

论教育评价的治理功能及其自反性立场^{*}

周作宇

(北京师范大学教育学部, 北京 100875)

摘 要: 从主体间性的角度看教育评价, 教育评价是泛在的价值操作过程, 也是意义建构的过程。在教育与评价之间, 存在正式和非正式的教育与正式和非正式的评价之间的组合模式。正式教育的正式评价起着专业化发展的引领作用。教育评价在特定场域展开, 受场域的影响。教育评价具有治理的功能, 是治理的工具。从个体到组织、国家乃至国际, 治理的层级不同, 评价的内容也不同。善治既是对国家治理的评价, 也具有教育治理的指导价值。评价具有导向性作用, 评价也是治理的对象。从评价治理的角度上说, 教育评价要持自反性的立场, 对评价保持再评价的开放思维。一方面增强元评价意识; 另一方面, 还要不断对现实问题的场域特征保持警觉, 警惕教育评价陷入行为主义、经济主义、地方主义和机会主义的陷阱。

关键词: 评价; 元评价; 教育评价; 善治; 教育治理

在治理体系与治理能力现代化成为制度建设的焦点任务之后, 治理现代化迅速起到话语引导的格式化作用。作为意向性语词, “现代化”内含“价值”意蕴, 具有融目标于过程的统摄力量^①。治理体系也好, 治理能力也罢, 以现代化为鹄的, 都有不断进步的过程性标度。对体系和能力水平的监测, 需要发挥评价的作用。合理的评价是现代化的工具, 与专业主义和透明度等价值有关(Dahler-Larsen & Boodhoo, 2019)。通过评价促进治理现代化, 进而实现人类社会的更大目的, 越来越为国际社会所关注。有的组织就将善治与评价结合, 提出通过政策评价改善治理(Schlagenhauf, 2020)。反过来看, 作为普遍存在的, 具有显在和潜在社会意义的活动, 评价也是治理的对象。在教育领域, 我国颁布了专门针对教育评价改革的文件, 将评价作为治理的重要内容, 凸显了评价所具有的利害攸关的政策地位(中共中央, 国务院, 2020)。一般而言, 在世界范围内看, 治理与评价之间存在着密切关系已受到不同程度的关注和重视。虽然如此, 在整体和局部、意图和行动、知识和操作等方面, 还有很大的理论探讨空间。本文从主体间性的角度挖掘教育评价的意义, 在对教育评价和教育治理的关系梳理的基础上, 阐述教育评价的治理功能。

一、教育评价: 意义建构、主体间性与场域影响

(一) 评价是泛在的行动主体进行价值操作的活动

评价是人类日常生活中普遍存在的与价值相关的活动。受笛卡尔影响, “主客二分”长期盘踞现代哲学话语体系。二分法具有一定的解释力, 但是也存在局限。为了克服二分法的局限, 马克思倡导实践哲学。古典哲学停留在认识世界的层面, 实践哲学则指向改造世界。杜威以“经验”统领其实用主义思想, 为其“哲学的改造”张目。沿着实践哲学和实用哲学的路径发展, 行动科学登场。行动学习与

^{*} 基金项目: 教育部哲学社会科学研究重大课题攻关项目“现代大学治理结构中的纪律建设、德治礼序和权力配置协调机制研究”(15JZD044)。

行动研究、行动科学具有相同的理论意境。从学习和改变的角度看,行动科学和行动研究也是行动学习。凸显主体能动性的实践(马克思意义上的实践)、经验(杜威意义上的经验)和行动(Argyris等人的行动科学意义上的行动),具有旨趣和内涵上的通约性。价值操作是实践、经验和行动的交集。操作包括无有操作($0 \rightarrow 1$, 或 $1 \rightarrow 0$)、增减操作(同质: $X \rightarrow nX (n > 0, n \neq 1)$)和改革操作(变质: $X \rightarrow X^y, y \neq 1$, 此处记号 y 没有幂指数的意义,只是表达 X 的质)。由认识到价值操作(实践、经验、行动)有一个中介环节——评价。评价是人类的基本行为(Worthen et al., 1997)。“所有行为举止,只要不是盲目地仅凭情感冲动行事,或只是机械地例行公事的话,似乎都包含评价”(杜威, 2007, 第5页)。

广义评价是日常性的、泛在的人类内隐或外显行为。无论是自我发动还是他者发动,指向自我还是指向他者,无论是有意识的还是无意识的、专业的还是业余的、清晰的还是模糊的、直接的还是间接的,从交谈中的语词选择到作文中的遣词造句,从出行路径的考量到交通工具的选择,从对刺激的感受对刺激的反应,种种活动都和评价具有这样那样的联系。质言之,评价无处不在、无时不有。狭义上讲,评价是关于“对象”价值的活动,是“主体”根据一定“目的”,通过特定“方法”收集关于某“对象”在某“价值领域”或“价值向量”的“数据”,根据特定指标、标准或“参照系”对数据进行分析,以识别对象的价值、品质、质量、效用、效率和重要性的过程。评价需要就对象的品质优劣做出评判(Worthen et al., 1997, pp. 5-23)。^②一般而言,评价对象包括政策、规则、战略、计划、流程、项目、干预、服务、组织、程序、事件、技术和材料等内容。当然,人也是评价的对象。评价有事前预判、事中伴随和事后反思三种视角。评价的目的是获取有关评价对象的知识,利用这些知识进行问责、决策、改进和管理。激发利益相关人持续学习也是评价的目的之一。评价既是过程,也是结果。^③

评价活动是人与世界关系的一种特殊反映。对于人与世界的关系向来有不同的总结概括,如概括为认识与被认识、改造与被改造、利用与被利用等类型。认识的目的一方面是揭示世界规律,另一方面是在世界和人之间建立意义关联,建构世界对人的意义,以及人对世界的意义。人类思想史上“天人感应”“天人合一”“人是万物的尺度”“我思故我在”“征服自然和改造自然”等表述,就是对这种意义关系的表达。评价具有认识和实践双重属性。评价首先是一种认识活动。它“以把握世界的意义或价值”为目的,揭示“世界对于人意味着什么,世界对人有什么意义”(冯平, 1995, 第30页)。意义建构使认识和实践得以联通。评价不仅仅是认识活动,同时也是一种实践活动。它根据人与世界关系的意义锚定和价值计算,实施价值操作,指导和支撑实践行动。评价为实践提供价值引领和行动动力。所以,人与世界关系类别中,还有“评价与被评价”的关系。

(二) 教育评价是在教育场域进行意义建构的过程

教育评价是评价的一种,是在教育领域中实施的评价活动。教育场域的意义建构具有特殊性。如何为教育领域划界,影响对教育评价对象的设定。就正式教育而言,教育的形式化最小单位为 $E = (T \cdot S \cdot C)$ 。其中, E 代表教育, T 代表教师, S 代表学生, C 代表课程内容, $()$ 代表三者组合在一起构成教育单元的特定场域。场域是个体的生活空间,它以个体为中心,具有不同的层级和半径,是层层相套彼此渗透相互影响的。构成教育单位的三要素,都是具有一定价值方向的矢量,或者说是意向性的变量。差别在于,教师和学生是动态的变量,而内容在一定时期是相对稳定的。虽然如此,对内容的理解、解释和说明,在一定程度上受人的影响。比如,同样是莎士比亚的《哈姆雷特》,在不同读者心中具有不同的寓意。即使数学这样的高度抽象和形式化的学科领域,对于其中同样的形式化表达,人们的理解也会有所不同。^④

因此,确切地说,教育的最小组织单位是 $\bar{E} = (\bar{T} \cdot \bar{S} \cdot \bar{C})$ 。教育是沿着教育目标方向,包含三要素在特定场域环境下相互作用,在时间区间 $(t_0 \rightarrow t_1)$ 运动变化的过程。在这个过程中,教育的变化量记为: $\bar{E}_c = \Delta \bar{E} = \bar{E}_1 - \bar{E}_0$ 。设 \bar{E}_i 为理想的教育状态, \bar{E}_r 为现实的教育状态, $\bar{E}_i = \bar{E}_r$ 的情况几乎是找不到的。显而易见,现实的教育矢量方向并非由单一要素决定。即 $\bar{E} \neq (\bar{T}); \bar{E} \neq (\bar{S}); \bar{E} \neq (\bar{C})$ 。教师和学生都是具有自由

意志的独立个体。他们的价值观始终处在变化发展的过程中,表现为价值行动。课程内容所具有的价值意蕴,不完全由教材决定。只有进入教师和学生的认知系统,教材内容才有实质的教育意义。在最为基本的教育目标上看,教师和学生共同指向课程目标。教育目标由课程目标构成。课程目标又需以课程内容为支撑。在教育目标、课程目标、课程内容之间,理论上具有逻辑一致。但这样的逻辑一致,在现实中并不完全成立。教师的言传身教在传统教育理论中占有重要地位。事实上,学生不是容器,不是被动的接纳器,教育目标的实现是师生交互作用的结果。这个过程不是封闭的,而是开放的。教师和学生的意义建构(sense making)是课程抵达教育现场的重要过程,也影响最后的价值实现。个体是在和他者的对话过程中建构自我的。在意义建构过程中,三要素有形式上的目标排列差别。以教师为主体,有教师通过内容抵达学生 $T_c^s(T \rightarrow C \rightarrow S)$ 和教师通过学生抵达内容 $T_s^c(T \rightarrow S \rightarrow C)$ 两种类型。前者以学生为目标,后者以内容为目标。本来,教育的目的是促使学生发展,而内容是中介。所以,就教育目的而言, $T_s^c(T \rightarrow S \rightarrow C)$ 存在“目标置换”。即,将学生发展性目标替换为工具性目标。在现实教育情境下,这种情况是普遍存在的。以学生为主体,学生是自身发展的责任主体。不论是通过教师还是内容 $[S_c^c(S \rightarrow T \rightarrow C \rightarrow S), S_c^t(S \rightarrow C \rightarrow T \rightarrow S)]$,最终要回到自身的发展上来。

