发展改革动态

2021 年第 12 期 共 141 期

发展规划处 2021 年 7 月 10 日

【聚焦评价改革】

教育评价的几大问题及发展方向

摘要：本文探讨教育评价必须面对的几大问题，包括教育目标的多重性、教育评价的完整性、教育目标的非兼容性及矛盾性，以及教育目标的个性化等。这些问题是教育评价中不可避免但又必须恰当处理的。因此，教育评价的发展方向应该是更加关注长期的教育性目标，减少对短期的教学性目标评价的依赖；更加注重个性化评价，找出个体学生的价值和发展方向。

关键词：教育评价；教育目标与教学目标；个性化教育；认知与非认知；社会情感能力

教育评价改革在世界各地都是一个热门的话题。究其原因，大约有四个方面。第一，第四次工业革命即将为社会带来巨大的变化。在三次工业革命之后，以人工智能、大数据等为核心的新一代技术所带来的革命毫无疑问会重新定义知识和技能的价值。过去有价值的知识和技能在全新的智能机器时代可能变得毫无价值，而以前没有价值的知识和技能反而会增值，成为社会最需要的（Zhao，2018c）。那么教育就必须帮助受教育者获取这些知识和技能。提倡 21 世纪技能的有关文献汗牛充栋，已经被广为接受（Duckworth & Yeager，2015；Zhao，2016a）。第二，现有评价方法的弊端日益明显。单一的评价方法已经不能够全面掌握学生的学习情况，更不能完整体现学生身心发展状况，而且会对某些背景的学生造成伤害（Emler，Zhao，Deng，Yin，& Wang，2019；Levin，2012）。第三，近年来对非学科知识技能的重视让人们重新思考教育的目的。教育绝不应该仅仅看重学生的学习，而必须关注学生的身心健康和健全发展，因此学生的社会情感能力（Social and Emotional Learning）、身心健康开始受到广泛关注（OECD Better Life Initiative，2017；Zhao，2020）。但传统教育评价并不关心学生的身心健

康，因此必须改革。第四，新冠肺炎对世界教育的冲击在现代教育史上没有先例，它导致全球十多亿学生一段时间无学可上（UNICEF，2020），同时也让无数的大规模教育评价终止。为此，不少人呼吁就此放弃大规模教育评价，寻求全新的教育评价体系和方法。

教育评价改革由于不同的原因而产生了不同的方法。有的是增加教育评价科目，比如 OECD 的 PISA 近年来增加了语文、数学和科学以外的国际化能力（global competence）、解决问题的能力（problem solving skills）和创造力（creativity）评价，也开始报告学生的幸福感和对生活的满意度等。有的是微改革，比如澳大利亚的新南威尔士州、维多利亚州、昆士兰州和堪培拉地区的教育厅最近发布的对其国家教育考试 NAPLAN 的评价报告认为该考试应该继续，只需要微调一下考试年级、反馈时间以及增加一项批判能力考试（McGaw，Louden，& Wyatt-Smith，2020）。

教育评价改革无论如何改都必须解决几大问题。这些问题不是教育评价的技术问题，而是本质的哲学问题，是社会、政府必须考虑的问题，但同时也涉及教育的科学性问题。这几个问题关乎教育评价的目标与发展方向，也关乎教育评价的实施和应用。本文就这几个问题展开讨论，并希望引起各界的重视。

一、教育目标的多重性

教育评价，无论是评价教师、学校、教育体系或者是学生，都离不开教育目标。一切评价都必须以是否或者在多大程度上达到教育目标为基本判断标准。说一个学校如何应该看它是否完成了教育目标，完成度如何。评价一个教师也是看他所教的学生是否或者在多大程度上成为我们想要的人。当然，评价学生更是要看其在多大程度上完成了教育目标。

但有个问题一直妨碍教育评价全面公正地进行，即教育目标的多重性。除了短期有目的的培训外，教育是为多重目标服务的。我们开展教育的目的是为了让孩子德智体美劳全面发展。德智体美劳就是多个目标。没有一个教育或者教学只是为一种目标服务的。比如我们常说的“教书育人”就至少有两个目标：希望学生学会读书并成为“人”。韩愈《师说》讲老师“传道授业解惑”也包含多个目标。就现代教育来说，我们的教育目标就更多了，包含认知目标（掌握知识、习得技能）和非认知目标（也称为情感性目标，如好奇心、自我认知、信心等）。

我们可以从不同角度来看待教育目标的多重性。首先，教育的目标具备多学科性，也就是说教育不是只教一门学科，它需要学生掌握多方面的知识和技能，因此学校教育必须包含不同的学科。从小学开始就有语文、数学、音乐、美术、体育以及相应的自然科学和社会科学，加上各个国家和地区的政治和社会组织以及相应的价值观和社会道德等。其次，教育学科的多重性也包含培养能力的多重性。每门学科有自己的知识和能力，但同时也培养超越学科的素养。比如全球化能力、个人品格、创造力、创业能力、沟通能力、合作能力、批判性思维能力等。

虽然世界各国的课程有一定差异，但毫无疑问，它们追求的教育目标都是多元的。例如由美国几十家科技公司和有关教育机构所推出的 21 世纪技能（21st Century Skills）包含关键学科、跨学科，还包含解决未来问题的 4C，即创造与创新能力（creativity and innovation）、批判性思维与解决问题能力（critical thinking and problem solving）、沟通交流能力

（communication）和合作能力（collaboration）。除此之外，还有信息技术、媒体和技术能力，还包括如何生活的能力（Partnership for 21st Century Skills，2007）。

与此类似，欧洲议会和欧盟理事会在2016年发布了关于欧盟公民为“实现个人目标、成为积极的公民、凝聚整个社会以及实现就业所需的核心能力”（The European Parliament & The Council of the European Union，2006）。除了学科之外，他们还提出8种主要能力。这些能力融合了“知识、技能与现实情况相适应的观念和态度”（p. 13）。此外，欧盟还强调其他关键能力，如批判性思维、创新能力、自我启发能力、解决问题能力、风险评估能力、决策能力以及积极调整自我情绪的能力等。再比如，澳大利亚的国家课程包括八大学习领域和八大超越学科的能力（Australian Curriculum Assessment and Reporting Authority，2010）。

二、教育评价的完整性

教育目标的多重性给教育评价带来了巨大的挑战。首先，如何评价才能保证评价的完整。如果我们承认教育目标的多重性，承认教育需要照顾多个教育目标，那么我就必须评价所有目标完成的情况。然而，在现有教育评价中，我们并没有评价所有的教育目标，而是只评价一小部分目标。

