

价值反思与学生评价

Value Reflection And Student Assessment

苏启敏 · 著



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PRESS GROUP
北京师范大学出版社

价值反思与学生评价

苏启敏 著



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

目 录

CONTENTS

导 论 价值反思：学生评价与评价标准 1

- 一、学生评价的“测不准现象” 2
 - 二、评价标准的“软”“硬”之争 8
 - (一) 学业抑或素质：一个无法
走出的怪圈 8
 - (二) “硬心肠”的评价标准 11
 - (三) “软心肠”的评价标准 13
 - (四) 关于两种评价标准的评价 16
-

第一章 学习成就：学生评价标准的

价值意蕴 19

- 一、学生评价标准的内涵 19
 - (一) 标准 20
 - (二) 评价标准 22
 - (三) 价值取向 28
 - (四) 价值观念 28
- 二、学生评价标准的意义 30
- 三、学生评价标准的类型 33
 - (一) 内容标准 33
 - (二) 表现标准 35
 - (三) 成就标准 38

第二章 目的之镜：内容标准的	
价值追求	41
一、国家主义	41
(一) 国家与国家主义	42
(二) 国家主义与教育目的	44
(三) 国家教育目的与内容标准	48
(四) 内容标准的意识形态	51
二、精英主义	55
(一) 谁是精英	56
(二) 作为再生产媒介的内容标准	61
(三) “底线标准”的“隐身”	65
三、学业主义	68
(一) 学业标准的“给定性”	68
(二) 学业至上	74

第三章 身心之维：表现标准的	
价值旨趣	79
一、绩效主义	80
(一) 绩效主义的表征	82
(二) 表现标准的“失真”	85
(三) 表现标准的“异化”	88
二、重智主义	90
(一) 智力优先原则	91
(二) 多元智能的神话	97
三、生源决定论	101
(一) 个人天赋决定论	104
(二) 家庭环境决定论	110
(三) 学校声誉决定论	115

第四章 人才之变：成就标准的	
价值转型	121
一、第一次浪潮：选举社会与选拔型	
标准	122
（一）竞赛隐喻及其人才假设	124
（二）杰出人才论与选拔性考试	128
二、第二次浪潮：工业社会与测量型	
标准	129
（一）生产隐喻及其人才假设	130
（二）科学管理运动与标准化测验	134
三、第三次浪潮：人本社会与真实型	
标准	137
（一）旅行隐喻及其人才假设	139
（二）回归真实与另类评价	142
第五章 观念冲突：学生评价标准的	
价值分歧	146
一、功利价值观与博雅价值观	147
（一）功利价值观及其对评价	
标准的影响	150
（二）博雅价值观及其对评价	
标准的影响	152
（三）从“文理分科”的争议看	
两种价值观的分歧	154
二、管理价值观与发展价值观	156
（一）管理者视角下的评价标准观	157
（二）利益相关者视角下的评价	
标准观	161

三、科学价值观与人文价值观	167
(一) 定量评价标准的价值中立 原则	168
(二) 定性评价标准的价值关联 原则	171
四、一元价值观与多元价值观	175
(一) 价值一元论与普适评价 标准	175
(二) 价值多元论与多元评价 标准	179

结 语 多元向度：学生评价标准的 价值选择	188
一、价值选择的三种路径	188
(一) 强制性的价值选择	189
(二) 竞争性的价值选择	190
(三) 协商性的价值选择	192
二、基于标准的学生评价改革之路	193
(一) 从“方法中心”到“问题 中心”	195
(二) 从“阶段评价”到“终身 评价”	196
(三) 从“同而不和”到“和而 不同”	198

参考文献	200
后 记	209

导论 价值反思：学生评价与评价标准

教育改革是当前社会各界颇为关注的热点问题。在诸多教育改革领域当中，教育评价改革所遭遇的阻力不小。一直未有定论的应试教育与素质教育之争，其实质就是社会各阶层对教育评价现状所引起的抱怨和不满的体现。由于教育评价涉及教育活动最为根本的人的培养问题，它往往以“指挥棒”的形式引导着教育活动的方向。在教育评价方面的任何一个改革措施，都会在整个教育改革进程中牵一发而动全身，引起教育系统的“蝴蝶效应”。据实而论，教育评价改革无疑是教育改革的重中之重。

