DOI: 10.16538/j.cnki.fem.20211101.101

因果关系评估的准实验设计与实证经济学的可信性革命

——2021年度诺贝尔经济学奖得主主要经济理论贡献述评

李宝良[1]，郭其友[2]

（1. 华侨大学 经济与金融学院，福建 泉州 362021；2. 厦门大学 经济学院，福建 厦门 361005）

摘　要：如何使用观察数据对变量之间的因果关系进行可信评估是实证经济学要面对的一个普遍问题。瑞典皇家科学院将2021年度诺贝尔经济学奖授予大卫•卡德（David Card），以表彰其对劳动经济学实证研究的贡献，以及约书亚•安格里斯特（Joshua D Angrist）和吉多•因本斯（Guido W Imbens），以表彰其对因果关系分析的方法性贡献。他们借助自然实验对因果关系评估进行准实验设计，显著提高了从观察数据中进行因果关系评估的可信性。对准实验设计方法和实证研究的回顾表明，各种准实验设计都有相同的基本逻辑，即在社会经济中找到一种场景也就是“自然实验”来模仿随机控制实验，评估反事实结果从而推断经济变量之间的因果关系；准实验设计的质量是决定实证研究质量和结论可信性的关键，高质量的准实验设计来自对自然实验背后相关制度运作过程的深刻理解和透彻分析。国内经济学者在借助准实验设计以科学规范的经济学研究方法讲好中国故事时，要注重对自然实验背后的相关制度进行深度的调查分析，规范准实验设计过程，避免P值操纵，并要加强影响机制的理论研究；政策制定者在利用准实验研究结论时，必须关注准实验设计研究结论的局限性和外部有效性。

关键词：大卫•卡德；约书亚•安格里斯特；吉多•因本斯；因果推断；准实验设计；实证经济学

中图分类号：F270　**文献标识码：**A　**文章编号：**1001-4950(2021)11-0140-13

一、引　言

因果关系评估是一切科学研究的核心问题。随机控制实验（randomized controlled trials，RCTs）是因果关系评估最有效的工具，被誉为因果推断的黄金标准。社会科学的许多领域也采用这一方式对因果关系进行评估。在经济学研究领域中，弗农•史密斯对实验经济学的研究和阿比吉特•班纳吉、埃丝特•迪弗洛和迈克尔•克雷默对减轻全球贫困问题所进行的实地实验研

收稿日期：2021-10-31

作者简介：李宝良（1980—），男，华侨大学经济与金融学院副教授，经济学博士；
郭其友（1963—），男，厦门大学经济学院教授，经济学博士（通讯作者，qiyouguo@163.com）。

究最具代表性(李宝良和郭其友,2019)。但是,随机控制实验有其局限。因其耗时、花费不菲,以及伦理道德的限制等因素,该方法并不总是行得通的。例如,经济学家难以甚至无法通过随机控制实验来评估最低工资、移民及教育政策对就业和收入等的影响。对于这些事关劳动者收入以及收入不平等的问题,能够用于因果关系评估的数据通常只有观察数据。然而,观察数据不同于实验数据,其中最大的问题是观察数据中的"内生性"问题,观察数据是人们在受约束条件下最优化的结果,这导致观察数据受各种可观测或不可观测因素的影响,使得研究者难以像使用实验数据一样,在保持其他影响因素不变的情况下来评估因果关系。

早期的实证经济学(empirical economics)[①]研究因缺乏解决内生性问题的有效办法,导致实证研究的结论缺乏可信性。以最低工资政策为例,即使如教科书所预测的那样,最低工资与失业率之间有正相关关系,人们却无法确认是因最低工资的提高而导致了失业率增加,还是因政策制定者为了回应失业率增加下低收入者的呼声而提高最低工资。因此,要使用观察数据对经济变量之间的因果关系进行可信的评估,就要找到解决观察数据中的内生性问题的办法。借助自然实验进行准实验设计,成为解决观察数据中的内生性问题的切入口。准实验设计基于这样一种思想:自然、政策以及制度等的变化有时候会提供一种场景,就像随机控制实验一样,它会将研究对象进行随机分组而排除内生性的影响。这些场景是在社会经济中自然发生的,因而被称为自然实验;它们又类似于随机控制实验,也被称为准实验。通过自然实验的选择,能够较好地排除观察数据中的内生性问题,从而很大程度上提高了因果关系评估的可信性。

本年度三位获奖者是借助自然实验对因果关系评估进行准实验设计的先锋。从20世纪90年代以来,卡德借助自然实验对最低工资、移民和教育政策问题中的因果关系进行准实验设计研究,他的许多实证结论颠覆了传统观念,并且引发了经济理论的发展和新一轮的实证研究,使人们对劳动力市场的运作有了更深刻的了解;安格里斯特和因本斯将经济学中的工具变量框架和统计学中的潜在结果框架结合起来,考虑了自然实验中普遍存在的异质性和不完全依从问题,在满足一组最小假设的条件下提出了局部平均处理效应的估计方法,澄清在自然实验中进行可信的因果关系评估所需要满足的关键假设,使准实验设计更加透明和可信。

以卡德、安格里斯特和因本斯为代表的因果关系评估的准实验设计方法,与以班纳吉、迪弗洛和克雷默为代表的实地实验方法一起奠定了实证经济学(empirical economics)的方法论基础,并在过去三十多年里掀起了实证经济学可信性革命的浪潮。诺贝尔经济科学奖委员会主席彼得•弗雷德里克森称:"卡德对社会核心问题的研究以及安格里斯特和因本斯的方法论贡献表明,自然实验是丰富的知识来源。他们的研究大大提高了我们回答关键因果问题的能力,这对社会大有裨益。"

本文对他们如何借助自然实验对因果关系评估进行准实验设计进行梳理和述评,主要结论如下:第一,因果关系评估的各种准实验设计方法万变不离其宗,有着相同的基本逻辑,都是借助自然实验的场景模仿随机控制实验的方式,寻找具有相同或者相似特征的研究对象作为对照组,对照组构成了因果关系评估的反事实结果。第二,对实证经济学的经典文献的研究发现,因果关系评估的可信性,它不是依赖于数据和计量方法本身,而是依赖于社会经济体制运行机制的深入理解和透彻分析,因而对自然实验的深入调查和探讨是必不可少的。事实上,实证经济学家不太可能仅依赖计量经济学方法对实证结论的因果关系进行解读。高质量的实证经济学研究要根据问题的背景,包括制度特征的深入分析以及数据的可得性选择合适的方法。

