心 理 学 报 2008,40(2):228~239
Acta Psychologica Sinica

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

蔡华俭 林永佳 伍秋萍 严 乐 黄玄凤

(中山大学心理系,广州 510275)

摘 要 以生活满意度量表为例,运用实证性因素分析,考察在中国文化下网络测验和传统纸笔测验之间的测量不变性。结果显示,网络测验和纸笔测验之间存在弱不变性,即网络测验和纸笔测验有着相同的测量单位;但网络测验和纸笔测验只存在部分的强不变性和部分的严格不变性,测验实施环境对结果的影响不可忽视。该研究表明,恰当设计的网络测验是可靠的,同时还提示,当一个测验在不同情境下运用时,检验测量不变性十分必要。

关键词 测量不变性,实证性因素分析,网络测验,纸笔测验,生活满意度量表。

分类号 B841.7

1 引言

1.1 网络心理测验及其可靠性研究

在过去十多年中,随着网络技术的发展和快速普及,心理学领域越来越多的研究者开始运用网络收集数据进行研究[1~3]。和传统通过纸笔测验收集数据相比,采用网络收集数据有很多好处:1)利用网络很容易获取非常大的样本数据[4];2)可以运用计算机自动计分并方便地进行数据转换[5];3)能减少很多无关因素的影响,比如:期望效应(demand characteristic)、观察偏差(observer bias)和反应偏差(response bias)等[6]。但是,由于研究者不能完全自由选择网络测验的参加者,也不能完全控制网络数据收集的情境和条件,因此,实践中经常碰到的一个问题就是:同一个测验通过网络实施和采用传统的纸笔测验形式实施得到的结果是否对等(equivalent);或者说,通过网络模式和纸笔模式实施的测验分数是否可以同样地解释。

事实上,心理测量的有关研究表明,同一个测验在不同的情境下应用时,很可能会导致测量特性的改变[8]。对于网络测验和传统纸笔测验是否具有测量对等性,迄今已有不少研究。不过,绝大多数都是基于经典心理测量理论展开的。这些研究有的侧重比较观测分数的均数和变异是否对等[9~19],有的侧重于检验估计的信度是否对等[19~23],有的侧重比较和外在效标的相关是否对等[16,19]。总的说来,这些基于经典测量理论的研究基本表明,网络测验和传统的纸笔测验有着对等的信度和效度,能同样地揭示某些组间差异(比如,性别差异,年龄差异等)。但是,由于这些测验都是基于观测分数进行比较的,在观测分数层面上的对等并不意味着测验就具有相同的含义,也不能认为就具有可比性。因为,同一个测验在不同情境下测量的潜在结构可能并不完全相同(比如,自尊量表在中国测量的除了自尊本身以外,可能还有谦虚),潜在结构和测量项目之间的关系未必完全相同,测量的残差变异(包括特异因素和测量误差)也未必相同,等等。因此,基于经典测量理论的分析得到的对等性并不意味着潜变量以及潜变量和观测变量之间的关系的对等。

因此,要深入探讨网络测验和纸笔测验是否对等,必须进行基于潜变量分析的测量不变性研究。理论上讲,测量不变性是指在给定潜变量的情况下,观测分数的条件分布的跨组的不变性[24]。通俗地讲就是测验在不同的情况下应用时不存在与特定组相关的测量偏差[25,26]。迄今,相关研究还很少,结果也不一致。有些基于实证性因素分析的研究发现网络测验和纸笔测验基本是对等的[27,28],但另一些却发现二者不完全对等[9,29];一项基于项目反映理

收稿日期:2007-05-25

* 本研究得到中山大学"985 工程"队伍建设"百人计划"引进人才科研启动基金资助。

通讯作者:蔡华俭,E-mail: huajian.cai@gmail.com

论的研究也发现,网络测验和纸笔测验只是部分对等[30]。最近,有研究者试图把基于潜变量的项目反应理论用于检验网络测验数据和纸笔测验数据的对等性[31],但是,这种方法只适用于小样本的二值计分情况,实际应用中局限性很大,总体上尚处于方法探索阶段。总之,关于网络测验和纸笔测验的测量不变性问题远没有一致的结论,相关研究只是刚刚起步,还需要大量的研究[32]。

1.2　在中国进行网络测验可靠性研究的重要性及现状

本研究将运用比较成熟的实证性因素分析第一次在中国研究网络测验和传统纸笔测验的测量不变性,即在多大程度上网络测验可以和纸笔测验进行同样的解释。在目前的中国进行这样的研究有着特别的重要性。首先,中国是一个网络大国,网民数量在过去十年中急剧上升,据最新统计,网民已达1.37 亿[33],网络调查和测评已经十分普遍,网络心理测验也发展迅速,已经有不少网站提供免费或收费的心理测验(比如:http://www.psy-test.net/;http://www.psych.gov.cn/tests/),但是,关于网络数据可靠性的研究却十分滞后。虽然已经有很多关于网络调查的可靠性的论文发表[34~39],但是,绝大多是都是在介绍和论述有关优缺点,多数是对国外相关研究和论述的转述,较为肤浅;极少有实证性研究,直接针对网络测验和纸笔测验测量不变性的研究尚未见发表。因此,开展相关研究在目前显得尤为迫切。

如前所述,国外已经有不少相关研究发表,但是这些研究基本都是在西方文化背景下进行的,研究结果未必能适用于中国。因为东西方文化有着巨大的差别,西方文化崇尚个人主义(individualism),而东方文化崇尚集体主义(collectivism)[40]。不同文化背景下的个体的心理和行为也有着巨大的差异,尤其是个体的自我差异巨大。西方个体的自我更多地与他人相对独立(independent),而东方个体的自我则更多地和他人相依(interdependent)[41];西方个体追求自我的一致性(consistency)[42~44],东方则更注重自我的灵活性(flexibility)[41,45];西方个体通常努力改变环境适应自我[46],东方个体则更多地改变自我适应环境[47];等等。而目前的网络测验绝大多数为自陈式测验。自陈测验在某种程度上就是要求个体通过自我报告的方式来表露自我。相对于西方个体,集体主义文化下的个体将对情境反应更为敏感,个体在作答的过程中更容易受外界环境变化的影响,容易受社会期许(social desirability)的影响,更不愿意真实地表露自己,在不同的情境下将更具有不一致性。和常见的纸笔测验情境相比,网络测验环境通常更为隐秘和多样化。那么,在中国文化环境下,网络情境下的测验性能将会如何?对此作深入的研究,将不仅在一般意义上拓展和丰富对网络测验和纸笔测验的测量不变性的理解,特别地,将对中国网络测验的发展和实践也有着重要的意义。

