·流行病与统计方法·

量表条目的统计学筛选方法研究进展

茅范贞[1]，韩耀风[1,2,3]，方亚[1,2,3]
1.厦门大学公共卫生学院卫生经济与政策研究中心，福建 厦门 361102；2.福建省卫生技术评估重点实验室，福建 厦门 361102；
3.分子疫苗学和分子诊断学国家重点实验室，福建 厦门 361102

摘要：量表在现代医学研究中的应用越来越广。对条目池中的条目进行筛选是量表研制的重要过程，而量表条目的统计学筛选方法是条目筛选的重要方法。目前比较常用的统计学筛选方法大都基于经典测量理论、概化理论和条目反应理论，如经典测量理论中的离散趋势法、相关系数法、因子分析法等；概化理论中的概化系数法；条目反应理论中的信息函数和项目功能差异法。本文将对以上三种理论下的量表条目筛选方法进行综述。

关键词：量表；条目筛选；经典测量理论；概化理论；条目反应理论
中图分类号：R195 文献标志码：A 文章编号：1003-8507（2015）01-0001-03

A Review of statistical methods on item selection for scales

MAO Fan-zhen, HAN Yao-feng, FANG Ya
Department School of Public Health, Xiamen University, Xiamen, Fujian 361102, China

Abstract: Scales are more and more extensively used in medical research nowadays. Item selection, which is conducted among an item pool, performed as the critical process of developing a scale. Statistical methods play an important role in item selection. Most of those methods were based on classical test theory (CTT), generalizability theory (GT), and item response theory (IRT) such as discrete tendency, correlation coefficient, and factor analysis from CTT; generalizability coefficient from GT; information function and differential item functioning from IRT. This review takes an overlook at the methods based on the above three theories.

Keywords: Scale; Item selection; Classical Test Theory; Generalizability Theory; Item Response Theory

量表是由若干问题或自我评分指标组成的标准化测定表格，主要用于测量被试者的某种状态、行为或态度[1]。随着医学模式的转变，量表在现代医学研究中的应用愈来愈广，研制合适的量表在现代医学研究中具有越来越重要的意义。量表的基本元素是条目，条目是针对测评特征的某方面提出的问题[1]。条目筛选是量表研制过程中的重要步骤，如何采用科学的方法筛选条目对保证量表的科学性和有效性至关重要。条目筛选包括专业筛选和统计学筛选。前者是从专业的角度考虑量表条目的代表性和重要性；后者是从统计方法的角度分析条目的代表性、独立性、确定性等，目前比较常用的有基于经典测量理论、概化理论和条目反应理论的筛选方法。本文主要针对量表条目的统计学筛选方法进行综述。

1 基于经典测量理论（Classical Test Theory，CTT）的筛选方法

CTT，又称真分数理论（True Score Theory），始于 19 世纪末，是最早形成的测验理论，应用十分广泛[2]。它假设真分数不变和误差完全随机，包括信度、效度、条目分析、常模、标准化等基本概念。基于 CTT 的条目筛选方法主要有：离散趋势法、相关系数法、因子分析法、区分度分析法、逐步回归法、判别式分析法、克朗巴赫 α 系数法、重测信度法等，主要从敏感性、代表性、重要性、独立性、区分性、内部一致性和稳定性的角度筛选条目[3]。

离散趋势和相关系数法较为简单：前者根据变异系数或标准差、后者根据 Pearson 积矩相关系数或 Spearman 等级相关系数等统计量的大小进行条目筛选；因子分析法利用条目在公因子上的载荷来确定，一般以≥0.4 作为入选标准；区分度分析法是根据不同组别如健康组和患病组的某条目得分差异有无统计学意义（如 t 检验），选取能够区分组别的条目；逐步判别分析法与区分度分析法类似，筛选对不同组别鉴别能力有较大贡献的条目；逐步回归法以对总评分（被试者对其目标值进行的总评分）变异贡献的大小筛选条目，剔除贡献小或未进入方程的条目；克朗巴赫 α 系数和重测信度法是信度评价的常用方法，用于条目筛选中是分别从内部一致性和稳定性的方面考虑条目去留，如果某一条目去除后所