①这里实证经济学指的是empirical economics而非positive economics。实证经济学(positive economics)是相对于规范经济学(normative economics)而言的,empirical的含义是源于实验和观察而非理论,因此,empirical economics是相对于理论经济学(theoretical economics)而言的,更准确的翻译是经验经济学,之所以翻译为实证经济学,是因为考虑到国内的翻译惯例。

第三，中国的经济改革为准实验设计提供了可资利用的丰富的自然实验资源，研究者以科学规范的经济学研究方法讲好中国故事时，必须注重对自然实验背后的相关制度进行深度的调查分析，规范准实验设计过程，避免P值操纵，并且要加强影响机制的理论研究；而政策制定者在借鉴参考准实验设计评估政策效果时，要关注准实验设计研究结论的局限性和外部有效性。

二、学术生涯与主要论著概述

卡德1956年出生于加拿大圭尔夫（Guelph），拥有加拿大和美国双重国籍。卡德1978年毕业于加拿大女王大学（安大略省金斯敦），获文学学士，1983年在普林斯顿大学获博士学位，后先任教于芝加哥大学、哈佛大学等，现任加州大学伯克利分校经济学教授、劳动经济学研究中心主任、计量经济学实验室主任、国家经济研究局劳动研究计划主任。他曾担任《美国经济评论》《计量经济学》《劳动经济学杂志》等期刊的联合编辑或副主编，并担任加拿大统计局、罗素圣哲基金会等的咨询顾问。卡德因对劳动经济学研究的杰出贡献而获得了许多荣誉，其中最重要的或许是获得1995年约翰•贝茨•克拉克奖。此外，他还获得伊扎劳动经济学奖（2006）、计量经济学会授予的弗里希勋章（2007）、BBVA前沿知识奖（2015）。

卡德师从奥利•阿申菲尔特（Orley Ashenfelter）而走上劳动经济学研究之路。阿申菲尔特是著名的劳动经济学家，是他第一个将双重差分法引入经济学的实证研究之中[①]，也是他指出了随机实验是实证研究的可行之路（Ashenfelter，1978）。在阿申菲尔特的引导和启发下，卡德的研究兴趣包括最低工资、移民、教育政策以及劳动力市场中与性别、种族相关的收入差异，研究议题涉及税率和劳动力供应、最低工资、粘性工资、罢工模式、工会和不平等、基于技能的技术变革、教育回报、移民等社会热点问题。卡德发表了上百篇的期刊论文和书籍章节。与艾伦•克鲁格合著的《神话与衡量：最低工资的新经济学》（1995）一书，集中体现了他对最低工资实证研究的成果。与艾伦•克鲁格合著的《工资、学校质量和就业需求》（2011）一书，是他对教育问题的实证研究成果。与史蒂文•拉斐尔合编的《移民、贫困和社会经济不平等》（2013），集中收录他对移民问题的研究。

安格里斯特1960年出生于美国俄亥俄州哥伦布，拥有美国和以色列双重国籍。他在1982年获奥伯林学院经济学学士，1987年和1989年在普林斯顿大学获经济学硕士和博士学位，先后任教于哈佛大学、希伯来大学、哥伦比亚大学等，现任麻省理工学院福特经济学教授、麻省理工学院蓝图实验室主任、国家经济研究局副研究员。他是美国艺术与科学院院士，美国经济协会、美国统计协会、计量经济学会等的会员，担任《劳动经济学杂志》的联合编辑。他因对计量经济学方法和实证研究的贡献而受到广泛赞誉，1999年获《经济学季刊》授予的格里利克斯（Griliches）荣誉奖，2011年被布达佩斯拉吉克•拉兹洛高等研究学院授予约翰•冯•诺依曼奖。

国内计量经济学界称安格里斯特为“安神”。“安神”的神奇人生与他不安分的性格分不开。他成长于犹太家庭，贪玩的秉性让其父母甚为担心他的人生。好在他在奥伯林学院时写了篇优秀论文，得到奥利•阿申菲尔特的赏识，后者有意收他为博士生。然而，安格里斯特却赴以色列当一名伞兵。在见识了战争的残酷之后，安格里斯特写信给阿申菲尔特想回到他名下攻读博士，由此重归学术之路。他的研究兴趣包括教育经济学和学校改革、社会计划和劳动力市场、移民、劳动力市场监管和制度的影响、项目和政策评估的计量方法。他对准实验设计方法的推广体现在与约恩-斯特芬•皮施克合著的《基本无害的计量经济学：实证研究者指南》《精通计量：从原因到结果的探寻之旅》两本教科书中，书中也汇集了他对劳动经济学的实证研究成果。

因本斯1963年出生于荷兰，拥有荷兰和美国双重国籍。因本斯1983年通过荷兰鹿特丹伊拉

①双重差分法最早可以追溯到约翰•斯诺（John Snow）对霍乱传播机制的研究（1855）。

斯姆斯大学的计量经济学候选资格考试(相当于学士学位),1986年以优异的成绩毕业于英国赫尔大学,获得经济学和计量经济学硕士学位,并于1989年和1991年在布朗大学获经济学硕士和博士学位,先后执教于哈佛大学、加州大学洛杉矶分校和伯克利分校等。2012年,他转职斯坦福大学商学院,现任应用计量经济学教授和经济学教授、经济政策研究所高级研究员。他是美国艺术与科学学院院士、经济计量学会会员、斯坦福大学商学院信托教师委员会成员、瑞士圣加仑大学名誉博士、美国国家经济研究局研究员。自1993年起,他担任多届计量经济学学会大会的组织委员,曾任《商业与经济统计期刊》《计量经济学》《计量经济学期刊》等的副主编,曾获得阿尔弗雷德•P•斯隆研究奖学金(1995—1998)、因将贝叶斯分析用于"在随机激励设计中评估流感疫苗的效果"研究而获得2001年度的米切尔奖。