1.3　用实证性因素分析研究测量不变性的基本原理

我们知道,心理测量的基本原理是通过个体在测验项目上的表现来推测其对应的内部或潜在的心理结构的特性。通常,每一个潜在的结构或结构的维度由一个或多个具体的项目来测量。在实证性因素分析的框架下,测验项目和对应的潜在测量的结构之间的关系可以理解为观测变量(或称标识变量(indicator)或显变量(manifest variable))与对应的潜变量(latent variable)(或称公共因素(common factor))和残差(residual)之间的关系。假如一个测验有 n 个项目,对应着 r 个潜在心理结构,按照实证性因素分析模型,上述关系可以表述如下:

$$X_m = \lambda_{mp}\xi_p + \delta_m \quad (1)$$

其中,X_m 为观测变量或标识标量,$m = 1, 2, \cdots, n$,ξ_p 为潜变量(latent variable)或公共因素(common factor),$p = 1, 2, \cdots, r$,δ_m 表示残差(residual),包括特异因素(unique factor)和测量误差(measurement error),λ_{mp} 表示观测变量 X_m 在潜变量 ξ_p 上的回归系数或因素负荷(factor loadings)。这样测量的观测分数和真分数之间的关系就转为观测变量和潜变量或公共因素之间的关系。假如采用矩阵,上述公式可以表达如下:

$$X = \Lambda\xi + \delta \quad (2)$$

其中,X 是一个($n \times 1$)的列向量,表示第 i 个人在 n 个观测变量上的得分,Λ 是一个($n \times r$)的因素负荷矩阵,表示 n 个观测变量在 r 个公共因素上的负荷,ξ 是($r \times 1$)列向量,表示第 i 个人在 r 个潜变量的因素得分,δ 为一个($n \times 1$)列向量,表示第 n 个观测变量不能被潜变量解释的测量残差。

需要提醒的一点是,方程式(1)和(2)中所有的观测变量都是中心化后的离差分数(即与均数之差),相应地,潜变量也是中心化的变量,这相当于剔除了均数的影响。如果要考察测验在不同的情境下是否有相同的测量起点,则需要把均数考虑在内。当把观测变量和潜变量的均数都包括在内时,即用

原始的观测分数来标识潜变量时，方程(2)则可以重新表述为：

$$X = \Gamma + \Lambda(\alpha + \xi) + \delta \quad (3)$$

其中，Γ 为一($n \times 1$)列向量，表示由潜变量预测观测值时的截距，α 为一($r \times 1$)列向量，表示潜变量的均数，其他符号和前述相应符号含义相同。

当测量应用于多个不同的组时，假如用 g 表示组别，方程(3)则可以写为：

$$X^g = \Gamma^g + \Lambda^g(\alpha^g + \xi^g) + \delta^g \quad (4)$$

在两个组的情况下，则对于每一组，观测分数都可以同样地进行线性表达：

$$X^g = \Gamma^{g1} + \Lambda^{g1}(\alpha^{g1} + \xi^{g1}) + \delta^{g1} \quad (5)$$

$$X^g = \Gamma^{g2} + \Lambda^{g2}(\alpha^{g2} + \xi^{g2}) + \delta^{g2} \quad (6)$$

假如我们要对某一心理结构进行跨组或情境的比较时，尽管我们感兴趣的是真分数或潜变量的变异，但是，实际上通常是通过观测分数的比较来实现的。由于我们的比较通常是基于观测分数，这样，只有当方程(5)和(6)中的相应参数对等时，观测分数的均数和变异的跨组差异才可以真实地反映对应的潜变量的均数和变异的差异，或者说，只有测量具有跨组的不变性时，基于观测分数的比较才是合适的和可靠的。

在实证性因素分析的框架下，检验一个测量在两个不同的组中是否具有不变性，通常要对上述方程涉及的 4 个成分的对等性进行跨组检验，这包括：潜变量 ξ 的构成或模式是否不变；观测变量在潜变量上的因素负荷是否不变，即是否有 $\Lambda^{g1} = \Lambda^{g2}$；由潜变量预测观测变量的截距 Γ 是否不变，即是否有 $\Gamma^{g1} = \Gamma^{g2}$；残差 δ 的变异(variance)是否不变，如果用 Θ 表示残差的变异，即是否有 $\Theta^{g1} = \Theta^{g2}$。根据检验的对象的不同，测量不变性由低到高构成四个水平：结构不变性(configural invariance)，弱不变性(weak invariance)，强不变性(strong invariance)和严格不变性(strict invariance)[48]。

结构不变性 结构不变性是最基本的不变性，如果在不同组之间，潜变量的数目相等并且都是由同样的项目来测量，或由同样的标识变量标识，即 ξ^{g1} 和 ξ^{g2} 具有对等的因素结构(factor structure)，则表明测量在不同组之间具有结构不变性。结构不变性只要求潜变量、显变量之间的基本结构关系对等，不要求对应参数相等。结构不变性的确认意味着同一测量在不同组内反映了类似的心理结构。但是，即使有了结构不变性，并不意味着测量的结果就可以进行有意义的相互比较，因为，项目对潜变量的重要性在不同组之间不一定相同，使得潜变量的含义未必一致。一旦结构不变性得到确立，我们就可以继续检验是否存在更为高级的不变性：弱不变性。

弱不变性 弱不变性检验主要是检验不同组之间的因素负荷是否相等，即是否有 $\Lambda^{g1} = \Lambda^{g2}$。如果每一个显变量在对应潜变量上的负荷在不同组之间都相等，则表明，测量的显变量和潜变量之间具有对等的关系，或者说每一个显变量在不同的组之间具有相同的单位，潜变量每变化一个单位，显变量在不同组中都会产生相同程度的变化，这样潜变量在不同组间可以同样地被测验项目定义，因而将有着相同的含义；反过来，每个测验项目在不同的组之间也有着相同的含义。因此，弱不变性有时又称为单位不变性(metric invariance)。如果测量弱不变性不成立，通常有两种可能。一是测验的某些或全部项目在不同组之间具有不同的含义，另一种可能是某些项目在某一组或多组存在系统的反应偏差。

强不变性 强不变性检验主要是检验不同组之间观测分数在由潜变量预测时截距是否相等，即是否有 $\Gamma^{g1} = \Gamma^{g2}$。弱不变性的确立只是表示测量在不同的组之间具有单位对等性，因素或潜变量具有相同的含义，但是，量表的参照点却可能不一致。一个量表即使具有跨组的弱不变性，具有相同的单位，但是如果没有相同的参照点，不经过转换，分数依然是不能直接比较的。因此，强不变性意味着测量在不同组之间具有对等的参照点，这样，估计的因素得分将是无偏的，并且，观测分数的跨组差异将可以完全反映所测量的潜变量的跨组差异，也就是进行跨组的均数比较是有意义的。如果强不变性不能得到确认，则有两种可能，一种可能是测验在不同组之间存在系统的反应偏差[49]，比如，中国人说“好”可能就相当于美国说“很好”；另一种可能是反映了具有理论意义的差异。前一种可能通常是不好的，需要尽可能控制和排除，因为会污染观测分数；但后一种可能是反映了真实的差异，是合乎预期的。比如，在自我评价的测量中，西方人在负性表述项目上(比如，“我觉得我没有什么值得自豪的”)的得分通常比东方人高，其实反映了不同文化下的个体的认知差异[50]，是有理论意义的。因此，不具有强不变性并不总是不好的，通常需要根据具体情况具体讨论。