主体间性理论认为,儿童之能够成为独立的主体,仅当他们的照护人认可其具有自己独立的心智、意志和感情。反过来,儿童之能够成为独立的主体,仅当他们能够认可其照护人也是独立的主体。主体间性就是由相互认可对方的独立主体性并努力理解对方的心智而建构的。主体性的形成是对话的过程,也是对话的结果。一方面坚持独立意志,另一方面以被人认可为前提。如果不能被认可,个人的主体性、自主性和自我意识就可能失去意义,缺乏确立的基础。意识到自己具有独立意志的瞬间,恰恰是依靠别人而获得的。没有主体性就没有主体间性。没有他者就没有自我(Auerbach & Blatt, 2001)。相对于学生发展而言,教师和内容都是工具性的。学生不承担教师发展的责任。学生进入教育现场的目的不是为了教师的发展。尽管教师发展对学生发展具有重要意义,但这不是学生的义务。在教师和学生主体性问题上,向来有不同的理论主张。所谓的教师中心论和学生中心论,认为教育过程中总有一个主体处于中心位置,或者教师是中心,或者学生是中心。“双主体”理论认为教师是教的主体,学生是学的主体。教师是教学生(向学生教),还是教内容?学生是学教师(向教师学),还是学内容?在潜意识里,这样的论断将学生看作是变量,而把内容和教师作为常量。事实上,教师也是变量。在师生互动过程中,变化的不仅仅是学生,教师也在变化。教师和学生都有其最近发展区。笼统地说,在知识储备上,教师的知识存量(K_t)比学生的知识存量(K_s)应该更大一些,即, $K_t > K_s$ 。但是,一旦涉及具体知识,问题就复杂起来。舒尔曼通过分析美国 1875 年加州小学教师考试和 20 世纪 80 年代的教师测试标准发现,经过一个世纪,内容方面的要求被忽视了(Shulman, 1986)^⑤。19 世纪下半叶教师的胜任力判据是内容知识。一个世纪后,教师测试的重点放在了教学论中如何教的问题上。根据他牵头开展的“教师知识增长”项目的研究成果,他提出了学科内容知识、教学论内容知识和课程知识三种不同的知识类型。进一步细分,他还提出了命题性知识、案例知识和策略性知识。命题性知识再细分,又可分为原理、箴言和规范(Shulman, 1986)。^⑥

其中,策略性知识具有超越原理性知识而达致实践智慧的特征。将命题、案例和规范融合起来形成超越性策略,是对教师的高阶能力要求。在规则不确定的具体情境下,和手艺人比较,专业人员不仅仅具备“怎么做”的技能型知识,还“知其然”并且“知其所以然”。导向自我知识的反思和元认知能力,以及将个人理解与他人分享沟通的能力,是教师专业化的重要特征。舒尔曼针对教师的胜任力和教师发展问题所推出的教师知识结构是有指导意义的。但是,教师的专业性不同于其他职业,与将物作为对象的职业不同,与其他将人作为对象的职业(比如医生)也有很大不同。教师所面对的是以发展为目的、在发展中并且经受发展评价的个体。在意义的世界“生成意义”是发展的要义。人具有目的和工具双重属性。“人是目的”就是对人的“目的性”的强调。教育理论中对“人力教育”的批评,不是看不

到人的工具性(利用与被利用属性)。事实是,从劳动分工看,居于一定职业位置的人,在劳动力市场和劳动结构中,无疑凸显的是其工具性。社会分工、职业差别、经济收入、物质生活无疑对个人的意义建构具有重大影响。但是,生存需要具有不同的层级和追求,人的发展远远超出了技术性维度。在教育现场,教师是职业性的,而学生(传统意义上的学生)则是前职业性的。对学生而言,目的是现实的,而工具则是潜在的。教师的意义建构是和教师职业紧密联系的。而学生的意义建构则既和当下个人目的性“价值感”(以人为目的的价值感)密切关联,也和未来的职业性工具价值关联。

意义建构是一个综合的动力学过程。“尊重感”和“认可感”既是教师意义建构的驱动力,也是学生意义建构的驱动力。“尊重感”和“认可感”不同于“成就感”。成就感源自个人对计划目标实现的体认,当然也受外部评价的刺激。但是尊重感和认可感是处在“身份”与“成就”的“无知之幕”下,对人之为人的基本尊严的尊重和认可。因此,从根本上说,尊重感和认可感是关系性的。在以发展为目标的教育单位,人与人之间的相互尊重应该是无条件的。当然,相对于学生来说,教师因为职业赋予的“权力”,在这种关系中占有主动地位。将教师作为观察对象,在最小教育组织形式 $[E = (\bar{T} \cdot \bar{S} \cdot \bar{C})]$ 之外,还有以班级组、年级组、学科组为一级,以及以学校、地区、国家为单位的更大的组织形式。以单位容量为参照,乍看起来,距离教师越近,对教师的意义建构影响越直接,影响也越大。但从大图观察,近距离“影响域”还受更大的意义空间的影响。不同层级对包括注意力在内的资源分配,事实上起着重要的作用。这就是场域的影响。从意义建构的角度观察教育评价,旨在建立广义的评价观,对日常师生互动中的评价行为及其影响予以关注,使教育评价从内容(C)的世界走向人(T, S)的世界。评价不仅仅是意义建构的工具,评价本身就是人的存在形式。

(三) 教育评价受社会场域的影响

教育不是在与世隔绝的真空中进行的。教育评价不能不受社会场域的影响。根据柯特·勒温的观点,个人行为(B)是个性(P)和环境(E)的函数。即, $B = (P \cdot E)$ 。教师的行为为 $B_t = (P_t \cdot E_t)$,学生的行为为 $B_s = (P_s \cdot E_s)$ 。自我与他者互为外在的环境因素。教师是学生外在的环境因素。学生也是教师的外在环境因素。与其他外部环境因素比较起来,人的因素具有主体意义建构的性质。感知、反应、理解、解释、施动等要素,是个体行为的基本构成。交互性是人际沟通的一般形式。师生之间的影响是相互的。个体的教育行为(B_e)是个体行为的一个子集。教育行为包含于个体行为,教育行为属于个体行为。即, $B_e \subseteq B$, 或 $B_e \in B$ 。如此细分意味着,个体的有些行为是教育行为,有些属于非教育行为。教师和学生都是以整体进入教育单位的。很难在教育行为和非教育行为之间划一条泾渭分明的界线。区别教育行为与非教育行为的意义在于,为个体留出一定的“自为空间”,即,不受教育责任和义务约束的任由个人自由支配的行为集合。

教师的自为空间和学生的自为空间有所不同。教师是一种社会职业,而学生不是真正意义上的职业。对教师来说,“学为人师,行为世范”是一种理想召唤,也是社会角色的公共期待。而“天高任鸟飞,海阔凭鱼跃”,则是对学生面向未来的无限可能空间的写照。教育的交互性表明,不能孤立地、割裂地看待教师和学生各自的行为,而应将其置于交互关系之中。一般而言,教育行为(B_E)首先是指向特定教育目标的教师行为和学生行为,即, $B_E = (B_t \cup B_s)$ 。但教师行为和学生行为的并集仅是构成教育行为的必要条件。只有师生相互作用,教育行为才真正发生。由交互性界定,交互性的教育行为(B_{Ei})乃是教师行为和学生行为的“交集”,记为: $B_{Ei} = (B_t \cap B_s)$ 。所谓交集,并不是简单的人数集合,而是指在“主体间性”(subjectivity)基础上教育目标的动态实现过程。通常认为,学生的发展是直接的教育目的。通过学生参与社会活动而使社会得到发展,是间接的教育目的。问题是,除了学生发展外,在教育行为中,教师有没有自己的需要和目的?教育过程是意义建构过程,不仅仅是学生在进行意义建构,教师也在进行意义建构。教育就是在意义建构过程中形成的关系。在认知、情感、体能方面的依附、独立、相互依存是对关系的一种划分。

我国传统上有丰富的论及“我”与“他者”相互关系的价值命题。比如，“仁者爱人”“己欲立而立人，己欲达而达人”“己所不欲，勿施于人”“古之君子，其责己也重以周，其待人也轻以约”“三人行必有吾师；吾日三省吾身”“行有不得皆反求诸己”，如此等等，不一而足。无论如何，明确教育是关系性的存在，是师生意义建构的过程，为讨论评价问题提供新参照。在认识者被认识、改造者被改造、利用者被利用之外，存在着的教育者被教育、评价者被评价的关系，反映出了教育和评价所具有的自反性特征。师生意义建构是在特定的场域中进行的。从场域中抽离出来，不受环境影响的教育和评价是没有的。场域是由权力结构、惯习、传统、制度、法律等个体外在的秩序结构和在该结构中行动的群体构成的。场域是多元价值、多元意象、多元主张等混成的“流体”。既固定又流动，既有刚性又有弹性，是场域的基本特征。完全固化的场域是不存在的。就教育而言，前文通过“还原论”分析了最小的组织单位，这个单位就是最小的场域，这个场域又是更大场域中的一个局部。层层嵌套的场域逐层对个体施加影响。也就是说，无论对学生而言还是对教师而言，影响其意义建构的因素，不限于最为直接的教育关系 $[E = (\bar{T} \cdot \bar{S} \cdot \bar{C})]$ 。场域远端的影响是通过政策、制度、资源等形式直接或间接施加的。根本上说来，或直接通过评价，或间接通过评价施加影响。从泛在的广义教育意义上说，教育评价不仅仅局限在最小教育单位，也不仅仅局限在学校这样的教育组织。它弥漫在整个社会，无处不在，无孔不入。

教育有正式教育(Ed_f)与非正式教育(Ed_i)之分，或者说，在第一系统(国民教育体系)和第二系统(国民教育体系之外泛在的教育行为)之间存在差别。评价有正式评价(Ev_f)与非正式评价(Ev_i)之别。将教育和评价组合在一起，形成四类教育评价域(表1)。(1)非正式教育-非正式评价(1: $Ed_i @ Ev_i$)；(2)非正式教育-正式评价(2: $Ed_i @ Ev_f$)；(3)正式教育-非正式评价(3: $Ed_f @ Ev_i$)；(4)正式教育-正式评价(4: $Ed_f @ Ev_f$)。现实社会中，在家庭、职场以及其他物理的或虚拟的公共空间，非正式教育都在发生。对非正式教育的非正式评价也非常普遍。虽然是非正式的，但是对个体和社会环境的影响还是客观存在的。影响社会分层的社会评价，对个体思想定向和行为选择具有潜移默化的作用。第1型教育评价尽管普遍存在且发挥作用，但缺乏有效的控制手段和工具。正式教育具有不同的层级，从以课堂教学为中心的教学，到班级(年级、学科组)、学校再到区域乃至国家，是一个相对独立的系统。以此为对象，各类社会主体也常常做出非正式的评价。线上线下对教育的或正面或负面的评论，具有非正式评价的性质。“人人都可以对教育评头论足”，乃是社会中存在着的非正式评价正式教育的写照(第3型)。在网络社会时代，这种非正式评价影响着公众对正式教育的认识和信心。第2型教育评价将非正式教育纳入评价视野，通过正式评价做出价值判断。由于非正式教育的泛在性，除非设计特定的项目，否则很难实施评价。所以，开展面向非正式教育的项目评价势在必行。随着督导制度的建立以及准官方评价机构的出现，第4型教育评价成为质量保证的焦点，越来越受到重视。

表1 正式与非正式的教育与评价

	非正式教育 Ed_i	正式教育 Ed_f
非正式评价 Ev_i	1: $Ed_i @ Ev_i$	3: $Ed_f @ Ev_i$
正式评价 Ev_f	2: $Ed_i @ Ev_f$	4: $Ed_f @ Ev_f$

