那道题还没有被出出来

王德生 · 德麦国际 SDE 学派 · SDE教育学 · 2026年8月10日 · 约 27,000 字 · 27 条参考文献 · 三种阅读模式

摘 要

“教育是什么”这个问题，今天有三个互不相容的前沿回答。符号学一支主张教育是符号过程，看学习者产生了什么读法就够了；受布兰顿影响的推论主义主张教育是把人引入理由之网，看他能走哪些推论步就够了；统计学习理论主张学习是给定假设类与分布下的归纳，看条件配得上配不上就够了。三家对同一个学生是否”学会了”经常给出互不相容的裁定，而三家共同假定了一件从未说出口的事：学会与否是这个学习者身上的一项属性，三家只是从不同侧面去读它。推翻这条前提的材料在第三家自己手里——无免费午餐定理说，脱离目标分布，学习者的泛化能力没有内容；而假设类从哪里来，这门理论明写不在自己管辖之内。本文由此提出：教育不是把一份内容送到接收端并在那里被读出来，不是掌握一张由共同体给定的理由之网，也不是在给定条件下把泛化误差降下来，而是与学习者迄今全部表现相容的那些做法的收缩；收缩到哪一支，只在一批**尚未被任何人出出来的题**上才显形。本文把这批题命名为**界题**，并给出配套读数**分法未定率**。判据不含程度词：这个学生到目前为止的全部作答，有几种彼此不同的做法能够全部答对？本文给出四条形式命题，其中两条是新的：分辨力只依赖互异的技能要求向量，因而与题数无关；一道题分辨做法的能力随它所要求的技能数按指数下降，最大值恰在”只考一件事”的那一档。本文用一份公开三十余年、被反复分析的诊断测验数据在真实数据上实算了一次：二百五十六种技能掌握模式在二十道题上只能分成五十八组，最大的一组有六十四种模式一道题也分不开；二十道题里只有十五个互异的技能要求向量，其余五道的分辨贡献严格为零；一道只考一件事的题若被撤掉，最大那一组立刻涨到一百二十八，而该测验最难的一道题分辨力只有它的八分之一。逐题分辨力的实测与闭式逐格相等。实算同时给出两个对本文不利的结果，均写入正文，并据此收窄了本文的主张。

关键词：界题；分法未定率；做法集合；等价类；混合策略模型；教育本体

Abstract

Three incompatible frontier answers to “what is education” are currently on offer. Semiotic accounts hold that education is semiosis, and that it suffices to look at the interpretants a learner produces. Brandom-inspired inferentialism holds that education is initiation into a web of reasons, and that it suffices to look at which inferential moves a learner can make. Statistical learning theory holds that learning is induction under a given hypothesis class and distribution, and that it suffices to look at whether the conditions match. The three routinely deliver incompatible verdicts on whether the same student has learned, yet all three quietly presuppose that “having learned” is a property of the learner which they merely read from different sides. The material that overturns this presupposition belongs to the third family: the No-Free-Lunch theorem states that generalization has no content apart from a target distribution, and the theory explicitly places the origin of the hypothesis class outside its own jurisdiction. This paper therefore

argues that education is neither the delivery of content to a receiver, nor the mastery of a communally given web of reasons, nor the reduction of generalization error under given conditions, but the contraction of the set of ways of proceeding still compatible with everything a learner has done so far — and that which branch survives becomes visible only on **items that have not yet been set by anyone**. We name this class of items boundary items, with the accompanying reading rate of undetermined ways. The criterion contains no modal terms: given everything this student has answered so far, how many mutually distinct ways of proceeding would get all of it right? We prove four formal propositions, two of them new: separating power depends only on the set of distinct skill-requirement vectors, hence not on the number of items; and an item's power to separate ways of proceeding decays exponentially in the number of skills it requires, peaking exactly at single-skill items. We run the test on a public dataset analysed for over thirty years: 256 skill-mastery profiles collapse into 58 classes over 20 items, the largest containing 64 profiles that no item separates; only 15 of the 20 items carry distinct requirement vectors, the remaining five contributing exactly zero; removing a single one-skill item doubles the largest class to 128, while the test's hardest item carries one eighth of that item's separating power. Observed per-item separating power matches the closed form cell for cell. Two results unfavourable to our thesis are reported in full and used to narrow it.

Keywords: boundary items; rate of undetermined ways; separating power; equivalence classes; ontology of education; non-identifiability

目录

1. 一 · 导论：三个不该同时为真的判断
2. 二 · 三家各自说了什么
3. 三 · 冲突定位：同一个人，三种互不相容的裁定
4. 四 · 三家共有而未说出的那一条，以及推翻它的那件材料
5. 五 · 界题：一批由差决定其存在、而尚未被出出来的题
6. 六 · 分辨力：四条形式性质
7. 七 · 实算：一份公开三十余年的数据
8. 八 · 十二处兑现
9. 九 · 与最容易被混为一谈的几种说法的分界
10. 十 · 四位从不同方向逼近同一件事的先行者
11. 十一 · 命题实践里一处已经在发作的矛盾
12. 十二 · 三家为什么各自到不了这一步
13. 十三 · 三处反过来削弱本文的约束

14. 十四 · 不适用的边界

15. 十五 · 可以推翻本文的条件

16. 十六 · 伦理限制

17. 十七 · 本文对自身的定位

18. 十八 · 结论

注 释 · 参考文献

一 · 导论：三个不该同时为真的判断

期末考试结束，两份卷子摆在教师面前。两个学生都是满分，二十道题一题不错，连步骤都写得同样清楚。教师签了字，两份记录进入学籍系统，从此这两个人在所有账面上是同一个人。

三年之后，其中一个在一门后续课程上遇到一类没见过的题，几乎立刻找到门路；另一个在同一类题上反复绕。回过头查，他们当年的差别是能查得出来的：一个把分数减法理解成”先把两个数变成同一种东西再减”，另一个理解成”按位置对齐、不够就借”。二十道题里没有一道能把这两种理解分开，因为出题人从来没打算把它们分开——出题人要分开的是会与不会。

这不是一个关于测量精度的故事。把卷面分从百分制换成千分制，把阅卷改成双评，把选择题换成解答题，都不会让这两个人分开，因为这二十道题在他们两人身上给出的答案本来就完全相同。差别是真的，它有后果，它在三年后显形，而它在毕业那天是可以算出来的——只要有人去算。没有人去算，不是因为算不动，是因为学籍系统里没有那一栏。

“教育是什么”，今天至少有三个各自成体系、彼此不相容的回答，它们各自都能解释上面这个故事的一部分，而它们合起来才逼出了这个故事真正奇怪的地方。

第一个回答来自符号学。教育是符号过程：教师交出的从来不是意义，只是可被读的东西；意义在接收的那一端重新发生。要看一个人有没有学，就看他现在把世界读成什么样了。这一支在教育理论内部已经形成一个有自己名字的分支，它主张学习者本身就是一个还在生长的符号，而不是一个被填充的容器（Stables & Semetsky, 2015）。

第二个回答来自逻辑学，具体说是受布兰顿语义学影响的推论主义。教育是把人领进”给出理由与索取理由”的那个游戏：掌握一个概念，就是掌握它承诺了什么、排除了什么、可以从什么推出来。数学教育研究里明确提出，学习应当被理解为对**理由之网**的掌握，而不是对表征的构造（Noorloos, Taylor, Bakker & Derry, 2017）。要看一个人有没有学，就看他能不能在那张网上走动。

第三个回答来自数学，具体说是统计学习理论。学习是在一个假设类里、面对一个分布做归纳；能学到什么，由这个类与这个分布配不配得上决定。这一支有一条硬定理：平均到所有可能的目标上，任何一个学习算法都不比另一个好（Wolpert, 1996）。要看一个人有没有学，就看他在没见过的样本上错多少，而这个数只有相对于一个分布才有意义。

三家在很多场合可以并存，但一到裁定具体的人就分岔。一个学生把每道题都用同一个套路做对，换个说法就垮：符号学说他的读法没有变，没学会；推论主义说他给不出理由、不知道什么被排除，没学会；统计学习理论会说，“换个说法”意味着测试分布变了，那不是学习失败，是问题换了——在原分布上他的误差是零。三条裁定不能同时为真，而它们各自都不是糊涂话。

本文的路线是这样的：先把三家各自的单一主张摆清楚（第二节），再找出它们在同一个学生身上给出互斥裁定的那个结构（第三节）；然后指出三家共同假定而从未说出口的那一条，并用其中一家自己的定理把它推翻（第四节）；由此提出一个新的对象与一个配套读数（第五节），给出它的形式性质（第六节），用一份公开数据实算一次并如实报告两个不利结果（第七节）。此后是跨领域的兑现

（第八节）、与最容易被混为一谈的十种说法的分界（第九节）、四位从不同方向逼近同一件事的先行者（第十节）、命题实践里一处已经在发作的内部矛盾（第十一节），以及三家为什么各自到不了这一步（第十二节）。最后是三处反过来削弱本文的约束（第十三节）、不适用的边界（第十四节）、可以推翻本文的条件（第十五节）、伦理限制（第十六节）与本文对自身的定位（第十七节）。

二 · 三家各自说了什么

选这三家不是因为它们都对，是因为它们各自把一样东西当成了单独够用的那一样，而它们当成够用的那三样彼此不同。挑源的规矩只有一条：三家必须真的打架，且不能靠”判谁对”化解。

2.1 符号学：看它被读成什么就够了

皮尔斯的符号定义里有一个位置常被略过：一个东西之所以是符号，是因为它对某人而言代表另一个东西，而”对某人而言”这一段要靠**解释项**来兑现。解释项不是接收者本人，是接收者身上被这个符号引发的那个后续的符号。于是符号过程没有终点：每一个解释项本身又是一个符号，需要下一个解释项。

把这条用在教育上，得到的是一个相当彻底的主张：**教师从来没有传递过意义**。他交出的是可被读的东西——一段话、一个板书、一个手势；意义只在学生那一端、作为一个新的解释项发生。因此”讲清楚了没有”这个问题在原则上不由讲的人裁定。

洛特曼把这条推得更远。他关心的是文化如何产生新东西，而他的答案不是”传得越准越好”。在他看来，精确的传递只是信息的搬运，真正生出新意义的是**传不准的那一部分**——两套符号系统之间那点无法互相翻译的余量，才是新意义的发生器（Lotman, 1990）。这句话对教育的含义是刺耳的：一堂课里最有产出的，可能恰恰是学生没有按教师的意思接住的那一段。

教育符号学这一支把这些主张收拢成一个明确的立场：学习者不是被输入的接收器，他本身是一个正在生长的符号，教育就是这个生长本身（Stables & Semetsky, 2015）。