严格不变性 严格不变性主要是检验测验的每一个项目在不同的组间残差是否具有相同的变异，即是否有 $\Theta^{g1} = \Theta^{g2}$。由于观测分数的变异通常有两部分组成：旨在测量的潜变量的变异和与其无关

的残差变异,所以如果测量的潜变量的残差的变异具有对等性,则观测分数的变异的跨组差异完全反映潜变量的变异差异。考虑到严格不变性是在强不变性的基础上进行的,因此,严格不变性的确立意味着观测分数的均数和变异的跨组差异完全由潜变量的均数和变异的跨组差异决定。

在统计上,四个水平的不变性具有层级嵌套关系,后一水平总是嵌套在前一水平上。反过来,只有在低一级的不变性得到证实以后,进行高一级的不变性检验才有意义。结构不变性是进行跨组比较的最基本的要求,通常是作为不变性检验的基线模型(baseline model)而存在;弱不变性的建立使得测量在不同的组之间具有对等的单位;强不变性的建立表明测量在不同组之间不仅具有对等的单位,还具有对等的参照点,潜变量的均数的跨组差异完全可以由显变量表现出来;严格不变性的建立则表明观测变量的跨组差异完全可以由公共因素解释。很多时候我们需要对潜变量的均数进行跨组的检验,这样的检验要求测量至少具有强不变性,否则,差异的含义难以解释。

在实践中,基线模型确立后,高一级的不变性并不总是能够得到证实。一旦高一级的不变性不成立,有时人们会终止不变性检验,但有时会进一步寻找导致不变性不成立的原因,并探索是否存在部分不变性(partial invariance),然后在此基础上进一步检验更高一级的不变性[51]。部分不变性可以存在于基线模型以上的任何一个水平上,即可以有部分弱不变性(partial weak invariance)(即部分因素负荷具有不变性),部分强不变性(partial strong invariance)(即部分截距具有不变性),部分严格不变性(partial strict invariance)(即部分残差变异具有不变性)。

1.4　目前的研究

本研究的主要目的是研究在中国文化环境下,网络测验和传统的纸笔测验是否具有测量不变性。我们将以生活满意度量表作为研究对象,通过探讨生活满意度量表(Satisfaction with Life Scale,SWLS)在网络情境下和纸笔情境下的测量不变性来探讨网络测验的测量不变性。生活满意度量表由 Diener 等于 1985 年编制[52],迄今已在包括中国在内的全球 150 多个国家应用过,被广泛地证明具有良好的信度和效度[53]。我们选择生活满意度量表,主要出于以下几方面的考虑:1)鉴于大多数网络测验属于人格和社会心理测验,因此所选量表应该具有常见的人格和社会心理测验的典型特征。生活满意度量表和大多数人格和社会心理测验一样,采用 Likert 量表评分,通过自我报告来完成,以各项目观测分数的总和作为量表总分,符合经典测量理论的真分数模型等。2)基于前述东方个体自我的特点,量表测量内容应该是与个体自我密切相关,这样个体的反应将会对情境较为敏感,此时检验测量不变性将更有实际意义。3)所选量表结构最好相对简单清晰,这样可以尽可能地降低由测验本身结构不明确性导致的误差,从而使得由不同测验实施模式导致的效应最好地表现出来。

2　研究方法

2.1　被试

参加纸笔测验的被试共 284 人,参加网络测验的大学生被试为 419 名。纸笔条件被试全部来自中山大学,平均年龄为 20.20 岁,范围为 18 岁 ~24 岁,标准差为 0.81;网络数据收集时间为 2007 年 3 月 14 日至 3 月 24 日,被试来源包括广东、北京、上海、四川、浙江、湖南、湖北、陕西、重庆等全国 26 个省份/地区,平均年龄为 21.31 岁,范围为 18 岁 ~24 岁,标准差为 1.38。

2.2　量表

本研究采用的生活满意度量表共有 5 个项目,分别是:1)在大多数情况下我的生活是接近我的理想的(swls1);2)我的生活条件非常好(swls2);3)我对我的生活感到满意(swls3);4)到目前为止我已经得到了生活中我想得到的重要东西(swls4);5)如果可以再活一次,我还愿意过现在这样的生活(swls5)。问卷要求被试在一个 7 点量表上表明自己对上述各陈述的同意程度,“1”表示非常不同意,“7”表示非常同意。这样,整个量表分数分布范围为 5 ~35 分。

2.3　数据采集

纸笔测验的采集:纸笔测验在中山大学完成,全部被试来自选修某选修课的学生。施测时分 3 个班同时在教室进行,每个班大约 100 人,测验为匿名,个人信息仅要求提供性别和年龄。问卷完成后,再输入计算机。最后有效数据为 284 名。

网络数据的采集:网络数据是通过一个专业数据收集网站完成的。为了吸引被试,在开始的指导语中明确告诉被试,测验后将可以马上得到一个简单的免费反馈,如果留下邮件地址,整个研究完后将可以得到个人的结果和一个简单的解释。测试完后,马上给被试关于该测验以及生活满意度的简单

介绍，并谢谢被试的参与。测验同样为匿名，个人信息要求提供年龄、所在地区、是否在校大学生等。测验网页面向全体网民，数据收集集中在 2007 年 3 月 14 日至 3 月 24 日之间，共有 418 名在校大学生完成了测验。

3 结果

3.1 数据的有关描述性统计

网络测验和纸笔测验的信度分别为 $\alpha = 0.84$ 和 $\alpha = 0.75$，两组的均数分别为 20.47 和 19.46，方差分别为 5.85 和 4.82，两组的变异没有显著差异，$F(1,702) = 1.34$，$p > 0,05$，但是网络组的得分显著地高于纸笔组，$t = 2.29$，$p < 0.05$。项目均数及项目间的相关见表 1.