基金项目：教育部人文社会科学研究项目资助（12YJA790030）；福建省自然科学基金计划资助项目（2012J01303）；厦门大学基础创新科研基金（CXB2011）
作者简介：茅范贞（1991-），女，在读硕士，研究方向：老年健康
通讯作者：方亚，E-mail：fangya@xmu.edu.cn

属维度的克朗巴赫 α 系数有较大上升，则说明该条目降低该维度的内部一致性，应该去掉，反之则保留；重测信度是计算前后两次测量的相关性，越大越好，一般要求相关系数≥0.7。可见，因子分析法、区分度分析法、克朗巴赫 α 系数和重测信度有较明确的入选标准[4]。但这些方法未见有严格的使用条件，如果从量表的敏感性出发，可考虑使用离散趋势法；如果从代表性和独立性出发，则可考虑相关系数法；如果从量表区别不同特征人群的能力出发，可使用区分度法，等等。

国内有许多医学量表的研制是采用基于 CTT 的方法进行条目筛选的，如生命质量 8 条简明量表、心理健康量表等，而且通常是将相关系数法、因子分析法、区分度分析法等多种方法结合起来进行条目筛选，从较为综合的角度判断条目的去留。这些筛选方法简便易懂，适用性广，用 SPSS 软件即可完成分析，是研究者热衷的分析方法。但 CTT 存在着条目统计量受样本的抽样变动影响大、被试的量表得分依赖于条目难度、平行测验假设无法实现、所有被试的测量标准误很难实现均等等问题[2]。

2 基于概化理论（Generalizability Theory，GT）的筛选方法

GT 是 20 世纪 60 年代由克朗巴赫等人在 CTT 的基础上提出的，它是用方差分析的思想对误差进行分解，控制误差方差以达到提高“信度”的目的。GT 中“信度”用“概化系数”表示[5]。

用 GT 理论研究时分两步进行：即 G 研究（Generalizability Study，拓广研究）阶段和 D 研究（Decision Study，决策研究）阶段[6]。估计出各方差成分相对大小的过程叫 G 研究阶段，在 G 研究阶段的基础上通过实验性研究，进一步考察不同测量条件下（如人的侧面、时间侧面、不同领域的条目）概化系数的变化，从而探索控制误差的最佳方法，做出最佳决策，称为 D 研究阶段。在量表研制中，在总条目数一定的情况下，考查条目池中不同组合模式的概化系数的变化，求得概化系数最大的组合模式，以此达到筛选条目的目的；或者，为探求量表的合适长度，可将条目数量作为测量侧面进行决策研究，分析哪种条目数量下的概化系数最大，以此达到确定最佳条目数的目的[7]。

潘海燕[6]等根据 D 研究尝试不同条目数的概化系数变化，得出慢性病患者生存质量共性模块的最佳条目数。安哲锋[7]等通过运用多元概化理论考察其测评表各因子条目数量与概化系数变化的关系，展示多元概化理论在量表条目筛选和量表编制中的作用。GT 的优点在于放宽 CTT 中的假设，用方差分析的思想分解误差，能够将多种测量条件共同引起的信度变化反映出来，其计算可通过 GENOVA 软件实现。GT 对于分析量表的信度虽有一定优势，但它从 CTT 发展而来，并未消除 CTT 的局限性。

3 基于条目反应理论（Item Response Theory，IRT）的筛选方法

IRT 也称潜在特质理论或潜在特质模型，是针对 CTT 的局限性提出来的，并基于潜在特质、能力单维和局部独立的假设。IRT 认为被试对测验（或量表）条目的反应与他们的潜在特质（即潜变量）有特殊的关系，这种关系用图像表示就是条目特征曲线，对应的函数表达式为条目特征函数。IRT 可用于指导条目筛选和量表编制。

