第三方公共服务绩效评价的评价:一项比较案例研究

马　亮　于文轩

摘　要　作为以目标责任考核为主的政府内部绩效考评体系的重要补充,第三方政府绩效评价自2000年以来发展迅速,但学术界对其研究却相对缺乏,对第三方政府绩效评估的发展、演进以及方法论和质量评估问题关注不够。本文对两项在中国影响较大、研究对象和内容近似的第三方政府公共服务绩效评价项目进行比较研究,并重点通过统计分析考察两项研究的信度,发现尽管两个项目的评价结果呈显著正相关关系,但仍存在一些重要的差别。通过对这些差别的比较和分析,对完善和提升第三方政府绩效评价质量并促进其健康发展提出了研究和政策建议。

关键词　第三方测评;公共服务绩效;政府绩效;公民满意度;信度

中图分类号　D035－3　　**文献标识码**　A　　**文章编号**　1001－8263(2013)05－0055－09

作者简介　马亮,新加坡南洋理工大学南洋公共管理研究生院研究员　新加坡 639798;于文轩,新加坡南洋理工大学公共政策与全球事务系助理教授　新加坡 639798

一、引言

公共服务绩效评价是加强政府问责,提升政府绩效的重要管理举措,是服务型政府建设的重要手段①。越来越多的政府将公共服务绩效纳入目标责任考核体系中,并通过公众满意度调查等方式获取公共服务绩效信息。较为典型的有政风行风测评、"万人评议"、"开门评议"等,为公众参与政府绩效评价提供了渠道,但也存在诸多问题②。由于测评通常将公众满意度与部门绩效挂钩,甚至进行末位淘汰和一票否决,一定程度上激励政府部门采纳逆向行为③。政府部门在开展公众满意度测评时也存在问卷设计不合理、问卷发放随意性强、问卷统计不恰当等方面的问题④。有学者认为,这反映了政府绩效评价的政治理性与技术理性之间的冲突,服务于政治理性的公众满意度测评牺牲了技术理性⑤。

在政府组织开展的公共服务绩效评价相对缺乏信度、效度和公信力的情况下,第三方政府绩效评价成为社会关注的焦点。人们普遍认为独立的第三方可以开展客观、专业的绩效评价,对政府绩效进行强有力的外部监督⑥。2012年7月国务院印发的《国家基本公共服务体系"十二五"规划》也指出,要开展全国、各地区和各行业的基本公共服务水平监测和评价,"积极开展基本公共服务社会满意度调查。鼓励多方参与评估,积极引入第三方评估。完善基本公共服务问责机制,增加基本公共服务绩效考核在政府和干部政绩考核中的权重"⑦。一些地方政府已经开始尝试委托第三方独立开展公共服务满意度测评,如兰州、北京、青岛、深圳、汉中等城市。还有越来越多的学术机构开始进行政府公共服务绩效评价,如华

南理工大学、新加坡南洋理工大学、中国社会科学院等⑧-⑬。

尽管第三方绩效评价被认为是中国政府绩效评价的未来发展方向之一，由于中国非营利组织(NGO)和公民社会发育相对不良，政府对第三方绩效评价存在潜在的干预和影响。第三方绩效评价是否可以自律，保持其独立性，采用科学的研究方法保证其评价的信度和效度，决定着其前途和命运⑭⑮。本文以两项针对中国主要城市公共服务绩效开展的第三方测评项目为案例，通过比较分析对其评价信度进行测评。我们的研究路径类似于“评价的评价”或元评价(Meta-evaluation)，即对评价结果的可靠性和有效性等的验证和检验⑯。我们希望我们对这两项典型第三方政府绩效项目的比较和分析，有助于我们对第三方政府绩效评价的质量和健康可持续性发展的进一步思考。

信度是政府绩效评价的一个重要质量指标，同效度、全面性、可理解性、功能性等共同构成了衡量“组织报告卡”等组织绩效评价体系的标准⑰。信度即可靠性，指对同一个对象进行重复多次测量的结果之间的一致程度⑱。就第三方公共服务绩效评价而言，信度指两个及以上的评价结果之间的一致程度⑰⑱。信度的测量有两个维度，一个是跨时间的信度测量，一般说来同一测量主体在相邻时间的测量结果应该呈现出一定的一致性。也就是说如果测评主体 A 对客体 X 的在相邻时间的测量结果的相关性高于测评主体 B 对相同客体 X 在同样相邻时间里的测量结果的相关性，测评主体 A 测评比测评主体 B 信度高；另一个是横截面的信度测量，也就是说如果测评主体 A 和测评主体 B 的测评都是客观科学的话，测评主体 A 的测评和测评主体 B 对同一客体的测评应该高度相关。如果两者间的测评有较大的差异，两个测评至少有一个信度较低。如果公共服务绩效测评的信度不高，也就是说测量的结果不是可靠的和稳健的，有可能导致各测评产生不同的结果，“公说公有理，婆说婆有理”，使评价对象接收到差别较大甚至自相矛盾的“绩效信号”，左右不是，无所适从，不仅无法指导政府绩效改进，而且可能产生对测评主体和测评活动的负面看法。因此，研究第三方政府绩效评价的信度对第三方政府绩效评估的发展有非常重要的意义。

