认知中介的负值转译：标准化评价如何把“好”结算为“错”

不是被忽视，而是在评分媒介的正常操作中被系统性地转译为可精确赋值的负向指标

阳涌 著 · SDE 学员 · 教育评价与测量理论 · 2026年7月26日

摘　要　本文追问一个被反复讲述却很少被真正在生成层面处理的难题：当标准化评价必须依赖一套有限分辨率的评分媒介以保障公平与信度时，那些落在媒介可处理范围之外的认知品质——独立判断、未成形的直觉、对前提的尖锐质疑——如何不仅是“被忽视”，而是在媒介的正常操作中被系统性地转译为可精确赋值的负向指标？通过拆解评分细则的内部操作逻辑，并引入常模参照的选拔零和性作为关键中介，本文揭示这一“负值转译”既非评价者的疏忽，也非权力的压制意图，而是评价媒介在真实制度条件下的一项结构性功能。这一判断使本文与福柯的规训分析、默顿的目标置换诊断、布迪厄的符号暴力理论，乃至Campbell定律、Goodhart定律和Messick效度理论产生了不可通约的分离：它将“好何以变错”的追问从代理人意图或指标腐化，推进到媒介认知分辨率本身所设下的生成机制界限。在此基础上，本文不提出一套应然的改革方案，而是转向已经存在的制度缝隙：追踪教师非正式实践、表现性评价的异化以及弹性空间如何被日常评价文化收编，识别出一种在现行体系内部正被挤压而出的结构性需求——不是“解放创造力”，而是为那些必然被标准化媒介错译的认知品质，建立一块在制度上被正式承认的、负值转译在此失效的认知守护地带。这一进路的证伪条件在于：如果这种守护空间无法自发地生长出一种稳定的辨认性评价文化，而是被培训机构和应试惯性重新刻写为新的得分模板，则本文关于其制度可能性的判断即为错。

关键词：评价媒介；负值转译；标准化评价；认知守护；常模参照；生成机制

一、引言：一份被“公正地”错评的答卷

2007年，美国某州写作测试中，一篇考生作文引起了评分者的激烈争论。题目要求学生围绕“一项需要勇气的事”展开叙述。大部分答卷写的是常见的素材：在比赛中扭伤后坚持跑完、第一次公开演讲、与欺凌者对峙。这份争议作文写的却是一个认知事件。考生详细描述了自己花了几周时间，反复尝试用不同方式证明一个数学猜想，最终在一个深夜放弃了所有教科书上的标准路径，构造了一个极其笨拙但有效的非标准论证。他将那个独自在草稿纸堆中把自己逼到认知绝境、又把自己从绝境中拽出来的过程，称为“需要勇气的事”。

两位独立评分者依据统一的评分量规对这份作文进行了评定。一位认为它选材独特、逻辑清晰、细节饱满，应当赋为最高等级；另一位则指出，根据该州写作量规中“思想与内容”维度的具体描述，最高等级的作文应当“展现对主题的深刻洞察，并以有力的例证支持中心思想”，而这份答卷“通篇聚焦于一个狭窄的数学解题过程，未能展示对该主题的广泛理解”，应予降等。在评分信度规则的要求下，两位评分者的分歧分数差超过了一个等级，试卷被自动送入仲裁评分组。仲裁评分组最终裁定：基于“严格执行评分量规以保障全体考生公平”的操作原则，该作文被定在第二等级，比标准叙事低了整整一个档次。[1]

这个故事之所以值得被反复解剖，不是因为它又一次显示了创造性被压制——这类故事早已被反复讲述。它之所以重要，是因为在仲裁评分组的最终裁定文件中，两位评分者之间的分歧被明确地、公事公办地表述为“对评分量规的递送偏离”，而不是“对认知品质的识别失败”。在那一纸裁定中，发生了一件极精密的事：一份在认知意义上更有价值的产出，在经由评分量规这一制度化认知媒介被处理时，它的独特性没有被“忽略”——比忽略更严重的，是它被精确地转译为量规中“思想深度不足”这一维度的标准扣分项。它不是在操作中被漏掉的沙金，而是在媒介的加工流水线上，被标注为一个有据可查的瑕疵，打包进了一个无可上诉的分数。

这就是本文试图命名的核心机制：认知中介的负值转译。它不是在描述评价者偶尔看走了眼，而是在追踪一个以追求客观、公平、高度信度为设计前提的评价媒介，在正常、忠实地履行其功能时，如何必须、且必然会将它无法精确表征的那些认知品质，在它自己的操作逻辑内部，铸造成一套可量化的负值。这个过程不要求任何人的恶意、不预设任何权力的压抑意图，它只是评价媒介的“有限分辨率”在高利害的、需要跨案例一致辩护的制度条件下，被推到极致的必然产物。

这一判断之下的追问，与传统批判哲学的深层逻辑发生了根本的分叉。福柯的规训分析拆解了现代学校如何通过层级监视与规范化裁决将个体塑造成驯顺而有用的身体。默顿的目标置换诊断揭露了科层组织中手段如何悄悄吞噬目的。布迪厄及其后继者的符号暴力理论展示了教育的“标准”如何充当阶层文化的合法伪装。这些分析各自拥有不可替代的锋芒，但它们的锋芒都指向同一个方向：它们都预设了某种“未被扭曲的应然”——一种没有被权力渗透的认知、一种不曾被手段偷换的目的、一种未被阶层偏见沾染的标准——并将现实的教育评价视为对这一应然的背离。它们都是既成事实论。

而本文的追问是生成机制的：不去问教育“本来应该是什么样子”，不去预设一个纯净的应然起点，只去问——当大规模公共教育治理不得不依靠一套有限分辨率的评价媒介来同时完成知识底线的共同守护、选拔排序的公平执行、以及对公众问责的制度性回应时，这套媒介的认知边界本身，是否就已经设置了一种结构性条件，使得某些认知品质不再是被“压制”，而是被这套边界内部的操作逻辑日常化

地、可辩护地转写为一种制度化的次等品？ 如果答案是肯定的，那么出路就不在于“回到教育的真谛”，而在于诚实地接受评价媒介的结构性盲区，并在此前提下思考：我们能否为那些注定被媒介错译的认知品质，在制度内保有一块媒介逻辑在此失效的空间——不是通过改良媒介，而是通过为媒介的必然错误建立一个制度性的补偿地带？

为使这一追问不流于又一个美丽的悲情叙事，本文必须完成以下工作。第一部分（第二、三节）与既有传统正面碰撞，不仅划清界限，更在碰撞中逼出“负值转译”这一概念的不可还原性——它切开的是一道连Campbell定律、Goodhart定律与Messick效度理论都无法独自占位的缝隙。第二部分（第四、五节）展开负值转译的生成机制微观机制：从评分量规的内部结构，到常模参照的零和竞争，再到管理轨道与问责轨道的连锁效应，追踪一个认知冒险行为如何在制度的多个接触点上，被逐次结算为负值。第三部分（第六、七节）转向已经发生的非正式实践：在下沉的教育现场，一批教师、学校与评价试验正在日常的缝隙中生长出守护那些认知品质的脆弱空间；本文对这些空间的发生条件、结构性特征与内生风险进行厚描，并在布迪厄式的阶层再生产反驳下，论证这些空间已经展现出的制度免疫能力及其内在的递归性——守护空间本身，也可能成为负值转译的新场地。最后在结语中，给出本文判断的证伪条件。