如果说学校评价、校长评价、教师评价、课程与教学评价是教育评价的“保护伞”，那么，学生评价（student assessment）无疑就是教育评价的“硬核”。因为教育评价归根结底是为了保证学校教育目的的顺利达成，而学生在学习或学习活动中所取得的发展与变化，又是衡量教育目的是否实现的核心标志。只有清楚了解学生在经过系统学习后的发展与变化，才有可能对教育目的的实现程度作出客观的判断。而要了解学生的发展和变化，就需要进行学生评价。所以，各个层面的教育评价改革，其成败均建立在学生评价改革能否顺利实施的基础之上，都围绕使学生发生社会所期望的变化这一根本目的而展开。只有学生评价改革取得长足的进展，才有可能以点带面地推进其他领域的教育评价改革。

有见于此，美国在 20 世纪 90 年代的基础教育改革就提

出了“基于标准的教育改革”(standard-based educational reform)的主张,力图通过建立高标准的课程与学生评价体系来提升教育质量。1990年,美国前总统老布什(Bush, G.)签发了一份名为《美国2000年教育战略》的文件,确立了美国教育要达到的六个目标。^①1994年,前总统克林顿(Clinton, B.)宣布了《2000年目标:美国教育法案》,提出了教育改革的基本原则。这两个文件都把“建立教育高标准”作为核心内容,要求学校把学生培养成为能够在全世界经济竞争中获胜的现代化人才。为了实现国家的教育目标,从联邦教育管理部门的立法和规定,到州和学区签发的各种文件,都把“高标准的质量”作为改善教育的宗旨。其中,高标准的学生评价是不可或缺的一个环节。这从一个侧面也反映出学生评价改革的重要性所在。

一、学生评价的“测不准现象”

在学生评价改革过程中,如何设计合适的评价标准是一个无法回避的问题。因为学生评价标准关系到“用什么来评价学生”“评价学生什么”和“如何评价学生”等诸多方面,没有合适的评价标准对此作出回应,学生评价就无从谈起。可是,学生评价具有极强的间接评价性质。在物理测量中,可以用尺、秤、温度计对物体直接进行测量,而学生的发展受遗传、环境和主观能动性等诸多因素的影响,处于不断变化之中。与此同时,各种各样的价值观念也影响着人们在学生评价中所做的判断。所以学生评价难以像物理测量那样确立具有客观性和普适性的评价标准。相同的学生评价标准可能难以对所有学生都作出准确的判断,而不同的评价者对同一位学生的评价可能又会

① 这六大目标是:(1)所有学龄儿童做好入学准备;(2)中学生的毕业率至少应提高到90%;(3)美国的学生在学满4、8和12年时,应当在有相当难度的课程——其中包括英语、数学、科学、历史及地理等科目中,学习成绩优秀,考试合格;(4)美国学生在数学与科学成就方面将是全球第一,名列前茅;(5)每个成年人都具有文化知识和在国际经济活动中的竞争能力;(6)每所学校将成为无毒品、无暴力的场所,还将成为秩序井然而又富有浓厚学习空气的园地(Educational Resource Information Center <ERIC>. America 2000: An Education Strategy. Sourcebook [R]. Washington, DC.: Department of Education, 1991: 8)。

采用不同的标准。这种特点造成了学生评价“测不准现象”的产生。^①

学生评价的“测不准现象”在时间和空间两个维度——有着不同的表现。在时间维度上，人们通常认为在学校学业成绩优秀的学生步入社会以后工作和学习必然也会表现优异。然而有研究表明，学生通过学校教育所取得的学习成绩与其未来在社会上的表现并没有表现出必然的因果关系。曾引起强烈社会反响的“第十名现象”是一个比较典型的例子。浙江省杭州市天长小学的周武老师担任小学班主任和语文任课教师近 20 年，发现了耐人寻味的“第十名现象”：当年就读小学时成绩在 10~20 名的部分学生，在后来的学习和工作中成绩出色；相反，有些当年成绩在前三名的优秀学生反而后来业绩平平。这一现象引起了周老师的关注。周老师有意识地对 1990 年前后毕业的 150 名学生做了追踪调查。调查数据显示，54% 的小学成绩在前三名的学生在以后的学习工作中仍能保持领先地位，而 80% 的“第十名”学生在以后的学习和工作中业绩有所上升。^② “第十名现象”引起了教育界的广泛关注和反思。