本文以下部分安排如下：首先，我们在第二部分介绍两个第三方公共服务绩效评价项目的背景、评价方法和评价结果，并对二者的异同进行初步比较和分析。其次，我们对两项评价的结果进行定量分析，通过实证研究考察评价结果的稳健性和可比性，并探讨两项评价异同的背后原因。最后，我们对本文的研究贡献和政策启示进行讨论，并指出本文的研究不足和未来研究方向。

二、第三方公共服务绩效评价的两个案例

(一) A 大学服务型政府调查

A 大学自 2010 年受某慈善基金的资助，开始启动中国城市公共服务质量调查项目，对中国主要大城市的公共服务质量进行测评和排名。2011 年 A 大学将调查主题拓展为中国城市服务型政府建设，采用计算机辅助电话访问(CATI)技术进行随机电话访问调查，提高了抽样调查的代表性和精确度。该项目认为，单纯的公共服务质量无法完全涵盖服务型政府的真实内涵，因此将测评焦点转向服务型政府，对包括公共服务满意度、政府信息公开、公众参与、政府效能和政府信任等在内的诸多维度进行考察。该项目还独创性地提出了“三位一体”的评价体系，从公众视角、企业视角和客观视角等三个方面对服务型政府进行全景扫描。

(二) B 机构基本公共服务力评价

B 机构与国内某咨询公司合作的基本公共服务力评价项目从 2011 年启动，主要对中国 38 个大城市的基本公共服务提供能力进行评价。虽然 B 机构将其测评内容称为“能力”调查，但其调查内容仍然主要以公民对整个领域公共服务的满意度为主，因此可以归为公共服务绩效评价的范畴。B 机构的调查范围和样本量与 A 大学类似，但在调查方法方面以面对面问卷调查为主。B 机构主要专注于公民对公共服务质量的评价，并收集了各城市的客观公共服务供给数据，作为衡量各城市公共服务能力的主要标准。

(三)案例比较

表1对二者的调查要点进行了初步比较,从中我们可以发现二者有许多共同点,但也存在不少差异。

表1 第三方公共服务绩效评价的案例比较

比较项目	A大学项目	B机构项目
启动时间	2010	2011
资助来源	慈善基金	政府课题资助
覆盖城市	32个(2010-2011)、34个(2012)	38个
评价体系	主观(公民)+主观(企业)+客观	主观(公民)+客观
评价指标	各领域公共服务满意度(公民+企业)	各领域公共服务满意度(公民)
指标数量	75(2011)、68(2012)	84(2012)
加权方法	等权重、主成分分析	等权重和基于公众关注度加权
调查方式	面对面问卷调查(2010)、计算机辅助电话访问(CATI)(2011-2012)	面对面问卷调查
问卷题项	57(2012)	45(2012)
问卷选项	5级、10级(不包括"不清楚"、"拒答")	2个、4个、5个(不包括"不清楚")
样本量	28425(25222+3203,2011)、27529(23923+3606,2012)	19058(2011)、25115(2012)
原因分析	量化研究	个案研究
结果使用	论文、图书、新闻发布会	图书、新闻发布会
发表时间	每年11月	7月(2011)、12月(2012)

注:本表资料来源为两项调查的公开发表的研究报告和相关新闻报道。B机构2012年的三级指标数量为84个,但实际使用的仅为45个主观指标,客观指标未使用。

首先,两项调查都由学术机构主持完成,其中A大学与国内高校合作,B机构与咨询公司合作完成。人们一般认为第三方学术机构有专业知识储备,追求学术研究的独立性和客观性,是进行第三方政府绩效评估甚至政府绩效评估的最佳主体之一[19][20]。

