我们的教育评价能促进学生的发展吗？

夏正江

上海师范大学教育学院　　教授、博士

摘要：凭常识我们知道，教育评价具有强大的导向功能，有什么样的教育评价，就有什么样的教育实践；有什么样的教育实践，就有什么样的学生发展。在基础教育阶段，我国中小学生的学习或发展质量令人满意吗？种种迹象和证据表明，我国基础教育的质量并不像有些人所说的那样，同国外相比“赢在起点、输在终点”，可能在某些重要方面，我们从一开始就输了！造成这种局面的原因，固然与人们脑子里秉持的落后的、片面的基础教育质量观有关，但长期以来我国中小学刻板、僵化的教育评价（学生评价）系统，亦难辞其咎。尽管教育评价系统对中小学生的学习与发展质量不能负全责，但要说负部分的或不可推卸的责任，总是可以的。基于这种立场，本文对我国中小学普遍流行的教育评价方式对学生的学习与发展造成的影响进行了系统的反思，并提出了相应的改进建议。论文共分三部分，分别围绕中小学应当考什么、评什么；如何考、如何评；如何解释和使用评价结果这三个问题，进行了从理论到实践的探讨。论文分析指出，我国中小学流行的教育评价实践主要存在以下三大问题：（1）“该考的不考、不该考的考了导致教育过程的扭曲与变形”，具体表现为：在评价范围的分布上，偏重认知类教学目标的评估；在知识获得的评估上，偏重陈述性知识和显性知识的检查；在智力发展的评估上，偏重于低层次认知能力或单一类型智力的评估；有的东西该考，但考的时机与场合不对（如小学入学招生考知识），有的东西压根就没有必要去考。（2）“评价的标准、方式方法不当导致评价结果的失真与无效”，具体表现为：评价标准泛政治化和泛科学化、迷信标准答案；纸笔测验一统天下，评价方式过于刻板、僵化，学生评语脸谱化、模式化、空洞化。（3）“对评价结果的解释与使用不合理导致评价功能的窄化与异化”，具体表现为：对评价结果的解释过于简单化，认为“好分数等于好学生”；采用竞争取向的评价视角，为分等、选拔、鉴别而评价；评价游离于教学过程之外，与教学相分离。

关键词：教育评价、学生发展

“现在考试，用对付敌人的办法，搞突然袭击，出一些怪题、偏题，整学生。这是一种八股文的办法，我不赞成，要完全改变。我主张题目公开，由学生研究、看书去做。例如，出二十个题，学生能答出十题，答得好，其中有的答得很好，有创见，可以打一百分；二十题都答对了，也对，但是平平淡淡，没有创见的，给五十分、六十分。考试可以交头接耳，无非是自己不懂，问了别人懂了。懂了就有收获，为什么要死记硬背呢？人家做了，我抄一遍也行。”——毛泽东在1964年春节座谈会上的讲话

上面这段话是毛泽东在1964年春节座谈会上的讲话，看了这段话，许多人可能会感到吃惊、感到有些“不可思议”。不是因为别的什么原因，而是这段话里包含的、传递的教育考试观念，同我们当中绝大多数人习以为常的、或当下教育界主流的、正统的教育考试观念相比，似乎是太过于“离经叛道”了！考题公开、即便题目没有做完只要答得有创见可以得满分、考试可以交头接耳、可以“问”、可以“抄”，诸如此类的观点，即便是今天看来，依然显得太过“大胆”、太过“超前”！但仔细想想，这里面难道就没有合理的、足以让我们对如今的教育评价系统进行深刻反思的成份吗？！

一、该考的不考、不该考的考了导致教育过程的扭曲与变形

在社会上、在教育界，经常会听到有关我国基础教育办得好与不好的各种评论，其中有这么一种看法相当普遍、相当流行：即认为我国的基础教育质量相当不错，只是高等教育、研究性教育比较落后；同美国相比，中国的教育“赢在起点”、“输在终点”。其理由是：同国外的学生相比，中国学生的基础知识、基本技能掌握得比较“扎实”、比较“牢固”，还有那么多的中国学生在国际奥赛上频频获得大奖，也是一个明证。这种看法有多大程度的合理性呢？国内教育界的学者、专家早有中肯的评论。例如，有学者从方法论的角度指出，要对不同社会或不同国家中小学的教学质量进行比较，首先必须要解决一个人才观或教学质量观的问题。如果仅仅用学生掌握“基础知识”和“基本技能”的扎实程度、牢固程度，作为衡量中小学教学质量高低的主要标准的话，那么，我国中小学的教学质量的优势是显而易见的。但如果我们把教学质量观转换一下，从培养学生的创新素质和实践能力的角度看。那么，美国中小学的教学质量就会远远高于中国。以科学教育为例，我国的科学教育在帮助学生掌握科学知识方面虽然是比较有效的，但由于我们的科学教育在很大程度上忽视了对学生进行科学态度与精神的教育，忽视了学生对科学过程与科学方法的体验与认识，因此，我们的科学教育是残缺不全的、存在重大缺陷的科学教育。此外，我国中小学生学习负担重，学习效率低，对学习和求知缺乏内在的兴趣，也很普遍。可以说，学生虽然掌握了大量的基础知识与基本技能，但为此却付出了诸如牺牲美好的童年、青春和健康，个性受到压制，创造性趋于泯灭等一系列高昂的代价。[1]又例如，有学者从思考基础教育的“基础性”角度指出，基础教育的“基础性”并不仅仅体现在，让学生掌握各门学科的“双基”，更重要的是，要为学生的终身发展打下良好的基础，这个“基础”主要体现在：终身学习的意识、学会学习的能力、探究精神、责任心以及适应社会和创造社会的能力；后一个“基础”之所以更为重要，原因就在于它是面向明天的，“仅仅靠昨天的知识与技能无法应对未来社会的挑战”。因此，根据这种理解，中国的教育同美国相比，我们在“起点”上就已经“输”了，理由是：我国的中小学教育是压制学生个性的、短视的、不可持续的教育，我们在强调“双基”的同时，却忽视了对知识的运用能力尤其是高层次思维能力的培养与发展；我国的基础教育引导学生

[1] 详见扈中平、刘朝晖着，《我国中小学教学质量真比美国高吗？》，载《中小学管理》，2001年第11、12期。

把大部分的时间与精力放在记诵昨天的知识上，而不是放在利用昨天的知识去创造明天的努力上；就书本知识的掌握与习得而言，中国的基础教育是“过剩”的，但如果就学生的“思维能力”、“人文精神”的培养而言，中国的基础教育却又是匮乏的、不足的。[2]

以上表明，有些人在对中美两国基础教育的质量进行比较时，之所以会得出错误的结论，根源就在于他们的基础教育质量观出了问题。基础教育质量观出了问题，意味着我们的教育评价系统肯定也跟着出了问题，因为有什么样的教育质量观，就会有什么样的教育评价；反过来也是如此，有什么样的教育评价，就会有什么样的教育质量观，二者是相互印证、相互参照的。基于这种理解，我们不妨从分析和反思我国现有的教育评价系统入手，看看我国中小学事实上正在追求什么样的基础教育。也许有人要说，对这个问题没有多大的讨论必要，因为我国的教育方针说得很清楚，中小学的教育目的是要使学生在智、德、体、美等几方面得到全面发展，或者说我们反对应试教育、追求素质教育。的确，这些说法并没有错，从理论上讲，从观念上讲，教育界、学界对基础教育应当干什么、追求什么并不胡涂，但无论专家学者对基础教育质量观的认识有多么的深入和正确，他们的认识毕竟只是停留在理论上、观念上，停留在“应然”的判断上，“应然”的判断代替不了对客观“现实”的描述。我们真正感兴趣的、真正关心的是，中国的基础教育“事实上”正在干什么、正在追求什么。要回答这一问题，只要我们仔细地去看看现实中我国中小学在对学生的学习进行评价时，倾向于“考”什么、不考什么，或者“评”什么、不评什么，就很容易从中探测出我国中小学“事实上”看重什么、关注什么、追求什么。

根据笔者长期的观察，愚以为，我国中小学教育评价在“考什么”、“评什么”的问题上，主要有两大问题，一是该考的东西没有去考，二是不该考的东西却考了。前一个问题主要体现在以下几个方面：

（一）在评价范围的分布上，偏重认知类教学目标的评估

同培养健全的人格、良好的个性心理质量相比，我们的教育评价系统更加看重的是文化知识的掌握和认知能力的发展。说我国中小学存在重“知”（智）轻“德”、重“理”轻“情”（情感、态度）的倾向，多少还是切中时弊的。因为我们很容易从学校教什么、考什么、评什么得出这种印象：课堂上教师一味地教学生知识，考试中一味地考知识，评价中一味地根据学生考得如何来评学生，这是绝大多数学校普遍存在的常态现象。尽管教育官员、学者们、教师们在理论上承认培养学生良好的品格，对学生进行情感教育很重要，但流行的教育实践、尤其是我们的教育评价系统却告诉我们，在学校教育中，真正重要的是掌握知识、发展智力，其它的东西虽说也很重要，但它们的重要性是第二位的。学生往往根据学校“考”什么、“评”什么来理解什么东西重要、什么东西不重要，教师也是如此。由于现实中绝大多数学校只是根据学生的考试成绩，来对教师的工作进行考核，来判断教师教得好还是不好；又由于学生的品行、情感发展不在学校考试之列，或者学校觉得这些东西难以考评，因此，本来非常重

[2] 详见华东师大陈玉琨教授 2002 年 9 月在由教育部中学校长培训中心与《文汇报》联合举办的“全国百名校长论坛”上的讲演“基础教育的再认识：兼论人材培养模式的改革与创新”。

要的东西，最后反而沦为可有可无之物，成为“说”起来重要“做”起来不重要，“忙”起来可以不要的东西。

情感、态度、价值观、行为习惯的学习，既是学校教育教学的重要目标，也是影响教育教学效果的重要变量。衡量一种教学有效与否、优劣与否，不能仅仅从认知的角度去观察、去判断，还必须从道德教化和审美体验的角度去观察、去判断。从某种意义上讲，学生在情感、态度、价值观、行为习惯方面的学习与发展，比认知方面的学习与发展来得更为重要。首先，从育人目标来看，不注重学生德性养成、情感发展的教育，是丧失“灵魂”的教育，“知识”只能教学生如何“正确地做事”（doing things right），却不能教会学生如何去“做正确的事”（doing the right things），前者只是针对“才”的教育，唯有后者才是“人”之为“人”的教育。其次，从学习取得成功的角度看，像动机、情感、态度、兴趣等非认知因素直接制约或影响着学习的成败。有许多人虽然智力上谈不上“天才”，但由于动机强烈、勤奋执着、持之以恒、永不放弃，最终也取得了巨大的成功。相反，有许多人本身很能干、很聪明，却由于缺乏兴趣或坚持，不敢迎接挑战，最终陷入平庸。所以，只教学生知识，不教学生如何做人；只关注学生认知的发展，忽视学生情感或品格的发展，这样的教育及评价系统是不完整的、残缺不全的，也是扭曲的教育。