它把什么当成单独够用的那一样：学习者当下表现出的读法。看他现在把什么读成什么，就够了。至于这个读法是经由什么路径长出来的、在什么条件下才成立，都是次要的。

这一家自己知道、但不往那个方向用的一件事：解释项永远不是最后一个。既然每一个解释项还要下一个解释项，那么”他读到了正确的意思”这句话在原则上就没有一个可以停下来核对的时刻。这一支用这条来论证意义的无限生长，从没有用它来问：那么谁有资格说他读对了。

2.2 逻辑学：看他能走哪些推论步就够了

推论主义的起点是对表征主义的拒绝。一个概念的内容不是它指向的那个东西，而是它在推理中的位置：接受了它就承诺了什么，它与什么不相容，它可以从什么推出来。布兰顿把这套讲成一个社会性的实践——“给出理由与索取理由的游戏”，参与者互相**记分**：谁承诺了什么，谁有资格主张什么（Brandom, 1994）。

这套东西进入教育研究是最近十几年的事，而且进得相当具体。巴克尔与德里把它用在统计教育上，主张教学设计应当围绕”让学生的每一步都落在理由之网里”来做，而不是围绕”先给表征再给应用”（Bakker & Derry, 2011）。诺洛斯等人则把它抬成对社会建构主义的替代方案，其核心判断句可以直接引用：这一路径”以掌握理由之网的方式，发展出一个更清楚的学习概念”（Noorloos et al., 2017, p. 437）。

对教育而言，这一支最有力的地方是它给出了一个可操作的教学动作：不要问学生”会不会做”，要问他”你凭什么这么做”、“如果不是这样会怎样”、“这一步排除了什么”。它也给出了一个诊断：一个学生能做对而说不出所以然，在这套框架里不是”理解得浅”，是**根本没有掌握这个概念**——因为概念就是那张网上的位置。

它把什么当成单独够用的那一样：把学习者组织成现在这样的那条推论路径。看他能在理由之网上走哪些步，就够了。

这一家自己知道、但不往那个方向用的一件事：布兰顿的规范性是”态度设立地位”的——什么算正确的推论，不是先于实践就摆在那里的事实，而是由互相记分这个实践本身设立的。他用这条来解释规范何以有约束力而不必诉诸柏拉图式的实在，从没有用它来问：那么”这张网”到底是谁的网，以及两个人各自掌握了一张不同的网时，怎么判。

2.3 数学：看条件配不配得上就够了

统计学习理论把”学习”写成一个可以证明定理的对象。给一个假设类、一个未知但固定的分布、一批独立同分布的样本，问：能不能以高概率把泛化误差压到某个值以下。经典结果把可学性与假设类的容量绑定在一起，容量由能被打散的样本集的大小刻画（Vapnik & Chervonenkis, 1971）。

这一支最反直觉、也最重要的定理是无免费午餐定理：如果不对目标做任何限制，那么平均到所有可能的目标上，任何算法都不优于任何别的算法，包括不优于乱猜（Wolpert, 1996；亦见 Shalev-Shwartz & Ben-David, 2014, Thm 5.1）。它的推论是硬的：**一切泛化能力都来自事先放进去的限制**，也就是归纳偏置。

教科书对这条的处理很坦白：定理告诉你在给定假设类之后会发生什么；假设类从哪里来，被称作”先验知识”，明确地不在这门理论的管辖之内（Shalev-Shwartz & Ben-David, 2014, Ch. 5）。

它把什么当成单独够用的那一样：学习赖以成立的那个条件场——假设类、表示、分布。看条件配不配得上，就够了。学习者”怎么想的”、“读成什么”，在界的推导里不出现。

这一家自己知道、但不往那个方向用的一件事：正是无免费午餐定理本身。它说的是泛化能力不是算法的属性，是**算法与分布这一对**的属性。这条被用来论证”必须有归纳偏置”，从来没有被用来说：那么”这个学生学会了”这句话，作为一个只提到学生的句子，是没有内容的。

2.4 三家为什么不能靠判谁对来化解

如果三家是在同一个问题上给出相反的答案，那么摆证据、判谁对就行了，不需要新框架。但它们不在同一个语域里作答：一家谈的是当下表现出的读法，一家谈的是把它组织成这样的那条路径，一家谈的是它赖以成立的那片条件。三条各自都不能被另外两条驳倒，因为它们回答的是不同的问句。

正因为不能判谁对，唯一的出路是造一个能把三者关联起来的说法。这也是本文接下来要做的事。

三 · 冲突定位：同一个人，三种互不相容的裁定

三家真正打起来，是在要对一个具体的人下一个是非判断的时候。下面三种学生在实际课堂里都不罕见，三家对每一种的裁定各不相同。

第一种：能做对，读法没变。 一个学生把每道分数减法都做对了，方法是”先通分、再减分子、最后约简”，一步不错。问他”为什么要通分”，他说”因为要通分”。让他把 5/3 − 3/4 画成一根数轴上的两段，他画不出来。

- 符号学：他产生的解释项与三个月前没有差别，他仍旧把分数读成”上下两个数”。**没学会**。
- 推论主义：他做不出任何一步说明性的推论，给不出理由，不知道什么被排除。**没学会**。
- 统计学习理论：在这一类题构成的分布上，他的误差是零。你说”换个表述他就垮”，那意味着你在另一个分布上测他，那不是学习失败。**学会了**。

第二种：说得出理由，做不对新题。 另一个学生能把通分讲得头头是道，会说”分母不同就不能直接减，因为它们的单位不一样”，也知道约简是为了什么。但把题换成 4又1/10 − 2又8/10，他卡住了，因为要借位。

- 推论主义：他在理由之网上走得动，能承诺、能辩护、能说出不相容项。**学会了**。
- 统计学习理论：在包含借位的分布上他的误差很高。**没学会**。
- 符号学：他的读法确实变了——他现在把分数读成”带单位的量”。**学会了，只是没走完**。

第三种：读法变了，还说不出、也还做不快。 第三个学生忽然说了一句”哦，分数就是把一根线剪成一样长的段”，此后他解题变慢了，因为他开始每道题都想一想；正确率暂时下降。

- 符号学：这是本轮最重要的一次事件，一个新的解释项出现了。**学会了**。
- 统计学习理论：他的经验误差上升了。**退步了**。
- 推论主义：他还不能把这条讲成理由、还不能用它作辩护。**尚未学会**。

三家不是”侧重不同”，是**对同一个人给出了互不相容的归属**。而更要紧的是它们打架的那个共同对象：三家争的其实是**“同一道题的又一次”该怎么算**。统计学习理论说”换个表述就是换了分布，那不是同一道题”；另两家说”那当然是同一道题，只是换了说法”。谁对，不能由证据裁定，因为”算不算同一道题”正是各自框架的前提，不是各自框架的结论。

于是问题就位了：这三家争的是**什么算同一件事的又一次**，该由谁裁、按什么算。

四 · 三家共有而未说出的那一条，以及推翻它的那件材料

4.1 那条前提

把第三节的三种裁定并排看，会发现一件三家从没争过的事。它们争的是**该看哪一维**——看读法、看推论、看条件；它们从没争过一件更靠前的事：

> **“学会了”是这个学习者身上的一项属性，只不过三家各自从一个侧面去读它。**

这条前提在三家的行文里都不出现，因为它不需要出现。符号学问”他现在读成什么”，这是在读他；推论主义问”他能走哪些步”，这是在读他；统计学习理论问”他在未见样本上错多少”，这也是在读他。三家像三台仪器围着同一个对象，争的是哪台仪器更靠得住。

把这条念给三家听，三家都会说”这还用说吗”。这正是它是共有前提的标志：会引起争论的东西不是共有前提，是第四家的立场。

这条前提有一个直接的制度后果，而这个后果我们每天都在用：**成绩是可以归属给个人的**。一份成绩单上写的是”张三，92 分”，不是”张三与这批题，92 分”。学籍系统、升学系统、招聘系统全都建立在”能力可归属给个人”这条上。

4.2 推翻它的那件材料，在第三家自己手里

推翻一条共有前提，从外面搬一个理由来是不算的——那只是加入争论，成了第四家。要取消争论，用的材料必须是三家之一自己已经掌握、已经发表、只是没往这个方向用的。

这件材料是无免费午餐定理。

它的内容可以不带任何技术地讲清楚：**如果对目标不作任何事先限制，那么任何一个学习办法在所有可能的目标上平均起来，都不比任何别的办法好**。 一个办法之所以好，只能是因为它事先假定的东西恰好与这一次的目标合得上。

推论是硬的：**“这个学习者的泛化能力”这个短语，作为一个只提到学习者的短语，没有内容**。 有内容的是”这个学习者—这个目标分布”这一对。换一个分布，同一个学习者的能力读数可以从最高变成最低，而他本人一点没变。

统计学习理论完全知道这一条，而且知道得比谁都清楚。它用这条来论证归纳偏置的必要性，用来警告”不要指望有普适的学习算法”。它从来没有用这条去说：那么**在教育里，“学会了”这句只提到学生的话，也是没有内容的**。它没有这么说，是因为它回答的是另一个问题——它问的是”什么条件下泛化界成立”，不是”教育是什么”。

这就是这件材料的力量所在：它不是我们从外面找来反驳三家的，它是三家之一自己的定理。符号学与推论主义无法退回自己的阵地说”那是数学的说法”，因为它们各自在裁定学生时用的正是同一个句式——一个只提到学生的句式。

4.3 于是前提被改写成什么

前提被推翻之后，剩下的不是”学会与否不可知”这种虚无的结论。剩下的是一个更具体的东西：

> **“学会了”是一个关于（学习者，一批任务）这一对的判断。而在教育实践里，那”一批任务”永远是已经出出来的那些题；没有出出来的那些，不进入这个判断。**

一份成绩单看上去是关于人的，实际上是关于**人与一份题目清单**的。换一份清单，同一个人的读数会变，而且不是变得”更准”或”更不准”，是**变成了另一个判断**。

这一步一走出来，第一节那两份满分卷子的怪异之处就有了位置：两个人在这二十道题上不可分辨，不是因为测得不准，是因为**这二十道题在他们两人身上的分辨力为零**。而分辨力是题目与做法这一对的性质，不是学生的性质。

4.4 第二层：所有人共同站着的那块地

上面那一条是三家共有的。还有一层更宽的，要从最容易与本文混淆的那些既有说法里逼出来——逐个问它们”结构性地看不见什么”，然后看这些看不见的东西背后是不是同一块地。

