

它读不出来的，它也造不出来

王德生 · 德麦国际 SDE 学派 · SDE教育学 · 2026年8月10日 · 约 27,426 字 · 36 条参考文献 · 三种阅读模式

摘要

「教育是什么」通常被当成一个关于人的问题。本文换一个落点。工程学、生物学与艺术学各持一条关于「一样东西是什么由什么单独决定」的强主张，三条互不相容：输出相同即状态相同；决定它的是它长成的那条路径；决定它的是它所站进去的那片条件。三家争的是一个人「学到了什么」该由哪一边裁，而共同假定了：那是这个人身上的属性，找对读法就读得出来。推翻它的材料来自工程学自己，但取法必须先收紧——可观测性与可控性的对偶连的是一个系统与它的对偶系统，给不出「同一个系统读不出即造不出」这一步，本文因此不从定理推出任何一步；承重的那一步是经验的：判定与培养共用一份成文规格时，评价标准与培养目标是一张表的两次使用。据此本文提出：教育的产物不是能力，是不可分辨性；一套判定装置装上的那一刻，同时切出一类它既读不出也造不出的能力，这一类在边界画出之前不构成一类，换一条边界就整批更换成员。读数为张成重合率。正文做了两次预注册真跑，均用公开数据、可复算，合计五条结果对本文不利，其中两条各逼出一次实质收窄。判据是一句不含程度词的问话：新旧两套读数，能不能互相算出来。

关键词：不可分辨性；未张成能力；张成重合率；判定装置；空无课程；教育本体

Title: What It Cannot Read, It Cannot Make

Subtitle: On indistinguishability as the product of education, and how the boundary of an assessment apparatus simultaneously carves out a class of capabilities it can never grow

Abstract: This paper relocates the question “what is education”. Engineering, biology and art each hold an incompatible claim about what alone settles the identity of a thing: engineering identifies states that produce identical outputs; biology holds that the developmental path decides; art holds that the surrounding conditions decide. All three presuppose that “what a person has learned” is a property of that person, readable if only one finds the right reading. The refuting material comes from engineering itself, but must be taken with care: observability–controllability duality relates a system to its dual, not the reading and the making of one and the same system, and therefore supplies no step of the argument here. The load-bearing step is empirical — where judging and cultivating are carried by a single apparatus, the assessment standard and the cultivation goal are two uses of one document. Hence: the product of education is not capability but indistinguishability; an apparatus, at the moment it is installed, carves out a class of capabilities it can neither read nor make — a class that does not exist as a class before the boundary is drawn, and whose membership is wholly replaced when the boundary moves. The reading is the span-overlap ratio. Two pre-registered runs on public data are reported; five of their results run

against the paper, and two force substantive narrowings, including withdrawal of the claim that the ratio is comparable across domains.

Keywords: indistinguishability; unspanned capability; span-overlap ratio; assessment apparatus; null curriculum; ontology of education

一·三条不该同时为真的判断

有三件事，各自所在的行当都视为常识，而它们不能同时为真。

第一件在工程学。一个系统的内部状态若对一切输入都产生完全相同的输出，那么在系统理论里，这两个状态不是「难以区分」，是**被判为同一个状态**。最小实现这个操作会把它们合并掉，合并之后的模型与原模型有逐点相同的输入输出行为。这不是权宜，是定理：既读不出又推不动的那一块，对外部行为没有任何贡献，因此它可以不在模型里。

第二件在生物学。二〇二〇年一项工作用可遗传的条形码给造血祖细胞逐个做标记，让它们分裂、分化，再回过头把成熟细胞的去向与它当初的转录组对上。结论里有一句很硬的话：姐妹细胞之间存在内在的命运偏倚，而单细胞转录组测序读不到这个偏倚。也就是说，两个在当下一切可读维度上完全一样的细胞，它们后来会走去不同的地方，而这个差别在它们分开之前已经存在，只是不在读得出的那一层。

第三件在艺术学。二〇二三年三月，美国版权局发布了一份关于含人工智能生成内容作品的登记指引；同年九月，复审委员会维持了对一幅获奖数字绘画的登记驳回。这里的关键不在谁对谁错，在于那个判定的形式：**两幅在像素层面无法分辨的图像，可以一幅受保护、一幅不受保护，而全部差别一处也不在图像上**。差别在它是怎么来的、来自谁、站在哪一段历史里。

把这三条摆在一起，它们对同一个问题给出互不相容的答案。问题是：**两个人在考场上、在工位上、在评审前表现得完全一样，他们是不是学到了同一样东西？**

工程学的答案是「是」，而且这个「是」不是让步，是它的立身之本——若表现相同还要追究内部差别，整个系统辨识与状态估计的学问就没有地基。生物学的答案是「不一定」，而且它有实测：读不出来的差别照样决定后续。艺术学的答案是「不」，而且它把差别的位置指得很明确：不在物上，在物之外那一圈条件里。

这三条不能靠「各有各的道理」和稀泥。若工程学对，另两家在追究一种对外部行为无贡献的差别，那是形而上学不是科学；若生物学对，工程学的最小实现在有发育的系统上是错的；若艺术学对，前两家都把一件由条件决定的事误当成由对象自身决定的事。三条各有硬证据，各自的领域也确实靠它吃饭。

本文认为这三条的不相容不是可以调停的，它指向一个三家都没有说出口的共同假定。找出那条假定，用其中一家自己的材料把它推翻，得到的判断三家都不会同意，而三家的证据合起来只能得出它。

路线是这样的：第二节把三家各自站的位置写实，第三节让它们在一个具体的人身上分岔，第四节说出那条共同假定，第五节交出推翻它的材料。第六到第九节是本文的正面部分。第十节是一次真跑，它的结果有两条对本文不利。第十一节到第十九节是兑现、约束、分界、边界、证伪与自查。

二·三家各自把什么当成单独够用的那一样

三家不是三种口气，是三个位置。逐家写实。

2.1 工程学：看它在输出端表现出什么就够了

工程学这一家的核心判断句能引到原文。卡尔曼在一九六三年那篇《线性动力系统的数学描述》里把系统按可控性与可观测性做了四分解，其中既不可控又不可观测的那一块在最小实现里根本不出现；同一分解在一九六二年的短文《线性动力系统的正则结构》里已经给出。更早，一九五六年摩尔在《序贯机器的思想实验》里就写死了机器状态的不可分辨关系：若两个状态对一切输入序列产生相同输出，它们不可区分，而不可区分的状态在化简后的机器里被合并为一个。

这条主张在今天仍是活的，不是教科书里的旧账。二〇一三年一篇工作把可观测性推广到复杂网络，问的是需要监测哪些节点才能重构整个系统的状态；结构可辨识性分析这一整片文献问的是同一件事的参数版本——给定这组测量，模型里哪些参数原则上不可能被定出来。这两支都默认同一件事：**不能被定出来，是一个关于测量配置的问题，而定不出来的那部分，工程上就当它不存在。**

工程学自曝的那一处是它自己的常识：**可观测性从来不是系统单独的性质，是系统与一组测量的联合性质。**判据里那个矩阵，一半是系统的动力学，一半是传感器怎么装的；换一组传感器，不可观测子空间就换一块。工程学天天用这条做设计可测性，而它从不把这条当作关于「这个东西是什么」的证据——因为在工程里，被测的那台机器先于测量而存在。

2.2 生物学：看它经什么路径被组织成这样就够了

生物学这一家的奠基句在一九九八年卡米洛夫-史密斯那篇《发育本身才是理解发育障碍的关键》里：不能拿成人神经心理学那套模块剖面去读一个发育中的系统，必须看轨迹本身。同一条主张的经典侧还有格里诺等人一九八七年来对经验期待型与经验依赖型可塑性的区分——同一段经验，来早来晚，长出来的不是同一样东西。

二〇一五年之后的实证正是前面提到的那一份：条形码谱系追踪把「状态」与「命运」分开，明确报告了转录组测不到的内在命运偏倚。这一支这些年长得很快，从谱系条形码到与单细胞组学的合流，方法学综述已经成片。

生物学自曝的那一处也很干脆：**那个偏倚是靠加装条形码才读出来的。**不加条形码，它就不存在于任何数据里。生物学从这一处得出的结论是「我们需要更好的标记技术」，而不是「路径差别是否存在，取决于装了什么标记」——因为在生物学里，细胞先于标记而存在。

2.3 艺术学：看它站进了什么条件就够了

艺术学这一家的奠基句在一九六四年丹托那篇《艺术界》里：把某物看成艺术，需要眼睛看不见的东西——一套艺术理论的空气、一段艺术史的知识。一九八一年那本《寻常物的嬗变》把它推到极致：

感知上无法分辨的两件东西，可以一件是艺术品一件不是。同一路的还有一九八〇年莱文森关于音乐作品的上下文主义——两份声学上相同的乐谱，若创作于不同的艺术史位置，是两部不同的作品。

当代实证不必去哲学期刊里找，它已经变成制度事实。美国版权局二〇二三年那份指引与随后的两次复审驳回，把「人类作者身份」写成了登记的门槛；同一幅图，人做的受保护，机器做的不受，而两者在像素上不可分辨。这是丹托命题第一次被写进一个国家的行政规则里。

艺术学自曝的那一处最要命：**它有读法而没有造法**。它能判定一件东西是不是艺术、是不是这一件而不是那一件，但它从不主张自己能据此造出艺术家。这一处自曝，本文第十二节会拿它反过来削自己。

2.4 三家的位置，一句话摆开

	它把哪一样当成单独够用	核心判断的出处	它自己自曝的那一处
工程学	它在输出端表现出的行为	摩尔 1956；卡尔曼 1962、1963	可观测性是系统与测量的联合性质
生物学	它被组织成这样的那条路径	卡米洛夫-史密斯 1998；温赖布等 2020	路径差别要靠加装标记才读得出
艺术学	它所站进去的那片条件	丹托 1964、1981；莱文森 1980	它有读法而没有造法

三句指向三处，不指向同一处。工程学落在「输出端表现出什么」，生物学落在「怎么被组织成这样」，艺术学落在「站在什么里」。这三处不能互相化约，也不能靠判谁对来化解——它们根本不在同一个语域里作答。

还要说明一件事：**教育学不在这三家里**。教育是本文这道题落地的地方，不是一个持有立场的位置。教育测量、学习科学、能力本位教育这些说法后面确实各有其人其年，但它们对本文这道题的回答都要先借上面三家中的某一家。教育学在本文里出现的位置是第十三节的分界，不是第二节的来源。

三·让三家在一个人身上分岔

抽象地摆立场看不出分量。取一个具体的人。

小林与小周同一年从同一所学校的同一个专业毕业，成绩单几乎一样，毕业设计的评级一样，实习评价一样，入职测评一样。用人单位手上的一切读数，把这两个人放在同一格。

工程学的处置：把他们合并。这不是省事，是正确——既然一切可得的输入输出对上都相同，两人在本模型下是同一个状态，对他们做不同的安排缺少依据，而做不同的安排会引入不可辩护的差别对待。工程学在这里其实站在公平的一边：**不可分辨即同一，是拒绝暗中歧视的一条硬规则。**