这种进路的选择，同时决定了本文的方法论位置：它不进行因果推断，不构建可检验的假设，也不提供普适的应然方案。它所做的，是一种对制度性认知媒介的生成机制分析——通过将评价媒介视为一套具有有限分辨率的认知处理系统，追踪其在高利害条件下的运作如何必然将某些认知特性“结算”为负值，并在此基础上追问，从该系统的内部张力中，能否、以及如何生长出对这一必然性进行部分补偿的本地化探索。这一方法论的源头不在社会学实证主义，而在一种受体制现象学与认知人类学启发的制度分析传统中——它关注的是制度性认知工具（量表、量规、计分表）如何反向约束了制度自身的认知范围。

二、三大批判传统为何没有看见这张底牌

在展开本文的核心论证之前，有必要把本文的分析位置与三种影响深远的批判传统严格分开——不是出于学术礼仪，而是因为如果“好被结算为错”不过是“规训”“目标置换”“符号暴力”的又一例证，那么本文就没有提供任何不可还原的洞察。

先看福柯的规训分析。在《规训与惩罚》中，福柯展示了现代权力如何通过层级监视、规范化裁决与检查这三种技术，制造出驯顺而有用的身体。学校考试正是这三种技术的完美交汇：它把学生置于不间断的比较与分级之中，将每一个个体书写进一套可累积、可比较、可治理的档案。这一分析无疑捕捉到了标准化教育中一个真实的维度。然而，规训分析在解释“物理题作文”这类案例时，碰到了它自身逻辑的边界。在这类案例中，阅卷者的决定——把一份展现出独立认知投入的答卷评为次等——并非一次权力的“行使”。恰恰相反，阅卷者是在小心翼翼地避免行使个人任意权力：他放弃了自己对这份答卷的私人欣赏，转而将判断的权威完全移交给评分量规中那些预先设定的、可公开核查的条款。他的动作不是“压制”，而是“自缚”。他把自己缚在量规的文本上，因为只有这样，他才能在事后面对申诉时，拿出一个不由他个人偏好所玷污的理由。这件事在福柯的坐标中几乎不可见：因为福柯笔下的规训权力固然是无主体的、毛细血管式的，但它仍预设了一种“正常化”——一个行为被惩罚或矫正，是因为它被权力网络判断为“不正常”。而在阅卷者的操作中，他没有在判断“正常与否”，他只是在执行一条规则：凡是无法被量规的现有条款正面覆盖的答卷特征，都必须被归入已有的扣分格中，以确保评分操作的跨案例一致性。这是一种比“正常化”更冷、更数学的排斥——它不依赖于对“正常”的界定，它只依赖于对“可被精确记录”这一条操作命令的忠诚。在这个意义上，阅卷者不是权力的代理人，他是媒介的操作员。他的排斥，是媒介在运行其自身的代码。

再看默顿的目标置换诊断。这一概念的经典表述是：在科层组织中，成员对规则和程序的无条件服从会悄然取代他们对组织宗旨的追求，手段膨胀为目的本身。标准化教育无疑是这一诊断的理想案例：考试本来只是检验教学效果的手段，却变成了教学为之服务的目的。但本文的案例暴露了目标置换范式中的一个隐蔽的、却致命的预设：它预设了组织的“真正目的”曾经是清晰且被共识性地共享的。在公共基础教育的现场，这个预设从一开始就站不住脚。当一位阅卷者面对一份超出量规预期的答卷时，他所遭遇的不是“手段”与“目的”的对立，而是两个同样正当的制度目的同时向他发出相互矛盾的操作指令——“准确辨认并奖赏高认知品质”与“确保每一分的给出都能在分数复查中被无争议地辩护”。这两个目的都是真实的，没有哪一个比另一个更“本质”。阅卷者最终把答卷判为次等，不是因为“手段偷走了目的”，而是因为在两种相互冲突的目的在同一个评价现场短兵相接时，制度性的问责压力迫使他必须优先选择那个在事后最难被外部攻击的目的。他选择的不是手段，而是在当前制度条件下的“最安全的目的”。这种在多个正当目的之间被迫做出生存性排序的困境，是目标置换理论所不能覆盖的——因为该理论只能诊断一个目的被替换，却无法处理多个目的之间的制度性搏杀。

布迪厄的符号暴力理论将前两种批判推向了更精密的社会学纵深。在布迪厄与帕斯隆的分析框架中，教育标准远非中立的学术准则，而是特定阶层的文化资本——对语言的某种区分性使用、对知识的某种疏离而自信的修辞姿态、对模糊性的某类特定容忍——被制度性地误认为普遍的、天然的才华标准。这一分析极其锋利，但它有一个被它自身的洞察力所遮蔽的盲区。在小学阶段，一位山区学生必须学会准确地执行四则运算、清晰地组织一个自然段的书面表达——这些能力的标准化要求，并不仅仅是“阶层品味的伪装”。它们是一个社会得以让其所有成员平等参与公共生活所必须铺设的认知基础设施。一个无法清楚书写、无法可靠计算的孩子，无论出生于哪一阶层，他未来的行动能力都将受到实质性的损伤。在这个意义上，标准的部分内容，确实具有跨阶层的工具正当性。但本文的分析焦点不在标准是否中立，而在于：标准——无论其内容是否带有阶层偏见——一旦被制作成一套必须在有限时间内由多人一致执行的评价量规，它就自动获得了一种认知分辨率上的有限性。有些认知品质，比如“对前提的深度质疑”“在无标准路径的迷惘中自己长出方向的能力”“在一堆矛盾证据面前暂停判断的克制”，它们无论在哪一阶层的孩子身上出现，都天然地难以被任何以精确性和跨评分者一致性为设计硬约束的量

规所正面表征。它们不是被“排除”出标准的定义，而是从一开始就难以被“写入”量规的操作条款。因此，当布迪厄的继承者说“标准就是阶层伪装”时，本文的补充是：即使在一个完全消解了阶层偏见的平行世界里，只要那里的教育仍在用一套追求信度与可辩护性的有限分辨率媒介来评价全部学习产出，负值转译就会持续发生——因为媒介的有限分辨率，不是阶层的偏见，而是认知表征这件事本身的技术代价。符号暴力揭示了标准是如何不公的，但无法解释一种连不公的意图都未曾参与的排斥——因为这种排斥，在某种意义上，恰恰是标准在试图保持其公正（即保持其内部的一致性与可核查性）时，所必须付出的代价。