3.2 测量不变性的检验

基于引言部分的有关论述，接下来我们逐步对网络数据和纸笔测验的结构不变性、弱不变性、强不变性和严格不变性进行检验。检验用的软件是 Mplus3.12，估计方法为极大似然法。

表 1 项目均数及项目相关矩阵

项目	网络组		纸笔组		相关矩阵				
	均数	标准差	均数	标准差	swls1	swls2	swls3	swls4	Swls5
Swls1	4.177	1.417	3.982	1.400		0.43	0.53	0.29	0.43
Swls2	4.391	1.383	4.120	1.541	0.56		0.44	0.21	0.29
Swls3	4.456	1.350	4.306	1.471	0.65	0.63		0.42	0.45
Swls4	3.465	1.567	3.028	1.608	0.50	0.37	0.55		0.35
Swls5	3.981	1.718	4.028	1.704	0.53	0.42	0.57	0.48	

注：表中所有相关都非常显著（$p < 0.01$），其中，左下三角部分为网络组，右上三角部分为纸笔组。

3.2.1 结构不变性——基线模型 首先我们运用多组实证性因素分析检验量表是否具有结构不变性，同时也为下一步检验设定基线模型。在基线模型中，对每一个组，所有从潜变量到显变量的负荷以及残差或独特因素都为自由参数（见图 1）。但是，由于两组中的负荷参数和潜变量的方差具有不确定性（indeterminacy），因素负荷的值的大小取决于潜变量的定义量表。因此，为了能顺利估计因素负荷大小，必须先设定因素的量尺。我们依据 Reise 等的建议[48]，1）把纸笔测验组设为参照组；2）把参照组的潜变量标准差设为 1，均数设为 0（即标准化），而另一组的相应参数都设为自由估计参数；3）把两组的第一个项目的负荷设为相等；4）两组的第一个显变量与潜变量的线性模型中的截距（intercept）设为相等。这样，所有的参数都锚定在一个稳定的量尺上，都和第一组标准化的潜变量相比。在此基础上再进行多组实证性因素分析。结果表明，模型拟合良好，拟合参数见表 2 中的模型 1。这表明，生活满意度量表在两组间具有结构不变性，即网络组的生活满意度和纸笔组的一样可以由五个同样的项目来标志。结构不变性的建立为接下来更为严格的不变性检验设定了基线模型，将作为严格的不变性模型的比较标准。

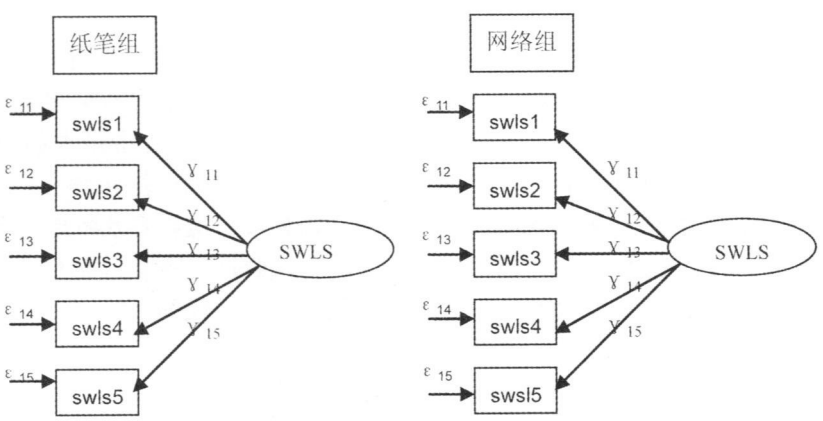

图 1 基线模型

3.2.2　弱不变性　基线模型仅仅表明网络环境和纸笔测验用同样的项目是合理的,这里,我们将进一步检验作为潜在变量的生活满意度在各项目上的负荷是否一致,也就是在不同的施测方式下,潜变量是否在同样程度上预测着显变量,或者说,各项目受所测量的潜变量的影响是否在不同组之间具有不变性。为此,我们在基线模型的基础上,设定不同组之间的所有负荷都相等,即 Λ_1 等于 Λ_2。这样,我们得到的模型拟合参数见表 2 中的模型 2。首先,我们可以看到该模型依然拟合良好;由于该模型是模型 1 的嵌套模型,拟合的差异直接反映模型改变是否合理。和基线模型相比,$\Delta\chi^2 = 2.02$, $df = 4$, $p > 0.05$,模型拟合并没有显著的改变,表明我们设定因素负荷在两组间相等是合理的。这样,不仅潜变量在不同的组间都可以有同样的项目测量,并且和各项目间的关系也具有对等性,潜变量改变一个单位,对外显项目的影响在不同组之间是一样的。

3.2.3　强不变性　在弱不变性得到确立的情况下,我们进一步检验不同的量表在不同的组别中是否具有强不变性,各显变量在由潜变量预测时截距是否相等。如果两个组的测量模型的负荷和截距都相等,则表明所测量的心理结构的潜变量在不同组之间的均数差异完全可以显变量的均数来表示。如果强不变性不成立,即各截距在不同组之间可能不等,这表示不同情况下可能存在系统的反应偏差(response bias)[48]。这样,在模型 2 的基础上,我们进一步设定两组模型所有的截距都相等。模型拟合见表 2 中的模型 3。虽然模型本身拟合良好,但是,和模型 2 相比,拟合却是显著地差,$\Delta\chi^2 = 14.7$, $df = 4$, $p < 0.05$,可见,不是所有的截距都相等,数据不支持完全的强不变性。根据 Vandenberg 和 Lance 推荐的程序[51],此时,我们可以继续寻找导致强不变性不成立的原因,检验是否存在部分强不变性。从和截距相关的修正指数可以看出,第四题在两组中修正指数都最大,按照 Reise 等介绍的方法[48],我们把两组的第四题的截距都设定为自由估计,得模型 4。重新估计的结果显示,所有的拟合指数都有提高,此时和模型 2 相比,模型拟合没有显著改变,$\Delta\chi^2 = 6.71$, $df = 3$, $p = 0.08$. 这样,除了第四题以外,其余题目的截距都具有不变性。或者说,除了第四题外,对其他的项目反应在不同组之间不存在系统偏差。

3.2.4　严格不变性　基于模型 4,最后我们继续检验严格不变性,即模型中残差方差的不变性。我们把两个组中的模型残差方差设为相等,得到模型 5。估计结果显示,模型拟和不好,和模型 4 相比,模型显著地差,$\Delta\chi^2 = 53.4$, $df = 5$, $p = 10^{-11}$,意味着至少不是所有残差方差都具有不变性。我们继续探索是否存在部分严格不变性。根据修正指数,逐步将第 2 题、第 3 题、第 4 题的残差方差在两组中设定为不等,最后得到模型 6。重新估计的结果显示,该模型拟合良好,和模型 4 相比,变化不显著,$\Delta\chi^2 = 3.57$, $df = 2$, $p = 0.17$。这样,部分严格不变性得到确认,项目 1 和 5 的残差方差具有不变性。

表 2　不同模型的拟合指数

模型	χ^2	df	χ^2/df	RMSEA	TLI	CFI	$\Delta\chi^2$	Δdf	p
1	30.344	10	3.034	0.076	0.965	0.983			
2	32.363	14	2.312	0.061	0.977	0.984	2.019	4	0.7321
3	47.078	18	2.615	0.068	0.972	0.975	14.72	4	0.0054
4	39.076	17	2.299	0.061	0.978	0.981	6.713	3	0.0817
5	92.491	22	4.204	0.095	0.945	0.939	53.42	5	10^{-11}
6	42.646	19	2.245	0.06	0.979	0.98	3.57	2	0.1678