两项调查启动时间相近,A大学启动稍早一年,B机构紧跟其后。两项调查的资助来源略有不同,A大学由非营利慈善组织资助,而B机构则依靠科研项目和合作伙伴资助。

两项调查关注的对象是中国的主要城市,以直辖市、省会城市和副省级城市为测评对象。在城市覆盖方面,A大学在2010-2011年包括4个直辖市、23个省会城市、5个计划单列市(非省会的副省级城市)以及1个地级市(苏州)。2012年未包括苏州但增加了西宁、银川和呼和浩特,覆盖城市34个,但囿于少数民族语言问题而未将乌鲁木齐和拉萨纳入调查。B机构的调查一直都涵盖38个城市,其中包括了2个非计划单列市的经济特区(汕头、珠海)。两项调查之所以都选择中国大城市作为调查对象,主要是它们的公共服务提供涉及面广、示范作用明显,影响力大、媒体关注度高且数据质量和可获得性较高,便于比较和分析。但在公共服务均等化和城乡公共服务一体化呼声日高的情况下,如何进一步涵盖更多的样本城市并确保其可比性,是一个有待研究的重要课题。

两项调查都采用主观与客观相结合的方式进行调查。B机构通过调查公民获取感知信息,通过政府统计资料获取公共服务供给的客观数据。A大学有所不同的是认识到企业也是公共服务的重要对象,将企业也纳入调查。尽管主客观结合的"双元综合评估"得到了学者们的推崇[21],但也存在一系列的理论和方法论问题,比如说公众满意度和公共服务供给之间并不一定存在对应关系,二者之间的显著相关性一直都无定论。此外两者测量的是不同构念(Construct),采用的是不同的度量指标,能否将其加总是一个值得商榷的问题[22][23]。更为重要的是,目前客观数据获取存在问题,通常来说滞后1-2年,也就是说,将不同年份的主观和客观数据汇总,也存在时间归因偏误的问题。只是考虑到客观投入转化为主观满意度需要一定的时滞,这种方法是有一定道理的,学界也通常倾向于选择采用最近年份的数据作为替代。但是,这种数据处理和汇总方式仍然使跨年比较较难,因为不同年份的数据汇总在一起已经很难明确说明究竟是数据年份不同导致的差异还是公共服务绩效本身发生了变化。更为重要的是,由于客观数据获取方面存在严重滞后、公布不全、统计口径不一致等方面的问题,通常很难将其与主观数据汇总分析,这也是为什么B机构在2012年没有将其作为最终评估依据,而仅采用主观感知指标衡量的原因所在。

对指标进行加权汇总方面,两项调查的做法既有相似之处,也存在差异。B机构采用等权重法进行对底层测量指标的数据进行汇总,而A大学除了使用等权重的方法以外,对三个以上指标

进行了主成分分析。在更高维度上进行汇总方面，A 大学先采用的是主成分分析法，然后在最后三个大维度（公民、企业和客观）采用了等权重法。B 机构调查了公民对不同公共服务领域的关注程度，据此设计了权重进行加总。不同的赋权方法是否会影响测评结果的稳健性和信度是评估研究中特别值得关注的问题，也是我们下面研究的要点之一。

从两项评估收集主观信息的方法上看，A 大学最初采用的是面对面问卷调查，但随后就选择采用 CATI 作为主要方法，因为后者在提高样本代表性、降低成本和大范围覆盖方面优势明显。B 机构一直采用的是面对面问卷调查，并配有其他方法（如电话访问），但这种多渠道收集信息的方式可能使不同来源的信息在可比性方面大打折扣，也会影响调查回复率和样本代表性。

两项调查都开展了大范围的问卷调查，每个城市都涵盖了 700 个左右的公民调查样本量，A 大学则另外在每个城市调查了约 100 个企业。从概率论的角度来看，这种大规模的调查能够确保样本的代表性和抽样误差控制在可以接受的范围。但与此相伴随的一个问题就是，这种大范围的调查成本较高，如何确保调查项目的可持续发展是一个重要问题。

当对城市公共服务供给水平和质量的评价结果进行解释时，两项调查也采取了不同的策略。A 大学主要依靠量化研究进行解释，而 B 机构则依靠其团队进行典型城市和公共服务领域的案例研究。一个着重积累实证研究知识，另一个则通过“讲故事”进行解释，两种不同的研究策略在结果呈现和受众接受度等方面都会产生不同的影响。

最后，两项调查都通过出版著作和论文以及召开新闻发布会等形式，获得了广泛的媒体关注，并得到政府部门的重视，为政府绩效信息使用提供了渠道，也促进了第三方公共服务绩效发挥其应有的管理效益和社会效应。两项调查也都选择下半年作为测评结果发布时间，因为跨年调查不利于各城市横向比较，而选择在上半年调查则有利于下半年进行信息发布。但这种时间安排也使采集同一年份的主客观数据面临困难，因此下半年调查而次年上半年发布则有可能克服这一障碍，但却有可能牺牲调查的时效性。

