资料如个体反应和分测验分数一般应用来说明或澄清某一点。报告应该仔细地加以组织和综合,使其简单、便于交流而不使人模糊。有关如何准备评定报告的书籍常常包括了一些有用的提示和一些可参照的标准写作方式的指南;一本特别有趣的、使写作对作者和读者都不太费力的小书是斯特伦克和怀特(Strunk & White, 1979)的《文体基础》(*The Elements of Style*)。

### 计算机在心理评定中的作用

近几十年已经有了包括自陈人格问卷在内的多种测验的计算机计分程序。大多数的计算机程序也进行惯常的统计分析,并提供诸如各种派生分数、据 SEM 而产生的分数距离、分数剖面图等进一步的信息。陈述性的咨询报告把计算机技术更复杂地应用于测验数据的报告和使用中(Batcher, 1987; Moreland, 1992)实质上,这些程序使用了大量依附于量化分数的特定水平或模式上的定性的解释性陈述数据库。除了节约临床工作者的时间外,这种程序还有其他优点。计算机能够系统而一致地搜寻广泛得多的数据资料,这对某一个临床工作者来说是不可能的。在选择恰当的陈述中,它能可靠地把更复杂的规则应用于统计数据中。它能插入许多其他相关变量资料,如不同正常群体的人口统计学资料。<sup>[51]</sup>

毫无疑问,计算机对心理评定的潜在作用给人的印象是深刻的(Butcher, 1987; Eyde, 1987; Gutkin & Wise 1991; Moreland, 1992),其许多潜力现在才刚刚开发(埃布莱逊, 1992) 例如分支技术和适用性测验(branching techniques and adaptive testing)的应用,其优点在能力测验中已被广泛认识到,现在刚刚找到其应用于人格测验中的方法(Ben-Porath & Butcher, 1986; Jackson, 1985, 1991),但仍然没有产生适合于临床应用的工具。

另一方面,计算机技术导致了认知机能评定的许多新工具的不断更新和快速增加,它们已在临床神经心理、学习不能和注意障碍评定中加以使用(参见 Krug, 1988, 1993; Stolf & Couch, 1992),在不久的将来,神经认知功能的适当评定虽然不可能只由计算机化测验施测的方式来完成(Golden, 1987),但计算机确实可能更精确地改变任务呈现的条件以便评定不同任务成分的操作水平。它们也以纸笔测验或个别测验不可能的方式记录和评价反应参数,如时间等。米克洛库格(MicroCog)认知功能评定工具(Assessment of cognitive Functioning, MicroCog, Powell 等, 1993; Powell & Whitla, 1994a, 1994b)是一种最近发展起来的、旨在检测成人认知障碍的可能性症状的计算机施测工具;麦克洛库格测验由注意和心理控制、记忆、推理与计算、空间加工与反应时等领域的 18 种分测验组成,使用了计算机技术特有的能力。这种相对快捷便宜的测验工具的潜在的使用领域包括:老年人轻度至中度水平认知机能退化的评定;可能暴露于有害环境的雇员认知操作的监测;一些要求精确评价神经认知变化的评定和监测。

许多评定特殊机能的新工具已经编制完成,毫无疑问它们还将继续出现。最早的这种性质的测验是几种连续作业测验,如注意变化测验(Test of Variables of Attention, TOVA, Lark,

Dupuy, Greenberg, Corman & Kindschi, 1996), 包括有听觉和视觉的版本。其他的例子还有夜班连续操作测验 (Vigil Continuous Performance Test, Cegalis, Cegalis & Bowlin, 1993) 和定步系列增加听觉测验 (the Paced Auditory Serial Addition Test, PASAT, Cegaliz & Birdsall, 1995)。两种测验都含有多媒体软件组合用于视觉和听觉通道的注意评定。图 17-5 显示了一名儿童在进行 TOVA 视觉版的测验。

从另一个角度来看, 计算机有可能综合来自于各种类型的测验、个案历史、直接行为观察等许多信息源的资料 (参见 Watkins & McDermott, 1991)。尽管计算机能够承担临床工作者对某个个案进行的综合工作, 但目前还没有为开发和维持这种



图 17-5 儿童在参加注意变化测试 (TOVA), 他通过按压一个和计算机一个端口相连的外部微型开关机械装置来反应。

## 综合程序所需的计算机数据库

在当前的应用中，存在一些必须考虑的严重的潜在问题。大多数可利用的计算机辅助测验解释系统把临床和统计程序结合起来。在不同的系统中，量化资料和临床判断的具体结合是不同的；数据库的技术性质和判断的临床专业知识也经常在发生变化。此外，因为各系统特有的关注问题不同，所以评定某一具体系统所需的信息常常不能得到。主要由于缺少必要的技术信息，计算机化测验解释程序的可能误用引起了广泛的关注 (Eyde & kowal, 1987; Fowler & Butcher, 1986; Matarazzo, 1986a, 1986b. MoreLand, 1987)。某些程序已经达到了充分的科学和专业标准或正在修订以达到这一目标，有些则性质不明，从未受到资深专业人士的评述，还有很多通过其宣传文献中未经证实的资料进行过分的吹嘘。第一份关于计算机化解释服务程序的评定及使用的指导纲要，是由美国心理学会于 1986 年颁布的。对这一指导纲要的补充和修订已整合到现在正在编写的新《测验标准》( *Testing Standards* ) 中。其他有关计算机化的心理测验工具使用的指导可以在另一些出版物中得到 (参见 Bersoff & Hofer, 1991 ; MoreLand, 1992)。

总评总之，由各种专业人员参与的心理评定领域如本章讨论的其他领域一样正在发生快速变化。除了以计算机为基础的评定系统正在不断加速发展，以及在本章最后一节已经显示的其他一些趋势之外，现在重新强调需要定向于积极的心理健康而不是心理治疗的评定工具。一个最近的例子是生活质量问卷 (Quality of life Inventory, QOLI, Frisch, 1994) 。它是一种生活满意度的测量，可用于治疗计划和治疗结果的评价中。这种类型的其他几个量表现在处在不同的发展阶段。另一种有意思的、独特的新工具是跨文化适应性调查表 ( Cross-Cultural Adaptability

Inventory, CCAI – C, Kelley & Meyers, 1993) ；正如其名字暗示的一样，这种自我评定工具其目的在于帮助个体确定适应其他文化的准备状况。大学生适应性问卷（The Student Adaptation to College Questionnaire, SACQ – R, W. Baker & Siryk, 1989）是另一种工具，像 CCAI 一样，代表了心理测验在个体自我理解和自我提高方面的应用；这种应用是咨询心理学影响的直接派生物，在将来可能得到不断扩大。

## 注 释

- [1] 在 1985 年的测验标准(AERA, APA, NCME)和他们提出的修订版(见第一章)都有一章专门讨论测验在教育中的使用问题。
- [2] 关于测验称谓发生变化的一个明显的事例是于 1994 年生效的大学入学考试(College Board)考试的新名词。众所周知的 SAT 的缩写体现在代表学业评定测验(Scholastic Assessment Test)，而不是学业能力倾向测验(Scholastic Aptitude Test)。新的 SAT 分为两个组成部分即 SAT – I 推理测验、代替以前的学业能力测验。SAT – II 学科测验(Subject Test)代替以前的成就测验。这些变化与测验的其他实质性革新相伴出现。这一点我们将在以后的章节中讨论。
- [3] 这种重叠性可以通过检测具有这些称呼的测验之间的内容相似性及其相关程度而得到经验性的说明（参见 W. Cloeman & Gureton, 1954; Cooley & Lohnes, 1976）。
- [4] 在数学、科学教育领域（参见 Penner, Batsche, Knoff & Nelson, 1993）和加强思维技能方面(Mulchay & Andrews, 1991)，最近出版的论著说明了良好的、解决问题定向的工作可以来自于这种批评性的检测。
- [5] 必须注意许多大规模的教育测验项目中所应用的工具，如 SAT、NAEP 和其他的标准化成就测量，有时一直采用问答题和其他开放式问题的形式。
- [6] 当然，测验项目的分类除了建构或选择的维度外，还可以从其他不同的维度进行。项目类型的两种不同分类的实例，还可以在本纳特（R.

- E. Bennett, 1993)和 R.E.斯依(Snow, 1993)的著作中找到。
- [7] 尤其见本纳特和瓦特 R.E. Bennett & Ward, 1993)的著作。
- [8] 这种趋势的例子是 CTB/麦克格劳-希尔公司(CTB/McGraw-Hill)最近出版的特纳诺瓦(TerraNova)系列,这个系列的组成成分包括新的基本技能综合测验;多重评定版,它把选择反应和建构反应项目结合在一起;行为评定版提供了由各地自己评分或由出版者评分的较长的开放式任务;常规成分版提供了补充项目试图评定特殊课程的目标。
- [9] 关于最低能力测验运动和其中的心理测量技术问题的全面讨论,见贝 克 1986 年的著作。另一种关于使用测验来证实学生能力的观点和问题的详尽评述见杰格尔(Jaeger)1989 年的著作。
- [10] 这一讨论隐含了“机能性读写算”(functional literacy)的概念(Sticht, 1975)。这一概念已经扩展至用口头和书面语言使用的理解和使用各种文件,以及算术计算来表示能力。机能性能力是由实践情况中的要求来界定的,如完成特定工作所需的阅读难度水平和数量或更广泛的、在现代社会中处理自己生活所需的基本教育技能。
- [11] 关于课堂测验和其他测验程序准备的进一步指导,请看林和格兰特(1995,第三章至第五章)的著作。
- [12] 测验说明的例子可以在安娜斯塔西(Anastasi, 1988b, p. 431)、林和格兰特(1996, p. 122)的著作中找到。
- [13] SAT 的进一步的历史背景见安娜斯塔西(1988b, p. 328 ~ 331)和唐隆(Donlon, 1984)的论著。
- [14] 到 90 年代早期, SAT 的言语和数学两个领域的平均分从 500 分别下降到 424 和 478。
- [15] 因为在选拔大学生中的作用, SAT 测验经常容易受到批评性的检查,如最近有报道说,参加测验的人可以根据 SAT 阅读任务而不参照伴随问题出现的信息就能正确地回答许多多项选择问题。这一发现重新引起了关于外部背景知识对 SAT 分数影响程度的争论(参见 S. Katz & Lautenschlager, 1995)。
- [16] 在 1982 年之前,一般测验称为能力倾向测验,学科测验则称为高级测验(Advanced Test),因为有 SAT 改变了这些测验的名称以避免对其目



的可能产生的误解。

- [17] 这些因素中最后一点的结果表现在全国教育目标的颁布中,它规定到 2000 年所有的美国儿童必须为学校学习作好准备(教育标准和测量全国委员会,1992)。
- [18] 实例请看:贝克和奥尼尔(E. I. Baker & O'Neil's, 1994)的教育和训练评定技术革新集 斯依和朗曼(R. E. Snow & Lohman, 1989)的认知心理学在教育测量中的应用;欧克兰德和汉布尔顿((Okland & Hambleton, 1995)的学术评定的当代国际发展。
- [19] 事实上,集教学、评定、管理于一体的能适用于个体的计算机化的系统很快出现了,如由 ETS 为需进一步提高学业技能的成人的学习者开发的附加学习软件(Learning Plus Program),IBM 的 K-12 年级的学校展望软件(IBM's School Vista software for K-12)就是当前这类工具的两个例子。
- [20] 特恩迪斯、邓内特和洪(Triandis, Dunnette & Hough, 1994)最近主编的手册,涉及关于世界各地不同文化背景中的工业和组织心理学问题。
- [21] 安娜斯塔西(1988b, pp. 468~474)讨论了职业测验问题,包括用于确定心理学家的资格和颁发心理学家执照的测验。关于一般执照和资格测验的效度分析和其使用问题的最近考察,请看由拉杜卡(LaDuca, 1994)编辑的关于这一问题的特刊:《健康保健专业人员的评定》(Evaluation the Health Professions)。
- [22] 工作分析方法的进展和问题的概述可以在哈维(Harvey, 1991)的论著中找到。克纳普、卢瑟尔、坎贝尔(Knapp, Russell & Campell, 1993)描述了工作分析在军队服务人员的选拔和分类中的特殊应用。
- [23] 有关预测作为人员选拔的一种范式的局限性的激烈讨论,请看得沃尔弗(de Wolff, 1993)的著作。
- [24] 如想简单了解影响个人和情境相互作用成功的变量多重性,请看斯滕伯格(Sternberg, 1994a)的著作。
- [25] 见盖恩(Guion, 1991)关于人事评定、选拔和安置过程的评述,包括各种测验和其他预测工具的效度和公正性的有关结果。
- [26] 这一部分的背景知识,见第十、十一章。

- [27] 对 GAPD 中的临界分数 SATB 组提供了一种替代途径。第九章详细地讨论了多重临界策略。与咨询相反,在人事选拔领域,临界分数最适当的用处是初步甄别求职者的一种或多种关键技能。
- [28] 还有计算机化的 ASAVB 测验版本 (CAT - ASAVB),自 1979 年来,它一直在发展,现在用于军队人员的筛选 (T. L. Russell, Reynolds & Campell, 1994)。对其发展的描述,见威斯柯夫和斯库拉兹 (Wiskoff & Schratz, 1989) 的著作。
- [29] 为开发编制基于操作的测量而进行的概念和方法的研究工作对, JPM 计划是有贡献的。JPM 需要在教育测评和其他领域证明其可应用性。
- [31] 关于有身体要求的职业的人事选拔程序的综述由霍根 J. C. Hogan, 1992 的论述。
- [32] 这些测验和几种关于计算机的测验工具来自麦克格劳-希尔公司 (McGraw-Hill/London House) 出版的 SRA 系列。
- [33] 见阿纳斯塔西 (1985e) 工业中的人格测验的历史回顾和与其相连的几种基本方法和应用问题的讨论。康弗等人 (Kanfer et al., 1995) 从人格和智力理论到工业和组织心理学,评述了各种概念应用的进展。
- [34] 由心理学家写的关于谎言检测和测谎仪问题的最早的评论之一是莱克恩 (Lykken, 1981) 的著作。较为简短的和近年的一些概述和讨论可以在德保罗 (DePaulo, 1994)、霍斯 (Honts, 1994)、柯其尔和拉斯克 (Kircher & Raskin, 1992) 以及萨克斯 (Saxe, 1994) 的著作中找到。
- [35] 具体的例证见美国心理学会组成的调查此问题的专门工作组的报告 (L. R. Goldberg, Grenier, Guion, Sechrest & Wing, 1991)。从 K. R. 墨菲 (K. R. Murphy) 1993 年编的工作诚实问题专集中可以得到进一步的信息。人事选拔中诚实测验的简短而有启发性的讨论,见卡纳拉和斯切内得 (Canara & Schneider, 1994)、萨克特 (Sackett, 1994) 的著作以及《美国心理学家》1995 年 6 月的评论部分。
- [36] 对当今工作中发生的转变和对人事工作可能带来后果的讨论。见兰迪、桑克斯特和柯勒·摩兰 (Landy, Shankster & Kohler Moran, 1995) 的著作。
- [37] 司法心理学是心理学在法庭中的应用,涉及法律、判决、监禁等问题。

- [38] 为了便利起见,在本节中用临床工作者这一术语指从事心理评定实践的任何专业人员。但是,必须记住这些专业的教育背景可能是几种专业领域中的任何一种,因而他们的理论取向和特长各不相同。关于临床、咨询、工业/组织、学校心理学家在他们传统领域中使用测验的有关信息,见 P.S.威斯(P.S. Wise)1989 年的著作。
- [39] 心理评定研究的当前方法问题的代表性论文可以在《心理评定》一期特刊(1995 年 9 月, Vol. 7, # 3)中找到。
- [40] 见阿纳斯塔西(1985a)关于评价这一研究时必须考虑的一些方法上的要求的讨论。
- [41] 关于考夫曼(Kaufman)方法的各种批评和反驳的有关信息,见考夫曼(1994,第一章)。关于考夫曼方法几种否定评论的一个问题是它们假定临床工作者单单依靠分数大小和差异作出决定。的确,剖面图分析技术的机械应用会产生误导,而这种假设和考夫曼的建议以及良好评定实践的原则是很不相同的。
- [42] 在《心理测量年鉴》一书中有关于这类测验的经常性评论(见 TIP - LV, P. 1116, 关于当前使用的测验)。
- [43] 卢瑟尔和斯塔卢(Russell & Starkey, 1993)编制了一套整合的计算机化系统,用于编辑 HRB 及其他工具中的许多测验原始分数,再把它们转化为根据年龄和教育情况校正的统一量表;为了有助于解释,这一程序也描绘结果剖面图并总结了结果。赫顿、格兰特和马修斯(Heaton Grant & Matthews, 1991)也准备了一套相似的系统,最近,实现了计算机化。
- [44] 对 P. L. 94 ~ 142 和其在诊断 LD 中分类的进一步讨论,见雷诺尔斯(C. R. Reynolds, 1990, 1992b)。对于学习不能及其评定,萨特勒(Sattler, 1958, pp. 598 ~ 617)的著作提供了一个简短而具有启发性的概述。
- [45] 这一类型的工具在《心理测量年鉴》中有定期的评述(见 TIP - IV, pp. 1095 ~ 1096, 有当前的列表)。有关计算机技术在各种类型行为评定程序中的运用,见柯拉乔威尔等人(Kratchowill et al., 1991)的著作。
- [46] BDI 的广泛的文献和早期版本的概述,见 MMY 第 11 期。
- [47] 见布朗、布鲁克等人(Brown, Brooks et al., 1996)关于这些和其他较好

的关于职业选择理论的概述，以及一些会引起争论的文章。萨维卡斯和莱恩特（Savickas & Lent）的《职业发展理论》是另一个关于这些理论和它们是如何相似、区分和相互补充的有价值的材料。这两本书中的许多章都是理论家的原著。

- [48] 一些旨在评价基于评定程序的决策或预测的效度的大规模的计划正在酝酿当中。其中一项已在进行中的研究是在人格评定协会的负责下，对使用人格评定程序或工具预测治疗的各个方面，预测医疗、立法或生活结果的研究进行完整的文献综述和随后进行的元分析（Handler, & Meyer, 1996 春/夏）。
- [49] 萨特勒（1988，第二十三章）综述了报告写作中的问题和易犯的错误，提出了许多实用性建议，并总结了关于报告写作的研究。其他伴有案例报告说明的资料包括欧恩柏（Ownby, 1991）和塔尔雷特（Tallent, 1993）的著作。
- [50] 见《美国心理学家》1996 年 6 期，由美国心理学会心理测量和评定专业委员会编写的“测验资料公布的声明”（Statement on the Disclosure of Test Data）。
- [51] 有关测验剖面图报告、陈述报告和其他计算机辅助测验解释软件的大多数出版者和发行者，为可能的使用者提供了样本报告和演示光盘。此外，心理软件信息手册（Psychware Sourcebook, krug, 1993）是当前计算机化评定成果的一种最完整的信息资料，其中包含了许多样本报告的复制品。

## 第十八章 测验中的伦理和社会问题

在测验程序研究和其实际应用中,心理学家长期以来一直关注着职业伦理问题。这种关注的具体例证是 50 年代早期开始的为发展第一套正规职业伦理规范而进行的系统经验研究计划。这一广泛的研究产生了被美国心理学会 (APA) 正式采纳,并于 1953 年首次颁布的一套标准。这些标准不断地得到评论和提炼,导致了定期颁布其修订版。当前的版本《心理学家的伦理原则和行为规范》( *Principles Ethical of Psychologists and Code of Conduct*, APA, 1992),<sup>[1]</sup> 是由序言和 6 个一般原则组成的。它们用来指导心理学家迈向职业的最高理想,也为不同领域工作的心理学家提供了 8 条强制性准则的伦理标准。

