# 教学相“涨”:高校学生成绩和评教分数双重膨胀研究*

哈 巍 赵 颖

**提要:**在经历了“重科研轻教学”的发展阶段后,近年来很多国家都开始关注高等教育人才培养的质量。在回归大学“立德树人”之根本的同时,也要警惕简单机械地使用学生评教这一工具的弊端。本文使用国内某大学2010-2016学年度的课程成绩和评教分数数据,利用该大学在2015年春季学期对课程成绩优秀率放松控制这一自然实验,探究学生成绩与评教分数之间的因果关系。研究发现,优秀率的放松带来了课程成绩和评教分数的双重膨胀,而且对于实验组中那些确实发生了分数膨胀的课程,学生在评教分数上给予了更加慷慨的回馈,课程分数每提高1分,学生评教分数显著提高2分。因此我们有理由相信,学生和老师之间围绕学生评教分数和课程分数产生了合作或者说“共谋”。

**关键词:**学生评教 分数膨胀 场域理论 委托代理模型 工具变量法

## 一、问题的提出

教与学两者之间存在相互依存、相互促进的深刻的共生关系。《礼记·学记》有云:“是故学然后知不足,教然后知困。知不足然后能自反也,知困然后能自强也。故曰教学相长也”。这句话意指在教与学的互动中,教师和学生共识、共享、共进,共同发展和提高。然而,随着现代高等教育的发展,以追求效率、强化管理为目标的“科层制”行政机构逐渐在崇尚自由的大学里落地生根,数量化、可操作化的考核评分制度作为组织管理的一个重要实践策略,追求让每个学校、学生和老师的能力都具有“计算性”和“审计性”(李松林,2007)。教学作为高

* 赵颖系本文通讯作者。感谢教育部人文社会科学研究青年基金项目“博弈论视角下新型师生关系:学生成绩和评教分数双重膨胀研究”(编号18YJC880142)和“学生评教的影响因素研究及其与分数膨胀的博弈关系分析”课题(编号2016YB09)的资金支持。感谢匿名评审人的宝贵意见。文责自负。

校组织管理的关键环节,为了对其进行监督和考评,相应的可量化的评价制度应运而生,即学生评教制度。作为"学"的主体,学生参与了大部分的教学过程,是教学效果的主要接收者,让其来评价教师的教学具有理论上的合理性,学生评教似乎是一个立意甚佳的评价方式。这一制度因其便捷有效且能让学生最广泛地参与其中,已成为最普遍使用的教学评价制度(Murray,2005)。

应当注意到,在过去的几十年中,有两个重大的发展趋势深刻地改变了教学在高等教育系统中的地位和角色。一个是二战后高等教育的大众化和普及化(Cummings & Santner,2015)。高等教育的普及化对教学产生了冲击,因为一方面新入学的学生质量下降,教学的成本上升(Trow,2007),另一方面教学资源并没有随规模的扩大而同步增加,因此造成了大班额等教学质量问题(Hornsby & Osman,2014)。另一个变化是科研在高等教育中的重要性日益凸显。20 世纪 80 年代出现的国家创新体系理论突出了科研在国家经济发展中的重要作用,大学被更多地赋予研究职能、创新职能和科技孵化职能(李兰、哈巍,2017)。20 世纪 90 年代世界一流大学概念的提出及其实践,以及后来世界一流大学排行榜的推出,极大地强化了科研对于大学自身的重要性(Shin & Kehm,2013)。世界三大大学排行榜中科研相关的权重都超过了 50%。在世界知名的研究型大学中,"重科研轻教学"已成为公认的事实。而作为个体的教师,也因为科研更能体现其学术贡献和个人成就,更倾向于把科研放在首要位置(刘振天,2017;Shin,2015)。这些变化都在影响着教与学之间的关系,从而影响着人才培养质量。

有鉴于此,很多国家和国际组织开展了一系列有益的尝试,重新关注人才培养质量(宋伟新,2016;Ryan,2015)。欧洲教学质量框架(European Qualifications Framework,EQF)为欧洲各国提供了一个教学质量保障的基准。联合国教科文组织与世界银行合作共同推出了"全球教育质量保障能力建设行动计划"(Global Initiative for Quality Assurance Capacity, GIQAC)(方乐,2017)。而经济合作与发展组织(OECD)的"高等教育学习成效评估"项目(Assessment of Higher Education Learning Outcomes, AHELO)更是开发了测试工具,从"通用能力"(generic skills)和"学科专业能力"(discipline-specific skills)两个维度来测量高校学生学习成果。但是这些努力在如何定义学习成效上还没有达成共识,在开发能够适用于不同文化背景下的测量工具方面

也还处于尝试阶段,而且这些努力过分强调了学生的重要性,未能从整体上营造出一个有利于提高学习成效的针对利益相关者(包括教师、管理者和领导等)的有机体系(Jacob & Gokbel,2018)。因此,虽然对教学质量的评价手段很多,包括学生评教、领导评价、教师自评、同行评价等,但学生评教(Student evaluations of teaching, SET)因其便捷有效、学生参与度高而被广泛使用(Murray,2005;Miller & Seldin,2014)。已有研究证实,学生评教制度的引入能够测量教师的教学有效性(Murray,2005),有助于提高教师教学投入(Stratton et al., 1994),增加学生对教师和课程的满意度。但也有学者认为,学生评教属于价值判断,所有参与者的评价都受到本身学识、经验、偏好等影响(Hou et al., 2017)。因此,学生评教也可能同时产生一些负面的结果,如一部分学者认为,学生评教最主要的功能是测量学生的满意度,而学生的判断能力存在局限,很难对教师教学进行评价(Uttl et al.,2017)。此外,过度使用学生评教是造成分数膨胀现象(学生的学业成就并没有获得增长而学生成绩提高)的主要推手之一(Kanagaretnam et al.,2003;Ehlers & Schwager,2016)。如下一节所示,教师和学生有可能发生"共谋",即教师给学生打高分,以换取学生在评教中的好评价,从而使教与学的关系异化成"教学相'涨'"的尴尬局面。

"教学相'涨'"的矛盾在我国高等教育发展中也客观存在。我国高等教育自1999年起经历了前所未有的扩张,高等教育毛入学率从1991年的3.5%迅速增长至2016年的42.7%,已经接近高等教育普及化阶段。① 在这个过程中也出现了教育资源短缺、师资力量缺乏、学生质量下降、管理制度落后等问题(谢安邦等,2005;王洪明、佟曾,2007)。与此同时,"重科研轻教学"之风也影响到了我国高校。1995年的全国科技大会上,江泽民同志正式提出"科教兴国战略"。1998年5月4日,时任国家主席江泽民在庆祝北京大学建校100周年大会上代表党和政府向全世界宣告建设世界一流大学,科研的重要性进一步上升,而教学工作的相对地位开始下降(李斐,2015;于佳鑫,2017)。如今,这一现象越来越引起社会各界的重视,高校管理者和学者们开始呼吁大学应该回归人才培养的职能,提高高校教学质量(黄达人等,

