■ 教育数字化·高等教育大数据专题　　DOI:10.15998/j.cnki.issn1673-8012.2022.04.005

大数据时代的因果推断
——教育政策评估的新路径

郭　娇[1],吴寒天[2]
(1.华东师范大学　高等教育研究所,上海　200062;2.浙江大学　教育学院,杭州　310058)

摘　要:在研究方法与数据来源不断更新迭代的当下,对新兴路径与范式的探索业已成为教育政策研究者亟须开展的工作。蓬勃发展的"数据密集型科学发现"被界定为科学方法革命的"第四范式",表征科学探究的基本范式在当前大数据时代前所未有的变化。这一深刻变革不仅发生在自然科学与工程领域,也同样作用于以教育学为代表的广义社会科学领域。然而,这一新范式与科学研究中既有的因果推断逻辑似乎存在一定程度的互斥,如何将二者有机融合或可被视为中观及宏观维度教育政策研究亟待探索的"蓝海"。聚焦教育政策评估中既有的因果推断范式,系统梳理国外运用随机实验、准实验设计以及机器学习等方法在该领域所开展的前沿研究,继而探讨如何构建"大数据-因果推断"这一新型研究路径,为教育政策实施效果评估提供依据。

关键词:大数据;因果推断;教育政策;政策评估;科学研究范式

[中图分类号]G640　[文献标志码]A　[文章编号]1673-8012(2022)04-0039-10

一、大数据时代的新范式及其对教育政策研究的意义

长期以来,研究方法的创新型与适切性一直是中外教育研究者反复讨论的问题,对广义社会科学其他领域研究方法的借鉴也一直伴随着教育学的发展。在研究方法与数据来源不断更新迭代的当下,对新兴路径与范式的探索业已成为教育研究者(尤其是宏观及中观维度教育政策研究者)亟待开展的工作。数据被喻为"21 世纪的石油",已成为和土地、劳动力、资本、技术并列的五大生产要素之一。为区别于数据"采集—存储—分析—应用"这一传统路径,美国高德纳咨询公司(Gartner)副总裁兼分析师莱尼(Douglas Laney)于 2001 年提出了具有海量(volume)、多样(variety)、迅捷(velocity)三

修回日期:20220506
基金项目:教育部人文社会科学研究规划基金项目"基于 Trace-SRL 的本科生在线学习行为分析"(21YJA880014)
作者简介:郭娇,女,四川德阳人,华东师范大学高等教育研究所副研究员,教育学博士,主要从事高等教育政策评估研究;
吴寒天,男,浙江杭州人,浙江大学教育学院研究员,博士生导师,教育学博士,主要从事国际与比较高等教育研究。
引用格式:郭娇,吴寒天. 大数据时代的因果推断:教育政策评估的新路径[J]. 重庆高教研究,2022,10(4):3-48.
Citation format: GUO Jiao, WU Hantian. Causal inference in the age of big data: a new approach to education policy evaluation [J]. Chongqing higher education research,2022,10(4):3-48.

大特征的大数据概念[1]。不同于工业界以减少能耗、提高效率、扩大收益等为目标的内生创新动力,科学共同体的主要兴趣在于揭示事物表面特征下的本质规律,即利用新的数据来源与数据结构、运用新的数据分析方法,探究未知领域和应对新的伦理挑战。图灵奖得主、著名数据库科学家格雷(James Gray)于 2007 年将上述过程归纳为"数据密集型科学发现"(data-intensive scientific discovery)。格雷将这一新型研究范式视为科学方法革命历程中的"第四范式"(the fourth paradigm),认为其意义不亚于前 3 次科学方法革命,即分别以伽利略的实验科学、牛顿的模型推演和冯·诺伊曼的计算机仿真为代表的颠覆性研究范式革新[2]。2012 年,美国《纽约时报》发表《大数据时代》(The Age of Big Data)一文[3],在一定程度上标志着对数据的关注进入了更为广阔的公共视野。本文在使用广为人知的大数据时代一词时,将其内涵狭义化为格雷定义下的"数据密集型科学发现",关注科学探究在当前研究范式变革背景下的变化。

已有文献大多聚焦自然科学或工程技术领域对大数据运用的探索,国际科技数据委员会(CODATA)中国全国委员会于 2014 年编著出版的《大数据时代的科研活动》一书同样以自然科学为主要关注对象,涵盖物理学、天文学、生物学以及医学等领域,其中仅有一章是关于社会科学的,内容仅涉及经济、管理、金融等领域。不难发现,国内外对于广义社会科学领域大数据运用的研究相对滞后。因此,探讨上述科研范式革新对于社会科学领域显性或潜在的影响显得尤为必要,其中包括教育学(教育政策研究)领域的大数据应用。

