

德麦国际 · SDE UNIVERSES

拒绝的社会生产

——教育评价的测度外壳与外壳再分泌律

王 德 生

今日长文 · 教育专栏 × 学术创造专栏 · 期刊论文频道

王德生 著 · SDE UNIVERSES · 约 4.8 万字 · 发表于 2026年7月18日 · 修订稿 · 三种阅读模式

本文的位置 · 2026 年 7 月 18 日

本文已被《反事实自我的社会制造——一步守恒、不可结清性与稀缺分配装置的产品》吸收为它的一个推论。那篇论文从被拒者的处境出发做了一条不引用任何人的推导，并由此订正了本文三处：**其一**，第六节第三小节的二约束交集不完整，规格是三条；**其二**，第十节第三小节的干预落点被撤回，反例是复读；**其三**，第四节第五小节把稀缺标为"可能才是根"，这一条被注销。本文正文一字未改，订正以标注形式插在原处——**一篇被吸收的论文应当留在原地被指着，而不是被悄悄改成对的。**

摘 要

现代教育评价研究共享一个从未被审视的预设：评价装置是为了测量学习而被发明的，它在当代的失灵是一种退化。本文论证这一预设历史与结构两个层面都不成立。从科举到普鲁士文官考试再到美国的标准化测验，大规模评价装置的诞生场景无一不是认知需求，全部是分配危机——需要在稀缺机会面前产生一种能被落选者接受的落选。据此，本文提出一个方向相反的判断：**评价装置的社会产品不是关于学习的信息，而是可接受的拒绝；测度不是这一产品的目的，而是它的生产条件。**测度功能是安抚功能为使拒绝可接受而分泌的准确性表象，本文称之为"测度外壳"。外壳在结构上必须看起来像测量，但不必是测量，因为它所生产的从来不是准确性而是不可争辩性。这一重新定位产生四个既有理论无法产生的推论。其一，分数与能力的脱钩不是病症而是常态，人工智能所揭穿的不是脱钩本身，而是社会曾经相信过耦合这一集体记忆。其二，任何承担稀缺分配功能的替代性评价装置，将在其优化产业成熟的时刻分泌出同构的测度外壳并与被测对象脱钩——替代不是替代，是宿主更换，本文将这一机制凝缩为"外壳再分泌律"。其三，该律有一个精确的边界条件——排序型装置必然分泌外壳，阈值型装置不必，这一区分为干预给出了一个非同义反复的落点。其四，脱钩时滞由优化产业的成熟速度决定，而人工智能同时压低了优化知识的生产成本与传播成本，因此时滞将坍缩到耦合记忆的形成时间以下；由此**外壳将换代而非消亡——从结果的不可争辩换到过程的不可争辩**，当前被普遍读为"防作弊"的那一整套技术，是外壳的换代，不是外壳的修复。本文以此正面对置波特（Porter）的机械客观性理论这一最强竞争解释，并在比较政治学中兑现该律的一次强制不一致：民主指数本身是排序型装置，它必然分泌外壳，因而该学科无法同时保留"反应性为真"与"民主指数测量被测国家的真实属性"这两个它目前都持有的命题。本文最后以自身在同行评审中的处境完成对该理论的自反检验，并解剖出本文自己分泌的一片外壳作为标本。

关键词：可接受的拒绝；测度外壳；外壳再分泌律；脱钩时滞；排序型与阈值型装置；教育评价；通约化

目 录

一、一个问错了的问题	4
二、四条抵近而未合拢的脉络	6
三、可接受的拒绝、测度外壳与再分泌律	9
四、方法与它的边界	13
五、产品倒置：三个装置的诞生场景	16
六、外壳为什么不必准确，又为什么仍需一点准确	19
七、外壳再分泌律：一条会禁止事情的律	22
八、选举：律在比较政治学中的强制不一致	30
九、与最强对手的正面交锋	33
十、两个落点，都不在评价内部	36
十一、结论	39
十二、自反：这条律适用于评价本文的装置吗	41
十三、参考文献	44

一、一个问错了的问题

(一) 一个不可解释的事实

2023 年以来，随着大语言模型能够在数分钟内生成合格的课程论文，全球教育界爆发了一场关于评价合法性的持续争论。争论的形式高度一致：如果机器能够在不具备学科素养的情况下拿到分数，那么分数还能证明什么？由此衍生出的对策也高度一致——加强监考、回归纸笔、开发检测工具、重新设计题型，或者更激进地，用项目制评价、能力档案、微证书来替代标准化分数。

这些对策彼此对立，却共享同一个前提：评价系统是一个测量装置，它现在测不准了，所以需要修理或更换。争论的双方——保守派要修理旧仪器，改革派要更换新仪器——在"这是一个仪器问题"这一点上从未分歧。

本文认为这个前提是错的，并且它错得很彻底：它不是把答案搞错了，是把这个装置是什么搞错了。

考虑一个在现有框架中无法被消化的事实。2008 年金融危机以一种无可辩驳的方式证明，三大信用评级机构对结构化金融产品的评级不具备预测效度——被评为 AAA 的证券大规模违约。按照"测量装置失灵则被更换"的逻辑，评级机构应当在此后被边缘化。实际发生的是相反的：标普、穆迪、惠誉在危机后的十余年里营收创下历史新高，其评级被更深地写入了各国的金融监管条款。一个被公开证伪的测量装置，不但没有消亡，反而获得了更强的制度地位。

如果测量的准确性是这类装置的存在理由，这个事实就是不可解释的。如果它不是，那么我们需要问的就不是"这个装置测得准不准"，而是"**这个装置到底在生产什么，以至于它的测不准无关紧要**"。

本文的回答是：它生产的是可接受的拒绝。

(二) 三种既有解释及其共同的方向错误

围绕教育评价的困境，既有研究提供了三条主要解释路径。

第一条是技术优化路径。这一路径将评价视为一个测量工程问题，其核心议程是提高信度、效度与公平性（Black & Wiliam, 1998; Stiggins, 2002）。它对当前危机的诊断是"工具跟不上时代"，处方是更好的工具。这一路径的盲区是自明的：它无法解释为什么一个多世纪的测量技术改进，并没有使公众对评价的信任上升，反而与之同步下降。

第二条是批判再生产路径。这一路径指出评价装置是社会不平等的再生产机制——它并非中立地测量预先存在的能力差异，而是把阶级结构的不平等转译为看似天然的分层数据（Bourdieu & Passeron, 1977; Bowles & Gintis, 1976; Au, 2009）。这一路径的解释力极强，但它有一个自身难以消化的经验困难：如果评价装置是统治结构施加于被统治者的符号暴力，那么最积极维护"分数是公

平的"这一命题的，本应是既得利益者。经验上恰恰相反——最激烈反对降低标准化考试权重的，往往是中下阶层家庭。批判路径通常用"误识"（*méconnaissance*）来处理这个困难，即被统治者内化了统治者的标准而不自知。但这一处理引入了一个不可证伪的辅助假设：任何反例都可以被解释为更深的误识。

第三条是制度惯性路径。这一路径强调旧制度在教材体系、教师培训、录取算法、社会认知上的巨大沉淀，用路径依赖与锁定来解释改革的反复失败（David, 1985; Arthur, 1994; Pierson, 2004; Mahoney & Thelen, 2010）。这一路径能够解释"改起来慢"，但它预测的是一条持续衰减的阻力曲线，而非教育改革所呈现的"启动—反弹—回到原点—再启动"的波形。更关键的是，它把评价装置的当前功能当作既定事实接受下来，只研究它为何难以改变，从不追问它一开始为何被建立。

三条路径的解释力各不相同，但它们共享一个方向：它们都把"测量学习"当作评价装置的目的，把"正当化分配"当作它的效应、副产品或被劫持的结果。技术路径要修好这个目的；批判路径揭露这个目的被权力污染；制度路径解释为什么这个目的难以被重新校准。

本文提出的判断是：方向反了。**正当化分配不是效应，是目的；测量学习不是目的，是这个目的得以运作的条件。**这不是强调重点的调整，而是因果箭头的调转，其后果贯穿本文全篇。

（三）本文的判断与结构

本文的核心判断可以先行给出：

大规模教育评价装置的社会产品，是可接受的拒绝——一种能够让被排除者在被排除后继续参与社会合作、而不诉诸暴力、退出或指认具体施害者的排除方式。为使拒绝可接受，装置必须分泌一层看起来像测量的表层，本文称之为测度外壳。外壳的产品不是准确性，而是不可争辩性；它必须借用测量的语法，因为测量的语法自带不可争辩性，但它不必真的测量任何东西。

这一判断的四个直接推论构成本文的主体：

推论一（脱钩常态）：外壳与被测对象的脱钩不是退化，是稳态。所谓"人工智能揭穿了分数与能力的脱钩"这一说法本身包含一个错误——被揭穿的不是脱钩（它一直存在），而是社会曾经相信过耦合这件事。被摧毁的是记忆，不是效度。

推论二（外壳再分泌律）：任何承担稀缺分配功能的排序型评价装置，其测度外壳都将在其优化产业成熟的时刻与它声称测量的对象脱钩，而装置本身不会因此消亡；试图以新装置替代旧装置的努力，会使新装置在获得同一社会功能的同时分泌出同构的外壳。替代不是替代，是宿主更换。

推论三（排序/阈值分界）：外壳再分泌律有一个精确的边界条件。装置分泌外壳的压力，来自它必须让被拒者接受被拒；而"被拒"有两种语法。排序型装置（你比别人差，所以你出局）必然分泌外壳，因为它的拒绝理由是相对的、因而永远可被争辩（"凭什么是他不是我"）。阈值型装置（你没到线，所以你出局）不必分泌外壳，因为它的拒绝理由是绝对的、可被被拒者自己验证（"我确实不会开

飞机")。这一区分解释了一个长期被忽视的经验规律：职业资格考试的测度效度显著高于升学考试，且不随时间衰减。它同时给出了本文唯一的、非同义反复的干预落点。

推论四（时滞坍塌与外壳换代）：脱钩时滞不是一个历史常数，它是优化产业成熟速度的函数。人工智能同时压低了优化知识的生产成本与传播成本，因此时滞将坍塌到耦合记忆的形成时间以下。当没有任何装置来得及生产出"它曾经是准的"这个记忆，结果型外壳就失去了可归因性——而分泌压力不会因此消失。因此外壳将换代：从结果的不可争辩，换到过程的不可争辩。当前被普遍读为"防作弊"的那一整套技术（监考录像、过程录屏、击键记录、AI 使用日志、现场追问），不是外壳的修复，是外壳的换代。

本文的结构安排如下。第二节梳理文献，精确定位本判断与最接近的既有理论——波特的机械客观性——之间的分岔点。第三节提出理论框架，操作化核心概念并给出命题系统。第四节说明方法。第五节从装置的诞生场景论证产品倒置。第六节论证外壳的分泌机制与其"不必准确"的结构原因。第七节论证外壳再分泌律，给出脱钩时滞的机制，并处理其边界条件。第八节把该律在比较政治学中兑现为一次强制不一致。第九节正面对置最强竞争解释并给出检验设计。第十节从该律推出干预路径。第十一节结论。第十二节承担自反检验。

二、四条抵近而未合拢的脉络

（一）通约化研究：把不可比的东西变成可比的

与本文最相关的第一条脉络，是关于通约化（commensuration）的社会学研究。埃斯佩兰和史蒂文斯（Espeland & Stevens, 1998）在一篇奠基性综述中把通约化定义为"把不同性质的属性转化为一个共同度量"的社会过程，并指出它是一种被严重低估的权力形式：通约化通过决定什么被纳入度量、什么被排除在度量之外，重塑了它所度量的世界。他们特别强调，通约化的核心效应是信息的丢弃——它把大量异质信息压缩为一个可排序的标量，而被丢弃的部分不会留下痕迹。

这一脉络在埃斯佩兰和绍德（Espeland & Sauder, 2007; Sauder & Espeland, 2009）对法学院排名的研究中获得了最有力的经验兑现。他们发现，《美国新闻与世界报道》的法学院排名并非中立地反映法学院的质量，而是反应性地（reactively）改变了法学院的行为：院校为了提升排名而重组资源配置、操纵录取数据、改变奖学金策略，最终使排名所测量的对象向排名的度量标准靠拢。他们把这一机制称为"反应性"，并指出其结果是排名与它声称测量的东西之间形成一个自我实现的闭环。

这一脉络对本文极其重要，但它止步于一个关键处：它解释了通约化如何改变被测对象，没有解释通约化为什么被社会需要。埃斯佩兰与史蒂文斯在综述结尾承认，通约化的兴起动因是一个未解问题，他们列出了效率、可比性、问责等若干候选，但没有给出一个统一机制。本文提供的正是这个机制：通约化被需要，是因为排序必须被排除者接受，而排序只有在被表述为共同度量上的位置差异时才可能被接受。**通约化不是权力的工具，它是拒绝的语法。**

（二）波特的机械客观性：最接近的、也是最强的竞争解释

本文必须正面处理的最强竞争解释来自波特（Porter, 1995）的《相信数字》。这是与本文判断最接近的既有理论，两者的分歧点必须被精确划出，否则本文的贡献将被消解为一次换词。

波特的论证是：定量方法在现代社会的兴起，不能用“科学进步”来解释，因为在许多领域，量化取代专家判断反而降低了决策质量。动因是信任的匮乏。他的核心经验对置极其有力：法国的国立路桥学校工程师拥有牢固的社会地位与不受挑战的专业权威，他们从不使用成本效益分析；美国陆军工程兵团处在国会、地方利益集团与公众舆论的持续质疑之中，他们发展出了世界上最精密的成本效益分析体系。波特由此提出“机械客观性”（mechanical objectivity）概念：数字提供了一种非个人的、按规则运行的、看起来不依赖任何人的判断的决策形式，它的功能不是让决策更准确，而是让决策能够抵御质疑。他的凝缩是：量化是弱者的策略。

波特已经说了“量化不是为了准确”。那么本文还剩下什么？

分歧点有三处，且每一处都产生相反的预测。

第一，因果方向相反。 波特的机制是防御性的、自上而下的：专家或官僚在被质疑的处境中，采用量化作为盾牌。承担需求的是决策者。本文的机制是需求性的、自下而上的：被拒绝者在必须接受拒绝的处境中，需要一个能让自己接受拒绝的理由。承担需求的是被决策者。这不是同一个机制的两种说法。它们对同一个变量给出相反的预测：波特预测专家权威越高，外壳越薄；本文预测拒绝越不可逆，外壳越厚，与专家权威无关。第九节将以此设计检验。

