

第七章 研究工具性能的测定



学习目标

通过本章内容的学习，学生能够：

◆ 识记

1. 陈述研究工具的信度和效度的概念。
2. 陈述列举研究工具的信度和效度的常见类型。

◆ 理解

1. 分析研究工具的信度和效度的高低。
2. 总结信度和效度的计算方法。
3. 说明国外量表的翻译和应用的基本过程。

◆ 运用

根据研究工具的特点与所测概念的性质对研究工具的性能进行评价，并在研究报告中报道信度和效度水平。

案例 7-1A

随着血液透析技术水平的不断提高，维持性血液透析患者的生存时间也得到了显著的延长，但与慢性肾衰竭和血液透析本身相关的近期和远期并发症严重影响着患者的整体健康状况。尤其是随着透析年限的增加，患者会出现诸多并发症，如贫血、心功能衰竭、营养不良、钙磷代谢紊乱、高血压、关节淀粉样变、神经系统病变、皮肤瘙痒等躯体症状。因此，血液透析患者不仅需要遵医嘱规律透析，而良好的自我管理行为（包括严格限制液体和饮食中钾、钠和磷的摄入，正确按照医嘱服用药物，以及养成良好的生活方式或生活习惯）对于患者保持良好健康状况和提高生活质量同等重要。

某研究者欲通过研究评价血液透析患者的自我管理行为水平以及患者自我管理行为水平与其健康状况的相关性，而且在寻找研究工具的过程中发现国内学者王爱平研制了血液透析患者自我管理行为问卷。该研究者是否可以直接使用该问卷？如何判断这种测量工具的质量高低？

测量是在资料收集环节根据某一法则赋予某个指标、事件或状态具体数值的过程。而研究工具（instruments）是在科学理论的基础上应用具体标准或法则发展而成的测量工具。对于研究中的具体或抽象指标，我们采用相应的直接或间接的测量方法。例如，对身高、体重、心率、血压等生理指标进行直接测量。当某一抽象概念（如态度、观念等）不能够直接测量时，则需对可以解释此概念的某些指标或属性进行测量。这种情况下，护理研究中最常用的研究工具为问卷或量表。根据研究的目的和内容，研究者可以自行编制问卷或量表，也可以引用公认





的或成熟的工具。我们知道低质量的研究工具会影响所收集资料的准确程度和可靠性,进而威胁到整个研究的科学性,因此在开展研究之前应该对研究中的每个研究工具进行质量评价。而且需要注意的是,将以往研究中报道的高质量的研究工具应用于新的研究时也应重新评价其质量,因为每一份问卷或量表均是原作者根据特定的研究目的、针对特定的人群而编制的,不一定完全适用于新的研究目的和研究人群。

信度和效度是用来评价研究工具质量的两个指标,同时具备高信度和高效度的研究工具是良好科研的必备条件。本章主要介绍研究工具的信度、效度的概念和计算方法,以及国外量表的应用过程。

第一节 信 度

一、信度的概念

信度(reliability)是指使用某研究工具所获得结果的一致程度,即测量工具能否稳定地测量到所测的属性或变量;同时,信度也是反映测量结果受到随机误差影响程度的指标。当使用同一研究工具重复测量某一组研究对象时所得结果的一致程度越高,则多次测量的随机误差越小、该工具的信度越高,也就越能准确地反映研究对象的真实情况。例如,使用同一体温计、在同一环境下、应用相同的测量方法每隔5分钟对同一组受试者的体温进行重复测量,共测量10次。倘若每次测量的结果均有变化,即有时比前一次的数值高,有时低于前一次测量值,那么你就会质疑该体温计的信度了。

二、信度的测定方法

信度检验集中在研究工具三个方面的特征:稳定性、内在一致性与等同性。量表的不同特征对应着不同的信度计算方法。具体选择哪种特征来描述研究工具的信度主要取决于研究工具的特性和研究者所关注的研究工具的特征。我们在评价信度的高低时常采用相关系数的形式来表达。信度系数取值范围在0~1,其值愈趋近于1,信度愈高,即表示该测量结果愈一致、稳定与可靠。

(一) 研究工具的稳定性

1. 定义

(1) 稳定性(stability):是指使用同一研究工具在不同时间对同一组研究对象进行重复测量,所得结果间的一致程度。常用重测信度来衡量研究工具的稳定性。

(2) 重测信度(test-retest reliability):又称稳定性系数,即采用重测法,使用同一研究工具在同样条件下对同一组受测者前后施测两次,求两次结果间的相关程度。重测信度越高,测量结果越一致,这也表明受测环境中日常随机因素的影响越小。

2. 计算方法 重测信度可以通过公式或使用统计学软件如SPSS进行计算。当测量的变量是分类变量时,可用Cohen's Kappa系数来评价重测信度。当变量为等级资料时,可采用Spearman秩相关分析法计算相关系数。当变量是连续变量时,多采用Pearson积差相关分析法,所得的Pearson相关系数(r)即为重测信度。目前也有人推荐对于连续变量或等级变量,可使用基于方差分析的内部相关系数(intraclass correlation coefficient, ICC)来评价工具的重测信度。大部分信度的计算也可通过SPSS统计分析软件中Analyze分析菜单下的Reliability Analysis信度分析模块来实现,其调用过程见表7-1。





