

教师评价多元化发展的国际前沿考察

刘 骥 李海茜

(陕西师范大学 教育学部 西安 710062)

摘 要: 建设教育强国是实现中华民族伟大复兴的基础工程,而教师是立教之本、兴教之源,对学习品质、思想道德、性格养成、心理健康等方面有着举足轻重且多元多样的影响。面对高质量发展阶段的全新教育需求,教师评价创新势在必行。科学有效的教师多元评价体系能激发教师全面育人的工作积极性,提升教师工作效能与职业满意度,从制度设计上为提高教学质量和促进学生全面发展提供有利基础。在工具选择上,教师评价多元化发展,形成了教师增值评价、学生成长百分位、利益相关者反馈、教学活动观察等四种主流评价手段。结合国际实践经验总结认为教育评价的多元化与教育发展的现代化息息相关,不仅反映了人才培养理念的现代化趋势,同时在教学评价方法论上提供了新的科学启示,有益于我国在新时期释放教育评价多元化发展的价值潜力。

关键词: 教师评价; 多元评价; 破五唯; 评价工具; 国际镜鉴

中图分类号: G40-058.1 **文献标识码:** A **文章编号:** 1001-5957(2022)03-0088-08

党的十九届五中全会明确提出到 2035 年中国将建成教育强国的宏伟目标。^①2020 年 10 月 21 日,中共中央、国务院联合印发了《深化新时代教育评价改革总体方案》(以下简称《总体方案》),吹响了新时代教育评价改革与教育现代化跨沟迈坎的冲锋号。^②《总体方案》不仅强调教育评价对教育发展方向的指引与引领作用,还针对当前教育评价的突出问题作出精准设计,指明了我国教育评价继续向科学化、事实化、多元化的方向发展。^③在教育现代化的新历史征程上,教育评价现代化不应独立于教育事业发展的总体现代化,而应积极衔接教师评价多元化发展的国际方法论前沿,以服务中国特色社会主义教育发展道路为基本遵循,坚持培养德智体美劳全面发展的社会主义建设者和接班人,充分考虑教师在组织和实施教学活动中五育并举的多维贡献,积极探索教师评价方式的多样性方案,进一步深化现有教育评价改革、优化现有教师评价模式。基于此,本文立足新时代教师评价改革的重大现实需求,通过国际比较视角,梳理教师评价多元发展的国际前沿进展,为构建符合我国新时代教育发展特点的教师评价体系总结有效国际镜鉴。

一、教师评价多元化发展的样态沿革

近些年教育在国家和地区发展中所扮演的角色愈发重要,各国对于教师评价也愈发重视,特别是占据全球技术与价值链优势位置的西方高收入发达国家,在教师评价的多元化改革探索中积累了大量实践经验。早在 1966 年,美国学者科尔曼(James Coleman)在美国联邦教育部门的经费支持下开展了对近 4000 所学校

收稿日期: 2022-02-15

基金项目: 国家社会科学基金教育学项目(CJA200256); 陕西省创新创业项目(S202110718337)

作者简介: 刘骥 陕西师范大学教育学部教授, 哲学博士; 李海茜 陕西师范大学教育学部项目助理。

① 陈宝生《锚定 2035 年建成教育强国的目标,全力抓好贯彻落实五中全会大学习》,《中国青年报》2020 年 11 月 9 日。

② 杜玉波《吹响新时代教育评价改革的号角》,《人民日报》2020 年 10 月 22 日。

③ 顾明远《教育评价改革,关键在落实》,《人民日报》2020 年 10 月 21 日。

共计 65 万名学生的调查研究。这被认为是迄今为止最早且规模最大、基于实证调研的多因素教育评价探索之一。^① 此后 欧洲国家也纷纷开展教师与学校评价改革 特别是自 20 世纪 90 年代以来英国开始推行的成长型教师评价。^②

新世纪以来 伴随着全球化背景下以教育质量提升为主旨的教育改革浪潮席卷各国 教育政策研究与实践愈发关注对教师评价多元化、科学化、模式化发展的探索。^③ 在我国 完善教育评价制度是新时代实现国家治理与教育发展双现代化进程中的一项重要战略。但深化教育评价改革的现实需求与教师评价技术发展不充分的矛盾日益凸显 因此国际新兴的教师评价多元化模式研究在评价维度、数据获取、分析思路等方面 正通过引入新视角推动教师评价的科学性与客观性发展。有学者认为教师评价领域亟须深度挖掘教师自身发展潜能并积极提升其实践性知识与创造性智慧 进而强化评价促进教师全面进步的发展性功能。^④ 针对如何选拔、培育、激励优秀老师等核心问题 国际上形成了以教师增值评价、学生成长百分位、利益相关者反馈、课堂教学观察为主要样态的新兴教师评价形式。