现代教育评价的专业化发展，就是从对正式教育的正式评价开始的。教育评价和心理测量在源头上建立了不解联系。^⑦撇开非正式教育不谈，就正式教育而言，它是一个多层级主体参与建构的体系。体系(S)是体系(S^n)的嵌套，是一个由诸多层级(n)子系统构成的复杂系统。相应地，教育评价也有一个体系问题。教育评价体系即从课堂教学到国家教育各个层级以形塑评价功能以及它与内外环境的关系的程序性、制度性和政策性安排(Hanberger, 2016)。从社会学意义上看，社会分层与评价有关。社会分层是制度化了的不平等的社会结构和安排。不平等有两个要件：差异和评价。不平等源于差异，但是差异只是不平等的必要条件，并非充分条件。对差异的评价，将差异分出等级，并依等级分配社会资源，经结构化沉淀，形成社会分层体系。评价与价值判断和社会分层的内在联系，是讨论教育

评价问题必须考虑的前提性因素。换句话说,现实地看,教育评价和社会评价紧密纠缠,不可能有纯粹的、独立的教育评价。社会评价是一个非常宽泛的概念。社会评价有的是自发的,融入经年累月积淀的文化传统之中;有的是人为的,为各级各类组织有意创制实施。比如品牌项目、人才称号、重点工程等考试选拔之外的项目,也起着评价导向的作用。无论是哪一类,对于师生双方的意义建构,都产生或大或小的影响。同时,对各种教育机构也产生实质性的影响。在微观层面(最小教育单位)教育评价就涉及价值问题。在组织和系统层面,教育的价值复杂性增加。层级多样化和主体多元化,使教育评价远远超出微观层面的价值操作行为。对包括评价在内的教育行动的治理,为组织、地区、国家和跨国组织所需要。

二、教育评价的治理功能:场域权力、价值操作与工具选择

(一)评价与治理具有内在联系

广义的评价和日常生活中的行动策略相联系。有的行动产生公共影响,有的行动是私人性的,不产生公共影响。凡具有现实或潜在公共影响的行动,都是治理的对象。当然,治理本身就是公共性的,因此也是治理的对象。正像广义的评价伴随人们的日常生活,广义的治理也覆盖人们的日常生活。一般而言,在公共空间,无评价不治理。反过来,无治理不评价。评价需要治理的眼光,治理也需要评价的视角。

评价和治理具有内在的联系。在评价和治理之间,存在相互纠缠的意义与互渗的功能关系。评价和治理都为一定的价值观牵引,都是价值行动,都是利用一定手段和工具实施的价值操作过程。两者都涉及价值和工具,都是在特定的场域实施的。场域是权力线分布的空间。权力线的分布方式决定了场域特性。暂时撇开场域不谈,仅就治理与评价的形式关系分析,可以就它们的价值与工具性质建立二维表格,从而形成治理价值-评价价值(G_vE_v)、治理价值-评价工具(G_vE_m)、治理工具-评价价值(G_mE_v)和治理工具-评价工具(G_mE_m)四种不同的组合(表2)。评价和治理都涉及主体、对象、方法等问题。谁评价、谁治理、评价谁/评价什么、治理谁/治理什么、怎么评价、怎么治理,在形式上具有一致性。超越特定主体和对象的评价与治理涉及层级跃迁。在最为一般的层面,存在着超越个体和组织的超主体。超主体将评价者、治理者作为对象。超主体的评价和治理分别表现为元评价和元治理。在特定的场域,主体构成及其评价和治理行动的影响存在层级上的差异。在不同领域也会因目标和任务的不同而有不同表现。表2所建构的二维形式关系,在具体场域获得内容性的动力学“生命”。

表2 治理与评价(价值与工具)

		治理G	
		治理价值(G_v)	治理工具(G_m)
评价E	评价价值(E_v)	$E_v \otimes G_v$	$E_v \otimes G_m$
	评价工具(E_m)	$E_m \otimes G_v$	$E_m \otimes G_m$

从治理的角度看,评价是改进决策,管理知识,提供关于行动的关联性、效率、效益、影响和持续性的可靠证据的重要工具。在组织层面,就是是否需要外力介入、什么管用什么不管用,以及在什么时候管用等方面,评价提供鉴定和证明。^⑧治理有的在“约定俗成”的秩序下实施,有的本身就在改变或创建新的秩序或规则。审慎的治理,在改造和创建秩序的时候保持理性的价值关照,实施价值操作。即使对约定俗成的秩序,也保持批判性审察的态度。套用苏格拉底“未经审察的生活是不值得过的”的话,可以做这样的类比:“未经审察的治理是不值得做的”。将治理作为审察对象,需要评价作为中介。评价具有治理的功能,是治理的工具,是治理实现目的的手段。反过来,治理也是评价的对象。治理本身就包含着评价行为。评价的治理功能表现在:(1)对治理进行评价,通过改造治理实现善治,针对的是治理的总体。(2)对治理中的对象领域进行评价,实现对象领域的善治,针对的是治理的局部。

其中,政策评价是对治理评价的重要组成部分。^⑨毫无疑问,对治理的整体评价不同于局部评价。局部表现好,并不一定整体表现好。整体表现好,也不意味着每一个角落都好。无论是整体还是局部,评价做不到全覆盖无死角。所以,现实中必然会存在“评价的盲区”。

治理和评价互为工具。治理通过评价实现其追求的价值。评价通过治理,建立专业的评价标准和评价生态,实现其价值鉴定和判断的操作目标。治理体系和治理能力现代化,以对体系和能力的评价为前提。更上位的价值目标就是现代化。现代化是目标词项,具有统摄治理的话语力量。但“现代化”的价值与意义,也需要建构与评价。不经过评价就无法了解治理的现状,因而也就不可能知道治理是不是在沿着正确的道路进行,从而无法判断善治与否。

(二)“善治”是国家治理水平的评价参照

教育治理与国家治理和社会治理相联系,不存在游离于国家治理与社会治理之外的教育治理。在全球化时代,教育治理还有国际社会的参照背景。教育评价需要以善治的视角观照。我国古代“治理”多指国家治理,体现于公秩民序。^⑩在现代社会,治理概念的内涵和外延有了一定的变化。跨国和跨境组织及其功能的发挥,使相对封闭的治理主体有了突破。治理问题已经超越了国家的界线,层级日益增多。全球治理建立在实体性主体和机制上。相应地,权力场域发生了变化,场域结构变得更加复杂了。特别是“二战”之后,联合国及其下属组织以及一些区域性国际组织纷纷成立。这些组织不仅仅在国际事务中发挥重要作用,同时也影响到国家内部的治理问题。

进入新世纪之后,联合国就发展问题先后发布重要报告,影响成员国的国家行动。2002年秋,联合国成员国承诺通过全球协作推进实现8大千年目标。^⑪2015年联合国大会第七十届会议上通过了《2030年可持续发展议程》(以下简称“《议程》”)。《议程》提出了17项可持续发展目标,^⑫凝聚了人类共同愿景,达成了世界人民的社会契约,开出了集体行动的清单,绘制了世界发展蓝图。这些目标不仅仅是技术层面和资源方面的,它们极大地挑战成员国动员人力和资源、制定和实施艰难政策、带领公民投身塑造未来的行动能力。与此同时,《议程》对全球治理也提出了挑战。在国际组织推动下,治理和“善治”日益成为发展计划中必须考虑的因素。联合国亚洲及太平洋经济社会委员会(ESCAP)将治理界定为:在组织、地方、国家和国际情境下决策制定和实施决策的过程,也是来自公私领域不同层级的政府部门、工商界、媒体、工会、影响比较大的人物等多重主体相互作用、集体行动的结果。对治理水平的评价,影响到发展项目的开发和实施。治理当然有好坏之分。善治就是良好的治理,是公共利益最大化的社会管理过程。^⑬世界银行认为善治是发展的必要前提。它通过六个维度(声音和问责、政治稳定和少有暴力、政府有效率、监管质量、法律规则、腐败控制)对200个国家治理水平进行监测,实际上就是对国家治理的一种评价。《人类发展报告》将善治界定为:为了消除社会腐败,给予人民参与影响他们生活的活动的权利、手段和能力,并能对政府的所作所为问责。

综合起来看,善治是社会良性运转的基本要素,体现为资源分配和使用的高效率、服务提供的公平性。善治在不同国家因文化传统、政治结构、经济发展水平的不同而有差异。虽然如此,在尊重人权、法律规则、公共部门讲求效率和效益、公共事务行动负责透明等方面还有共识。其基本特征包括:合法性、透明性、责任性、法治性、响应性和效率性(俞可平,2000,第9—11页)。善治不限于国家权力或政治意志。法治、透明度、问责不仅是行政程序或制度设计的技术性问题,它们也是民主过程的结果。驱动民主程序的力量不仅是尽职尽责的领导力,还包括社会利益群体的广泛参与和相互之间的争论。权力合法化和有效制度供给是民主过程的保障。权力的行使需要成文的标准支持,标准本身则是需要得到普遍支持的社会价值(Johnston, 2016)。国际组织的发展项目主要针对欠发达国家和地区,但善治的概念适合所有国家和地区。对全球治理的普遍性原则和标准而言,每个国家不仅仅是“消费者”,还是“生产者”。从外部看,全球治理中要防止普罗科拉斯提斯(Procrustes)床或“铁床匪”效应。从内部看,丧失自我的主体性而按照抽象原则进行“削足适履”的裁剪,必然会造成跛足的悲剧。

对普遍性的贡献和对特殊性的坚持,是国家治理的动力学特征。路径依赖和传统粘连,使任何一种体系无法做外科手术式的切割移植,也不可能指望内源性的自发剧变。虽然如此,在全球化背景下,拥有全球眼光,以国际组织凝练的善治标准为参照,对推动国家善治还是非常有意义的。

(三) 评价具有治理的功能,是实现善治的工具

从治理到善治,有一个评价环节。治理是评价的对象,治理本身也是治理的对象。治理好与不好,公共利益能不能得到最大化实现,善治能不能得到充分保证和体现,无疑需要通过评价做出鉴定和判断。对治理进行评价的前提,是治理具有可评价性。一方面,作为整体的治理体系和治理能力是可评价的。另一方面,构成治理体系的各个子系统及其要素也是可评价的。各个层级的目标和价值、制度和政策、行动和结果,无不构成评价的对象。不良治理(bad governance)越来越被看作是社会中的万恶之源。一些国际组织逐步将其援助或贷款建基于目标国家或地区的“善治”,为此也推动目标地的治理改革(UNESCAP, 2009)。善治概念的广泛传播,在一定意义上看,是国际组织在实施各项援助计划的过程中对项目成效进行评价的反推产物。项目执行情况是治理体系和治理能力的反映。对项目实施评价在一定程度上也是对治理水平的评价。

