

教育实证研究的 “德国视野”与“德国经验”^{*}

——底特里希·本纳与李政涛、彭韬的对话

底特里希·本纳¹ 李政涛² 彭 韬³

(1. 柏林洪堡大学,德国柏林;2. 通讯作者,华东师范大学教育学部,上海 200062;
3. 柏林洪堡大学,德国柏林)

摘 要:底特里希·本纳,柏林洪堡大学教育科学研究所荣休教授,当代德国最杰出的教育学家之一,主要研究教育、教养理论和普通教育学。围绕“德国学者如何看待教育实证研究”的问题,本纳教授与李政涛教授、彭韬博士展开了对话,向我们展示了何谓教育实证研究的“德国视野”和“德国经验”。

关键词:底特里希·本纳;教育实证研究;德国视野;德国经验

李政涛(以下简称“李”):您在华东师大举行的第二届中德教学论对话上,谈到了德国教育科学研究的经验转向,对我们很有启发。您提到,德国的教育实证研究者成立了一个自己的研究协会,以与传统的教育理论研究者分庭抗礼,您能否具体说一说这个情况?

底特里希·本纳(以下简称“本纳”):教育实证研究热是在 20 世纪末随着诸如 TIMMS、PISA 等国际学生成绩比较研究的开展并受到人们关注而兴起的。人们开发各种工具来测量学生在如读、写、数学、外语、自然科学与技术等科目上的成绩。这样的研究之所以可能,是因为丹麦数学家乔治·拉什(George Rasch)开发了一种新的统计学模型。在此之前的统计学方法能够用平均值来对整体民众的情况进行评估,但是不能对单个学习者进行评估,也不能对单个学习者接下来的教育过程进行预测。而拉什模型可以做到把各测试题联系起来,同时将各测试题与任意数量的受试成绩联系起来。例如,我们可以用 200 道测试题来测 5000 个受试,然后把所有结果放进模型中进行运算,以此可以得出一个按难度排列的试题序列,在这个序列中可以用解释学的方法进行分级。这样就可以形成一个能力层级模型,每个学生在这个模型中都处于一个很确定的数值上。这个数值说明的是,所有在此难度以下的试题,学生都有 50% 或更大的可能性做对。最简单的题可以保证 100% 做对,随着试题难度的增加,他依次有 90%、80%、70%、60% 和 50% 的可能性做对。在这个难度之上的试题,他则按试题难度的增加依次有 40%、30%、20%、10% 的可能性做对。借助于这个模型,人们可以按学生在某学科上所具备和展现出来的能力所在的水平,来进行国际比较。这是一个学生成绩比较研究的突破,这种研究能对每个学习者做出可信的、可校验的评估,并以此能够对不同的教育体系做出可信的比较。

李:这正是传统的教育研究难以做到,而只有教育实证研究能胜任的,对吗?

本纳:传统研究不能对学生成绩做出预测,而只能给出平均值,传统研究是不对单个学习者进行评

^{*} 基金项目:教育部人文社会科学重点研究基地华东师范大学基础教育改革与发展研究所重大项目“基于学校变革实践的理论原创和学派建设”(16JJD880017)。

估的。教育实证研究不仅调查学生的能力水平,而且也调查学生成长和受教育的环境。教育实证研究要调查学生家长的受教育水平、职业、学生的居住地、上学时间、学校类型、城区环境等。这样,人们就不仅能够对单个学生做出评估,也能够对学生进行分组:既可以按性别、户籍(例如是外国人还是本国人)来分,也可以按出身于文化水平高的阶层或文化水平低的阶层(看家中的藏书量)来分。德国学生在PISA2000的测试中成绩低于OECD平均水平,在数学、阅读和书写的方面成绩尤其差,这被称作“PI-SA震动”。过去人们以为,德国的教育系统是优质有效的,但这样看来并非如此。中国和芬兰的教育系统就比德国的更有效。这次测试的结果使过去的德国教育看上去变成了一场“教育灾难”,于是人们就想利用这种研究方法来提升德国教育系统的质量。德国在20世纪60年代的教育改革也源于一场教育灾难。那时德国只有很少的高中毕业生,而这些毕业生大多都是来自有教养阶层的,只有极少的工人子弟、女生和农村孩子取得了高中毕业证。这就是德国六十年代的教育灾难,它蕴含了极大的教育不公平。这回的教育灾难不是高中毕业生,而是学生在完成了义务教育时(十五六岁)所达到的能力水平,远低于最初的目标和期望。现在,国家大力推进这种研究,因为这个研究承诺能开发出工具来提高学生成绩。

彭韬(以下简称“彭”):那么可以说,这是从对教育的反思延伸到对教育研究的反思,是这样吗?