(一) 教师增值评价

教师增值评价(Teacher Value-Added Assessment ,TVA) 是利用教育统计技术手段对教学效能与学习效能转化效果所开展的增值性评价 其核心目的是运用实证数据衡量单位教学时间内教师教学活动对学生学习效果所产生的增量影响。从方法论起源上看 教师增值评价可溯源至美国学者科尔曼于 1966 年向美国国会提交的《教育机会公平报告》(Equality of Educational Opportunity Report) 。该报告虽然没有直接提出教师增值评价这一概念 但其基于教育生产函数的研究方法在事实上开启了多因素教育评价的增值效应评估范式 进而启发了有关教育增值性评价的广泛探索。^⑤ 受科尔曼大规模实证研究所产生的广泛社会影响的启发 美国各州纷纷开展教育增值性评价探索 其中较具应用规模的是由田纳西州开发的 TVAAS 体系(Tennessee Value-Added Assessment System) 。自 1992 年以来 田纳西州政府签署《教育提升法案》(Education Improvement Act) 该州在全域内开展教师增值评价 长期运用大规模学生学业表现数据监测单位学年内教师教学效能与学生学习效能的转化增量。随着教师增值评价在田纳西的试点 以增值性评价为实质内核的多元增值评价进入快速发展的时期。如今 类似 TVAAS 的教师增值评价体系已被引进至美国 30 余州的教育问责机制中。^⑥ 而在英国 也于近些年开展全面的教育增值评价的实践探索 并将教师增值评价结果作为一项重要的制度创新纳入现有的教师考核体系中。^⑦ 从方法论实质上看 教师增值评价的核心是通过追踪调查的实证统计方法衡量学生在单位教学时间内学习结果发生的变化 并通过建立多因素回归分析模型 分离出人口学因素、家庭背景、生源情况等不可控的非教学因素对学生学习效果的影响 然后单独考察教师教学活动对学生学习效果影响的“净效应”(net effect) 实现对教学效能与学习效能转化关系的科学、客观评价。^⑧ 教师增值评价自田纳西试点以来 不仅广泛影响了美国本土的教师评价模式 也成为其他发展中国家和地区对教师评价进行探索、研究与实践的重要参照。例如在智利 有研究通过教师增值评价的计算 从教育公平的视角估测优秀教师对学生学业成长的贡献 进一步剖析优秀师资资源分布情况与实现教育公平之间的密切联系。^⑨

① 杨文杰、范国睿《教育机会均等研究的问题、因素与方法〈科尔曼报告〉以来相关研究的分析》,《教育学报》2019 年第 2 期。

② 雷万鹏、马红梅、黄华明《基于学生成绩残差分解技术的教师教学质量评价》,《华东师范大学学报(教育科学版)》2021 年第 7 期。

③ 刘骥、黄少澜《教师技能对教育优质均衡的重要作用:基于跨国数据的实证》,《教师发展研究》2020 年第 4 期。

④ 刘志军、王洪席、张红霞《促进教师不断发展的评价体系构建》,《清华大学教育研究》2015 年第 6 期。

⑤ 谌启标、柳国辉《西方国家教师效能研究发展述评》,《教育研究》2011 年第 1 期。

⑥ 岳伟、陈俊源、胡祥恩、莫伦《从 TVAAS 到 Individual Growth: 美国田纳西州增值性教师评价发展新趋势》,《现代教育管理》2019 年第 9 期。

⑦ 崔秀兰、姜君《英国教师专业标准评价指标及价值取向的变迁》,《黑龙江高教研究》2019 年第 4 期。

⑧ R. Chetty, K. Friedman & J. Rockoff, “Using Lagged Outcomes to Evaluate Bias in Value-Added Models”, *American Economic Review* 5(2006) : 393 - 399.

⑨ R. Torres, “Tackling Inequality? Teacher Effects and the Socioeconomic Gap in Educational Achievement: Evidence from Chile”, *School Effectiveness and School Improvement* 3(2018) : 383 - 417.

经过多年来教师增值评价模型不断改进,当前国际上应用较广泛的增值评价模型可以划分为两大类。其一为教师增值模型(Value-Added Model,VAM),也就是以学生前后两次成绩的差值为因变量,运用教师固定效应预测单位教学时间内任课教师对学生分数增长的影响。在具体计算中,VAM模型常通过运用教育行政口径的面板追踪数据,加入学生层面协变量(X_{isjt})、学校层面协变量(S_{isjt})针对学生*i*于年份*t*在学校*s*任课教师*j*的指导下所产生学习增量($[A_{isjt} - A'_{isj(t-1)}]$)估计的偏误进行矫正,旨在较为准确地计算每名教师的固定效应(T_{isjt}),而教师固定效应所对应的系数 θ 也就是VAM模型中的计算核心:教师增值。从数量关系上看,VAM模型的具体估算常采用如下形式:

$$[A_{isjt} - A'_{isj(t-1)}] = \beta_0 + X_{isjt} \cdot \beta_1 + S_{isjt} \cdot \beta_2 + T_{isjt} \cdot \theta + \varepsilon_{isjt}$$