① 数据来源:中华人民共和国教育部网站(http://www.moe.gov.cn/jyb_sjzl/s5990/201711/t20171110_318862.html)。

2017)。

我国高校在20世纪80年代中期就引入了学生评教制度(刘妙龄,2005)。1990年原国家教委正式颁布了《普通高等学校教育评估暂行规定》(中华人民共和国国家教育委员会令第14号),学生评教活动步入正规化发展。如今,在我国高校的管理之中学生评教已经成为一项制度化的活动。然而,作为一项保障教学质量的重要制度,学生评教在运行过程中却一直备受质疑。教师认为学生评教结果成了行政考核的工具,偏离了保障教学质量的根本目的(周继良等,2017),且学生受到学识、视野的限制,不具备全面评价教师教学质量的能力(朱富强,2016)。学生担心给老师差评会遭到报复,认为评教制度仅仅是个形式,并不值得认真评价(李亚娟,2016)。

时至今日,建设世界一流大学在经过了早期重点支持吸引人才、加强基础设施建设和学科建设后进入人事管理制度改革的深水区。高校在人事改革中普遍实施“因科研而聘、因科研而升”的管理办法(张端鸿,2016),但是在设置教师晋升标准时也开始加入教学量和教学成果的硬性要求,如规定学生评教低于一定排名的教师当年职称晋升暂停申请,甚至失去授课资格等(赖勇强等,2017;曹辉、钮菁菁,2017)。这使得高校教师对学生评教分数日益看重,甚至出现了“讨好”学生的倾向,具体表现为顺应学生喜好进行教学、降低课程难度、放宽对学生的要求以及给学生更高的成绩等等。这一制度的引入加上人事制度的改革,让教与学二者的关系发生了始料未及的变化。如果教师放松给分就能换来更高的评教分数,那么不仅说明学生评教制度在有效性方面存在缺陷,而且更值得担忧的是,在这项制度下,教师和学生的关系可能会发生异化,形成所谓的“合作”和“共谋”。

根据布迪厄的场域理论,教育场域的自主性在于运行其中的“通货”或媒介是文化资本,因此主导该场域的权力形式应该主要为学术权力(或文化权力)。然而吊诡的是,本应发挥核心作用的文化资本却常常受制于社会资本(权力)或经济资本(金钱),造成教育场域的异化。本文所关注的学生评教制度亦可看作逐渐成为重要行动主体的学校(行政管理者)在教学及整个教育场域之中所展开的一系列微观权力实践的体现。本文将以学生评教制度为着力点,一方面从理论上建构教育场域下利益主体间的委托代理模型,另一方面使用某研究型大学2010–2016学年度的课程分数及教学评估数据,利用该大学一个外

生的政策变化来分析学生成绩与评教分数之间的因果关系,以期为该领域的实证研究做出有益的补充。

## 二、教育场域下的教与学关系:“委托—代理—客户”[①]模型

自20世纪70年代,学者们开始关注学生评教制度的引入对师生关系的影响,并开始使用博弈论作为分析工具(Correa, 1987)。研究者大多聚焦在师生之间的博弈上,如卡纳哥特那姆(Kanagaetnam et al.,2003)通过构建教师和学生的效用函数,[②]发现适度使用学生评教可同时提高老师教学投入和学生学习投入与收获,但过度使用学生评教会造成教师过分重视评教得分,从而放松课程要求和给分标准,导致分数膨胀。朗本(Langbein,2008)注意到教与学的关系实际关乎学校、教师、学生三个利益相关者,学校是委托人,教师是代理人,学生是“客户”,由于委托人学校对代理人教师的教学行为和效果无法直接观测,委托人(学校)便引进了一个评估机制即学生评教制度,让客户(学生)来评价代理人(教师)的表现。

本文尝试将布迪厄的社会学理论和朗本的三者委托代理模型相结合,来考察教育场域中高校、教师和学生三个利益主体围绕教与学展开的互动与博弈(如图1所示)。文化资本作为教育场域的“通货”,既是利益主体进行教育实践活动的基础,又是他们彼此之间展开微观权力实践的载体(吴越、李惠娜,2016;何晓芳,2012)。因此,教师和学生的教学互动、教师的科研活动以及学校的管理服务活动,都是为了获取不同层次或程度的文化资本而采取的策略和实践。

### (一)委托人:学校

对于高校而言,其关注的文化资本体现为大学声望或者说大学的排名,因为这与优质生源和师资、社会捐赠以及政府的资源配置息息相

① “委托—代理”理论虽然一般被认为起源于经济学,但实际上也可以追溯到社会学的交易理论,一些社会学家指出这种理论有潜力也应该更广泛地应用于社会学研究之中(Eisenhardt,1989;Shapiro,2005)。

② 所谓“效用函数”,是表示消费者在消费中所获得的效用与所消费的商品组合之间的数量关系的函数。

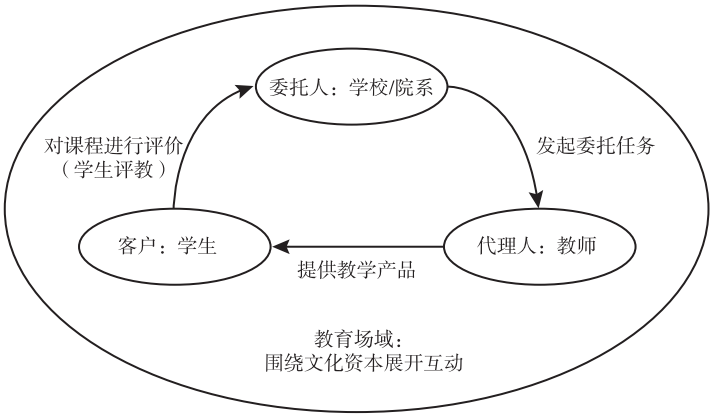


**图1　教育场域中的委托代理关系**

关(谢小燕等,2014)。教学和科研是高校的重要职能,也是其声望形成的手段。相较于教学成果,科研成果(R)更容易评价,能更快更好地提升学校的文化资本(袁祖望,2002;Langbein,2008),因此高校、教师和学生三者的利益是一致的,不存在委托代理问题。而教师的教学努力与教学效果无法直接而精确地被监督和评判,或者说这样做的成本太高,因此出现了委托代理问题。教学质量虽然不会在短期内影响学校的经费收入,但从长期来看,人才培养的质量最终将影响大学的声望,因此近年来大学对教学给予了适当关注,希望教师能够在教学数量和质量两个方面保证一定的投入,并通过学生评教(T)来体现。对于学校来说,学生成绩(g)的提高能够提高学生升学的质量或数量,以及提升就业率。委托人的效用函数为:

$$P = f(g, T, R)$$

**(二)代理人:教师**

教师在整个教育场域中处于中间人的位置。教师既是文化资本的保有者,承担着知识传承的任务(教学);又是文化资本的获取者,承担着知识创新的任务(科研)。教师关注的是自己的职称晋升和工资。假定教师的总努力水平(C)是恒定的,则教师花在教学上的努力和花在科研上的努力存在竞争关系。对教师来说,需要衡量教学产出(用学生评教分数T来衡量)和科研产出(R)的重要性,做出一个努力水平