三、数据与方法

本文以两项评价在 2011 年和 2012 年的调查结果为依据，使用的数据分析来自两个项目公开出版和发行的资料。以此为基础我们对两个测评项目的总体评价结果和各领域评价结果的相关性进行分析[24]。由于 2012 年 A 大学调查了公民和企业，B 机构仅调查了公民，我们以公民调查结果为依据进行比较分析。由于两个测评项目涵盖的城市范围略有差异，因此 2011 年可比较的城市有 31 个，2012 年可比较的城市为 34 个，包括 4 个直辖市、25 个省会城市（不包括乌鲁木齐和拉萨）和 5 个计划单列市。

由于两个评价项目采取的量纲和加权方法不同，我们无法直接进行均值差异检验（如 T 检验），本文中我们主要使用相关性分析对两个测评项目的结果进行信度检验。

四、主要研究发现

（一）总体公共服务绩效

尽管两个测评项目存在些许差异，但两者关注的对象都是政府公共服务提供的能力和效果，测评的对象又是同一时期的中国主要城市，两者具有相当的可比性。由于 A 大学测评的公众视角维度与 B 机构的公共服务满意度测量基本是同一个构念（公共服务满意度和政府效能），我们将二者进行统计分析和比较。

首先，我们分析两个测评项目横截面上的信度。表 2 显示两者在 2012 年的相关系数为 0.486，2011 年为 0.564，均在 0.05 的水平上通过统计显著性检验。一般来说，皮尔逊相关系数在 0.8 以上属于相关程度极高，0.6－0.8 属于高度相关，0.4－0.6 属于中度相关，0.2－0.4 属于低度相关，而 0.2 以下属于相关程度极低[25]。这说明两个测评结果尽管存在一些差别，还是有一定的相关性。

为了检验是否是不同的加权方法，影响了两

个测评结果之间的一致性,我们采用机构 B 加权的方法将 A 大学的测评数据重新加权,并和 B 机构的结果进行相关性分析,2012 年两者测评结果之间的相关性为 0.557,略有提高,但是提高的幅度不大。

然后,我们检验了两个测评的跨时间信度。A 大学两年评价之间的相关系数(0.821)远高于 B 机构的(0.425),这从一定程度上表明前者的评价稳健性较高。从公共服务提供的角度上说,城市的表现具有相当大的路径依赖性,很难在一年的时间里有比较大的变化。如果对同一城市的测评在前后两年发生了非常显著的变化,我们对这一测评的结果就需要进行仔细的审视。

表 2　总体公共服务绩效评价结果的相关关系矩阵

	B12	B11	A12	A11
B12	1			
B11	0.4254*	1		
A12	0.4862*	0.5236*	1	
A11	0.3786*	0.5641*	0.8210*	1

注:变量尾号 12 表示 2012 年数据,样本量为 34 个;11 表示 2011 年数据,样本量为 31 个。* 表示在 0.05 的水平上统计显著。

为了进一步考察两项测评结果的相关关系,我们在图 1 绘制了两项评估在 2012 年的二维散点和线性拟合图。

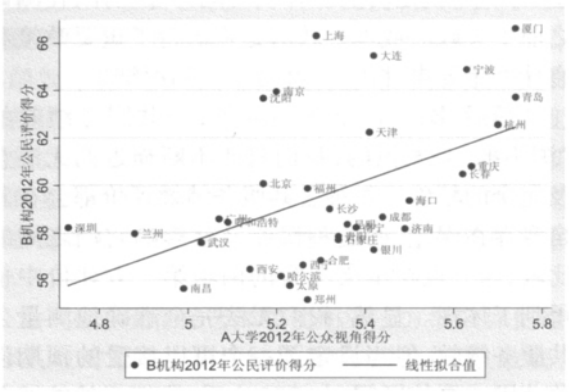

图 1　总体公共服务绩效评价比较

结果显示二者在许多城市的得分和排名方面都存在一些差异。较为典型的差距体现在南京(在两项排名之间差距 20 名,下同)、沈阳(19)、上海(17)、银川(16)、北京(15)。结果说明,即便都是第三方独立公共服务绩效评价,二者对同一组城市的排名也有较大差异,可以使一些城市可以从"先进"一落千丈成为倒数的"后进",也可产生相反的结果而使一些城市从一个排行榜上的"差生"摇身一变成为"优等生"。例如,南京在 B 机构排名第 5 名(仅计算 34 个样本城市),而在 A 大学的排名为第 25 名,即倒数第 10 名;银川在 B 机构排名第 27 名,而在 A 大学的排名升为第 11 名,与大连并列"十佳城市"。与此同时,一些城市在两个排行榜中都保持较为一致的排名,如厦门、宁波、南昌、西安、长沙等,排名变化不超过 2 名,按照 95% 的置信区间是可以接受的波动范围。