当然，可能有人要说，认为我国中小学不重视学生的情感或品格教育，这是不对的，因为我国的教育历来都是强调“德育”为先，把培养学生良好的品格放在第一位。的确，我们承认这一点，但这种强调绝大多数时候只是停留在理论上、观念上。有时候，我们愈是在理论上、观念上强调德育的重要性，实践中的德育可能就愈是不受重视、不受关注。究其原因不外乎是，许多人认为德育是“软”性的、是不考的、也是无法考的东西，只有重大考试的分数才是“硬”性的东西。在应试教育的大环境下，“唯分数论”几乎是必然的。要改变这一点，光靠倡导“素质教育”起不了多大的作用，我们的教育评价制度必须改革、必须跟上才行。事实上，学生心理是否健康、人格是否健全、是否喜爱学习、敢于探索、具有社会责任感，像这样一些宝贵的质量，虽然很难通过传统的纸笔测验加以检测与评估，但只要我们对它予以足够的重视，总是能够找到适当而有效的办法对它进行评估的。比如，教师可以使用利克特自陈量表，让学生以匿名的方式对问卷中的提问作出反应；教师还可以对发生在自然情境中的、真实的学生行为表现进行观察，把他观察到的、他认为特别重要的或有意义的、典型的、发生在学生身上的事例或事件片段，用简短的文字记录下来；教师也可以运用同伴提名法，给学生呈现一系列简短的有关学生的行为描述，要求学生写下最适合每项描述的同学姓名（如我们班上愿意同其它人分享物品的同学是____，愿意帮助他人完成作业的同学是____，总是鼓励其它人做得更好的同学是____）；教师还可以运用访谈法，与学生进行面对面的、一对一的个别交谈，以获得有关学生行为的自述信息。在获得各种有关学生行为表现的评价信息后，教师就可以对这些信息进行综合判断，撰写有关学生行为表现的评语。这些评语年复一年地积累起来，放入专为每个学生建立的成长记录袋，就可以作为高校录取新

生的一个重要参考依据。

据说，美国大学在录取新生时，除了要看考生的SAT成绩、高中阶段的平时成绩外，还要看学生的综合素质，包括高中阶段的选修课程、参加的课外活动或社会活动、推荐信、开卷作文、以及做义工或打工的经历。美国虽然是一个崇尚个人自由、讲究个人权利的社会，但同时也是一个非常注重个人为小区提供服务的社会。美国的高中生能否获得毕业证书，除了学分、成绩等要求外，还要求学生能自觉自愿地参加社会公益活动，无偿地为社会服务一定的时间。比如，某高中规定学生提供社会服务的时间为：高一 5 个小时，高二 10 个小时，高三 10 个小时，高四 35 个小时。学生做义工的形式多种多样：有的到教会做事，有的做洗车工，有的到敬老院陪老人聊天，有的帮穷人修建房子，有的到建筑工地做搬运工，有的到幼儿园做“男阿姨”，有的为伤残人提供服务，有的甚至自掏腰包到贫穷国家搞扶贫活动，有的为灾区举办义捐、义演活动。除了做“义工”外，还要有一定的打工经历。在大学招办审核人员的眼里，穷人的孩子不打工，是要“扣分”的（因缺乏责任感），而富人的孩子打工，则可以“加分”（因自立精神）。[3]由此可见，对学生品格方面的发展进行评估，并不是做不到的问题，而是一个我们愿不愿意去做的问题。

值得特别强调指出的是，在对学生的情感和品格发展进行评估时，如何去评估只是一个手段和技术问题，把握好评估的价值导向，才是我们首先要解决的大问题。比如，我们讲，教育不仅要教给学生知识，更要教会学生做“人”，那么，我们要教会学生做什么样的“人”呢？在这个问题，不同的时代、不同的社会有着不同的回答。就现状而言，我国中小学在评价学生时，普遍倾向于把是否“听话”、“顺从”，当作评判学生好坏的一个主要标准。在家里，家长希望孩子乖巧、听话，到了学校，老师希望学生顺从、听话，那些有主见、有想法，遇事能独立思考、不盲从，敢于对书本、对教师、对权威的话提出质疑，凡事总要问个为什么、喜欢争辩的学生，往往被学校看作是“刺头儿”，看作是“不成熟”的表现。那么，听话的孩子就真的是好孩子吗？读一读历史上那些英雄豪杰、伟大人物的传记，你会发现，他们当中性格“叛逆”的好像并不在少数。其次，听“谁”的话，对“什么”顺从，如果家长、教师讲的话、制定的规则是错误的，孩子也要“听”吗、也要无条件地“顺从”吗？试想一下，如果我们把“听话”作为学生评价的取向，会导致什么样的后果呢？从社会方面来看，以“听话”、“顺从”为导向去培养学生，一定会造就一大批弱智无知、头脑简单、容易被操纵与控制、毫无创造力的“群氓”，这样的社会也一定是个保守、封闭、没有生气与活力、暮气沉沉，普遍趋于平庸、注定会走向衰败和没落的社会，因为社会的革故鼎新、进化与改良，最终总是与少数个体敢为天下先、特立独行、标新立异联系在一起；从个体方面看，以“听话”、“顺从”为导向培养出来的学生终究难成大器。过于听话的、对家长或教师言听计从的孩子，往往没有主见，“唯上”、“唯书”，缺乏独立思考与判断力。难道人们常说，“淘气的孩子有出息”，“听话的孩子没啥出息”。当然，说“听话”的孩子不见得就是好孩子，

[3] 黄全愈着，《“高考”在美国》，北京：北京大学出版社，2003 年版，第 54-58，92-93 页。

并不意味着“叛逆”的孩子，就一定是好孩子，那些任性妄为、横蛮不讲道理、处处以自我为中心，动不动就对别人发号施令、发脾气的“小霸王”，谁也不能说这样的孩子就是好孩子。所以，真正地说来，是否“听话”、“顺从”，根本就不能作为判断孩子好坏的标准，问题的关键是，我们要让孩子从小就养成服膺“理性”与“真理”、人人平等、相互尊重、民主协商、以理服人的意识与行为习惯。有了这样的意识与行为习惯，不论是“听话的孩子”，还是“不听话的孩子”，都是好孩子。

（二）在知识获得的评估上，偏重陈述性知识和显性知识的考察

1、偏重陈述性知识的评估

“陈述性知识”（declarative knowledge），主要是相对于“程序性知识”（procedural knowledge）而言的，它主要说明事物“是什么”、“为什么”和“怎么样”，是个人可以有意识地回忆出来的关于事物及其关系的知识，常以事实、概念、原理、法则、命题等形式出现。“程序性知识”则是关于“怎么办”或“如何做”的知识，或者说是关于完成某项任务的行为或操作步骤的知识，有时候，人们又把这种知识称为“实践知识”或“过程知识”。例如，关于怎样下棋、怎样钓鱼、怎样修剪花木、怎样弹奏一种乐器、怎样开车的知识，即是一种程序性知识。程序性知识常常镶嵌在特定的实践情境之中，并通过个体外在的行为显露出来，因此，人们常常把“程序性知识”称为根据某人会做什么而推知某人所具有的知识。人们通常所讲的各种操作步骤、实践技能，包括各种学习策略、问题解决策略，本质上都属于程序性知识的范畴。

陈述性知识与程序性知识的区分，与英国分析哲学家莱尔关于“know-how”与“know-that”的区分十分相似。莱尔所讲的“know-that”，往往表述为关于事物的各种命题，而“know-how”则往往表现为各种做事的知识。莱尔指出，“愚蠢”不等于“无知”，前者是由 know-how 方面的欠缺所引起的，后者则仅仅意味着个体缺乏 know-that 方面的知识。在莱尔看来，个体“智力”的核心在于拥有“know-how”方面的知识，而仅仅拥有“know-that”方面的知识，很可能是一个愚蠢的人。莱尔还认为，know-how 相对于 know-that 具有逻辑上的在先性，无论是发现还是拥有一种 know-that 知识，都要以 know-how 为前提；试图将“know-how”还原为、归结为 know-that 是错误的。由此看来，让学生掌握和拥有“程序性知识”，从某种意义上讲比掌握和拥有“陈述性知识”更为重要一些（因为后者才是培养和发展学生实践能力的关键因素），退一步来讲，两者至少具有同等的重要性。

然而，从现实的角度看，相对而言，我国中小学更加注重和强调的是，学生对陈述性知识的掌握，其主要依据是，在我国中小学，占绝对支配地位的教学方式仍然是教师的直接讲授与传递，这种教学方式对学生获得“陈述性知识”可以说是最适合不过的，而“程序性知识”的习得则不然，绝大部分程序性知识的习得都必须诉诸实践性的教学或活动性教学。后文将要讲到的，我国中小学普遍倚重纸笔测验来对学生的学习进行评估与考核，这一点进一步强化了上述倾向。从教育测量学的角度看，对陈述性知识的检测，采用

传统的纸笔测验（包括填空题、是非判断题、选择题、匹配题、简答题、辨析题、论述题等），是最有效的，但采用纸笔测验却无法有效地评估和检测学生对程序性知识的掌握。

2、偏重显性知识的评估

1958 年，英国化学家、哲学家波兰尼在《人的研究》这本书中首次提出了“隐性知识”（tacit knowledge）的概念。在波兰尼看来，人类所有的知识基本上可分为两类，即“显性知识”与“隐性知识”，所谓“显性知识”，是我们能够用语言、文字、数字和图表清楚地予以表达的知识，“隐性知识”则是一种“只可意会不可言传”的知识，是一种经常使用却又不能通过语言、文字或符号予以清晰表达的知识。隐性知识常常隐含在个体的行动之中，表现为个体的经验、眼光、趣味、技巧、诀窍、灵感、信念、习惯等，其主要特点是，难以编码与度量、不能以正规的形式加以传递、不易大规模储存和传播，具有非逻辑性、非公开性、前言语性、模糊性、个体性等特点；虽然它不能用言语明确地加以表达，但却可以在行动中被展现、被觉察、被意会。

波兰尼进一步指出，隐性知识是显性知识的基础，一切显性知识都有其默会的根源；人类所有的知识不是隐性知识，就是植根于隐性知识；显性知识不过是大树上结出的果实，而给大树提供营养的树根，则是隐性知识。从教育的角度看，教学不仅要致力于让学生掌握书本上那些明确的、客观的公共知识，还要致力于让学生掌握那些镶嵌于实践情境之中的、难以言表的经验知识或行为知识。因为一方面隐性知识是个体获得显性知识的基础，另一方面隐性知识还是个体增强实践能力的关键。要培养和发展个体的实践能力，光靠掌握外显的书本知识是远远不够的，还必须掌握大量的“实践知识”~ 即隐性化的程序性知识才行。人们常说的“高分低能”现象，在很大程度上就是由于学生掌握的知识主要局限于显性化的书本知识，缺乏相应的实践经验造成的。前文说到，我国中小学教育实践过于注重陈述性知识的习得及其评价，这样倾向势必也会导致中小学对显性知识的掌握及其评价的“一边倒”，因为“程序性知识”在很大程度上也往往表现为隐性知识。当然，隐性知识因其自身的特点难以检测与评估，也间接地导致了教育者对它的忽视，但这绝不意味着隐性知识就不能被检测与评估，只不过检测与评估的手段不同于传统的纸笔测验罢了。