需要注意的是,本文所讨论的应用并非指在微观层面预测学习者的努力程度、对某一个知识点的掌握情况或某一门课程的教学效果等,而是指在中观和宏观层面对某一学校(机构)所实施的制度、某一地区(学区)所开展的项目或某一个国家/经济体所制定的政策等(在本文中统称为教育政策)进行量化分析。运用大数据对微观维度教育活动进行预测的研究者多有计算机、实验心理学、脑科学等学科背景,业已形成了教育数据挖掘(educational data mining)和学习分析(learning analytics)两个全新的研究领域[4]。这两个教育研究的细分领域已经形成了较为成熟的学术共同体,拥有固定的学术年会与期刊,本文在此不作赘述。运用大数据进行中观或宏观维度教育政策评估的研究者多来自经济学、管理学等领域,人数相对少且较为分散,尚未形成紧密而成熟的共同体。考虑到教育政策的作用范围、资源配置的力度以及跟踪调查的难度,这一类研究/评估尤为需要大数据的助力。然而,这一大数据时代的新型范式似与教育政策实证研究中既有的因果推断逻辑存在一定程度的互斥。本文将系统阐述中观及宏观维度教育政策评估中的既有因果推断路径,继而分析这一研究路径在大数据时代的挑战与机遇。

本文重点关注西方发达国家运用随机实验、准实验设计以及机器学习等科学研究方法在相关领域开展的前沿研究,探讨如何在因果推断的基本逻辑之下运用大数据为教育政策的制定、实施以及评估提供依据。高等教育学这一细分研究领域里同样业已形成较为紧密而成熟的学术共同体,关注特定的研究对象、探究特定的研究问题、运用特定的研究方法、构建特定的理论模型、观测与分析特定的现象与规律,并形成特定的政策建议。在充分尊重上述特定情境的前提下,大数据时代的因果推断这一新型研究路径在高等教育领域的运用前景值得期待。

二、教育政策评估中的因果推断路径

因果推断(causal inference)的哲学基础最初由英国实证主义哲学家与经济学家穆勒(John Stuart Mill)于 1851 年在其所著的《逻辑体系》一书中提出。在该逻辑体系下,判定变量之间的因果关系需要满足 3 个条件,即时序性(假定的"因"要在"果"之前发生)、共变性(只要"因"改变,"果"即会随之变化)以及排他性(假定其他变量都不变,"果"仍然会随着"因"的改变而改变)[5]。若违反上述任一

条件,所得出的研究结论都不能称为因果推断,而只能称为相关关系,甚至是一种带有误导性的伪相关(spurious correlation)。“冰淇淋与鲨鱼攻击”的例子就是一种典型的伪相关,即并不是吃冰淇淋引来了鲨鱼,而是气候炎热时吃冰淇淋的人与下海游泳(继而遭遇鲨鱼攻击)的人都有所增多,因而两者的概率才会同时上升。

作为教育经济学研究基本预设之一的人力资本理论(human capital theory)也面临着类似挑战。以高等教育阶段大学生就业相关的政策研究为例,高学历与高收入之间的关联时常被质疑究竟是一种因果关系抑或仅仅只是一种相关关系。如要同时满足因果推断的三大基本条件,首先,教育经历要发生在工作之前,因此这一类基于人力资本理论探讨教育-收入关系的实证研究多采用大学生毕业之后的起薪而非数据采集时段的当前年薪评估个体教育投入的经济回报,以此排除在职培训、继续教育等干扰因素。其次,只要学历发生变化,收入就要随之变化。然而,收入变化的方向及大小在不同的实证研究中存在着分歧。求职者的个体偏好(例如倾向于从事公益机构工作)或雇主的薪酬结构(例如体制内就业的潜在福利优于现金收入)等都会抑制高学历带来的收入回报,甚至出现与“高学历-高收入”预设相悖的案例,即过度教育(overeducation)[6]。最后,3 个条件中最难验证的一条是排他性。个人动机、毅力等影响收入的潜在因素极为多样,无法如自然科学实验一般加以严格控制。如何引入自然科学中随机实验的设计思路来检验两个变量之间的因果关系,一直是广义社会科学领域量化研究者致力于回答的核心问题之一。

在梳理科学研究方法之前,尚有一个不容忽视的问题,即为何(教育)政策评估不能单纯依赖相关分析,而需要依靠逻辑更为严密的因果推断? 西方发达国家政策评估的兴起可追溯至 20 世纪 60 年代,当时的美国约翰逊政府推行一项名为“向贫穷开战”(War on Poverty)的社会改革。这一改革覆盖教育、医疗、社保等众多公共领域,耗时长且投入大,但取得的成效却明显低于预期。这一失败引发了 20 世纪 70 年代西方国家对于公共政策的一系列反思:公共财政经费与其他社会资源具有稀缺性,显然不能满足上述各领域的所有需求,而政府应如何在各种公共资源配置方案中做出合理选择,以及如何在政策干预结束之后评价其结果并向全社会公示。伴随着上述反思与追问,基于证据进行公共政策决策逐渐成为潮流。

20 世纪 90 年代,英国布莱尔政府声明要将“以证据为基础的公共政策”(evidence-based policy,EBP)奉为圭臬[7]。这里所说的“证据”指通过实证研究得出的科学发现,而这里所说的实证研究既包括量化分析,也包括质性研究,即广义的实证研究。就研究问题而言,既可以包括描述性问题(例如:发生了什么? 预期目标达到了吗? 谁获益,谁损失?),也可以包括干预性问题(例如:如果发生了 A,那么结果是 B 吗?)。相关分析尽管能就上述描述性问题提供描述性证据,但这显然不足以说服政府(或其他决策部门)投入本已十分稀缺的公共资源。干预性问题只能通过因果推断才能建立完整而严密的逻辑链条,帮助决策部门找准实现预期目标的着力点。

