“在评价科学化方面，我们一方面要注重评价人才队伍建设，评价专家要兼有基础研究和应用研究的经验，同时还要专门培养教育政策评价方向的研究生；另一个方面要加强实证研究及教育数据库的建设。美国教育评价研究的发展离不开健全的、数据详实且及时更新的、公开的数据库资源。

多元化评价：中美比较的视角

■严文蕾

有人存在的地方就有竞争，有竞争就需要有评价。然而，对人进行评价并不是一件容易的事，在高等教育领域中也不例外。教师评职称、评头衔要数论文，引人才又要看头衔……论文不够怎么办？学术“造假注水”便滋生出来。

这样的恶性循环，引起了党中央的注意，在2018年的两院院士大会上，习近平总书记就指出，“人才评价制度不合理，唯论文、唯职称、唯学历的现象仍然严重”。近年来，为解决这一问题，中共中央、国务院和教育部等部门发布的系列重大政策将“四唯”“五唯”清理作为突破口，推动着我国高校和科研院所科研评价制度由一元走向多元的重大转型与改革。

下面，笔者就将通过中美比较的视角，在对中美高等教育科研评价相关问题的优劣特征进行分析比较的基础上，探寻各自特点，力求促进和实现中美高等教育相互取长补短。

理解评价的本质

要对中美高等教育科研评价相关问题的优劣特征进行分析比较，明确评价概念在中美语境下的差异是前提。

教育评价在西方主要对应三个英文概念：Testing、Assessment和Evaluation。Testing，即考核、考试。Assessment，即各种能力测评。依据美国三大权威教育组织（美国教育研究会、美国心理学会、美国教育测量全国理事会）联合编制的《教育与心理测试标准》，Testing即通过一种系统的方法，获取有关人或项目的样本信息，从而推断出学生的知识、特征或倾向。Evaluation则侧重对教育干预效果的测定，包括微观层面教学策略效果的测定，以及宏观层面国家教育政策效果的测定。

这三者间，考试为评价提供收集证据的工具，测评是各项考试的综合，而考试和测评等多方面形成的证据可以支持有效的评价，三个概念间相互联系，环环相扣。因此，一个完整的教育评价过程包括了考试、测评和评价三个阶段。

那么，被人们广泛讨论、纠结的评价究竟是什么？该如何理解、剖析？

事实上，评价的本质是基于材料和证据的搜集与分析，对教育各个环节及其特征和结果进行判断的过程。比如，一所企业要想招聘一个人才，它需要该应聘者的简历，此外还要进行笔试、面试等，这都是根据材料和证据进行判断的环节。

评价有三个基本要素，分别是判断、标准、利益相关者。其中，作出判断是评价过程的终端环节；评价标准则是进行判断的根本依据。而评价标准的制定往往很难统一，它取决于价值观。因此，价值观的不同是导致评价标准产生争议的根源所在。

另外，任何评价过程都关涉多元的具有相互利益关系的主体。由于利益相关群体的多样化和差异化、资源及时间的有限性，教育评价往往很难同时满足各方利益诉求。但教育评价必须明确主要利益相关者，才能确定评价的价值导向、制定出符合利益相关主体需求的评价标准，继而作出合理的服务利益相关主体的价值判断。

评价工具——考试的诞生

在2300多年前的中国，科举考试制度诞生了。没有人能够预想到，由此诞生的考试制度竟一直延续至今。而西方在教育测评领域的历史则要比中国晚得多，以桑代克在1904年出版的教材《教育测量》和1923年出版的第一个斯坦福成就测验（SATest）为其教育测评领域最早的里程碑式的标志。

除了考试产生先后的差异外，美国考试发展的历程也与中国很不相同。

美国自上世纪30年代开始实施SAT考试，上世纪50年代开始实施区一级的标准化考试，上世纪70年代开始实施州一级的标准化考试，上世纪80年代扩大到全国考试，上世纪90年代后开始尝试国际考核。SAT在发展至今的八十余年里，其形式和内容基本上没有改变，仅在写作题目方面有所增添。考试发展的总体趋势是实施的范围和规模越来越大。可见，美国考试发展呈现自下而上的特征。