其二为学生固定效应模型(Pupil Fixed-effect Model,PFM),也就是通过学生固定效应的实施,将不同任课教师在同一名学生所学不同科目中所产生的分数增长进行比较评价。简言之,PFM模型是通过运用学生固定效应尽可能排除学生及学校层面协变量的偏误影响,进而将教师增值得分计算固定在同一名学生不同科目表现间的差异,由此实现教师增值相对准确的估计。与运用教师固定效应估算所不同的是,PFM模型的学生固定效应主要利用不同科目*k*间的学习分数变化开展对学生*i*于年份*t*在学校*s*任课教师*j*的指导下所产生学习增量($[Y_{isjkt} - Y'_{isjkt(t-1)}]$)的估算,也就是利用同一名学生的家庭背景、班级状况、校园环境等因素均相同的特点开展不同任课教师的增值比较。在计量样态上,PFM模型的运算逻辑常采取如下思路:

$$[Y_{isjkt} - Y'_{isjkt(t-1)}] = \gamma_0 + X_{isjkt} \cdot \gamma_1 + S_{isjkt} \cdot \gamma_2 + T'_{isjt} \cdot \theta' + \mu_i + \varepsilon_{isjkt}$$

(二) 学生成长百分位

虽然近些年西方发达国家教育财政经费、高学历教师比例不断提升,但民众却普遍抱怨教育质量不佳、水平提升停滞。新世纪以来,美国小布什政府大力推行《不让任何孩子掉队》法案(No Child Left Behind Act)。在这样的大政策环境下,学生成长百分位模型(Student Growth Percentiles,SGP)被赋予了推动教育评价改革的重任。^①该模型最初由美国教育统计学者贝特本纳(Damian Betebenner)提出,率先于美国科罗拉多州全面推广试行,其模型的计算实质是比较不同教师所指导的相似学生是否在单位教学时间内存在学习增量上的相对差异,从而对教师教学表现进行评价。^②与VAM模型关注学生在单位教学时间内产生的学习增量不同的是,SGP模型所估算的是任课教师对相同基础的学生所产生的相对学习增量,通常以每名任课教师所有学生相对学习增量的中位数或平均值作为具体得分的依据。^③近些年,SGP模型在英国、澳大利亚等国家得到迅速推广,其中最重要的原因之一是其将数据统计与教学实际中所关心的学生成长进行了有效结合,也就是帮助教育管理者更直观地理解教师在单位时间内,其指导学生学习结果位次的相对变化。在实践中,SGP模型的应用广度与工具本身的精度也在不断发展,例如有研究利用教育生产函数框架开展验证性测算发现师生关系品质对学生成长百分位的变化有重要影响。^④此外,新近研究发现随着学生学业表现数据观测点数量的增加,SGP模型的偏误逐渐减小,意味着SGP模型本身的精度将有潜力随着学生数据量增大而得到提升改善。^⑤

从方法论实质上看,SGP模型通过运用分位数回归方法,将学习增量定义为在条件密度函数下学生所处百分位的变化,以学生在不同时间点的前测成绩作为条件变量,对学生成长百分位的计算采取基于与学生在时间点*t*下成绩相关的条件密度函数开展估计。^⑥简言之,SGP模型估计的核心在于对基线水平相似的学生群体中开展针对学习增量的横向比较,而由于所有百分位的变化都是较相应条件密度函数而言,因此学生百

① A. Ismail, "Impact of Contextual Predictors on Value-Added Teacher Effectiveness Estimates", *Education Science*, 10 (2020): 390.

② D. Betebenner, *A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections Trajectories* (Dover, New Hampshire: The National Center for the Improvement of Educational Assessment 2011).

③ D. Betebenner, "Norm- and Criterion-Referenced Student Growth", *Educational Measurement: Issues and Practice* 4 (2009): 42-51.

④ D. Dennie, P. Archarya, D. Greer & C. Bryant, "The Impact of Teacher-Student Relationships and Classroom Engagement on Student Growth Percentiles of 7th and 8th Grade Students", *Psychology in Schools*, 5 (2019): 765-780.

⑤ K. Castellano & D. McCaffrey, "Estimating the Accuracy of Relative Growth Measures Using Empirical Data", *Journal of Educational Measurement*, 1 (2020): 92-123.

⑥ J. Lockwood & K. Castellano, "Alternative Statistical Frameworks for Student Growth Percentile Estimation", *Statistics and Public Policy*, 1 (2015): 1-9.