生物学的处置：不合并。一个是从初中起自己一步步摸出来的，一个是三年补习班一路推上来的；两条轨迹不同，即便当前一切可读维度重合，后续可塑性也不同。生物学会说：给我一个新情形，差别就会出来。它无法预先说出是哪个新情形。

艺术学的处置：也不合并，但理由完全不同。一个的作品是在一个有师承、有争论、有观众的场里做出来的，一个的作品是照着评分表凑出来的；两件东西在成品上不可分辨，而它们不是同一类东西。艺术学也无法预先说出这个差别会在什么时候兑现——它甚至不认为它必须兑现，因为在它看来「是什么」不必等到「会怎样」来证明。

三条处方在这里正面分岔，而且分岔是可执行的：合并、按轨迹分档、按来源分类，是三种不同的人事安排，通不过「各有道理」这一关。

现在把问题往前推一步。假设用人单位听了生物学的，加装一项新测评，专门测「面对没见过的问题时的重组能力」。加装之后，小林与小周分开了。

这时候发生的到底是什么？两种读法：

读法甲：原本就存在的差别被测出来了。测评是发现工具，差别先在那儿。 **读法乙：**一条新的边界被画出来了，边界一画，两人分处两侧。

三家都会选甲。工程学会说这是加了传感器、原不可观测的模态变成可观测的；生物学会说这是加了标记、原读不出的路径差别显形了；艺术学会说这是这件事终于被放进了合适的条件里去看。三家在这里第一次意见一致，而这正是共同假定露头的地方。

四·三家共同假定了什么

把三家的争论压成一个句式。三家争的是：「这个人学到了什么」这件事该由哪一边来裁——由他在输出端表现出的行为、由他长成的路径、还是由他所站进去的条件。

填在横线上的那样东西是「他学到了什么」。三家吵的是归属，而它们共同假定了：这是那种可以有归属的东西——它是这个人身上的一个属性，在裁定的那一刻已经确定，读法只是把它取出来。

把这句话念给三家听，三家都会说这还用说吗。工程学：状态当然是系统的状态，传感器只是读它。生物学：命运偏倚当然在细胞里，条形码只是把它显影出来。艺术学：作品的身份当然由它的来历确定，鉴定只是把来历查清楚。没有一家会为这句话争起来，这是它是共有前提的验收标志。

这条前提有一个更一般的名字，而它比三家都老。莱布尼茨一六八六年在《形而上学论》里写下不可分辨者的同一：若两样东西的一切属性都相同，它们就是同一样东西。三百多年来，这条被反复攻打，但被攻打的一直是它的形而上学版本。它的工程学化身从来没有被当成一条主张看待过——「凡是对现行读数没有差别的，就不是一样东西」，这句话在今天不是一个立场，是各行各业的默认设置。

这里要区分两层。第一层是三家共同假定的那条：那件事是这个人身上的属性。第二层更宽，是所有相关的人都站着的那块地：一样东西的存在，以它在某个现行读数上留下差别为条件。第二层不能用第一层顶替。三家的那条被推翻，不等于这块地被动过；本文第五节推翻的是第一层，第十三节末尾才回来处理第二层。

五·推翻它的那件材料，来自工程学自己

推翻共有前提的材料必须来自三家之一自己已经掌握、已经发表、只是没往这个方向用的事实。从外面搬一个理由来推翻，得到的只是第四家的立场——那样只是加入了争论，没有取消它。

材料是**可观测性与可控性的对偶**。

内容是这样：对一个线性时不变系统，输出矩阵与状态矩阵这一对的**可观测性**，等价于它们转置之后那一对的**可控性**。这条在卡尔曼那批工作里就已经写明，此后成了控制理论的基本工具。龙伯格观测器的设计与状态反馈的设计是同一个问题的两面；卡尔曼滤波与线性二次型调节器的方程互为对偶。教科书里做完一边，直接把另一边写下来。

它在工程学里回答的问题是：**怎样把观测器设计得又快又稳**。这是一个纯粹的设计问题。没有人把这条定理当作关于「一样东西是什么」的证据，因为在工程学的问题结构里，被观测的系统先于观测配置而存在，观测配置改的是我们的知识，不是那台机器。

这里必须先拆掉一处很容易顺手写下、而本文初稿确实写下过的错话。**对偶连的是一个系统与它的对偶系统，不是同一个系统的读与造**。定理说的是：这一对的**可观测性**，等价于**转置之后那一对的可控性**。对同一个系统，不可观测子空间与不可控子空间一般不是同一块；正因为不是，卡尔曼的分解才有四块而不是两块——若读不出就等于造不出，那张分解表上有两块永远是空的。

所以「读不出的方向同时造不出」这一步，**对偶定理给不了**。初稿里那句话已删，删掉的理由记在这里而不是悄悄改掉：一条被引来当地基的定理，若被读成它没说的意思，后面所有的严整都是假的。

定理仍留在本文里，位置降了一级：它只提供一个**提问的形状**——「一个方向能不能被读出」与「一个方向能不能被造出」有可能不是两件独立的事。这个形状不构成论证。论证要另找，而且必须在判定装置这一侧找，不能在机器那一侧找。

现在把这条定理从机器搬到判定装置上，而搬动是有条件的，条件必须先写死。

限度声明（本节最要紧的一句）：那是线性时不变系统的定理，需要有限维、需要时不变、需要线性。学校、医院、产线、乐团一个也不满足。本文不主张这条定理适用于它们，也不从它推出本文的任何一步。本文的承重全部落在下面这一段，而这一段不引用任何定理。

承重的那一步是这样的：在一个**判定与培养由同一套装置承担**的系统里，读与造共用的不是一条数学性质，是一份文件。课程要教什么，由考试要考什么决定；训练要练什么，由验收要验什么决定；一个人往哪个方向使劲，由他的努力在哪个方向上能被看见决定。这不是类比，这是同一份文件的两次使用：一套**判定装置的规格书**，读的时候叫**评价标准**，用的时候叫**培养目标**。它们逐条对应，因为它们本来就是一张表。

这一步是经验主张，不是定理，因此可以是错的，而且错法很具体：**只要能找出一项写进了培养目标、不在任何一项验收读数上、而仍被稳定地培养出来的能力，本文这一步就塌了**。它被列为第十六节的 F6，是全部证伪条款里最便宜的一条——取一份课程标准与它配套的验收标准逐条对表，报缺口

比例即可。本文没有做，因为手上没有配套齐全的成文材料。这是本文欠下的第一条真跑，且它比第十节那两条都更致命。

于是共有前提垮掉了。「这个人学到了什么」不是一个先在的属性等着被读出来。装上一套读法这个动作，同时做了两件事：它规定了什么算被读出，也规定了什么算能被造出。读法不是找出来的，是装上去的；而装什么读法，本身是这套教育的一部分，不是这套教育之外的观察。

第三节末尾那个岔口现在有答案了。小林与小周在新测评之后分开，不是甲也不完全是乙——是那条新边界同时创生了一批「现在能被造出来的方向」和一批「现在造不出来的方向」，而两人被分到边界两侧这件事，是这次划界的产物，不是它的发现。

六·未张成的那一类

6.1 命题

教育不是能力的传递，不是发育轨迹的塑造，也不是意义条件的接入，而是不可分辨性的生产；它每成功一次，就把一批差别从「两个人之间的差别」改写为「同一格里的噪声」，同时在能力空间里切出一类它在原则上既读不出也造不出的能力。

三个被否掉的说法，各是一家的划界标准。能力的传递是工程学的：受过教育＝输出通过验收。轨迹的塑造是生物学的：受过教育＝能力沿着一条被塑造的路径长出来。条件的接入是艺术学的：受过教育＝被接进某个共同体的解释场，动作因而有了意义。三条都真，都不够。

被切出的那一类，本文称为**未张成能力**——简称未张成的那一类。

6.2 五条判定

一项能力属于某套判定装置的未张成的那一类，当且仅当五条同时成立：

1. **不在张成之内**：它在这套装置的全部读数上不留下任何可分辨的差别。判据不是「测不准」，是「有它和没它，这套装置的读数逐项相同」。
2. **对偶地不可达**：这套装置所配的培养安排，无法系统地把人推向它。判据是——找不出一条「做什么、做多久、怎么算做到了」的规程；找得出，说明它已经落在张成之内。
3. **由这条边界切出**：它成为一类，是因为这条边界被画出来了。换一套装置，这一类的成员整批更换。判据是——说得出口「换成哪一套装置，它就不再属于这一类」。
4. **无持有者**：没有任何一个岗位、科室、部门或个人的职责里写着它。它不属于谁没做好，它不属于任何人。
5. **无字段**：三家的账本上都不设这一栏。工程学把它并进不可观测子空间然后商掉；生物学把它记进残差与发育噪声；艺术学把它推给「艺术界」，而艺术界不被计量。

第三条是与所有邻居分开的那一条，也是最容易被读漏的一条。它说的不是「有一批能力被埋没了」，是**这一批之所以成为一批，是划界的产物**。

6.3 一个能立刻走一遍的例子

某职业学院把「设备维修」这门课的验收从「按标准流程完成八个规定故障的排除」改为「在教师随机制造的故障上，说出你判断故障位置的依据并排除它」。

改之前：张成之内是「按流程执行」的各个维度。未张成的那一类里有「在没有流程可依时构造一个判断依据」这一项。它不留下任何可分辨差别，因为八个故障都有标准流程；也没有任何规程能把人推向它，因为把它写成规程就取消了它；没有哪位教师的职责里写着它。

改之后：这一项进入了张成之内。而**新的一批进了未张成的那一类**：比如「判断这台设备值不值得修」——因为新验收里所有故障都是被制造出来给人修的，从来没有一个「不修」是正确答案。

注意这里的关键：改革**不是把未张成的那一类变小了，是换了一批成员**。这是本文与「教育有盲区、应当补上」那一路说法最硬的分歧，第十三节会逐条处理。

七·为什么它是一类存在物，不是一个新读数

这一节回答一个必须回答的质疑：所谓未张成的那一类，会不会只是「没被测到的能力」的一个新叫法？没被测到的能力当然存在，命不命名它都存在，那么本文做的就只是操作化。

判别只有一条：**删掉本文提出的这个读数之后，那样东西是不是就不存在了？** 若它照样存在、只是不好测，本文做的是操作化；若它是由这个读数造出来的，本文做的才是别的事。

答案是后者，理由有三层。

第一层：没有边界就没有外部。 「没被测到的能力」这个说法，预设了一个先在的、完整的能力总集，测评是从中取出一部分。未张成的那一类不是这个总集的补集——它是**这一条边界的补集**，而补集只有在边界画出之后才是一个集合。边界没画，无所谓内外；边界一移，成员整批更换。你不能说「小林身上有三项未张成能力」，只能说「相对于这套装置，小林与小周的差别落在未张成的那一侧」。