3.1 信息函数[8] 在进行量表编制时，常用信息函数来描述一个测验（量表）或一个条目测量的有效性，包括测验（量表）信息函数和条目信息函数。其含义是从一项测验或一个条目中所得到的最大信息量；其实质是用该测验（量表）或条目去测量被试者的某种潜变量提供的信息。信息函数随被试特质水平变化，受条目自身特性影响，并具有可加性，其平方根的倒数是该点特质水平估计值的标准误[9]。信息函数值越大，反映测验分数对能力估计的精度越高。应用 IRT 中的信息函数进行条目选择的原则是应用最少的条目数，使测验信息函数达到最大。

以信息量大小来筛选条目尚未见有统一的标准。有学者根据特质水平估计值的标准误估计必须达到的测验信息量，并按条目信息量从大到小依次选取，直至信息量累计值刚好达到或超过估计的测验信息量[10]，也有研究在每个领域保留某一确定条目数的前提下保留条目平均信息量最大的条目[9,11]。在进行条目筛选的时候如果有统一的标准说明信息量低于多少时不能保留，将会使条目筛选的结果更可靠。

3.2 条目功能差异（Differential Item Functioning，DIF）[12] DIF 是指具有同一潜在目标特质的两组平行被试组（不同年龄、性别、教育水平、地区等）选择某条目的某一选项的概率不同，即同一条目具有不同的特征曲线。如果某条目有功能差异时，来自两个不同群体但特质相等的人选择该条目同一选项的概率有差异。这不仅与条目特征曲线的原理不符，而且直接影响到对量表效度的评价。因此，保留不存在 DIF 的条目，删除或修改存在 DIF 的条目，对量表的编制和修订具有重要的指导意义。

在 IRT 中，当条目在不同被试下的参数相等时认为该条目不存在 DIF，反之则存在 DIF。有时也不

那么严格限定参数相等，如同一条目在两个亚组中的平均阈值差异大于 0.5，则可认为该条目存在 DIF[13]。在新量表的编制中可以考虑对存在 DIF 的条目进行删除或者重写；而对于一个被证实具有较好信度、效度、反应度的成熟量表，在不能删除或重写的情况下，可以对具有 DIF 的条目进行校正，如有人提出用 IRT 的等值技术来校正。

就 IRT 本身，它为题库建设、量表编制、条目与测验质量的分析评估等提供了新的理论观点和方法工具。它虽然克服了 CTT 和 GT 的传统缺陷，但单维性假定有时难以满足，并且它建立于更复杂的数学模型之上，对测验条件要求严格等，同时，IRT 的模型复杂，计算量大，应用时必须借助专门的软件，如 BILOG、MULTILOG、RUMM 等。例如用 MULTILOG 进行等级反应模型分析来获取条目的信息量、用 RUMM 进行 DIF 检测。可以看出 IRT 对研究者和测验都提出了更高的要求。

4 其他筛选方法

随着多学科的交叉发展，信息学也被应用于条目或因素筛选中，如结合 Filter 过滤式和 Wrapper 容器式的特征选择方法、贝叶斯网络方法、关联规则和 Apriori 算法、Bootstrap 等方法进行筛选。但这些方法计算过程复杂，对研究者提出更高的要求，且在量表研制中的应用还不太成熟。

不同理论下的统计学筛选方法各具优缺点。一般来说，量表条目的统计学筛选极少采用某种理论下的一种方法，而是某种理论下的多种方法，或者结合不同理论下的不同方法。不管使用什么方法，条目筛选的过程都要基于抽样的数据，样本量的大小对量表条目筛选有影响，但涉及这种内容的文献尚少。方积乾[14]等认为，从资料的特点和分析目的来看，有一些估计样本量的原则和变通方法，如根据测评目的估计或用 Louter 的多变量多组比较的样本含量估计法计算。

