METHODS FOR MEASURING SCHOOL EFFECTIVENESS

衡量学校效能的方法

Joshua Angrist, Peter Hull, and Christopher R. Walters
NBER Working Paper No. 30639
November 2022
JEL No. C11, C21, C26, I20, I21, I24

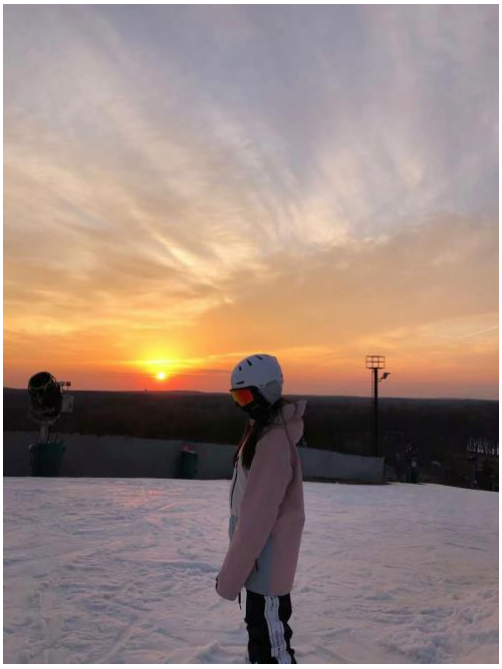
摘要

许多个人和政策的抉择都取决于对学校效能的看法，本文将学校效能定义为就读于特定学校或一系列学校对学生测试分数和其他结果的因果效应。一些广泛使用的学校评价框架比较了不同学校的学生的平均成绩，但平均成绩里的不受控制的差异可能更多地是因为选择偏差，而不是因果效应。这样的选择问题促使了一波计量经济学的创新，出现了很多使用随机和准实验的变化因素来衡量学校效能的研究。本文回顾了这些研究中的实证策略，强调了已解决和未解决的问题。全文使用了实证的例子。

翻译：赵雨（zhaoyu1998@ruc.edu.cn）

中国人民大学农村发展专业硕士研究生，研究兴趣：发展经济学，教育政策评估

校对：王恺赟（kaiyunw@umich.edu）



密歇根大学公共政策&应用经济学硕士在读，学术小白+叶老师（？？？）铁粉一枚，对基于因果推断的 Post-secondary student access and succes、Long-term outcomes of education 等方向很感兴趣，希望在自己的不断努力下未来可以做出一些好的研究。

原文：

Angrist, J., Hull, P., & Walters, C. R. (2022). **Methods for Measuring School Effectiveness.** NBER Working Paper No. 30803. Forthcoming in *Handbook of Economics of Education*, Volumn 6.

1 引言

广泛的个人和政策的决定都取决于对学校质量的看法。一方面，父母在选择学校和社区时，会权衡学校效能和其他因素，如住房成本和通勤时间；另一方面，学校质量问题也推动了高风险政策的决定，如学校的关闭、重新组织和扩建。为了满足对学校质量信息的需求，基于成绩的公立学校效能的衡量标准已经大量涌现。这些衡量标准包括由一些州和地区公布的"学校报告书"，以及由GreatSchools.org等私人组织发布的学校质量评级。

如何才能可靠地衡量学校质量？本文回顾了最近用于估计学校效能的计量经济学方法，其定义是上某所学校或一组类似的学校（如特许学校[1]）对学生成绩的因果效应。为了更好地衡量学校质量，必须面对来自选择偏误的基本挑战：学校与学校之间的比较同时会反映学生的能力和家庭背景信息，甚至多于学校效能方面的信息。经济学家已经为这种偏误设计了一系列的解决方案，并且越来越多地利用现代学校分配方案中的随机因素来设计有说服力的自然实验。

我们对于文献的梳理从一个相对简单的问题开始，即如何衡量通过抽签(lottery)提供入学机会的单一学校或学校部门的效能。例如，Angrist等人（2010）研究了新英格兰地区第一所知识就是力量计划（Knowledge is Power Program，KIPP）特许学校的效果。原则上，在被超额申请的特许学校进行抽签确定入学机会，像这样的方式能够识别在这所学校就读的因果效应。这种识别策略是利用有关学校入学机会的（也许是有条件的）随机提供作为入学率的工具变量（IV）来实现的。

虽然从概念上理解是简单的，但在使用school-lottery IV的过程中会出现一些复杂的问题，甚至掩盖了估计一个学校或地区的质量。我们对抽签基础知识的讨论涵盖了与协变量、等待名单以及随机分配、学校注册和结果测量之间的延迟有关的问题。我们还讨论了单个学校抽签的断点回归（RD）类似问题，即根据考试分数是否通过分数线来录取学生，而不是有条件地随机分配。这里的主要例子是像 Abdulkadiroğlu 等（2014）研究的那些选择性招生的"考试"学校。最后，如Angrist等（2016a）一样，单一学校（single-school）

[1] 美国特许学校（charter school）是经由州政府立法通过，特别允许教师、家长、教育专业团体或其它非营利机构等私人经营公家负担经费的学校，不受例行性教育行政规定约束。——译者注

IV 的设定被用来审查学校对考试分数分布的影响的估计方法。

当然，许多关于学校质量的问题涉及到不止一个部门或学校。一个希望测量多种因果效应的学者必须建立一个共同的反事实，以确保估计是真正的苹果对苹果的比较。集中分配学校的做法有利于识别多学校和多部门的模型，如波士顿、丹佛、新奥尔良和纽约市的学区（仅举几例）。集中分配计划在相关地区的大多数学校中随机分配了相当份额的席位，这类地区的随机分配是有条件的：不同的学生被分配到特定学校的席位，其比例由偏好和优先级决定。Abdulkadiroğlu 等（2017）和 Abdulkadiroğlu 等（2021）推导了集中式分配中产生的分配概率，并说明分配率如何为多个地区或学校的因果推断奠定基础。我们综合解释了这项工作的实际经验。

然而，对学校效能的关注显然早于集中分配的出现。常规的增值模型（value-added models，VAMs）通过回归控制滞后项和其他协变量来衡量学校效能。由于单个学校的增值模型估计值往往是有噪声的，经济学家长期以来一直采用经验贝叶斯（EB）方法来减少取样方差[2]。最近，Angrist 等人（2016b，2017，2021）开发了一套 EB 方法，使用集中分配来衡量学校效能，并减少传统 VAM 估计中的选择偏差。这些方法旨在平衡传统 VAM 的相对精度和集中分配带来的偏差减少和模型验证。这里通过对马萨诸塞州、丹佛市和纽约市的学校的应用说明了新的 VAM 模型和方法。如同 Angrist 等（2021）一样，这些应用表明，即使在没有随机分配座位的学校，集中分配也能阐明学校效能。

本章结构如下。第 2 节回顾了基本的 IV 框架，因为它适用于单一学校或部门的学校质量衡量。第 3 节扩展了该框架，以涵盖具有异质效应的模型，对多个地区和年份的有效性的测量，以及使用 RD 式的席位分配的识别策略。第 4 节描述了使用集中分配来联合估计多个不同部门的学校质量。这里的一个主要例子是对不同类型的特许学校的分析。第 5 节讨论了回归控制的 VAM 估计，概述了分析单个学校质量的 EB 框架，并展示了如何利用准实验性录取或分配变化来验证和改进传统的 VAM。最后，我们强调了学校效能测量方面的其他挑战和研究前沿。

[2] Raudenbush 和 Bryk（1986）是第一个在这种情况下应用 EB 方法的。

2 学校抽签基础知识

2.1 单一学校的影响

在美国，有大量的实证文献使用随机招生抽签来衡量 K-12 学校对学术成果的影响。例如，特许学校的研究（Hoxby 和 Murarka，2009；Angrist 等，2010，2012，2013，2016a；Abdulkadiroğlu 等，2011；Dobbie 和 Fryer，2011，2013，2015；Clark 等，2015；Davis 和 Heller，2019；Cohodes 等， 2021；Setren，2021），学校优惠券（Chingos 和 Peterson，2015；Mills 和 Wolf，2017；Abdulkadiroğlu 等，2018），小型学校（Bloom 和 Unterman，2014），精英学校（Engberg 等，2014），寄宿学校（Curto 和 Fryer，2014 年），以及学校选择的各个方面（Cullen 等，2006；Hastings 等，2009；Deming，2011，2014；Deming 等，2014）[3]。我们首先回顾了这项工作所考虑的最简单问题的计量经济学方法，即单一学校的效能。