第二层：它的成员资格是关系性的，不是内在的。 一项能力不会本质上属于未张成的那一类。同一项「判断这台设备值不值得修」，在一套装置里未张成，在另一套里是核心指标。这一点与「潜能」「隐性知识」这类说法根本不同——那些是人身上的东西，本文这一类是**装置与人之间的一个位置**。

第三层：它有一个别的说法给不出的后果。 若只是「没被测到」，处方自然是「多测几维」。而按对偶，读的边界与造的边界是同一条，所以：**扩张读法不等于把埋没的东西挖出来，而是同时换掉造得出与造不出的两批。** 一次评价改革不是补漏，是换一副骨架。这个后果不能从「没被测到」推出来，它只能从本文这一类推出来。

还要交代一处诚实的让步。本文并不主张一切读不出的差别都无关紧要，也不主张它们必然在将来兑现。读不出的差别是否起作用，取决于有没有人装上能读它的那套读法，而装不装是一个制度选择，**不是一个自然过程。** 这一让步来自第十二节生物学给的那条约束，它削掉了本文原本很想写的一句更强的话——「真正的能力终会显现」。那句话本文不写，因为它没有依据。

八·辨别装置

一个名字不是辨别维度。把它升成一张两轴的表。

第一轴：这套装置的各项自身指标，在这次改动之后是**改善**还是**未改善**。所谓自身指标，指信度、施测效率、评分一致性、公平性审查、题目区分度这一类，也就是一次改革在结项报告里会写的那些。

第二轴：**张成重合率**——新装置的读数中能被旧装置读数重构的比例，高则两套读的是同一批方向，低则真的张成了新方向。定义与清点规则在第九节。

	重合率低（真换了方向）	重合率高（同一批方向）
自身指标改善	扩口径：读得出与造得出的两批同时更换，代价通常也真实发生	同口径精修（要打的一格）：一切形式审查通过，可分辨方向一个没增
自身指标未改善	混乱了：换了方向而没做好，通常被判为改革失败并回滚	退：既没换方向又做差了，很少能推行下去

要打的是右上那一格。它必须是全部形式审查都能通过的那一格，否则不需要本文这套东西——一眼就能看出的坏，任何人都看得见。四条签名，逐条对照：

签名一·短期指标全面改善。 同口径精修恰恰是技术上最成功的一类改革：题目更精、施测更省、评分更稳、争议更少。所有能报的数都变好。

签名二·失效不产生任何信号。 没有任何现行指标会因为重合率接近一而变差，因为没有人计算重合率。这一格的失效是静默的，而且静默不是疏漏——现行指标体系里就没有这个量的位置。

签名三·越正规化越严重，而且是数学上的。 这一条比前两条硬得多，值得单写。教育测量最成熟的两项技术——等值与纵向量尺——它们的存在理由就是让新旧两套分数可以互相换算。而「可以互相换算」正是重合率等于一的定义。一次改革越是做得专业、越是保证了年度可比与跨版本可比，它的重合率就越是在设计上被钉死在一。一套装置越专业，它越不可能张成新方向——这不是从业者不努力，这是可比性要求的直接推论。

签名四·误诊阻挡。 前三条会误伤一类正当情形：有些系统本来就该保持高重合率。安全关键岗位的资格认定、法定计量、司法鉴定，它们要的正是「同一件事在不同时间不同人手里得出同一个数」。所以第四条签名是一个反向条件：若这套装置的产物只被用于判定「合不合格」而不被用于分配稀缺机会，高重合率不是病。一旦它被用于排序、择优、分层，高重合率才落进要打的那一格。这一条把学历筛选与安全资格分开，本文只谈前者。

九·张成重合率与它的清点手册

9.1 定义（全文只此一份）

张成重合率 $\lambda(B|A)$ ：读法 B 的全部读数中，能被读法 A 的读数线性重构的方差比例。

设 A 为读法 A 的 $n \times p$ 读数矩阵，B 为读法 B 的 $n \times q$ 读数矩阵，各列先标准化。对 B 的每一列在 A 上做最小二乘回归得到决定系数， λ 为这些决定系数的平均值。 λ 趋近一，表示 B 落在 A 张成的空间内，换了题面没换口径； λ 趋近零，表示 B 与 A 正交，真的张成了新方向。

9.2 清点手册

项	写死的规定		
读数名	张成重合率，记 λ		
定义	B 的读数中被 A 线性重构的方差比例，取各列决定系数的均值		
粒度	一次「读数」= 一次进入记录的、对该对象的数值化判定。三次期中成绩算三个读数，不合并为一个总评		
边界例子	一份卷子的六道大题分数：若六个分数各自入档，算六个读数；若只有总分入档，算一个		
编码规则一	各列先标准化，否则量纲大的项主导		
编码规则二	二值项按零一计入，不剔除		
编码规则三	不做变量筛选。筛选会同时抬高或压低 λ ，且方向不可预知		
编码规则四	要求样本量至少是 A 的列数的十倍，否则 λ 因过拟合虚高，判为不可用		
编码规则五	λ 不对称。报告时必须写明方向， $\lambda(B \setminus$	A) 与 $\lambda(A \setminus$	B) 一并给出
编码规则六	必须同时报出同口径内部 λ 作为该领域的天花板 （A 各列互相重构、B 各列互相重构）。 λ 的绝对值不可跨领域比较，只可与本领域天花板并读；缺此项的 λ 判为不可读		
表头	对象编号 读法 A 各列 读法 B 各列 每列决定系数 $\lambda(B \setminus$	A) $\lambda(A \setminus$	B) n p q
不适用	A 的列数接近样本量时；两套读数不在同一批对象上时；读数是纯类别而无序时		

三处一致性自查：本文正文第九节、证伪条款第十六节、赌注第十六节末，引用的是同一个定义、同一份粒度规定。

9.3 四条性质

无量纲。可事后清点，不需要新采集，只要两套读数落在同一批对象上。它把一个印象（「这次改革是不是换汤不换药」）变成一个数。

第四条最要紧，也最容易漏：**它与既有的主要指标正交。**一次改革可以在信度、效度、区分度、公平性上全面改善而 λ 接近一。举一个「既有指标最好看而这个数最难看」的真实例子：**计算机自适应测验**。它的测量精度、施测效率、题目信息利用率全面优于纸笔定长测验，是测量学的技术高地；而单维项目反应理论的模型假设本身就规定了只有一个可分辨维度，题库通常也就是原题库。**在这类设计里 λ 不是碰巧高，是被写进模型的。**举不出这样的例子，一个新读数就只会被读成既有指标的代理；这一条举得出来。

△ 这里要撤掉初稿写下的一句话。初稿在本节开头写的是「无量纲，因此三个领域可以并排比较」。第 10.5 节的第二次真跑把它推翻了： **λ 的量表两端会随领域移动。**在教育那份数据里，同口径内部是零点八五、跨口径是零点零三，两端相距二十四倍；而在机器学习评测那份数据里，同口径内部只有零点六五，跨版本是零点五九，两端几乎贴在一起。**同一个零点五九，在前一个领域里是低得刺眼，在后一个领域里几乎顶到天花板。**因此本文改判： λ 的绝对值不可跨领域比较，报告时必须同时报出该领域的同口径内部 λ 作为天花板，否则这个数读不出大小。这一条已写进第 9.2 节清点手册的编码规则六。

9.4 三个领域的同一个比率

领域	A 是什么	B 是什么	λ 高意味着
教育	旧评价方案的各项分数	新评价方案的各项分数	新方案没有增加可分辨方向
生物学	一套表型测定	另一套表型测定	新测定没有触到旧测定之外的性状
工程学	旧传感器配置的可观测子空间	新配置的	加装的传感器没有扩大可观测子空间
艺术学	比赛评分表各分项	策展人评述编码后的各维	两套评审读的是同一批方向
机器学习评测	换版前的评测集各项	换版后的评测集各项	新一版没有张成新方向（见 10.6）

四处量纲不同，比率相同。

十·真跑两条：预注册、公开数据、以及五条对本文不利的结果

证伪条款不许全是纸面的。本节把最便宜的一条用公开数据跑通，预注册在看结果之前冻结，全部脚本与数据坐标随文附出。

10.1 数据与预注册

数据是加利福尼亚大学欧文分校机器学习数据库里的学生表现数据集，两所葡萄牙中学，数学课样本三百九十五人，葡萄牙语课样本六百四十九人，含三次成绩与二十六项行为与背景记录，公开可下载可复算。

预注册写死了三条假设与阈值：

P1：两门完全不同学科的成绩，在同一批学生上，控制两科共有的背景记录之后，残差相关仍达到零点四以上。含义是——把两科绑在一起的**不是学科内容**。**P2**：档案口径（学校已记录的行为与背景）能被考试口径线性重构的比例达到零点三以上。**这一条是本文原始版本的预测**：若教育装置只有一个口径，学校记录的其余东西应当大体落在它张成的空间里。**P3**：按考试口径取前两成与按档案口径取前两成，两份名单的重合度低于零点四。

10.2 结果

先给量表的两端，否则一个孤零零的比率读不出大小。

比较	葡萄牙语课	数学课
同一口径内部：第三次成绩由前两次重构	0.848	0.822
同一口径内部：后两次成绩由第一次重构	0.716	0.684
跨口径：档案由考试重构	0.035	0.027
跨口径：考试由档案重构	0.341	0.252
档案中行为类由考试重构	0.041	0.034
档案中出身类由考试重构	0.029	0.019

P1 命中。按数据集自带的合并规则，两科唯一匹配到的重叠学生为三百五十八人（数据集原始说明给的数字是三百八十二，本文这里更严，只保留键唯一的配对，差额如实记在此处）。两科最终成绩的原始相关为零点四八二；控制掉二十六项共有背景记录之后，残差相关为零点四五二，置换检验的显著性小于万分之一。**控制掉家庭教育程度、学习时间、缺勤、家庭支持、通勤、饮酒、健康这一整套之后，那条把两科绑在一起的带子几乎原封不动。**

P2 未命中，而且是大幅未命中。 阈值零点三，实测零点零三五与零点零二七，低了一个数量级。这一条直接推翻了本文原始版本的一个预测。 学校记录的行为与背景，百分之九十七的方差读不出考试口径；反过来考试口径倒有约三分之一能被档案重构。这个不对称是本节最有信息量的一处。

P3 命中。 两份前两成名单的重合度为零点三一五与零点三六七，即两套口径选出的人有三分之二不同。但这一条的代理很弱，见下。

10.3 这些结果改了本文的什么

第一处修订，也是本文因真跑而做的最大一次改动。 P2 未命中说明「机构只有一个口径」这个图景是错的。学校手里明明记着大量与考试口径几乎正交的东西。于是本文的定义必须收窄：**未张成的那一类，要相对于「入判口径」定义，不是相对于「机构记录的一切」。** 入判口径指真正进入分配决策的那一套读数——录取、评优、分班、推荐、录用所依据的那些。学校记着缺勤、家庭支持、课外活动，而这些几乎不进入任何一次分配。