(二)各领域公共服务绩效

针对各领域公共服务绩效的评价更有利于比较,因为它们在涵盖的范围方面更加准确和细分,比较的结论也更为可靠。A 大学的公民满意度评价包括公共教育、医疗卫生、住房保障、社会保障、环境保护、公共安全、基础设施、文体设施、公共交通等 9 个领域,B 机构的调查包括公共交通、公共安全、住房保障、基础教育、社会保障和就业、基本医疗和公共卫生、城市环境、文化体育、公职服务等 9 个领域。两项调查较为一致的有 8 项,A 大学对政府效能的测评同 B 机构对公职服务的调查较为一致,因此总共包括 9 项子领域可供比较。

表 3　各领域公共服务绩效的评价

A 大学项目		B 机构项目		相关系数
服务领域	指标数量	服务领域	指标数量	
环境保护	1	城市环境	6	0.7663*
公共安全	1	公共安全	5	0.6916*
公共交通	1	公共交通	5	0.6021*
政府效能	2	公职服务	6	0.4671*
公共教育	1	基础教育	4	0.4326*
社会保障	1	社会保障和就业	4	0.3589*
文体设施	1	文化体育	4	0.3533*
医疗卫生	1	基本医疗和公共卫生	6	0.1527
住房保障	1	住房保障	4	0.1288
基础设施	1	/	/	/

注:表中为 2012 年数据,样本量为 34 个。* 表示在 0.05 的水平上统计显著。

表 3 最后一栏显示,两项评估在 9 个子领域的相关系数均值为 0.439,属于中度相关。其中相关系数最高的是环境保护领域,为 0.766;相关系数最低的是住房保障领域,为 0.129。分析结果显示,我们考察的 9 个公共服务领域的相关系

数均为正，且有 7 个通过了 0.05 水平上的统计显著性检验，仅有住房保障和医疗卫生领域的相关系数低于 0.2 且统计不显著。环境保护、公共安全和公共交通等三个领域都属于高度相关的公共服务领域，相关系数均在 0.6 以上。公职服务和公共教育属于中度相关，而社会保障和文体设施属于低度相关。为什么不同领域的公共服务在两项评估中的相关程度不尽相同呢？我们将在后文对此进行详细讨论。

五、讨论与结论

(一) 原因探析

为什么同样是第三方公共服务绩效评价，两项针对同一组中国大城市的独立评价却得出了较为不同的结果。结合前文对两项评价的初步比较，我们可以将二者的结果差异归结为以下几个原因。

首先，二者的测评指标体系存在差异，直接导致了测评结果的不同。表 3 显示，在 2012 年 B 机构的调查题项较为细致，平均每个公共服务领域包括 5 个题项，共计涉及 45 个问题; A 大学的调查都是一个公共服务领域一个题项，简便可行，但代价是很难做到细致深入。二者的量纲也不同，B 机构采用的是 2、4、5 等三种选项，而 A 大学采用了 10 级和 5 级两种量纲。由于公众受制于知识水平和阅历等方面的影响，他们对不同量纲的精度把握能力不同，并会对评价结果产生一定的影响。更为重要的是，二者对各调查题项的操作化方法不同。A 大学采用的是“满意不满意”和“同意不同意”的尺度，而 B 机构则采用公民对某公共服务的实际感知状况衡量，如距离文体休闲设施的远近、打车等候时间的长短等。

其次，二者的抽样和调查方法不尽相同，使对同一组城市同一个领域的测评结果不同。A 大学的调查采用 CATI 方法，能够根据电话号段对公众进行随机抽样，其样本代表性也较佳。B 机构的调查主要采用面访自填问卷，可能会存在样本代表性偏低的问题。B 机构在报告中指出，在样本特征分布方面，特别是在学历、工作单位性质和收入等方面存在高学历人员、“体制内人员”和高收入者的比例偏高的问题。这种样本特征可能会影响测评结果，因为这些社会强势群体对公共服务的满意度显然不同于社会弱势群体，而后者恰恰是更加需要公共服务关注的社会群体。