与之相反，中国考试的发展路径则呈现自上而下的特征，往往始于国家统一考试，继而逐渐放权到省和市。

虽然，中美教育和历史文化背景不同，但是不同的考试发展路径没有优劣之分，它们均服务于学生的发展和考试制度的不断完善，也是完成评价的工具之一。

面向问责的教育评价

教育评价的主要功能之一是问责。以美国为例，其最重要的教育法案——《不让一个孩子落后法案（NCLB）》即规定以考试结果作为问责的依据。根据NCLB法律要求，各州开发了州级统一考试，要求所有学生参加，并以测评结果为依据对教育管理者进行问责。以麻州为例，这一考试即马萨储塞州（以下简称麻州）综合评估系统。依据这一系统的测评结果，麻州学校被评定为五个等级：1级代表优异；2级代表合格；3级和4级代表较差（排名后20%的学校）；5级代表“长期表现不佳”。其中，3~4等级的学校会获得额外支持与援助，5级学校将由麻州基础教育部接管。同时，各个学校的管理者会接受相应的问责。

事实上，基于评价的问责制度对于教育质量的提高有较为显著的效果。通过波士顿公立学校NCLB问责结果统计（2013~2016），我们或许可以有更加直观的感受。根据该统计，2013年，波士顿地区被统计的公立学校中1级21所、2级12所、3级59所、4级7所、5级2所；2014年，被统计的公立学校中，1级14所、2级22所、3级54所、4级7所、5级2所；2015年，被统计的公立学校中，1级14所、2级23所、3级53所、4级8所、5级2所；2016年，被统计的公立学校中，1级21所、2级24所、3级46所、4级9所、5级2所。从统计数据中可见，实行问责制度后，波士顿地区1级和2级的合格与优质公立学校总数基本呈现逐年增加的趋势，3级和4级需要改进的学校总数逐渐减少，可见，以测评驱动问责可在一定程度上提高教育质量。同时，测评也是实现教育公平的重要手段。考核不合格的学校多是弱势群体学生集中的学校，通过考核问责，这些学校被动提高了学生的学业成绩和教育质量。

学业考试是评价的重要组成部分和依据，但并不等同于评价。中美两国的考试在综合评价中占据的权重具有显著的差别，按照学习阶段（幼儿园、小学、初中、高中、大学），根据相关数据，将中美学生考试在评价中的权重做成函数分布图（如图1所示），差异一目了然。

从图中不难看出，中国学生在接受高等教育前各级考试、考核随学段增长而逐年加码，到了高中达到顶峰，大学后却降下来，呈缓慢下降趋势。相比之下，美国一直呈持续上升趋势，直到博士研究生阶段，其中，虽然在高中及以前一直低于中国，但是到大学以后高于中国。

由此可见，考评应符合人的发展规律，即随年龄增长，对学生的考试要求、责任心期望等应相对增加。然而，在中国高等教育阶段，考试没有严格执行或者效果没有充分发挥，这可能是造成满意度偏低的原因之一。

科学化的教育评价设计

除了在现行的考评上存在显著差异外，在对于构建科学化的教育评价设计上，中美两国侧重也不尽相同。

笔者基于对八本中国权威教育类综合期刊筛选出的近三年评价主题相关论文的分析来看，大多数文献侧重于评价的基本理论探索和理论框架的构建、引介及运用。这与美国相关文献侧重于以评价解决实际教育问题，及教育政策和干预效果评价的实证研究有一定的差异。

为了了解美国当前教育评价的目的与内容、主要功能和科学方法，笔者对从美国教育评价领域最权威的学术期刊《教育评价与政策分析》中筛选的近三年来的81篇实证论文进行了分析。

从搜集的81篇论文来看，当前美国评价的主要内容有：NCLB执行效果的深入评价和持续问责、弱势群体学生数学成绩的提升、低收入家庭学生大学入学机会、校园突发事件对学生学业成绩的影响等。这些文章也反映了美国教育评价中存在的两个钟摆现象：一是质量和公平之间的平衡，另一个是知识和能力之间的均衡。评价的直接目的在于衡量学生的能力水平，而其终极目的是服务政策和教育公平。因此，美国教育评价更重视对政策干预效果的评价，探寻国家资助项目对教育公平起了多大作用，尤其是对弱势学生群体（移民学生、西班牙裔学生、黑人学生、英语非母语的学生、特殊教育的对象、来自低收入家庭的学生、学业成就低的学生、女学生等）的干预效果如何。