KIPP 经常出现在公众视野中，假设研究者对在这个网络中的特许学校就读的效果感兴趣。需要知道的是，KIPP 学校是 "没有任何借口（No excuses）"的公共教育方法的象征，这是一种被广泛复制的城市特许学校模式，其特点是上学时间长，学年延长，选择性地雇用教师，为教师提供广泛的数据驱动的反馈，学生的行为规范，以及对传统阅读和数学技能的关注。不久前，在新英格兰只有一所这样的学校——马萨诸塞州林恩市的 KIPP 中学。这所学校有多好呢？Angrist 等人（2010，2012）利用 KIPP 抽签申请人的数据来衡量了林恩市这所中学的学校效能。

抽签策略使用 IV 来确定伯努利处理的因果效应，$D_i \in \{0,1\}$，这里表示学生 i 的 KIPP 入学。如果一个学生在林恩市 KIPP 上学，用 $Y_i(1)$表示他的潜在六年级成绩水平；否则，就用 $Y_i(0)$表示他的成绩。对于这个学生成绩的观察结果 Y_i，是这两种潜在结果中的一种，这取决于 D_i。因此，我们可以这样写：

$$Y_i = D_i Y_i(1) + (1-D_i)Y_i(0) = Y_i(0) + (Y_i(1) - Y_i(0))D_i.$$

为了关注选择偏差的问题，我们首先假设一个恒定的因果效应，对于所有的学生 i，$Y_i(1) - Y(0) = \beta$。这时，观察到的结果可以这样

[3] 在美国之外，基于抽签的学校评估包括 Angrist 等（2002）、Lee 等（2014）、Zhang（2014）、Behaghel 等（2017）、Lee 和 Nakazawa（2021）、Oosterbeek 和 de Wolf（2021），以及 Romero 和 Singh（2021）。

表示：

$$Y_i = \mu + \beta D_i + \varepsilon_i. \quad (1)$$

其中 $\mu \equiv E[Y_i(0)]$，$\varepsilon_i = Y_i(0) - \mu$。$\varepsilon_i$ 可以被认为是对学生能力、家庭背景和学习动机的综合衡量。我们把这个衡量标准简称为 "能力"。

估计 β 的核心挑战来自于这样一个事实，即能力和 KIPP 的入学可能是相关的。与典型的城市学生相比，KIPP 的学生可能特别有动力，或者来自父母受教育水平更高的家庭（Rothstein，2004），这样会使得 $E[\varepsilon_i \mid D_i = 1]$超过了 $E[\varepsilon i \mid Di = 0]$。在这种情况下，等式(1)中的因果参数 β 不太可能与 Y_i 对 D_i 回归中的斜率系数相符，或者说，在 D_i 取 0 和取 1 的情况下，Y_i 的条件平均值的差异与因果参数 β 不一致。选择偏差很可能导致按 KIPP 入学情况进行的平均成绩比较超过 β：

$$E[Y_i \mid Di = 1] - E[Y_i \mid D_i = 0] = \beta + E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0] > \beta.$$

多元回归估计可以通过控制家庭背景和申请前（即滞后）成绩改善这一问题，但不一定能消除这种选择偏差。

IV 使用随机招生抽签解决了选择偏差带来的问题。假设有 n 个学生申请 KIPP 的 $m < n$ 个六年级的席位，让虚拟变量 Z_i 表示申请者获得了一个席位。我们假设抽签既公平又是有结果的。从形式上看，这意味着：

假设 1A：抽签获得的 offer 与学生能力无关，$\varepsilon_i \perp Z_i$。

假设 1B：抽签赢家比输家更有可能注册入学，$E[D_i|Z_i = 1] > E[D_i|Z_i = 0]$。

假设 1A 是合理的，因为抽签的 offer 是随机分配的，所以与学生的能力无关，不过 1A 也要求抽签的 offer 与特许入学以外的渠道无关，这是一个排他性的限制。假设 1B 是可信的，因为抽签机会为入学打开了大门：虽然不是所有的抽中的学生都入学，一些未抽中的申请人可能通过其他渠道找到他们的方式来入学，但入学是由抽签的 offer 强烈预测的。

这一对假设使 Z_i 成为等式(1)中 D_i 的有效工具变量。由于这里的工具变量是 Bernoulli，IV 估计值是 Wald（1940）型的均值差异比率：

$$\beta_{IV} \equiv \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]} = \beta. \quad (2)$$

β 的识别来自于等式(1)以及假设 1A 要求的 $E[\varepsilon|Z_i]=0$，而假设 1B 确保等式(2)的分母非零。

等式(2)的分子和分母可以计算为成绩和特许学校入学率对 Z_i 的双变量回归的斜率系数。

$$Y_i = \gamma + \rho Z_i + u_i, \tag{3}$$

$$D_i = \psi + \pi Z_i + v_i. \tag{4}$$

传统上，等式(3)被称为简约式（reduced form），而等式(4)则是相应的第一阶段（first stage）。我们有时使用这些术语来表示斜率系数 ρ 和 π，而不是包含它们的方程。等式（2）表明，$\beta_{IV} = \rho/\pi$，这与假设 1A 和 1B 下的因果参数 β 相一致。

2.2 对分配风险和协变量的控制

学校通常每年都进行招生抽签，许多学校还为不同的入学年级和已经入学的学生的兄弟姐妹以及其他特殊群体进行单独抽签。这种多重抽签的情况可以被视为自然发生的类似于用于衡量药物疗效的分层随机对照试验（RCTs）。在分层 RCT 中，受试者在不同的阶层中以不同的比率接受治疗，治疗分配只在阶层内独立于潜在的结果。

区分抽签阶层的信号特征是获得 offer 的条件概率，这个特征我们称之为分配风险。例如，当一所学校刚开学时，申请人的分配风险可能是 0.9，需求很弱，几乎所有的申请人都能被录取。但在以后的几年里，随着学校的建立和普及，分配风险可能下降到 0.5 或更低。同样地，在某一年，兄弟姐妹获得入学席位的比率通常比非兄弟姐妹高得多。一组具有相同分配风险的学生被认为是构成了一个风险组(risk set)。即使在分层的 RCT 中，分配风险的不同也是选择偏差的一个可能来源。因此，除了最简单的单层抽签研究设计外，我们对所有的分配风险进行控制。

在风险组内，单一学校抽签的 IV 框架同样适用。用 $R_i \in \{1, ..., K\}$ 来编码申请人 i 的风险组，$R_{ik} = 1[Ri = k]$ 表示 i 在 k^{th} 集。风险组 k 中的申请人的条件 Wald 估计值是：

$$\beta_{IV,k} \equiv \frac{E[Y_i \mid Z_i = 1, R_i = k] - E[Y_i \mid Z_i = 0, R_i = k]}{E[D_i \mid Z_i = 1, R_i = k] - E[D_i \mid Z_i = 0, R_i = k]}. \tag{5}$$

如果假设 1A 和 1B 在 $R_i = k$ 的条件下成立，那么 $\beta_{IV,k}$ 捕捉到该组申请人参加特许学校的因果效应（以下均简译成“特许效应”）。

在实践中，基于 $\beta_{IV,k}$ 的相似样本的特定风险组估计可能是有噪声的。但我们可以将这些条件性的估计汇总成一个更精确的总结性估计。

一个控制了所有风险组的两阶段最小二乘法（2SLS）估计式方便地将 $\beta_{IV,k}$ 集合在一个单一的加权平均值中。将(1)所描述的因果模型和(4)所描述的第一阶段等式与风险控制相结合，可以得到一个 2SLS 的设置：

$$Y_i = \beta D_i + \sum_{k=1}^{K} \delta_k R_{ik} + \eta_i, \tag{6}$$

$$D_i = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + \upsilon_i. \tag{7}$$

风险控制的因果模型(6)中的参数 δ_k，可以看作是等式(1)中 $\mu+\varepsilon_i$ 对 R_{ik}，以及相关残差 η_i 的回归系数。第一阶段等式(7)中同样通过包括 R_{ik} 作为回归因子来控制风险（这是一项规则，2SLS 第一阶段需要包括出现在与之相关的因果模型中的控制变量）。

