

大数据驱动的教育研究新范式

王战军 乔 刚

(北京理工大学 研究生教育研究中心, 北京 100081)

一、引言

随着现代信息技术在经济社会领域的渗透,特别是“互联网+”、“大数据”、“人工智能”的发展上升为各国国家战略之后,新技术已经从支持性工具变成行业变革的重要驱动力量。信息技术正在与制造业、交通、医疗、金融、教育等传统行业深度融合,并催生出行业新形态、新模式。信息社会对教育的影响核心是,借助信息技术基础,使人与人、物与物、知识与知识空前广泛连接,并引发组织管理决策、资源开放共享、知识生产与传播的深刻变化,推动教育管理和决策的依据从经验驱动转向数据驱动。这主要是因为信息技术直接推动了教育领域的大数据生产,这些数据规模更大、速度更快、类型更杂、价值隐含更深;而大数据的深层含义在于海量数据的集聚提供了看待世界的新视角,蕴含了以往凭经验无法获取和发现的信息,如今可以通过现代信息技术和手段得以呈现。在信息社会时代,数据不仅仅是资源、力量和竞争力,更是生产力。数据显示,2015年全球大数据储量已达到8.61ZB,并且每年增长速度在40%左右。^[1]据Intel预测,到2020年,全球数据量将达到44ZB(1ZB=10亿TB=1万亿GB),中国产生的数据量将达到8ZB,占全球的五分之一。^[2]今后国家竞争力将部分体现为一国拥有数据的规模以及解释、运用数据的能力,数据主权将是继边防、海防、空防之后另一个大国博弈的空间。同时,数据作为一种特殊的资本,虽然没有排它性和消耗性,但是却能通过数据之间的整合实现1+1大于2的目的。数据竞争将成为今后商业、国家竞争的一种重要形式。有学者将数据比作“价值巨大的矿石”与“璞玉”,一方面说明了数据本身内涵的潜在价值,另一方面也间接反映了数据开发利用方式的重要性。当前,大数据的理念、思维和方法已经被广泛引入企业、政府、医疗、银行等领域,人们希望借助大数据的东风实现

作者简介:王战军,男,北京理工大学研究生教育研究中心主任,教授。

乔刚,男,北京理工大学博士后,延安大学教育科学学院教师。

基金项目:教育部哲学社会科学研究重大课题攻关项目(16JZD044)。

领域的改革和创新,带来发展的新动力和创新的突破口。正如美国著名管理学家 W. 爱德华·戴明(W. Edwards Deming)所指出的,今后除了上帝一切都要用数据说话。在信息社会,既要发现、认识、获取信息,也要在此基础上挖掘、分析、运用信息,二者缺一不可,前者是后者的前提和基础,后者则是产生实际效益的关键。

根据图灵奖获得者吉姆·格雷(Jim Gray)的划分,有史以来的科学研究可以分为四种范式:实验科学(Empirical Science)、理论科学(Theoretical Science)、计算科学(Computational Science)和数据密集型科学(eScience 或 Data-intensive Science)。^[3]其中,数据密集型科学以大数据为基础,通过对大量有关教育教学和科研数据进行收集、整理、分析和挖掘,以可视化的工具和方式呈现教育教学状态。信息社会的发展为科学研究提供“数据密集型研究”这一新的范式,也为教育教学管理和决策提供“数据”这一新的视角和出发点。与此同时,信息技术的发展推动了我国教育领域大数据的形成,推动着教育管理和决策从主观臆断转为依赖数据说话、靠数据决策的常态机制,助推教育研究范式从传统向大数据的转向。随着信息技术的发展、社会环境的变化,教育研究范式的转变也充分反映出在中国特色社会主义新时代进程中教育研究的重新定位,以适应新的社会现实。新旧范式的更新运动既是人类解释自然的逻辑知识体系的更换,又是科学共同体成员价值观念、思维模式、语言规则、实践标准的“量子跃迁”。^[4]

信息社会下的教育研究范式强调教育研究应从数据出发,发现教育状态的要素特征和结构关系。这一教育研究范式不仅仅是利用数据作为研究结论的证据,更重要的是通过数据发现教育状态、反映教育活动规律、引导教育研究行为更加科学化和理性化。随着信息技术与教育的融合不断深化,以数据及所产生知识在社会各主体间流动为基础的社会生态系统正在逐步形成,大数据驱动的教育研究、教育管理与决策等成为时代主题。可以说,某一研究范式的出现代表了某一学术共同体对本学术领域内关键问题的认定和发展方向的把握,并据此形成比较一致的话语规则和价值取向,成为学科内部认同的基础。^[5]信息社会的教育研究范式必将由局部性转向系统性,由线性化转向扁平化,由经验判断转向数据驱动,由静态转为动态。