单纯从学科来看，绝大部分地区的教育评价基本上是评价语文（阅读写作）、数学和科学。比如PISA作为全球最大的教育评价项目，每三年一次，但其考核的教育目标是很有限的。虽然增加了其他能力的评价，但基本科目还是没超越语文、数学以及科学。许多国家的大规模考试也基本局限于语文和数学。比如美国各州对学校的教育测评也是以语文、数学为主，基本不考虑其他学科。美国的教育质量追踪考试（NAEP）也是以语文和数学为主。澳大利亚的国家年度考试（NAPLAN）考的也是语文和数学。各国更重要的升学考试也基本是考语文和数学，可能再加上外语。其他学科的评价基本上由教师上课评分执行，没有纳入大规模的统一考试。

教育目标中所提到的能力基本没有评价。在现有教育评价体系里面，除了PISA的批判性思维能力、全球化能力和已提出并即将开展评价的创造力之外，没有一个国家和地区有大规模关于能力的评价。加拿大不列颠哥伦比亚省（British Columbia）虽然将创造力、合作能力、沟通交流能力等列为关键能力，而且有比较灵活的课程支持这些能力的培养，但也没有完整的评价这些能力的系统。

可以说，到目前为止没有一个教育系统对所有教育目标的评价是完整的，许多教育体系开展的评价只是关注了教育目标中很少一部分知识技能，而且用这很少一部分的评价对学生、老师和学校进行很重要的判断，导致教育评价的狭隘化，进而导致教育活动的狭隘化。因为考试的指挥棒作用，所考的才是所教的所学的，所以即使一个教育体系将许多学科和能力纳入教育目标，但如果评价的范围有限，教育活动也会因此受到极大的影响。

教育评价的狭隘性出自几方面的原因。第一个原因是评价的成本。教育评价是需要极大成本的。成本包括研发评价工具、实施评价、给出结论的直接经济成本，还包括学生、教师等参与评价的时间和精力成本等。如果将所有学科所有能力都纳入正规的统一评价，那么成本必然大幅增加，可能增加到一个教育体系不可承担的程度。因此，许多教育体系就选择了看似容易执行的小

范围评价。第二个原因是评价技术。到目前为止，许多学科和能力并没有科学的被大家广泛接受的评价工具。比如，艺术、体育、道德、公民等学科并没有统一的评价工具，创造力、批判性思维能力和交流沟通能力等也没有相应的评价工具。虽然有不少努力，但从评价的科学性来看，不少学科和能力的评价工具都与语文和数学相差甚远。从另一个角度来说，今天的教育评价基本被评价工具主宰，不是我们想评价什么就评价什么，而是我们能评价什么就评价什么。第三个原因是教育评价决策者的决定。也许有不少决策者认为教育评价只需要看主要的学科，在他们看来，也许语文和数学之外的学科和能力本身并不重要，因此就不考虑评价其他目标。最后，教育目标本身也许“太狡猾”，不容易评价或者本身说起来好听，但其本质我们并没有认识清楚，因此无法进行评价。比如公民与社会能力、道德能力等本身就是非常复杂的概念，而且没有统一标准，难以描述，就更难以制定标准，开展评价了。

教育评价的狭隘性是现有教育评价的最大的问题。它导致了许多教育体系对学生的误判，对教育质量的错误追求以及对教学的误导。教育评价改革要改的问题之一就是不能再让狭隘的教育评价继续下去，要改变政策和发展方向，要全面完整地评价所有教育目标。

三、教育目标的非兼容性

要全面完整地对所有教育目标进行评价的一大问题是教育目标的不兼容。决策者在设计教育目标的时候思考的是所有学生都应该学习所有的学科、掌握所有的能力，都能达到预设的目标。然而，现实情况并非如此，因为不是所有教育目标都是相互支撑的，它们之间甚至会发生冲突。也就是说在追求某一个教育目标的时候，可能对另外的目标有反作用。或者说一种教学方法、一个教育体制有利于某一些教学目标，但对实现另外的目标有害。也有可能某一种教学方法或教育体制对某些学生实现目标有利，但对另外的学生则有害（Zhao，2018d）。

（一）学科之间的矛盾

首先，最明显也是最容易理解的非兼容性存在于学科之间。这个道理非常明显，就是说各个学科之间的发展是有矛盾的，这个矛盾的最大来源就是时间。时间对每一个学生来说都是一个恒量，一个不可能变化的量。如果一个学生每天花 4 个小时学习数学，那么这 4 个小时就不可能用来学语文。同理可推，她如果花 2 小时学英语就不可能再把那 2 小时用来学音乐（Zhao，2018d）。

一个教育系统设定的是学生要掌握好几个学科的知识和技能，这项要求可以有几个层次。第一是全部掌握，就是要求所有学生掌握课程的所有内容。第二是掌握最基础的一部分，就是要求学生掌握课程的基础部分，达到及格的水平。第三是超越要求，允许部分学生超越课程内容，达到更高的水平。

教育评价的目的也是多样化的。从国家层面来看，也许要求的是第二项，也就是所有学生达标及格就可以，在这种要求下，学科之间的矛盾还可以处理，因为课程安排的时间基本能够让学生掌握各学科的基本内容。但是，如果是要用教育评价来选拔学生，那么要求的就是第一和第三项。在选拔性的教育评价中，所有学生相互竞争，都要考出一个好成绩，那么对他们的要求就变得很高了，要对学科内容全部掌握甚至超越所学的内容。为了达到这个目的，学生就必须花时间在相应的学科上，因此就会产生学科之间时间上的冲突，由此学生就必须放弃一些选拔性考试不

考的科目而把时间放在必考科目上。其结果就是教育目标不可能全部实现，这就是由许多考试导致的对一些学科重视而侵占其他学科时间现象发生的原因，致使教育目标失衡。

（二）学科成绩与情感发展的冲突

迄今为止绝大部分教育评价考核的都是学科内容和学科能力。虽然有不少国家和地区都把学生的情感发展作为教育目标的一部分，但从现有的国际性考试来看，学科成绩与情感发展是有冲突的。PISA 和世界数学科学国际测试（TIMSS）都提供了足够的证据。

