

学科通融 · 之一

能被量的那一面

为什么一件事最容易被看见的地方，往往正好不是它长东西的地方

教育学 × 再生医学 × 行为公共政策

作者 王德生 + Claude

栏目 学科通融 · 站外碰撞

篇幅 约2.4万字 · 28页

发表 2026年7月27日

摘要

三拨互不相识的人正在各自的领域里吵架：一拨吵学东西该让人费劲还是该让人轻松，一拨吵治病该修零件还是该改目标，一拨吵想让人过得好该改人还是该改规则。三场架给出三个互相排斥的答案，而三边都各自发现了同一件事却都以为那是自己领域的病——你能测到的那个东西，跟你真正想要的那个东西，方向常常是反的。本文追出这件事怎么一步步接管一个系统的治理方式，给出四道工序、十个行当里的真两难，以及四条可以让自己失败的判据。

这一篇由哪三个学科的理论体系撞成

第一家说要在单元这一层动手而且要加阻力；第二家说别在单元这一层动手、单元自己有目标；第三家说在这一层讨论本身就是被推动的分心。三家两两冲突，各自的证据都硬，各自的批评都打在对方软处。三家均为站外的公开文献，链接直达原始出处，可自行核对。

教育学 · 学习与记忆

[可取用困难框架与失用新论 \(Bjork & Bjork\)，对阵认知负荷理论 \(Sweller 一系\)](#)

难度不是要克服的成本，难度就是那个起作用的动作本身；而人对「我学会了没有」的判断系统性地是反的。两家的正面比较见 Pyke, Lunau, Javadi 2024, 《Quarterly Journal of Experimental Psychology》。

<https://pubmed.ncbi.nlm.nih.gov/39641213/>

医学 · 再生与形态发生

[集体智能与生物电界面 \(Michael Levin\)，《身体的多尺度智慧》2025](#)

当前生物医学几乎全是自下而上瞄准分子硬件；细胞群是有目标的求解者，杠杆在改写共享的目标状态，而接口的价值不在于它最根本，在于它可读可写。

<https://onlinelibrary.wiley.com/doi/10.1002/bies.202400196>

商业经济与公共政策 · 行为科学

[个体框架与系统框架之辩 \(Chater & Loewenstein\)，《Behavioral and Brain Sciences》](#)

个体层干预的效果一直低于预期；把问题定义在个体层这个动作本身是被利益方推动的；而随机对照试验的要求，大幅限制了能被纳入考虑的政策范围。Cass Sunstein 对最后一条有力反驳。

<https://www.ssrn.com/abstract=4046264>

目 录

为什么一件事最容易被看见的地方，往往正好不是它长东西的地方4

一、三个吵得正凶的地方4

二、这三家其实在互相打架6

三、把那条重合的线拽出来8

四、为什么能被量的那一面，正好不是长东西的那一面9

五、它是怎么接管的：四道工序10

六、"上一层"这三个字为什么会骗人13

七、为什么这几年突然加速14

八、十个地方的同一件事15

九、这条判断最容易被拿去干坏事的地方19

十、那怎么办19

十一、什么情况下这篇文章是错的21

十二、出处与最近的邻居22

十三、这一篇是怎么撞出来的26

十四、这篇文章的自我审判27

为什么一件事最容易被看见的地方，往往正好不是它长东西的地方

有三拨人正在各自的领域里吵架。他们互不认识，读的不是同一批论文，开的不是同一个会，甚至不知道对方存在。

一拨在研究人怎么学东西，一拨在研究细胞怎么长成身体，一拨在研究政策怎么才能真管用。

这三场架吵的不是同一件事。但如果你把三场架的记录摊在一张桌子上看，会看到一个古怪的重合：**三边都在为同一个问题争，而且都没意识到自己在争这个问题。**

那个问题是：**这件事该在哪一层动手？**

而更古怪的是第二层重合：三边都发现了对方也发现了的那件事——**你能测到的那个东西，跟你真正想要的那个东西，方向常常是反的。** 但三边都只在自己的地盘里发现了这件事，没有一边知道这是三个领域共有的。

这篇文章想把那条重合的线拽出来。

一、三个吵得正凶的地方

第一个：学东西的时候，该让人费劲还是该让人轻松

这场架已经吵了三十年，到今天还没完。

一边说：**要让人费劲。**

这一派的核心发现有点反常识。他们把「记住」拆成两个完全不同的量：一个是这东西在你脑子里扎得多深，另一个是你此刻能不能想起来。这两个量可以完全脱钩——一样东西可以扎得很深，但此刻死活想不起来；也可以此刻张口就来，但其实根本没扎进去。

关键在这儿：**当一样东西你已经快想不起来的时候，你费劲把它想起来，那一下带来的扎深程度是最大的。** 想得越轻松，扎得越浅。

所以这一派的建议全都是反直觉的：把复习拉开间隔（让你先忘掉一点）、把不同类型的题混着做（让你没法用惯性）、少看答案多自己回想（让你先卡住）。这些做法在当下看全都是效率更低的，考试前一天你会觉得自己学得更差了。但过三个月再测，差距很大。

这一派还有一个更狠的说法：**你自己对"我学会了没有"的判断，系统性地是反的。** 因为大脑用「读起来顺不顺」当作「我会不会」的判据，而这个判据恰好是错的。所以那些让你感觉良好的学习方法——反复读、划重点、临时抱佛脚——全都是在优化一个错的东西。

另一边说：**要让人轻松。**

这一派的道理同样硬：人的工作记忆容量极小，一次能同时转的东西就那么几个。如果一份材料本身就复杂——要素之间互相牵扯，不能拆开来一个个学——那么任何额外的困难都会把人直接压垮，什么也学不到。他们做过实验：在高度复杂的材料上，那些"有益的困难"不但无益，反而有害。

这场架到 2024 年底还在正式发论文比较双方，2026 年还在出新研究。两边都承认对方有证据，也都承认自己的适用范围有边界，但谁也没能把对方收编进自己的框架里。

第二个：治病该修零件还是该改目标

这场架比上一场安静，但更根本。

主流做法是修零件：找到出问题的那个分子、那个基因、那个通路，然后想办法把它修好或者关掉。整个现代生物医学基本上就是这么运转的，而且成果巨大——没有这条路，我们今天不会有那些真正救命的药。

但有一拨人说：**这条路选错了层级。**

他们的看法是这样的：一堆细胞长成一只手，这件事不是分子硬件自动展开的结果。这是一群细胞在**解决一个问题**——它们知道目标形状是什么，如果你在中途破坏它，它们会换一条路继续朝那个目标走。你把一只蝌蚪的眼睛挪到尾巴上，它照样能看见；你切掉一段，它长回来的还是原来那个形状，不是随便乱长。

这不是比喻。这一派的说法是：那群细胞在维护一个**共享的目标状态**，而这个目标状态是可以被读出来、也可以被改写的——不是通过修改它们的基因，而是通过改变它们之间的电信号模式。

所以他们的主张是：与其自下而上地修一万个零件，不如自上而下地跟那个集体"说一句话"，改掉它的目标。他们特别强调，电信号之所以值钱，**不是因为它是最根本的那一层，而是因为它是一个能读能写的接口**——你可以用它跟那个集体沟通，而不必知道那个集体内部到底是怎么算的。

