学得好不好，不是比出来的——策展评估：从“谁更合适”到“什么是这一个”的范式转换

本文从一个被反复目击却从未被充分追问的教育裂隙出发：教师在日常教学中能清楚地“看见”一个学生独特的认知特质——那个解题慢但空间直觉极强的孩子，那个总是用笨方法但每次都能自己走通的孩子——然而这种“看见”，在涉及升学决策的制度层面，几乎从不被算数。制度承认的唯一证据，是经过标准化翻译后的可比较分数。本文论证了这一裂隙并非评估技术的不足，而是“可比较度量预设”的制度硬化结果：现代学校体系在规模压力与公平诉求的双重挤压下，将学业评价的地基浇筑为“比较—排序—分配”，使得认知的独特构造在制度层面变得不可见。

作者 阳涌 · 约 1.6 万字 · 发表于2026年7月24日

摘要

本文从一个被反复目击却从未被充分追问的教育裂隙出发：教师在日常教学中能清楚地“看见”一个学生独特的认知特质——那个解题慢但空间直觉极强的孩子，那个总是用笨方法但每次都能自己走通的孩子——然而这种“看见”，在涉及升学决策的制度层面，几乎从不被算数。制度承认的唯一证据，是经过标准化翻译后的可比较分数。本文论证了这一裂隙并非评估技术的不足，而是“可比较度量预设”的制度硬化结果：现代学校体系在规模压力与公平诉求的双重挤压下，将学业评价的地基浇筑为“比较—排序—分配”，使得认知的独特构造在制度层面变得不可见。本文在这一历史张力的追踪中，逼出“策展评估”这一概念：学业评价不是将学习者按同一把尺子排序，而是由学习者在评估者协作下，策展出一份能够忠实呈现其独特认知构造的认知档案，评估者依此为其匹配最适宜的发展路径。策展评估不可被第四代评估（Guba & Lincoln, 1989）的动态协商或档案袋评价的多源证据等已有范式所覆盖或1:1替换——其不可还原的内核在于：它要求评估暴露每个学习者“在无模板条件下独立试错”的生成性痕迹，以此作为识别认知构造的核心证据，而非依赖利益相关者的解释性共识。这一内核将评估从“比较谁更好”转换为“识别这一个，并为其找到最适配的去处”。本文以这一新命名为核心组织论证，追踪可比较度量预设的历史发生机制、剖析其在东西方两个制度样本中的硬化与圣化过程、正面回应布迪厄式的资本批判并以开放的制度设计保持张力，最后以一个可供检验的微观案例给出策展评估的操作图景与明确的证伪条件。

关键词：策展评估；可比较度量预设；学业评价；认知构造；教育公平

一、教师看得见，制度看不见：一道水沟题揭开一个被反复目击的裂隙

一个初中数学老师在课堂上给出一题：“一块长方形菜地，长12米，宽8米。王大爷沿着菜地四周挖了一条宽1米的水沟。水沟的面积是多少平方米？”

大部分学生能算出来。他们先扩大长方形的长宽各2米，面积140平方米，减去菜地面积96平方米，得44平方米。答案正确，步骤清晰。老师接着问：“为什么减出来的就是水沟面积？”一个孩子最终说：“大长方形减去小长方形，就是水沟。”

这个答案在语词上是对的，但它不是对图形关系的捕捉，而是对操作步骤的复述。那个孩子没有在心里把绕着菜地挖了一圈的水沟看成一个自己可以走进去的空间。他只是复述了一遍减法的规则。

同一间教室里，另一个孩子被问到同一道题。他愣了好一会儿，开始说：“我走过去。就是我沿着水沟走一圈……但是四个角会被走两次……”他在草稿纸上画了很久，画得很乱，把水沟切成四段直条和四个角上的小正方形，逐块加总，最后也得出44平方米。如果这是限时考试，他极有可能做不完。但他在一件事上远远超过了前一个孩子：他把那段抽象的文字，在自己心里重建成了一个自己可以走路的世界。

两个都算对了。如果只比较分数，他们都“学会了”。但任何一位有经验的数学老师都能感到：这两个孩子在认知深处完全不一样。前一个孩子“学会”的，是怎样把一道题翻译成一组已知操作序列——“拿到长方形套公式，条件变就修正参数”。后一个孩子“学会”的，是如何把一个陌生的文字情境重建成一个他可以进入、可以行走、可以亲手拆分的空间。前一种认知被当前的考试评价充分辨认并奖励：它标准化、高效率、易评分。后一种认知不仅不被特殊辨认，反而因为过程的“低效”而可能被整个评价体系惩罚——在限时的卷面上，他来不及走完他的路。

这个裂隙——教师看得见而制度看不见——不是个案逸事，不是某个教师不够努力去“看见每一个孩子”，不是某种可以通过“加强过程性评价”来弥合的疏忽。它根植于一个比任何教学法都更深的层面：我们如何判断一个人“学得好不好”这件事。在现有教育制度中，对一个学生“学得好不好”的判断，几乎无一例外地通过一道固定的工序来完成：统一任务、标准答案、数字评分、横向排序、按序分配机会。这道工序已经在全球教育系统中运行了一个多世纪，深入人心到我们几乎不再把它看作一种“可以被质疑的预设”，而把它看作“评价”这个词的自然含义。

本文要做的，正是把这套工序从“自然含义”的位置上拉下来，把它重新放回“一种特定的历史建制”的位置上，然后追问：这套工序在它被发明出来的时候，偷偷塞进去了一个什么样的、从未被充分质疑的预设？