“伦理规范”由负责对学会会员的投诉进行调查和裁决的 APA 伦理委员会 (APA Ethical Committee) 具体执行。该委员会的运作规则和程序及其年度报告都发表在 APA 的官方期刊《美国心理学家》( *American Psychologist*, APA 伦理委员会, 1995, 1996) 上。参与“伦理规范”最新修订工作的上届伦理委员会成员准备了一本必不可少的指导手册,其中包括了对伦理标准应用的评论和解释 (Canter, Bennett, Jones & Nagz, 1994)。最近由伯索福 (Bersoff, 1995) 编辑的另一本关于心理学伦理的历史和当代文献集,讨论了不同情形下伦理的两难处境。波普和瓦斯奎兹 (Pope & Vasquez) 的《心理治疗和咨询中的伦理》( *Ethics in Psychotherapy*

and Counseling)一书,用了一章的篇幅讨论评估问题,提出了一些很好的应用建议。韦纳(Weiner,1995a)最近在一本书中专列一章讨论了如何预见人格评估中的伦理和法律的挑战。

在90年代我们看到了:联邦和州政府层面的立法行动、法庭判决、不同角度的职业指导纲要的大量增加。其中许多是针对心理学的一般应用和特定心理测验的使用。其中有一些我们在前面有关特定类型的测验问题及其应用部分的章节中已经讨论过。这些指令的组合效果对实践者常常是不明确的或不一致的,甚至是相互冲突的。

最近大家的注意力愈来愈集中于心理服务的提供者需要在职业伦理原则、法律、义务法则,以及他们工作的组织机构的政策之间取得平衡。因此,APA通过它所属的各种办事处和委员会,对有关的发展趋向进行监测,针对应用心理家可能遇到的问题发布标准、指导和说明,努力为其成员提供指导和信息。《心理服务提供者的一般指导纲要》(*General Guidelines for Providers of Psychological Services*, APA, 1987a)和《服务项目引介的专业指导》(*Specialty Guidelines for the Delivery of Services*, APA, 1981,正在修订中)的颁布,为几种领域的心理学专业实践提供了帮助。还有一些具体问题的指导,由另外一些文件提供,如专业实践和标准委员会起草的《离婚程序中儿童监护评价指导》(*Guidelines for Child Custody Evaluations in Divorce Proceeding*, APA, COPPS, 1994)和本章详细阐述的几种文件。此外,80年代后期以来,APA准备了一系列的丛书,总结和分析了各州法律对心理健康专业人员的影响,其中约有12本已经出版发行,另一些,最近也进行了修订(参见Caudill & Pope, 1995; Dettrila & Otto, 1995; Shuman, 1990; 1993; Wulach, 1991)。

APA心理测验和评定委员会(CPTA)具体致力于探讨良好

测验和评定实践中的问题,为其实践和其他 APA 专业组织提供技术建议。关于和测验相关的问题及实践的指导,也已颁布了几种文件,我们将在本章的后面部分提及。CPTA 还评述了测验实施联合委员会(the Joint Committee on Testing Practices, JCTP)的工作。它是一个由 APA 和其他关心测验的专业组织建立的团体。联合委员会编制了《教育中公平测验应用准则》( *Code of Fair Testing Practices in Education*, JCTP, 1988)和其他文件,其目的在于改善测验应用的状况和防止它的滥用 ( Eyde et al, 1988, 1993)。目前 JCTP 正在准备一份关于“测验被试的权利和责任”( *Rights and Responsibilities of Test Takers* )的声明。

第一章我们讨论了被试的看法和测验使用者作用的某些问题;本章我们将注意转向影响不同领域中测验使用的伦理和社会问题。此外,关于职业能力的问题,我们将简单讨论测验发行者的责任、被试的隐私权、诚实问题、不同背景和能力水平个体的测试问题。虽然,在某种程度上,我们将涉及立法的影响,但有关测验使用法律方面多重性的详细讨论超出了本书的范围,对于这一问题感兴趣的读者可以参考本章和本书其他章节所提到的文献(见第九章和第十七章)。

## 心理测量和评定中的伦理问题

自 70 年代以来,人们不仅一直关注着理论和应用心理学各领域的伦理问题,而且也关注着较为广泛的价值问题 ( Bersoff, 1995; Diener & Grandall, 1978; Tacob & Hart Shorne, 1991 Pope & Vasquez, 1991)。在测验领域,埃迪和奎因坦斯 ( Eyde & Quaintance, 1988) 和梅西克( Messick, 1980b, 1989, 1995)等人对各种实

践价值的作用和伦理原理进行过极具思考性、启发性的分析。在较为具体的水平上,APA“伦理规范”包含了许多应用于心理测验的准则,其中之一(如评价、评定或干预)是直接关于心理评定技术发展和使用的,另一种(如司法活动)则以一节的篇幅专门说明法律背景中的评估问题。此外,隐私和诚实的伦理标准,尽管其范围广泛,像其他一般原则和几种伦理标准一样,和测验紧紧相连(APA,1992)。“伦理规范”中讨论的一些问题和第一章阐述的‘测验标准’中的有关要点是紧密相关的,事实上,《测验标准》本身的内容有助于确定测验使用的职业责任。

除了APA,其他一些相关专业团体和学会也形成了自己的一套伦理规则和准则。从测验角度来看,其中最有关的是由美国咨询协会(American Counseling Association,ACA)于1989正式通过的题为“标准化测验使用者的责任”的声明(Responsibilities of Users of Standardized Tests, the“RUST Statement”)。另一个有用的文件是工业和组织心理学会(Society for Industrial and Organizational Psychology, SIOP, 1987)为专门的目的而编制的《人事选拔程序的有效性和使用的原则》(Principles for the Validation and Use of Personnel Selection Procedures, 见第七章)。

详细说明测验在现代社会中的地位的影响事件是《能力测验使用、结果及争论》(Ability Testing: Uses, Consequences, and Controversies, Wigdor & Garner, 1982)的出版。这两卷书是一项历时4年,关于标准化能力测验在学校、进入高级学校和就业测验中的使用情况的研究项目的最终报告。从公开广泛争论测验的价值之时起,这一项目就在国家研究院的主持下由多学科委员会指导实施。从80年代早期开始,其他关于测验问题的重要研究和报告陆续出版(参见Hartigan & Wigdor, 1989, Office of Technology Assessment, 1992)。一般来说,不同研究小组的研究



果，再次证实那些重申的关于能力测验的作用及其可能误用的结论。

政府越来越普遍地卷入心理测验和其他评定工具的应用之中，导致了测验和评定委员会(BoTA)的建立。BoTA是在美国国防部、教育部、劳工部的支持下于1993年创立的。BoTA是国家研究院正在进行的一项活动，其主要目标是帮助政策制定者理解、评价测验以及其他一些作为公众政策评估的工具。这一委员会重视多种情境中测验和评定的交叉问题(cross-cutting issues)，还出版了关于这些主题的报告，如教育2000目标的立法动议(Feuer & Kobes, 1995)、一般能力倾向测验的改善计划(BoTA, 1995)、工作性质变化的评定和教育含义(Black Feuer, Guidroz & Lesgold, 1996)。

## 使用者资格和专业能力

能力的‘伦理规范’准则为‘心理学家所能提供的服务和使用的技术只能在其教育、训练或经验所允许的范围内’(APA, 1992, p. 1599)。至于规定测验只能由那些具有恰当资格的测验者使用，是防止被试接受不恰当使用测验的一项措施。<sup>[2]</sup>当然，所需的资格随测验的类型而变化。因而，相对较长时间的集中训练和指导下的工作经验，是恰当使用个别智力测验和大多数人格测验的要求，而教育成就或工作熟练性测验所需要的专门心理学训练相对较少。必须注意，为了教学目的，在课堂上参加测验的学生通常不能对他人实施测验，也不能恰当地解释这些测验分数。

受过良好训练的测试者，根据测试的具体目的和施测的对

象选择恰当的测验。他们熟悉所选定的测验的有关研究文献，能评价诸如该测验常模、信度、效度这些特征方面的技术优点；在施测中他们对可能影响测验成绩的因素——如第一章说明的那些——具有敏感性；他们根据个体其他相关信息慎重考虑测验分数，而后才得出结论或提出某些建议。总之，他们必须熟知人类行为科学知识，防止在测验分数的解释中作出没有根据的推断当测验由心理学技术员或助理员或缺少心理测量原则专业培训和适当评估实践的人实施时，需要一位完全合格的心理学家，至少是咨询师，为适当解释测验成绩提供必需的方法原理，这是很重要的。

谁是合格的心理学家？显然，因为学科的差异和训练的专业不同，没有一个心理学家在所有的领域中都是同样合格的，甚至在范围较小的心理测量和评定领域也是如此（见第十七章）。在认识到这一事实的前提下，“伦理规范”要求心理学家认识到他们的特殊能力的界限和他们的专长的局限（APA, 1992, p. 1599）；这一伦理义务的含义在前面所阐述的能力原则中进行了清楚的阐述。

在提高专业标准和帮助公众鉴别合格的心理学家方面有意义的一步是，心理学家州执业执照和资格证书法律的颁布。包括哥伦比亚特区在内的所有的州，现在都有这类法律。加拿大的大多数省也颁布了规范心理学应用的法律（所有这些法律的总结见 APA, 1993, pp. xlii - xlv）。“执照”和“资格”的术语经常交互使用，然而在心理学中，“资格”一般指心理学家头衔的法律保护；而“执照”控制着心理学的应用而不管开业者所确定的头衔是什么。“执照法”需要包括心理学应用的定义。大多数州开始时只有比较简单的资格法，但正在努力向执照法迈进。在任何一种法律中，这些要求一般是心理学博士学位、指导下的一定

量的工作经验、在资格考试中有满意的成绩。执照法一般包括对心理学家训诫处理的根据，其范围是从罚款、惩戒到暂停执业和吊销执业执照。许多管辖部门直接或间接地把 APA 的伦理要求整合到它们的法令中，所以违反“伦理规范”也经常成为一些进行训诫处理的根据。APA 编制了“州颁发心理学家执照范式条例”(Model Act for State Licensure of Psychologists, APA, 1987b)，作为各州起草规范心理学应用法律参照的标准。

更高水平的专业证书由美国职业心理学理事会 (ABPP) 提供。获得这项证书要求在指定的专业领域里有高水平的训练和经验。ABPP 通过其分支专业理事会颁发资格证书的领域有临床、咨询、工业/组织、学校心理等等。APA 的人名录包括每一专业当前证书持有者的名单，该名单也可从 ABPP 获得。作为专业领域里私人组成的理事会，对州执照和资格法的实施部门不具有强制的权威。

在过去的十年里，健康关心服务系统的转化和职业市场的其他变化，使得心理学应用中的可信问题愈加紧迫。因此，美国心理学会已经采取步骤，以便更有序地处理存在于当前情况中的许多潜在冲突，其中的一个步骤就是创立 APA 职业心理学学院 (APA College of Professional Psychology Sleek, 1995)。它将采用一种包括考试以及教育和经验先决条件的程序，对心理学中各种程度的熟练水平颁发证书。另一个步骤是发展一种正式认定心理学应用实践中各种专业和熟练程度的程序 (APA, Joint Interim Committee for Identification and Recognition of Specialties and Proficiencies, 1995a, 1995b)。毫无疑问，对有关测验使用者资格的特殊问题的进一步指导以及颁发评估专家证书的新程序也即将出现。

## 测验出版者的责任

购买测验一般限于某些具有最低资格的人。几个主要测验出版者目录册具体说明购买者需满足的要求，一般来说，需要具有心理学硕士学位或同等的资历。一些出版者参照测验使用者的资格把测验分为不同的水平，从教育成就和职业熟练程度测验、团体智力测验和兴趣调查，到诸如个体的智力测验和大部分人格测验；并对测验的个体购买者和被认可的机构购买者进行了区分。需要特定测验以进行分班或研究的研究生，必须持有负责恰当使用这一测验的心理学指导教师签署的订购单。<sup>[3]</sup>

限制测验发行的努力有着双重目的：测验材料的可靠和防止测验的滥用。我们必须注意，测验发行者可能作出了真诚的努力来实现这些目标，然而他们能够实现的控制必然是有一定限度的。在某些情况下，不可能调查和证明测验购买者所声称的资格(参见 Oles & Davis, 1977)。而且，形式上是否具有资格只是一种粗略的筛选手段，显然，如心理学硕士学位甚至是博士学位、政府执照、ABPP 的证书，并不能说明该个体一定具有使用特定测验的资格，也不能说明他所受的训练和合理解释测验所获得的结果有关。恰当使用测验的主要责任最终还是在测验使用者或有关的机构。

另一种职业责任是和心理测验作者和测验的销售者有关。测验不能过早地作为一般使用来发行，在缺乏足够的客观证据时，对于测验的优点不能作出任何声明；当只是为了研究的目的而过早地发行测验时，必须进行具体的说明，并相应地控制其发行。测验手册必须提供足够的材料，以便能对测验本身进行评

价；还需要提供关于其施测、记分和常模的充分信息，手册应该是对测验已知事实的说明，而不是把测验说得如何好的促销手段。要经常修订测验和常模以防止其过时也是测验作者和出版者的责任。当然，各种测验过时的速度，随测验本身性质的差异而有很大的不同。

一些测验在选拔、安排或诊断决策中使用时需要保密，不能部分或全部在公众媒体上公布，这是有明显理由的。某些测验的公开化会使该测验在以后对其他人使用时失去效度。此外，在媒体上公布测验，会导致一般公众成员进行自我评估，从而造成心理上的损伤。另一种几乎总是非专业的测验实践是通信测试，这种测试方法不仅没有控制测试条件，而且通常在没有个体其他相关信息的条件下对测验分数进行解释。除了少数几种可能的例外——例如对相对老练和有良好动机的个体进行兴趣或价值调查——这些条件下的测试结果可能比毫无用处还要糟糕。<sup>[4]</sup>

值得注意的是自 80 年代以来，测验出版者已开始采取步骤，保证他们出版和发行的测验得到恰当地使用，测验分数得到正确的解释。为了达到这一目的，他们努力扩大和改进与顾客关于某些测验的交流，并加强公众对测验的一般性理解。测验出版者和 APA 及其他全国性的组织一起参加测验实施联合委员会的测验使用者资格和培训的项目（Eyde et al., 1988, 1993）。此外，他们建立了测验出版者协会（ATP），其组织成员致力于提高评定服务和结果的正直性，努力增强它们的社会价值。ATP 最近出版了第 2 版《职前诚实测验的规范指导纲要》（*Model Guidelines for Preemployment Integrity Testing*, 1996）。

## 隐私的保护

和人格测验有关的一个特殊问题是对隐私的侵犯。在一项题为“隐私和行为研究”(1967)的报告中,隐私权定义为“某人决定和其他人分享自己关于个人生活的想法、情感、事实的多少程度的权利。这一权利进一步描述为“保障自由和自我决定的关键”(p.2)。因为一些情绪、动机或态度特征的测验必然是伪装的,在该种测验的过程中,被试可能在不知道的情形下显露一些特征。为了测验效果的目的,可能需要被试不知道对某种测验的反应进行解释的具体方式。尽管如此,一个人仍不应该在欺诈条件下接受任何测验。在这一方面最重要的是被试对测验结果的用途应该有清楚的了解。

虽然对侵犯隐私的关注最普遍地表现在人格测验方面,但它也合乎逻辑地适用于其他任何类型的测验。任何种类的智力、能力倾向或成就测验都有可能揭示个体不愿意表明技能和知识方面的局限性;而且,对个体行为的任何观察活动,如面谈、非正式交谈或其他个人接触,都可能得到个体试图掩盖和不愿意被揭露的信息。在有关侵犯隐私的讨论中,心理测验常常被选取出来讨论,这一事实可能反映了关于测验有一些流行的错误概念,以及它们经常被错误地用来作为对个体作出决策的唯一基础。如果所有测验都被界定为对行为样本的测量,除了行为本身外,没有任何神秘的东西,那么公众对测验的恐惧和怀疑将会减轻。同样,如果测验只在综合评定的范围内加以解释,那么在对个体作出利害攸关的重要决策时,就不会过分强调任何特定的测验结果。

我们也必须注意到在所有的行为研究中,不管是采用测验还是其他观察方法,都有侵犯隐私的可能性。但是,作为科学家的心理学家在致力于提高对人类行为的认识这一目标时,是可能会出现价值冲突的,对此他们必须针对具体情况加以解决。<sup>[5]</sup>这一问题显然不那么简单,它也是广泛争论的课题之一。<sup>[6]</sup>对个人隐私的保护不可能形成统一的规则,只可能提供一般性的指导。解决问题的方法只能是根据案例的具体情境,根据心理学家个体的伦理意识和职业责任而定。

另一个相关的因素是测验的目的,究竟是在于个体咨询,还是在于机构的人员选拔和分类决策,或者是在于研究。例如,在临床或咨询的情形下,顾客经常乐意暴露自己的问题以便获得针对性的帮助。不管测验目的如何,隐私的保护包括两个关键概念:“适当”和“告知的同意”。要求个体披露的信息必须和测验声明的目的有关。这一原则的重要含义是所有实际的努力必须是确定测验在特定诊断或预测目的上的效度。最近法律的发展,如“索洛卡与戴顿·赫德森”(Soroka v. Dayton Hudson)案件(参见 Merenda, 1995)和“1990年美国残疾人法案(P.L. 101~336)”强调了职前测验中使询问的侵犯性维持在最低水平的重要性,以及确保这类询问与工作成绩明显有关的重要性(参见 Bruyere & O'Keeffe, 1994; D. C. Brown, 1996; Herman, 1994, 第二章)。在索洛卡案件中,求职者对使用筛选测验提出了异议,其根据是询问关于宗教信仰和性喜好方面的问题(这些问题摘自 MMPI 和 CPI),这是具有侵犯性和偏见性的。尽管这一案件不了了之,但某些测验编制者,包括最近 MMPI 和 CPI 修订版的作者,已经把这些项目从自我报告问卷中删除了(见第十三、十七章)。

“告知同意”的概念也要求清晰地加以说明。它在个案中的

应用需要进行重要的判断(AERA、APA、NCME,1985)。尽管当前的“伦理规范”包括了一种明确的标准:只是在治疗上而不是在评估上要求“告知同意”,但这一要求也隐含在其他有关职业情境中评价和诊断的标准以及《规范》的其他几个部分中。而且,州心理学理事会的章程、案件法律、组织规则、和/或流行的实践标准,一般在评估和干预活动中都要求“告知同意”(Canter et al.,1994,p.67)。被试当然应该知道测验的目的、所收集资料的种类和分数的用途,然而,没有必要事先向他呈现测验项目或告知具体反应是如何记分的,在未成年人的案例中,测验项目也不应该给父母看。<sup>[7]</sup> 这些信息常常使测验失效。在测量和评估情境中出现的关于“告知同意”及相关的具体问题,在关于被试权利的“测验标准”一章进行了探讨。

## 保 密

像隐私保护一样,和隐私的保护有关的测验资料的保密问题也是多方面的。其基本问题是谁将接触测验结果?在具体情境中有几种考虑会影响其答案,其中包括测验内容的可靠性、错误理解测验分数的危险性和各种人要知道结果的需要。