政策评价是治理评价的重要组成部分。评价对于判断决策依据是否可靠、对政策执行结果进行公共问责以及考察公民对政府信任度方面起着重要的作用。“评价者的任务是尽力将证据压缩成一个词:好还是差”(Scriven, 1971)。评价为政府提供证据和经验,以支持基于需求的政策规划、项目改进和组织学习(Agrawal, 2017)。评价是促进项目计划、执行与管理的有效工具。评价有助于项目负责人和其他参与人员了解项目是不是根据日程在进行,项目的投入、活动、输出和外部因素是不是按照计划在运转。评价帮助计划者和管理者了解项目目标的达成度。^⑩可以说,评价是实现善治的重要手段,好的评价有利于治理趋善。另一方面,所有评价都在特定的治理情境或治理体系中展开。评价的实施需由治理营造氛围、保证条件、提供行动结构。善治为评价创造良好发挥作用的的环境。善治为好的评价提供路径,使其不断进化。作为治理的特殊领域,教育治理和国家治理、社会治理不可分割。教育评价乃是教育善治的工具。教育善治需要好的教育评价。

(四) 教育评价与教育治理存在层级对位

《深化新时代教育评价改革总体方案》(以下简称“《总体方案》”)指出“教育评价事关教育发展方向”,发挥着“指挥棒”作用。教育评价改革的目的是“提高教育治理能力和水平,加快推进教育现代化、建设教育强国、办好人民满意的教育”(中共中央,国务院,2020)。文件将教育评价和教育治理水平相联系,正是看到了评价和治理的密切联系。教育由第一系统和第二系统构成。第一教育系统以国民教育体系为边界,第二系统是第一系统之外的所有教育场域(称第二系统是为了讨论问题方便)。两个系统教育行为的表现形式和作用方式有所不同。在现实世界,二者是交叉发挥作用的。教育具有基础性、先导性和发展性的特点。一方面,第一系统的教育评价影响全社会的治理水平;另一方面,第二类系统的泛在教育评价,对第一类教育评价具有直接或间接的影响。现代思想界创造的许多诸如集体意识和集体无意识等概念,皆指向为个体所有但又超越个体的心理倾向和意向偏好。

这些集体意向对个体的影响有的是潜在的,有的则是清楚表达的。在所有意向族中,评价起着牵引或导向作用。评价和场域结构有关,存在权力关系问题。“评价权”既是对评价的权利界定,也是对评价的权力限定。拥有评价权利不一定拥有评价权力,拥有评价权力也未必拥有评价权利。评价权力不是无限的,也不是无所不包的。在特定场域,主体依据评价权,选择一定评价工具做价值操作,制造“评价事件”(Evaluation Event)。从最小教育单位到国家教育场域,评价事件发生的频度和影响度有所不同。在评价事件中,“最近他人”的评价作用最为直接,尤其是在个体心理层面影响直接且持久。讨论教育评价的治理功能,乍看起来是一个宏观治理问题,实际上宏观治理如果不考虑微观操作,就会流于抽象扰动而不解决一线问题。

第一系统的教育场域有多个层级。为了方便讨论,可以简化为五个层级的锥形结构(表3):由师生最小教育单位、学校、地区和国家(其中,学校有学段和类型上的差别,而地区也有更细致的区别,如地州县乡等)以及国际构成。在权力、价值和工具之间存在着不同的组合。公权力越大,公共价值的普遍性和代表性越大,可供选择的工具越多。相应地,教育评价的影响面越大,影响越深远。从权力大小看,层次越高,权力越大。从人口数量来看,正好相反。权力层次越高,人数越少。就价值操作(价值判断、价值计算、价值整合、价值定位、价值执行)而言,国家层面的治理目标是价值最大化,前提是价值共识最大化。在地方层面,地方既为凝聚国家价值做贡献,也有地方的价值定位和地方性的价值共识。地方价值原则上必须体现国家价值,也有地方特色。学校受国家和地方双重价值指引。学校追求特色办学,也有校本价值主张,也需要校本价值共识。就工具(技术、方法、手段、机制)选择而言,不同层面的工具箱内容有所不同。国家、地方和学校的治理工具主要是制度、政策和资源。当然,在具体执行的时候,人际关系也发挥作用。这些工具的适应范围受不同层级的权力影响。从纵向的权力关系看,越往基层,工具越具体,政策空间和资源容量越小,影响半径也就越小。由于国家具有独立主权,国际层的评价范围和效力是有限的。在国际交流不断拓展深化的背景下,国际评价越来越发挥重要作用。

表3 场域与评价

场域层级与权力关系	价值操作	工具选择
国际	国际价值共识及其实现	国际评价工具
国家(横向、纵向)	国家价值共识及其实现	国家评价工具 (制度、政策、资源、人际)
地方(横向、纵向)	国家价值共识+地方价值共识及其实现	地方性评价工具 (制度、政策、资源、人际)
学校(横向、纵向)	国家价值共识+地方价值共识+学校价值共识及其实现	学校评价工具 (制度、政策、资源、人际)
师生(基本单位)	国家价值共识+地方价值共识+学校价值共识+师生价值共识及其实现	人际互动中的评价工具

师生关系是最小的教育治理单位。“由管理到治理”的基本精神就是“权利平等”和“利益相关人全纳”。最小单位看似简单,实际上蕴含着一切治理的核心议题:“自我与他者之间的关系”。这个最小单位的工具有什么?当然有教育的物化工具,比如教育设施设备、现代化的教育技术、教科书等。可以做这样一个思想实验:假如所有的场域层级消失、所有的教育设备被付之一炬,只剩下教师和学生(自学除外),即, $\bar{E} = (\bar{T} \cdot \bar{S} \cdot \bar{C})$ 变成了 $\bar{E} = \bar{T} \cdot \bar{S}$,教育还有哪些工具?教师自身就是唯一的工具。教师既是评价的工具,也是教育的工具。当然,教育本身就包含着评价。教育治理与评价须臾不能忽视教师的评价行为。教育评价所具有的治理功能,与评价事件的性质相关。评价事件是价值操作和工具选择的组合。在什么场域、什么样的评价在实现什么样的治理功能,需具体问题具体分析。

就教育的最小单位而言,师生教育关系既受外力影响,师生彼此也相互影响。日常教育行为中的评价事件是总体教育评价的基本构成。从基本教育单位外推逐级对照,在微观教育评价和宏观教育评价之间需要价值上的一致,或价值共识。层级越低,对象越具体,个性越丰富,评价感受性越强。教师具有一定的自由裁量权。各个层级的权力意志自上而下抵达教师之后,必然要经受教师的“理解”“解释”和“转移”。因此可以这样说,师生间评价是实现教育治理的基础,是治理目标实现的“最后一米”。“指挥棒”是对教育评价的治理功能的一种比喻。不同层级通过教育评价指向教育目标,实现治理的功能,所使用的工具有所不同。最终抵达“教育”的落脚点,无疑是教育的基本单位。学生如何评价自己和教师,教师如何评价自己和学生,体现教育评价的最小单位特征($T_v S$)。在最小评价单位之外,学生和教师还分别受其他层级主体的评价。除了第一系统的评价之外,他们还受到第二系统的评价。在诸多评价中,影响最大的是决定“个人收益”的评价。

对学生而言,评价有的有形,有的无形。有的产生即时短期的影响,有的影响长远。考试成绩、“小红花”“三好学生”等,作为履历材料对其后续发展的鉴定提供基础信息。师生交互作用过程中出现的评价事件,构成学生体验到的最为直接的评价力量。教师的评价性语言行为(聪明、天才、笨蛋、“猪脑子”)和非语言行为(“关注”“忽视”“冷脸”),有意识的和无意识的评价,对学生有着直接的影响。学段越低,直接的评价行为的影响越大。学校层面开展的“差生”“双差生”帮扶,表面上看是为了促进学生发展,事实上作为一种特殊的评价论断,可能会对学生产生消极影响。对于教师而言,除了自我发展和专业发展外,称号、评优、职称、奖金等外在的评价影响其社会地位、经济收入和个人名誉。

这些外在的评价对于最小教育单位而言,无疑产生着实质性的影响。任何教育理想的话语表达,不经教育行动的“知识转移”,就会停留在抽象的理想层面。而师生对外部评价的解读,必然影响其主体间性的建构,必然对教育行动产生直接影响。作为最为核心的发展主体,学生的切身感受,尤其是对评价的体认,是教育治理抵达的根本判据。有质量包容性的教育供给,立德树人的核心任务,在抵达最小教育单位的时候,经层层评价主体的价值操作,或者越来越增强,或者越来越衰减。体现公共利益的教育评价当然希望逐层递增,即价值共识越具体越强烈:纵向共识保持一致,横向共识实现最大化。宏观层面的价值大词,在微观层面有一个理解和解释的环节。价值的超越性特质无疑使价值整合从微观向宏观跃迁成为可能。但一切理解和解释都是在特定场域进行的。意义建构是人和环境相互作用的产物,也是人际互动的结果。不同层级的职业身份和职位权力,使个人的角色有所分化。在特定的场域,对个人的角色整合而言是一种挑战,对于不同个体结成的组织,不同个体的不同角色之间的协调,就更是挑战了。实际上,在不同场域、不同层级和不同个体之间,无疑存在着价值缝隙甚至是价值鸿沟,也存在着工具和价值之间的错位。这就需要对教育评价实践加以治理。

三、教育评价的自反性:评价陷阱、评价治理与元评价

(一) 教育评价要坚持自反性立场

教育评价和研究是近邻。教育评价需要面对人文社科领域无法回避的方法论问题。教育评价改革不能不触及评价理论与实践背后的深度假设,不能不触及评价的范式。检讨和反思教育评价的深度假设和方法论范式,对评价力量和局限保持清醒的意识,自觉为评价结果的应用范围划界,是教育评价应有的自反性立场。^⑮评价在研究和行动之间架起桥梁。评价是好是坏,涉及评价“供需双方”的态度和品质。不是所有的评价都值得实施和采用,也不是所有的用户都具备应用评价结果的应有态度或能力。早在1969年,古巴和林肯就曾总结过评价失败的七大原因:一是缺乏对评价的充分界定;二是缺乏充足的理论支撑;三是缺乏决策的过程性知识;四是缺乏做出判断所需要依靠的标准;五是缺乏对评价对象复杂度的区分方法;六是缺乏组织、加工、报告评价信息的机制;七是缺乏训练有素的评价人员(Guba & Lincoln, 1981, 第10页)。20世纪80年代初期,他们针对美国的实践,对评价的供需两方予以批评。以政策制定者为代表,需求方存在着无知(ignorance)、怠惰(laziness)和政治塞责(political sidestepping 塞责)的陋习。正是因为有这些特点,他们无视评价的结果,不能真诚应用之。

