心理评定量表的编制和修订中存在的一些问题

戴晓阳，曹亦薇
（深圳大学师范学院，广东 深圳 518060）

【摘要】 目的：对当前我国心理评定量表的编制和修订中存在的一些问题进行分析和讨论。方法：通过《中国期刊网》检索到公开发表的心理评定量表410个，从量表编制理论、样本、信效度等方面进行分析。结果：①77.5%的自编量表(N=365)在编制前缺乏清晰的理论框架；②样本量普遍较大，但人群分布较局限；③条目分析报告较少且存在方法学问题；④量表的内部一致性中位数为0.872(α系数)和0.861(分半信度)，平均各维度中位数为0.774(α系数和分半信度)，量表和平均各维度重测信度分别为0.810和0.745。但是近一半的研究未报告重测信度；⑤在运用因素分析方法做效度研究中存在一些问题，40%的研究未报告效标效度，只有22%的研究做了实证效度。结论：我国心理评定量表的研制工作期得了显著的成果，但仍然存在许多问题值得注意。

【关键词】 心理测验；评定量表；信度；效度

中图分类号：R395.1　　文献标识码：A　　文章编号：1005-3611(2009)05-0562-04

Some Defects in Development and Revision of Psychological Rating Scales

DAI Xiao-yang, CAO Yi-wei
Normal College, Shenzhen University, Shenzhen 518060, China

【Abstract】 **Objective:** Some defects in the development and revision of scales were analyzed and discussed. **Methods:** 410 scales were selected from published Chinese journals in the China National Knowledge Infrastruct(CNKI). **Results:** ① For 365 self-made scales (77.5%), the goals and/or dimensions of scales were not clear before construction; ② The distribution of samples were limited comparatively; ③ The results of item analysis were reported only in few articles and there were problems of methodology; ④ The medians of internal consistence of the full scales were 0.872 (α coefficient) and 0.861 (split-half coefficient), and the medians of internal consistence of the average facets were 0.774 (both α and split-half coefficient). The stabilities of scales and the average facets were 0.810 and 0.745, but half of the stabilities were not reported. ⑤ The method of factor analysis was used inappropriately in some articles. The criterion-related validities were not reported in 40% articles.

【Key words】 Psychological tests; Rating sales; Reliability; Validity

从上世纪九十年代中期以来，我国的心理学工作者自行编制和修订了许多心理评定量表，取得了丰硕的成果，为研究者、临床工作者、教育工作者、管理者和其他人员提供了各种有效和实用的心理和行为特征评估工具。然而在取得了显著成就的同时，在量表的编制和修订的方法和过程中也存在一些问题，本文试图对一些比较突出、具有共性的问题进行探讨，以达到促进我国心理测量和测验事业发展的目的。

1　对象与方法

对《中国期刊网》进行全面检索，从1993年至2007年中期被《中国期刊网》收入的公开发表在国内77种心理学和其它相关专业期刊上的评定量表达410个（由于不同杂志被收入的时间不同，实际数目应大于此数），量表内容涉及心理卫生、学习与教育、管理心理与人才选拔、社会心理学、司法鉴定等多个领域，尤以前三个方面的量表最多。

本文拟针对量表编制理论、样本构成、项目分析、信度、效度以及相关论文的发表这几个方面，对我们所看到的问题逐个进行分析和讨论。在信度分析时，对于有总量表和分量表（维度）信度者分别计算，后者只用平均值表示；没有总量表者只计算平均分量表信度。

2　结果和讨论

2.1　量表编制理论存在的问题

在编制一个测验或评定量表时，编制者首先会确定一个总体目标，即所编量表主要用于评估什么心理特征。但是仅确立了量表的总体目标是不够的，接下来编制者需要思考这种（类）心理特征具有哪些行为表征，也就是说我们可以从哪些方面（角度）对其进行客观评估，这些方面或角度通常就构成了所编测验或量表的测量维度。正如美国著名的认知心理测量学家Embretson[1]指出：首先要界定测验的结构，提出测验的认知模型，这种认知模型的详细特征