第二，解释范围不同。 波特解释了量化为何被采用，没有解释量化在被证伪后为何不被放弃。他的框架内含一个预测：如果机械客观性的客观性外观被戳破，它的防御功能就失效，从而应被放弃。这一预测被 2008 年后信用评级机构的境遇直接证伪——外观被彻底戳破，制度地位反而上升。本文的框架能够消化这个事实：承重的从来不是客观性，是可接受性；而可接受性只要求“存在一个非人格的程序”，不要求“这个程序是对的”。

第三，理论形态不同。 波特提供的是一部历史社会学的谱系，其命题形态是“在什么条件下量化会兴起”。本文提供的是一条带边界条件的律（外壳再分泌律），其命题形态是“任何承担 X 功能的装置必然出现 Y，除非满足 Z”。律必须能被证伪，本文在第七、九节给出了它的证伪条件与边界。

一个更大的家族。 波特不是孤例。斯科特（Scott, 1998）关于国家可读性（legibility）的论证具有同样的形态：国家为了管理而把复杂的、地方性的、只有当事人才懂的社会现实，简化为可以从中心被读取的网格——地籍、姓氏、标准度量衡、单一栽培林。斯科特与波特分享同一个因果方向：简化是自上而下的需求，被简化者是承受方。本文与这个家族的分岔是同一处：本文主张，在评价这一类装置上，简化的需求首先来自下面——来自那个必须接受自己被简化的人。不是国家需要把他读成一个数字，是他需要被读成一个数字，才能接受那个数字对他做的事。

（三）组织社会学的脱耦：分离，而非分泌

迈耶和罗恩（Meyer & Rowan, 1977）的经典命题指出，组织采纳某些正式结构并非为了技术效率，而是为了与制度环境中的合法性标准保持一致，而这些结构可能与组织的实际运作完全脱耦

（decoupled）。布罗姆利和鲍威尔（Bromley & Powell, 2012）进一步区分了“政策—实践脱耦”与“手段—目的脱耦”，指出当代组织越来越多地只是执行规则的形式，而不再关心规则是否实现其宣称的目的。

这一脉络为“学校维持严格评价并非为了促进学习，而是为了展示公正”提供了精确的制度主义描述，本文承其洞见。但脱耦理论描述的是一种分离关系：两样东西本来在一起，后来分开了，或者一开始就是两张皮。本文描述的是一种生成关系：其中一样是另一样分泌出来的。这个差别不是修辞。

脱耦理论预测：如果外部合法性压力消失，正式结构会被放弃，因为它此时纯属冗余。分泌关系预测：即使外部压力消失，只要装置仍需产生拒绝，外壳仍会被分泌——因为外壳不是给外部看的，是给被拒者看的。取消排名后立即出现的自发影子排序（见第九节），是脱耦理论无法预测而本文可以预测的现象。

（四）文凭主义与信号理论：两个被误读的经典

科林斯（Collins, 1979）的文凭社会理论论证，教育扩张并未提高社会流动性，而是制造了文凭通胀的螺旋；文凭是身份群体争夺职业垄断地位的武器。斯宾塞（Spence, 1973）的信号模型则论证，教育的价值不在于所传递的技能，而在于高能力者获取文凭的成本较低这一分离均衡条件，使文凭成为可靠信号。

这两个经典在当代讨论中被反复引用，但都被误读为“文凭的功能是筛选”。实际上科林斯的论证已经非常接近本文：他明确指出文凭的技术功能（培养生产力）在经验上几乎不成立，其真实功能是冲突中的地位划界。本文与科林斯的分歧在于：科林斯是冲突论的，他的世界里有争夺者与被排斥者，装置是前者压制后者的武器；本文的世界里，被排斥者自己需要这个装置。这不是补充科林斯，是与他相反。一个纯冲突论的世界无法解释为什么落选者游行要求的是“考试要更严格更公平”而不是“取消考试”。

斯宾塞的模型则给出了一个本文可以直接使用的技术结论：一旦获取信号的成本对所有类型的行动者都趋近于零，分离均衡崩溃，信号的信息含量归零。生成式人工智能正在做的就是这件事。斯宾塞的模型由此预测文凭应当失去价值。它没有失去。这是信号理论无法消化、而本文可以消化的事实：**信号的信息含量与它的社会功能无关。**

（五）拒绝的社会学：一条被忽视的线索

有一条与本文直接相关、却几乎从未被引入教育研究的线索：埃尔斯特（Elster, 1989）在《所罗门式判决》中系统研究了社会如何在理性无法决定的情况下做出决定——器官分配、子女监护权、征兵、稀缺住房的分派。他的核心发现是：在这些场景中，社会经常选择那些明显不是“最优”的机制（抽

签、排队、先到先得、机械的资历规则），而选择它们的理由恰恰是它们不做实质判断。埃尔斯特指出，一个机制的价值有时正在于它切断了争辩的可能性。

埃尔斯特研究的是分配，没有研究评价；他的机制是“不判断”，本文的机制是“看起来在判断”。但他给出了本文所需的、既有文献中最硬的一块地基：存在一整类社会装置，其功能是产生可被接受的分配结果，而不是产生正确的分配结果。教育评价属于这一类装置，只是它选择了与抽签相反的策略——不是承认自己不判断，而是分泌一层判断的外壳。为什么？因为抽签生产的是可接受的运气，而现代社会需要的是可接受的应得。这个差别是第六节第三小节的主题。

（六）缺口的精确定位

综上，既有文献在四个方向上分别抵近了本文的判断：通约化研究看到了排序如何重塑被排序者，波特看到了量化不为准确，脱耦理论看到了形式与实质的分离，埃尔斯特看到了存在一类以可接受性而非正确性为产品的装置。

但没有任何一条脉络把这四点合起来问一个问题：**如果一个装置的产品是可接受的拒绝，那么它的“测量”部分是什么？**

本文的回答——测量部分是这个产品的分泌物，即测度外壳——不能从上述任何一条脉络单独推出：通约化研究会说它是共同度量，脱耦理论会说它是仪式性外壳（这个词最接近，但它假设外壳是给外部环境看的），波特会说它是抵御质疑的盾牌，埃尔斯特会说它是切断争辩的规则。这四个说法各自解释了外壳的一个侧面，没有一个解释它为什么必须看起来像测量而不是像抽签、像资历、像神谕。

答案在第六节第三小节。

三、可接受的拒绝、测度外壳与再分泌律

（一）核心概念

概念一：可接受的拒绝（Acceptable Rejection）

本文将“可接受的拒绝”操作化定义为：一种排除方式，它使被排除者在被排除之后，仍然继续参与产生该排除的社会合作体系，而不采取以下三类行动之一——诉诸暴力、退出体系、指认一个具体的施害者。

这一定义有三个要件，缺一不可。

要件一：被排除者继续参与。 一次拒绝如果导致被拒者永久离场，它在社会意义上是失败的——不是对被拒者失败，是对体系失败，因为体系需要落选者继续在其中劳动、纳税、供养下一代考生。可接受的拒绝的成功标志是：被拒者第二年再来，或者他的孩子再来。赫希曼（Hirschman, 1970）

关于退出、呼吁与忠诚的经典三分法为此提供了精确的坐标：可接受的拒绝所要生产的，正是在遭受损失之后仍然选择忠诚而非退出、且不诉诸呼吁的那种状态。赫希曼关心的是组织如何因缺乏反馈而衰败；本文关心的是一类装置如何专门生产这种无呼吁的忠诚——在本文的框架中，那不是衰败的症状，是产品。

要件二：不指认具体施害者。 这是最关键的要件，也是理解全部现象的钥匙。当一个人被拒绝时，他会寻找责任的落点。如果落点是一个具体的人（考官、校长、招生办主任），拒绝就转化为怨恨，怨恨积累为冲突。可接受的拒绝要求责任的落点是非人格的——不是任何人拒绝了你，是你的分数拒绝了你。**分数的社会功能，首先是吸收怨恨。**

要件三：可接受性由被拒者承担，不由拒绝者承担。 这是本文与所有批判理论的分岔点。拒绝者（大学、雇主、国家）需要的只是决定；他们不需要被拒者接受。是被拒者自己需要接受——因为不接受的代价由他自己承担：一个认为自己被不公正地毁掉了的人，无法组织自己剩下的人生。接受不是被灌输的，是被拒者为了继续活下去而自己生产的。

由要件三推出一个反直觉但可检验的命题：对“分数是公平的”这一命题的信念强度，应与被拒风险负相关，而非与既得利益正相关。最需要相信考试公平的，是那些最可能考不好的人。这一命题同时是本文与批判再生产路径的判决性分岔。

概念二：测度外壳（Measurement Shell）

本文将“测度外壳”操作化定义为：为使拒绝可接受而分泌的、借用测量语法但不承担测量义务的表层结构。

外壳有三个结构特征。

特征一：借用测量的语法。 外壳必须以分数、等级、排名、指标等形式呈现，必须有命题、评分、统分的完整流程，必须能被表述为“某人在某个维度上的位置”。它必须看起来像一次测量。

特征二：不承担测量义务。 外壳不需要与它声称测量的对象保持稳定关联。它的效度可以长期为零，且这一事实可以被广泛知晓，而不影响它的运作。

特征三：产品是不可争辩性，不是准确性。 外壳所生产的是“这个结果无法被争辩”这一状态。它之所以必须借用测量的语法（特征一），正是因为测量的语法自带不可争辩性——一个数字不能被反驳，只能被重新计算；而重新计算需要挑战整套程序，其成本远高于任何个体所能承担。

特征三解释了特征一与特征二为何可以共存：如果外壳的产品是准确性，那么特征二是它的失败；如果外壳的产品是不可争辩性，那么特征二根本不是它的性质之一——准确与否与它无关。

一个必须被区分开的日常混淆。 被拒者说的是“至少高考是公平的”，不是“高考是准确的”。这两句话在既有文献中被系统性地混用，它们的差别却是整个理论的枢纽。“准确”是关于装置与世界的关系（效度问题）；“公平”是关于装置与被拒者的关系（可接受性问题）。一个装置可以完全不准确而完全公平——只要它对所有人施加同一套非人格的程序。这正是被拒者所要求的东西。他们从未要求过准确。他们要求的是“没有人在暗中动手脚”，即：没有可指认的施害者。

概念三：外壳分泌 (Shell Secretion)

分泌关系不同于寄生关系、脱耦关系、共生关系，必须被精确区分。

- **寄生**：A 依附于 B 并消耗 B，B 先于 A 存在。若测度被安抚寄生，则测度曾独立存在过。本文否认这一点。
- **脱耦**：A 与 B 并存但不连接。若测度与安抚脱耦，则两者是并列的两张皮。本文否认这一点。
- **分泌**：B 为了自身的运作而产生出 A，A 从未独立存在过，且 A 的形态完全由 B 的功能需求决定。贝壳不是贝的寄生物，也不是与贝脱耦的另一样东西；贝壳是贝分泌出来包住自己的东西，它的形状由贝的需要决定，它从未有过自己的生活。

测度是安抚分泌出来的壳。 这句话的全部内容是：教育评价装置中的测量成分，其形态、精度、覆盖面、技术演进方向，全部由“使拒绝可接受”这一功能需求决定，从未由“认识学习”这一认知需求决定。第五、六节将逐条兑现这个断言。

概念四：外壳再分泌律 (The Law of Shell Re-secretion)

本文的核心律，其完整表述为：

任何承担稀缺分配功能的排序型评价装置，其测度外壳都将在其优化产业成熟的时刻与它声称测量的对象脱钩，而装置本身不会因此消亡。试图以新装置替代旧装置的努力，将使新装置在获得同一社会功能的同时分泌出同构的测度外壳。阈值型装置不受此律约束。

律的三个组成部分需要分别说明。

(a) 脱钩的必然性：不是因为人们作弊，也不是因为指标被优化目标反噬（这是坎贝尔定律 [Campbell, 1979] 与古德哈特定律的内容，本文承其洞见但主张它们只是表层机制），而是因为外壳从一开始就不承担测量义务。脱钩不需要机制来产生，它是默认状态；需要机制来解释的反而是耦合——为什么在某些时期外壳与被测对象保持了一定关联？答案见第六节第四小节：那是外壳借用测量语法所必须支付的最低保真成本，而耦合期的长度由优化产业的成熟速度决定（第七节第二小节）。

(b) 装置的不死性：装置不因脱钩而消亡，因为脱钩没有伤害它的产品。这是本文对信号理论、技术优化路径与制度惯性路径的共同回应：它们都预测装置会因失效而承压，只是对承压后果的判断不同。本文预测装置不承压。

(c) 再分泌：这是律的核心，也是它最锋利的地方。它的推理链是：安抚需求内生于被拒者 → 需求不因旧装置被废除而消失 → 新装置一旦承接稀缺分配功能，立即面对同一需求 → 新装置必须让被拒者接受被拒 → 因此必须提供非人格的、不可争辩的拒绝理由 → 因此必须分泌测度外壳 → 因此在其优化产业成熟时与它声称测量的对象脱钩。

这条链的每一步都不依赖任何人的意图。 没有人策划它。设计新装置的人真诚地想要一个更好的评价方式；他们分泌外壳，是因为他们的装置一旦承接了分配功能，就必须回答"凭什么是他不是我"，而这个问题只有一种社会上可行的答案格式。

（二）命题系统

P1（产品命题）：大规模教育评价装置的社会产品是可接受的拒绝，不是关于学习的信息。测度是这一产品的生产条件，不是产品本身。

P2（分泌命题）：测度功能是安抚功能为使拒绝可接受而分泌的准确性表象。其形态由可接受性的需求决定，不由认识的需求决定。外壳的产品是不可争辩性，不是准确性。

P3（脱钩常态命题）：外壳与被测对象的脱钩是稳态而非退化。人工智能所揭穿的不是脱钩，而是社会曾相信过耦合这一集体记忆。被摧毁的是记忆，不是效度。因此，一切以"恢复分数的效度"为目标的对策（监考、检测、回归纸笔、重设题型）都是范畴错误——它们试图修复一个从未存在过的东西。

P4（再分泌命题）：任何承接稀缺分配功能的排序型替代装置，将在其优化产业成熟时分泌出同构的测度外壳。替代不是替代，是宿主更换。

P5（分界命题）：排序型装置必然分泌外壳；阈值型装置不必。区别在于拒绝理由的语法：排序型的拒绝理由是相对的（你比别人差），因而永远可被争辩，因而需要外壳来切断争辩；阈值型的拒绝理由是绝对的（你没到线），可被被拒者自己验证，因而不需要外壳。