量表能通过测量某些表征或调查对象的自我感受来测评许多无法精确测量的变量，如疼痛、生存质量、日常生活能力等，因此是医学中常用的工具。量表中的条目直接影响一个量表的信效度，因此条目筛选至关重要。随着多学科的交叉发展，医学量表的条目筛选不再是单单以专家经验的方法，而是结合数学理论、心理测量学理论、信息学理论等领域的方法来实现。CTT 简单易理解、体系完整，在量表条目筛选中仍有着重要的地位；GT 是在 CTT 的基础上结合方差分析的思想，在研究测量误差方面有更大的优越性，但统计计算较 CTT 繁杂；IRT 是在克服 CTT 和 GT 的局限性基础上发展起来的一种现代测量理论，数理逻辑严密，测量精度高，但对使用者的素质和客观条件都有很高的要求，应用范围受到一定限制，针对 IRT 的理论缺陷，多维度 IRT、非参数 IRT 和认知诊断理论成为 IRT 在理论研究上的着力点。最早在心理测量和考试领域得到应用的 IRT 也逐渐用于指导医学量表的编制。尽管如此，IRT 代表着现代测量理论的发展方向，随着 IRT 的不断成熟、统计理论的发展和计算机技术的普及，IRT 在量表条目筛选中的使用可能会更加成熟，特别是多维模型、非参数模型的应用等。

参考文献

[1] 孙振球. 医学统计学 [M]. 第 3 版. 北京：人民卫生出版社，2010：430-431.

[2] 韩耀风，郝元涛，方积乾. 项目反应理论及其在生存质量研究中的应用 [J]. 中国卫生统计，2006，23 (6)：562-565.

[3] 郝元涛，孙希凤，方积乾，等. 量表条目筛选的统计学方法研究 [J]. 中国卫生统计，2004，21 (4)：209-211.

[4] 秦浩，陈景武. 医学量表条目的筛选考评方法及其应用 [J]. 中国行为医学科学，2006，15 (4)：375-376.

[5] Iramaneerat C，Yudkowsky R，Myford CM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement [J]. Advances in Health Sciences Education，2008，13 (4)：479-493.

[6] 潘海燕，丁元林，万崇华，等. 概化理论在慢性病生命质量测定量表共性模块评价中的应用 [J]. 现代预防医学，2012，39 (12)：2927-2928，2931.

[7] 安哲锋，骆方，张厚粲. 多元概化理论在评定量表编制中的作用——以音像教材测评数据分析为例 [J]. 心理科学，2008，31 (5)：1192-1194.

[8] Normand S，Belanger AJ，Eisen SV. Graded response model-based item selection for behavior and symptom identification [J]. Health Services and Outcomes Research Methodology，2006，6 (1/2)：1-19.

[9] 杨铮，戚艳波，万崇华，等. 慢性病生命质量量表共性模块项目反应理论分析 [J]. 中国公共卫生，2012，28 (11)：1477-1480.

[10] 刘全. 基于等级反应理论的民意类调查问卷选项策略 [J]. 统计与决策，2013 (9)：20-23.

[11] 林岳卿，方积乾. 世界卫生组织生存质量老年人量表简化版的研制 [J]. 中国临床心理学杂志，2011，19 (1)：27-30，34.

[12] Makransky G，Glas C. Modeling differential item functioning with group-specific item parameters：A computerized adaptive testing application [J]. Measurement，2013，46 (9)：3228-3237.

[13] Lai JS，Cella D，Chang CH，et al. Item banking to improve，shorten and computerize self-reported fatigue：An illustration of steps to create a core item bank from the FACIT-Fatigue Scale [J]. Quality of Life Research，2003，12 (5)：485-501.

[14] 方积乾，郝元涛. 生存质量研究的设计与实施 [J]. 中国肿瘤，2001，10 (2)：69-71.

收稿日期：2014-04-30