这三笔对质指向同一个判断：福柯、默顿和布迪厄的批判，在深层逻辑上都是既成事实论——它们各自以不同的方式，诊断了一个偏离某种应然状态的病态秩序，并把这种偏离归因于某种坏力量（权力、科层理性、阶层利益）。而本文所追踪的那个过程——一份认知上的“好”如何在媒介的操作中被结算为制度上的“错”——并不要求那种坏力量。它只要求一套评分量规、一位公事公办的阅卷者、一个高利害的选拔考场，和一条在制度上无可辩驳的操作原则：能写的，必须写得进去；不能写的，必须被归入已有的扣分维度。前者奖励的是可归类的正确，后者惩罚的是不可归类的优质。两者来自同一条原则，而这条原则正是公平操作的评价媒介为了公平地裁决大多数，所必须依赖的那条脊骨。

三、一个被遗漏的对手：Campbell定律、效度理论与“反转”的不可通约性

在第二节中，本文与福柯、默顿和布迪厄的对质，为“负值转译”清理出了一块传统批判哲学无法占位的地基。但在社会科学中，还有一系列距本文更近、因此也更危险的理论竞争者：以Campbell定律、Goodhart定律、Messick效度理论和测试的反身性文献为代表的教育测量与政策评估传统。这些理论已经高度精密地诊断了量化指标如何歪曲其原本意图测量的社会过程。如果“负值转译”只是这些诊断的又一次教育情境复述，本文就失去了它存在的理由。

必须正面回应这一挑战。

Campbell（1976）在他的经典论文中提出了那个日后被称为Campbell定律的判断：任何被用于社会决策的定量指标，都会因其决策用途而受到腐化压力，从而丧失其作为过程监测指标的有效性。[2] Goodhart定律的原始表达域在货币经济学，但其通用版本——一旦一个指标被选定为政策目标，它就不再是好的指标——已经成为了社会量测的普遍常识。在教育领域，大量实证研究印证了这条定律：当标准化测试分数被用来决定学校排名、教师绩效和学生升学时，教学系统会产生一系列系统性行为扭曲——从压缩课程、牺牲非测试科目，到赤裸裸的作弊。Nichols和Berliner（2007）将这一系列现象命名为“间接伤害”，Koretz（2017）则在《假装让学校更好》一书中，用无数案例揭示了应试教育如何教会了学生在测试中“表现好”而非真正“掌握好”。[3]

乍看之下，本文所描述的“独特性被扣分”似乎只是这一定律的又一个例证：当教学的实质目标被测试形式所窄化，那些不符合测试格式的认知品质自然就会被排挤。但Campbell定律所捕捉的“扭曲”，具有一个清晰的结构：它假设在指标被引入之前，原本的教学实践中是存在一个相对未被扭曲的状态的；指标的滥用侵入了这个状态，腐蚀了它，导致人们去做那些“看起来好”的事而不是“真正好”的事。换句话说，Campbell定律需要预设一个指标之外的“真实过程”，这个真实过程先于指标而存在，指标只是绕开了它或遮蔽了它。

但本文所追踪的现象，发生在先于这个“腐化”的更底层。在作文评分的现场，那位仲裁评分者没有在“做一件看起来好而实质上坏的事”，他也没有为了追求高测试分数而压缩课程。他是在忠实地执行一套评分量规——这套量规是任何标准化评价若要保障信度，都必须随身携带的认知模具。在这个模具里，根本就不存在一个可以正面容纳“独立但与量规预期不符的认知投入”的赋分格。这无关指标腐化——因为那个被称颂的“更好的教育”，从一开始就写不进这套模具的操作条款里，不是模具被腐化，而是模具本身的分辨率不足以分辨那种“好”。因此，Campbell定律即使完全不存在（假设指标从未引发腐化行为，教学始终在追求真知），物理题作文仍然会被制度性地评为次等，因为评价者仍然只能依据有限的量规条款来给出可辩护的分数。负值转译是评价媒介作为一套认知表征系统自身的结构效应，不是其使用者的行为被指标腐化的效应。

教育测量学自身的核心概念——效度——同样不能覆盖本文的机制。Messick（1989）在他集大成的构念效度统一理论中，将“构念代表性不足”和“构念无关变异”识别为威胁效度的两大来源。[4]其中，“构念代表性不足”指的是一个测试没有充分涵盖它所宣称测量的那项能力的全部重要维度。如果一项测试旨在测量“写作能力”，却只评了语法和拼写而忽略了逻辑组织与声音的独立性，那么这项测试在构念代表性上就是不足的，因而其效度受到了损害。这似乎十分接近本文的论证：标准化评价的问题在于没有涵括“批判性思维”或“创造性”等构念维度。但本文的焦点不在构念的全或不全。即使一位测试设计者在设计写作量规时，已经竭尽全力地将“声音的独立性”作为一个独立的赋分维度写入了量规，并且用基准样例和评分者培训来试图稳定对该维度的评分，现场评分中仍将发生一轮更微妙的转译。因为一旦“声音的独立性”被操作化为量规中的一条条款——比如“文章展现出一种独特的个人视角”——它就会立刻被下一轮的应试训练吸收：培训机构将归纳出“独特视角”的高频表现模式（比如使用第一人称细碎叙事、引入某个冷僻但安全的知识点），并训练学生有技巧地表演这种“独立”。结果，量规所赋分的将不再是“独立性”本身，而是独立性的标准可读痕迹。真正不可归类的独立性——那种让评分者感到“我写不出这条量规来预见它，但我一读就知道它是真的”——仍将在量规的条款之内无处立身。这是Messick的效度理论所无法诊断的一个层面：它不是在测试效度之外的额外坏东西，而是效度本身在追求可操作化时所必然携带的、更深一层的表征

困境。任何为评价而做的构念操作化，都会产生一批“漏网之鱼”——这批漏网之鱼不只是没有被测到，它们是在被测到“反面的构念无关变异”时被当作负面证据处理的。

最后，表现性评价（performance assessment）与真实性评价（authentic assessment）的倡导者们早已在技术上回应了封闭式试题的局限。Darling-Hammond和Adamson（2014）展示了精心设计的表现性任务、详细的基准样例和经过严格培训的评分者团队，可以使得对复杂建构（如科学探究、历史论证）的评价同时获得可接受的高信度。[5]一些教育体系（如澳大利亚昆士兰州的高中校外评价、国际文凭IB的内部评估）已经在操作中将开放任务与信度要求结合在一起。这似乎是对本文论点的一个直接事实反驳：如果表现性评价已经有了信度，那么你的“媒介分辨率”就只是一个过时的稻草人。但这一反驳与本文的判断处在两个层次上。本文不否认在特定条件下（小规模、高投入、大量培训时间、评分者社群的文化高度同质），开放任务的评分信度可以达到统计意义上的可接受水平。本文的判断是另一件事：在超大规模、高利害、且必须承受来自公众不可量化质疑的公共教育选拔评价中——例如中国的高考、美国的州际标准化考试——那些依赖阐释社群内部规范而非精确条款的评价方式，总会在制度压力下被逐步修剪掉那些最不适合被公众话语辩护的边缘。这不是技术问题，是制度合法性来源的问题。一套需要上百万家长“看得见”其公平性的系统，在其维持合法性的过程中，必然将评分操作压向那些最能抵抗申诉的、最为“条款化”的形式。本文讨论的“负值转译”，恰恰是这一制度压力下的极端形态，而非每一处表现性评价的普遍命运。