这个预设，可以这样表述：教育的任何重要后果，必须首先被翻译成某种可以在不同个体之间进行统一度量和比较的量值，才能被公正地用于分配后续的稀缺机会。本文称之为“可比较度量预设”。这个预设并不声称“只有可度量的学习才是重要的”。但它规定了一件后果深远的事：即使教师在日常中看见了一个学生某种独特的认知特质，只要这个特质没有被翻译为一个可与他人比较的分数或等级，它就不能在决定这个学生的升学、分轨、获得何种教育机会的制度决策中具有效力。

这不是一个抽象的理论命题。它精确地描述了那位看到两个孩子差异的数学老师面对的现实：她知道那个用笨方法走通水沟的孩子拥有一种珍贵的空间重建能力，但她只能把他期中卷面上的72分填进成绩单——和他同班的另一个孩子考了94分，把所有标准解法都写得整齐无误，但从未在任何一次作业中暴露过自己的困惑。制度承认的，是那两个可以比较的数字。老师“看见”的东西，在制度层面不算数。

这个裂隙的哲学后果已被教育评价领域的批判传统反复触及。早在1970年代，Stake（1975）就提出“响应式评估”，主张评估不应遵循预设的标准化程序，而应响应被评估方案的独特特征与利益相关者的关切，以“个案叙事”取代“通用量规”。1980年代末，Guba与Lincoln（1989）在其“第四代评估”中更彻底地指出：评估不是测量某个先在的“真实”，而是评估者与利益相关者之间围绕多元价值进行建构性协商的过程；评估的终点不是分数，而是各方达成的共识性解释。这些批判已经瓦解了可比较度量作为评估“自然形式”的合法性，但它们始终在一道门槛前停下：它们从未真正回答，当评估必须用于升学这类高利害决策时，如何避免“不可比较的个案叙事”被决策者弃用、或退化为某种更隐蔽的权力运作。第四代评估可以把一个孩子的认知叙事说得很丰满，但招生官打开面前的两份申请时，他需要的仍然是一个裁决——而裁决一旦做出，他就已经在比较。

策展评估要进入的，正是这道被留下的窄门。它不是第四代评估的一个教育情境变体，也不是档案袋评价的更精致版本。它尝试解决一个第四代评估回避了的问题：如何在不把不同的认知构造翻译为可比较量值的前提下，仍然为升学决策提供可操作的、经得起外部审计的匹配依据？它的回答不是更精巧的叙事建构，而是要求评估过程本身暴露一个信号——学习者在“无模板条件下独立试错”的生成性痕迹——这既不是利益相关者协商出来的，也不是按量规评定出来的，而是从学习过程中涌出的、尚未被既有分类抹平的认知纹理。

二、可比较度量是怎样被炼成地基的：制度的代价

如果可比较度量只是一套可供选用的评价技术，那么它的局限早就可以被教学法上的改良所弥补——表现性评价、过程性评价、真实性评估、档案袋评价，每一种都试图超越标准化测验，捕捉更多维度的学习痕迹。但这些改良在涉及升学决策时，无一例外地被收编回了同一套逻辑：档案袋里的作品最终需要被压缩成一张量规上的等级分数，才能放进成绩单；真实性评估中学生的复杂表现，最终被标定为从“初级”到“精通”的某个可比较等级。这不是因为教育研究者缺乏想象力，而是因为可比较度量不是教学层面的懒惰，而是制度层面的硬性要求：当稀缺的升学机会必须在有限时间内被分配给大量申请者时，决策者需要一套能对外部质疑说“这是公平的”操作工序。可比较度量之所以无法被改良绕过，是因为它承载的不是教学功能，而是制度的合法性功能。

要理解这一承载是如何历史地发生的，必须回到可比较度量被制度化浇筑为地基的那个关键时期——不是去翻找某个“被设计出来的蓝图”，而是去追踪一套制度工序在特定历史张力的挤压下，如何从诸多可能的方案中被一步步逼出来，又在事后被使用者逐渐接受为“理所当然”的过程。

2.1 普鲁士：标准件的诞生

18世纪末的普鲁士，在拿破仑战争中的溃败逼出了一个前所未有的张力。邦国要重建军队与行政体系，急需大批能读能写、能听懂命令并能独立处置简单事务的公民。这意味着必须把基本教育推向此前从未受过教育的大批下层阶级子弟。然而当时的普鲁士没有足够的合格教师，没有统一教材，甚至没有足够的校舍。在这两条硬约束——国家功能对认知能力的需求、与教育基础资源的极度匮乏——的挤压下，任何精耕细作的个别化教学方案都是空中楼阁。

制度做出的结算，在事后被称作“义务教育体制”，在它发生的当时，是一整套被逼出来的次优方案：将知识从工匠行会与家庭情境中剥离，编制成统一的、可以在任何一间教室里重复使用的教学单元；将教师从一个需要长期师徒制训练的专门职业，重新定义为可以批量培训的“教学技工”；将学生的学业成果，用统一考试的“过”与“不过”来判断。这套方案的核心发明不是教学法，而是一道制度工序：把学习成果翻译成可比较、可汇总、可按行政需求统一处理的量值。

这不是任何人预先设想的“理想教育”。它是资源极度匮乏的条件下，为了达成最低限度的国民教育覆盖率而做出的制度性妥协。但这个妥协一旦被固定为制度常态，它就反过来塑造了后继几代人对“教育应该什么样”的想象：班级授课被接受为教学的自然形式；统一考试被接受为检验学习的自然手段；用分数对学生进行分层，被接受为管理学生群体的自然前提。“标准件”不是普鲁士教育改革的“过错”，它是那次改革在历史约束下被结算出的稳态。但这套工序的核心前提——学习可以被拆成标准单位进行传递与测量——在此后的扩散中逐渐被当成不言自明的地基。