为项目编制提供了具体说明。Smith 等人在考察临床评定量表的质量时，也将量表是否有一个明确清晰的理论基础和科学合理的量表结构列为重要的标准[2]。在编制量表的实践中，不一定理论结构的每个方面都可以用测验或问卷的方式进行评估，但在着手编制条目之前弄清楚自己准备测量的心理特征以及该特征的主要方面(或维度)是否能通过测验或问卷方法客观、准确地进行评定，这是量表编制能否成功的关键。要把握好这一点，一方面需要查阅相关的文献和研究相关的理论，在此基础上提出该心理特征的理论及结构模型；另一方面还需要通过认真的预备实验对假设的理论结构进行求证性探索。

然而从调查的结果来看，在查找到的 365 篇自编评定量表中，只有 22.5%的研究者在量表编制前已经有比较清楚的理论模型（包括维度水平）；而 68.9%只有一个初步的理论概念，其测量维度是通过因素分析方法事后引出来的；还有 8.6%的量表甚至在编制之初连理论概念都没有，作者只是根据自己的感觉将可能是评价某种心理特征的条目收集在一起，给一群被试实施，然后通过统计方法引出几个相应的维度，再形成量表的结构维度，这样的做法显然是不可取的。我们可以设想，如果量表编制者在设计条目的阶段忽略了测量目标的某个方面，没有编制相应的条目，那么在统计处理时无论采用什么方法也不可能生成相应的测量维度。因此，该量表并没有达到编制者最初的设计目的，也不能够完整准确地评估被测者的相应心理特征。

2.2 量表的样本量及其适应人群的问题

调查结果表明，大部分研究样本在 200 人以上，少数样本在万人以上，但也有少数量表采样低于 100 人，这种情况多出现在一些用于特殊人群(如精神病人、管理者等)的量表。从统计学的角度来说，样本量过少容易导致较大的抽样误差。尤其是心理测量工具本身的精确度就不及物理、化学等测量工具，其误差相对而言明显较大，因此更需要较大的样本含量才能保证结果的稳定性。关于每个变量究竟需要多少个样本量才能保证结果的稳定性？由于研究目的、分析方法不一样，目前无法设立统一标准。但是一般来说，若要报告相关系数，需要 200 个以上的样本量才可视为结果是稳定的。对于方差分析等方法，通常认为样本量最好在 20 以上。也有研究者认为每个变量最低要求 5~10 个样本点[3]。探索性因素分析通常要求样本量是变量的 5~10 倍。但是本研究调查中发现有个量表有 5 个因子共 24 题，但被试人数仅为 30 人。

另一个问题是：构成样本的对象以在校大学生和研究生为多（占 36.0%），其次是中学生（占 27.1%），其他年龄段的人群相对较少。我们推测主要原因是大、中学生群体被试比较容易获得，而寻找已经工作的成年人作为研究被试的难度大多了。由此带来的问题是大部分评定量表的使用范围受到很大的局限，特别是一些受群体特征影响较大的评定量表，如与工作效能、职业应激等有关的量表。

Smith 和 McCarthy 指出，研究者通常在同一个样本中进行条目分析，将不好的条目删除后再计算内部一致性和因素分析。这样虽然能够得到较好的信度和因子结构，但也可能存在较大的误差。他们认为在一个好的临床评定量表研制过程中常需要三个样本，两个平行样本中一个用于做条目分析，确定量表的维度；另一个平行样本用于做探索性因素分析和信度研究；第三个样本用于重复验证量表的心理测量学指标[2]。

