量表的信度及其影响因素

安胜利, 陈平雁
(第一军医大学 卫生统计学教研室, 广州, 510515)

中图分类号: B841.7 文献标识码: A 文章编号: 1005-3611(2001)04-0315-04

Measurement Scales: Reliability and Affecting Factors

AN Sheng-li, Chen Ping-Yan
Fist Military Medical University, Guang Zhou

【Abstract】 **Objective**: As a tool of measurement, clinical scales are widely used in the research of medicine. Reliability is one of the most important criteria for good clinical scales. In this review, methods to assess the reliability of scales and the factors that may affect reliability are summarized. Application and evaluation of statistical reliability are also briefly discussed.

【Key words】 Scale; Reliability; Affecting factors

量表作为测量工具广泛应于医疗卫生研究领域。评价量表的两个重要指标是信度(reliability)和效度(validity)。一个好的量表应既是可信的(reliable)又是有效的(valid)[1-3]。虽然高信度并不意味着高效度, 但是, 高效度一定是以高信度为前提的[4]。由此可见信度在量表评价中的重要性。本文将就量表设计对信度的影响进行回顾。

1 量表的信度

信度是对相同条件下重复测量结果的近似程度的一种度量, 它反映了一测量工具在多大程度上能辨别出被测者之间的差异[2,5,6]。信度的评价指标是信度系数 R, 理论上可以表达为真值方差(σ_T^2)与测量值方差(σ_M^2)的比值, 即被测者真值的变异在测量值的总变异中所占比例。

$$R = (\sigma_T^2)/(\sigma_M^2)$$

信度分析大致可分为两类, 即内部一致性分析(internal consistency analysis)和稳定性分析(stability analysis)或重复性分析(reproducibility analysis)。分半信度(split-half reliability), Cronbach α 系数(Cronbach' α coefficient)和 θ、Ω 系数用于评价内部一致性; 重测信度(test-retest reliability), 复本信度(alternate-form reliability, parallel-form reliability), 评分者间信度(inter-rater reliability)和评分者内信度(intra-rater reliability)用于评价稳定性或重复性。

1.1 分半信度

用一量表测量某一人群, 然后将量表中的条目随机分为数量相同的两半, 则每个被测者各有两个半表的合计分, 其积矩相关系数或秩相关系数为分半信度。最常用的方法是将条目按奇、偶数分半。分半信度只是半个量表的信度, 整个量表的信度须用 Spearman-Brown 公式校正,

$$R = 2R_1/(1+R_1)$$

式中 R1 为两个半表得分的相关系数。以上是在假定两个分半量表方差齐的条件下进行的。若两方差不齐, 则须用下式[7],

$$R = 2(1-\frac{\sigma_O^2+\sigma_E^2}{\sigma^2})$$

其中 σ_O^2、σ_E^2、σ^2 分别为奇数量表(或其中一个分半量表)、偶数量表(或另一个分半量表)和整个量表的方差。由于分拆量表的方法很多, 不同的分拆方法可能得出不同的分半信度, 故其稳定性较差, 这是该法的明显不足。

1.2 Cronbach α 系数

这是目前常用的评价内部一致性的方法, 几乎应用于所有的信度分析中[5],

$$a = \frac{K}{K-1}(1-\frac{\Sigma\sigma_i^2}{\sigma^2})$$

若量表中的条目以二分法(dichotomously)记分, 可用 KR-20(Kruder-Rechardson 20)公式计算 α,

$$\alpha = \frac{K-1}{K}(\frac{\sigma^2-\Sigma p_i q_i}{\sigma^2})$$

以上二式中, K 为量表中条目总数, σ_i^2 为第 i 个条目得分的方差, σ^2 为总分的方差, p_i 为第 i 个条目被测者回答为 0(或 1)所占的比例, $q_i = 1-p_i$。α 系数避免了分半信度的缺点, 它对量表的内部一致性作了更为慎重的估计, 因为它相当于以所有可能的组合分拆量表所得分半信度的平均值[4]。但 Patrick

认为除非量表中的条目数精确地反映了欲测内容，否则α系数将低估信度[8]，Waltz 也认为α系数可能低估了量表的信度[9]。一般地，α系数以大于0.7为好，但 Streiner 和 Norman 认为它不宜高于0.9，以避免人为地通过增加条目数量的方法达到提高α系数的目的，而这种条目数量的增加仅仅是靠同一问题以差异甚微的不同方式多次出现于量表中而实现的[9]。