本纳:当然与此有关。教育政策制定者希望借助于这样的研究成果来改善教育体制。教育实证研究已经在德国开展了二十年,但是在社会选拔方面并没有产生很大的变化。即便在教育实证研究的条件下,无论是代际传承,还是好的教育机会,在不少德国的下一代那里,仍然存在缺失和匮乏的现象。教育实证研究的确能够有效测量学生能力,但却很难证明,什么样的教育改革能够带来更好的学生成绩。

彭:这说明,即使在德国,依然有对PISA这样的教育实证研究的质疑和批判?

本纳:是的。首先,教师本身就认为,从心理测量学的研究成果中没有产生对他们的教育实践的启发。第二,对教育实证研究的批判也来自教育哲学领域,他们批判对学习的测量、对教育的测量和对教育的标准化。第三,这种批判还来自教育政策领域,他们感到失望的是,教育实证研究虽然做出了很多数据,表明德国学生成绩发生了改变或改善,但是对教育改革的具体措施却没能给出任何建议。

李:这些批评和质疑或多或少体现出了教育实证研究的界限,您怎么看教育实证研究的界限问题?

本纳:实证研究的界限原本在于,它并不能测量教学的教育理论和教学论层面上的质量,而仅仅对学习成绩进行标准化了的测量。但这个界限如今已经发生了移动。由于在教育实证研究中加入了教育理论和教学论的思考,因此,教育实证研究可以做得更有水平了。我从两个方面来谈这个界限是怎么移动的。第一个方面,就是教师进修和德国各州的教育研究所所进行的研究。现在德国的每个联邦州都建立了教育研究机构,并将教育实证研究与专业教学论、学校评估、学校教育理念建设等方面结合起来。这些机构会在对学校进行过测试之后,对单个学校给出关于他们学校每个班学生成绩的反馈,以显示出这所学校各班学生在哪些科目上的成绩是怎样的。学校得到反馈后,任课教师、班主任、年级组长、校领导等就知道应该在哪些方面加强教学力度。这种对单个学校采取的指导措施是卓有成效的,比如具体的研究结果得以深入到每个班、每门课以及学校文化,这对教学质量的改善是很有启发和帮助的。在此过程中,专业教学论的观点、普通教育学的思想和教化理论的反思也发挥着重要作用。这一切都意味着,教育实证研究不是自说自话,而是与教育科学的其他分支学科联合起来进行,并且渗透到教育实践的各个不同层面——上课的教师、师资培训的机构、学校教育督导、学校发展部门等。而他们的成果显示,这些基于实证的学校发展研究,确实对教育实践的改善起到了一以贯之的支持作用。例如,跟我们一直有项目合作的汉堡教育质量监测研究所,他们就是在学校测试之后给单个学校发去反馈,呈现他们的学生在各科的成绩统计结果,并且还用纵向数据对学生能力发展进行监测。他们所研究的不是整个教育系统内的学生成绩,也不是国际学生成绩比较,而是同时也在纵向数据中比较同一批学生、同一所学校、同一个城区。这种教育实证研究的措施被证明是非常有价值的。这是第一个界限的变动,它表明,教育实证研究的运用和渗透,扩展到多个具体教育实践领域。

第二个界限的改变在于,人们一直在争论,教育实证研究的对象领域是应该缩小还是扩大。如果这个领域缩小,产生的危险在于可以被测量的科目的地位就变得愈加重要,而不能测量的科目地位就会逐渐降低。而目前我带领的团队研究的项目,走的是另一条道路。我们尝试用教育实证研究的方法,去研究至今没有被实证研究染指过的那些科目领域,在这些领域中也建立起“能力量表”,同时也尝试改善试题的性质,开发出具有反思性的测试题,使其带有教化(bildend)的力量。我们的目标是:这种反思性试题要能够测量和评判教学是否激发起了学生的教化过程。总而言之可以说,教育实证研究的界限是处于运动之中的。

彭:那么您认为当下教育实证研究的界限在哪里呢?