2.3 量表条目分析的现状及其存在的问题

条目分析是测验和量表编制过程中一个非常重要的环节，当测试样本收集回来后，研究者会在量表的条目分析上面做大量的工作，但是呈现研究结果时它往往只占很小的篇幅。调查发现仅一半的研究呈现了条目分析的结果，大部分是条目与总分(或分量表总分)的相关(题总相关)，只有极少的研究呈现了鉴别指数(或区分度)。从严格意义上来说，题总相关系数不能算区分度。正确做法是先将条目分数变换成 0-1 分数后再算与总分相关即点二列相关系数。另一方面，大多数研究者在计算区分度时往往疏忽在总分中除去该条目得分的步骤，因此会造成点二列相关系数“虚高”，在条目数量较少时尤为明显。建议对条目较少的测验进行区分度分析时应将总分中除去该条目的得分后再计算点二列相关或二列相关。

2.4 量表的信度及其存在的问题

信度是衡量一个测验或量表可靠性(误差大小)的指标，主要包括考查项目之间一致性的指标如分半信度和 α 系数，考查量表在不同时间测试时结果的稳定性指标如重测信度，其它信度指标还有评定者之间的一致性和平行本信度等等。

2.4.1 量表或分量表内部条目的一致性

调查结果表明超过 90%评定量表均做了这个指标，其中 75.2%的研究计算了全量表 α 系数，35.8%的研究计算了全量表分半相关系数，还有 32.4%的研究同时

采用了两种方法。部分量表由几个独立的分量表组成,所以只计算了分量表信度。还有一些研究者同时计算了分量表和全量表的信度。我们计算分量表的平均相关系数,结果见附表。结果显示:大多数公开发表量表的全量表内部一致性都达到了较高的水平,分量表(或维度)的内部一致性也达到中等以上水平。因此我们推荐 0.8 作为评价一个评定量表全量表内部一致性的标准,0.75~0.8 作为评价评定量表的分量表(或维度)内部一致性的标准。

附表　量表的信度和分量表平均信度

信度系数	全量表信度			平均分量表信度		
分布范围	α系数	分半信度	重测信度	α系数	分半信度	重测信度
小于 0.4	0	0	0	0	0	2
0.4~0.499	0	0	4	1	0	4
0.5~0.599	3	2	6	11	5	10
0.6~0.699	19	9	13	58	19	33
0.7~0.799	47	31	56	137	35	63
0.8~0.899	118	64	75	101	31	33
0.9 以上	120	40	18	22	5	12
量表个数	307	146	172	330	95	157
平均值	—	0.842	0.793	—	0.764	0.741
中位数	0.872	0.861	0.810	0.774	0.774	0.745
最大值	0.992	0.999	0.990	0.950	0.930	1.000
最小值	0.500	0.550	0.420	0.450	0.580	0.270

2.4.2　量表的重测信度　结果表明，只有 229 个(51.9%)量表进行了重测信度研究。重测样本平均为 85.8 例(标准差 79.3);两次测量之间的间隔天数从 1~180 天,平均间隔为 23.5 天(标准差 20.9)。从附表可以看出，大多数公开发表量表的全量表重测信度都达到了较高的水平,分量表(或维度)的重测信度也达到中等以上水平。因此我们推荐 0.8 作为评价一个评定量表全量表重测信度的标准,0.75 作为评价评定量表的分量表(或维度)内部一致性的标准。值得指出的是:在本次调查中发现近一半的研究没有做重测信度研究，对于这样一个基本且重要的指标被量表的编制或修订者以及发表这些论文的编辑或审稿者所忽略,这不能不说是一个极大的遗憾。

2.5　量表效度的调查结果及其存在的问题

效度是衡量测验或量表的有效性，即在多大程度上达到了设计目的指标。Anastasi 指出:“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,.....”[1]。效度分析过程应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,然后才准备符合这种结构定义的测验项目。一般包括内容关联效度、结构关联效度和效标关联效度三类。

2.5.1　内容关联效度　大多数研究者在其研究论文中都报告采用专家或同行评议的方法进行了内容效度研究，但由于多数论文在发表时不会附上量表以及各维度的具体条目，因此本调查无法对其内容效度情况做出进一步的评价。