1.3　基于因子分析的θ和Ω系数法

用各条目的得分构成一相关矩阵，求其特征方程的解，计算θ系数，

$$\theta = \frac{K}{K-1}(1-\frac{1}{\lambda})$$

式中 K 为条目数，λ为最大特征值。进一步利用主成分因子分析法求出各条目共性因子方差，计算Ω系数，

$$\Omega = 1-\frac{K-\Sigma h_i^2}{K+2r}$$

式中 r 为各条目间相关系数的总和，h_i^2 为第 i 个条目的共性方差。θ系数法要求分析的条目数在5个以上，以得到较稳定的结果；因子分析中各条目的共性方差 h_i^2 等于该条目在各因子上的载荷值的平方和[10]，所以，理论上Ω系数综合了各条目对欲测概念的个别贡献，且对量表的性质无特殊要求。目前，有关此方法应用于信度评价的报道尚少。巫秀美曾以"中老年预防结肠癌社区干预试验的健康行为问卷"为例，比较了α系数、θ系数和Ω系数，结果为α系数最小，Ω系数最大[11]。这一结论是否有普遍意义还有待于进一步的研究。

对某量表要作出正确的信度判断，单靠内部一致性分析是不够的[6]。

1.4　重测信度

用同一量表在同一人群中先后测量两次，其测量结果的积矩相关系数或秩相关系数为重测信度。两次测量可以由不同的调查者进行，也可以由同一调者进行，前者由于误差还可能来源于调查者对量表的理解差异及被测者影响不同，所以对其信度要求较后者高[12]。重测信度主要受两个因素的影响：首先，被测者的特征可能随时间发生变化，那么两次测量的差异就不单纯由随机误差引起，除非每个被测者都发生了同样的变化；其次，第二次测量受前一次的干扰，而不同的被测者有着不同的记忆力，从而发生残留效应(carry－over effect)。专家们认为间隔时间应据具体的调查内容而定，短可一小时，长可一年，一般为14天左右[6]。该法适于研究某些比较稳定的特征。

1.5　复本信度

设计出两个在形式、内容及难度上高度类似的量表，将两个量表同时或先后测量同一人群，两次结果的相关系数为复本信度。其中，先后测量的方法同样可出现重测信度评价中遇到的问题。该信度评价方法最接近于平行测试模型，但要设计出真正可互相替代的量表是非常困难的。

1.6　评分者间信度和评分者内信度

一般用于非自评量表的信度评价。前者是用于度量不同调查者间的一致性，后者是度量同一调查者在不同的场合下(如不同时间、地点等)的一致性。

1. 连续性资料　两名调查者的评分者间信度和测量两次的评分者内信度的评价可用积矩相关系数(或秩相关系数)，也可用组内相关系数 ICC(intra-class correlation coefficient)[6,13,14]，

$$ICC = \frac{MS_{区组} - MS_{误差}}{MS_{区组} + (K-1)MS_{误差} + \frac{K(MS_{处理} - MS_{误差})}{n}}$$

式中 $MS_{处理}$、$MS_{区组}$ 和 $MS_{误差}$ 分别为随机区组方差分析中的处理组(即调查者)间的均方、区组(即被测者)间的均方以及误差的均方，K 为调查者人数(此处为2)，n 为被测人数。对于总体的 ICC 是否等于0的假设检验可用下式进行，

$F = MS_{区组}/MS_{误差}$，$V_1 = 1$　　$V_2 = (n-1)(k-1)$

可见该检验实际上就是方差分析中对区组(即被测者)间是否相同的检验。在实际工作中，当误差均来源于随机变异时，积矩相关系数于 ICC，否则，前者大于后者。Streiner 认为 ICC 较积矩相关系数更真实地反映了信度水平[6]。对于多名调查者的评分者间信度和测量多次的评分者内信度的评价用 ICC。

2. 分类资料　两名调查者对被者作两分类的评定结果写成如附表形式，其信度评价用未加权 Kappa 系(K)[15]。

附表　配对记数资料的一般格式

		调查者甲		合计
		＋	－	
调查者乙	＋	a	b	a＋b
	－	c	d	c＋d
合　计		a＋c	b＋d	n

对于两名或多名调查者进行多分类评定的情形，则分别需用加权 Kappa 系数(weighted Kappa)和

广义 Kappa 系数(generalized Kappa)以评价调查者间的一致性,此不详述,请参阅有关文献[16-18]。

$$k = \frac{P_o - P_e}{1 - P_e}$$

式中 P_o 为实际符合率,

$$P_o = \frac{a + b}{n}$$

P_e 为期望符合率,

$$P_e = \frac{1}{n^2}[(a + c)(a + b) + (b + d)(c + d)]$$

对总体 Kappa 系数是否为零的检验见下式,

$$Z = k / \sqrt{V_k}$$

式中,Z 为标准正态离差,V_k 为方差,

$$V_K = \frac{P_O(1 - P_E)}{n(1 - P_E)^2}$$

评分者间/内信度小于0.4属较差;在0.4至0.75之间属于尚可;大于 0.75 属满意[5]。

值得一提的是 ICC 既可用于分类资料又可用于计量资料的信度评价[6,17]。对于同一分类资料,未加权 Kappa 值等于将被测特征编为 0,1 计算所得 ICC;当资料为多分类时,用标准权重算出的加权 Kappa 值等于 ICC。