若说，美国教育评价的核心内容是质量与公平，那么其主要功能则是问责和改进。

依据对81篇论文的分析，笔者发现59%的教育评价旨在完善政策和干预措施，32%的评价指向问责，其他9%的评价则意在引起政府关注、促进管理加强。例如，布莱恩·雅各布等学者对密歇根优秀课程（MMC）的效果进行了评价与问责，发现MMC所包含的较高期望对学生的学习成绩影响不大。

事实上，在美国教育评价的问责和改进功能往往是同时实现的。《每个学生都成功法案（ESSA）》就要求各州通过评价问责找出陷入困境的学校，继而通过制定以证据为基础的资助政策，扭转其弱势局面。

在美国，教育评价的科学化设计是学者们关注的重点之一。

依据筛选的文献可见，美国教育评价科学化设计有两个特点。一是由于教育的滞后性特征，美国所有教育干预都要做到长期跟踪，否则教育效果不能显现。二是强调使用实验方法（随机实验和准实验法）。所谓随机实验，就是将研究对象随机分组，对不同组实施不同的干预，以对照效果的差异，具有能够最大程度地避免实验设计、实施中可能出现的各种偏倚，平衡混杂因素，提高统计学检验的有效性等诸多优点，被公认为是评价干预措施的金标准。例如：凯瑟琳·M·布罗顿等学者利用一项随机实验发现，威斯康星州的低收入家庭学生获得额外助学金后，可以改善学生的学术成绩和发展前景，从而得出了经济资助促进大学成功的方式之一是通过资助来减少学生兼职工作的时间从而提升其学习效果的结论。

而常用的准实验研究设计则有标准或目标比较、等组对照、统计控制（前测和后测或只后测）、统计控制—后测控制组设计、其他前测—后测控制组设计、其他后测，仅从单个受试者设计中选取对照组的设计等。由于教育实验对象是学生，要符合伦理原则，很难严格控制所有无关变量，因此常常采用准实验法，即在实验中未按随机原则来选择和分配被试，只把已有的研究对象作为被试，且只对无关变量作尽可能控制的实验。笔者筛选的文献中，也是此类研究较多。

事实上，不论采用何种评价方法，评价最核心的还是提供证据。美国教育研究院按照是否采用对比的科学研究方法、是否有真正的控制组和实验组、是否随机、是否能复制等标准区分了对“证据”“可能是证据”“没有证据”的判定（详见表1）。

在美国教育部和国家自然研究基金的每一个项目规划中，必不可少的就是评价，且是第三方评价，重点是通过评价搜集数据以衡量项目干预的效果。干预的效果可以用效应量来表达，效应量越大说明效果越好。影响效应量的因素包括：干预的时间、参与者数量、开始时间（在学前班或幼稚园、一年级或以上）、结束时间（从干预结束到评价之间的时间间隔）、干预主题（阅读、数学、语言、拼写、其他科目）等。这值得中国学者借鉴，在评价设计中要注意控制好上述因素，最大程度地提高效应量。

总之，在笔者看来，美国的经验要批判性地借鉴。

首先，美国在对教育干预的及时性、过程性、客观性、第三方评价方面的经验可以为我们提供有益借鉴。尤其对于一些中美共存的教育问题，如能力分班等，美国已经做了半个多世纪的探索和研究，并对每一种干预都进行了评价，其成果非常值得我们参考。

其次，在评价科学化方面，我们一方面要注重评价人才队伍建设，评价专家要兼有基础研究和应用研究的经验，同时还要专门培养教育政策评价方向的研究生；另一个方面要加强实证研究及教育数据库的建设。美国教育评价研究的发展离不开健全的、数据详实且及时更新的、公开的数据库资源。