因本斯对计量经济学理论研究情有独钟。自1982年起,他一边从事计量经济学的学习和研究,一边在大学兼任研究助理或讲师。他早期的研究兴趣是基于选择的抽样问题的推断和优化研究。20世纪90年代中期后,他与安格里斯特、鲁宾合作,研究重点和研究兴趣逐步转向准实验设计的因果关系评估方法,包括匹配方法、回归断点估计以及双重差分法的改进。因本斯的代表作是与反事实框架的提出者唐纳德•鲁宾合著出版《统计学、社会学和生物医学的因果推断导论》(2015),该书总结了因果推断领域的最新理论进展,被誉为因果推断理论应用研究领域最经典的教科书。

三、因果关系评估的准实验设计

准实验设计在本质上是对随机控制实验的模仿,要把握因果关系评估的准实验设计,首先要理解随机控制实验的精髓。接下来,我们将以随机控制实验作为比较的基准,就安格里斯特和因本斯对工具变量法所需满足的前提假设(Imbens和Angrist,1994),或者更一般地说,使用自然实验进行因果关系评估的准实验设计时需要做出的假设进行探讨;同时回顾安格里斯特及其合作者运用准实验设计方法对教育回报率(Angrist和Krueger,1991)以及教育投入与学习表现(Angrist和Lavy,1999)的实证研究,展示从自然实验中能够进行哪些因果推断,或者不能进行哪些因果推断。

(一)作为基准的随机控制实验

因果关系评估是要评估某个处理对个体的影响,它被称为处理效应(treatment effect)估计。要准确地估计处理对个体的影响,就需要知道受处理的个体,假如它没有受到处理会怎么样?这就是所谓的反事实结果(counterfactual outcomes)。然而,因果关系评估的难点在于,无法让同一批病人在服药的同时又不让他们服药。如果仅是比较一批服药的和另外一批不服药的病人,那么两批病人结果的差异可能不是药品的效果,而是两批病人之间的差异导致的。简言之,要准确地评估处理效应,需要在除了是否接受处理的不同之外,还要保持其他各个方面的条件不变,才能准确地估计反事实结果,进而准确地进行因果关系评估。

随机控制实验就是一种评估反事实结果的有效手段。仍以药品有效性检验为例,研究人员通常采用精心设计的双盲实验评估药品的有效性。具体是,首先将病人随机分组,随机分组确保了这两组人在性别构成、平均年龄、嗜好的比例等各个方面具有相同或者相似的特征;然后随机选择一组作为处理组施以服药的处理,另一组作为对照组则服用安慰剂。在这个过程中,参与实验的病人和医生不知道其服用的是药品还是安慰剂。之所以这样做,是因为常识和生理学理论指出,药物的效果会受到人体自身有免疫力和自我恢复的能力及心理暗示的影响。

随机控制实验中的处理组和作为反事实结果的对照组之间均值的差异,提供了药品效果的可信估计,这个处理效应也称为平均处理效应(average treatment effect,ATE)。平均处理效

应等于处理组的平均处理效应(average treatment effect for the treated,ATT)和未处理组(对照组)的平均处理效应(average treatment effect for the untreated,ATU)的加权平均,其权重是处理组和对照组的比重。由于异质性的存在,ATT通常不等于ATU,因而通常也不等于ATE。一般而言,在随机控制实验中,可以估计ATE,但是在自然实验中,能够估计的只是ATT。

虽然随机控制实验的设计和实施比较麻烦,但是实验数据的分析则相对简单。在上述药品有效性检验的例子中,可以通过两组之间的均值的*t*检验进行判断。但是在现实中,许多因果关系评估是难以通过随机控制实验来检验的。研究人员能借助的就是观察数据,而这些观察数据通常是人们选择的结果,这就存在内生性问题,计量经济学就是为了解决观察数据中的内生性问题发展而来的。这些方法在计量经济学教材中有详细的介绍(安格里斯特和皮施克,2012,2019)。接下来的重点是,以随机控制实验作为基准,归纳自然实验在模仿随机控制实验进行因果关系评估中所面临的挑战,阐述各种准实验设计方法,它们如何在保持其他条件不变的情况下来估计反事实结果,从而进行因果关系评估。这是精髓所在。

(二)自然实验与准实验设计

在经济社会发展过程中,一些自然发生的事件会导致类似于随机控制实验的场景或自然实验,这些事件不在人们的掌控之中,会将人们随机分成不同的群体、接受不同的处理。准实验设计的优势在于节约了设计和实施随机控制实验所需的时间和资源。但是,自然实验毕竟不是随机控制实验,利用自然实验对因果推断进行准实验设计也给研究者提出了新的挑战。一个挑战是异质性(heterogenity)问题,即处理组和对照组对同样处理的反应可能是不同的,即ATT可能并不等于ATU;另一个挑战是不完全依从(compliance)问题,即当处理效应因人而异时,人们在进行选择时很有可能不会完全依从自然实验。这两个挑战在随机控制实验中也同样存在。例如,实验对象可能退出实验,这时研究人员通常采用最终完成实验的处理组和对照组去估计,被称为意图处理效应(intention to treatment,ITT)。很明显,如果不遵从实验而退出的人具有某种系统性特征的话,那么ITT很可能不同于ATE。

在随机控制实验中,可以通过对实验的过程进行严格把关来减少两者的偏差。但是,不完全依从问题在自然实验中普遍存在,其与异质性的结合给因果推断带来了新的问题。早期的研究人员试图对研究对象的行为施加严格的假设,以此来对总体或处理组的平均处理效应进行因果推断,但是,这些假设往往过于严格而缺乏实际用途。要解决内生性问题,就得找到一个工具变量,对是否接受处理进行随机分组。因本斯和安格里斯特试图探讨的问题是,如何在不对无法观察的研究对象行为施加额外的假设下,能够从自然实验中进行哪些因果推断。

1.工具变量法

以安格里斯特和克鲁格对教育回报率的估计的里程碑式实证研究(Angrist和Krueger,1991)为例,他们选择出生日期作为工具变量。这个工具变量来自于出生日期的随机性以及美国的教育法规产生的自然实验。在美国,早出生的学生将比其他晚出生的学生更早到达从高中辍学的合法年龄。假设有两个孩子的出生日间隔一天,即一个出生于12月31日,另一个在1月1日。由于开学的日期是统一的,譬如都在9月1日,这样,出生于12月的孩子在开学时的实际年龄是5岁8个月大,而出生于1月份的孩子是6岁8个月大[①]。如果法律规定年满18周岁可以辍学,那么,只有1月份出生的孩子才能在高中毕业前辍学。可见,出生日期的随机变化导致教育程度的变化,这是由法律法规引起的,与学校教育的其他决定因素无关。因而,这样就可以采用是否将入学截止日期前出生的哑变量作为工具变量。