本纳:这很难说。我相信,教育实证研究现在已经能够讨论规范问题了。我们现在所做的 ETiK 项目就是一个实例,我们用实证研究的方式观察学生如何评判规范问题。

彭:比如说,教育实证研究没法讨论价值问题,这算是它的一个界限吗?

本纳:我更宁愿说,这是价值讨论的界限。价值讨论只关注价值和无价值,不涉及教育,因而教育实证研究不涉及价值讨论。我不认同一种“价值-经验”的二元论。

彭:那么是否存在一种经验和其他事物的二元关系呢?

本纳:不存在。

李:自孔德以来,实证研究已经经历了漫长的发展与演变的历史,您怎么看 20 世纪以来实证研究的发展和转变?

本纳:20 世纪初,有的教育研究者试图研究教育和社会化的“自然法则”。他们相信,只要能够用因果关系来澄清教育的功能,那么也就有办法改善教育了。这种用单维因果关系来解释教育过程的研究设想,后来被扬弃了。教育过程是由教师的行动支持的,但我们所追问的那种因果性是教学论建模、教学理念、学生学习成果之间的互动,它总是一种教师行动、学生学习之间的关联,这里不可能只存在单维度、纯线性的因果关系,而是一种多维度交织的复杂情况。因而,在我看来,这种扬弃和认识,是教育研究的一个进步。

教育实证研究的对象领域也扩展到了伦理学-道德教育、政治-法律教育、宗教教育,甚至还初步涉及了审美教育的领域。有人已经开始尝试在美育领域进行能力导向的观察,评估学生在美育方面的感受力、行为能力和阐释能力。我们在伦理学-道德教育的领域进行了长达十年的教育实证研究。伦理学作为一门中小学的课程,是最近才开设起来的,在此之前,关于道德教育,主要是宗教课在进行,或者师生在学校中就某些观点进行一般性地讨论。到 21 世纪初,德国各州都陆续在中小学开设了伦理学课。我们基于伦理学课的教学大纲建立了伦理学课的能力模型和测试工具,以此,学生道德能力得以在模型中得到分化,并得以用测试工具对道德能力进行分层分类地测量。在对道德能力的建模中,我们将学生能力区分为伦理学-道德基础知识、伦理学-道德判断能力和道德行为设计能力。这种划分基于一种普遍性的学生能力三维结构:学生在任何领域的能力都首先涉及由教学所传授的基础知识,然后是该科目(专业)专门的判断能力,以及将这些知识和判断能力运用于个人生活和公共生活的能力。

彭:教育的实证研究有没有其特殊性?

本纳:有的。教育实证研究不关注学生从课堂教学中学到了什么基础知识,它们能测量的是学生在读写方面达到了怎样的能力水平,而不问学生学过什么文法知识。这种研究是我们拒绝的。既然学生是通过教学而受教育的,既然知识是通过教学来传授的,那么我们就不能不关心学生学习知识的情况而只测量能力,我们必须测量学生学到了哪些知识、掌握到了什么程度。在写作方面,不具备正字法知识的学生往往比具备正字法知识的学生犯更多错误。因而,我们做测试总是要测量学生的基础知识、判断能力和在行动中运用知识的能力。这样的研究更加接近于一种教育科学的研究。教育科学的教学研究,能够发现和阐明教学与学生达到的教育水平之间的关联,因为这种研究考察基础知识,考察

在教学中处理了什么内容。我们努力要把教育实证研究发展成教育科学的研究,而我们在此运用的是教育实证研究的工具。

李:实证研究的方法和其他研究的方法如何实现弥合和对接?