2.2 三个咬合的代价维度

标准件化的工序本身并不“坏”——它确实在极短时间内历史性地扩大了教育的覆盖面。但它付出的代价需要被精确地识别，因为这些代价正是后续任何试图超越标准化评价的努力都会碰到的硬壁。

第一，知识从情境中被脱钩。在集体化的教学单元出现之前，学习大多发生在情境之中：学徒在作坊里跟着师傅边做边学；农家少年在田边跟着父辈边量边学。在这些情境中，知识不是抽象公式，而是卷入身体感知、空间直觉、生活经验与特定工具的活体。正因为知识活在情境里，每个学习者的认知构造就自然地留下痕迹：一个学徒同一块材料试了三种不同锻打方法，三种都失败了，但旁观的老铁匠能从每一次失败的位置和响声中识别出他哪种火候没过、哪种路数还没摸到门。这些痕迹本身就是“认知档案”的自然形态。

当知识被标准化地呈现在黑板上的那一瞬间，情境被抽走了。勾股定理从那块年年被洪水淹没的土地上，被剥离成一个用直角三角形和三个字母表达的等量关系。这层抽离本身不是“错误”——它是规模化的必然代价——但它带来一个深远后果：知识的标准化呈现，不再能接住不同认知构造的差异。一个用身体学的人，一个必须看到故事才能把住的人，一个需要自己先画一遍图才信的人——他们面对黑板上的同一个公式时，各自无法进入。在情境学习中，这种差异会被师傅自然地识别——他会换一种方式给这个学徒讲解，让他先做再讲；但在标准化课堂中，这些差异被统一归为“学习能力不足”。

第二，错误从路标变成宣判。在情境学习中，错误不是失败，而是关于“你离那地方还差一座桥”的信息。一个孩子说“我昨天放了很多面包给鸭子”，他用了“放”而不是“喂”。母亲没有纠正他，她顺着他的话说“对呀，你喂了鸭子”——她在搭桥。在母语学习与学徒训练中，错误恰恰是暴露学习者当前思维构造的最灵敏探针：一个学生反复在同一个类型的题目出错，不是因为他“笨”，而是因为他的认知构造恰好在这个位置留了一道缝——而这道缝，正是他最需要被识别的地方。

标准化考试中最致命的一刀，不是让学生多刷题。它把错误从路标变成了宣判。一个学生花了十分钟走了一条弯路，最后答案错了——在他的试卷上，这道题只得到零分。那个零分告诉他：你错了，但完全没有告诉他：你其实是在哪儿拐错的、前面那段思路其实是走得通的、你离正确的路径只差半步。制度工序需要这个零分，因为它要把所有路径收束为可比较的单向排序。一旦制度把错误标记为惩罚而非信息，学生就不再愿意暴露自己的真实错误——他在可比较排序中的生存策略，是把“看起来像正确答案的东西先背下来，确保自己的答卷上不出红叉”。路径收束没有停止行走，它把行走变成了对着脚印模板反复踏步——学生的考卷上留下了大量正确步骤，却没有一道题暴露出他是怎么想的。

第三，错误的暴露从协作资源变成竞争风险。在没有比较压力的学习共同体中，同伴之间的互问、互评、相互暴露彼此的认知路径，是识别各自构造的最自然通道。一个孩子在草稿上展示了另一种更笨的解法，旁边的孩子看见了，说“你这个地方卡住了”，两个孩子就此开始一次自发的讨论——这次讨论不仅让那个被卡住的孩子走通了，也让那个说出来的孩子更清楚地看见了自己已经掌握的东西。他们的认知在碰撞中各自显影。

一旦学业成果被纳入零和的次序——我进入前10%就意味着你失去一个进入前10%的机会——协作通道就被封上了一半。一个学生暴露自己的困惑，不是在争取帮助，而是在可比较的排序博弈中暴露了自己不够“快”。策展所需要的“错误暴露”——让评估者看见那些被真正走进去的弯弯路——在这种生态中是不可得的。

这三个维度不是时间上的先后步骤。它们是同一道制度工序从不同侧面咬合的后果：知识的脱钩让认知无法天然留下个人印记；错误的惩罚化让本已稀薄的印记也被打磨成正确率的面具；关系的竞争化让学习者之间彼此看见认知构造的协作通道被关上。它们一起完成了同一件事：让学习者的认知构造在制度层面变得不可见。这不是偶然的副作用，而是制度运行的一个必要前提：如果每一个学习者的认知构造都清晰可见且各不相同，那套“统一进度 + 统一考试 + 统一分层”的机器就根本无法运转——因为你没法把各不相同的东西排进一条单列里。

这三道代价不是制度的“缺陷”，而是它选择承担的一种负担：制度为了满足“在最稀缺资源下、以最可审计的公平方式、在最短时间内处理最大规模分配决策”这一功能，付出了“看不见认知构造”的代价。这道交易在它发生的年代是理性的、甚至是唯一可行的次优解。但当教育资源从极度匮乏走向相对充裕，当社会对人才的需求从批量达标走向多样态适配，这道旧交易是否仍然必须被继续遵守——这个追问，才是策展评估出现的发生学条件。

三、制度化圣化：高考“公平”叙事的东方样本

普鲁士锻造了可比较度量作为规模化教育的制度基座，中国则在另一条更深的文明谱系中，赋予了这套工序一种几乎无法被理性触及的神圣性。解剖这个东方样本，不是为了讲述一个地方性故事，而是要展示策展评估在分析上能够切开的一个旧概念切不开的层次：不是高考制度本身的“弊病”，而是“公平”这个被推至制度合法性顶点的话语，是如何在特定历史张力中被一步步铸造成了制度自我锁定的锁芯的。