2SLS 使用等式(7)产生的第一阶段拟合值作为(6)中 D_i 的工具变量[4]。 Kolesár(2013)表明，这种 2SLS 估计值的特点是一个二值的内生变量，对离散协变量的控制是饱和的，它捕捉了风险组内 IV 系数的加权平均，可以写成：

$$\beta_{2SLS} = \sum_{k=1}^{K} [\frac{\omega_k \pi_k p_k (1-p_k)}{\sum_l \omega_l \pi_l p_l (1-p_l)}] \beta_{IV,k}. \tag{8}$$

其中，$\omega_k = Pr[R_{ik} = 1]$是风险组 k 中申请人的份额；$\pi_k = E[D_i | Z_i = 1, R_{ik} = 1] - E[D_i | Z_i = 0, R_{ik} = 1]$ 是这些申请人相应的第一阶段；$p_k = Pr[Z_i = 1 | R_{ik} = 1]$，所以 p_k (1-p_k)是条件 offer 方差。如果假设 1B 对每次抽签都成立，那么每个 $\beta_{IV,k}$ 的权重就是非负的，这样 $\pi_k > 0$，对所有 k 来说都是如此。这种加权方案有可能产生 β 的精确估计，因为它给了有更多学生、更强的第一阶段和更多工具变量变化的数据以更多的权重[5]。

另一个 2SLS 估计式将入学工具变量和风险组虚拟变量之间的

[4] 这个 2SLS 估计的简约式方程可以写成：

$$Y_i = \rho Z_i + \sum_{k=1}^{K} \gamma_k R_{ik} + \xi_i ,$$

其中 γ_k 是简约形式的风险组效应。因为这个模型是恰好识别的（即工具变量的数量等于要使用工具的变量的数量），2SLS 与间接最小二乘法的估计相一致，即用 OLS 对简约式的系数 ρ 的估计除以(7)中第一阶段的系数 π 的估计。

[5] Goldsmith-Pinkham 等（2022）研究了这种基于回归的加权方案的效率。

相互作用添加到排除工具列表中，导致一个具有完全饱和的第一阶段的过度识别模型[6]。具有饱和第一阶段的 2SLS 产生了一个不同的加权平均数，在(8)中用 π_k^2 代替 π_k，详见 Angrist 和 Imbens（1995）。在(6)中残差恒定和同方差的情况下，这种过度识别模型的 2SLS 估计是有效的，因为它产生了（渐进地）最精确的估计，利用收集到的数据可以计算。另一方面，严重过度识别模型的 2SLS 估计可能会比恰好识别模型的估计受到更多的有限样本偏差的影响（Andrews 等，2019；Angrist 和 Kolesár，2021）。

协变量控制

除了风险组控制之外，学校质量的 2SLS 估计通常还包括描述学生背景的额外协变量。虽然不需要消除选择偏差，但非风险协变量通常会提高 2SLS 估计的精确度。一个带有协变量的 2SLS 设置是这样的：

$$Y_i = \beta D_i + \sum_{k=1}^{K} \delta_k R_{ik} + X_i'\mu + \eta_{i,} \tag{9}$$

$$D_i = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + X_i'\psi + \upsilon_i. \tag{10}$$

协变量向量 X_i 可能包括申请人的人口统计学特征，如种族、性别和免费午餐状况，以及在申请人参加抽签之前测量的基线考试分数。

假设入学 offer 的抽签是在风险组内随机分配的，Z_i 和 X_i 在分配风险的条件下是不相关的。这意味着在 Z_i 对风险组虚拟数值和 X_i 的回归中，X_i 的系数应该为零。因此控制 X_i，2SLS 的估计值就不会改变。这个事实是 2SLS 的 Frisch-Waugh-Lovell 定理的一个版本。然而，如果控制 X_i 可以减少结果的残差，那么包括 X_i 的 2SLS 模型的估计值可以预期比省略 X_i 的模型的估计值更精确。

一个部门(sector)的多所学校

重要的是，风险组的想法可以延伸到对涵盖一个以上学校的部门进行分析。假设林恩市有两所 KIPP 中学，KIPP A 和 KIPP B。五年级的申请者可以申请这两所特许学校中的一所。在多校/单一部门的分析中，如果申请者在任何一所学校就读，就被编码为曾在 KIPP 就读；我们没有试图分别识别 KIPP A 和 B 的因果效应。在单一部门的分析中，我们追求共同的 KIPP 处理效果。此外，在这个例子中，林恩的 KIPP 效应也是一个特许学校的效应，因为林恩没

[6] 这个第一阶段是饱和的，因为它包括 $E[D_i \mid Z_i, R_i]$ 的每个值的一个参数。

有非 KIPP 的特许学校。

多校方案为其他 IV 策略打开了大门。例如，我们可以使用两个 offer 虚拟变量作为工具变量，每个学校一个。由于只有一个内生变量表示任何 KIPP 的入学，2SLS 的估计值被两个工具过度识别。在这种情况下，每个工具变量都会产生不同的分配风险。然而，通常情况下，一个恰好识别的单一工具变量的估计是有吸引力的——无论是出于教学的原因，还是因为有许多工具变量的模型（在更复杂的现实世界的应用中）可能会有细小的样本偏差。

典型的恰好识别的设定使用了一个单一的任意 offer 工具变量。在两所学校的情况下，任意 offer 工具变量表示申请人收到一所或两所学校的录取通知，分配风险是两个基本录取事件的联合概率。例如，如果 A 和 B 的抽签是独立的，申请人 i 在学校 A 有风险 $p_i(A)$，在学校 B 有风险 $p_i(B)$，那么 KIPP 的 offer 的风险是 $p_i(A)+ p_i(B)- p_i(A)p_i(B)$。只申请一个或另一个学校的申请人的 $pi\,(A)=0$ 或 $pi\,(B)=0$。因为分配风险仅由申请学校的情况决定了三个值，所以有三个风险组。我们下一节通过对马萨诸塞州城市特许学校的分析来说明这种情况。

2.3 马萨诸塞州城市特许学校的影响

对马萨诸塞州城市特许中学效果的调查突出了学校效能的基本抽签分析的关键因素，使用的是 Angrist 等人（2013）分析的样本。这里的 2SLS 设置包括风险组控制和基线（非风险）协变量。风险组控制包括所有申请的特许学校组合的指标，按申请年份分开。所关注的干预效果是参加马萨诸塞州城市特许中学之一的影响，主要是在波士顿。这项调查将城市特许学校与所有其他公立学校进行比较，包括传统学校和特许学校。

表 1 报告了这个样本的描述性统计。第(2)列显示了 2002 年至 2011 年期间，在 9 个有抽签记录的城市特许学校中的 6038 名申请五年级或六年级的学生的基线（四年级）特征。抽签样本只保留了每个学生最早的申请年份，并排除了获得保证录取的申请者（例如，有兄弟姐妹入学的申请者），从而得到了一组随机抽签录取的学生。为便于比较，同一时期的全部马萨诸塞州城市学区人口的特征显示在第（1）列。

描述性统计结果显示，抽签申请人与更广泛的学生群体之间存在一些明显的差异。特别是，抽签申请人更有可能是黑人，不太可能被划分为英语学习者，而且基线测试分数更高。虽然这两个群体在四年级时的成绩都低于州平均水平，但特许学校申请人的数学和

英语成绩比城市平均水平高出大约 0.1 个标准差（σ）（这里马萨诸塞州每个年级的所有测试分数都被归一化为平均值为 0，标准差为 1）。这些差异说明了选择进入特许学校的可能性，强调了基于抽签的实验的价值。

平衡性和减员

有两个经验性的检查来探究基于抽签的学校效能分析的假设的有效性。第一个是协变量平衡性检查：在随机抽签的情况下，抽中者和未抽中者的基线特征应该是相似的。表 1 第(3)列显示了四年级学生的特征与抽签 offer 虚拟变量的回归系数，抽签 offer 虚拟变量的定义是：如果学生收到任何特许学校的录取通知，则指标等于 1，并控制分配风险。如果学生在抽签日收到录取通知，或在等待名单上收到后来的录取通知，则被编码为录取。下文第 3.4 节将讨论这些录取类型之间的区别。

表 1 第(3)列报告的平衡性系数估计值显示，抽中者和未抽中者之间的差异一致都很小，在统计上不显著，与风险组内随机分配特许学校 offer 的情况一致。对所有特征平衡的原假设的联合检验未能在常规水平上进行拒绝（p=0.69），同样地在风险组内建立了对假设 1A 的经验支持。