“基于证据的政策”率先出现在医疗、健康以及公共卫生领域,随后被应用于教育、扶贫等其他领域。最具说服力的证据来自医学临床实验中的随机控制实验(randomized control treatment,RCT),即病人被随机分成实验组(treatment group)和对照组(control group),分别服用含有有效成分的药物或安慰剂。由于两种药剂外观完全一致,病人、家属及其主治医生的主观反应都不会干扰对服药后治疗过程的观察,从而不会影响对药物效果的跟踪研究。这类随机对照实验同样被应用于教育政策评估,尤其是基础教育阶段。其中最具代表性的案例之一是对美国田纳西州 20 世纪 90 年代初小班化改革,即该州“师生成就比例计划”(Student-Teacher Achievement Ratio,STAR)的成效研究。该州11 600 名就读于学前班至小学三年级的学生被随机分配至小班(实验组)、传统班(对照组 A),或是增加了一名助教的传统班(对照组 B)。长期跟踪研究表明,小班化教学提高了实验对象参加 SAT 或 ACT 考

试(即美国高中毕业生学术能力水平考试)的比例和分数,而这一效果对于来自少数族裔家庭的学生更为显著[8]。

进入 21 世纪后,美国小布什政府在 2002 年推出的《不让一个孩子落后》法案(即 NCLB 法案,或译为《有教无类》法案)中明确要求,在出台教育政策施加干预前需得到“科学研究”(scientifically-based research)的支持,而这类研究需要满足两个条件:其一,教育行为或项目的相关信息必须通过严格、系统、客观的程序获取;其二,研究设计需采用随机实验或准实验方法,且在多种评估方法中优先承认随机实验所得出的结果[7]。显而易见,上述教育政策评估中的因果推断依赖随机实验或准实验方法实现,其目的在于提高政府决策过程的科学化,减少人为干扰,继而实现公共资源配置的优化。

值得注意的是,随机对照实验在教育政策评估中的应用有其局限性。一方面,出于科研伦理等因素考虑,随机实验的参与者通常需自愿报名,然而相当比例的学生和家长不愿意作为“小白鼠”参与实验。在科研伦理监管机制较为完备的国家,学生或家长在参与此类实验前享有知情权,同时还需要签署信息披露文件(否则研究者就不能采集、保存或使用其个人信息),并有权随时中止或退出实验。另一方面,就实施过程而言,即使上文所列举的经典案例也不能完全排除人为影响的干扰。与医学实验不同,学生、老师以及家长都清楚地知道自己究竟身处实验组抑或对照组,并会因此改变自身的行为,例如高等教育阶段就读于北京大学元培学院、清华大学“姚班”、浙江大学竺可桢学院等各种实验班的学生。被分到实验组(实验班)的师生通常会产生霍桑效应(Hawthorne Effect),即因为有机会参与实验而受到鼓舞,努力好好表现;被分到对照组(传统班)的师生通常会产生约翰-亨利效应(John Henry Effect),即因为无缘参与实验而加倍努力以证明自己[8]。

当标准随机实验难以实施时,研究者可采用准实验设计进行因果推断,具体方法包括断点回归(regression discontinuous)、工具变量(instrumental variables)、倾向得分(propensity score)和倍差(difference-in-difference)等,国内外文献均已对上述方法进行过详尽评述[5,9]。本文在此仅以工具变量为例,介绍采用准实验设计进行教育政策评估的思路。引入工具变量来判定两个变量之间因果关系这一策略,其核心在于先识别出原因变量中的随机成分,继而检验这一随机成分的改变是否带来结果变量的变化,基于上述逻辑建立的两阶段最小二乘(two stage least squares, TSLS)回归模型即可用来推断整体的因果关系。

好的工具变量具有外生性(exogeneity),这是准实验性质的集中体现。外生性不能通过实证检验,只能从逻辑上来论证,这是运用工具变量进行因果推断的成败关键。河流、山脉、地震等自然现象都是教育政策研究中常见的选择,因其难以被人为干扰。以基础教育阶段的学校布局为例,霍克斯比(Caroline Hoxby)于 2000 年使用美国不同学区内河流的分布情况作为工具变量,用以推断学校数量与教育质量之间的关系。这一研究即满足判定因果的 3 个条件:就时序性而言,河流存在于学校建立之前;就共变性而言,河流改变了学校数量以及各校之间的竞争关系;就排他性而言,河流本身显然不直接影响教学质量,而只能通过学校数量发挥间接作用[10]。生老病死同样具有自然规律的不可控性,在教育政策评估中可以巧妙地加以利用。例如,美国在发动越南战争期间采用基于生日的抽签形式来决定年轻男性是否需要服兵役,这就产生了带有随机性的工具变量。在这套机制中,每个出生日期对应一个从 1 到 365 的随机序列号。只有当该序列号小于美国国防部每年决定的一个特定取值时,这些男性才会被征召入伍。将抽签结果与社保局的薪资记录相结合进行分析,其结果表明 1970 年抽签入伍的白人男性在 1984 年的年薪相较于无须入伍的同龄人低 1 100 美元左右,这意味着在越战期间服兵役这一随机事件对收入水平产生了长期的负面影响[11]。此外,诸如空间距离、社会政策、集聚数据等都可用于构建类似的工具变量[5]。

