16 PF问卷网络与纸笔施测方式的比较

吴瑞林 王建中 马喜亭
(北京航空航天大学 心理与行为研究所, 北京 100083)

摘 要: 为检验网络测试与纸笔测试方式是否具有同等的信效度, 16PF问卷被用于对两个样本分别进行网络(n=213)和纸笔(n=2801)施测; 并从网络测试样本中随机抽取 47人, 随后进行纸笔重测。在α系数、测题同质性和次级因素结构三项检验中, 各个人格因素的表现各有优劣, 未得出统一的结论, 但可以确定网络测试的信效度较纸笔测试没有明显下降。配对 t检验的结果显示, 两种施测方式下同一批被试的结果在部分因素上有显著性差异, 不能将纸笔测试获得的常模直接用于网络测试。

关键词: 心理测验; 网络测试; 纸笔测试; 人格; 16PF
中图分类号: B841.2 文献标识码: A 文章编号: 1003-5184(2010)05-0078-06

1 引言

计算机和网络技术的发展已经彻底改变了人类的活动方式, 并渗入到生活、学习和工作的各个方面。在网络进行心理测试可以由计算机直接计分、统计和管理, 减少了处理流程, 被试可以分散到不同地点、自选时间进行测试, 具有便捷性、灵活性、广泛性等优点[1]。因此, 很多心理测验都从传统的纸笔形式转移到网上, 近年来, 利用网络进行心理测验的情况在国内也越来越普遍。但纸笔测验可否直接搬到网上, 这种直接使用后的测验属性(主要指信度和效度)能否达到心理测验的要求, 尚处于争论之中。

在国外, 从事心理测量学研究的学者很早就开始关注这一问题。对于网络测试和传统纸笔测试的信效度以及均值和标准差是否等同, 一些基于经典测量理论(classic test theory)的研究得出了不完全相同的结论。Meier[2]和 Cohen[3]等人都认为心理测验能否从纸笔方式直接搬到网上不能一概而论, 需要逐个量表进行考察。Buchanan和 Smith通过对一个人格测验——自我监控量表(SRS-R)的实验发现, 网络测试的信效度优于纸笔测试[4]; Carlbring等人在恐怖症病人中使用了一些恐怖障碍量表进行测试, 发现网络测试和纸笔测试的信度指标几乎相同, 但某些因素的均值和标准差却差异显著[5]; Vallejo等人在大学生中使用 SCL90-R和 GHQ28量表进行调查, 发现纸笔和网络两种施测方式的数据结果呈现高度的相关[6]。

现有国内公开发表的论文中, 对该问题的实证研究还非常少。能够检索到的有蔡华俭等人[7]于 2008年发表的《网络测验和纸笔测验的测量不变性研究》一文。该研究以 5道题目生活满意度量表为例, 发现网络测试和纸笔测试之间存在弱不变性, 即两种施测方式具有相同的测量单位; 但只存在部分的强不变性(节距相等)和部分严格不变性(残差相等)。

因此, 在中国本土背景下研究心理测验在网络施测方式下的测验属性, 具有十分重要的现实意义, 对中国网络心理测试系统的开发和应用也有一定的借鉴和指导意义。卡特尔十六种人格因素问卷(以下简称为 16PF)是目前被应用最多的人格测验之一[8]。比较其在两种施测方式下的信效度指标, 有一定的典型性和代表性。同时, 人格作为较为稳定的特质, 在短期内不会发生较大变化, 前后测间受到的其它因素干扰小。所以, 研究选定 16PF作为测试工具, 从经典测量理论的信效度角度对网络和纸笔两种施测方式进行检验和比较。

2 方法

2.1 工具

调查使用的量表为李绍衣 1981年修订的 16PF中译本[9], 该版本是在 16PF英文第三版的基础上翻译、修订而来[10]。据统计, 它是我国高校对新生进行心理普查时使用频率最高的人格测验[11]。量表共包括 187道题目, 其中有 3道为提示和测伪用, 其它 184道题构成 16个人格因素, 其中的 15个人格因素又可以组合成 8个次级因素。