表7-1 SPSS软件信度分析的调用过程表

信度的测量方法	调用过程	分析结果
重测信度	Analyze → Scale → Reliability Analysis → Intraclass correlation coefficient 选项 Analyze → Descriptive Statistics → Crosstabs → Kappa 选项	内部相关系数 ICC (测量变量为连续变量 或等级变量) Kappa 系数 (测量变量为分类变量)
Cronbach's α	Analyze → Scale → Reliability Analysis → Model 中 Alpha 选项	Cronbach's α 系数
折半信度	Analyze → Scale → Reliability Analysis → Model 中 Split-half 选项	折半信度系数
复本信度 / 评定者 间信度	Analyze → Correlation → Bivariate → 根据 需要可选择 Pearson、 Kendall's tau-b 和 Spearman 选项	Pearson 积差相关系数、Kendall 等级相关 系数和 Spearman 秩相关系数 (适用于两 列资料)



知识链接

Kappa 系数：是由 Cohen (1960) 提出的一种判断一致性程度的指标。当评估的变量是分类变量时，可用 Kappa 系数来评估重测信度。Kappa 值介于 0 ~ 1 之间，数值越大，信度越高。一般认为， $Kappa \geq 0.75$ ，信度良好； $0.75 > Kappa \geq 0.4$ ，信度一般； $Kappa < 0.4$ ，信度较差，需考虑修改或删除项目。Kappa 系数可通过公式计算： $k = (P_a - P_e) / (1 - P_e)$ 。

其中， P_a 是两次测量的实际一致率， P_e 是两次测量的期望一致率。

Kappa 系数的计算也可通过 SPSS 软件实现 (表 7-1)。



知识链接

内部相关系数 (Intraclass Correlation Coefficient, ICC) 最先由 Bartko 于 1966 年用于测量和评价可信度的大小。ICC 等于个体变异度除以总变异度，即：

$$ICC = (MS_{\text{区组}} - MS_{\text{误差}}) / [MS_{\text{区组}} + (m-1)MS_{\text{误差}} + m(MS_{\text{处理}} - MS_{\text{误差}}) / n]$$

其中 $MS_{\text{区组}}$ 、 $MS_{\text{处理}}$ 和 $MS_{\text{误差}}$ 分别是随机区组 (被观察对象) 间的方差、处理组 (重复次数间) 的方差和误差的方差； m 是处理组数 (重复次数)； n 是被观察对象的例数。

ICC 的计算也可以通过统计学软件来进行 (表 7-1)。

ICC 既可用于定量资料，又可用于等级资料和分类资料。当 ICC 用于分类资料时， $ICC = Kappa$ 。

ICC 值介于 0 ~ 1 之间，数值越大，信度越高。一般认为， $ICC \geq 0.75$ ，信度良好； $ICC < 0.4$ ，信度较差。对于定量资料常需要更高的 ICC 值。





案例 7-1B

在实施血液透析患者自我管理行为水平的研究前，研究者在学者王爱平研制的血液透析患者自我管理行为问卷（65 个条目）的基础上经专家反复审阅和指正、*t* 检验法和相关系数法调整和修改以及因子分析方差最大旋转法，最后形成 25 个条目的终量表即血液透析患者自我管理行为评定量表。在研究中研究者再次通过预实验对该量表的重测信度进行评价。

使用量表对 10 名维持性血液透析患者进行重复测试，前后两次测试的时间间隔为 2 周，第一次得分为 *X*，第二次得分为 *Y*（见表 7-2），求其重测信度。

表 7-2 10 名维持性血液透析患者的两次测试得分

研究对象	1	2	3	4	5	6	7	8	9	10
<i>X</i>	70	65	78	80	66	65	91	77	85	71
<i>Y</i>	68	63	82	72	71	70	88	73	90	66

由于测试得分属于计量资料，所以使用 Pearson 积差相关分析法计算公式：

$$r = \frac{n \sum XY - \sum X \cdot \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2] \cdot [n \sum Y^2 - (\sum Y)^2]}}$$

n：样本例数

$\sum X$ ：第一次测试 *n* 名研究对象的得分之和

$\sum Y$ ：第二次测试 *n* 名研究对象的得分之和

$\sum XY$ ：*n* 名研究对象各自第一次与第二次测试得分乘积后的和

$\sum X^2$ ：第一次测试 *n* 名研究对象各得分平方之和

$\sum Y^2$ ：第二次测试 *n* 名研究对象各得分平方之和

由公式可计算得出 $r=0.86$ ，即量表在该研究中的重测信度系数为 0.86。

3. 使用过程中的注意事项 重测信度常用于物理或技术测量工具以及纸笔版量表的信度评价。通常在使用前需做出假设：两次测量时所测因素是保持不变的，两次测量结果的任何差异是由随机误差导致的。这就要求在使用重测信度时注意以下几点：

（1）所测变量的性质必须是稳定的：护理研究中的很多现象或研究因素如希望、应对、焦虑或健康行为等变量，可能会在较短时间内发生改变，因此重测后两次结果的变化主要是这些因素确实发生了变化而非工具不稳定所引起；所以这种情况不适合评价其重测信度。而对于性质相对稳定的问题，如个性、价值观、自尊、生存质量、成人的身高等变量，可用重测信度来评价研究工具的信度。