（三）在智力发展的评估上，偏重于低层次认知能力或单一类型智力的评估

1、偏重低层次认知能力的评估

前面讲到我国中小学比较注重书本知识的获得、认知能力的发展，这只是相对于学生健全人格的形成、正常情感的培养而言得出的结论。其实，就知识的获得或认知能力的发展而言，我国中小学的教学及其评价系统也存在相当的局限。为了说明这个问题，不妨先来回顾一下美国学者布卢姆有关教育认知目标的分类思想。布卢姆关于教育认知目标的分类早在上个世纪六十年代就已经提出来了，不过早期的分类在实践过程中被证明存在各种缺陷。后来，安德森（Anderson,L.W.）等人对布卢姆原来的分类框架作了较大的修订，新的分类框架不同于原来一个维度的分类（即把所有的认知类教育目标分为“知识”、“领会”、“应用”、

“分析”、“综合”和“评价”），而是采用了“知识”和“认知过程”二个维度的分析框架。“知识”涉及学习的内容或结果，依据“从具体到抽象”的顺序分为四个类别，即事实、概念、程序和元认知（学习策略）；“认知过程”涉及学习的过程及心理能力，依据认知的复杂程度不同，“从低到高”分为六个类别：即记忆、理解、应用、分析、评价和创造。这套分类框架对教师如何搞好课堂教学，在许多方面都具有启示意义。例如，对不同层次、不同类型的认知目标，需要采用不同的教学方法（例如若要帮助学生达到和实现高层次认知目标，就不能仅仅采用像“讲授”这样单向交流的教学方法，必须采用师生双向互动交流的教学法）。又例如，在课堂提问方面，教师若要培养学生高层次的认知思维能力，在课堂上他就不能只提一些低认知水平的问题。所谓低认知水平的问题，就是那种较少要求投入思考，对学生思维的挑战性不大，学生只需要记住教师在课堂上所讲的内容，或者只需要翻翻课本，就能从书本上找到现成答案的问题。大量的课堂观察和课例研究表明，在我国中小学，传统的“满堂灌”（教师一讲到底）正在被“满堂问”（边讲边问）所替代，高密度的提问（一节课教师的课堂提问数竟然高达一百多个）已成为课堂教学的重要方式，但这种高密度、高频率的课堂提问并没有真正激发和调动学生高水平的思维技能。上海的知名学者顾泠沅教授把教师在课堂上的提问分为五类，即常规管理性问题、记忆性问题、推理性问题、创造性问题和批判性问题，据他观察，在绝大多数的中小学课堂上，教师的课堂提问记忆性问题居多（74.3%），推理性问题（21%）次之，极少有创造性、批判性问题；提问后让学生齐答或举手回答的比例很高，学生主动表达自己观点或向教师提出问题的极少出现，课堂几乎完全为教师所控制。[4]同样的现象在教师课后布置给学生的作业中也有明显的体现：绝大多数教师布置给学生的课后作业，往往侧重于对课堂所讲知识的巩固与运用，学生只要提前预习一下，上课时认真听讲、记好笔记，课后及时复习，不需要投入多少真正的思考，就能轻而易举地完成课后作业。当然，在课堂提问和布置作业上，教师能针对事实、概念、原理或规则的记忆、理解与运用进行提问、布置作业，还算是不错的，毕竟这也是常规教学一个主要的任务。但问题是，如果所有的教学都停留在记忆水平或理解水平上，学生的自主学习能力、探究能力、创造能力，包括批判性思维能力，就难以得到刺激、锻炼与发展。根据教学对学生思维投入与运用的挑战性程度的不同，人们常把教学分为三种水平，即记忆水平、说明理解性水平和自主探究水平，可以说我国中小学绝大多数的课堂教学都停留在前两种水平上，只有极少数达到了第三种水平，这对我们培养高层次的创新人材是极为不利的。

要改变上述倾向，就必须对教师的课堂提问、作业布置及各种考试（包括课堂测验、单元测验、期中期末考试、乃至中考高考等）进行联动性的改革。课堂提问、作业布置本质上也是一种教育评价，教师喜欢提一些什么样的问题、给学生布置什么样的作业，对学生的回答或提交的作业倾向于做出怎样的响应，这些都在发挥评价导向作用。当然，正规考试的导向作用对学生的影响最大。为了引导学生全面达到和实现各种层次的认知目标，在编制有关

[4] 沈兰、郑润洲编，《变革的见证：顾泠沅与青浦教学实验30年》，上海：上海教育出版社，2008年版，第121-124页。

课程的教育测验时，试题的编制者就必须注意和做到，在一张试卷中，用来考察高、中、低三种不同层次认知目标的试题都要占有一定的比例，只有这样才能引导教师搞好教学，也只有这样才能通过考试，真实、全面地反映学生的认知发展水平以及教师的教学绩效。有些教师在课堂上很注重通过课堂提问、作业布置等教学环节，培养学生高水平的思维技能，但如果我们的教育测验和考试只偏重对学生低认知水平教学目标的考察，那么，这种教育评价就会对教师的教学评价起到一种误导作用，即真正的好教师、优秀教师反而受到较低的评价，而那些平庸的、中规中矩的教师反而受到较高的评价。据说在课新改中，那些按新课改的要求去上课的教师，其学生反而考不过那些按老方法（教师讲学生听、大量做习题做练习等）教出来的学生，这只是一种假像，问题很可能出在我们的教育评价系统过于偏重低认知水平教学目标的考察上。

2、偏重单一类型智力的评估

第三条中讲到，我国中小学的教育评价系统通常只能用来考察学生对低层次认知类教学目标的达成度，而不能用来考察学生对高层次认知类教学目标的达成度。相应地，我国中小学的教学只适合用来培养学生中、低层次的认知技能，不能有效地用来帮助学生获得高层次、高水平的认知技能，从这个意义上讲，我国中小学是不太重视学生智力发展的。当然，这个现象在上个世纪八十年代有所改变：当时教育界在讨论教学的目的与任务时，就十分明确地提出教学不仅要让学生掌握各门学科的“双基”，还要发展学生的智力，再到后来还提出了培养和发展学生“非智力因素”的问题。其实，从理论上讲，对教学应发展学生的智力这一点，大家都有共识。问题在于，大家对教学所要发展的“智力”本身究竟是什么，并不是十分的清楚。在我国，心理学界一般认为，“智力”是一种综合的认知方面的心理特征，是人认识、理解客观事物并运用知识、经验等解决问题的能力，包括注意力、记忆力、观察力、想象力、思考力等。不过，在国际上，人们对智力的看法并不一致。比较典型的智力理论有：卡特尔的流体智力与结晶智力划分理论、伽德纳的多元智力理论、珀金斯的真实智力理论、斯藤伯格的成功智力理论、萨洛维和梅耶尔的情绪智力理论等，所有这些智力理论尽管对智力的看法不一，但也有一些共同的倾向：它们都对传统的智力概念提出了批评，认为传统的智力概念只涉及内涵宽广、结构复杂的人类智力的极小一部分，而且也是相对来讲不太重要的一部分，这种智力概念不能反映人类心理能力的多样性与复杂性。相反，它们都认为，人类的智力并不是某个单一的实体，而是多元化的，存在不同类型、不同性质的智力；不能简单地用某个单一的、统一的标准去衡量学生智力的高低，用伽德纳的话来讲就是，“问题不在于一个人有多聪明，而是他怎样聪明、在哪个方面聪明。”

如果我们用这种观念去看待学生智力发展的话，就不难发现，传统学校教育实践所奉行、所采纳的智力观，基本上是一种单一的智力观，这种单一智力观主要体现在：以言语智力、数理逻辑智力为核心，强调分析思维。按伽德纳的观点，在每个个体身上都存在七到八种不同类型的智力，如言语智力、数理逻辑智力、视觉空间智力、身体动觉智力、音乐智力、人

际智力、内省智力、自然观察智力；不同的个体擅长不同类型的智力，每个个体都有自己的智力强项或弱项；但令人遗憾的是，当前的学校教育实践，不论是教学系统还是评价系统，对学生而言，真正受到关注、鼓励和强化的、得到发展的只是传统的、经典的言语智力和数理逻辑智力，其它类型的智力往往受到忽视。斯腾伯格也有类似的看法：他认为，人类的思维本来有三种类型，即“分析思维”（司法型思维）、“实践思维”（执法型思维）和“创造思维”（立法型思维），“分析思维”主要用来分析、评论与评判，“实践思维”涉及到知识的使用与运用，主要用于（方案或计划的）执行与实施，“创造思维”主要用于规划、设计、发明与创造。斯腾伯格指出，个体能否在社会生活中取得成功，关键取决于他能否平衡使用上述三种思维能力。因此，学校教育必须关注和强调上述三种思维能力（或智力）全面的、均衡的发展，而不能只钟情于其中的某一种。但现实的情况却是，当前的学校教育实践倾向于不断强化和奖励学生的分析思维（分析能力），忽视了对人生的成功恰恰十分重要的实践思维（实践能力）与创造思维（创造能力）的培养与发展。在学校里，学生们花大量的时间从事阅读、听讲和记笔记，消化和巩固老师和书本告诉给他们的信息内容，并在测验中回馈出来。教师根本不关心学生的智力在真实生活中的运用，也不关心学生对新知识的探索。这一点不仅反映在教学上，也反映在各种类型的学习评估与测验上。通常我们的教育测验只能反映出学生对静态的、死的书本知识的掌握程度，不能反映学生运用知识和使用知识的能力，也不能反映学生发现新问题、提出新看法、解决新疑难的能力。

偏重经典的言语智力或数理逻辑智力，忽略其它类型的智力，偏重分析思维，忽视实践思维和创造思维的教学及其评价取向，其后果是相当严重的。本来，天才有许多种，不同的学生有着不同的聪明方式，如果我们的学校只推崇某个单一类型的智力，用一个标准去衡量学生，势必会导致对相当一部分学生真实智力的严重低估，并给他们烙上与其实际情况不符的“愚笨”标记，从而导致这些学生内在的优势潜能得不到承认、开发与利用，从而，上天赐与给他们的优异才能被我们的教育系统给活活地“掩埋”和“扼杀”了，这无论是对学生个体，还是对整个社会都是一个巨大的损失与浪费。要改变这一点别无它法，只有从改变评价任务和评价标准入手。比如，在课堂提问的教学环节上，教师可以有意识地将书本学习与现实生活联系起来，问一些涉及知识使用的应用性问题，或者问一些没有固定答案、较为复杂的开放性问题。又比如，在作业布置的教学环节，教师可以尝试在自己的学科教学中，给学生布置涉及各种类型思维运用的作业任务。以语文课为例，在这门课的教学，教师可以要求学生运用想象力，给《呼啸山庄》安排一个现代结局；或者要求学生读完《斧头》一书后，考虑如何运用从作品中学到的技巧，帮助自己在野外求生；或者让学生分析一下《夏洛特的网》这本书的作者，为什么要把书名取为“夏洛特的网”。上述三个作业任务要求的思维类型各不相同，它们分别要求学生运用创造思维、实践思维和分析思维，这样的作业布置既能让学生展示其思维的长处，又能让学生思维的“短板”得到弥补，两者的结合即是《学记》中所讲的“长善救失”。除作业布置以外，在各门学科的考试测验中，教师有意识地加入少