P6（时滞命题）：脱钩时滞是优化产业成熟速度的函数，不是历史常数。优化产业成熟的条件是：装置的度量规则稳定到可被逆向工程、优化外壳的回报率超过提高被测属性的回报率、且优化知识的传播成本低到足以使它从个体经验变成可售商品。当耦合期短于耦合记忆的形成时间，结果型外壳失去可归因性，分泌压力将转向另一类外壳。

P7（干预命题）：由 P4，一切在评价端的干预（改测量、换指标、建替代认证）都是同义反复——它们只更换宿主。唯一非同义反复的干预有两个落点：（i）语法落点：把排序型装置改造为阈值型装置；（ii）需求落点：降低单次拒绝的终身不可逆性，从而降低安抚需求的产生率，从而降低外壳的分泌压力。两者都不在教育评价内部。

命题之间的逻辑关系：P1 是地基；P2 说明地基如何产生上层建筑；P3 是 P1+P2 对当前危机的直接应用；P4 是 P1+P2 的时间展开，是本文的核心律；P5 是 P4 的边界条件，同时是它的可证伪面；P6 给出 P4 的时间参数并使 P3 产生一个新推论（外壳换代）；P7 是 P4+P5 的实践推论。

（三）本框架产出而既有框架产不出的推论

一个理论的份量不在它能解释多少已知现象，而在它能否产出竞争理论产不出的推论。本文列出五条，每一条都在后文被展开并给出经验落点。

推论一（信念的逆向分布）：对“评价是公平的”这一信念的强度，应与被拒风险正相关，而非与既得利益正相关。批判再生产路径预测相反。这是一个可以用现有调查数据直接检验的分歧。

推论二（评级机构的不死）：一个被公开证伪的排序装置，其制度地位不会下降，可能上升。技术优化路径、信号理论、制度惯性路径都预测下降或至少承压。

推论三（影子排序）：取消一个排序装置的显式度量后，若其分配功能未被取消，将立即出现自发的、由参与者自建的影子排序。脱耦理论预测取消后正式结构变为冗余而被放弃；批判路径预测取消后焦虑下降。两者都不预测影子排序的自发涌现。

推论四（阈值型装置的保真）：不承担排序功能的资格型装置（飞行执照、外科手术资格、职业焊工认证），其测度效度显著高于升学考试，且不随时间衰减。既有框架中没有任何一个预测这一分野——它们的机制（通约化、机械客观性、脱耦、路径依赖、坎贝尔定律）全部应当同等地作用于这两类装置。

推论五（时滞坍塌与外壳换代）：在人工智能条件下，新建立的排序型装置的脱钩时滞将坍塌到它的第一个周期之内，短于耦合记忆的形成时间。因此结果型外壳将失去可归因性，而分泌压力不会消失，装置将转向过程型外壳。当前被读为“防作弊”的一整套技术是外壳的换代而非修复。既有框架无一预测这一转向——它们把这些技术读为对效度的防御，而本文读为对可归因性的重建。

推论四与推论五是本文最重的两笔。前者把一条“什么都能解释”的犬儒判断（评价都是骗人的）转化为一条有边界、有反例、可被证伪的律；后者把这条律从一个关于过去的谱系转化为一个关于正在发生之事的、可以在数年内被检验的预测。

四、方法与它的边界

（一）研究设计

本文是一项概念分析与机制建构研究，以历史谱系追溯与跨域结构同构为双重校验工具，其产出是一条带边界条件的律及其可证伪的检验设计。本文不提供一手数据。

选择这一设计的理由是：本文所挑战的不是既有理论的某个结论，而是它们共享的前提（评价装置是测量装置）。挑战前提不能靠在前提内部收集更多数据——那样收集到的一切都会被前提消化。它只能靠两件事：（i）追溯前提的历史来源，看它在装置的诞生场景中是否成立；（ii）寻找前提在其他领域中的同构失效，看它是否是一个跨域的系统性错误。本文两者都做。

（二）历史谱系追溯与它的最坏情况

第五节对三个独立起源的大规模评价装置——中华帝国晚期的科举、19 世纪普鲁士—德意志的文官考试、20 世纪美国的标准化测验——的诞生场景进行结构化追溯。追溯的目标不是叙述制度史，而是回

答一个单一问题：在每一个装置被建立的时刻，建立者试图解决的问题是"我们不知道谁有能力"，还是"我们不知道如何让落选者接受落选"？

选择这三个案例的理由是它们的独立性：若三者的诞生场景呈现同一结构，这一结构就不能被归因于文化传播。

这个理由需要面对一个它自己的弱点，而不是把它放进括号里。

关于普鲁士文官考试是否受科举影响，存在一条经由耶稣会士报告与十八世纪欧洲汉学传播的论证线索，它在史学上有争议且未被定论。本文不判断这条线索的真伪，而是直接接受它的最坏情况：

假设这条线索成立。那么本文只剩两个独立案例：科举与美国的标准化测验。

这两者之间的独立性无争议：美国标准化测验的直接前身是一战期间的陆军 Alpha/Beta 团体测验，其技术谱系来自比奈—西蒙量表与心理测量学（Lemann, 1999），与科举没有任何传承关系。

两个独立案例是否够用？对于排除文化传播这一个竞争解释，够——两个没有共同祖先的案例呈现同一结构，文化传播就不能解释这个同一。对于建立普遍性，不够——但本文从未从这三个案例归纳普遍性。**普遍性由第七节的机制承担，历史案例只承担一件事：证明这个机制在装置的诞生时刻就在运作，而不是后来被附会上去的。**

因此本文明说最坏情况：本文的历史论证在最坏情况下站在两个案例上，且这不影响它承担的那一件事。普鲁士案例在此情况下降格为佐证而非论据——但它的内容仍然值得保留，因为韦伯对非人格性的分析在本文框架下的重读，不依赖于普鲁士与科举是否独立。

追溯依据的是这三个领域中已被广泛接受的二手史学研究（Elman, 2000; Miyazaki, 1976; Lemann, 1999; Karabel, 2005），本文不提供新的档案工作。这构成本文的一个明确局限。

（三）跨域结构同构

第七节第四小节将外壳再分泌律在五个教育之外的领域中兑现，第八节将其中一个（选举）展开为一次强制不一致。

跨域调用有被批评为"隐喻不肖"的风险，因此本文对每一次调用设定了统一的准入条件：**该领域必须存在一个承担稀缺分配或不可逆决定功能的装置，且该装置的测度效度已被独立研究证伪，且该装置未因此消亡。**只有同时满足这三条的领域才被纳入。不满足的领域被显式排除并说明理由。

这一准入条件的作用是使跨域地图可被证伪：如果有人能指出某个满足三条准入条件的领域，其装置在效度被证伪后确实消亡了，本文的律就受到实质挑战。

准入条件必须对本文自己同样严格。本文在第七节第四小节逐行标注每个领域是否严格满足第二条（"效度已被独立研究证伪"，而非"存疑"）；一个只满足"存疑"的领域不得入表。

（四）与本文的自我关系

本文的对象是一类装置，而本文自身正在被这类装置评价（同行评审、期刊分级、引用计数）。这不是一个可以被回避的修辞尴尬，而是本理论必须面对的最硬的检验：如果本理论正确，它必须能够解释它自己的命运。第十二节承担这一检验。这一检验的纪律是两条：其一，如果本文能够论证该律适用于一切装置却唯独不适用于评价本文的装置，那么该律就不是一条律，而是一个指向他人的标签，应被作废；其二，**自反检验必须指出至少一处本文自己分泌的、具体的外壳，否则它只是一个姿态**——一个只在总体上承认“本文也分泌”的自认，无法被反驳，因此也无法被兑现。

（五）方法局限

其一，历史追溯依赖二手研究。 本文对装置诞生场景的解读建立在既有史学成果之上，未做档案工作。一个持不同框架的史学家可能从同样的材料中读出不同的诞生动机。本文在第五节的处理是：只使用那些在各自领域内已被广泛接受、且不由本文的框架所引导的史学结论，在每处标注该结论的争议状态，并在存在竞争读法时退到一个在所有竞争读法下都成立的更弱命题。

其二，判决格只有一个实例，且是事后选取的。 第九节第二小节的 2×2 设计给出了一个能区分本文与波特的格子，并给出了该格子里的一个实例（法国 concours）。这不构成一次判决，构成一次演示——它演示了两个理论在这个格子里给出相反预测，并给出一个实例的方向。判决需要这个格子里的多个实例被系统测量。本文列出可以进入这个格子的候选（日本的东大入试、韩国的修能、印度的 IIT-JEE 与 UPSC、法国的 concours、英国牛剑入学在其考试主导时期），但本文没有测量它们。这是本文检验方案的一个明确缺口。

其三，“优化产业成熟”需要一个操作化判据。 第七节第二小节把脱钩时滞交给了这个变量，因此它必须可观察。本文给出的判据是：存在以优化该装置的输出为唯一目的的商业服务，且该服务的市场规模在若干年内稳定或增长。这个判据是可操作的，但它的信度未经检验，且它在优化知识不被商品化而被非市场机制扩散的场景中（如家族内部传承、单位内部经验共享）会失灵。

其四，阈值/排序的分界在现实中是连续谱。 现实装置很少是纯粹的排序型或纯粹的阈值型（医学院录取既有分数线又有排名）。本文在第七节第五小节给出了一个定性的判别形式，但它是定性的，且本文在同一处说明了为什么它拒绝写出一个更好看的定量形式。

其五，本文没有解释稀缺本身。 本文把稀缺当作外生变量。但一个更彻底的追问是：为什么现代社会必须把如此多的终身后果压在如此少的位置上？这个问题在本文的框架之外，且它可能才是根。

订正三·本条局限被注销

稀缺不是根，也不在框架之外。 可结清性与稀缺正交：德国的二元制有相当程度的稀缺而账可结清（职业资格由固定标准定义、可抵达、可兑现，且被市场独立定价）；而无名额、无录取线、无正式拒绝的自我优化场域，其“我本可以更好”完全不可结清——因为它的内容由一个持续更新的他人分布定义——毒性不低于高考。名额只是把那个“本可

以的我"锚定到他人位置上的一种方式，不是唯一的方式，也不是根。见[律篇第四节第二小节](#)。

五、产品倒置：三个装置的诞生场景

本节回答一个只需要历史材料就能回答、却几乎从未被提出的问题：大规模评价装置被建立的那一刻，建立者要解决的是什么问题？

（一）科举：一个以落第者为产品的装置

科举在隋唐之际的确立动机，在史学上并非单一读法。"皇权对付门阀士族的工具"是其中一种；另有把它读为察举制在人口流动与地方失控条件下的技术性替代，也有把它读为选官制度长程官僚化的一步。本文按第四节第五小节的纪律标注这一争议状态，并且不使用其中任何一种读法作为论据。

本文需要的是一个更弱、但在所有竞争读法下都成立的命题：

没有任何一种关于科举起源的史学读法，把它的诞生动机读为"我们不知道谁有才能"。

这个命题不需要在竞争读法之间选边。皇权对付门阀是一个分配问题；察举失效是一个分配程序的失效；选官官僚化是一个分配机制的形式化。三种读法在动机的内容上分歧，在动机的**类型**上一致：全部是分配，没有一个是认知。而本文的论证只需要类型，不需要内容。

门阀内部对谁有才能的判断相当清楚——争的从来不是判断的内容，是判断权的归属。这一点在三种读法下都成立。

说明问题的是科举在其成熟期的运作数据。艾尔曼（Elman，2000）对明清科举的系统研究给出了一组常被引用却极少被消化的数字：明清两代，通过乡试者约占应试者的百分之一到二，会试的录取比例更低；一个士人从童生到举人可能耗费三四十年，绝大多数人终身停在最低一级或根本不入级。这意味着，这个装置在其全部运行历史中，百分之九十九以上的产出是落第者。

按照"选拔人才"的功能定位，这是一个荒谬低效的装置——它耗费了整个社会几个世纪的智力资源，产出的绝大部分是失败。但如果把落第者本身视为产品，一切就顺了。艾尔曼指出，落第者并未离场：他们成为塾师、幕友、讼师、地方乡绅、宗族领袖，构成了帝国基层治理的实际承担者，并且继续把自己的儿子送进同一个装置。他们不但没有成为体系的敌人，反而成了体系最坚定的维护者与最有效的传播者——因为他们是唯一有能力向下一代解释"为什么这是值得的"的人。

一个装置若以选拔为目的，落第率是它的成本。若以生产可接受的拒绝为目的，落第者就是它的产品，而落第者继续参与就是它的成功指标。明清科举的运作数据只在第二种读法下是可理解的。

（二）糊名与誊录：一项针对"施害者"而非"误差"的技术

科举史上最能说明外壳性质的，是宋代确立的糊名（弥封）与誊录制度。糊名将考生姓名籍贯密封，誊录则由专人将试卷重抄一遍再交考官评阅，以防笔迹辨认。

这两项技术在评价技术史上通常被记为"提高公正性的进步"。但需要精确地问：它们提高的是什么？

它们不提高效率。糊名不会让一份八股文更准确地反映考生的治国能力——它对测量精度的贡献为零。誊录甚至可能降低精度（抄错、书法信息丢失）。它们所消除的，是考官偏袒的可能性，即可指认的施害者。

这是一个可以推广到整个评价技术史的观察。逐条检视升学评价技术的演进——统一命题、统一考场、统一评分细则、双盲阅卷、机读答题卡、多套试卷随机分发、成绩复核程序——每一项引入时的官方理由与社会诉求，几乎无一不是"这样能更准确地测量学生"，全部是"这样就没有人能做手脚"。**技术的演进方向是朝向不可争辩性，不是朝向准确性。**

这是本文 P2 的第一个直接经验兑现：外壳的形态由可接受性的需求塑造，不由认识的需求塑造。

（三）普鲁士文官考试：非人格性作为产品

普鲁士在 1806 年耶拿战败后的施泰因—哈登堡改革中系统建立文官考试制度，其场景同样是分配危机：贵族对官职的世袭垄断与新兴受教育市民阶层的参与要求之间的冲突，叠加战败后国家对行政效能的紧迫需求。

（按第四节第二小节，若普鲁士考试受科举影响的线索成立，本节降格为佐证。本节的内容不依赖该线索的真伪。）

韦伯（Weber, 1978）对这一制度的分析给出了本文所需的关键概念。他把官僚制的核心特征概括为 *sine ira et studio*——不带怨恨与偏爱地处理事务，并把"非人格性"（Unpersönlichkeit）列为官僚制合理性的首要来源。韦伯是在赞美这一特征的行政效率，但他所描述的机制，在本文的框架下是另一件事：**非人格性的首要功能不是效率，是让被处理者无法怨恨任何人。**