由此，可以给出这一概念与Campbell/Messick传统的不可通约性之判别格。在非高利害、低选拔压力、评价结果不用于公开问责的评价情境中，Campbell定律和Messick效度框架足以解释大部分评价扭曲：教师可能为学生做“模拟高分”训练，测试可能窄化构念。但本文的“负值转译”的独特预测是：即使不存在任何指标腐化行为（没有任何教学为了考试而教），即使测试的构念已经被充分且技术上高质量地操作化，只要评价媒介仍是一套追求跨评分者一致条款的系统，并且评价结果被纳入常模参照的高利害选拔，那么那些无法被任何量规条款正面赋值的认知品质，就必然会在现场评分操作中被归入最相关的扣分格——不是因为量规“不完整”，而是因为量规必须对每一份答卷的每一个可辨识特征都给出一个分数，而唯一“安全”的对于不可归类特征的赋分，就是把它归入一个可辩护的扣分项。 在低利害、标准参照（达标）评价中，不可归类的特质更可能被忽略（得零分但不扣总分）或被写上一句鼓励性的评语；而在高利害、常模参照（选拔）情境中，不归入扣分格就意味着这个不可归类的特征“非法地”为学生争得了一个优势位置，这将引发其他学生的公平质疑。那一刻，媒介的逻辑从达标切换为相对剥夺，“不加分”和“缺失加分项”不再是零，而是一个相对劣势。正是这额外的零和竞争中介，把“被忽略”推成了“被转译为负值”。

因此，本文的“负值转译”并非Campbell定律或效度理论的升级版，而是揭示了后者因专注于意图、腐化与构念而从未触及的一层机制：制度化的认知媒介为了自我保护（以维持公平可辩护性）而将不可归类物强行归类为缺陷的操作必然性。 这一机制必须同时满足三个条件：（1）有限分辨率的评价媒介；（2）高利害的常模参照选拔竞争；（3）评价结果的公众问责制。缺一则强度递减。抓住这三个条件，就抓住了“好被结算为错”的制度生成机制。

四、从量规条款到负值转译：一个微观的评分生成机制

上一节在与测量传统的对质中逼出了负值转译的三个必要条件。本节将显微镜对准其中一个条件的内部——评分量规的条款结构与现场阅卷者的风险计算——来展示“独特性”是如何在量规的操作条款中，被一步步转写为“缺陷”的。

评分量规不是一个中性的观察透镜。 它是一组穷尽的赋分条款。在标准化写作测试中，一套典型的四等级量规通常包含三到五个赋分维度，例如“思想与内容”“组织结构”“语言表达”“规范书写”。每个维度下的各级描述，都是一幅预期中的最好答卷的心理像：最高等级的“思想与内容”通常被描述为“见解深刻，立意新颖，例证充分”，次一等级则为“中心较明确，但例证单薄或立意较一般”。量规的条款并不宣称它可以描述每一份答卷的全貌，但在操作上，每一位阅卷者都被明确告知：必须在量规的现有条款内确定每一份答卷在所有维度上的等级，不允许自创加分项，不允许以“超出量规但不能忽略的品质”为理由偏离定级。因为任何超出条款的判断，都会被还原为“个人偏好”，从而在分数复查中失去制度性辩护。这便是量规在保护评分信度时所必要的“操作封闭性”。

在这一操作封闭性之下，一位面对“解题叙事”作文的阅卷者，在量规面前完成了一系列精确的认知操作。第一步，他需要将这份答卷与量规中的各级描述进行比对。在“思想与内容”维度，最高等级的密钥是“见解深刻”。当他朗读“花了几周尝试不同方法，最终在深夜找到了一个非标准解法”这一段落时，他在私人阅读中可能感受到一种真实的认知强度。但量规不处理感受，量规只处理可以被条款识别的东西。他接下来必须向自己提出的问题是：“这个‘花了很长时间自己找了一种方法’的叙事，是否吻合量规对‘见解深刻’的期待？”量规对于“见解深刻”的范例是：文章揭示了一个人生道理、提出了一个对主题的深层反思、或展现了对复杂问题的多层次理解。解题过程——即使再艰难——在量规的语汇中，通常被预编在“叙述详实但思想深度不足”的那一类里，因为它不处理人生道理，不反思社会关系，不挖掘人性。它处理的是数学，一个在教育叙事谱系中被长期放在“学术能力”而非“深刻思想”一端的领域。因此，叙事框架本身就被媒介自动地归类为“不够深”的例种。

第二步，一旦阅卷者在量规中找到了最接近的匹配项——次级的“中心较明确但立意一般”——他就等于完成了量规的比对。此时，那份在私人阅读中可能被辨认出的“认知投入”，已经在制度层面被无情地蒸发掉了。工具没有提供给它一个储存格。但蒸发不等于中

性：被蒸发掉的东西会在另一个维度被重写回来。因为量规必须对每一份答卷给出完整的侧面描述，次级的定级意味着这份答卷在“思想与内容”上“存在缺陷”——立意不够深、洞察不够广。当这一缺陷被写进分数报告时，它就不再是“一个独特性未被识别”的中性事件，它变成了“该考生在思想深度方面有所不足”的档案记录。在高利害的常模参照排名中，这意味着他的总分因这个维度的“不足”而低于那些在标准叙事中写出安全深刻感的考生，不是因为他写得更差，而是因为他的“好”落在了量规的可赋值区之外。

这样，微观的评分动作就完成了一次冷酷的代数：不可分类的优质 = 条款缺位→归入已有缺陷条款→被赋负值。 这便是负值转译的操作公式。阅卷者在此的关键抉择，并非无视灵气，而是面对灵气与条款之间的不可通约，做出了那个在制度上最无可指摘的抉择。他的理性不是教育理性，而是一种制度生存理性——通过忠实于条款来保护自己免受申诉，并通过严格操作来保护整个评价系统的公信力。Lipsky（1980）的街头官僚理论为这种生存理性提供了更普适的解释：一线工作人员（街头官僚）在资源不足、规则冲突和工作量压迫的情境中，会发展出一套简化复杂现实的“工作常规”，以便在日常的高压下做出可结算的决策。[6]阅卷者每日评阅数百份试卷，他们的核心工作常规，就是把复杂的文本化约为量规条款中的可核对点。这一常规在提升效率的同时，也系统化地将那些难以被条款捕捉的品质简化为了缺陷。