量用以考察学生批判性思维、创造性思维的开放性试题，看看学生思维的广度与深度、思维的严密性与独特性如何，并据此对学生的学习质量做出评判，也不失为一个好的评价办法，至少它比原来那种只考一些封闭性的、有着固定答案的书本知识，对促进学生思维类型的多样化发展要好得多。当然，要做到这些，各种教育测验的评价标准也要跟著作相应的调整才行。比如，在美术课中，让学生画一幅画，我们不能光看学生的“画”技如何，还要看学生画得有没有创意、有没有想象力。

现在，让我们回到第二个问题~“不该考的东西却考了”。这个问题主要体现在两个方面：第一是，有的东西属于该考的范围，但考的时机与场合不对。比如，许多家长或幼儿园急于求成，在孩子很小的时候就开始教给他们大量的知识（如识字、阅读、计算），美其名曰注重早期的“智力开发”，不让孩子输在“起跑线”上。一些所谓的名牌小学招生时也要考知识，逼得幼儿园也不得不大量地教知识，这种做法在国内早已司空见惯，而且有愈演愈烈之势。尽管中国古代早就有“揠苗助长”的故事，但在现实中，许多家长、幼儿园都不认为他们的做法是“揠苗助长”。“在学前阶段，孩子面临的最重要的挑战是发展感情和社会技能，即怎么和别人相处、怎么在陌生人的环境中保持情绪的稳定，而非读写算的能力。”[5]第二个表现是，在各种外力因素的综合作用下，学校里大大小小、各种各样的教育测验与考试的内容，开始趋于琐碎化或无意义化。自从引入计算机改卷和标准化考试后，试题越出越偏、越出越怪、越出越难、越出越烦琐，让学生摸不着头脑，出题的人好像故意要为难学生、折磨学生、捉弄学生似的。比如，在语文课上，要求学生分析某个句子有什么深刻的含义啦，为什么用这个词不用那个词啦，加点的词有什么作用啦，某个东西象征着什么啦；这篇短文可以分为几层、几段，每一层、每一段的意思是什么啦；这是什么词性、什么修辞、什么语法结构啦，诸如此类的考题，对学生来讲究竟有多大的意义呢？还有一些题目莫名其妙，简直没法做，像解释“灌”与“溉”有何不同、“酸溜溜”是什么意思、用李大钊的“钊”字组词，诸如此类的题目足以让人抓狂、感到恐怖！为什么会出现这种情况呢？有些人把这种现象归结为标准化考试，在笔者看来，这倒在其次。真正的原因是，教材的编选者、试题的编制者包括我们的教师，对学校里各门学科究竟应当教什么、应当考什么，对各门学科独特的教育价值，对我们究竟要培养什么样的人，缺乏深刻的认识与理解，因而，编教材的人把大量没有多大教育价值的平庸之作选入教材；出试题的人拿一些根本没有多大价值、毫无意义的东西去考学生，也就不奇怪了。

二、评价的标准、方式方法不当导致评价结果的失真与无效

（一）关于评价标准的不当

1、评价标准泛政治化

评价标准泛政治化，在语文科的教学中体现得最为明显。应当承认，语文这门学科既是

[5] 薛涌着，《一岁就上常青藤》，北京：中国青年出版社，2009 年版，第 24 页。

一门语言类的工具性学科，同时也是一门对学生的情感、态度、价值观具有重要影响的审美学科或人文学科。作为一门人文学科，入选教材的课文内容理应具有较高的人文价值，但这里所讲的“人文价值”不能仅仅从意识形态的灌输方面去理解，我们必须把对学生进行人文教育，提升学生的“人文素养”，与对学生进行思想政治教育区别开来，二者不是一回事。但在现实中，我们常常把二者混为一谈。一个最明显、最典型的表现就是，我们常常自觉或不自觉地用一套固定的、刻板的、教条化的话语系统，去阐释和解读课文作品的意义。比如，无论是什么文学作品，我们都倾向于用反对封建主义、批判资产阶级、同情人民大众等来概括，或者从政治的角度和高度去分析它所包含的阶级局限或消极面等。比如，学朱自清的《荷塘月色》时，教师强调说，这里有一点小资产阶级知识分子的苦闷情绪，要正确对待；学苏轼的《赤壁怀古》时，老师往往强调指出，“人生如梦”调子过于低沉，过分悲观，不可取；学莫泊桑的《项链》时，老师告诉学生，这篇作品是对小资产阶级的虚荣心或享乐主义思想的讽刺；讲到老舍的《骆驼祥子》时，老师就告诉学生，这篇作品告诉我们，个人奋斗没有出路。即便是对鲁迅作品的阐释，也存在类似的倾向：不少分析都是基于上个世纪五六十年代的“鲁迅观”，其中不少分析是对鲁迅作品的简单化、机械化、庸俗化的理解，甚至是曲解。有的作品即便本身是描写大自然的，我们也要人为地给它赋予某种政治上的意味。比如，李健吾的《雨中登泰山》这篇文章，中间有一段文字是描述泰山上的“松树”的，说有的松树“把根扎在悬崖绝壁的隙缝，身子扭得像盘龙柱子，在半空展开枝叶，像是和狂风乌云争夺天日，又像是和清风白云游戏”；有的松树“望穿秋水，不见你来，独自上到高处，斜着身子张望”；有的松树“像一项墨绿大伞，支开了等你”；有的松树“自得其乐，显出一副潇洒的模样”。这段文字本身没啥深奥的，作者运用比喻、拟人等修辞方法，表现了松树的千姿百态、各具情趣的自然景色，但在一个有关这段描述的考试选择题中，凡是学生作这种理解的都被判为不正确，标准答案是，“作者运用比喻、拟人等修辞方法，表现了松树在逆境中奋斗的自豪感和旺盛的生命力”。[6]可见，在试题的编制者眼里，泰山上的松树都是政治生物，都在“逆境”中“奋斗”、“自豪”！像这种牵强附会的解释，在语文课的教学和考试中，绝对不在少数。

评价标准的泛政治化，另一个突出的体现是学生的作文。有人说，我国中小学生的作文是“假话、假感想、假故事大全”，不仅写作的题材大同小异，而且，连话语表达方式也都差不多。不信的话，请看以下这些我们小时候写作文惯用的话语：

无数革命先烈抛头颅,撒热血，才换来了我们今天的幸福生活……

今天,我们怀着无比沉痛的心情来到了烈士陵园……

“小朋友，谢谢你，你叫什么名字”,“我叫红领巾”。

“每当遇到困难想退缩时，脑海中忽然闪过 xx 的身影，和他们比起来，我的这点困难算什么!”

[6] 王丽编，《中国语文教育忧思录》，教育科学出版社，第 14-15 页。

我们一定要继承和发扬 xx 的革命精神与优良传统，……

自从十一届三中全会以来……改革开放的春风吹遍了神州大地……

我暗暗下定决心，长大以后一定要……

我一定要努力学习，不辜负 xx 的期望，做一个又红又专的少先队员（或社会主义的接班人）……将来为四化建设（或共产主义事业）添砖加瓦！

上面列举的这些典型的作文话语，相信许多人都不陌生。成年后再回过头来看看当年写的这些东西，一定会觉得很可笑，但当时浸润于其中的“我们”，可能并不觉得这有什么问题。也许有人要说，这是当时那个特定的历史时代造成的，时代在前进、社会在发展，今天我们的孩子已经不这样写作文了。确实，我们要承认这一点，今天中小学生的作文话语系统比以前更加多样化、个性化，但在某些时候、某些方面，历史的“遗风”还会时不时地再现。比如，今天绝大多数中小学还会倾向于把“思想健康”、“积极向上”作为评判作文写得好与不好的一个重要标准。一个孩子写的作文哪怕从语言运用与表达的角度写得再好，如果“思想境界”不高，没有与“主旋律”挂钩，肯定是拿不到高分的，有时候甚至还会得不及格的分数。试看以下的例证：

[例 1]：《记一件有趣的事》

第一篇作文题目是《记一件有趣的事》。小男孩酷爱足球，他开篇就说他认为踢足球是最有趣的事，然后描写他踢球时的愉快，球场上一些精彩的细节，还穿插着写了两个他崇拜的球星。看起来他对这些球星的情况了如指掌，写得津津有味，如数家珍。

男孩的这篇作文写得比较长，语言流畅，情真意切，还有一些生动的比喻。看得出他在写作中投入了自己的感情。虽然整个文章内容与标题框定的处延略有出入，总的来说属上乘之作。我从头看到尾正要叫好时，赫然看到老师给的成绩居然是“零”分，并批示要求他重写。

我万分惊讶，不相信作文还可以打零分，况且是这样的一篇佳作。

赶快又往后翻，看到男孩又写了一篇相同题目的。他妈妈在旁边告诉我，这就是在老师要求下重写的作文。

这次，“一件有趣的事”变成了这样：踢球时有个同学碰伤了腿，他就停止踢球，把这个同学护送到医务室包扎伤口，又把同学送回家中，感觉做了件好事，认为这是件有趣的事。这篇文章的字数写得比较少，叙事粗糙，有种无病呻吟的做作。老师给出的成绩是 72 分。

朋友告诉我，这一篇内容是儿子编出来的，因为孩子实在想不出该写什么。但凡他能想到的“有趣”的事，除了足球，都是和同学们搞恶作剧一类的事情，他觉得老师更不能让他写那些事，只好编了件“趣事”。[7]

上述例证在现实中绝非孤立的个案！通过这个例证不难发现，我们仍然在用过去那种泛政治的、追求“高大全”的思维定势在要求学生。其实，对一个孩子来讲，他的任何表达只

[7] 尹建莉着，《好妈妈胜过好老师》，北京：作家出版社，2009 年版，第 101-102 页。

要是真实的自我表达，就根本不存在“健康”与“不健康”的问题。儿童有儿童的思维特点、心理特点，儿童思想的成熟、品格的发展有它自身的节律与规律，我们不能用成人的标准去衡量他们、要求他们，脱离儿童精神发展的实际需要，强迫他们去写一些既不理解、又不感兴趣的话题（比如，写一件反映社会主义新风尚的事），或者引导他们用流行的意识形态话语去表达自己的思想情感，以符合“思想健康”、“积极向上”的标准，只会引导孩子从小就学会说假话，这样做的后果是，不仅作文写不好，连人格的培养也不要了。