（2）两次测试的间隔时间应合适：在评价物理测量工具或仪器的重测信度时，可以在第一次测量之后立即行第二次测量，或者可以在使用仪器一段时间后进行再测量，均可获得较客观的重测信度，而且后者还可以用以确定校准频率。但使用如上方法对纸笔版量表进行重测信度的检验时并未达到起初设想的效果，其原因可能是：此过程存在“记忆效应”，即重测时受试者可能记得上次的回答而高估了重测信度；或者第一次的测量引起了受试者对某些测量项目的反思，在重测时改变对相应项目的回答，进而低估了重测信度。因此，对于纸笔版量表，确定间隔时间的原则是时间的间隔要足够长，以使第一次测量对第二次测量结果不会产生影响，但也不能过长，以免客观情况发生改变；通常建议间隔时间为 2 ~ 4 周。

（3）两次施测的间隔期内被试者应没有接受过与所测特质 / 变量相关的学习和训练。





(4) 测量环境应一致：在两次测量时应尽量保证测量环境相同，如相同的测试者、测量程序、测量时间以及相似的周围环境等，以减少外变量的干扰。

(二) 研究工具的内在一致性

1. 定义 内在一致性 (internal consistency) 是指组成量表 / 问卷的各项项目之间的同质性或内在相关性，主要反映的是项目之间的关系。内在一致性越高，同质性就越好，说明组成研究工具的各项项目都在一致地测量同一个属性或指标。例如某问卷是用于测量护士对患者的同情心情况，如果组成这个问卷的所有项目均与同情心有关，则说明此问卷的内在一致性较好，信度高；如果其中有一项或几项是用来测量护士对患者病情的评估能力的，则此问卷的内在一致性就差，信度就低。内在一致性的测量是信度测量中应用最多的一种，一方面与重测信度相比它只需进行一次测量，经济且方便；另一方面它更适用于心理社会学方面指标的测量。

2. 计算方法 折半信度、Cronbach's α 系数与 KR-20 值常用来衡量研究工具的内在一致性。

(1) 折半信度 (split-half reliability)：是测定内在一致性最古老的方法之一。具体做法是将组成研究工具的项目 (即“组成量表 / 问卷的条目”) 分成两部分 (常用前后折半法、奇偶折半法或完全随机法)，对某一人群测量后分别加以计分，对这两部分的得分进行相关分析，计算 Pearson 相关系数 (r)。但这样只反映了两部分的相关程度，整个测量的实际信度需要用 Spearman-Brown 公式校正后得出：

$$r_{xx} = 2r_{hh} / (1 + r_{hh})$$

其中， r_{xx} 代表研究工具的折半信度， r_{hh} 代表两折半组间的相关系数。

在采用折半信度分析时，两半的题项数目尽可能相等。折半信度虽然比较经济和简便，但存在着内在的不足之处：首先，它反映的只是两半测试项目之间的一致性；再者，不同的折半方法会导致不同的结果，使得折半信度的计算带有一定的随机性。

(2) 克朗巴哈系数 (Cronbach's α 系数) 与 KR-20 值：是通过计算研究工具中所有项目间的平均相关程度，以反映该工具的内在一致性，它们相当于所有可能组合的折半法信度系数的平均值，因此避免了折半信度计算的缺点。Cronbach's α 系数是目前最常用的信度系数，适用于测验结果以多值计分的情形。很多量表 / 问卷测量的内容包括几个领域或维度，宜分别计算 Cronbach's α 系数。

$$\text{Cronbach's } \alpha \text{ 系数的计算公式为: } \alpha = \frac{k}{k-1} \left(1 - \frac{\sum S_i^2}{S_t^2} \right)$$

其中， k 为工具的条目数， S_i^2 为每一项目得分的条目内方差， S_t^2 为全部条目测试总得分的方差。

KR-20 值是 Cronbach's α 系数的一种特殊形式，适用于回答“是”或“否”、“对”或“错”等的二分制答案的研究工具。

Cronbach's α 系数与 KR-20 值的计算较为复杂，常通过统计学软件如 SPSS 来进行。而且当问卷条目的应答方式是二分类时，Cronbach's α 系数即相当于 KR-20 值；即仅通过“Reliability model”中“Alpha”选项就可以实现 Cronbach's α 系数与 KR-20 值的计算，具体调用过程见表 7-1。

(三) 研究工具的等同性

1. 定义 等同性 (equivalence) 是指不同观察者使用相同工具同时测量相同对象，或两个大致相同的研究工具同时测量同一对象时所得结果的一致程度。评定者间信度和复本信度均可以用来表示研究工具的等同性这一特征。

2. 计算方法

(1) 评定者间信度 (interrater reliability)：是指不同评定者使用相同工具，同时测量相同





对象时,不同评定者间所得结果的一致程度。一致程度越高,则该工具的等同性越好,信度越高。如使用观察法收集资料时,不同观察者使用同一评定工具进行观察时会产生观察者偏倚,这种情况下研究工具应包括所使用的评定工具及观察者。因此,当研究过程中存在不同资料收集者使用评定工具进行收集资料时,有必要评估资料收集者间评定结果的一致性,即评定者间信度。

当评定者为两人时,根据评定结果的资料类型的不同,评定者间信度可通过 Pearson 积差相关系数(连续变量)、Spearman 秩相关系数(连续等级变量)或 Kendall 等级相关系数(两个分类变量均为有序分类)表示。如果评定者在两人以上,需采用等级记分(其他形式的资料需转化为等级资料),然后采用 Kendall 相关系数来评价评定者间信度。评定者间信度的数值介于 0 ~ 1 之间,数值越趋近于 1 说明研究工具的评定者间信度越高。一般认为评定者间信度至少应达到 0.6,当 ≥ 0.75 时认为该工具的评定者间信度非常好。