注:模型 1:基线模型,纸笔组为参照组;模型 2:弱不变性模型,在模型 1 的基础上,设因素负荷在两组间相等;模型 3:强不变性模型,在模型 2 的基础上加设截距在两组间相等;模型 4:部分强不变性模型,在模型 3 的基础上设项目 4 的截距为自由参数;模型 5:严格不变性,在模型 4 的基础上,设残差方差在两组间相等;模型 6:部分严格不变性,在模型 5 的基础上,设两组中项目 2、3、4 的残差方差为自由参数。

3.3　潜变量均数和变异的跨组比较

通常有两种方法可以检验不同组之间的潜变量的均数和变异的差异。一种方法是直接通过比较潜变量均数的估计值和它的标准误进行检验,因为,研究两组中的一组,即纸笔测验组的潜变量已经被设为均数为零、标准差为 1,另一组的对应值都是和他们相比较而获得的;另一种方法是,通过把两均数限制为相等和不等从而比较不同模型间的拟合参数的差异来进行,如果差异显著,则表明二者差异显著,否则,则可认为没有显著性差异。由于两种方法结

果通常是一致的，参照 Reise 等的做法[48]，这里我们选用第一种方法。我们的检验以部分部分严格不变性模型为基础。在模型 6 中，我们发现，网络测验对应的潜变量的均数为 0.151，标准误为 0.089，二者之比，即 z 值为 1.691，$p > 0.05$，不显著，表明网络测验对应的潜变量和纸笔测验对应的潜变量之间均数是一致的。对于网络测验对应的潜变量的方差，均数为 1.206，标准误为 0.161，95% 的置信区间为（0.890，1.521），1 被包括在其中，表明网络组和纸笔组之间潜变量的变异差异不显著。

3.4 参数估计评估

通过前面的分析显示，网络测验和纸笔测验所获得的数据最后达到部分的严格不变性。具体地，测验项目在不同组之间对应着类似的潜变量，并且有类似的单位，5 个项目中的 4 个在有潜变量预测时具有相同的起点，5 个项目中的两个具有相同的残差变异。基于最后的部分严格不变性模型，有关参数估计见表 3。

表 3 基于模型 6 的参数估计

	截距				因素负荷		残差方差			
	参数		标准误		参数	标准误	参数		标准误	
变量	纸笔	网络	纸笔	网络			纸笔	网络	纸笔	网络
swls1	4.009		0.073		0.992	0.062	0.881		0.062	
swls2	4.219		0.071		0.879	0.062	1.646	0.972	0.154	0.078
swls3	4.296		0.077		1.081	0.065	0.937	0.425	0.110	0.057
swls4	3.334	3.028	0.086		0.868	0.066	1.932	1.482	0.178	0.112
swls5	3.908		0.082		1.022	0.072	1.745		0.107	
因素的均值和标准误										
公共因素										
均数	0		0.151		0	0.089				
方差	1		1.206		0	0.161				

从表 3 中可以看出，所有估计的截距非常显著，至少在 30 个标准误之外。同样，估计的因素负荷至少在 13 个标准误之外，说明测量的潜变量对项目的预测能力都很高；残差也都非常显著的大于零，至少在 7 个标准误之外，表明各项目并不能完全由潜变量预测，残差不可忽视；并且，对于三个变异不具有不变性的残差，基于网络数据的残差变异都要显著地小于纸笔测验。

4 讨论

本研究采用跨组的实证性因素分析，以生活满意度测验为例，对网络测验和纸笔测验的测量不变性进行了研究。结果发现，网络测验和纸笔测验具有完全的结构不变性和完全的弱不变性，部分的强不变性和部分的严格不变性。潜变量的均数和变异都不存在显著性差异，这些结果表明测验在网络和纸笔两种不同的实施模式下测量特性既有相同的地方、又有不同的地方，不同模式的模式内的测验分数的个别差异和组别差异可以进行类似的解释，但是，在模式间进行比较时需要考虑反应偏差等。

4.1 网络测验和纸笔测验的测量不变性

本研究中，以生活满意度为例的网络测验和纸笔测验具有完全的弱不变性，每一个项目和对应的潜变量之间都具有对等的关系，具有完全对等的单位，也就是说测验分数每一个单位的变化在网络实施模式下和纸笔测验有着相同的含义。这表明，相同的分数差异在网络测验和纸笔测验上可以进行同样的解释。比如说，20 和 25 之间的差异，在网络版和纸笔版上是意味着同样的差异。但是，尽管如此，这并不意味着可以把网络测验和纸笔测验的分数进行直接的比较。比如，网络测验的 20 分并不一定和纸笔测验的 20 分对等，因为，量表在两种条件下未必有着同样的量表起点。比如在生活中，我们计算年龄，有时用虚岁，有时用实岁，如果大家都用虚岁或都用实岁，则不同的人的年龄是可以比较的，但是，如果一个人用实岁，一个人算虚岁，则两个人的年龄不可以直接比较，因为按照两种不同算法得到的年龄虽然单位相同，但是起点不一样。不同实施模式下测验分数要可以直接对比，二者必须要有对等量表起点，即要具有强测量不变性。

研究还发现，测验在两种实施模式间只表现出

了部分的强不变性,5 个项目中的 4 个具有相同的起点,第 4 个项目"到目前为止我得到了生活中我想得到的重要东西"在网络和纸笔两个不同的实施模式下有着不同的截距,分别为 3.33 和 3.03。由于测验分属通常是各个项目分数之和,这样,在不同版本间的分数的差异就不是完全反应了对应的潜变量之间的差异,或者说不是完全由潜变量的变化引起。前面说过,这种差异有可能是纯粹的反映偏差,也有可能是由其他有意义的因素导致。本研究中,这种差异可能反映了社会赞许性导致的谦虚反应倾向。因为,本研究中网络测验的隐秘性比集体的纸笔测验要好,纸笔测验是集体施测,大约 100 人一个班,虽然测验也是匿名,这种集体的环境很可能导致被试相对更为谦虚的反应。

强不变性意味着观测分数的跨情境差异可以反应对应的潜变量之间的差异,严格的不变性则意味着观测分数的均数和变异的跨情境差异完全是有潜变量引起,因为在不同的情况下测量的残差变异都是对等的。如果只有部分的严格不变性,则意味着某些项目的残差在不同的实施模式下包含了不同的特异因素。本研究中,第 2 题(我的生活条件非常好)、3 题(我对我的生活感到满意)、4 题(到目前为止我已经得到了生活中我想得到的重要东西)的残差不等,纸笔测验在这些题目上的残差都要大于网络测验。这可能反映了在不同实施环境下个体对情境的不同反应。本研究中的纸笔测验是集体施测,在这么多人的公共环境下,由于中国人的自我是情境依赖性的,环境的噪音在很大程度上反映在测量的残差里。而网络测验则相对地要更为隐秘,个体受环境的影响应该更为少,所以,最终测验分数里不能由旨在测量的结构解释的残差要小。这说明网络测验在某种程度上要优于集体施测的纸笔测验。