网络测试工具为北京航空航天大学心理与行为研究所开发的网络心理测验平台。该测试平台采用网页(Web)形式, 分页呈现测试题目; 每页显示 10

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

道测试,使用键盘和鼠标完成答题操作;测验无时间限制。

2.2　对象和过程

被试全部为大学一年级的新生,根据其所在院系被分为两组,一组参加纸笔测试(以下称为纸笔测试样本),另一组参加网络测试(以下称为网络施测样本)。参加纸笔测试的同学由各院系分别组织集中施测,参加网络测试的同学被分在 3个计算机房同时集中施测。纸笔测试样本含有被试 2801人,其中男生 2259人(占 80.6%),女生 542人(占 19.4%);平均年龄为 18.52±2.07岁;独生子女 1476人(占 52.7%)。网络施测样本共有被试 213人,其中男生 161人(占 75.6%),女生 52人(占 24.4%);平均年龄为 18.87±2.41岁;独生子女 119人(占 55.9%)。

参加过网络测试的同学中又有一小部分被抽出,在一周之后的相同时间集中进行了纸笔测试,这部分样本被称为网络测试后纸笔测试样本。该样本共有 47人,其中男生 31人(占 66.0%),女生 16人;平均年龄为 18.38±0.64岁。纸笔测试后,又对这 47名同学分别进行了个别访谈,被试均表示在两次测试相隔的一周内,没有发生过对自己产生重大影响的生活事件。

在数据分析过程中,网络施测和纸笔测试的内部一致性信度、测题同质性效度、以及次级人格因素的构想效度将被分别检测和比较;最后,还将对网络测试后纸笔测试样本 47名同学的前后测进行配对 t检验,以判断两种施测方式的结果是否相等。测试数据使用 SPSS11.0进行 α系数、相关系数和配对 t检验的计算,使用 Bentler开发的结构方程模型软件 EQS6.1进行多组验证性因素分析。

3　结果与分析

3.1　内部一致性信度

三个样本中,16个一级人格因素的 Cronbach's α系数见表 1。网络施测样本的 α系数从 0.019至 0.774,16个因素 α系数的均值为 0.405;纸笔测试样本的 α系数从 0.007到 0.753,平均值为 0.436。网络施测样本的 16个 α系数中,有 7个高于纸笔测试,平均值相差 0.058;9个低于纸笔测试,平均相差 0.100。网络测试后进行纸笔测试样本的 α系数整体上较网络测试样本和纸笔测试样本的 α系数都高,这可能与该样本的样本量较小有关。

表 1　16个人格因素的 α系数

因素名称	因素代号	网络施测(n=213)	纸笔测试(n=2801)	网络与纸笔之差	网络测试后进行纸笔测试(n=47)
乐群性	A	0.644	0.586	0.058	0.671
聪慧性	B	0.378	0.339	0.039	0.533
稳定性	C	0.538	0.550	−0.012	0.431
恃强性	E	0.494	0.538	−0.044	0.525
兴奋性	F	0.774	0.753	0.021	0.815
有恒性	G	0.429	0.458	−0.029	0.494
敢为性	H	0.557	0.684	−0.127	0.743
敏感性	I	0.076	0.404	−0.328	0.579
怀疑性	L	0.344	0.367	−0.023	0.538
幻想性	M	0.019	0.193	−0.174	0.511
世故性	N	0.062	0.007	0.055	0.213
忧虑性	O	0.586	0.611	−0.025	0.488
实验性	Q1	0.177	0.084	0.093	0.230
独立性	Q2	0.617	0.513	0.104	0.450
自律性	Q3	0.214	0.348	−0.134	0.434
紧张性	Q4	0.570	0.535	0.035	0.606

3.2　测题同质性效度

为了探讨构成测验的各个测试题的有效程度,即为查明每种人格因素量表中的测试题是否有价值,且是否测量同一人格特质,辽宁省教育科学研究所的李绍衣[9]和华东师范大学的祝蓓里、戴忠恒[13]使用过 8道测试题与 4个因素之间的相关来检验量