①这个问题类似于国内幼儿园报名的出生日期以8月31日为界限,不同之处是美国在某一年出生的人在同一年的9月1日入学,因而12月31日出生的人一到年龄就可以上学,1月1日的人只能下一年才能入学。

按照出生日期与是否完成高中学业可以将总体分成4个子样本，不同子样本的行为是异质性影响的反应：第一个子样本是不管出生日期为何都会完成高中学业的，这一类人也称为始终接受者（always-taker）；第二个是不管出生日期为何都会辍学的，这一类人也称为从不接受者（never-taker）；第三个是出生日期在入学截止日期前就完成高中学业、在入学截止日期后就辍学的，这一类人被称为依从者（compliers）；第四个是出生日期在截止日期前就辍学、在截止日期后就完成高中学业的，这一类人被称为排斥者（defiers）。由于始终接受者和从不接受者的行为不会因工具变量而发生变化，因而无法用来进行因果推断；能够借助工具变量用于因果推断的实际上是第三个子样本，也就是依从者。但是，这会受到排斥者的干扰。

为了能够使用工具变量法进行因果推断，因本斯和安格里斯特（Imbens和Angrist，1994）认为，一个有效的工具变量至少应该满足如下的假设：（1）随机分配假设：工具变量应该和随机分配一样好，即工具变量应与所有潜在结果不相关。这个假设与随机控制实验一样，使得处理组和对照组具有相似的特征。（2）相关性假设：工具变量与被处理与否相关，它可以通过是否中签与是否服兵役两者之间的相关性检验来判断。（3）排他性假设：工具变量影响潜在结果的唯一途径是通过处理。从理论上来讲，这个假设是不可检验的，只能通过经济学理论或者常识进行论证。例如，抽签是根据出生日期随机抽取的，出生日期影响未来收入是通过是否服兵役可能是一个合理的假设。假设（1）和（3）意味着工具变量是外生的，也即外生性假设由随机性和排他性两个假设组成，但这是两个不同的假设。此外，为了排除排斥者的干扰，他们还提出了第四个辅助的假设，即（4）单调性假设：所有的个体都以同一个方向受影响，要么不受影响，这个假设意味着排斥者很少或者不存在。

在满足以上假设的基础上，借助两阶段最小二乘法，因本斯和安格里斯特提出了局部平均处理效应（local average treatment effect，LATE）定理：对于任何随机分配的、第一阶段估计值不等于零的工具变量，如果满足单调性和排他性约束，那么局部平均处理效应就是简约式估计值与第一阶段估计值之比，也就是处理对依从者所产生的平均因果效应。所以，局部平均处理效应也称为依从者平均因果效应（complier average causal effect，CACE）。由于异质性的存在，CACE可能不同于始终接受者和从不接受者的平均因果效应，这也是这个平均处理效应被称为局部的原因。

在教育回报率的估计（Angrist和Krueger，1991）中，依从者是这个实证研究主要对象，出生在入学截止日期前完成高中学业的依从者是处理组，出生在入学截止日期后辍学的依从者是对照组。这个工具变量满足随机分配的假设，因为出生日期是随机的；也满足相关性假设，因为这个可以通过工具变量与高中毕业相关性进行检验；还可以合理地假设排斥者不存在，也即基本满足单调性假设；剩下的问题是排他性假设，由于工具变量也会产生入学年龄的变化，如果工资收入与入学年龄有关系，那么排他性假设可能会被违背。假如工资收入与入学年龄无关，那么这个工具变量就满足了四个相关的假设。在这种特殊情况下，安格里斯特和克鲁格使用工具变量法估计了教育的回报率。他们的研究发现，额外受教育一年的回报率大约为9%。这个估值高于早期的基于普通最小二乘法的回归估计结果。这是由于处理的异质性导致的。事实上，准实验设计对因果关系的评估针对的是那些尽快辍学的可能性很高的人，因此所估计系数代表的是确定了受工具变量影响的群体的平均教育回报率，这个群体是教育回报率研究的“依从者”，他们的平均处理效应可能不同于那些未受自然实验影响的“非依从者”。

2.其他准实验设计方法

安格里斯特和因本斯对局部平均因果效应的研究表明，即使存在异质性和不完全依从问题，工具变量也可以在一组最少但是在许多情况下合理的假设下识别因果治疗效应。其所确定

的因果效应是依从者之间的平均因果影响，即由于随着工具变量的变化而改变行为的人群子集的因果影响。这使得因果关系评估准实验设计所需满足的假设的性质更加透明，并且提供了一个探讨诸如双重差分法以及回归断点设计等其他准实验设计方法所要满足的条件的基本框架。事实上，其他的准实验设计方法与上述工具变量法有着相同的基本逻辑，就是借助自然实验将研究对象进行随机分组：在工具变量法中，依从者按照工具变量分成是处理组和对照组；在回归断点设计中，断点附近的一侧是处理组，另一侧的附近是对照组；在双重差分法中，受政策影响的是处理组，不受政策影响的是对照组等。

这些准实验设计方法要能对因果关系进行可信的评估，其假设前提类似于与工具变量法的前提假设，都是要通过外生的变化来进行因果关系评估。以回归断点设计[①]（regression discontinuity design，RDD）为例，它考虑的是这样一种情况，当沿着某个运行变量移动时，被处理的概率发生“跳跃”，这个跳跃点就是一个的断点。以断点作为分界，断点两边的研究对象可能存在着比较大的差异；但是如果在断点的附近两侧截取一个子样本，那么这两组人之间除了是否接受处理之外，其他各个方面具有相似的特征。可以看出，回归断点设计是最接近随机控制实验的一种准实验设计方法。回归断点设计的关键前提假设是断点的外生性，也就是断点要能够将断点附近的研究对象随机分组，这与工具变量法中随机分配假设一样；断点影响潜在结果的唯一途径是通过处理，这实际上就是工具变量法中的排他性假设。因而，回归断点设计也可以看成是一种工具变量法，断点就是一个工具变量（安格里斯特和皮施克，2019）。