个体逐渐意识到他们有权得到自己测验报告中的结果。被试应有机会评论报告的内容,如有必要可以证明或纠正有关事实的信息。咨询者现在愈来愈努力使顾客成为自己评定中的一个主动的参与者。因此,测验结果必须以不用技术行话或术语,且以容易理解的形式呈现,必须针对测验的即时目的,必须有适当的自我保护以防止测验结果的误用及错误解释。

对测验记录保密的讨论,常常要涉及除被试(或未成年人的



父母)和施测者以外的第三者是否可以接触。作为其基础的原则是测验记录在被试不知道和不同意的情况下不能公开,除非这种公开是法律所规定的或为了某种为法律所允许的有效的目的。当测验在某种组织中实施时——如学校系统、法院或就业场所——应该在测验时告知个体测验目的、测验结果将如何使用、它们将提供给哪些有合法需要的组织人事部门。当外界其他人,如一个未来的雇主或大学要求得知测验结果时,必须有被试本人同意公布数据的说明。同样的要求也适用于临床和咨询情形中的测验施测,或适用于某种研究目的。关于这一问题更进一步的指导,可以在 CPTA 发表的“关于测验资料披露的声明”(APA, 1996)中找到;它帮助心理学家处理有关测验记录披露的问题。此外,APA 法律问题委员会(COLI)为那些在实践过程中必须处理关于顾客的记录或测验数据的传票或强制性证明问题的心理学家,设计了一些策略(APA, COLI, 1996)。

另一问题是组织机构中记录的保存问题。一方面,个体的纵向资料不仅对于研究,而且对于理解个体和对个体进行咨询都是很有价值的。情况常常是如此,这些优越性必须以测验结果的恰当使用和解释为前提。另一方面,旧记录的可利用性为其错误的使用打开了方便之门,如从过时的资料中进行不正确的推论,未经许可使用于测验的原始目的之外的其他目的。例如当为高等学校入学资格评定某一个体时,引用其小学三年级所获得的 IQ 或阅读成绩分数,显然是十分荒唐的。然而,当测验记录保留许多年时,就存在这样的危险,它们可能被用于被试(或其父母)从未预料和从未同意的其他目的。当为了有利于个体而合法地长期使用或为了可接受的研究目的而保留记录时,为防止这种滥用,对它们的提取必须加以非常严格的控制。任何类型的组织机构关于个人记录的销毁、保存和提取应该形成

一套明确的政策。“记录保存指导”的阐述 (APA, COPPS, 1993) 就包括了更多有关这个问题的信息。

## 测验结果的交流

近年来,心理学家开始更多地考虑测验结果对接受者有意义和有用的交流形式。当然,信息不应例行公事地传递,而应该提供恰当的解释。操作的一般水平和用简单语言进行的质的描述比具体的数字分数好,除非是和受过良好训练的专业人员交流。甚至受过良好教育的业外人士也会混淆百分位和百分数、百分位和 IQ 分数、常模和标准、兴趣等级和能力倾向分数。但是更严重的错误解释是从测验分数得出的结论,甚至当它们的技术意义被正确理解时,也是如此。一个熟悉的例子是这样的流行假设: IQ 表示个体一种固定的特性,预先决定了个体生活中的智力成就水平。

除了测试者本人,测验结果可能的接受者中有未成年人父母、教师和学校的其他人员、雇主、精神分析者、法院和行为矫正人员。在所有与测验相关的交流中,需要考虑接受者的特性。这不仅包括接受者的一般教育情况和有关心理学和测验的知识,而且还适用于他们对信息的预期情绪反应,例如父母或教师对儿童的个人情绪牵连可能干扰他们平静而理智地接受事实信息。

当和被试本人,不管是儿童还是成人,<sup>[8]</sup>交流测验结果时,也碰到同样的问题。当和第三者交流时,防止错误解释的同样一些一般保护原则也适用于此。在这一方面,“测验标准”强调在临床和咨询应用中使用测验的人,应该为被试提供有关测验结果和根据结果而提出的建议作出恰当的可理解的解释。

当一个人正在了解其自身优点和缺点时，对测验信息的情绪反应的考虑是十分重要的。当个体得到关于他的测验结果时，不仅对其数据应该由具有适当资格的人进行解释，而且当他可能被该信息导致情绪波动时能够采取适当的对策。例如，一位大学生得知其在学术能力倾向测验上的不良成绩时，可能变得非常气馁；一位天才学龄儿童如果发现她比同伴要聪明得多，她可能会形成懒惰的习惯，或许会变得无合作性和难于管理。不管分数本身正确与否，这种有害的效果当然可能出现。即使某测验得到了正确施测、记分和恰当的解释，在知道其分数而没有进一步讨论的机会的情况下，很可能有害于个体。

咨询心理学家特别关注的问题是发展有效的方法，把测验信息传递给他们的顾客(参见 Hood & Johnson, 1997, 第十七章)。尽管这一过程的详细细节超出了我们现在讨论的范围，但两种主要的准则是特别重要的：第一，要把测验报告看成咨询过程这一整体的一部分，必须将其融入整个“咨询者—顾客”的关系中去；第二，只要有可能，咨询者需按照顾客提出的具体问题使顾客参与测验结果的解释。咨询中需要慎重考虑的是顾客对呈现给他或她的信息的接受情况。在咨询中，如果个体不管出自何种理由不接受任何信息，那这些信息有可能白白浪费。另一方面，接受经过恰当解释的测验结果对顾客具有治疗价值，尤其是在‘认知—定向’的治疗中。

## 对不同人群的施测

背景 自 50 年代以来的几十年中，我们看到了公众越来越关注少数民族、妇女和残疾人以及其他一些少数人群的权利。[9]

这种关注也反映在联邦政府和州政府有关公民权利法案的颁布中。与改善来自这些不同群体的个体的教育和职业机会的机制相关,心理测验也成了注意的焦点(Gifford, 1989a, 1989b)。心理学文献包含了许多关于这一课题的讨论,有些影响显著,有些影响则尚未显现出来。其中作用最为清晰的是专业学会提供的几个状况报告和指导纲要(参见 ACA, 1989; APA, 少数民族事务委员会, 1990; APA, 评估、测量和统计分会, 1993; Prediger, 1993; Sackett & Wilk, 1994)。此外对不同人群进行恰当测验的指导正日益增加(见第九章; Dana, 1996a; Sattle, 1988, 第十九、二十章; Sazuki et al., 1996; Valencia & Lopez, 1992)。

在国家研究委员会、技术评估处和其他这类组织赞助下准备的研究报告(在本章前面部分已有摘引),根据当前的社会情境探讨了有关测验的争论,对测验的作用表现出了一种平衡的观点。现在许多关注集中在文化状况可能影响少数人群成员的能力倾向、兴趣、动机、态度和其他心理特征的发展,因而降低了测验分数这一问题上。对这一问题提出的一些解决办法反映了对心理测验性质和功能的误解。群体或个体不同经验背景的差异不可避免地表现在测验成绩中,每一个心理测验测量的只是一个行为样本。因为文化影响行为,其影响要由测验进行检测,并应该由测验进行检测。如果从测验中能排除文化的差异,我们可能会降低它作为某一行为度量的效度,而这一行为正是测验所要评定的。这样,测验就不能提供为纠正妨碍成绩的条件所需的信息。

对不同种类的特殊群体的理论原理和测验程序,我们在第九章和第十二章进行了更为充分的讨论。在第六章把“测验偏见”这一概念和测验效度联系起来进行了技术分析。本章的兴趣主要集中于少数群体测验的专业问题和社会内涵。

法规自 1960 年以来,少数人群的教育和就业测验得到了迅速的发展。这些发展包括立法行动、执行命令和法庭判决等。有关教育测验的立法在第九章和第十七章进行了引述和评论;对外部强制测验的当前趋势和问题的评述可见林和格洛鲁德(Linn & Gronlund, 1995,第八章)的论著中。<sup>[10]</sup>

在就业领域,法庭在公民权利法律的解释和应用中起着越来越重要的作用。几个著名法庭案件的涵义,在人事和测验的文献中得到了心理学界和法学界人士的广泛讨论(参见 APA, CPTA, 1988; Berioff, 1983, 1984; Bouyere & O'keeffe, 1994; Bollander, 1982; Merenda, 1995; Meyers, 1992; Wigdor, 1982)。关系最为密切的联邦立法是 1964 年公民权利法案第七项(P.L. 88-352, 1964, 也称为就业机会均等法案)及其以后的修订法案,即 1991 年的公民权利法案(P.L. 102-116),还有 1990 年的美国残疾人法案(P.L. 101-336)。就业机会均等委员会(EEOC)主要负责这些法案的实施和执行,该委员会为这一目的编制并发行了指导纲要。1978 年为了简化程序和改善合作性,EEOC、民事服务委员会(现在是美国的人事管理处)和司法部、劳工部、财政部联合采用了“雇佣人员选拔程序的统一指导纲要”,<sup>[11]</sup>

就业机会均等法案在雇佣决策的选拔程序中禁止基于种族、肤色、宗教、性别或原国籍的歧视。这些法则适用于雇主(包括私人 and 政府)、劳工组织、职业中介机构、执照和资格管理协会。当测验或其他选拔程序的使用导致了对少数人群体求职者拒绝率高于非少数人群体求职者时(“不利影响”),<sup>[12]</sup>其效用性必须由所求工作的效度证据来证实。

从历史上来看,可接受的测验效度的要求一直由“测验标准”人事选拔程序的效用和使用原则”(SIOP, 1987)和大家知道的专业领域的其他该类文件来确定。然而在过去的 20 年里,有

几个案件显示外在的法律考虑已经融入心理测验实践中，尤其是那些和公民权利相连的案件；其中之一是“戈尔登·鲁尔”（Golden Rule，也见第七章）协议。这一协议解决了戈尔登·鲁尔保险公司和教育测验服务处（ETS）之间关于由 ETS 准备的保险执照测验的争端。这一争端的解决要求当多数人群体和少数人群体的正确答案出现不同的比例时，要优先使用那些具有最小组间差异的测验。虽然其目的是在于提高公平性和使不利影响最大程度的减少，但戈尔登·鲁尔争端的解决引起了关于项目偏向的性质的假设和关于经验证据在何种程度上能论证解决办法所建议的程序等问题的激烈争论（APA, CPTA, 1988; Bond, 1987; Linn & Drasgow, 1987; Rooney, 1987）。

在肯定行动（affirmative action）的讨论中，1978 年的“统一指导纲要”提出：即使当选拔程序被证明有满意的效率，但是如果结果对少数人群体有相当大的拒绝率时，也必须采取步骤尽可能减少这种差距。肯定行动隐含着一个组织要做的不仅仅是避免歧视。从心理学上讲，肯定行动计划（在过去的几年里，在政治舞台上它不断受到攻击）可以认为是努力弥补过去社会不公平的后效。GATB 在 80 年代实施的亚群体常模化工作，其目的是为白人、黑人、说西班牙语人的求职者提供具有类似的工作分派比率，尽管他们的能力倾向分数有较大的差距（见第十七章，也见 Hartigen & Wigdor, 1989）。这是肯定行动的具体例证，旨在减少职前测验的不利影响。然而，这一实践产生的争论导致了 1991 年公民权利法案的通过，它明确禁止基于种族、肤色、宗教、性别或最初国籍的任何形式的分数调整。在心理测验范围内，大家认识到这一法案的结果“比国会所预想的更为深远”（D.C. Brown, 1994, p. 927）。它严格限制使用分别具有男女两性常模的人格和身体能力测验（也见 L. S. Gottfredson, 1994; Ke-

hoe & TenoPyr, 1994; Sackett & Wilk, 1994)。事实上,一些测验的作者和出版者已经采取步骤提供替代性的记分程序,废除分开的性别常模(参见 Gough & Bradley, 1996)。

另一个意愿良好的,为所有人清除障碍,以敞开就业机会的努力是 1990 年的“美国残疾人法案”(ADA, P. L. 101 ~ 336)。它引起雇主及其他对恰当的职前测验感兴趣的人的关注。ADA 的雇佣条款规定,阻止雇主在提供某种工作前使用医学测验或询问物质滥用历史或过去的精神病状况。EEOC 关于职前与能力缺陷有关的询问和医学检测的准则及规则,迄今为止还是一个悬而未决的问题:什么样的心理和人格测验在职前情境中是允许使用的?

专业、法律和伦理之间规范的不一致性在将来将继续出现(参见 D. C. Brown, 1996)。毫无疑问,它们使得测验在所谓高风险的就业和教育领域中的应用更为困难。在很大程度上,这种情形表示了一种进步,因为它强调需要一种清晰的认识,不管是在科学研究还是在实践领域,价值都包含在任何重要的决策中。用梅西克的话来说,就是“价值将融入测验的意义和结果中去,并一直如此……这使得一直潜伏的东西变得清楚、明了,即效度的判断就是价值判断”(1995, p. 748)。但是,甚至一些善意的、讲道理的人,可能会而且一定会对价值有明显的不同意见;这就是困难所在。

与测验有关的因素 在对各种不同的个体施测时,把对测验和效标行为都有影响的因素和只对测验发生影响的因素区分开来,这是十分重要的。后者是“与测验有关的因素”,能降低测验的效度。这些因素包括以前参加测验的经验,测验时表现良好的动机,与施测者的关系,过分地强调速度,其他影响特定测验的成绩与所探讨的广泛行为领域无关的变量。当测试的个体

具有不同的文化背景或具有某种能力缺陷时，应进行特殊的努力以减少那些‘与测验有关因素’的作用(见 Sattler, 1988 第十九、二十章)。为这一目的而设计的适当地参加测验定向程序、准备性的练习和其他步骤都是很理想的(第一章和第九章)。

具体测验内容也可能影响测验分数，这种影响和测验所要评估的能力无关。例如，在算术推理测验中，在某种特定文化环境中不熟悉的客体名称或图形，会表现为一种限制测验的障碍。某些特殊测验内容可能会通过被试的情绪和态度反应，以一种更为微妙的方式欺骗性地影响成绩。例如，描述典型郊区中产阶层家庭情境的故事或图片，会使生活在低收入城区家庭的儿童感到非常陌生。测验内容的性别刻板印象的存在，如男医生或男飞行员和女护士或女乘务员的画像，也可能产生不利效果。根据这些考虑，大多数测验出版者现在作出了具体的努力来废除不恰当的测验内容。事实上，关于测验内容可能对少数人群体的被试具有负面含义的检查是测验建构过程中一个共同的步骤(见如 ETS 标准, 1981/1987)。

对不同文化背景和经验历史的人以及具有残疾的个体的测验，是整个“测验标准”普遍关注的问题。这种一般性的定向反映在测验发展和测验使用的几个标准中。此外，一些具体章节论述了具有不利条件和语言差异人群的测验问题，这个群体代表着相当一部分美国人。

**测验分数的解释和使用** 在不同组群及所有的测验中最需要考虑的问题是测验分数的解释。有关少数人群成员测验中最经常性的忧虑来自对分数的错误解释。如果一个少数民族被试在能力倾向测验的得分较低或在人格测验中获得异常的分数，就必须调查他或她所以如此的原因。例如，某一算术测验中的不良分数，可能是被试的低动机、阅读能力差、算术知识的不足



等等其他原因导致的。此外，还应该考虑评估个体测验分数中所采用常模的类型。<sup>[13]</sup>

测验是用来显示个体在某一确定的时间内他可以做什么，而不能告诉我们个体为何有这样的表现。为了回答这个问题，我们需要调查个体的背景、动机和其他相关环境。测验也不能告诉我们：一个以前一直是在良好环境中成长、现在却处在文化和教育条件均差的情形下的儿童可能会具有什么样的能力。而且，测验也不能通过排除文化剥夺对分数的影响来弥补文化的剥夺；相反测验应该揭示这种文化剥夺的影响，以便采取适当的补救措施。通过拒绝测验或设计一些对这种效果不具有敏感性的测验来掩盖文化条件不良的效果只能阻碍社会问题的真正解决。

把分类和标记作为取代理解的一种捷径的倾向仍然非常盛行。经典精神病学的诊断分类是这种倾向著名的例子，比如给病人加上诸如‘妄想性精神分裂症’或‘躁狂抑郁病患者’这样的标记。因为知道了这种分类系统的许多缺陷，最近精神疾病诊断手册的作者描述了各种类型的障碍，把称呼加在条件上而不是遭受其害的个体身上（见美国精神卫生学会，1994）。而且，心理学家自身也逐渐转向人格描述；不像诊断性的标记，这些描述集中在异常行为的起源和它的个体意义，并为治疗提供更有效的基础。然而传统的标记不是那么容易去掉的。

另一种分类倾向的具体例证是 IQ 的错误解释。根据流行的错误概念，IQ 是先天智力潜能的标志，表现有机体固定的特质。正如我们在第十二章所看到的一样，这种观点在理论上是站不住脚的，也缺乏经验数据的支持。当进行恰当的解释时，智力测验分数不应助长对个体进行固定的分类；相反，智力测验和其他测验都可以看成一张地图，在其上可以确定个体现在的位置。当与其他有关经验背景的信息结合起来时，测验分数应

有助于个体最佳发展的有效计划。

**测验的客观性** 当社会刻板印象和偏见可能扭曲人际间的评价时,测验提供了一种对偏好、专制和任意决策的防御。当公民权运动势头正盛时,几个观察家唤起了对标准化测验所起的积极作用的关注。加德纳(J. W. Gardner, 1961 pp.48 ~ 49)在评论学校中使用测验时写道:“测验既不着眼于年轻人是衣着破烂还是华丽漂亮,也听不到贫民窟中发出的呻吟,但测验能从任何水平的人群中发现智力天才。”

即使测验被取消,但个体或组织决策的需要仍然存在。决策将不得不依靠于我们熟知已久的其他替代方法如推荐信、面试和成绩平均数。现在,这些数据资料经常和测验分数结合使用,但没有代替测验。事实上,标准化测验是作为这些无信度的主观性的和可能有偏见的传统程序的补偿手段而引入的。一般已经证明这些测验的替代方法在预见学业成绩和工作绩效方面比测验的准确性要差(Wigdor & Garner, 1982, Pt. I, 第一章)。最近发展的替代方法,如绩效和业务评估,最终可能证明与传统的测验相比具有某些优点。然而到目前为止,对这些技术的研究表明:与它们所要补充或替代的标准化测验相比,它们并不更为有效,对不同的人群也并不更具公正性(见第十七章)。

对测验的攻击往往不能将其在决策公平性中所起的积极作用与其作为慎重判断的便捷替代物的滥用区分开来。从测验的社会性背景来看,能力测验委员会(Wigdor & Garner, 1982, Pt. I)强调:测验既不能视为解决社会问题的灵丹妙药,也不能视其为替罪羊。为特定少数人群成员增加机会的社会目标不应和测验过程有效性混淆在一起。能力测验委员会在其结论性的陈述中评述道:“对更公正社会的追求已将能力测验置于争论的中心,这既夸大了其好的声望,又夸大了其坏的名声(p.239)。”这一陈

述现在仍然是正确的，且由于缺少有生命力的替代物，在将来很长的一段时间里极有可能仍是正确的。

总之，就像误用在其他任何人身上一样，测验的确也能误用于少数人群体。但是，当使用恰当时，它们在防止不恰当的和不公正的歧视方面起着重要作用。当评价测验的社会性结果时，我们必须慎重评价没有使用测验，因而不得不依靠其他公平性差于测验的决策程序而得到的社会结果。此外，在确定测验的结果中，我们必须仔细区分测验恰当使用的结果和测验误用的结果，并把测验的直接结果和受测验以外的因素影响的间接结果分开(Tenopyr, 1995)。否则我们可能会完全因错误的理由抛弃一种尽管需要不断改进、但可能证明是不可替代的工具。