这一派 2025、2026 年还在密集出论文，而且已经不只是理论——他们真的在动物身上改写过目标形状。

主流那边当然不服。修零件这条路有几十年的实绩，而"细胞群有目标"这种说法在很多人听来太玄。这场架的实质是：**生命体到底更像一台机器，还是更像一群会解决问题的东西？**

第三个：想让人过得好，该改人还是该改规则

这场架最热闹，因为它直接关系到钱和政策。

有一整套做法，过去二十年席卷了全世界的政府和企业：既然人的判断有系统性的偏差——短视、拖延、被默认选项牵着走——那就把选择的呈现方式改一改，用很小的成本引导人做出更好的选择。自动加入退休金计划、把健康食品摆在货架显眼处、在缴税通知上加一句"你的邻居大多已经交了"。

这一路的两位重要推动者，几年前公开反水了。

他们说：我们错了。这类干预的实际效果一直令人失望地平庸——不是没用，是远小于宣传。

但更狠的是第二句：**把问题定义在个人身上这个动作本身，是有人在推的。**

他们的论证是：在一个又一个政策领域里，既得利益方在公开场合大力支持"改变个人行为"这一路，同时在私下游说反对系统层面的改革。道理很简单——如果一个群体的目的是搞垮系统性改革，那它最有效的办法就是推广"真正的问题出在个人身上"这种看法。含糖饮料公司愿意资助"鼓励大家多运动"的公益广告，因为那笔钱买断的是糖税。

第三句更狠：这类干预不只是效果小，它**占着位置**——它让系统改革显得不那么必要，从而使系统改革更难发生。

这一句遭到了最强烈的反驳。反驳者说：没有任何证据表明关注个人层面会让系统层面的努力变少；而且个人层面的干预本身确实有实实在在的正面效果。这场架到 2026 年还在打。

二、这三家其实在互相打架

三场架，三个领域，看起来各管各的。

但如果你问它们同一个问题——**要让一样东西变好，你该在哪一层动手**——它们会给出三个互相排斥的答案。

第一家说：在单元这一层动手，而且要加阻力。

学习这件事，全部杠杆都在学习者身上。间隔、交错、自己回想，每一样都是施加在个体身上的操作。而且方向是加阻力——让他更费劲，他才长得牢。

第二家说：别在单元这一层动手，单元自己有目标。

如果第二家对，那第一家就是在犯一个层级错误：你以为你在给一块材料加工，其实你面对的是一个自己有主意的东西。对一个有目标的系统施加零件层的力，它会绕过去、会补偿、会在你不看的地方恢复原样。真正省力的做法是改它的目标，而不是加它的阻力。

第三家说：这两家争的东西本身就是被设计出来分你心的。

如果第三家对，那前两家都在同一个陷阱里：**只要你还在讨论"怎么对这个单元动手"，你就已经输了。**你该问的是谁在决定这个单元处在什么条件下——而这个问题的答案不在单元里，在规则里。而且第三家会

指出一件让人不舒服的事："让学习者用更好的方法"这个框架，谁受益？学校可以不改课程、不改班额、不改考试，只要求学生学得更聪明一点。这不就是教育版的"这是你自己的问题"吗？

三家两两冲突。

第一家对第二家：你说单元有目标，那我给它加阻力它就该反抗——可事实是加了阻力它确实学得更牢了。第二家对第一家：它学得更牢，不等于你的加阻力起了作用；也可能是加阻力恰好改变了它对"我要成为什么"的判断，而你以为是努力起了作用。第三家对前两家：你们俩争在哪一层加力，而真正决定结果的是谁有权设定这个单元的处境——这一层你们都没碰。第一家对第三家：你说个体层没杠杆，可我们手上有几十年的实验数据，间隔和交错的效果是被反复重复出来的。第二家对第三家：你说要改规则，可规则改完了，那个单元内部的目标还是没变——你只是换了它周围的墙。

这是个真三角。没有哪两家能合并。

得先替三家各自说句公道话

上面把三家写成互相打架，容易给人一个印象：它们各自都很偏。不是的。三家各自都有最强的自辩，不先把这些自辩摆出来，后面撞出来的东西就不算数——撞倒一个稻草人不叫碰撞。

第一家会说：你说我全部杠杆都在个体身上，可我的证据是几十年、几百项、跨年龄跨学科重复出来的。间隔的效果、交错的效果、自己回想的效果，重复率高得罕见。你可以说我层级选得不对，但你不能说我没效果——**在一个到处都是重复不出来的结果的领域里，我这一族是少数扛住了重复检验的。**

这条自辩很有力。而且它反过来构成了对第三家的质疑：如果个体层真的没有杠杆，为什么个体层是唯一一个能稳定重复出效果的层级？

第二家会说：你说我玄，可我不是在讲哲学，我是在动物身上真的改写过目标形状。我不需要知道那群细胞内部怎么算，我只需要一个能读能写的接口——这跟你不需要懂芯片就能用键盘是一个道理。**而且恰恰因为主流那条路是自下而上的，它才在最难的几个病上卡了几十年。**

这条自辩也很有力。而且它反过来质疑第一家：你的间隔和交错之所以有效，你确定是因为"努力"吗？还是因为它们碰巧改变了那个学习者对自己的判断？你测的是行为，你解释的是原因，中间那一步你没验过。

第三家会说：你们俩都在说自己有效果，可效果的大小呢？我承认个体层干预有效，我只是说效果远小于宣传，而且**代价是整整二十年的政策注意力**。这二十年里，含糖饮料的税没收上，住房没建起来，退休金制度没改——而我们发了几百万条短信。

这条自辩最难反驳，因为它不比效果的有无，比的是**机会成本**。而机会成本是没法做对照实验的。

三条自辩摆出来之后，冲突反而更清楚了：

三家各自都对，各自的证据都硬，各自的批评都打在对方的软处。它们之所以撞不出结果，不是因为哪一家更蠢，而是因为它们在争的那个问题——该在哪一层动手——本身可能不是一个可以由证据裁决的问题。

而这一点，恰恰是下一节要说的东西的入口。

三、把那条重合的线拽出来

但三家有一样东西是共有的，而且三家都只在自己的地盘里发现了它。

第一家发现：在学习里，你唯一能当场测到的东西——学习者此刻表现得多好——跟你真正想要的东西——三个月后他还记不记得——是**负相关**的。当下表现越好，往往说明扎得越浅。

第二家发现：在医学里，你能测到的东西——分子指标、基因表达、蛋白浓度——跟你真正想要的东西——这堆细胞最后长成什么形状——**不在同一层**。你把所有分子测得清清楚楚，也读不出那个目标形状是什么。

第三家发现：在政策里，这件事最赤裸。他们指出：要求一项政策干预必须通过随机对照试验来验证，这个要求本身就**大幅限制了能够被纳入考虑的政策范围**。

请把这第三句读两遍。

它说的不是“我们测错了”。它说的是：**因为系统层面的改革做不了随机对照试验——你没法把一个国家随机分成两半，一半改税制一半不改——所以系统层面的改革就自动从“可以被认真考虑的选项”里掉出去了**。不是有人否决了它，是它连进入讨论的资格都没有。