2 影响信度的因素

任何能导致测量过程中产生误差的因素均可影响信度。

2.1 导致产生不一致的事件

(1)被测者的实际特征发生了变化。该现象在进行重测信度和一些平行量表信度的评价时会出现,前已述及。

(2)随机变化。这在所有信度评价中都可能发生。在被测者方面,如被测者在这一次测试时有病,在另一次 测试时无病;或在量表的这个版本测试中误解了某问题,而在另一版本测试中却没有误解;或对量表中容易理解的问题通过考虑回答,而对不易理解的问题却进行猜测等。另一方面,不同的调查者对被测者的影响不同;或在不同测试进行过程中一些可能的因素(如外环境)不同或不稳定都将会影响到信度的评价。

2.2 量表的设计

(1)应答条目的级数。许多被测特征是连续性变量,因而设行应答条目时应设法使之量化。一种方法是连续区间标度法(visual analogue scale,VAS),即一条通常为 100mm 的直线,两端表示极端状况(如极好、极差);另一种方法是将被测症状分为数个级(step,point)。显然,对连续性变量只分为两级会导致信息的损失。由此推广,如果条目中级的水平数低于被测者的判断能力,仍会损失信息。虽然在特定的条件下,人的辨别力似乎具有很大的偶然性,但有证据表明事实并非如此。大量研究认为,我们经常遇到的信度(0.4~0.9)将随着所用的级数的减少而降低。Nishisato 和 Torii 经验性地研究了该问题,认为"七级式"至"十级式"量表的信度较实际信度相比损失极少;"五级式"量表减少约 12% 的信度;"两级式"量表减少约 35% 的信度[6]。Nagata 用"四级式"、"五级式"、"七级式"和 VAS 等四种具体的量表调查了病人,结果是信度相似[19]。对于级数更多的情形尚未见报道。这方面的问题还有待于更深入全面的研究。

(2)条目的数量。对于测量目的相同但条目数量不同的量表,比较其信度高低时,应考虑条目数量因素。一般来说量表的条目数越多,信度越高,似但随着条目数的增加,被测者可能会出现疲劳或注意力不集中,反而使误差增大,信度降低。有时为权衡二者,在一定的条件下可将一个大的量表设计为几个分量表,以每个被测者在几个分量表中的总分或平均分作为最后得分以对该组合量表的信度进行较为准确的评价[4]。假设新增条目同原有条目肯同等程度的代表性,则当条目数增至原来的 K 倍时,利用 Spearman-Brown 公式,所得信度为,

$$R = \frac{Kr}{1 + (K - 1)r}$$

式中 r 为增加条目前的信度。

(3)条目的代表性。一个高信度的量表,其条目的陈述应客观并具有良好的代表性,否则将增加被测者应答时猜测的可能性,从而导致误差增大。信度降低。

(4)得分范围。量表得分范围(range)也会影响信度。将分组水平不同的被测者的得分合并以扩大得分范围,会增大信度。其根本原因是由于量表得分的总变异由被测者真实值变异和误差变异两部分组成,若总变异增大而误差变异度不变或变化很少,则真实值变异所占比例必然增大,从而使信度增高。但这种提高的方法在一些情况下是不合适的:例如,将 5、10、15 岁年龄组的能力测试得分合并,并只算出一个信度系数,由于不同年龄组的能力水平明显不同,导致得分范围扩大,该信度必然较高,但当该量表应用其中特定年龄组人群时,其实际信度并不会

这么高。另外，在某量表测试结果中，若其中绝大多数被测者的得分要么都很高，要么都很低，以致于平均得分接近于最大可能值（ceiling）或最小可能值（floor）时，应设法更换条目或改动条目的应答方式，使平均得分接近中间值，从而提高信度[6]。