二、动态范式及其基本特征

在此背景下,教育研究日益注重“用数据说话、用数据研究、用数据管理、用数据决策、用数据创新”。一种新的教育研究新范式逐渐形成,即动态范式(dynamic paradigm)。

动态是相对于静态而言的,指事物变化的状况。其含义是运用现代信息技术的思想、思维、理念和方法,对教育现象进行全样本、全过程、全景式研究,揭

示教育规律、动态呈现教育现状,持续提高教育教学质量,以适应社会发展和人类发展。这一范式以大数据驱动为核心,强调数据来源的广泛性,以分析、诊断、描述、研判等方式,通过数据之间的交互,发现数据的价值,进而直观呈现教育教学状态,预测教育教学发展趋势、预警教育教学质量。动态范式以现代信息技术保证研究的科学性,依靠数据处理的技术规则保证研究的规范性,通过深度的数据挖掘发现研究结论。因此,动态范式能够克服传统教育研究和决策的局部性、线性化和静态性以及依靠专家的经验 and 理性思考去判断教育状态的局限性,能够实现教育管理与决策中的有限理性,推动教育决策的科学化。在动态范式下,不仅数据的收集具有动态性、互动性,同时由于数据存储技术、数据处理技术的不断更新,处理数据的工具也呈现动态性的特征。动态范式的产生是新时代教育发展的必要结果,也是现代信息社会推动的时代产物。

动态范式作为教育研究的新范式,不是对传统教育研究范式的全盘否定,而是在继承和发展传统教育研究范式的基础上,适应社会经济和教育发展进入新阶段出现的一种新范式,它是现代信息技术在教育中的鲜明映射。全面理解动态范式,既需要明确其基本特征,也需要明晰其具体的行动框架。动态范式是信息社会的产物。较之于传统的教育研究范式,动态范式以大数据驱动为核心,具有全样本、交互式、全景式、全过程、全透明的特征。

1. 全样本

动态范式下的教育研究数据是全样本。传统的统计学观点认为,数据处理特点是通过局部样本进行统计推断,从而了解总体的规律性。^[6]但是,由于各种抽样调查工作是在事先设定目的前提下展开工作,不管多完美的抽样技术,抽到的只是总体中的一部分,样本都只是对总体片面和部分的反映。^[7]在我国教育领域,以业务管理数据库为代表的结构化数据源的海量积累加以图文影音形式记录的半结构化、非结构化数据及实时流数据源快速增长。教育管理涉及的大数据非常广泛,来源于教育管理系统、在线学习平台和课程管理平台等方面管理者、施教者、学习者的行为数据乃至日常教育活动中人类的所有行为数据,都可以称作教育大数据。^[8]目前,我国在国家、省级、校级三级层面都积累了大量的数据。国家层面有教育部基础教育质量监测中心数据平台、高等教育监测数据平台、教育部学位与研究生教育发展研究中心有关研究生教育的数据库等,各省也相继建立中小学教育数据库,中小学的教育信息化、数字化校园建设重点从基础建设、数据积累向数据的应用转移。高校内部的大规模开放在线课程、移动终端、社交媒体、即时通讯在教育管理与教学活动中的使用(如“雨课堂”)、校园一卡通等,都积累了大量的数据资源。随着国家加大信息公开的力度,除了内部业务数据和国家教育统计数据,直接相关的行业发展、人才需求、教育舆情等外部数据泛在分布,既有结构化的数据,又有半结构化、非结构化数据和实时流数据,这些数据通过公开渠道就可以获得。从国家到各学校积累的大量数据为我国教育研究提供了大数据来源,推动了我国教育研究从抽样走向

全样本,进而为实现教育研究走向更加科学化奠定了基础。

2. 交互式

动态范式具有灵活、直观的特点。它强调通过教育数据与研究者的交互,帮助研究者及时修正研究偏差,调整研究方法,并最终实现研究目标。随着现代信息技术的不断发展,数据之间的实时、动态交换已经变成可能。如汽车行驶中所采集的导航数据就是数据之间交互的结果。特别是虚拟现实(virtual reality)、空间隐喻(spatial metaphor)等交互式技术的快速发展,教育研究者可以将这些技术和手段以图形图像等形式呈现,实现数据与研究者的交互,从而直观呈现教育教学状态。交互式强调研究者利用现代信息技术,对研究所需的数据进行便捷与快速的用户分类、分析、维护和查找信息,进而帮助研究者进行数据分析和数据处理,并最终实现研究目标。动态范式的交互式虽然强调数据与研究者彼此之间信息的交流,但它是以人为中心的。以人为中心,教育研究者一方面应及时掌握现代信息技术的基本知识和基本技能,拓展自己的核心信息素养;另一方面应具备数据观念、数据思维,并不断增强数据意识。同时,研究者必须了解数据的特征、结构等各个方面,才能更好地实现数据与研究者的交互,从而为教育研究活动服务。

