·讲座·

心理测验的编制与心理测量学分析

曲成毅

随着医学模式的转变,在流行病学研究中采用心理测验作为评价指标的研究与日俱增。仅就2003－2005 年中华流行病学杂志发表的论文统计,涉及心理测验的研究论文共计105 篇,平均每期2.92篇,而 10 年前同期(1993－1995 年)相同性质研究论文仅 3 篇,平均每期0.17篇,增幅 17 倍。流行病学研究中采用的心理测验有多种,按其功能主要可概括为:能力测验(智力、认知功能、体力活动等)、人格测验(性格、气质、情绪、态度、信念、价值观等)和学绩测验(如健康知识等)。知识-信念-行为(KAP)、生存质量(QOL)、健康状态、心理压力、自杀意念、不良行为、感觉统合等均可看作是多种心理现象的综合评定。

由于大多数心理现象的定义均不十分明确,特别是相同刺激物在同质个体中可能产生完全不同的心理反应,各种心理现象难以客观、精确评价。尽管许多研究者努力探索借助仪器设备通过实验对心理现象进行测量,但是迄今为止心理测验的主要方式还是通过观察人的少数有代表性的行为,对贯穿在全部行为活动中的心理特点做出推论。在这一过程中,量表成为保证测试结果是否真实的关键。流行病学工作者对心理测验量表的编制、修订高度重视,仅近 3 年(2003－2005 年)中华流行病学杂志发表有关量表编制、修订的论著就有 9 篇。无论自己研制或是引用成熟的量表,测试项目的选定、结果的解释、项目分析、信度和效度评价等均是判定量表优劣的重要依据。现就心理测验的编制与心理测量学分析中应当注意的几个重要环节简述如下。

1.测试项目的选定:心理测验的量表(scale)由各种不同类型的项目/题目(item)及为完成题目所必需的器具组成,量表比问卷(questionnaire)更加严谨,量表的结果要能对被试者的心理现象进行量化评价,组成量表的题目要反复推敲,精心选择。测试项目选定的主要依据是:①测量对象的特点如年龄、智力、文化背景、阅读水平等,同一测验目标不同测验对象选用题目的内容和形式大不相同,凡新创建量表的题目需经多次试测修改以完全适合测试对象,引用他人的量表应当首先关注该量表是否适合本研究的人群。国外引进的量表应当剔除或替代那些因文化传统的差别不适合中国国情的项目,要努力避免因翻译生硬致使理解困难甚至发生歧义,或翻译不当未能准确表达作者原意。②测量的目标如智力、体力、情绪、性格等,这是选定题目最重要的依据。心理学指标不同于生理学指标的特点之一是它的不确定性,心理测量的目标只是一个概念化的抽象定义,不同研究者依据这一定义可以具体化为若干不同的操作定义。比如智力测验表编制者可以将"智力"具体化为语言和操作两个能区(Wechsler),组合性智力、经验性智力、实用性智力三种成分(Sternberg),感觉运动期、前运算期、具体运算期、形式运算期 4 个阶段(Piaget),语文理解、语词流畅、数字运算、空间关系、机械记忆、知觉速度、一般推理 7 个因素(Thurstone)等[1]。依据不同理论编制的不同智力测验量表测定的"智商"只是依据量表作者设计的题目测查结果所反映的那部分"能力"。又如心理学理论认为,"态度"指个体对一组社会事物而产生的具有一致性的反应群,主要包括评价性的认知、对事物的情感和好恶和对事物的可观察的行动倾向三种成分,态度量表由针对上述三种成分的若干题目组成,每一题目的设计可采用针对某一事物的陈述,根据个人对题目所作反应给予分数,以代表他(她)对该事物所持态度的方向和强弱程度[2]。因此无论量表的编制者或使用者均应明确该量表所测定的心理现象是基于什么样的心理学理论,又具体化为什么样的操作执行定义。③测量的用途如群体评价、一般描述、探索病因、疾病诊断、选拔或预测等等,目的不同,编制测验的取材范围及题目的难度应当有所不同。