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表的测题同质性效度。具体办法为:选取量表中分属于4个因素的8道测试题(因素A的第51,151题;因素C的第4,30题;因素F的第58,108题;因素Q3的第73,98题),计算它们与其所属因素及其他因素的皮尔逊相关系数。这里也采用该方法来比较两种施测方式下的测题同质性效度,结果见表2。

无论哪种施测方式,题目与所属因素间的相关系数均在统计上显著,且与所属因素间的相关均强于与非所属因素的相关。网络测试中因素C的题目和因素F的58题与所属因素间的相关系数大于纸笔测试的结果,纸笔测试中因素A和因素Q3的题目与所属因素间的相关系数大于网络测试的结果;两种施测条件下,各题目与非所属因素的相关系数趋势基本一致。

表2 测题同质性的检验结果

所属因素	题目	纸笔测试(n=2801)				网络测试(n=213)			
		A	C	F	Q3	A	C	F	Q3
A	51	0.51**	0.04*	0.18**	0.04*	0.47**	−0.05	−0.01	−0.11
	151	0.45**	0.03	0.26**	−0.01	0.39**	0.08	0.04	−0.14*
C	4	0.09**	0.46**	0.23**	0.29**	0.07	0.51**	0.36**	0.07
	30	0.11**	0.43**	0.19**	0.22**	0.00	0.47**	0.26**	0.08
F	58	0.21**	0.10**	0.50**	0.03	0.14*	0.25**	0.58**	0.09
	108	0.27**	0.35**	0.58**	0.26**	0.20**	0.40**	0.57**	0.05
Q3	73	0.10**	0.32**	0.22**	0.54**	0.09	0.40**	0.27**	0.42**
	98	0.05**	0.27**	0.07**	0.48**	−0.08	0.25**	0.07	0.46**

3.3 次级人格因素的结构效度

16PF的次级人格因素是由15个人格因素组合而成的,可以使用结构方程模型对其构想效度进行验证性因素分析,以判断其是否符合原来的理论构想。纸笔测试样本的验证性因素分析的结构方程模型如图1左侧所示,拟合指标列于表3的第二行;网络测试的拟合指标列在第三行。两个样本的拟合指标虽然可以接受,但并不非常理想,尤其是CFI指数,远没有达到大于0.9的一般要求,这说明现行16PF中译本的次级因素结构需要被重新探讨。由于两个样本的样本量相差较大,这里对χ^2/df指标的比较没有意义,纸笔测试样本的GFI指标好于网络测试样本,网络测试的CFI和RMSEA指标好于纸笔测试样本。

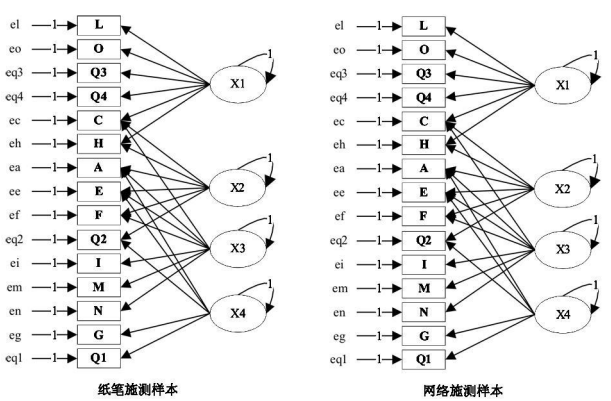

图1 次级因素结构不变性验证模型

在研究中,还希望了解该次级因素结构在两种施测方式下是否一致。所以,在前面的基础之上又利用测量不变性(measurement invariance)的概念来检验该构想结构在两种施测方式下有没有显著的差异。

首先,构建如图1所示的结构方程模型作为基线模型(baseline model),在这里该模型被称为模型1。根据Reise提供的三种锚定基线模型的方法之一[14],这里设置所有次级因素的方差为1,两组样本的一级因素到次级因素的因素载荷(factor loading)设为自由估计。接下来,在模型2中强制设定两个样本中对应的因素载荷相等,就是说让纸笔样本模型的次级因素X1到一级因素I的因素载荷等于网络样本的因素X1到因素I的因素载荷,依次类推,使得每一对对应的因素载荷都相等。如果模型1与模型2间χ^2值的变化没有达到统计上显著的范围,就可以认定16PF的次级因素结构在两个样本间是相同的。