知识链接

Kendall 等级相关系数又称和谐系数,是表示多列等级变量相关程度的一种方法。这种资料的获得一般采用等级评定的方法。例如,让 K 个临床带教(评价者)依据某一评价标准对 N 名实习护生的评判性思维能力进行评定,每个评价者都能对 N 名护生能力的高低排出一个等级顺序,最小的等级序数为 1,最大的为 N。这样, K 个评价者便可得到 K 列从 1 至 N 的等级变量资料。为评价评判性思维能力评定表的评定者间信度,将这类 K 列等级变量资料综合起来求相关,就可用 Kendall 相关系数。Kendall 相关系数也常通过统计软件计算,其 SPSS 调用过程见表 7-1。

(2) 复本信度(copy of reliability):是指采用两个或两个以上大致相同的研究工具/复本同时测量同一研究对象所得结果的一致程度。复本信度反映的是测验形式的一致性程度,此方法要求两个复本除表述方式不同外,在内容、格式、难度和对应题项的提问方向等方面要完全一致。复本信度主要用于标准的知识性测试工具的开发与评价,在临床研究中较少被使用。当设计方案中有重复测量这一环节时,使用复本工具可以改善研究质量。评估复本信度要用两个复本对同一群受试者进行测量,然后计算两种复本测量结果的相关系数。复本信度系数的计算方法与评定者间信度相同。测定复本信度时需要考虑复本实施的时间间隔,一般来说,复本应几乎在同一时间实施,以剔除时间的影响。在实际操作时,可让受试者连续进行作答这两份测量问卷或试题,但两份问卷或试题被回答的先后顺序是随机确定的,然后再计算出两个测试结果的相关系数,即复本信度系数。

三、信度的评价与报告

(一) 信度的评价

1. 评价过程中样本量的选择 量表的信效度评价常用于量表编制过程中筛选条目、根据预试验的评价结果修改与完善量表,以及正式研究前通过预试验评价研究工具的质量。在发展量表或进行条目筛选时,所需的样本量为条目数的 5 ~ 10 倍。本章主要介绍通过预试验评价研究工具的质量。在进行预试验时,研究者可以选取 10 ~ 20 例样本或总样本量的 10% 进行信度测试。

2. 信度高低的判断 信度用信度系数来表示,信度系数越大,表明测量的可信程度越高。但目前尚无统一的信度标准。学者 DeVellis (1991) 认为,0.60 ~ 0.65 (最好不要); 0.65 ~ 0.70





(最低可接受值)；0.70 ~ 0.80 (相当好)；0.80 ~ 0.90 (非常好)。因此，一般认为对于一个已发展成熟的心理社会学测量工具，在新的研究中其信度值至少应达到 0.80；如果是生理学测量工具，所要求的信度值会更高。而对于一个新研制的心理社会学工具，信度值达到 0.70 即可接受。目前护理学研究中很多成熟的量表为多维度的，通常情况下各维度的信度系数最好在 0.70 以上，0.65 ~ 0.70 之间尚可以接受。若总量表或维度的信度不理想时，则应考虑重新修订量表并再次进行信效度评价。

(二) 信度的报告

1. 评价研究工具的质量时，一定要同时报道出其信度与效度水平。

2. 介绍研究工具的信度时，应该报告出研究工具的信度数值以及相应的计算方法，从而为其他研究者判断研究工具的适用性提供参考。例如在研究报告中明确说明“学龄前儿童饮食行为量表的 Cronbach's $\alpha=0.85$ ”。

3. 每一种信度只是反映调查研究中所产生误差的一部分或某个方面，并不代表该调查研究全部的可靠性。但并非所有的研究工具均需报告出其稳定性、内在一致性和等同性。研究者应根据工具的特点以及所测变量的性质来决定。例如对于所测变量相对稳定（如人格）的工具，研究者可同时报告出其重测信度与 Cronbach's α 系数。而当所测变量不稳定（如健康行为、满意度等）时，不适合报告工具的重测信度，此时可以报告该工具的 Cronbach's α 系数。如果在研究设计中有不同的研究者参与资料收集的环节，则应报告研究工具的评定者间信度。

第二节 效 度

案例 7-2A

乙型肝炎病毒感染相关歧视态度严重影响了 HBV 感染者的物质和精神生活质量、用药依从性、自愿咨询与检测，以及获取卫生服务的意愿和能力。因目前尚未发现全面、客观地评定他人对 HBV 感染者歧视态度的研究工具，某研究者在阐释歧视态度的概念并广泛阅读文献的基础上研制了乙型肝炎病毒感染相关歧视态度量表。那么，如何判断该量表能否达到测量的目的？或者说如何判断该量表能否有效衡量他人对 HBV 感染者的歧视态度？

一、效度的概念

(一) 定义

效度 (validity) 是指某一研究工具能够测量出研究者想要测量的概念或特质的程度。它反映测量结果的有效性或正确性。效度越高表示研究工具的测量结果越能代表所要测量对象的真正特征。如研究者欲探讨自我效能训练对提高血液透析患者自我管理水平的效果，研究者在正式研究之前就应确保测量工具“自我管理行为评定表”所测的结果能有效而准确地反映血液透析患者自我管理行为的水平。