心理测量是通过外显行为推导心理现象,题目应当选择那些易于观察或判断的行为或动作,资料收集越齐全,设计项目则越顺利,测验的内容才不致有所偏颇,才能提高行为样本的代表性。如编制人格测验量表时应注意收集人格的主要理论,用于描述人格的术语,临床观察到的不同人格类型的行为表现,其他人格测验的项目等。收集资料时应当注意行为的普遍性,即该行为无论职业类别、职务大小、文化高低甚至性别差异,是一般人群都会表现出的普遍现象。收集题目的数量应当比最终量表所需数量多两倍以上,以便经试测后筛选及编制复本。一般地说,量表包括的题目越多,测试结果越接近真实,由于各种人格测验目标相对更不明确,组成量表的题目也相对较多,比如常用于测定成人人格特征的量表艾森克人格问卷(EPQ)有 90 题,16 种人格因素测验(16PF)187 个题,明尼苏达多项人格测验(MMPI)多达 566 个题目。在流行病学研究中,由于重点考察心理现象的人群特征,在保证调查结果真实性的前提下量表的题目尽可能精简,比如"一般健康问卷"(GHQ)有 60 题、30 题、28 题和 12 题四个版本[3],在筛查人群心理障碍时往往选用 12 题(GHQ-12),在我国的使用适应性研究结果证明,GHQ-12的

作者单位:030001 太原,山西医科大学公共卫生学院流行病学教研室

各项评价指标甚至优于 GHQ-30[4]。

由于行为和心理现象具有社会性，题目的表达形式可分为自评(self-reports)和他评(external assessments)两类。对题目的回答依据测试内容不同其要求有所不同，能力或学绩测验测量的是最高行为，应当鼓励被试努力发挥最高水平。各种人格测验测量的是典型行为，题目的设计应当诱导被试表现出自己对某一事件的真实想法。

经过试测和项目分析，选择区分度(differential)高，难度(difficulty)适宜的题目经过适当编排合成测验量表。题目的编排应当遵循由易到难，先一般问题后敏感问题的原则，这样可以避免被试在难题上耽误时间太多，或因涉及隐私影响后面问题作答。题目的编排可采取先将题目按照内容或形式不同分为不同的分测验，每一分测验均可独立施测，若干分测验联合组成一个量表，通常称此类编排为并列直进式。另一类先将所有题目依难易程度分为不同层次，再将不同性质的题目予以组合，作交叉式排列，难度逐渐上升，称为混合螺旋式。

2. 计分和结果的解释：每一题目的计分可采用定性和定量两种形式，定性以是/否、有/无、通过/失败、同意/不同意、平均水平以上/平均水平以下等形式记作 0，1。问题可以有单选或多项选择，前者问题单一，每一问题只有一个答案，对于能力测验，通常是以是否能够完成某一操作或是否具备某种行为表示；后者一个问题有多种答案，但其中有一个是最适宜的最佳答案，其余为与问题相关但是非正确(或非最佳)的干扰答案，此类题目在知识调查中较为常用。定量计分一般采用等级数据，依据心理状态的不同等级计分，目前编制的各类人格测验量表项目计分由3～11 个等级不同。比如某研究者设计对艾滋病患者态度问卷题目之一："政府出钱治疗艾滋病患者，还不如拿这笔钱救助贫困地区失学儿童"。答案可以在非常同意/比较同意/无所谓/比较不同意/很不同意中选择，按正向或负向计做 5，4，3，2，1。

每一题目的计分通常称为原始分，全量表得分的计算最简单的方法是总加量表法(Likert 量表)，即将每题得分简单算数相加。但这要求每一个题目计分方式相同，每一题在总量表得分中是等值的。比如上述态度量表，如果全量表有 20 个题，每个题在总的态度中所占比重应当相同。它的优点是制作过程简单，凡与研究主题有关的项目均可列入量表内，测量范围较宽，可以通过增加项目提高信度，可以充分表达受试者态度的强烈程度，但缺点也很明显：相同态度分数者可能持有不同的态度模式，有些人喜欢以激烈的方式表达其轻微的不同意程度，有人则以温和的方式表达其轻微的不同意，同样的态度可能获得不同的分数[2]。