最后，笔者建议我国不妨也创办一本权威的、国际化的教育评价期刊，这将有利于集中中国教育评价的成果，同时也有利于国际交流与传播。

（作者系美国马萨诸塞大学波士顿分校终身教授、教育领导学系主任）

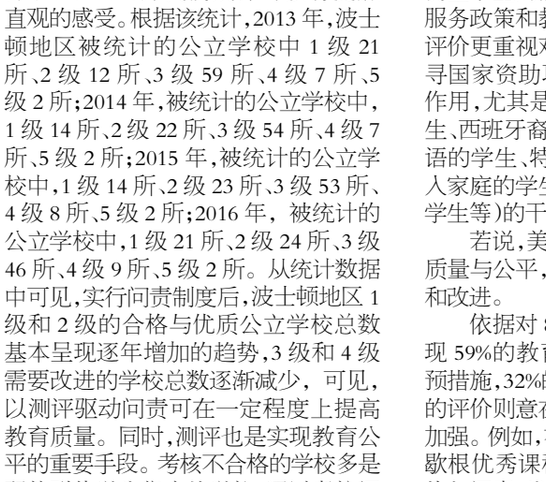

图1 中美学生考试在评价中的权重对比示意图

	证据	可能是证据	没有证据
对比	在两个或以上组之间进行对比：控制组和实验组	在两个或以上组之间进行对比：控制组和配对组	只有一个组（缺少实验组或配对组）
随机分配到各组	对象是随机分配到各组	对象是大致随机分配到各组	对象没有随机分配到各组
一致的处理	对不同组的处理是一致的	对不同组的处理达到最小的差异	对不同组的处理存在不可接受的差异
多个“地方”进行	多个“地方”进行	在一个或多于一个“地方”进行	在一个“地方”进行
复制	研究可使用相同程序复制（多种情形或多个案例）	研究可重做并得到相同结果，但不是通过完全一致的方式	研究不可复制；仅从一套数据中得出结论
控制	对于实验组和对照组外部影响因素得到控制	对于实验组和对照组大多数外部影响因素得到控制	外部影响因素极少或没有得到控制，可能干扰研究结果

表1 关于证据质量的分类

域外传真

无协议脱欧将对英国大学造成严重威胁

■珍妮特·比尔

随着脱欧审议结束日期的加速临近，在无协议脱欧即将成为现实的状况下，英国大学面临前所未有的危机。在英国，受到上述问题的影响，来自欧盟的5万教职工、13万学生，在新一年的初始就面临着自身未来的极大不确定性，更不要提那1.5万名在欧盟留学的英国学生将会如何不安了。

事实上，如果在3月29日英国真的无协议脱欧，那么英国大学还会面对更多其他的隐患。这包括了采购、数据维护、互相的质量认证、知识产权。而大学通过提供就业、社会服务、支持区域供应链等形式对地区经济发展做出贡献也将会陷入不利的境地。更为重要的一点是，无协议脱欧将会给英国重要的研究链条带来损害，而这一链条中涉及了从癌症治疗到对抗全球气候变化技术等，关系到广大社会群体利益的研究。这也是为什么在上一周，大学相关领导会集体向英国下议院议员提交文书，强调我们正处于怎样的危机中。我们强烈呼吁政府和议会正确的措施和保证要到位，特别是要让大学顺畅地转变，而不是在3月29日时无协议的脱欧。

大学的研究和教学，对于英国脱欧后能否繁荣发展是极为重要的。英国大学是世界上最优秀的研究体系之一，吸引了大量的优秀学者、顶尖学生，以及全球合作伙伴，这些对于英国来说都不能够失去。研究经费问题就很好地展示了我们面临的风险有多大。在欧洲研究理事会和“玛丽·斯克沃多夫斯卡·居里行动计划”（欧盟“地平线2020”科研规划中的一部分，是欧盟资助个人科研最高奖项之一，也是欧盟投资最多、内容最丰富的全球性科研计划）的资金支持下，英国大学有着引以为傲的创新能力和解决突破问题的过往。高风险往往会带来大收获，正是这些资金流使得英国的科学和研究茁壮成长。

在承接欧洲研究理事会基金项目方面，英国目前是世界上最成功的国家。我们的估算显示，在2007年到2017年间，英国得到了1850笔项目资金，与我们数目最相近的竞争对手德国也只有1330笔。在这些基金项目的支持下，英国学者获得了多项负有盛名的奖项，包括6个诺贝尔奖、4个菲尔兹奖、5个沃尔夫奖。根据2017年的一项独立研究表明，在欧洲研究理事会的项目中，超过70%取得科学突破或获得重大进展。当然，政府已经保证一些研究资金的持续投入，但是最为关键的是，政府要将这一保证覆盖范围扩大，并且不延期，尽快取代欧洲研究理事会和“玛丽·斯克沃多夫斯卡·居里行动计划”的资金支持。

如果没有得到可靠的保证，据估算，英国大学在接下来的两年内，将会损失近12亿英镑的研究经费，这对于大学雇佣和维持教职工的能力会带来巨大的影响，而这些员工正是大学的生命线。