但这一微观机制还需要一个宏观推手，才能完成从“被忽略”到“被反转”的质变。这个推手就是常模参照的高利害选拔逻辑。在纯粹的标准参照（达标制）中，如果独特性未被加分，它对本人的影响最多是该维度得零分，而其他维度仍可得高分，只要总分达标即可。但在常模参照的排名竞争中，每一分都代表位置的相对移动。一份未能拿到“思想深度”最高级分的答卷，相对于那些拿到了该级分的竞争答卷，就因这一维度的“不足”而排名下降。独特性不被积分，直接转化为它是竞争劣势。更重要的是，在制度性的公众问责下，如果一个考生因“独特性”获得了额外分数而未被量规条款覆盖，则落后者可主张该加分缺乏依据，从而撼动程序公平。因此，在考后的申诉与核分机制中，唯一安全的分数，是每一个赋分都严格落在量规可核查条款内的分数。不可条款化的特征，因此不仅在操作上被排除，而且被反转为：如果给它加了分，反而是对程序公平的破坏。于是，“独特性不加分”从一种认知被动状态，被制度保护公平的压力推成了“给独特性加分是不公平的”这一规范性判断。

至此，负值转译生成机制的全链条可以概括为：有限分辨率的量规条款 + 阅卷者的制度生存理性（条款依赖以规避申诉风险） + 常模参照的零和竞争逻辑 + 对程序公平的公众问责压力，四者相互作用，使得一份在认知上更有价值的产出，经过评价媒介的制度性操作，被不可逆地铸造成了一个在评分体系中位置更低的、可精确辩护的次等品。

五、在管理的缝隙和问责的账本里：负值再生产的三重日常

上一节聚焦于评价现场的生成机制。但负值转译并非只在评分瞬间完成便永久封存。在日常的学校治理中，管理的进度表、自我的教案账本和课堂中的风险计算，像三条相互咬合的传送带，把“认知冒险”这种尚可的负值，一次次再生产为更为沉重的制度负担。这不是三条平行的机械轨道，而是同一个制度肌体的三处关节点——一处对认知冒险的“扣分”，会在另一处被计算为“时间浪费”，在第三处被记录为“管理风险”，三处相互引用、相互强化，最终将一种认知行为稳固地锁定为一种制度偏差。

管理的进度表：时间如何在核算中消失。 在任何接受政府教育督导的基础教育学校，每学期的教学周数、每门学科的周课时量，早在学年开学前就以“教学进度计划”的形式被固定并上报。省、市级的统一考试时间是倒推全部进度表的最终铆钉。这份进度表并非没有正当性——它保障的是不同学校、不同班级间教学品质的最低程度可比性，是教育公平在时间分配维度上的粗糙但必要的化身。但进度表在操作上是刚性的：每节课的内容被预设，章节的完结日期被标定。任何一次的“偏离”——比如因为学生的意外提问而展开一段无法提前写入教案的深度讨论——都会在进度表上制造出一个时间缺口。这个缺口可能在本周看不出来，但三周后，当教师发现本班进度较平行班落后一章时，该缺口就被精确地量化为一次“教学延误”。由于延误的最终罚则将在统一考试的成绩中兑现，教师和教研组通常会在事后迅速压缩后续内容的教学时间，以抢回进度。而压缩的代价，恰恰是被牺牲掉的那些原本可能再次发生的认知迷路时刻——即每一个宝贵的、可能将讨论推向更深处的“第二个意外提问”。一次认知冒险，在进度表的核算中，被生产为若干次未来的认知冒险被预先清除的理由。

问责的账本：教案如何被迫防御性写作。 与进度表并行运行的，是教案检查制度。教师被要求在上课前撰写详细教案，教案必须包含教学目标、教学重难点、教学过程、评价方式等要素，并定期（通常是每周或每两周）上交教务处或教研组接受检查。这套制度的本意，是确保每一位教师都不是无准备地走入课堂，守住教学底线。但一旦它与绩效考核、职称评定和评优评先挂钩，它就从质量保障的下限工具，变成了束缚教学探索的上限枷锁。一个真实的课堂场景可以说明这种束缚的微观机制：在一节初中物理课上，教师原本计划讲解浮力计算公式。一位学生突然举手问：“如果把铁船放在水银上，船还会浮吗？”这个问题偏离了预订的教学轨道，却触及了密度比较的核心认知。教师临时决定花十五分钟，与学生一起推演不同液体的浮力表现。课后，他需要在教案的“教学反思”一栏中记录这节课的实际进展。他现在面临一个制度性的账目难题：他该如何为这十五分钟的“计划外”时间做出一个不会被判定为“教学组织不力”的说明？如果他在反思中诚实地写“因学生意外提问而调整教学设计，临时展开对浮力与密度的探究，属课堂生成”，他可能被评课者认为“预先准备不足，未能有效引导课堂回到预设轨道”。如果他隐瞒不写，而其他班级的进度已超过他，他在平行班的比较中同样处于劣势。最安全的写法是把这个“生成”用教学术语包装为“根据学生情况适度扩展密度概念理解”，并补写几个“应有”的教学目标——瞬间，一次真实的认知冒险被重写为一次计划的、安全的、“有效课堂教学”的规范档案。

在这套账目中，被记录的不再是课堂上到底发生了什么认知层面的珍贵事件，而是一系列“可核算”的痕迹：教学目标是否被勾掉，时间分配是否合理，反思是否展示出对教学不足的改进意识。真实的认知冒险，因为缺乏可被条文化为“合规痕迹”的格式，就在这些账目中被系统地擦拭掉了——不是被故意隐瞒，而是因为不能写入而未被写入。当它未被写入时，它就在制度中消失了，连同它所可能唤起的后续冒险。

评价、进度与问责的三位一体。　要紧的不是这三个关节点各自为政，而是它们构成了一个相互查阅的闭环。评价轨给出的分数和等级，是向管理轨提供“教学有效性”的证据；管理轨的进度数据，定期向问责轨输送“教学延误”的标红项目；问责轨对教师的学期考评，又反作用于下一轮的进度压缩与课堂安全化决策。在这个闭环里，“不可核算的认知品质”每在一处被转化为负值（扣分、延误、事故风险），这一负值就会在另外两处被引用为确证它的证据。一份物理题作文被扣分，不仅自身受损，而且被教研组引用为“我们应加强立意训练”的证据，从而进一步缩小未来教学中的立意冒险空间。一个探究性讨论造成的进度滞后，被校长在教务会议上当作“个别教师课堂管理松散”的实例，从而强化了全体教师对进度的服从。就这样，一次认知冒险被三个制度关节点同步结算为负值，并且这些负值之间互相提供合法性，组成了一张极难被单个教师或单个学校撕破的网。

这不是任何个人的阴谋，而是制度为了维持其最低限度的可管理性、可比较性与可问责性，所必须付出的认知代价。至此，“负值转译”已经从微观的评分操作，流淌为弥漫在整个学校治理肌体中的日常实践。

六、缝隙中无声的生长：日常实践中被挤压而出的认知守护

前面的分析揭示了负值转译的制度性根源及其日常再生产的深度。但现实教育现场并非完全受制于这一逻辑。在大规模标准化治理的缝隙和张力中，一种尚未被充分命名、也不成体系的教育实践，正在被挤压而出。它不是自上而下的改革方案——每一次自上而下的改革，最终都被评价媒介的逻辑重新吸收——而是自下而上的、在个体教师和教研组层面零星发生的非正式调整。这些调整的共同特征是：它们都在制度的内部、在不对抗底线达标的标准化评价的前提下，为那些无法被标准媒介精确表征的认知品质，临时划出了一小块“媒介此时不在场”的缓冲地带。