由此也不难理解为什么两个项目在住房保障、医疗卫生、文体设施、社会保障、公共教育等领域测评的结果的相关度较低，因为这些公共服务在很大程度上属于“俱乐部服务”，排他性较强，即受到户籍、经济状况和社会地位等因素的制约，因此不同样本的感知和评价水平会差异较大。相对来说，环境保护、公共安全和公共交通都属于普适性的公共服务，任何人都享有同等权利，而不受上述特征的“门槛效应”影响，因此它们在两项调查中的相关度非常高。

再次，二者的加权平均方法不同，也对结果的差异产生了影响。B 机构在各公共服务领域内部采用简单平均法，而在形成公共服务整体得分时采用各公共服务领域的公众关注度作为权重进行加总。A 大学的加权方式略有不同，在对总体公共服务满意度的加权时采用了主成分分析法。我们的分析显示，当将二者的加权方式归一化后，总体公共服务绩效的相关系数得到了一定程度的提升，表明加权方式不同可能是其结果差异的一个原因。另外，还有一个决定两个测评项目差异的重要因素，就是时点。什么时间开展测评，在测评之前中央政府或地方政府是否出台了重要政策和有什么重要事件出现，都会影响测评结果。

最后，除了以上等因素外，公共服务绩效的“测不准”是永恒的，我们只能不断逼近而无法企及完全的精确。每种公共服务绩效评价都会在其自身存在误差，而这种误差又会影响它们之间的比较，这一点在本文考察的两项第三方评价中也得到了体现。显然，我们无法完全准确地测量公共服务绩效，但当设定了一个可以接受的预期误差范围或置信区间(如 95%) 后，我们仍然能够获得一个令决策者满意的测量结果，而这也是有限理性情境下的现实决策的逻辑。

需要说明的是，尽管存在上述差异，两项评估的整体结果仍然是显著正相关的，而且大部分公共服务领域的测评都是正相关的，不存在负相关

的情况。两者在对公认的公共服务提供较好的城市的测评方面呈现出相当的一致性和稳健性。这表明两个项目对样本城市公共服务绩效做出了较为可靠的评价,对城市政府衡量和改进公共服务绩效具有相当的参考价值。

(二)研究贡献与政策建议

国际学界有关组织计分卡的研究较多,对日益增长的外部绩效评价开展了大量研究[17]。有关世界治理排名的分析也越来越多[18]。总体来说,目前有关这方面的研究仍然以美国和国际组织为主,缺少来自发展中国家的证据。本文以两个中国案例为基础,对有关第三方公共服务绩效评价的问题进行了初步研究,提供了来自中国的证据,对于我们进一步考察这些问题大有裨益。本文对第三方公共服务绩效评价项目结果的衡量和比较进行了初步尝试,并发现和分析了第三方绩效评价在公共服务提供测评上可能出现的问题和面临的挑战,为学界进一步理解和探讨第三方绩效评价的相关问题提供了启示。

本文的研究发现,由于进行社会科学研究的种种局限,两个相互独立的第三方公共服务绩效评价,尽管各自采用较为专业的研究方法,也可能产生比较不同的结果。现在越来越多的政府委托第三方机构开展独立的政府绩效评价或公共服务满意度测评。许多高校、学术机构、咨询公司、新闻媒体等也积极响应,组织开展了大量第三方公共绩效评价和排名活动,为政府管理决策和政策执行提供了参考依据和政策建议。从政府的角度而言,如何看待这些评估活动的专业性、科学性以及相互之间可能出现的冲突和不一致,对政府完善自身绩效考评和第三方绩效评估的发展意义重大。

一方面政府在选择和使用第三方绩效评估结果时,应对第三方测评结果在独立性、信度、效度、全面性、可理解性、功能性等方面的表现作出分析和评估,选择最适合或最能反映其绩效表现状况的第三方评价结果。另一方面,政府也应该认识到绩效测评的局限性和测不准原理,不能过分偏倚任何一个第三方绩效评价来源,而应该综合利用多种来源和渠道的评价结果,对公共服务绩效作出一个公允的评判,并用于指导公共服务绩效改进。这种做法也符合"循证公共管理和公共政策"的诉求,即博采最新、最全面的证据,据此作出管理决策和公共政策[26]。

尽管政府绩效信息使用被认为是联结绩效测量与绩效管理的关键纽带,但过度使用、误用和滥用政府绩效信息却可能导致难以预料的负面结果。政府应该正确看待和使用绩效测评信息,过度依赖绩效评估结果来进行管理,比方说许多地方政府将政府绩效评价结果运用到严格的干部奖惩中,可能会导致适得其反的效应。此外,更为重要的是对绩效改进予以奖惩,即不仅关注政府绩效的存量,更应关注政府绩效的增量,因为存量在多数情况下更多会受制于区域和政策领域本身固有的资源禀赋和发展基础,而增量则更多反映了政府部门的能力和努力程度。