另一方面,作为供给方,传统评价也有自身的缺陷,即所谓“评价的贫困”。尽管评价结果有统计学上的意义,但评价者不关注实际用户的需要和关切。对用户而言,并未产出真正有价值的知识(Guba & Lincoln, 1981)。作为对实证主义评价忽视用户需求倾向的调整,他们从建构主义和自然探究的立场出发,提出具有里程碑意义的第四代评价思想。第四代评价超越了自然科学的范式,将人文、政治、社会、文化和其他情境性因素纳入评价视野,主张知者和被知者之间的统一。评价使评价者和利益相关人在互动过程中联合起来,共同创造出评价的结果。评价目的指向赋能解放,评价过程强调协商对话,评价的出发点坚持行动导向,是第四代评价的特征。他们曾经组织持后实证主义、建构主义、批判理论等方法论取向的学者开展对话,以《范式对话》的书名汇总了对话各派的观点(Guba, 1990),终结了所谓“方法论大战”(Denzin, 2010)^⑯。古巴和林肯的理论触及评价的基本范式和深度假设的方法论问

题。他们关注的是正式的专业化评价问题。事实上,评价充斥于日常的人类生活。对评价的范式自觉,不仅仅是专业化评价必须拥有的自省态度,也应该是日常评价的观察视角。

在世界范围内看,一方面各类评价不断出现,影响也在不断扩大(比如针对15岁学生的PISA测试),跨国评价经验的参考日显重要;另一方面,尚缺乏对评价经验的系统研究。评价本身就是评价的对象。元评价是对评价的设计、实施、效果进行系统、透明的分析和鉴定的过程,是系统收集已有评价信息,对评价的深度假设、价值预设、评价工具、评价执行、价值裁定进行再评价的活动。对于政策部门而言,元评价是对评价功能理解和改进评价功能的有益工具。^⑭为了规范和指导评价实践,许多国际组织或评价协会开发了关于“好评价”的标准或原则。英国评价协会的指南概括了好评价的八项原则:目的明确(Clarity)、结论真实(Integrity)、独立开展(Independence)、结果可及(Accessibility)、公正可信(Trust)、平等守正(Equity)、原理透明(Transparency)和多样性观照(Diversity)。^⑮这八项原则的借鉴意义是明显的。斯塔夫姆(Daniel L. Stufflebeam)曾指出,“元评价的价值在于保证评价所提供的结论是合理的。实施元评价符合公共的、专业的和机构的利益。元评价对于评价实践的持续改进、各类机构的有效评价系统的建立和管理是有益的。专业化的评价者逐步严肃对待其元评价的责任。当然,好的评价实践还需要其他的工具和程序协同发挥作用”(Stufflebeam, 2001)。美国开发评价标准的工作可以上溯到1954年。当时的背景是,为了防止测评机构的腐败,美国心理学会在克伦巴赫(Lee Cronbach)的带领下研制测验标准。20年后美国教育评价联合委员会(the Joint Committee on Standards for Educational Evaluation)出现。这个委员会是在20世纪60和70年代美国对其“反贫困项目”的评价失败后成立的。其最初使命是将各种各样的利益相关者(包括教师、顾问、统计学家、心理计量学家、评价专家、政策制定者和行政管理人员)集合起来,检讨和反思项目评价的意义和真谛。在联合委员会制定的标准出台之前,美国教育者普遍认为,好的评价应该满足实验设计所需要的内部和外部的效度要求,以及测量所需要的信效度要求。将评价质量的鉴定仅仅局限于信效度而不考虑实用性、可行性和正当性,反映了当时的狭隘评价观(Stufflebeam, 2004)。委员会成立后凝聚了各界对评价的共识,扩大了评价标准的视野,逐步消除分歧,统整专业性的话语体系,先后推出项目评价标准(1981)、人事评价标准(1988)和学生评价标准(2003)。其中项目评价标准不断更新调整。在1994版《标准》中,将元评价作为“准确性”的一项子指标(Accuracy, A12)。其原因在于“评价也需要按照评价标准(委员会开发的四个标准:实用性、可行性、准确性、正当性)对自身实施形成性的和终结性的评价。如此,评价行为才能得以恰当实施。评价结束后,利益相关者可以检查评价的优缺点”。

2010年版的委员会标准在原有四项标准的基础上增加了“评价问责(Accountability)”一项。元评价由原来归属准确性转归评价问责。新版本为什么要做这样的调整?这实际上是将元评价和治理联系起来了。治理中的评价是公器,由评价的治理功能反推对评价的治理,自然是需要对评价进行评价的,这就是自反性在评价中的体现。“元(meta-)”具有“在……之后”和“在……之上”意思。元治理、元评价还有其他“元X”,都反映了人类认识世界和改造世界过程中的“自反性”特征。自反性的复杂度随主体和场域变化而变化。无论什么场域,自反性最后要落到个体上来。评价者个体的价值观、信念和经验,不可避免地会带到评价事件中来。消除评价者的偏见和知识错觉,需要评价者超越自我,跳出自我中心的视角,克服自恋情结,由自我到他者,由个体到组织、地方和国家,最后在全球治理的角度不断反观自身,重审立场,划清边界,沿“元”的路线提升认识。越是远离个人空间,越需要更为复杂丰富的“场知识”。一方面是包括自身在内的半径不同的场域知识;另一方面是评价对象所处的场域知识。显而易见,场知识(历史、传统、权力结构、人际结构等)是个无底洞。因而,评价标准无论怎么“精致科学”,都是在有限理性基础上形成的有限标准。这就向评价者循环上升的自反性意识和能力提出挑战。

(二) 教育评价要警惕“评价陷阱”

从治理的角度看教育评价,首先需要检讨历史积淀下来的和现实建构的评价陷阱。评价陷阱是有

意无意设置的“评价圈套”,或通过诱因设置使评价参与者陷入评价的恶性循环。^⑭教育实践中实际存在着的行为主义、经济主义、地方主义和机会主义,是评价陷阱的思想根源。立场模糊、认知偏差与利益冲突则是评价陷阱在内容上的表现。

首先要警惕行为主义陷阱。行为主义是学习心理学的一个流派。心理学家通过动物实验(白鼠、猫、狗等)研究人的心理现象,将刺激反应建立起来联结,界定为学习的过程。强化是行为主义的一个重要概念。通过正强化建立并且巩固联结,通过负强化阻止联结建立或使已有联结消退,是行为主义的基本主张。对于许多人类行为而言,行为主义的学习观具有解释力。《卖油翁》的技能描写、迈克尔·乔丹的一万次投篮训练的励志故事、“熟能生巧”等成语,都在揭示重复训练可使一项技能达到熟练程度的道理。教学中通过大量重复练习以达到熟练,将学习挤兑为“大量刷题”的“题海战术”。对于应试而言,实践证明,重复训练是有效的。当然,题海战术也有“战术”或“技能”上的差别。学校出现善于猜题押宝并带领学生做针对性、适应性训练的“应试名师”,社会上出现了专门研究题型和解题策略的“应试医生”“应试咨询师”或“应试教练”,各类应试机构和学校应试教育交相辉映。考试成绩凝聚了教师和学生、学校和家庭的共识。教育管理中设立各类花样不断翻新的奖项,作为正强化诱因,牵引学校、教师和学生“入套”。认知学派心理学、人本主义心理学、心理咨询和脑科学研究的发展,极大丰富了人们对人类学习的理解。多元智能、成功智力、情商、变商、“X商”(胆商、财商)等新语汇,开辟了人类发展的新语境。虽然如此,行为主义心理学在教育和管理领域仍大有市场。在最小教育单元,行为主义的课堂管理表现为通过表扬、漠视和批评这样的强化工具激发学生的学习动机。在表扬和批评之间的“无视”或“冷落”,看上去无关强化。但是,如果将教师的注意力看作是评价的工具,被无视者,恰恰是获得注意力最少的。显然,即使在最为基本的教育单元,无论是聚焦教育的质量价值,还是教育的公平价值,都必须正视行为主义的正负面影响。以辩证的眼光看,行为主义过头了,就会造成对全面发展的损害。

其次要警惕经济主义陷阱。在更大的场域观察,经济主义是行为主义的变种。经济主义重视和强化经济的基础性地位。通过聚焦经济指标和经济绩效凝聚共识,通过经济刺激和物质奖励拉动生产。将经济杠杆作为社会化大生产的价值操作的主要工具。“GDP主义”是经济主义在社会资源配置和政策激励中的集中体现。在经济主义的思想动员下,社会各领域向经济看齐。特别是在公共资源配置的时候,经济表现亦占有相当地位。“一个口号哪怕非常平庸,如果重复千百遍,且与奖赏的许诺或惩罚的威胁动人地结合在一起,那么,人这种动物就无可奈何地确立起所需要的条件反射了”(贝塔朗菲,拉威奥莱特,1989,第13页)。贝塔朗菲曾警告:“我们已经征服了世界,但是却在征途中的某个地方失去了灵魂”。“我们已经失掉了或已经看不见生活方式的指路明灯,即人的价值了”(贝塔朗菲,拉威奥莱特,1989,第19页)。在教育领域,经济创价、经济红利、经济利益与教育绩效挂钩。各种名义的“称号”,成为国民收入再次分配或再再次分配的砝码。人才称号、项目称号、机构称号等各种“品牌”和经济收入一起,在社会分层中占据了举足轻重的位置。经济学中按可量化的“优”“绩”分配奖励原则的边界消失了。在许多领域,都以经济筹码度量和分配。当“称号”的荣誉性附加经济价值后,荣誉激励就会被解读和演化为荣誉后面的利益计算。包括经济收益的“溢价”利益和自由流动的“议价”利益在内,经济主义一方面刺激了人性中的自大和贪婪,另一方面搅动了教育的生态,最终演变成一个人和所有人的战争。公共价值和共同体意识在追逐个人或小集体利益的散户行动中衰竭。《总体方案》中批评的“短视行为”和“功利化倾向”由谁来“买单”和“接盘”,需要从各级各类教育主体的深层思想中探寻“买卖关系”或“交易关系”。

第三要警惕“地方主义”与“机会主义”陷阱。在教育场域,地方(省区、地市、县、乡及同级的不同机关部门)是重要的价值操作层级。各层级之间的纵向权力和利益配置以及同级横向的权力和利益配置,对地方的不同行为主体无疑具有结构性影响。纵向的权力配置使社会在分布上出现“中心与边缘”的结构性差异。权力中心和社会结构的中心同构。中心城市具有“抽水机”一样的资源汲取能力,