表3给出了模型1和模型2的各项拟合指标,同样,两个模型的拟合指标也都不很理想,其中CFI指数不到0.5,RMSEA指数也没能小于0.05。模型2与模型1相比,χ^2值增大38.02,同时因为强制两个样本的因素载荷相等,模型的自由度增加了24,进行卡方检验,P值为0.035,小于0.05的显著性级别。因此,不能认为16PF的次级因素结构在两个样本中一致。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　次级因素结构测量不变性验证的拟合指标

模型	χ^2	df	χ^2/df	GFI	NFI	CFI	RMSEA	$\Delta\chi^2$	Δdf	p
纸笔样本	1460.46	81	18.03	0.930	0.467	0.476	0.078			
网络样本	162.14	81	2.00	0.898	0.412	0.525	0.069			
模型 1	1623.05	162	10.02	0.928	0.462	0.479	0.077			
模型 2	1661.07	186	8.93	0.926	0.449	0.474	0.073	38.02	24	0.035
模型 3	1655.46	185	8.95	0.927	0.451	0.476	0.073	32.41	23	0.092

在模型 3中,根据 EQS提供的修正指数,对模型 2进行了修改。从次级因素 X3到一级因素 M的因素载荷不再设置为强制相等,而是分别加以估计,其它因素载荷依旧设为对应相等。拟合指数列在表 3中最后一行,模型 3与模型 1相比,χ^2值增加了 32.371,自由度增加 23,p值为 0.093,没有达到 0.05的显著性。将 X3与 M间的因素载荷设为自由估计后,两个样本的次级因素结构达到部分一致性。通过三个模型的比较,可以得出这样的结论,16PF的次级因素结构没能在两种施测方式下保持完全的一致;除 X3与 M间的因素载荷外,其它一级因素与次级因素间可以认为具有相同的测量单位,也就是说次级因素每改变一个单位时,除 M外的一级因素在两种施测方式下的变化相同。

3.4　配对 t检验

表 4给出的是 47名同学网络施测结果与一周后纸笔测试得分的配对 T检验。从相关系数看,除幻想性因素外的 15个因素的相关系数为 0.584到 0.835,且相关均达到了非常显著可信的水平;只有幻想性因素的相关系数偏低,为 0.390,但也达到了显著可信的水平;16个因素的平均相关为 0.690。从 t检验的结果看,在恃强性、世故性、兴奋性、有恒性、敢为性、敏感性和忧虑性 7个因素上,也就是说在接近半数的因素上,前后测的结果表现出显著的差异。

表 4　纸笔测试和网络测试的配对 T检验

因子名称	因子代号	网络测试得分 (M±SD)	纸笔测试得分 (M±SD)	相关系数 (r)	t值
乐群性	A	8.36±4.51	8.77±4.68	0.83***	−1.04
聪慧性	B	8.64±2.49	9.11±2.03	0.67***	−1.71
稳定性	C	16.51±3.34	16.85±3.52	0.69***	−0.87
恃强性	E	9.32±3.55	12.23±4.07	0.65***	−6.16***
兴奋性	F	14.74±5.19	16.32±5.91	0.84***	−3.31**
有恒性	G	12.30±3.18	13.36±3.44	0.67***	−2.71**
敢为性	H	11.28±4.18	13.04±5.53	0.79***	−3.59**
敏感性	I	11.49±3.01	10.68±4.13	0.79***	−2.18*
怀疑性	L	8.45±3.46	8.53±3.63	0.67***	−0.20
幻想性	M	13.49±3.48	14.53±4.19	0.39**	−1.67
世故性	N	7.89±3.14	9.34±3.05	0.58***	−3.52***
忧虑性	O	7.94±3.52	6.79±3.41	0.64***	−2.68*
实验性	Q1	10.64±3.00	11.21±3.36	0.69***	−1.55
独立性	Q2	12.55±4.10	11.83±3.66	0.71***	−1.66
自律性	Q3	13.64±2.77	14.04±3.09	0.67***	−1.16
紧张性	Q4	9.70±4.23	8.94±4.13	0.76***	−1.79