从第三方公共服务绩效评价方的角度而言,我们的研究也提出了值得注意的方面。第三方政府绩效评估需要客服研究中的困难,不断提高研究的信度和效度,提升研究的权威性。中国第三方政府绩效评估刚刚兴起,这方面我们面临着更严峻的挑战。如果第三方公共服务绩效评价服务于"排行榜"的排名目的,更需要特别慎重,因为测量和评价由于各种各样的客观和主观原因,存在较大的误差,唯一准确的是"测不准"现象的存在。来自经济合作与发展组织(OECD)、世界银行等国际组织的经验都指出,应该将排名的局限予以报告,特别是说明排名的置信区间[27]。如果第三方公共服务绩效评价的主要目的在于通过评价识别有待改进的领域并提出有针对性的改进措施,那么这种评价的排名功用就需要弱化[28]。

与政府自行组织的公共服务绩效评估相比,第三方评价具有明显的制度和技术优势。但第三方评价之间也存在一定的竞争关系,都在"政府绩效信息市场"上努力争夺政府的有限注意力[29]。当缺乏来自其他竞争对手的强有力挑战时,第三方也可能出现一定的自律问题[14]。因此,我们呼吁更多的学术机构能够参与到第三方公共服务绩效评价的行列中,通过提供更多的绩效信息选择,丰富"政府绩效信息市场",为政府提供更全面多

角度的绩效测评信息，推动第三方公共服务绩效评价事业的健康发展。这一点同大学排名的发展规律是一致的，即越来越多的大学排名涌现，为“消费”这些排名的大学提供了更多的选择，也为优秀的大学排行榜脱颖而出提供制度基础[30]。

（三）研究不足与未来研究方向

本研究是对第三方绩效测评项目进行研究的初步尝试，不可避免存在一系列的局限性。首先，我们的分析主要是横截面比较，还缺少更为具体的纵贯比较。此外，受制于信息的可获得性，我们只是对已公开的信息进行了初步的统计分析。深入的分析有赖于更多原始数据的获得。我们的研究发现，一是需要更深入的统计分析来进行检验，二是需要不同的研究者用相同的数据进行交叉验证。第三方公共服务绩效评价还处于起步和探索阶段，已经公开或披露的案例还较为有限，无法支撑更加深入细致的比较研究。这也说明进一步提升第三方政府绩效评价的信息公开和透明度，可能是未来该领域发展的重要取向。公开原始数据并交由公众和学界检验，可以大大提高第三方绩效评估的可信度、独立性和权威性。

其次，本研究的样本量受制于第三方评价的覆盖范围，很难拓展到其他城市和政府层级。相对来说，区县和乡镇（街道）层级的政府承担了大量基本公共服务的直接提供职能，公众与它们的接触也更多，对它们的研究可能更有助于我们认识第三方公共服务绩效评价的问题。

此外，除了进行统计分析，本文也期望在未来通过深度访谈等方法，获取更丰富的信息以了解第三方公共服务绩效项目的开展情况和面临的挑战。与此同时，也期望未来研究能够对其他类似的第三方公共服务绩效评价进行“元评价”，以推动第三方政府绩效评价健康有序发展。

注：

①孟华《推进以公共服务为主要内容的政府绩效评估——从机构绩效评估向公共服务绩效评估的转变》，《中国行政管理》2009 年第 2 期。

②吴建南、高小平《行风评议：公众参与的政府绩效评价研究进展与未来框架》，《中国行政管理》2006 年第 4 期。

③Gao J. Governing by goals and numbers: A case study in the use of performance measurement to build state capacity in China, *Public Administration and Development*, 2009, 29 (1): 21 – 31.

④Yang Y, Wu J. Are the “Bigger Fish” Caught? China's Experience of Engaging Citizens in Performance Management System, *Public Administration Quarterly*, 2013, 37 (2): 143 – 173.