分位的变化也同时反映了学生的绝对学习增量和教师的相对学习贡献。由此可见,SGP模型就是通过评价任课教师在单位时间内给同一水平的学生所产生的额外学习增量用于评价教师表现,从而更客观地反映教师对不同起点学生的学业影响。在具体的分位数回归计算中,SGP模型依托每名学生的前测成绩所对应的条件密度函数以此考察时间点 t 下的成绩相对位置。在实践中,时间点 t 下的学生成绩相对位置可用于衡量该段时间内所产生的学习增量是否符合预期、低于预期或超过预期。在实际数据处理中,条件分位数函数的估算往往要求至少收集两次前测成绩,即时间点 $t \geq 2$,方可估计时间点 t 下学生成绩 Y_t 所对应的条件密度函数。如下公式所示,每 τ 个百分位的条件密度函数分别由每一项前测成绩 Y_j 的 φ 基函数决定。

$$Q_{Y_t}(\tau | Y_j) = \sum_{j=1}^{t-1} \sum_{i=1}^3 \varphi_{ij}(Y_j) \beta_{ij}(\tau)$$

(三) 利益相关者反馈

作为量化数据与质性数据的交叉结合产物,利益相关者反馈调查(Stakeholder Feedback Survey,SFS)是一种通过问卷和访谈形式针对教育利益相关者的教师评估形式,通常被认为是对于现有基于学生成绩单维考量方式的重要补充。目前许多发达国家都已开始利用SFS进行教师工作效能的多维考察,其中以美国各州为例,SFS主要用于收集包括学生、家长、教师、教育管理者等多方教育利益相关者的评估意见。从历史脉络上看,SFS已有较长的发展历程。早在1896年,美国艾奥瓦州率先开始向2至8年级的学生收集有关教师教学情况的调查意见。^①此后各州纷纷效仿,将教师他评信息逐步纳入教育问责改革措施中,特别是在《每一名学生成功法案》(Every Student Succeeds Act)通过后,美国联邦教育部门格外重视SFS在补充标准化测试作为教师教学度量主要维度中的关键作用。截至2018年底,美国至少有34个州已明确立法要求将SFS纳入教师评价的核心工具库。伴随各国(地区)教育评价体系对于加强教育质量的公信监督的发展需要,SFS的应用在近些年迅速拓广。例如在英国,基于数字化测评的艾杜罗平台(Edurio)通过收集以学校为核心的教育社会网络的教育评价信息,以模块化、信息化、可视化手段将教育利益相关者的实践体验呈现于教育管理部门,据此为学校课程设计、就业能力追踪、教育质量提升提供参考指南。^②在我国香港,特区教育主管部门对于教师评价改革过程也关注SFS对学校自我评估机制与教育质量保证架构的重要作用,鼓励一线学校通过搜集家长、教师、教辅人员及校工对学校不同教育教学范畴的建言,开展自查自检,推动全教学链质量的提升。^③

在实施模式上,SFS又可细分为学生反馈、同行反馈、家长反馈三大类,其调查对象、调查目的、调查设计各有不同。首先,学生反馈是以学生为调查对象,以考察学习体验、师生互动、教学效果为调查目的,通常选择运用李克特量表对教师表现的不同维度进行评价。例如,美国各州教师评价系统中应用广泛的Tripod工具将学生评教按照“7c”维度分类加以体系化,即关心(care)、协商(confer)、吸引力(captivate)、清晰度(clarity)、知识巩固(consolidate)、挑战性(challenge)、课堂掌控(classroom management)。^④其次,同行反馈,即教师对彼此教学能力和教学表现的客观评议也是SFS的重要组成部分。教师的同行反馈由来已久,特别是资深教师评教、向年轻教师提供业务指导在许多学校已经是“帮扶”机制的重要组成部分。然而随着教学理念、内容、技术的快速发展,年轻教师在分享新教学观点、方法等方面发挥着越来越重要的作用。因此,许多教育管理者逐渐开始发现同行反馈作为治学治教工具的重要潜力,特别是应用于考察师资管理中非显性却又重要的评价维度,如课堂准备、专业素养等方面。最后,家长是教育活动不可或缺的利益相关者,可以站在较全面的视角观察儿童发展,特别是对课堂教学成效、家校教育联动效果有着切身体会。然而在相当长的一段时期,家长对教育活动的看法在教师评价体系中并未得到充分体现,直到近些年,家长反馈才逐步成为

① A. Tuma, L. Hamilton & T. Tsai, *A Nationwide Look at Teacher Perceptions of Feedback and Evaluation Systems* (Santa Monica CA: RAND Corporation 2018).