这次收窄的代价与收益都要说清。代价：本文的适用面变窄了，它不再是关于「一个机构能读到什么」的命题，只是关于「一个机构据以分配的那一套读到什么」的命题。收益比代价大得多：**它给出了一个便宜得多的处方。** 若未张成相对于入判口径而言，那么「扩张口径」在很多场合根本不需要新采集——数据已经在档案里躺着了，缺的只是把它接进判定。这个处方绕开了第十二节工程学那处约束（多装传感器本身有代价，且代价由被测者承担），因为它不增加任何采集。

第二处修订：粒度不是小事。 敏感性检验显示，把三次成绩合并成一个总评之后， λ 从零点零二七变到零点零一七二，在数学课样本上差了将近一点六倍。同一批数据、两种都合理的编码，读数差一半以上。这直接逼出了第九节清点手册里那条粒度规定，而那条规定在真跑之前只是一句原则。

第三处：P1 命中，但它证明不了本文最想要的那件事。 控制掉背景之后两科残差仍高度相关，这排除了「家庭与行为背景」这个解释，但剩下两个候选本数据分不开：一是一般认知能力，二是两科共用的那套评价装置。本文倾向后者，而**本数据没有资格支持这个倾向。**要分开它们，需要的是同一批人在两套评价装置下的读数，而不是在两门课下的读数。这一处本文写在正文里，不写在脚注里。

第四处：P3 的代理是弱的。 档案口径的合成分用了档案变量的第一主成分，而它只解释了百分之十点八与十一.五的方差——档案变量本身高度分散，没有一个主导方向。换句话说，「档案口径」在这份数据里其实不太成其为一个口径。这一处削弱了 P3 的证据力度。

第五处：代理变量明写。 这份数据里的「档案」是研究者事后用问卷收集的，不等于学校日常记录；本文用它替代学校档案，是一处弱代理。真正该跑的是学校自己的学籍系统，而那份数据不公开。

10.4 这次真跑证明了什么、没证明什么

证明了： λ 是可算的，量表两端是分开的（同口径零点八以上，跨口径零点零三），粒度会显著改变读数，而机构确实持有大量与入判口径正交的读数。

没有证明：本文的核心命题。核心命题说的是一次评价改革前后的重合率，而这份数据里没有一次改革。第一次真跑做不到这一点，于是有了第二次。

10.5 第二次真跑：找到了一次有日期的改革

核心命题要的那种数据——同一批对象，在一次口径更换前后各有一套读数——在教育里几乎不公开。但它在另一个领域是公开的。

一个公开的大模型评测榜单在二〇二四年六月做过一次整体换代。旧读法六项（ARC、HellaSwag、MMLU、TruthfulQA、Winogrande、GSM8K），新读法六项（IFEval、BBH、MATH Lvl 5、GPQA、MuSR、MMLU-Pro）。换版的公开理由是旧集已被刷到饱和且存在污染，需要更难、更新的题。这是一次有日期、有理由、由专业团队执行的口径更换，**而且它与教育里的改革有一处关键不同：它不要求新旧可比，它是故意把可比性打断的。**按第八节签名三的逻辑，这正是最有可能张成新方向的那一类改革。若连它都张不成，签名三就不只对教育成立。

对象是两版都跑过的模型，按名称匹配，**共同样本三百九十七个**（旧版去重后六千八百一十二，新版四千四百九十六）。预注册在运行前冻结，阈值直接取自本文原有的证伪条款 F4，未作改动。

一处执行偏离，如实记录：预注册写了「两版都剔除标记为可疑的模型」，而旧版存档快照里该字段七千二百六十行全为真，即整份榜单退役时被统一标记，字段不携带区分信息。处置是旧版不剔除、新版照剔，并加做一次两版都不剔除的敏感性检验；两种做法的结果到小数点后三位相同。阈值、匹配键与 λ 定义均未动。

10.6 结果

读数	数值
同口径内部：旧六项互相重构（留一均值）	0.867
同口径内部：新六项互相重构（留一均值）	0.649
跨版本：新六项由旧六项重构，λ 方向为「新由旧」	0.590
跨版本：旧六项由新六项重构， λ 方向为「旧由新」	0.652
只用新读法总分（粒度粗化）	0.793
前两成名单重合度	0.418
两版平均分的秩相关	0.827

逐项看新六项各自被旧六项重构：MMLU-Pro 零点八三三、BBH 零点七六四、GPQA 零点五四九、IFEval 零点五三六、MATH 零点五二九、MuSR 零点三三〇。

10.7 这一跑对本文不利的三处

不利之一 · 主件落在不确定带，不能算本文赢。 预注册写死： λ 低于零点三则本文第十一节第十条与 F4 为伪； λ 高于零点六则这次改革在结构上没有换方向；中间判为不确定，如实报告且不作有利解释。实测零点五九〇，落在中间，且差零点零一没够到那条对本文有利的线。按预注册，本文不能据此说「连故意打断可比性的改革也没换掉方向」。这一条只能这么写。

不利之二·量表两端在这个领域没有分开，这直接推翻了本文的一条性质。第一次真跑里，同口径零点八五对跨口径零点零三，两端相距二十四倍， λ 的大小一看就懂。这一次，新读法自身的内部 λ 只有零点六四九，而跨版本是零点五九〇——旧读法预测每一道新题的本事，几乎与新题的亲兄弟预测它的本事一样大。两端贴在一起，孤立的 λ 值就读不出大小。本文因此撤掉第 9.3 节「无量纲，因此三个领域可以并排比较」那一句，并在清点手册加了编码规则六。这是本文因真跑而做的第二次实质收窄，代价是 λ 从一个可以横向排队的数，降为一个只能与本领域天花板并读的数。

不利之三·一条方向性预测只对了一半。预注册预测「MMLU-Pro 最高、GPQA 与 MuSR 最低」。MMLU-Pro 确实最高（零点八三三），MuSR 确实最低（零点三三〇），而 GPQA 落在中间（零点五四九），高于 IFEval 与 MATH，与预测不符。GPQA 是专门为「专家才答得出、网上搜不到」而造的题，本文原以为它最该是新方向；它不是。本文没有解释，如实列为第十九节的洞四。

10.8 这一跑立住的两处

立住之一·加难不等于扩口径，这次有数了。新六项里被旧六项重构得最好的是 MMLU-Pro（零点八三三），而它正是旧读法里 MMLU 的加难重制版；被重构得最差的是 MuSR（零点三三〇），它换的是任务形态而不是难度。同一次改革内部，加难的那一项与换形态的那一项，重合率差两倍半。第八节说「细化不是扩张」、第十一节第十二条说「把十项细化为三十项重合率通常接近一」，此前都只是推断；这里第一次有了同一次改革内部的对照数。

立住之二·粒度效应在第二个不相干的领域复现。第一次真跑里，把三次成绩合并成一个总评， λ 从零点点二七变到零点点一七，差一点六倍。这一次，把新读法六项压成一个总分， λ 从零点点五九〇涨到零点点七九三，差一点三四倍。两份毫无关系的数据、两个相反的方向、同一个结论：粒度一动，读数就动，而两种编码都合理。清点手册里那条粒度规定，至此不再是一份数据的偶然。

还有一处不好归类的： λ 零点五九对应的前两成名单重合度是零点点四一八，即换版之后前两成里换掉了将近六成。单一个点画不出剂量-反应曲线，所以这不作为 F1 的证据，只作为 F1 那条曲线上的第一个坐标点记在这里。

10.9 两次真跑合起来证明了什么、没证明什么

证明了： λ 是可算的；粒度会显著改变读数，且这一点跨领域复现；一次专业团队执行的、故意打断可比性的口径更换，其新读数仍有近六成方差落在旧读数张成的空间里；同一次改革内部，加难的那一项比换形态的那一项高出两倍半的重合率。

没有证明：本文的核心命题仍然没有在教育里被跑过一次。第二次真跑找到的那次改革不在学校里。模型不是人，也不会因为知道自己被这样评而改变准备方向；恰恰因为不会，它才排除掉了回路效应那一路的解释。但也正因如此，它对判定与培养共用规格这一步毫无检验力：榜单只判定，不培养。按第十二节第二条，纯判定系统本来就不在本文适用范围内。所以第二次真跑检验的是 λ 这个工具，不是本文的核心命题。这一点必须写在正文里，不能留在脚注里。

十一·十二处兑现

兑现不是类比。类比是「像」，兑现是「在那个领域里能据此预测出一件具体的事」。逐条给出预测与去哪儿看。

一·**学位与职业资格的分层**。预测：一个专业的入判口径越窄，其毕业生在入职三年内的表现方差越小、而五年后跨岗位迁移的成功率越低。去处：同一行业不同院校的校友追踪数据，看的是方差不是均值。

二·**招聘中的算法面试**。预测：把算法题面试换成系统设计面试，若两套读数的重合率高，则被录取者名单变动小于百分之十五。这可在任何一家换过面试形式的公司的历史数据上事后清点，不需要新采集。

三·**师范生的实习评价**。预测：实习评价表若由课堂观察量表逐项派生，其重合率接近一；而实习指导教师的自由评语若单独编码，能张成正交方向。检验：把评语文本编码后与量表分数做一次重构，看落在哪一端。

四·**医学教育的双轨**。客观结构化临床考试当初被造出来，理由正是笔试张不成临床方向。预测：若这两套读数的重合率确实低，那么两套读数各自选出的前两成学生重合度也应当低。**这一条本文没有去查，列为待查，不作为证据使用。**

五·**体育选材**。预测：一套以体能测试为入判口径的选材体系，其未张成的那一类里长期包含「比赛中改变打法的能力」——因为体能测试里没有对手。检验：看该项目的选材标准里有没有任何一项的执行需要一个会反应的对对手。

六·**音乐比赛与音乐教育**。预测：某比赛评分表若十年未改，则获奖者的演奏在可编码维度上的多样性单调下降，而在评分表未覆盖的维度上不下降甚至上升。录音是现成的，编码是可做的。

七·**工程认证**。预测：以「毕业要求达成度」为核心的专业认证，其达成度矩阵的秩远小于条目数。这可直接在任何一份认证材料上算，材料是公开的。

八·**单细胞生物学的正面对照**。谱系条形码之所以是一次真的扩口径，是因为它张成了转录组之外的方向——这正是本文所说的低重合率。预测：任何一种新的表型测定，若其读数能被既有测定高比例重构，它不会改变任何一个已有结论。这一条在方法学论文里是可以直接检验的，而现在很少有人报告。

九·**工程学的设计可测性**。预测：一块电路板的可测性设计若只是把既有测试点加密，故障覆盖率的提升会在某一点后趋于平台；真正的提升来自加入与既有测试点正交的观测通道。这是可测性文献里的既有经验，本文只是给它一个跨领域的名字。