2、评价标准泛科学化

把人文学科当作理工科考，强调标准答案只有一个，忽视了人文理解的多元性、多义性与模糊性特点，这是当前我国中小学教育评价中一个突出的特点。例如，语文这门学科就具有很强的人文性、审美性，要学好这门学科，仅仅诉诸科学的分析思维是远远不够的，更多的时候必须诉诸艺术形象思维，诉诸个人的情感体验与移情理解，诉诸人与人之间的互动交流。相应地，在学生学习的评价上，我们就不能用适用于理工科的科学思维及其评价方式（其主要特点是追求思维的确定性与严密性，强调标准答案），简单地搬到、套到象语文这样的人文学科上，因为科学知识与人文知识分别代表不同的意义类型，它们的获得与证明分别建立在不同的思维形式、不同的检验标准之上。例如，历史上包括到现在还有不少人以为，将童话、神话和寓言故事纳入课程教材是不妥的，因为这些东西充斥着“猫话狗话”、“鸟言兽语”，有违科学的精神，会把儿童培养成“猫化狗化的国民”，对儿童的身心发展不利，因此，建立在教材中取缔它们。这种观点当然是杞人忧天，当然是荒谬的，因为它根本不了解儿童，不了解儿童在如何成长与发展的，不了解儿童的精神世界、思维方式与成人世界的差异！其实，童话、神话才是儿童真正需要的精神食粮,童话的世界和儿童的精神世界有一种天然的契合，它可以满足儿童深层的心理需要（比如，以浪漫的、游戏的、富于幻想的眼光和心态去看待世界）。[8]由此可见，我们不能用“科学”的眼光去评判文学作品。但可悲的是，在现实中，有些教师经常犯这样的错误，试看以下例证：

[例 2]：为什么“青蛙”和“蛇”没有出来？

在小学低年级的一节语文课上，教师正在带领学生学习“小画家”一课。该课文的主要内容是说，冬天下雪了，大雪将整个原野都覆盖起来。清晨，小鹿、小鸡等小动物们都出来了，纷纷用自己的足或爪子在雪地上画出了美丽的图画。教师在完成了教学任务以后，向学生们提了一个问题：为什么“青蛙”和“蛇”没有出来？不一会儿，有一个学生站起来回答说，“老师，因为青蛙和蛇没有毛衣服，怕冷，所以呆在家里没出来。”老师听了以后很不高兴，用非常严厉的口吻说，“不知道就不要乱说！”

[例 3]：造句[9]

在某小学，教师要求学生用“想”、“活泼”、“悄悄”、“丢”这几个词造句。学生给出的

[8] 刘晓东：《“猫话狗话”、“鸟言兽语”：儿童成长的精神食粮》，载《天津师范大学学报（基础教育版）》2003 年第 4 期。

[9] 扈中平着：《教育目的论》，武汉：湖北教育出版社，1997 年版，第 259-260 页。

造句是：想——我想听到开花的声音；活泼——河里的水很活泼；悄悄——我们听不懂小鱼的悄悄话。丢——上街时，毛毛把爸爸丢了。结果，教师都给它们打了一个大大的“×”。

在上述两个案例中，教师都犯了同样的错误，那就是把“文学课”（语文课）混同为“科学课”。在第一个案例中，教师对“青蛙和蛇为什么没有出来”给出的答案是，“青蛙和蛇是冷血动物，冬天需要冬眠”；在第二个案例中，教师之所以不认可学生的造句，理由无非是：开花不可能有声音，活泼只能用来形容人，小鱼不可能讲话，孩子把爸爸丢了不符合生活常识，这些理由都是站在“科学”的立场上来讲的，全然不顾语文课的学科独特性。从文学的角度看，前后两个案例中的学生反应，都充分体现了儿童的童心、童趣和特有的想象力，判它们为错，实在是对语文课学科性质的误读。

3、迷信标准答案

对于教育考试与评价，在许多人的脑子里都存在这么一种思维的定势：即凡是作为教育考试和评价的内容都必须具有客观的必然性与确定性，只有那些具有唯一正确答案的问题，才能纳入到中小学的教育考试与评价中去。对许多人来讲，把答案不固定或具有多种可能答案的问题，纳入到中小学的教育考试与评价中去是不可思议的，那样的话必然导致无效的教育评价。我们认为，这种思维的定势必须破除：要真正检测学生在学校里学到了什么、发展得怎样，既要用那些稳定的、客观的、具有确定性的知识去考他们，同时也要留出适当的空间，用那些在书本上找不到现成答案，学生必须综合运用所学知识，根据实际情况作出灵活应对的开放性试题去考他们。从某种意义上讲，采用后一种考试更能考出学生的实际发展水平。我们根本没有必要去担心：用开放的、答案不确定的问题去考学生，会导致考试效度与信度的丧失。理由是：用开放性的问题去考学生，尽管学生对问题的解答存在某种程度的不确定性，但我们依然可以制定出相对合理的“评价标准”，来对学生反应的高下、优劣作出评判。比如，从学生的反应中，我们可以看出，一个学生思考的视野比另外一个更广阔；论证更清晰、更严密、更有说服力；提出的观点更独特、更新颖、更有价值。其次，面对一个答案不确定的开放性问题，学生可以站在不同的立场、从多个不同的视角去思考它，得出不同的结论，承认这一点并不意味着学生得出的任何结论、作出的任何反应都一样“正确”、一样“合理”，答案的“多元化”与“随意化”是两个完全不同的概念，承认一个问题的答案是不固定的、是多元化的，并不意味着这个问题的答案是可以任意给定的。比如，在语文课中，教师可以鼓励学生对文学作品的意义作多元化的解读，但这种多元解读绝对不是随意化的、胡乱的解读，因为任何一种解读都必须有它的根据，都必须联系“文本”产生的特定社会历史条件以及作者个人特定的生活经历来进行解读。这也就是说，即便是在人文学科的学习中，学生对“文本”意义的自由解读也是有边界的。

据此，在人文学科的教学及评价中，适当加入一些开放性的考试内容，是完全可行的。我们根本不用担心，以开放性问题作为考试内容会影响到评分的客观性，我们可以采取许多技术手段来避免这一点。比如，要求出题者事先拟好合理的评价标准，列出各种可能的答案，

采用多人评分制避免评分的主观性，等等。其次，不论是在哪一门学科的教学及评价中，尤其是在像语文、历史、社会等这样的人文学科中，教师应当鼓励学生对所学文本作出异质的、多样化的反应，只要这种反应是有根据的、能够自圆其说的。久而久之，学生就会慢慢地知道，在这个世界上，并不是所有的问题都有答案；在某些领域，真理并不只有一个，真理可以是多元的、不确定的。令人遗憾的是，现实中有这种意识的老师并不多见：不要说积极主动地去引导学生运用不确定性思维、相对性思维，去思考各门学科领域的开放性问题；有时候，学生自发地表现出对开放性问题的探索兴趣，教师还会阻止学生去思考它们，理由是这样做浪费时间，对考试提高分数没啥好处。更为恶劣的是，有时候，学生对某个问题的回答明明是正确的、合理的，教师也将它判为“错误”，惟一的理由仅仅是，与标准答案不符！试看以下的例证：

[例 4]：《三个抄写员》

黎锦熙（1890-1978）是我国著名的国学大师。民国头十年他在湖南办报，当时帮他誊写文稿的有三个人。

第一个抄写员沉默寡言，只是老老实实地抄写文稿，错字、别字也照抄不误，后来这个人一直默默无闻。

第二个抄写员非常认真，对每份文稿都先进行认真仔细地检查然后才抄写。遇到错字、病句都要改正过来。后来，这个抄写员写了一首歌词，经聂耳谱曲后命名为《义勇军进行曲》，他就是_____。

第三个抄写员则与众不同，他也仔细看每份文稿，但他只抄与自己意见相符的文稿，对那些意见有悖的文稿则随手扔掉，一句话也不抄。后来，这个人建立了以《义勇军进行曲》为国歌的中华人民共和国。他就是_____。

在阅读上述短文的基础上，回答以下的问题：

（1）还有一个抄写员我们不知道他的名字，那是因为他_____。

（2）简要概括出三个抄写员的特点：_____。

对问题（1），某学生给出的回答是，“后来一直默默无闻”，教师判“错”，因为标准答案是，“老老实实地抄写文稿，错字、别字也照抄不误”。对问题（2），某学生给出的回答是，“第一个抄写员沉默寡言，老老实实；第二个抄写员对待工作认真仔细；第三个抄写员与众不同，十分有志气”，教师同样判“错”，因为标准答案是，“第一个抄写员老老实实地抄写文稿，错字、别字照抄不误；第二个抄写员认真地抄写，把错字、别字都改过来；第三个十分有主见，意见有悖的文稿一律不抄。”

（二）关于评价方法的不当

1、纸笔测验一统天下

教育评价是一门科学，有什么样的评价内容或对象，就必须采取与之相适应的评价方法。例如，对于陈述性知识的评估，采用传统的纸笔测验方式，包括填空题、判断题、选择

题、匹配题、简答题、辨析题、论述题等，是比较有效的。但纸笔测验并不适合用来评估和检测学生对“程序性知识”的掌握。对“程序性知识”，常见的评估方法采用“表现性评估”（performance assessment），也就是对学生实际从事某项活动、或完成某项任务的具体行为表现进行观察，然后依据事先制定的评价标准对学生的学习进行评价。“例如，了解莎拉能否骑车的惟一方式就是看她骑车的表现，仅仅让她告诉你如何骑车并不足以判断她是否会骑车，你说你知道如何管理课堂行为？那好，让我们看看你到底是如何做的！你说你可以教会学生如何做减法，那么请演示给我们看看！”[10]这便是人们常讲的所谓“表现性评价”。

表现性评价的根本目的在于，考察学生将知识和理解转化为实际行动的能力。学生能够背诵课堂规则，并不意味着他们在课堂环境里能够执行课堂规则；学生能够写出一连串确保实验安全的操作步骤，但这并不意味着在特定的环境里，他们就能够真正地展示这种操作技能；学生会描述如何去做某件事与他们真正会做那件事，是有区别的。用文字描述如何做各种跳水动作，或者回答一些关于跳水规则的选择题，都不是评估跳水动作的合适方式，有效的跳水评估只能看跳水者的实际表现。表现性评价最大的优点在于，它可以在真实情境中检测出学生对程序性知识（也即实践技能）的掌握程度。象“交流技能”（写作、朗读、演讲、辩论）、“操作技能”（握笔、组装仪器、使用剪刀、解剖青蛙、打字）、“运动技能”（如罚球、蛙泳、跨栏等）、“社会技能”（如共享玩具、小组合作、遵守校规、自我控制等）等技能学习的评估与检测，只能诉诸表现性评估，采用纸笔测验是无效的。