一个被贵族排斥的市民之子，知道自己被谁排斥了，他的怨恨有落点，落点会积累为政治力量。一个考试不及格的市民之子，不知道被谁排斥了——排斥他的是一个程序。他的怨恨没有落点。它只能返回自身，成为"我不够格"。

这不是一个副作用。这是这个装置在 1806 年的普鲁士被建立起来所要解决的、最紧迫的那个问题。

（四）关键实验：当测度威胁到分配需求时，谁让步？

前三小节论证了诞生场景，但一个批评者可以合理地反驳：诞生动机不决定后续功能，一个为分配危机而建的装置，完全可能在运行中发展出测量功能。这个反驳是有力的，它要求一个能够区分两者的实验。

历史提供了这个实验。卡拉贝尔（Karabel, 2005）对哈佛、耶鲁、普林斯顿录取史的研究记录了一个精确的自然实验。1920 年代，这三所大学的录取以学术考试为主。这一装置产生了一个它的设计者不想要的结果：犹太学生的比例迅速上升——因为他们在考试上表现更好。哈佛的犹太学生比例在 1920 年代初期达到百分之二十以上。

面对这一结果，这三所大学做出的选择被完整地记录在案：它们改掉了装置。1926 年前后，三校陆续引入"品格" (character)、"领导力"、"全面发展" (well-roundedness) 等无法被标准化考核的维度，引入校友面谈、推荐信、体育特长、家庭背景问询。这套新标准的公开理由是教育理念的升级——培养完整的人，而非书呆子。其实际效果与实际目的（卡拉贝尔的档案工作在这一点上是明确的、并已被广泛接受）是把犹太学生比例压回可接受的水平。

这个实验的结构是：**测度与分配需求发生冲突。谁让步？**

答案是测度让步。而且让步的方式极其说明问题：它没有被替换成一个更准确的测度，它被替换成了一个更模糊的测度——因为模糊性正是分配需求所需要的。

如果测量学习是这个装置的目的，1926 年的常春藤应当欢呼——它们的测量工具正在成功地识别出最有能力的学生，只不过这些学生恰好是犹太人。它们没有欢呼。它们改掉了工具。

一次改动可以是偶然。但这个结构在评价史上反复出现：智力测验因为"测量了无法通过努力改变的东西"而被逐出升学录取；SAT 从 Scholastic Aptitude Test 一路更名，最终在 1990 年代彻底放弃 aptitude 一词，并在 2016 年的改革中明确宣布向高中课程内容对齐——即，向可以通过努力提高的东西对齐。

这一条演进线的方向，用"提高测量准确性"完全无法解释：如果目标是准确地测量学生的认知能力，从能力倾向转向课程掌握度是一次纯粹的倒退，因为后者混淆了能力与教学质量、家庭资源、备考投入。用本文的框架，它是一次必然：**外壳必须测量可归因于被拒者自身的东西，否则拒绝不可接受。**"你天生不行"是一个无法被接受的拒绝理由——它不满足要件一（被拒者会离场）也不满足要件三（他无法用它组织自己剩下的人生）。"你不够努力"则是一个完美的拒绝理由：它把责任交还给被拒者，从而完成了怨恨的吸收。

外壳向可努力性靠拢，不向准确性靠拢。 这是 P2 最锋利的经验兑现，也是既有框架中没有任何一个能够产出的推论。

六、外壳为什么不必准确，又为什么仍需一点准确

（一）不可争辩性的生产

外壳的产品是不可争辩性。这需要通过机制化地说明，而不是断言。

一个拒绝要成为可接受的，被拒者必须无法组织出一个有效的反驳。反驳的组织需要一个对象——某个可以被指认为“错了”或“不公”的东西。外壳通过三重结构消除了这个对象：

第一重，非人格化。 拒绝你的不是人，是分数。你无法怨恨一个数字。

第二重，程序化。 分数不是被判定的，是被计算的。计算意味着，任何人处在评分者的位置上都会得出同一个结果。这切断了“换个人评就不一样”这条反驳路径。

第三重，通约化。 你与其他人被放在同一个标量上。这切断了最后一条反驳路径——“我和他不一样，不能这么比”。通约化的社会功能，正是让这句话变得不可说。埃斯佩兰和史蒂文斯（Espeland & Stevens, 1998）把这称为信息的丢弃；从本文的角度，被丢弃的信息恰恰是争辩的材料。**丢弃不是通约化的成本，是它的产品。**

三重结构合起来，产生了一个被拒者无法着力的表面。他伸手去抓，抓到的只有一个数字、一套程序、一个刻度。这就是不可争辩性。

请注意：这三重结构中，没有一重需要准确性。非人格化不需要，程序化不需要，通约化更不需要——它只需要一个刻度存在，不需要这个刻度测量了任何真实的东西。

（二）效度与不可争辩性的系统性冲突

外壳不仅不需要准确，它与准确之间存在结构性冲突。

考虑那些能够提高效率的评价方式：长期的过程性观察、多元证据的三角互证、深度的个别面谈、了解学生的导师做出的整体判断。教育测量学在信度效度的技术意义上并不否认这些方式能够更接近“这个人到底怎么样”。

这些方式在升学评价史上被系统性地淘汰。淘汰的理由从来不是“它们测不准”。理由永远是：可以被操纵，不透明，依赖判断。

翻译成本文的语言：**它们无法生产不可争辩性。** 一个导师的整体判断，无论多么准确，都留下了一个可指认的对象——那个导师。被拒者可以说“他对我有偏见”。这句话一旦可说，拒绝就不再可接受。

因此，评价史上反复上演的是这样一场交易：用效度换取不可争辩性。而这场交易的方向是单向的——从来没有一次改革是用不可争辩性换取效度。

这解释了一个既有文献长期无法解释的现象：为什么一个多世纪的测量技术改进，没有带来公众信任的上升？因为技术改进的方向从来不是准确性。它是不可争辩性。而不可争辩性的提高，恰恰以效度的下降为代价。信任没有上升，是因为改进的东西不是被信任所需要的东西——但这一点从来不重要，因为运作所需要的从来不是信任，是不可争辩。

（三）为什么不干脆抽签

如果外壳不需要准确，为什么社会不采用一个更彻底、更便宜、也更诚实的方案——抽签？

埃尔斯特（Elster, 1989）证明了抽签在许多分配场景中被实际采用，且被接受。抽签生产的不可争辩性是完美的：它没有人格，没有判断，没有可指认的对象，且成本近乎为零。按前述逻辑，抽签应该是最优解。

它不是。这个反例必须被解释，否则本文的机制是不完整的。

答案在于两类可接受性的差别。**抽签生产的是可接受的运气；现代社会需要的是可接受的应得。**

这个差别不是道德偏好，是结构需求。一个基于运气的分配可以被单次接受（我这次运气不好），但它无法支撑一个跨代的社会叙事。如果机会分配是抽签，那么：家长无法向孩子解释为什么要读书；学校无法解释自己为何存在；社会无法要求任何人努力；成功者无法解释自己凭什么占据位置。整个合作体系的动员机制会失去着力点。

这一点在关于绩效社会的经典论述中已被从另一个方向触及，只是从未被连到评价装置的机制上。杨格（Young, 1958）在发明“绩效主义”（meritocracy）这个词时是把它当作反乌托邦讽刺来写的：他的虚构社会之所以比世袭社会更残酷，正是因为它成功地让失败者相信失败是自己应得的——世袭社会的底层可以恨贵族，绩效社会的底层只能恨自己。桑德尔（Sandel, 2020）在半个世纪后把这一诊断推到当代：绩效叙事的暴政不在于它分配不公，而在于它剥夺了失败者对自己失败的任何外部解释。

杨格与桑德尔都把这看作绩效社会的代价——一个我们应当设法减轻的道德损害。本文的读法相反：**那不是代价，那是产品。**一个让失败者恨自己的装置，正是一个成功地生产了可接受的拒绝的装置；杨格所描述的那种残酷，是这个装置正常工作的标志，不是它的故障。

因此，装置必须生产一种拒绝，它同时满足两个要求：**不可争辩**（否则怨恨有落点）**且可归因于被拒者自身**（否则努力失去意义）。

抽签满足第一条，违反第二条。导师判断满足第二条（他判断的是你），违反第一条。

只有一种东西同时满足这两条：**一个非人格的、程序化的、通约化的、且测量某种被拒者自己控制的属性的装置。**这就是测度外壳的完整规格。它必须借用测量的语法，不是因为社会想知道什么，而是因为这是唯一能同时满足两个约束的形式。

订正一·规格是三条，不是两条

这一段是本文最不可替代的一笔，而它不完整：**它推不出外壳**。飞行执照考试非人格、程序化、通约化、且测量某种被拒者自己控制的属性——它满足上面两条约束的每一个字，而它不分泌外壳，其测度效度高且不衰减。一个满足全部规格却不产生规格所预测的后果的装置，是对规格的直接证伪。缺的第三条是**不可结清**：飞行执照的拒绝可结清（去学、学会、过、离场），高考的不可结清。前两条是交付条件，第三条是持有条件，而**分泌压力全部来自第三条**。这也意味着本文第七节第六小节的排序/阈值分界不是这条规格的"边界条件"——它就是规格的第三条。推导与四装置判决表见**律篇第三节**。

这是本文对"为什么必须是测量"这个问题的完整回答——第二节第六小节指出，既有四条脉络没有一条能回答它。

（四）最低保真成本：外壳的唯一保真压力

规格的最后一条（测量某种被拒者自己控制的属性）给外壳施加了它唯一的保真压力，本文称之为最低保真成本。

外壳必须与某种真实的个人差异保持最低限度的相关。如果分数与考生做过的任何事情都完全无关（比如按身份证尾号排序），表象立刻崩溃，可归因性消失，拒绝变回运气。所以外壳必须测量点什么。

但这个"点什么"的要求极低，而且它的方向不是效度。它只要求：**存在一个被拒者可以事后认领为"我没做够"的东西**。

刷题量满足这个要求。记忆强度满足。应试技巧满足。它们全都与"学习"这个装置声称测量的东西关系稀薄，但它们全都完美地满足最低保真成本——因为它们是被拒者自己控制的、可以事后认领的。

这就是脱钩的完整机制，也是 P3 的论证：脱钩不需要坎贝尔定律来解释。坎贝尔定律（Campbell, 1979）说的是"当一个量化指标被用于社会决策，它就会承受腐蚀压力并扭曲它所监测的过程"——这个机制成立，但它是表层的：它描述了外壳被使用后如何腐蚀，它假设外壳一开始是准的。本文的机制在它之下：外壳从一开始就只需要满足最低保真成本，而不是效度。坎贝尔定律解释的是外壳如何从"有点准"滑到"完全不准"；本文解释的是为什么这个滑动不构成问题——因为最低保真成本在滑动全程中都被满足着。刷题量始终是被拒者自己控制的。

由此得到 P3 的完整形式：**脱钩是稳态。需要解释的不是脱钩，是耦合**。而耦合只是最低保真成本的副产品——在装置刚建立的那段时间里，还没有人知道如何专门优化外壳，因此外壳与被测对象碰巧保持了一定相关。这个相关不是装置的成就，是它的尚未成熟。

"那段时间"有多长？ 本文在此前的表述中把它归纳为"一到两代人"，并把这个数字标为最脆弱的一处，因为它来自案例归纳而非机制推导。第七节第二小节收回这个归纳，把它换成一个机制——而这个机制不是新引入的，它就在上一段里。

七、外壳再分泌律：一条会禁止事情的律

（一）律的推理链

律的核心推理链已在第三节给出，此处兑现其每一环：

安抚需求内生于被拒者（要件三）→ 该需求不因旧装置被废除而消失，因为它的来源不是装置，是稀缺 → 新装置一旦承接稀缺分配功能，立即面对同一需求 → 新装置必须让被拒者接受被拒 → 因此必须提供非人格的、不可争辩的、且可归因于被拒者自身的拒绝理由（第六节第三小节的完整规格）→ 因此必须分泌测度外壳 → 由最低保真成本机制（第六节第四小节），外壳在优化产业成熟后与它声称测量的对象脱钩。

链条的每一步都不依赖任何人的意图。这是律的关键性质，也是它区别于一切阴谋论式解释的地方。设计新装置的人真诚地想要一个更好的评价方式。他们分泌外壳，是因为他们的装置一旦承接了分配功能，就必须回答"凭什么是他不是我"，而这个问题在现代社会只有一种可行的答案格式。

（二）脱钩时滞：从归纳到机制

链条的最后一环留下了一个空位：**"优化产业成熟后"是多久？**

一条不给时间窗的"必然"不可证伪——它可以被无限期地推迟。因此这个空位不是一个细节，是律的可证伪性所在。

本文此前把这个空位填为"一到两代人"，并诚实地标注它是三个案例的粗略归纳（科举在八股定型后的形式化约两代人；SAT 从 1926 年设立到 1980 年代应试产业成熟约两代人；中国高考从 1977 年恢复到应试体系完全成熟约一代半）。一个从案例归纳出来的数字不是机制，它是一个装饰在律上面的常数。而一个装饰性的常数，恰好是本文自己所定义的那种东西。

本节收回这个归纳。收回的方式不是引入新东西，是认领一个已经在第六节第四小节里的机制。

第六节第四小节说：耦合"只是最低保真成本的副产品——在装置刚建立的那段时间里，还没有人知道如何专门优化外壳"。这句话已经把机制说完了。耦合期不是一个历史常数，它是**优化知识从不存在到成熟所需的时间**。脱钩不是慢慢发生的，它发生在优化产业成熟的那一刻——因为在那一刻之后，理性的被拒者会把全部努力从"提高被测属性"转向"优化外壳"，而外壳与被测属性的相关随之归零。

因此：

脱钩时滞 T = 优化产业从零到成熟所需的时间。

优化产业成熟需要三个条件同时满足：

条件 a（规则可逆向工程）：装置的度量规则必须稳定到足以被反推。规则每变一次，已积累的优化知识贬值一次。因此规则的变更频率决定优化知识的折旧率——一个每年重写全部规则的装置，其优化产业永远不成熟，代价是它自己也失去了不可争辩性（规则不稳定 = 程序化被破坏 = 三重结构缺一）。这个代价解释了为什么没有装置采用这条防线。

条件 b（优化回报率超过学习回报率）：把一单位努力投入优化外壳的收益，必须超过投入提高被测属性的收益。这个不等式一旦成立，理性的被拒者就全部转向优化外壳。**注意这一步不需要任何人作弊，也不需要任何人犬儒。它只需要理性。**