3. 全景式

在传统教育管理与决策活动、教育活动中,无论宏观、中观还是微观层面,由于普遍受限于条块分割的管理部门职能和权力,管理者往往从职能划分出发,决策过程“只见树木、不见森林”,缺乏系统性。由于管理体制和职能条块分隔,我国教育数据的收集和应用也呈现出严重的碎片化,将数据作为私有资源独占,“数据垄断”现象十分突出,“数据孤岛”普遍存在,教育系统内部数据源与外部数据源的整合不足,这也使得数据对于教育管理和决策过程的支持严重不足。在管理决策的旧范式中,直觉和经验是决策的主角,数据仍然是决策辅助或支持手段。大数据是更多、更快、更杂、更深的数据,能够为全景式管理决策提供全样本、全周期、全过程、全透明的支撑。动态范式的全景式在时间上贯通过去、现在与未来,在空间上覆盖宏观、中观和微观,具有全局性观察、全周期沉浸、全过程跟踪等特点。在这一模式下,教育管理和决策的主体、过程、依据等都将发生重要变化。当前,教务、科研、人事、学生、财务、后勤管理等海量数据记录着学校的管理活动和行为痕迹,国家及各省市有关教育发展的大量数据反映了我国教育的总体状况。动态研究范式通过运用大数据的技术和方法为教育研究提供全景式视角,从而为全面呈现我国教育发展战略、教育教学管理水平、教育治理能力和治理水平等提供支持。

4. 全过程

在动态范式研究下,研究者可以运用教育大数据,对研究对象从多角度、多层次进行全过程、系统化的分析和研究。研究者可以在受教育者个体从幼儿

园、小学、初中、高中到大学整个教育阶段,对其上课、课后练习、社会实践整个教育教学环节的活动进行全面监测,全面掌握受教育者个体的思想水平、政治觉悟、道德品质、文化素养、日常行为等,从而为学校教育实现全程育人、全方位育人构建全方位、全要素的教育教学体系提供数据支撑。研究者可以运用大数据分析的技术和方法,对各级政府的教育投资从投入到产出的整个过程实施全过程、动态化监控,为教育管理者有效掌握和调控我国教育效益提供数据支持。研究者还可以利用教育及我国社会经济发展的大数据,对全国或区域教育发展现状及趋势进行全方位的监测和预警,从而为教育管理和决策服务。

5. 全透明

动态范式下的数据来源可信、可溯,数据处理过程可复制,数据处理的结果实行公开化和实时共享。大数据驱动下的动态范式数据来源,不仅包括学校公开的信息、国民经济和社会发展统计数据、有关政府机构和社会组织的调查数据,而且包括联合国教科文卫组织发布的数据调查等。动态范式研究所需要的数据不再过度依靠传统的问卷调查等方法获取,研究者通过公开途径就可以获取。这一方面克服了传统研究过程中教育研究者为获取大量研究数据而带来的诸多不易,另一方面也可以保证了研究数据来源的真实性和可靠性。动态范式下的教育研究活动不仅是数据来源的透明,而且包含研究方法和研究结果的透明,即动态范式采用在线、近线以及离线技术,通过对收集到的数据进行分布式、非结构化以及实时数据处理等,进而深度挖掘数据之间的相关性。同时,动态范式通过运用各种简洁的图表等方法,一目了然地把数据之间的关联特征和相互关系直观明了地呈现出来,方式简单直接,为教育管理者了解教育状态、走向及变化趋势提供了一种形象、直接的解读和展现数据方式,有助于透彻地理解数据所包含的价值。这些可视化的数据呈现方式具有直观化、关联化、艺术化和交互性等特点,有着非凡的艺术吸引力和强烈的色彩感染力与穿透力,能够强化多元主体的认知效果和记忆力,更有利于信息的传递。^[9]

三、动态范式的行动框架

动态范式不仅仅是一种理论层面的建构,需要明晰其具体的行动框架,才能使之在教育研究中真正发挥作用。为此,本文构建了动态范式的基本行动框架图(见图1)。它通过对数据的收集、清洗和挖掘,客观呈现我国教育教学状态,为提升我国教育管理和决策水平服务,为引导我国教育政策的制定和教育事业的发展服务。

数据采集是动态范式的基础。动态范式中的数据收集主要解决研究中所需数据的来源问题。动态范式下采集的数据不是传统统计学意义上的小样本数据,而是大规模、全覆盖、深层次的全样本数据;在采集方式上,动态范式强调