安格里斯特及其合作者使用了回归断点设计研究了班级规模对成绩的影响（Angrist和Lavy，1999）。教育投入对学习表现影响的实证研究经常面临教育的投入的内生性问题的困扰，比如班级规模如果是按照学习表现来确定，将学习表现差的学生分到小班去，那么使用学习表现对是否小班的哑变量进行回归可能会得到班级越小，学习表现越差的荒谬结论。针对这个问题，他们借助以色列的分班情况作为自然实验。以色列的班级人数上限为40人。如果是41人通常分为两个小班，而39人就为一个大班。这创造了一个类似的实验场景：将入学人数略高于或低于40、80或120的学校进行比较，其中班级规模差异很大。在这种情况下，学生人数不同的学校在其他方面可能非常相似。因此，随着学校入学人数的增加，一个描述学生数量和学业成绩之间关系的回归应该在这些断点处表现出不连续性。使用以色列数据的回归断点估计表明，当班级规模下降时，成绩显著提高。

以上回归断点估计的关键假设是，在断点的两侧，个体在其他方面都是相似的。例如，在安格里斯特的研究中，这要求注册人数在35—39之间的学校的学生（对照组）和注册人数在41—45之间的学校的学生（处理组）具有相似的家庭背景。这可以通过处理组和对照组的均值检验来判断，还可以通过对断点附近学生背景特征分布。如果存在某种聚集特征，那么有可能表明一些家庭存在策略行为，会将自己的小孩分到班级比较小的学校，因而可能不是随机抽样，这会影响回归断点设计的有效性。当然，这个因果关系评估针对的是断点附近两侧处理组和对照组之间的差异，这个处理效应也是局部平均处理效应。

（三）比较与小结

纵观往届诺贝尔经济学奖获得者的贡献，有多位获奖者与实验或者因果关系评估有关系。他们或在实验室开展随机控制实验对经济理论进行检验（史密斯，2002年度），或者在实际的经济环境中进行实地实验（班纳吉、迪弗洛和克雷默，2019年度）；或者通过计量经济学方法的设

①回归断点设计最早出现于1960年教育心理学家唐纳德•坎贝尔（Donald Campbell）的研究报告（Thistlehwaite和Campbell，1960年）。坎贝尔曾就极力倡导回归断点设计，但是那时并没有得到学术界的重视。1972年，戈德伯格（Goldberger）将其引入在经济学领域中，但是真正使回归断点设计受到经济学家关注的是1999年安格里斯特和拉维在《经济学季刊》合作发表的论文（Angrist和Lavy，1999）。

计，例如金融学中用于市场有效性检验的事件研究法（法码，2013年度）以及用于宏观经济中的因果关系实证研究的结构宏观计量经济学和向量自回归模型（萨金特和西姆斯，2011年度）。安格里斯特和因本斯对因果关系分析的方法论贡献在于，他们在不能做实验的情况下借助经济社会变迁中制度、政策变化、自然变化产生的自然实验模仿随机控制实验进行准实验设计，由此丰富了因果关系评估的工具箱。

就准实验设计方法本身而言，安格里斯特和因本斯的贡献在于，他们将经济学中的工具变量法与统计学中的潜在因果关系模型结合，提供了一个理解各种准实验设计方法的统一框架。工具变量法本质上是借助工具变量对研究对象进行随机分组，其他的准实验设计方法也有相同的基本逻辑，都是使用自然实验来模仿随机控制实验的方式，将研究对象分为处理组和对照组。他们的研究指明了工具变量所必须具备的条件，或者更一般地，利用自然实验对因果关系评估进行准实验设计所必须满足的前提假设。安格里斯特和因本斯通过实证研究示范了如何寻找自然实验进行准实验设计。他们指出，最好的工具来自对某些计划或干预的制度细节的深入了解（Angrist和Krueger，2001）。

四、实证经济学的可信性革命

将随机控制实验方法引入实际经济环境中进行实地实验，或者借助自然实验对观察数据的因果关系评估进行准实验设计，都为实证研究提供了突破口，由此掀起了近三十多年来的实证经济学可信性革命的浪潮。其中，劳动经济学实证研究的可信性革命最为显著。这一部分重点关注卡德对最低工资、移民和教育问题的实证研究。1990年以来，以卡德及合作者创造性地借助自然实验，比如新泽西州和宾夕法尼亚州最低工资政策差异、1980年的马列尔偷渡事件以及州一级教育支出的变化等分别对最低工资、移民、教育等一系列问题的因果关系评估进行准实验设计，创造出近似于理想的实验环境，从而极大地提高了实证研究结论的可信性。

（一）最低工资与就业

最低工资政策的初衷是为了帮助低收入者提高收入。然而，最低工资增加企业的用工成本，从而降低了低收入和低技能劳动力的需求，进而可能减少就业。这个经典的预测也受到广泛的认同。有79%的经济学家同意最低工资法提高了年轻人和不熟练工人中的失业（曼昆《经济学原理》）。其实，最低工资政策是否使低收入者受益仍然是不清楚的。其中，最低工资政策对就业的影响是问题的关键。早期的劳动经济学家借助时间序列数据或截面数据进行实证研究，这些实证研究结论因为缺乏解决观察数据中的内生性问题而缺乏可信性。虽然实证研究发现最低工资的提高与失业率的增加呈正相关关系，但是研究者无法区分这种正相关关系是最低工资的提高导致了失业率的增加，还是失业率的增加激发了提高最低工资的呼声进而导致最低工资政策的出台。

1992年，卡德使用最低工资政策没有发生变化的州作为对照组，研究了加利福尼亚州的工资和就业演变（Card，1992）。研究发现，虽然加利福尼亚州在1988年最低工资提高了27%，与此同时的加利福尼亚州青少年的工资与对照州相比增加了10%，没有证据表明青少年就业率下降。卡德还发现，与对照州相比，加利福尼亚州的就业人口比增加了4%，这似乎是由劳动力参与率的增加推动的；换句话说，加利福利亚州最低工资政策对青少年就业的影响除了最低工资的影响之外，还可能受到不同州劳动力市场条件差异的影响。为了解决这个问题，必须排除不同州劳动力市场条件差异的影响。卡德和克鲁格尔借助新泽西州与宾夕法尼亚州的最低工资政策差异作为“自然实验”，通过双重差分设计估计了最低工资政策对就业的影响（Card和Krueger，1994），这是双重差分法的经典应用。