#### 注 释

- [1] 为了尽可能广泛地普及“心理学家的伦理原则和行为规范”——以后简称为“伦理规范”，APA 对于任何索取者免费赠送一份单行本。
- [2] 关于测验使用者的作用和由 JCTP 工作组实施的测验使用者资格和使用培训项目(Eyde et al., 1988, 1993; Moreland et al., 1995)的讨论见第一章。加拿大心理学会和英国心理学会也采取一些步骤以发展确定测验使用者资格的系统(D. C. Brown, 1995)。
- [3] 美国心理学会心理测验和评定专业委员会(1995)已经准备了一项声明，为心理学研究生和本科生的教育中安全使用心理测验提供指导。
- [4] 作为对有关把 MMPI 带回家进行施测适宜性调查的答复，APA 伦理委员会最近准备了一项关于带回家“测验”的政策性声明(1994, pp. 665 ~ 666)。
- [5] 《人类参与的研究行为的伦理准则》(APA, 1982)一书在这一方面提供了一些指导。
- [6] 见阿伦·汉森(F. Allan Hanson, 1993)的“对测验的评论及其在现代社会中的作用”，尽管可以清楚地看到这篇论文的思想观点和测验是相对

立,并远非公正的,但从人类学的角度看,读者可能会感兴趣。

- [7] 未成年人心理评价中的评估和其他伦理及法律问题的有关同意问题的指导,见坎普豪斯和弗里克(Kamphaus & Frick, 1996,第四章)的著作。
- [8] 1992年9月的《心理评估》所包括的一个特殊部分是关于为顾客提供心理测验反馈这一问题的。在波普(K. S. Pope)的一篇特别有价值的论文中,讨论了他认为在评估中最被忽视的关于反馈的十个基本问题(1992, p. 265)。
- [9] 尽管在美国人口中妇女占有统计上的大多数,但在法律上、职业上和其他方面与少数民族有许多共同的问题,所以当“少数人”这一术语在这一方面时,它包括妇女在内是可以理解的。
- [10] 关于心理教育评估领域中一些重要的法庭决议的讨论可以在艾耶斯、戴和罗塔托里(Ayers, Day & Rotatori, 1990)的著作以及雷斯奇里(Reschly, 1988)的著作中找到。
- [11] “统一指导纲要”已经变得过时,显然需要进行评述和修订。“指导纲要”的修订版预期在90年代后期出版的测验标准之后出版(见第一章)。
- [12] B. 勒纳(B. Lerner)分析了不同法庭案例中不利于影响评估方式的不一致性(1980a; Ironson, Guion & Qstrander, 1982)。
- [13] 1994年12月出版的《心理评估》有一个部分专门提供了常模评定诸方面的信息和指导。

## 附 录 测验和其他评价工具

本表列举了本书中提及的所有测验，但是下列三种情况除外：

- (1) 由于历史原因而引证的尚未出版的测验；(2) 尚未出版的测验；  
(3) 文献中描述的测验和作者提供的测验。若想了解进一步的信息，请参阅《心理测量年鉴》( *Mental Measurements Yearbook* )、布罗斯心理测量研究所( *Buros Institute of Mental Measurements* )提供的其他文献《测验评论》( *Test Critiques* , Keyser & Sweetland, 1984—1994 年)。此外，还有第一章所援引的其他资料。

### 测验名称(缩写)/出版者的编码

|                                                                   |                           |
|-------------------------------------------------------------------|---------------------------|
| AAMR Adaptive Behavior Scale (ABS)/PRO - ED                       | AAMR 适应行为量表(ABS)/PRO - ED |
| ACT Assessmeng/ACT                                                | ACT 评定/ACT                |
| Adjective Check List (ACT)/CPP                                    | 形容词核查表( ACT)/CPP          |
| Alcohol use Inventory (AUI)/NCS                                   | 酒精使用调查问卷( AUI)/NCS        |
| Armed Forces Qualification Test (AFQT)/U.S. Military              | 陆军资格测验( AFQT)/美国军方        |
| Armed Services Vocational Aptitude Battery ( ASVAB)/U.S. Military | 军队服务职业能力侧向成套测验/美国军方       |
| Basic Personality Inventory (BPI)/Sigma                           | 基本人格调查表( BPI)/Sigma       |

- Bayley Infant Neurodevelopmental Screener(BINS)/TPC **贝利婴儿神经发展筛选量表( BINS)/TPC**
- Bayley Scales of Infant Development-Second Edition (Bayley - II)/TPC **贝利婴儿发展量表第 2 版(贝莱 - II)/TPC**
- Beck Depression Inventory (BDI)/TPC **贝克抑郁问卷(BDI)/TPC**
- Behavior Assessment System for Children (BASC)/AGS **儿童行为评定系统(BASC)/AGS**
- Bender Visual Motor Gestalt Test (Bender-Gestalt)/WPS **本德视觉运动格式塔测验 (本德一格式塔)/WPS**
- Bennett Mechanical Comprehension Test (BMCT)/TPC **本纳特机械理解测验 (BMCT)/TPC**
- Benton Visual Retention Test, Fifth Edition (BVRT)/TP **本顿视觉保持测验第 5 版( BVRT)/TPC**
- Boehm Test of Basic Concepts-Revised (Boehm - R)/TPC **波恩基本概念测验修订版(波姆 - R)/TPC**
- Bracken Basic Concept Scale (BBCS)/TPC **布莱克恩基本概念量表**
- Brief Symptom Inventory (BSI)/NCS **简明症状问卷(BSI)/NCS**
- British Ability Scales (BAS)/NFER - Nelson **英国能力量表( BAS)/NFER - Nelson**
- Bruininks-Oseretsky Test of Motor Proficiency/AGS **布鲁宁克斯一奥西里基动作水平测验/AGS**
- California Achievement Tests-Fifth Edition (CAT)/CTB **加利福尼亚成就测验第 5 版(CAT)/CTB**
- California Child Q - Set/CPP **加利福尼亚儿童 Q - 定向/CPP**
- California Diagnostic Mathematics Tests (CDMT) CTB **加利福尼亚数学诊断测验( CDMT) CTB**
- California Diagnostic Reading Tests (CDRT)/CTB **加利福尼亚阅读诊断测验( CDRT)/CTB**

- California Psychological Inventory-Third Edition (CPI-3)/CPP **加利福尼亚心理调查卷第3版(CPI-3)/CPP**
- California Q-Sort Deck/ CPP **加利福尼亚 Q 分类卡片/ CPP**
- Campbell Interest and Skill Survey (CISS)/NCS **坎贝尔兴趣和技能调查(CISS)/NCS**
- Career Assessment Inventory-The Enhanced Version (CAI-EV)/NCS **职业评定问卷改进版(CAI-EV)/NCS**
- Career Assessment Inventory-The Vocational Version (CAI-VV)/NCS **职业评定问卷—职业版(CAI-VV)/NCS**
- Career Beliefs Inventory(CBI).CPP **职业信念问卷(CBI).CPP**
- Career Development Inventory (CDI)/CPP **职业发展问卷(CDI)/CPP**
- Career Directions Inventory (CDI)/Sigma **职业方向问卷(CDI)/西格玛**
- Career Interest Inventory (CII)/TPC **职业兴趣问卷(CII)/TPC**
- Career Planning Program (CPP)/ACT **职业计划纲要(CPP)/ACT**
- Children's Apperception Test (C.A.T)/CPS **儿童统觉测验(C.A.T)/CPS**
- Cognitive Abilities Test (CogAT, Form 5)/Riverside **认知能力测验(CogAT,表5)/里佛萨德**
- College Level Examination Program (CLEP)/ETS **大学水平考试大纲(CLEP)/ETS**
- Columbia Mental Maturity Scale (CMMS)/TPC **哥伦比亚心理成熟量表(CMMS)/TPC**
- Comprehensive Tests of Basic Skills-Fourth Edition(CTBS/4)/CTB **基本技能综合测验第4版(CTBS/4)/CTB**
- Computer Programmer Aptitude Battery (CPAB)/SRA **计算机程序员能力倾向测验(CPAB)/SRA**
- Crawford Small Parts Dexterity Test (CSPDT)/TPC **克洛福德身体部分**

**灵巧性测验(CSPDT)/TPC**

Cross-Cultural Adaptability Inventory (CCAI)/NSC **跨文化适应性调查表(CCAI)/NSC**

CRT Skills Test/SRA **CRT技能测验/SRA**

Culture Fair Intelligence Test/IPAT **文化公平智力测验/IPAT**

Das-Naglieri Cognitive Assessment System (CAS)/Riverside **达斯—纳格利里认知评定系统(CAS)/Riverside**

Differential Ability Scales (DAS)/TPC **区别能力量表(DAS)/TPC**

Differential Aptitude Tests-Computerized Adaptive Edition (DAT Adaptive)/TPC **区别能力测验—计算机化适应版(DAT适应)/TPC**

Differential Aptitude Tests-Fifth Edition (DAT)/TPC **区别能力测验第5版(DAT)/TPC**

Draw-a-Man Test (see Goodenough-Harris Drawing Test) **“画男人”测验(见古登诺夫—哈里斯画图测验)**

Edwards Personal Preference Schedule (EPPS)/TPC **埃德华个人偏爱表(EPPS)/TPC**

Embedded Figures Test (EFT)/CPP **包含图形测验(EFT)/CPP**

Fagan Test of Infant Intelligence/Infantest **费根婴儿智力测验/婴儿测验**

Fleishman Job Analysis Survey (F-JAS)/MRI **弗莱希曼工作分析调查(F-JAS)/MRI**

General Aptitude Test Battery (GATB)USES **一般能力倾向成套测验(GATB)USES**

Goodenough-Harris Drawing Test/TPC **古迪纳夫—哈里斯画人测验/TPC**

Graduate Record Examinations (GRE)/ETS **研究生资格考试(GRE)/ETS**

Group Embedded Figures Test/ CPP **团体包含图形测验/ CPP**



- Guide to the Assessment of Test Session Behavior for the WISC – III and the  
WIAT/TPC WISC – III 和 WIAT 测验期间行为评价指南/TPC  
Guilford-Zimmerman Temperament Survey (GZTS)/CPP 吉尔福德—齐  
默尔曼气质调查表(GZTS)/CPP
- Halstead-Reitan Neuropsychological Test Battery (HRB)/RNL 霍尔斯特  
德—瑞坦神经心理成套测验(HRB)/RNL
- Haptic Intelligence Scale/Stoelting 触觉智力量表/斯多尔汀
- Harrington-O'Shea Career Decision-Making System-Revised (CDM – R)/  
AGS 哈林顿—奥西职业决策系统修订版(CDM – R)/AGS
- Hogan Personality Inventory-Second Edition (HPI)/HAS 霍根人格问卷  
第 2 版(HPI)/HAS
- House-Tree-Person (H – T – P)/WPS 房屋—树木—人(H – T – P)/  
WPS
- Infant-Toddler Developmental Assessment(IDA)/Riverside 婴幼儿发展  
评价(IDA)/Riverside
- Iowa Tests of Basic Skills/Riverside 依阿华基本技能测验/Riverside
- Iowa Tests of Educational Development/Riverside 依阿华教育发展水平  
测验/Riverside
- Jackson Personality Inventory-Revised(JPI – R)/Sigma 杰克森人格调  
查表修订版(JPI – R)/Sigma
- Jackson Vocational Interest Survey(JVIS)/Sigma 杰克森职业兴趣调查  
表(JVIS)/Sigma
- Kaufman Adolescent and Adult Intelligence Test (KAIT)/AGS 考夫曼青  
少年和成人智力测验(KAIT)/AGS
- Kaufman Assessment Battery for Children (K – ABC)/AGS 考夫曼儿童  
评定成套测验(K – ABC)/AGS
- Kaufman Brief Intelligence Test(K – BIT)/AGE 考夫曼简短智力测验  
(K – BIT)/AGE

- Kaufman Test of Educational Achievement (K - ABC)/AGE **考夫曼教育成就测验(K - ABC)/AGE**
- Kuder General Interest Survey (KGIS)/CTB **库德一般兴趣调查(KGIS)/CTB**
- Kuder Occupational Interest Survey (KOIS)/CTB **库德职业兴趣调查表(KOIS)/CTB**
- Kuder Preference Record-Vocational (KPR - V)/CTB **库德爱好档案—职业版(KPR - V)/CTB**
- Leiter International Performance Scale-Revised(LIPS - R)/Stoelting **利特国际操作量表 修订版(LIPS - R)/Stoelting**
- Luria-Nebraska Neuropsychological Battery (LNNB)/WPS **鲁利亚—耐布拉斯卡神经心理成套测验(LNNB)/WPS**
- Machover Draw-a-Person Test(D - A - P)/Thomas **马乔弗尔“画人”测验(D - A - P)/Thomas**
- McCarthy Scales of Children's Abilities (MSCA)/TPC **麦卡锡儿童能力量表(MSCA)/TPC**
- Metropolitan Achievement Test-Seventh Edition (MAT)/TPC **城市成就测验第 7 版(MAT)/TPC**
- Metropolitan Readiness Tests-Sixth Edition (MRT)/TPC **城市准备状态测验第 6 版(MRT)/TPC**
- MicroCog: Assessment of Cognitive Functioning/TPC **米克洛库格认知功能评定/TPC**
- Millon Adolescent Clinical Inventory (MACI)/NCS **米隆青少年临床调查表(MACI)/NCS**
- Millon Adolescent Personality Inventory (MAPI)/NCS **米隆青少年人格调查表(MAPI)/NCS**
- Millon Clinical Multiaxial Inventory-III (MCMI - III)/NCS **米隆临床多轴调查表(MCMI - III)/NCS**

- Millon Index of Personality Styles (MIPS)/TPC **米隆人格类型指标 (MIPS)/TPC**
- Minnesota Clerical Test (MCT)/TPC **明尼苏达神职人员测验 (MCT)/TPC**
- Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A)/UMP **明尼苏达青少年多相人格调查表 (MMPI-A)/UMP**
- Minnesota Multiphasic Personality Inventory-2 (MMPI-2)/UMP **明尼苏达多相人格调查之二 (MMPI-2)/UMP**
- Multidimensional Aptitude Battery (MAB)/Sigma **多维能力倾向成套测验 (MAB)/Sigma**
- Myers-Briggs Type Indicator (MBTI)/CPP **梅耶斯—布里格斯类型指标 (MBTI)/CPP**
- Occupational Aptitude Survey and Interest Schedule-Second Ed. (OASIS-2)/PRO-ED **职业能力倾向调查表和兴趣表第2版 (OASIS-2)/PRO-ED**
- Ordinal Scales of Psychological Development/UIP **心理发展顺序量表/UIP**
- Orleans-Hanna Algebra Prognosis Test/TPC **欧利恩—汉纳代数诊断测验/TPC**
- Otis-Lennon School Ability Test-Seventh Edition (OLSAT7)/TPC **奥蒂斯—莱农学校能力测验第7版 (OLSAT7)/TPC**
- Paced Auditory Serial Addition Test (PASAT)/For Thought **定步系列增加听觉测验 (PASAT)/For Thought**
- PDI Customer Service Inventory/PDI **PDI顾客服务问卷/PDI**
- PDI Employment Inventory/PDI **就业问卷/PDI**
- Peabody Picture Vocabulary Test-Revised (PPVT-R)/AGS **皮博迪图片词汇测验修订版 (PPVT-R)/AGS**
- Personality Assessment Inventory (PAI)/PAR **人格评定调查表 (PAI)/PAR**

PAR

Personality Inventory for Children-Revised (PICR)/WPS **儿童人格调查表修订版(PICR)/WPS**

Personality Inventory for Youth (PIY)/WPS **青年人格问卷(PIY)/WPS**

Personality Research Form (PRF)/Sigma **人格研究表( PRF)/Sigma**

Porteus Mazes/TPC **波蒂厄斯迷津/TPC**

Quality of Life Inventory (QOLI)/NCS **生活质量问卷( QOLI)/NCS**

Raven's Progressive Matrices (PRM)/Oxford (U.S. Distributor: TPC) **瑞文渐进推理测验( PRM)/Oxford (U.S. Distributor: TPC)**

Revised Minnesota Paper Form Board Test (RMPFBT)/TPC **明尼苏达纸笔测验修订版( RMPFBT)/TPC**

Revised NEO Personality Inventory (NEO PI - R)/PAR **NEO 人格调查表修订版( NEO PI - R)/PAR**

Roberts Apperception Test for Children (RATC)/WPS **罗伯特儿童统觉测验 (RATC)/WPS**

Rotter Incomplete Sentences Blank (ISB)/TPC **罗特不完全句子填充 (ISB)/TPC**

Schaie-Thurstone Adult Mental Abilities Test/CPP **谢伊-瑟斯顿成人心理能力测验/CPP**

Schedule for Nonadaptive and Adaptive Personality (SNAP)/UMP **不适应性和适应性人格表/UMP**

Scholastic Assessment Test (SAT)/ETS **学业评定测验 (SAT)/ETS**

Self-Directed Search (SDS)/PAR **自我定向研究( SDS)/PAR**

Senior Apperception Test (S.A.T)/CPS **高级统觉测验(S.A.T)/CPS**

Sixteen Personality Factor Questionnaire-Fifth Edition (16PF)/IPAT **16 人格因素问卷第 5 版( 16PF)/IPAT**

Social Climate Scales/CPP **社会气氛量表/ CPP**

Social Skills Rating System (SSRS)/AGS **社会技能评定系统( SSRS)/**

AGS

Stanford Achievement Test-Eighth Edition/TPC **斯坦福成就测验第 8 版/TPC**

Stanford-Binet Intelligence Scale-Fourth Edition (SB - IV)/Riverside **斯坦福一比纳智力量表第 4 版(SB - IV)/Riverside**

Stanford Diagnostic Mathematics Test-Third Edition (SDMT)/TPC **斯坦福数学诊断测验第 3 版(SDMT)/TPC**

Stanford Diagnostic Reading Test-Third Edition (SDRT)/TPC **斯坦福阅读诊断测验第 3 版(SDRT)/TPC**

Stanford Writing Assessment Program/TPC **斯坦福写作评定项目/TPC**  
State-Trait Anger Expression Inventory (STAXI)/PAR **状态-特质愤怒表现调查表(STAXI)/PAR**

State-Trait Anxiety Inventory (STAI)/CPP **状态-特质焦虑调查表(STAI)/CPP**

State-Trait Anxiety Inventory for Children(STAIC)/CPP **状态-特质儿童焦虑调查表(STAIC)/CPP**

Strong Interest Inventory (SII)/CPP **斯特朗兴趣调查表(SII)/CPP**

Structure of Intellect Learning Abilities Test/WPS **智力结构-学习能力测验/WPS**

Student Adaptation to College Questionnaire (SACQ)/WPS **大学生适应性问卷(SACQ)/WPS**

Student Self-Concept Scale(SSCS)/AGE **学生自我概念量表**

Student Styles Questionnaire/TPC **学生风格问卷/TPC**

Symptom Checklist 90-Revised (SCL - 90 - R)/NCS **症状检查 90 项一修订版(SCL - 90 - R)/NCS**

System for Interactive Guidance Information-Revised (SIGI - PLUS)/ETS **交互式指导信息系统修订版(SIGI - PLUS)/ETS**