如果同一个量表中的题目采用了不同的计分方式，比如有 0，1 记分，又有等级计分，有些题目满分为 8，有些满分为 40 等，总分的合成就不能将原始分简单相加，需要将原始分转换为具有一定参照点和单位的导出分数。依据解释分数时的参照标准不同，导出分数分为常模参照分数和标准参照分数两类。所谓常模参照是将受试者的测量值与同质总体作比较，根据受试者在该人群中的相对位置来报告他(她)的结果，这是心理测量的主要评分方法，正如 Anastasi 所说：事实上，心理测验就是针对行为样组进行的客观的和标准化的测量[2]。此类心理测验量表的编制重要环节是应当提供一系列能代表同质总体的样本人群(或称标准人群)的分数，相当于生理指标的正常值，不同性质标准人群的正常值组成常模(norms)。能力测验 0，1 计分的正常值常以某一题目在标准人群中通过率为50％～75％时确定，多重计分的正常值则常采用公式：$Z=(x-\bar{x})/s$ 进行转化，称为标准分或量表分。式中 x 为某人的原始分数，$\bar{x}$ 为标准人群该项目得分的平均分数，s 为标准差，可以想象，Z 值的范围一般由 -3～$+3$，此时已涵盖了 99％以上的人群。由于 Z 值经常出现小数点和负数，常采用 $A+BZ$ 的形式再次转化，这样无论原始分满分值是多少，也无论该题目得分的分布为何种类型，均转换为以 A 为均数，B 为标准差的正态分布。比如韦氏智力测验的量表分是 $10+3Z$，智力测验的最终结果智商为 $100+15Z$，以此分值的高低判断行为或心理水平。如果研究者无法判定该指标总体分布是否为正态，或理论推导不可能呈正态，此时可用百分位数或其他方式进行标准化。所谓标准参照分数是将测验结果与某种特定的标准比较，一种标准是对测验材料实际掌握的程度，称内容参照分数，另一种是用预期的效标成绩来解释测验分数，称结果参照分数，此类计分常用于知识类调查量表设计。

某一心理特征往往不是一种单一的心理现象，比如智力就是由若干不同性质的能力组成，即使仅仅测定一个人的记忆力，就可分为色彩、图形、言语、文字、数字、事件、方位、空间、概念等不同性质的记忆，还要分瞬时记忆、短期记忆、长期记忆等不同记忆的形式。人格测验的结构就更加复杂，比如自杀意念的调查按照一般行为发展理论分为愿望(wish)、态度(attitude)、计划(plan)三个阶段，意念的产生涉及抑郁、孤僻、猜疑、犹豫、淡漠、冲动、幼稚、病态人格等不同品质的心理现象，组成自杀意念量表的题目就应当反映被试者以上不同心理现象[5]。所以量表的编制者首先应当明确本量表测查的心理特征由几个心理现象组成，哪些题目的结果归类于哪些心理现象(有些量表被称为维度)，不同性质的心理现象往往构成一个个分测验(或分量表)，这些分测验相对独立，又互相联系，这一过程经常采用统计学因子分析的手段完成。在由不同题目的原始分合成分测验得分，由各分测验得分合成总量表分时要考虑的问题是各题目在分测验中以及各分测验在总量表得分中是等值？还是有不同的权重。事实上采用标准分就将每个变量作了与它的标准差成比例的加权，有时还需要采用专家判断(Delphi 法)综合考虑确定权重。

20 世纪 80 年代后期，计算机在心理测量的实践中逐渐发挥重要作用，使用范围涉及实施测验、建立题库、适应性测验、评分和解释等，特别是将常模数据储存于计算机，使评分

过程变得更加精确和便捷，但是，解释报告只能与专家判断结合起来使用，否则受试者个人生活背景和测验当时的情况可能被忽略[6]。

3. 测验的项目分析：在试测的基础之上对各个项目进行分析是编制和修订测验的重要环节。在定性分析的基础之上，难度和区分度是两个重要的评价指标。

难度指测验题目的难易程度，对于采用二分法计分的项目，难度以该题目在标准人群中的通过率表示，$P=(R/N)\times100\%$，式中 P 为通过率或称难度指数（difficult index），R 为答对或通过该项目的人数，N 为全体被试人数，通过率越高表明难度越小。当被试人数较多时，可以先将被试按照测验总分从高到低排序，将总分最高的 27% 和最低的 27% 被试定为高分组和低分组，分别计算两组在某一项目上的通过率，采用公式 $P=(P_H+P_L)/2$ 计算通过率，式中 P_H 和 P_L 分别为高分组和低分组的通过率。对于选择题，由于允许猜测，备选答案数目越少，机遇的作用越大，为平衡机遇对难度的影响，可用公式校正：$CP=(KP-1)/(K-1)$，式中 CP 为校正后的通过率，K 为备选答案数。对于等级或其他多重计分的项目，可用公式 $P=\bar{x}/x_{max}$ 计算难度，式中 $\bar{x}$ 为全体被试在某一项目上的平均分，x_{max} 为该项目的满分。