3.1 科举的势能蓄水池

科举制度运行了一千三百年。它在一个前现代社会里，建立了“按统一考试的成绩分配政治权力与社会地位”的制度常规。这个常规并不是说从前的中国人相信考试是天经地义的——而是说，经过一千三百年的反复操作，“凭本事考出来改变命运”这个想象，已经长成了这个文明对“教育是用来做什么的”这一问题的默认答案。这套默认答案是一种历史势能：它使得此后任何一个时代的改革者，都只能在“考试—分层—改变命运”这条主轴上做加减法，而不能把整条轴换掉。

3.2 恢复高考：一次爆发性力量与圣化的完成

1905年废科举后，中国经历了半个多世纪的制度震荡——西式学制的引入、新文化运动的教育实验、民国时期的精英办学、新中国初期的工农速成教育与推荐入学——每一次都试图在这一断裂带上摸索替代性方案。但它们都无法解决同一个根本张力：在稀缺资源条件下大规模选拔人才时，究竟用什么样的规则才能让被淘汰的人服气？1977年恢复高考，在这个意义上不是一次简单的“恢复考试”，而是一次被蓄能了二十余年的制度的爆发性复位。对经历文革的一代人而言，“分数面前人人平等”这七个字带着历史创伤的慰藉与被剥夺感的释放，一齐涌回来。它不是一项选拔制度，它是一道公平的圣谕。

这种情感强度将“公平”这两个字在中国的教育话语中推到了一个几乎不能被理性讨论的位置。质疑标准化测试的任何环节，都可能被解读为质疑公平本身——因为在这套群体的心理基建上，标准化等于可被外部审计，可被外部审计等于不能被权力私下操作。这一叙事链条的硬度，不是来自任何人的刻意设计，而是来自文革记忆投下的那一道长长的伤疤。标准化考试——作为公平最易被审计的技术形态——获得了一份无形的免检执照。此后的教育政策史反复印证了它的硬度：1990年代素质教育改革试图加入更多非考试维度的评价，迅速被“这影响分数吗”的质问弹回；21世纪核心素养与综合素质评价的试点，触及综合评价的升学权重时，立刻被“凭什么你的评价比我的评价算数”的合法性焦虑质疑。政策文本可以写“多元评价”，但制度骨骼拒绝改变自己的受力结构。

3.3 不被看见的认知构造

在公平叙事的圣化效应之下，落入了可比较度量盲区的，不仅仅是教学的艺术层面。一个高三学生，他的认知档案呈现出“对空间直觉极强、但对形式化方程演算偏慢”的特质。他在遇到需要自己先构造出物理情境模型才能动笔的题目时，可以在草稿上画出极富洞察力的结构关系草图；但在高考物理卷面上，大量是给定公式、给定待入变量的计算题——因为只有这类题的答案才不产生评分争议。他的强项没有出现，他的弱项被充分测量。他在总分排序中掉到了重点线之外。

不是因为他不适合学物理。事实上，他那种不被考卷承认的空间构造直觉，恰是实验物理和工程设计领域极为稀缺的认知特质。他输给的，不是一个比他更会思考物理的人，而是一套从未打算测量他那种认知维度的度量工序。在制度的规范定义里，“公平”在此处是被兑现了的——所有考生

答同一张卷子，在同一套评分标准下被数字打分。但在更深的一层意义上，一个不测量某种认知特质的度量工序，不可能对该种认知特质的持有者是公平的。他被塞进一套工序里，这道工序从头到尾没有问过他“你是谁”，最后对他宣读了排位。他不是被比下去了，他是没有被看见。

这就是被圣化后的可比较度量所隐藏的最深的代价：制度可以认为自己是公平的，而同时——完全合法地——错失它所不测量的那些认知构造。

四、策展评估：从裂缝中发生的命名与操作构造

前三节追踪了可比较度量预设的历史硬化过程，直到高考制度中“公平”叙事的圣化效应。这不是为了诊断“制度的弊病”，而是为了逼出一个问题：如果制度为了满足规模化的分配功能而付出了“看不见认知构造”的代价，而这一代价在今天已经从“理性的无奈”变成了“对多样化认知的沉默否定”，那么在机制上，是否可以从现有的裂缝中逼出一种不同的评价方式——一种不以“比较—排序—分配”为工序核心、而以“识别—策展—匹配”为工序核心的评价方式？

本节不再先发明一个概念再铺它的定义，而是沿着前文铺开的裂缝，跟着问题走，直到“策展评估”被逼出来。

4.1 从裂缝中长出来的命名

前述裂缝有一个共同特征：它们不是“评价技术不够敏感”造成的——档案袋评价可以收集满一整个学期的过程性作品，形成性评价可以在每个教学单元后给出及时的反馈——但这些信息在遭遇升学决策这座硬壁时，总是被重新翻译为可比较的等级或分数，因为制度需要一个对外可审计的答案：凭什么录取A而不是B？改革在这道硬壁面前的反复弹回，暴露的正是可比较度量预设的一个结构性盲点：它把“公平”与“可审计”绑死在了一起。公平必须是“能在最短时间内向质疑者出示同一把尺子量出的数字”。在这一绑定下，任何不可比较的认知构造信息，即使再丰富、再真实，也无法在制度层面被接受为决策依据。