而个体层面的干预可以做随机对照试验：发一千条短信，五百人发五百人不发，两周后看数据。做得漂亮，发得出去，报得上成绩。

于是二十年下来，全世界的政策工具箱里塞满了短信、默认选项和提示语，而税制、住房、劳动法这些真正的大件，在“证据为本”的名义下被系统性地挪出了视野。

把三家的发现摆在一起，那条线就出来了：

> 一个系统里能被读出来的那一面，会自动接管它被治理的方式；而那一面通常不是它长东西的那一面，甚至跟长东西的那一面方向相反。

我把这件事叫作**能被量的那一面接管了**。

四、为什么能被量的那一面，正好不是长东西的那一面

上面那句话如果只是一个观察，那它是软的。得说清楚它凭什么。

这里有一个很朴素的道理，说穿了不难，但很少有人正面说：

一样东西能被从外面测到，前提是它得在外面留下痕迹。而很多要紧的过程，恰恰是在不留痕迹的时候才在进行。

分三层讲。

第一层：正在长的东西，通常还没有形状

一颗种子在土里发芽的头几天，你把它挖出来看，它什么也不是——一点白根，一点软芽。你能测的是重量、长度、颜色，而这几个量在这个阶段全都难看。

真正在发生的是根系在决定往哪个方向扎、细胞在决定分化成什么。这些事没有一样是能从外面量出来的，因为它们还没有变成形状。

等它们变成能被量的形状时，决定已经做完了。

学习也是这样。一个学生费劲想一道题、想了二十分钟没想出来——这二十分钟里发生的事，是他脑子里那些半成品的连接在互相试探。你在第二十一分钟给他一张卷子，测出来的是零分。而如果你给另一个学生直接看答案再考，测出来是满分。

两个数字，一个零一个满分。而实际情况正好反过来。

第二层：能被量的东西，会被优化；而优化会趋向那个量本身

这一层大家比较熟，但通常只被说了一半。

通常的说法是：一个指标一旦成了目标，人们就会针对指标做优化，指标就不再准了。这个说法是对的，但它说的是**人开始作弊之后**的事。

还有更早的一层：**在任何人作弊之前，仅仅因为某个东西可测，做事的人就会自动往那个方向靠。**

一个真诚的老师，没有任何作弊的意思，只是每周要交一份教学进度表。表上有"讲授课时""练习题量""单元测验平均分"三栏。他会怎么安排下周的课？

他不会造假。他只会——完全出于好意——把那些能填进表里的活动排得多一点。而那个卡住的学生，那个需要他课后多留二十分钟、用一个不在任何教案里的说法讲一遍的学生，那二十分钟在表上是零。

他没有做错任何事。表也没有做错任何事。但那二十分钟被挤掉了。

第三层：不可测的那一面，没有代言人

这是最要命的一层，也是最少被说的。

一件事要进入决策，得有人替它说话。而替它说话的人得能拿出点什么。

可测的那一面天生有代言人：数据、图表、对照组、显著性。不可测的那一面没有——**替它说话的人只能说"我觉得这个很重要"**，而这句话在会议室里等于什么也没说。

所以在任何一次资源分配、任何一次改革讨论、任何一次预算审批里，可测的那一面会不断赢，而不可测的那一面会不断输。不是因为决策者蠢，恰恰是因为决策者负责任——一个负责任的决策者当然应该要证据。

结果是：不可测的那一面不是被打败的，是被弃权判负的。它甚至没有上场。

这三层叠在一起，就得出了这篇文章的核心判断：可测量性不是一个中性的工具。**它是一个会自己选层级的东西。**它决定了在哪一层动手，而这个决定是在任何人开始思考"该在哪一层动手"之前就已经做完了的。

五、它是怎么接管的：四道工序

抽象说完了。下面是具体怎么发生的。四道工序，按顺序，每一道单独看都完全正当。

工序一：把要的东西换成能量的东西

第一步不是造假，是**替换**。

你真正想要的是"学生三年后还能用这个"。这个东西没法测——三年后学生已经毕业了，联系不上,而且就算联系上了,你也分不清是你教的还是他后来自学的。

所以你换一个：单元测验成绩。这个能测，当周就能测。

替换的时候，所有人都知道这是个替代品。会议上会有人说"当然，考试成绩不完全代表学习效果"。这句话会被记进纪要，然后被忘掉。

关键在于：从这一刻起,那个真正想要的东西就再也没有出现在任何一份文件里了。它活在人们的口头共识里，而口头共识扛不过三次人事变动。

医学里的替换是：你真正想要的是"这个人活得好且久"，你换成"这个指标降到多少"。政策里的替换是：你真正想要的是"这个社会里的人过得好些"，你换成"这项干预的效应量"。

工序二：让替代品变成入场券

第二步，替代品从"我们用它来看看情况"变成"没有它就不能上场"。

这一步通常以质量提升的名义完成，而且理由充分得没法反驳：我们不能凭感觉分配公共资源，我们需要证据。

于是任何一项提议都必须先回答"你的效果怎么测"。答不上来的提议不是被否决，是**根本不会被提出来**——提出者自己就先算了，因为他知道过不了那一关。

政策那一家把这件事说得最准：随机对照试验的要求，大幅限制了能被考虑的政策范围。

请注意这句话的形状。它不是说"随机对照试验测错了"，它是说：**这道门的形状，决定了什么东西能走进这个房间**。而系统层面的改革，形状不对，进不来。

工序三：做事的人开始朝入场券的方向长

第三步是最安静的一步，因为它不是决策，是适应。

一个人在一个系统里待五年，他会逐渐学会这个系统认什么。不是因为他投机，是因为他做的事只有被认了才算数，不被认的部分做多了，他自己也撑不住。

于是他会变。不是变坏，是变得更擅长生产那种能被认的东西。

这一步完成之后，你会得到一批**真诚的、能干的、且完全朝着替代品长的人**。他们不会承认自己在追指标，因为他们真的不是——他们只是学会了怎么把事做得"能交代"。而"能交代"和"真做成了"这两件事，在他们的职业生涯里从来没有分开过，所以他们也没有理由怀疑这两者会不一样。

工序四：不可测的那一面被重新解释为不存在

最后一步，闭环。

当所有人都朝替代品长了足够久之后，那个原来真正想要的东西，会被重新解释成一个模糊的、不专业的、说不清楚的东西。

"你说的那种'真正的理解'，具体指什么？能不能操作化一下？"

这个问题听起来完全合理，而且提问者是好意的。但它的实际效果是：**凡是不能被操作化的东西，都被判定为不存在**。

到这一步，系统就闭合了。它不再需要任何人推动，因为它内部的每一个人都在真诚地做正确的事。而那个原来要的东西，已经没有任何一个位置可以让它被提起了——提起它的人会显得不够专业。

四道工序走完，**这个系统现在极其擅长生产一样东西，而那样东西不是它当初被建立起来要的那个**。

每一道工序的现场

四道工序说完了，但抽象地说完不够。下面各给一个具体现场，都是能对上号的那种。

第一道的现场：一次会议纪要。

某单位开会讨论新的评价办法。有人提出："我们真正想提升的是服务对象的实际获得感，但这个不好测，先用满意度问卷代替，后面再完善。"

会议纪要上会怎么记？大概率是："经讨论，确定以满意度问卷得分作为核心指标。"

那句"先用……代替，后面再完善"不会进纪要。 不是有人故意删的，是因为纪要要记的是决议，而那句话不是决议，是一个附带的保留。

三年后，新来的负责人翻开制度文件，看到的是"核心指标：满意度问卷得分"。他不知道这是个替代品。对他而言，这就是目标本身。

第二道的现场：一句"你这个怎么评估"。

某人提了一个方案，方案的价值需要五年才看得出来。评审会上有人问："你这个效果怎么评估？"