2.5.2　结构关联效度　因素分析是对量表结构进行验证的一种常用方法,调查发现 85%的研究者采用了这种分析方法，绝大多数研究者在需要使用因素分析方法来验证结构效度的研究中均采用了这种方法，统计发现 73.5%的研究使用了探索性因素分析方法,32.0%的研究使用了验证性因素分析方法。其中 102 个研究同时使用了两种分析方法,39 个研究只使用了验证性因素分析方法。结果说明在评定量表的编制和修订中因素分析已成为一种常规的分析手段。但是我们也发现在使用因素分析方法时存在一些问题：①少数研究对同一个样本先做探索性因素分析，再做验证性因素分析，这样做是毫无意义的。但是在另一个样本中验证前一个样本中所得到的因子结构，这样可以用来证明量表因子结构的稳定性。②单个因子所含条目太少,例如有的研究平均一个因子只含 2~3 个条目。例如有个研究自编量表 4 个维度,只有 10 个条目。侯杰泰等认为:“每个因子应至少应有 3 个或更多的指标（条目）”[3]。而 Fabrigar 等建议：当研究者认为因子的特征比较清楚时,每个因子至少应有 4 个条目;如果因子的特征不太确定时则应包含 6 个条目[4]。而芝祐顺认为每条目得到 10 个左右的公共因子是比较合理的[5]。③有的研究抽出的因子所解释的方差太少，最低者仅解释 17%的总体方差,这说明这一组变量的共同方差较少，而其独特性方差和误差所占比例较大。Streiner 认为对于一个好的因子结构而言,抽出的因子应能够解释总体方差的 50%以上[6]。④在对因子命名的时候最好遵循分类学原则，分类的标准尽可能保持一致,使得不同类别之间具有相对独立性。⑤许多有关量表编制或修订的论文在呈现因子结构时只列出了符合主要因子数值，而将其它因子的负荷值省略，这样做使得读者和使用者无法完整地判断其因子模型。其它分析结构关联效度的方法包括会聚效度和区分效度,可能由于发表文章的篇幅限制,大部分研究均没有呈现这些结果，仅 18.7%的研究有报告这类结果。

2.5.3　效标关联效度　调查发现约 60.2%的论文报告了量表的效标效度,其中 54%研究报告了一个效标效度,24.3%的研究报告了两个效标效度,最多者报告了 9 个效度指标。但令人不能满意的是近 40%的研究未进行任何效标效度的研究。选择一个良好的效标是验证量表效标效度的前提,Anastasi 认为:

“对于许多测验的目的来说,最令人满意的效标度量是实际工作表现的追踪记录”,“在编制某些人格测验时,精神病诊断既是选择项目的基础,也是测验效度的证据”[1]。Anastasi 所说的这些指标就是人们通常说的“金标准”,而以这些“金标准”为效标所做的效度研究被称为实证效度。但是在本调查中发现,只有 22.9%的研究做了实证效度,而 37.3%的研究则用其它量表的得分作为效标。我们分析部分原因可能因为准备测量的某些心理特征不容易找到相应的“金标准”,研究者只好选择了这些间接指标替代,但更多的情况下可能是因为这些具有“金标准”特征的效标较难以获得,需要花更多的时间和精力才可能得到。我们认为对于那些用于辅助诊断、决策的心理评定量表,如异常心理、教育诊断量表、人才选拔量表等,应当尽可能进行实证效度研究。

3 小结与建议

近 10 多年来,我国的心理工作者以编制和修订并且公开发表了许多用于各种目的的评定量表,取得了非常显著和卓越的成就。但是在这些量表的编制或修订过程中还存在一些问题和不足,特别是在测验的信、效度研究方面。为此我们特提出几点建议:①在准备编制心理评定量表的开始阶段就应当重视理论构想和测量维度的设计。②扩大样本的来源,使所编制的量表能用于更广的群体。③注意选择正确的条目分析统计方法。④注意重测信度研究。⑤在效度研究中应特别注意加强实证效度的研究,同时也应注意采用正确的因素分析方法和表述方式。⑥专业期刊的编辑和审稿者要加强相关论文的质量审查。