此外，我们现有的世界领先的学者和研究人员可能会离开，前往欧洲研究理事会资金获得不存在风险的国家，而那些原本想要来英国的研究者可能也会望而却步。最近的诺贝尔物理学奖获得者邓肯·霍尔丹就表示，他曾经考虑从普林斯顿回到英国，但是如果英国从那些有名望的机构获得基金资助的机会被取消的话，他可能会改变主意。而这样的事件并不是个例，这映照出了整个英国大学中会出现的状况。

我们只剩下几周时间供英国政府和议会来寻找出路，避免无协议脱欧的结果。否则，可以毫不夸张地说，英国的研究、文化、科技会面临一个需要大学和整个国家花费几十年时间来恢复的大退步。

（作者系利物浦大学校长、英国大学联盟主席，许悦编译）

四海游学

我在卡内基梅隆大学

■阳少轩

来到卡内基梅隆大学已经一个学期了，但每次看到学校墙上那句“My heart is in the work”的校训时，我都会回想起第一次看到这句话时的心情。要是有人问我在美国卡内基梅隆大学上学的感受，我想最好的答案莫过于“My heart is in the work”。

初到美国

当我拖着行李箱走出匹兹堡机场时，心情是复杂的：既有踏上异国他乡的激动，又有远离祖国开始独自生活的不安；既有对未来研究生学习的憧憬，又有对美国大学繁重课业的担心。都说留学是“洋插队”，这话一点也不假。

在国外，除了要完成紧张的学习任务，还要在生活上照顾好自己。记得第一次面对空空如也的出租屋，突然感到这也许是每个留学生都会遇到的挑战。

睡过地铺，抱怨过为什么选择辛苦的求学方式；坐很久的车去购买家具，再汗流浃背地组装；给自己做难吃的晚餐……在来美国的初期遇到了很多在国内没有做过的事情。有时候也会感到不知所措，但是每完成一件任务都会有一种成就感，渐渐地会感觉到自己真的有所成长。

在校苦读

在美国计算机排名第一的学校学习，我最大的感受就是同学们十分刻苦。不管在学期的哪个阶段，最热闹的校车一定是半夜的那一班，凌晨的图书馆还是如白天一样，第二天清晨的教学楼角落里经常有各种睡姿的同学……

还记得网上曾经有人开玩笑说：“在美国上学，你只能学习、睡觉、玩耍，三选二。但是在CMU，很多时候你只能选择学习，尤其是当你选的课比较繁重的时候，你甚至都没有时间去超市，没有时间刷新各种社交软件上朋友们的动态，没有时间认真地坐下来吃一顿饭。”

也许有人会觉得这样被学习充斥的日子很疯狂，而当你身处学校时，就会发现疯狂无处不在：遍布全校的计算机机房；随处可见能让人推导公式的白板；走廊里时常出现最新研究的机器人；无论在路上还是健身房，总能听到有人在讨论一些前沿的技术。

而学校的课程从教授到助教，甚至从每一次作业和实验中，你都能感受到一种认真的态度。

很多课的授课教师都是相关领域里的佼佼者，但他们在课上不仅没有架子，还非常和蔼可亲，遇到不懂的问题，他们都会耐心地解释。平时助教和教授还会有大量开放的时间与学生一对一、面对面地沟通与交流，他们会认真地为你解答所有的问题，直到你理解为止。每一次的作业也都会有详细的反馈，如果你认真地投入其中，一定会有一种从一无所知到一知半解，最后融会贯通的感觉。在这所学校留学，痛并快乐着，在这里能让人真正领悟到那句“My heart is in the work”的含义。

感受匹兹堡

匹兹堡是一座让人感觉很舒适的城市，它曾经被评为美国最宜居的城市之一。站在市中心你能感受到城市的繁华，而在远离中心的居民区你又能体会到小镇的宁静。

由于钢铁行业的衰落，这座美国曾经的“钢铁之城”已经不再充满重工业的气味，散布在各个角落的画廊和艺术馆让这座城市更有艺术气息。居民区路边各种各样的花朵也为这个城市增添了许多色彩。周末闲暇之余你可以和朋友去爬山，听演唱会。

然而，在这座充满魅力的城市。除了享受美景，也要为学习疯狂。

（作者系卡内基梅隆大学留学生）