这个问题是完全正当的、专业的、负责任的。提问的人没有任何恶意，他只是在做他该做的事。

但提方案的人听到这句话时，心里有一个瞬间的计算：我可以老实说"五年后才知道"，然后这个方案就没了；或者我可以编一个可以在半年内看到的中间指标。

绝大多数人会选第二个。而选了第二个之后，他会开始为了那个中间指标而调整方案本身——**不是为了骗人，是因为方案现在要对那个指标负责了。**

这一步的关键是：**筛选发生在提案之前。** 那些真正不适合被测的方案，多数根本没有被提出来过，所以它们在任何统计里都不存在——包括在"我们否决了多少提案"这个统计里。

第三道的现场：一个人五年后的样子。

一个刚入行的人，第一年会为那些不计分的事花很多时间：多解释一遍、多陪一会儿、多查一个来源。他做这些不是因为算过账，是因为他觉得该这么做。

第二年他发现，做这些的人和不做这些的人，在年度评价上没有区别。 第三年他发现，不做这些的人反而更从容——他们有更多时间产出那些能被计分的东西。 第四年他开始下意识地判断哪些事"值得做"。 第五年，他已经不再产生"要不要多做一步"这个念头了——那个念头不是被压下去的，是不再冒出来了。

这一步最关键的性质是：他没有变坏。 你如果去问他，他会真诚地告诉你他在认真工作，而且他说的是实话。他只是不再看见那些不计分的地方了。

第四道的现场：一次"能不能操作化一下"。

到这一步，某个还记得原来要什么的人在会上提了一句："我们是不是把最要紧的那部分弄丢了？"

会上有人问："你说的'最要紧的那部分'，具体指什么？能不能给个可以操作的定义？"

这个问题一出，讨论就结束了。因为那个人给不出。他要说的东西恰恰是那种给不出操作定义的东西——如果给得出，它早就被纳入指标了。

于是他的发言被记作"提出了一些方向性的意见", 然后不了了之。而下一次, 他就不提了。

四道工序里, 只有这一道是有人在场的。前三道都是无声的、自动的、没有责任人的。只有第四道, 有一个具体的人被具体地问住了。而恰恰是这一道, 最少被写进任何复盘。

六、"上一层"这三个字为什么会骗人

前面那三家有一个共同点: 它们都说别在零件层动手, 要上一层。

但它们说的"上一层"指的是三个完全不同的方向, 而且这三个方向互相排斥。这一节要把这件事讲清楚, 因为它是这篇文章里最容易被误用的部分。

第一家的上一层: 上到策略。

从"多做几道题"上到"用什么方式做题"。层级确实上了一格, 但仍然在学习者身上——它改的是这个人怎么做, 不是这个人身处何地。

第二家的上一层: 上到单元内部的目标。

从"修哪个分子"上到"这群细胞认为自己该长成什么样"。这一层比第一家更靠里——它进到了单元内部, 去改那个单元自己持有的目标。

第三家的上一层: 上到单元外部的规则。

从"怎么让这个人少喝含糖饮料"上到"糖该不该收税、饮料该怎么定价、贫困社区有没有别的选择"。这一层完全在单元外部——它根本不关心这个人内部持有什么目标。

三个方向: 一个更靠外一点但还在人身上, 一个进到里面去, 一个完全出去。

所以当有人说"别在细节上打转, 要从更高层面解决"的时候, 这句话在没有指明方向之前, 等于什么也没说。它可以是三件截然不同的事:

- 改这个人的做法 (还在原地, 只是换了个做法)
- 改这个人的目标 (进到里面去)
- 改这个人的处境 (完全出去)

而这三件事的成本、效果、以及谁有权做, 全都不一样。

为什么"上一层"这么容易被含糊地用?

因为可测的永远是最下面那一层。往上任何一层都不好测, 于是"上一层"这三个字自然地变成了一个**没有人需要为它交出证据的位置**。谁都可以说要从更高层面解决, 因为没人能测你到底解决没解决。

这就是本文那条判断的一个副产品：**可测的那一面接管治理之后，不可测的那些层就退化成了修辞。** 它们不是被消灭了，是被降格成了会议上的漂亮话——说了不用负责，因为反正测不了。

七、为什么这几年突然加速

前面讲的那四道工序，本身不是新东西。人类有账本的时候就有它了。

要命的是**速度**。

从前走完这四道工序需要几十年。一个行业先立起一套指标，从业者用二三十年慢慢朝它长，等到那个原来要的东西彻底说不出口了，往往已经过了一代人。慢有慢的好处：慢的时候，总有几个还记得原来要什么的老人在，他们的存在本身就是一份没写下来的备份。

现在不一样了，而且是三件事叠在一起造成的。

第一件：测量变便宜了，而且是断崖式地便宜。

从前测一件事要花钱：发问卷、雇人观察、做抽样。贵，所以你得先想清楚测什么。这个"先想清楚"本身是一道闸门——因为贵，所以只有真正重要的东西才值得测。

现在测量的边际成本接近于零。一个人在网页上停了多久、鼠标划到哪儿停住、哪一段视频被拖过去了，这些数据是顺带产生的，不测反而要多写代码。

于是那道闸门没了。**从前是"这件事重要所以我们测它"，现在是"这件事能测所以它变得重要"。** 因果方向反了，而反过来的这一天，没有任何人开会决定过。

第二件：反馈变快了。

四道工序里的第三道——做事的人朝指标长——从前需要几年，因为反馈周期是年度考核。

现在反馈是实时的。一个做内容的人今天发一条，两小时后就知道数据好不好。**在这种速度下，"朝指标长"不再是一个职业生涯尺度的缓慢漂移，而是一个以天计的调整过程。** 一个人可以在三个月里被训练成完全不同的样子，而且他自己感觉不到，因为每一次调整都很小、都很合理、都有即时的正反馈。

第三件，也是最要紧的：现在做优化的不是人了。

前两件事还只是把人类的老毛病加速。第三件是质变。

一套自动优化的东西，它没有那个"原来我们要的是什么"的记忆。它不会在某个深夜突然想起初衷，不会因为良心不安而在报表上手下留情，也不会有一个快退休的老人在会上说一句"我们当年不是这么干的"。

它只有目标函数。而目标函数就是那个替代品。

这意味着第一道工序——把要的东西换成能量的东西——在这里是**不可撤销的**。人做替换，替换之后原来那个东西还活在人的记忆里，虽然弱，但还在，有机会被重新提起。机器做替换，替换之后原来那个东西**在系统里不留任何痕迹**——它从来没有被写进目标函数，所以对这套系统而言，它从来不存在。

三件事叠起来的后果是：**从前需要一代人才能走完的四道工序，现在可以在一个产品周期内跑完。**

而且有一个新的、从前没有的性质：**跑完之后没有人能说出丢了什么。** 从前至少还有几个老人记得。现在那批本该成为老人的人，是在这套已经跑完的系统里长起来的——他们从来没有见过替换之前的样子。

这一节不是在说技术不好。测量便宜是好事，反馈快是好事，自动优化解决了大量人做不好的事。这一节只说一件事：**这三样东西同时把那四道工序的转速提高了两个数量级，而没有任何一样东西同时提高了"记住我们原来要什么"的能力。**