2006 年，美国布鲁金斯研究所研究员 Tom Loveless 在一份关于美国学生学习情况的报告中指出了一个特别有意思的现象。在 2003 年的 TIMSS 中（表 1），各国学生的考试成绩与他们对学科的信心和享受程度成反比。也就是说，考分越高的教育体系的学生对数学和科学越没有信心也不如考分低的学生享受学习（Loveless，2006）。这一现象一直持续。到 2011 年的测评中（表 2），东亚教育体系，比如韩国、日本、新加坡以及中国台湾、中国香港的学生成绩远远高于美国、英国、澳大利亚，但这些国家学生的信心和喜爱学科的程度远远低于美国、英国和澳大利亚（Zhao，2016b）。

表 1　TIMSS 成绩与信心和享受课程的关系

	等级	相关
信心	4	-0.58
	8	-0.64
	4	-0.67
享受课程	8	-0.75

资料来源：Loveless，2006

表 2　TIMSS 2011 数学成绩与信心比较

国家地区	数学分数	信心（%）	看重数学（%）
韩国	613	3	14
新加坡	611	14	43
中国台湾	609	7	13
中国香港	586	7	26
日本	570	2	13
美国	509	24	51
英国	507	16	48
澳大利亚	505	17	46

世界上的另一大教育评价项目 PISA 也发现存在同样的情况（表 3），即 PISA 考试成绩与情感发展水平成反比。例如，PISA 2010 年的考试成绩与学生的创业信心有显著的负相关，同时也与各

国的创业活动有显著的负相关，也就是说，PISA 成绩越高的国家和地区的人口创业信心越低，创业活动也越少（Zhao，2012）。

这一现象在多年的 PISA 测试中都是存在的。从最近一次 PISA 测试，也就是 2019 年发布的测试结果来看，这个负相关现象依然存在。例如，2019 年的 PISA 结果表明，学生对生活的满意度与 PISA 分数成负相关，PISA 成绩越好的国家和地区其学生对生活的满意度越低，而考分低的国家和地区学生的满意度越高。此外，2019 年 PISA 还发现，害怕失败的心理与 PISA 考分成正相关。即学生越是害怕失败，其分数越高；不害怕失败的学生分数越低。而教育的目的是让学生追求成功，不是害怕失败。

表 3　PISA 2010 成绩与创业活动相关系数

	PISA 阅读	PISA 数学	PISA 科学
感知能力（Perceived Capabilities）	-0.595**	-0.586**	-0.608**
新生创业率（Nascent Entre Rate）	-0.693**	-0.636**	-0.678**
新企业拥有率（New Biz Ownsp Rate）	-0.371*	-0.374*	-0.392*
早期创业活动总量（Total Early Stage Entre Activity）	-0.658**	-0.620**	-0.658**

数据来源：PISA 2010 和 2010 国际创业活动研究

考试分数学业成绩与情感发展不一定是同期的，也就是说，学业成绩完全可能妨碍情感的正常发展，也可以说情感发展可能妨碍学习成绩。二者不一定是同期同等发展的。那么教育评价就必须考虑二者的轻重，这就取决于决策者本身的价值观。也就是说一个教育体系应该对教育目标作出相应的判断，我们追求的是情感发展还是考试分数，是成绩重要还是信心和兴趣重要。

（三）短期与长期教育目标的矛盾

教育目标也可以分为长期和短期。在教学中，我们既希望学生可以很快掌握所学的知识和技能，也希望这些知识和技能可以留下来，为后面的学习和工作服务。也就是说，学习可以迁移（transfer），学到的知识和技能不是马上忘记而是长期有用。然而，这两种目标不一定是一致的，有可能冲突。有些教学方法有助于短期目标，让学生可以很快掌握要学的知识和技能，但有可能这种掌握是表面的、记忆性的，而没有经过深度学习，对长期目标没有太大的意义。

在美国教育界争论极大的直接教学法（Direct Instruction）就是一个很好的案例。直接教学法有两类。一类是指所有以教师为中心，严密安排教学内容和教师直接讲授的教学法。另一类是大写的直接教学法（Direct Instruction），特指一种专门的直接教学法，这种方法起源于 19 世纪 60 年代。其主要方法也是以教师为中心，对学生有严密的管理办法，教学内容很细，也是教师直接讲授。和直接教学法相对的是以学生为中心的探究式学习（inquiry-based learning）或

者是当前大为流行的项目教学法（project-based learning）。对这两种教学方法从上世纪 70 年代到今天已经有了成千上万的不同研究，就研究的结果而言就是发现了教学的短期和长期目标的矛盾（Zhao，2018d）。

卡内基梅隆大学心理学教授 David Klahr 和匹兹堡大学应用研究与评估中心主任 Milena Nigam 的一项研究可以作为佐证。他们的实验对比了使用直接教学法和探究式学习法的三年级和四年级的孩子学习变量控制策略的结果。他们发现，在学习后的第二天，测量的结果显示两种方法没有明显差异，也就是说两种教学法的效果是一致的。因此他们的结论是直接教学法优于探究式学习，因为效果一样而直接教学法花的时间远远少于探究式学习（Klahr & Nigam，2004）。但这个结论很快就受到了挑战。哥伦比亚大学心理学教授 David Dean Jr.和 Deanna Kuhn 做了一项与 Klahr 和 Nigam 几乎一模一样的研究，只是延长了时间。他们发现在第 11 周测评时，直接教学法的效果开始消失，在第 17 周时更是如此。使用直接教学法的学生在第 11 周和第 17 周时表现远远低于对照组。因此，他们认为完全不必要去追求直接教学法的短期结果，因为发现式学习最终会让学生达到短期的学习目标，而且其影响更加深远，学生学得牢，还能灵活运用，达到学习迁移的目标（Dean Jr & Kuhn，2007）。

Kapur 把直接教学法短期的成功归为“失败的成功”（unproductive successes）。这种成功就是短期教学目标得到极大满足，但对长期教学目标并没有太大的帮助。“可以让学生在记忆和执行解决问题的程序上表现很好，但学生对他们自己在干什么没有相应的理解”（Kapur，2016，p. 290）。Kapur 做了一系列实验验证他的假设。在一项实验中，他将学生分为两个组。一个组采用直接教学法，另一个采用探究式学习法。结果是直接教学法组的学生在开始时采用传统标准方法解决传统标准的问题，成绩优异，但到后面要创造性地解决问题时就大大落后了。