总而言之,工具变量被喻为社会科学中因果推断的“圣杯”,这充分反映了其寻觅过程不仅需要灵感,而且充满艰辛[12]。大数据时代为这一探索以及其他基于随机实验或准实验设计的因果推断提供了更多的想象力,同时也带来了全新的挑战。

三、大数据时代因果推断路径的挑战与机遇

广义社会科学领域实证研究既有的因果推断路径在大数据时代显然面临着全新的挑战。就数据来源而言,除政府、国际组织、大学、研究机构等遵循传统路径收集的行政数据或调查数据外,以美国互联网三巨头(谷歌、脸书、亚马逊)和我国的 BAT(百度、阿里、腾讯)为代表的私人企业掌握了数量惊人的个人偏好与行为数据。高等教育阶段的产学研合作在大数据时代具有更大的想象空间与社会价值,集中体现为大学拥有的专家团队与企业拥有的海量数据“强强联手”,从而把传统科研项目推进到一个前所未有的层面。以反映通货膨胀的消费者物价指数(CPI)为例,美国麻省理工学院(MIT)的“十亿价格项目”(The Billion Prices Project, BPP)从 2008 年至 2016 年每天跟踪 60 个国家1 000多家网店的 1 500 万件商品及服务的价格,并与各国统计局公布的传统物价指数进行对比。BPP 的更新速度快于传统的物价指数,其估算过程中的跨国对比所涵盖的内容更丰富,估算值甚至比部分国家的官方指数更为可靠。例如 2008 至 2010 年,阿根廷官方公布的年均通胀率为 11%,而 BPP 估计的阿根廷年均通胀率则在 20% 以上[13],后者显然更接近一般民众的主观感受。就数据采集而言,新的采集方式更为全面,也更为隐蔽,甚至触及生日、银行卡账号等个人隐私及敏感信息,由此引发了一系列伦理追问(例如,某种类型的数据是否该被采集、保存或公开?数据应该被谁拥有?如果数据丢失或泄露,又应该由谁负责?)[14]。数据特征与数据伦理固然重要,但本文论述的重点在于数据分析,尤其是服务于教育政策领域因果推断的科学分析。如前所述,因果推断在教育政策评估中扮演着日益重要的角色,而大数据时代则重新形塑了其发展趋势。

就某种程度而言,大数据给因果推断带来了质疑与颠覆,即出现了本文开头所担心的专家思维与数据逻辑之间的互斥或割裂。牛津大学教授迈尔-舍恩伯格(Viktor Mayer-Schonberger)在其 2013 年出版的《大数据时代》一书中倡导 3 种思维,即要全体不要抽样,要效率不要绝对精确,要相关不要因果[15]。随着机器学习在图像识别、无人驾驶等领域不断取得突破,强调相关分析而非因果推断的趋势不断加强,似乎科学研究已经(或在不久的将来)不再需要人类专家厘清逻辑结构或找出关键变量加以干预。这种基于机器学习的相关分析又被称为关联分析(association)、以数据为中心(data-centric)的分析、不用建模(model-free or model-blind)的分析或“黑盒子”(black-box)分析。借用图灵奖得主、美国加州大学伯克利分校(University of California, Berkeley)教授珀尔(Judea Pearl)的比喻,这种分析的本质就像达尔文所描述的自然选择,并不能替代人类思维建立因果模型并打造精妙的工具[16]。2017 年,斯坦佛大学经济学教授阿西(Susan Athey)也在《科学》(Science)期刊上发文指出,用大数据进行的相关分析只是一种预测,并非决策,而只有了解这些行为背后的前提假设,才能基于这些数据来优化决策[17]。

如何在大数据时代通过因果推断来提供决策依据?珀尔提出 7 个要点:(1)与阿西的看法一致,他首先强调要让前提假设变得透明,且可检验。作为贝叶斯网络(Bayesian network)的奠基者,他建议采用图模型来让假设可视化,指出哪些假设可用数据检验,哪些只能从逻辑上论证。(2)混淆变量(confounding variables)需要加以控制。例如父母不仅影响子女的受教育程度,也影响子女的择业及其收入,这就是一个典型的混淆变量。珀尔在图模型里用“后门”(back-door)来解决这一问题,近似于在回归模型里加入父母的学历、职业、收入等控制变量。(3)用反事实推理(counterfactuals)来设计