- **项目反应理论**把”同一作答模式即同一能力”当作给定，因此看不见同分而不同法。一旦承认它，原始分作为充分统计量的地位就塌了，而那是整套等值与量表化的地基。
- **认知诊断的属性模式**把”一种做法就是它所需技能的集合”当作给定，因此看不见”所需技能相同而使用次序不同”的两种做法。一旦承认它，Q 矩阵这个装置本身就不足以刻画做法，因为它只记要哪些技能，不记怎么用。
- **推论主义**把”理由之网由共同体给定”当作给定，因此看不见两个人各自掌握了一张不同而都自洽的网。一旦承认它，“掌握理由之网”就不再是一个可裁定的成就。
- **符号学**把”意义在解释项里发生”当作给定，因此看不见”没有任何人取用而差别已经存在”的情形。一旦承认它，符号过程就不是意义的唯一发生处。
- **统计学习理论**把”假设类给定”当作给定，因此看不见表示的选择本身就是学习的主要部分。一旦承认它，全部泛化界的条件句前件就落空了。
- **软件工程的变异测试**把”原程序是对的、变异体是错的”当作给定，因此看不见两个都对的实现。一旦承认它，变异得分就不再是测试套件质量的读数。

这六行背后是同一块地：

> **判对错的装置，同时被当成了判异同的装置。**

因为对错是二值的，所以”都对”被读成”一样”。这句话所有人都站在上面，包括互相批判的各派——批评标准化考试的人也站在上面，因为他们的处方是”换一套更好的题去测”，而更好的题仍然是用来分对错的。谁也没法回头看它，不是因为没人看，是因为看的人都站在上面。

五 · 界题：一批由差决定其存在、而尚未被出出来的题

5.1 命名

设一个题域（比如”两位数分数减法”），设一份**做法清单** R——该题域内被公认为可行的那些做题办法，每一种都是一条完整的处理次序。对任一道题 i 与任一条做法 r ∈ R，做法 r 在题 i 上给出一个确定的答案 a(r, i)。

设一个学习者到目前为止做过的题构成集合 I，他的作答构成向量 x。定义

R(I, x) = { r ∈ R : a(r, i) = x_i 对所有 i ∈ I }

即**与他迄今全部表现相容的那些做法**。教育实际做的事，是让这个集合收缩。

现在定义本文的对象。对 R(I, x) 中任意两条做法 r, r′，令

B(r, r′) = { i ∈ I : a(r, i) ≠ a(r′, i) }

B(r, r′) 里的题，就是界题。

界题不是难题——它可能极其简单。不是偏题——它可能就在考纲正中。不是知识点——它不对应任何一条要求。不是错题——两条做法在它上面都可能给出正确答案中的一个，或者一条对一条错，但这不是它成为界题的理由。它成为界题的唯一理由是：**这两条做法在它上面分岔**。

要紧的是它的存在方式。界题不在任何题库里，不在任何考纲里，没有人出过它。它由两条做法之间的差**定义**出来，而不是从一个现成的清单里挑出来。撤掉这个定义，它不是变得难找，是**不存在**——它不是一批被遗漏的题，它是一批还没有被造出来的题。

一个具体的例子。第一节那两个满分学生，一个把分数减法处理成”化成同一种东西再减”，一个处理成”按位对齐、不够就借”。它们的界题里有这样一道：

> 2 − 1/3 = ?

第一条做法要先把 2 写成 6/3，第二条做法要从整数部分借一个 1 变成 1又3/3。两条都能得到 1又2/3，所以这道题**不是**界题——除非把要求改成”写出中间那一步”。改成写中间步骤之后，它立刻成为界题，而且只要一道就够。这说明界题不是题目本身，是**题目加上什么算作作答**这一对。

5.2 读数：分法未定率

u(I, x) = |R(I, x)|/|R|

分法未定率：与学习者全部作答相容的做法数，占该题域做法清单总数的比例。

它满足一个可用读数该有的四条：无量纲，因此在不同学科、不同年级之间可以并排比较；可事后清点，只要有清单和一张做法与题目的对照表，任何人都能重算；它把”这孩子到底懂了没有”这种印

象变成一个数；**而第四条——与既有主要指标正交——本文在第七节里发现它在现有唯一可算的编码下并不成立，那一节会详细交代这处失手，以及它把本文的主张改成了什么。**

配套的清点手册必须写死，否则读数不是读数：

读数名：分法未定率　符号：u 定义：u = |与学习者迄今全部作答一致的做法| ÷ |该题域做法清单总数| 粒度：什么算"一条做法"——一条从题面到答案的完整处理次序；两条做法不同，当且仅当存在至少一道该题域内的题，两条在其上给出不同的作答（含中间步骤，若中间步骤被计入作答）边界例：先约简再通分 vs 先通分再约简 —— 在只记最终答案时是同一条，在记中间步骤时是两条。粒度由"什么被记为作答"决定，必须先写死。　编码规则：　① 做法清单由该题域的教学文献与错误分析文献汇编，不得由单个编码者临时增补；② 一条做法必须能在该题域全部题目上给出确定作答，给不出的不入清单；③ 只在部分题目上有定义的做法，按"该子题域"单独立表，不与全域表混算；④ 学习者的作答按原始记录编码，不得据"他大概是这么想的"回填；⑤ 相容判定用严格相等，不用近似。 表头：学习者 | 已做题号 | 作答向量 | |R(I,x)| | |R| | u | 最大不可分辨组大小 不适用：做法清单不存在的题域；作答只记对错而不记内容的记录。 本篇三处引用一致性：正文 §5.2 = 实算 §7 = 证伪 §15，同一定义、同一粒度。

5.3 判据：一句不含程度词的问话

> **这个学生到目前为止的全部作答，有几种彼此不同的做法能够全部答对？**

这句话里没有”应当”“充分”“真正”“实质性”。它可以直接拿去问一份题库、一份作答记录、一份评分细则，而不必去问任何人的判断。

把它在五个场景里跑一遍，五个答案互不相同，其中两个是查出来的、而且反过来改了本文：

1. **分数减法期末卷（第七节实算）**：答案是一个可以算出来的整数。落在最大那一组里的学生，二百五十六种技能掌握模式中有六十四种与他的作答完全一致。**这是查出来的。**
2. **编程课的自动评测**：答案是”通过全部测试用例的实现有多少种”。这个数在软件工程里有现成的近似办法，叫变异得分——它衡量的正是一套测试能分开多少种不同实现（DeMillo, Lipton & Sayward, 1978）。**这是查出来的，而且它改了本文**：本文原以为这个读数在任何行当里都没有实现，事实是软件工程用了将近五十年。
3. **驾照路考**：答案接近于”很多”。路考记录只有扣分项，没有做法项，因此与一份满分记录相容的驾驶方式在原则上无法清点。这一格给出的不是一个数，是”清单不存在”。
4. **医学的鉴别诊断训练**：答案偏小。这是本文找到的唯一一个做法清单**事先就存在**的行当——鉴别诊断表本身就是一份做法清单，而住院医的训练明确地以”排除到剩下几个”为目标。
5. **一节语文课上的课文理解**：答案是”清单不可能穷举”。这一格暴露了读数的适用边界，见第十四节。

第二个与第四个是本文没有预料到的：**这个读数在两个行当里已经是标准工序，而在教育的主干里连字段都没有。**　这一发现改变了本文对自身贡献的定位——本文补的不是一个没人想过的想法，是把两个行当里的成熟工序搬到一个至今没有它的位置上，并说明为什么那个位置一直空着。

5.4 一张两轴的辨别格

只有一条轴的东西是一个名字，不是一个辨别维度。第二条轴取”卷面表现”，第一条轴取”剩余做法”。

	剩余做法已收窄到一支	**剩余做法仍有多支**
全部作答正确	**已定之会**：会做，且做法唯一确定。这是教学想要的结果。	**未分之会（本文要打的那一格）**：全部形式审查通过，卷面无可挑剔，而”他到底在用哪条做法”从未被任何一道题分开。
有错	**已定之误**：一个确定的错规则。这一格是最容易教的——错误本身把做法钉死了。	**在摸**：既有错，做法也未定。这是学习中途的正常状态。

这张格子里最要紧的是右上那一格。它必须满足三个签名，否则第二条轴就选错了——若要打的那一格是一眼就能看出的坏，那不需要任何框架，谁都看得见。

签名一：它在短期指标上表现为改善。 未分之会的学生卷面全对，班级平均分因他上升，教师的教学评估因他好看。没有任何一个现行指标会在这一格上报警。

签名二：它的失效不产生任何信号。 差别要到界题被出出来的那一天才显形，而那一天可能永远不到；即便到了，那时的失败会被归给”后续课程太难”或”这孩子后劲不足”，不会被归给三年前那份满分卷。

签名三：它自我加固。 这是三条里最要紧的一条。题库越标准化、越对齐考纲、越经过预试筛选，它的题目就越是被挑成”会的人答对、不会的人答错”——而这恰恰是在挑那些**对不同做法给出相同答案**的题。第十一节会指出，这不是命题实践的疏忽，是它的目标本身。

三条签名齐了，右上格才成立。它成立的意思是：**未分之会不是失职的产物，是尽职的产物。**

六 · 分辨力：四条形式性质

本节把上面那些话写成两条可以检验的命题。用的模型是教育测量学里最常用的合取型模型，它假定一道题要答对，必须具备该题所要求的全部技能；每道题要求哪些技能，写在一张矩阵里。

设有 K 项技能、J 道题，矩阵 Q = (q_jk)，q_jk = 1 表示第 j 题需要第 k 项技能。一个学习者的技能掌握模式记作 α ∈ {0, 1}^K。**理想作答向量**为

$$\xi_j(Q, \alpha) = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$$

即”具备该题全部所需技能则答对”。两个模式**可分开**，当且仅当它们的理想作答向量至少在一道题上不同。

定义一道题 j 对一个做法集合 R 的**分辨力**为：它把 R 划分成的作答等价类的数目减一，除以 |R| − 1；分辨力为零，意思是这道题对 R 里的所有做法给出同一个答案。

命题一（并列正确做法的分辨力为零）

> 设两条做法 r, r′ 在题目集合 I 上的作答完全相同。则 I 中每一道题对 {r, r′} 的分辨力都是零，且这一结论与 I 的大小无关。

证明是同义反复级的，但它的含义不是。它说：**加题不解决问题**。　　如果两条做法在一批题上作答一致，那么把这批题从二十道加到二百道、从一次考试加到一学期的全部作业，只要新加的题仍然落在”两条做法作答一致”的范围内，分辨力就仍旧是零。分辨力不随题量增长，它只随**题的选取**变化。

这条对合取模型有一个尖锐的特例。设某一条做法对应”具备全部 K 项技能”，另一条做法对应另一套技能划分下的”具备全部技能”——也就是两个各自精通一套不同解法的学生。两人的理想作答向量都是全 1。于是：