本纳:这种对接要通过试题开发来保障。比如我要开发数学的测试题,那么我当然要在试题中设置数学的视角,但我还可以加入公平正义的视角。把一个数除以4和把一定量的物品公平地分给4个人,这是两种问题。我们可以加入生活世界的视角,例如,我们可以反思数学如何改变了人的生活,数学的历史与人类历史的关系等等,这都是我们可以在开发试题的时候考虑的方面。这样,就能开发出很有价值的试题,这些试题能符合教育理论和教学论的要求。我们的ETiK测试题就是这样的,在这些试题中考虑到了不同的伦理学体系,而且这些体系并不处于一种等级性的关系中,它们是不同的判断形式,但其中并没有哪一种优于另一种。然后将这些不同的伦理学体系与自己的道德、陌生的道德和公德相联系,通过这样的联系,就形成了我们开发测试题的一个网络系统,它们具有教育理论的意义、具有与教学的连接性、具有理论的反思性,而我们同时运用实证研究的方法和工具来处理这些试题,并且以实证研究的方法来分析通过这些试题得到的数据,从而形成对学生道德能力和水平的评估与鉴定。

李:这个例子具体说明了您所倡导的“做教育实证研究时不能忽视基础理论研究”,我非常赞同您这个看法。那么,要做好实证研究,需要什么样的基础理论?

本纳:我想以“道德能力研究”为例来说明。道德判断发展的经典理论模型,是科尔伯格的道德判断发展模型,他提出了纯形式的三水平六阶段模型。他的测试主要是看受试的道德判断形式处于哪个水平或阶段,但受试者的道德判断是否用于行动却不在测试范围之内,而且受试的判断是否属于“道德判断”也没有得到考虑,他只看判断的心理测量学的水平。我们在ETiK项目中研究得出的结果是:学生在道德基础知识、判断能力和行为参与能力上呈现的能力水平结构,跟科尔伯格的模型是不一样的。我以“道德判断能力”这个范畴为例。在这个分支能力中的最低水平是:“学生能够在与经验相近的情境中对道德问题进行判断,并根据所学的道德论证方式和思维方式进行判断。”学生能在自我已经习得的、熟悉的道德中进行判断。第二级水平是:“学生能够对自我尚不熟悉的经验世界的道德冲突和道德问题进行反思和处理。”学生能把自己的道德和陌生的道德进行关联,可以对自己的道德和陌生的道德间的冲突做出判断,而不仅仅是根据自己的道德来进行判断。第三级水平是:“学生能够对不同道德标准间因为存在分歧而产生的道德问题进行反思和判断。”第四级水平是:“学生能够在各种伦理学问题中、在各种伦理学视角下反思社会价值和规范,并探寻其有效范围。”在最高水平上,学生能够对社会上存在争议的事情进行判断。达到这个判断能力水平的学生,有能力“有理有据”地参与讨论那些社会上还在争论的事情,这并不意味着,他们会复述教师告诉他们的观点,而是他们自己能够将不同的论述形式加以运用,并在此过程中理解和尊重别人的观点。这说明,他们具有一种讨论性的道德判断能力。我们所做的实证研究的结果表明,修过伦理学课的学生达到的能力水平是比没有修过的学生高的,由此揭示了伦理学课的真实情况。而且,我们通过实证研究所建立的能力层级结构,并非像科尔伯格那样形成前习俗、习俗、后习俗水平,而是在自己的道德、对待陌生道德以及参与对争议性问题的讨论等新的维度结构上进行分析,这样的分析讨论是紧密贴近伦理学、紧密联系教学内容的。

彭:这就说明,在道德能力测量的背后有一套道德教育的理论。

本纳:是的。这背后是有一套道德教育的理论,有一种带着各种不同的基础模型的伦理学,而且这个体系中没有一种唯一的正确性,而是多元的、带有协调性和论争性的正确性。

李:我有这样一个感觉,德国的实证研究跟英美的实证研究是不同的,而这里的原因是德国具有普通教育学的传统,而英美没有,您怎么看?