算法。他指出,针对一个具体的研究对象,只能观察到一个结果(例如,考研的结果要么是“录取”要么是“落选”),因此需要借鉴已有的随机实验或准实验的思路,估算一组研究对象的均值。(4)通过中介效应分析(mediation analysis)来区分直接与间接影响,分析工具包括图模型与结构方程模型(structural equation model)等。(5)注意外在效度(external validity)与抽样偏差。珀尔指出,机器学习的研究者已经认识到了这一点的重要性,但仅凭相关分析无法保证结论的稳健性,即不受抽样影响而适用于不同人群。(6)缺失值(missing data)需处理。无论是研究对象退出实验或拒绝回答调查问题,都会造成数据缺失。他建议了解这些缺失值产生的原因,再有针对性地采取删除、插补等措施。(7)通过可以验证的假设,系统地构建一系列模型,最后再把这些因果推断整合成科学发现。如果以上 7 点无法实现,珀尔认为数据规模再大,分析过程再复杂,也不能得出因果结论,因为“数据本身并不是科学”[16]。

目前在广义社会科学领域运用机器学习来进行因果推断的研究较少,且集中在计量经济学领域。2014 年,美国加州大学伯克利分校教授兼谷歌首席经济学家瓦里安(Hal Varian)出版了《大数据:计量经济学的新窍门》(Big Data: New Tricks for Econometrics)一书。他在该书中指出,因果推断是机器学习与计量经济学最重要的合作领域,并举例用贝叶斯结构时间序列(Bayesian structural times series)来评估广告投放对网站访问量的影响。瓦里安同时指出,目前的机器学习还是以预测为主。显而易见,与传统的线性回归相比,机器学习的优势在于海量的数据与灵活的模型,更适于拟合非线性相关,可以通过正则(regulation)避免过度拟合,可以把数据分成训练集(training set)与测试集(test set)来进行交叉检验(cross-validation),还可以通过集成法(ensemble)提高预测准确度。机器学习的主要不足则在于上文提到的“黑盒子”分析路径,即没有假设检验、不提供回归系数及标准误,其中尤以集成法最难以解读[18]。

2017 年,美国芝加哥大学经济学教授穆莱纳森(Sendhil Mullainathan)及其团队共同撰写了《机器学习:一种计量经济学的应用方法》一书。基于从 2011 年美国住房调查中随机抽取的 10 000 套房屋信息,穆莱纳森及其团队用 150 个变量(包括其非线性以及交互作用)来预测房价,并事先抽出另外 41 808 套房屋作为测试集来进行检验。在比较最小二乘法、回归树(regression tree)、LASSO、随机森林(random forest)以及集成法这 5 种数据分析方法中,无论是训练集(即用于建模的 10 000 套房屋信息)、测试集,还是按五分位分组对比,随机森林与集成法的预测质量(用 R^2 衡量)都明显好于最小二乘法。除了这一具体案例,穆莱纳森还梳理了这一全新领域在过去 4 年的发展,包括应用于政策评估中的因果推断,以及对随机实验/准实验设计的改进[19]。以随机实验为例,传统方法只能比较实验组与对照组的实验效果均值(average treatment effect),无法反映个体差异。阿西团队用决策树、随机森林等机器学习方法来处理实验干预效果的异质性(heterogeneity)。2017 年,该团队在《美国经济评论》(American Economic Review)期刊发文,建议除了回归系数及标准误之外,还需补充 4 种检验结果,即比较不同的模型、调整变量的取值范围、采用不同的抽样方法以及基于半数的可重复取样交叉验证[20]。

另一项潜在的改进可能与上文提到的工具变量有关:究竟应该使用虚拟变量、连续变量、平方项、还是自然对数来识别因变量中的随机成分?这在本质上是一个预测问题,而这恰恰是机器学习所擅长的领域。美国麻省理工学院经济系与统计学系教授切诺祖科夫(Victor Chernozhukov)团队于 2018 年发文,提出采用双重机器学习方法(double machine learning)同时解决关于“因”与“果”的两个预测问题。这一路径适用于基于随机实验及工具变量等的因果推断,用于验证的 3 个政策评估案例,分别是美国宾夕法尼亚州失业保险金实验,401(K)养老金参与资格对个人财富净值的影响,以及以早期移民死亡率作为工具变量推断 64 个欧洲国家的个人产权制度对人均 GDP 的影响[21]。

在此基础上,阿西团队仍在继续改进基于实验数据或观察数据的因果推断。以针对美国加州提供就业培训项目 GAIN(the great avenues for independence)的随机实验为例,阿西团队通过控制性别、学历、参与实验之前的收入等 28 个变量,对 4 个县 19 170 人在参与实验之后 9 年里的平均收入进行分析。由于每个县选择实验组与对照组的方法与标准不同,导致该项实验的分组随机性受到质疑,这也是研究者在社会科学领域开展随机实验所共同面临的难题之一。阿西团队的贡献在于巧妙运用倾向性得分(propensity)构建了一个权重来优化实验组与对照组的分配——既非机械地平均分配(即各占 50%),也避免了各县随意地进行分配,同时还能达到实验整体效果的最大化(让能从实验中获益最多的群体尽可能多地参与进来)。基于深度学习的决策树,GAIN 的优化分配方案是让参与实验前 3 个月内有收入的群体占总参与者的四分之三,然后再根据学历(例如是否高中毕业)或家庭结构(例如是否有子女)来进一步细分[22]。由此可见,因果推断路径在机器学习的助力下能得到更为精准的实施——不仅分组更为客观,且实验效果也能惠及更多的潜在受益者。这种通过机器学习进行动态优化的实验设计或政策执行被阿西团队命名为自适应实验(adaptive experiment)或政策学习(policy learning)[23]。