对散布的优质教育资源(师源、生源、财源)实施吸附。导致优质资源配置中出现地域上的“倒金字塔”分布。在高等教育普及化和国际化背景下,有大学文凭的“文凭社会 1.0 版”升级换代,出现有中国名牌大学文凭的“文凭社会 2.0”和有世界名牌大学文凭的“文凭社会 3.0”新版本。不断扩大的文凭需求驱动社会的心理预期和社会冲动。在文凭社会牵引下,招生配额和入学机制吸引全社会的注意力。教育机会的激烈竞争在不平衡不充分的矛盾现实面前凸显。在这个过程中,形形色色的地方主义得以表现。地方主义是一个相对概念。它随着场域层级的变化而变化。尽管地方参照有大有小,但是地方主义有一个共同的特点,就是以地方为中心进行价值操作。理想的状况是地方在凝聚公共价值的时候,在国家的公共价值前提下动员。地方主义的价值参照则基于地方,在价值操作中持“地方先于国家”的立场,为了地方利益不惜牺牲更大的公共价值。“圈子文化”“码头文化”“山头主义”“部门主义”“裙带主义”“族群主义”等,乃是地方主义的变种。不看大图的地方主义有其思想上的近亲,即“机会主义”。机会主义与“原则中心”的精神相对立。在权衡得失计算利益的时候,机会主义首先考虑的是一项选择对自己是否有利,而不考虑其他人的利益。只要有获利的机会,即使背离原则、违章越线、放弃操守也在所不惜。如果说地方主义以地方为中心的话,机会主义以利益和结果为中心。机会主义对教育生态的破坏尤为严重。

（三）教育评价治理需要加强评价能力建设

对教育评价的治理需要从两个方向用力。即,自上而下和自下而上。相对于自上而下,自下而上的治理具有很大的分散性。因治理的组织层级的限制,它不具有强大的动员力。一般而言,对于各不同层级的任一评价对象,在治理和评价之间有四种形式化的匹配方式。以评价为对象,无论自我评价还是他者评价,都成为内外治理的内容(表 4)。其中,(1)自我评价和自我治理是最基本的形式。意即无论有没有外部评价和外部治理,自我都要对自我评价实施治理。(2)他者评价和自我治理意在自我对他者的评价进行治理。(3)自我评价和他者治理则是根据他者治理自我评价。(4)他者评价和他者治理的意义在于,不管自我内在的评价和治理如何,治理主体是外部的,治理的对象“评价”也是外部的。他者与自我既可能具有垂直的权力关系,也可能是横向平行的权力关系。

表 4 治理与评价（主体与对象）

		治理G		
评价E	自我评价 (E _s)	自我治理 (G _s)	他者治理 (G _o)	元治理 (G _m)
		(1) E _s ®G _s	(3) E _s ®G _o	
	他者评价 (E _o)	(2) G _s ®E _o	(4) G _o ®E _o	超主体治理
	元评价 (E _m)	超主体评价		

理论上可以做这样的区分,但是实际上,所有外部的他者治理或评价,只有经过个体的意义建构才有可能转化为其行为策略。如果个体内在价值定向与外在价值操作不能达成一致,即使表面上可以实现外部价值遵循,但在内心深处,仍然存在价值矛盾,甚至是价值对立。经验表明,人们的价值立场是模糊的,也是在不断变动调整的。认识也是动态的过程。知的错觉是常常发生的。在不同主体之间,除了价值立场和认知水平存在差异之外,在利益上也有竞争,甚至是对立的。价值立场模糊、认知偏差和利益冲突问题,需要通过评价治理加以解决。价值澄清、认识升维和利益平衡,乃是对教育评价进行治理的焦点。对治理的治理(元治理)和对评价的评价(元评价),是评价治理的超主体构成。

《总体方案》是对教育评价进行治理的上位文件。以“意见”“方案”等命名的文件是实施治理的一种政策工具类型。从逻辑上讲,文件具有结果与过程的二重性。一方面是价值操作的结果,或价值共识的达成;另一方面,文件又是价值共识动员和操作的起点。文件不是一般的政策工具,而是包含价值内涵的工具。文件开发的发动、成形和执行,循自上而下的路径。文件的权威性和影响力随权力层次的变化而变化。以文件的颁布为节点,文件的效力体现在前文件时期和后文件时期的对象结构、内容

和性质的变化。《总体方案》将教育改革的突破口锁定在教育评价改革,文件形成的参与主体、理论基础、实践根据、政策主张如何,文件所涉及的价值共识的基础如何,可供选择的工具如何,文件对各级各类教育行为的实质影响如何,诸如此类的问题是将政策工具作为评价对象提出的。不言而喻,尽管共用“评价”一词,但是评价的位阶是不同的。以第三只眼睛看政策工具的开发和应用,从一开始即拥有“元评价”视镜,在治理过程中对治理本身保持“元治理”的警醒,既有助于政策执行过程的监督,也有利于因地制宜地适时调整。

评价所涉及的基本问题包括:谁来评价、评价谁、评价什么、怎么评价、评价的结果如何使用、评价有什么影响和后果等。这几个问题的核心是两个:评价主体及评价方法。二者结合,焦点就是评价胜任力。依据评价胜任力水平,可以将评价简单地分为专业化评价和非专业化评价。专业化评价的主体分为个体和组织,即专业的评价人士和专业性评价机构。专业化程度取决于两方面的知识和能力;一是对象性的,二是评价性的。前者与所要评价的对象有关,后者与作为一项专门知识领域的评价能力有关。比如环境评价,既需要在环境领域有必要的科学知识,也需要在评价方面有专门的训练。对象越复杂,对象性知识难度系数越大,越需要关于对象方面的知识储备。没有受过专门评价训练的工程师、医生、律师、科学家、艺术家、文学家等专家,在评价实践中应邀做同行评价的现象非常普遍。其根本原因就在于,他们在对象性的知识领域是专家,或“行家里手”。一个不懂数学但是经过评价训练的人,和一个懂数学但是没有经过评价训练的人,哪个更具数学评价能力?显然是后者。不过,就教育而言,大学教育中某一学科的教授对本领域的高深学问很有研究,但是在评价学生发展的时候可能会作出荒唐的评价。反过来看,一个在评价方面训练有素的评价专家,不可能期望他对所有学科领域都有高深学问。但是,一个既懂数学、又受过评价训练的人,和前者比较起来,岂不是更胜任评价工作?因此,可以说评价具有专业二重性,即评价的专业性和评价对象性素养的专业性。以正式评价和非正式评价为一维,以专业化评价和非专业化评价为另一维,将两维联立,形成关于评价的四类组合。即,(1)正式的专业化评价;(2)正式的非专业化评价;(3)非正式的专业化评价;(4)非正式的非专业化评价(表 5)。如果将对象性因素考虑在内,以对象性能力为一维,以评价能力为另一维,根据能力水平做二维建模,形成四种类型:(1)高对象性能力-低评价能力;(2)低对象性能力-低评价能力;(3)低对象性能力-高评价能力;(4)高对象性能力-高评价能力(图 1)。

表 5 评价的专业性

	专业化评价	非专业化评价
正式评价	(1) 正式的专业化评价	(2) 正式的非专业化评价
非正式评价	(3) 非正式的专业化评价	(4) 非正式的非专业化评价

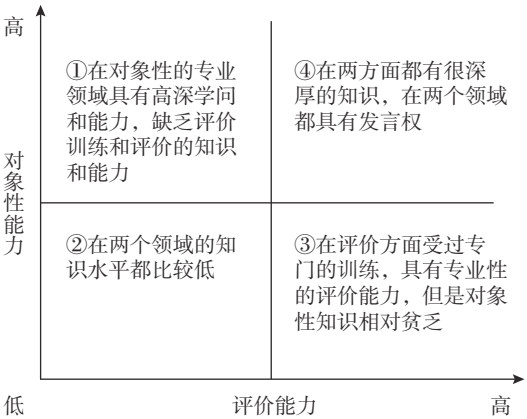


图 1 对象性能力与评价能力的二维建模

《总体方案》针对“五类主体”的评价任务提出要求。五类主体在教育事业中分别承担着不同的责任和任务。他们是事业的推动者,是事业发展的评价者,也是被评价者。每类主体都面临着“他者评价”和“自我评价”的评价考验。特别是自我评价,对各类主体提出评价胜任力的要求。在最为微观的层面,需培养学生的“自我评价”能力。教师也需要有评价胜任力。学校、地方、国家等评价活动的参与者,都需要具备评价胜任力。无论是正式评价还是非正式评价,都需要评价胜任力。随着评价专业化的发展,由学术、学科和学者构成的专门性机构网络,呈现出共同体不断扩大的气象。学生评价、教师评价、校长评价、教学评价、学科评价、管理评价、学校评价、行政评价、政策评价等围绕教育活动而开展的各类评价,构成广义泛在的评价中的教育子集,与其他领域的评价不同。教育评价所面对的组织的特点是松散耦合的。根据卡尔·维克(Karl Weick)的界定,所谓松散耦合,是指这样一种情境,在其中,构成系统的要素是相互联系和响应性的,但它们仍然保持其物理上的或逻辑上的独立性和特殊身份。要素之间的相互联系是间歇的而非连续的、偶尔的而非经常的、微弱的而非显著的、逐步的而非即刻的(Weick, 1976)。

为什么行政人员或教师们将从学术报告或讲座中听来思想拒之门外,其行政管理和课堂教学顽固地坚持原样?为什么尽管教育改革花样翻新,但是教育结果的变化总是不尽如人意?为什么一项意图良好的改革收不到预想的效果?在一定程度上看,这正是因为教育组织具有松散耦合的特点。将学校管理混同于大楼的建筑或工厂的运营,混同于行政机关或企业部门,显然是不适当的。“学校更像是农场,而非工厂”。虽然松散耦合的要素关联变化莫测、似无还有、若隐若现,但它是将组织“粘连”起来的潜在的重要成分。教育组织和其他类型的组织一样是一个信息加工系统。它为人的感知和理解创造环境。在一定意义上说,感知和理解本身也是环境的构成要素。个体的意义建构对组织认知具有重要的作用。诸如决策、计划、制定战略等组织活动的复杂性、模糊性和灵活性,可以从个体的意义建构中获得解释。有效的意义建构是组织可靠性的基础(Starbuck, 2015)。教育评价不同于环境评价,也不同于产品的质量检测。教育评价的最大特点是:人是评价的中心。人既是评价者,也是被评价者。“论断者被论断”。人人都是评价链上的一个接口。教育以人的发展性为前提。人的复杂性和发展的动态性,决定了教育评价的“高利害”“高风险”特征。人的许多方面不可度量,以可测之维度量人的整体,自然是“测不准”。教育评价不是“称体重”,以有限的工具探测无限的可能,当然存在风险。因此不能轻易给人或组织定等级下结论。现实是,并非所有的评价参与者都能意识到评价的高风险性。随意的不经慎思的评价事件随处可见。并非所有的评价者都具有专业化的评价能力。即使是专业评价机构和专业评价者,也还需要不断提升和改进。《总体方案》提出要“加强专业化建设。建构政府、学校、社会等多元评价体系,建立健全教育督导部门统一负责的教育评估监测机制,发挥专业机构和社会组织作用”。反向观之,这正是看到了我国专业化评价力量不足的现状。