即使是在随机分配的情况下，抽签分析也会因赢家和输家的非随机差异而受到影响。例如，一些学生在没有获得特许学校 offer 时可能会退出公立学校系统，但在有机会时却入学特许学校。这种选择性减员可以改变留在公共系统的抽签赢家和输家的构成，可能会产生选择偏差。因此，值得研究抽签赢家和输家之间的后续参与率的差异。当这两个群体的后续参与率相似时，选择偏差不太可能成为一个问题（Lee，2009）。

表 1 的底部报告了在有风险组控制的模型中，抽签后第一年的结果测试分数可用性指标对 offer 指标的回归系数。Angrist 等人（2013）样本的后续参与率为 80.1%，赢家和输家之间的后续参与情况差异不大。这表明在这种情况下，减员并不是一个主要的问题。Engberg 等人（2014）和 Abdulkadiroğlu 等人（2018）讨论了在差异性减员令人担忧的情况下，学校抽签处理效果的界限。

2SLS 估计

一个完整的 2SLS 分析汇报了第一阶段的估计值，以及对感兴趣的因果效应的 2SLS 估计。表 2 列出了基本的特许学校抽签分析的内容。这里的处理变量是在提交申请的那年秋季开始的学年中是

否入学特许学校的指标；结果变量是在年末（在五年级或六年级）的测试成绩。表中的结果包括风险组虚拟变量、学生人口统计学特征、以及基线（四年级）数学和英语分数。

波士顿特许学校的 offer 使特许入学率提高了近 60 个百分点，这是一个很大的第一阶段估计值，在表 2 的第一列中可以看到。这一结果确定了风险组内第一阶段假设 1B 的有效性。表中第二列第一行汇报了 2SLS 的估计值，数据显示特许学校的入学率将数学成绩提高了 0.45σ，令人印象深刻。在这个样本中，非特许学生的分数比州平均水平低 0.32σ。因此，在波士顿特许学校就读一年，预计会将数学成绩提高到高于州平均的水平。相应的 OLS 估计显示在第（1）列，产生的特许效应系数为 0.33σ。显而易见，尽管有正向选择偏差的担忧，但未经测试的模型低估了 IV 效应。一个可能的解释是处理效果的异质性，这是我们接下来要讨论的问题。

3 抽签 IV：实施细节和扩展

3.1 异质性的影响

表 2 的因果模型规定，至少在抽签风险组内，特许学校入学的因果效应是不变的。当然在实践中，特许效应对不同的申请人可能有所不同。个人变化的特许效应是由潜在的结果定义的：$\beta_i = Y_i(1) - Y_i(0)$。不同的学生同样也会对特许学校的 offer 做出不同的反应。异质的第一阶段可以用一组潜在的处理方法来描述，D_i（1）和 D_i（0），分别表示当 $Z_i=1$ 和 $Z_i=0$ 时 i 的特许入学状态。

在一个潜在结果异质的世界里，IV 捕捉到了对特许抽签 compliers 的因果效应。这一点在 Wald 估计值的局部平均处理效应（LATE）解释中得到正式体现。具体来说，Imbens 和 Angrist（1994）和 Angrist 等人（1996）表明：

$$\beta_{IV} = E[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)] .$$

这一结果是在假设独立、排他、存在第一阶段和单调性的情况下得出的，定义如下：

假**设 2A**：独立/排他，$(Y_i(1), Y_i(0), D_i(1), D_i(0)) \perp Z_i$。

假**设 2B**：第一阶段，$E[D_i | Z_i = 1] > E[D_i | Z_i = 0]$。

假**设 2C**：单调性，$D_i(1) \geq D_i(0)\ \forall_i$。

假设 2A 将我们先前的假设 1A 调整为异质性处理效应框架，而假设 2B 与我们先前的假设 1B 一致[7]。新的要求是假设 2C，即抽签 offer 工具变量必须能微弱地提高所有学生的特许学校入学（不仅仅是平均）。这一单调性限制似乎是合理的，因为很难想象抽中 offer 会降低入学的情况。

LATE 理论对特许学校申请者们进行了划分。"Never-takers"是指即使有机会也拒绝上特许学校的学生，因此 $D_i(1) = D_i(0) = 0$，这可能因为一个学生是否参与抽签是由于父母的热情而不是他们自己。"Always-takers"是指即使没有录取通知书也会上特许学校的学生，因此，$D_i(1) = D_i(0) = 1$。这类学生可能会反复申请，或以其他方式找到进入的途径。特许抽签的 "compliers"是那些只有在收到录取通知书时才入学的学生，这意味着 $D_i(1) > D_i(0)$（即 $Di(1) = 1$，$Di(0) = 0$）。LATE 理论表明 β_{IV} 提供了 compliers 上特许学校的平均因果效应。在有多个抽签的情况下，等式(8)意味着带有风险组控制的 2SLS 确定了抽签特有的 LATE 的加权平均值。因此，表 3 中的估计

[7] 假设 2A 结合了独立性和排斥性，这是 LATE 框架中的不同假设。

值应该被解释为衡量申请人通过随机 offer 诱导参加特许学校的因果效应。这些效应可能高于或低于特许学校的总体平均效应，这取决于 compliers 与其他人的比较。

3.2 对 compliers 的描述

Complier 协变量均值

谁是特许抽签者中的 complier？值得注意的是，虽然在任何数据集中都没有对个人抽签者进行编码（因为我们从未见过同一个学生 i 的 D_i（1）和 D_i（0）），但抽签者的特征可以被描述。这方面的工具在 Abadie（2002，2003）中有详细介绍。

Complier 分析的基础是，根据假设 2C，对 IV 分析有贡献的人群只包含 always-takers、never-takers 和 compliers。此外，与 compliers 不同的是，always-takers 和 never-takers 是可以识别的：$D_i = 0$ 和 $Z_i = 1$ 的学生一定是 never-takers，而 $D_i = 1$ 和 $Z_i = 0$ 的人一定是 always-takers。D_i 和 Z_i 的另外两个组合涉及到 compliers 和这些其他群体的混合：$D_i = Z_i = 1$ 的人群是 always-takers 和 compliers 的混合物，而 $D_i = Z_i = 0$ 的群体是 never-takers 和 compliers 的混合。这三组群体的相对规模是可以确定的，因为 never-takers 的比例是由被拒绝的那一部分提供的，always-takers 的比例是由没有收到 offer 的学生中的一部分提供的，而 complier 的比例等于第一阶段的规模。因此，我们可以结合观察到的 always-takers 和 never-takers 的信息，从混合分布中找出 compliers 的特征。

Abadie(2002)用一个简单的 2SLS 等式来实现这一逻辑，用于描述 compliers 的特征：

$$g(X_i, Y_i) \times 1\{D_i = d\} = \psi_d + \gamma_d 1\{Di = d\} + v_{id}, \tag{11}$$

$$1\{D_i = d\} = \phi_d + \pi_d Z_i + e_{id}, \quad d \in \{0,1\}. \tag{12}$$

这里 $g(X_i, Y_i)$ 是学生基线特征 X_i 和抽签后结果 Y_i 的任何函数。在(11)和(12)中设置 $d = 1$ 意味着我们在 IV 程序中使用 Z_i 作为 D_i 的工具变量，$g(X_i, Y_i)$ 乘以 D_i 作为结果变量；设置 $d = 0$ 则表示使用 Z_i 作为 $(1 - D_i)$ 的工具变量，$g(X_i, Y_i)(1 - D_i)$ 作为结果变量。在假设 2A、2B 和 2C（以及 Z_i 独立于 X_i 的假设）下，Abadie（2002）表明，这些非常规的 IV 程序可以恢复被处理和未被处理的 compliers 的特征。

$$\gamma_d = E[g(X_i, Y_i(d)) | D_i(1) > D_i(0)], \quad d \in \{0, 1\}.$$

这个结果有很多有用的影响和应用。首先，通过设置 $g(X_i, Y_i) = X_i$，IV 程序产生了抽签 compliers 的任何前置协变量 X_i 的平均数，

方便了抽签 compliers 与其他群体在可观察层面的比较。注意由于 X_i 不受特许学校干预的影响，在这种情况下，IV 系数 γ_1 和 γ_0 都能恢复 compliers 的平均 X_i，所以这两个参数应该相等。可以通过报告任何一个参数的估计值，或两个估计值的平均值来总结 compliersd 的特征，这两个估计值之间的差异与表 1 中用作平衡检查的按 offer 状态划分的 X_i 的差异成正比。因此，我们能够把这种平衡性检查看成是嵌入在部分遵从性抽签中的隐性 complier RCT 中的干预组和控制组的协变量平衡的等效检查。