整合上述最新研究进展,教育政策评估的"大数据-因果推断"新路径的基本架构设计如图 1 所示。大数据时代的因果推断范式具有数据密集性和场境依赖性两个本质特征,反映为图中的两大支柱(即基于数据的因果推断与嵌入真实的教育场境)。教育政策评估的路径创新体现在数据、技术以及应用 3 个层面。具体而言,在数据层面不仅打通了宏观社会经济结构、中观院校机构以及微观师生个体数据,而且通过增设的大数据中心及其数据采集、清洗、挖掘、可视化等功能助力决策咨询、专家分析、管理实施以及公众问责;在技术层面则在成熟的量化(如调查问卷)与质性研究工具(包括访谈、案例、课堂观察等)基础上加入了随机控制实验、准实验(如工具变量)、机器学习(包括决策树、随机森林等)、自适应实验或政策学习等最新技术手段;就应用层面而言,除动态监测、过程挖掘等功能创新之外,还可对教育政策评估的原有重要功能(如高等教育阶段的学情调查、学科评价等)进行升级或拓展。

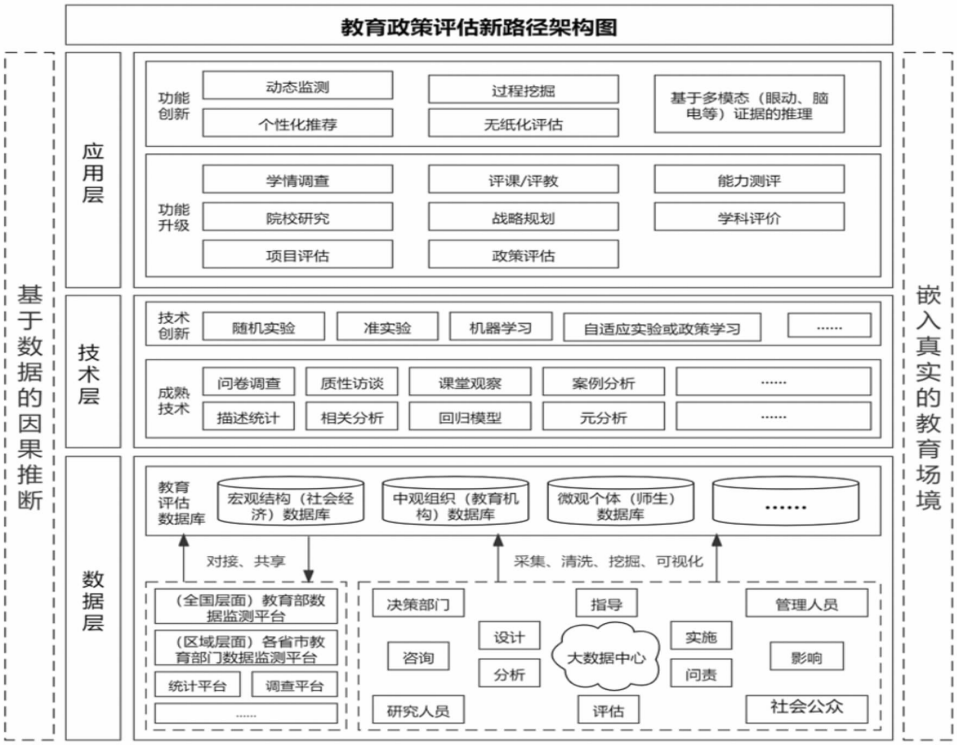

图 1　教育政策评估新路径架构图

四、“大数据-因果推断”范式:教育政策评估的新路径

科学研究范式的演进不仅仅作用于科学共同体对纯粹知识的探索,包括教育政策在内的公共政策及其制定与实施过程同样受到深刻而持续的影响。从古希腊城邦的陶片放逐法(Ostracism),到形成于19 世纪中后叶的“罗伯特议事规则”(Robert's Rules of Order),人类社会始终在不断追求决策过程的合理化与理性化。自 20 世纪 90 年代以来,基于科学证据的决策成为趋势,而因果推断则在教育等公共领域的政策评估中扮演着日益重要的角色。通过随机实验与准实验设计(运用工具变量、断点回归、倾向得分、倍差等)取得的研究结论相较于相关分析更具有说服力。显而易见,相关分析只能回答描述性问题,因果推断才能对政策干预的效果进行严谨的评估(例如,如果提高大学教师的工资,他们的教学质量会发生变化吗),甚至可以进行反事实的推理(例如,如果大学没有扩招,大学生就业会出现什么局面)。对问题的探究、对方法的改进以及对决策过程显性或隐性的影响,多重动因共同推动了广义社会科学领域的因果研究,产生了美国田纳西州 STAR 改革成果研究等教育政策评估领域的经典案例。珀尔把这种研究范式的快速发展称为因果革命(Causal Revolution)[16]。