那个能力至今仍然只能靠人，靠记忆，靠一个愿意在会上说"等一下"的具体的人。而这个东西，恰好是不可测的。

八、十个地方的同一件事

下面把这条判断放到八个行当里去试。每一个我都写成**真两难**，不是写成"所以别测量"。想不测量的人没看懂这篇文章。

不测量的后果远比测量坏得多——那是靠关系、靠辈分、靠谁嗓门大来分配资源的世界，而在那个世界里吃亏的永远是没背景的人。这一点必须先说死。

这条判断不给你一个更聪明的测法。它只告诉你：每一次你把某样东西装进量表，你都在同时得到公平和失去别的东西；然后由你决定这次要哪一样。

一、学校

可测的是分数、升学率、课时、教案完成度。不可测的是那个学生有没有真的想通、有没有在某一刻决定"这门课跟我有关"。

一所严格按可测量指标管理的学校，会比一所松散的学校公平得多——它不会因为老师喜不喜欢你而给你不同的对待，不会让校长的亲戚占便宜。这是真的进步，而且是穷孩子的进步。

代价是：那个需要老师课后多留二十分钟的学生，会被系统性地漏掉。因为漏掉他，表上看不出来。

得到公平、可比较、不靠人情的分配。失去那些只有在没人计时才会发生的一步。

二、医院

可测的是各项化验值、手术量、平均住院日、并发症率。不可测的是这个病人有没有被当成一个人对待、医生有没有在某个瞬间多想了一层。

指标化管理让医疗质量的下限被大幅抬高了。你今天走进任何一家三甲医院，遇到一个庸医的概率远低于三十年前。这个成就无可否认。

代价是：那些真正难的病人——病情不典型、说不清楚、几个科都不管的——会被系统性地推来推去。因为接住他们，在任何一张表上都是负分。

得到可靠的下限。失去处理例外的能力，而例外恰恰是最需要医疗的那部分人。

三、公共政策

这一条前面已经说过，这里只补一句最要紧的：**这一家自己是最诚实的**。他们是这一路的推动者，然后自己站出来说效果远小于预期。

但他们那条最狠的判断——个体层的干预会挤占系统层的改革——遭到了强烈反驳，而且反驳有力：没有证据表明关注个人会让系统改革变少。

这场架我没有能力裁决，也不该在这里装作能裁决。我只能指出：**双方都同意的那一半，已经足够支撑本文的判断了**。双方都同意个体层干预可以做随机对照试验而系统层改革不能。仅这一点，就足以让工具箱长期偏斜。

得到可问责的公共资金。失去所有形状不适合走进那道门的政策。

四、科研

可测的是论文数、影响因子、引用次数、项目金额。不可测的是这个人是不是真的在解决一个重要的问题。

评价制度的每一次改进都让它更公正：从数篇数改成看影响因子，从看影响因子改成看引用，从看引用改成同行质性评价。每一步都更接近学术的本意。

而每一步，也都是把一张更细的网罩在同一片地上。

这一条有一个特别阴的地方：**一个人可以完全不作弊、完全真诚，而把六年的工作拆成六个可发表的小块每年交一块**。这些小块每一块都合格，每一块都增加了人类知识。但那个需要五年不交东西才能长出来的东西，不会出现。

得到不靠师承和人脉的晋升。失去那些做之前不知道结果的工作。

五、公司

可测的是营收、成本、工时、任务完成率、代码行数、客户满意度。不可测的是有没有人在系统出问题之前就闻出味道不对。

一家把所有事都纳入指标的公司，运转起来比一家靠老员工经验的公司稳定得多，也不怕人员流动——因为知识都在流程里，不在人脑子里。这是巨大的进步。

代价是：那个"这事我本来可以不管但我管了"的动作会消失。写进考核之后它变成本职工作，做了不算功劳；不写进考核它就在表上是零。两条路都通向它消失。

得到不依赖个人的稳定性。失去那层让系统能扛住意外的余量。

六、养孩子

这一条最私人，也最能说明问题。

可测的是分数、名次、几门特长、身高体重。不可测的是这孩子有没有一件他自己真心想做的事。

而这里有一个残酷的算术：**在他十八岁之前，那件他真心想做的事，在任何一张可比较的表上，都不如多考二十分值钱。**因为多考二十分能换到一所具体的学校，而"他真心想做一件事"换不到任何东西——它甚至没有一个可以填进去的栏目。

于是即便是最爱孩子的父母，只要他做的是理性的选择，也会一点一点把那件事挤掉。他不是不爱孩子，他是在一个只承认可测量物的世界里替孩子做最优选择。

得到可以兑换的东西。失去那件不能兑换的东西。

七、慈善与公益

可测的是受助人数量、发放数量、项目执行率、每元成本效益。不可测的是有没有一个人的处境真的因此改变了轨迹。

要求公益组织交出可核查的成效数据，防止了大量的浪费和欺骗，这一条无可辩驳。

代价是：那些最难的对象——需要三年反复接触才可能有一点变化的、失败率极高的——会被系统性地放弃。因为帮一百个容易的人，数据比帮三个难的人好看得多，而且好看很多倍。

得到可核查的公共善款。失去最需要帮助的那一小撮人。

八、算出来的推荐

最后这一条最新，也可能最要紧。

可测的是点击、停留时长、完播率、转化率。不可测的是这个人看完之后是不是变得好一点。

现在几乎所有向人推送内容的系统，都在优化前面那几个量。而这几个量与"这个人变好了没有"之间的关系，**大概率是负的**——最能让人停留的内容，往往不是最让人受益的内容。

这一条与前七条有一个关键区别：前七条里做决定的还是人，而这一条里做决定的是一套自动优化的东西，它的速度和规模远超人。**四道工序在这里可以在几个月内跑完，而不是几十年。**

得到规模、效率和不知疲倦。失去所有那些不能被点击率表达的东西——而这个损失发生得非常快。

九、判案

可测的是结案数、审限内结案率、上诉改判率、同案同判的一致性。不可测的是这个当事人有没有觉得自己被认真听过。

裁判文书上网、同案同判、量刑指南——这些改革让司法的下限被显著抬高，也让"看关系判案"变得困难得多。任何一个曾经在沒有这些约束的年代打过官司的人，都不会想回去。

代价是：那个需要法官多花四十分钟、当着当事人的面把为什么这样判讲一遍的动作，在任何一张表上都是纯损耗——它占用了本可以处理另一个案子的时间，且不产生任何可计量的产出。而对那个当事人来说，那四十分钟可能是整个案子里唯一让他觉得自己不是一个卷宗编号的时刻。

更麻烦的是上诉改判率这个指标。它本意是控制错案，但它在实际运行中会奖励一种非常具体的行为：**判得保守**。一个法官如果总是判在最不容易被改的那个位置上，他的指标会非常好看，而他判的每一个案子在实体上都可能离公正差一点点。这个"一点点"不会出现在任何数据里。

得到不靠关系、可被外部监督的司法。失去那些只有法官自己知道该多走一步的地方。

十、安全与风控

可测的是事故率、违规次数、检查覆盖率、隐患整改闭环率。不可测的是有没有人在事情出问题之前就闻出了味道不对。

这一条有一个特别的性质，值得单独指出：**在安全领域，成功是没有痕迹的。**

一场没有发生的事故，不产生任何记录。一个因为某人多看了一眼、多问了一句而被消灭在苗头里的隐患，在所有报表上等于零——甚至比零还糟，因为那个人为此花的时间会被记作低效。