对项目的难度分析主要目的是为了筛选项目，如果测验的目的是为了区分个体之间的差别，对于通过率为 100% 或 0 的题目就失去意义，此时应当选择通过率接近 50% 的题目，因为此时区分度最高；如果测验的目的是为了考核能力或知识，对于特定的测试对象，允许有很高的通过率。

区分度是指测验项目对被试心理特征的区分能力，区分度的简单计算常采用高分组和低分组在某一项目通过率的差值表示：$D=P_H-P_L$，此处 P_H 和 P_L 的计算方法同前，D 称为鉴别指数，D 值越大，项目区分度越高，一般认为，具有良好鉴别力的量表，所有项目区分度应达到0.3以上。

计算区分度最常用的方法是相关法，即以某一项目分数与效标分数或测验总分的相关作为该项目区分度的指标，相关系数越高，该项目的区分度越高。对于两个连续变量的资料，可计算 Pearson 积差相关系数，对于一个变量为连续变量，另一个变量为二分变量的资料，如项目计分通过为 1，未通过为 0，量表总分为连续变量，可采用点二列相关计算，公式如下：

$$r_{pb}=\frac{\bar{x}_p-\bar{x}_q}{s_t}\sqrt{pq}\quad 或$$

$$r_{pb}=\frac{\bar{x}_p-\bar{x}_t}{s_t}\sqrt{\frac{p}{q}}$$

式中 $\bar{x}_p$ 为二分变量通过组对应的连续变量的平均数，如该题回答正确者总分的平均数，$\bar{x}_q$ 为二分变量未通过组对应的连续变量的平均数，如该题回答错误者总分的平均数，$\bar{x}_t$ 为总的连续变量的平均数，s_t 为连续变量的标准差，p 为通过组人数与总人数之比，q 为未通过组人数与总人数之比。此外还可以采用 t 检验比较二分变量通过组与未通过组对应的连续变量的均数之间差异是否有统计学意义，如均数差异显著，则相关系数也显著。

对于多项选择题，还要分析对每一个备选答案的反映情况，比如正确的被选答案被所有被试选中，说明该题目太容易或题目中可能提供了某种暗示；如果某个错误答案没有一个被试选择，说明该题目不具迷惑力；如果所有被试都选择某一错误答案，说明编制测验时定错了答案，或者全体被试获取了错误信息；如果高分组与低分组在某一答案的选择相等甚至高分组低于低分组，说明该题目可能与考察的心理特征无关；如果一个题目被试未答人数过多或各备选答案人数相等，说明题目过难或题意不清，被试凭猜测作答等。

在知识调查中常以某一已知的标准作为参照，无论题目难度如何，凡是在要求达到的知识范围之内均应保留。有时候采用教育前后两次相同题目测查，如果同一对象两次测查结果分别为 F 和 P（F 为失败，P 为通过），说明题目难度适宜且效果较好，如果为 PF，说明题目编制错误或效果太差，如果为 FF，说明题目太难或效果太差，如果为 PP 说明题目太易。两次测查结果某题目通过率之差，或后测结果达标组和未达标组某题目通过率之差可视为区分度。

4. 测量的信度：信度（reliability）指重复测量的一致程度，表示测量的一致性或可靠性，影响信度（以及效度）的因素是在测量和评价时可能出现各种误差。心理测量的误差主要包括随机误差（测量误差）与系统误差（偏倚）。前者是与测量目的无关偶然因素引起而又不易控制的误差，既影响信度又影响效度。后者是与测量目的无关变量引起的恒定而有规律的效应，只与效度有关。故此，心理测量实得分数的变异或称观察分数的总变异数（S_O^2）包括由测量工具引起的，与测量目的有关的变异数（S_V^2）、恒定的与测量目的无关的变异数（S_I^2）及测量误差（S_E^2）三部分组成。

在测量理论中，信度被定义为一组测量分数的真变异数（包括与测量目的有关的和无关的变异数）与总变异数（实得分数的变异数或称观察分数的变异数）的比率，计算时以信度系数 r_{xx}^2 表示：$r_{xx}^2=(S_V^2+S_I^2)/S_O^2$，由于 $S_O^2=S_V^2+S_I^2+S_E^2$，所以 $r_{xx}^2=1-S_E^2/S_O^2$