十·**大模型的评测集**。预测：一个新评测集若与既有评测集在同一批模型上的重合率高，它不会改变模型排名。**这一条可以今天就跑**——公开榜单上多套评测集在同一批模型上的分数是现成的，算一次重构只需要一张表。而如果它确实高，那么「我们又造了一个更难的评测集」这件事在结构上没有改变任何东西。

十一·训练目标本身。 一个系统若被显式要求提高一致性、降低随机性，它在结构上被推向重合率为一的那一侧；而这个推动在全部标准评测上读不出来，**因为评测集本身就在张成之内**。这一条对本文自己适用，第十八节会回来处理。

十二·公共部门的绩效考核。 预测：一次「指标体系优化」若把原有十项细化为三十项，重合率通常接近一（细化不是扩张）；而加入一项由外部主体给出的读数，重合率会显著下降。检验：把新旧两套指标在同一批被考核单位上的历史数据做一次重构，多数机构手上就有。

十二处里有五处所需数据已经存在于某个机构或某个公开榜单里，缺的只是把两套读数并到一张表上算一次。

十二·三家反过来给本文加的约束

三家不只是本文的材料，它们也各自削掉了本文一块。这一节写的是被削掉的那些。

12.1 工程学削掉了本文的价值排序

本文原本会写下一句更强的话：**教育应当持续扩张它的可分辨维度**。这句话被工程学削掉了。

理由是工程学自己天天付的代价。传感器不是免费的：观测通道本身是失效点，装得越多系统越贵越脆越容易被攻击；而更要命的是探针效应——**测量装置会改变被测对象**，这在电路里叫负载效应，在人身上叫「因为这项被考所以只练这项」。所以「多测几维」不是一个中性的好处方，它是一次加重，而加重的代价由被测者承担。

因此本文不主张把重合率做低。 **λ 是诊断读数，不是操作杆**。本文能主张的只剩两件：一是把它算出来（这不增加任何采集），二是第十节那个修订带来的便宜路子——把已在档案里的正交读数接进判定，而不是新增采集。这一削很重，它把本文从一份改革方案降格为一次清点。

12.2 艺术学削掉了本文的适用范围

对偶的前提是判定与培养由同一套装置承担。艺术学恰恰不是这样。

艺术学有读法而没有造法，而这不是它的欠缺，是它的立场。康德在《判断力批判》里把天才规定为**为艺术立规则而自己说不出规则的那种才能——一门明确拒绝「可教」的学问，不可能有一份两用的规格书**。评奖、评论、鉴定、策展都是纯判定系统：它们判断，但不培养，也不承认自己能培养。

于是本文的核心主张在艺术学这一家里不成立。适用范围因此收窄为：**判定与培养由同一套装置承担的系统——学校、训练营、职业资格、企业内部的晋升体系**。纯评价系统不在内。这是一处真削弱，动的是适用范围本身。

这一削还有一个副作用，对本文有利也有不利：一旦一个纯评价系统开始承担培养职能（比如某比赛开办大师班、某榜单发布备考指南），它就落进本文的范围了，而这个转变是有日期、可查的。**这既是一条可检验的推论，也是一处诚实的窄化**。

12.3 生物学削掉了本文最想说的话

生物学讲等终性：不同的路径可以到达同一个终态。若终态真的相同，那么路径的差别在下游就不再起作用。这条直接顶住本文「**读不出的差别仍在起作用**」这一主张，逼本文交出一个机制说明。

本文交不出一个自动的机制说明，只能交出一个受限的：**读不出的差别在有人装上能读它的读法时起作用，不在别的时候自动起作用**。那份条形码工作正是这个形状——命运偏倚不是自己冒出来的，是加装了条形码之后才被显影，而在此之前它在任何数据里都不存在。

代价是本文必须放弃一句很想说的话：「**真正的能力终会显现**」。本文不写这句，它没有依据，而且它恰恰是本文所反对的那种前提（差别是先在的属性，只等被读出）的一个抒情版本。

十三·与最近的几种说法的分界

这一节决定本文是不是换皮。选十三位最容易与本文混为一谈的说法，逐条给出它说到哪一步、分界落在哪里、以及一条能分出对错的对照预测。前十位来自方法学与社会科学，后三位是教育学与指标研究自己的正主——它们比前十位都近，初稿把它们漏了，这是本文修订前最大的一处缺口。

一·不可分辨者的同一（莱布尼茨，一六八六）。它说：一切属性相同即同一。分界：它对「一切属性」量化，本文对「某一组已装上的读法」量化，而后者是可设计、可更换的。判定形式：若两个对象在扩充读法之后仍不可分辨而下游行为分开，则本文对而它的工程学化身错；若每次扩充读法都能把已分开的下游行为对上，则相反。这一条本文当作地基处理，不当作对手，见本节末。

二·机器状态的不可分辨性（摩尔，一九五六）。它说：对一切输入序列输出相同的两个状态可合并。分界：摩尔的合并是对给定输出函数而言的，而他不问输出函数从哪儿来；本文问的正是输出函数怎么被装上去的，以及装的人是谁。判定形式：若一套装置的输出函数由与被测者无关的外部约束确定，则本文的「装置是教育的一部分」不成立；若输出函数由培养方自己选定并同时用作培养目标，则本文成立。学校属于后者。

三·可控性（卡尔曼，一九六三）。站内已有一篇文章用可控性处理过教学里「下一步该做什么有几个答案」这个问题，那篇把可控性当作占位者推开，理由是可控性是系统性质而倒推的唯一性是知识分布性质。本文与那篇的关系是互补而不是竞争：那篇处理的是一条路径上倒推有没有解，本文处理的是读与造共用同一条边界。判定形式：一个教师群体对「下一步」意见高度一致（那篇的读数很高）而这套装置的重合率接近一，则两读数正交、两篇各说各的；若两读数在实测中相关超过零点六，两篇必有一篇是代理，届时保留可清点成本更低的那一个。

四·多质多法矩阵（坎贝尔与费斯克，一九五九）。这是本文这个读数的方法学正主，必须点名。它问的是：同一特质用不同方法测是否收敛、不同特质用同一方法测是否区分。分界：多质多法矩阵预设了特质表——它要先有「这是哪几个特质」，才能填那张矩阵。本文的读数不需要特质表，它只问两套读数张成的空间是否相同。判定形式：若两套装置在特质层面被判为测同一批特质而它们的重合率很低，则本文对而多质多法的框架不足以覆盖；若重合率总是随特质重合而变，则本文被它吸收。

五·典范相关分析（霍特林，一九三六）。这是本文这个读数的算法正主。分界：本文不主张算法上的新意， λ 就是两组变量之间共享方差的一个汇总。本文主张的是它的用法——把它当作一套判定装置的本体读数，用来判「这次改革在结构上有没有改变什么」，而不是当作两组变量之间的关联度。判定形式：若一次改革的重合率很高而它确实改变了下游的人选与能力分布，则本文这个用法是错的。

六·增量效度（塞克雷斯特，一九六三）。它问：新工具对某个效标的预测力有没有增加。分界：增量效度相对于效标，而效标本身通常来自旧口径；本文的读数不需要效标，只需要两套读数落在同一批人身上。判定形式：一个新工具若在增量效度上为零而重合率很低，则本文对而增量效度漏掉了它（它张成了新方向，只是那个方向没被现有效标覆盖）；若两者总是同升同降，则本文被它吸收。这一条是本文与教育测量学最近的接触面。

七·效度作为论证（梅西克一九八九，凯恩二〇一三）。它说：效度不是测验的属性，是对分数解释与使用的论证。分界：本文与它高度相容，但走得更远一步——若效度只相对于用途，那么被测的那个「能力」也只相对于用途，而这一步现代效度理论没有走，它仍然把能力当作被测的实体。判定形式：见第十四节那处不一致。

八·国家的可读性（斯科特，一九九八）。这是本文最危险的一位。它说：国家为了治理而把社会改造成可读的形式，而可读性正是大规模社会工程能力的前提。它已经把「读」与「造」联在一起了。

分界有两层。第一层：斯科特的联系是单向且有先后的——先要可读，于是简化现实，可读性是改造的前提；本文的联系是同一条边界的两侧，因而扩大可读性不是获得了改造能力，而是换掉了两批。第二层更要紧：斯科特那本书的核心概念是地方性知识，它在国家视野之外继续存在、继续被传承；本文说的那一类在边界画出之前不构成一类，边界一动就整批换人。

判定形式：若某种被现行入判口径压掉的能力，在口径不变的情况下能由其他机制（师徒、行会、私下）持续培养出来，则斯科特对而本文错；若它只在口径改变之后才出现，而口径一改就整批换成另一批，则本文对。可查的场合：一次纯技术性的口径变更——比如录取从总分改为分维、且未对外公告——下游的能力分布是否随之整批更换。

诚实交代一处：地方性知识的存在是本文对偶主张的一个真实反例。师徒制里判定与培养确实由同一批人承担，而它长出了大量从不出现在任何读数上的东西。本文只能靠第十二节那处窄化来应对——本文谈的是判定与培养共用一份成文规格的系统；师徒制没有成文规格，因而不在于内。这个应对是有代价的：它把本文的适用面进一步缩到了官僚化的培养体系。

九·分类造人与回路效应（哈金，一九八六）。它说：一种分类被造出来之后，被分类的人会知道自己被这样分类，并因此改变，于是分类的对象也改变了。分界：回路效应需要被分类者知道——它的机制是自我认识与自我塑造。本文这一类不与任何人互动：没有人认领它，没有人知道自己属于它，它的存在完全由边界的划定所蕴含。判定形式：若一类能力的分化只在它被命名并被人认领之后才出现，则哈金对而本文错；若一次无人知晓的、纯技术性的口径变更就整批更换了成员，则本文对。这一条与斯科特那一条共用同一个可查场合，两条一起判。

十·站内已有的一篇《能被量的那一面》。这是自家最近的一位。那篇说：能被读出来的那一面会自动接管治理方式，而它通常不是长东西的那一面。分界：那篇讲被读出的那一面如何接管（读出者的命运），本文讲读不出的那一批如何被创生（未读出者的存在地位），而本文多出的那一步是对偶——读不出等于造不出，因此不能靠「它还在，只是被忽略了」来安慰。判定形式：换一套完全不同的读法之后，若原先被埋没的那一面浮现出来，则那篇对；若浮现的是一批全新的东西而原先那一面并没有变得可造，则本文对。

另外两位自家的近邻一并交代。一篇处理「无人要求时段的自发变异」，那篇的量在时段上定义，本文的类在方向上定义，两者可同批测且互不蕴含。另一篇用细胞学、工程学与艺术学三门写「这一节课结束时这个班里多了什么」，与本文用了同样三门学科中的两门。同一门学科被两篇拿走的是两件不同的东西：那篇的工程学取的是过程能力与「质量不能检进去」，本文取的是可观测性与对偶；那篇的艺术学取的是作品自足论，本文取的是不可分辨物与条件决定；而那篇是一道关于怎么办的题，本文是一道关于这是什么的题。三家的落位也全部对调过。