进入大数据时代以来,因果推断遇到了全新的挑战,以提倡“要相关不要因果”的迈尔-舍恩伯格为代表,在过往居于领先地位的新锐研究范式似乎未经普及就已经“过时”。借助海量数据、灵活模型、快速调优(tuning)以及交叉检验,机器学习在预测上具有显著的优势。然而,其不足之处同样十分明显:无假设检验,不提供回归系数及标准误,尤其是采用随机森林或集成法的分析结果就像“黑盒子”一样难以解读[17]。因此,因果推断或许是广义社会科学与信息科学最应联手实现突破的领域。这类大跨度的跨学科科研合作也是当前各国政府资助的重点。2016 年,美国国家科学基金会(National Science Foundation, NSF)所列出的重点科研前沿即包括在大数据支持下“开发和评价创新型学习和教学机制方式”[14]。同年,瑞士国家科学基金会(Swiss National Science Foundation, SNSF)资助了“基于大数据的因果推断”(Causal Inference with Big Data)科研项目,该项目通过机器学习的方法来评估对失业工人进行就业培训的效果[24]。2018 年,中国自然科学基金增设了“教育信息科学与技术”这一申请代码,其资助领域包括“教育大数据分析与应用”,聚焦于教育学与信息学的深度合作与前沿探索。

目前,广义社会科学领域运用大数据进行因果推断的研究数量较少,且以计量经济学为主。如前所述,机器学习等分析方法可被用于改进随机实验以及工具变量等准实验设计。这些改进可以通过不同的模型、不同的变量取值范围、不同的抽样方式、不同的实验分组进行补充检验,也可以如同设立防火墙(firewall)一样把数据分成训练集与测试集进行交叉检验[23]。这些前沿研究尽管数量不多,但已陆续发表在《科学》(Science)、《美国经济评论》(The American Economic Review)和《计量经济学》(Econometrica)等权威学术期刊上,并在最近十余年中形成了“计算社会科学”(computational social science)这一交叉学科,其研究领域可被界定为“开发和应用计算方法分析复杂的、海量的(包括模拟的)人类行为数据”[25]。这一新现象值得我国相关领域的研究者加以关注。此外,2015 年,阿西在美国国家经济研究局(NBER)暑期培训班主讲《机器学习与因果推断》;2016 年,机器学习国际会议(International Conference on Machine Learning, ICML)开设了因果推断工作坊。就我国的教育政策研究者而言,实现大数据背景下因果推断的应用,不仅意味着完善自身业已熟悉的研究方法,还包括勇于迈出进行跨界探索的关键一步。

社会的现实需求、科研经费的支持、研究方法提升的路径以及学术发表的途径,这些因素共同描绘了因果推断在大数据时代的发展前景。目前,广义社会科学领域中基于机器学习的因果推断尚以经济学研究为主,近两年零星出现了若干教育领域的应用研究(例如智利全国与美国纽黑文全区的

中小学智能择校大数据平台[26]、美国976节小学英语课的视频逐字转录的海量文字记录的研究[27],以及以美国7个学区84所小学为实施单位的学生行为随机干预等[28]),但尚未出现高等教育学领域与政策研究相关的经典文献——这既是遗憾,更是极大的鞭策。面对这片社会科学研究的"蓝海",高等教育学的学术共同体对教育政策的评估如何在这场大数据时代的"因果革命"浪潮中不掉队,继而跻身世界知识生产体系的前列,或已成为迫在眉睫的课题。本文以这一核心问题作结,希望上述探讨仅仅是抛砖引玉,拉开序幕。因果推断的逻辑、想象及诠释,与机器学习的海量、灵活及效率,二者相互结合,必能为教育决策机构提供更具有说服力且更能满足异质性需求的科学依据。

参考文献:

[1] BEYER M,LANEY D. The Importance of "Big Data": A Definition[EB/OL]. (2012-06-21)[2022-05-11]. https://www.gartner.com/en/documents/2057415.

[2] TOLLE K M, TANSLEY S, HEY T. The fourth paradigm: data-intensive scientific ciscovery[J]. Proceedings of the IEEE, 2011,99(8):1334-1337.

[3] LOHR S. The Age of Big Data[EB/OL]. (2017-02-11)[2022-05-11]. http://faa.unm.edu/P200.040.FA17/Resources/Reflections/Steve%20Lohrfeb.pdf.

[4] SIEMENS G,BAKER R. Learning analytics and educational data mining: towards communication and collaboration [R]. Proceedings of the 2nd international conference on learning analytics and knowledge,2012:252-254.

[5] 黄斌,方超,汪栋. 教育研究中的因果关系推断:相关方法原理与实例应用[J]. 华东师范大学学报(教育科学版),2017,35(4):1-14,134.

[6] CAPSADA-MUNSECH Q. Overeducation: concept, theories and empirical evidence[EB/OL]. (2017-09-15)[2022-05-11]. https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12518 .

[7] 陈霜叶,孟浏今,张海燕. 大数据时代的教育政策证据:以证据为本理念对中国教育治理现代化与决策科学化的启示[J]. 全球教育展望,2014,43(2):121-128.