② L. Rose, M. Manolis & C. Mutlu, “Edurio: Deep Analytics of Stakeholder Feedback for School Networks” (2017-04-07) [2022-02-23]. <https://www.ucl.ac.uk/ioe/case-studies/2017/apr/edurio-deep-analytics-stakeholder-feedback-school-networks>.

③ 香港特别行政区教育局《学校自我评估机制之持份者问卷调查》,中国香港:香港特别行政区教育局 2021年。

④ L. Desimone, T. Smith & D. Frisvold, “Survey Measures of Classroom Instruction Comparing Student and Teacher Reports”, *Educational Policy*, 2(2009): 267-329.

SFS 校外评价的主要途径之一。例如 2011 年 纽约儿童权益倡导组织(Action for Children) 首次将家长反馈纳入纽约市教师评价系统并展开倡议,认为家长反馈对现有的评教方式有重要补充的作用。^① 自 2012 年以来,越来越多的教育主管机构开始制定收集家长反馈用于教师评价的方案,如修订后的《教学框架》(Framework for Teaching) 认为,教师的教学表现不仅仅体现在课堂观察中,其专业责任(professional responsibility) 还应包括建立有效家校沟通、促进课堂外学习等措施,因此建议全面且公正的教师评价体系应考虑到优秀教师可能对学生课外表现所产生的重要贡献。^②

(四) 课堂教学观察

教师评价发展历程中历史最悠久的评价形式便是课堂教学观察(Classroom Teaching Observation ,CTO) ,但长久以来,CTO 一直未形成系统化、标准化的评价范式,缺乏客观统一的评价标准,因此被认为其有效性有待商榷。然而自上世纪 60 年代以来,英美等西方发达国家对于教学结果的关注并未完全解决教育质量提升的困境,因此近些年学者和管理者们再次将目光投向教学过程性评价,期望从课堂教学中找出优质教学发生的机制。例如,美国在奥巴马政府时期,联邦教育部所实施的“力争上游”计划(Race To The Top) 将课堂观察式教师评价作为考核各州教育绩效的重要指标之一。^③ 而田纳西州于 2013 年开始推行“全面有效性测评”改革(Overall Effectiveness Assessment) ,即通过课堂教学观察的科学化发展实现对教师教学表现综合评价与提升的目标。^④ 此外在荷兰、韩国、南非、印尼、巴基斯坦等国,CTO 的课堂应用被认为有助于促进教师教学的公平性,特别是通过观察式与评价式指导相结合帮助教师适应教室中愈加丰富的多元化学习需求,以尽可能有效的教学方式支持不同类型的学生成长。^⑤

综合地看,国际上无论是强调标准化教学的教改框架,又或是以教学结果为导向的评教系统,在近些年都趋向于增加针对课堂实践、教学过程的权重。相应的,CTO 发展至今已延伸发展出不同类型与形式,虽然评价工具有别,但其评教实质都在于探究教师在教学发生与学习发生中所产生的过程性价值(formative value) 。^⑥ 第一,设计基于 CTO 的评教方案必须是基于深度理解教师课堂实践在促进学习发生过程中的基础作用,并明确清晰的教学标准和与其相配套的观察量表是描述、评价和提升教师教学实践的核心路径。例如,智利所推行的“优异教学框架”(Marco para la Buena Enseñanza) 制定了详尽的课堂观察方案,包含了课堂呈现方面,如时间和课堂管理、差异化教学、师生互动以及学科教学方面,如学科知识讲授、教学与学习适配情况等。^⑦ 第二,有关 CTO 的国际研究认为教师在教学活动中所产生的过程性价值一定是结合具体学情、促进学习发生、具有复杂特征的教学行为,因此教学评价方案必须在实际考量中允许教师结合教学实境、因材施教的灵活创意。例如,新加坡教师“表现提升管理系统”(Enhanced Performance Management System) 就明确提出将教学创造力列为 CTO 评教过程中衡量教学品质的重要维度,鼓励教师根据具体学情状况开展针对性教学策略。^⑧ 第三,高品质 CTO 的实施有利于通过评教识别教学短板、发现教学问题,从而精准促进教学水平的提升,形成教学相长的正反馈循环。在我国先秦儒家思想巨著《礼记·学记》中以“君子既知教之所由

① J. Grissom & P. Youngs , *Making the Most of Multiple Measures: The Impacts and Challenges of Implementing Rigorous Teacher Evaluation Systems* , (New York , N. Y. : Teachers College Press 2015) .

② D. English , J. Burniske , D. Meibaum & L. Lachlan-Haché , *Uncommon Measures: Student Surveys and Their Use in Measuring Teaching Effectiveness* , (Washington , D. C. : American Institutes for Research 2015) .

③ U. S. Department of Education , *Race to the Top Program* , (Washington , D. C. : U. S. Department of Education 2009) .