十一·空无课程（艾斯纳，一九七九）。这是全部十三位里最近的一位，近到不处理就等于换皮。他把学校教的东西分成三种：明写的、隐含的、以及空无的——学校不教的那一批，而他明确主张它照样起作用，因为学生因此终身缺少某些看待事物的方式。同一路上还有更早的一位：隐性课程（杰克逊，一九六八）。

分界有三层，逐层给。第一层，定义的锚点不同。空无课程锚在内容清单上——哪些科目、哪些智力过程没有被排进课表；本文这一类锚在入判口径的边界上——哪些方向不在任何一次进入分配的读数里留下差别。一门课可以开着而它的成果从不进入任何判定（本文的未张成，艾斯纳的非空无）；也可以某个内容从没进课表而它在验收里被反复读到（艾斯纳的空无，本文的非未张成）。两者交叉，不互相蕴含。

第二层，也是本文多出的那一步：对偶。空无课程是一份缺席清单，它蕴含的处方是明确的——把缺的补进去。本文这一类不能这样处理，因为读不出与造不出是同一条边界的两侧：把一项补进课程内容，并不等于它变得可造，除非它同时被补进入判口径。艾斯纳的空无可以靠加内容来消除，本文的未张成不能。

第三层，成员资格的来源不同。空无课程有一份可以先列出来的候选清单（艺术、法律、经济学、伦理推理……），它是相对于「一个受过良好教育的人应当具备什么」这个先在标准的缺口；本文这一类在边界画出之前不构成一类，没有先在标准，换一套装置就整批更换成员。

判定形式（三层共用一次可查）：取一项被判为未张成的能力，只把它写进课程内容与教学计划，不动入判口径，一段时间后看它是否被稳定地培养出来。若培养出来了，则艾斯纳对而本文错——那说明缺的只是内容，读与造不共用一条边界；若课程里明写了、教师也教了，而它在任何进入分配的读数上仍不留差别、也仍长不稳，则本文对。这一条与第十二节末尾的 F6 是同一件事的正反两面，可一次做完。

十二·指标一旦被用于决策就会变形（坎贝尔，一九七九；古德哈特，一九八四；斯特拉森，一九九七）。坎贝尔那条写得最狠：一个量化社会指标越是被用于社会决策，就越会受到扭曲的压力，也越会扭曲它本要监测的那个过程。古德哈特的短句更广为人知：一个量一旦被当成目标，它就不再是一个好的量。

分界只有一句，但要说准：这三位讲的是变形，本文讲的是划界。变形的机制是被考核者的应对——刷分、掐尖、报表美化；它预设了那样东西本来在那里，只是被扭曲了。本文这一类不涉及任何人的应对，它在没有任何人作弊、所有人都尽力做好本职的情况下照样被切出来，因为它是边界的补集，不是行为的产物。一句话：坎贝尔的病靠诚实可以缓解，本文的这一类靠诚实缓解不了。

判定形式：找一个没有任何人有动机应对的场合——被考核者不知道指标存在、或指标不带任何后果——看那一类是否仍然存在。若指标一旦无后果、被压掉的能力就长回来了，则三位对而本文错；若照样长不出来，则本文对。这一条与第八条、第九条共用同一个可查场合（一次未公告的纯技术性口径变更），三条一起判。

十三·排名的反应性（埃斯佩兰与索德，二〇〇七）。他们做的是法学院排名：公开的量度不只是描述世界，它会把世界改造成它所描述的样子——学校重新分配预算、改招生策略、改就业统计口径，最后排名成了自我实现的。

分界：**反应性**讲的是被排名者向指标靠拢，本文讲的是指标画完之后有一批东西既读不出也造不出。前者是一个动力学过程，有速度、有滞后、可以观测到「靠拢」；后者是一个划界的即刻后果，没有过程，边界画出的那一刻它就在。更要紧的是方向：反应性预测被压掉的维度会**萎缩**（资源被抽走），本文预测的不是萎缩而是**更换**——旧的一批被读进来的同时，新的一批被切出去，总量并不减少。

判定形式：一次扩张口径的改革之后，看未被覆盖的方向是继续萎缩（反应性对），还是**换了一批**、总量不变（本文对）。这一条是本文与前两位分得最开的地方，也是最难查的一条，因为它要求对「未被覆盖的方向」有独立于口径的清点手段——而本文自己论证了这种手段不存在。**此处困难本文解决不了，列为第十九节的洞五。**

13.11 划完之后：那块所有人都站着的地

逐个问这十三位「它结构性看不见什么」：

- 多质多法矩阵把**特质表**当作给定，因此看不见「特质表本身由装置决定」——一旦承认这一点，矩阵的行列就没有先在的名字，它自己就塌了。
- 增量效度把**效标**当作给定，因此看不见「效标来自旧口径」——一旦承认，它的增量就是循环的。
- 效度作为论证把**用途**当作给定，因此看不见「用途与被测实体是同一次划界的两面」。
- 可读性把**被简化掉的现实**当作先在的、完整的，因此看不见「补集只有在边界画出后才是一个集合」。
- 回路效应把**被分类者的自我认识**当作机制入口，因此看不见无人认领的那一类。
- 摩尔的合并把**输出函数**当作给定。
- 莱布尼茨把**属性总集**当作给定。
- 空无课程把「一个受过良好教育的人应当具备什么」这份清单当作给定，因此看不见「清单本身由装置写就」。
- 指标变形那三位把**被扭曲之前的那个量**当作给定，因此看不见「没有被扭曲的时候它也只是一次划界」。
- 反应性把**被排名者会靠拢**当作机制入口，因此看不见没有人靠拢的那一批。

这几行背后是同样东西被当作给定：**一样东西的存在，以它在某个现行读数上留下差别为条件。**这是第二层，也是本文第四节末尾预告的那一块地。它比第一层宽，因为它不只被这三家共享，也被互相批判的各派共享——正是因为大家都站在它上面，谁也没法回头看它。

必须写明这块地的界限，不能夸大。它在**以读数为分配依据的领域**里成立：评价、测量、考核、认证、绩效。它在有些领域不成立——比如环境流行病学里，一个从未被任何人读到过的暴露量仍被当作真实原因；比如地质学里，没有任何记录的构造事件照样被承认。本文不主张这块地是普遍的，只主张在教育及一切据读数分配机会的场合，它是共同地基。

十三点五· 问题七篇的互切

本文与另外六篇问题文章出自同一批工作：七篇都问「教育是什么」，七篇都以「某个口径读不到、算不到或根本没放东西的那一类」作答。这一节把七篇摆在一张表上逐列分开，因为不分开，七篇看上去就是一个判断写了七遍。

篇	三门学科	它命名的那一类	读数	相对于什么被定义	对象是什么	有没有时间结构
它读不出来的，它也造不出来	工程学·生物学·艺术学	未张成能力	张成重合率 λ	入判口径的边界	能力空间里的方向	无
那道题还没有被出出来	符号学·逻辑学·数学	界题	分法未定率 u	已出题目的集合	尚未被出出来的题	无
同许	混沌学·舞蹈学·地理学	同许族	带宽·同许率·分离视界	可用观测的粒度	人与人之间的配对	有，会自己散开
不计的那一遍	戏剧学·经济学·法律学	不计遍	零计比 z ·首计序 k	计数装置的字段表	做过的那些遍	有，被记录即消灭
本来就是这样	圣经学·佛学·道家	未判之域	判形滞后期	传承体系的留造分配	所传之物的一部分	有，被指出即消失
先空着的那一处	营养学·水力学·园林学	无认空	留空清单·认领时刻	责任清单的认领状态	接受端的位置	有，被认领即开始填
设对	历史学·正义学·体育学	虚对	设对清单	比较判断的另一侧	被比较的那一个	无，它根本不存在

七篇的对象类型互不相同——方向、题、配对、遍、内容的一部分、位置、被比较的那一个——这是最干净的一刀。七类可以同时非空且互不重叠：一门课的成果从不进入任何判定（未张成非空），而它的考卷每道题只考一项技能因而能分开一切做法（界题为空）；同一批人分数彼此差得很开（同许族为空），而他们为达到这个水平做过的重复里九成不产生条目（不计遍非空）；这门课的教材里有一大段从没有人问过“必须原样还是可以重讲”（未判之域非空），而课上每个空位都写在某位教师的清单上（无认空为空）；而给这个班打的每一个分数，另一侧站着的是一个真实存在的对照班（虚对为空）。七个读数在同一批人身上可以取到任意组合。

判决性对照：一个动作，七种不同的预测。 取一批本来不进入分配决策的记录，只把它接进判定，其余一概不动：

- **未张成能力**：被切出的那一类整批更换成员，总量不变。
- **界题**：分法未定率不变——接进判定的记录并不增加能把做法分开的题，除非它恰好只考一项技能。
- **同许族**：带宽变窄、同许族变小，而分离视界不随之改变。

- **不计遍**：那批重复的构成功能下降——记了就换格。
- **未判之域：触发一次判形**——被接进判定的那部分内容从此有了留造归属，未判之域缩小一块，而其余部分不动。
- **无认空**：外部来源的进入率下降——一旦有了责任人，填它的通常就是认领它的人。
- **虚对：读数不变**——设对清单问的是另一侧站着谁，接进判定的是这一侧的记录，两者不相干；除非新记录本身规定了对照来源。

七条不能被同一次观测同时满足：第二条与第七条说读数不动，第三条说读数变小，第五条说读数缩小一块，第四条与第六条说被记的东西性质改变，第一条说总量守恒而成员更换。**任何一次真实的口径变更都会至少判掉其中两条。**七篇因此是可以互相证伪的七个判断，不是一个判断的七种说法。

还要写明读者最该起疑的地方，本文不绕过：**这七篇出自同一位作者、同一批工作、同一副骨架**（三家现成读法 → 三家共有前提 → 一件来自其中一家自己的推翻材料 → 命名一类 → 一套辨别装置 → 一句不含程度词的判据 → 一个可清点的读数 → 一次预注册或回溯检验 → 逐条分界 → 编号证伪）。骨架相同不构成七篇内容相同的证据，但它确实是七篇最可能出错的地方：**一副好用的骨架会让七个不同的问题看起来长得一样。**判据交给读者：上表七列里，若有两篇能在五列以上取到相同的值，那两篇应当合并。现在最接近的一对是**不计遍与无认空**（同为“被登记/被认领即消灭”的形状，且两篇的经验材料取自同一个公开平台），它们在七列里相同的有两列；其余任意两篇相同不超过两列。这条判据可以拿去核对，也可以拿去推翻。

十四·一处强制不一致：可迁移能力这个框架的两条主张相撞

一个判断若只是新增了一种看法，它不逼任何人改口。本节要说明的是：**不接受本文这条判断的那一门学问，它自己已经在同时持有两条不能同真的主张，而这处不一致就在它现行的处方里发作。**

靶子是**可迁移能力框架**——各国课程改革里的核心素养、通用能力、二十一世纪技能这一族，欧盟的关键能力建议与经济合作与发展组织的学习指南针都属此列。它同时持有两条：