[8] KRUEGER A B,WHIRTMORE D M. The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from project star[J]. Economic journal,2001(468):1-28.

[9] 张羽. 教育政策定量评估方法中的因果推断模型以及混合方法的启示[J]. 清华大学教育研究,2013,34(3):29-40.

[10] HOXBY C. Does competition among public schools benefit students and taxpayers? [J]. American economic review, 2000(5):1209-1238.

[11] ANGRIST J. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records [J]. American economic review,1990(3):313-336.

[12] 陈云松. 逻辑、想象和诠释:工具变量在社会科学因果推断中的应用[J]. 社会学研究,2012,27(6):192-216,246.

[13] EINAV L,LEVIN J. Economics in the age of big data[EB/OL]. [2022-05-11]. https://web.stanford.edu/~leinav/pubs/Science2014.pdf.

[14] 刘三女牙,杨宗凯,李卿. 教育数据伦理:大数据时代教育的新挑战[J]. 教育研究,2017,38(4):15-20.

[15] 维克托·迈尔·舍恩伯格,周涛. 大数据时代生活、工作与思维的大变革[J]. 人力资源管理,2013(3):174.

[16] PEARL J. Theoretical impediments to machine learning with seven sparks from the causal revolution[EB/OL]. (2018-02-05)[2022-05-11]. http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf.

[17] ATHEY S. Beyond prediction: using big data for policy problems[J]. Science,2017(355):483-485.

[18] VARIAN H R. Big data: new tricks for econometrics[J]. Journal of economic perspectives,2014,28(2):3-28.

[19] MULLAINATHAN S, SPIESS J. Machine learning: an applied econometric approach [J]. Journal of economic perspectives,2017,31(2):87-106.

[20] ATHEY S, IMBENS G, PHAM T, et al. Estimating average treatment effects: supplementary analyses and remaining challenges[J]. American economic review,2017,107(5):278-281.

[21] CHERNOZHUKOV V. Double/debiased machine learning for treatment and structural parameters[J]. The econometrics journal,2018,21(1):1-68.

[22] ATHEY S, WAGER S. Policy learning with observational data[J]. Econometrica,2021,89(1):133-161.

[23] HADAD V, HIRSHBERG D, ZHAN R, et al. Confidence intervals for policy evaluation in adaptive experiments[EB/OL]. (2021-04-05)[2022-05-11]. https://doi.org/10.1073/pnas.2014602118.

[24] KNAUS M, LECHNER M, STRITTMATTER A. Heterogenous Employment Effects of Job Search Programmes: A Machine Learning Approach[EB/OL]. (2020-03-26)[2022-05-11]. https://arxiv.org/abs/1709.10279.

[25] LAZER D M, PENTLAND A, WATTS D J, et al. Computational social science: obstacles and opportunities-data sharing, research ethics and incentives must improve[EB/OL]. (2020-08-28)[2022-05-11]. https://doi.org/10.1126/science.aaz8170.

[26] ARTEAGA F, KAPOR A J, NEILSON C A, et al. Smart Matching Platforms and Heterogeneous Beliefs in Centralized School Choice [EB/OL]. (2021-06-12)[2022-05-11]. https://christopherneilson.github.io/work/documents/Warnings/AKNZ_June_2021.pdf.

[27] LIU J, COHEN J. Measuring teaching practices at scale: a novel application of text-as-data methods[J]. Educational evaluation and policy analysis,2021,43(4):587-614.

[28] SCHOCHET P Z. A Lasso-OLS Hybrid Approach to Covariate Selection and Average Treatment Effect Estimation for Clustered RCTs Using Design-Based Methods[EB/OL]. (2020-05-05)[2021-12-17]. https://arxiv.org/abs/2005.02502.

(编辑:杨慷慨 校对:张 腾)

Causal Inference in the Age of Big Data: A New Approach to Education Policy Evaluation

GUO Jiao[1], WU Hantian[2]

(1. *Institute of Higher Education*, *East China Normal University*, *Shanghai* 200062, *China*;
2. *College of Education*, *Zhejiang University*, *Hangzhou* 310058, *China*)

Abstract: With the continuous updating and iteration of research methods and data sources, the exploration of emerging paths and paradigms has become an urgent work for the educational policy researchers. The booming "data-intensive scientific discovery" is defined as the "fourth paradigm" of the scientific method revolution, which represents the unprecedented change of the basic paradigm of scientific inquiry in the current "big data era". The profound change takes place not only in the field of natural science and engineering, but also in the field of broad social sciences represented by pedagogy. However, there seems to be a certain degree of mutual exclusion between the new paradigm and the existing causal inference logic in scientific research. How to organically integrate and apply the two can be regarded as the "blue ocean" to be explored in the study of meso and macro educational policies. Focusing on the existing causal inference paradigm in education policy evaluation, the frontier research carried out in the field was systematically clarified by using random experiment, quasi experimental design and machine learning, and then how to build a new research path of "big data causal inference" was discussed, so as to provide a basis for the evaluation of the implementation effect of education policy.

Key words: big data; causal inference; education policy; policy evaluation; paradigm of scientific inquiry