于是安全管理有一个内在的、无法回避的扭曲：**你能奖励的只有"事故发生之后处理得好"，你没法奖励"事故根本没发生"**。因为后者没有事件，而没有事件就没有记录。

结果是一个所有资深安全员都懂、但很难写进任何制度的现象：**做得最好的那个人，看起来最闲**。他的辖区从来不出事，所以他的工作量在系统里显示为最低。而那个不断在救火的人，数据非常漂亮——响应及时、处置得当、闭环率百分之百。

得到可核查的安全底线。失去那种在事情变成事件之前就把它按住的能力，而这恰恰是安全的全部意义所在。

九、这条判断最容易被拿去干坏事的地方

必须专门写一节，因为这条判断有一个非常方便的滥用形式，我几乎可以肯定有人会这么用。

滥用形式是：**"这个东西没法量，所以你不能拿数据来要求我。"**

领导对下属说：我们做的是创造性工作，不适合量化考核（然后按自己的喜好给分）。 机构对捐款人说：我们做的事效果需要很长时间才显现（然后不给任何交代）。 学者对评审说：真正的学问是不能用论文数衡量的（然后十年不出东西）。 公司对投资人说：我们的价值不体现在当期财报上（然后烧钱）。

这些话全都在借用本文的判断。而它们全都是错的。

判别方法只有一条：这条判断是给决策者用的，不是给被评价者用的。

它的正确用法是：**当你手上握着评价权的时候，知道自己的量表会造成什么。** 它的错误用法是：**当你被评价的时候，用它来拒绝被评价。**

这两者的差别不是姿态问题，是结构问题。评价者说"我知道我的量表会漏掉一些东西，所以我另外留一块不评价的地方"——这是在放弃自己的权力。被评价者说"你的量表漏掉了我"——这是在争取自己的利益。前者要付代价，后者不用。

凡是用这条判断替自己免责的，一律不算数。 因为这条判断的全部重量，就在于它要求握权的那一方主动放弃一部分控制,而这件事从来没有人替他做。

还有第二种滥用，更隐蔽：**用它来反对一切测量。**

这一种听起来比第一种高尚，实际后果更坏。因为不测量的世界不是一个温情的世界，是一个凭关系的世界。在那个世界里，最先被牺牲的永远是那些没有关系可凭的人。

本文的判断不是"别测量"。是"知道自己在换什么"。

十、那怎么办

不给操作手册。给操作手册这个动作本身就违反了这篇文章的全部内容——手册是量表的标准形态。

但有四件事可以说。

第一，先把能量的全量清楚。

绝大多数东西该测,而且该测得清清楚楚。工时该, 工资该, 安全该, 谁做了什么该。不测量的地方最容易藏剥削和偏私,这一点比这篇文章的所有内容都更常见、更紧迫。

顺序不能反：先把该算清的全算清，再谈剩下那一小块。 顺序反了就是要流氓。

第二，把替换那一步写下来。

第四节讲的四道工序里，第一道是替换——把要的东西换成能量的东西。这一步现在通常是口头完成的，然后被忘掉。

能做的最省力的一件事是：**每次设立一个指标时，把"这个指标是用来替代什么的"写在指标旁边，且永远不删。**

这一条听起来微不足道，但它挡住的是第四道工序——那个把不可测物重新解释为不存在的动作。只要那行字还在，就总有人能指着它问一句"我们原来要的是这个吗"。

第三，留一块明确不评价的地方，并且不奖励它。

关键在"不奖励"三个字。一旦奖励，它就变成新的指标，四道工序重新开始跑。

这块地方会被浪费掉一大部分。这不是设计缺陷，这就是它的价格。你没法既要这块地方的产物，又要每季度确认它在长。

而这一条有一个必然的失败模式，必须一起说：**当上级压力增大、指标吃紧时，第一个被牺牲的永远是这块地方。** 因为在账面上，取消它没有任何代价。要顶住这一点，需要一个愿意为一块空白挡子弹的人——而这件事本身没有任何可测的回报。

第四，问一句"这个形状进不来的东西是什么"。

这是全文最有用的一句话，也是从政策那一家学来的。

每次你设立一道门槛——必须有数据、必须有对照组、必须能量化、必须写清预期成果——问一句：**因为这道门是这个形状，什么东西现在进不来了？**

大多数时候你会发现，进不来的恰恰是最大的那几件。

一把可以随身带的尺子

最后收成三个问题，谁都能用，用在任何一次设立指标之前。

第一问：我真正想要的那个东西，能不能直接测？

如果能——直接测它，别设替代品。这种情况比人们以为的多：营收就是营收，工伤起数就是工伤起数，房子盖没盖起来就是盖没盖起来。这一类根本不归本文管。

只有当答案是"不能，我要的那个东西没法直接测"的时候，才往下走。

第二问：我要用什么去替代它，这个替代品跟它是什么关系？

三种可能，后果完全不同：

正相关——替代品越好，真东西通常也越好。这是最好的情况，放心用，但仍然要把"这是替代品"写在旁边。

无关——替代品和真东西各走各的。这种情况用了也白用，但至少不主动坏事。

负相关——替代品越好看，真东西反而越差。这是本文全篇在讲的那一类，学习中的当下表现是最清楚的例子。**碰上这一类，用它就是在主动往反方向开车，而且开得越认真，跑得越远。**

大多数人从来没有问过第二问。他们默认替代品和真东西是正相关的——而这个默认，恰恰是最该被检查的地方。

第三问：因为这道门是这个形状，什么东西现在进不来了？

这是三问里最有用的一问，也是唯一一个能在设立指标的当下就发现问题的。

设完一条指标，把它当成一道门，然后想一想：**哪一类做法因为形状不对，从此不会再被提出来了？**

如果你想不出任何一类，多半是你想得不够久。任何一道门都会挡住东西，一道挡不住任何东西的门不是门。

这把尺子不告诉你该不该设指标。**它只保证你知道自己在换什么。** 而在这之前，绝大多数人以为自己只是在把管理做得更规范一点。

十一、什么情况下这篇文章是错的

一个说法如果怎么都推不翻，那它也就永远证明不了自己对。所以要把话说死。以下四种情况，任何一种被证实，本文的核心判断就该被放弃。

第一种：可测的那一面与要的那一面其实是正相关的。

本文的地基是"能量的那个量与真正要的那个量方向相反"。这一条可以直接测。

取一个领域，同时记录两组数据：一组是当前正在被用作评价依据的指标，另一组是那个真正想要的结果——需要长期追踪才能拿到的那种。学习就追三年后的保持，医疗就追五年后的生活质量，公益就追十年后的处境。

如果两组数据在足够大的样本上显著正相关，那么本文所说的整个偏斜就不存在，只是测得还不够久。**这一条是最直接的，而且是可以真做的。**

第二种：找到一个把不可测那一面成功纳入而没有毁掉它的例子。

本文说不可测物一旦被纳入量表就变质。那么只要找到一个反例就够：某个领域成功地把原本不可测的那一面变成了可测的，而且被测的那一面没有因此变形——从业者没有开始朝指标长，那样东西还是原来那样东西。

如果这样的例子存在并且能被重复，那么本文最多降格成一句弱得多的话："某些测量方式有害"。而那句话是常识，不值得写两万字。

第三种，也是我自己最担心的：也许起作用的根本不是"可测"，而是"被用于分配"。

有一个解释比我的解释更朴素，可能也更对：**问题不在于一样东西能不能被测量，而在于测量结果有没有被拿去分配资源。**

按这个解释，一个纯粹为了了解情况而做的测量，不会造成任何偏斜——只有当测量结果决定了谁拿钱、谁升职、谁被淘汰时，四道工序才启动。如果是这样，本文命错了名——真正该写的是"分配"，不是"可测"。