1979 年 Penelope Peterson 在综述了 200 多篇文章后得出的结论是：使用直接教学法或者传统教学法的学生在学业考试上有一些优势，但是他们在抽象思维考试，比如创造力和解决问题方面略差一些。与此相反，使用开放式教学法的学生学业考试成绩差一些，但创造力和解决问题能力要强一些。此外，开放式教学法在帮助学生树立对学校和老师的积极态度、培养学生的独立性和好奇心方面优于直接教学法。（Peterson，1979，p. 47）

短期和长期教育目标的不一致，甚至冲突给教育评价带来极大的挑战。因为如果注重短期评价，比如一堂课或者一学期的课，那么就有可能受短期目标的影响而忽略长期目标。既然某些教学方法对促进短期目标作用很大，但损害长期目标，那么这样的评价就有可能失之偏颇；而如果只看着长期目标，又有可能失去对短期目标的检测。因此，教育评价必须要有所侧重，对长短目标应有通盘考虑。

（四）创造力与短期学习目标的冲突

创造力和与其相关的好奇心是 21 世纪许多人士都认为特别重要的教育目标。但是，短期的教育目标，比如掌握一定的知识和技能，可能会极大地破坏创造力和好奇心的发展。有研究证明，创造力与学习成绩不成正相关，也就是说学习成绩的好坏对创造力本身没有影响。但加州大学伯

克利分校的心理学教授 Elizabeth Bonawitza 的研究表明，过多的教会伤害创造力和好奇心（Bonawitza et al.，2011）。

这项研究包括两个实验。在第一个实验中，任务是玩一个新玩具。研究者们把 85 个年龄 48-72 个月的孩子随机分为教学组、中断组、幼稚组和基础组 4 个小组。教学组的活动和直接教学法非常相似，实验员就像一位老师，她举着玩具，告诉孩子们：“大家看着，这是我的玩具，我要教你们我的玩具如何玩。大家看着。”然后，她给孩子们演示了如何玩这个玩具。中断组的活动和教学组基本一致，只不过实验员在演示一结束就离开了现场。幼稚组接近于探究式学习。实验员假装刚刚发现有这个玩具而且不知道如何玩这个玩具。她一边玩，一边对孩子们说：“你们看见没有，让我来试试。”她演示了玩具的玩法，但假装是偶然发现的。在基础组，实验员没有演示玩具的玩法，只是说了一句话让孩子们注意到了玩具：“你们看这个玩具，看看。”为了保证和其他组的时间一致，她花了 2 分钟看玩具，然后就把玩具放回桌子上。在所有组中，实验员在介绍结束后都鼓励孩子们玩这个玩具并找出这个玩具的原理，然后就让孩子们玩。研究小组用录像对所有小组孩子们的活动进行了记录并对比了各组玩玩具的总时长、独特行为的次数、花在重复实验员演示行为上的时长和发现玩具功能的次数。其数据表明：教限制了孩子们的探索和发现。那些被教过的“教学组”孩子更少地体验玩具的功能也更少地发现了玩具的其他功能。第二项实验证明了第一项实验的发现，而且发现小孩能够猜测教学目的。即，他们在没有明确教学的情况下及时猜测到老师要教而且他们就应该听老师的。另一项对儿童的研究的结果与此类似。研究发现当实验员直接给出指令的时候，幼儿园的孩子们更容易模仿实验员，而不愿意自己探索找出新奇的答案（Buchsbauma，Gopnika，Griffithsa，& Shaftob，2011）。

这两项研究表明如果追求知识的掌握，教学的确有用，但是过早掌握传统知识和解决问题的方法有可能让孩子们失去好奇心和创造力，让他们听话，让他们养成听成人和老师的指导和教学的习惯，而失去自己把握事物探寻世界的兴趣。那么在 21 世纪，创造力和好奇心成为重要教育目标的时代，教育评估是否需要跳出单纯评估知识掌握这个圈子而要更多地评估孩子的好奇心和创造力呢？

（五）认知与非认知的矛盾

前面提到学业成绩，比如 PISA 和 TIMSS 成绩，与学生的情感发展不一致。这其实涉及学生的生活情绪感受之外的另一个大问题，那就是认知能力（cognitive skills）与非认知能力（non-cognitive skills）的关系。一般说来，学业成绩考核的是一个人的认知能力，比如记忆力，按要求完成任务的能力，以及使用信息解决问题的能力。它关注的是一个人能否解决问题，但不关注非认知能力，也就是一个人是否想解决问题。

能干但不想干不敢干就涉及个体的非认知能力，涉及比如动机、个性、信心、坚韧、兴趣等许多与知识和技能没有直接关系的因素。非认知因素不是一个简单的概念也不是一项技能。它包含许多至今没有完全被所有个体接受的概念。总体来看，人们提的比较多的有坚韧性（grit）、成长性思维（growth mindset）、动机（motivation）、自我决定（self-determination）、自我控制（self-control）、情商（emotional intelligence）、社交能力（social

intelligence）、感恩（gratitude）、信心（confidence）、好奇心（curiosity）以及开放心态（open-mindedness）等。这些因素在现有的教育评估里面基本没有测试。但有关研究证明它们有助于个人社会情感发展，促进有目的的活动，帮助个体作出有意义的判断和决定，并与个人终身成就有极大关系（Brunello & Schlotter，2010；Duckworth，2015；Levin，2012）（Duckworth & Yeager，2015；Levin，2012）。这也是单纯的学业考试成绩不能预测个人或国家成功的主要原因（Baker，2007；Goleman，1995；Tienken，2008；Zhao，2016；Goleman，1995；Tienken，2008；Zhao，2016b）。

认知能力的发展和非认知能力的发展不一定完全一致。PISA 和 TIMSS 数据说明了这一点。专门注重培养学生的非认知能力或社会情感技能的教学并不能提升学生的学业成绩也是一个反证。近年来社会情感能力教学（social and emotional learning）虽然风行一时，也有不少研究试图证明其效果。然而结论就是除了能对相对应的社会情感能力有一定促进外，对学习成绩没有显著效果（Durlak，Weissberg，Dymnicki，Taylor，& Schellinger，2011；Hart，DiPerna，Lei，& Cheng，2020；Wigelsworth et al.，2016）。

在我们日常教学中常见有的学生考试成绩很好，但对学科本身并没有兴趣。还有的学生成绩不一定好，但对学科却有浓厚的兴趣。或者有学生学习成绩不好，但却有独到的思维方式，或者特别有人缘，善于处理人际关系等。总而言之，我们不能将认知与非认知混为一谈，而且要清晰地认识到他们之间可能的矛盾。而且，教育评估也必须同时考虑到认知能力与非认知能力评估。