最后,我们还检验了网络测验和纸笔测验所测量的潜在变量的均数和变异的跨组差异。结果显示,均数不存在显著性差异,这不难理解,因为本研究中的对象都是大学生,所以具有类似的生活满意水平。在引言部分,我们已经讲过,测验观测分数的跨组差异未必反映旨在测量的潜在结构的差异,很可能还反映了反应的偏差等其他无关因素的干扰。在本研究中,如果仅看观测分数,网络组的均数要显著地大于纸笔组,但是,如果看潜变量,则两组的均数不存在显著性差异。因为观测分数均数的组间差异不仅包括了由潜变量导致的差异,还包括了量表参照点的差异。从前面的结果可以看出,测验在两组间并不具备完全的强不变性,第 4 题的量表起点有显著的不同,当考察排除了这种起点差异后的潜变量的均数差异,就发现两组间的差异是不显著的。可见,本研究中观测分数的显著跨组差异并不代表着对应的心理结构的真实差异,如果我们依据观测分数做出最后的推断,则会得出貌似正确但实际上是错误的结论。这也说明,在没有确认测量不变性的情况下,进行基于观测分数的跨组或跨情境比较将可能导致错误的结论。

总之,本研究中网络测验和纸笔测验至少具有弱不变性,即具有相等的单位和含义,部分的强不变性和部分的严格不变性则可能反映了测验实施环境的变化的影响。这些都显示,在集体主义文化下,测验环境是影响测验结果的一个重要因素,而对来自不同样本的测验结果进行比较时,也必须非常小心。收集测验数据,最好不要进行集体施测。

本研究对四个水平的不变性都进行了检验,理想的情况应该是在四个水平上都存在不变性。但是,在实际应用中,未必需要测验满足全部的不变性要求,具体要视研究的目的和要求而定。如果研究只是想知道测验在不同的情境或应用于不同的对象时是否具有相同的单位或含义,或者,不需要进行跨组的比较,则只需要弱的不变性就可以了。如果要进行均数的跨组比较,则需要强不变性。如果还要进行跨组的变异的比较,则需要严格的不变性。在大多数情况下,对测验只要求具有弱不变性,至多严格不变性,因为,多数的研究只需要进行模式内的比较,少数进行均数的跨组差异检验,极少需要对分数的变异进行跨组的比较。

4.2　研究的意义和局限性　本研究是第一个在中国,也是第一个在集体主义文化下进行的关于网络测验和纸笔测验的测量不变性的研究,因此将有着重要的意义。首先,本研究表明,在集体主义文化下,网络测验和纸笔测验具有相当程度的测量不变性,具有相等的单位,测验分数的差异有着相等的含义,在某种程度上讲,网络测验更为可靠;同时,研究也揭示,测验实施环境对测验结果可能会带来影响,这不仅将在一般意义上丰富了对网络测验和纸笔测验的不变性、网络测验可靠性的理解,还将对中国网络测验的发展实践有着重要的指导意义。

本研究提示,至少和传统的集体施测的纸笔测验相比,网络测验是可靠的,甚至更好。鉴于利用网络收集数据的诸多优势,我们完全可以利用网络收集数据开展有关研究。国外已经有很多利用网络收

集的数据开展的研究发表，还有大量进行各种测评服务的网络测验提供。在国内，已经有不少网站提供网络测验服务，但是，利用网络数据开展研究非常少，目前只有一例研究报道[54]。期待在不久的将来，在国内能看到越来越多的基于网络数据的研究发表。

但是，本研究也提示，测验在不同的实施模式间得到的结果可能不完全对等。由于国内绝大多数心理测验的常模都是以纸笔测验为基础制定的，在没有确认网络测验和纸笔测验具有强不变性之前，即确认量表有相同的单位和起点之前，把网络测验获得的分数和基于纸笔测验的常模相比较很可能是不恰当的。

由于心理学研究中进行跨组或跨情境的比较是很常见的，比如，不同性别、不同文化、不同处理之间的比较等，本研究显示在进行跨组或跨情境的比较之前，对测验工具进行测量不变性的检验是非常有必要的。只有建立了相应的测量不变性，比较才是有意义的。否则，结果很可能是有偏的，甚至可能是完全错误的。在国内，近十年来心理学发展非常迅速，心理测验在科学研究和实际生活中得到了越来越广泛的运用，虽然国内已有少量关于测量不变性的介绍文章出现[55-56]，但是，却少有实证研究关注测验在不同情境或样本下的测量不变性，大家似乎依然默认同一个测验在不同的情况下理所当然地具有测量不变性，同一个测验用在哪里都可以进行同样的解释。本研究表明，这是不对的，测验环境在一定程度上影响着测验的特性，测验未必总是有测量不变性，基于观测分数得到的结果很可能是误导的。希望本研究能引起中国的心理学研究者和实践者对测验在不同情境下应用时的测量不变性的重视和研究兴趣，促进心理测验的科学运用。

此外，不同于国内已有的少量介绍[55,56]，本文从一个新的视角并以尽可能通俗的语言对测量不变性的概念、原理、检验程序及结果解释等都作了比较全面的介绍，并详细展示了如何在实践中进行不变性研究，相信这些介绍在方法上对国内未来的研究者开展测量不变性的研究也会有一定的帮助。

最后要指出的是，虽然本研究具有重要的意义，但不能不注意到其局限性。第一，本研究根据理论预期证实了测验在网络和纸笔模式下具有弱不变性，但是只发现了部分的强不变性和部分的严格不变性。从理论上讲，部分不变性的探索过程是一种事后的检验，所获结果是否具有跨样本的可推广性尚有待进一步探讨。对此，Bryne 等曾建议进行交叉效度（cross－validation）的检验；但是，他们同时也指出，交叉效度的检验不是万能药，也不能保证绝对的可推广性，并且，由于样本所限很多时候未必能实现[57]。实际上，在实践中，包括 Bryne 本人等的研究在内的大多数不变性的研究在涉及部分不变性时都没有进行交叉效度的检验[48,51,57]。类似地，本研究也没有进行这样的检验。第二，本研究只是以生活满意度量表为例，只考察了人格和社会心理的测验，没有考察能力测验、成就测验等等，对于其他测验情况如何，这有待更多的研究来回答。最后，本研究的样本不是完全随机对等，研究揭示的某些不对等性不单是反映了测验模式本身的影响，还在某种程度上可能反映了样本的不对等性的影响，因此，要想把各种不同的因素的影响予以区分，必须进行设计更为严密的研究。不过，尽管本研究基于不完全对等的样本，我们依然发现测验具有完全一样的因素结构（结构不变性），各项目在因素上的负荷也都一样（弱不变性），甚至大部分的截矩（部分强不变性）和部分的残差都一样（部分完全不变性），这也从另一个方面说明了网络测验和纸笔测验相当程度的对等性。因为如果样本真的是完全对等的话，可以预期不变性的水平将会更高。不过，理论上讲完全对等的样本通常只能在实验室得到，目前的取样应该更符合实际，所获结果也更能反映实际情况，具有更高的生态效度（ecological validity）。