那么，是否有可能在保留“可审计”这一合法性要求的前提下，将“审计的对象”从比较出的差值，换成匹配的理据？换句话说，决策者能不能不只是对质疑者出示一个分数，而是出示一份关于“这个学生——而不是别的学生——为什么被匹配到这一路径”的辨明书？

这个问题的发生语境，就是“认知构造的不可通约性”被严肃承认的历史时刻。当认知科学四十余年的研究反复告诉我们——从Chi、Feltovich与Glaser（1981）观察到专家与新手的知识是按完全不同的深层原理而非表面特征来组织的，到教育神经科学对个别学习者不同认知路径的实证追踪——知识在被不同的人学会之后，不再是同一个东西；它们被每个人的先前经验、认知倾向、思维惯习和情感记忆，改造成了独一无二的认知构造。这个认知构造不是一条单轴上的高与低，而是一个形态——有人抽象压缩能力极强，有人在具体情境中直觉超常；有人靠逐层推演走得稳，有人靠

跳跃试探撞得远。这些形态之间的差异，是类别的差异，不是等级的差异。它们无法被翻译为同一根度量轴上的高下而不丢失其本质信息。

一旦承认这一点，可比较度量预设之下的“公平分配”逻辑就露出了一道更深的裂缝：如果公平分配的前提是“把所有人放在同一根轴上比高低”，而认知构造根本就无法被放在同一根轴上——那么整套工序从一开始就不是在公平评价，而是在做格式转换。它不是先判断再排序，而是在排序之前，已经把“不同”格式化为“可比”。

从这道裂缝中长出的命名，就是策展评估。

4.2 策展评估的临时定义与操作构造

因为它是从裂缝中长出来的，这个命名在被给出的此刻，它不该是一个被钉死的定义。但它需要一张可辨识的轮廓，以便后续论证能够运作。不妨先给出这样一段临时定义：

策展评估是一种学业评价范式，它不旨在将不同学习者的认知成果翻译为可比较的度量或等级，而是要求评估结果呈现一份能够忠实记录该学习者“在无模板条件下独立尝试、暴露真实修正轨迹”的生成性证据集，评估者依此辨明该认知构造作为一种独特形态的轮廓，并为其匹配最适宜其认知构造发展的后继路径。

这个定义的临时性质在于：它不是在宣告一种成熟的制度蓝图，而是在标定一个方向——这个方向区别于两个最邻近却必须分清的既定范式。

第一个是第四代评估。如前所述，Guba与Lincoln（1989）的建构主义评估已经跨出了颠覆性的一步：他们拒绝将评估对象视为等待测量的实在，主张评估是多元利益相关者围绕各自的价值关切进行协商建构的过程；评估报告不是一组分数，而是一个案例叙述，呈现多种视角的交叉与建构出来的共识。这一步已经瓦解了可比较度量预设的认识论前提——它证明评估不必先有一个统一的尺度，而可以从个案内部的张力中生成“这情形到底怎么样”的判断。

但第四代评估没有解决一个制度上致命的问题：当两个“案例叙述”被放到招生官的桌上，他必须做出裁决时——录取谁——他的裁决实际上已经在比较两个叙述。第四代评估提供的“协商共识”，一旦进入高利害分配决策的场域，要么被重新压缩为决策者私下所作的隐性比较（从而丧失它宣称的反标准化性格），要么就根本承担不了分配功能而被制度弃用。

策展评估要接住这个掉落的球。它不同于第四代评估的那个点，不是它提供了“另一种更精确的叙事”，而是它在“叙事”中锁定了某一个特定类型的信号：学习者“在无模板条件下独立试错”的生成性痕迹。这个信号不是利益相关者“认为”这个学生有什么特质——第四代评估可以做到这一点，但它的叙事可以被善于表达的一方操纵成一个漂亮的故事。策展评估要求的是“暴露”：要求评估材料的收集过程包含那些无法被预先准备好的时刻——当学生面对一个没有例题可参照、没有熟手可模

仿、没有已训练过的标准步骤可直接套用的任务时，他是怎么入手的？他卡在哪里？他怎么从卡住的地方再往前走？这些痕迹，不是谈话谈出来的，是学习者在一个“被迫自己构造路径”的任务中留下的行走记录。

4.3 在碰撞中逼出的控制变量：生成性无模板暴露

这里到达了策展评估与最近邻理论最需要切开的那条分离线。第四代评估的承重变量是什么？不是“协商”这个表面操作，而是它所依靠的证据生成方式：利益相关者在互动中各自提出自己的解释，评估者综合各方理解建构一个共识性的案例描述。这个过程的输出，是“被解释过的学习”——是各个观察者和参与者已经赋予了意义的、经过了语言整理的学习痕迹。第四代评估虽然反对标准化量表的裁判权，但它并没有要求证据本身必须包含一个特定的结构：学习者在自己尚未给经历赋予意义时那种尚未被规整过的摸索痕迹。

策展评估所依靠的核心信号，在这一点上不同于第四代评估。它所识别的那种生成性痕迹，不是事后被学生或老师叙述出来的，而是在被看见的那一瞬间，尚未来得及被翻译成一个“我能阐述清楚”的故事。它的可靠度，不取决于利益相关者的叙事一致性，而取决于那个被暴露的卡住、弯曲、再走通的过程是否真实地发生在评估材料的收集现场。在这个意义上，策展评估将评估所依据的证据，从“建构出来的看法”推前了一步，推到了“被逼着露出来的生成纹理”。