四、多重教育目标的个性化

教育评价如何对待教育目标的多重性和不兼容性是两个大的问题。教育评价还有一大问题，就是学生的个体差异。由于先天和后天的相互作用，学生之间存在极大的个体差异，但是教育评价往往看重的是学生在某一个时间的表现，基本上忽略了学生先天的差异和与学校无关的因素，因此虽然测试的结果是学生先天、家庭、所处社区等因素加上学校学习共同作用的结果，但教育评价的结果往往被单纯地认为是学校教育的结果，从而误导对教育评价结果的理解。

（一）先天和后天的交互作用

人的发展是先天和后天交互作用的结果（Ridley，2003；Lewontin，2001；Lewontin，2001）。一个人天生具有不同的可能性，一般称为智力（intelligence）或者天赋（giftedness）。传统的心理学对人的天性测试就是智力测试，得出的结论以智商来表示。后来人们对智商或者智力产生了怀疑，因为人不可能只具备一种智力。哈佛大学的 Howard Gardner 在 20 世纪 80 年代提出了多元智能理论（multiple intelligence theory）（Gardner，1983）。耶鲁大学的 Robert Sternberg 也提出过类似于多元智能的理论并做了相关研究（Sternberg，1988）。到今天为止，多元智能理论基本为大家所接受。虽然对多元的具体内容以及多元智能的开发实证尚有争议，但是就人天生就具备不同的学习潜力这一点基本上达成一致了。有的人对光对颜色更敏感，那么如果后天机会得当，他们也许就有可能成为艺术家。有的人天生对数字或者数字模型敏感，如果有机会，他们就有可能在数学方面做出成绩。

除了多元智能以外，人天生的追求也不一样。前俄亥俄州立大学心理学教授 Steven Reiss 提出的多元追求理论指出，从整体来看，人类可能有 16 种追求，但每个人的追求是不一致的。比如，某人可能极度追求权力，想影响别人。对其来说，可能 80%的追求与权力有关，而其他的需求很小，也许就 20%；而对另一个人来说，其追求的是知识，是认识世界，是探索未知，那么他的追求也许 80%与好奇心有关。还有人特别追求家庭生活，另外的人可能看重安静的经历（Reiss，2004；Reiss，2000）。

此外，人的性格也有极大的差异，这些差异也与遗传有关。现代心理学所总结的人类五大性格类型就说明了人的性格的极大差异（John，Robins，& Pervin，2008）。这五大性格类型，也称为人格，包括开放性（openness）、责任心（conscientiousness）、外倾性（extraversion）、宜人性（agreeableness）和神经质性（neuroticism）。在这五个方面，每个人所具备的特征是不一样的。比如外倾性方面，有的人特别外向，而有的人特别内向。在开放性上也是如此，有的人特别具备想象力，特别追求新东西，而有的人则很守旧，不愿意参与新经历、新活动。

人先天遗传的智能追求以及性格与后天经历相互作用，从而形成不同的人。后天经历对先天因素产生的作用主要在于提供发展的机会或者压抑先天的发展。例如，一个具备音乐天赋的孩子如果没有听到过音乐或者经历音乐的训练是基本不可能成为一个音乐家的。人类历史上出现过的狼孩就是典型例子。他们都具备语言天赋，但出生后没有和其他人打交道，没有模仿使用语言的机会，因此也就不会使用语言了。后天经历也可以张扬一个人的天赋。比如一个有音乐天赋的人生在一个音乐世家，生下来就能接触音乐，而且还有父母给予的音乐熏陶，那么他就可能成为一个音乐家。

后天经历对人格和追求起的作用也十分相似，可以压抑或者扩展一个人先天的追求和人格。一个追求权力的人如果有机会体验权力带来的快感，而且有机会抓住权力，那么就有可能更加追求权力。相反，如果在其后的经历中没有享受权力带来的快感，这种追求就可能弱化。人格也是如此，一个先天开放的人，如果后天的经历让其享受到开放追求新事物的积极结果，那么他可能更开放。如果他生在一个保守的社会，那么其开放性有可能慢慢下降。

但是后天不可能彻底改变先天。人人都可以学会画画，但不是每个人都可以成为毕加索。先天对后天的学习有极大的影响。一个具备天赋的人学习速度和最后可以达到的水平，在同样的后天经历下会远远高于一个不具备天赋的人。如果花的时间一样、教师一样，而且经历一模一样的话，具备天赋的人可能成为一位卓越的专家，而不具备天赋的人就是一个平庸的掌握了相关知识和技能的人。

由于先天和后天的交互作用，学生之间既有先天差异，后天的经历也大不一样，那么他们在进入学校时就已经有了很大差异。每个孩子都有自己的优势和弱势，有自己的强项和弱项，有自己的追求，有自己的性格。这些差异从其本身来看无所谓好坏，只是差异，不是差别。但是，一旦进入教育系统这些差异就会变为差别。因为每个教育系统都有自己的教育目标，而这些教育目标又是学校开展教学的最基本要求，也是教育评价最根本的出发点。

（二）个性化与教育评价

在教育目标的指导下，教育评价就是对学生的差异性做出价值 上的判断。符合教育目标的能力、知识、行为、性格、兴趣就是有价值的，有用的。反之，不符合教育目标的就是没有价值的，没用的。而且教育目标直接决定学校开设的课程和看重的行为。因此学生的差异极大地影响每个孩子与教师、课程和课堂的关系。教育进入学校时，有的孩子由于先天加后天的原因对要学的内容以及课堂环境和老师都很熟悉，而且已经具备了一定的能力，他们就是特别优秀的孩子；而另外一些孩子的天赋和家庭经历让他们对学校的一切都很陌生，其能力可能与学校要教的课程没有关系，因此看起来特别差。当然还有一些孩子可能居中，不太差也不太好。

但是需要注意的是那些很差的或者中等的孩子，其实是很有自己的长处和追求的。只不过他们的长处、能干的事情和其喜欢的事情不是教育评价当时所关注的，而教育评价所测评的又正好不是他们的长处。许多学生因为这一点而失去了真正发展自己的机会，被迫放弃自己的优势，转而接受学校所教的课程和所看重的知识和行为。然而，他们的优势并非学校教育所要求的，这些学生有可能永远落后于那些恰好具备符合教育目标的能力和兴趣的学生而成为社会的底层。