实际操作时，采用同一批受试者间隔一定时间前后两次测定结果的相关系数表示称再测信度（test-retest reliability），此为量表稳定性方法即判定两次相关施测的一致性，这是一种最简单、最直接的计算信度的方法，缺点是第一次施测时所产生的效应会对第二次施测有影响，称为“练习效应”。

用同一批受试者不同测试或评分者间测定结果的相关系数表示称评分者信度（scorer reliability），两个评分者间的信度评价可采用 *Kappa* 指数，多个评分者间的信度则采用和谐系数（coefficient of concordance），计算公式

$$W=\frac{\sum R_i^2-\frac{(\sum R_i)^2}{N}}{\frac{1}{12}K^2(N^3-N)}$$

式中 W 为和谐系数，K 为评分者人数，N 为被试人数，R_i 是每一被试评分的等级和。

信度测验的另一种方法是编制同样性质的题目两份对同一批受试者测定，计算其结果的相关系数称复本信度(alternate-form reliability)，此为等值性方法，每个复本都要包含和另一个复本相同数量和类型的题目，有同样的平均数、标准差及题目之间的相关度。由于编制完全相同的两份量表难度较大，有人将一个完整的测验结果分为两半，比如一个分数来自奇数题目，另一半分数来自偶数题目，计算两个分半分数之间的相关特称分半信度(split-half reliability)。分半信度计算的相关是半个测验的信度，因此常用 Spearman 公式加以校正：$r = 2r_{hh}/(1 + r_{hh})$，此处 r_{hh} 为两个分半数的相关系数。

将一份量表分为对等的两半有多种方式，不同的分半方法可能得到不同的分半系数，于是研究者在评价心理测量量表信度时采用另一种"分半"的方法即判断量表题目(行为领域)的一致性，称同质性信度(homogeneous reliability)。当题目为二分变量时，采用 Kulder-Richardson 公式计算相关系数

$$r_{kk} = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum p_i q_i}{s_x^2}\right)$$

式中 K 表示测验的题数，p_i 为通过某题之人数比例，q_i 为未通过该题人数比例，s_x^2 为测验总分的变异系数(方差)，当已知各个项目的难度时，使用这一公式十分方便。当采用下列公式时，可不必逐题计算通过率

$$r_{kk} = \frac{Ks_x^2 - \bar{x}(K - \bar{x})}{(K-1)s_x^2}$$

式中 $\bar{x}$ 为测验总分的平均数。当题目为多重计分时，采用 *Cronbach's* α 系数

$$\alpha = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum s_i^2}{s_x^2}\right)$$

式中 s_i^2 为每题的变异系数，s_x^2 为总分的变异系数。如果测量误差变异相对于观察的总变异增加，则信度下降。标准参照测验题目分数的变异一般较小，计算信度时需要对相关法计算的信度系数予以校正，一般采用 Livingston 提出的公式

$$r_{CR} = \frac{r_{NR}s^2 + (\bar{x} - C)^2}{s^2 + (\bar{x} - C)^2}$$

式中，r_{CR} 为标准参照测验的信度，r_{NR} 为采用相关法计算的信度系数，s 是分数的标准差，$\bar{x}$ 是分数的均数，C 是达到标准的分数线。对于所有题目均采用 0,1 计分的量表，两次测查结果的一致性还可采用 Lindeman-Merenda 提出的公式

$$C = \frac{nb - sf}{nb + v(n + b + v)}$$

式中 C 为一致性，n 为两次施测中均计分为 0(未达标)的人数，b 为两次施测中均计分为 1(达标)的人数，f 为仅在第一次施测中计分为 1(达标)的人数，s 为仅在第二次施测中计分为 1(达标)的人数，v 为 f 或 s 中较小的值。

信度系数的大小依据测量工具的用途而定，如果我们的兴趣在于区分不同的群组，信度系数一般达到0.7即可接受，如果针对个体进行诊断，信度需达到0.85[7]。一般地说，心理测量的题目越多，测验越长，信度越高。这是因为一方面题目越多，反应心理特征就越全面，另一方面，测验的项目多，在每个项目上的随机误差就可以互相抵消。题目的难度也会影响信度，难度过高或过低会使测验分数的范围过窄或过宽，区分度小的题目信度较低。量表的不稳定性还可能由于在不同场合提出的问题不同造成，例如在面谈时使用不同的语气，不同的兴趣，谈话环境中的温度、湿度、照明度、通风、是否有其他干扰等均会影响提问者和回答者。回答者的心身状态如饥饿、干渴、疲劳、心境、合作态度、注意力不集中等也会影响施测结果。如果误差的来源以同样的方式和程度影响测查，则对信度不会产生影响，但会影响效度，如果上述误差以随机的方式出现，则会增加测量误差。