条件 c（优化知识的传播成本足够低）：优化知识必须能从个体经验变成可复制、可出售的商品。一个只存在于个别塾师脑子里的诀窍不构成产业；一个能被印成书、开成课、写成脚本、打包成服务的诀窍才构成。

在条件 b 满足的前提下， T 由条件 a 与条件 c 共同决定：

$T \approx$ 规则稳定化所需的时间 + 优化知识由传播成本决定的扩散时间

三个历史案例在这个式子下不再是三个巧合，而是同一个机制在三种传播成本下的三次读数：

装置	规则稳定化	优化知识的传播成本	观察到的 T
科举	八股定型（明中期）	高：师徒口授、抄本、闱墨选集	约两代人
SAT	快（题型早期固定）	中：印刷业 + 面授补习机构	约两代人
中国高考	快（1977 后迅速定型）	中偏低：全国教辅出版 + 补习产业	约一代半
平台声誉信号	极快（算法公开）	近乎零：脚本可分享、服务可购买	数年

第三小节的案例三（GitHub star 与 Kaggle 排名）此前被写为“时间尺度不等，取决于优化产业成熟的速度”——那是把机制当作了一个注意事项。它不是注意事项，它是变量。数年与两代人之间的差距，不是装置类型的差别，是传播成本的差别。

这个改写做了三件事。

第一，律不再依赖那个数字。“一到两代人”退位为一个在特定传播成本下的读数，不是律的内容。

第二，可证伪面变得干净。旧的证伪条件是"找到一个排序型装置，其外壳在两代人之后仍与被测对象保持稳定耦合"——这个条件可以被无限期地拖延（"再等等，还没到"）。新的证伪条件是：

找到一个排序型装置，其优化产业已经成熟（存在以优化该装置输出为唯一目的的商业服务，且该服务的市场规模稳定或增长），而其外壳仍与它声称测量的对象保持稳定耦合。

这个条件不能被拖延，因为"优化产业成熟"是一个可观察的、有日期的事件，不是一个可以再等等的期限。本文没有找到这样的装置；这不构成证明，只构成邀请。

第三，也是最重的一件：它产出了一个既有框架无法产出的推论。

（三）推论五：时滞坍缩与外壳的换代

生成式人工智能同时压低了条件 c 的两端：优化知识不再需要被某个人发现、写成书、卖出去；它可以被现场生成。同时它压低了条件 b 中"优化外壳"的成本，使不等式在更多场景中成立。

因此：

在人工智能条件下，任何新建立的排序型装置，其脱钩时滞将坍缩到该装置的第一个周期之内。

具体含义：一个 2026 年建立的新评价装置，不需要等待一到两代人。它将在它的第一届参与者毕业之前脱钩。

这句话给出了本文对"AI 冲击"的第二个诊断，它与第一个诊断（P3：被摧毁的是记忆，不是效度）咬合但独立：

人工智能没有揭穿外壳，也没有摧毁外壳。它做的是把再分泌的周期从两代人压缩到一个学年。

旧的节奏是：装置建立 → 耦合期（长到足以让一代人相信耦合）→ 脱钩 → 记忆残留 → 危机 → 建立新装置 → 新的耦合期。这个循环的每一步都依赖耦合期足够长，长到能生产出"它曾经是准的"这个集体记忆。而那个记忆，正是可归因性的燃料：被拒者之所以能把责任交还给自己，是因为他相信这个装置至少在原则上测量了某种他自己控制的属性。他不需要这个装置现在是准的；他需要相信它本来是准的。

当耦合期坍缩到一个学年，这个记忆来不及形成。**将不会再有任何人相信任何装置曾经是准的。**

而由要件三，安抚需求不会因此消失——它的来源是稀缺，不是记忆。于是分泌压力仍在，但结果型外壳已经无法承载它。

由此得到本文最重的一个推论：

推论五（外壳换代）：当脱钩时滞坍缩到耦合记忆的形成时间以下，结果型外壳失去可归因性，而分泌压力不消失。因此装置将转向另一类外壳——一类不依赖任何关于耦合的记忆、可以在单次交互中当场完成可归因性生产的外壳。

这类外壳必须满足第六节第三小节的完整规格：非人格、程序化、通约化、且测量某种被拒者当场可以认领的属性。它不能依赖"这个装置曾经是准的"。

只有一种东西满足：**过程的可见性**。

不是结果准确，是过程被看见。监考录像、答题过程录屏、写作过程的击键记录、版本历史、AI 使用日志、"你是怎么想出来的"的现场追问、答辩式的当场复述。

这正是当前正在大规模发生的事情，而它在既有框架下被普遍读为"防作弊技术"——即，读为对效度的防御。本文的读法相反：

它不是防作弊，它是外壳的换代。从结果的不可争辩，换到过程的不可争辩。

因为过程是被拒者当场可以认领的（"我确实没有自己想"），它不需要任何关于耦合的记忆。它的三重结构完整：非人格（拒绝你的是日志不是人）、程序化（同一份日志任何人看都是同一份）、通约化（过程被切成可比较的片段）。而它的最低保真成本被支付得极其便宜——它确实测量了某种真实的、被拒者自己控制的东西（他做了什么），只是这个东西与"学习"的关系，不比刷题量与"学习"的关系更近。

这个推论与本文自己的 P3 咬合，并修正了 P3 的一个不完整之处。P3 说"一切以恢复分数效度为目标的对策都是范畴错误——它们试图修复一个从未存在过的东西"。推论五说：**那些对策会成功，但不是以它们声称的方式成功。**它们不会恢复效度（效度从未存在）；它们会成功地重建可归因性。而由于成功的判据在社会上从来是可接受性而不是准确性，它们会被记录为一次成功的改革。

这个推论可以被证伪，且很快：如果在人工智能条件下，过程型装置没有取代结果型装置；或者它取代了但不满足三重结构（比如它保留了一个可指认的判断者）；或者结果型装置在时滞坍缩之后仍然维持了可归因性——推论五就受损。检验的窗口不是二十年，是三到五年。

（四）正在发生的再分泌：三个可追踪的案例

案例一：美国大学的 test-optional 转向。2020 年前后，因疫情与公平性争议，美国大量高校取消了 SAT/ACT 要求。这是一次彻底的装置废除——外壳被拿掉了。按照"外壳是压迫工具"的理解，其后果应当是阶层偏倚下降。

实际发生的是：录取的权重转移到了课外活动、个人陈述、推荐信、面试、"影响力叙事"。围绕这些维度，一个升学咨询产业在数年内成熟，其单价远高于 SAT 备考（SAT 备考可以靠一本题册与自

律，个人陈述的包装不能）。多项研究指出，新组合的阶层偏倚高于它所替代的标准化考试——因为标准化考试是三重结构最完整的外壳（非人格、程序化、通约化），而叙事型材料的三重结构全部残缺，从而给资源留下了更大的作用空间。

请注意这个案例的结构：改革的意图是消除外壳的不公，结果是分泌了一层更不公的外壳。这不是执行不力，是律。也请注意它的时滞：数年，不是两代人——因为叙事型材料的优化知识传播成本极低（一个升学顾问的诀窍可以在一个季度里被复制到全行业）。

案例二：中国的综合素质评价。 作为降低"唯分数"的政策工具引入。其后果是可观察的：素质档案的填写迅速标准化，出现了模板、代写服务与量化折算表（志愿服务多少小时折多少分、竞赛奖项如何换算）。**新装置在数年内长出了分数。** 它必须长出分数——因为不长出分数，它就无法回答"凭什么是他不是我"。时滞同样是数年而非两代人，原因同上。

案例三：开放平台的声誉信号。 GitHub 的 star 数、Kaggle 排名、开源贡献记录，曾被广泛论证为"不依赖学历的真实能力证据"。它们一旦被雇主用于筛选（即承接稀缺分配功能），刷星服务、贡献注水、榜单优化立即出现。这一装置的时滞是三个案例中最短的，因为它的规则公开（条件 a 瞬时满足）且优化知识可以被写成脚本（条件 c 的成本近乎零）。

三个案例的共同结构：承接分配功能 → 分泌外壳 → 脱钩。 而三个案例的时滞全部是数年而非两代人——这不是因为这些装置比科举脆弱，是因为它们的优化知识的传播成本比明清低了几个数量级。这三个案例因此不只是律的兑现，它们是第二小节那个式子的三次读数。

（五）跨域兑现：满足准入条件的领域

按第四节第三小节的准入条件（存在承担稀缺分配或不可逆决定功能的装置 + 该装置的测度效度已被独立研究证伪 + 该装置未因此消亡），以下四个领域被纳入本表。选举因为需要展开为一次强制不一致，单列为第八节；同行评审因为本文自身在其中，留待第十二节。

领域	装置	效度证伪（准入条件二）	是否消亡	律的兑现
金融	信用评级	强 ：2008 年结构化产品评级的大规模违约	否，制度地位上升	生产的是可接受的资本拒绝：被拒的融资方无法指认施害者，只能指认自己的评级
法律	正当程序	中 ：冤案率研究表明程序不保证事实准确	否，程序要求持续加严	生产的是可接受的判决：程序的加严方向是消除"法官偏见"而非提高事实发现率
组织管理	绩效考评	弱 ：绩效评级与客观产出的相关在元分析文献中偏低	否	生产的是可接受的晋升与解雇：被解雇者面对的是评级，不是主管
医学	精神障碍诊断分类	强 ：见下	否，形式化程度反而提高	生产的是可接受的拒绝：被拒者是痛苦真实但拿不到标签的人

逐行标注证伪强度，是准入条件对本文自己的执行。 组织管理一行是本表最弱的一行：本文没有审计那批元分析，"相关偏低"距离"效度已被独立研究证伪"有距离。它入表是勉强的，而地图不依赖它——删掉它，本节的结构不变。本文保留它并标注它的弱，而不是把它写得和金融一行一样硬。

医学一行需要单独说明，因为它的稀缺物不是位置，是承认。

诊断标签分配什么？病假额度、保险赔付资格、伤残认定、雇主的调整义务、家人的耐心、以及"你的痛苦是真实的"这一社会承认。这些全部有限，且全部被标签调节。

被拒者是谁？那些痛苦真实但拿不到标签的人——慢性疲劳、纤维肌痛、长新冠的早期患者，以及每一个被告知"检查结果都正常"的人。

拒绝的三个要件在这里全部兑现，且兑现得比教育更干净：

- **要件一（继续参与）**：被拒者不离场。他继续看病、继续找下一个医生、继续等待下一版诊断手册。
- **要件二（不指认施害者）**：不是任何医生拒绝了他，是标准拒绝了他。医生本人常常是同情的——这恰恰是外壳工作的证据：非人格性把医生从施害者的位置上摘了出来。
- **要件三（责任交还给自己）**：被拒者生产的接受是"也许我只是没那么严重"、"也许是我心理上的问题"。责任回到了他自己身上。

证伪强度为什么是强的。 DSM-5 的田野试验 (Regier et al., 2013) 公布了若干主要诊断的评定者间信度：重度抑郁障碍与广泛性焦虑障碍的 kappa 值落在 0.2–0.3 区间。需要一步论证才能把它接到效度上：评定者间信度是构念效度的上限——两位受过训练的精神科医生无法一致做出的诊断，不可能稳定地对应任何构念。因此田野试验的 kappa 是对构念效度的一个上限约束，而这个上限对若干主要诊断低到该学科内部无法辩护。这不是外部批评，是该装置自己公布的数据。更硬的一笔来自制度层面：美国国家精神卫生研究所时任所长在 DSM-5 出版前夕公开声明该手册缺乏效度，并宣布该所的研究资助将不再以其分类为依据 (Insel, 2013)。**装置的效度被该领域最有权威的资助机构公开否定，而装置在同一年出版，并继续担任临床与保险的标准。**

律的兑现，以及这一行为什么是全表最强的一行。

它给出了要件三最硬的一个经验兑现，硬于教育：**被拒者主动要求进入外壳。**

患者社群为某个诊断进入 DSM/ICD 而进行的持续动员——慢性疲劳综合征、纤维肌痛、长新冠——其诉求从来不是"取消诊断分类"，是"把我纳入分类"。他们要求的不是废除装置，是被装置接纳。

这与落选者游行要求"考试要更严格更公平"而不是"取消考试"（第二节第四小节对科林斯的分岔）是同一个结构，且它在医学场景中更不可能被误读为符号暴力的内化：一个要求自己的痛苦被编码的人，不是在内化压迫者的标准。他是在为自己生产一个可以用来组织剩下的人生的东西。批判路径读不了这个现象，它只能把它读为更深的误识。本文的读法是直接的：**外壳是被拒者需要的东西，因此被拒者会为进入外壳而斗争。**

而诊断手册在信度被自己的田野试验证伪、效度被自己领域的资助机构否定之后，形式化程度反而提高——更多的说明符、更细的严重度分级、更结构化的访谈工具。这些改进的方向不是提高构念效度，是提高不可争辩性：一个用结构化访谈工具做出的诊断，比一个凭临床印象做出的诊断更难被争辩，而它未必更准。

四个领域的共同结构：装置的效度被证伪，装置未消亡，且证伪之后装置的形式化程度反而提高。最后一点是律最反直觉、也最容易被检验的推论——它预测的是与所有既有框架相反的方向。

为什么某些常被类比的领域不在表中。 生态防火、核威慑、金本位这类场景常在制度分析中被用作类比，本文将它们排除，因为它们不满足第一条准入条件：它们不生产"可接受的拒绝"，没有一个被排除的个体需要接受一个关于自己的判断。排除它们不是保守，是使地图可被证伪：一个由"什么都像"的类比构成的地图不能被任何证据推翻。

（六）边界条件：排序型与阈值型的分界

律的可证伪性依赖于它有一个明确的边界。本文主张这个边界是拒绝理由的语法。

排序型装置的拒绝理由是相对的：你出局，因为有 N 个人排在你前面。这个理由永远可被争辩，因为它包含一个比较，而任何比较都可以被质疑（"凭什么这么比"）。因此排序型装置必须分泌外壳来切断争辩。

阈值型装置的拒绝理由是绝对的：你出局，因为你没达到标准。这个理由可被被拒者自己验证。他不需要相信任何人，他自己知道他不会开飞机。因此阈值型装置不需要外壳。

决定语法的不是装置的意图，是名额。 一旦存在名额上限，装置必须排序，无论它自称在做什么。

这里必须精确说明本文能支撑什么、不能支撑什么，因为这一处是本文最容易分泌自己的外壳的地方。

本文的机制支撑三个定性判断：

- (i) 分泌压力是名额约束紧度的递增函数；
- (ii) 分泌压力是拒绝不可逆性的递增函数；
- (iii) 名额约束紧度为零时（无名额，达标即过），分泌压力为零。