的分配。教师的教学水平、教学投入和其他外生变量一起决定了教师的学生评教得分(T)。教师的教学和科研成果共同决定了其职称晋升的机会。教师的效用函数为:

$$F = f(T,R,C)$$

**(三)客户:学生**

学生选择接受教育,一方面是为了获得知识、提高能力,另一方面是为了给就业市场(或研究生院)一个高能力的信号,从而把自己和低能力的学生区分开,获得更高的薪酬(或升学机会),这个信号就是成绩。成绩在相当程度上代表了学生在学校里获得的文化资本,进而形成学生自身人力资本的信号。同时,学生作为被服务对象,一方面给出了对教学过程和自己的学习体验满意度的短期评价(体现为评教分数T),另一方面也会对在学校的整个学习体验有长期的评价(校友忠诚度)。在学习过程中,学生付出了时间和金钱等成本(E)。因此,学生的效用函数为:

$$S = f(g,T,E)$$

**(四)学生评教导致分数膨胀的内在机制**

由上文的分析可知,在教育场域中,三者存在利益一致性,都看重学生评教分数,这是他们各司其职、相互合作的基础。学生评教制度犹如知识传递效果的标尺,在实际使用中替代知识,成为实际意义上的"通货",将三个利益主体联系在一起。但如果老师给学生的课程成绩会影响到学生给老师的评教分数,那评教分数 T 则变为 T(g),即学生成绩的函数,从而使学生成绩和评教分数同时成为三个利益相关者都看重的变量。在这种互动关系中,教师有动机采取降低课程要求、给学生打高分的策略来换取更高的学生评教;学生也乐见这种能以更少的努力获取更高分数的形式,对那些给分"厚道"的教师给予更高的教学评价作为回馈;高校作为委托人看到的是课程分数和学生评教分数同时上涨,进而在职称晋升上奖励那些评教分数高的教师。这样一来,通过分数的膨胀,学生获得了更好的成绩,教师获得了更高的工资或晋升机会,学校获得了更高的升学率或就业率,但教育质量却发生了下滑。

已有实证研究普遍发现学生的课程成绩和他们给教师的评教分数

正相关(Gorry,2017),但是可能有其他原因可以解释这种正相关,比如说在教学上投入大量精力的老师改善了学生的学习效果,进而获得了学生的好评。因此需要一个学生课程成绩的外生变化来考察学生在教学评价上会如何反应。现有文献中使用比较严格的因果推断方法的有两篇。布彻(Butcher et al.,2014)使用双重差分方法来评价美国卫斯理学院的课程成绩上限政策对分数膨胀和学生评教分数的影响,发现在受上限政策影响的院系中老师给 A 的比例下降了 18 个百分点,同时这些院系选修课程的学生数量也下降了 18%,学生强烈推荐其他学生选修课程的比例下降了 5%。戈里(Gorry,2017)模仿上文的实证设计,利用美国一所州立大学商学院的数据来分析课程分数上限政策对于分数膨胀和学生评教的影响,并且采用工具变量法来进一步考察学生评教和课程分数的互动关系,结果表明老师所给的课程分数的下降引起了学生评教分数的下降。国内鲜有研究考察课程分数和学生评教分数之间的关系,使用严谨的因果推断方法的研究则更为少见。

## 三、背景和数据

本文使用国内公立研究型大学 A 大学 2010 - 2016 学年度共 12 个学期[①]的本科生课程平均成绩和学生评教数据,其所属学科涵盖理科、工科、人文、社科、医学等 5 个大类。因体育课和实验课与理论课的教学过程有别,这里去掉全部体育课程和实验课程,只保留理论课程。

A 大学自 2007 年以来长期实行课程分数的优秀率控制(大于等于 85 分为优秀,优秀率即为一门课程中大于等于 85 分的学生人数比例,取值区间为 0% - 100%),优秀率超过 40% 的课程需要任课教师递交正式申请详细说明突破原因,并由院系和教务部审批通过后才可录入课程成绩,且对全部课程适用。虽然在这种控制下有些课程仍然会突破 40% 的优秀率,但整体的优秀率得到了一定的抑制。从 2014 - 2015 学年第二学期(即第 10 学期)开始,A 大学在实际操作中放松了对优

① A 大学每个学年度共有秋季、春季、夏季三个学期,因夏季学期开设课程明显少于春、秋二季且具有一定的特殊性,夏季学期课程不参与学生评教,因此本文的分析中只保留春、秋二季学期的课程。

秀率的控制，任课教师可以自行设定优秀率，这可能导致更多的课程突破40%的优秀率。由图2a可以看出，在优秀率控制放松之前的第1－9学期，全部课程优秀率整体呈缓慢上升的态势，在第10学期放开优秀率控制后，全部课程优秀率陡然上升到47.9%，比上一学期上升了3.2个百分点。伴随着优秀率的变化，课程分数也在相应变化。在第1－9学期，全校课程平均分的变化较为平稳，总共增长了0.5分。而第10学期放开优秀率控制后，仅仅一个学期全校课程平均分就增长了0.5分，并保持着继续增长的趋势。