参 考 文 献

1 Nosek B A，Banaji M，Greenwald A G. Harvesting implicit group attitudes and beliefs from a demonstration Web site. Group Dynamics，2002，6：101～115

2 Rentfrow P J，Gosling S D. The do re mi's of everyday life：The structure and personality correlates of music preferences. Journal of Personality and Social Psychology，2003，84：1236～1256

3 Srivastava S，John O P，Gosling S D. Development of personality in early and middle adulthood：set like plaster or persistent change? Journal of Personality and Social Psychology，2003，84（5）：1041～1053

4 Buchanan T. Potential of the Internet for personality research. In：M H Birnbaum（Ed.）. Psychological experiments on the Internet. San Diego，CA：Academic Press，2000. 121～265

5 Cook C，Heath F，Thompson R L，et al. Score reliability in Web－or－Internet－based surveys：Unnumbered graphic rating scales versus Likert－type scales. Educational and Psychological Measurement，2001，61：697～706

6 Davis R N. Web－based administration of a personality questionnaire：Comparison with traditional methods. Behavior

Research Methods, Instruments & Computers, 1999, 31: 572 ~577
7 Mezzacappa E. Letter to the Editor. APS Observer, 2000, 13: 10
8 Cronbach L J. Essentials of psychological testing. New York: Harper & Row, 1990
9 Buchanan T, Johnson J A, Goldberg L R. Implementing a five – factor personality inventory for use on the Internet. European Journal of Psychological Assessment, 2005, 21: 115 ~127
10 Buchanan T. Internet research: Self – monitoring and judgments of attractiveness. Behavior Research Methods, Instruments & Computers, 2000, 32: 521 ~527
11 Cronk B C, West J L. Personality research on the Internet: A comparison of Web – based and traditional instruments in take – home and in – class settings. Behavior Research Methods, Instruments & Computers, 2002, 34: 177 ~180
12 Epstein J, Klinkenberg W D, Wiley D, et al. Insuring sample equivalence across Internet and paper – and – pencil assessments. Computers in Human Behavior, 2001, 17: 339 ~346
13 Krantz J H, Ballard J, Scher J. Comparing the results of laboratory and World – Wide Web samples on the determinants of female attractiveness. Behavior Research Methods, Instruments, & Computers, 1997, 29: 264 ~269
14 Pasveer K A, Ellard J H. The making of a personality inventory: Help from the WWW. Behavior Research Methods, Instruments & Computers, 1998, 30: 309 ~313
15 Pettit F A. A comparison of World – Wide Web and paper – and – pencil personality questionnaires. Behavior Research Methods, Instruments & Computers, 2002, 34: 50 ~54
16 Pomplun M, Frey S, Becker D F. The score equivalence of paper – and – pencil and computerized versions of a speeded test of reading comprehension. Educational and Psychological Measurement, 2002, 62: 337 ~354
17 Smith M A, Leigh B. Virtual subjects: Using the Internet as an alternative source of subjects and research environment. Behavior Research Methods, Instruments & Computers, 1997, 29: 496 ~505
18 Stanton J M. An empirical assessment of data collection using the Internet. Personnel Psychology, 1998, 51: 709 ~725
19 Webster J, Compeau D. Computer – assisted versus paper – and – pencil administration of questionnaires. Behavior Research Methods, Instruments & Computers, 1996, 28: 567 ~576
20 Barak A, Cohen L. Empirical examination of an online version of the self – directed search. Journal of Career Assessment, 2002, 10: 387 ~400
21 Buchanan T, Smith J L. Using the Internet for psychological research: Personality testing on the World Wide Web. British Journal of Psychology, 1999, 90: 125 ~144
22 Gati I, Saka N. Internet – based versus paper – and – pencil assessment: Measuring career decision – making difficulties. Journal of Career Assessment, 2001, 9: 379 ~416
23 Pasveer K A, Ellard J H. The making of a personality inventory: Help from the WWW. Behavior Research Methods, Instruments & Computers, 1998, 30, 309 ~313
24 Mellenbergh G J. Item bias and item response theory. International Journal of Educational Research, 1989, 13: 127 ~143
25 Lubke G H, Dolan C V, Kelderman H, et al. Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. British Journal of Mathematical and Statistical Psychology, 2003, 56: 231 ~248
26 Meredith W. Measurement invariance, factor analysis, and factorial invariance. Psychometrika, 1993, 58: 525 ~543
27 Potosky D, Bolsko P. Computer versus paper – and – pencil administration mode and response distortion in noncognitive selection tests. Journal of Applied Psychology, 1997, 82: 293 ~299
28 Stanton J M. An empirical assessment of data collection using the Internet. Personnel Psychology, 1998, 51: 709 ~725
29 Fouladi R T, McCarthy C J, Moller N P. Paper – and – pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. Assessment, 2002, 9: 204 ~215
30 Donsonvan M A, Drasgow F, Probst T M. Does computerizing paper – and – pencil job attitude scales make a difference? New IRT analyses offer insight. Journal of Applied Psychology, 2000, 85: 305 ~313
31 Ferrando Pere J, Lorenzo – Seva, Urbano. Irt – related factor analytic procedures for testing the equivalence of paper – and – pencil and internet – administered questionnaires. Psychological Methods, 2005, 10(2): 193 ~205
32 Thompson L F, Surface E A, Martin D L, et al. From paper to pixels: Moving personnel surveys to the Web. Personnel Psychology, 2003, 56: 197 ~227
33 中国互联网络信息中心. 第 19 次中国互联网络发展状况统计报告. 2007 – 01 – 23
34 Fang J M, Shao P J, Su J, et al. A empirical study on response probability of web – based survey (In Chinese). Management Review, 2006, 18(10): 12 ~17
(方佳明,邵培基,粟婕等. 基于网络的问卷调查回复率影响因素实证研究. 管理评论, 2006, 18(10): 12 ~17)
35 Li R, Song T Y. An analyses of web – based survey in China (In Chinese). Information Science, 23(6): 891 ~895
(李锐, 宋铁英. 国内网络调查研究分析. 情报科学, 2005, 23(6): 891 ~895)
36 Li L. Some thoughts about web – based survey (In Chinese). Shanghai Statistics, 2002, 6: 27 ~28
(李岚. 对网络调查的思考. 上海统计, 2002, 6: 27 ~28)
37 Sun L, Cai L. A new method to study sensitive social issues (In Chinese). Forum of statistics and information, 2000, 13(2): 43 ~44
(孙蕾, 蔡亮. 敏感性问题的统计调查新方法——网上调查. 统计与信息论坛, 2000, 15(2): 43 ~44)
38 Wang J, Zhang Y Y. The characteristics, problems and solutions of web – based survey in China (In Chinese). Jinan Academic Journal, 2001, 23(2): 49 ~52
(王军,张云云. 我国网络调查的特点、问题及其对策. 暨南学报(哲学社会科学), 2001, 23(2): 49 ~52)