(5)样本容量　在其他条件不变的情况下，样本容量越大，估计出的信度越准确。假设研究前给定量表的信度为 R，总体信度可信区间的一半宽度为 CI，用正式估计相应的所需样本容量 n，

$$n = \left[\frac{Za/2}{Z(R) - Z(R + CI_H)}\right]^2 + 3$$

式中 Z 为 Fisherz 变换（Fisher's z transformation），

$$Z(R) = \frac{1}{2} ln \frac{1+R}{1-R}$$

影响信度的因素可出现于量表的设计、测试、分析和样本的，抽样等各个环节，实际工作中需要尽量考虑全面，并据具体情况运用合适的信度评价方法。值得注意的是，不同的人群使用同一份已被证实具有较高信度（和效度）的量表，其结果仍可能有所不同。所以，我们不能单说某个量表的信度合适与否。信度的高低只有结合特定的人群才有意义，正如爱因斯坦谈到时间一样，信度也是相对的。

参　考　文　献

1 Aiken LR. Questionnaire and inventories. surveying opinions and assessing personality. New York: John Wiley &Sons. 1997

2 Cox DR, Fletcher AC, Gore SM, et al. Quality of life assessment: can we keep it simple? J R Statist Soc A, 1992, 155: 353－393

3 Greco LD, Wolop W and McCarthy RH. Questionnaire development : validity and reliability. CMAJ 1987, 136: 699－700

4 Friedenberg L. Psychological testing : design , analysis and use. Boston : Allyn and Bacon, 1995

5 陈平雁，黄浙明. 病人满意度的调查与分析. 中国医院管理，1999，19(7)：19－22

6 Streiner DL and Norman GR. Health measurement scales: a practical guide to their development and use . Oxford, New York, ToKyo, Oxford University Press, 1995

7 Guttman L. A basis for analyzing test－retest reliability. Psychometrika, 1945, 10: 255－282

8 Patrick E. Shrout and Thomas J. Yager. Reliability and validity of screening scales: effect of reducing scale length. J Clin Epidemiol, 1989, 42(1): 69－78

9 Waltz CF. Measurement in nursing research. Philadelphia: F. A. Davis Company, 1991, 143

10 金丕焕主编. 医用统计方法，第 2 版，上海：上海医科大学出版社，1993，281

11 巫秀美，倪宗瓒. 因子分析在问卷调查中信度效度评价的应用. 中国慢性病预防与控制，1998，6(1)：28

12 McDowell I and Newell C. Measuring health: a guide to rating scales and questionaire. New York , Oxford University Press, 1991. 112

13 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep, 1966, 19: 3

14 Morton AP and Dobson AJ. Assessing agreement. Med J Aust 1989, 150: 384

15 左积乾主编. 医学统计学与电脑实验. 第一版. 上海：上海科学技术出版社，1997，245

16 颜文伟. 检验一致性的统计方法. 中华神经精神科杂志，1986，19(6)：367

17 Fless JL. Statistical methods for rates and proportions. USA: John Wiley &Sons. Inc, 1981. 225－234

18 Cohen J. Weighted kappa: nominal scale agreement with provision for health measurement or partial credit. Psychological Bulletin, 1968, 70: 213－20

19 Nagata C, Ido M, Shimizu H, et al. Chocie of response scale for health measurement : comparison of 4, 5 and 7－point scales and visual analog scale. Journal of Epidemiology, 1996, 6(4): 192－7

（收稿日期：2000－08－29）

（上接第 314 页）

10 Belsky J. The effects of infant day care reconsidered. Early Childhood Reaearch Quarterly, 1988, 3: 235－272

11 Domingo M Keppley S, Chambliss C. Relations of early maternal employment and attachment in introvertive and exreovertive adult. Psychological Reports, 1997, 81: 403＝410

12 Cohn DA, Cowan PA, Cowan CP, et al. Mothers' and fathers' working models of childhood attachment relationships, parents styles and child behavior. Development and Psychopathology, 1992, 4(3): 417－431

13 Hill EM, Young JP, Nord JL. Childhood adversity attachment security and adult relationships: A preliminary study. Ethology and Sociobiology, 1994, 15(5－6): 323－338

14 Davies PT, Cummings EM. Exploring children's emotional security as a mediator of the link between marital relations and child adjustment. Child Development, 1998, 69(1): 124－139

15 依田明（日）. 家庭关系心理学. 天津：天津人民出版社. 1987. 68－87

16 Cook WJ. Understanding attachment security in family context. Journal of Personality and Social Psychology, 2000, 78(2): 285－294

17 刘　芳，孟昭兰，胡　平. 母婴依恋类型及其在社会参照作用上的差异. 中国儿童发展，1993(4)：45－48

（收稿日期：2001－04－25）