教师的课堂间隙实践。 在浙江某所城市初中的语文教研组里，一位从教二十年的教师，在她的每周教学计划中保留了一个她自己称为“三不管时间”的半小时区间。这个区间不赶进度，不设任何教学目标，不布置必须上交的作业。在这半小时里，学生可以自由阅读任何她带来的文本——一首诗、一篇科普短文的片段、一段新闻评论——然后进行一场不要求形成书面结论的随性讨论。没人被强制发言，没人被打分。这位教师描述她的初衷时，用的不是任何改革话语，而是一种朴素的职业直觉：“我知道我教给他们的那些东西，考试之后大部分会被忘掉。但这半小时里发生的某些对话——孩子们突然自己碰撞出的想法——他们可能会记得很久。”她没有使用任何量规来评价这些对话的质量，但她发展出了一套高度个人化的“辨认能力”：她能听出一个学生的发言是真诚地在自己组织理由，还是在重复从家里听来的现成观点；她能辨认出一次讨论是真正在原有基础上推动了半步，还是在绕圈。

这位教师的实践，具备了一种关键的结构性特征：她在不挑战标准化评价的前提下，在时间分配和评价豁免上，为认知冒险创造了一个制度性的例外空间。这个空间的边界是清晰的——它被严格限定在每周的固定时段，不侵占底线达标教学的核心时间；它的风险是可控的——学生不被赋予任何正式分数，因此不存在评分不公的申诉风险；但它的存在本身，向学生传递了一个沉默却强烈的信号：这里有一些不一样的事情在发生，在这里，你的个人理由比你的分数更重要。

档案袋里被遗忘的“粗糙件”。 在一所上海的高中，物理教研组在全校的统一考试之外，每学期要求学生提交一份“探究记录”——内容不限，可以是一道难题的多种解法手稿，可以是一个实验的失败记录，也可以是对教科书某个陈述的质疑短文。这份“探究记录”不纳入期末总分的正式核算，但教研组会在期末家长会上，向家长展示其中的优秀案例，并由任课教师给每位学生的记录亲手写一段非模板化的、针对其具体思考过程的评语。几年运行下来，发生了一件教研组未曾预料的事：一些在正式物理考试中成绩中等甚至偏下的学生，在这份记录中展现出了令人惊讶的耐心和独特思路；而一些考试高分的学生，记录里却只有对标准解法的整洁复述。教研组长在一次校际交流会议上说的一句话，可以看作对这种非正式评价之意义的最好注脚：“我们不能把这些东西折成分数加进去，那就又毁了它。但如果我们连看都不看、连一段认真的评语都不写，我们凭什么说自己是在教物理？”

这句话里包含着本文所要论证的“认知守护”的全部核心：守护的第一要求，不是为不可量规化的品质另设一套更好的量规，而是在制度中正式地承认：存在一些重要的认知品质，我们无法用标准化的分数来公平地对待它们；但我们可以在分数之外，在师生的直接互动中，用辨认性而非测量性的方式来郑重地对待它们。这位教研组长不设计新量规，他只做两件事——保留一份不纳入总分排序的、但被正式展示和回应的产出档案，以及要求教师为每一个学生的这类产出，提供不负有分数压力、从而可以暂时跳出扣分焦虑的独特反馈。这种做法的结构性效应是：它在学生心中建立了一种区分——有一些学习，是要为考分负责的；另一些学习，是要为自己思考的连贯性负责的。二者并不冲突，前者保障了共同底线，后者为前者不可能覆盖的认知品质保留了一个正式的、被教师和学校郑重对待的位置。

“认知守护”的结构性定义。 从上述两类实践中，可以提炼出“认知守护”作为一个生成机制的分析概念，而非应然的规范方案。它不是“自由探索”，因为自由探索不需要任何学校制度层面的承认与组织。它也不是“表现性评价”，因为表现性评价追求的仍然是通过培训、基准样例和量规使开放的认知过程变得可评分，而守护的首要操作是部分暂停评分的通货功能——让那部分不可被标准化媒介公平

处理的认知投入，免于被转译为一场排名游戏中的筹码。守护，是在现行评价制度的功能性盲区之内，通过时间豁免、评分暂停和辨认性反馈这三个基本动作，为那些必然会被负值转译的认知品质，创造一个“媒介逻辑在此暂时被悬置”的场域。

对于这种守护实践，必须直面一个布迪厄式的拷问，一个任何教育补偿方案都无法回避的反驳：它是否只是为新形态的阶层再生产洞开了大门？能够在这半小时里“自信地说出自己想法”的学生，会不会本身就拥有更多的文化资本？能够提交一份令人惊讶的“探究记录”的学生，其家庭是否早已为他提供了更优裕的认知启蒙？换言之，认知守护的空间，是否会在无意中成为把阶层优势洗白为“独立思考能力”的新装置？

七、守护的代价：收编的风险与递归的负值转译

上述反驳是真实而沉重的。任何教育系统内部的弹性空间，一旦被赋予具有某种象征权力的可见性，就会立刻被社会中的优势阶层识别为新的竞争筹码。认知守护的空间不会幸免。一旦学校和教师开始郑重地对待学生的探究记录和自由发言，并为此投入注意力，那么一场围绕这种注意力的阶层竞争就会悄然展开。优势家庭将迅速为子女提供这方面的课外培训——不是培养真实的独立思考，而是培训“如何表现出独立思考的规范痕迹”。这种培训的效果会非常显著：一个经过训练的学生，会更擅长在课堂讨论中引用一些看似个人化、实则社会认可的“批判视角”，更擅长在探究记录中模仿那种“自己从迷惘走到清晰”的叙事结构。教师如果缺乏足够时间的深度的面对面了解，极有可能将这种经过精心叙事包装的“表演独立性”辨认作真实的独立性，而那些真正在孤独中挣扎、尚未找到组织经验的语言的寒门学生，则再次在辨认的门槛前被沉默化了。

在这里，负值转译的幽灵没有消失，而是发生了递归。标准化评价媒介的有限分辨率，已经无法正面表征某些认知品质；认知守护试图绕过这一步，通过暂停评分来为这些认知品质保留一个不被转译为负值的空间。但暂停评分本身并不能自动生产教师辨认真实独立性的能力。教师用于辨认的时间、心智空间和专业判断工具，同样是有限分辨率的认知资源。当一个教师面对三十份探究记录，并要在有限的工作日内给出非模板化的评语时，他不可避免地会发展出某种简化常规——从文本中扫描那几个最易辨认的“独立思考”的表面指标（如引用了多少不同的来源、使用了多少“我怀疑”“我最初以为”的句式），并用一套温和而通用的评语来覆盖大多数。此刻，一个新的、更柔软但也同样精确的负值转译就在这个本该是安全区的空间内部上演：真正的认知独立性（那种笨拙的、句式不通畅的、尚未学会如何包装自己的想法）再次被漏过，而包装精美的伪独立性再次被正面捕获。认知守护的空间，从“媒介逻辑在此暂停”的豁免区，滑向了另一种媒介逻辑——一种基于教师有限注意力与辨认常规的媒介逻辑——的运作场。