Teamwork-KSA Test/SRA **团队工作 - KSA 测验/SRA**

- Tell -Me-A-Story (TEMAS)/WPS “给我讲个故事”(TEMAS)/WPS  
TerraNova series/CTB 特纳诺瓦系列/CTB
- Test Anxiety Inventory (TAI)/CPP 焦虑测验问卷(TAI)/CPP
- Test of Cognitive Skills-Second Edition (TCS/2)/CTB 认知技能测验第  
2 版(TCS/2)/CTB
- Test of Nonverbal Intelligence - Second Edition(TONI - 2)/PRO - ED  
非言语智力测验第 2 版(TONI - 2)/PRO - ED
- Test of Variables of Attention (T.O.V.A)/UAD 注意变化测验(T.O.  
V.A)/UAD
- Tests of Achievement and Proficiency/Riverside 学习成就和效率测验/  
Riverside
- Tests of Adult Basic Education (TABE)/CTB 成人基础教育测验  
(TABE)/CTB
- The Holtzman Inkblot Technique (HIT)/TPC 霍兹曼墨迹技术(HIT)/  
TPC
- Thematic Apperception Test (TAT)/Harvard 主题统觉测验 (TAT)/  
Harvard
- The Rosenzweig Picture-Frustration Study (P - F Study)/PAR 罗森茨威  
格图片一挫折研究(P - F 研究)/PAR
- The Scenotest/H&H 戏剧测验/H 和 H
- The Values Scale/CPP 价值量表/CPP
- Vigil Continuous Performance Test(VIGIL)/For Thought 夜班连续操作  
测验(VIGIL)/For Thought
- Vineland Adaptive Behavior Scales (VABS)/AGS 瓦因兰适应性行为  
量表(VABS)/AGS
- Washington University Sentence Completion Test (WUSCT)/Erlbaum 华  
盛顿大学不完全测验 (WUSCT)/Erlbaum
- Wechsler Adult Intelligence Scale-Revised (WAIS - R)/TPC 韦克斯勒

**成人智力量表修订版( WAIS - R)/TPC**

Wechsler Individual Achievement Test(WITA)/TPC **韦克斯勒个人成就测验( WITA)/TPC**

Wechsler Intelligence Scale for Children-Third Edition ( WISC - III)/TPC  
**韦克斯勒儿童智力量表第 3 版( WISC - III)/TPC**

Wechsler Preschool and Primary Scale of Intelligence-Revised ( WPPSI - R)/TPC **韦克斯勒学龄前及小学生智力量表修订版 ( WPPSI - R)/TPC**

Wonderlic Personnel Test/Wonderlic **温德利克人格测验/ Wonderlic**

Woodcock-Johnson Psycho-Educational Battery-Revised ( WJ - R)/Riverside **伍德科克—约翰逊心理教育成套测验修订版 ( WJ - R)/Riverside**

Woodcock Reading Mastery Tests-Revised/AGS **伍德科克阅读掌握测验修订版/ AGS**

Work Keys/ACT **工作要素/ ACT**

## 主题索引

本索引页码均指英文版原著页码，请按中文版边码检索。

- AAMR Adaptive Behavior Scale(ABS) AAMR 适应行为量表(ABS)  
251 – 252, 551
- Ability, use of term 能力, 术语的运用 475
- Ability levels, reliability and 能力水平, 信度 107
- Ability Testing: Uses, Consequences, and Controversies 能力测验：运用，  
结果及争论 535 – 536
- Absolute scaling 绝对量表的制订 174 – 177
- Academic achievement, indices of 学业成就, 指标 120
- Academic intelligence, role of 学术智力, 的作用 497 – 498 (See also  
Intelligence; Intelligence Tests) (参见智力, 智力测验)
- Achievement need (n – Ach) 成就需要 (n – 阿赫) 421
- Achievement tests 成就测验  
description of 的描述 475 – 478  
development of 发展 42 – 44  
difference between aptitude tests and 能力测验之间的区别 475  
general batteries 一般成套测验 481 – 483
- Acquiescence response set 默认反应定势 377 – 378
- ACT Assessment ACT 评定 486, 551



- Adaptive testing **适应性测验**
- Computerized **计算机化** 192, 276 – 278
- individually tailored tests **适应个体的测验** 274 – 275
- Adaptive treatments, decision theory and **适应治疗, 决定论** 154
- Adjective Check List (ACL) **形容词核查表** 457 – 458, 469, 551
- Adult intelligence, problems in testing **成人智力, 测验中的问题** 331 – 337
- Advanced Placement Program (APP) **高级编班项目** 487
- Affirmative action **肯定性行动** 546 – 547
- Age decrement, in intelligence test scores **随年龄而发生的衰减下降, 智力测验分数中的** 331 – 333
- Age differentiation **年龄差异** 127
- Age norms **年龄常模** 55 – 56
- Alcohol Use Inventory **酒精使用调查问卷** 524, 551
- Alternate form reliability **复本信度** 93 – 95
- American Association on Mental Retardation (AAMR) **美国智力落后学会** 247, 251 – 252, 555
- American Board of Professional Psychology (ABPP) **美国职业心理学委员会** 537 – 538, 555
- American College Testing (ACT) Program **美国大学测验 (ACT) 方案** 43, 486, 487, 494, 526, 555
- American Council on Education **美国教育理事会** 43
- American Counseling Association (ACA) **美国咨询协会 (ACA)** 535, 555
- American Educational Research Association (AERA) **美国教育研究协会** 29, 555
- American Guidance Service **美国指导服务中心** 74
- American Psychiatric Association **美国精神病学会** 367

- American Psychological Association (APA) **美国心理学会 (APA)** 29,  
38, 532, 533 – 534, 556
- Committee on Psychological Tests and Assessment (CPTA) **心理测验  
和评定委员会** 534
- cultural diversity and **文化多样性** 341 – 342, 346
- ethics and **伦理学** 533 – 550
- Golden Rule case and **戈尔登·鲁尔案** 200
- Joint Committee on Testing Practices (JCTP) **测验实施联合委员会**  
534 – 535, 539
- Science Directorate of **科学理事会** 28
- Task Force on Intelligence **智力工作小组** 295
- Technical Recommendations for Psychological Tests and Diagnostic Tech-  
niques **心理测验和诊断技术的方法推荐** 126
- (See also Standards for Educational and Psychological Testing) **(参见教育  
和心理测验标准)**
- Americans with Disabilities Act (ADA) (1990) (P.L. 101 – 336) **美国残  
疾人法案** 253, 359, 389, 541, 547
- Anchor Test Study **参照测验研究** 70
- Answer sheets, test scores affected by **答题纸, 受答题纸影响的测验分  
数** 14 – 15
- Aptitudes Research Project **能力倾向研究计划** 314
- Aptitude testing. See also Multiple aptitude batteries/tests **能力测验。  
也见多种能力测验**
- difference between achievement tests and **成就测验之间的差异**  
475
- special **特殊的** 39 – 42, 501 – 505
- for special programs **针对特殊程序的** 498 – 501
- Aptitude Testing (Hull) **能力倾向测验 (赫尔)** 493

- Armed Forces Qualification Test(AFQT) 陆军资格测验(AFQT) 163,  
271,499,551
- Armed Services Vocational Aptitude Battery(ASVAB) 军队服务职业能  
力倾向成套测验(ASVAB) 192,271,278,499-501,551
- Army Alpha test 军队甲种测验 39,271
- Army Beta test 军队乙种测验 39,234,271,343
- Arthur Performance Scale 阿瑟操作量表 254
- Assessment-center techniques 以评价为中心的技术 452-454,492  
-493
- Assessment Systems Corporation 评价系统公司 277,556
- Association of Test Publishers(ATP) 测验出版者协会 539-540,556
- Attitude measurement 能力测验  
nature of attitudes 态度的性质 404-405  
type of scales 量表的类型 405-409
- Autobiographical memories 自传体记忆 427-429
- “Average deviation” 平均离差 53
- Barnum effect 巴纳姆效应 529
- Basal age 基础年龄 55
- Basal level 基础水平 210
- Base rate 基础率 146-149
- Basic Personality Inventory(BPI) 基本人格调查表(BPI) 358,373-  
374,551
- Bayesian estimation 贝叶西评价 79
- Bayley Infant Neurodevelopmental Screener(BINS) 贝利婴儿神经发展  
筛选量表 239,551
- Bayley Scales of Infant Development 贝利婴儿发展量表 237-239,  
551

- Beck Depression Inventory (BDI) 贝克抑郁问卷 524, 551
- Behavioral assessments 行为评定
- assessment techniques 评定技术 523 - 524
  - observer reports 观察者报告 463 - 469
  - situational tests 情境测验 450 - 454
- Behavior Assessment System for Children (BASC) 儿童行为评定系统 (BASC) 467, 524, 551
- Bell Curve, The (Herrnstein and Murray) 钟形曲线 (赫斯坦因和莫莱) 295
- Bender-Gestalt Test (BGT) 本德一格式塔测验 517
- Bender Visual Motor Gestalt Test 本德视觉运动格式塔测验 517, 551
- Bennett Mechanical Comprehension Test 本纳特机械理解测验 503 - 504, 551
- Benton Visual Retention Test (BVRT) 本顿视觉保持测验 (BVRT) 517, 551
- Berkeley Growth Study 伯克利成长研究 237
- Binet scales 比纳量表 37 - 38
- Binet-Simon Scale 比纳一西蒙量表 37 - 38, 55, 205
- Biodata 个人传记资料 469 - 471
- Biserial correlation, item discrimination and 二列相关, 项目鉴别度 187
- Bivariate distribution 双变量分布 142 - 143
- Blind Learning Aptitude Test (BLAT) 盲人学习能力倾向测验 (BLAT) 256 - 257
- Board on Testing and Assessment (BoTA) 测验和评价委员会 536, 556
- Boehm Test of Basic Concepts 波恩基本概念测验 490, 551

- Bracken Basic Concept Scale **布莱克恩基本概念量表** 490,551
- Bain damage, diagnosis of **脑损伤,诊断** 514 – 518
- Brief Symptom Inventory (BSI) **简明症状问卷 (BSI)** 350,552
- British Ability Scales (BAS) **英国能力量表 (BAS)** 226,552
- Bruininks-Oseretsky Test of Motor Proficiency **布鲁宁克斯—奥西里斯基动作水平测验** 252,552
- Buros Institute of Mental Measurements **布鲁斯心理测量研究所** 27, 556
- California Achievement Tests **加利福尼亚成就测验** 192,482,552
- California Child Q-Set **加利福尼亚儿童 Q – 定向** 469,552
- California Diagnostic Reading and Mathematics Tests **加利福尼亚阅读和数学诊断测验** 488,552
- California Guidance Study **加利福尼亚指导性研究** 326
- California Psychological Inventory (CPI) **加利福尼亚心理调查表** 359 – 360,552
- California Q-Sort Deck **加利福尼亚 Q分类卡片** 469
- Campbell Interest and Skill Survey (CISS) **坎贝尔兴趣和技能调查 (CISS)** 396,552
- Career assessment **职业评定** 525 – 527
- Career Assessment Inventory-Enhanced Version (CAI – EV) **职业评定调查表—改进版(CAI – EV)** 400n,552
- Career Assessment Inventory-Vocational Version (CAI – VV) **职业评定调查表—职业版(CAI – VV)** 400,552
- Career Beliefs Inventory **职业信念问卷** 527,552
- Career counseling **职业咨询** 525 – 527
- Career Development Inventory **职业发展问卷** 527,552
- Career Directions Inventory **职业方向问卷** 402 – 403,552

- Career Interest Inventory (CII) **职业兴趣问卷 (CII)** 289n, 396, 525, 552
- Career maturity **职业成熟** 526 – 527
- Career Planning Program (CPP) **职业计划纲要** 526, 552
- Carl D. Perkins Vocational and Applied Technology Education Act (1990)  
**卡尔·D·帕金斯职业和应用技术教育法案** 389
- Carl D. Perkins Vocational Education Act (1984) **帕金斯职业教育法案 (1984)** 389n
- Carnegie Corp **卡耐基公司** 43
- Carnegie Foundation for the Advancement of Teaching **卡耐基教学发展基金会** 487
- Cattell's Culture Fair Intelligence Test **卡特尔文化公平智力测验** 311, 552
- Ceiling level **上限水平** 210
- Central tendency **集中趋势** 51 – 52
- Certification of psychologists **心理学家证书** 537 – 538
- Character Education Inquiry (CEI) **性格教育咨询 (CEI)** 451 – 452
- Childhood, early. See Also Infant and preschool testing **童年早期, 参见婴儿和学前测验**
- educational testing **教育测验** 489 – 492
- implications for intervention programs **干预计划的涵义** 330
- intelligence in **智力** 327 – 330
- longitudinal studies of intelligence **智力的追踪研究** 323 – 327
- Children's Apperception Test (CAT) **儿童统觉测验 (CAT)** 422, 552
- Chronicity **长期性** 515
- Classification, tests used for **分类, 用于分类的测验** 160 – 164
- Classroom tests **课堂测验** 484 – 485
- Clerical aptitude testing **文书能力测验** 504 – 505

- Clinical and counseling testing **临床和咨询中的测验** 509
- behavioral assessment **行为评定** 523 - 524
  - career assessment **职业评定** 525 - 527
  - intelligence tests and **智力测验** 512 - 514
  - judgment in **判断** 527 - 529
  - learning disabilities, identifying **学习失能,鉴别** 518 - 522
  - neuropsychological assessment **神经心理评定** 514 - 518
  - psychological assessment **心理评定** 510 - 511
- Clinical Versus Statistical Prediction (Meehl) **临床预示对统计预示 (米尔)** 527 - 528
- Coaching, test scores affected by **辅导,受辅导影响的测验分数** 23 - 24
- Coefficient alpha  $\alpha$  **系数** 97 - 99
- Cognitive Abilities Test (CogAT) **认知能力测验 (CogAT)** 280, 281, 284, 286, 552
- Cognitive psychology **认知心理学**
- contributions to validity **对效度的影响** 134 - 135
- Cognitive styles **认知风格** 444 - 448
- Cognitive task analysis, factor analysis and **认知任务分析, 因素分析** 320
- College Board **大学董事会** 17, 24, 487. See also Scholastic Aptitude Test; Scholastic Assessment Test (Taking the SAT) **参见学业能力测验 学业评价测验(采用 SAT)**
- College Entrance Examination Board (CEEB) **大学入学考试委员会 (CEEB)** 24, 43, 556
- College Level Examination Program (CLEP) **大学水平考试大纲 (CEEB)** 487, 552
- College level tests **大学水平测验** 485 - 487

- Columbia Mental Maturity Scale (CMMS) 哥伦比亚心理成熟量表 (CMMS) 259,552
- Combining tests, methods for 综合测验,方式 156 – 160
- Communicating test results 测验结果的交流 543 – 544
- Competency tests in basic skills 基本技能胜任测验 483 – 484
- Complete Guide for Occupational Exploration (CGOE) 职业探究全面指南 (CGOE) 526
- Comprehensive Early Memories Scoring System (CEMSS) 早期记忆综合记分系统(CEMSS) 428
- Comprehensive Tests of Basic Skills 基本技能综合测验 192,552
- Computer(s) 计算机
- guidelines for using 应用的指导 75 – 76
  - item analysis and 项目分析 200
  - multimedia and interactive technologies in testing 测验中的多媒体和相互作用技术 453n
  - role in psychological assessment 心理评定的作用 529 – 532
  - scoring 评分 74
  - used to interpret test scores 用于解释测验分数 74 – 76
- Computerized Adaptive Testing (CAT) 计算机化适应性测验 (CAT) 192,276 – 278
- adapted to Differential Aptitude Tests (DAT) 适应于区别能力倾向测验(DAT) 288
- Computer Programmer Aptitude Battery (CPAB) 计算机程序员能力倾向测验 505,552
- Computer-related aptitudes 与计算机有关的能力 505
- Concept Assessment Kit-Conservation (CAK) 守恒概念评定成套测验 (CAK) 242 – 243.552
- Concurrent validation 同时效度 118 – 119



- Confidentiality issues 保密问题 542 - 543
- Consequences of testing 测验结果 137 - 138
- Constructed-response tasks 建构反应的任务 479, 480, 481
- Construct-identification validation procedures 结构鉴别程序 126 - 135
- personality inventories 人格问卷 362 - 374
- Construct validity 结构效度
- cognitive psychology and 认知心理学 134 - 135
  - convergent and discriminant 幅合效度和区别效度 129 - 130
  - correlations with other tests 与其他测验的相关 127 - 128
  - developmental changes 发展性变化 127
  - experimental interventions 实验干预 130 - 132
  - factor analysis 因素分析 128 - 129
  - internal consistency 内部一致性 129
  - structural equation modeling 结构方程建模 132 - 134
  - use of term 术语的运用 114, 126
- Content-description, validity and 内容描述, 效度 114 - 118
- Content meaning, domain-referenced testing and 内容意义, 领域参照测验 77 - 78
- Content-related approach to self-report personality inventories 自陈人格调查表的内容关联方法 349 - 350
- Content sampling 内容取样 94
- Convergent and discriminant validity 幅合效度和区别效度 129 - 130
- Cornell Medical Index 康奈尔医学索引 350
- Correlation coefficient 相关系数 85 - 91
- meaning of correlation 相关的含义 85 - 87
  - statistical significance 统计显著性 88 - 92
- Counseling. See Clinical and counseling testing 咨询, 参见临床和咨询

## 测验

- Crawford Small Parts Dexterity Test **克洛福德身体部分灵巧性测验**  
502,552
- Criterion **效标**  
     in item analysis **项目分析中的** 179 – 181  
     in test validation **测验效度中的** 8 – 9, 118 – 120
- Criterion analysis **效标分析** 124
- Criterion contamination **效标污染** 119 – 120
- Criterion keying **效标关键作用**  
     biographical items **生物学项目** 470  
     interest inventory items **兴趣调查表项目** 389, 392 – 394  
     personality inventory item **人格调查表项目** 350 – 351
- Criterion-perdiction, validity and **效标预测, 效度** 118 – 126, 492
- Criterion-referenced testing **效标参照测验** 76
- Critical incident techniques **效标事件技术** 493
- Cross-Cultural Adaptability Inventory (CCAI) **跨文化适应性调查表**  
(CCAI) 532, 552
- Cross-cultural testing. See Multicultural testing **跨文化测验。见多种文化测验**
- Crossroads in the Mind of Man (Kelley ) **人类智力十字路口 (凯利)**  
311
- Cross-sequential design **交叉系列设计** 333
- Cross-validation **交叉效度** 194 – 196, 470
- CRT Skills Test **CRT 技能测验** 505, 552
- Cultural deprivation **文化剥夺** 343
- Cultural diversity. See also Multicultural testing **文化多样性。也见多元文化测验**  
     differences versus cultural handicap **文化差异与文化障碍** 342 –

- 343
- effects of, on testing situation **影响,测验情境** 345 – 346
- language in transcultural testing **跨文化测验中的语言** 343 – 345
- Cultural psychology, field of **文化心理学,场** 341 – 342
- Cultural stereotypes **文化刻板印象** 343
- Culture and Psychology **文化与心理学** 342
- Culture Fair Intelligence Test, Cattell's **文化公平智力测验, 卡特尔的**  
311, 552
- “Culture-free” tests **“跨文化”测验** 342
- Cutoff scores **临界分数**
- profile analysis and **剖面分析** 159 – 160
- reliability and **信度** 112
- use of **应用** 80 – 83
- Das-Naglieri Cognitive Assessment System (CAS) **达斯—纳格里里认  
知评定系统(CAS)** 233, 552
- Decalage, in Piagetian scales **滞差,皮亚杰量表中的** 244
- Decision models, fair use of tests and **决策模型,测验的公平应用**  
170 – 171
- Decision theory **决定论**
- adaptive treatments and **适应性处理** 154
- models for fair use of tests **公平应用测验的模型** 170 – 171
- moderator variables and **调节变量** 155 – 156
- sequential strategies and **序列决策** 153 – 154
- utility and **效用** 151 – 153
- validity and **效度** 144 – 156
- Derived scores **导出分数** 49
- Descriptive statistics **描述统计** 53n