表 3 对 Angrist 等（2013）的马萨诸塞州城市特许学校申请者样本说明了这种方法。第（1）列和第（2）列显示了未干预州和干预州对一系列基线特征的两个可用估计值的 complier 平均值。这些结果来自于对(11)和(12)的 2SLS 估计，并增加了可加的风险组控制。根据等式(8)，我们可以将估计值解释为风险组特定 complier 平均数的加权平均值。正如表 1 中的平衡检查所预期的那样，在所有特征方面，经过干预的和未经过干预的 complier 估计值都非常相似。第(3)列显示了一个更有效的对每个 complier 平均数的单一估计，通过叠加(11)和(12)中两个等式的数据并进行 2SLS 估计，在各等式中施加一个共同的系数 γ，并在学生层面上聚类标准误[8]。

表 3 的第(4)和(5)列列出了 always-takers 和 never-takers 的平均特征，以便与 compliers 进行比较。如上所述，这些群体的特征可以分别根据($D_i = 1, Z_i = 0$)和($D_i = 0, Z_i = 1$)的数据单元计算出来。为了与用于计算 complier 平均数的 2SLS 加权法相比较，我们将 $X_i\ D_i\ (1 - Z_i)$ 与 $D_i\ (1 - Z_i)$ 进行回归来估计 always-takers 的均值，将 $X_i\ (1 - D_i)Z_i$ 与 $(1 - D_i)Z_i$ 进行回归来估计 never-takers 的均值，并都控制了风险组变量。这些程序通过基于回归的权重在各风险组之间自动加总，正如第(1)-(3)列中的 IV 估计值一样。

表 3 中的比较展示了马萨诸塞州城市特许学校抽签的行为反应类型之间的差异。与 always-takers 或 never-takers 相比，compliers 不太可能是黑人，更可能是白人或西班牙裔，而且更可能被归类为英语学习者。从基线考试成绩来看，always-takers 是成绩最低的群体。正如第 3.7 节中进一步讨论的那样，compliers 的特征为一组基于抽签的 IV 估计的外部有效性提供了部分引导。

[8] 原则上，通过对叠加进行三阶段最小二乘法（3SLS）估计，利用各等式之间的协方差结构，可以形成一个更有效的估计，这就是编译器平均值（Zellner and Theil, 1962）。

Complier 潜在的结果分布

Abadie (2002)方法的第二个应用产生了 compliers 的边际潜在结果分布[9]。通过在等式(11)和(12)中设置 $g(X_i, Y_i) = Y_i$，$d = 1$，我们得到一个 IV 系数 γ_1，等于特许效应 $Y_i(1)$的 complier 平均数。同样，设定 $g(X_i, Y_i) = Y_i$，$d = 0$，得到的 γ_0 等于非特许效应 $Y_i(0)$的比较者平均值。这表明，平均潜在结果的水平是确定的，而不仅仅是它们之间的差异（即 LATE）。这些平均潜在结果可以与 never-takers 的平均 $Y_i(0)$和 always-takers 的平均 $Y_i(1)$进行比较，这是直接观察到的。这种比较可以作为选择干预的测试基础（Angrist，2004），也可以作为从 LATE 推断其他处理效应参数的参数化建模方法的输入（Bringch 等，2017；Kline 和 Walters，2019）。

事实上，两个潜在结果的全部边际分布都被确定了，而不仅仅是平均值。通过设置 $g(X_i, Y_i) = 1\{Y_i \leq y\}$，对于常数 y 和每个 d，我们得到在 y 处评估的 $Y_i(1)$和 $Y_i(0)$的 compliers 累积分布函数[10]。这反过来意味着，量化的处理效应(QTEs)被确定为 compliers。compliers 的 QTEs 可以通过倒置的 CDF 或通过加权量化回归来计算，这种方法在 Abadie 等人（2002）中有详细介绍。

密度通常比 CDFs 更容易解释，特别是在图形分析中。在一个点 y 的 compliers 密度可以通过在等式(11)设置 $g(X_i, Y_i) = \frac{1}{h}K(\frac{Y_i - y}{h})$ 来进行估计，$K(\cdot)$是一个对称的在零处最大化的核函数，h 是一个渐进地缩减到零的带宽。用这个 $g(\cdot)$的选择来估计等式(11)和(12)，可以得出对 compliers 潜在结果密度的估计值。在应用于私立学校抽签时，Abdulkadiroğlu 等人（2018）提出了一个带宽选择程序，该程序适应 Silverman（1986）的经验法则，以产生适合 compliers 密度估计的带宽[11]。

图 1 展示了 Angrist 等人（2013）对马萨诸塞州城市特许学校申

[9]对 complier 的干预效果的联合分布 $Y_i(1)$-$Y_i(0)$一般没有点识别(point identified)。Frandsen 和 Lefgren (2021)展示了如何使用特许抽签来形成这种联合分布的界限。

[10] 以这种方式获得的候选者潜在结果 CDF 不必是弱增加的。递减的 CDF 预示着违反了基本独立性、排他性或单调性假设。Huber 和 Mellace(2015)以及 Kitagawa(2015)使用了这个想法来测试工具变量的有效性。

[11] Silverman 对高斯核 K( )的经验法则设定 $h=1.06 \times N^{-1/5}\sigma$，其中 N 为样本量，σ 为 结果的标准差。Abdulkadiroğlu 等人（2018）插入了对 compliers 的这些数量的一致估计，通过在等式(11)中设置 $g(X_i, Y_i) = Y_i$ 和 $g(X_i, Y_i)=Y_i^2$，估计 compliers 的数量为第一阶段乘以总样本量和 compliers 潜在结果的前两个矩。

请者样本的这种compliers密度评估技术。一个值得注意的结果是，城市特许学校机构为非白人学生带来的考试分数收益比白人学生大得多，减少了种族成绩差距。图 1 分别报告了白人和黑人学生的数学分数密度，左边是 $Y_i(0)$的密度，右边是 $Y_i(1)$的。这些结果来自于对等式(11)和(12)的 2SLS 估计，每个等式都加入了风险组控制。根据等式(8)，估计值反映了风险组内 complier 密度的加权平均数。上图显示了抽签前一年的基线分数的密度。正如预期的那样，干预组和控制组在基线上的分数分布是相似的。黑人学生的基线分数密度相对于白人学生来说是向左偏移的，这表明存在巨大的种族成绩差距。通过bootstrap Kolmogorov-Smirnov（KS）测试，拒绝了未接受干预和接受干预的 compliers 的基线分数分布的种族平等（$p < 0.01$）[12]。

图 1 的后续部分表明，上城市特许学校消除了抽签者的种族成绩差距。右边的处理密度显示，黑人分布在抽签后明显向白人分布靠拢，这两个分布在七年级时几乎没有区别了。在七年级，KS 检验未能拒绝黑人和白人的 $Y_i(1)$分布的相等（p=0.94）。相比之下，左边的分布显示，随机进入传统公立学校的 complier 存在持续的成绩差距。七年级 $Y_i(0)$分布的种族差距与基线时的差距相似，并且未干预的结果分布的种族平等被拒绝接受（$p < 0.01$）。

3.3 多年级情况

上述分析仅限于抽签后单一年级的结果，并有一个衡量抽签后一年的入学率的处理变量。当一个年级以上的结果可用时，它们可以在一个集合分析中合并。一个叠加年级的 2SLS 设置可以用以下方式描述：

$$Y_{ig} = \beta D_{ig} + \sum_{k=1}^{K} \delta_k R_{ik} + X_{ig}^{'} \mu + \eta_{ig}, \tag{13}$$

$$D_{ig} = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + X_{ig}^{'} \psi + \upsilon_{ig}. \tag{14}$$

其中 Y_{ig} 是学生 i 在 g 年级的结果，协变量向量 X_{ig} 包括年级和日历年的影响以及其他基线特征。由于无论观察到什么年级的学生，分配风险都是固定的，所以风险控制不必因年级而异。