这一步的后果，是让第四代评估与策展评估在同一个场景下会产生不同的判断。设想一个学生，口才极好，非常擅长自我叙述。在第四代评估的协商现场，他可以就自己的认知特质做出极富洞察力的解释，让评估者深感信服。但另一个学生不善言辞，说不出自己是怎么想的，只能在给他一道没有例题参照的陌生任务时，在草稿上走出一条弯弯绕绕但最终自己走通了的路——他没办法用语言把这条路解释出来，但他确实走了。第四代评估可能会认为前一个学生呈现出更强的“认知反思能力”并据此在叙事中赋予他更高的匹配推荐。策展评估则必须认定后一个学生那份说不出但走得通的草稿，才是认知构造更真实的线索——因为那份草稿上的痕迹不是被语言建构出来的，而是在一种被逼着自己走的状态下自然留下的。

这就是两个范式在匹配决策上的分离点。策展评估的标准提问句式不是“你怎么看自己的学习”，而是“给我们看看，在一个你从未见过的东西面前，你做的第一件事是什么”。

4.4 认知档案的构造原理

到这里，可以给出“认知档案”在策展评估中的构造原理了。认知档案不是一袋随意收集的作业。它由评估者按照一个特定的识别逻辑来组织：只选入那些在“无模板条件”下生成的、暴露了学习者不被预设路径遮掩的真实认知运动的痕迹。

这样的痕迹在现实教学生活中其实大量存在——一个孩子周末自己琢磨一个不会做的题目时，在草稿本上画的那些谁也看不懂的歪歪扭扭的线；一个孩子在课堂上脱口而出的一个“不对”的答案，但他说的方式暴露了他是在一个方向上试图自己往前推；一个孩子在已经上交的作业本边上，自己又用铅笔记下的一次更简的、未经老师批改的自证尝试。这些材料在现有评价体系中毫无价值——它们不在统一评分标准测量的范畴里。但在策展评估的构造原理中，它们恰恰是最有价值的原料。因为只有在没有标准答案预期、没有评分导向的环境中，一个学习者的认知构造才会褪去伪装。

当然，不是所有学生都拥有同等的留下这些痕迹的机会。这种不平等，正是策展评估必须内置防操纵设计的根本原因——它必须保证评估材料是从一个标准化设计的“陌生任务”中当场生成，而不是从日积月累的不对等课外准备中挑选。这一制度设计将是下一节回应的核心。

五、来自布迪厄的质疑与一种开放性的制度设计

策展评估在完成不可还原校验后，立刻面临一个无可回避的质疑：即使取信于“生成性无模板暴露”作为辨识认知构造的核心信号，这件事本身难道不是一种符号暴力吗？

5.1 认知构造的分类是否也是符号暴力？

布迪厄批判最锋利之处，不在于“精英操纵评估结果”——那是批判的表层。他的深层判断是：任何一种被制度承认的、用于区分人的分类体系，本身就是一种符号暴力，因为它把某一特定社会阶层的感知图式、认知倾向和话语习惯，默认为区分人的合法标准。布迪厄在《国家精英》（1989）中追踪法国重点高校的招生，揭示的正是这一机制：那些看似中立的“综合素质”评价维度——品味、修辞、从容、广泛的文化旨趣——恰恰是中上阶层生活方式的延伸。它不是精英刻意操纵了评估，而是评估所承认的“值得被看见的东西”，本来就是在精英家庭中被养出来的东西。

这个判断不会因为策展评估把维度从“文化旨趣的广泛性”换成“在陌生任务面前的独立试错倾向”，就自动失效。一个在资源丰沛家庭长大的孩子，从小被鼓励“做错了没关系，试试看”，他在陌生任务面前的从容度与被暴露的探索链路长度，本身就是一种家庭惯习的痕迹。另一个在匮乏环境中长大的孩子，从小被环境的容错率逼得只能给出最稳妥的答案——他在陌生任务面前的第一反应可能是等待指示，而不是跳进自己摸索。

策展评估确实面临被这种优势捕获的风险。它声称通过“强制暴露无模板任务中的生成痕迹”来绕过精英的精致叙事，但连这种“敢在陌生面前暴露自己”的倾向本身，都不是均匀分布在社会光谱上的。这是一个必须诚实承认的张力。本文的回答不是“它不会被权力捕获”，而是“它给权力捕获增加了什么标准化考试所没有的阻力”。

5.2 制衡设计：让捕获成本高于刷题

标准化考试中，精英再生产的主要路径是资源堆叠：更好的补习、更强的刷题训练、更多的模拟考。这套操作的致命优势是它的可重复性与可预测性——有钱有闲就能做，而且做完就确实能提高分数。在可比较度量框架下，这种策略是完全合法的：它只是让一个学生在同一根轴上跑得更快。

策展评估如果要做制度设计，它的核心逻辑就是把“精英操纵的成本”提高到至少不低于、最好远远高于标准化考试中的刷题成本。这可以通过以下几项强制性的制度要求来实现。

第一，任务的不重复性与非预先准备性。策展评估所采用的“陌生任务”，必须是每一轮评估中重新生成的结构开放型任务，且不公开任务池。它不是让学生去背诵一套预先可能的陌生任务题型——因为一旦它能被题型化，它就退回了可比较度量。学生真正被置于一个“你必须现场让它对你发生”的任务里，他要调用的是他真实的认知取向，而不是已经训练好的应对模板——那种模板，正是精英家庭最擅长用课外辅导提前为孩子铺设的。陌生性在这里不是一个形容词，而是一项被制度强制保证的现场条件。