关于“第十名现象”的研究说明了这样一个事实：早期优良的学业成绩并不能保证学生在成年以后具备杰出的才能或作出卓越的贡献。它为学生评价带来的启示就在于：应该以什么样的标准评价学生，才能尽可能地减少这样的误差，使学生得到更为全面的发展，在社会上真正发挥自己的天赋。如果学校教育评价学生的标准仅仅是知识的掌握，以考察学生的记忆力、接受力作为在评价的主要内容，那么自然就很难预测在评价中获得“第一名”的

① 物理学（量子力学）上的“测不准原理”又称为“不确定性原理”（uncertainty principle），该原理表明：一个微观粒子的某些物理量（如，位置和动量或方位角与动量矩，还有时间和能量等）不可能同时具有确定的数值，其中一个量越确定，另一个量的不确定程度就越大。学生评价的“测不准现象”则表明：学生的学习成就与社会期望，以及不同文化背景下对学生学习成就的评价，不可能具有精确的一致性，甚至有较大误差。但是，在特定条件下，通过采取手段校正误差，它们是有可能逐渐趋向或接近一致的。如果说物理学上的“测不准”是一种原理，其存在本身就是客观的。那么，学生评价的“测不准”则只是一种现象，其存在既有客观原因（如，社会为个人提供的资源、机遇的充裕程度），也有主观原因（如，评价标准的适切性），本文主要探讨后者。

② 王曦.“第十名现象”的调查与研究 [J]. 教育研究与实验, 2001 (1).

学生步入社会以后，在学习和工作上遇到需要综合运用智力与非智力因素解决问题的时候，是否还能保持领先优势。而“第十名”的学生虽然在学生评价中的名次并不靠前，但是也一样存在着成材的可能性。如果仅仅根据以知识记忆为主要内容的评价标准，这些学生的优点就可能被忽视乃至埋没。因此，学生评价需要正视因为评价标准不当所造成的“测不准现象”及其消极作用。

学生评价的“测不准现象”也同样体现在空间维度上，不同社会文化背景下的学生评价也会造成学生在学习成就上的差异。2001年4月2日的《中国青年报》刊登了一篇署名端木的文章，该文详细介绍了中美两国教师在学生评价上的种种差异。这些差异，在某种意义上来说更多的是学生评价标准方面的差异，而这些差异又对学生评价实践具有价值导向意义，尤其对学生的身心发展影响甚巨：

仅仅在4个月前，端木送走的那个女儿，是一个被老师批评为“没有数学脑子”、垂头丧气地对父亲说“我厌学了”的孩子。4个月，并不足以让一个学生本身的素质发生天翻地覆的变化，不同的可能是她所处的教育环境，以及她得到的评价和激励。女儿的美国老师说：“她在任何校园都会受到珍视。”当一个学生受到“珍视”的时候，与不被珍视，甚至被忽略、被歧视的时候所表现出来的样子，是会有所不同的吧。

我读过不少教育学和心理学方面的书，深知“评价”和评价的方式在一个孩子成长中的作用。因此，哪怕孩子在某次考试中成绩不好，我也总是说：“这没关系，没准儿比全对还好，因为错过的东西印象更深！”但是我也感到，家长的鼓励对孩子的影响越来越小，因为你不“专业”、不“权威”，孩子做不出来的题，你通常也做不出来；你指导的作文，通常还会被判低分。很明显，在学校里，老师对孩子的评价具有决定性的影响。每个学期结束后，成绩单上照例有老师的评语，本来这是一个使家长了解孩子的大好机会，可惜每次都是那寥寥无几的二三十个字，几乎是年年相同的套话——两三个词儿的优点和两三个词儿的缺点——有一次的评语绝对是学生的字体，看来老师自己根本就没有写评语的兴趣，干脆让学生干部代劳了。