纸笔测验的局限性还不只这些。近几十年来，心理学界有关学习机制的研究表明，个体的学习有两种基本形式，即外显学习与内隐学习。前者是指那种有目的、有意识、需要付出努力的学习，后者则是一种诉诸无意识认知加工的学习过程。例如，儿童无需系统地、有意识地学习语法规则，就能在母语的氛围中不知不觉地学会说母语，这就是内隐学习的典型体现。内隐学习的主要特点是，学习是在无意识中自动发生的；学习的结果主要表现为个体觉察不到或虽然觉察到了、但难以言表的缄默知识或隐性知识；学习过程不受或较少受年龄、智力、心理异常和脑损伤的影响，具有高效性与抗干扰性。从有效的教学角度看，好的教学应当想办法引导学生在课堂上同时使用上述两种认知机能系统进行学习，但问题是学生内隐学习的结果（表现为隐性知识），我们有什么办法对它进行评估与检测呢？采用传统的纸笔测验肯定不行，唯一可行的办法还是诉诸实际的“操作”：从内隐学习的角度来看，所有要求学生用言语“说出来”的评估方法都属于直接测验，它只能测出学生外显学习的结果，而不能测出学生内隐学习的结果。对内隐学习的结果，只能用间接的方式去检测，也就是通过观察个体实际上会“做”什么、或“做”得怎样，来间接地推测他所掌握的隐性知识。更重要的是，在诉诸言语表达的测试中取得优异成绩的人，其实际的操作能力并不一定高，反过来也是如此，因为人们所做的与所回答的很可能没有多大的相关。所谓“高分低能”、“低

[10]（美）Gary D.Borich & Martin L.Tombariaf 着，国家基础教育课程改革“促进教师发展与学生成长的评价研究”项目组译，《中小学教育评价》，北京：中国轻工业出版社 2004 年版，第 126 页。

分高能”，在此均可以得到合理的解释。[11]这就提示我们，书面考试并不是衡量学生学习质量高低的唯一手段，必须把书面测试与诉诸“实际操作”的表现性评估有机地结合起来。

2、评价方式过于刻板、僵化

评价方式的刻板、僵化主要体现在，学校里流行的教育测验与考试绝大多数都是封闭的、限时的、去情境化的纸笔考试。这种考试当然在某些范围内确实是有效的，但如果学校所有的重要考试都是这种类型的，肯定有问题。先看限时的闭卷考试，这种考试要求学生快速答题，这对那些反应快，习题做得多、经过反复训练的学生是很有利的，但这种考试最大的弊端是，它不允许学生深思熟虑、仔细地思考。至于闭卷考试，从好的方面讲，它很适合用来检测那些稳定的、客观的、具有明确结论的书本知识（即“已知的知识”），但不好的地方是，它不能考查学生对未知世界的探索，容易诱导学生死记硬背和考试作弊。而开卷考试则不同，它可以最大限度地避免死记硬背和考试作弊，因为开卷要求学生事先要去看书，如果一个学生平时不看书，到考试的时候，很可能连答案到哪里去找都不知道，翻书也没用；更何况真正的开卷考试，考的题目往往比较灵活、比较开放，不是学生通过翻书就能直接从书本上找到现成答案的。许多人以为开卷考试的难度比较小，其实不然。真正的开卷考试是一种难度更大、要求更高的评价，它比闭卷考试更注重对知识的灵活运用，更能体现学生思维的深度或高水平的思维技能（如归纳、概括、分析、推理、评论、说理等）。试想，世界上哪个学者、哪个科学家的研究成果是在限时、封闭（不看任何参考数据、不向任何人请教、不作任何调研）的情况下搞出来的？我们自己写一篇学术论文，从选题、构思、成文到修改，哪一个不是建立在大量的搜集数据、与他人讨论交流、仔细思考、反复修改的基础之上？惟独在学校的教育考试中，我们却要求学生在限时的、封闭的情境下，独立解答某个指定的问题，岂不是咄咄的怪事！试看以下的例证：

[例 5]；美国高考的开卷作文[12]

美国 3000 多所大学中，大部分不要求提交作文。不过，档次高的学校，绝大部分都要求提供 1-2 篇开卷作文。开卷作文的题目都是从上一届申请者中征集的。作文题目往往既不失新颖活泼，又不失思考的深度与广度。例如，美国西北大学 2003 年的作文题有四个，要求考生从四个题目中任选一个。这四个侯选题目分别是：（a）谁是你们这代的代言人？他或她传达了什么信息？你同意吗？为什么？（b）有种理论认为，伟大的领袖人物都是在他们所处的具体的时代产生的。照你的看法，伟大人物的产生，是由于所处的环境，还是由于个人的特质？试举出一位人物来支持你的观点。（c）在愚蠢的错误和聪明的失误之间总是存在着重大的不同。请说一说你的一个聪明的失误，并且解释它怎么给你或他人带来益处。（d）罗马教皇八世要求艺术家 Giotto 画一个完美的圆，来证明自己的艺术技巧。什么看似简单的行为能表现你的才能和技巧？为什么？又比如，普林斯顿的作文题是：（a）你认为什么思想、发明、发现或创造到目前为止对你的人生产生了最大的影响？请简单说明。（b）假如你

[11] 郭秀艳，《内隐学习在学科教学中的应用与思考》，载《上海教育科研》2004 年第 7 期，第 26-27 页。
[12] 黄全愈着，《“高考”在美国》，第 165-169 页。

得到一年的时间为别人提供自己的服务，你将选择去干什么？为什么？（c）什么是你曾经不得不作出的最困难的决定？你是怎么做的？（d）到目前为止，你取得的什么成功给你带来了最大的满足？

传统教育考试的另一个弊端是，考试往往是在脱离具体生活情境的背景下进行的，我们把这种考试称为“去情境化”的考试或评估。这种考试之所以盛行，原因就在于，学生早已习惯于在脱离日常生活背景的环境下学习书本知识。现代学习科学研究表明，个体的认知是情境化的，有些人在脱离实际生活情景的标准化测验中，被认为缺乏计算或推理能力，但在日常生活情景中，如缝制衣服、在超市购物、往卡车上装牛奶箱、在争端中维护自己的权利时，却能准确地表现出上述测验中所需要的计算能力或推理能力。“有许多任务，人们在一般性的情况下无法胜任，但是在某一特定的情境中却能够出色完成。有很多一个人无法完成的事情，在与同伴的互助协作之下，却能够顺利完成。”人类学家发现，“当人们在自己本土文化环境中工作时，他们经常会有一些意想不到的出色表现，这些在测试中是无法预测到的。那些在试验中或是实验室中好像‘哑巴’的人，在他们熟悉的环境中完成难度相当的任务时，往往表现得非常‘聪明’。即便是在我们的社会中也是如此，卡车司机和超市售货员在他们所熟悉的情境中，能够几乎直觉性的解出相当难的算术问题，而让他们在实验室内解决这些问题时，他们却解不出来。一个貌似有学习障碍的孩子借助同伴的互补性，却可以在一个集体制作蛋糕的复杂活动中担任领导角色。”[13]

既然有些人特别擅长在限定的时间里和没有人际接触的、非情境化条件下接受测试，有些人却只善于在需要较长时期的努力才能完成的专题作业中，或者在情境化的评估中才能更好地表现自己、展示自己，那么，传统的教育考试与评价制度（即限时的、封闭的、去情境化的考试），就必须有所调整与改变。调整的办法主要有两个：一是在某些学科（如政治、历史、社会等）的教育评价中，增加开卷考试的比重；另一个办法就是，将学生置于真实、自然的评估情境中，通过观察学生解决疑难问题、完成真实任务、创造文化产品的实际表现，来对学生的学习和发展状况进行评估，这种评估也叫基于真实情境的展示性评估或表现性评估，以下两例即是十分典型的情境化评估。

[例 6]：夏令营[14]

去年，老师带一组学生去一个森林保护区参加夏令营活动，去之前老师就告诉他们到了那儿，首行要搭一座供他们居住的房子，并告诉他们要带什么、怎么搭、该选择什么样的地方搭。到那儿后，老师领着他们一起搭。而今年，老师带同一组学生去夏令营，只是去的地方不同。这次，老师把搭房子的任务完全交给了学生，什么提示也没给。面对这样的问题情境，学生便自己想办法了。他们首先研究了当地人的住房、当地的气候、地形条件，然后

[13] （美）朱迪思·H·舒尔曼主编，《教师教育中的案例教学法》，华东师范大学出版社，2007 年版，第 23-24 页。

[14] （美）Linda Torp, Sara Sage 着，刘孝群、李小评译，《基于问题的学习》，中国轻工业出版社，2004 年版，第 32 页。

兼顾自己的居住习惯，最后他们为自己搭好了一座令他们满意的房子。后来，他们总结说，这次不仅搭好了房子，而且还学到了许多知识，在其它方面也得到了锻炼，而且还觉得比上次更有趣。

3、学生评语脸谱化、模式化、空洞化

并不是所有的教育评价都要给学生打一个分数，评价除了“评分”外，还有写“评语”。后者往往使用日常化的生活语言，对学生某段时间以来的表现进行客观的描述，并以此为依据对学生表现的优劣、好坏作出评判。在我国中小学，给学生写期末评语（或称操行评语）的事，一般是由班主任来完成的。从理论上讲，这种做法确有它的必要性，因为它可以让学生更好地认识自己，让家长更加全面、系统地了解自己的孩子在校学习的情况，包括取得的成绩、发生的变化、存在的问题等等，以便更好地与学校沟通、配合，通过家校合作让自己的孩子发展得更好。但这只是理论上的一种假设，实际效果如何取决于教师如何去写这个评语。从普遍的意义上讲，常见的学生操作评语毛病主要有以下几点：一是客观、如实的叙事式描述偏少，主观的判断过多，给人的印象是许多“判断”缺乏“事实”的支撑；二是对学生的评语千人一面，抽象、笼统、空泛的“大”词或“套”语随处可见，能写出每个学生的特点与个性的评语，记忆中少之又少！倒是像“热爱祖国、关心集体、尊敬师长、团结同学、对人有礼貌、学习认真刻苦，讲究卫生，爱护公物，遵守校纪校规，希望今后再接再厉、戒骄戒躁、更上一层楼”这一类的用词，至今还记忆犹新！第三个常见的毛病是，评语的用词“死板”、“生硬”、“不生动”，缺乏人情味，过分强调和突出评语的“鉴定”功能，忽视了通过评语对学生进行“激励”的功能。当然，除了上述这些毛病外，还有一些问题可能更为严重一些，比如，有的教师敷衍了事，由学生代写评语；有的教师从网上下载现成的评语，直接抄或贴到学校的成绩报告单上。这就不是一个简单的评语写得好与不好的水平问题，而是一个缺乏专业精神的态度问题。