⑤付景涛、倪星《地方政府绩效评估的政治理性和技术理性——以珠海市万人评议政府为例》，《甘肃行政学院学报》2008 年第 6 期。

⑥徐双敏《政府绩效管理中的“第三方评估”模式及其完善》，《中国行政管理》2011 年第 1 期。

⑦国务院《国务院关于印发国家基本公共服务体系“十二五”规划的通知（国发〔2012〕29 号）》，中央政府门户网站，http://www.gov.cn/zwgk/2012 – 07/20/content_2187242.htm，2012。

⑧郑方辉《2012 中国政府绩效评价红皮书》，新华出版社 2013 年版。

⑨吴伟、于文轩、林挺进、王君《提升城市公共服务质量，打造服务型政府——2010 连氏中国城市公共服务质量调查》，《城市观察》2011 年第 1 期。

⑩于文轩、林挺进、吴伟《提升政府治理水平，打造服务型政府——2011 连氏中国服务型政府指数及中国城市服务型政府调查报告》，《华东经济管理》2012 年第 7 期。

⑪⑫侯惠勤、辛向阳、易定宏《公共服务蓝皮书：中国城市基本公共服务力评价（2010 – 2011）》，社会科学文献出版社 2011 年版。

⑬Gao J. How Does Chinese Local Government Respond to Citizen Satisfaction Surveys? A Case Study of Foshan City, *Australian Journal of Public Administration*, 2012, 71 (2): 136 – 147.

⑭包国宪、张志栋《我国第三方政府绩效评价组织的自律实现问题探析》，《中国行政管理》2008 年第 1 期。

⑮曹惠民《生态学视角下的政府绩效评价研究——以第三方政府绩效评价为例》，《太平洋学报》2010 年第 8 期。

⑯Hanssen CE, Lawrenz F, Dunet DO. Concurrent Meta – Evaluation: A Critique, *American Journal of Evaluation*, 2008, 29 (4): 572 – 582.

⑰Gormley WT, Jr, Weimer DL. Organizational Report Cards, Cambridge, MA: Harvard University Press, 1999.

⑱Hood C, Dixon R, Beeston C. Rating the Rankings: Assessing International Rankings of Public Service Performance, *International Public Management Journal*, 2008, 11 (3): 298 – 328.

⑲Yang K & Holzer M. The Performance – Trust Link: Implications for Performance Measurement, *Public Administration Review*, 2006, 66 (1): 114 – 126.

⑳吴建南、阎波《谁是“最佳”的价值判断者：区县政府绩效评价机制的利益相关主体分析》，《管理评论》2006 年第 4 期。

㉑陈振明、刘祺、蔡辉明等《公共服务绩效评价的指标体系建构与应用分析——基于厦门市的实证研究》，《理论探讨》

2009 年第 5 期。

㉒王佃利、刘保军《公民满意度与公共服务绩效相关性问题的再审视》,《山东大学学报》(哲学社会科学版) 2012 年第 1 期。

㉓曾莉《公共服务绩效主客观评价的一致性论争:来自不同的声音》,《东南学术》2013 年第 1 期。

㉔Piotrowski SJ, Ansah ESI. Organizational Assessment Tools: Report Cards and Scorecards of the Federal Agencies, *Public Administration Quarterly*, 2010, 34 (1): 109 – 142.

㉕Evans JD. Straightforward Statistics for the Behavioral Sciences, Pacific Grove, CA: Brooks/Cole Publishing Company, 1996.

㉖Heinrich CJ. Evidence – Based Policy and Performance Management: Challenges and Prospects in Two Parallel Movements, *The American Review of Public Administration*, 2007, 37 (3): 255 – 277.

㉗Arndt C, Oman C. Uses and Abuses of Governance Indicators. Paris: OECD Development Centre, 2006.

㉘Nardo M, Saisana M, Saltelli A, et al. Handbook on Constructing Composite Indicators: Methodology and User Guide, Paris: OECD Publishing, 2005.

㉙Rogers EW, Wright PM. Measuring organizational performance in strategic human resource management: Problems, prospects and performance information markets, *Human Resource Management Review*, 1998, 8 (3): 311 – 331.

㉚Aguillo I, Bar – Ilan J, Levene M, et al. Comparing university rankings, *Scientometrics*, 2010, 85 (1): 243 – 256.

(责任编辑:宁　岩)

Assessing Third-Party Public Service Performance Assessments: A Comparative Case Study

Ma Liang & Yu Wenxuan

Abstract: Although since 2000 third-party government performance assessment projects have been mushrooming, a dearth of studies has been conducted to study the burgeoning phenomenon, particularly studies examining its emergence, evolution, particularly, its methodology and quality. In this article, we compare two influential third-party government performance projects with similar subjects and assessment schemes. We also conduct statistical analysis on the reliability of the two projects. We found that although the assessment results of the two projects are significantly and positively correlated, there are some important discrepancies. By comparing and analyzing these discrepancies, we provide our research and policy suggestions to the development of third-party government performance evaluation, improving its quality and promoting its healthy development.

Key words: Third-party performance assessment; public service performance; government performance; citizens' satisfaction; reliability