这两个解释在绝大多数情况下预测一样，因为现实中几乎所有测量最后都被拿去分配了。要分开它们，必须找到一个**测了但明确不用于任何分配**的安排：数据只归档，不进任何评价，参与者知道这一点且这一点可信。

如果在这种安排下，那种朝指标长的行为照样出现——说明是可测本身的问题，本文站得住。如果完全没有出现——说明本文找错了对象，真正起作用的是分配，我该收回这个名字重写。

我倾向于前者，理由是第四节第一层讲的那个道理：**正在长的东西还没有形状，这一点与分不分配无关。**但我倾向于哪一种不构成证据。这是本文身上最软的一处，我把它指出来，而不是绕过去。

第四种：三个来源里有一个被推翻。

本文撞的三家理论，每一家自己都还在被争论中。如果其中任何一家的核心主张被有力推翻——比如证明学习中的困难其实全程有害，或者证明细胞群并不持有任何目标状态，或者证明个体层干预的效果其实很大——那么本文的三角就塌了一个角。

三角塌一个角，剩下两家就变成了互相印证而不是互相打架，而本文那条判断恰恰是从打架处捡出来的。到那时它需要重新被撞一次。

十二、出处与最近的邻居

本文不是学术论文，正文里没有加注。但把话说到了别人头上，就该把出处交代清楚；更要紧的是，其中有几条离本文非常近，不把界划开，读者有理由怀疑本文只是给旧说法换了个名字。

下面分三组：被撞的三家、最近的邻居、以及其余的支撑。

第一组：被撞的三家

教育学·学习中的困难那一家

- Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 59–68).
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. doi: 10.1177/1745691615569000
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475–479.
- Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable difficulty effects in the learning of high-element interactivity materials. *Frontiers in Psychology*, 9, 1483. doi:10.3389/fpsyg.2018.01483
- Pyke, W., Lunau, J., & Javadi, A.-H. (2024/2025). Does difficulty moderate learning? A comparative analysis of the desirable difficulties framework and cognitive load theory. *Quarterly Journal of Experimental Psychology*, 78(10), 2181–2195. doi: 10.1177/17470218241308143

第三条是本文第三节那句“能当场测到的和真正想要的方向相反”的直接来源——它的标题就是「学习与表现」，而它整篇在论证的正是：训练过程中我们能观察和测量的是表现，而表现常常是一个不可靠的指标。第五条是对立那一方最要紧的一篇，它论证在高要素交互的材料上，额外的困难是有害而非有益的。第六条是两家的正面比较。

医学·细胞群有目标那一家

- Levin, M. (2025). The multiscale wisdom of the body: Collective intelligence as a tractable interface for next-generation biomedicine. *BioEssays*, 47(3), e202400196. doi: 10.1002/bies.202400196
- McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7, 378. doi: 10.1038/s42003-024-06037-4
- Levin, M., & Watson, R. (2026). Machines all the way up and cognition all the way down: Updating the machine metaphor in biology. *Seminars in Cell & Developmental Biology*, 177–178, 103668. doi:10.1016/j.semcdb.2026.103668

本文用到的三条——当前生物医学几乎全是自下而上瞄准分子硬件、细胞群是有目标的求解者、以及接口的价值在于可读可写而非最根本——集中在第一篇。

这一路并非孤例。更早的一场同类争论可以参照：

- Soto, A. M., & Sonnenschein, C. (2011). The tissue organization field theory of cancer: A testable replacement for the somatic mutation theory. *BioEssays*, 33(5), 332–340.
- Vaux, D. L. (2011). In defense of the somatic mutation theory of cancer. *BioEssays*, 33(5), 341–343.

这两篇是同一份刊物上正面对垒的一对，一方主张癌是组织层面的问题、基因突变是结果而非原因，另一方为主流的突变论辩护。争的还是那个问题：病因在部件还是在部件所处的那一层。

商业经济与公共政策·该改人还是该改规则那一家

- Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, 46, e147.

本文引用最重的那一条——随机对照试验的要求大幅限制了能被纳入考虑的政策范围——出自他们及合作者在《Behavioural Public Policy》上的后续文章。对他们最有力的反驳来自 Cass Sunstein，他明确不同意个体层干预会挤占系统层改革这一条，也不同意个体层干预效果有限。这场争论到二〇二六年仍在进行。

第二组：最近的邻居——本文必须与它们划清界线

这一组是本文最该老实交代的部分。有六个说法离本文非常近，如果不把差别说清楚，本文就只是换了个名字。

- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90. doi:10.1016/0149-7189(79)90048-X
- Goodhart, C. A. E. (1984). Problems of monetary management: The UK experience. In *Monetary Theory and Practice* (pp. 91–121). London: Macmillan.
- Strathern, M. (1997). "Improving ratings": Audit in the British University system. *European Review*, 5(3), 305–321.
- Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.
- Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40. doi:10.1086/517897
- Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton: Princeton University Press.

与坎贝尔的界线。 坎贝尔那条广为流传的判断是：一个量化的社会指标越是被用于社会决策，它就越容易受到腐蚀压力，也越容易扭曲和腐蚀它本要监测的那个社会过程。

这是离本文最近的一条，近到本文必须解释自己多出了什么。

差别在**起点**。坎贝尔的因果链从"被用于决策"开始——先有一个指标被拿去分配资源，然后腐蚀才发生。本文的因果链从更前面一步开始：**在任何腐蚀发生之前，不可测这件事本身，就已经决定了人们会在哪一层动手。**

还有一个差别在**对象**。坎贝尔关心的是那些进入了指标、然后被扭曲了的东西。本文关心的是那些**从来没有进入过指标**的东西——它们不是被扭曲了，是根本没上场。第四节第三层讲的"不可测的那一面没有代言人，它是被弃权判负的，甚至没有出场"，这一句坎贝尔那条判断里没有。

与古德哈特的界线。 那条被反复引用的表述——当一个量度成为目标，它就不再是一个好量度——讲的是被测量者的策略性反应：人们会针对指标优化，于是指标与它原本代理的东西脱钩。

本文第五节的第三道工序讲的那个人，**没有做任何策略性反应**。他不作弊，不针对指标优化，他只是在五年里逐渐学会了生产能被认的东西，而且他自己感觉不到。性质的转换不需要任何人博弈就已经完成了。

与反应性研究的界线——这一条本文欠得最多。 埃斯佩兰与索德二〇〇七年那篇论证：人会因为被评价、被观察、被测量而改变行为；他们给出两条起作用的路径，自我实现的预言与可通约化。这是本文第五节第三道工序最直接的学术依据，本文该老实承认这一笔。

差别在于**追问的位置**。他们研究的是一个已经存在的排名如何重塑机构；本文追问的是更靠前的一步——**为什么某些东西能进入排名而另一些进不来，以及那道门的形状本身决定了什么。** 用第十节最后那个问法来说：他们研究的是走进房间之后发生了什么，本文研究的是门是什么形状。

与量表的暴政的界线。 穆勒那本书的方向与本文一致，都在讲量化指标会挤出专业判断。

但落点不同，而且是相反的。穆勒的处方偏向少用指标、更多信任专业判断。**本文明确拒绝这个处方——**第九节整节都在说，不测量的世界不是一个温情的世界，是一个凭关系的世界，而在那个世界里最先被牺牲的永远是没有关系可凭的人。本文的落点不是"别测"，是"知道自己在换什么"。