(二) 信度与效度的关系

信度主要反映研究工具的测量结果的一致性、稳定性或可靠性，而不涉及结果是否有效或能否正确测量到所测的特质；效度则针对工具的目的，重点考察测量结果的有效性。它们





之间的差别在于所涉及的误差不同,信度衡量的是随机误差的影响,效度是反映由于测量了与测量目的无关的变量所引起的系统误差。对研究工具而言,效度是其首要条件,而信度是效度的必要条件,有效的工具必是可信的,但可信的工具未必是有效的。因此,研究者在选择研究工具时,既要评价该工具的信度,也要考察其效度,这样才能完整地评定研究工具的质量。

二、效度的测定方法

根据研究工具测量的目的与角度,可以通过内容效度、效标关联效度、结构效度等来评价研究工具的效度。

(一) 内容效度

1. 定义 内容效度(content validity)是指一个研究工具实际测到的内容与所要测量的内容之间的吻合程度,即量表/问卷能否反映所要测量的特质、能否达到测量的目的或要求。内容效度是根据理论基础及实践经验来对研究工具是否包括有足够的项目,而且有合适的内容分配比例而作出判断。常用来评价量表/问卷内容效度的方法是专家评议法,最广泛的指标是内容效度指数。

2. 评价过程与计算方法

(1) 专家评议:请3~10位(一般为5位)对研究工具所测定或涉及的领域熟悉的专家组成专家组。如研究者编制了含12个条目的乙型肝炎病毒感染相关歧视态度量表,所请的专家应对乙型肝炎病毒感染的医学知识、歧视态度概念较为熟悉,同时也应包括一位在工具构建和心理测量学方面有丰富经验的专家。专家组对量表条目与原定内容范围的吻合程度(相关性)及代表性作出判断。可以采用表格的形式对专家的意见进行整理。

(2) 计算内容效度指数(content validity index, CVI):

内容效度指数分为两类:①条目水平的内容效度指数(item-level CVI, I-CVI),即对各个条目的内容效度作出评价;②量表水平的内容效度指数(scale-level CVI, S-CVI),即对整个量表的内容效度进行评估。

I-CVI的计算方法:在内容效度评价的专家咨询问卷中要求专家就每一条目与相应内容维度的关联性或代表性作出评价。通常,可采用4等级评分:1=不相关,2=弱相关,3=较强相关,4=非常相关。这样,就每一条目,给出评分为3或4(反映关联性较好)的专家人数除以参评的专家总数即为相应条目的I-CVI。

量表所有条目I-CVI的平均值即为S-CVI,可表示为S-CVI/Ave。

(3) 内容效度的评价与再评价:

当I-CVI ≥ 0.78 , S-CVI/Ave ≥ 0.90 时,即可认为研究工具的内容效度较好;当I-CVI < 0.78 时,提示研究者需要根据专家意见认真修改、删除或增补条目;在对量表进行修改后进行的第2轮专家评议中,专家人数可以减少为3~5人;然后再次计算I-CVI与S-CVI。

案例 7-2B

在研究者编制了含12个条目的乙型肝炎病毒感染相关歧视态度量表后,为了确定该量表是否有足够的项目反映HBV感染者歧视态度,以及其内容分配比例是否合适,该研究者请A、B、C、D和E共5位专家对其内容效度进行评价,结果见表7-3。





表7-3 12条目乙型病毒性肝炎相关歧视态度量表的内容效度指数

条目	专家评分					评分为3或4 的专家人数	I-CVI
	A	B	C	D	E		
1	4	4	4	4	4	5	1.00
2	4	4	3	4	4	5	1.00
3	4	3	3	4	3	5	1.00
4	4	2	3	4	4	4	0.80
5	4	3	3	4	4	5	1.00
6	4	4	3	4	4	5	1.00
7	4	3	4	4	4	5	1.00
8	2	3	4	4	4	4	0.80
9	4	4	4	4	4	5	1.00
10	2	3	3	2	4	3	0.60
11	4	4	4	4	3	5	1.00
12	4	4	3	3	3	5	1.00

由以上结果可见,该量表中11个条目的I-CVI大于0.78,提示这些条目的内容效度优秀。条目10只得到5位专家中3位的认可,I-CVI为0.60,提示条目水平的内容效度较低,应考虑进行修改或删除。对各条目I-CVI计算均数可得 $S-CVI/Ave=0.933$,大于0.90,量表水平的内容效度较好。

对量表中条目10进行进一步修改,然后实施第2轮专家评价,专家人数为5人。结果显示:12个条目的I-CVI范围在0.80~1.00, $S-CVI/Ave$ 为0.95,该量表具有较好的内容效度。

(二) 效标关联效度

效标关联效度(criterion-related validity)主要反映的是研究工具与其他测量标准之间的相关程度,但未体现研究工具与其所测概念的相符程度。相关系数越高,说明工具的效度越高。根据使用效标测量的时间,可以将效标关联效度分为同时效度(concurrent validity)和预测效度(predictive validity)。

1. 同时效度 同时效度是指研究工具与现有标准之间的相关程度。例如,在“血液透析患者摄水依从性现状分析”的课题中使用了“血液透析患者摄水依从性量表”作为研究工具,一方面测定研究对象的依从性水平,另一方面研究者可以根据量表中各条目的得分情况对血液透析患者依从行为或导致依从性较差的原因进行进一步分析。在评价该研究工具的效度时可以对量表得分与透析间期体重增加百分比(IWG%,临床上常用来反映患者两次透析间期进水量的指标)进行相关性检验,相关系数即为同时效度值。一般认为相关系数 ≥ 0.7 比较理想。