尽管我对女儿很有信心，觉得她是一个心智很正常的孩子，品行也没有什么必须纠正的缺陷，但她的理科成绩确实有江河日下的趋势，老师对女儿的评价开始影响我，“也许女儿真是缺乏理科方面的才能？”“也许她真是缺乏逻辑思维能力？”“也许她学习真的不够努力？”我自己都能感到在给女儿打气时有些言不由衷了。

终于有一天，女儿迟疑地对我说：“爸，我厌学了……”“是吗？”我沉默无语，内心涌起一片可怕的绝望。我明白错不在女儿，是什么让她苦苦挣扎到高三，却突然丧失了自信和学习兴趣？！她曾是一个多么快乐和不甘人后的女孩儿呀……所有这些，大概就是当有可能脱离这种教育的机会来临时，我和女儿都没有犹豫的原因。

几乎丧失了学习自信的女儿，在美国会怎么样呢？她在陌生的英语教育环境中将遇到巨大困难，这是我可以想象的，也许她一开始就会沦为班上成绩最差的一名，美国老师会对这样的学生作何评价呢？我简直不敢想下去，并且做好了让女儿再上两年高中的思想准备。

不久，似乎发生了奇迹。女儿那边好消息不断传来，“突破”先从法语开始，女儿首次得了满分！女儿说，一次法语课，有别的老师进来找我们的老师，我们老师正忙着，她就跟我聊了两句，我们老师进来以后说“她刚从中国来”。那个老师说“我知道”。我们老师又说“刚来，法语就已经得100了”。那个老师感叹：“Jesus（耶稣呀）！”这声惊叹让女儿心花怒放。

这些赞扬有点像兴奋剂，不知为什么，女儿开始自信得让我感到陌生，一些我原来并未发现的特质似乎开始呈现。我不大有把握有资格“指导”她了。果然，3个月过去了，女儿不经我同意，干脆地告诉我“准备今年就申请大学”，她甚至已经试着考了一次“托福”。我愣了一会儿，觉得女儿有点好高骛远，托福能考个四五百分就了不起了，她什么时候见过这个世面呀！不久成绩出来了，着实让我大吃一惊，她竟考了600多分。这个成绩申请美国大学绰绰有余。

女儿真的开始申请大学了。她告诉我其中一个必要程序是中学老师的推荐信。与中国的“一考定终身”完全不同，美国对申请大学的学生采用综合评价，不仅要看你大学入学资格考试（SAT）的成绩如何，你平时的成绩也

占一定的比例，还要看你有什么特长，甚至做过多少时间的社会公益工作，中学老师的评价也是其中重要的一环。

指导老师的评语：斯蒂芬表现得很完美。在我做顾问的经历中，还没有听说过有外国学生比她更快地完成了学术转型。谦虚的斯蒂芬甚至不愿意接受她应得的高分数。数学和其他理科方面的科目对她来说很轻松，遥遥领先于她的同班同学。她喜欢语言，学起法语来是个明星。然而在英语和美国历史方面，她的阅读和写作水平还需要努力。她的所有老师都有共同的想法，“她太不可思议了，请再给我们20个像斯蒂芬这样的学生！”他们一致赞扬她的勤奋、学术好奇心、专心学习和愿意帮助小组中其他同学的行为。平时斯蒂芬在课堂上很安静，但一被叫到回答问题时总是清楚无误，显示出极强的理解力。

斯蒂芬在学业上越来越自信，同时她也开始交朋友，在社交方面开始轻松自如。她爱好运动，希望能参加我们学校的春季田径比赛。她开始意识到自己和沙龙高中的同龄人有许多共同之处。最近，她的父亲问她美国教育和中国教育有什么区别时，她说：“老师不一样。美国老师非常亲切、友好。考完试他们进行评论；而在中国，我们只是拿到分数。在中国我们需要死记硬背，而在美国你不得不学习思考，学会表达思想。”