第二，试错痕迹的不可伪造性。标准化答题中，学生可以靠反复刷题把正确答案的路径背下来，交出一份看不出任何犹豫的完美答卷。但在一个设计得当的陌生任务中，一个未曾真正自己进入问题的学生，无法凭空制造出逼真的“走弯再走通”的痕迹——因为那种痕迹包含着局部的迷路、自相矛盾的尝试、在某个环节突然切换思路的断裂。这些非连续的特征，靠事后的自我叙述是编织不出来的，或者即使被试图模仿，其僵硬程度在评估者的专业识读中也很容易被从真实的生成纹理中区分出来。任务的陌生性在这里和痕迹的现场性构成了双重保险——你无法提前准备，也无法事后表演。

第三，评估者共同体的跨阶层构成。策展评估的匹配建议，不由单个教师也不由单个学校单独决定。制度设计要求认知档案的识读，必须经过一个至少包含来自不同社会阶层背景的评估者参与的共同体审核。这个设计不是出于身份政治的象征性包容，而是因为认知构造的辨识本身，最容易被识读者自身的阶层惯习所误读——一个在中产家庭中长大的识读者，可能下意识地把“善于用语言描述自己思路”等同于“认知构造更清晰”。来自不同背景的识读者可以对彼此的误读进行交叉制约，使匹配建议必须在一个被多方审查的环境中通过。

第四，匹配建议必须附带“异见档案”。这是策展评估制度设计中最独特的一笔。任何一份匹配建议，如果评估共同体内部存在实质分歧——例如，一部分评估者认为该学生的认知档案呈现出某种构造形态，另一部分评估者认为其中更显著的信号指向另一种形态——那么这些分歧本身必须被记录为匹配建议附带的“异见档案”，一同提交给升学决策机构。这意味着决策者看到的不是一份被裁定的“客观评估”，而是一组带有内部张力的识读结果。异见档案的存在，使得任何试图私下操纵评估结果的人，都面临被内部异见暴露的风险。这同时也使得策展评估在认识论上保持诚实：它承

认自己对认知构造的识读是有限度的、可争议的，而不是一张宣称已经看清全部真相的判决书。

这四项设计共同构成了一种张力结构——策展评估不是在消除权力对评估的渗透（这不可能），而是在更换权力渗透的路径成本。标准化考试中，精英通过堆叠训练资源来渗透路径，这条路成本虽高，却是透明的、成熟的、被普遍接受的。策展评估使得精英如果想维持同等程度的渗透力，就必须支付一种更难以规模化的成本——不是“多付补课费”，而是要持续地在每一个评估现场绕开陌生任务的不重复性、试错痕迹的不可伪造性、评估者共同体的跨阶层交叉审核与异见档案的制衡。《国家精英》不是策展评估的对立面；它恰恰为策展评估提供了必须被置于制度核心的警惕装置。真正的诚实，不是在理念层面自证不受权力影响，而是在制度层面布置权力无法一次性清扫的阻力，并允许这些阻力在未来实践中被持续地调整。这个博弈是开放性的，没有终点——但这恰恰是策展评估比标准化考试“更不伪装公平”的原因。

六、一份认知档案的长相：从水沟题走进来的男孩

前面五节完成了从理论到制度设计的论证。但有一个要求尚未兑现：让读者“看见”一份认知档案是什么样的，策展评估的识别与匹配是怎么操作的。本节给出一个基于真实观察启发的微观案例，不是为了提供通用模板，而是为了展示策展评估在操作层面如何被构造。

6.1 案主与情境

案主正是开篇那个在水沟题上用笨方法走通的孩子——姑且称他为小路。小路当时初二，数学成绩在年级中游晃荡，期中期末都在70分上下。他不调皮，上课听讲，作业完成——但他的正确率波动很大：看似简单的计算题偶尔出错，期末卷的最后一道综合题却有时能给出一个半对的、思路极偏但并非全错的答案。他的数学老师在对他的日常观察中有一种直觉：这个孩子不是“学不好数学”，他是“用和别人不一样的方式在学数学”。但这话没法写进成绩单。在一次家长会后，她决定对小路的数学学习做一次深度的跟踪识读——这个动作，成为了策展评估的原型实验。

6.2 档案的采集与构成

在接下来约半个学期的跟踪中，老师有意识地采集了四类材料。

第一类是被迫生成的陌生任务记录。老师说服了学校，在下学期的数学期中考试中为小路单独换掉最后一道综合题——不是降低难度，而是把题目类型从“给定模型，按套路代入”换成“描述一个陌生的、需要自己先在纸上建出一个结构才能进入的问题”。小路拿到的题目是这样的：“一个正方体被切成两个完全一样的多面体，切面必须经过正方体的中心。画出来，并解释你是怎么切分它的体积的。”这不是一道训练过的题型。小路在草稿区画了一个又一个透视图——第一个只是几条随手画的线，第二个标上了他认为的中心点的位置，第三个开始做辅助面。他的正式作答只写了一页，但他的草稿纸留了四页。那四页草稿纸被老师完整保留下来，作为档案的核心基础材料。

第二类是课堂追问的录音转录。老师征得小路同意后在几次课堂讨论中做了针对他的录音，事后请他一起回听。在一段讨论中，老师问全班：“两个三角形面积相等，一定全等吗？”小路在底下低声说了一句，老师当时没听清。回听时发现他说的是“要压扁的”——他在脑子里把一个三角形的一条高往下压，面积不变但形状变了。他没用“等底等高”这个标准术语，但他已经把那个关系在自己脑子里做成了一幅可以扳动的手柄。

第三类是修订历程的时间线。老师把小路一学期中三次重要作业的原始提交、批改反馈、以及小路的修订稿按时间排列。其中一次，小路在作业旁用铅笔轻轻画了一道自己想的简化解法——没按老师教的步骤，但思路是对的。这个铅笔痕迹以前谁也没注意过。老师在小路的修订记录边注了一行字：“不是偶然一次。他习惯在用规定步骤做完后，自己去试试有没有别的路。”