多年级模型通常引入一个年级变化的处理方法，以反映对感兴趣的学校或部门的接触。Abdulkadiroğlu 等人（2011 年）和 Angrist

[12] 关于这个测试程序的细节，请看图中说明。

等人（2013 年）的特许学校研究实现了这一点。具体来说，内生变量 D_{ig} 被定义为在抽签和观察成绩的年级之间的特许学校入学年份。在我们只需要知道在特许学校就读的总时间以满足相关的排他性的假设下，这就调整了因重新申请或辍学而造成的暴露差异。例如，我们假设不存在时间细节上的差异，如在特许学校就读的特定年级。假设上学的总时间调节了 offer 的效果，等式(13)中的 2SLS 估计值捕捉了 Angrist 和 Imbens（1995）中定义的那种平均因果反应（ACR）。ACR 将 LATE 泛化到具有可变强度的处理的情况。在这种情况下，ACR 衡量的是学生入学选择因抽签 offer 而改变的每年影响的加权平均值。

马萨诸塞州城市特许学校申请者的叠加多年级设置产生的估计值出现在表 2 的第（3）-（6）列。这些估计是通过在第一年的样本中加入抽签后的结果到八年级来计算的。第（4）列的估计值使用了与单年分析相同的虚拟内生变量。2SLS 的估计值为 0.58σ，超过了第（2）列的估计值，但这个估计值的大小因申请人经历了长达 4 年的特许入学而变得复杂。第（6）列的估计值是用暴露年限作为内生变量计算的。这一情况下的第一阶段略大于 1，而由此得出的 ACR 估计值为 0.31σ，这意味着在各年级和抽签中，平均而言，每一年的特许入学都会使 compliers 的数学成绩提高大约三分之一的标准差。与单年级分析一样，第（3）列和第（5）列的 OLS 估计值是正的，但比相应的 2SLS 估计值要小，这表明在回归调整后的特许学生和非特许学生的比较中存在适度的向下偏差。

3.4 编码抽签 offer 工具变量

到目前为止的分析，我们还没有考虑过时间问题。在实践中，抽签录取的学校通常会进行第一轮的初步录取，而最初没有被录取的学生则被列入候补名单，由学生按抽签号排序。随着一些初步录取的学生被拒绝，在等待名单上的学生也会被录取。初次录取和候补录取的信息可用于构建特许入学的多种工具变量。

一个自然的双工具变量策略是使用 2SLS 将一个初始录取的虚拟变量和一个表示等待名单录取的虚拟变量结合起来。由于初始录取和等待名单录取的情况可能不同（例如，如果等待名单上的学生在收到录取通知之前在其他地方注册），过度识别的 2SLS 可能会产生比使用单一录取虚拟变量的恰好识别模型更精确的估计。Abdulkadiroğlu 等人（2011 年）对波士顿特许学校的分析显示，双工具设置的效率提升不大。然而，值得注意的是，对过度识别的

2SLS 估计的 LATE 解释需要一个更强的单调性假设，而不是通常在单一 IV 中引用的单调性假设（Mogstad 等，2021）。

de Chaisemartin 和 Behaghel (2020)讨论了在每次抽签学生人数不多的情况下使用等待名单 IV 的设定。当学校以班级规模为目标时，在申请者很少的抽签中，从等待名单上的 offer 构建的 IV 可能与潜在的结果相关联。这个问题源于这样一个事实，即在班级规模的目标下，最后收到录取通知的学生必须是一个 complier，从而导致 compliers 在录取中的比例过高。de Chaisemartin 和 Behaghel（2020）提出了一个加权的 IV 估计，改善了使用等待名单上的 offer 的 IV 估计中的偏差。最初的 offer IV 也通过使用不取决于录取情况的预先确定的录取截止点来规避这个问题。

原则上，随机排序的抽签名单的数据可以用来构建更精确的估计，同时避免使用已实现的等待名单 offer。用 L_i 表示在特许抽签中分配给申请人 i 的顺序，申请人从{1, ..., $\bar{L}$}中随机排序。一个初始 offer IV 是一个指标，用于表明 L_i 低于固定的截止点 C。更一般地说，作为工具使用的 L_i 的有效函数是每个抽签号码的预期特许入学率，可以在给定的录取模型中计算出来。例如，假设一所学校计划在招收到 C 级学生之前一直提供 offer，offer 在不同的学生之间是独立的，而且没有 always-takers。那么，offer 的总数等于 C，再加上一个成功率为 C 的负二项式随机变量，成功概率等于 offer 的接受率 π。这意味着抽签名单上的学生 $L_i = l$ 最终进入该特许学校的可能性是:

$$\Pr[D_i = 1 \mid L_i = l] = \pi[1 - 1\{l > C\}(1 - \frac{\int_0^{\pi} u^{l-C-1}(1-u)^{C-1}du \times (l-1)!}{(l-C-1)!(C-1)!})]$$

这个等式显示，$L_i \leq C$ 的学生可以保证得到录取，在这种情况下，入学的概率是符合率 π；对于 $L_i > C$ 的学生，第二个因素捕捉到在等待名单上的位置至少还有一个席位要提供给他们的概率。我们还没有看到将这种最佳的 IV 方法应用于抽签准实验。在实践中，当抽签规模较大时，使用等待名单录取的问题可能并不重要。

3.5 多部门模式

抽签框架可以延伸到一次性捕捉多个部门效应的模型和方法。例如，我们可能对区分 KIPP 特许学校和属于其他网络的学校的效应感兴趣。在介绍了多部门框架后，第 4 节概述了利用集中的算法分配来识别和估计地区效应的方法。第 5 节扩展了这一内容，概述

了增值模型，允许许多学校中的每所学校有不同的因果效应，而不考虑部门。虽然大体上相似，但每个分析领域都提出了独特的概念和实施挑战。

抽签 compliers 的反事实命运

第 2.3 节中概述的对城市特许学校的分析提出了一个重要的概念问题：与什么相比？在某一特许学校的申请者中，抽签落选者可能就读于传统的公立学校、属于另一部门的特许学校，或一些考试学校，仅举这几个备选。因此，描述各地区的招生分布情况是有帮助的，可以进行比较。正如 Abdulkadiroğlu 等人（2014）和 Chabrier 等人（2016）一样，我们将其称为反事实命运的分布[13]。

Cohodes 等人（2021）对波士顿新的和老的特许校园的研究数据说明了反事实命运的想法。这项分析估计并比较了新开的特许学校（在 2010 年投票倡议取消波士顿特许学校上限之后）和老学校的影响。在马萨诸塞州特许学校的本地话中，允许开设新校区的学校被称为 "成熟的提供者"，而与之相关的新学校是 "扩张校区"。这里的分析也考虑了其他不受扩张影响的特许学校和试点学校的影响，是波士顿特许学校的一个替代方案。

表 4 总结了改革后每个特许学校类型的 compliers 的反事实的命运。这些估计值是通过对等式的 2SLS 估计计算出来的。等式(11)和等式(12)中，对于申请某个特许学校类型的人，d=0。根据该特许学校类型的 offer 和入学对 Z_i 和 D_i 进行编码，并将 $g(\bullet)$设为特定地区入学率的一个指标(跟前面一样控制风险组）。第(1)列显示，虽然 53%的没有收到主校区抽签 offer 的人参加了传统的公立学校，但另外 27%的人在扩张校区入学。同样，在抽签的其他特许学校中（既不是主校区也不是扩张校区），23%的抽签失败者最终被扩张的特许学校录取。相反，第(3)列显示，在扩张型特许学校抽签中，很少有未经处理的 compliers 进入其他特许学校类型，因此，这一组的反事实主要是由 BPS 地区学校（传统公立学校或试点学校）组成。这些结果表明，了解反事实的构成对于解释每种特许学校类型的抽签结果非常重要。