④ F. Martinez , S. Taut & K. Schaaf , “Classroom Observation for Evaluating and Improving Teaching: An International Perspective” , *Studies in Educational Evaluation* 49(2016) : 15 – 29.

⑤ R. Maulana , S. Andre , M. Helms-Lorenz , J. Ko , S. Chun , A. Shahzad , Irinidayani , O. Lee , T. De Jager , T. Coetzee & N. Fadhilah , “Observed Teaching Behavior in Secondary Education Across Six Countries: Measurement of Invariance and Indication of Cross-national Variations” , *School Effectiveness and School Improvement* 1(2020) : 64 – 95.

⑥ G. Whitehurst , M. Chingos & K. Lindquist , “Getting Classroom Observations Right: Lessons on How from Four Pioneering Districts” , *Education Next* 1(2015) : 62 – 68.

⑦ F. Martinez , “Combining Multiple Measures of Teacher Practice and Performance: Technical and Conceptual Considerations for Teacher Evaluation” , *Pensamiento Educativo Latinoamericano* 1(2013) : 4 – 20.

⑧ S. Scalfani & E. Lim , *Rethinking Human Capital in Education: Singapore as a Model for Teacher Development* , (Washington , D. C. : Aspen Institute 2008) .

兴,又知教之所由废,然后可以为人师也”指明了为师者科学掌握教学基本规律的重要性。可见,教师不仅需要知道如何教得好,更需要理解教学改进的方略。在当代的教学改革实践中,大量国际研究也表明依据教学观察记录开展教学诊断、提供实境反馈、制定提升方案可以有效地助推教师专业性成长。^① 第四,测量工具的信效度作为设计 CTO 实施方案时的重要考量方面,不仅决定着科学评教的有效性,更影响着教师评价体系的公信度。因此,许多国家都格外重视 CTO 实践中“谁来观察”和“如何观察”的现实问题。例如在美国洛杉矶市,教育主管部门有着对教学观察员十分严格的资格准入制度,要求教学观察员在获得评教资格前至少接受 20 小时的方法理论培训和教学实境随训,而在芝加哥市,则选择运用课堂教学视频录像等电子信息采集手段,由资深的持证教学观察员开展集中评教评审并提出相应改进建议。^②

二、教师评价工具评析

在宏观战略层面,《总体方案》明确指出教育评价事关教育发展方向,有什么样的评价指挥棒,就有什么样的办学导向。教育评价改革事关教育现代化的总体进程,是引导教师立德修身、潜心治学,促进学生全面发展、教育全面现代化的基础制度保障,而贯彻落实习近平总书记关于教育的重要论述和全国教育大会精神,探索更加科学多元、符合国家育人需求的教师评价方式是重点攻坚任务。为了更好地服务改革实践,本研究深入开展教师多元评价工具的文献研究,积极对接国际前沿的教师评价改革经验,以分类归纳评价工具的专题形式呈现(见表 1),尝试勾勒融通各类工具之长、符合我国教育实情、适宜新时代教育需求的教师评价多元化发展图景。

表 1 国际教师评价工具评析

	教师增值评价	学生成长百分位	利益相关者反馈	课堂教学观察
优势	- 内部效度高 - 关注学习增量	- 概念直观 - 关注相对表现	- 多维度、多渠道 - 易于实施	- 实践经验成熟 - 利于教师成长
挑战	- 学科适用度 - 数据要求高	- 受班级构成影响 - 易造成无效竞争	- 受主观态度影响 - 基础教育应用少	- 对观察者要求高 - 受霍桑效应影响

第一,TVA 的优势主要表现在其内部效度高、统计方法成熟,通过分析学生成绩变化轨迹,将教师贡献从非教学因素影响中分离出来,单独考察教师对学生学习增量的“净效应”,因此对学习增量的估计结果具有较强的客观信度,并能够相对有效地排除非教学因素对学习增量估计的偏误干扰。自上世纪 70 年代发展至今,TVA 已经形成了相对完善的基础模型构建,涵盖 VAM 和 PFM 两大类模型,可以有效满足对分科教学、团队教学等多种组织形式的分类评价需求。然而,TVA 在教师评价中的应用仍面临着一些现实挑战。其一是 TVA 的学科适用度相对较窄。无论是 VAM 或 PFM 模型,均需以学生成绩为教师增值估算的结果变量,因此其适用度与标准化测试的应用高度重叠,这也意味着如何将 TVA 实践于美术体育音乐等缺乏标准化评价测量工具的术科是当前最主要的应用研究重点。其二是 TVA 的统计算法对数据要求较高,其有效估算依赖于学生学业追踪面板数据库的构建,特别应涵盖对每名学生的跟踪式多时间点测评才能对其任课教师开展增值累计评价,因此建设 TVA 所适配的数据库是一套涉及学生、教师、学校等多层次、多维度与多观测点复杂相关信息准确对接的系统工程,需要评价维度、测量工具、数据管理等不同视角下的协同顶层设计与实施。