本文的机制不支撑任何具体的函数形式。写成

分泌压力 $\propto$ 名额紧度 $\times$ 不可逆性

看起来更硬。但那个乘号是从哪来的？不是从机制来的。乘法给出了 (iii)——任一因子为零则乘积为零——而 (iii) 恰好是本文要的结论。于是函数形式是被结论选中的，不是被推导出来的。加法给不出 (iii)，所以加法被排除；但 $\min(\cdot)$ 也给得出 (iii)，而本文没有任何理由在乘法与 $\min(\cdot)$ 之间做

选择。除此之外，那个乘号还偷运了两样本文完全没有论证的东西：两个因子可以被通约到同一个标量上，以及它们的效应是线性可乘的。

因此本文在这里只写它能写的：

分泌压力 = Φ (名额紧度, 拒绝不可逆性), 其中 Φ 对两个自变量都严格递增, 且 $\Phi(0, \cdot) = 0$ 。

这个式子比乘积式弱得多，看起来也不像一个发现。它是本文在这一处能诚实支撑的全部。

那个更好看的乘积式，本文把它留在了第十二节——作为一个标本。

（七）判决性案例：USMLE Step 1 的改制与压力转移

这一分界不是思辨的产物。它有一个近乎实验室条件的现实案例，且这个案例在本文提出之前就已完整发生。

美国医师执照考试（USMLE）Step 1 在设计上是一个纯粹的阈值型装置：它的功能是判定考生是否具备行医所需的基础医学知识，达标即过，无名额约束。按本文的律，它不应分泌外壳，其测度应长期保真。在其历史的大部分时间里，情况正是如此。

然后发生了一件事：住院医培训项目开始把 Step 1 的三位数分数用作申请筛选的排序工具。这一挪用没有改变考试本身的任何技术属性——同一套题库、同一套评分、同一个通过线。它只改变了一件事：这个装置开始承接稀缺分配功能。

本文的律对此有一个明确的预测：分泌压力立即出现，外壳开始形成，测度开始与被测对象脱钩。

实际发生的完全是这个：一个被称作 "Step 1 mania" 的现象出现了——医学生把大量时间从临床学习转向应试训练，围绕 Step 1 的商业备考产业（题库、闪卡、押题课程）迅速成熟，考试分数与临床能力的关联受到医学教育界的持续质疑。2020 年，USMLE 宣布 Step 1 改为通过/不通过（pass/fail），并于 2022 年 1 月实施。改制的官方理由被明确表述为：降低分数被用作排序筛选所带来的扭曲。

这是本文的律在现实中被一个完全不知道本文存在的专业共同体所独立执行的一次干预：切断排序挪用，解除分泌压力。

这个案例的价值在它的后续。 改制之后，住院医申请的排序需求没有消失——名额还在那里。压力立即转移到了 Step 2 CK：这门原本地位次要的考试，其分数在改制后迅速成为新的核心筛选指标，围绕它的应试产业开始成熟。

这就是再分泌，而且是在两年内、在一个可以被完整观察的封闭系统里发生的再分泌。**需求没有被消灭，它换了个宿主。**

这个案例同时兑现了 P4 与 P5，且它给出了 P7 的具体形式（见第十节）。它也再次读出了第二小节的机制：Step 2 CK 的时滞是两年而非两代人，因为它的优化产业不需要从零建起——Step 1 的商业备考产业只需要把已有的题库、闪卡、课程平移到另一门考试上。**优化基础设施的可移植性，是传播成本的一个特例。**

本文认为这个案例是这条律目前最硬的经验锚点，也是最容易被用来推翻本文的地方——如果对 USMLE 改制后的追踪数据显示 Step 2 CK 并未成为新的排序装置，或其应试产业并未成熟，本文的再分泌命题就直接受损。

八、选举：律在比较政治学中的强制不一致

跨域兑现如果只是在别的领域里找到相似的结构，它就还是一张类比地图，而类比地图不禁止任何事情。本节把该律在一个学科中推到底，看它是否能强制那个学科内部不一致。选择比较政治学，是因为它是所有候选领域中把自己的核心分类法押在一个装置上押得最重的一个。

（一）该学科的核心分类法及其判据

比较政治学以选举为轴心组织自己的核心分类。达尔（Dahl, 1971）把多头政体定义为竞争性（contestation）与包容性（inclusiveness）的组合，二者都通过选举兑现。此后关于混合政体的整个文献——选举型威权（Schedler, 2002）、没有民主的选举（Diamond, 2002）、竞争性威权（Levitsky & Way, 2010）——都以同一个判据划界：**选举是否产生真实的不确定性**。普沃斯基（Przeworski, 1991）把这一判据凝缩到极致：民主是一个结果不确定的制度。

这个判据同时承担两件事：它是分类的依据（哪些政体是民主的），也是测量的对象（Polity、V-Dem、Freedom House 的核心指标全部围绕它构造）。在这个学科里这两件事是同一件事：分类依据是可测量的，因为它对应一个潜变量；潜变量有一个质变点，那个质变点就是民主化。

（二）本文的判断与它强制的不一致

由 P1，选举首先是产生可接受的败选的装置。它的社会产品是：败选者接受结果、准备下一轮、不诉诸暴力、不指认施害者。**这正是选举与政变的全部区别——政变的失败者知道是谁干的。**

由 P2，选举的"真实不确定性"是外壳的最低保真成本。它必须存在某种真实的、参与者自己控制的差异（得票），否则可归因性消失，败选变回运气，接受立刻崩塌。但它只需要满足最低保真成本，不需要达成"选出更善治理者"这一声称的目标——而关于选民信息水平与议题投票的经典研究（Converse, 1964）以及对回顾性投票的系统检讨（Achen & Bartels, 2016）已经把这一目标的达成度压到了很低的位置。按照"测量装置失灵则被更换"的逻辑，这一文献应当动摇选举的制度地位。它没有。这是第一节那个不可解释的事实在政治学中的重演。

接受这两条，比较政治学面临一个它无法在内部消化的不一致。它同时持有：

- (a) 选举的功能是产生真实选择；偏离此功能的选举是威权的。
- (b) 民主指数测量偏离此功能的程度。

如果本文正确，(a) 是错的：功能是产生可接受的败选，真实选择是外壳保真度上的一个量。那么 (b) 测量的就不是它声称测量的东西——它测量的是外壳保真度。而 (b) 是 (a) 的唯一经验支撑：这个学科对 (a) 的信心，来自 (b) 的指标在跨国面板上运转良好。

于是它无法在放弃 (a) 的同时保留 (b)，也无法在保留 (b) 的同时解释为什么 (a) 从未被独立于 (b) 检验过。

更精确地：**民主与威权的分界，在本文的框架下不是种类差别，是外壳保真度这个连续量上的一个位置。**而这个学科整个方法论——二分类的政体转型研究、民主化的门槛、“民主倒退”的诊断——建立在那里有一个质变点这一假设上。本文说：没有质变点，只有保真度。

(三) 递归一击

上一小节的不一致还可以被读为“本文与该学科的分歧”，而一个学科可以拒绝一个外部判断。它尚未强制不一致。

强制来自下一步。

民主指数本身是排序型装置。它们承担稀缺分配功能吗？是的，且被分配的稀缺物是可枚举的：国际合法性、发展援助的资格条件、主权信用评级中的政治风险项、外国直接投资的国别评估、多边组织中的席位与话语、以及一个国家在国际交往中被对待的方式。这些全部有限，且全部被这些指数的排名调节。

按本文的律，它们必然分泌外壳，且在其优化产业成熟后脱钩。本文由此给出一个具体预测，且它的初步痕迹是可观察的：围绕指标条目的针对性达标行为将会出现并成熟——法条的表面修订以对齐特定指标项、选举观察机制的形式化接待、专门为指数编码者准备的国别文书、以及一个为改善指数排名提供咨询的服务市场。这与法学院为提升排名而重组行为（Espeland & Sauder, 2007）是同一个机制，且传播成本更低——指数的编码手册是公开的，条件 a 瞬时满足。

于是：

比较政治学用来测量选举外壳保真度的工具，本身就是一层外壳，且它正在脱钩。

这一步强制不一致，因为它不是一个外部判断。它是该学科自己的方法论工具，在该学科自己已经接受的机制下的必然结论。反应性（Espeland & Sauder, 2007）在政治学中被广泛引用且被接受。因此这个学科同时持有：

- (c) 反应性为真：排名改变被排名者的行为，直到被排名者向度量标准靠拢。
- (d) 民主指数测量的是被测国家的真实属性。

它无法同时保留 (c) 和 (d)。而它目前两个都持有，且把 (c) 用于分析别人的指数（大学排名、法学院排名、营商环境指数），把 (d) 留给自己的指数。这个豁免没有理由。本文的律不给这个豁免——它不认识"这是我们自己的指标"这个论据。

（四）一个可检验的、与该学科标准预测相反的命题

强制不一致本身不是经验命题。本文给出它的经验落点。

该学科的标准预测：成熟民主国家中围绕选举结果正当性的冲突，其强度应与该国选举的真实竞争性负相关——竞争越真实，结果越难被指为不公，冲突越少。这是"民主质量下降 → 冲突上升"这一读法的直接推论，也是当前关于"民主倒退"的整个诊断文献的隐含前提。

本文的预测相反：冲突强度与选举的真实竞争性无关，与**败选的不可逆性**正相关。败选不可逆性的操作化：权力的赢者通吃程度——行政任命权的集中度、司法任命的政治化程度与任期长度、选举周期的长度、联邦与地方之间的权力冗余度、下野者的政治生存空间。

两个预测在数据上可分离，因为这两个变量在成熟民主国家中并不共变：某些高竞争性国家的败选高度可逆（联邦制、比例代表制、短周期、司法独立、多层级的权力冗余），某些高竞争性国家的败选高度不可逆（赢者通吃的总统制、由胜选者提名的终身制最高法院、单一选区制、缺乏地方权力冗余）。

本文的预测是：控制选举竞争性之后，败选不可逆性对选举正当性冲突强度仍有显著的独立预测力；而控制败选不可逆性之后，选举竞争性的独立效应消失。

如果结果相反——竞争性有独立效应而不可逆性没有——本文对选举的整个读法受损，而 P1 的普适性随之受损。这是一个可以用现有跨国面板数据直接跑的检验，本文没有跑它。

（五）本节与选举型威权文献的关系

需要说明本节与那个文献的关系，因为表面相似度很高。

那个文献已经论证：混合政体中的选举同时承担挑选统治者与宣示合法性两种功能，当它不再产生真实选择时，取消它会引爆合法性危机，维持它则暴露功能空转。这与本文的判断非常接近。

但它有一个自我设限：**它把这一结构当作威权政体的特征，用以区别于有实质竞争的选举。**这个设限是这个文献存在的条件——如果这个结构是所有选举的结构，那么"选举型威权"就不是一个类别，是一个位置。

本文的主张就是这个：**所有选举首先都是产生可接受的败选的装置。**民主国家的选举与威权国家的选举的区别，不在于前者不生产可接受的败选，而在于前者的最低保真成本被支付得更足。一场败选之所以能被接受，从来不是因为败选者相信计票是准确的——他不核对。是因为他找不到一个可以指认的施害者。当他找到了，如同他相信选举被操纵时，外壳就破了；而外壳一破，接受立刻消失，冲突立刻具备落点。

近年若干成熟民主国家中围绕选举结果正当性的冲突，在本文的框架下不是"民主的倒退"，是外壳破裂：装置照常运转，票照常被计，但那个使败选可接受的表面被戳穿了。这与本文对教育的诊断（P3：被摧毁的是记忆，不是效度）是同一个机制在另一个领域的兑现。

而由推论五，这个领域的时滞也在坍塌：使外壳破裂所需的成本，随着争议叙事的生产与传播成本趋近于零而同步坍塌。这给出本节最后一个可检验命题：**外壳破裂的频率将与争议叙事的传播成本负相关，而与选举的真实竞争性无关。**

九、与最强对手的正面交锋

（一）竞争解释：波特的机械客观性

本文的最强竞争解释不是路径依赖，也不是批判再生产，而是波特（Porter, 1995）的机械客观性理论。理由在第二节第二小节已说明：波特已经论证了量化不为准确，本文若不能与他分岔，就只是一次换词。

波特的完整主张可以被表述为：**量化是弱者的策略。**当决策者的权威牢固时，他们凭判断决策，不需要数字；当决策者的权威受到质疑时，他们需要一个看起来不依赖任何人的判断的程序来抵御质疑。因此，量化的厚度与决策者的专业权威负相关。

（二）2×2 设计与它能承担的份量

两个理论对同一组变量给出相反的预测：

- **波特的预测：**外壳厚度 $\propto$ 决策者权威的脆弱程度
- **本文的预测：**外壳厚度 $\propto$ 拒绝的不可逆程度（与决策者权威无关）

在多数现实场景中，这两个变量共变（权威脆弱的地方往往也是拒绝激烈的地方），因此两个理论无法被区分。区分必须在两个变量分离的格子里进行。

	拒绝可逆性高	拒绝不可逆性高
决策者权威高	德国手工业师傅（Meister）资格：行会权威极高，可重考可转轨。两说都预测薄。不判决。	法国大学校入学考试（concours）：判决格
决策者权威低	美国社区学院开放录取：两说都预测薄。不判决。	中国高考：两说都预测厚。不判决。

判决格只有一个：决策者权威高 × 拒绝不可逆性高。

这一格的实例是法国大学校的 concours，而它恰好是波特自己的核心案例国的核心装置。

波特的经验支柱之一，是法国国立路桥学校的工程师：他们的社会权威牢固到不需要成本效益分析，这与被国会质疑的美国陆军工程兵团形成对照。这个对照是波特全书最有力的一笔。

但同一批法国工程师的入学装置，是世界上外壳最厚的装置之一：两年的预备班（classes préparatoires）、极度形式化的笔试与口试（colles）、精确到小数位的排名、名次决定校次的刚性对应。这套装置的形式化程度远超美国的大学录取。

波特预测薄。实际是厚。而本文预测厚——因为 concours 的拒绝是法国社会中最不可逆的拒绝之一：大学校的名次在很大程度上决定终身职业轨迹，且几乎没有中途转轨的通道。