多部门 2SLS

表 4 中记录的反事实入学模式促使我们在一个统一的框架内对多个学校地区进行分析，而不是像第 2.3 节中那样将单个地区与一

[13] 见 Kline 和 Walters（2016）和 Feller 等人（2016）对早期儿童项目背景下的反事实命运的分析。

个综合反事实进行对比。我们最初采用的是一个恒定效应的因果模型，该模型描述了在几个不同的学校类型中，上其中一个学校的后果。

假设一个地区的每个学生都在编号从 0 到 S 的 S 部门中的一所学校上学，其中 0 部门代表传统的公立学校。我们定义了一组互斥且详尽的虚拟变量来代表每个部门的入学，$D_{is} \in \{0, 1\}$表示在 s 部门上学，对每个学生来说$\sum_{s=0}^{S} D_{is} = 1$。那么，一个具有特定地区效应的因果模型就可以得到：

$$Y_i = \mu + \sum_{s=1}^{S} \beta_s D_{is} + \varepsilon_i,$$

参数 β_s 衡量衡量的是相对于传统公立学校被忽略的部门，在 s 部门的学校上学的效果。如前所述，ε_i 代表可能与部门入学选择有关的未观察到的学生异质性。

现在假设我们有一组抽签，用于每个部门的招生。和第二节一样，假设有 K 个相互排斥的抽签组，用 $R_{ik} = 1$ 表示学生 i 参加了抽签 k。这些抽签组应该被看作是对应于学生可能参加的特定学校抽签的所有组合。用 Z_{is} 表示，如果学生 i 至少收到 s 部门的一所学校的 offer，则指标等于 1。请注意，当学生可以申请多个抽签活动时，Z_{is} 不一定是互斥的，因为学生可能收到多个地区的 offer。将等式(9)和(10)扩展到多地区的环境中，就会出现以下具有多个内生变量的方程组：

$$Y_i = \sum_{s=1}^{S} \beta_s D_{is} + \sum_{k=1}^{K} \delta_k R_{ik} + X_i' \mu + \eta_i, \tag{15}$$

$$D_{is} = \sum_{m=1}^{S} \pi_{ms} Z_{im} + \sum_{k=1}^{K} \tau_{ks} R_{ik} + X_i' \psi_s + \upsilon_{is}, s \in \{1, ..., S\}. \tag{16}$$

与第二节相同，我们可以把等式(15)中的 δ_k 和 μ 看作是 $\mu + \varepsilon_i$ 对风险组指标和协变量的投射系数，而(16)中定义的 S 第一阶段等式是部门入学指标与风险组和协变量对抽签 offer 的投射。两阶段最小二乘法估计的过程是通过 OLS 拟合每个第一阶段等式，然后在代入第一阶段预测值 D_{is} 后通过 OLS 运行(15)。经过上述的简单扩展，这个 2SLS 程序将恢复地区效应的一致估计值 β_s，条件是提供 Z_{is} 独立于风险组内的能力 η_i，并为每个地区引起足够的入学率变化。如第 3.3 节所述，将这一设置直接扩展到多个年级的堆积结果是很简

单的。

表 5 再现了 Cohodes 等人（2021）对波士顿几种特许学校类型对数学成绩影响的估计。这项分析将主校区、扩张校区和其他非扩张的特许学校视为独立的部门，还区分了扩张改革前后的入学。按照第 3.3 节所述的方法，本分析将所有抽签后观察到的抽签申请人的分数堆积起来，并将内生变量 D_{is} 进行编码，作为在每个地区上学的年数。如第 3.4 节所述，工具变量是每个地区类型的初始和等待名单录取的指标，模型控制了所有学校逐年具体招生抽签的交叉点的风险组指标。

这个多部门模型的 2SLS 估计显示，在主校区和扩展校区就读的实质性处理效应为每年三分之一个标准差的程度。主校区的效果在扩张改革前后是相似的，这表明本体学校的有效性并没有因为扩张到新的地点而被稀释。其他非扩张型的特许学校的效果是正的，但比被选为进行扩张的学校的效果要小，表明马萨诸塞州指定了更有效的学校进行扩张。在一个单一的 2SLS 模型中包含这些特许学校类型，意味着我们可以把每个人的影响解释为相对于相同的传统波士顿公立学校的基准。

这种多部门方法有另外两个特点值得注意。首先，像等式(15)这样的多部门模型通常依赖于每个部门对学生的影响不变的假设。Imbens 和 Angrist（1994）的 LATE 结果并不适用于具有多个内生变量的模型，而且一般来说，对此类模型的 2SLS 估计的因果解释需要对效应的异质性或对工具变量的行为反应进行强有力的限制（Behaghel 等，2013；Kirkeboen 等，2016；Bhuller 和 Sigstad，2022）。第二，与此类似，这里描述的多部门模型将同一部门的所有学校视为同等有效。部门内学校质量的异质性使估计值的解释变得复杂，并产生了违反排他性的可能性。例如，一个学生从一个扩张的特许学校转到另一个学校，以应对抽签的变化，在等式(15)中的内生变量没有变化，但如果两所学校的质量不同，可能会经历一个分数的变化，从而违反了排他性。请注意，即使是对单一部门的评估，这种排他性的违反也是一个潜在的问题，因为它将干预编码为在该部门的任何特许学校机构就读，包括没有抽签的特许学校机构。

这些问题促使我们采取一种方法，进一步扩展部门模型，允许每个学校有自己的因果 "附加值"。然而，如果只有一部分学校持有超额的抽签，到目前为止描述的抽签方法不能直接应用于这样的模

型，因为抽签不能产生足够的工具变量来识别所有学校的影响。我们将在第 5 节回到个别学校增值和抽签超额的话题。

3.6 作为局部抽签的录取不连续问题

与上述抽签方法密切相关的研究设计利用了基于入学考试或其他标准的录取规则的不连续性。例如，在波士顿、纽约和其他地方，高度选择性的考试学校要求进行入学测试，并录取分数足够高的学生。我们可以把随机抽签看作是这种具有随机分配的录取分数的录取规则的一个特例——正如第 3.4 节所讨论的，抽签 offer 被分配给那些在随机排序的名单上的位置低于阈值的学生。当录取分数不是随机分配的时候，高于和低于录取门槛的学生一般不会有可比性。这个问题可以用回归不连续（RD）的方法来解决，该方法在录取门槛的一个小范围内归零，将偶然高于或低于门槛的类似学生隔离出来进行比较[14]。

我们通过回到第 3.1 节的潜在结果模型来介绍录取 RD 的方法，这次描述的是参加考试学校的影响。让 $Y_i(1)$和 $Y_i(0)$表示学生 i 在考试学校或传统公立学校就读的结果，让 $D_i(1)$和 $D_i(0)$表示有无考试学校录取通知书时 i 的就读选择，而不是像在特许学校抽签中那样被随机分配。考试录取通知书 Z_i 是根据观察到的考试分数 T_i 中的一个临界点 c 分配的：

$$Z_i = 1\{T_i \geq c\}.$$

我们假设潜在的结果满足以下假设：

假设 3A： 平均潜在结果在门槛上是平稳的，

$\lim_{t\to c-} E[Y_i(d)\,|\,T_i = t] = \lim_{t\to c+} E[Y_i(d)\,|\,T_i = t]$, $\lim_{t\to c-} E[D_i(z)\,|\,T_i = t] = \lim_{t\to c+} E[D_i(z)\,|\,T_i = t]$ ，对于 $(d,z) \in \{0,1\}^2$.

假设 3B： 越过门槛会增加入学率，$\lim_{t\to c+} E[D_i\,|\,T_i = t] > \lim_{t\to c-} E[D_i\,|\,T_i = t]$.

假设 3C： 局部单调性，$Pr[D_i(1) \geq D_i(0)|T_i = c] = 1$.

这三个假设是假设 2A、2B 和 2C 的局部版本，只适用于分数在录取门槛附近的学生。假设 3A 中的平稳性条件要求学生不能精确

[14] 关于从录取分数线得出的教育项目评估的其他例子，见 Hoekstra（2009）、Zimmerman（2014）、Card 和 Giuliano（2016）、Kirkeboen 等（2016）、Dustan 等（2017）、Hastings 等（2019）、Zimmerman（2019）、Anelli（2020）、Sekhri（2020）、Jia 和 Li（2021）。(2017)，Heinesen（2018），Hastings 等（2019），Zimmerman（2019），Anelli（2020），Sekhri（2020），Jia 和 Li（2021），Bleemer 和 Mehta（2022），Beuermann 和 Jackson（2022），以及 de Roux 和 Riehl（2022）。

地操纵他们与门槛有关的分数，因此那些刚刚超过和刚刚低于门槛的学生在预期中是相似的。在这些条件下，我们可以认为录取门槛为 T_i 接近 c 的学生定义了一个局部随机抽签，并将前面介绍的基于抽签的方法应用于这个子群体。这个想法与最近将 RD 设计分析为局部随机试验的研究相一致（Cattaneo 等，2016）。