2. 预测效度 预测效度是测量工具作为未来情况预测指标的有效程度。预测效度在医学研究领域常见于发展具有预测功能的评估工具或诊断工具。以“应激控制与应对能力对健康状况的预测效度研究”为例,通过分析目前评估工具(应激控制与应对能力量表)的测量结果与未来数年后测量对象的实际健康水平之间的相关程度,评价该量表的预测效度。

(三) 结构效度

结构效度(construct validity)指研究工具所能衡量到理论上期望的特征的程度,即量表/问卷与其所依据的理论或概念框架间相符合的程度。结构效度的建立比较复杂,目前常用的评





价方法是因子分析 (factor analysis)。

在进行因子分析以前,必须先进行因子分析适合性的评估,以确定所获得的资料是否适合进行因子分析。一般采用 KMO 检验或 Bartlett 球形检验来进行适合性分析,KMO 值(介于 0 ~ 1 之间)越大,越适合于作因子分析,或者当 Bartlett 球形检验的 P 值 < 0.05 时才适合做因子分析。Kaiser (1974) 指出,当 KMO 值 > 0.9 时,非常适合进行因子分析;当 KMO 值 < 0.6 时,不适合进行因子分析。

因子分析可以检验量表/问卷不同条目间的关系,相关性比较密切的几个变量归于同一类,代表了一个基本结构,即构成一个因子。因此,因子分析的目的是想了解属于相同概念的不同问卷项目是否如理论预测的那样集中在同一公共因子中。通过因子分析可得到若干个因子,其意义类似于组成“结构”的领域。所得因子的因子负荷反映了条目对领域的贡献,因子负荷值越大说明与领域的关系越密切。良好的结构效度应满足以下标准:

(1) 因子应与问卷设计时的结构假设的组成领域相符,且因子的累积方差贡献率至少 40% 以上;

(2) 每个条目都应在其中一个因子上有较高负荷值 (> 0.4),负荷值 < 0.4 的条目为拟删除条目;

(3) 因子方差均应 > 0.4,该指标表示每个条目的 40% 以上的方差都可以用因子解释。

因子分析需要借助统计学软件进行,在 SPSS 的调用过程为: Analyze → Data Reduction → Factor,并在“Factor Analysis Desprtive”界面中勾选“KMO and Bartlett’s test of sphericity”,即同时获得因子分析适合性检验的结果。

案例 7-1C

在血液透析患者自我管理水平的研究开展之前,研究者对学者王爱平编制的“自我管理行为问卷(65 个条目)”进行了评价和修改。研究者在广泛阅读相关文献的基础上结合自我管理行为概念及血液透析相关知识,确定了研究工具应涵盖以下方面:饮食与饮水依从行为、自我护理行为(包括服药、内瘘保护、病情监测等)以及维持正常角色与功能(如培养兴趣爱好)。在经专家反复审阅、指正和小样本预试验(50 例)的基础上,通过 t 检验法和相关系数法调整和修改不合适的条目,并应用因子分析方差最大旋转法,删除因子荷载 < 0.4 的条目,最后形成 25 个条目的终量表,即该研究所用的血液透析患者自我管理行为评定量表。而且在研究中再次通过预实验对该评定量表的重测信度进行了评价。

因子分析方差最大旋转法的结果显示: KMO 值 = 0.65,且 Bartlett’s 检验 $P < 0.001$,可认为该问卷适合做因子分析。因子分析结果显示,有 10 个公因子(F1 ~ F10)被提取,累积方差百分比为 69.78%,说明 25 个条目的总方差中有 69.78% 是由这 10 个公因子来决定的。每个条目的因子负荷值在 0.41 ~ 0.89 之间。而且 10 个公因子根据专业解释分别属于三个维度:饮食与饮水依从行为维度(F1、F3、F2,共 13 个条目)、自我护理活动维度(F10、F6、F4、F8,共 7 个条目)与维持正常角色与功能维度(F5、F7、F9,共 5 个条目),与理论构想基本符合(表 7-4)。

表 7-4 自我管理行为评定量表的因子分析结果 (%) (n=50)

因子	饮食与饮水 依从行为维度	自我护理 活动维度	维持正常角色 与功能维度
F1 钠与蛋白质摄入依从行为	11.941		
F3 离子摄入依从行为	8.359		
F2 液体摄入依从行为	9.133		





续表

因子	饮食与摄水 依从行为维度	自我护理 活动维度	维持正常角色 与功能维度
F10 服药依从行为		4.214	
F6 内瘘保护行为		5.916	
F4 自我病情监测		7.296	
F8 知识寻求行为		5.601	
F5 形成良好的生活方式			6.548
F7 维持身心健康			5.710
F9 培养兴趣爱好			5.509
解释变量量	29.433	23.027	17.317
累积变量量	29.433	52.460	69.777