根据前文所作的理论建构,对教育改革实践而言,可以概括出如下的结论:(1)表1建构了教育与评价的正式与非正式组合模型。教育评价改革不能忽视非正式教育和非正式评价的存在。通过主体间性建构主体性,对评价的所有参与者都有意义。特别对学生来说,主体性建构更是教育目的的重要构成。杜威说“沟通就是教育”。只要学生与周遭的人们开展交流互动,泛在的教育就会产生。教育评价要成为引导学生回归自我、建构主体性,使他成为自己的过程。教育者的任务就是帮助学生成其所是,为其所能,而非他者的复制、替身,也非家长或教师的傀儡(Osuji, 2015)。(2)教育评价具有治理功能,是治理的工具。反过来,治理也是评价的对象。根据表2和表3,评价是在特定场域选择工具进行的价值操作。“现代的发明只不过刚做到用原子弹代替步兵的程度。然而,作为最后一着,起决定作用的,总是价值体系、观念体系和思想体系。”(贝塔朗菲,拉威奥莱特, 1989, 第17页)。在全球化时代,跨境交流加深,国际组织影响力在扩大。国际合作倒逼国家层面推动“善治”。在国内,不同层级的价值操作和工具选择不同。评价和治理的价值统一,是评价的治理功能得以发挥的重要条件,当

然,也需要治理提供实现功能的环境。在价值和工具之间,需要警惕“目标置换”现象。“遵守规章制度,原本被设想为手段,竟变成了目的”,“从而手段的工具性价值变成了最终价值”。要防止“对可以量化的具体目标过分强调,而对比较抽象、较难衡量的目标却予以忽视”(卡斯特,罗森茨韦克,1985,第188—189页)。(3)评价胜任力的提升是一个漫长的过程。对象性能力和专业化评价能力是支撑评价的两个轮子。哪个轮子“缺血”,都可能使评价的“战车”失衡。任何评价都是在特定场域实施的。“场知识”对于评价者而言,是必不可少的素质要求。任何具体的场域都有一个“隐形评价者”,这就是该场域经年累月积淀的传统的的评价惯习。“过去未去”,评价的历史传统在意义建构中发挥着不可回避的作用。所以,评价胜任力包含着对传统评价的感知力和改造力。(4)教育评价需要持自反性立场,强化在各个层面的自反性意识,重视元评价的自反性作用,提升督导与评价的行政地位,警惕评价陷阱,建立联合评价标准组织势在必行。

参考文献

- 贝塔朗菲,拉威奥莱特.(1989).《人的系统观》(张志伟等译).北京:华夏出版社.
- 杜威.(2007).《评价理论》(冯平,余泽娜等译).上海:上海译文出版社.
- 冯平.(1995).《评价论》.北京:东方出版社.
- 卡斯特,罗森茨韦克.(1985).《组织与管理——系统方法与权变方法》(李柱流等译).北京:中国社会科学出版社.
- 俞可平.(2000).《治理与善治》.北京:中国社科文献出版社.
- 中共中央,国务院.(2020).《深化新时代教育评价改革总体方案》.取自 http://www.gov.cn/zhengce/2020-10/13/content_5551032.htm.
- Agrawal, R., Kalugampitiya, A., Rinzin, J., and Hashim, K. (2017), Leveraging Monitoring and Evaluation Systems for Good Governance. Retrieved from <https://ideas-global.org/wp-content/uploads/2017/12/Chapter-11.pdf>
- Auerbach, J. S., & Blatt, S. J. (2001). Self-Reflexivity, Intersubjectivity, and Therapeutic Change. *Psychoanalytic Psychology*, 18(3), 427—450.
- Boonwaat, L. (2001). An overview of alternative development and illicit crop eradication policies, strategies and actions in the region. In: Proceedings of a regional seminar on Alternative Development: Sharing Good Practices Facing Common Problems. Bangkok: UNODC-Regional Centre for Asia and the Pacific, 129—148. Retrieved from https://www.unodc.org/documents/alternative-development/Manual_MonitoringEval.pdf.
- Bowman, M.L. (1989), Testing Differences in Ancient China, *The American Psychologist*, 44(3), 576—578.
- Dahler-Larsen, P., & Boodhoo, A. (2019). Evaluation culture and good governance: Is there a link?. *Evaluation*, 25(3), 277—293.
- Denzin, N. K. (2010). Moments, Mixed Methods, and Paradigm Dialogs. *Qualitative Inquiry*, 16(6), 419—427.
- Guba, E. G. (1990). *The paradigm dialog*. Sage Publications, Inc.
- Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. Jossey-Bass.
- Hanberger, A. (2016). Evaluation in Local School Governance: A Framework for Analysis. *Education Inquiry*, 7(3), 217—236.
- Johnston, M. (2016), Good Governance: Rule of Law, Transparency, and Accountability. Retrieved from https://www.researchgate.net/profile/Michael_Johnston26/publication/267974525_Good_Governance_Rule_of_Law_Transparency_and_Accountability/links/570d0b1908aed31341cf00c6/Good-Governance-Rule-of-Law-Transparency-and-Accountability.pdf?origin=publication_detail
- Martin, W. A. P. (1870), Competitive examinations in China, *The North Atlantic Review*, 111(228), 62—77.
- Osuji. (2015). Shifting Paradigm in our Educational Assessment. *Global Advanced Research Journal of Educational Research and Review*, 4(1), 001—005.
- Rerkasem, K. (2000). Regional perspectives on alternative development (AD). Paper prepared for UNODC seminar on Cooperation for Alternative Socio-Economic Development to Replace Opium Poppy Cultivation. Hanoi: Vietnam, 7. Retrieved from https://www.unodc.org/documents/alternative-development/Manual_MonitoringEval.pdf.
- Schlagenhauf, J. (2020). Improve Governance with Policy Evaluation, OECD Public Governance Review.
- Scriven, M. (1971). *Evaluating Educational Programs*. In F. G. Caro (ed.), *Reading in Evaluation Research*. New York: Russell Sage Foundation.
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4—14.
- Starbuck, W. H. (2015). (2015). Karl E. Weick and the dawning awareness of organized cognition. *Management Decision*, 53(6), 1287—1299.

- Stufflebeam, D. L. (2001). The Meta evaluation Imperative. *American Journal of Evaluation*, 22(2), 183—209.
- Stufflebeam, D. L. (2004). A Note on the Purposes, Development, and Applicability of the Joint Committee Evaluation Standards. *American Journal of Evaluation*, 25(1), 99—102.
- UNESCAP. (2009). What is Good Governance? Retrieved from <https://www.unescap.org/sites/default/d8files/knowledge-products/good-governance.pdf>
- Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative Science Quarterly*, 21(1), 1—19.
- Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program Evaluation: alternative approaches and practical guidelines* (Second Edition). USA: Longman Publishers.

注 释:

①十八届三中全会提出“推进国家治理体系与治理能力现代化”。十九大确立了 2035 年我国制度建设和能力建设的目标,即“各方面制度更加完善,国家治理体系和治理能力现代化基本实现”。

②我国最新的关于教育评价的改革文件还将增值评价和综合评价纳入政策话语体系。

③*Evaluation Standards of the Swiss Evaluation Society, Adopted (in German and French) by the General Assembly on 9 September 2016.*

④儿童很小的时候就知道 $1+1=2$ 。但是,如果将这个数学计算表达式分别再做抽象化和具体化处理。有的表达在普遍性上没有变化,受认识者的影响相对小一些。还有的表达的普遍性比较有限,受认识者的影响要更大一些。其中,(1)(2)具有普遍性,是常量,差别在于(2)引入了指数的概念和运算规则。(3)开始出现具体的事物,尽管在数量计算上和(1)同,但是由于有了具体化指称, $1+1=2$ 的性质已经发生了变化,出现了选言判断和不等式。不过,除了牲畜主人和兽医为行医需要会“具体牛马具体分析”外,人们对牛马的认识是不会受到李家或张家的影响的。(4)和(3)具有形式一致性。由于认识对象是孔子和苏格拉底这样有名有姓的人,两个表达式发生了“质的变化”。(5)(6)由以人为单位变为以人的思想为单位,情况就更为复杂了。以后来人第三者立场上观察,(5)(6)分别表达了师生关系,各对师生在生活空间(场域)上具有较大的交集,思想上具有传承性。在数学表达上,出现了交集的符号(\cap)。沿着(5)(6)的线索,发展到(7),我们可以发现差异在于两者的场域交集。孔子和苏格拉底生活在不同的生活空间(时空)。由(7)再做区域上的抽象,形成(8),即所谓的中西之别。(1) $1+1=2$ (自然数,加法) (2) $1^x+1^y=2$ (自然数,指数,加法,乘法) (3) $1_{\text{张家牛}}+1_{\text{李家马}}=1_{\text{牛}}+1_{\text{马}}$ (或 \neq ,白马非马,黄牛非牛) $=2_{\text{牲畜}}=2_{\text{动物}}=2_{\text{生物}}=2_{\text{物质}}$ (自然数,加法,分类) (4) $1_{\text{孔子}(c)}+1_{\text{苏格拉底}(s)}=2_{\text{具体人}}=(\text{或}\neq)2_{\text{人}}=2_{\text{动物}}=2_{\text{生物}}=2_{\text{物质}}$ (5) $1_{\text{孔子思想}}+1_{\text{孟子}(m)\text{思想}}=2_{\text{中国人思想}}\{\neq 2_{\text{思想}}(\text{如果}, c\cap m\neq 0);\neq 2_{\text{思想}}(\text{如果}, c\cap m=0)\}$; (6) $1_{\text{苏格拉底思想}}+1_{\text{柏拉图}(p)\text{思想}}=2_{\text{希腊人思想}}\{\neq 2_{\text{思想}}(\text{如果}, s\cap p\neq 0);\neq 2_{\text{思想}}(\text{如果}, s\cap p=0)\}$; (7) $1_{\text{孔子思想}}+1_{\text{苏格拉底思想}}=2_{\text{人思想}}\{\neq 2_{\text{思想}}(\text{如果 } c\cap s\neq 0);\neq 2_{\text{思想}}(\text{如果}, c\cap s=0)\}$; (8) $1_{\text{中国人思想}}+1_{\text{希腊人思想}}=2_{\text{人群思想}}$,及其变体: $1_{\text{中国人思想}}+1_{\text{西方人思想}}=2_{\text{人群思想}}$; $1_{\text{东方思想}}+1_{\text{西方思想}}=2_{\text{区域思想}}$ 。