A大学采用学生网上评教的方式，在考试周开始前的两周内进行，在考试周开始时结束，由于此时教师尚未对学生进行最后的考核，因此

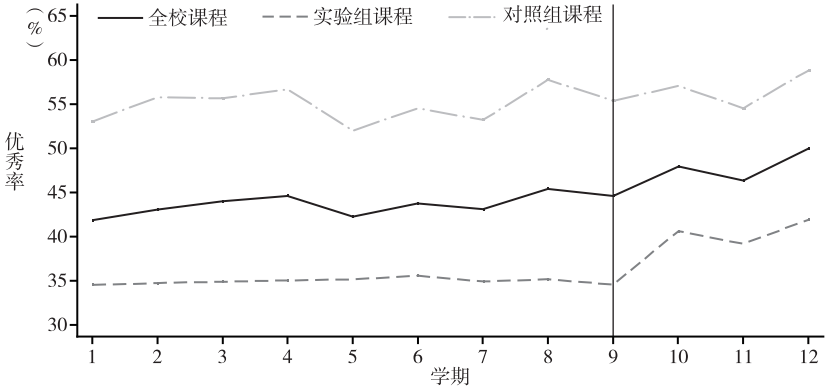


图2a 课程优秀率变化

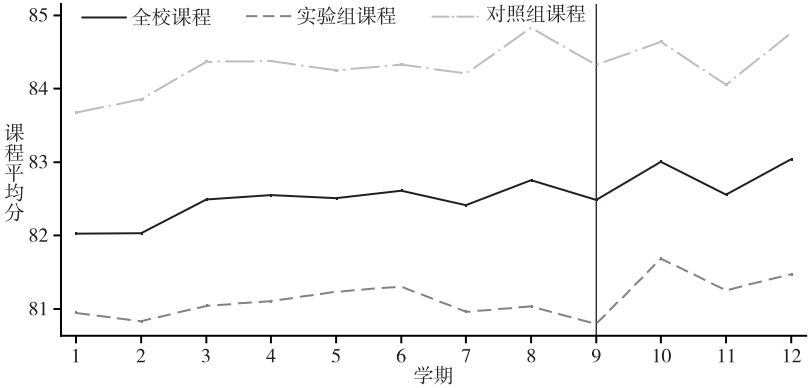


图2b 课程分数变化

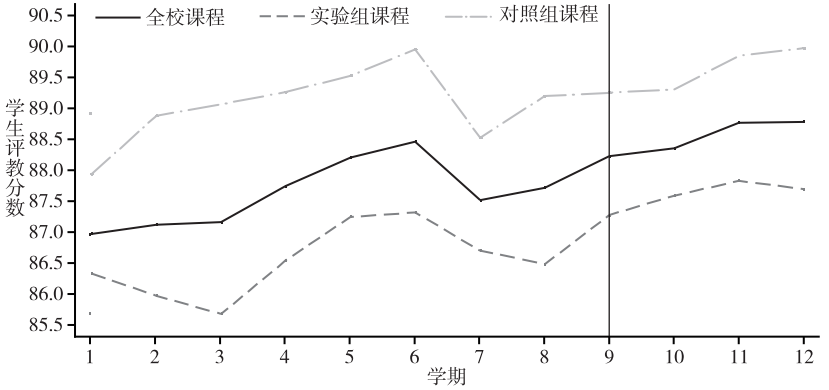


图2c 学生评教分数变化

**图2　优秀率和学生评教分数的变化趋势**

在评教时学生不知道自己最终的课程成绩。学生自愿对课程和教师进行评价,并且该评价完全匿名,教师不会知道具体某位学生的打分。在本文的数据中,虽然在评教时学生并不知道自己最后的课程成绩,但实际上在教学过程中,老师向学生释放了给分高低的信号,学生形成了对课程给分的预期。首先,有些课程由平时分、出勤分、期中测验、期末考试构成,学生在平时的学习过程中已经取得了一部分分数;有些课程没有平时分,但教师会在上课的过程中明确考核要求,这也使学生对自己的成绩有一个基本预期;有些学生会打听某门课程以往的给分情况,形成了学生对最终成绩的预期;教师在教学过程中可能会对某些学生产生某种主观态度并使学生感知到,学生可借此对自己的成绩形成预期。因此,学生评教的分数可以通过预期分数与教师的给分形成关联。很多实证研究都发现,无论是预期分数还是实际分数都会对学生评教产生影响(Krautmann & Sander,1999;McPherson et al.,2009),朗本(Langbein,2008)通过实证研究发现预期分数比实际分数对评教分数的影响更大。由图2c可以看出,在第1-9学期,参评课程的学生评教分数呈上升趋势,在政策改变后,上升趋势更为明显,全校平均学生评教分数由第1学期的87.0上升到第12学期的88.8。

当然这种课程分数和评教分数的总体上涨未必能说明放松优秀率限制必然导致课程分数和评教分数的膨胀。可能的解释是在这期间学

生的质量或者学生在学习的投入上都发生了显著的提高;另外老师也可能在此期间在教学上更努力,因此学生的学习效果更好,学生对于课堂体验也更加满意。为了对政策效果进行因果推断,本文将所用数据分为实验组和对照组,使用教师和课程两个变量联合定位个体(即同一位老师讲授的同一门课程),数据被处理为非平衡面板数据。计算在政策改变前每个个体的平均优秀率,若该平均优秀率低于或等于40%,则该个体纳入实验组,否则放进对照组。若某个体只在政策变化前或政策变化后出现,而不是在政策变化前后同时出现,则该个体实际不参与计算,但因这些个体影响描述统计,所以此处去掉不同时在政策变化前后出现的个体。从图2中可以看出,这两类课程在优秀率、平均分和学生评教分数上的差异都在缩小,那些受优秀率控制放松影响的课程(实验组)在迅速追赶对照组。全部课程描述统计见表1。在政策前后,实验组的优秀率和平均分皆有所上升,而对照组的优秀率和平均分整体保持稳定;实验组的上课人数一直高于对照组,两组在春、秋二季学期的分布大致相当。从开课教师的职称来看,对照组教授开课的比例略高于实验组。① 实验组全校必修课、专业必修课和通选课的比例较高,专业选修课的比例低于对照组。

**表1　　实验组与对照组课程描述统计**

| | 实验组 | | | | 对照组 | | | |
|---|---|---|---|---|---|---|---|---|
| | 政策前(n=4942) | | 政策后(n=1489) | | 政策前(n=4205) | | 政策后(n=1285) | |
| | 均值 | 标准差 | 均值 | 标准差 | 均值 | 标准差 | 均值 | 标准差 |
| 优秀率(%) | 35 | 8 | 41 | 14 | 57 | 21 | 57 | 21 |
| 课程平均分 | 80.91 | 3.43 | 81.49 | 4.20 | 84.46 | 3.92 | 84.49 | 4.74 |
| 学生评教分数 | 86.66 | 6.01 | 87.73 | 5.97 | 89.21 | 6.22 | 89.73 | 6.64 |
| 评估率(%) | 66 | 14 | 69 | 15 | 67 | 15 | 68 | 17 |
| 上课人数 | 78 | 68 | 68 | 65 | 46 | 58 | 41 | 54 |
| 秋季学期(%) | 57 | 50 | 34 | 47 | 56 | 50 | 34 | 47 |

① 政策后两组的教授比例都有所下降,这是因为自2014年起,该校对教师授课量的要求变得更加严格,若授课量不够将不能职称晋升。因此,政策后非教授的开课量增加了。

续表1

| | 实验组 | | | | 对照组 | | | |
|---|---|---|---|---|---|---|---|---|
| | 政策前(n=4942) | | 政策后(n=1489) | | 政策前(n=4205) | | 政策后(n=1285) | |
| | 均值 | 标准差 | 均值 | 标准差 | 均值 | 标准差 | 均值 | 标准差 |
| 教师职称(%) | | | | | | | | |
| 教授 | 50 | 50 | 47 | 50 | 51 | 50 | 48 | 50 |
| 课程类别(%) | | | | | | | | |
| 全校必修 | 10 | 30 | 10 | 29 | 2 | 15 | 4 | 19 |
| 专业必修 | 44 | 50 | 44 | 50 | 44 | 50 | 40 | 49 |
| 专业选修 | 31 | 46 | 32 | 47 | 42 | 49 | 44 | 50 |
| 通选课 | 15 | 36 | 14 | 35 | 11 | 32 | 12 | 32 |