> **两个各自精通一套不同解法的学生，在任意多道题上都无法被分开**。分辨力为零，与题量无关。

分数减法这批题确有多套解法，教育测量学自己早已指出并专门建模（de la Torre & Douglas, 2008）。而在只记对错的作答编码下，这两个人是同一个人。

命题二（完备性的代价）

> 在合取模型下，矩阵 Q 能分开全部 2^K 个技能掌握模式，当且仅当 Q 的行向量里包含全部 K 个单位向量——也就是说，测验里必须为每一项技能各配一道**只考它一件事**的题。

这条是教育测量学的既有结果（Chiu, Douglas & Li, 2009；证明与推广见 Köhn & Chiu, 2017）。本文要指出的是它的一个后果，那个后果在原文献里没有被读出来：

唯一能把做法分开的题，是最”简单”的那种题。 一道只考一件事的题，在真实考生总体上通过率很高、区分度很低，而经典项目分析里”区分度低”正是剔除的标准理由。于是测验的品质控制程序与测验的分辨能力之间，存在一个方向相反的作用。下面两条把这个”方向相反”从一句观察变成一个可以算的量。

命题三（分辨力只依赖不同的题型，不依赖题数）

> 在合取模型下，两个技能掌握模式可分开，当且仅当它们所支配的**不同技能要求向量**的集合不同。因此：**两道技能要求向量完全相同的题，第二道对分辨力的贡献严格为零。**

证明只需一步。理想作答向量 $\xi(Q, \alpha)$ 完全由集合 $S(\alpha) = \{ q \in R_Q : q \leqslant \alpha \}$ 决定，其中 R_Q 是 Q 的**互异**行向量之集；两个模式的理想作答向量相同，当且仅当 $S(\alpha) = S(\alpha')$。于是全部分辨结构只由 R_Q 决定，与每个行向量重复了多少次无关。**证毕。**

推论一：等价类数不超过 $2^{|R_Q|}$，与题数 J 无关。推论二（更实用）：**一份测验里，凡技能要求向量与已有题目相同的题，无论出多少道，都不增加任何分辨力**——它们只增加信度，不增加分辨。第七节会报告这一条在实测数据上的一个相当难看的数字。

命题四（逐题分辨力的闭式，及它的单调性）

设一道题要求 m 项技能（$m \geqslant 1$），则在 K 项技能、2^K 个模式的全体中，它能分开的模式对数恰为

$$\mathbf{N(m) = 2^{(K-m)}(2^K - 2^{(K-m)})}$$

且 $N(m)$ 在 $m \geqslant 1$ 上**严格递减**；$N(0) = 0$。

证明：能答对该题的模式，是那些在该题所要求的 m 个位置上全取 1 的模式，共 $2^{(K-m)}$ 个；答不对的有 $2^K - 2^{(K-m)}$ 个。该题分开的恰是”一个答对、一个答不对”的那些对，故其数目为二者之积。令 $x = 2^{(K-m)}$，则 $N = x(2^K - x)$，在 $x < 2^{(K-1)}$ 时随 x 递增；$m \geqslant 1$ 时 $x \leqslant 2^{(K-1)}$，故 m 增大（x 减小）时 N 严格减小。$m = 0$ 时 $x = 2^K$，$N = 0$。**证毕。**

这条闭式把第六节末尾那句话变成了一个数量关系，而且方向是硬的：

> **一道题分辨做法的能力，随它所要求的技能数按指数下降；最大值恰好落在 $m = 1$，也就是”只考一件事”的那种题上。**

$m = 0$ 的那一档也值得看一眼：一道谁都会做的题分开的对数是零。**这与直觉一致。**真正反直觉的是它的另一端——一道要求五项技能的难题，分辨力只有一道单技能题的八分之一（当 $K = 8$ 时）。**难题不是更有信息量的题，是更没有信息量的题**，只要问的是”他用的是哪条做法”而不是”他会不会”。

命题四同时说明了为什么第七节的技能层面结果会是那个样子：一项技能读得出来还是读不出来，几乎完全由它有没有一道 m = 1 的专属题决定，而 m = 1 的题恰是命题实践最不看重的那一类。

七 · 实算：一份公开三十余年的数据

本文不满足于纸面上的推导。本节用一份任何人都可以下载复算的公开数据，把最便宜的那条检验跑一遍，并如实报告两个对本文不利的结果。

7.1 数据与预注册

数据是塔苏卡的分数减法测验：二十道题、八项技能（Tatsuoka， 1990）。它是认知诊断领域被分析得最多的一份数据，三十余年里被反复用于模型比较。所用的技能矩阵取德拉托雷与道格拉斯的版本（de la Torre & Douglas，2004），逐字取自张、德卡洛与英的论文表十三（Zhang，DeCarlo & Ying，arXiv:1303.0426）。八项技能是：把整数化成分数；把整数部分与分数部分分开；先约简再相减；通分；从整数部分借位；竖式借位；分子相减；把结果化为最简。

统计之前先把假设与阈值写死存档，事后不得调整：

- **H1**：该矩阵不完备，即存在不同技能掌握模式产生完全相同的理想作答向量。判定：多成员等价类数大于零。
- **H2**：存在”只由一道题分开”的模式对；移除该题，等价类数严格下降。判定：至少一对。
- **H3**（本文最愿意押、也最可能失手的一条）：不可分辨不只发生在低掌握一侧。在掌握技能数不少于六项的模式中，落在多成员等价类里的比例大于百分之二十。判定：不大于百分之二十则本条为伪。
- **H4**：在合取模型下，理想作答全对的技能掌握模式只有一个。这一条若成立，说明”全对”在这套编码里等同于”唯一做法”，那是编码规则的后果而不是关于学生的发现。**这一条成立对本文不利，成立也必须写进正文**。

计算是确定性的：列出全部 2^8 = 256 个模式，算出每个模式的理想作答向量，按向量分组。脚本四份（分划、补完备、正交性自检、两条命题的核验），可复算。

7.2 结果

H1 成立。 二百五十六个技能掌握模式，在这二十道题上只能分成 **五十八** 个等价类：三十二个是单元素的，二十六个含多个模式；**最大的一组含六十四个模式**。这个数字与已发表的结果一致，本次是一次独立复现。

换成大白话：**这份被用了三十多年的测验，把二百五十六种可能的会法压成了五十八种可分辨的结果**。 落在最大那一组里的学生，有六十四种彼此不同的技能掌握情形与他的作答完全一致，而这二十道题一道也分不开它们。

分辨力的分布。二百五十六个模式两两配对共 32,640 对：

- **4,224 对（12.9%）** 二十道题一道也分不开；
- **4,600 对（14.1%）** 只被一道题分开；

· 其余的对被两道及以上分开。

也就是说，**超过四分之一的做法对，靠这份测验只有零道或一道题在支撑其可分辨性**。

H2 成立，而且比预期严重。 第九题（3又7/8 − 2，只需要”把整数部分与分数部分分开”这一项技能）**独家分开了 4,096 对**——没有任何别的题能替它。撤掉这一道题：

> 等价类 58 → 57，**最大的那一组从 64 个模式涨到 128 个**。

一道题。这道题在真实考生上几乎人人做对，是任何一份项目分析报告里都会被标为”过易、区分度低”的那种题。**测验一半的分辨结构挂在一道随时可能被删掉的题上**。

H3 成立。 不可分辨不是差生专有：

掌握技能数	模式个数	其中不可分辨	比例
⩾ 5 项	93	67	**72.0%**
⩾ 6 项	37	21	**56.8%**
⩾ 7 项	9	3	33.3%

掌握了六项以上技能的模式里，超过一半仍与别的模式不可分辨。这条对本文重要：它说明”分不开”不是一个只在起步阶段出现的现象。

技能层面，最刺眼的一条。 逐项算”该技能读不出来”的模式占比：

技能	读不出来的模式占比
分子相减	0.0%
把整数部分与分数部分分开	0.0%
把整数化成分数	50.0%
通分	50.0%
把结果化为最简	50.0%
竖式借位	68.8%
从整数部分借位	75.0%
先约简再相减	**87.5%**

两个读得完全清楚的技能，恰恰是这份测验里配了”只考它一件事”那种题的两项——第九题只考”分开整数与分数”，第六、第八题只考”分子相减”。**可读性完全由有没有一道专属的题决定，与那项技能本身有多重要毫无关系**。

而读得最不清楚的那一项是”先约简再相减”，八项里唯一一项**不是必做步骤而是一个选择**——做了省事，不做也对。八项技能中最像”做法”而不像”工序”的那一项，正是这份测验百分之八十七点五读不出来的那一项。

命题三的实测：五道题的分辨贡献恰好为零。 这二十道题里，**只有十五个互异的技能要求向量**。第三题与第二题、第八题与第六题、第十六题与第十四题、第十七题与第十一题、第二十题与第四题，各自完全相同。按命题三，去掉这五道题，分辨结构应当分毫不变。实算的结果正是如此：

> 二十道题：等价类 58，最大组 64。**只留十五道题：等价类 58，最大组 64。**

也就是说，**这份测验四分之一的题目，对”他用的是哪条做法”这个问题的贡献严格是零**。它们对信度有贡献——重复测量降低随机误差；对分辨力没有任何贡献。这不是命题人的失误，重复设题在测量学上是正当做法；它只是说明，**信度与分辨力是两个可以背道而驰的东西，而现行的品质控制只管前一个**。

命题四的实测：闭式与实测逐格相等。 把每道题实际分开的模式对数与闭式 N(m) = 2^(K − m)(2^K − 2^(K − m)) 比对：

该题要求的技能数 m	闭式 N(m)	实测分开的对数	本测验中的题号
1	16,384	16,384	6, 8, 9
2	12,288	12,288	2, 3, 12, 14, 15, 16
3	7,168	7,168	1, 7, 11, 17
4	3,840	3,840	4, 5, 10, 13, 18, 20
5	1,984	1,984	19

逐格相等，无一例外。**这份测验里最”难”的那道题（第十九题，要求五项技能）的分辨力，是最”易”的那三道题的八分之一。** 而在任何一份常规的项目分析报告里，第十九题会被表扬，第六、第八、第九题会被列为候删。

补完备的代价。 要让这份测验能分开全部二百五十六种模式，按命题二至少还要再加**六道只考一件事的题**（二十道加到二十六道）。补齐后等价类从五十八升到二百五十六，最大组从六十四降到一。六道题换来分辨力翻两番——而这六道题恰恰是命题实践最不愿意放进正式测验的那一类。