参 考 文 献

1 Anastasi A, Urbina S. Psychological testing (7th Ed). New Jersey: Prentice Hall Inc, 1997. 121-138

2 Smith GT, McCarthy DM. Methodological considerations in the refinement of clinical assessment instruments. Psychological Assessment, 1995, 7(3): 300-308

3 侯杰泰,温忠麟,成子娟. 结构方程模型及其应用. 北京:教育科学出版社,2004. 126-127

4 Fabrigar LR, Wegener DT, MacCallum RC, et al. Evaluating the use of exploratory factor analysis in psychological research. Psychology Methods, 1999, 4(3): 272-299

5 曹亦薇译. 芝祐顺著. 因素分析法. 北京:人民教育出版社,1999. 151

6 Streiner DL. Figuring out factors: The use and misuse of factor analysis. Canadian Journal of Psychiatry, 1994, 39: 135-140

(收稿日期:2009-04-25)

(上接第 558 页)

值得注意的是,探索性因素分析保留的 31 个项目与原量表的项目归属完全一致。英文版问卷 3 个因素共同解释了变异的 37.95%,而修订后 PhoPhiKat 问卷中文版 3 个因素共同解释了总变异的 35.91%,与英文版问卷基本相近[6]。验证性因素分析结果均在可以接受的范围之内,虽某些指标偏低,但在前人的量表修订研究中也不乏先例,因此可以说 PhoPhiKat 问卷中文版具有良好的结构效度。以社交焦虑和羞怯作为效标,发现被笑恐惧与其都有较高的相关,而被笑愉悦和笑他愉悦两个维度与校标的相关系数也达到了显著水平,说明问卷也有较高的效标关联效度。(致谢:对瑞士苏黎世大学 Willibald Ruch 教授和 René T Proyer 博士在本文写作过程中所给予的帮助表示感谢。)

参 考 文 献

1 喻丰,郭永玉. 与笑有关的三种个体差异:被笑恐惧、被笑愉悦与笑他愉悦. 心理学探新,2009, 29(1): 82-86

2 Titze M. The Pinocchio Complex: Overcoming the fear of laughter. Humor and Health Journal, 1996, 5: 1-11

3 Salameh WA. Interview with Dr. Michael Titze. Humor and Health Journal, 1996, 5: 12-20

4 Ruch W, Proyer RT. The fear of being laughed at: Individual and group differences in gelotophobia. Humor: International Journal of Humor Research, 2008, 21(1): 47-67

5 Ruch W, Proyer RT. Who is gelotophobic? Assessment criteria for the fear of being laughed at. Swiss Journal of Psychology, 2008, 67(1): 19-27

6 Ruch W, Proyer RT. Extending the study of gelotophobia: On gelotophiles and katagelasticists. Humor: International Journal of Humor Research, 2009, 22(1-2): 183-212

7 Department of Personality Psychology and Diagnostics, University of Zurich. Gelotophobia, the fear of being laughed at/Multi nation study on gelotophobia. Retrieved February 8, 2009, from http://www.psychologie.uzh.ch/perspsy/gelotophobia/

8 汪向东. 心理卫生评定量表手册(增订版). 中国心理学卫生杂志社,1999. 244-248

9 陈思远,管阳阳,张伟,等. Hewitt 多为完美主义量表在 529 名大学生中的试用与修订. 中国临床心理学杂志,2008, 16(7): 823-825

10 侯玉波,张梦. 青少年应对方式量表的修订. 中国临床心理学杂志,2006, 14(6): 566-568

11 郭素然,辛自强,耿柳娜. 事件影响量表修订版的信度和效度分析. 中国临床心理学杂志,2007, 15(1): 15-17

(收稿日期:2009-02-28)