注:(1)优秀率是百分数,课程平均分和学生评教分数均为百分制。(2)n为课程门数。

## 四、方法和模型

本文借鉴了戈里(Gorry,2017)分析学生成绩和评教分数时所用的方法,利用优秀率控制政策的变化,将全部课程分为实验组和对照组。对于在政策改变前优秀率已经突破40%的课程,优秀率控制的放松对这类课程并没有太多影响,因此,受政策变化影响较大的是在政策改变前严格遵守优秀率限制的课程。通过建立实验组和对照组,本文利用双重差分法来排除其他因素,评估该政策的变化对课程分数和学生评教分数的影响。此外,更为重要的是探讨学生评教分数是否受课程分数的影响。这涉及对学生评教有效性的评估,即考察在现有学生评教制度下,教师和学生是否会发生共谋。为了回答这个问题,本文使用工具变量法来评估课程分数对于学生评教分数的影响。

### (一)双重差分法(Difference-in-differences Model)

双重差分法假设在没有外生变量干扰的情况下,实验组和对照组课程的学生成绩和评教分数有着相同的发展趋势,政策变化后,实验组的趋势受政策的影响发生了变化,而对照组仍然延续之前的变化趋势。通过对比实验组和对照组,可以构造出实验组的反事实(counterfactuals),得到真实的处理效应。由图2可知,本文的实验组和

对照组在政策改变前课程分数和学生评教分数的变化趋势都大致相同,满足双重差分法的基本假设。①

建立双重差分固定效应模型如下:

$$Y_{ij} = \beta_0 + \beta_2 Treated_i + \beta_2 Postpolicy_j + \beta_3 Treated_i \times Postpolicy_j + \beta_4 X_{ij} + \gamma_i + \theta_j + e_{ij} \quad (1)$$

其中,$Y_{ij}$代表某个体 $i$(教师和课程联合变量)$j$ 学期的课程平均分或学生评教分数。$Treated_i$ 是虚拟变量,该个体属于实验组时取值为 1,该个体属于对照组时取值为 0。$Postpolicy_j$ 是虚拟变量,当它等于 1 时表示政策改变后,等于 0 时表示政策改变前。$Treated_i \times Postpolicy_j$ 是交互项,其系数即为本文关注的处理效应。$X_{ij}$是一系列控制变量,包括上课人数、上课人数的平方。$\gamma_i$ 代表教师和课程联合变量的固定效应,它控制的是不随时间变化的课程难易程度或者老师给分的严格程度(当结果变量为学生的课程分数时),或者教师是否是个以学生为中心的老师等(当结果变量为学生评教分数时)。$\theta_j$ 是学期固定效应,控制的是不随个体(老师和课程)而变化的总体时间趋势,如该学校不同年级学生群体在学业能力上的差异,该校整个教师或者学生群体对于教学或者学习的重视程度的总体变化趋势等。当这两个固定效应同时使用时,*Treated* 和 *Postpolicy* 这两个变量将被吸收掉,因此没有必要再展示其系数。$e_{ij}$是随机误差项。

**(二)工具变量**(Instrumental Variables)

如前所述,学生的课程成绩和他们给教师的评教分数正相关并不能说明学生和老师之间围绕教与学出现了共谋,另一种可能是在教学上投入大量精力的老师改善了学生的学习效果,进而获得了学生的好评。因此本文使用优秀率控制政策变化这个外生政策冲击作为课程分数的工具变量。该政策放松了教师给分的限制,直接影响学生分数,但对被解释变量评教分数的影响应该只是以学生的课程分数或者说在课堂与老师互动后形成的期望分数为中介,因此可作为合适的工具变量。研究对象为实验组课程。

工具变量法第一步如下:

① 本文在做 DID 之前已经做过"平行趋势检验",受篇幅限制,没有体现在正文中。

$$AverageGrade_{ij} = \alpha_0 + \alpha_1 Postpolicy_j + \alpha_2 X_{ij} + \gamma_i + e_{ij} \quad (2)$$

其中$AverageGrade_{ij}$为某个体 $i$(教师和课程联合变量)$j$ 学期的课程分数,$Postpolicy_j$为工具变量。$X_{ij}$代表一系列控制变量,包括上课人数、上课人数的平方和是否秋季学期。这里学期固定效应不再加入,因为这会吸收掉优秀率放松这一政策的影响。$\gamma_i$仍为个体固定效应。

工具变量法第二步如下:

$$Y_{ij} = \beta_0 + \beta_1 \widehat{AverageGrade}_{ij} + \beta_2 X_{ij} + \gamma_i + e_{ij} \quad (3)$$

其中$Y_{ij}$代表某个体 $i$(教师和课程联合变量)$j$ 学期的学生评教分数。通过该外生政策变化形成的分数变化,考察学生评教分数是否会对分数变化产生呼应。

## 五、结果和分析

### (一)政策变化对课程平均分的影响

表 2 是双重差分法的计算结果[公式(1)],反映了优秀率政策的变化对课程分数的影响。总体而言,政策的放松使得实验组课程的分数相对于对照组课程平均提高了 1.0 分(这是一个非常大幅度的提升,相当于从第 1 到第 9 学期全校课程平均分增长幅度的 2 倍,相当于对照组改革前课程分数的标准差的 1/4);上课人数每增加 1 人,课程平均分约降低 0.02 分。① 各类职称教师的给分都受到了政策的影响,教授比政策前打分提高了 1.2 分,非教授比政策前打分提高了 0.8 分。② 分课程类别来看,全校必修课和通选课都没有明显受到政策变化的影响,而专业课程受到的影响比较显著。相对于专业必修课,专业选修课面临吸引学生前来选课的压力,在政策放松后分数的提高幅度最大,高达 1.6 分。