第三节 国外量表的应用过程

量表是用来量化主观或抽象概念的一种测量工具,在护理科研、教育、管理及临床工作中发挥着越来越重要的作用,为收集可靠资料、准确评价干预效果提供了科学依据和参考。但国内护理界对量表的研制工作开展较晚,量表相对缺乏,而国外的量表发展相对比较成熟,因此,国内护理研究者更愿意借鉴国外发展成熟的量表对概念进行测量,不仅省时、节力,而且在引入过程中的修改和调适可以使研究工具更加可靠、有效。此外,多中心临床试验的广泛开展也推动着护理研究的国际化,科学地引进和使用国外量表受到越来越多护理研究者及学者的重视。在量表的引进过程中不仅要严格遵循一定的量表翻译规则,还应考虑到社会文化差异等因素,而需要对量表进行跨文化调适,并进行性能的评价,以保证目标量表和源量表的等同性。一般来说,国外量表的翻译可以按照以下步骤进行。

一、翻译

在翻译之前需要与量表原作者联系,取得其同意后方可翻译和使用量表。翻译是将源语言量表翻译为目标量表的过程,包括正向翻译与回译。典型的翻译模型是由 Brislin 于 1970 年提出的,翻译过程如图 7-1 所示,翻译终止的标志是翻译人员认可目标语言版本可以准确地传递源量表的信息。主要有两种翻译方式:一人法和双人法。本文以常使用的双人翻译回译的方法为例,介绍翻译的基本过程。

1. 正向翻译 即将源语言量表翻译为目标语言(中文)量表。由母语是目标语言的两名双语翻译者独立地进行单向翻译,即翻译为中文,得到两个翻译版本(A1 和 A2)。要求这两名译者必须有不同的经历和文化背景。一名译者能够理解量表所测量的概念,可以从临床和量表测量学两个角度保证量表的对等性。另一名译者被称为单纯的翻译者,最好没有医学背景,也不具备测量学知识,但语言及翻译水平高于前者,即更能准确理解源量表条目的意思,对语言能够字斟句酌,准确识别语言表达的细微差别。



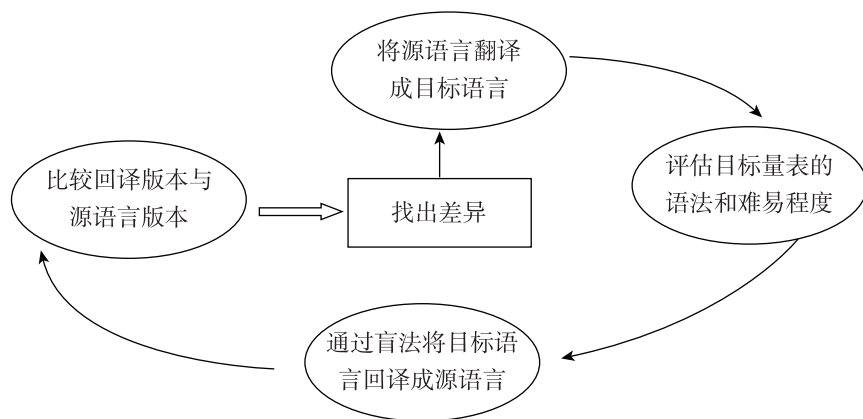


图 7-1 Brislin 翻译模型

2. 比较两个翻译版本 (A1 和 A2), 形成综合版本 (A3) 第三名会双语且有双文化背景的、有一定评判能力的人员对两个版本 (A1 和 A2) 进行比较。对于不一致之处或存在模棱两可的表达时, 需要与前两名译者以及研究人员一起通过讨论解决。最后将翻译合并, 形成目标语言的综合版本 (A3)。例如在田庆秀等人翻译引进“支气管哮喘用药依从性量表”的过程中, 两名译者在条目 “I take it less than instructed” 的翻译上存在不一致, 译为“我服药会少于指示用药”较难理解, 不符合中文表述习惯; 因此改为“我吸入的药比医生要求的少”, 即在源量表的基础上采取了意译。

3. 由另外两名译者将目标语言综合版本 (A3) 回译成源语言量表 (B1 和 B2) 这两名译者的母语最好是源语言, 而且均未接触过源量表, 也不知道其在进行量表的回译过程。这两名译者的特点类似于正向翻译的前两名译者的特点。

4. 比较两个回译量表 (B1 和 B2) 同样由另一名双语人员比较、组织讨论, 确定最终的回译的源语言综合版本 (B3)。

5. 组织专家委员会, 对源量表与源语言综合版本 (B3) 的等同性进行考察, 并得到终稿前的目标量表的版本 (C)。专家委员会的组成成员应包括: 方法学专家、卫生保健专家、语言学专家和所有翻译者 (包括参与正向翻译、回译和综合的翻译者)。该过程需要提交量表选择和翻译过程中的所有材料, 专家根据问题提出建议, 统一翻译过程中的分歧意见, 审核量表的指导语, 确定翻译的量表条目是否准确反映了所测概念, 是否具有可读性 (小学 6 年级水平); 还应详细记录专家意见, 并请所有专家对最终版本进行评分, 用以评估翻译效度和内容效度。另外, 还需根据现有人群特点结合源量表规则统一评分标准和操作标准, 指导量表的应用。

二、量表的文化调适与预试验

为了使引入的量表适用于目标文化背景, 需要在翻译的基础上对某些条目进行修订, 使之尽量符合目标语言的习惯用法。例如, 在欧美国家, “nursing manager” 和 “nursing executive” 分别对应于我国的 “护士长” 和 “护理部主任”, 如果直译为 “护理管理者” 和 “护理首席执行官” 就会造成费解。另外, 中国台湾和香港汉语书面语使用繁体字, 内地使用简化字; 而且三个地区的措辞、专业术语、名称均有差异, 所以在引入量表时同样要进行文化调适。