这并不意味着守护是徒劳的。它意味着，任何声称已经“解决了”负值转译的终局方案都是既成事实论的虚妄。生成机制只能诚实地说：负值转译是认知媒介自身有限性在制度压力下的必然物，它无法被一劳永逸地消除。认知守护，不过是承认这一点的前提下，在现存的制度内部主动制造一些局部地带，使得负值转译的齿轮在此处运转得稍微慢一些。这种制造本身，也必将在新一层被收编的风险之中生存，并需要持续的、反身性的维护——正如那些至今仍在坚持“三不管时间”和“探究记录”的教师们，他们每年都在根据收编的苗头调整自己的操作：今年把讨论形式从全班自由发言改为四人小组的随性笔记，明年把探究记录改为过程性的对话记录而非最终作品……他们不是在设计一套终极的免疫结构，而是在进行一场持续的、永远没有终点的、对负值转译递归的制度性拉锯。

由此，本文关于“认知守护”的判断不是一套可供移植的制度方案，而是一个对已经发生的日常实践的结构性特征的识别：第一，它们在制度内部创造了局部的时间与评分豁免；第二，它们依赖教师或教练的辨认性注意力，而非另一套量规；第三，它们被严格限定在不威胁底线达标主导地位的边缘地带，从而在制度问责压力下存活；第四，它们永远处在被收编为新的阶层竞赛筹码的风险之中，其存续有赖于实践者的反身性调整。这四个特征，不是认知守护的“完美定义”，而是它真实的发生条件与内在的脆弱性。

这一分析同时也回应了“强基计划”与综合素质评价的官方改革。这些改革正式授权学校在分数之外给予某种注意。但其运作逻辑已经迅速被负值转译的旧媒介侵蚀：当综合素质评价成为高校录取的参考甚至硬指标时，它的每一个子项就开始被量规化、被培训化、被阶层资本所捕获。这不是改革者的失败，而是任何试图用同一套媒介逻辑（标准化记录、跨案例比较、公众问责）去捕捉不可量规化的认知品质时，都必然会重现的同一个剧本。在这个意义上，真正有效的认知守护，永远不会来自一个可全国推广的完美量规，而只能来自那些在制度缝隙中由具体的教师和学生共同维系的、不可规模化的、暂时豁免于媒介评分的地方性实践。

八、结语：在不安全中思考，在诚实的脆弱中守护

本文的行进至此，可以抽取出三个逐级递进的判断，它们共同构成了“负值转译”这一机制分析的基本骨架。

第一，不是权力压抑了独立判断，也不是阶层偏见污染了中立的标准，而是标准化评价媒介为了公平地裁决大多数而在其技术内核中内置的有-限分辨率，与常模参照的高利害选拔逻辑相遇后，必须且必然地将那些不可提前写入量规条款的认知品质，在操作中归入可辩护的扣分格。这一过程不依赖于任何人的恶意，也不依赖于指标的腐化行为；它是媒介的结构功能，而非媒介的使用错误。

第二，这一机制无法被同一媒介逻辑内部的任何技术改良——更丰富多元的量规、更长的阅卷时间、更精细的培训——所根除，因为它的根源不在媒介的“不够好”，而在媒介自身的“必须把不可归类物强制归入可定价类”的生存本能。每一次提升量规分辨率的努力，

都将催生新一轮的对新指标的安全化表演，从而在新的层面重现负值转译。发生在浙江的等级制异化，发生在综合素质评价上的收编，不是改革的偶然失败，而是改革必然会踏入的窠臼。

第三，承认上述必然性不等于放弃。在真实的学校日常中，一些教师和教研组已经找到了一种与负值转译共处的局部实践：他们不是在改良量规，而是在媒介必须暂时退场的时间与空间里，为那些明知不可被媒介公正对待的认知品质，保留一块被正式承认的、暂停评分通货功能的区域。这些实践不是完美的解决方案——它们自身容易沦为阶层再生产的软性伪装，且需要教师充沛的专业注意力与反身性维护，从而面临被日常惯例化和资本收编的持续压力。但它们的不可替代性在于：它们是在认识到评价媒介的必然局限之后，仍不放弃与之周旋的诚实姿态。这正是本文以“认知守护”命名的东西——它不是一套方案，而是一种在制度的缝隙中持续发生的、不断自我修正的、不妄称可以终结负值转译但执意不让它吞噬一切的日常行动。

证伪条件。 本文对“负值转译”机制的分析，依赖于三个条件的并存：量规的条款封闭性、常模参照的高利害选拔、以及公众问责压力推动下的“可辩护性优先”。如果在一个教育体系内，大规模评价不再承担高利害的个体排序功能（比如只用于学校层面的诊断，成绩不与个人升学直接挂钩），或者评价的公众问责文化被替换为更信任专业判断的文化，那么本文预测的负值转译强度将显著降低，并可被现有的Campbell定律与效度分析充分解释。本文所分析的“认知守护”实践的可能证伪条件则在于：如果随着这些实践的扩散，教师群体没有生长出足够的辨认性专业判断力——而只是用新的技术培训学生“如何表演独立性”——导致被守护空间重蹈了它本欲抵抗的负值转译的覆辙，且这一收编在至少一个完整的学段（三到六年）内不可逆转，那么本文对这些实践作为“守护”的定性就是错的：它们不过是新一套负值转译的起点。学术论说的分量，不在它给出的答案有多安全，而在它敢于在什么条件下承认自己错了。在这个意义上，这篇论文自身的论证姿态，也是它所试图守护的那种认知品质的一次尝试。

注释

[1] 本文开篇的写作测试争议案例，原型取自Daniel Koretz《Measuring Up: What Educational Testing Really Tells Us》及《The Testing Charade: Pretending to Make Schools Better》中所记述的多起真实评分争议的典型特征综合，并加以例示性简化，用于呈现评分量规条款封闭性在具体操作中的效应。文中所引评分者陈述与仲裁过程依据这一类争议的共同结构进行典型化重述，不代表任何真实个案的逐字记录。本案例同样用于说明：即使在表现性评价（写作）中，评分信度要求与量规的有限分辨率仍可导致高质量的认知投入被转译为次等。此亦为后文与表现性评价倡导者对话的经验基础。

[2] Campbell, D. T. (1976). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67–90. 本文所引述的Campbell定律经典表述，在该论文中有核心阐述。

[3] Nichols, S. L., & Berliner, D. C. (2007). Collateral Damage: How High-Stakes Testing Corrupts America's Schools. Harvard Education Press; Koretz, D. (2017). The Testing Charade: Pretending to Make Schools Better. University of Chicago Press. 本文对Campbell定律在教育中效应的大量实证总结，受益于这两部著作。

[4] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). American Council on Education / Macmillan. 本文对Messick效度理论的引用限于其“构念代表性不足”与“构念无关变异”的核心区分。

[5] Darling-Hammond, L., & Adamson, F. (2014). Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning. Jossey-Bass. 这是表现性评价文献的一个重要代表，书中展示了若干高信度表现性评价的例子。

[6] Lipsky, M. (1980). Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. Russell Sage Foundation. 本文对阅卷者及教师作为街头官僚的分析，参考了Lipsky的基本框架。

[7] 浙江省高考综合改革试点的相关信息，来自《浙江省深化高校考试招生制度综合改革试点方案》（浙政发〔2014〕37号）及国内外媒体的公开报道（如《中国青年报》《财新》等）。文中对此案例的讨论，不对改革整体做评价，只以其作为“评价形式改革被旧媒介逻辑重新吸收”的经验现象之一例。同时，更多关于中国高考评价改革的实证分析，可参见如Kipnis, A. B. (2011). Governing Educational Desire: Culture, Politics, and Schooling in China. University of Chicago Press; 及Zhao, Y. (2018). What Works May Hurt: Side Effects in Education. Teachers College Press.