① 此处虽有上课人数的二次项,但因系数很小,仍可认为上课人数每增加 1 人,评教分数约降低 0.02 分。

② 此处将教授和非教授两个样本在政策前后给分提高的幅度进行了比较,并无显著差异。篇幅所限,比较过程未写入正文。

**表 2　　优秀率变化对课程分数的影响**

| 变量 | 全校课程 | | | |
|---|---|---|---|---|
| | 总体 | 教授 | 非教授 | |
| 处理效应 | 1.026*** (0.160) | 1.207*** (0.246) | 0.847*** (0.214) | |
| 上课人数 | -0.0180*** (0.00276) | -0.0129*** (0.00410) | -0.0215*** (0.00385) | |
| 上课人数的平方 | 4.13e-05*** (6.93e-06) | 2.91e-05*** (9.50e-06) | 5.03e-05*** (1.03e-05) | |
| 常数项 | 83.39*** (0.181) | 82.71*** (0.311) | 83.89*** (0.222) | |
| 个体固定效应 | 是 | 是 | 是 | |
| 学期固定效应 | 是 | 是 | 是 | |
| 观测值 | 11800 | 5830 | 5850 | |
| $R^2$ | 0.708 | 0.699 | 0.722 | |
| 变量 | 课程类别 | | | |
| | 专业必修 | 专业选修 | 全校必修 | 通选课 |
| 处理效应 | 0.673*** (0.230) | 1.611*** (0.311) | 0.617 (0.504) | 0.396 (0.413) |
| 上课人数 | -0.00974 (0.00638) | -0.0448*** (0.00831) | -0.0119 (0.00941) | -0.0126** (0.00498) |
| 上课人数的平方 | 2.95e-05 (2.21e-05) | .000135** (5.57e-05) | 2.85e-05 (3.03e-05) | 2.38e-05*** (8.91e-06) |
| 常数项 | 82.35*** (0.339) | 84.37*** (0.368) | 82.19*** (0.591) | 84.78*** (0.416) |
| 个体固定效应 | 是 | 是 | 是 | 是 |
| 学期固定效应 | 是 | 是 | 是 | 是 |
| 观测值 | 5153 | 4317 | 782 | 1548 |
| $R^2$ | 0.756 | 0.672 | 0.711 | 0.652 |

注：(1) * $p<0.1$, ** $p<0.05$, *** $p<0.01$。(2)括号中为稳健标准误。

## (二)政策变化对学生评教分数的影响

表 3 是双重差分法的计算结果[公式(1)]，反映了优秀率政策的变化对评教分数的影响。总体而言，政策变化后，实验组的评教分数相对于对照组平均提高了 0.5 分，相当于改革前对照组的 1/12 个标准

差;上课人数每增加 1 人,评教分数约降低 0.03 分。① 教授获得的评教分数并没有显著的变化,而非教授的评教分数则显著上升了 0.84 分。结合上一节教授和非教授都显著放松了分数,这可能反映的是教授并没有很好地把要放松分数这个信息传递给学生,进而影响了学生期望成绩的形成。分课程类别来看,专业必修课、全校必修课和通选课都没有明显受到政策变化的影响,只有专业选修课的评教分数提高了 0.9 分。与上一节表 2 的结果对照来看,专业选修课的分数提高是最显著的,因此学生评教的正向反馈也最为强烈。一个可能的解释是:相对于专业必修课,在专业选修课上教师面临着吸引学生选课的压力,因而可能更愿意将降低课程要求和给分尺度的信号发送出来,从而影响学生的成绩预期。

**表 3　　优秀率变化对评教分数的影响**

| 变量 | 全校课程 | | | |
|---|---|---|---|---|
| | 总体 | 教授 | 非教授 | |
| 处理效应 | 0.518 ** (0.248) | 0.375 (0.352) | 0.839 ** (0.363) | |
| 上课人数 | −0.0288 *** (0.00389) | −0.0282 *** (0.00567) | −0.0284 *** (0.00557) | |
| 上课人数的平方 | 5.85e−05 *** (9.88e−06) | 5.48e−05 *** (1.34e−05) | 5.97e−05 *** (1.51e−05) | |
| 常数项 | 0.518 ** (0.248) | 0.375 (0.352) | 0.839 ** (0.363) | |
| 个体固定效应 | 是 | 是 | 是 | |
| 学期固定效应 | 是 | 是 | 是 | |
| 观测值 | 11837 | 5857 | 5857 | |
| $R^2$ | 0.667 | 0.667 | 0.680 | |
| 变量 | 课程类别 | | | |
| | 专业必修 | 专业选修 | 全校必修 | 通选课 |
| 处理效应 | 0.446 (0.382) | 0.892 * (0.465) | 0.0217 (0.623) | −0.373 (0.544) |

① 此处虽有上课人数的二次项,但因系数很小,仍可认为上课人数每增加 1 人,评教分数约降低 0.03 分。

续表3

| 变量 | 课程类别 | | | |
|---|---|---|---|---|
| | 专业必修 | 专业选修 | 全校必修 | 通选课 |
| 上课人数 | -0.0165 **<br>(0.00828) | -0.0658 ***<br>(0.0123) | -0.0135<br>(0.0158) | -0.0200 ***<br>(0.00728) |
| 上课人数的平方 | 1.20e-05<br>(2.73e-05) | 0.000182 ***<br>(6.88e-05) | 5.68e-05<br>(5.11e-05) | 3.38e-05 ***<br>(1.30e-05) |
| 常数项 | 88.18 ***<br>(0.496) | 88.94 ***<br>(0.509) | 85.21 ***<br>(1.089) | 88.41 ***<br>(0.599) |
| 个体固定效应 | 是 | 是 | 是 | 是 |
| 学期固定效应 | 是 | 是 | 是 | 是 |
| 观测值 | 5161 | 4339 | 783 | 1554 |
| $R^2$ | 0.720 | 0.610 | 0.771 | 0.642 |

注:(1) * p<0.1, ** p<0.05, *** p<0.01。(2)括号中为稳健标准误。

**(三)分数对学生评教的影响**

本文将优秀率控制政策放松作为工具变量,用以评估学生成绩对于评教分数的影响。表4呈现了工具变量法的两阶段最小二乘法计算结果[公式(2)、(3)]。课程平均成绩提高1分,学生评教分数显著提高2分,相当于对照组改革前的1/3个标准差。这意味着相对于教师给分的提高,学生的回馈更加慷慨。分课程类别来看,课程平均成绩对于专业选修课程的评教分数影响比专业必修课来得大,专业选修课的给分提高1分,学生评教分数能够提高2.6分,但是这种差异并不显著。分教师类型来看,学生对于非教授的成绩放松更加敏感,几乎是教授的2倍,而且在统计上是显著的。

Cragg-Donald Wald F和Kleibergen-Paap rk Wald F两个第一阶段统计量用于检验本文所采用的工具变量是否为弱工具变量。根据经验,本文认为两个统计量的数值大于10,可视为通过了弱工具变量检验。对于表4中的前四列,该工具变量都通过了弱工具变量的检验。而对于专业选修而言,虽然只在边缘上通过了弱工具变量的检验,但使用Anderson-Rubin Wald统计量时所有样本都通过了弱工具变量检验。