量表的文化调适可以通过专家委员会讨论和预试验进行, 同时这两个过程也分别从专家和研究对象的角度评价了量表的效度。预试验是将终稿前的目标量表版本 (C) 在目标语言背景下对将要适用的目标人群进行试验性的测试, 也就是从研究人群的角度评估量表能否准确反映所测的概念。例如在 “哮喘患儿生活质量量表” 的引入中, 预试验选择的样本就应该为哮喘儿





童。在测量后还可以采用访谈的方法了解受试者对量表内容的理解程度、填写量表的感受,并记录填写时间。预试验的样本量一般在 10 ~ 40 人。依据预试验的结果,对版本 C 进行修订及文化调适,形成暂定的最终版本(D)。

三、测量源量表与中文版量表的等同性

测量源量表与中文版量表的等同性需要在一定数量的双语样本(既懂源语言又懂目标语言的受试者)中进行。让这些双语受试者对两种语言版本的量表进行作答,然后比较两种量表所得总分之间的相关性以及各条目得分的相关性。相关程度越高,说明两版本量表的等同性越好。但是,实际研究工作中获取双语样本的难度较大,也可选取一定数量(样本含量一般为条目数目的 10 倍)的、母语是目标语言的受试者进行预实验,评定中文版本量表的信度和效度,如果与源量表的信效度接近,说明两个版本量表的等同性比较好。

小 结

1. 信度和效度是用来评价研究工具质量的两个指标,同时具备高信度和高效度的研究工具是好科研的必备条件。因此,考察研究工具的质量时,一定要同时报道出其信度与效度水平。
2. 介绍研究工具的信度和效度时,应该报告出研究工具的信度或效度数值以及相应的计算方法或评价方法。
3. 信度是指使用某研究工具所获得结果的一致程度或准确程度。信度检验集中在研究工具在三个方面的特征:稳定性、内在一致性与等同性。重测信度常用来评价研究工具稳定性的大小。可以通过计算 Cronbach's α 系数或 KR-20 值来评价内在一致性。研究工具的等同性常使用评定者间信度来表示。
4. 对于一个已发展成熟的心理社会学测量工具,在新的研究中其信度值至少应达到 0.80;而对于一个新研制的心理社会学工具,信度值达到 0.70 即可。而且量表各维度的信度系数最好在 0.70 以上,0.65 ~ 0.70 之间可以接受。
5. 效度是指某一研究工具能够测量出研究者想要测量的概念或者特质的程度。根据研究工具测量的目的与角度,可以通过内容效度、效标关联效度、结构效度等来评价研究工具的效度。
6. 国外量表的引进或翻译可以按照正向翻译、回译、文化调适以及测量两个版本量表的等同性等几个步骤进行。在量表的引进过程中,应保证量表既适合中国的文化特点,又不偏离原文的意思,同时要确保翻译后量表具有良好的信度与效度。

自 测 题

一、名词解释

1. 重测信度 2. 内在一致性 3. 内容效度

二、填空题

1. 用来反映研究工具质量高低的两个指标是_____和_____。
2. 信度的三个主要特征为稳定性、_____与_____。





3. 一般认为新研制的工具信度应达_____, 广泛使用的工具在新的研究中应达_____。

三、选择题

1. KR-20 值常用来反映研究工具信度的特征是
A. 稳定性
B. 敏感性
C. 等同性
D. 内在一致性
E. 正确性
2. 下列有关信度与效度关系的描述, 错误的是
A. 只要研究工具存在, 就必然有它的信度和效度
B. 信度高的研究工具其效度必然高
C. 信度低的研究工具其效度必然低
D. 效度低的研究工具其信度可能较高
E. 信度是效度的必要条件
3. 下列各项中属于反映工具与其所依据的理论或概念框架的相结合程度的效度是
A. 表面效度
B. 内容效度
C. 同时效度
D. 预测效度
E. 结构效度
4. 在评价评定者间信度时, 当评定结果的资料类型为定量资料时, 可采用的方法为
A. Pearson 积差相关系数
B. Spearman 秩相关系数
C. Cohen's Kappa 系数
D. Kendall 等级相关系数
E. 内部相关系数

四、简答题

1. 试述使用重测信度时的注意事项。
2. 试述内容效度的评价过程与计算方法。
3. 简述将国外量表翻译为中文版本量表的基本步骤。

五、案例题

指出以下论文中有关研究工具质量报告的不妥之处。

论文题目: 糖尿病患者自我管理行为水平的调查研究。

测量工具的报告:

本研究采用的研究工具是“糖尿病患者自我管理行为问卷”。该问卷是根据自我管理行为概念及糖尿病患者的特点, 查阅相关文献而设计的, 内容主要包括饮食、药物治疗、运动锻炼、血糖或尿糖检测、病情控制、糖尿病知识的理解与自评等几个方面, 共 25 个条目。问卷中每个条目的回答分 4 个等级, “1”表示从不这样, “2”表示偶尔这样, “3”表示有时这样, “4”表示总是这样, 计分由 1 分到 4 分。问卷总分范围为 25 ~ 100 分, 总分越高, 提示患者的自我管理行为水平越高。

通过预试验对问卷质量进行评定, 结果如下: 重测信度系数为 0.89; 内容一致性评价结果为 0.65; 评定者间信度为 0.62。说明该问卷质量良好, 可以用于研究。

(余燕朝)