针对上述这些毛病，近些年来各地纷纷倡导学生评语的改革，主张把过去那种“鉴定式的评语”改成“激励式”、“谈心式”、“描述式”的评语。应该说，从改革的大方向上讲，这样做是对的，但具体的“描述式”、“谈心式”、“激励式”评语究竟该怎么写，并不是一个很简单的问题。首先，我们必须建立几条最基本的有关评语好坏的评价标准，比如：①评价应以事实为依据，在描述的基础上作判断。这就要求教师必须有意识地观察和记录学生的日常行为表现，通过各种管道（如家访、个别谈心、日常观察、听取科任教师或学生的意见与反映等），深入细致地了解学生的思想动态及个性特点，建立好学生个人成长的记录袋，为期末撰写学生评语提供充足的档案数据。②评价应全面、具体，具有针对性：好的学生评语应是全方位的评价，不能仅仅评价学生学习成绩的好坏或学习能力的高低，还要对学生的学习态度、人际交往、心理质量、行为习惯、身体健康等进行全面的评价，具体指出学生突出的优点与长处在什么哪方，主要的缺陷与不足又在什么地方，还要给学生提出明确具体的、切实可行的期望与建议。③作判断时应恰如其分，把握好分寸：好的学生评语应实事求是，优

点要讲足、讲充分，不要吝惜赞美之词，但对学生身上存在的突出问题或明显的缺点，绝不回避，而且，要点中要害。当然，对年龄幼小的孩子，评价一般来讲应以肯定、鼓励为主。比如，讲缺点与不足时，用提建议的方式来表达对孩子的信任与期望；或者用比较委婉、温和、商量的语气来提出希望；或者使用“肯定之中有否定”、“否定之中有肯定”的方式，对学生的行为表现发表评论，引导学生学会以辩正的眼光，去看待自己的优点或缺点。切忌使用轻率武断的、判决式的语言，乱给学生贴标签，这是最要不得的。④评价应因人而异：由于每个学生的个性特点、发展方式各不相同，因此，写评语也应具有一定的弹性，做到因人而异。比如，在给“差生”写评语时，不妨适当地“放大”他的优点与长处，这样做有利于学生树立自信心，建立正向的心理循环。相反，给“优生”写评语时，不妨稍微苛刻一点，有意地挑挑“毛病”，这样做有助于“优生”克服自大、自负的倾向。

为了写出学生、家长爱看，对学生发展有益的评语，这里有两条建议：一是学校应当为每个学生建立一个详细的成长记录袋，用以收集有关学生学习、成长、发展的所有相关数据信息；与此同时，每个教师都应当有一本班级观察日志，用来详细地记录在自己的课堂上每天所观察到的学生日常行为表现；还有，每个班级都应建立一个由值周班干部轮流记录的“班级日志”，用于详细地记录班上每天发生或出现的、有意义的重大事件、活动。有了上述这三个方面的数据数据，教师撰写学生评语就不再是一件多么困难的事了。第二条建议是，学生评语不应只由某个单一的角色主体（如班主任）来包办，最好是让不同的角色主体（如科任教师、学生同伴以及学生自己）都参与进来。比如，学期末，先让学生自己给自己写评语，对自己一个学期以来的学习情况进行总结；然后，在以小组为单位的学生民主评议会上，每个学生向小组成员宣读自己的总结，小组成员当场进行评议。由于同学之间朝夕相处，彼此了解，通过民主评议，不仅使每个学生的自我鉴定得到矫正、趋于完善，也使同学之间在充满团结、和谐、民主的气氛中交流思想，互相学习、受到教育。最后，班主任教师再根据自己平常撰写的班级观察日志，结合学生的自我鉴定、同伴的评议意见，对学生整个学期的表现进行全面、综合的评定。必要的时候，还可以征询各科任教师的意见，以便让评语更加客观、可信，更好地发挥其诊断、鉴别和激励功能。

三、对评价结果的解释与使用不合理导致评价功能的窄化与异化

（一）一个未公开承认但实际采用的假设：好分数等于好学生

正如仅凭学生参加高考的考试成绩，无法断定一个学校办学质量的好坏一样，仅凭学生参加正规考试的分数，也无法断定学生的优劣。但在现实中，人们常常对学生的考分作出错误的解释，把好分数等同于好学生。对于该等式的荒谬性，我们有以下的反驳：

1、学生在学校里的学习有许多类型，并非所有类型的学习都可以量化，用一个简单的分数来衡量，而这些不能量化的学习类型其重要性可能丝毫不亚于可以量化的学习类型。比如，学生情感的发展、品格的养成，就很难量化处理，但又有谁能否认它的重要性呢？正因

如此，绝大多数学校在对学生的在校学习表现进行评价时，除了采用正规的考试以外，还要给学生写详细的评语。仅看考试成绩，不看评语，就无法全面地了解学生。

2、学校里流行的各种教育测验与考试，究竟考了些什么，考的内容是否必要、是否有价值、是否有意义，这对我们判断学生的考分意味着什么，有着至关重要的意义。如果学校的考试内容，仅仅指向陈述性知识或书本显性知识（而非程序性知识或隐性知识），或者低层次、低水平的思维技能（而非高层次、高水平的思维技能），那么，学生考一个好分数又能说明什么呢？顶多能说明，他具有较好的接受学习或记忆能力。

3、在通常情况下，学校教育的考试内容往往是依据国家颁发的课程标准（或教学大纲）、围绕指定的教材内容来出题的。这也就是说，超出大纲、课本和教材的内容，学校通常是不会考的。在这种情况下，如果一个学生利用寒暑假、节假日或课余时间读了大量的课外书籍，参加过许多有意义的课外或校外活动，那么，他所知道的、会做的许多东西，很可能就不会纳入到学校教育的考试范围中来，因而，仅凭学校的考试分数，我们很难看出学生真实的认知结构或能力水平。

4、由于学校通行的教育考试往往采取封闭、限时、去情境化的纸笔考试（或称标准化考试），如前所述这种考试方式存在很大的局限性，它只适合一部分学生（如擅长言语智力或数理逻辑智力的学生）。这就意味着，一个在传统的教育考试中考得不好的学生，很可能在开放的、不限时的、情境化测试中表现优异，考出高分。因此，光看一个孤零零的分数，不看学生是如何参加考试的，得出来的结论很可能就有问题。

5、即便学校考试的内容与方式都没有问题，学生考分的高低与其真实水平或潜能的高低依然不能直接划等号。学生在校的学业成绩除了取决于他们所接受的教学质量（即取决于教师教得好不好）之外，还取决于许多其它的变量：如学校环境状况（如教室照明充分吗？温度适中吗？周围的噪音大吗？学生在校感到安全吗？）、家庭环境状况（如学生家庭的收入状况、父母的职业及受教育程度、亲子互动方式等）、学生个人因素（如学生的母语、原有基础、智力水平、努力程度、平时的健康状况、考试当天的健康及情绪状况）等，所有这些变量都会对学生在校的学业成绩造成直接的影响。所以，真实地说来，只有极少的学生学业成绩不好，是由于智力水平低造成的，绝大部分学生考得不好是因为贪玩、不好好学习，或者学习方法不当，或者教师教学质量差，或者健康状况不好等原因造成的。所以，仅凭学生在某一特殊时刻的一次考试表现，去判断学生实际达到的心智发展水平，要冒极大的风险。试想，学生考得不好，明明是由学校和家庭环境不好，或者教师教得糟糕引起的，我们还要把学业失败的原因归罪于学生本身，并且给学生贴一个“差生”的标签，世界上还有比这更残忍的事情吗？

6、根据考分的高低来评判学生的好坏与优劣，暗含这么一个预设，即学校里的学生肯定有好有坏、有优有劣。这个预设本身合理吗？在真正的教育家眼里，学生没有好、坏之分，所有的学生都是好学生。因为每个学生都有自己的特长，有的这方面好，有的那方面好，有

的今天表现好，有的明天表现好。有的学生刚开始时，基础较差、考得不好，经过一段时间的努力，有了很大的进步，尽管他的考分在班上仍处于中下游，对这样的学生，能说他不是好学生吗？有的学生天分很高，由于不努力、不尽力，导致其学业成就低于其能力水平，像这样一类未能充分发挥学习潜力的学生，更不能说他是“差”生。因为像这样的学生，只要学习动机得到激发，往往就会一飞升天，一鸣惊人，爆发出惊人的能量！正所谓暂时的“差”，不等于永远的“差”，暂时的“落后”，不等于永远的“落后”。不同的个体有着不同的发展步调、节奏、轨迹：有的起步早，有的起步晚；有的先快后慢，有的先慢后快；有的对这种事物敏感，有的对那种事物敏感；有的偏爱这种学习方式，有的偏爱那种学习方式，面对如此千差万别的学生，如果我们采用同一种教育方式去对待他们，只会人为地制造出许许多多假性的“差”生或学业失败者。

（二）竞争取向的评价视角：为分等、选拔、鉴别而评价

1、用一把尺子去衡量所有的学生

在动物世界，有的动物擅长游泳、有的擅长挖洞、有的擅长飞翔，如果采用同一标准（比如“游泳”）去考所有的动物，或者要求所有动物都要学习同一种技能，达到同一个标准，当然是荒谬的，但在人类的世界中，这样的事经常发生，我们并不感到奇怪。在教育上，我们规定所有的学生都要学习同样的教材内容，达到同样的发展水平，然后用统一的试卷、统一的评价标准、统一的分数线去评价所有的学生。之所以这样做，原因就在于我们假定，社会需要单一规格的人材；所有学生具有相同的潜能或能力倾向；所有学生都愿意学习相同的内容，都能够达到相同的发展水平，这大概就是人们常说的“标准化教育”吧，它把教育看成是一种类似奥林匹克式的竞赛活动，而不是把教育看作是发展个体独特才能的事业。这样做的后果是什么呢？由于优秀的标准只有一个，大家都往一条道上挤，最终比来比去，把学生人为地分为三六九等，其结果就是，我们把本来可以取得“成功”的一大批学习者，变成了“失败者”，让他们带着“失败者”、“被淘汰者”的心态进入社会、进入生活。当年报考清华大学的吴晗、钱钟书，如果按现在的高考录取体制，恐怕连一般的本科院校都上不了。所以，虽然现在的教育口口声声讲以人为本，以学生发展为本，但骨子里并不是这样，因为社会的各行各业是分层的、人们的身份地位是分等的，因此，学校教育也要迎合这种分层、分等的需要，为它们提供合法的依据，这其实是现代教育的一种异化。须知，教育的真正目的在于，开发人的潜能，促进人自由而全面的发展。“自由发展”的结果必须是，“万类霜天竞自由”，人人都做自己想做的事情，人人都找到自己的坐标与归宿，人人都以欣赏的眼光去看待他人，相互尊重、相互支持、和睦共处，最终达到所谓“各美其美，美人之美，美美与共，天下大同”的境界。要达到这种境界，就必须以多元化的优秀标准，去取代单一的优秀标准；用差异化的评价去取代整齐划一的评价。杭州天才小学尝试取消传统的评选“三好生”的做法，取而代之的是评选“进步生”、“特长生”、“全能生”；并提出“天长没有差生，只有有差异的学生”的口号，这种改革的尝试朝着正确的方向迈出了可喜的步伐，值得推广