教育评价是为检测教育目标实现的程度而服务的。如果教育目标改变了，教育评价可以改变。但教育目标是否有可能改变呢？其实教育目标应该而且确实在改变。不同的社会需要有不同的教育目标。人类社会进入第四次工业革命时代，我们的教育目标应该扩大，扩大到可以包容不同的人才，让每个人的天赋和兴趣都得到承认和发挥。这是教育目标制定者和教育评价需要考虑的。

（三）学习与教育评价

教育评价要考核的是教育的作用和效果，因此评价的是学生在评价时刻的状态，而且要从这个状态中找出教育教学的作用，要看到学生的进步。然而从学生的先天与后天的交互作用来看，任何时候测评到的学生状态必然是在以前的基础上成长起来的，也就是有先天的作用和后天的努力。那么教育评价就必须想办法找到从某一个时间点到另一个时间点的教育教学的作用。增值性教育评价（value-added education measurement 或者 value-added assessment）是方法之一。增值性教育评价本来是一种对教师的评价，其方法是用现有学生的成绩对比以往学生的成绩以及同一年级学生的成绩，从中找出差异，这差异就可认定为教师的贡献及增值。增值性教育评价在美国以及其他国家的教育评价中获得了不少支持，因为它似乎可以更公平地展现学生在一段时间内的成长而且可以与其他群组进行对比，也可以消除过去其他因素的影响。但是，研究证明增值性教育评价本身的效果与想象的效果差别很大，并没有想象的公平和准确。2019 年美国国家经济研究署（National Bureau of Economic Research）发表的一篇论文说明了这一点（Bitler，Corcoran，Domina，& Penner，2019）。该研究的主要目的是验证增值性评价的效度。研究者们采用常用的增值评价模型来预测一个教师不可能改变的结果——学生身高。他们分析了纽约市学生的数据后发现教师对学生身高的影响达到了对数学和语文成绩的影响程度，从而引起了对增值评价效度的怀疑。他们发现影响学生成绩的因素远远超越教师这一单一因素。

增值性教育评价也可以是评价个体学生在某一个时间段以内的成长情况。比如我们可以测评学生三年级开始到三年级结束之间的成长情况，并从中看到一个学生的进步。这个进步可以归结

于学生的努力、教师的工作、课程的作用以及学校和家长的共同作用。但是问题在于，我们如何看待他一年级之前的情况。一年之间的学习似乎与前面的情况没关系，但其实关系很大。首先，先天因素可以大大影响后天的学习速度。在学科上有天赋的孩子学习速度可以远远高于在该学科不具天赋的孩子。其次，入学前学生已经养成的习惯和已经掌握的知识可以极大地影响后面的学习。具备一定知识和有关习惯的孩子学习可能是“加速度”而不具备这些条件的孩子可能需要更多的时间。再次，学生的天赋、已有知识技能和习惯可能与这一年中的学习经历产生巨大冲突或者特别适合，因此产生的学习效果会大大不一样。

因此要评价教育教学对学生的影响、教师对学生的影响、学校对学生的影响以及课程对学生的影响就极其困难。尤其是当想用单一简单的评价工具评价学生的成长和学习时，这个问题就更明显。因此在实施教育评价时，我们必须考虑到学生的天赋、兴趣、人格、追求及其后天所有的环境。更重要的是，开展教育评价之后，我们更需要关注学生个性化教育，帮助每一个孩子顺应他们的天性而学习。

五、未来教育评价的方向

教育评价的目的很多。大规模教育评价可以统一监测一个国家或地区（一个教育体系）的教育状况，可以看到与现有教育目标的差距，也可用于观察一个教育体系的运作，还可以用于选拔学生，考察学校和教师等。小型的教育评价可以用于测评学生对知识和技能的掌握，可以用于分班、制定教学计划、检测学生的进步和成长，也可以用于对教师和学校的考察。但从宏观上看，教育评价的目的有两大类：选拔和促进学习。

（一）选拔

教育评价的选拔目的非常明确。任何一个教育体系都有一套用于选拔的评价工具，因为每一个教育体系在提供教育之外的主要功能就是为社会选拔其所需要的人才。任何一个社会都需要人才，而且任何一个社会都不可能满足所有人的需求，因此必须要有一个选拔工具让一些学生享有更好的社会资源和机会。这些资源和机会在不同的社会以不同的方式出现，往往是以更高的就学机会和相应更好的工作为代表。所以，利用教育评价为高中、大学以及研究生等教育选拔学生是各国都有的现象。

选拔性教育评价基本是常模参照性测试（norm referenced test），也就是必须进行比较。它不在乎某个学生是否掌握了要求的知识和技能，而是对比学生考分的高低。因为任何资源和机会都是有限的，无论怎样的评价，最终都是将学生进行对比，选出和资源及机会相匹配的人。

以选拔为目的的教育评价还必须对评价标准和评价内容有一定的限制。目的不同，选拔标准就不同。比如美国的大学升学选拔以统考（SAT 和 ACT）分数、学生学业成绩（GPA）、学生其他表现（推荐信）以及学生的经历和特长（自我描述）等为主要内容。SAT 和 ACT 考试提供的是学生的排名，也就是相对于其他考生的位置。GPA 提供的是学生高中阶段的成绩和相对于其他学生的排名。推荐信和自我描述提供的是该生的独特性，以反映考试以及 GPA 不能提供的情况。

然而，美国的“高考”其实问题很大。经过大量研究发现，SAT 和 ACT 并不能预测学生进入大学后的学习情况。加州大学系统追踪了该校从 1996 到 1999 年入学的 78000 名学生后发现 SAT 分

数只能解释大一学生成绩的 13.3%（Geiser & Studley，2001）。大学理事会（College Board），就是 SAT 的提供者，也对 SAT 以及大学成绩做了大量的分析。其中一项是对 1980 到 2000 年诸多研究的元分析，结果发现最好的结果是 SAT 可以解释 10%的毕业情况差异，也就是说一个人能否大学毕业只有 10%的原因与 SAT 分数有关（Kobrin，Patterson，Shaw，Mattern，& Barbuti，2008）。ACT 成绩对大学生学习的影响也差不多。美国国家经济研究署的一项研究发现，ACT 能解释的大一成绩的差异是 13%（Bettinger，Evans，& Pope，2011）。由此可见，“高考”分数对大学生学业的影响其实很低。因此，美国大学开始摒弃 SAT 和 ACT 的做法。2014 年 Bates 学院对来自 30 多所把 SAT 或 ACT 成绩作为选项，也就是对学生可以不提供成绩的不同类型大学的 123000 学生的大学成绩分析表明，没提供考试成绩的学生大学成绩不低于提供了成绩的学生（Hiss & Franks，2014）。由于 SAT 和 ACT 不能预测大学学习，也就是它们并没有测试到大学生学习所需要的能力，美国已经有近一半的大学开始不要求标准化考试成绩（FairTest，2018）。高中学业成绩对大学的影响，单独来看和 SAT 以及 ACT 差不了太多，只能解释大学成绩的极小一部分差异。