甲：这些能力是跨情境可迁移的一般能力。这是它全部价值所在——若不能迁移，就没有理由把它从学科内容里单拎出来作为一层目标。**乙：**它们必须被可测量地界定，才能进课程、进评价、进达成度报告。这是它被做成框架的全部理由——不能测的目标进不了任何一份文件。

两条在本文的读数下正面相撞。乙一旦执行，这些能力就落进某一套入判口径张成的空间里；而甲要求它在口径之外仍然是同一样东西。**若它真的可迁移，它就不可能被任何一套口径穷尽；若它被一套口径穷尽了，它就只是那套口径里的一个方向，而不是跨情境的能力。**

现行的化解办法是：声称测的是「表现的样本」而不是能力本身。这个办法要付一个代价，而这个代价没有被说出来：**「素养」在甲乙两条里已经不是同一样东西了**——甲里它是人身上的一个属性，乙里它是任务表现的一个类。两处用同一个词，而所指不同。

已经在现行处方里发作。同一份文件里通常同时要求两件事：培养可迁移的通用能力；用标准化量表评估其达成度。而这两件事之间需要一个判据——这两套读数张成的空间是不是同一个——**这个判据在任何一份现行框架里都不存在。**没有它，「达成度百分之八十」这句话既可能意味着能力真的长出来了，也可能只意味着量表被满足了，而两者在文件上完全同形。

退路只有两条，各有代价，写明如下：

其一，承认素养只是任务表现的一个类。代价：「可迁移」失去内容，整个框架退回为一份跨学科的任务清单，而那不需要新概念。其二，坚持它是人身上的属性。代价：它不可能被任何测评穷尽，达成度评估在原理上无从谈起，现行的评估报告失去依据。

两条退路都可接受，而现行框架两条都不选，同时持有甲乙。这就是本文所说的不一致，它是可查的——去看任何一份核心素养评估方案，看它有没有回答「量表张成的空间与所欲培养的能力张成的空间是不是同一个」。

顺带交代第二处不一致，它在生物学一侧，本文不展开：发育科学同时持有等终性（不同路径可达同一终态）与轨迹依赖（路径决定后续可塑性）。若终态相同而后续不同，那么「相同的终态」这个说法就必须收窄为「在所测的那些维度上相同」，而一旦收窄，等终性就不再是关于终态的主张，是关于测量的主张。

十五·不适用的边界

五条，逐条写明。

一·**纯判定系统不适用**。见第十二节第二条。评奖、评论、鉴定、司法鉴定、法定计量，它们判断而不培养，对偶的前提不成立。

二·**无成文规格的培养不适用**。师徒制、家学、非正式的传帮带，判定与培养虽由同一批人承担，却没有一份两用的规格书。本文不覆盖它们，而它们的存在是对本文的一处真实限制。

三·**安全资格类的高重合率不是病**。见第八节第四条签名。一套装置的产物若只用于判合不合格、不用于分配稀缺机会，重合率高正是它该有的样子。

四·**不适用于个体诊断**。 λ 是装置的读数，不是人的读数。说「这个学生落在未张成的那一侧」是错的用法；正确的说法是「相对于这套装置，这两个人的差别落在未张成的那一侧」。任何把它写成个人评语的做法都超出了它的定义。

五·**不适用于能力不足型的失败**。一所什么都没做好的学校，它的问题是执行，不是口径。本文这套东西专门用在**一切形式审查都通过**的那一类系统上；用错地方会把简单问题复杂化。

十六·分层引用、证伪条件与赌注

16.1 本文可以分三层引用

第一层（最强，最易倒）：未张成的那一类是一类存在物，其成员由入判口径的边界所创生，边界一移整批更换。**第二层（分析工具）：**第八节那张两轴表与第九节的读数 λ 。第一层若倒，第二层仍可单独使用——即便未张成的那一类只是一个比喻，「这次改革有没有换掉可分辨方向」仍是一个值得算的问题。**第三层（最稳，也最便宜）：**机构手里已经持有大量与入判口径正交的读数，而它们不进入任何一次分配。这一层由第十节第一次真跑直接支持（档案与考试的重合率零点零三），不依赖前两层，今天就可复核。**修订后的分层还要加一句：**第二次真跑检验的只是第二层这件工具，不检验第一层；第一层至今仍是一条推论，没有一次落在教育里的实测。

引用纪律：只采用 λ 作为流程指标的人，不得据此声称发现了一类新的能力。

16.2 最强的竞争解释，以及排除它的设计

最朴素、最难反驳的那句是：**其实不就是能力本来就是一维的吗？**若一般认知能力主导一切，那么任何两套装置的重合率都会高，与装置怎么设计无关，本文所说的一切就都是这个事实的一个复杂说法。

排除设计：一般能力的解释预测**重合率与装置设计无关**；本文预测**重合率是装置设计的函数**。因此关键检验是——在同一批人上，一套专为张成新方向而设计的装置（如临床技能考核之于笔试）应当给出显著更低的重合率，而两套同源设计的装置（如同一题库的两种施测形式）应当给出接近一的重合率。**两者之差若不显著，本文为伪。**

第十节的数据提供了这个检验的一半：同一口径内部零点八以上，跨口径零点零三，差了二十四倍。但那不是两套评价装置，是一套评价装置与一份档案，所以只算半个证据。

16.3 证伪条件（八条，编号）

F1·主件。控制施测时长、题量、样本量与被试构成后，若「两套装置的重合率」与「两套装置各自选出的前两成名单的重合度」之间不存在剂量-反应关系，则本判断为伪——届时重合率只是一个统计量，不携带关于分配后果的信息。样本要求：同一制度环境内至少六个单元（同省的多个地级市、同一行业的多家企业、同校的多个院系），关键混淆变量同质；**严禁用两个国家或两个地区充数。**

F2·正向读数，写成顺序而非相关。若一次改革确实换了方向，则**重合率的下降应当早于下游能力分布的改变出现**。若观察到分布先变、重合率后降，则那是分化倒逼口径（回路效应那一路对），本文错。

F3·已跑过一次，结果部分不利。见第十节。P2 未命中，本文据此收窄了定义；粒度敏感性检验显示同一数据两种合理编码差近一点六倍，本文据此写死了清点手册。

F4 · 已跑，结果落在不确定带。 见第 10.5 至 10.8 节。实测 $\lambda(\text{新}|\text{旧}) = \text{零点五九〇}$ ，未低于零点三（故该推论未被证伪），也未达到零点六（故本文不能宣称改革什么也没换）。按预注册，这一条判为不确定，不作有利解释。**同一跑推翻了本文另一条性质（ λ 可跨领域比较），并复现了粒度效应。**

F5 · 无人知晓的口径变更。 找一次纯技术性的、未对外公告的入判口径变更（录取算法、评优权重、晋升打分表）。若下游的人选构成随之整批更换，则本文对回路效应与可读性两路不足；若不变，则本文错。

F6 · 对偶的直接检验。 在一个判定与培养共用规格的系统里，找出一项写进了培养目标却不在任何一项验收读数上的能力。若这类项存在且它们确实被稳定地培养出来了，则对偶主张为伪。

F7 · 反向。 若一套装置持续保持接近一的重合率而其毕业生在从未被考过的新情形上表现不劣于低重合率装置的毕业生，则本文关于「造不出」的主张为伪。

F8 · 第三层的独立检验。 抽查至少二十份公开的评价改革方案文本，看有没有一份报告了新旧口径的重合度或与之等价的量。若超过两成报告了，第三层主张为伪。

若被证伪，我们会学到什么。 若 F1 落空，说明分配后果由别的东西决定，那么把力气放在口径结构上是错的，应当去查分配规则本身；若 F5 落空，说明能力分布的塑造确实要经过被塑造者的自我认识，那么无人知晓的技术性改动是安全的，而这会是一条对制度设计很有用的好消息。

16.4 赌注

到二〇三二年十二月三十一日为止，至少三个国家级或大学级的评价体系，在其改革文件中公布「新旧口径重合度」这一项读数，含定义、算法与实测值。

写死什么不算命中：一，只说「新方案更全面」「更能反映综合素质」，不算。二，报告了新旧分数的相关系数但用于论证等值与可比，不算——那是用它来证明重合，不是用它来警示重合，方向相反。三，只报告新方案内部的维度数或因子数，不算，必须是新旧之间。四，由研究者事后计算而机构未在决策文件中公布的，不算。五，本文作者推动写入的，不算。

十七·反噬，与三条伦理

反噬预言。 这条判断一旦被制度采用，最省事的达标办法会摧毁它要保护的那样东西。

具体路径：若某个上级部门规定「每次评价改革必须证明重合率低于零点五」，最省事的做法不是去接入那些正交的档案读数，而是**新增一批与旧口径无关、也与任何东西无关的噪声维度**。重合率会漂亮地掉下去，而新增的方向不承载任何能力。更省事的做法是调粒度——第十节已经实测过，同一批数据换一种合理编码，读数差近一点六倍；一个想把数做低的人不需要造假，只需要选一种对自己有利的编码。

我看不到出口。 任何用来证明新方向「有内容」的检验，都必须诉诸某个效标，而效标本身来自旧口径——这正是第十三节第六条说的那个循环。本文能给的唯一底线是：**验收看事后清点的实际读数与它所依据的编码规则是否公开，不看有没有设立制度**。这不是出口，只是让作弊需要多写一份文件。

三条伦理，缺一条本文不该被拿去用。

一，**这个读数指向装置，不指向人**。任何把 λ 写进个人评语、教师考核或学生档案的做法，都超出了它的定义，见第十五节第四条。

二，**不得为了把这个数做好看，而在关键场合安排变动**。具体地说：不得为了降低重合率而在安全资格认定、法定计量、司法鉴定这类需要跨时可比的场合更换口径。这类场合的高重合率是正确的。

三，**已经据此作出不利安排的，必须公开编码规则并给出复核通道**。编码规则指第九节那份清点手册里的粒度与五条编码规定。不公开编码规则的读数，不能作为任何不利决定的依据。

十八·本文自己

自反不是说「本文也有局限」，是用本文自己的判据给本文定位，并说出这个定位的后果。本节换一个落点：不检本文的写作，检**提出这道题的那条指令**，以及**用来给本文打分的那套标准**。

第一层：这道题的问法。「教育是什么」这个问法，预设了教育是一样东西、有一个本质、可以被说出来。而按本文自己的判断，「是什么」这件事从来不是先在的属性，它随着谁在读、用什么读而定。因此本文对这道题的正确回应本来应当是先问「谁在读」，而本文没有这么做，它直接给出了一个答案。这一处不一致本文承认，只能这样处理：本文给出的不是教育的本质，是在一个判定与培养共用规格的系统里，教育这套装置在做什么——落点从对象移到了装置，而问法没有跟着改。

第二层：给本文打分的那套标准。本文是应一条要求「达到某个水平」的指令写成的，而那个水平由一套五维评分标准判定。按本文自己的读数：**这套评分标准就是一套入判口径**。它有五个维度、有权重、有封顶规则。于是本文这一篇的处境与它所描述的学生完全一样——本文在这五个维度上的表现，与本文在这五个维度之外的一切，是两个正交的方向，而只有前者入判。