- Developmental norms 发展常模 54 - 58
- Developmental transformations 发展转换 329
- Deviation IQ 离差智商 64 - 66
- Deviation response set 离差反应定势 378
- Diagnosis versus prediction 诊断对预测 5
- Diagnostic and Statical Manual of Mental Disorders (DSM) . DSM III, DSM  
- IV 心理障碍诊断与统计手册(DSM) 248n,367
- Diagnostic testing 诊断测验 488 - 489
- Differential Ability Scales (DAS) 区别能力量表 (DAS) 42,192,236,  
318,552
- description of 描述 227 - 228
- development of 发展 226 - 227
- reliability 信度 230 - 231
- scaling and norming 量表和常模 228 - 229
- validity 效度 231 - 232
- Differential Aptitude tests (DAT) 区别能力测验 288 - 290,552
- computerized version of 计算机化版本 288,522
- for Personnel and Career Assessment 人事和职业评定 288
- role of 作用 110,288 - 290,503,509,525
- Differential item functioning (DIF) 区分项目的功能 (DIF) 196 - 200
- Differential validity 区分效度 161
- Difficulty, measurement of 难度,的测量 173 - 179,187 - 192
- Discriminant validation 区别效度 129 - 130
- Distribution curves 分布曲线 50 - 51
- Domain-referenced testing 领域参照测验 76 - 80
- Draw -a-Man Test “画人”测验 19,264 - 266,552
- Draw -a-Person (DAP) “画人”(DAP) 266,429 - 430,438
- Drawing techniques 绘画技术 429 - 431

- Dynamic assessment **动力评定** 488 - 489, 521 - 522
- Early Memories Procedure (EMP) **早期记忆方法 (EMP)** 428 - 429
- Educational testing **教育测验**
- achievement tests **成就测验** 475 - 478
  - assessing young children **年幼儿童评定** 489 - 490
  - college level **大学水平** 485 - 487
  - construction versus choice structure **建构对选择建构** 478 - 481
  - diagnostic and prognostic **诊断和预后** 488 - 489
  - general achievement batteries **一般成就测验** 481 - 483
  - minimum competency in basic skills **基本技能中的最低能力** 483 - 484
  - teacher-made **教师自编的** 484 - 485
- Educational Testing Service (ETS) **教育测验服务中心 (ETS)** 25, 28, 557
- Anchor Test Study **参照测验研究** 70
- creation of **创造** 43
- Golden Rule case and **戈尔登·鲁尔案** 546
- Learning plus program **学习加计划** 490
- National Adult Literacy Survey **全国成人识字调查** 483 - 484
- 1990 Annual Report of 1990 **年年度报告** 481
- Educational use of tests **测验的教育运用** 474 - 490
- Education for all Handicapped Children Act (1995) (P.L.94 - 142 ) **残疾儿童教育法案(1975年)** 237, 247, 518 - 519
- Edwards Personal Preference Schedule (EPPS) **埃德华个人偏爱表 (EPPS)** 370 - 371, 377, 552
- Ego development **自我发展** 455
- Embedded Figures Test **包含图形测验** 446 - 447, 552

- Empirical criterion keying **经验准则确定**
- California Psychological Inventory **加利福尼亚心理调查表** 359 – 360
- defined **限定的** 350 – 351
- Minnesota Multiphasic Personality Inventory **明尼苏达多相人格调查表** 74, 351 – 359
- Personality Inventory for Children **儿童人格调查表** 360 – 362
- Employee Polygraph Protection Act (1998) (P.L. 100 – 347) **雇员测谎保护法案(1988)** 507
- Employment tests, validation of **就业测验, 效度** 491 – 496
- Enhanced Guide for Occupational Exploration (EGOE) **职业探究增强型指南(EGOE)** 526
- Environment, multicultural testing impacted by **环境, 受环境影响的多元文化测验** 269 – 270
- Environment, measures of **环境, 的测量**
- assessment of environment **环境的评定** 269 – 270
- perceived environment **环境感** 461 – 463
- Equal Employment Opportunity Act (P.L. 88 – 352) **就业机会均等法案** 545 – 546
- Equal Employment Opportunity Commission (EEOC) **就业机会均等委员会** 545 – 546, 556
- Equipercntile method **同等百分位数方式** 70
- Error of central tendency **趋中误差** 466
- Error of estimate **估计误差** 141 – 144
- Error of measurement **测量误差** 84, 107 – 109
- Errors, rating **误差, 评定** 466 – 467
- Error variance **误差方差** 84 – 85
- Ethical Principles in Conduct of Research with Human Participants (APA)

## 人类参与的研究行为的伦理准则 540n

Ethical Principles of psychologists and Code of Conduct (APA) 心理学家的伦理原则和行为规范 (APA) 533 - 534

## Ethics 伦理学

code of 规范 533 - 534

communicating test results 交流测验结果 543 - 544

confidentiality issues 保密问题 542 - 543

issues described 描述的问题 535 - 536

privacy issues 隐私问题 540 - 541

qualifications and professional competence issues 资格和职业的胜任问题 536 - 538

responsibilities of test publishers 测验出版者的责任 538 - 540

testing of diverse populations 对不同人群的测验 544 - 550

## Examiner 主试

preparation 准备 13 - 14

qualifications 资格 10 - 11, 536 - 538

reliability 信度 99 - 100

variables 效度 17 - 19

variance 方差 99

Exner's Comprehensive System 埃克斯纳综合系统 414 - 416

Expectancy tables/charts 期望表/图 81 - 83, 141

Experimental psychology influences on early test development 实验心理学, 对早期测验发展的影响 34

Face validity 表面效度 117 - 118

Factor analysis 因素分析 41, 288

cognitive task analysis and 认知任务分析 320

composition of a test 测验的因素构成 307 - 308

- construct validity and **结构效度** 128 - 129
- factor loadings and correlation **因素负荷和相关** 308 - 309
- factor matrix **因素矩阵** 303 - 304
- of intelligence **智力的** 303 - 309
- interpretation of **解释** 306 - 307
- oblique axes and second-order factors **斜交坐标轴和二级因素** 309
- references axes **参照坐标轴** 304 - 306
- self-report personality inventories and **自陈人格问卷** 362 - 367
- Factorial validity of a test **测验的因素效度** 128, 306
- Factor matrix **因素矩阵** 303 - 304
- Fagan Test of Infant Intelligence **费根婴儿智力测验** 245, 552
- Fair use of tests, decision models for **测验的公平应用, 决策模型** 170 - 171
- Faking **伪装**
- projective techniques and **投射技术** 433 - 434
- self-report personality inventories and **自陈人格问卷** 374 - 376
- False acceptances **错误接受** 144
- False rejections **错误拒绝** 145
- Feedback, test scores affected by **反馈, 受反馈影响的测验分数** 19
- Fels Research Institute **菲尔斯研究所** 325, 326
- Field dependence **场依存性** 446 - 448
- Finding Information **寻找信息** 28
- Five-Factor Model (FFM) **五因素模型 (FFM)** 364 - 367, 506
- Fixed reference group **固定参照组** 71 - 73
- Fleishman Job Analysis Survey (F-JAS) **弗莱希曼工作分析调查 (F-JAS)** 494
- Flexibility of closure **闭合的灵活性** 445
- Forced-choice technique **迫选技术** 376 - 377



- Forensic psychology 司法心理学 509n
- Free association test 自由联想测验 44, 425 - 426
- Frequency distribution 频数分布 49 - 51
- Frequency polygon 频数多边形 50
- Functional literacy 机能性读写算 483n
- G factor g 因素 310 - 311
- Gender-related issues 与性别有关的问题 407 - 408
- General Aptitude Test Battery (GATB) 一般能力倾向成套测验 (GATB) 26, 159, 496 - 499, 525, 552
- General factor 一般因素 310
- Generalizability theory of reliability 信度的概化理论 85n
- Generalized linear item-response theory (GLIRT) 概括的线性项目一反应理论 (GLIRT) 192n
- Gerontological Apperception Test 老年人统觉测验 423
- Gesell Developmental Schedules 格塞尔发展顺序量表 57, 236 - 237
- Golden Rule case 戈尔德·鲁尔案 199 - 200, 546
- Goodenough-Draw-a-Man Test 古迪纳夫“画男人”测验 19, 264 - 266
- Goodenough-Harris Drawing Test 古迪纳夫—哈里斯画人测验 264 - 266, 430, 552
- Gottschaldt Figures 戈特沙尔特图形 446
- Grade norms 年级常模 56
- Graduate Record Examination Board 研究生资格考试委员会 24, 25 - 26, 487
- Graduate Record Examinations (GRE) 研究生资格考试 (GRE) 487, 552, 557
- Group(s) 团体

- item discrimination and extreme **项目鉴别度和首项** 182 - 183
- item discrimination and small **项目鉴别度和小项** 183 - 185
- validity and contrasted **效度和对照的** 122
- Group Embedded Figure Test **团体包含图形测验** 446, 447, 448, 552
- Group factor **团体因素** 311
- Group testing **团体测验**
- adaptive testing **适应性测验** 274 - 278
  - advantages of **优点** 273
  - disadvantages of **缺点** 273 - 274
  - measuring multiple aptitudes **测量多重能力倾向** 287 - 293
  - multilevel batteries **多种水平成套测验** 278 - 287
  - origin **起源** 138 - 139
  - role of **作用** 38 - 39
  - versus individual testing **对个别测验** 272 - 274
- Guided assessment **指导性评价** 489
- Guides to the Assessment of Test Session Behavior for the WISC - III and the WIAT, WISC - III **和 WIAT 测验期间行为评价指南** 513, 552
- Guilford-Zimmerman Temperament Survey **吉尔福德—齐默尔曼气质调查表** 363, 552
- Guttman scale **格特曼量表** 57n, 77n, 406
- Halo effect **光环效应** 466
- Halstead-Reitan Neuropsychological Test Battery (HRB) **霍尔斯特德—瑞坦神经心理成套测验** 517, 552
- Haptic Intelligence Scale **触觉智力量表** 257n, 552
- Harrington-O'Shea Career Decision-Making System-Revised (CDM - R) **哈林顿—奥西职业决策系统修订版(CDM - R)** 396, 526, 552

- Hayes-Binet 海耶斯—比纳 256
- Head Start 海德—斯塔特 325, 330
- Hearing impaired, testing of 听觉受损, 的测验 254 – 255, 552
- Heritability of intelligence 智力的遗传力 297 – 300
- Heteroscedasticity 异质变异性 143
- Hierarchical theories 层次理论 316 – 318
- Hiskey-Nebraska Test of Learning Aptitude 希斯基—内布拉斯加学习  
能力倾向测验 255
- Histogramms 直方图 50
- Hogan Personality Inventory (HPI) 霍根人格问卷(HPI) 508, 552
- Holtzman Inkblot Technique (HIT) 霍兹曼墨迹技术 (HIT) 418 –  
419, 554
- Home Observation for Measurement of the Environment (HOME) 环境测  
量的家庭观察(HOME) 270
- Homogeneity of test items 测验项目的同质性 97 – 98
- Hopkins Symptom Checklist 霍普金斯症状检查表 350
- House-Tree-Person 房屋—树木—人 430, 552
- How to Take a Test: Doing Your Best (Dobbin) 如何参加测验：尽力而  
为 26
- Human figure drawings (HFDs) 人像图画 (HFDs) 430
- Human Intelligence International Newsletter 人类智力国际新闻通讯  
299
- I – E Scale (Internal-External Scale) I – E 量表(内部—外部量表)  
408 – 409
- Illusory validation 错觉效度 438
- Impression management 印象处理 375
- In-basket test 篮子测验 493

- Incremental validity **增值效度** 148
- Index of discrimination **鉴别度指数** 185 – 186
- Index of reliability **信度指数** 100
- Individuals with Disabilities Education Act (IDEA) **残疾人教育法案 (IDEA)** 247, 252 – 253
- Individual testing, group testing versus **个别测验, 团体测验** 272 – 274
- Industrial use of tests **测验在工业领域的应用** 490 – 509
- Infant and preschool testing, See also Childhood, early **婴儿和学前测验, 参见童年早期**
- administration of **指导语** 235 – 236
- Bayley Scales of Infant Development **贝利婴儿发展量表** 237 – 239
- current trends in **当前的倾向** 244 – 247
- historical background of **历史背景** 236 – 237
- McCarthy Scales of Children's Abilities (MSCA) **麦卡锡儿童力量表 (MSCA)** 239 – 240
- Piagetian scales **皮亚杰量表** 240 – 244
- predictive validity of **预测效度** 328
- types of tests **测验的类型** 236
- Infant Psychological Development Scales **婴儿心理发展量表** 241 – 242, 552
- Infant-Toddler Developmental Assessment (IDA) **婴幼儿发展评定 (IDA)** 245 – 246
- Inferential statistics **推断统计** 53n
- Information Bulletin (Graduate Record Examination Board) **信息公告 (研究生资格考试委员会)** 25
- Informed consent **获得信息后的同意** 541
- Inkblot techniques **墨迹技术**

- alternative approaches **选择方法** 416 - 418
- Exner's Comprehensive System **埃克斯纳综合系统** 414 - 416
- Holtzman **霍兹曼** 418 - 419
- Rorschach **罗夏** 5, 195, 411 - 414
- Instability of intelligence test performance, reasons for **智力测验成绩的不稳定性, 原因** 326 - 327
- Institute for Personality Assessment and Research (IPAR) **人格评定研究所(IPAR)** 452, 457 - 458
- Integrity tests **诚实测验** 507
- Intelligence **智力**
- Defined **界定** 294 - 295, 320 - 322, 497
  - in early childhood **童年早期的** 327 - 330
  - factor analysis of **因素分析** 303 - 309
  - heritability and modifiability **遗传力和可变性** 297 - 300
  - longitudinal studies of children's **儿童的追踪研究** 323 - 327
  - motivation and **动机** 300 - 302
  - personality and **人格** 302
  - problems in testing adult **成年智力测验中的问题** 331 - 337
  - trait organization theories **特质组织理论** 309 - 318
  - traits, role of **特质的作用** 318 - 322
  - use of academic **学术智力的运用** 497 - 498
- Intelligence quotient (IQ) **智商 (IQ)**
- Deviation **离差** 64 - 66
  - first use of **首次运用** 38
  - meaning of **意义** 295 - 297
- Intelligence Test for Visually Impaired Children (ITVIC) **视力受损儿童的智力测验(ITVIC)** 257
- Intelligence test performance **智力测验成绩**

listability of **不稳定性** 326 – 327

population change in **人群变化** 337 – 340

stability of **稳定性** 324 – 326

Intelligence tests. See also under name of **智力测验, 参见智力名下的各条内容**

Binte **比纳** 37 – 38

clinical and counseling testing and **临床和咨询中的测验** 512 – 514

group **团体** 38 – 39

use of **运用** 205

Interactive computer systems **交互式计算机系统** 75

Intercept bias **截距偏向** 168 – 170

Interest tests **兴趣测验**

Career Assessment Inventory-Vocational Version **职业评定调查表—职业版** 400

current setting **现状** 387 – 389

Jackson Vocational Interest Survey **杰克森职业兴趣调查表** 397 – 398

Kuder Occupational Interest Survey **库德职业兴趣调查表** 398 – 399

opinion surveys and attitude scales **意见调查和态度量表** 404 – 409

overview of **总的看法** 396 – 402

Self-Directed Search **自我向向的探索** 400 – 402

sex fairness and **性别的公平性** 388

Strong Interest Inventory **斯特朗兴趣调查** 389 – 396

trends **倾向** 402 – 404

Interitem consistency, measures of **项目间的一致性, 测量** 97 – 98

- Internal consistency 内部一致性 129
- Internal-External Scale (I-E Scale) 内部—外部量表(I-E 量表)  
408-409
- Interval scales, item difficulty and 间距量表,项目难度 174
- Interviewing techniques 访谈技术 464-465
- Iowa Test of Basic Skills 依阿华基本技能测验 482,553
- Iowa Test of Educational Development 依阿华教育发展水平测验  
482,553
- IQ. See Intelligence quotient IQ. 智商,参见智商
- Item analysis 项目分析
- cross-validation 交叉效度 194-196
  - current developments in 当前发展 200-202
  - differential item functioning 项目的不同功能 196-200
  - item difficulty 项目难度 173-179
  - item discrimination 项目鉴别度 179-187
  - item response theory 项目反应理论 73-74,187-192
  - role of 作用 172
  - of speeded tests 速度测验 193-194
- Item characteristic curve (ICC) theory 项目特征曲线(ICC)理论 189-191
- Item difficulty 项目难度
- distribution of test scores 测验分数的分布 177-178
  - interval scales 间距量表 174
  - percentage passing 通过百分率 173-174
  - testing purpose and 测验目的 179
  - Thurstone absolute scaling 瑟斯顿绝对量表 174-177
- Item discrimination 项目鉴别度
- biserial correlation 二列相关 187

- choice of criterion **效标的选取** 179 – 182
- extreme groups used in **用于项目的首项题组** 182 – 183
- index of discrimination **鉴别度指数** 185 – 186
- phi coefficient  $\varphi$  **系数** 186 – 187
- small groups used in **用于项目的小项题组** 183 – 185
- statistical indices of **统计指数** 182
- Item information functions **项目信息函数** 191
- Item response theory(IRT) **项目反应理论** 73 – 74
- basic features **基本特征** 189 – 191
- current status of **当前状况** 192
- item -test regression **项目 — 测验回归** 187 – 188
- models **模型** 191 – 192
- Item sampling, and test reliability **项目取样, 测验信度** 94 – 97
- Item -test regression **项目 — 测验回归** 187 – 188
- Item writing **项目编制** 484
- Jackson Personality Inventory-Revised(JPI – R ) **杰克森人格调查表—修订版( JPI – R)** 373,553
- Jackson's Basic Personality Inventory **杰克森基本人格调查表** 358, 551
- Jackson Vocational Interest Survey (JVIS) **杰克森职业兴趣调查表** 397 – 398,553
- J-coefficient J- **系数** 495
- Job analysis method **工作分析法** 493 – 494
- Job element method **工作要素法** 493 – 494
- Job performance **工作成绩**
- criteria **效标** 121 – 122,495 – 496
- predicting **预测** 494 – 495



- Job samples 工作样本 492
- Joint Service Job performance 联合军种工作绩效  
Measurement/Enlistment Standards (JPM) project 测量/入伍标准  
(JPM)计划 500 - 501
- Journal of personality Assessment 人格评定杂志 401,435n
- Kautman Adolescent and Adult Intelligence Test (KAIT) 考夫曼青少年  
和成人智力测验 (KAIT) 224 - 225,553
- Kautman Assessment Battery for Children (K - ABC) 考夫曼儿童评定成  
套测验 (K - ABC) 222 - 224,236,553
- Kautman Brief Intelligence Test (K - BIT) 考夫曼简短智力测验 (K -  
BIT) 225,553
- Kautman scales, development of 考夫曼量表,的发展 222
- Kautman Test of Educational Achievement (K - TEA) 考夫曼教育成就  
测验 (K - TEA) 520,553
- Kent -Rosanoff Free Association Test 肯特—洛沙诺夫自由联想测验  
426
- Kinetic Family Drawing (KFD) 动态家庭图画 (KFD) 430 - 431
- Kuder General Interest Survey (KGIS) 库德一般兴趣调查 (KGIS)  
339,553
- Kuder Occupational Interest Survey (KOIS) 库德职业兴趣调查表  
(KOIS) 398 - 399,553
- Kuder Preference Record-Vocational 库德爱好档案—职业版 398 -  
399,553
- Kuder-Richardson reliability 库德—理查逊信度 97 - 99
- Kuhlmann-Binet 库尔曼—比纳 38
- Lambda coefficient 拉姆达系数 399n
- Language, in transcultural testing 跨文化测验中的语言 343 - 345