7.3 两个不利结果，以及它们改了本文什么

不利结果一：H4 成立。 在合取模型下，理想作答全对的技能掌握模式**只有一个**。也就是说，在这套编码里”全对”等同于”唯一确定的做法”。

这直接顶到了本文第五节那张辨别格的右上角——**“未分之会”这一格在这套编码里是空的。**

本文不打算含糊过去。这个结果说明的不是本文错了，而是**这套编码装不下本文的对象**。合取模型把”一种做法”定义成”掌握的技能子集”，那是一个只承认缺失的本体：两个学习者的差别，在这套编码里只能是”谁少了什么”。一个什么都不缺的人，按定义只有一种。**“两个人都会做、而做法不同”在这套编码里不是难以测量，是无法表达**。

于是本文原来打算写下的那句更强的话——“分法未定率可以直接在现有诊断数据上算出来，且全对的学生剩下的做法最多”——被这次实算削掉了。在现行编码下，这句话是假的，而且不可表达。

不利结果二，比第一个更重。 本文对读数做了正交性自检：分法未定率与卷面得分是不是同一个数的影子？在这二百五十六个模式上算：

> **皮尔逊相关 r = −0.792；斯皮尔曼秩相关 ρ = −0.932**。

按理想得分分组，全 0 作答的学生 u = 0.2500（64/256），全对的学生 u = 0.0039（1/256），中间几乎单调下降。**在唯一有公开数据可算的编码里，分法未定率不是一个独立读数，它是分数的一个近乎单调的影子**。

一个与既有主要指标高度相关的新读数，会被立刻读成那个指标的一个代理，白造。本文必须承认这一点，而不是绕开它。

但这次的原因是可以说清楚的，说清楚之后它反而支持本文的主张：**负相关是编码造成的，不是对象造成的**。 在只承认缺失的本体里，做法数当然随掌握数单调下降——因为”做法”被定义成”技能子集”，掌握得越多，子集的自由度越小。要让分法未定率与分数正交，做法清单里必须包含**并列正确的做法**：同样全对而次序不同、表征不同、依据不同的那些。而那样的清单，在现有的任何一份公开教育数据里都不存在。

这两个不利结果合起来，把本文的主张改成了下面这一句，而这一句比原来那句硬：

> 不是”教育的记录看不见剩余做法”，而是**教育的记录里没有它的位置**——现行编码把”做法”定义成”缺了什么”，因此”两条都对而不同”在其中不是一个可以为真或为假的命题，是一个说不出来的句子。

这正是界题作为一个对象的判据：撤掉这个说法，它不是变得难测，是不存在。

7.4 这份数据回答不了的一件事

必须写清楚：本节的全部计算都是在**理想作答**上做的，也就是不考虑失误与蒙对。真实作答里有噪声，噪声会让”可分开”变成”分得开但要更多样本”，而”不可分开”仍旧是不可分开——因为没有任何样本量能弥补一个恒等于零的差别。所以噪声只影响本节结论的一半：**它使本节报告的不可分辨比例成为下界，不是上界**。

另有一件本节完全无力回答的：本文真正关心的是并列正确的做法，而这份数据的编码只记录缺失。要检验本文最核心的那条，需要一份**做法清单在先、作答含中间步骤**的数据。据本文所知，这样的公开数据目前不存在。造出它的最小设计写在第十五节。

八 · 十二处兑现

一个说法要算数，不能只是”像”，得能在别的行当里据此预测出一件具体的事。下面十二处，每一处给出预测，不给类比。

一 · 软件工程的变异测试。 把程序当学生、测试套件当考卷：变异测试故意在源码里制造小改动，看测试能不能发现。杀不死的变异体，就是”两个不同的实现通过了同一套测试”。变异得分正是分法未定率的补（DeMillo, Lipton & Sayward, 1978；Jia & Harman, 2011）。**预测**：任何一个只按覆盖率验收的项目，其存活变异体比例应显著高于按变异得分验收的项目，即使两者行覆盖率相同。这一条已被大量实践证实，本文借它说明读数可行，不主张原创。

二 · 大语言模型的评测。 评测集衡量的是”在这些题上对不对”，而两个内部机制完全不同的模型可以在整个评测集上给出相同答案。对照集这一手法正是人工制造界题：对原样本作最小改动使标签翻转，看模型跟不跟得上（Gardner et al., 2020）。**预测**：在任一公开评测集上并列前两名的模型，其在对照集上的分歧率应显著高于它们在原评测集上的分歧率；差额越大，说明该评测集的分辨力越低。这个数现在就能算。

三 · 医学的鉴别诊断。 鉴别诊断表就是一份做法清单，而”再做哪项检查”这个决定，就是在选一道界题——选那项能把当前还剩的几个诊断分开的检查。**预测**：住院医培训中，“排除到剩几个”这个读数应比”最终诊断正确率”更早地区分出后来成为好医生的人；正确率的区分要等到罕见病例出现，而剩余诊断数每一次问诊都能读。

四 · 飞行员的模拟机复训。 复训科目是固定的，因而两个各自形成了不同应对习惯的机长可以在全部科目上表现相同。**预测**：在标准科目外临时插入一个未列入大纲的组合故障，两名标准科目全优的机长，其处置次序的分歧率应显著高于零；分歧率越高，说明标准科目的分辨力越低。

五 · 司法的量刑一致性。 两位法官在全部已判案件上量刑相同，不意味着他们用的是同一条裁量思路。**预测**：把两位一致率极高的法官的历史判决输入一个只含边界情节的假设案件集，其分歧率应远高于历史一致率；这个集合就是他们之间的界题。

六 · 招聘的结构化面试。 结构化面试提高了信度，其做法是把问题固定下来。**预测**：结构化程度越高的面试流程，两位评分完全一致的面试官对同一候选人后续绩效的预测分歧越大——一致的是打分，不是判断。

七 · 工业的良率与工艺。 两条产线良率相同、检验项全过，工艺参数的组合可以完全不同。**预测**：当原材料批次发生一次未预告的变动，两条历史良率相同的产线的失效率应出现显著分岔；变动前的检验项对这个差别的分辨力为零。

八 · 体育的技术评分。 体操与跳水的评分表是一份对错清单加扣分表。**预测**：两名在现行评分表下得分相同的运动员，其在动作被规则修订后的适应速度应出现系统性差别；评分表越细，这个差别越读不出来。

九 · 语言测试。 一份四选一的阅读理解，两个用完全不同策略答题的考生（一个读文本、一个做排除）可以全对。**预测**：在题干保持不变、仅打乱选项内在逻辑关系的一份平行卷上，两名原卷同分的考生分数分岔应显著；这是最容易做的一次界题实验。

十 · 教师资格的课堂观察量表。 量表把课堂行为拆成可勾选的条目。**预测**：两位在同一份量表上得分相同的教师，其学生在下一学年遇到新类型任务时的表现应出现分岔；量表条目数越多，反而越读不出这个差别，因为条目是按”做没做”而不是按”怎么做”设的。

十一 · 农业的栽培规程。 两个按同一份规程操作、产量相同的农户，其应对一次反常气候的处置可以完全不同。**预测**：反常年份的减产率差异应远大于常年产量差异；规程本身对这个差别的分辨力为零。

十二 · 本文所处的这一类工作。 两篇在全部现行评审项上得分相同的论文，其被后续研究实际使用的方式可以完全不同。**预测**：同一批同分论文的引用**用途**分布（作为方法、作为反例、作为背景）分歧度，应远高于其评审分分歧度。这一条本文自己也逃不掉，见第十七节。

反向的一处必须写明：**并非所有行当都缺这一栏**。医学的鉴别诊断与软件工程的变异测试都有它。教育没有，不是因为教育更落后，是因为另外两个行当有一个教育没有的东西——一份先于个案就存在的做法清单。第十五节把”造出这份清单”作为最稳的一层主张。

九 · 与最容易被混为一谈的几种说法的分界

下面十种说法与本文极易混淆。每一条写四件：它是什么、它说到哪一步、分界在哪、以及一个能判出谁对的对照观察。凡是只能写出”本文更强调”“更深入”“视角不同”的，本文视同承认被它覆盖。

一 · 版本空间（Mitchell, 1982）。 机器学习里最贴近的一个：与已见样本一致的全部假设构成一个集合，学习即收缩它。**分界**：版本空间预设假设类给定，且其中只有一个是对的目标；本文说的做法集合里，**多条同时是对的**，收缩到哪一支不是逼近真相，是选定一条。**判定**：若在某个题域里，任意两条与全部作答相容的做法在该题域全部题目上必给出相同答案，则本文的对象退化为版本空间，本文多余；若存在两条都对而在某道未出的题上分岔，则版本空间的”唯一目标”设定在此不适用。

二 · 委员会查询与主动学习（Seung, Opper & Sompolinsky, 1992；Hanneke, 2014）。 用多个假设之间的分歧来挑最值得标注的样本，分歧系数是成熟读数。**分界**：它把分歧当作**采样信号**，目的是尽快消灭分歧；本文把分歧当作**被教育者的当前状态**，且不主张越小越好。**判定**：若在某个教学场景里，分歧被消灭得越快、学习者后续对新任务的适应越好，本文的价值排序错而主动学习对；若存在分歧被快速消灭而后续适应变差的情形，则分歧不只是待消灭的噪声。第十三节给出了这一条的实际让步。

三 · 认知诊断的等价类与边际可识别率（Zhang, DeCarlo & Ying）。 与本文距离最近的一位——它明确地把不可分辨的技能掌握模式分成等价类，并给出可一致估计的可识别率。**分界**：它把不可分辨判为**模型识别性的技术缺陷**，处方是标出”未分类”以控制误分类；本文判为**教育的常态与产物**，且它的模式是”掌握与未掌握”的组合，装不下”都会而做法不同”。**判定**：若把某份测验补足全部单技能题、使全部模式可分辨之后，两个模式相同的学生在后续新任务上的表现分歧降到与随机相当，则不可分辨确实只是技术缺陷，本文错；若分歧仍显著高于随机，则被补足的那套编码本身漏掉了一个维度。**这一条是本文最需要正面处理的，第十节给出处置。**

四 · 规则空间法（Tatsuoka, 1983）。 教育测量里最早把”学生用的是哪条规则”作为诊断对象的工作，本文所用数据正出自这一脉。**分界**：规则空间的规则是**错误规则**——它诊断的是”他用了哪条错的做法”；本文关心的是**都对而不同**。**判定**：取一份规则空间诊断报告，看它的规则清单里有没有两条都能在全部题目上给出正确答案的；若一条也没有，则该框架的清单在构造上排除了本文的对象。

五 · 深层结构与表面特征的分类（Chi, Feltovich & Glaser, 1981）。 专家按原理给物理题分类，新手按表面特征分类——这是”什么算同一道题”这个问题最经典的实证。**分界**：它把分类方式当作专长的**征兆**，是专家与新手之间的差；本文说的是**两个都算专家的人之间**的差，且关心记录能不能承载它。**判定**：在两名分类结果完全相同的被试之间，若其在新构型问题上的表现分歧不高于随机，则本文的对象可被”专长水平”这一维吸收；若显著高于，则分类相同不蕴含做法相同。