1992年11月，新泽西州将最低时薪从4.25美元提高到5.05美元，但相邻的宾夕法尼亚州的最低时薪将保持在4.25美元。由于两州的地理位置临近，可以认为两州处于统一的劳动力市场，具有相同的劳动力市场条件。这创造了一个天然的实验场所，可以排除劳动力市场条件的影响。通过比较这两个州在最低工资调整前后的情况，评估最低工资提高对就业的影响。卡德与克鲁格尔分别对两个州实地调查，收集了1992年2月（最低工资调整之前）和11月（最低工资调整之后）大约400家快餐店在这两年前后一系列数据，包括工资水平、雇用人数以及产品价格。选择快餐店作为调查对象，是因为快餐店是受最低工资政策影响最大的行业之一。

他们采用双重差分法进行估计。第一重差分分别估计了这两个州的快餐店就业状况在最低工资提高前后的差异。这个差异代表了最低工资和宏观经济的影响。例如，他们发现新泽西州的就业在最低工资政策实施后确有下降，但这主要是受美国东部地区持续恶化的宏观经济形势所致，而非提高最低工资的影响。第二重差分通过将新泽西州就业状况在最低工资提高前后的差异，然后减去宾夕法尼亚州就业状况在最低工资提高前后的差异，这样就排除了宏观经济形势对就业的影响。可以看出，这里必须有一个关键假设，那就是宏观经济形势对两个州就的影响相同，即所谓的平行趋势假设。这个可以通过比较新泽西州（处理组）和宾夕法尼亚州（对照组）在最低工资提高前快餐店就业的变化进行检验。

卡德和克鲁格尔的研究发现，没有证据表明新泽西州最低工资的提高减少了快餐店的就业。这项研究结论挑战了经济学家和普通大众有关最低工资对就业，特别是低收入人群就业有负面影响的共同信念。尽管该研究结果饱受争议，但是卡德和克鲁格尔对反事实结果估计的研究设计，引起人们对早期的研究结论和教科书上模型的预测结果的怀疑，进而引发了人们对最低工资问题的重新探讨。经济学家提出了很多新的解释，如劳动成本、生产力、价格反应、劳动力市场结构等，对最低工资的提高为何没有导致降低就业影响的机制提出了解释。例如，卡德和克鲁格尔认为，是因为快餐店可以将最低工资提高导致的负担转嫁给消费者。这些新的解释又引发了新一轮的实证研究，从而极大地促进了人们对劳动力市场运行机制的理解。

（二）移民的影响

移民问题也是广受争议的政策议题。一方面，大量移民特别是低技能移民的流入对当地的劳动力供给产生正向冲击，影响本地居民的就业前景并压低本地居民的工资；另一方面，大量移民的流入又增加了本地商品和服务的需求，这可能改善本地居民的就业前景。因此，事先并不清楚低技能移民流入如何影响居住本地居民。具体来说，移民的劳动技能与本地居民是替代还是互补，以及企业在面对移民涌入时是否要增加针对移民劳动技能的投资，这些问题人们并不清楚，因此需要对移民问题进行实证研究。研究的最大难点在于，移民可能会选择对劳动力需求不断增长的地方，这些地方即使没有移民，对本地居民的需求也与其他市场不同。这就是移民问题实证研究中的内生性问题。因此，实证研究的挑战在于，研究者要找到办法估计出，在没有移民涌入的情况下本地居民就业状况和工资收入的反事实结果。

1980年的马列尔偷渡事件（Mariel Boatlift ）提供了一个解决上述问题的机会。卡德极其敏锐地捕捉到这个事件，借助这个自然实验来评估移民的影响。事件的背景是，1980年，古巴首脑菲德尔•卡斯特罗宣布，任何希望离开古巴的人都可以移民。在古巴政府的支持下，自1980年4月至10月的6个月期间，大约125 000名古巴人从古巴玛丽尔港偷渡到美国佛罗里达州（主要是迈阿密），结果使得迈阿密的劳动力增加了7%。这就是马列尔偷渡事件。

马列尔偷渡事件为评估移民影响提供了一个理想的自然实验（Card，1990）。一是该事件导致的迈阿密的劳动力供给曲线是外生的；二是这些移民主要集中在迈阿密相对狭小的地区，没有对其他区产生外溢影响，这使得研究者可以用其他相似的地区作为对照组。卡德采用迈阿密

人口调查的个人失业数据，并基于人口统计和经济状况选择了四个具有可比性的城市（亚特兰大、洛杉矶、休斯顿和坦帕-圣彼得堡）作为对照组。在此基础上，卡德通过双重差分法估计了移民对本地居民工资和失业率的影响。研究结果令人诧异，大量移民的涌入对本地居民工资和失业率没有显著影响。卡德认为，这是因为迈阿密的劳动力市场有能力吸收这些移民。卡德的研究受到质疑，原因之一是他在选择对照组城市时存在主观性；有学者利用合成控制法解决了这个问题，他们使用合成控制法复制了卡德的研究并得到类似的结论（Peri和Yasenov，2018）。

卡德的这项研究再次挑战了经典的完全竞争劳动力市场理论模型的预测，深化了人们对移民影响的理解。在随后的研究中，很多的研究也发现移民对本地居民的影响不大。除此之外，在对移民影响本地人的机制的理论和实证研究中也发现，本地居民通过进入需要当地语言沟通技能等与移民竞争较少的职业来避免负面后果，实际上本地居民受益于新移民的涌入；相比起来，受移民影响最大的是先前的移民，新移民的涌入对先前的移民构成了竞争；企业通过技术投资适应移民流入，可以减少新涌入的移民对先前移民的不利影响。

（三）教育投入与教育回报

关于教育投入如何影响学习表现以及教育回报率的早期实证研究受到内生性的困扰。例如，不可观察的能力因素同时影响受教育程度和工资收入，如果基于工资收入对教育程度的回归来估计教育回报率，实际上是混杂了教育和能力的共同影响，因而未能对教育的回报率作出可信的估计。实证研究者需要找到受教育程度的外生变量来解决这个问题。安格里斯特借助出生日期导致的教育程度的外生变化来估计教育的回报率，以及使用回归断点设计估计小班教育对学习表现的影响，这是教育问题实证经济学的两个里程碑式的经典研究。卡德和克鲁格（Card和Krueger，1992a；Card和Krueger，1992b）这两篇论文研究了学校质量对劳动力市场结果的重要性，也是教育回报率估计的重要转折点。