这个定位有一个可以立刻算的后果：**若把这套评分标准与另一套（比如某刊物的审稿意见表）在同一批文章上做一次重构，重合率是多少？**本文没有算，因为拿不到同一批文章的两套读数。这是**本文自己欠下的那条真跑**，而它恰恰是第十节说「公开数据里几乎没有同一批对象的两套读数」的又一个实例。

第三层：本文的写法本身落在哪一格。本文有摘要、有关键词、有编号章节、有参考文献、有可证伪条款、有写死日期的赌注——**这一整套形式，是学术写作这套装置的入判口径**。本文在这套口径里做得越标准，它就越落在第八节那张表的右上格：形式审查全过，而是否张成了新方向，读者无从从形式上判断。本文对此无法自证，只能把判据交出去——第十三节的十三条分界与第十六节的八条证伪，都是为了让别人能判，而不是为了让本文显得可信。

第四层，本次修订本身就是一次可以拿本文去读的事件。这一版与初稿的差别有五处：删掉第五节那句错用对偶的话、补上第十三节漏掉的三位、加做第二次真跑、撤回第 9.3 节一条性质、多列两个洞。这五处不是作者重读时想到的，是一份**按五个维度打分的评审读数点出来的**——而那五个维度正是本文所说的入判口径。于是可以直接问本文自己那句判据：这一版与初稿，两套读数能不能互相算出来？答案是不能，因为它们本来就不是两套读数，是同一套口径读出来的一次前后。**本文这一版之所以比初稿好，恰恰是因为它更贴合那套口径；而它在那套口径之外变好了没有，本文没有办法知道。**这不是一句谦辞，这是本文核心命题在本文自己身上的一次兑现，也是它唯一一次可以被读者当场核对的兑现。

十九·结论，以及三个本文解释不了的洞

回到开头那个问题：两个人表现完全一样，他们是不是学到了同一样东西？

本文的答案不是「是」也不是「不是」，是：**这个问题问错了对象**。「一样不一样」不是那两个人身上的事实，是那套判定装置的产物；而这套装置在把两人判为同一的同时，也就规定了它此后能长出什么、长不出什么。教育不是往人身上加东西，是**制造不可分辨性**——它每成功一次，就把一批差别改写成同一格里的噪声，并在能力空间里切出一类它既读不出也造不出的能力。这一类在边界画出之前不成其为一类，边界一动，成员整批更换。

因此关于「教育是什么」，本文能给的最短回答是一句不含程度词的问话，它可以拿去问任何一份评价方案、任何一次改革结项报告、任何一套考核办法：

新旧两套读数，能不能互相算出来？

算得出来，那次改革在结构上什么也没做；算不出来，它换掉的是两批——能读的那批和能造的那批，同时。

交出三个本文解释不了的洞。

洞一·分不开那两个解释。第十节 P1 命中之后，剩下的两个候选——一般认知能力与共用的评价装置——本文的数据没有资格分开它们，本文倾向后者也没有依据。要分开，需要同一批人在两套评价装置下的读数，而这种数据现在拿不到。

洞二·说不出边界为什么会动。本文说边界一动成员就整批更换，却说不出边界什么时候会动、由什么推动。观察到的少数几次真扩口径（临床技能考核、谱系条形码、航线运行审计、以及第 10.5 节那次换代）看上去都是被外部事故、外部技术或旧口径饱和逼出来的，但本文没有系统地查过这件事，这只是印象。

洞四·GPQA 那一项不该在中间。第 10.7 节那条半错的预测。一套专门为「专家才答得出、检索不到」而造的题，本文原以为它最该张成新方向，实测它被旧读法重构了一半以上，反倒不如一套换了任务形态的题。本文给不出解释，也不打算给一个事后的：若「**新方向**」不能从题目的设计意图预测出来，那么本文关于「**什么样的改革才算扩口径**」的全部说法都少了一块。

洞五·分不开更换与萎缩。第十三节第十三条那处困难。本文预测未被覆盖的方向是被整批更换而不是萎缩，而要把这两者分开，需要一套独立于任何口径的清点手段——本文自己论证了这种手段不存在。这不是一处待办，是一处结构性的自我封锁：本文最想要的那个证据，按本文自己的说法就不该拿得到。唯一的出路可能是间接的：不去清点未覆盖的方向，而去看两次不同改革之后被覆盖方向的成员表是否互不包含。本文没有做。

洞三·没有处理师徒制。第十三节那处让步是靠「无成文规格」这一条挡回去的，而这个挡法很勉强——师徒制里判定与培养确实由同一批人承担，它长出的东西也确实从不出现在任何读数上。若有人

能说明师徒制里的「读」与「造」是怎么分开的，本文的对偶主张就该收窄一次，而不是靠一句范围声明绕过去。

注释

一、第一节所述工程学的四分解与最小实现，见卡尔曼一九六二年与一九六三年两篇；机器状态不可分辨性的原始表述见摩尔一九五六。

二、第一节所述条形码谱系追踪的结论，原文表述为姐妹细胞分析揭示出单细胞转录组测序检测不到的内在命运偏倚。本文所引为该结论，不涉及其方法细节。

三、第一节所述著作权登记，指美国版权局二〇二三年三月十六日的政策声明，以及同年九月五日复审委员会对一件数字绘画的驳回维持。本文只取「不可分辨而归类不同」这个判定形式，不对该政策本身表态。

四、第五节的限度声明是本文最要紧的一处自我限制，且本版比初稿更紧。对偶原理连的是一个系统与它的对偶系统，它给不出「同一个系统读不出即造不出」这一步；初稿把这一步当成定理的推论，本版已删，并在正文写明删的理由。本文不从该定理推出任何一步，承重全部落在「判定与培养共用一份成文规格」这条经验主张上，它由 F6 单独承担证伪。

五、第十节全部脚本、预注册文本与结果文件随文附出，数据坐标在正文中已给全，可复算。预注册在查看任何结果之前冻结，事后未改阈值。

六、第十节所报的重叠学生数为三百五十八，低于数据集原始说明给出的三百八十二，差额来自本文只保留合并键唯一的配对。

七、第十一节第四条（临床技能考核与笔试的关系）本文未核实，正文已标为待查，不作为证据使用。

八、第十节第二次真跑的预注册文本、执行偏离记录与脚本随文附出。数据为两版公开评测榜单的结果表，按模型名称匹配，共同样本三百九十七。阈值直接沿用正文原有的 F4，未因数据而改。

九、第十三节第十一至十三条是本版新增。初稿的十条分界漏掉了教育学与指标研究自己的三位正主，这是初稿最大的一处缺口，本版补上并各给一条判定形式。

参考文献

1. Beeching, E., Fourrier, C., Habib, N., et al. (2023). Open LLM Leaderboard (2023–2024 存档版) . Hugging Face.
2. 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
3. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
4. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of the 5th FUTURE BUSIness TEchnology Conference*, 5–12. 数据集: UCI Machine Learning Repository, Student Performance, <https://archive.ics.uci.edu/dataset/320/student+performance>
5. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
6. Danto, A. C. (1964). The Artworld. *The Journal of Philosophy*, 61(19), 571–584.
7. Danto, A. C. (1981). *The Transfiguration of the Commonplace: A Philosophy of Art*. Harvard University Press.
8. Eisner, E. W. (1979/1994). *The Educational Imagination: On the Design and Evaluation of School Programs*. Macmillan. (空无课程见 “The Three Curricula That All Schools Teach” 一章)
9. Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40.
10. European Union (2018). *Council Recommendation on Key Competences for Lifelong Learning (2018/C 189/01)*.
11. Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., & Wolf, T. (2024). Open LLM Leaderboard v2. Hugging Face. (2024 年 6 月换版)
12. Goodhart, C. A. E. (1984). Problems of monetary management: The U.K. experience. In *Monetary Theory and Practice* (pp. 91–121). Macmillan.
13. Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child Development*, 58(3), 539–559.
14. Hacking, I. (1986). Making Up People. In T. C. Heller et al. (Eds.), *Reconstructing Individualism*. Stanford University Press, 222–236.
15. Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.
16. Jackson, P. W. (1968). *Life in Classrooms*. Holt, Rinehart and Winston.

17. Kalman, R. E. (1962). Canonical structure of linear dynamical systems. *Proceedings of the National Academy of Sciences*, 48(4), 596–600.
18. Kalman, R. E. (1963). Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2), 152–192. doi: 10.1137/0301010
19. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
20. Kant, I. (1790). *Kritik der Urteilskraft*, § 46–47.
21. Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, 2(10), 389–398.
22. Leibniz, G. W. (1686). *Discours de métaphysique*, § 9.
23. Levinson, J. (1980). What a Musical Work Is. *The Journal of Philosophy*, 77(1), 5–28.
24. Liu, Y.-Y., Slotine, J.-J., & Barabási, A.-L. (2013). Observability of complex systems. *Proceedings of the National Academy of Sciences*, 110(7), 2460–2465. doi:10.1073/pnas.1215508110
25. Luenberger, D. G. (1971). An introduction to observers. *IEEE Transactions on Automatic Control*, 16(6), 596–602.
26. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). American Council on Education / Macmillan, 13–103.
27. Moore, E. F. (1956). Gedanken-experiments on Sequential Machines. In C. E. Shannon & J. McCarthy (Eds.), *Automata Studies*. Princeton University Press, 129–153.
28. OECD (2019). *OECD Learning Compass 2030: A Series of Concept Notes*. OECD Publishing.
29. Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
30. Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23(1), 153–158.
31. Strathern, M. (1997). ‘Improving ratings’ : Audit in the British University system. *European Review*, 5(3), 305–321.
32. U.S. Copyright Office (2023). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. 88 Fed. Reg. 16190 (Mar. 16, 2023).
33. U.S. Copyright Office Review Board (2023). *Second Request for Reconsideration for Refusal to Register Théâtre D’opéra Spatial* (Sept. 5, 2023).
34. Villaverde, A. F., Barreiro, A., & Papachristodoulou, A. (2016). Structural identifiability of dynamic systems biology models. *PLoS Computational Biology*, 12(10), e1005153.

35. Wagner, D. E., & Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21, 410–427.
36. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), eaaw3381. doi:10.1126/science.aaw3381

人机分工声明

本文的问题、三门学科的指定与目标水平由王德生提出，文献查证、结构组织、两次真跑的预注册与脚本、正文撰写由 Claude 执行，全文由王德生定稿。两次真跑的数据均为公开数据集，脚本与预注册文本随文附出，可复算。

第五节的限度声明与那处删改、第十节两跑合计五处对本文不利的结果、第 9.3 节撤回的那条性质、第十二节三处约束与第十九节五个洞，均为写作与修订过程中被材料推翻或削弱后如实写下的部分，不是事后补写的谦辞。本版相对初稿的五处改动，其来源已写在第十八节第四层，不另设修订说明。