六 · 规则跟随的欠定（Wittgenstein, 1953，§185–201；Kripke, 1982）。 任何有限的训练都与无穷多条继续方式相容——这是本文形式层面的祖宗，且早了七十年。**分界**：它停在**原则上不可能**，因而在实践上无所主张，也不给读数；本文给的是一个有限清单上的有限比例，以及一条制度性后果。

判定：这一条不需要判定，本文承认被它在形式层面完全覆盖；本文可主张的只是把它从”原则上的不可判定”降为”某个题域上一个可清点的数”。**若有人认为一个可清点的有限读数不构成对那条原则性论证的任何增量，本文这一部分的贡献即为零**。

七 · 经验对理论的决定不足（Quine, 1960）。 同一批观察相容于多套互不相容的理论。**分界**：奎因谈的是理论选择，主体是科学共同体；本文谈的是个体在一个有限题域上的做法，且给出的是可执行的清点。**判定**：同上，本文承认在形式层面被覆盖。

八 · 极限内的语言辨识（Gold, 1967）。 仅凭正例，某些语言类在极限内无法辨识。**分界**：戈尔德的结论是关于**可辨识性**的，且以无穷样本为背景；本文的结论是关于**有限测验的分辨力分布**的，并指出加样本无用而换题有用。**判定**：若在某个题域里，增加同类题目能使不可分辨对的比例单调下降到零，则本文关于”加题无用”的主张在该题域为伪。

九 · 形成性评价（Black & Wiliam, 1998）。 它的处方是”多收集过程中的信息、及时反馈”。**分界**：形成性评价关心的是**及时性**，它默认收集到的信息种类是够的，缺的是时点；本文说的是**种类不够**——再及时的反馈，若仍以对错为单位，分辨力仍为零。**判定**：在两个反馈频次相同、而一组的反馈只报对错、另一组要求学生写出中间步骤的班级之间，后者对做法的分辨力应显著更高；若无差别，则本文关于”记录单位”的主张为伪。

十 · 迁移与情境（Gick & Holyoak, 1983）。 远迁移失败是老问题，本文的许多例子读上去像迁移失败。**分界**：迁移研究把失败归给**表面相似度**这一变量，处方是提高结构提示；本文把它归给**记录单位**，处方是造清单与出界题。**判定**：若在充分提示结构相似性之后，两名原测同分学生的分歧消失，则本文的对象可被迁移框架吸收；若提示后分歧不变（因为两人本来就都能做对，只是路径不同），则迁移框架未触及本文所指。

十一 · 同样从符号学、逻辑学与数学入手的一项先行研究。 本文之前已有一项工作以同样这三个学科为入口，追问同一性的来源，提出”个体化”这一路径，并把三个学科重述为对同一场生成行使的三种权能（王德生与协作者，《个体化的生成》，二〇二六）。两者共用三个学科名称，因此必须分清。**分界**：那项工作问的是”一个东西的同一性从哪里来”，答案落在本体层，且不产出任何可清点的读数；本文问的是”什么算同一件事的又一次”，答案落在记录层，产出一个有限清单上的比例与一条制度性预测。取用点也无一重合——那项工作取的是生物符号学与同伦类型论的单价公理，本文取的是教育符号学与统计学习理论，双方引用的文献零交集。**判定**：若在某个题域里，两条做法的差别可以由”它们所指对象的同一性判据不同”完全解释，则本文的对象可被那项工作吸收；若两条做法指向的对象同一性判据相同（同一道题、同一个答案），差别只在处理次序，则那项工作的框架在此不产生任何判断。第七节的分数减法数据落在后一种情形：两条做法算的是同一个式子，得的是同一个数。

最近的一位是第三条。 它离得最近，因为它算的正是本文要算的那种集合，而且算得比本文早。本文之所以仍不能被它吸收，只有一个理由：它的清单里只有缺失，没有并列。这不是它的疏忽——它的问题结构决定了它只需要缺失。第十节给出对它的处置。

十二、混合策略的项目反应模型（米斯利维与费尔黑斯特，一九九〇）—— 心理测量学自己最近的一位，本稿修订前漏掉了它。 它说到哪一步：不同被试可能用不同解法，而**只有作答可以被观测，解法**

本身不能；于是用实质理论写出各策略下作答向量的似然，把作答概率表达为「每套策略各自的题目参数 ＋ 用该策略的人所占比例 ＋ 该策略内部的能力分布」，再反过来估出某个具体的人用了哪套策略的概率。这一支后来长成混合项目反应模型的一整片文献。

它与本文近到必须逐句分开：**两边都从「作答相同而做法不同」出发**。 分离线只有一条，但很硬——**它假定策略清单是给定的**。 要写出那个混合似然，必须先有 K 套候选策略；模型能做的是把人分到已有的某一支，不能告诉你清单上还缺哪一支，更不能告诉你「哪一道题还没有被出出来」。本文问的正是这一件事：清单上任意两条做法之差，定义了一道题；那道题在不在已出的题目里。

判定形式：给定一份策略清单与一批被试，若混合模型的分类不确定度随着题数增加而单调下降，则它对，而本文所说的界题只是「样本量不够」；**若在题库里加入任意多道同型题、分类不确定度不降，则本文对**——因为新增的题若不是只考一项技能的那种，它一道也分不开新的模式。这一条不必新采集，它是本文第六节命题四的直接推论，可以在本文第七节那份 Q 矩阵上当场算：那二十道题里只有十五个互异的行向量，第三、八、十六、十七、二十题的分辨贡献严格为零，把它们复制一百遍，等价类仍然是五十八个。

十·四位从不同方向逼近同一件事的先行者

本文最要紧的一次自查，是承认自己不是第一个。把上一节里最危险的几位摆在一起看，会看到一个整齐的形状：四条互不来往的脉络，从四个方向逼近了同一件事，而每一条都在最后一步停住了，且都是因为各自问题结构的缘故。

软件工程（一九七八年起）已经有了这个读数的完整实现——变异得分衡量的正是”一套测试能分开多少种不同的实现”。它停住的地方是：它只用这个读数评价**测试**，从不用它描述**被测者的状态**；而且它的目标是消灭变异体，即以差别为敌。

主动学习（一九九二年起）已经有了用假设间分歧驱动决策的完整算法。它停住的地方是：分歧对它是**采样信号**，用完即弃；它假定目标唯一，因此不承认”多条都对”。

认知诊断（一九八三年起）已经有了做法集合、等价类与可识别率。它停住的地方是：它把不可分辨读成**要修的缺陷**，且它的做法只能表示为技能的缺失组合。

规则跟随与决定不足（一九五三年、一九六〇年）已经有了最彻底的形式论证。它停住的地方是：它把结论留在**原则层**，既不给有限清单，也不问制度记不记这一栏。

四位停住的地方各不相同，但停的原因是同一个：**在它们各自的问题里，“两条都对而不同”要么不存在（软件工程与主动学习假定有唯一正确目标），要么不需要（认知诊断只需诊断缺失），要么无从下手（哲学论证不以清点为目的）**。

本文补的是那一步，而且只是那一步：**把”还剩哪些做法”从测试质量的读数、采样的信号、模型识别的缺陷、原则上的不可判定，改判为教育的产物本身**；并据此指出一个可以去查的制度后果。除此之外，本文没有新方法，也没有新数据——第七节用的是三十年前的公开数据，用的是别人早已给出的算法。

诚实地说，这意味着本文的位置比它初看上去要窄。如果有人认为”改判”不算一项贡献，那么本文剩下的只是第十一节那一条可查的预测。本文接受这个评价，并把那一条单独列为第三层主张。

十一 · 同题七篇的互切

本文与另外六篇同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，另一侧站着的都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。

十二 · 命题实践里一处已经在发作的矛盾

上一节说本文最实的东西是一条可查的预测。这一节把它写出来。

一份正规测验的题目要经过预试与项目分析。两条最通行的筛选标准是：难度适中（通过率过高或过低的题被剔除），以及区分度足够（与总分相关低的题被剔除）。这两条都有充分的技术理由，它们让测验的信度更高、量表更稳。

同时，测验理论的另一支——效度论证——要求测验分数能支撑关于**未来表现**的推论（Kane, 2013）。一份测验不是为了描述考生在这些题上的表现，是为了据此说些别的话：他能不能上这门课、能不能做这份工作。

这两条要求在通常情形下并不打架。但第六节的两条命题把它们逼到了正面：命题二说，能分开全部做法的充要条件是每项技能各配一道只考它一件事的题；命题四给出了更细的量——**一道题的分辨力随它所要求的技能数按指数下降，最大值恰在只考一件事的那一档**。而只考一件事的题在真实考生总体上通过率很高——它考的是基础步骤，大多数人会。通过率高，方差小，与总分的相关自然低。于是：

> **能把做法分开的题，正是项目分析最先剔除的那一类。**

于是一份测验越是经过严格的品质控制，它对做法的分辨力越低，而它被要求支撑的那个关于未来的推论，恰恰依赖做法。这不是命题人的疏忽，是两条各自正当的规范在同一份试卷上相向而行。

第七节的数据把这件事变成了具体的。第九题只考一件事，通过率在模型下是最高的一档，它独家分开了四千零九十六对做法；撤掉它，最大的不可分辨组从六十四涨到一百二十八。同一份测验里最”难”的第十九题要求五项技能，分辨力只有它的八分之一。**在任何一份常规项目分析里，第九题是候删项，第十九题是范例**。

把命题四与项目分析的两条标准并排放，矛盾的形状就完全清楚了：**项目分析按”难度适中、与总分相关高”选题，选的是 m 大的那一类；而分辨力按 m 指数下降。两条准则在同一份试卷上，一条按 m 增而选，一条按 m 增而减**。这不是程度上的张力，是方向上的相反。

可以现在就去查的一条：取任何一个有预试记录的题库，把被以”区分度低”或”过易”剔除的题与保留的题分成两组，比较两组中”只考一项技能”的题所占比例。本文预测前者显著高于后者。所需数据在多数命题机构的既有记录里已经存在，一个人一周内可以做完。若两组比例无显著差异，本文这一条为伪。

现行的化解办法是把这个差别整个记进误差项。但误差项的定义是随机的、无系统方向的，而做法差别是系统的、且方向固定——它总是使那些”另有一套自洽做法”的人被记成与主流做法者相同。**把一个有方向的量记进一个被定义为无方向的项，是这一处矛盾目前的全部处置方式**。