这一笔能承担多少份量，必须说清楚。 判决格里只有一个实例，且它是在结果已知之后被选取的。**因此这不是一次判决，是一次演示。** 它演示了两个理论在这个格子里给出相反预测，并给出了一个实例的方向。它不排除选择性——本文选中 concours，是因为它显眼且方向对本文有利。

判决需要这个格子里的多个实例被系统测量。本文列出可以进入这个格子的候选：日本的东大入试、韩国的大学修学能力试验、印度的 IIT-JEE 与 UPSC、法国的 concours、英国牛剑入学在其考试主导时期、以及若干国家的司法考试在其名额受限时期。这些案例中，决策者（各自体系中的学术权威）的权威强度与拒绝的不可逆性可以被分别编码。本文没有编码它们。这是本文检验方案的一个明确缺口，且它是本文自愿承认的缺口——一个把 $n=1$ 写成“判决”的检验设计，正是本文自己所描述的那种借用了科学语法而不承担科学义务的东西。

（三）本文对波特的理论吞并（而非并列）

上一小节的结果不应被读为“波特错了”。更准确的读法是：**波特的机制是本文机制的一个特例。**

工兵团为什么需要量化？因为他们的项目可能被国会否决。**他们自己就是潜在的被拒者。** 他们需要的，与考场里的学生需要的是同一样东西：一个非人格的、不可争辩的、可归因于自身的拒绝理由。他们不是在用数字说服国会，他们是在为自己制造一个能够接受否决的形式。

而法国路桥工程师在决策时不需要量化，是因为在那个位置上他们不面临拒绝；同一批人在十八岁面对 concours 时，则是纯粹的被拒者，于是外壳达到了极致的厚度。

同一批人，两个位置，两种结果。波特把这两个位置读成了"权威高 vs 权威低"，因此他的变量是决策者的属性。本文认为实际的变量是位置：**谁在承受拒绝的风险，谁就需要外壳。**

由此，本文与波特的关系不是并列的两种解释，而是包含关系：把波特的"弱者"精确化为"面临不可逆拒绝的位置"，他的全部经验发现都可以被本文的框架消化，而本文能够多解释他解释不了的四件事——评级机构在证伪后的强化、法国 concours 的厚度、USMLE 改制后的压力转移、以及诊断手册在效度被公开否定后形式化程度的提高。

（四）次要检验：影子排序（排除批判再生产路径）

批判再生产路径预测：排序装置是统治结构施加于被统治者的符号暴力，因此取消该装置的显式度量，被统治者的焦虑应下降。

本文预测相反：安抚需求内生于被拒者，若分配功能未被同时取消，取消显式度量将导致自发的影子排序——由参与者自己重建的、非官方的排序装置。

这一分歧有大量现成的经验落点：

- 取消班级排名的学校，家长自建的成绩统计群与私下互换分数的行为
- 社交平台隐藏点赞计数的实验（Instagram 于 2019—2021 年在多个市场测试隐藏点赞数），随后出现的第三方计数工具，以及平台最终改为用户可选而非全面隐藏
- 取消 GPA 排名的高校，用人单位与研究生项目自行重建的院校—课程折算表

三个案例的共同结构：度量被拿掉，需求还在，参与者自己把度量造回来。

批判路径无法预测这个结构：如果度量是被施加的，那么施加被撤除后，没有理由预期被施加者自己重建它。脱耦理论也无法预测：如果正式结构是给外部环境看的，那么外部要求撤除后，它应当变为纯粹冗余而被放弃。

只有一个前提能预测影子排序：**需要它的人就是被它拒绝的人。**

（五）本理论的证伪条件

本文在五个层面上是可证伪的，逐条列出：

（1）P1 的证伪：若能证明某个大规模评价装置的诞生场景确实是认知需求而非分配危机——即建立者面对的问题是"我们不知道谁有能力"而非"我们不知道如何让落选者接受落选"——且该装置在无名额约束的条件下仍然发展出了排序与外壳，则产品命题受损。

（2）P4 + P6 的证伪（律的核心）：找到一个同时满足以下三条的装置——承担稀缺分配功能、排序型、且**其优化产业已经成熟**（存在以优化该装置输出为唯一目的的商业服务，且该服务的市场规模稳定或增长）而外壳仍与它声称测量的对象保持稳定耦合。找到一个，律即被推翻。

这个 falsifier 与本文早期版本的差别值得指出。早期版本的条件是"两代人之后仍保持耦合"，它可以被无限期地拖延——"再等等，还没到"。现在的条件不能被拖延，因为"优化产业成熟"是一个可观察的、有日期的事件。**一条律的证伪条件如果被它的辩护者用等待来消解，那条律就不是律。**本文没有找到这样的装置；这不构成证明，只构成邀请。

(3) P5 的证伪（边界）：若阈值型装置（无名额约束）也系统性地表现出与排序型装置同等的脱钩与应试产业成熟，则排序/阈值的分界不成立，律退化为一句"所有评价都会腐败"的犬儒判断——那样它就失去了全部价值。USMLE 案例的后续追踪是这一条最直接的检验场：若 Step 2 CK 在改制五年后并未成为核心排序装置、其应试产业并未成熟，本文的再分泌命题直接受损。

(4) 推论五的证伪（外壳换代）：若在人工智能条件下，过程型装置没有取代结果型装置；或它取代了但三重结构不完整（保留了一个可指认的判断者）；或结果型装置在时滞坍塌之后仍然维持了可归因性——推论五受损。检验窗口是三到五年，不是二十年。这是本文最快可被检验、也最容易被推翻的一条。

(5) 第八节的证伪：若跨国面板显示，控制败选不可逆性之后，选举竞争性对选举正当性冲突强度仍有显著独立预测力，而不可逆性的独立效应消失，则本文对选举的读法受损，P1 的普适性随之受损。

需要处理的内生性。第二小节的设计有一个必须承认的内生性问题：一个社会的拒绝不可逆性与它的外壳厚度可能互为因果——正因为有了极厚的外壳，社会才把如此多的终身后果押在一次拒绝上。本文无法在观察数据中完全切断这个回路。可行的部分补救是利用外生冲击：某些国家因劳动力市场结构改革（而非教育政策）而改变了转轨的可逆性，可以观察其考试烈度的滞后变化。这一设计的可行性尚待验证，与第二小节的 $n=1$ 一起构成本文检验方案的两个最大弱点。

十、两个落点，都不在评价内部

（一）为什么评价端的一切干预都是同义反复

由 P4，任何在评价端的干预——改进测量、更换指标、引入多元评价、建设替代认证——都只是更换宿主。它们不能不失败，因为它们瞄准的是外壳，而外壳不是承重墙。

这个结论对教育政策具有直接的、且相当难以接受的含义：**教育评价改革是不可能的，因为它不是评价问题。**过去四十年全球教育改革的反复失败，不是因为改革者不够努力或不够聪明，而是因为他们修了一个不承重的部件。

这也是本文与其前身框架（把困境诊断为测度与安抚的双重锁定，并主张"先分流安抚、后释放测度"的三阶段改革）的分岔点。那个方案的第一阶段要求在旧系统外建立独立的认可通道，并要求该通道必须具备"公平可见"——即必须提供一种类似旧分数的可比较性。这一要求在方案内部是自治的：不提供可比较性，新通道就无法分流安抚。但由本文的律，提供可比较性的那一刻，新通道就已

经开始分泌外壳，它在一个时滞之后就是新的旧系统。**那个方案在自己的第一阶段里埋了自己的反例。**它没有错在执行细节上，它错在它仍然把安抚当作一种可以被"分流"到别处去的负荷，而没有看到安抚会在任何落脚处重新长出同一层壳。

而由推论五，这个诊断还要更硬一层：在人工智能条件下，新通道分泌外壳所需的时间从一到两代人坍缩到数年。**一个改革方案的设计周期，现在长于它所设计的装置的脱钩周期。**这意味着评价端改革不只是注定失败，它是注定在自己完成部署之前就失败。

（二）落点一：语法改造（排序→阈值）

由 P5，若分泌压力来自名额约束下的相对拒绝，则把装置从排序型改造为阈值型，可以解除分泌压力。这不是取消评价——阈值型装置仍然评价，仍然拒绝，仍然可以很严格。它改变的是拒绝理由的语法：从"你比别人差"改为"你没到线"。

USMLE Step 1 的改制是这一落点的一次真实执行，它同时暴露了这一落点的极限：压力转移到了 Step 2 CK。

这个失败给出了一条精确的设计原则，而这条原则只能从本文的律中推出：

阈值化必须是全谱的，否则只是转移。若一个分配场景中存在任何一个仍然产出连续标量的装置，排序需求会立即在它上面重新落脚。

这条原则对教育政策是可操作的、也是苛刻的：把高考改为等级制而保留其他连续指标（学业水平考试原始分、竞赛名次、学校排名），排序压力会转移到保留的那个上面。彻底的阈值化要求在同一分配场景中的所有装置同时放弃连续标量。这在名额约束保持不变的条件下几乎不可能实现——因为名额约束本身要求某处必须有一个排序。

由此推出一个更硬的结论：**语法改造的可行性上限，由名额约束决定。**名额不动，语法改造只能转移压力，不能解除它。

这把问题交给了第二个落点。

（三）落点二：需求端（降低拒绝的不可逆性）

订正二·本小节的落点被撤回

反例摆在中国的教育现场，且规模巨大：**复读**。复读是可逆性的极限形态——没有任何一次拒绝是终局的，单次拒绝的终身不可逆性被降到接近于零。按本小节的预测，分泌压力应当因此降低。实际相反：复读是这个装置的病理最纯粹、最集中的形态。**可逆性让那个"本可以通过的我"保持可抵达，同时不让他终止——这不是缓解，是最优运行条**

件。它不清账，它续期。本小节把"德国二元制更温和"读为可逆性的功劳，读错了变量：起作用的是本小节自己写下的那句话的后半段——"职业资格在劳动力市场上具有独立的、不被学历折算的价值"，那说的是可结清与被承认，不是可逆。订正后的落点只剩一条，且更窄：让账可以被结清，且结清必须被收账方承认。

安抚需求的产生率是稀缺性与拒绝不可逆性的函数。稀缺性在多数场景中不可由教育政策调整。拒绝的不可逆性可以。

这里必须精确地区分本文的主张与一个近似的、但不同的主张。老左派的方案是"消除不平等"——即降低稀缺性。这不是本文的方案。本文的主张更窄、也更可操作：**降低单次拒绝的终身不可逆性。**

这两者可以分离。一个社会可以保持相当程度的不平等，同时使拒绝高度可逆——被拒者在三十岁、四十岁仍有实质性的重新进入通道。也可以相当平等而拒绝极不可逆——所有人得到的差不多，但那一次决定了是哪一份。

可逆性的操作化是可能的，且不依赖任何新的数据基础设施：

- 三十岁及以后完成职业轨道实质性转换者占同龄人的比例
- 通过非传统路径（非全日制、职业教育、成人教育、企业内训）进入高收入职业的比例
- **首次教育分流决定与四十五岁时职业地位的相关系数**——这是最直接的一个指标：它直接测量"那一次"决定了多少

本文的预测是：这一相关系数与该社会的考试烈度（备考时长、校外辅导支出占家庭可支配收入比、考试相关的青少年心理危机发生率）呈正相关，且在控制人均收入、不平等程度（基尼系数）、教育扩张程度之后仍然显著。

德国的二元制与东亚的单轨升学构成了这一预测的一个粗略对照：德国的分流发生得更早（十岁左右），按"分流早晚决定压力"的常识预测应当更残酷，实际是更温和——因为它的分流是可逆的（职业教育通道与高等教育通道之间存在实质性的横向转换机制，且职业资格在劳动力市场上具有独立的、不被学历折算的价值）。**决定压力的不是分流的早晚，是分流的可逆性。**这是一个可以用现有的跨国数据直接检验的命题。

（四）这一节的诚实位置

本文的干预建议在实践上是弱的，这一点必须直说。它没有给出"做什么"的清单，它给出的是"不要做什么"的判决——不要在评价端做任何事，那是同义反复。而它指出的两个落点，一个（全谱阈值化）在名额约束下几乎不可行，另一个（可逆性）不在教育部门的管辖范围内，它属于劳动力市场制度、职业资格体系、社会保障结构。

这不是本文的谦虚，是本文的结论：**教育评价的困境不能在教育内部被解决，因为它从来不是一个教育问题。**一个理论如果诚实，它就必须允许自己得出"你管辖的范围内没有解"这样的结论，而不是为了显得有用而发明一些在自己的框架下注定失败的建议。

十一、结论

（一）核心发现

本文的中心判断可以凝缩为一句：

教育评价装置的社会产品不是关于学习的信息，而是可接受的拒绝。测度不是这个产品的目的，是它的分泌物——一层必须看起来像测量、却从不承担测量义务的外壳，它所生产的是不可争辩性而非准确性。因此，分数与能力的脱钩不是病症而是常态，人工智能所摧毁的是"曾经耦合"这一集体记忆而非效度本身；因此，任何承接稀缺分配功能的替代装置都将在其优化产业成熟时分泌出同构的外壳——替代不是替代，是宿主更换。

这个判断与既有研究的关系不是补充，是方向的调转。技术优化路径要修好测量，批判路径要揭露测量被权力污染，制度惯性路径要解释测量为何难以被重新校准——三者都把测量当作这个装置的目的。本文认为测量从来不是目的。它是一个更基本的社会需求（在稀缺面前产生可被落选者接受的落选）为了运作而必须长出来的东西。

由此，本文对当前危机给出的诊断，与几乎全部现行对策相反：监考、检测、回归纸笔、重设题型这一整类"恢复分数效度"的努力是范畴错误，它们试图修复一个从未存在过的东西；而项目制评价、能力档案、微证书这一整类"更换更好的测量"的努力是同义反复，它们只是给同一个需求更换宿主。

但本文对这一诊断加了一条它此前没有的推论。由 P6 与推论五：那些"恢复效度"的努力会成功，只是不以它们声称的方式成功。**它们不会恢复效度，它们会重建可归因性——而社会对成功的判据从来是可接受性，不是准确性。**因此它们会被记录为一次成功的改革，而外壳完成了一次换代：从结果的不可争辩，换到过程的不可争辩。

（二）理论贡献

其一，对评价社会学：一条带边界的律，而非又一个诊断。 现有的评价批判——通约化改变被测对象（Espeland & Stevens, 1998）、量化是弱者的策略（Porter, 1995）、指标一旦成为目标就腐败（Campbell, 1979; Muller, 2018）——都成立，但它们共享一个形态上的弱点：它们什么都能解释，因而什么都不禁止。本文的外壳再分泌律有一个明确的边界条件（排序型必然分泌，阈值型不必）和一个明确的时间参数（时滞由优化产业成熟速度决定），因而它禁止一些事情：它禁止阈值型装置系