5.测量的效度：效度指实际测量值和真值之间的一致程度，即一个测验对它所要测量的特质准确测量的程度，在测量理论中，效度被定义为与测量目的有关的真实变异数与总变异数的比率。计算时以效度系数 r_{xy}^2 表示

$$r_{xy}^2 = S_V^2/S_O^2$$

由于 $S_O^2 = S_V^2 + S_I^2 + S_E^2$，$r_{xx}^2 = (S_V^2 + S_I^2)/S_O^2$，所以 $r_{xy}^2 = r_{xx}^2 - S_I^2/S_O^2$，也就是说，效度总是受信度制约，要想提高效度，先决条件是有较高的信度，但有较高的信度，却不一定有较高的效度，效度的高低还要取决于恒定的与测量目的无关的变异数(S_I^2)在总变异中的比例(S_I^2/S_O^2)。

由于大多数心理现象是建立在一个概念基础之上的定义，什么是欲测定的心理现象的"真值"往往很难确定，这给心理测量效度评价带来困难。有时候研究者不得不声明他(她)所研究的就是采用该量表测定的那种心理特质。正如 Wolfe 所说：那些看起来测的是它想要测得东西的测量工具，总要比那些表面上看不出来想要测什么的测量工具更有效一些[7]。想要测得的东西是什么？这是心理测量的基本点，也是难点。焦虑、抑郁、压力、敌意、沮丧等心理测量的指标或变量是研究者对人的心理活动一种构想，它是建立在观察基础上并希望有助于解释人类行为原因的一种抽象事物。

效度评价实际操作时，常常采用由专家判断测验题目是否符合原定内容范围的方法，这称为内容效度(content validity)，尽管这种评价带有较大的评价者本身的主观性，但如果参评的专家具有较高水准，评价程序得当，还是可以得出有价值的结论。具体评价方法和 Delphi 法相似，首先应当确定欲测定内容的范围，比如有关知识或技能的轮廓，编制不同层次的纲目及加权比例，制定评定量表，约请一定数量的相关专家作出评定。内容效度对标准参照测验更为准确，因为在标准参照测验中研究者主要考核被试的知识，而哪些是被试应当掌握的知识范围专家判断时最有把握。所谓表面效度(face validity)从技术层面看并非真正含义的效度，它的参照物不是测到的真值而是测得的表象[3]。比如采用一个具有内容效度的测量儿童能力的量表施测于成年人，可能

因为题目的表述或测评的行为对成年人不相适宜甚或幼稚可笑，导致因缺乏表面效度而影响了内容效度，这一点在儿童量表和成人量表互用时尤为关注[8]。

审查测验结果是否符合心理学上的理论见解称构想效度或结构效度(construct validity)，比如从理论上认为正常人群中的智力呈正态分布，某智力测验在标准人群中测定结果符合正态分布，此时可认为该量表具有构想效度。因此，确定构想效度的关键是研究者必须提出所测心理现象的理论构想，有时候这种构想可能只是一种假设，测验过程实际是在验证这一假设，从这一层面看，建立在假设基础之上心理测量结果的推理应当十分谨慎。考察测验的同质性是判定构想效度的方法之一，它可以提供测验所测的心理现象是单一特质还是多种特质，Kuder-Richardson 公式和 α 系数均可作为构想效度评价的指标。构想效度评价最重要的方法是作因子分析，找到影响测验分数的共同因素，计算每个测验在共同因素上的负荷，由专家判定此类负荷是否与设计时的构想相一致，这种效度有时称为因素效度(factorial validity)[2]。武阳丰等[9]在编制"国人生活质量普适量表"时采用因子分析的方法证明该量表较WHO-100和SF-36有更好的构想效度，其理论依据是各因子的方差贡献率相差越小越好。有时判断理论构想还可以用实验的方法，比如理论上认为学生在考试之前会出现焦虑，某一测量焦虑的量表在一场考试前后对同一批对象测查分值有显著差异，说明该量表具有这方面的构想效度。