- Latent trait models **潜在特质模型** 73
- Latent trait theory **潜在特质理论** 189 – 191
- Leaderless Group Discussion (LGD) **无领导团体讨论 (LGD)** 452
- Leadership, testing for **领导能力, 领导能力测验** 507 – 508
- Learning disabilities (LD) **学习失能**
- assessment techniques **评定技术** 519 – 521
  - defined **界定** 518
  - dynamic assessment **动力评定** 521 – 522
  - identifying **鉴别** 518 – 522
- Learning-potential assessment **学习潜力评定** 521 – 522
- Legal regulations **法规**
- civil rights and testing **公民权利和测验** 544 – 547
  - ethical issues and **伦理问题** 545 – 547
  - learning disabilities and **学习失能** 518 – 519
  - physical disabilities and **身体残疾** 252 – 254
- Leiter International Performance Scale **利特国际操作量表** 258
- Leiter International Performance Scale-revised **利特国际操作量表—修订版** 261 – 263, 553
- Licensing of psychologists **心理学家的执照** 537 – 538
- Lie detector **测谎仪** 425 – 426
- Life-history data **生活史的资料** 469 – 471
- Life Values Inventory **生活价值调查** 387
- Likert-type scale **利克特型量表** 406 – 407
- Linear transformations **线性转换** 61 – 62
- Local norms **地方常模** 71
- Locus of control **控制点** 408 – 409
- Longitudinal studies of children's intelligence **儿童智力的追踪研究** 323 – 327

Luria-Nebraska Neuropsychological Battery (LNNB) 鲁利亚-耐布拉斯卡神经心理成套测验(LNNB) 517 - 518, 553

Machover Draw-a-Person Test (D - A - P) 马乔弗尔画人测验 (D - A - P) 429 - 430, 553

Mastery testing 掌握测验

domain-referenced testing and 领域参照测验 78 - 79  
reliability and 信度 112

McCarthy Scales of Children's Abilities (MSCA) 麦卡锡儿童能力量表 (MSCA) 239 - 240, 553

Mean 平均数 51, 52

Mean square deviation 均方差 53

Measures of Personality and Social Psychological Attitudes 人格和社会心理态度的测量 407

Measures of styles and types 风格和类型的测量

cognitive styles 认知风格 444 - 448

personality types 人格类型 448 - 450

Mechanical aptitude tests 机械能力倾向测验 503

Median 中位数 51, 52

Mediated learning experience 中介学习经验 343

Mental age 智力年龄 37 - 38, 55 - 56

Mental level 智力水平 37, 38

Mental Measurements Yearbook (MMY) 心理测量年鉴 (MMY) 27 - 28, 204

Mental retardation 智力落后

assessment of 评定 247 - 252

early classification of 早期分类 33 - 34

Mental test(s) 心理测验

- early 早期的 36 – 37
- first use of term 术语的首次应用 36
- Meta - analysis 元分析 125 – 126
- Metropolitan Achievement Test 城市成就测验 70,553
- Metropolitan Readiness Tests 城市准备状态测验 490,553
- Micro CAT 微软 CAT 277
- MicroCog: Assessment of Cognitive Functioning 米克洛库格：认知功能  
评定 530,553
- Millon Adolescent Clinical Inventory (MACI) 米隆青少年临床调查表  
(MACI) 369 – 370,553
- Millon Adolescent Personality Inventory (MAPI) 米隆青少年人格调查  
表(MAPI) 369 – 370,553
- Millon Clinical Multiaxial Inventory – III (MCMI – III) 米隆临床多轴  
调查表 – III(MCMI – III) 367 – 370,553
- Millon Index of Personality Styles (MIPS) 米隆人格类型指数 (MIPS)  
370,553
- Minimum competency tests in basic skills 基本技能的最低能力测验  
483 – 484
- Minnesota Multiphasic Personality Inventory (MMPI, MMPI – 2, MMPI – A)  
明尼苏达多相人格调查表 (MMPI, MMPI – 2, MMPI – A) 74,  
351 – 359,553
- adolescent (MMPI – A) 青少年 (MMPI – A) 357
- comments about 的评论 357 – 359
- description of 的描述 351 – 352
- revision (MMPI – 2) 修改本 (MMPI – 2) 352 – 357
- Minorities, testing of. See Multicultural testing 少数民族,的测验,的多  
元文化测验
- Mode 众数 51,52

- Moderator variables 调节变量 155 - 156
- Motivation, intelligence and 动机,智力 300 - 302
- Motor impaired, testing of 运动受损, 的测验 257 - 259
- Multicultural testing 多元文化测验
- approaches to 方法 267 - 269
  - description of problem 问题描述 259 - 261
  - ethical issues 伦理问题 544 - 550
  - impact of environment on 环境对道德的影响 269 - 270
  - other names for 其他名称 260n
  - traditional methods 传统方式 261 - 266
- Multidimensional Aptitude Battery (MAB) 多维能力倾向成套测验  
290 - 293, 318, 553
- Multilevel batteries 多重水平成套测验
- overview of 总的看法 278 - 279
  - recognition of multiple aptitudes 对多重能力的认识 286 - 287
  - representative 代表性的 279 - 280
  - test content at different levels 不同水平的测验内容 280 - 286
- Multiple aptitude batteries/tests 多重能力倾向成套测验 41
- differential aptitude test 区别能力倾向测验 110, 288 - 290
  - measuring 测量 287 - 293
  - multidimensional aptitude tests 多维倾向能力测验 290 - 293
- Multiple choice items 多项选择题 272, 478
- Multiple correlation 多重相关 158
- Multiple discriminant function 多重鉴别功能 161 - 162
- Multiple-factor theories 多因素理论 311 - 314
- Multiple regression equation 多元回归方程 157 - 159
- Multitrait-multimethod matrix 多重特质-多重方法矩阵
- Myers-Briggs Type Indicator (MBTI) 梅耶斯-布里格斯类型指标

( MBTI) 377,449 - 450,553

- Narrative computer interpretation 叙述性的计算机解释 74
- National Adult Literacy Survey 全国成人识字调查 483 - 484
- National anchor norms 全国参照常模 70 - 71
- National Assessment of Educational Progress 全国教育发展评定 76, 478
- National council on Measurement in Education (NCME) 全国教育测量委员会 (NCME) 29,558
- National Reference Scale 全国参照量表 70
- National Research Council 全国研究委员会 536,545
- Naturalistic observation 自然主义的观察 463 - 464
- Navy Classification Battery 海军分类成套测验 19
- NEO Personality Inventory (NEO PI - R) NEO 人格调查表 (NEO PI - R) 365 - 366
- Neuropsychological assessment 神经心理评定 514 - 518
- Nominating technique 提名技术 467 - 468
- Nonlanguage tests, role of 非语言测验,的作用 234
- Nonlinear transformation 非线性转换 62
- Nonreading tests 非阅读测验 234 - 235
- Nonverbal tests, role of 非言语测验,的作用 234 - 235
- Normal curve 正态曲线 50 - 51,178
- Normalized standard scores 正态的标准分数 62 - 63
- Normal percentile charts 正态的百分位数图 60
- Norm -referenced testing 常模参照测验 79 - 80
- Norms 常模
- developmental 发展的 54 - 58
- grade 年级 56

- local 本地的 71
- projective techniques and 投射技术 434 - 435
- relativity of 的相关 66 - 74
- role of 的作用 6 - 7, 48
- statistical concepts 统计概念 49 - 54
- subgroup/specific 亚团体/特定的 71
- within-group 团体内 58 - 66
- Objective questions on tests 测验的客观问题 478
- Objectivity, in psychological tests 客观性, 心理测验中的 7
- Oblique axes 斜交坐标轴 309
- Observer reports 观察者的报告 463 - 469
- Occupational Aptitude Pattern (OAP) 职业能力倾向模式 (OAP) 498 - 499
- Occupational Aptitude Survey and Interest Schedule - 2 (OASIS - 2) 职业能力倾向调查表和兴趣表 - 2 (OASIS - 2), 396, 553
- Occupational testing. See also Career; Interest tests 职业测验。参见 职业 兴趣测验
- personality testing in the workplace 工作场所的人格测验 505 - 509
- use of 的应用 496 - 505
- validation of employment tests 就业测验的效度 491 - 496
- Occupational themes 职业主题 390 - 395, 398, 399, 400, 401
- Office of Strategic Services (OSS) 战略服务机构 (OSS)
- assessment-center program 评价中心技术 452
- On Your Own; Preparing for a Standardized Test 依靠你自己 准备参加 标准化测验 26
- Open-ended tasks 开放式的任务 479

- Opinion surveys and attitude scales **意见调查和态度量表** 404 – 408
- Ordinal scales **顺序量表** 56 – 58
- Ordinal Scales of Psychological Development **心理发展的顺序量表**  
241 – 242, 553
- Orleans – Hanna. Algebra Prognosis Test **欧利恩斯—汉纳代数诊断  
测验** 488, 553
- Orthogonal axes **直角坐标轴** 309
- Oseretsky Tests of Motor Proficiency **奥西里斯基动作水平测验** 252
- Otis –Lennon School Ability Test (OLSAT) **奥蒂斯—莱农学校能力测  
验** 280, 281, 286, 482, 553
- Otis Self-Administering Tests of Mental Ability **奥蒂斯心理能力自我测  
验** 272, 497
- Paced Auditory Serial Addition Test (PASAT) **定步系列增加听觉测验  
(PASAT)** 531, 553
- Paper-and-pencil tests **纸一笔测验** 503 – 504 – 507
- Pattern analysis. See Profile analysis **模式分析, 参见个人能力测验分析**
- PDI Customer Service Inventory PDI **顾客服务问卷** 508, 553
- PDI Employment Inventory PDI **就业问卷** 508, 553
- Peabody Picture Vocabulary Test (PPVT – R) **皮博迪图片词汇测验**  
251, 258 – 259, 553
- Pearson Product-Moment Correlation Coefficient **皮尔逊积差相关系数**  
88, 90, 141
- Peer assessments **同伴评定** 467 – 468
- Percentage passing, item difficulty and **通过百分率, 项目难度** 173  
– 174
- Percentiles **百分率的** 58 – 61
- Perceptual functions **知觉功能** 445 – 446



- Performance techniques. See also Job performance **操作技术。参见工作成绩**
- drawing techniques **绘画技术** 429 – 431
- global **全面的** 492 – 493
- play and toy techniques **游戏和玩具技术** 431 – 432
- Performance tests, role of **操作测验, 的作用** 45, 121, 234
- Perkins-Binet Test of Intelligence for the Blind **帕金斯一比纳盲人智力测验** 256
- Personal and Academic Self-Concept Inventory (PASCI) **个人和学业自我概念问卷 (PASCI)** 457
- Personal constructs **个人概念** 460 – 461
- Personal Data Sheet **个人资料调查表** 44, 349, 350 – 351
- Personality, intelligence and **人格, 智力** 302
- Personality Assessment Inventory (PAI) **人格评定调查表 (PAI)** 358 – 359, 553
- Personality inventories, self-report **人格问卷, 自我报告**
- California Psychological Inventory **加利福尼亚心理调查表** 359 – 360
- content-related procedures **内容关联程序** 349 – 350
- current status of **当前状况** 385
- empirical criterion keying **经验主义标准的关键** 350 – 362
- factor analysis **因素分析** 362 – 367
- Minnesota Multiphasic Personality Inventory **明尼苏达多相人格调查表** 74, 351 – 359
- Personality Inventory for Children **儿童人格调查表** 360 – 362
- personality theory **人格理论** 367 – 374
- situational specificity **情境特性** 379 – 384
- test-taking attitudes and response biases **参加测试的态度和反应偏**

- 向 374 – 379
- Personality Inventory for Children-Revised (PIC – R) **儿童人格调查表修订版(PIC – R)** 360 – 362, 468n, 553
- Personality Inventory for Youth (PIY) **青年人格问卷 (PIY)** 362, 468n, 553
- Personality Research Form (PRF) **人格研究表 (PRF)** 371 – 374, 553
- Personality tests **人格测验** 44 – 45
- in the workplace **工作场所的** 505 – 509
- Personality theory **人格理论**
- Edwards Personal Preference Schedule **埃德华个人偏爱表** 370 – 371
- Millon Clinical Multiaxial Inventory **米隆临床多轴调查表** 367 – 370
- Personality Research Form **人格研究表** 371 – 374
- Personality types **人格类型** 448 – 450
- Personnel selection, use of tests in **人事选拔, 测验的应用** 144 – 164, 490 – 509
- Phi coefficient, item discrimination and  $\phi$  **系数, 项目辨别** 186 – 187
- Physical disabilities, testing persons with **身体残疾, 残疾者测验**
- hearing impairments **听力受损** 254 – 255
- legislation and issues regarding **立法和有关立法的问题** 252 – 254
- motor impairments **运动受损** 257 – 259
- visual impairments **视力受损** 256 – 257
- Piagetian ordinal scales **皮亚杰顺序量表** 127
- Piagetian scales **皮亚杰量表** 240 – 244
- Pictorial classification tests **图片分类测验** 259
- Pictorial techniques **图片技术**
- Children's Apperception Test **儿童统觉测验** 422

- Gerontological Apperception Test **老年人统觉测验** 423
- Roberts Apperception Test for Children **罗伯特儿童统觉测验** 422
- Rosenzweig Picture-Frustration Study **罗森茨威格图片一挫折研究**  
423 - 425
- Senior Apperception Test **高级统觉测验** 423
- Thematic Apperception Test **主题统觉测验** 419 - 423
- Picture vocabulary tes **图片词汇测验** 258
- Pintner-Paterson Performance Scale **平特一佩特森操作量表** 254
- Placement, use of term **安置,术语的运用** 160
- Play and toy techniques **游戏和玩具技术** 431 - 432
- Populations **总体**  
intelligence test performance and changes in **智力测验成绩和变化**  
337 - 340  
samples versus **样本对总体** 68 - 70
- Porteus Mazes **波蒂厄斯迷津** 258, 553
- Portfolio assessment **代表作的评价** 479
- Positive manifold **正流形** 305
- Power test **难度测验** 102
- Practicing to Take the GRE General Test **参加GRE一般测验的练习**  
25
- Predictive validity in infant and Preschool Testing **婴儿和学前测验的  
预测效力** 328
- Preschool testing. See Infant and preschool testing **学前测验。参见婴  
儿和学前测验**
- Preselection **预先选拔** 141 - 142
- Primary mental abilities **主要心理能力** 311 - 313
- Principles for the Validation and Use of Personnel Selection Procedures **人  
事选拔程序的有效性和使用的原则** 535, 546

- Privacy issues **隐私问题** 540 – 541
- Productivity, validity and **生产率, 效度** 149, 151
- Profile analysis **剖面分析**  
     cutoff scores and **临界分数** 159 – 160  
     of intelligence test results **智力测验结果** 512
- Profiles, examples **剖面, 例子**  
     aptitude test **能力测验** 110 – 111  
     personality test **人格测验** 355 – 356
- Prognostic testing **预测测验** 488 – 489
- Project A **投射 A** 124, 501
- Project Head Start **早期教育项目** 325
- Projective techniques **投射技术**  
     autobiographical memories **自传体记忆** 427 – 429  
     as clinical tools **作为临床工具的投射** 441 – 442  
     evaluation of **评估** 432 – 442  
     inkblot techniques **墨迹技术** 411 – 419  
     performance techniques **表演技术** 429 – 432  
     pictorial techniques **图片技术** 419 – 425  
     as psychometric instruments **作为心理测量工具的投射** 440 – 441  
     role of **的作用** 45, 410 – 411  
     verbal techniques **言语技术** 425 – 427
- Protocol analysis **礼仪分析** 135, 320
- Psychiatric diagnosis **精神病诊断** 123
- Psychological assessment **心理评定** 510 – 511  
     role of computers in **计算机在投射中的作用** 529 – 532
- Psychological tests. See also under name of test administration of **心理测验**. 参见测验名下的心理测验 13 – 17  
     capacity **能力** 5

- defined 界定 4-9
- diagnostic/predictive value of 诊断或预测的价值 4-5
- history of 的历史 32-37
- misuses 误用 10-12, 547-549
- objectivity of 的客观性 7
- reasons for controlling the use of 控制使用的原因 10-13
- reasons for using 使用的原因 2-4, 474
- reliability of 信效度 7-8
- sources of information 信息源 27-29
- standardization 标准化 6-7
- uses 使用 2-4
- validity of 的效度 8-9
- varieties 多样化 37-45
- Psychologists 心理学家
- first experimental 首批实验心理学家 34
- qualifications and professional competence issues 资格和职业胜任问题 536-538
- Psychomotor skills and testing 心理运动技能和测验 502-503
- Q sort Q 分类 458-459, 469
- Quality of Life Inventory (QOLI) 生活质量问卷 (QOLI) 532, 553
- Range 全距 52
- Rapport 和睦关系
- establishing 建立 15-17
- projective techniques and 投射技术 433
- Rasch model 拉希模型 191-192, 288
- Rater relevant contact 评价者接触程度 466

## Ratings 评定

criterion measurements 效标度量 123

errors 误差 466 - 467

as an observation technique 作为一种观察的技术 465 - 467

Raven's Progressive Matrices (RPM) 瑞文渐进推理测验 (RPM) 263  
- 264, 311, 343, 553

Raw scores 原始分数 6, 49

Readiness tests 学习准备测验 489 - 490

Reference axes in factor analysis 因素分析中的参照坐标轴 304 -  
306

Reference groups, fixed 参照组, 固定的 71 - 73

Regression equations 回归方程 157 - 159

Rehabilitation Act(1978) 康复法案 253

## Reliability 信度

ability levels and 能力水平 107

alternate-form 复本信度 93 - 95

coefficient 系数 90 - 91

cutoff scores and 临界分数 112

defined 界定 7 - 8, 84

Differential Ability Scales 区别能力量表 230 - 231

index of 指数 100

Kuder-Richardson 库德-理查逊 97 - 99

mastery testing and 掌握测验 112

overview of 总的看法 100 - 102

projective techniques and 投射技术 435 - 436

retest 复测 436

scorer 评分者 99 - 100

speeded tests and 速度测验 102 - 104

- split-half 分半 95 - 97
- standard error of measurement 测量的标准误差 107 - 111
- Stanford-Binet 斯坦福一比纳 211 - 212
- test length and 测验长度 96
- test - retest 再测信度 91 - 93
- variability and 效度 105 - 107
- Wechsler Intelligence Scales 韦克斯勒智力量表 219 - 220
- Report writing, in psychological assessment 报告的写作心理评价中  
报告的写作 528 - 529
- Rep Test 个人构念项目测验 460 - 461
- Response capabilities 反应能力 497
- Response sets and styles 反应定势和类型 377 - 379
- Responsibilities of Users of Standardized Tests (RUST statement ) 标准化  
测验使用者的责任 ( RUST 的声明) 535
- Retest reliability 复测信度 436
- Revised Minnesota Paper Form Board Test (RMPFBT ) 明尼苏达纸笔测  
验修订版(RMPFBT) 503,553
- Revised NEO Personality Inventory, (NEO PI - R) NEO 人格调查表修  
订版(NEO PI - R) 365 - 366,553
- R - I - A - S - E - C scales R - I - A - S - E - C 量表 403 - 404
- Roberts Apperception Test for Children (RATC) 罗伯特儿童统觉测验  
(RATC) 422,553
- Role Construct Repertory Test (Rep Test ) 角色构念项目测验 (Rep  
Test) 460 - 461
- Roleplaying techniques 角色扮演技术 453 - 454
- Rulon formula 鲁隆公式 98n
- Rorschach 罗夏 5,195,411 - 414 - 553
- comprehensive system, Exner's 综合系统,埃克斯纳的 414 - 416