十三 · 三家为什么各自到不了这一步

三家没到这一步，不是因为疏忽。一个领域看不见什么，通常不是它的缺陷，是它的问题结构的影子。

符号学到不了，是因为它的对象在原则上不可清点。 它的核心洞见——每一个解释项还需要下一个解释项——使得”还剩几种读法”这个问句在它的框架里没有答案：符号过程是开放的，任何清单都是武断的截断。它必须放弃清点，才能保住它最重要的那条主张。因此它可以说”意义在接收端重新发生”，却不能说”重新发生成了几种”。

推论主义到不了，是因为它的规范来自共同体。 布兰顿的记分是社会性的：什么算正确的推论，由实践设立。一旦承认两个学习者各自掌握了一张不同而都自洽的理由之网，“记分”就失去了唯一的裁判席，而记分正是这套语义学的引擎。它必须把网当作单数的，才能让”掌握”是一个可裁定的成就。

统计学习理论到不了，是因为它的定理只在假设类给定之后才成立。 它把假设类的来源明确划在管辖之外，这不是回避，是保持定理有效性的必要条件。而”两条都对的做法”恰恰是关于假设类内部结构的事——在它的框架里，两个泛化误差都为零的假设是同一个等价类里的元素，不需要区分，也无从区分。

三家各自的核心问题，恰好使这一步成为不必要的。这也是为什么这一步只能在三家之外走出，而它用的材料又必须是三家自己的。

十四 · 三处反过来削弱本文的约束

跨领域不能只是本文去别处取用。下面三处是别的领域反过来给本文加的限制，每一处都削掉了本文原本打算主张的东西。

第一处，来自主动学习，削掉的是价值排序。 本文最自然的倾向是”剩余做法多是坏事，应当尽快收窄”。主动学习的实践给出的却是相反的一面：在算法里，把分歧压得太快会导致采样集中在一小片区域，最终泛化更差；保持一定的分歧是有价值的。这迫使本文放弃”越窄越好”这一整套价值排序。本文现在只能主张：**剩余做法是一个应当被记录的状态，而不是一个应当被最小化的指标。** 如果没有这一处，本文原本会写下的更强的话是”教学的目标是把分法未定率降到最低”——那句话现在不能写。

第二处，来自医学的鉴别诊断，削掉的是适用范围。 鉴别诊断之所以能清点，是因为它的做法清单由整个专业事先编纂、长期维护，而且清单本身是有限的。本文原以为做法清单可以由研究者按需汇编。医学的实践说明：一份能用的清单是一项持续数十年的集体工程，不是一次编码任务。这把本文的适用范围从”任何题域”收窄到**“已经有或愿意长期维护一份做法清单的题域”**。目前符合这个条件的教育科目，本文所知不超过数学与部分理科的少数几个子域。

第三处，来自符号学，动的是本文界定的边界本身。 洛特曼的那条——传不准的那一部分才是新意义的发生器——对本文是一记重击。本文把”两条做法在某道题上分岔”当作一个待读出的差别，隐含地把分岔当作静态的、已经在那里的东西。洛特曼提示的是：分岔的那一刻可能正是新东西发生的那一刻，而不是一个等待被测量的既成事实。若这条成立，则**界题不只是用来读出差别的工具，它可能是制造差别的场合**——出一道界题给学生，可能不是在测他，是在把他推向某一支。本文无力在现有材料上分开这两种读法，只能明写：**本文关于”界题读出既有差别”的说法，可能把一个生成过程误写成了一次测量。**

十五 · 不适用的边界

一 · **做法清单不存在且不可能存在的题域**。 一节讨论一首诗的语文课，“做法”不是有限的，本文的读数在那里没有定义。硬套只会得到一个假的数。

二 · **只记对错的记录**。 若作答记录里没有中间步骤、没有过程数据，本文的读数无法计算。这不是数据不够多，是数据的种类不对——把同一种记录再收集十年也算不出来。

三 · **目标确实唯一的技能**。 有些训练的目标就是让所有人做法完全一致：外科的无菌操作次序、飞行的检查单、核电的操作规程。在这些场合，剩余做法为零是**对的**，本文不主张记录它以外的任何事。第十六节的伦理条款会再说一次这一点。

四 · 学习的早期阶段。 一个刚接触一个题域的人，剩余做法当然多，那是正常的，不构成任何诊断。本文的对象是**表现已经稳定为全对之后仍然剩下的那些做法**，不是学习中途的不确定。

五 · 本文不主张剩余做法与任何后续表现之间存在已被证实的关系。 第八节的十二处全部是预测，不是结果。本文只证明了这个量在现有测验上是零或很大，没有证明它有用。**这是本文与一个已被验证的理论之间的全部距离**。

十六 · 可以推翻本文的条件

15.1 最强的竞争解释

在写证伪条件之前，先写出那句最难反驳的”其实不就是……吗”。它是：

> **其实不就是测验信度不够、题量不足吗？**

这个解释朴素、通行、而且解释力很强：两个学生分不开，多考几次不就分开了。本文必须能排除它，否则本文只是把一个测量学问题换了个说法。

排除的办法在第六节的命题一里：**分辨力不随题量增长**。 如果两条做法在一批题上作答一致，加同类的题不会改变任何东西。因此：

> **控制题量与重复次数之后，若不可分辨对的比例随题量单调下降到接近零，则本文为伪，那时”分不开”应归因于样本不足而非题目选取。**

第七节的数据在这一点上支持本文：二十道题里，撤掉一道使最大不可分辨组翻倍，而那一道是唯一一道单技能题——起作用的是**哪一道题**，不是**多少道题**。

15.2 编号的证伪条件

F1（本节最便宜的一条，已跑，见第七节）在任何一份公开的技能矩阵上，若不存在多成员等价类，则本文关于”现行测验分辨力不足”的主张在该测验上为伪。**已跑：不成立，五十八个等价类，最大组六十四。**

F2 若在某个题域内，两条被专家一致判为”都正确”的做法，在该题域全部题目上给出完全相同的作答（含中间步骤），则界题在该题域为空集，本文的对象不存在。

F3 控制题量、重复次数与作答时长之后，若”只考一项技能的题”在预试剔除记录中的占比不高于保留题，则第十一节的矛盾为伪。样本要求：同一命题机构、同一学科、同一学段的不少于六份题库，跨机构或跨国的对比不算数。

F4 若在两个反馈频次相同、一组只记对错、一组记中间步骤的班级之间，对做法的分辨力无显著差异，则本文关于”记录单位”的主张为伪。

F5 若把某份测验补足全部单技能题、使全部模式可分辨之后，两个模式相同的学生在后续新任务上的表现分歧降到与随机相当，则不可分辨确实只是模型识别性的技术缺陷，本文对第九节第三条的分界不成立。

F6 若剩余做法被快速收窄的学习者，其在后续新任务上的适应一致优于剩余做法保持较宽的学习者，则第十三节第一处的让步过头了，本文应恢复”越窄越好”的价值排序——那时本文的价值主张错，但描述性主张仍立。

F7（正向读数，写成顺序而非相关）在一门课程里，若”学生之间在中间步骤上出现分歧”这一事件，**并不早于**”学生之间在最终答案上出现分歧”这一事件，则本文关于分辨力次序的主张为伪。顺序预测比相关难被巧合满足。

F8 若第七节报告的负相关（$\rho = -0.93$）在一份包含并列正确做法的清单上仍然成立，则分法未定率确实只是分数的一个代理，本文的读数应被放弃。

15.3 最小的一次实验

上面第二条与第八条需要一份现在不存在的数据。造出它的最小设计如下，一个人一学期可以做完：

取一个已有做法清单的题域（分数运算或一元一次方程），从教学文献与错误分析文献里汇编出该题域的做法清单，只收录**能在全部题目上给出正确答案**的那些（这一条把错误规则排除在外，正是与规则空间法的分界所在）。设计一份测验，要求写出中间步骤。按清单对每份作答编码，算出每个学生的相容做法集合。然后在一学期后给一份只含界题的后测。

判定写死：若前测同分的学生在后测上的分数分歧不显著高于前测同分学生在一份普通平行卷上的分歧，本文的核心主张为伪。

15.4 一条写死日期的赌注

到二〇三三年十二月三十一日，至少有三个独立的教育机构（不含本文作者所在机构）在其正式的学生记录格式中，设立一个独立于分数与掌握概率的字段，用以记录”与该生作答相容的做法集合”或其等价物，并公开该字段的编码规则。

写死什么不算命中：机构自行声明”我们重视过程性评价”不算；把中间步骤留档而不做相容性编码不算；把它作为科研项目的临时数据而不进入正式记录格式不算；只在某一门课的某一次考试上做一次不算；由本文作者参与设计的不算；把它做成一个可用于排名的合成指数不算。

15.5 若本文被推翻，会学到什么

如果 F2 被触发——即在实际题域里，专家一致判为都正确的做法在全部题目上作答完全相同——那么学到的东西比本文成立更有意思：它意味着**教学内容的组织本身已经把并列做法消灭掉了**，即课程在教之前就已经完成了收窄。那时该问的问题不是”记录能不能读出做法”，而是”课程是在什么时候、按什么标准把其余的做法从教材里删掉的”。

十七 · 伦理限制

分法未定率是一个可以拿去考核人的读数，因此必须写死三条，缺一条本文不可被用于任何评价用途。

一 · 这个读数指的是题目与做法这一对，不是人。 一个学生的分法未定率高，说的是**这份测验对他的做法分辨力低**，不是说他学得差。把它读成对学生的评价是对定义的误用。它更适合放在测验的质量报告里，而不是学生的成绩单上。

二 · 不得为了把这个数做好看，而在安全关键的训练里制造做法多样性。 无菌操作、检查单、操作规程这些场合，剩余做法为零是正确的目标。任何以”提高分辨力”为名在这些场合鼓励做法分歧的做法，本文明确反对。

三 · 已经据此对任何人做出不利安排的，必须公开编码规则并提供复核通道。 由于做法清单的汇编包含判断，编码规则必须可核、可争议、可修订。

还有一条要写在这里，因为它是本文的反噬所在：**一个读数一旦进入考核，最省事的达标办法是在文书上做手脚。** 具体到本文：要让分法未定率好看，最省事的办法是把做法清单做短——清单里只留两三条做法，任何学生的相容集合都会显得很窄。清单由谁编、编多长，没有任何外部约束。

更麻烦的是界题本身的自毁结构：**一道题一旦被公布为界题，它就会被专门教，从而不再分辨任何东西。** 这与考试题库的保密问题同形，但更难办——保密可以靠轮换解决，而界题一旦被教，学生的做法本身就变了，题目不是失效，是它的对象消失了。

本文看不到出口。任何使剩余做法可核查的装置，都要靠出题来实现；而出题就是把界题变成已出的题。能做的只有一条底线：**验收看事后清点的实际读数，不看有没有设立制度**；并且明写，这条底线挡不住上面那个反噬，只是让它慢一点。