使用最多的评价效度的方法为效标关联效度或标准效度(criterion related validity)，它是采用一种参照标准如学业成就、临床诊断、工作表现、社会地位、经济水平或大家公认的某一量表作为效标来评价现有量表，有时称为实证效度。由于心理现象的复杂性，效标依研究者对所测心理现象的理解而定，效标仅仅可以看作是一个"准金标准"。比如，学业成就可以依教师评语，社会评价，工作表现而定；精神症状可以依据临床诊断而定；特殊训练成绩，年龄，职业等不同群体测定结果的差别等均可用以设定为效标。前述武阳丰等为证实国人生活质量普适量表的效度，采用各种慢性病患病率作为效标[9]。各类心理测量中尤以"人格"测验的效标最难确定，目前用于考验人格测验效度的方法包括：①能否识别在客观基础上选出的极端的一群人；②是否与专家评定结果一致；③能否预言将来的行为；④是否与生活资料或临床记录一致；⑤以理论上的推断去判断等，但这些都不是理想的效标。

影响效度的因素主要来源于系统误差，除信度之外，项目的质量和数量、施测的过程、被试的特点和状态均会影响结果。比如在人格测验中，施测者想要得到的是被试者的典型行为，但由于受到社会称许的影响，被试可能发生并未表述其真实心态的伪饰现象。为此，测试者需要调整题目的社会称许性，向被试阐明测验目的与个人品德无关，同时在量表中设计测量伪饰项目，以便对被试的反应态度做出评价。效度评价应当注意心理特征既具有稳定性且有可变性，比如智商经过长时期仍相对稳定，用测验作为诊断和预测，但能力可随年龄、教育、训练而改变，人格特征则更是随环境而变动。大量研究表明：人群智力分布曲线基本呈常态分布，但智力低的一端人次较多，可能是因为除按正常变异规律由遗传基因引起智力落后外，还可由疾病、脑外伤等造成的智力缺陷。由于不同群体对某些测验的项目有不同的经验，不同语言文化对测验结果可能发生影响，人们试图编制"超文化"的测验测查先天遗传的心理特质，非语言测验就是一种尝试。由于个体的一切行为都渗透着文化环境的影响，如果一个测验排除了所有与文化有关的材料，它就脱离了具体内容，因而不能测量任何有实际意义的东西。因此，人们又开始探索"文化公平"测验，使不同团体对所有测验项目都具有相同的经验或平均安排有利于不同团体的项目，以此提高量表的效度。

增加题目数量不仅提高信度，也能提高效度，改变后的效度值可用以下公式计算

$$r_{(nx)y} = \frac{nr_{xy}}{\sqrt{n(1 - r_{xx} + nr)}}$$

式中 $r_{(nx)y}$ 是测验增长为原来 n 倍的效度值，r_{xy} 为原测验的效度，r_{xx} 为原测验的信度。

信度和效度评价的指标大多数采用相关系数，从理论上讲，被试异质性越高，所测得的分值分布越宽，越容易出现相关。因此进行信度和效度评价时选用样本尽量选择心理特质差异悬殊者作为研究对象。

参 考 文 献

1 郑日昌，蔡永红，周益群. 心理测量学. 北京：人民教育出版社，1998. 9-26.

2 Anastasi A. Psychological testing. 5th ed. New York: Macmillan, 1982, 22-44, 103-183, 552-555.

3 Goldberg D, Williams PA. User's guide to the general health questionnaire. Berkshire, Nfer-Nelson, 1991. 64-67.

4 杨廷忠，黄丽，吴贞一. 中文健康问卷在中国大陆人群心理障碍筛选的适宜性研究. 中华流行病学杂志，2003，24：769-773.

5 Beck AT, Steer RA. Beck scale for suicide ideation. San Antonio, Psychological Corporation, 1991. 1-5.

6 Jackson C. 了解心理测量过程. 姚萍，译. 北京：北京大学出版社，2000. 35-47.

7 Aiken LR. 心理问卷与调查表. 张厚粲，译. 北京：中国轻工业出版社，2002. 172-175.

8 刘爱玲，马冠生，张倩，等. 小学生 7 天体力活动问卷信度和效度的评价. 中华流行病学杂志，2003，24：901-904.

9 武阳丰，谢高强，李莹，等. 国人生活质量普适量表的编制与评价. 中华流行病学杂志，2005，26：751-756.

(收稿日期：2006-03-06)

(本文编辑：张林东)