这两篇论文都利用了学校质量的外生变化，这些外生变化来自20世纪30年代至50年代美国特别是美国南部对教育的大量投入，他们用师生比、平均学期长度和教师的相对工资来衡量不同州的教育质量。但是，劳动收入除了受到不同州的学校质量的影响之外，还受到各州劳动力市场条件的影响，因而，需要一种能够区分这些不同影响的策略。1992年，卡德和克鲁格利用人们跨州迁移这个自然实验设计了一个策略，其基本思路是比较居住在同一个州但成长于不同州的人，由于他们成长于不同的州，他们的劳动收入会受到不同州的学校质量的影响，但是居住在同一个州因而他们所面临的劳动力市场条件是一样的。换句话说，可以将学校质量好的看成处理组，将学校质量差的看成对照组，这两组人面临的劳动力市场条件是一样的，由此可以排除劳动力市场条件的影响。在实证研究中，卡德和克鲁格采用这种跨州迁移策略分别按居住州、出生州和出生队列（cohort）估算学校教育的收入回报，然后挑出与给定队列在特定州长大相关的学校教育回报。他们利用1980年人口普查的收入数据，以1920年至1949年间出生的男性为研究对象，评估了学校质量的影响，发现在学校质量较高的州接受教育的男性，额外受教育年限的回报率较高（Card和Krueger，1992b）。另一项研究则利用州级迁移策略，考察1960年到1980年黑人与白人男子工资收入差异的下降是否受到相对教育质量的影响，他们利用搬到北方各州的个人信息表明，1960年至1980年间黑人和白人收入差距缩小幅度中有20%可以由黑人学生学校相对质量的提高来解释。

卡德和克鲁格的这两项研究对早期的实证研究提出了质疑，重新激发人们对教育投入和教育回报率的兴趣，引发了一场关于学校质量和学校资源对学校和劳动力市场结果是否重要的讨论。过去30年的研究得出的总体结论是，在工业化国家，学校资源似乎对劳动力市场的结果很重要（Jackson、Johnson和Persico，2016）。这并不意味着所有类型的教育投入的增加都会带

来成就和劳动力市场结果的改善。对处于劣势的学生来说，学校资源对学校成绩的影响往往更大，这表明他们的学校选择比来自优势背景的学生受到更大程度的限制。

（四）扩展与评价

准实验设计方法的引入点燃了劳动经济学领域实证研究可信性革命的导火线，这场可信性革命从劳动经济学开始，像涟漪一样波及制度经济学、经济史、健康经济学、保险、产业经济学、犯罪经济学、宏观经济学等经济学的众多领域，涌现了许多经典的研究。例如，在制度经济学领域和经济增长问题的实证研究中，阿西莫格鲁、约翰逊和罗宾逊（2001）对制度影响经济增长的研究是利用工具变量法进行准实验设计的经典。可见，卡德、安格里斯特和因本斯推动了过去30多年来经济学各领域的实证研究方式的改变。

从安格里斯特和卡德及其合作者对劳动经济学的实证研究可以看出，准实验设计给实证经济学研究带来了两大优势。第一个优势是，通过引入准实验设计，在对照组的选择和设定下，研究者明确了影响研究结论可信性的关键性假设是什么，从而使得讨论重点更加明确。第二个优势是，准实验设计使得研究者可以像随机控制实验一样，通过平衡检验或者均值检验，以判断处理组和控制组在各个方面是否相似，以及处理后以处理组和对照组均值之间的差异来估计处理效应，这有助于对研究的过程和结果进行直接的展示和简单的解释。

五、结　语

过去30多年来，准实验设计方法的引入和发展推动了实证经济学可信性革命的发展。安格里斯特和因本斯对工具变量法的发展，提供了一个理解准实验设计方法的基本框架，给实证研究者提供了利用观察数据进行因果关系评估的强大工具，他们不仅极大地丰富了实证研究的工具箱，而且推动和拓展了因果关系分析的应用领域。在我国改革开放不断深化的过程中，各种制度创新和政策试点不断推出，这些制度创新和政策试点创造了许多可资利用的自然实验。这些自然实验为有志于应用科学规范的经济学方法讲好中国故事的经济学家（陆毅和孙天阳，2021）提供了丰富的素材，利用准实验设计的方法对这些制度创新和政策试点的效果进行评估，也有助于相关政策制定者调整和完善已出台的政策措施及其规划。

首先，要借鉴准实验设计方法讲好中国故事，必须注意到运用准实验设计进行因果关系的可信评估并非易事。随机控制实验需要研究者对实验进行精心的设计，并密切跟踪实验的实施过程；基于自然实验的准实验设计虽然不需要研究者亲自实施实验，但是也需要了解自然实验的具体情况，根据自然实验的特征选择准实验设计的方法，判断是否满足这些方法使用的前提条件，特别是自然实验如何对研究对象进行随机分配，从而解决观察数据的内生性问题。即使经过如此细心的准实验设计，对因果关系评估结论的解读仍然要十分谨慎，因为从自然实验中获得的因果关系推断，通常只适用于总体的某个局部。

其次，要规范准实验设计的研究过程，避免P值操纵。针对准实验设计实证研究论文的综合分析发现，这些准实验设计实证研究中存在一种发表偏差（publication bias）现象（Brodeur等，2020），与显著性水平刚好大于0.05相比，有大量准实验设计研究的显著性水平刚好低于0.05。这表明，实证研究者在进行准实验设计研究时可能存在P值操纵问题。研究者可能通过对照组的选择、控制变量的选取、删除异常值等方式对数据进行调整，操纵实证研究结果的显著性水平。防范P值操纵要依赖实证研究者和经济学界的共同努力，实证研究者要注重和规范准实验研究的设计过程；对于经济学界而言，要建立实证研究设计的平台，提前开放研究设计和数据是防范P值操纵的有效方法。