本纳:既是也不是。我有一个美国学生,她在洪堡大学读了教育学博士,现在在苏格兰当教育学教授。她有这样的感受,英美是有对教育哲学的讨论的,英国也有Philosophy of Education杂志,但是他们

研究得很少。从“普通教育学”向“教育科学”的发展已经持续了很多年,比如在传记研究、在布迪厄式的或福柯式的研究中都有这样的过程。普通教育学使这样的研究成为了可能,也给这样的研究造成的困难。普通教育学的确为教育科学的研究提供了有用的理论,但同时,普通教育学在教育实证研究中被运用得太少了。例如,在开发测试工具时,应该以普通教育学理论为基础,应该联系不同的判断形式,应该区分知识、判断力和行动设计能力,等等。正如我们现在开展的研究这样,确实地将实践哲学、教育学、教学论和实证研究联合起来。可是这样做的教育实证研究太少了。而在教育实证研究中对教育学理论运用不足也导致了这种研究的结果跟教育教学实践的改善没有多大关系。从教育理论的视角来看,这种研既显得肤浅乏味,也与课标中那些高要求的教育目标不相符。

李:这是一种我一直期待的综合性的研究方式,这种综合需要注意什么呢?我总是担心,各种研究方法的“综合”,被等同于不同研究方法的简单“混合”“累加”或“堆砌”。

本纳:这种综合应从三个层面进行。第一个层面是,我们不仅测量学生的道德基础知识、判断能力和行为设计能力,我们还在每个分支能力的维度中形成了不同的能力水平。通过测量和统计,我们可以得知某一个班的学生在各个分支能力上分别处于什么样的水平。我们的测量对象,是与教学内容相符的内容,测量重心放在通过学校教学所培养的道德能力,我们测的是伦理学意义上的道德能力。第二个层面,我们严格区分了“教学练习题”“考试题”和“能力测试题”。一道能力测试题一般用时一分钟,而整个测试不会超过两个小时。它要测量的是“能力”,为此需要经过心理测量学的专业筛选。而教学练习题则是关于某个主题的,它有可能涉及一节课的内容,也可能涉及一系列课的内容,做一道练习题也许要一天或一周。考试题则负责考察学生是否能凭借自己的想象力、创造力或某种能力来解答问题。对这三种题目的区分使得我们能够跟教师解释这样的问题:如果你们的学生在测试中成绩不佳,那么应该用哪些教学练习题来加强他们的薄弱环节?这种从测试题到教学练习题的转化,不是一种从测试对教学的推导,而是测试和教学之间的“非等级性互动”。对教学练习题和考试题的建设其实就是测试题的功能,测试题应该促使教师在教学中不断改善其所布置的教学练习题。这就是一种网络,一种在与学校相关的场域中针对教学与教学法、考试与学生成绩和能力测试的网络。

李:非常高兴能与您再次交流,从中我感受到了三种力量:教育学的力量,德国教育学的力量,和您本人的学术思想的力量。今天的教育学世界,始终存在被其它学科蚕食的危机,始终存在自身根基逐渐漂移以致消散的危机,始终存在无法证实自身独特、不可替代的价值的危机,我们急需新的学术力量的灌注和生成。期待今后有更多机会与您交流!

作者简介:底特里希·本纳(Dietrich Benner),柏林洪堡大学教育科学研究所荣休教授,1941年出生于德国,1965年在维也纳大学获哲学博士学位,1970年在波恩大学取得教育学教授资格,之后在弗莱堡大学、明斯特大学担任教育学教授,1991年至今担任柏林洪堡大学普通教育学专业教授。本纳教授是当代德国最杰出的教育学家之一,他的主要研究方向是教育、教养理论和普通教育学。他于1990-1994年曾担任德国教育学会主席,数次被选为洪堡大学前哲学四院主任,担任德国教育学顶级期刊《教育学杂志》(Zeitschrift für Pädagogik)编委长达二十多年,在德国教育学界享有极高声誉。本纳教授著作等身,至今已发表论文近200篇,出版专著28本、合著22本。其代表作《普通教育学:教育思想和行动基本结构的系统的和问题史的引论》已出到第八版,被翻译成包括中文在内的多种语言。本纳教授拥有学术交流的国际眼光,与世界多国教育学者建立了合作关系,尤其与中国保持了二十多年友好而密切的关系。2004年,本纳教授被聘为华东师范大学荣誉教授。

李政涛,教育部长江学者特聘教授,华东师范大学教育学部副主任、“生命·实践”教育学研究院院长,华东师范大学基础教育改革与发展研究所研究员,专业方向为教育基本理论、教育人类学等。彭韬,柏林洪堡大学博士候选人,专业方向为普通教育学。

(责任编辑 董想文)