**表 4　　学生成绩对评教分数的影响**

| 变量 | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | 总体 | 教授 | 非教授 | 专业必修 | 专业选修 |
| 平均分 | 2.008*** | 1.330** | 2.588*** | 2.082*** | 2.595*** |
| | (0.404) | (0.522) | (0.638) | (0.773) | (0.983) |
| 上课人数 | -0.0108 | -0.0155* | 0.00219 | -0.0191 | 0.00663 |
| | (0.00811) | (0.00819) | (0.0182) | (0.0189) | (0.0316) |
| 上课人数的平方 | 2.27e-05 | 2.51e-05 | 9.01e-07 | 1.18e-06 | 4.73e-05 |
| | (1.96e-05) | (1.79e-05) | (5.11e-05) | (5.83e-05) | (7.92e-05) |
| 是否秋季学期 | -0.216 | -0.0383 | -0.294 | -1.280 | -0.163 |
| | (0.419) | (0.477) | (0.726) | (1.812) | (1.180) |
| 一阶段 Cragg-Donald Wald F 统计量 | 37.23 | 17.06 | 19.72 | 10.58 | 9.09 |
| 一阶段 Kleibergen-Paap rk Wald F 统计量 | 37.84 | 16.30 | 20.85 | 11.36 | 8.48 |
| Anderson-Rubin p 值 | 0 | 0.003 | 0 | 0 | 0 |
| 观测值 | 5788 | 2814 | 2880 | 2515 | 1768 |

注:(1) * $p<0.1$, ** $p<0.05$, *** $p<0.01$。(2)括号中为稳健标准误。

## 六、总结和讨论

在过去的几十年中,高等教育场域内部的分层不断加剧,学校卷入大学排名的漩涡,教师需要应对更加严格的职称晋升制度,学生要面临愈加严峻的就业形势。高等教育在数量和质量上扩张的同时,也面临着新公共管理主义影响下的日益严苛的绩效考评的压力。在高等教育场域内部,高校、教师和学生三者在这种日益激烈的竞争和量化考评标准下都采取了不同的策略和实践来获取自身的文化资本。高校关注的文化资本体现为大学声望或者说大学的排名,教师关注的是自己的职称晋升,学生要获得的是更有竞争力的文凭和更高的成绩。科研和教学是高校展开的主要活动,也是上述这些文化资本赖以形成的基础。在科研方面,由于科研成果比较容易客观测量和评价,因此高校、教师和学生的利益是一致的。但是在教学方面,由于高校对教师的教学行

为和效果无法直接观测,高校和教师之间的利益并不一致,出现了委托代理问题。因此高校作为委托人引进了一个评估机制即学生评教制度,让学生(客户)来评价教师(代理人)在教学上的表现。如果学生评教能够如实反映教师的教学投入和效果,那么这个委托代理问题就迎刃而解了。但是如果学生评教与老师给学生分数之间存在着正向回馈的机制,那么教师则有动机采取降低课程要求、给学生打高分的策略来换取更高的学生评教;学生也乐见这种以更低的努力获取更高分数的形式,对那些给分"厚道"的教师报之以更高的教学评价。对于学生课程分数和学生评教分数的双重膨胀,学校也睁一只眼闭一只眼,没有动力去打破这种默契,因为学校在这个过程中也得益于学生满意度的提高和学生更好的就业和升学前景。美国比较前沿的理论研究证明了当高校过分强调学生评教制度在教师评价中的地位时,老师更容易对学生放水(Kanagaretnam et al., 2003)。最新的实证研究也证明了这种师生共谋的可能性(Butcher et al., 2014; Gorry, 2017)。

目前,国内对于教学相"涨"的实证研究比较少见。本文使用国内某研究型大学 2010 – 2016 学年度的课程分数和教学评估数据,利用该大学在 2015 年春季学期放松优秀率控制的这一自然实验来分析学生成绩与评教分数之间的因果关系。在放松优秀率之前优秀率没有超过 40% 的课程现在可以自由地提高学生成绩的优秀率的比例,而那些优秀率已经超过 40% 的课程应该不受这次改革的影响,可以成为上述实验组的对照组。从描述统计来看,这两类课程在改革之前虽然在绝对水平上有显著的差异,优秀率更高的课程的学生评教结果也更好,但是两者的发展趋势并没有特别显著的差异。但是在改革之后,两者之间无论是在优秀率、课程分数和学生评教分数上的差距都缩小了,实验组的课程在快速追赶对照组的课程。双重差分回归为这个政策改革提供了更加精确的效果评估:优秀率的放松使得实验组课程的分数相对于对照组课程平均提高了 1.0 分(相当于对照组改革前课程分数的标准差的 1/4),使得实验组的评教分数相对于对照组平均提高了 0.5 分(相当于改革前对照组的 1/12 个标准差)。工具变量回归结果显示:对于那些改革后确实发生了分数膨胀的课程,学生通过学生评教分数给予了更加慷慨的回馈,课程分数每提高 1 分,学生评教分数显著提高 2 分。因此我们有理由相信,学生和老师之间围绕着学生评教分数和课程分数产生了"合作"和"共谋"。相比美国的文献,我们的研究更进

一步分析了这种“共谋”在不同身份的教师和不同类型的课程上的差异。具体而言,这种“共谋”在非教授开设的课程和专业课上体现得更为明显,这很可能是因为还未晋升教授的教师在面对学校人事晋升中的教学数量和教学评估要求时压力更大,因此更有动机与学生形成更加紧密的“共谋”关系。专业课上的“共谋”更加明显,很可能是因为在专业课上学生和老师能够获得多次博弈的机会,从而建立起彼此信任的关系,而不像全校公共课那样只有一次打交道的机会。这两个发现更加深入细致地刻画了学生与教师围绕教与学展开的共谋。对于高等教育政策制定者和管理者来说,应该更加全面地认识学生评教这个工具,不要机械地把学生评教作为一个硬性指标使用在教师职称的评审中,否则过于简单的量化指标可能会导致教师和学生采用不合理的策略来应对。一旦这种教学相“涨”的行为经过沉淀成为高等教育场域的一种“惯习”,那么就意味着学生评教制度将成为滋生异化教学关系的土壤,而不是破解委托代理问题的工具,高等教育“立德树人”的目标将难以实现。

本文也存在一定的局限。首先,本文的研究结论基于一个高校的情况,缺乏推广至我国高校整体情况的有效依据;其次,对政策改革后的观察只持续了三个学期,缺乏较长时段的观察;再次,限于篇幅,本文只将课程按性质加以分类,而没有按学科类型、院系等进一步划分。此外,本文使用的是课程层次的数据,未来可以进一步使用学生个人层次的跟踪数据来研究学生和教师之间的动态博弈关系。希望在后续研究中能对本文的局限进行补充,并对上述问题展开深入探讨。

参考文献:

曹辉、钮菁菁,2017,《高校学生网上评教:再评价与奖惩机制探索》,《改革与开放》第 3 期。
方乐,2017,《国际高等教育质量保障改革发展及其启示》,《上海教育评估研究》第 5 期。
黄达人等,2015,《大学的根本》,北京:商务印书馆。
何晓芳,2012,《大学治理场域中的资本、惯习与关系》,《大连理工大学学报(社会科学版)》第 3 期。
赖勇强、吕伟、叶逢福,2017,《地方院校学生评教文化现象透视》,《教育评论》第 1 期。
李斐,2015,《论我国高校教学与科研关系的演变与协调发展》,《高校教育管理》第 1 期。
李兰、哈巍,2017,《“百人计划”对中科院科研生产力的影响(1993 – 2004)》,《清华大学教育研究》第 5 期。
李松林,2007,《课堂场域中的权力运作》,《教育理论与实践》第 1 期。
李雅娟,2016,《高校学生评教:名存实亡》,《中国青年报》1 月 18 日。

刘妙龄,2004,《高校学生评价教师教学的有效性研究》,华中科技大学博士学位论文。
刘振天,2017,《教学与科研内在属性差异及高校回归教学本位之可能》,《中国高教研究》第6期。
宋伟新,2016,《国际与中国高等教育质量保障的新进展与发展方向——基于"高等教育质量保障:国际经验与中国探索国际研讨会"的分析》,《教育探索》第12期。
王洪明、佟曾,2007,《高校扩招与高等教育质量问题研究》,《理论界》第1期。
吴越、李惠娜,2016,《多元化巨型大学的权力结构关系新阐释——基于教育场域的视角》,《现代大学教育》第2期。
谢安邦、韩映雄、荀渊,2005,《高校扩招后教学质量调查与分析》,《教育发展研究》第15期。
谢小燕、顾来红、徐蓓蓓,2014,《新管理主义的评估问题剖析与"第四代评估"理论的借鉴》,《南京理工大学学报(社会科学版)》第2期。
于佳鑫,2017,《国家创新体系背景下辅助人员对科研生产力的影响——以中国科学院为例》,北京大学教育学院硕士学位论文。
袁祖望,2002,《高校重科研轻教学现象透视及纠偏》,《汕头大学学报(人文社会科学版)》第1期。
张端鸿,2016,《回归常识才能避免学生评教名存实亡》,《中国青年报》12月20日。
周继良、龚放、秦雍,2017,《高校学生评教的制度定位逻辑及其纠偏》,《中国高教研究》第11期。
朱富强,2016,《已成"毒瘤"的高校评教体系:言必称美国》,《中国社会科学报》3月29日。
Butcher, K. F., P. J. McEwan & A. Weerapana 2014, "The Effects of an Anti-grade-inflation Policy at Wellesley College." *Journal of Economic Perspectives* 28(3).
Correa, H. 1987, "Teacher-student Interaction: A Game Theoretic Extension of the Economic Theory of Education." *Mathematical Social Sciences* 13.
Cummings, W. K. & K. Santner 2015, "What Happened to Universal Education? In the West and in Asia." In I. Shin, G. Postiglione & F. Huang(eds.), *Mass Higher Education Development in East Asia*. Cham:Springer.
Ehlers, T. & R. Schwager 2016, "Honest Grading, Grade Inflation, and Reputation." *Cesifo Economic Studies* 62(3).
Eisenhardt, Kathleen M. 1989, "Agency Theory: An Assessment and Review." *Academy of Management Review* 14(1).
Franz, Wan-Ju Iris 2010, "Grade Inflation under the Threat of Students' Nuisance: Theory and Evidence." *Economics of Education Review* 29(3).
Gorry, D. 2017, "The Impact of Grade Ceilings on Student Grades and Course Evaluations: Evidence from a Policy Change." *Economics of Education Review* 56.
Hornsby, D. J. & R. Osman 2014, "Massification in Higher Education: Large Classes and Student Learning." *Higher Education* 67(6).
Hou, Y. W., C. W. Lee & M. G. Gunzenhauser 2017, "Student Evaluation of Teaching as a Disciplinary Mechanism: A Foucauldian Analysis." *The Review of Higher Education* 40(3).
Jacob, W. J. & V. Gokbel 2018, "Global Higher Education Learning Outcomes and Financial

Trends: Comparative and Innovative Approaches." *International Journal of Educational Development* 58.

Kanagaretnam, K., R. Mathieu & A. Thevaranjan 2003, "An Economic Analysis of the Use of Student Evaluation: Implications for Universities." *Managerial and Decision Economics* 24.

Krautmann, A. C. & W. Sander 1999, "Grades and Student Evaluations of Teachers." *Economics of Education Review* 18(1).

Langbein, L. 2008, "Management by Results: Student Evaluation of Faculty Teaching and the Mismeasurement of Performance." *Economics of Education Review* 27.

McPherson, M. A., R. T. Jewell & M. Kim 2009, "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classe." *Eastern Economic Journal* 35.

Miller, J. E. & P. Seldin 2014, "Changing Practices in Faculty Evaluation: Can Better Evaluation Make a Difference?" American Association of University Professors (http://www.aaup.org/article/changing-practices-faculty-evaluation#.VulYjE0UWpo).

Murray, H. G. 2005, "Student Evaluation of Teaching: Has It Made a Difference?" (https://www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of-Teaching1.pdf)

Organisation for Economic Co-operation and Development (OECD) 2013, "Measuring Learning Outcomes in Higher Education: Lessons Learnt from the AHELO Feasibility Study and Next Steps." (http://www.oecd.org/site/ahelo/)

Rojstaczer, S. 2016, "Grade Inflation at American Colleges and Universities." (http://www.gradeinflation.com)

Ryan, T. 2015, "Quality Assurance in Higher Education: A Review of Literature." *Higher Learning Research Communications* 5(4).

Shapiro, Susan P. 2005, "Agency Theory." *Annual Review of Sociology* 31.

Shin, Jung Cheol 2015, "Mass Higher Education and Its Challenges for Rapidly Growing East Asian Higher Education." In I. Shin, G. Postiglione & F. Huang(eds.), *Mass Higher Education Development in East Asia*. Cham:Springer.

Shin, Jung Cheol & Barbara M. Kehm(eds.) 2013, *Institutionalization of World-class University in Global Competition*. Dordrecht/New York:Springer.

Stratton, R. W., S. C. Myers & R. H. King 1994, "Faculty Behavior, Grades and Students Evaluations." *Journal of Economic Education* 25.

Trow, M. 2007, "Reflections on the Transition from Elite to Mass to Universal Access: Forms and Phases of Higher Education in Modern Societies since WWII." In J. J. F. Forest & P. G. Altbach(eds.), *International Handbook of Higher Education*. Dordrecht: Springer.

Uttl, B., Carmela White & Daniela Gonzalez 2017, "Meta-analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." *Studies in Educational Evaluation* 54.

作者单位:北京大学教育学院、教育经济研究所(哈巍)
北京大学政策法规研究室、社会学系(赵颖)
责任编辑:杨　可