4　讨论

4.1　两种施测方式信效度指标的优劣

比较 16个因素的 α系数,网络施测的结果在 7个因素上高于纸笔测试,其它因素则低于纸笔测试,没有得到确定的趋势。而测题同质性检验也呈现交错的结果,在部分题目上,网络测试更具优势,而在另一些题目上纸笔测试的相关结构更好。次级因素的结构方程模型拟合指标上,两个样本也互有优势,无法判断哪种施测方式的次级因素构想效度更好。

配对 t检验是判断两种测试方式可否等同的另一种方法,两种施测方式在各因素上较高的相关系数说明两种测试方式在测量 16个人格因素上的表现较为一致。同时,t检验的结果显示,两种测试方式在 7个因素上有不同程度的统计显著性差异。但

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

这种差异究竟是两种不同的施测方式造成的，还是前后测间其它因素影响所致，还需要进一步的交叉实验来验证。

根据上面的统计结果，很难得出统一的结论，判定哪种测试方式的信效度指标更优。这也验证了Meie等人的判断[2]：心理测验能否从纸笔测试迁移至网络，不仅要逐个量表的考察，其中的每个子量表也要分别加以验证。总体上看，可以确定的是网络施测的信效度并没有明显低于纸笔测试。

4.2 16PF较低的信效度指标

上节中的统计结果再次验证16PF各因子内部一致性系数不高的事实，最低的甚至达到0.007。这一结果与徐蕊等人在2006年发表的研究[12]一致，也是16PF中译本一直以来被诟病的地方。同样，研究对16PF次级因素结构进行检验性因素分析的拟合指标也不理想，这点也与徐蕊等人的研究发现相同。

从题目内容上分析，经过近30年的使用，16PF中很多题目的说法和效果都需要重新进行考量，尤其是测量“聪慧性”的有关题目。比如第177题为“一人__事，众人受累。”，要求选出空缺的字；对于绝大多数同学来说，并不知道这句话，该题目的信效度自然也就很低[15]。从这些方面看，现行使用的于上世纪八十年代初修订的16PF中译本亟待改进。

4.3 用测量不变性的方法检验量表的结构一致性

测量不变性的方法被用来判断16PF的次级因素结构在两种施测方式下是否一致，其基本原理是检验在强制两样本的因素载荷相等的情况下，模型的χ^2值有无显著性变化。统计结果显示，两种施测方式下16PF获得了部分结构一致性，即除一个因素载荷外，大部分的因素载荷都可设为相等。所以，可以认为16PF的次级因素结构在两种测验方式下大体上是一致的。

该检测方法适用于判定测验结构是否具有跨组的一致性。而测量工具的结构一致性也是心理学中进行跨文化研究、跨情景研究，以及比较性别间差异、种族差异、年龄差异等各种分组比较研究的基础。只有在确定了测量不变性的基础之上，进行跨组的比较才有意义；否则，得出的结果很可能是完全错误的。

4.4 被试对两种施测方式的主观感受

除对测试的数据进行分析外，研究还对同时参加两种测试方式的47名同学进行了访谈，以了解被试的主观感受。访谈结果显示，70%以上的学生表示更愿意采用网络的方式进行测试，因为其操作方便、节省时间；但也有10%左右的学生表示更愿意参加纸笔测试，因为长时间对着计算机屏幕让他们感到眼睛发酸，而纸笔测试感觉上更加舒适。这说明，网络测试的人机交互界面还需要进一步得到改进，以减缓被试的眼疲劳程度，使网络测试能够得到更好的推广。

5 结论

统计结果显示，16PF中译本在网络进行测试的α系数和测题同质性效度与纸笔测验相比，均没有明显的下降；而两种施测方式的次级因素构想效度大体上也一致，可以将16PF中译本移植到网络作为正式的心理测验使用。

但是，在部分因素上，两种施测方式获得分数的平均值存在显著性的差异。因此，不能简单的把通过纸笔测试获得的常模直接用于网络测试，有必要为16PF的网络测试版本建立专有常模。

参考文献

1 叶茂林. 网络心理测验法述评. 心理科学，2006，28(2)：423－425.