第二,SGP 的优势主要体现在两个方面。其一是 SGP 关注学生成长,并以百分位变化的直观形式度量同类型学生所处位置发生的相对变化、以学生成长进步幅度评价教师贡献,因此概念直观、易于理解,便于向学生、家长、教师等非教育统计专业人士进行宣传。其二是 SGP 所关注的核心变量为相对表现,也就意味着每名学生都按照自己的基准学业水平与相近的学生进行成长进步幅度的比较,从价值取向上引导教师关注每一名学生的学习品质,并由百分位的实际算法确保教师对每名学生成长的关照。与此同时,现有的国际实

① M. Kraft & F. Gilmour, “Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness”, *Educational Researcher* 5 (2017): 234–249.

② J. Hattie & H. Timperley, “The Power of Feedback”, *Review of Educational Research* 1 (2017): 81–112.

实践经验也表明SGP的实施仍面临着一些具体问题。一方面,SGP易受班级学业水平结构影响。由于SGP算法的核心在于对前测成绩所对应的条件密度函数的估计,因此班级整体学业水平的高低直接影响着条件密度函数的分布状态。由于SGP模型通常不会对学生前测成绩之外的非学业特征进行额外校正或调整,因而前测成绩所对应的条件密度函数分布状况差异往往又折射出班级构成在非教学因素上的异质性分布。另一方面,SGP虽然能相对准确估计“成长量”,但却往往无法回答“多少成长量是客观且适宜的”的现实问题。也正因此,由SGP所引入教师评价的学生成长概念,往往在实践中却可能造成无效竞争,促使课堂教学脱离教育基本规律,过度强化教学过程中对特定知识和技能的竞争性学习。因此许多已经开展SGP应用的国家,均开始制定科学客观的“成长标准”,以明确各年级、各科目间学生学习的规律性差异。^①

第三,SFS的优势主要表现在两个方面。一方面,SFS凸显了教师评价的多维度、多渠道特点。现有的SFS调查通常包括可量化与不可量化的综合教学信息,不仅仅只是考察学生成绩,而是尽可能多维度地收集学生、家长、教师、教育管理者等多方教育利益相关者对教学体验、教学品质、教学结果的评估意见,并且信息获取渠道形式多样,包括问卷调查、结构化访谈等多种采集样态。另一方面,SFS理念直观、易于实施。从数据收集的角度考虑,SFS并不需要复杂的算法模型与庞杂的学生数据库,更易于客观评教的实施开展,更益于在硬件条件不成熟的环境下推广。值得注意的是,当前SFS的评教实践仍面临诸多难点。首先,最为突出的是SFS易受主观态度影响,而现有的调查工具还存在较大局限性、缺乏普遍信服度基础。例如学生和家長在进行教师评价时可能带有主观色彩,而教师同行互评又可能受考核压力影响而产生无效竞争,因此SFS教师评价结果往往并非完全基于教学表现或贡献,而受非教学因素干扰较大。其次,SFS作为一种历史发展相对较短的评教工具,目前国际上的应用还主要停留在高等教育阶段,多用于高校教师的教学品质测评,仅有极少数国家将真正多维度、多渠道的SFS引入用于评价基础教育阶段教学,因此中小学评教的相关实例应用较少,相关经验还有待进一步挖掘。

第四,CTO的优势主要表现在应用范围广、实践经验成熟、教学针对性强、利于助推教师专业性成长等方面。首先,CTO的应用范围十分广泛,在各国的评教实践中均占有重要的位置,是许多教师十分熟悉且已接受的评价形式。作为当今教育评价学界最重要的共识之一,研究认为课堂教学实践是教育改革促进学习品质最关键且最有效的途径,事实上许多教师评价体系中将课堂教学观察视作识别优秀教师的“黄金准则”。^②其次,CTO具有教学针对性强、关注教学过程的特点,相较于前三种评教模式更利于“评与教”实现目标的统一。在CTO的实施中,教师评价不再尝试“从远处”抽象地理解优秀教学,转而以剧场式“身临其境”去感受高品质教学的发生全过程,用最直观与最微观的观察信息评价教师对学习发生的贡献,从而促进教师专业发展与课堂教学水平的实质性提升。当然,CTO在实践中还存在许多问题与优化的潜力。CTO的有效开展对于课堂教学观察者本身的教学水平要求很高,因为观察者必须要在优秀教学发生时能够有效识别并做出判断。更进一步看,观察者出现在课堂中本身也是对课堂教学环境的重要改变,而无论是被观察的教师又或是被观察的学生都易受“霍桑效应”影响而有意改变自身行为倾向,从而影响课堂观察的真实性。