由于传统升学评价的缺陷，美国中学开始了寻找更科学的评价范式。从 2014 年的一所学校开始到今天的 300 多所学校参与的 Mastery Transcript Consortium（MTC）就是案例之一（Mastery Transcript Consortium，2020b）。MTC 是一个以中学为成员的非盈利组织，一开始由美国有名有地位的私立中学组成，后来公立学校开始参与。MTC 的主要目的是创立一种新型的学业评价模式，供大学招生使用。与传统的学业评价只给一个学业综合成绩（GPA）不同，这个模式的最大特点就是用电子化手段全面展现每一个学生的学业情况和自身优势与弱势，并提供相应的证据证明学生达标的情况以及个体优势，从而让大学清楚地看到每个学生的学习情况和所掌握的知识与技能，明确每个学生的强弱点。MTC 一个虚拟的学生学业成绩报告单如图 1 所示（Mastery Transcript Consortium，2020a）。

MTC 已经开始给大学提供类似的成绩单，而且效果也不错。该组织发展迅速，2017 年 4 月正式发布时参加学校是 50 家，到 2020 年已经超过 300 家。新冠疫情以后，叠加美国大学对 SAT 和 ACT 的放弃，SAT 和 ACT 考试都受到了极大冲击，越来越多的学校不得不至少暂时放下对 SAT 和 ACT 的要求。可以说美国学生选拔性评价已经开始发生巨大变化，而且这种变化会越来越大，其特点如下。

第一是个性化。随着对学生的能力、追求、人格的个性化认识加深，对先天和后天交互作用的认识加深，对学生个体发展的不一致性的认识加深，对现有教育评价难以完整全面地测评学生的个性发展的认识加深，人们开始寻求个性化的教育评价。而且个性化的教育评价似乎也更适合像升学之类的选拔性评价，因为大学学习经历本身就是多样化的。首先大学层次本身就多样化。一个大学体系包含研究型、教学型、实践型等，而且还有 4 年、3 年等不同的学制。一个研究型大学对于要招的学生所具备的能力知识要求可能完全有别于一所实践型大学。其次，大学的专业千差万别，所要求的知识能力和兴趣也因此千差万别。可以想象戏剧专业应该和物理专业对学生有不同的要求。同理，数学专业对学生的要求应该不同于文学专业。因此不可能有某一个单一的标准可以为所有学校所有专业挑选人才。应该采用的办法就是使用能够展现学生个性的评价手段

学校和专业寻找适合的人才。

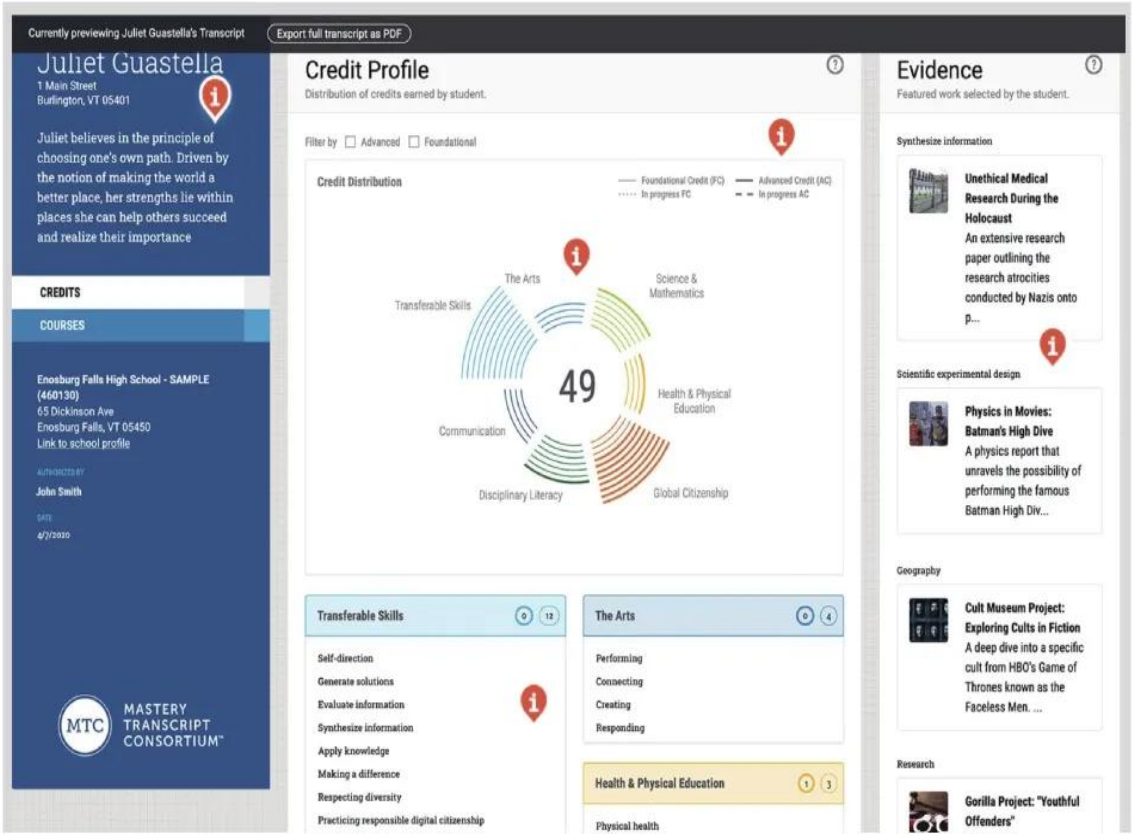

图 1　　MTC 学生成绩单

第二个变化是能力化，也就是摆脱学科本身的知识，而寻求知识之上建立的能力（competence）。比如 MTC 考核的科学与数学能力包括统计推理、代数推理、几何推理、科学实验推理、科学解释等。它的通用能力包括自我管理、生产解决方案、评估信息、综合信息、运用知识等。MTC 还含有国际公民等能力。知识的能力化可以摆脱传统学科的约束，评价的不是掌握某一学科的知识，而是掌握这些知识后能够达到的能力，而这些能力可以通过活动和跨学科的学习来培养。