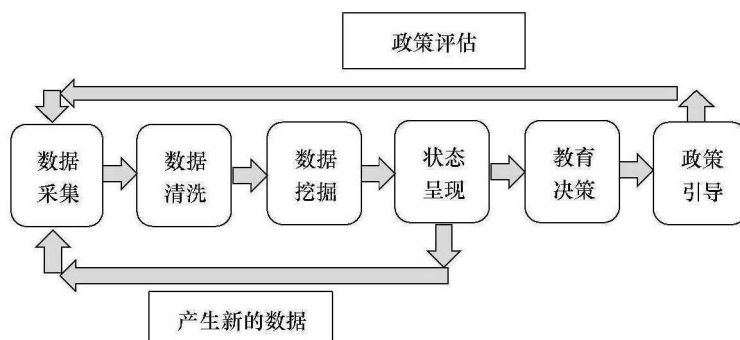


图1 动态范式的基本行动框架

采用数据爬虫等技术,在线高频次甚至是实时进行数据收集;在覆盖面上,既包括全国经济、人口、财政收入等有关社会经济发展水平的公开数据,也包括各级各类教育的教学管理数据、科学研究数据、满意度调查数据、社交网络数据等,还包含源于教育活动和网络媒体中的交互数据和日志数据等半结构化、非结构化和实时流数据;在深度上,既包括不同国家、地区的宏观数据,更包括各级各类学校内部的课程、教师、学生、学科建设等微观数据;在时效性上,通过每季度或每月、每周、每日甚至实时动态采集,保证数据的连贯性和完整性。首先将对采集到的数据进行抽象,形成教育大数据的元数据(metadata);在此基础上,针对基础数据进行横向处理,对分析数据进行纵向处理,进而进行数据的关联分析与数据融合;最终,形成“抽取—转换—装载”的数据生产线,为中国教育数据仓库提供持续的动态化数据。

运用大数据方法采集到的数据不是系统的数据集合,而是单个、分散的数据,具有庞杂性、割裂性、多源异构等特征,单独分散的数据块无法满足大数据分析的需要,必须进行关联融合才能产生价值。要发挥这些数据的功效,需要对这些大量未相互连接的、碎片化的数据进行清洗,以剔除无效与杂乱的数据,并把有效数据进行归纳、分类,通过数据组合、数据整合和数据聚合等三个层次的数据关联融合,由低到高,逐步实现数据之间的深度交互。数据挖掘的目标是在数据仓库中发现模型、趋势和规则,以评估(预测或评价)打算采取的战略。^[10]在动态范式中,研究者运用联机分析技术、关联分析技术、聚类分析技术、分类分析技术等多种分析技术,对有关数据的关联度进行多维度、多层次的分析,以直观呈现教育教学状态,为教育管理者持续改进教育教学质量,提高科学管理和精准决策提供有力支持。动态范式下状态呈现和政策引导等环节中还会持续产生新的数据,需要进行再收集、分析使新的数据为教育政策引导服务,动态范式是一个循环往复、螺旋上升的过程。因此动态范式下,在决策信息输入上,应将“小数据”变大,从全局而不是局部上考察决策问题;同时,在决策信息输出上,将“大数据”变小,从快速生成的混杂数据中通过敏捷分析挖掘出具

有决策参考价值的知识,使决策者不是靠主观猜测和经验推理而是依据深度数据分析的结果直接了解事实,进而扩大决策者的有限理性,实现即席决策和精准决策,推动现代教育管理的变革与创新。

参考文献

- [1] 2017 年全球大数据产业规模与竞争格局预测[EB/OL]. <https://www.qianzhan.com/analyst/detail/220/170411-c19f6837.html>.
- [2] 英特尔预测 2020 年全球数据量达 44ZB;中国占据五分之一[EB/OL]. http://www.sohu.com/a/192946472_114760.
- [3] Gray, J. (2009, October 16). Jim Gray on eScience : A transformed scientific method. Retrieved March 30, 2015, from http://research.microsoft.com/en-us/collaboration/Fourth_paradigm/4th_paradigm_book_jim_gray_transcript.pdf.
- [4] 张武升,廖敏.教育研究范式的变革与发展趋向[J].华中师范大学学报(人文社会科学版),2005(5):128—133.
- [5] 王洪才.论高教研究的四种范式[J].北京师范大学学报(人文社会科学版),2002(3):74—82.
- [6] 韦博成.漫谈统计学的应用与发展(1)[J].数理统计与管理,2011(1):85—97.
- [7] 朱建平,章贵军,刘晓薇.大数据时代下数据分析理念的辨析[J].统计研究,2014(2):10—19.
- [8] 周湘林.大数据时代的教育管理变革[J].中国教育学刊,2014(10):25—30.
- [9] 王战军,乔刚.以新发展理念引领高等教育质量监测平台建设[J].中国高教研究,2016(7):25—30.
- [10] [美]路易斯·戈麦斯-梅西亚,戴维·鲍尔金,罗伯特·卡迪.管理学——原理、案例和实践(第3版)[M].詹正茂主译.北京:人民邮电出版社,2009:14.

(责任编辑 宋朋洋)