十八 · 本文对自身的定位

本文不打算用”本文也有局限”来收尾。用本文自己的判据，检两样东西：提出这道题的那条指令，以及生产本文的这条流程。

先检那条指令。 本文起于一个明确的要求：回答”教育是什么”，从三个指定学科入手，产出一个新说法。用本文的判据问这条指令：**满足这条指令的、彼此不同的做法有几种？** 答案是：很多，而且这条指令里没有任何一处能把它们分开。它没有规定要哪一类答案，没有规定什么算作没达到，也没有规定用什么去检验。换句话说，**这条指令本身是一道分辨力极低的题**——它对一大批彼此差别很大的产物给出相同的验收结论。

这不是对提问者的抱怨，这是本文判据的一次自我应用，而且它有具体后果：由于指令的分辨力低，本文与”另一篇同样满足这条指令而内容不同的文章”之间，目前没有任何一道已出的题能分开。第十五节的证伪条件是本文为自己出的界题，出这些题的动机正在于此。

再检这条流程。 生产本文的流程有一个固定的形状：找几家互不相容的立场、找出它们共同的假定、用其中一家的材料推翻它、命名一个新东西、给一个读数、划一批界线、写一批证伪条件。用本文的判据问这条流程：**与这条流程迄今全部产出相容的”做法”有几种？** 答案是：目前无法区分。这条流程的每一次产出都通过了它自己的全部检查，而它是否在不同的题目上真的走了不同的路，从未被任何一道题分开——因为验收这条流程的题，全是它自己出的。

这条流程正落在本文那张表的右上格里：形式审查全过，短期读数好看，失效不产生信号，而且越正规化越严重——检查项越细，越容易用同一套骨架反复填空而通过全部检查。本文对此没有解法，只有一条不彻底的措施：本文换掉了这条流程惯用的几处做法——证伪的主件由一条写死日期的赌注换成了预注册加实算，自我定位这一节由”本文自己也落在那一格里”换成了检验指令与流程本身。**换做法只能降低同形的程度，不能证明这一次真的不同。** 要真的分开，需要一道由外部出的、这条流程未曾见过的题。

最后交出两个本文解释不了的洞。

第一个洞：本文没有说清楚，一份做法清单该由谁来编。清单里放什么、不放什么，本身是一个判断，而这个判断恰恰是本文要研究的那类判断。这里有一个回环，本文没有解开它，只是指出医学用了几十年建立鉴别诊断表这一事实，说明它至少不是不可能的。

第二个洞：本文全部的论证都假定”做法”是一个离散的、可枚举的东西。第十三节第三处已经指出，这个假定可能把一个生成过程写成了一次测量。如果做法不是离散的，那么本文的读数在原则上就不存在，第十四节第一条的边界就要扩大到覆盖大部分人文学科，也可能覆盖数学教育本身。本文没有办法判断这一点。

十九 · 结论

回到那两份满分卷子。

三个前沿理论各自都能说出它们的一部分：符号学会说要看两人的读法，推论主义会说要看两人给不给得出理由，统计学习理论会说要看两人在新分布上错多少。三家都对，而三家合起来暴露了一件三家都没说的事——它们都在读那两个人，而那两个人之间的差别，严格说来不是他们身上的属性，是他们与一批题目之间的关系。二十道题不是没测准，是这二十道题在他们身上的分辨力恰好为零。

于是”教育是什么”有了一个不同形状的回答。教育不是把一份内容送到接收端并在那里被读出来，不是掌握一张由共同体给定的理由之网，也不是在给定假设类与分布下把泛化误差降下来。**教育是与学习者迄今全部表现相容的那些做法的收缩；而收缩到哪一支，只在一批尚未被任何人出出来的题上才显形。** 这批题不是难题、不是偏题、不是知识点、不是错误，它由两条都对的做法之差定义出来，撤掉这个定义它就不存在。

第七节的实算给出的数字是具体的：一份被分析了三十余年的诊断测验，二百五十六种技能掌握情形只能分成五十八组，最大的一组有六十四种一道题也分不开；二十道题里有五道的分辨贡献严格是零；测验一半的分辨结构挂在一道随时可能被以”过易”删掉的题上；八项技能里唯一一项是”选择”而非”工序”的那一项，八成七的情形下读不出来。而第六节的闭式说明这不是偶然：**一道题分辨做法的能力随它所要求的技能数按指数下降，而命题实践恰恰按相反的方向选题。** 这两条准则不是有点张力，是方向相反。

实算同时给出了两个对本文不利的结果，本文没有把它们放进注释。它们说明现行的编码把”做法”定义成”缺了什么”，因此在那套编码里，“两个人都会做而做法不同”不是一个可以为真或为假的命题，是一个说不出来的句子。本文因此把主张收窄成一句更硬的话：**不是记录看不见剩余做法，是记录里没有它的位置。**

本文没有新方法，也没有新数据。软件工程用变异得分做了将近五十年，主动学习用分歧驱动采样做了三十多年，认知诊断算等价类算了十几年，而规则跟随的论证在七十年前就完成了。本文做的只有一件事：把”还剩哪些做法”从测试质量的读数、采样的信号、模型识别的缺陷、原则上的不可判定，改判为教育的产物本身，并指出一处已经在发作的制度矛盾。

如果这个改判不算贡献，那么本文剩下的是第十一节那一条可查的预测，与第十五节那份一个人一学期能做完的实验设计。本文认为那两样也够。

注释

1. 本文所用的”做法”一词，严格限定为”一条从题面到答案的完整处理次序”，且两条做法不同当且仅当存在至少一道该题域内的题，两条在其上给出不同作答。日常语义里的”思路”“方法”“策略”与之不完全重合，正文一律不混用。

1. 第六节的四条命题在合取型模型下陈述。命题三只依赖”理想作答由所支配的要求向量集合决定”这一点，故对任何满足该性质的模型成立；命题四的闭式依赖合取形式，析取模型下经作答取反后同形，一般模型下只保留单调性而闭式不再精确。对析取型模型，由于其与合取型模型的对偶关系，全部结论经作答取反后同样成立。对更一般的模型，命题一仍然成立（它只依赖作答一致的定义），命题二的充要性变为充分性（Köhn & Chiu, 2017）。

1. 第七节的全部计算基于理想作答，不含失误与蒙对参数。加入这两个参数后，“可分开”变为”需要更大样本才能分开”，“不可分开”仍为不可分开。故本文报告的不可分辨比例是下界。

1. 第七节所用脚本四份（等价类分划、补完备与两策略分辨力、读数正交性自检、命题三与命题四核验）均为确定性计算，不含随机成分；技能矩阵逐字取自公开文献，任何人可独立复算。

1. 关于第十节：把四位先行者写成”汇聚”而非”各有不足”，是因为四条脉络停住的位置各不相同，而停住的原因都可以从各自的问题结构推出。这是解释，不是评价。

1. 本文为理论论文，第八节的十二处兑现全部是预测，均未经检验。把它们读成结果是误读。

1. 人机分工声明：本文的选题、三家学科的指定、以及”给出一个新说法”的要求由人给出；文献检索、形式推导、第七节的全部计算与写作由人工智能系统完成，计算脚本与预注册文本随文可查。第十七节对这条流程的自我定位适用于本文全文。

1. 全文约二万四千汉字。本文可以分层引用，四层的稳固程度依次上升：第一层是”界题”这一构念（最强，也最易倒）；第二层是那张两轴辨别格（作为分析工具）；第三层是第六节的命题三与命题四（初等但精确，且与实测逐格相符，其成立不依赖前两层，也不依赖本文的任何主张）；第四层是第十一节那条关于预试剔除记录的可查预测。引用第二层时请连同第七节的两个不利结果一并引；命题三与命题四可以脱离本文的全部论点单独引用。

参考文献

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. Mathematical Thinking and Learning, 13(1–2), 5–26.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74.

Brandom, R. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment. Cambridge, MA: Harvard University Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121–152.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. Psychometrika, 74(4), 633–665.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics, 34(1), 115–130.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. Psychometrika, 69(3), 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. Psychometrika, 73(4), 595–624.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. Applied Psychological Measurement, 35(1), 8–26.

DeMillo, R. A., Lipton, R. J., & Sayward, F. G. (1978). Hints on test data selection: Help for the practicing programmer. Computer, 11(4), 34–41.

Gardner, M., Artzi, Y., Basmov, V., et al. (2020). Evaluating models' local decision boundaries via contrast sets. Findings of EMNLP 2020, 1307–1323.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15(1), 1–38.

Gold, E. M. (1967). Language identification in the limit. Information and Control, 10(5), 447–474.

Hanneke, S. (2014). Theory of disagreement-based active learning. Foundations and Trends in Machine Learning, 7(2–3), 131–309.

Jia, Y., & Harman, M. (2011). An analysis and survey of the development of mutation testing. IEEE Transactions on Software Engineering, 37(5), 649–678.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. Psychometrika, 82(1), 112–132.

Kripke, S. A. (1982). Wittgenstein on Rules and Private Language. Cambridge, MA: Harvard University Press.

Lotman, Y. M. (1990). Universe of the Mind: A Semiotic Theory of Culture. London: I. B. Tauris.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrika, 55(2), 195–215. doi:10.1007/BF02295283

Mitchell, T. M. (1982). Generalization as search. Artificial Intelligence, 18(2), 203–226.

Noorloos, R., Taylor, S. D., Bakker, A., & Derry, J. (2017). Inferentialism as an alternative to socioconstructivism in mathematics education. Mathematics Education Research Journal, 29(4), 437–453.

Peirce, C. S. (1931–1958). Collected Papers of Charles Sanders Peirce (C. Hartshorne, P. Weiss & A. Burks, Eds.). Cambridge, MA: Harvard University Press.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT ' 92), 287–294.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.

Stables, A., & Semetsky, I. (2015). Edusemiotics: Semiotics of Education — Language and Learning in the Era of Digital Media. London: Routledge.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4), 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold & M. Shafto (Eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition (pp. 453–488). Hillsdale, NJ: Erlbaum.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and Its Applications, 16(2), 264–280.

Wittgenstein, L. (1953). Philosophical Investigations (G. E. M. Anscombe, Trans.). Oxford: Blackwell.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), 1341–1390.

Zhang, S. S., DeCarlo, L. T., & Ying, Z. Non-identifiability, equivalence classes, and attribute-specific classification in Q-matrix based cognitive diagnosis models. arXiv:1303.0426. https://arxiv.org/abs/1303.0426

本文所用的四份计算脚本与预注册文本随文提供，数据为公开可下载的分数减法测验技能矩阵，全部结果可独立复算。