和效仿。[15]

2、只看最后的结果，不问过程。

竞争取向的教育评价，往往只看最后的结果，不问这个结果是怎么来的。以成败论英雄，分数考得好的人即是“英雄”，考得不好的人无论怎样辩解都没有用，都是“狗熊”。但事实上，这种只看结果、不问过程的评价取向，很容易诱导学生急功近利、投机取巧，就像社会生活中有些人为了达到自己的“目的”不择“手段”一样。须知，“结果”与“过程”有时候不对等的，“过程”可能是丰富的、曲折的、精彩的、富有意义的，但最终的“结果”可能是稚嫩的、粗糙的、不起眼的。大家可能都知道，在奥运史上有这么一个令无数人感动的故事：在 1968 年墨西哥奥运会上，坦桑尼亚的马拉松选手阿赫瓦里（John Stephen Akhwari），在途中因意外摔倒导致膝盖受伤、肩部脱臼的情况下，仍然坚持跑完全程，尽管在全部的参赛者中，他的成绩垫底，但他说的那回话~“我的祖国派我到这里来，不是为了让我开始比赛，而是要我完成比赛”~却让他成为奥运史最具影响力的参赛者，当年的奥运冠军许多人可能早已记不得了，但“阿赫瓦里”的名字却传遍了世界，为世人所铭记。不仅体育比赛如此，学生在校的学习也是如此。比如，新课改所倡导的“探究式学习”，与其去关注和强调探究结果的“正确性”，还不如关注和强调探究过程的“真实性”。只要学生参与了探究的过程，感受到了探究的乐趣，领悟到了探究的方法，就已经达到了探究学习的目的。所以，有人说，探究学习重在“体验”与“感受”，确有它的道理。试看以下的例证：

[例 7]：抄近路不如走弯路[16]

一次，一群其它学校的老师到 56 号教室参观。学生们正在安装火箭模型。有一组学生虽然做得很认真，但是他们弄错了飞弹部分的装置，于是来访的老师频频向那一组走去，想为孩子们示范正确的安装方法。雷夫果断而有礼貌地阻止了他们，他们之间有这样一段对话：

访客：（很小声地）雷夫，你都不知道啊，他们做错了。

雷夫：我知道啊。

访客：机翼都歪了。

雷夫：是啊，是歪了。

访客：发射架粘得太靠近火箭头了。

雷夫：确实如此。

访客：可你就眼睁睁地坐在这里？

雷夫：是啊。

访客：可他们的火箭会飞不起来呀。

雷夫：一会儿肯定会飞不起来……

访客：可是……

---

[15] 吕型伟总主编，天长小学实验组编，《直面差异：来自杭州天长小学的教育叙事》，人民教育出版社，2004 年版，第 13-14 页。

[16] 杨思卓：《第 56 号教室的奇迹》，载《读者》2010 年第 13 期，第 17 页。

雷夫对此早已深思熟虑：学生们接下来就得找出火箭飞不起来的原因，他们得回到教室自己好好想想。我们的科学家们一天至晚在做的事情不就是这个吗？

失败是由身为教师的我们自行认定的，在第 56 号教室，飞不起来的火箭不是失败，只有当学生停止解决问题的尝试才算失败。所以，最佳的教学不是老师和家长说个不停，而是像雷夫一样：我不想说，我很清醒。让孩子们展开探索的翅膀吧，让他们经历困惑、苦恼，去收获惊喜与领悟。

[例 8]：“陀螺”制作的展示与交流[17]

“有用木头做的陀螺吗？”

“有！”

一个清脆响亮的声音从后面传来，大家都循声望去，只见尤瑞之高高举起一个木陀螺。

教室里一阵哄堂大笑。

这哪是陀螺？黑不溜秋，下端凹凸不平、粗糙不堪，看上去就是一块丑陋的木头疙瘩。

“好难看啊！”有的同学尖声大叫。

“肯定转不起来。”又有同学说。

“是，是转不起来…”尤瑞之垂下手，尴尬地笑着。

在又一阵笑声中，我走过去，高高举起那个丑疙瘩，郑重地端详。学生们看我这样认真，都安静下来，注视我的反应。

“这块木头现在还是蛮难找的。”

“我是问楼下的工匠要来的。”尤瑞之抬起了头。

“你很善于和别人打交道。”我赞许道。

“他很热心，砂轮也是他给我的。”他声音响了一些。

“大家看这些刀印，要花多少工夫啊！”我指着陀螺上的坑坑洼洼说，“看得出，是你用自己稚嫩的小手一刀一刀削的。”

“我做了三次才做到这个样子。”

“我真佩服你顽强的毅力。”我说，“失败打不倒你。”

尤瑞之黑黑的眼睛放出光彩。“后来我就又做了个小的，这个能转起来！”他拿出了一个葡萄大小的小木珠，“这是我家的檀香珠。刚开始它是横转的，我寻思是重量不够，就在下面钉了粒小铁珠，转的时间就长了。”

“失败后你不泄气，而是冷静找原因动脑筋。”我又夸奖。

周围的许多双眼睛里慢慢也装满了敬佩。

尤瑞之更来劲了，指指点点，“这个小铁珠是我从楼下工匠那里要来的。是一辆旧自行车上的。可惜，下面的小铁珠我没有钉好，如果钉得正中心一点，还要好！”

“很好，你能懂得在必要的时候向适当的人求助，非常善于利用身边的人力资源和物力

[17] 见程艳，《课程变革在挑战中前行：小学综合实践活动课程实施的质的研究》，浙江大学硕士学位论文，2008 年 5 月。

资源。”我又点头赞许。

（三）评价背离其本意，与教学相分离

评价的结果要么体现为“评分”，要么体现为“评语”。本来，不论是评分还是评语，都只是教育、教学的一种手段，其本意都是为了看看学生通过学校的教育教学活动，究竟学到了什么、发展得怎样，然后对学生的学习与发展状况进行诊断、分析，找出其中存在的主要问题及其原因，然后提出有针对性的、切实可行的改进意见与建议，这才是教育评价的本意或本体功能之所在。这也就是说，真正的教育评价应当为促进学生的“学习”与“发展”而服务，而要做到这一点，就必须在提供“回馈”信息，对问题进行“诊断”，继而思考如何“改进”教学上做文章。但在现实中，“评价”的概念首先被窄化为“评分”，而在“评分”结果的使用上，又大多囿于“竞争”和“比较”。一个孩子刚入学时，本来是不知道分数为何物的，更不知道为追求分数而学习，正是在家长和教师的反复诱导下，在残酷竞争的压力下，他们开始懂得了分数的重要，并逐渐养成了为分数而学习的习惯。“分数”原本只是一个提供教学回馈的手段，但现在分数却成了“目的”本身、成了“主人”；原本为探索未知世界、满足求知欲而学习，现在却成了为分数而学习，这不是评价功能的异化，又是什么呢？仅仅给学生打一个分数可能是简单的、容易的，但要从学生的得分中“读”出其中暗含的意味（比如，学生的学习存在何种障碍与困难、这些障碍或困难是由什么引起的、有什么办法可以消除这些障碍与困难），却颇为不易！须知，“分数”本身是不会说话的，好的教育评价应当像医生给病人看病、开处方一样，如果一个医生只给一个生病孩子的健康状况打个分，之后什么也不做，既不开处方，也不告诉家长孩子得的是什么病，病因是什么，应该怎样照料他，就一走了之，这样的医生当然是不称职的。

要回归教育评价的本体功能，就必须把评价看作是教学本身的一部分。过去的教育评价之所以功能窄化、异化，根本的原因就是，评价与教学相分离，评价游离于教学过程之外；换言之，我们不是为了促进教学而评价，而是为了给学生分等、分级而评价。改进这个问题的办法有两个：

第一，当学生由于这样或那样的原因，没有能够掌握知识时，教师没有必要急于给学生评分。因为即使评分，打上一个不及格分数，对学生也没有什么好处，只会让学生感到苦闷和压抑。正确的办法是，当教师发现学生在作业或考试中错误较多时，可暂不评分，将作业或试题返还学生，让学生重做一遍，自己去发现并纠正错误，直到正确无误或学生感到满意时才提交给教师评分。正如苏霍姆林斯基所做的那样，“我从来不给小学的学生打不及格的分数。如果儿童有什么地方做得不好，我就对他说：‘你试一试重做一遍，只要下点功夫，你就一定能做好。现在还没有给你打分数，你再努点力，就一定能得到好分数。要是你有哪一道题不懂，明天上课前到学校里来，咱们一起想一想’”。[18]

第二，注重对学生学习的“初始性评价”与“形成性评价”。“初始性评价”一般发生在教学之初（开学头一周或头十天左右），其目的是确定学生学习的起点或基准线，帮助教师

[18] （苏）苏霍姆林斯基着，杜殿坤编译，《给教师的建议》，北京：教育科学出版社，1984 年版，第 317 页。

回答“我的教学应从哪儿开始？”。教师在开始新学期或新单元的教学时，通常会使用初始性评价，即采用“预备考试”的形式，获取学生对将要学习和讲授的内容已知的情况，这种信息为教师有效地设计教学（备课）提供了重要的参照信息。教师获取初始性评价信息的来源主要有：学生档案记录材料、前任教师的评价、入学预备考试、课堂上的观察与提问、师生个别谈话、家长的评价等等。“形成性评价”通常发生在师生互动的教学期间或课堂教学的某些阶段，是一种伴随教学过程而进行的经常性评估，其根本目的是为了给教师的教学提供及时的回馈信息，帮助教师及时地发现学生学习过程中存在的困难与问题，从而帮助教师及时地调整教学、改进教学。形成性评价通常有两种类型：一类是“非正式的课堂评价”，这种评价主要依赖学生的身体语言、面部表情、对课程的参与、对问题的反应以及他们提出的问题等线索，来对学生的学习状况进行实时的评估。“评估在教学期间进行；在课堂内展开；需要做出实时的决定；关注学生对教学内容和活动的反应；主要依赖非正式的学生信息和反应”，是这类评价的主要特征。19这种评价的独特性在于，教师既要一边从事教学活动，又要一边观察学生、解读学生的反应；在教学期间，教师几乎没有时间反思其所观察到的内容，或者没有时间收集额外的信息，教师必须根据有限的、残缺不全的、不确定的观察信息，做出实时的决策或反应。正因为如此，光靠这种非正式的课堂评价是不够的。教师还必须通过“正式的课堂评价”来收集有关学生的评价信息。所谓“正式的课堂评价”，主要是通过简短的课堂练习、课堂小测验、课后的家庭作业等，来对学生的学习状况进行评价。

[19] （美）Peter W.Airasian 着，徐士强等译，《课堂评估：理论与实践》（第四版），上海：华东师范大学出版社，2008 年版，第 125 页。