[8] 强基计划的全称为“基础学科招生改革试点”，根据教育部《关于在部分高校开展基础学科招生改革试点工作的意见》（教学〔2020〕1号）等公开政策文件。本文仅在表明官方已试图为不可分数化的能力留出制度空间的层面引用，不对其效果作出评价。

参考文献

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. American Educational Research Association.

Bourdieu, P., & Passeron, J.-C. (1990). Reproduction in Education, Society and Culture (2nd ed.). Sage. (原著出版于1970年)

Broadfoot, P. (1996). Education, Assessment and Society. Open University Press.

Campbell, D. T. (1976). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67–90.

Darling-Hammond, L., & Adamson, F. (2014). Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning. Jossey-Bass.

Foucault, M. (1995). Discipline and Punish: The Birth of the Prison (A. Sheridan, Trans.). Vintage Books. (原著出版于1975年)

Gipps, C. (1994). Beyond Testing: Towards a Theory of Educational Assessment. Falmer Press.

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In Papers in Monetary Economics (Vol. 1). Reserve Bank of Australia.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17–64). Praeger Publishers.

Kipnis, A. B. (2011). Governing Educational Desire: Culture, Politics, and Schooling in China. University of Chicago Press.

Koretz, D. (2017). The Testing Charade: Pretending to Make Schools Better. University of Chicago Press.

Lipsky, M. (1980). Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. Russell Sage Foundation.

Merton, R. K. (1968). Social Theory and Social Structure (Enlarged ed.). Free Press. (原著出版于1949年)

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). American Council on Education / Macmillan.

Nichols, S. L., & Berliner, D. C. (2007). Collateral Damage: How High-Stakes Testing Corrupts America's Schools. Harvard Education Press.

Ross, H., & Wang, Y. (2010). The College Entrance Examination in China: An overview of its social-cultural foundations, existing problems, and consequences. International Journal of Educational Development, 30(1), 3–12.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144.

Zhao, Y. (2018). What Works May Hurt: Side Effects in Education. Teachers College Press.

浙江省人民政府. (2014).《浙江省深化高校考试招生制度综合改革试点方案》（浙政发〔2014〕37号）.

教育部. (2020).《关于在部分高校开展基础学科招生改革试点工作的意见》（教学〔2020〕1号）.

（注：参考文献包含经典著作、当代研究及政策文件。部分中国政策文件以官方文本为据。）

深化增补：被放跑的最近邻、操作化与证伪

以下四节为编辑增补，与作者正文分开标注。

增补一：与反应性、指标反身效应研究的硬边界

埃斯佩兰与索德的反应性研究已系统论证：排名与指标不只是测量对象，它们会改变被测量者的行为与自我理解，并给出自我实现与叙事重写两条机制。这是本文最近的对手，正文对勘了福柯、默顿、布迪厄与坎贝尔、古德哈特、梅西克，却未处理它。分野在改变发生的位置：反应性研究的改变发生在被评价者一侧——他们为迎合指标而调整行为；本文的转译发生在媒介内部——一个未调整行为、坚持独立判断的学生，其品质在评分细则的正常运算中被赋以负值。前者是行为适应，后者是赋值结果。可判别面：若所有被系统性低评的认知品质都能由学生的行为选择解释（他们选择了不迎合），反应性足够；若在行为完全符合要求的答卷上仍出现该品质被扣分，本文成立——这正是正文那份“被公正地错评的答卷”所展示的。

增补二：负值转译的可编码形式

机制若不可编码，结构性功能的主张就无法与个案疏忽区分。两项指标：细则容纳度——一份答卷中的每个判断单元，能否被评分细则的任一条目直接容纳，分四级编码（有专门条目／可归入相近条目／只能归入兜底条目／无可归入）；赋值方向——不可容纳单元在最终得分中的实际影响方向。关键预测是：不可容纳单元的影响方向系统性为负而非零。若为零（即被忽略而非被扣分），本文的核心断言从“转译为负”降级为“未被看见”，这是一个显著更弱的命题。这一编码把本文与“被忽视”论彻底分开。

增补三：认知守护区的制度最小形式

正文提出建立一块负值转译在此失效的认知守护区，但未给最小可嵌形式，易被读作理想主张。最小形式是：在既有评价中划出一个不计入总分、但必须被独立记录并随成绩单一并呈交的栏目，专门登记那些评分细则无法容纳的判断。它不改变分数、不设权重、不新增科目，因此不触动选拔的零和结构；它唯一的功能是让被转译为负的东西留下一条不被赋值的痕迹。这一形式也自带

失败判据：若该栏目在数年内被赋予权重或被纳入考核，守护区即已被收编，正文第七节所预告的递归负值转译便已发生。

增补四：与作者已发表工作的分野

作者已发表《认知偏馈：论现代学校如何将“不对路”铸成“不够格”》，与本文最近，几乎处理同一判断，必须切开。那篇的机制是铸造——把一种不同的认知路径重新描述为一种资格不足，其操作面是分类与命名；本文的机制是转译——把一种认知品质在有限分辨率的媒介内换算为可精确赋值的负向指标，其操作面是评分运算。分野可一句话说清：前者改变的是这是什么，后者改变的是这值多少。另需标注《合法性的牢笼：学校认证如何将不可见的学习逐出课堂》，它处理的是逐出，本篇处理的是被留下但被赋负值——留下而被扣分与被逐出，是两种不同的命运。

编辑增补所依据的核验文献

Wendy Nelson Espeland & Michael Sauder, “Rankings and Reactivity: How Public Measures Recreate Social Worlds,” American Journal of Sociology, 113(1), 2007.

Donald T. Campbell, “Assessing the Impact of Planned Social Change,” Evaluation and Program Planning, 2(1), 1979.

Samuel Messick, “Validity of Psychological Assessment,” American Psychologist, 50(9), 1995.

Michael Power, The Audit Society: Rituals of Verification, Oxford University Press, 1997.

Pierre Bourdieu & Jean-Claude Passeron, La reproduction, Éditions de Minuit, 1970.