斯蒂芬是个不同寻常的女孩。她独立、灵活，非常善于适应生活中的变化。她以乐观的态度看待将来在大学的学习。尽管她想念父亲、其他亲人和在中国的朋友，但她肯定对大学的挑战有准备。在沙龙高中的极短时间内，她就证明了自己是优秀的学生，是积极进取、善于接受挑战的女孩。

我满怀热情地赞同最具竞争力的大学接纳她。（乔·贝克汉姆）

这像是一份学生评价吗？说实话，贝克汉姆先生的推荐信，我读起来就像是一个老朋友在与我娓娓而谈，亲切、从容，充满热忱。显而易见，他是如此负责，除了自己的观察，还逐一征询了女儿所有科目老师的意见，遣词用句非常谨慎——例如，我很担心女儿因为爱面子，不主动在上课时积极回答问题的习惯，而这是美国教育中非常被看重的品质，几次在邮件中叮嘱她要“改正这个缺点”。贝克汉姆先生也发现了这一点，但他使用的词是中性的——“安静”。他在小心翼翼地呵护着什么呢？呵护着学生的自尊！读完这

些老师极具个性的评价，那个“没有数学脑子”的、只能上文科班的、垂头丧气感到“厌学”的女儿，消失了；取而代之的是一个看起来正全面获得进步，甚至有点出类拔萃的女儿！仅仅三四个月过去，女儿的学习状态和自信简直就像换了一个人，是什么起了如此奇妙的催化作用？

我想说，是不同的评价标准使然！一个孩子如果总是受到负面评价，就会产生自我的“负驱动”，自暴自弃。处在学习过程中的学生就像一杯没倒满的水。在中国老师这里，通常看到“一半是空的”；而在美国老师那里，却总是看到“一半是满的”——前者否定，后者肯定，哪一种会对学生产生激励作用呢？当然是后者，这早已被心理学上著名的“罗森塔尔效应”所证明——仅仅是因为教师对学生的期待不同，一部分学生就会比另外的学生取得更大的进步。这难道不是我们的教育应当反省、深思的吗？^①

显然，在不同的社会文化背景下，学生评价所呈现的特征也存在不少差异。这种差异的实质是评价标准的差异。它意味着人们对学生评价目的的不同理解。是为了学生的升学而评价呢，还是为了学生的终身发展而评价呢；是消极地评价学生呢，还是积极地评价学生呢。不同的选择对学生的整个学习历程甚至整个人生都会产生不同的影响。由此可见，选择不同的学生评价标准也是造成“测不准现象”的重要原因，评价者需要正视这种差异对学生造成的影响。然而，再往深处想，难道中国的教师就不知道要在学生评价中采取激励的措施吗？究竟是什么原因使教师在评价中有意或无意地采取伤害学生自信心的评价标准。换句话说，当教育工作者运用学生评价标准实施评价时，也许应该有意识地问问自己：究竟什么是学生评价标准？是谁的标准？它们如何得来？这些标准背后的目的和意图是什么？

上述对“测不准现象”的分析，其目的并非是要说中国的学生评价一无是处，也并非为了证明美国的学生评价比中国更优越，而是想借此引起人们对评价实践的思考：评价实践总是围绕着特定的评价标准展开，不

① 端木：中美教师对同一个孩子的不同评价——“我以性命担保她行！”[N]．中国青年报，2001-4-2（2）。（因篇幅所限引文略有删节）

同的学生评价标准是导致“测不准现象”的主要原因，而“测不准现象”造成的影响有可能是正面的，也有可能是负面的。虽然学生评价的“测不准现象”是一种常态现象，但是人们仍然试图去消除这种因“测不准”而造成的负面影响。这是因为，一方面，人们总是希望学校教育所培养的人能够符合社会的期望，因此特定的学生评价应该与特定的社会期望相联系，并尽可能接近一致，如果学生评价与社会期望误差较大甚至完全脱节，学校教育在人的培养上的价值就无从体现；另一方面，学生评价关系到学生的切身利益，如果不对学生评价的误差进行修正，学生就有可能被不合时宜的学生评价引向歧途，延误学生通过学习获得发展与成长的时机。为减少“测不准现象”可能造成的负面影响，人们不得不通过变革学生评价标准来校正学生评价的误差，将负面影响控制在可以接受的程度。这是学生评价改革的根本目的，同时也是不同时代、不同社会背景下学生评价改革所要面临的首要问题。