2 Meier S. The chronic crisis in psychology measurement and assessment: A historical survey. San Diego: Academic Press, 1994.

3 Cohen R J, Swerdlik M E, Smith D K. Psychological testing and assessment. Mountain View, CA: Mayfield Publishing, 1992.

4 Buchanan T, Smith J L. Using the internet for psychological research: Personality testing on the world wide web. British Journal of Psychology, 1999, 90(1): 125－144.

5 Carlbringa P, et al. Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. Computers in Human Behavior, 2007, (23): 1421－1434.

6 Vallejo M A, Mananes G, Comeche M I. Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. Journal of Behavior Therapy and Experimental Psychiatry, 2008, 39(3): 201－208.

7 蔡华俭，等. 网络测验和纸笔测验的测量不变性研究——以生活满意度量表为例. 心理学报，2008，40(2)：228－239.

8 金瑜. 心理测量. 上海：华东师范大学出版社，2001.

9 李绍衣. 卡特尔十六种人格因素测验指导手册. 沈阳：辽宁教科所，1981.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

10 程嘉锡, 陈国鹏. 16PF第五版在中国应用的信度与效度研究. 中国临床心理学杂志, 2006, 14(1): 13－16.
11 邵利聪. 大学新生心理健康测试量表的选择与使用分析. 教育与职业, 2006, 8(24): 141－143.
12 徐蕊, 宋华淼, 苗丹民. 卡特尔 16种人格因素 (中国版) 构念效度的验证. 第四军医大学学报, 2007, 28(8): 744－746.
13 祝蓓里, 戴忠恒. 卡氏十六种人格因素中国常模的修订. 心理科学通讯, 1988, (6): 14－18.
14 Reise S P, Widaman K F, Pugh R H. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 1993, 114(3): 552－556.
15 王建中, 吴瑞林. 高校心理健康测评的现状与发展趋势. 高校心理健康教育新进展. 吉林: 吉林出版社, 2007: 1: 109－114.

Comparison between Internet and Paper-Pencil Administration of 16PF Test

Wu Ruilin　Wang Jianzhong　Ma Xiting
(Institute of Psychology and Behavior, Beihang University, Beijing 100083)

Abstract: A personality test (16PF Chinese Revised Version) is employed to conduct a comparison study of psychometric properties between Internet administration and paper-pencil administration, since personality is a stable trait. Two samples, the Internet test sample (n=213) and a traditional paper-pencil test sample (n=2801), is taken the test which is very popular in China. The 47 subjects from the Internet sample are selected randomly to participate in a parallel paper-pencil administration one week later. The results show personality factors have different performance on Cronbach's α, consistency of items, and confirmatory factor analysis of second order factors of two samples. It is obvious that reliability and validity of Internet administration are not worse than those of paper-pencil administration. The statistic of paired-samples t-test shows there are significant different between the data of Internet administration and paper-pencil administration on some factors. So there will be problems if applying the norm obtained via paper-pencil administration to Internet test directly.

Key words: psychometrics; internet test; paper-pencil test; Personality; 16PF

(上接第 71页)

A Study of the Construct-related Validity of Assessment Center By Multivariate Generalizability Theory

Wang Bo　Tian Xiaoxun　Shao Yanping　Che Hongsheng
(School of Psychology, Beijing Normal University, Beijing 100875)

Abstract: Assessment Center is a principal form of modern personnel assessment which is mainly used to assess middle and upper management by various simulative tasks. However seldom studies could prove that it has ideal construct-related validity. This study also fails to prove it according to the real data from an actual assessment center project of a financial firm. Then, through the analysis of multivariate generalizability theory, this study points out the relationships between dimensions and tasks and reveals the measurement reliability of each dimension, both of which may explain the paradoxical construct-related validity. Finally, this study discusses the approaches of optimizing the construct-related validity and improving the assessing reliability.

Key words: assessment center; construct-related validity; multitrait-multimethod; multivariate generalizability theory

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net