三、建设新时代教师评价体系的路径选择

《总体方案》吹响了新时代教师评价改革的冲锋号,积极领会党的十九届五中全会对经济社会发展的全面部署,准确把握经济社会发展新阶段对加快教育评价体系现代化的新需求,建设高质量、高水平的教育评价体系是新发展理念的集中体现,是建成教育强国必然之选择。在建设现代化教育强国的时代命题下,树立科学的教育发展观、构建教师评价的多元化发展方案显得尤为重要。在新时代,教师评价体系被赋予了更科学、更客观、更全面的高质量发展方向。作为课堂教学活动的核心组织与实施者,教师是满足高质量教育发展的主力军,其对于学生全面发展的贡献并不局限于学业成绩,而对学生品德培育、心理健康等方面有着举足轻重的影响。由此可见,单纯以升学或成绩作为教学活动成效的评价标准是不科学、不客观、不全面的,远

^① W. Elias & E. Isenberg, “How Does Value Added Compare to Student Growth Percentiles?” *Statistics and Public Policy*, 1 (2015): 1–13.

^② A. Schleicher, “Building a High-Quality Teaching Profession: Lessons from Around the World” *Educational Studies*, 1 (2011): 74–92.

不能满足五育并举高质量教育发展的新时代需求。当前大量国际实证研究认为,虽然教师教学水平对学习品质有着重要作用,但学生成绩也受诸多非教学因素影响,比如家庭学习环境、家庭教育投入,甚至是班级同伴差异等因素都可能对学习品质和学习效果产生重要影响。^①

在理念层面,教师评价作为衡量教学活动有效性的重要环节,评价的主体内容关注教学质量与学习品质,评价的主要目的为指导教学实践与提升学习效果。在价值层面,科学、客观、全面的教师评价体系不仅有益于激发教师职业效能与教学活力,帮助教师更好地从教学实践中成长,更有益于吸引优秀人才投身教学事业。目前在国际上,有关教师评价多元化发展的研究通常旨在突破传统教师评价模式的弊端,进而建议采取更加科学、更加全面的评价手段,在测量评价过程中因地制宜、因校而变、因课施策,多元化、多维度、多渠道地考察教师在组织和实施教学活动过程中对学习发生、学习品质、学习效果的重要贡献。现有的国际评教工具发展动态从包容性、客观性、科学性等方面为构建新时代教师评价体系提供了镜鉴启示。首先,多元评教工具对不同层次、不同能力水平学生的覆盖,对不同地域、不同教学环境教学实境的考量,对不同年级、不同科目的针对性方案,对学生、家长、教师、教育管理者等多方教育利益相关者的分类考察,为构建包容性教师评价体系提供了价值借鉴。其次,多元评教工具的客观性主要体现在如何通过科学严谨的技术实践有效反映不同教育利益相关者对教学体验、教学品质、教学结果的切身感受与评估意见,进而满足多样化的育人需求,相应的提升教育教学效果。最后,多元评教工具的科学性强调在机制繁杂的教师评价系统中,对教学因素与非教学因素开展更加精准的度量,更加科学地分离识别教师与教学因素对促进学习发生、优化学习品质、提升学习效果的实际贡献,提高教师评价结果的科学信效度和公众信服力。

(责任编辑:杨霞 责任校对:杨霞)

Examining the Global Frontier on the Multidimensionality of Teacher Evaluation

LIU Ji , LI Hai-qian

(Faculty of Education , Shaanxi Normal University , Xi'an , Shaanxi 710062 , China)

Abstract: Establishing a world-class education system is considered foundational groundwork for realizing the great rejuvenation of the Chinese nation , and teachers have a critical value to contribute. As key planners , implementers , and participants in educational activities , teachers influence a wide range of student outcomes , including learning , morality , personality , and mental health. Design of teacher evaluation systems should be reflective of teachers' wide-range influence. An effective multidimensional teacher evaluation system not only incentivizes job passion , efficacy , and satisfaction , but also provides institutional foundation that advocates for instructional and learning gain. This synthesis of global evidence on multidimensional teacher evaluation identifies teacher value-added assessment , student growth percentiles , stakeholder feedback surveys , and classroom teaching observation of promise. Considering the international evidence on multidimensional teacher evaluation and its value informing design of teacher evaluation systems in China , development of teacher evaluation systems is strongly tied to educational modernization , and reflects both new trends in educational philosophy and methodological soundness.

Key words: teacher evaluation; multidimensional evaluation; breaking “Five-Only”; evaluative instruments; international comparison

① 刘骥《如何应对全球学习危机?——世界银行〈2018 世界发展报告〉述评》,《全球教育展望》2018 年第 6 期。