- consensus Rorschach **罗夏测验的一致性** 417
- various approaches **各种方式** 416 – 418
- Rosenzweig Picture-Frustration Study **罗森茨威格图片—挫折研究**  
423 – 425, 554
- Rotter Incomplete Sentences Blank (RISB) **罗特不完全句子填充**  
(RISB) 426 – 427, 553
- Sample-free scaling **与样本无关的量表** 73 – 74
- Sample heterogeneity **样本的异质性** 141
- Samples, versus populations **样本, 对总体** 68 – 70
- Sampling **抽样**
- of behavior **行为的** 4 – 5
  - of content **内容的** 94
  - of persons **个体的** 68 – 70
- Scatter, age scales **分散, 年龄量表** 55
- Scenotest **戏剧测验** 431, 554
- Schaie-Thurstone Adult Mental Abilities Test **谢伊—瑟斯顿成人心理能力测验** 334n, 553
- Schedule for Nonadaptive and Adaptive personality (SNAP) **不适应性和适应性人格表(SNAP)** 369n, 553
- Scholastic Aptitude Test (SAT). 12. See also Scholastic Assessment Test  
**学业能力测验(SAT), 参见学业评定测验**
- declining scores **下降分数** 338 – 339
  - impact of coaching on **辅导的影响** 24
  - renaming of **重新命名** 72n, 475n
- Scholastic Assessment Test (SAT) **学业评定测验(SAT)** 62, 554
- change in **变化** 475n, 485 – 486
  - recentered score scale for **重新聚中分数的量表** 485 – 486



- scale continuity in 量表的持续性 72
- for the visually impaired 视力受损者的 256
- School readiness 入学准备 489 - 490
- School Vista software 学校展望软件 490<sub>n</sub>
- Science Research Associates (SRA) 科学研究会 (SRA) 505
- Scorer reliability 评分者效度 99 - 100
- Scores, types of 分数, 的类型 54 - 66
- Screening 筛选 160
- Seattle Longitudinal Study (SLS) 西雅图追踪研究 (SLS) 333 - 334
- Second-order factors 次级因素 309
- Selection-response tasks 选择反应任务 479, 480, 481
- Selection, use of term 选择, 术语的应用 160
- Selection ratio 录取率 146 - 148
- Self-concepts and personal constructs 自我概念和个人构念 454 - 463
- Adjective Check List 形容词核查表 457 - 458
- environmental perception 环境知觉 461 - 463
- Q sort Q 分类 458 - 459
- Role Construct Repertory Test 角色构念项目测验 460 - 461
- self-esteem inventories 自尊问卷 455 - 457
- Semantic Differential 语义区分 459 - 460
- Washington University Sentence Completion Test 华盛顿大学完成句子测验 427<sub>n</sub>, 454 - 455
- Self-Directed Search (SDS) 自我定向研究 (SDS) 400 - 402, 554
- Self-esteem inventories 自尊问卷 455 - 457
- Self-Esteem Scale (SES) 自尊量表 (SES) 456
- Self-exploration 自我探索 387
- Self-fulfilling prophecy 自我实现的预言 18

- Self-report, client **自我报告, 委托人** 524
- Self-report inventory, 44. See also Personality Inventories, self-report **自陈问卷, 参见人格问卷, 自我报告**
- Semantic Differential **语义区分** 459 – 460
- Senior Apperception Test **高级统觉测验** 423, 554
- Sentence completion **句子完成** 426 – 427
- Sequential patterning **序列模式** 127
- Sequential strategies, decision theory and **序列决策, 决定论** 153 – 154
- Significance levels **显著性水平** 90
- Simple structure **简单结构** 305
- Simulation tests **模拟测验** 492
- Situational specificity, self-report personality Tests and **情境特殊性, 自陈人格测验** 379 – 384
- Situational stress test **情境压力测验** 453
- Situational tests **情境测验**  
     assessment-center programs and roleplaying **以评定为中心的程序和角色扮演** 452 – 454
- Character Education Inquiry **性格教育调查** 451 – 452  
     role of **的作用** 45, 450 – 451
- Sixteen Personality Factor Questionnaire (16PF) 16 **人格因素问卷 (16PF)** 363 – 364, 554
- Skewed distribution curves **偏态分布曲线** 177 – 178
- Slope bias **斜率偏向** 165 – 168
- Social Climates Scales **社会气氛量表** 462 – 463, 554
- Social desirability, self-report personality inventories and **社会适宜性, 自陈人格问卷** 374 – 376
- Social Skills Rating System (SSRS) **社会技能评定系统 (SSRS)** 524,

554

Society for Industrial and Organizational Psychology (SIOP) 工业和组织

心理学会(SIOP) 491, 535, 559

Society for Personality Assessment 人格评定协会 528n, 559

Spatial aptitude tests 空间能力倾向测验 503 - 504

Spatial Learning Ability Test 空间学习能力测验 201

Spearman-Brown formula 斯皮尔曼—布朗公式 96, 98n, 101, 103,  
172

Special aptitude testing 特殊能力倾向测验 39 - 42, 501 - 505

Speed tests 快速测验

definition 界定 102

item analysis of 项目分析 193 - 194

reliability and 信度 102 - 104

Spiral-omnibus format 螺旋混合形式 272 - 273

Split-half reliability 分半效度 95 - 97

Standard deviation(SD) 标准差 53 - 54

Standard error of measurement (SEM) 测量的标准误差 (SEM) 107  
- 111

Standard error of estimate 估计的标准误 143

Standard error of score differences 分数差异的标准误 109 - 111

Standardization, test 标准化, 测验 6 - 7

Standardization sample 标准化样本 6

Standardized achievement test. See Achievement tests 标准化成就测  
验。参见成就测验

Standard scores 标准分数 61 - 64

Standards for Educational and Psychological Testing (APA) 教育和心理  
测验(APA)标准 29 - 31, 118 - 126, 474, 491, 510 535, 546

Stanford Achievement Test 斯坦福成就测验 42, 482, 554

- Stanford-Binet Intelligence Scales **斯坦福一比纳智力量表** 38, 40,  
42, 236, 512, 554
- administration and scoring **施测和评分** 207 - 210
- as a criterion in validating group tests **作为各种团体测验效度分析的效标** 123
- evolution of scales **量表的发展** 205 - 207
- fourth edition (SB - IV) **第 4 版斯坦福一比纳量表 (SB - IV)**  
207 - 214
- reliability **信度** 211 - 212
- standardization and norms **标准化和常模** 210 - 211
- validity **效度** 212 - 214
- for the visually impaired **视力受损者的** 256
- Stanford Diagnostic Mathematics Test **斯坦福数学诊断测验** 488, 554
- Stanford Diagnostic Reading Test **斯坦福阅读诊断测验** 488, 554
- Stanford Writing Assessment Program **斯坦福写作评定项目** 488, 554
- Stanine scale **标准九分量表** 63
- States versus traits **状态对特质** 384
- State-Trait Anger Expression Inventory (STAXI) **状态—特质愤怒表现调查表(STAXI)** 384, 554
- State-Trait Anxiety Inventory (STAI) **状态—特质焦虑调查表** 384,  
554
- State - Trait Anxiety Inventory for Children (STAIC) **状态—特质儿童焦虑调查表** 384, 554
- Statistical concepts **统计概念** 49 - 54
- Statistical significance **统计显著性** 88 - 90, 109 - 111
- Stereotype accuracy **刻板印象的准确性** 437 - 438, 450
- Stereotype vulnerability **刻板印象的脆弱性** 343
- Strong Interest Inventory (SII) **斯特朗兴趣调查表 (SII)** 554

- origins and development of **起源和发展** 389 – 390
- psychometric evaluation **心理测量学的评价** 395 – 396
- scoring and interpretation **记分和解释** 394 – 395
- Strong Vocational Interest Blank (SVIB) **斯特朗职业兴趣表 (SVIB)**  
389
- Structural equation modeling **结构方程建模** 132 – 134
- Structure of Intellect Learning Abilities Test **智力结构—学习能力测验**  
314, 554
- Structure-of-intellect model **智力结构模型** 314 – 316
- Student Adaptation to College Questionnaire (SACQ) **大学生适应性问卷(SACQ)** 532, 554
- Student Self-Concept Scale (SSCS) **学生自我概念量表 (SSCS)** 456  
– 457, 554
- Student Styles Questionnaire **学生风格问卷** 449, 554
- Study of Values **价值研究** 387<sub>n</sub>
- Subgroup/specific norms **亚团体/具体常模** 71
- Suppressor variables **抑制变量** 158 – 159
- Symptom Checklist – 90 – Revised (SCL – 90 – R) **症状检查 90 项—修订版(SCL – 90 – R)** 349 – 350, 554
- Synthetic validation **综合效度** 495
- System for Interactive Guidance Information (SIGI) **交互式指导信息系统** 75, 554
- T score **T 分数** 63
- Taking the SAT (College Board) **参加 SAT(大学董事会)** 25
- Task decomposition **任务分解** 135
- Taylor-Russell tables **泰勒—拉塞尔预期表** 146 – 148
- Teacher-made classroom tests **教师自编的课堂测验** 484 – 485

- Teamwork-KSA Test **团队工作 – KSA 测验** 505,554
- Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA) **心理测验和诊断技术的方法推荐 (APA)** 126
- TEMAS (Tell-Me-A-Story) **TEMAS (给我讲个故事)** 422 – 423,554
- TerraNova series **特纳诺瓦系列** 483n,554
- Test (s). See also under name of **测验。参见测验名下的各种条目**
- anxiety **焦虑** 20 – 22,383,480
  - classification and use of **分类和运用** 160 – 164
  - communication **交往** 12
  - environment **环境** 14 – 15
  - equating **相等** 74
  - information curve **信息曲线** 109
  - methods of introducing **引入方法** 15 – 17
  - sophistication **技巧** 24 – 26
  - sources of information about **关于测验的信息源** 27 – 31
  - specifications **说明** 115,484
- Test Anxiety Inventory (TAI) **测验焦虑调查表(TAI)** 22,383,554
- Test Anxiety Profile **测验焦虑剖面图** 383
- Test batteries **测验成套**
- defined **界定** 156
  - multiple regression equation **多元回归方程** 157 – 159
  - profile analysis and cutoff scores **剖面分析和临界分数** 159 – 160
- Test bias **测验偏向**
- differential item functioning **项目的不同功能** 196 – 200
  - intercept **截距** 168 – 170
  - self-report personality tests and response biases **自陈人格测验和反应偏向** 374 – 379
  - slope **斜率** 165 – 168

- statistical analysis of **统计分析** 164 - 171
- Test ceiling, effect on score distribution **测验难度上限,对分数分布的影响** 177 - 178
- Test Collection Bibliographies (ETS) **测验文献目录集(ETS)** 28
- Test construction, validity and **测验的构成,效度** 137 - 138
- Test content, security of **测验内容,的可靠性** 12 - 13
- Test examiners **测验的主试**
- advanced preparation of **主试的预先准备** 12 - 13
  - cultural diversity and **文化差异** 345 - 346
  - need for qualified **所需资格** 10 - 11
  - projective techniques and **投射技术** 434
  - test scores affected by **受主试影响的测验分数** 17 - 19
- Test fairness **测验公平**
- decision models for **针对主试的决策模型** 170 - 171
  - use of term **术语的应用** 169
- Test floor, effect on score distribution **测验的难度下限,对分数分布的影响** 177 - 178
- Testing the limits **测量限度** 521
- Test of Cognitive Skills (TCS) **认知技能测验 (TCS)** 280, 283, 285, 286, 554
- Test of Nonverbal Intelligence (TONI - 2) **非言语智力测验(TONI - 2)** 266
- Test of Variables of Attention (TOVA) **注意变化测验** 531, 554
- Test publishers, responsibilities of **测验的出版者,的责任** 538 - 540
- Test results, communicating **测验结果,交流** 543 - 544
- Test - retest reliability **再测信度** 91 - 83
- Test scores, factors affecting **测验分数,影响测验分数的因素**
- answer sheets, types of **答题纸,的类型** 14 - 15

- feedback 反馈 19
- test examiners 测验的主试 17 - 19
- training/coaching 训练/辅导 23 - 27
- Test scores , interpreting 测验分数, 的解释
- computers used for 计算机用于测验评分 74 - 76
- cutoff scores, use of 临界分数, 的使用 80 - 83
- domain-referenced 领域参照 76 - 80
- ethical issues 伦理问题 545 - 550
- norm -referenced 常模参照 79 - 80
- standard error measurement 标准误差测量 107 - 111
- Tests in Print 已出版的测验 28
- Tests of Achievement and Proficiency 学习成就和效率测验 482, 554
- Tests of Adult Basic Education (TABE ) 成人基础教育测验 ( TABE )  
484, 554
- Test takers, views of 测验参与者, 的观点 20 - 23
- Test -taking attitudes and response biases 测验参与者的态度和反应偏向 374 - 379
- Test user, role of 测验使用者, 的作用 11 - 12
- Test User Qualification Working Group (TUQWoG) 测验使用者资格工作小组 ( TUQWoG ) 11
- Test User Training Working Group (TUTWoG) 测验使用者训练工作小组 ( TUTWoG ) 11 - 12
- Thematic Apperception Test (TAT) 主题统觉测验 (TAT) 419 - 423
- Thurstone absolute scaling 瑟斯顿绝对量表 174 - 177
- Thurstone-type scale 瑟斯顿型量表 406
- Time -lag comparisons 时间延滞比较 333
- Training./coaching, test scores affected by 训练/辅导, 受训练/辅导影响的测验分数 23 - 27



- Trait organization theories 特质组织理论 309
- hierarchical theories 层次理论 316 – 318
- multiple-factor theories 多因素论 311 – 314
- structure-of-intellect model 智力结构模型 314 – 316
- two-factor theory 二因素论 310 – 311
- Traits 特质
- self-report personality inventories and situational specificity 自陈人格  
问卷和情境特殊性 379 – 384
- states versus 状态对特质 384
- T scores T分数 63
- Two-Factor Index of Social Position 社会地位的二因素指标 269
- Two-factor theory 二因素论 310 – 311
- Uniform Guidelines on Employee Selection Procedures (1978 ) 雇员选拔  
程序的统一指导纲要 (1978 年) 546
- U.S. Air Force 美国空军 63, 161
- U.S. Army 美国陆军
- Alpha and Beta tests 军队甲种测验和乙种测验 39, 234, 271
- Aptitude Area 能力倾向领域 163
- Selection and Classification Project 选拔和分类计划 124, 161,  
496, 501
- U.S. Civil Service Commission 美国文官选拔委员会 43
- U.S. Department of Defense 美国国防部 488, 500
- U.S. Department of Labor 美国劳工部 498 – 499
- U.S. Employment Service (USES) 美国职业服务中心 (USES) 26.  
159, 498, 525 – 526, 560
- U.S. Office of Education 美国教育部 70
- U.S. Office of personnel Management 美国人事管理处 488, 493

U.S. Office of Strategic Services (OSS) **美国战略服务局 (OSS)** 45, 452

University of California, Institute of Personality Assessment and Research,  
**加利福尼亚大学, 人格评定研究所** 452, 457 - 458

Utility, decision theory and **利用, 决定论** 151 - 153

## Validity **效度**

comparison of different types **不同类型的比较** 136 - 139

consequences of testing **测验的结果** 138 - 139

construct-identification procedures **结构鉴别程序** 126-135

content-description **内容描述** 114 - 118

convergent and discriminant **幅合和区别** 129 - 130

criterion-prediction **效标预测** 118 - 126

cross-validation **交叉效度** 194 - 196

decision theory and **决定论** 144 - 156

differential **区分** 161

Differential Ability Scales **区别能力量表** 231 - 232

early uses of **早期使用** 114

face **表面效度** 117 - 118

factorial **因素的** 128 - 129

impact of, on individuals and society **效度对个人和社会的影响**  
138 - 139

incremental **增值效度** 148

productivity and **效度和生产率的关系** 149, 151

projective techniques and **投射技术** 437 - 439

of psychological tests **心理测验的** 8 - 9

Test construction and **测验结构** 137 - 138

Validity coefficient **效度系数** 8 - 9

- conditions affecting **影响的因素** 141 - 143
- error of estimate and **估计误差** 141 - 144
- magnitude of **效度的量值** 143 - 144
- Validity generalization(VG) **效度的概化** 124 - 125, 499
- Values Scale **价值量表** 387, 554
- Variability **离散性** 52 - 53
- reliability and **信度** 105 - 107
- Variance **方差** 53
- Verbal techniques **言语技术** 425 - 427
- Vigil Continuous Performance Test **夜班连续操作测验** 531, 554
- Vineland Adaptive Behavior Scales (VABS) **瓦因兰德适应行为量表**  
249 - 251, 554
- Vineland Social Maturity Scale **瓦因兰德社会成熟量表** 249
- Visually impaired, testing of **视觉受损的测验** 256 - 257
- Vocational interests. See Career; Interest tests **职业兴趣。参见职业；  
兴趣测验**
- Washington University Sentence Completion Test (WUSCT) **华盛顿大学  
不完全测验(WUSCT)** 427n, 545 - 455
- Wechsler Adult Intelligence Scale (WAIS - R) **韦克斯勒成人智力量  
表(WAIS - R)** 75, 215 - 216, 291
- Wechsler-Bellevue Intelligence Scale **韦克斯勒一贝勒维智力量表**  
215
- Wechsler Individual Achievement Test (WIAT) **韦克斯勒个人成就测  
验(WIAT)** 521, 554
- Wechsler Intelligence Scales **韦克斯勒智力量表** 62, 65, 512
- abbreviated scales/short forms **简化量表/简短表格** 217, 219
- age decrement **年龄下降** 331 - 333

- description of scales **量表的描述** 216 – 217
- development of **量表的发展** 214 – 216
- norms and scoring **常模和评分** 219
- reliability **信度** 219 – 220
- validity **效度** 220 – 221
- for the visually impaired **视力受损者的** 256
- Wechsler Intelligence Scale for Children (WISC – R, WISS – III) **韦克  
斯勒儿童智力量表(WISC – R, WISS – III)** 75, 215, 216, 554
- Performance Scale for Deaf Children **聋儿的操作量表** 254 – 255
- Wechsler Preschool and Primary Scale of Intelligence (WPPSI, WPPSI – R)  
**韦克斯勒学龄前及小学生智力量表 (WPPSI, WPPSI – R)**  
216, 236, 554
- Within-group norms **团体内标准** 58 – 66
- Wonderlic Personnel Test **温德利克人事测验** 497, 554
- Woodcock-Johnson Psycho-Educational Battery-Revised **伍德科克一约  
翰逊心理教育成套测验修订版** 239n, 520, 554
- Woodcock Reading Mastery Tests-Revised **伍德科克阅读掌握测验修  
订版** 232, 554
- Woodworth Personal Data Sheet **伍德沃斯个人资料调查表** 44, 349,  
350 – 351
- Word association test **词汇联想测验** 425 – 426
- Work Keys system **工作要素系统** 494, 554
- Z scores **Z 分数** 61