⑤我国的情况和美国不同。20 世纪 80 年代中期推行教师资格考试,在中学阶段考查内容主要是学科性的。

⑥1875 年加州教育局对小学教师的考试内容共有 20 项,主要考查学科内容性的知识。教师考试,1000 分中只有 50 分是和教学论有关的。比如三个问题:为了跟得上教学论的发展,你将选修什么课程?你如何将孩子正确拼写?你如何使那些懒惰而漫不经心的孩子对学习产生兴趣?到了 20 世纪 80 年代,情况发生了很大的变化,考查的内容有 7 个方面:教学计划的组织、评价、对差异的体认、多元文化意识、理解青年、管理、教育政策与程序。教学论知识包括如何管理课堂、如何组织活动、如何分配时间和轮值、如何结构化布置作业、如何表扬或谴责、如何细分问题层级、如何做好教学计划、如何判断一般的学生理解程度等。从认知派学习论立场出发考虑教师发展和教师教育,不可避免地会触及内容的硬核。

⑦美国学者认为中国是评价的发源地。美国国家教育研究学会 1969 年出版的第 68 份年刊 *Educational Evaluation: New Roles, New Means* 中提到,菲利普·杜波伊斯在 1964 年测验问题大会上提交的 *A Test-Dominated Society: China, 1115 B.C---1905 A.D.* 中指出,公元前 2200 年的时候,中国开始每三年对官员的业绩进行检查,但那时还没有关于内容和方法的记录。到了公元前 1115 年,也就是在商代,测验制度变得清晰了。杜波伊斯的这个论断对后来的评

价专业群体影响比较大。这从一些比较重要的著作和教材可以看出。例如, Worthen, Sanders, Fitzpatrick(1997), 又见 Guba, Lincoln(1981)。对于杜波伊斯的论断, 1989年玛丽琳·鲍雯进行了追问。杜波伊斯的资料源于马丁(Martin, 1870), 他认为中国的文官考试从公元前2200年的禹舜开始。马丁又是从什么资料得出这样的结论, 不得而知。鲍雯猜测或许是公元前100年司马迁所作《史记》。其中提到在禹舜的统治下, 每隔三年对官员进行一次测验。由于缺乏夏代的文字记载, 关于夏代的历史还是一种传统或文学状态, 因此鲍雯对马丁和杜波伊斯的论断持怀疑态度(Bowman, 1989)。1904年, 桑代克出版了第一本教材《心理与社会测量理论导论》, 奠定了心理测验理论的基础。

⑧OECD有一个发展援助委员会(Development Assistance Commission: DAC), 很早就开始对援助项目进行评价, 考察受援国在公共管理、分权、参与、人权、法治等方面的治理情况。

⑨OECD在2020年发布了《通过政策评价改善治理》的报告。这份报告是根据对42个OECD和非OECD国家调查的结果完成。报告共分: 包括“走向政策评价的系统路径”“政策评价的开展: 制度基础为何?”“不同国家是如何回应提高评价质量、应用评价结果的挑战的?”

⑩如《荀子·君道》里说: “至道大形: 隆礼至法则国有常, 尚贤使能则民知方, 纂论公察则民不疑, 赏克罚偷则民不怠, 兼听齐明则天下归之; 然后明分职, 序事业, 材技官能, 莫不治理, 则公道达而私门塞矣, 公义明而私事息矣; 如是, 则德厚者进而佞说者止, 贪利者退而廉节者起。”《韩诗外传·卷五》中论及“道”与“君”“王”, 提到养生、辩治、显设和粉饰四统。“四统者具, 天下往之, 四统无一, 而天下去之”“省工商, 众农人, 谨盗贼, 除奸邪”即善养生。“天子三公, 诸侯一相, 大夫擅官, 士保职, 莫不治理”即是善辩治。“决德而定次, 量能而受官, 贤以为三公, 贤以为诸侯, 次则为大夫”就是善显设。“修冠弁衣裳, 黼黻文章, 雕琢刻镂, 皆有等差”即为善粉饰。《孔子家语·贤君》载孔子和宋君的对话。宋君问: “吾欲使长有国而列都得之, 吾欲使民无惑, 吾欲使士竭力, 吾欲使日月当时, 吾欲使圣人自来, 吾欲使官府治理, 为之奈何?”孔子对最后两个愿望的回答是: “崇道贵德, 则圣人自来; 任能黜否, 则官府治理。”道家《守道》里讲: “治人、事天, 莫若嗇。夫为嗇, 是为早服。早服谓之积德。重积德则无不克。无不克则莫知其极。莫知其极可以有国。有国之母, 可以长久。是为深根固蒂, 长生久亲之道。”其中“治人”即“人君治理人民”。法家《制分》里说: “是故夫至治之国, 善以止奸为务。是也何? 其法通乎人情, 关乎治理也。然则去微奸之道奈何? 其务令之相规其情者也。”

⑪消除极端贫困和饥饿, 普及化的小学教育, 促进性别平等和赋权妇女, 降低儿童死亡率, 改进母性健康, 抗击艾滋病、疟疾和其他疾病、保证环境可持续, 全球合作促进发展。

⑫17项指标涉及内容包括: 消除贫困、粮食安全与可持续农业、健康生活方式、包容与公平教育(SDG4)、性别平等、水与环境卫生、现代能源、经济增长与就业、基础设施和可持续工业、国家间不平等、城市和人类居住、消费和生产、气候变化、海洋资源、陆地生态、社会发展与司法、执行手段与全球伙伴关系。

⑬我国具有丰富的善治思想。《礼记·大学》中概括的“格致诚正、修齐治平”将修身和使命联系起来, 《孟子·寡人之于国也》中有关于“农时、鱼鳖、山林、衣帛、豚狗彘”和“庠序孝悌”等论述, 都蕴含着深邃的善治理念。关于治理的关系性和生态性, 有“水能载舟也能覆舟”的箴言和“近朱者赤近墨者黑”“蓬生麻中不扶而直, 白沙在涅与之俱黑”的认识。印度公元前的政治家考底利耶(Kautilya)认为, 统治者的善治表现为: 寓其幸福于臣民的幸福之中, 寓其福祉于臣民的福祉之中。

⑭A Manual on Monitoring and Evaluation for Alternative Development Projects. 这份报告是替代性发展项目的督导评价手册(Alternative development, AD)。《手册》编制的目的是深入理解AD的原理和概念; 了解AD的多样性与复杂性; 有能力开发本国AD督导与评价的分析框架; 在多样的生物物理、社会和文化环境下有效应用AD理念指导督导与评价。AD是最近出现的一种新发展范式。这个词出现在20世纪90年代中期, 由UNODC提出。替代性发展概念可以上溯到20世纪40年代晚期中国政府成功在西南省份消灭非法种植鸦片罂粟(Rerkasem, 2000)。替代性发展计划在UNODC的介入下正式出现。1998年6月, 联合国大会第20次特别会议正式赋予替代性发展的共同意义。根据联大特别会议(UNGASS)的定义, “替代性发展是通过特别设计落后地区发展的方法禁止和消除非法种植含有麻醉药和治疗精神病物质的植物, 以及在充分认识到目标社区和群体的社会文化特点基础上, 在全面持久的非法毒品解决方案的框架下, 采取有力措施反毒禁毒, 持续性国家经济增长和可持续发展的努力”。目标是消除非法种植, 提供具有法律依据的、经济合理的可持续的收入手段, 以及增进不发达地区社区的生活质量(Boonwaat, 2001)。本

文介绍替代性发展计划,看上去距离教育评价和治理比较远,但是,从发展的替代性选择和督导评价的借鉴意义看,具有一定的可比性。如,在教育评价改革过程中,如果发现有的评价是成问题的,需要改变的,那就不仅仅停留在改革的要求上,还要提供可供选择的替代性方案。贝塔朗菲曾说“凡是没有选择的地方,就只有必然而无价值”,不无道理。

⑮自反性(Reflexivity)是一个内涵丰富的概念,在不同语境具有不同的涵义。国内学者对自反性的理解,多受贝克、吉登斯和拉什所著《自反性现代化》的影响。如王能东的《“自反性现代性”的理论述评》(载《国外理论动态》2009年第7期);谢龙新的《“自反性现代化”与文学批评理论的“自反性”》(载《湖北师范学院学报·哲学社会科学版》2009年第6期);李庆霞的《现代性的反思性与自反性的现代化》(载《求是学刊》2011年第6期)。贝克等人将自反性和现代性联系起来。对现代性自反性的论述,主要从反思的自反性和自我消解的自反性两个方面展开。西方学者从不同的学科视角讨论自反性。比如,加拿大麦吉尔大学的飞利浦·萨尔兹曼就从人类学的立场讨论自反性。他对后现代立场进行了批判,对自反性的价值持怀疑态度。在他看来,后现代思想家强调主观性和个体理解,对知识和真理的获得是无所助益的。可见,自反性在西方世界仍然处在争论之中。本文从反思性、自我批判、自我修正、自我更新等意义上使用自反性。

⑯诺曼·邓金在梳理北美社会科学方法论发展线索的时候提出“方法论大战”。他认为从20世纪70年代开始,学术界经历了三次方法论大战。第一次发生在1970—1990期间,主要是后实证主义和实证主义之间的较量。第二次发生在1990—2005期间,社科领域出现了多种新范式,方法论大战在后实证主义、建构主义、批判理论、解释主义、自然主义、后结构主义、女权主义等流派之间展开。第三次大战从2005年开始,体现为基于证据的研究(evidence-based)与混合方法论、解释学、批判学派之间的争论(Denzin, 2010)。

⑰ *Metaevaluation as a Tool for the Improvement and Development of the Evaluation Function in Public Administration.*

⑱ *UK Evaluation Society Guidelines for Good Practice in Evaluation.*

⑲这里所谓的“陷阱”和“圈套”,只是借用比喻的说法描述观察到的现象,无意强化其贬义的成分。

(责任编辑 王 森)

On the Governance Function of Educational Evaluation and Its Self-reflexive Stance

Zhou Zuoyu

(Faculty of Education, Beijing Normal University, Beijing, 100875, China)

Abstract: In the perspective of inter-subjectivity, educational evaluation is a ubiquitous value processing, and sense-making as well. There are different combinations between formal and informal education and formal and informal evaluation. The formal evaluation of formal education plays the role of direction to the professional development in educational evaluation. Any educational evaluation is executed in a specific field. Therefore, it is context bounded. Educational evaluation is an instrument of governance, sharing the functions of governance. At the different governance level, from individual to organization, state and the international communities, evaluation has the different contents. Good governance is a kind of evaluation at the state level. It also has the directory value for the educational governance. Evaluation itself is the object of governance, while it has the function of governance. Educational evaluation should stick to the self-reflexive stance, keeping itself under re-evaluation. On the one hand, the meta-evaluation awareness should be raised. On the other hand, it should be on the alert against behaviorism, economism, localism and opportunism in the evaluation exercises.

Keywords: evaluation; meta-evaluation; educational evaluation; good governance; educational governance