二、评价标准的“软”“硬”之争

如果要变革学生评价标准，就不得不涉及对“什么样的标准才是合适的标准”的理解。在教育评价学者看来，“所有标准都是判断问题。然而，判断的基础从‘硬’到‘软’，不尽相同。谁的判断？这个问题是首要的和经常发生的问题”^①。从现有的研究文献来看，学者们对于何谓“合适的标准”的类型大致有两种看法，一种是以学业评价为主要形式的“硬心肠”的评价标准，另一种是以素质评价为主要形式的“软心肠”的评价标准。

（一）学业抑或素质：一个无法走出的怪圈

克龙巴赫（Cronbach, L. J.）把学生评价分为“最佳表现”（maximum performance）的评价与“通常表现”（typical performance）的评价两种类型。^②所谓最佳表现评价就是学生在动机得到充分调动的情况下，所能达到的最好表现。这类评价主要用于判定学生潜在能力或学习成果。能力倾向测验以及

① 瞿葆奎. 教育学文集·教育评价 [C]. 北京：人民教育出版社，1989：762.

② Cronbach, L. J. Essentials of Psychological Testing [M]. New York: Harper & Row, 1984: 35.

学业成就测验就属这一类型。能力倾向测验用于预测学生未来的学习能力。学生成就测验用于确定学生过去学习活动的成果。所谓通常表现评价，主要用于判断学生未来的学习能力、态度等个人素质。它的重点在于获得在日常情况下学生典型的反应。观察与日常行为记述是最常用的技术。在中国的学生评价活动中，习惯于把关于“最佳表现”的评价称为学业评价，把关于“通常表现”的评价称为素质评价。

学业评价，在英文中一般称为 achievement assessment 或者 performance assessment，它是对学生的学业成就与表现的评价。学业评价并不仅仅指考试成绩，还包括教师通过其他途径收集有关学生学业及其表现的各种信息和资料，以及在此基础上对学生的学业情况作出测评。这种测评的形式有很多，笔试、口试、教师课堂对学生掌握知识情况的观察等。学业评价并不等于学生评价，学业评价是学生评价的下位概念。学业评价也不等于考试，学业评价是考试的上位概念。三者的隶属关系为学生评价——学业评价——考试。由于学业评价是学生评价中一个最常用的评价类型，长期以来学业评价被等同为学生评价。所以，人们对传统学生评价的批评，在很大程度上是对学业评价的批评，尤其是选拔性考试的批评。

素质评价，顾名思义就是对学生素质进行的评价。何为素质，根据部分学者的界定，就是“个体的先天禀赋以及在此基础上，通过环境和教育影响所形成和发展起来的相对稳定的身心组织的要素、结构及其质量水平”^①。2002年12月下发的《教育部关于积极推进中小学评价与考试制度改革的通知》提出了学生评价的基础性发展目标，包括道德素质、公民素养、学习能力、交流与合作、运动与健康、审美与表现。^②这六个方面是在新课程改革背景下对学生素质的一种规范性定义。素质评价也主要围绕这六个目标展开。

从理论与实践情况来看，一直以来，人们对于学生评价的理解基本处于一种二分法层面之上，即把学生评价划分为“学业”与“素质”两种类型，

① 于建福. 素质教育 [M]. 北京：教育科学出版社，1999：2.

② 教育部关于积极推进中小学评价与考试制度改革的通知 [J]. 人民教育，2003 (Z1).