39 Yu W L. The challenge toward traditional survey from web – based survey (In Chinese). Information Methods, 2001, 11: 26 ~ 27
(郁伟龙. 网络调查对传统调查的挑战. 情报方法, 2001, 11: 26 ~ 27)
40 Oyserman D, Coon H M, Kemmelmeier M. Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta – analyses. Psychological Bulletin, 2002, 128(1): 3 ~ 72
41 Markus H R, Kitayama S. Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 1991, 98: 224 ~ 253
42 Feistinger L. A theory of cognitive dissonance. London: Tavistock, 1957
43 Heider F. The psychology of interpersonal relations. New York: John Wiley and Sons, 1958
44 Swann W B, Wenzlaff R M, Krull D S, et al. Allure of negative feedback: Self – verification strivings among depressed persons. Journal of Abnormal Psychology, 1992, 101: 293 ~ 306
45 Kanagawa C, Cross S E, Markus H R. "Who am I?": The cultural psychology of the conceptual self. Personality & Social Psychology Bulletin, 2001, 27: 90 ~ 103
46 Su S K, Chiu C, Hong Y, et al. Self organization and social organization: American and Chinese constructions. In T. R. Tyler, R. Kramer, & O. John (Eds.), The psychology of the social self. Mahwah, NJ: Lawrence Erlbaum, 1999. 193 ~ 222
47 Chiu C, Dweck C S, Tong J U, et al. Implicit theories and conceptions of morality. Journal of Personality & Social Psychology, 1997, 73: 923 ~ 940
48 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: 552 ~ 566
49 Bollen K A. Structural equations with latent variables. New York: John Wiley, 1989
50 Rodgers J, Peng K, Wang L, et al. Dialectical self and psychological well – being. Personality and Social Psychology Bulletin, 2004, 30: 1416 ~ 1432
51 Vandenberg R J, Lance C E. Are view and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 2000, 2: 4 ~ 69
52 Diener E, Emmons R A, Larsen R J, et al. The Satisfaction with Life Scale. Journal of Personality Assessment, 1985, 49: 71 ~ 75
53 Pavot W, Diener E. Review of the Satisfaction with Life Scale. Psychological Assessment, 1993, 5: 164 ~ 172
54 Wang J P, Xie W, Sun H W, et al. The effects of different coping strategy on people's behavior habits under the stressor of SARS (In Chinese). Chinese Journal of Clinical Psychology, 2004, 1: 41 ~ 44
(王建平, 谢伟, 孙宏伟等. SARS 应激下不同应对策略对人们行为习惯的影响. 中国临床心理学杂志, 2004, 1: 41 ~ 44)
55 Liu J, Wu W K. A Study of Measurement Invariance and a Numerical Case (in Chinese). Psychological Science 2005, 28 (1): 170 ~ 174
(刘军, 吴维库. 心理测量平衡性研究与实例. 心理科学, 2005, 28 (1): 170 ~ 174)
56 Bai X W, Chen Y W. Measurement equivalence: Concept and test conditions. Advances in Psychological Science, 2004, 12(2): 231 ~ 239
(白新文, 陈毅文. 测量等价性的概念及其判定条件. 心理科学进展, 2004, 12(2): 231 ~ 239)
57 Bryne B M, Shavelson R J, Bengt Muthen. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychological Bulletin, 1989, 105 (3): 456 ~ 465

Examining the Measurement Invariance between Paper-and-Pencil and Internet – Administered Tests in China

CAI Hua-Jian LI N Yong-Jia WU Qiu-Ping YAN Le HUANG Xuan-Feng
(*Department of Psychology, Sun Yat – Sen University, Guangzhou* 510275, *China*)

Abstract

Concerns about the quality of internet-based tests have been brought by the increasing applications of such tests in psychological research. Over the past years, a large body of studies has been conducted to examine the equivalence of internet-administered tests to their paper-and-pencil counterparts. Although studies based on Classic Test Theory (CTT) showed that internet-administered tests were trustable, studies based on measurement invariance tests produced mixed findings. What is more, most studies so far have been conducted in individualistic cultures. Given these,, the present study aimed to examine the equivalence between internet-based and paper-and-pencil tests in a collectivistic culture, particularly, in China. To this end, we employed Confirmatory Factor Analysis (CFA) to examine the measurement

invariance of the selected scale: Satisfaction with Life Scale (SWLS) across modes.

SWLS was administered via internet and the paper-and-pencil modes. Five items were rated on a 7-point likert scale ranging from 1 (" strongly disagree") to 7 (" strongly agree"). A total of 418 self-selected college students from 26 provinces in China took the internet-based test. And a total of 288 college students at Sun Yat-Sen University were sampled to take the paper-and-pencil test in classroom. For the internet sample, the age ranges from 18 to 24 years old with a mean of 21.31 (SD = 1.38); for the paper-and-pencil sample, the age ranges from 18 to 24 years old with a mean of 20.20 (SD = 0.81).

Multi-group CFA was employed to test measurement invariance between the internet administered and the paper-and-pencil SWLS. Results showed weak measurement invariance held across these two test modes, indicating metric similarity between the tests; partial strong measurement invariance and partial strict measurement invariance also held, suggesting that response bias existed in some items across modes; further analysis revealed that the paper-and-pencil test included more noise arising from administering environment.

In terms of mean comparisons, significant differences between modes were found in observed scores but not in latent scores. For the variances, no significant differences were found between modes in either latent scores or observed scores. These findings suggested that administering environments produced potential impacts on observed scores.

As the first examination of the measurement invariance in Chinese samples, the study provided initial evidence that internet-based tests have equivalent metrics with paper-and-pencil tests. Further, the results from the partial strong invariance and partial strict invariance may indicate the sensitivity of Chinese people to environments that may be resulted from collectivistic culture. Taken together, the findings from this study suggest that although internet-based tests are trustable in China, cautions of response biases should be kept in mind when conducting cross-groups (or modes) comparisons. Also, the findings underscore the importance of examining measurement invariance when a test is applied across multi-groups (or modes).

Key words　measurement invariance, Internet, paper-and-pencil, China, subjective well-being (SWB).