再次，在进行准实验设计的同时，要加强影响机制的理论研究。准实验设计方法更侧重的

是经济变量之间的因果关系评估，然而经济变量之间为什么会有这样或那样的关系需要进一步研究。正如对最低工资对就业影响的研究中，卡德通过准实验设计发现最低工资政策对就业的影响与教科书上基于完全竞争劳动力市场的预测不一致，颠覆了传统的观点。但是这并非问题的结束，而是提出了更多新的问题。经济研究不仅要知其然，更要知其所以然。这要求经济学者加强影响机制的理论研究，在最低工资问题中，经济学家通过劳动成本、生产力、价格反应、劳动力市场结构等方面解释了为什么最低工资政策对就业影响甚微，这实际上正是影响机制的研究。而对这些影响机制是否有效则需要进一步的实证研究。因此，经济学者在借助自然实验进行因果关系评估的同时，也要加强影响的理论机制的研究。当然，在这个过程中也要不断完善因果关系评估的准实验设计理论与方法。例如工具变量法中弱工具变量问题，工具变量分析依赖于强变量工具，如果工具很弱，估计就可能有偏差，因果关系评估就具有误导性。

最后，对于政策制定者而言，他们要意识到准实验设计研究结论的局限性。如前所述，准实验设计实证研究更多的是考虑因果关系评估的内部有效性问题，而且其研究结论通常只是针对受处理影响的子样本而言才成立。因而，准实验设计尽管极大地提高了因果关系评估的内部有效性，但是任何实证研究与特定的时间、地点和特定的研究设计有关，其因果关系的实证证据总是局部的。如果将其推广到新的环境中，就必须要考虑外部有效性的问题。这也是政策制定者在参考准实验设计研究结论制定或调整政策时需要注意的问题。

主要参考文献

[1]安格里斯特, 皮施克. 基本无害的计量经济学[M], 郎金焕、李井奎译, 格致出版社, 2012.

[2]安格里斯特, 皮施克. 精通计量: 从原因到结果的探寻之旅[M], 郎金焕译, 格致出版社, 2019.

[3]李宝良, 郭其友. 因果关系的实地实验与新实证发展经济学的贫困治理之道——2019年度诺贝尔经济学奖得主主要经济理论贡献述评[J], 外国经济与管理, 2019, 41(11): 136-152.

[4]陆毅, 孙天阳. 以科学规范的经济学研究方法讲好中国故事[J], 经济学(季刊), 2021, 21(5): 1877-1882.

[5]Acemoglu D, Simon J, and James A. Robinson. The colonial origins of comparative development: An empirical investigation[J]. American Economic Review，2001, 91(5): 1369-1401.

[6]Angrist J D, Krueger A B. Does compulsory schooling attendance affect schooling and earnings?[J]. Quarterly Journal of Economics，1991, 106(4): 979-1014.

[7]Angrist J D, Lavy V. Using Maimonides' rule to estimate the effect of class size on scholastic achievement[J]. Quarterly Journal of Economics，1999, 114(2): 533-575.

[8]Ashenfelter O. Estimating the effect of training programs on earnings[J]. Review of Economics and Statistics，1978, 60(1): 47-57.

[9]Card D. The impact of the Mariel boatlift on the Miami labor market[J]. Industrial and Labor Relations Review，1990, 43(2): 245-257.

[10]Card D. Do minimum wages reduce employment? A case study of California 1987–1989[J]. Industrial and Labor Relations Review，1992, 46(1): 38-54.

[11]Card D, Krueger A B. Does school quality matter? Returns to education and the characteristics of public schools in the United States[J]. Journal of Political Economy，1992, a,100(1): 1-40.

[12]Card D, Krueger A B. School quality and black-white relative earnings: A direct assessment[J]. Quarterly Journal of Economics，1992, b,107(1): 151-200.

[13]Card D, Krueger A B. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania[J]. American Economic Review, 1994, 84(4): 772-784.

[14]Imbens G W, Angrist J D. Identification and estimation of local average treatment effects[J]. Econometrica, 1994, 62(2): 467-475.

[15]Jackson K, Johnson R, Persico C. The effects of school spending on educational and economic outcomes: Evidence from school finance reforms[J]. Quarterly Journal of Economics，2016, 131(1): 157-218.
[16]Peri G, Vasil Y. The labor market effects of a refugee wave: Synthetic control method meets the Mariel Boatlift[J]. Journal of Human Resources, 2018. doi: 10.3368/jhr.54.2.0217.8561R1.

The Quasi-experimental Design Approach to Causality and the Credibility Revolution in Empirical Economics: A Review of Main Contributions by 2021 Nobel Economics Laureates

Li Baoliang[1], Guo Qiyou[2]
（1. *School of Economics and Finance*, *Huaqiao University*, *Quanzhou* 362021, *China*；
2. *School of Economics*, *Xiamen University*, *Xiamen* 361005, *China*）

Summary: How to evaluate the causal relationship between economic variables credibly using observational data is a common problem faced by empirical economists. The Royal Swedish Academy of Sciences awarded the 2021 Nobel Prize in economics to David Card for his empirical contributions to labor economics, and Joshua Angrist and Guido Imbens for their methodological contributions to the analysis of causal relationships. This paper reviews how they devise quasi-experimental designs for causality evaluation with the help of natural experiments and the credibility revolution in empirical labor economics with emphasis on the analysis of century-old questions: the employment effects of minimum wage, the labor market impact of immigration and educational investments. The main conclusions are as follows: Firstly, various seemingly different quasi-experimental designs of causality evaluation have the same basic logic. They all use the scene of natural experiments to imitate the way of randomized controlled trials (RCTs), and look for research objects with the same or similar characteristics as the control group that constitutes the counterfactual result of causality evaluation. Moreover, a survey of the classical literature of empirical labor economics shows that the credibility of causality evaluation does not depend on the data and econometric methods themselves, but on the in-depth understanding and thorough analysis of the operation mechanism of social and economic system. Therefore, the in-depth investigation and discussion of natural experiments is essential. Finally, China's economic reform provides abundant natural experimental resources for quasi-experimental designs. Researchers, who want to tell the Chinese stories with scientific and standardized economic research methods, should pay attention to the in-depth investigation and analysis of the systems behind natural experiments, standardize the quasi-experimental design process to avoid p-hacking, and strengthen the theoretical research of influence mechanism. Policy-makers should pay attention to the limitations and external validity of the research conclusions when learning from quasi-experimental designs to evaluate the effective of policy.

Key words: David Card; Joshua D Angrist; Guido W Imbens; causality; quasi-experimental designs; empirical economics

（责任编辑：宋澄宇）