然后制定相应的评价标准实施评价。这种标准制定的思路源于评价学界对评价性质的理解。传统的观点认为,评价是一种价值判断活动,价值判断是评价的根本属性。如格朗兰德 (Gronlund, N. E.) 就把评价表述为下面的公式^①:

评价 = 测量 (量的记述) + 价值判断

评价 = 非测量 (质的记述) + 价值判断

也就是说,评价是在量或质的记述的基础上进行价值判断的活动。由于学业评价通常以知识与能力的评价为主,评价主要指向学生易于表现、易于测量的发展与变化,因此又称为“硬评价”。素质评价通常以情感、态度、价值观的评价为主,评价主要指向学生内隐的,不易测量的发展与变化,因此又称为“软评价”。

根据教育评价学者的界定,“硬评价通常包括:(1)设计能够辨别因果关系的研究设计;(2)搜集客观的、可靠的和有效的资料;(3)运用精密的统计技术对资料进行分析。”与“硬评价”相对应,“软评价通常包括:(1)一项至多项表明相关关系的研究设计;(2)主观的和评判性的资料;(3)对资料缺乏精密的统计分析。”^② 鲍尔 (Ball, S.) 指出,硬评价阵营中的极端分子视软评价为评价者没有主见和缺乏训练的表现。他们进而一致认为软评价的结果难以复制,并认为由于其主观性很难控制,以致根本无法说出什么是合理的结论、什么是武断的或对机遇事件的正规陈述。软评价阵营中的极端分子对硬评价的反感也毫不逊色。他们声称,硬评价未抓住某个方案影响的实质,忽视了那些用目前现成的各种客观测量难以恰当评定的变量,而且十分机械,依据一些对相对来说无足轻重的资料所做的统计检验,而不是依据理性判断来得出结论。^③

这样,就会产生一个直接的问题:什么样的评价所得出的结果是人们认为可以接受的呢?如果想要对某一班组学生的发展和变化进行评价,一方面,

^① Gronlund, N. E., *Measurement and Evaluation in Teaching* [M]. New York: Macmillan, 1990: 6.

^② 瞿葆奎. 教育学文集·教育评价 [C]. 北京:人民教育出版社, 1989: 775.

^③ 同上书, 776.

运用标准化测验得出具有统计意义的高均值测验成绩，另一方面，全体学生一致报告自己学习进步缓慢，那到底哪一种评价结果更为可信呢？这无疑取决于评价者所持的学生评价标准。当评价者属于“硬评价阵营的一分子”，他也许会更倾向于采用“硬心肠”的评价标准；当评价者属于“软评价阵营的一分子”，他也许会更倾向于采用“软心肠”的评价标准。这两种基本观点形成了学生评价标准方面的两大主流思想，一直争论至今。

（二）“硬心肠”的评价标准

“硬心肠”的评价标准与现代教育评价学的形成有着极深的渊源。现代教育评价学发端于美国，在发展初期，教育评价的对象较为单一，一般为军人或学生。19 世纪末至 20 世纪初，美国兴起“教育测验运动”。在 1887 年到 1898 年间，赖斯（Rice, J.）在美国实施了第一次正式的标准化测验。他在多个学区进行学龄儿童拼音教学的比较研究，结果发现每周花 200 分钟在拼音练习上的学区和只花不到 10 分钟的学区，在学习成就上没有显著的差异。赖斯在测验中并没有专门制定量化的标准，但是他提出了一个很重要的论点：教育人员必须成为实验者和量化思考者，并应用比较研究设计来探究学生的成就。^① 1904 年，桑戴克（Thorndike, E. L.）发表了《精神与社会测验学导论》，成为测量学上的经典之作。书中标有测验运动信徒们所遵奉的信条：“凡存在的东西都有数量，凡有数量的东西都可测量”^②。这一思想成为当时教育评价发展的主流思想。

所谓“测量”（measurement），本来是自然科学的方法，即按一定的测量尺度（评价标准）获取数量结果的步骤。在这里，测量尺度是极其重要的工具。测量必须具备一定的测量单位和测量的基点作为参照。不过，智力测验和学力测验一类教育测量的尺度，不像测量重量、长度之类的物理测量尺度那样，具有严谨的精确性和普适性。例如，在教育测量中的成就分数，如五

① 斯塔弗尔比姆. 评估模型 [C]. 苏锦丽, 等译. 北京: 北京大学出版社, 2007: 8.

② 李聪明. 教育评价的由来 [A]. 瞿葆奎. 教育学文集·教育评价 [C]. 北京: 人民教育出版社, 1989: 63.