统性脱钩；它禁止排序型装置在其优化产业成熟后仍保持耦合。这些禁止使它可被证伪，也使它区别于“评价都是骗人的”这类犬儒判断。

其二，对波特的机械客观性：从并列到包含。 本文不主张波特错了，主张他的机制是本文机制的一个特例。他把变量识别为决策者的权威强度；本文认为实际的变量是位置——谁承受拒绝的风险，谁就需要外壳。工兵团需要成本效益分析，是因为他们的项目可能被否决，他们自己就是潜在的被拒者。这一重新定位使波特的全部经验发现被保留，同时使他解释不了的四件事被消化。

其三，对制度分析的一般贡献：分泌关系。 制度理论已有寄生、脱耦、分层、漂移、转化等一系列描述制度部件间关系的概念。本文补充了一种：**分泌**——B 为自身运作而产生 A，A 从未独立存在，A 的形态完全由 B 的功能需求决定。这一关系与脱耦的差别产生可检验的分歧：脱耦预测外部压力消失后正式结构因冗余而被放弃；分泌预测只要 B 的功能需求还在，A 就会被重新长出来。影子排序现象是这一分歧的经验落点。

其四，对比较政治学：一次强制不一致，而非一次跨域类比。 该学科同时持有“反应性为真”（排名改变被排名者）与“民主指数测量被测国家的真实属性”这两个命题，并把前者用于分析别人的指数、把后者留给自己的指数。这个豁免没有理由。本文的律不给这个豁免。由此，民主与威权的分界在本文框架下不是种类差别，是外壳保真度上的一个位置——这个结论不是一个外部判断，是该学科自己已接受的机制在自己的工具上的应用。

其五，一个时间机制，而非一个时间常数。 本文早期版本把脱钩时滞归纳为“一到两代人”，这是一个装饰在律上的常数，它使律的证伪条件可以被等待消解。本文把它换成了一个机制（时滞 = 优化产业成熟时间 = $f(\text{规则稳定化速度, 优化知识传播成本})$ ），由此得到一个不可拖延的证伪条件，并产生了一个此前不存在的推论：在人工智能条件下时滞坍缩到耦合记忆形成时间以下，外壳因此换代而非消亡。

（三）政策含义

本文的政策含义主要是负向的，且不易被接受：

第一，停止在评价端投入改革资源。 由 P4，评价端的一切干预都是更换宿主。这包括本文所批评的，也包括本文所继承的那些方案——包括“先分流安抚、后释放测度”这类看起来已经绕过了正面对抗的精巧设计。而由推论五，评价端改革现在还要面对一个新的算术问题：改革方案的设计与部署周期，已经长于它所设计的装置的脱钩周期。

第二，若要在评价端做事，唯一有意义的动作是语法改造，且它必须全谱。 把排序型改为阈值型可以解除分泌压力，但只要同一分配场景中只剩下任何一个连续标量，压力就会转移到它上面（USMLE 案例）。这条原则是苛刻的，且它的可行性上限由名额约束决定。

第三，实际的落点在教育之外。 降低单次拒绝的终身不可逆性——通过职业资格体系的独立价值、中年转轨通道的实质化、非传统路径的劳动力市场认可——可以降低安抚需求的产生率，从而降

低外壳的分泌压力。这不属于教育部门的管辖范围。这是本文最难被接受、也最不可回避的结论：教育评价的困境不能在教育内部被解决，因为它从来不是一个教育问题。

（四）局限

其一，历史论证依赖二手研究，且在最坏情况下站在两个案例上。 本文对装置诞生场景的解读建立在既有史学成果之上（Elman, 2000; Miyazaki, 1976; Lemann, 1999; Karabel, 2005），未做档案工作。若普鲁士考试受科举影响的线索成立，本文的独立案例从三个降为两个；本文明说了这个最坏情况并论证它够用（第四节第二小节），但“够用”的判断本身依赖于普遍性由机制而非案例承担——一个不接受这个分工的读者可以合理地认为本文的历史地基比它声称的薄。

其二，判决格 $n=1$ 且事后选取。 第九节第二小节是一次演示，不是一次判决。判决需要该格子里多个实例被系统编码，本文列出了候选但没有编码它们。

其三，判决格的内生性未被切断。 拒绝不可逆性与外壳厚度可能互为因果。本文提出的外生冲击设计（利用非教育政策导致的转轨可逆性变化）的可行性尚未验证。

其四，“优化产业成熟”的判据未经检验。 本文把律的时间参数交给了这个变量，并给出了一个可操作的判据（存在以优化该装置输出为唯一目的的商业服务，且市场规模稳定或增长）。但该判据的信度未经检验，且它在优化知识不被商品化而通过非市场机制扩散的场景中（家族内部传承、单位内部经验共享、封闭社群的口耳相传）会失灵。这是把常数换成机制之后新产生的脆弱点——本文认为这个交换是划算的（一个失灵条件明确的机制强于一个无机制的常数），但它是一个交换，不是一次纯粹的改进。

其五，排序/阈值在现实中是连续谱。 本文只给出了定性形式（ Φ 严格递增、 $\Phi(0, \cdot) = 0$ ），并在第七节第六小节说明了为什么它拒绝写出一个更好看的定量形式。这使边界条件的操作化程度低于它看起来应有的程度。

其六，本文没有解释稀缺本身。 本文把稀缺当作外生变量。但一个更彻底的追问是：为什么现代社会必须把如此多的终身后果压在如此少的位置上？这个问题在本文的框架之外，且它可能才是根。

十二、自反：这条律适用于评价本文的装置吗

一个关于评价装置的理论，如果能够论证该律适用于一切装置却唯独不适用于评价它自己的那个装置，那么它不是一条律，是一个指向他人的标签。本节承担这一检验。

（一）同行评审：律的第六个兑现

同行评审是不是排序型装置？ 期刊版面是稀缺的，存在名额约束。因此是。由 P4，它必然分泌外壳。

它的外壳是什么？ 是评审意见。评审意见必须以质量判断的语法呈现——方法论是否严谨、贡献是否充分、文献是否完备、论证是否成立。它必须看起来像一次对学术质量的测量。

它生产的是什么？ 由 P1，是可接受的拒绝：让被拒稿者接受被拒，并继续投稿（要件一）、不指认具体施害者（要件二——这正是双盲评审的功能，它与糊名制在结构上是同一样东西，且理由完全一致：它不提高判断的准确性，它消除可指认的施害者）、以及把责任交还给自己（要件三——"我的方法论确实不够严谨"）。

它脱钩了吗？ 关于同行评审的评审者间信度，已有大量证据表明其水平极低（Cicchetti, 1991）。彼得斯与切奇（Peters & Ceci, 1982）做过一个著名的实验：把已在顶级期刊发表的论文改换作者姓名与机构后重新投给同一批期刊，绝大多数未被识别出是已发表论文，且绝大多数被拒——理由是严重的方法论缺陷。这个实验精确地展示了外壳的性质：同一份稿件，同一批期刊，相反的判决，且两次判决都以质量测量的语法给出，都不可争辩。

它是否因此消亡？ 没有。它的形式化程度反而在持续提高——更多的评审人、更结构化的评审表、更细的评分维度、更严的利益冲突声明。这正是第七节第五小节表格中四个领域的共同结构：**效度被证伪之后，形式化程度反而提高**。同行评审是这条律的第六个兑现，而它是唯一一个本文自己身处其中的。

它的时滞呢？ 由 P6，同行评审的优化产业的传播成本正在坍缩：投稿信的撰写、审稿意见的应对策略、期刊偏好的逆向工程、乃至稿件本身的生成，其成本在数年内下降了几个数量级。本文的律因此对本文自己的宿主给出一个预测：同行评审的脱钩将在数年内完成，而它不会因此消亡；它将换代到过程型外壳——预注册、数据与代码的强制公开、贡献者角色分类、写作过程的留痕。这些正在发生，且它们被普遍读为"提高科研诚信"。本文的读法与第七节第三小节对教育的读法完全一致：**不是诚信，是换代**。

（二）一个标本：本文自己分泌的那一片外壳

上一小节的自认还是笼统的。而笼统的自认是廉价的：一个理论如果只在总体上承认"本文也分泌"，那个承认无法被反驳，因此也无法被兑现。它是一个姿态，不是一次检验。

本节给出一个标本。

本文在准备第七节第六小节时，写下过这样一个判别式：

分泌压力 $\propto$ 名额约束的紧度 $\times$ 拒绝的不可逆性

它看起来是本文最硬的一句话。它有一个正比号、两个变量、一个乘法。它可以被引用、被代入、被画成一张图。

而它是本文分泌的一片外壳，且它的三重结构是完整的：

- **非人格化**：判别式不是本文的判断，是一个式子的输出。你无法反驳一个式子的作者，只能重算这个式子。
- **程序化**：任何人代入同样的两个变量都得到同样的结果。这切断了"换个人判断就不一样"这条反驳路径。
- **通约化**：两个性质完全不同的东西——名额的紧度、一次拒绝对一生的覆盖面——被放在了同一个标量上相乘。而被丢弃的，恰恰是争辩的材料：一旦它们被通约，"这两个东西不能这么合起来算"这句话就变得不可说。

它的最低保真成本也被支付了：它确实与某种真实的东西相关（两个变量都会影响分泌压力）。它只需要相关到让读者可以认领"我确实觉得名额越紧压力越大"，就够了。它不需要准确。

而本文写下它，不是因为本文推导出了它。 是因为本文要在一个分配稀缺版面的装置里活下来，而这个装置对" Φ 对两个自变量都严格递增，且 $\Phi(0, \cdot) = 0$ "的接纳概率，低于它对" $\infty A \times B$ "的接纳概率。这两句话说的是同一件事。后者多出来的部分——那个乘号、那个可通约的标量、那个线性——全部是外壳。

本文把诚实的版本放在第七节，把这一片外壳放在这里。这不是因为本文自觉。是因为本文被自己的律逮到了，且逮到的时刻在提交之前而不是之后，纯属运气。

而这条律对此还有一句更不客气的话：**本文还有别的外壳，本文没找到它们，而本文找不到它们的原因与考生找不到自己的原因是同一个——分泌是为了让本文自己能接受本文自己的处境。**

（三）最后一笔

本文提出了一条律，给出了 2×2 设计、五层证伪条件、操作化指标、跨域准入条件、逐行标注的证伪强度。这一整套装备的功能是什么？

它使本文的主张不可争辩。它借用了科学的语法——正如外壳借用测量的语法。它做的是它所描述的那件事。

这里必须诚实到底，不能靠一句"但我们是自觉的"脱身。本文的处境是这样的：按本文自己的边界条件（P5），本文不该分泌外壳——因为本文不分配任何稀缺资源，它不拒绝任何人。但本文确实分泌了。为什么？因为它要在一个分配稀缺版面的装置里活下来。

这不是本文的失败，恰恰是律的又一次兑现，而且是最干净的一次：**一个自身不分配稀缺的东西，一旦进入一个分配稀缺的场域，就必须分泌外壳。** 本文的外壳不是本文的需求，是它的宿主的需求。

由此，本文能够对自己的命运做出一个可被观察的预测：**如果本文的判断被广泛接受，"外壳再分泌律"会成为一个可被引用的标签，人们会为了发表而使用它，它会与它所指称的东西脱钩。** 这条律预测自己会被吸收并脱钩。

而由 P6, 本文还能给出这个预测的时间参数, 且它比本文早期版本给出的短得多: 使用这条律的优化知识——如何在—篇论文里恰当地引用它、如何用它包装一个既有发现、如何让审稿人觉得作者读懂了它——其生产与传播成本在人工智能条件下近乎为零。**因此它的脱钩不需要二十年。它需要一个投稿周期。**

这不是自我拆台。一个预测自己会被免疫的理论才是不可证伪的; 一个预测自己会被吸收并脱钩、并给出脱钩时滞的理论, 给出的是一个可以被观察的时间序列。如果三年后"外壳再分泌律"成为教育社会学论文里一个用来提高录用概率的关键词, 而没有任何人再去检验它的边界条件——那就是这条律关于它自己的预言的兑现, 也是它作为一条律的最后一次、也是最彻底的一次经验支持。

十三、参考文献

-
- Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.
- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of Michigan Press.
- Au, W. (2009). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society and culture* (R. Nice, Trans.). Sage.
- Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America: Educational reform and the contradictions of economic life*. Basic Books.
- Bromley, P., & Powell, W. W. (2012). From smoke and mirrors to walking the talk: Decoupling in the contemporary world. *Academy of Management Annals*, 6(1), 483–530.
- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.
- Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1), 119–135.
- Collins, R. (1979). *The credential society: An historical sociology of education and stratification*. Academic Press.
- Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). Free Press.
- Dahl, R. A. (1971). *Polyarchy: Participation and opposition*. Yale University Press.
- David, P. A. (1985). Clio and the economics of QWERTY. *American Economic Review*, 75(2), 332–337.
- Diamond, L. (2002). Elections without democracy: Thinking about hybrid regimes. *Journal of Democracy*, 13(2), 21–35.

- Elman, B. A. (2000). *A cultural history of civil examinations in late imperial China*. University of California Press.
- Elster, J. (1989). *Solomonic judgements: Studies in the limitations of rationality*. Cambridge 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.
- Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24, 313–343.
- Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.
- Insel, T. (2013, April 29). Transforming diagnosis. NIMH Director's Blog. National Institute of Mental Health.
- Karabel, J. (2005). *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin.
- Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.
- Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge University Press.
- Mahoney, J., & Thelen, K. (Eds.). (2010). *Explaining institutional change: Ambiguity, agency, and power*. Cambridge University Press.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363.
- Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China* (C. Schirokauer, Trans.). Yale University Press.
- Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.
- Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2), 187–195.
- Pierson, P. (2004). *Politics in time: History, institutions, and social analysis*. Princeton University Press.
- Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
- Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge University Press.
- Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, 170(1), 59–70.
- Sandel, M. J. (2020). *The tyranny of merit: What's become of the common good?* Farrar, Straus and Giroux.
- Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63–82.

- Schedler, A. (2002). Elections without democracy: The menu of manipulation. *Journal of Democracy*, 13(2), 36–50.
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.
- Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, 83(10), 758–765.
- Weber, M. (1978). *Economy and society: An outline of interpretive sociology* (G. Roth & C. Wittich, Eds.). University of California Press.
- Young, M. (1958). *The rise of the meritocracy*. Thames & Hudson.