第四类是老师与小路的两次半结构化谈话的录音整理，老师事后写成田野笔记。谈话中老师没有直接问“你怎么学的”，而是把一张他之前画过的草图拿出来，问他当时卡在哪，怎么卡过去的。小路回答得很慢，有时答不上来就拿起笔画。有一处他说：“我当时觉得这个角一定要先画，不然转不过来。”老师事后在记录中写道：“他是按自己必须先在空间里找到一个落脚点，才能开始推理的路径在走的。不是乱，是有他自己的进入次序。”

6.3 识读与匹配

学期结束，老师把这几类材料摊在一起，邀请了一位长期从事认知学习研究的外部研究者、一位曾经教过小路但不知道这次实验背景的校内同事、以及一位来自本地一所职业技术学院的招生老师，共同做了一次协同识读。

他们注意到几条反复出现的主线。第一，小路对空间关系的进入方式高度依赖手绘与实物想象——他需要先在纸上“把东西画出来”，或者在心里“走进去”，才能开始推理。一旦任务缺了这个视觉化前奏，他的表现就明显下降。第二，他极少直线走通一个题，而是在草稿上反复试探，试探的路径不是随机的——他的错误有方向性，每次试错都离正确结构更近一步，但代价是远远比标准解法耗时。第三，他对标准术语的提取偏慢，但当他用自己的话命名过那个结构之后，他对该结构在新情境中的识别反而比一般学生更快——他的理解是活的，只是活在他自己的那套语法里。

三人共同体在评估建议上很快达成了主体共识：小路的认知构造呈现出一种“空间直觉驱动、需要自建视觉锚点方可进入形式推理、在容错条件下具有较强关系迁移力”的形态。这种形态在当前的标准化笔试中被系统性低估——因为考试不给画草图的时间，也从未在评分标准中承认草图的价值。在匹配建议上，三人给出的共同建议是：在接下来的初高衔接阶段，不要按常规路径把小路径送入刷题型重点班（那里他的弱项会被反复惩罚、强项则终身无法被辨认），而是将他匹配到一所侧重项目制学习的学校或课程轨道，在那里他可以在给定的开放性物理或工程问题中，用自己的视觉建构方式进入推理，并在项目推进中被逐步支持形式化表达的精炼化。

这份匹配建议，不是给小路的最终“定性”。它附带了由评估者共同体签名的一份简短异见备案：那位职业技术学院老师指出，小路在形式符号的慢速提取可能在实验室物理环境中会被有效代偿，但也不排除他在后期遇到对符号抽象能力要求更高的物理分支时遇到阻塞。建议匹配的同时应注明需定期重审，看他的形式化能力是否在适切的支架下获得了补偿性增长。这条异见，被原文保留在提交给家长和升学协调员的档案中。

6.4 这个案例说明了什么

这个案例不是要建立一个“策展评估就该这样做”的操作规程。它只是呈现了一种可能性：当评估过程的强制性要求不再是“交出一份标准答案”，而是“在足够陌生、无现成模板的任务下留下完整的认知运动痕迹”，并由一个跨阶层、带有异见保留的评估者共同体来识别这些痕迹时，一个在旧制度中只能被看到“70分”的少年，第一次在制度的意义上被看见了。这并不保证他走上坦途——策展评估不承诺童话结尾，它只是保证一件事：他的认知构造开始算数了。这个变化不是情感上的，是制度逻辑上的。

七、结论：一个可以被证伪的新范式

本文的论证路线，是从一道水沟题暴露的裂隙出发，追查到可比较度量预设的制度硬化过程，到它在东西方两个样本中的圣化与锁定，再到从裂缝中被逼出一套以“识别—策展—匹配”替代“比较—排序—分配”的评价方式，最后给出认知档案的构造原理、制度制衡的设计、以及一个可供检验的微观案例。策展评估这个概念，不是在前人成果上换一张新标签；它与第四代评估的分离点，在于它不依靠利益相关者的叙事共识来生产评估判断，而依靠评估过程中被逼出来的、尚未被语言规整的生成性痕迹。这个分离点，才是它作为认知构造识别范式的不可通约内核。如果替代性评估改革要接受本文的论证，它就必须回答：在升级决策中，你是否愿意承认那些“被逼着露出来的、走弯路再走通的痕迹”是比“解释自己学得怎样的话语”更可靠的评估依据，并愿意为此重新设计你评估材料的收集与管理方式——包括在招生中承认需要制度承重的异见，而不是虚设的多元。

但一个理论的可贵并不在于它坚固，而在于它告诉使用者在什么条件下它自己会失效。以下给出策展评估的证伪条件。

如果未来实践显示，在陌生任务中暴露的认知痕迹，与学习者在后续长期发展中的实际认知质量之间，不存在显著且可被独立复现的对应关系，那么策展评估的效力基础即告瓦解。如果制度实施后的持续观察发现，防精英操纵的四重制衡设计——陌生任务的不重复性、试错痕迹的不可伪造性、评估者共同体的跨阶层构成、匹配建议必须附带异见档案——被实践证明可以被系统性地绕过或收编，精英操纵的成本并无显著高于标准化考试中的刷题策略，那么策展评估作为一种范式，就不比它所批评的对象更值得被推广，应被重新评估乃至放弃。一个不能为自己的失败提前划定坐标的范式，还只是一个漂亮的说法。上述条件就是它的坐标。

教育被评价方式锁定的故事，讲了很久了。本文试着扳了一下锁舌——不是撬开它，是让你听清楚它里面什么东西在响。