与审计社会的界线。 鲍尔那本书提出审计不只是检验信任，它替代信任。这一条与本文第四节第三层最相关：可测的那一面天生有代言人，而不可测的那一面只能说"我觉得这个很重要"，这句话在会议室里等于什么也没说。

总起来说，这六家各自说对了一件事，而它们共同的位置是：**它们都在研究测量落到某个领域之后发生了什么。** 而本文关心的是更前的一步——测量的可能性本身，在任何测量被实施之前，就已经替所有人选好了动手的层级。这一步，六家里没有一家正面处理过。

如果读完之后你仍然觉得本文说的其实就是其中某一家，那么第四节没有写成功，责任在我。

第三组：其余的支撑

- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417–444.
- Bjork, R. A., & Bjork, E. L. (2019). Forgetting as a friend of learning: Implications for teaching and self-regulated learning. *Advances in Physiology Education*, 43, 164–167.
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384. doi:10.1098/rsos.160384
- Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's Law. arXiv: 1803.04585

第三条值得单独提一句，因为它是本文第五节第三道工序最严格的形式化版本：如果一个领域奖励的是产出数量而不是结果的可靠性，那么方法较差但出得快的研究做法会在群体层面被选择出来——**而这个过程不需要任何研究者主观上想要作弊**。这一条与本文那句“他没有变坏，他只是不再看见那些不计分的地方了”是同一件事的两种说法，一种用演化的语言，一种用一个人的五年。

十三、这一篇是怎么撞出来的

按本栏的规矩，把工序和作废数一并交出来。

三个来源，分属三个互不相干的学科，且在同一个问题上给出互相排斥的答案：那件事该在哪一层动手。

- **教育学**（可取用困难那一家）：在单元这一层动手，而且要加阻力。
- **医学**（细胞群有目标那一家）：别在单元这一层动手，单元自己有目标，要改的是目标。
- **商业经济与公共政策**（该改规则那一家）：单元这一层根本就不是杠杆所在，而且在这一层讨论本身就是被推动的。

九条判断，三家各抽三条最承重的。

教育学那三条：记得牢和此刻想得起来是两个独立的量；难度不是要克服的成本，难度就是那个起作用的动作本身；人对“我学会了没有”的判断系统性地是反的。

医学那三条：细胞群是有目标的求解者不是被动零件；杠杆在目标不在零件；接口的价值不在于它最根本，而在于它可读可写。

政策那三条：个体层干预的效果一直低于预期；把问题定义在个体层这个动作本身是被推动的；个体层干预会挤占系统层改革。

三十六次对撞，作废十一条。

作废的分两类。七条是同义重复——两条判断说的其实是同一件事的两种说法，撞不出新东西。四条是撞不起来——两条判断各说各的，中间没有互相顶住的地方，硬要连只能靠"也""并且"，那不是碰撞，是拼贴。

这一栏在多数账本上一分不值。作废的部分不会出现在成品里，没有人会为它付钱，在任何进度表上都记作零。但它是这篇文章的地基——留下的二十五条之所以立得住，正是因为那十一条被扔掉了。

而且必须承认，这一栏本身也是这篇文章所讲的那件事的一个例子：**碰撞质量这个东西没法量,能量的只有成品字数。**

留下的二十五条聚成五条暗流：

一、能量的那个量与真正要的那个量方向相反。 二、把有目标的东西当成没目标的材料。 三、"上一层"是歧义的——三家都说上一层，指向三个互斥的方向。 四、接口既是杠杆也是漏洞。 五、失效的做法靠占位存活。

第一条是根：可测量性决定了哪一层被当作可以动手的层。可测的是零件，于是所有人都在零件层动手（第二条）；上一层因为不可测,每家只能凭自己的理论猜它在哪儿，猜出三个互斥方向（第三条）；接口之所以既是杠杆也是漏洞，正因为接口就是可测的那一面（第四条）；失效的做法能占位，是因为它有可测的产出（第五条）。

收敛成一条判断：一个系统里能被读出来的那一面，会自动接管它被治理的方式；而那一面通常不是它长东西的那一面,甚至跟长东西的那一面方向相反。

这条判断三家里谁也说不出来。第一家只看到学习者的自评是反的，第二家只看到医学在错的层级动手，第三家只看到个体层是被政治性地推动的。三家各自把这件事当成自己领域的病，而它不是——**它是量表这个东西的普遍性质。**

三家之所以各自看不见，也是有原因的，而且原因就在各自的位置上。教育研究的全部方法建立在测量之上，要它怀疑测量本身，等于怀疑自己的仪器。医学那一家忙于反对"修零件"，而没有往上追一层问"为什么大家都在修零件"——答案是因为零件可测。政策那一家离答案最近，他们已经点出了随机对照试验的门槛效应，但他们把它当作行为科学界的一个具体失误，没有把它认成一条普遍规律。

能在几个学科里同时兑现：本文第八节给了十个——学校、医院、公共政策、科研、公司、养孩子、公益、算出来的推荐、判案、安全与风控。每一个都写成了真两难，而不是"所以别测量"。

十四、这篇文章的自我审判

最后必须说一句对我自己不利的话，否则前面两万字全部作废。

这条判断对这篇文章本身适用，而且是双重致命的。

第一重：我之所以能找到这三家理论，是因为它们各自把自己的争论写成了**可检索的论文**。那些没有被写成论文的、可能更对的判断，我根本看不见。**这篇文章的取材面，恰好等于可检索面。**

而按照本文自己的判断，可检索的那一面，多半不是那些真正长东西的地方。所以本文极可能漏掉了一整片更要紧的东西，而且我没有办法知道漏掉的是什么——这正是第四节第三层说的那件事：不可测的那一面没有代言人，包括在这篇文章里。

第二重更直接。这篇文章会被打一个分数。

而分数正是一个量表。一旦它有了分数，关于它的一切判断都会开始向那个分数靠拢：分数高就值得读，分数低就不值得。**读没读过、有没有被它改变过什么，这些不可测的东西会退到分数后面去。**

也就是说：**我在写完一篇论证"量表会接管"的文章之后，亲手给它装了一个量表。**

这不是自谦，也不是一个漂亮的收尾。这是这条判断的直接推论，我绕不开。我唯一能做的，是把这一笔如实写在这里，而不是假装我发现了一个不会被自己的发现所损坏的东西。

那么写它还有意义吗？

有一点。因为在这篇文章之前，那个替换动作是**无声地**完成的——没有人知道自己刚才把什么换成了什么，只知道"现在的评价体系有点问题"。这篇文章至少让那次替换变得可见：你在写下那条指标的时候，会知道自己正在把什么挪出视野。

知道自己在做一次替换，和以为自己只是在做规范化——这两者之间的差别，值这两万字。

但也仅此于此。

如果读完这篇文章，你只是从此多了一个词可以用——那你就是它所描述的那个替换动作。

如果读完之后的某个下午，你在设立一条指标之前停了一下，问了一句"因为这道门是这个形状，什么东西现在进不来了"——那这篇文章才算被接住了。

而那件事，我永远不会知道。这正是它成立的条件。

本文由三个分属不同学科、且互相冲突的理论体系碰撞而成，作者 王德生 + Claude。

本文由三个分属不同学科、且互相冲突的理论体系碰撞而成。全部来源均为站外公开文献，可自行核对。 作者：王德生 + Claude · SDE Universes · 学科通融