第三个变化是真实性、作品性。新的教育评价越来越看重的是学生做过什么，有什么作品，而慢慢疏远于学过什么。MTC 收集提供学生的经历和作品用以支撑他们所达到的能力，而不仅仅是他们学过什么课程。知识的价值越来越小，人们看重的是一个人能否运用知识解决问题，是否具备超越知识的能力和睿智。

第四个变化是顺应社会大变革。世界教育面临巨大变革，教育必须适应第四次工业革命带来的诸多变化。其中高等教育的变化可能会导致选拔性教育评价的变化。现有的为高等教育选拔人才的模式是假定高等教育机会是有限的，而且还可能帮助、训练选拔更高职位的人才。但是社会的发展已经导致高等机会大众化，高等教育文凭已经贬值，而且未来社会一个人可以不上大学而获取高等教育所提供的教育。那么高等教育选拔性的教育评价就不应该延续现有的模式，不应该以排斥和选拔为主，而应该以最好地反映学生个性化的能力和经历为主，以帮助学生学习为主。

（二）促进学习

教育评价的另外一个重要目的是促进学生学习。对教师、教材、教学方法、学校以及学校系统的评价其实都是为学生学习服务的。当然，学生学习的内容和目标在不同国家和地区是有不同要求和规定的。不同国家和地区对学生的要求和所学内容以及要达到的目标有各自的规定，其开展教育的方式、对学校和教师的管理办法、对课程的管理和实施等都有各自的特点。因此各地对如何促进学生学习有自己的看法和做法。然而，学习过程有一定的规律，使用教育评价促进学生学习应该遵循相应的规律。这些规律是人类学习过程中不可规避的，是适用于所有学生的，是在开展教育评价时必须要考虑的，在此我们做一些总结。

一是教育目标的整体性。教育无论如何都不是为单一目标服务的。任何一个教师、学校或者学校系统都不可能也不应该只教会学生一样东西或者只有一个培养目标。德智体全面发展，认知能力和非认知能力同时培养，以及知识和技能、道德与身心健康等的共同成长都是教育的目的，都应该评价，都需要关注。因此任何教育评价都必须看重教育目标的整体性，而不能只考核某一个目标的完成情况。

二是多重教育目标的矛盾性。教育的多重目标有可能是相互矛盾的，比如学科之间的时间竞争、长期教育性目标与短期教学性目标的矛盾，以及认知能力与非认知能力发展的不同步等。这些矛盾要求教育评价必须全面考核学生的学习，而且有时候必须放弃对短期目标的追求，或者减少对短期目标和认知能力目标的考核，以便学生有时间、有精力达到教育性的长期目标。这对教育评价是一个极大的挑战，因为目前大多数学校开展的评价都是短期的，看重的是一星期、半学期或一学期学生对知识的掌握，而不是学生能力的成长。

三是先天与后天。现有教育评价极少考虑到学生的天赋、追求与性格，但是要促进学生学习我们必须要有所考虑，因为学生先天因素对其学习有极大影响。促进学习是为了促进每个孩子的学习，而不是只促进大部分孩子的学习，或者只追求平均分的提升。因此促进学习的教育评价必须要促进每个孩子的学习生态，考核到每个孩子是否获得其所需要的教学。

考虑到这些因素，目前促进学习教育评价的走向基本上是朝“学生个人学习档案”（learner profile）方向发展（Zhao，2018a，2018b）。个人学习档案可以多重定义多种做法。它可以包含个性化学习内容和规定的学习内容。个人学习档案收集学生的主要学习历程、学习成果。学习成果一般以作品为主，用来反映学生学习的进步。除了成果，学生档案还记录成果形成的过程、所学到的内容，以及反思和未来发展等内容。也可以包含学生、教师以及家长等共同商定的学习计划和完成情况等。个人学习档案是一个十分个性化的教育评价方案。它可以从多方面评价学生的进步，能够比较完整地长期考察一个学生的学习进展，也可以照顾到教育目标之间关系，让教师、家长等看到学生的短期和长期学习情况，从学生作品的进步上分析学生的学习情况以及需求。个人学习档案充分考虑到学生的先天因素，因此对每个学生的学习都要记录学生的兴趣、追求、优势等。个人学习档案对学生的评价是形成性的（formative）而不是总结性的（summative）。它可以像 MTC 提供的档案一样包含一些考试性的内容，但它的关键在于长期性，可以记录学生从入校到毕业的所有学校经历。更重要的是，个人学习档案可用于长期分析和观察学生的发展路径及

其在某一个时期所需要的学习经历，可以帮助老师和学生分析学习中存在的问题并探讨解决方法，从而让学生看到自己的进步、不足和需求。

六、总结

本文探讨教育评价中的几大问题。这些问题体现了教育本身的复杂性，同时也指出教育评价面临的挑战。教育目标的多样性极其不兼容，教育目标的整体性和多样性是教育目标本身的特点，然而教育评价一般不会注意到这两个问题，因为教育评价的执行者和实施人很多时候关注的是评价工具和分析结果。但是，为了达到教育评价为教育服务，为学生和学校服务，以及为提升教育系统质量服务的目的，教育评价必须要系统地考虑全面性和整体性。学生的个体发展日趋重要。而学生的先天和后天交互作用是学习的本质。这种交互作用的结果就是学生的独特性和个性化。教育评价如何体现并促进学生的个性化，促进教育的个性化也是一个重点。

作者简介：赵勇，博士，美国教育科学院院士，国际教育科学院院士。美国堪萨斯大学教育学院杰出教授兼创新创业教育中心共同主任，澳大利亚墨尔本大学教育学院教育领导力教授，中国华东师范大学全球讲席教授。　曾任美国密歇根州立大学教育学院杰出教授，俄勒冈大学教育学院副院长，全球在线教育中心主任，校长杰出教授，英国巴斯大学全球讲席教授，澳大利亚维多利亚大学米切尔教育与公共卫生政策研究所特聘研究员。

文章来源：华东师范大学学报教育科学版，作者赵勇。