在假设 3A、3B、3C 的条件下，Wald 比率的局部版本可以识别接纳阈值的 compliers 的 LATE。具体来说，我们有

$$\beta_{RD} \equiv \frac{\lim_{t\to c+} E[Y_i \mid T_i = t] - \lim_{t\to c-} E[Y_i \mid T_i = t]}{\lim_{t\to c+} E[D_i \mid T_i = t] - \lim_{t\to c-} E[D_i \mid T_i = t]} = E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0), T_i = c].$$

这个表达式可以看作是在阈值 c 的附近，用门槛指标 Z_i 作为考试入学 D_i 的工具变量。

录取不连续设计的实证实施将第 2 节的基本抽签 2SLS 方法改编为局部抽签实验。在这里，风险控制涉及录取考试分数 T_i，在 RD 中被称为驱动变量。那些驱动变量值远离分界线的申请者有退化的风险：他们要么被处理，要么不被处理，概率是 1。驱动变量值接近分界线的申请者可能通过也可能不通过。

在这种情况下，最简单的风险控制策略是参数化的：包括作为协变量的驱动变量的多项式函数，在没有处理的情况下调整驱动变量和结果之间的关系。然而，在实践中，只关注相关截止点（cutoff）附近的申请人的非参数策略往往更有说服力。局部或非参数方法的正式动机来自于这样一个事实，在 Abdulkadiroğlu 等人（2021）中精确的限制意义上，在一个不断缩小的带宽或分界线周围的间隔中，申请者被提供席位的极限概率等于二分之一。

参数和非参数的 RD 估计策略可以通过以下 2SLS 设置来描述：

$$Y_i = \mu + \beta D_i + (1 - Z_i) f(T_i - c; \delta_0) + Z_i f(T_i - c; \delta_1) + \eta_i, \tag{17}$$

$$D_i = \psi + \pi Z_i + (1 - Z_i) f(T_i - c; \tau_0) + Z_i f(T_i - c; \tau_1) + \upsilon_i. \tag{18}$$

这些等式用驱动变量的平滑函数取代(6)和(7)中的风险组指标。$f(t;\delta)$，参数为 δ，满足 $f(0;\delta)=0$。在实践中，这个函数通常是一个多项式，因此，$f(t;\delta) = \sum_{k=1}^{K} \delta_k t^k$。决定多项式系数的参数允许在阈值的两侧，并且允许在第一和第二阶段的等式中有所不同。

参数化 RD 策略使用所有或大部分感兴趣的样本来计算该模型的 2SLS 估计值，通常有灵活的驱动变量控制。非参数 RD 降低权重或删除离截止点较远的观测值，并使用不太灵活的驱动变量控制

（通常是线性）。后者需要选择带宽来确定非参数估计中使用的样本量和权重。如何最好地选择带宽的问题刺激了大量的和正在进行的理论工作（Imbens 和 Kalyanaraman，2011；Caloneico，2014）。

图 2 取自 Abdulkadiroğlu 等人（2014）的研究，绘制了纽约市（NYC）高度选择性考试学校的 RD 第一阶段和简约式关系。A 组显示了纽约市三所考试学校--布鲁克林科技大学、布朗克斯科学大学和斯蒂文森大学的申请者的入学率，作为学生的录取分数与各学校录取分数线的距离的函数。该图显示了在整个分界线上入学率的大幅跳升，表明考试学校入学率的第一阶段很强。B 组显示了相应的简约式对 Regents 数学标准化考试分数的影响。Regents 的分数在考试学校的录取分数线上是平稳的，显示出上考试学校对 compliersde 考试分数没有影响。尽管考试学校和传统公立学校在成绩上存在巨大差异，但这些零影响还是发生了。这说明了基于不连续的实验在区分因果效应和选择偏差方面的力量。

3.7 外部有效性

本节的结果表明，基于抽签的研究设计可以识别 compliers 的平均处理效应，这是一个定义明确且可解释的子群。基于抽签的估计也与评估某些政策改革有关。Kline 和 Walters（2016）表明，在没有外部性和溢出效应的情况下，LATE 抽签申请人中可用席位数量边际增加的政策相关参数。人们往往有兴趣问，抽签 LATE 的外部有效性是否会延伸到其他子群或政策变化。至少有四种形式的外部有效性是值得考虑的[15]。

首先，在抽签申请人中，对抽签者的影响可能不同于对 always-takers 和 never-takers 的影响。虽然这些其他群体的申请者的处理效应没有被确定，但可以用第 3.2 节的方法来评估他们观察到的特征或潜在结果的水平是否与 compliers 的不同。这种分析可以提供一种感觉，即 compliers 是否能代表所有的申请者。值得注意的是，在许多学校的抽签中，always-takers 可能很少或不存在，因为学生可能很难在没有收到录取通知的情况下进入学校。在这种片面的不服从的情况下，LATE 等于申请者中的处理效应（TOT），这是项目评估文献中的一个传统目标参数（Bloom, 1984）。

第二，对抽签申请人的影响可能与选择不申请抽签的学生的影响不同。表 1 显示，在马萨诸塞州城市特许学校的例子中，申请者

[15] 见 List（2021）和 List 等人（2021）对教育项目的外部有效性和规模的相关讨论。

和非申请者的特征不同，这表明处理效应也可能不同。为了推断非申请者的人群，利用其他类型的实验来改变抽签申请人的组成是很有用的。在波士顿特许学校的一个此类应用中，Walters（2018）将基于特许学校距离的工具变量与随机抽签结合在一个广义的 Roy 模型框架中（Roy，1951；Eisenhauer 等，2015）。直观地说，由于特许学校附近的学生更有可能申请，那些从附近申请的学生比从远处申请的学生更少被选中。在假设距离与其他观察到的特征一样是随机分配的条件下，按距离划分的抽签 LATE 的变化因此可以用来区分申请倾向和处理效应之间的关系。Walters（2018）的结果表明，特许学校申请者在成绩提升方面被负向选择，因此，如果特许学校教育扩大到新的人口，处理效应可能会更大。与这一发现相一致，Cohodes 等人（2021）表明，波士顿的特许学校在改革后继续产生巨大的成绩提升，因为改革后的申请人群更能代表波士顿的整体情况。沿着类似的思路，Abdulkadiroğlu 等人（2016 年）将抽签记录与基于传统公立学校的特许学校接管的替代研究设计相结合，显示出对抽签申请人和被特许学校认可的学生的类似影响。

第三，有抽签记录的学校可能与非抽签学校不同。例如，申请人数多于席位的热门学校可能比不太热门的学校更有效，或者有行政能力保留有组织的抽签记录的学校可能更有效[16]。Abdulkadiroğlu 等人（2011）和 Angrist 等人（2013）报告的非实验性估计表明，马萨诸塞州超额申请的特许学校比没有抽签的学校更有效，Baude 等人（2020）表明，在德克萨斯州，更有效的特许学校随着时间的推移获得市场份额。另一方面，Ab-dulkadiroğlu 等人（2018）报告说，入学率下降的私立学校更有可能选择加入抽签的凭证计划。Abdulkadiroğlu 等人（2020）表明，在纽约市的高中里，学校的受欢迎程度与学校效能关系不大。Abdulkadiroğlu 等人（2014）和 Dobbie 和 Fryer（2014）显示，在波士顿和纽约，备受追捧的考试学校的效果有限。因此，不清楚我们是否应该期望超额申请的学校在总体上更有效。

最后，教育市场中存在的抽签式学校选择方案可能会通过竞争或其他渠道对其他学校的学生产生溢出效应。在这种情况下，抽签可以衡量内部有效的对申请人的部分均衡影响，但可能会失去更广泛的一般均衡效应。识别这种溢出效应通常需要从市场结构的变化而不是学生层面的入学机会的变化中获得替代的研究设计。这个模

[16] 这个想法是 Allcott（2015）研究的 "地点选择偏差"的一个版本。

式的研究例子包括 Figlio 和 Hart（2014），Gilraine 等（2021），以及 Campos 和 Kearns（2022）。

4 集中分配

4.1 延期接受与单一决胜（Tie-breaking）

许多城市学区利用集中分配的算法在全区范围内实施选择[17]。就像上面讨论的分散式学校招生抽签一样，许多集中式分配系统也有随机因素，以打破具有相同匹配标准的学生之间的联系。然而，与简单的学校抽签不同，集中式匹配中的基本风险组的性质通常被一个看似复杂的迭代过程所掩盖。Abdulkadiroğlu 等人（2017）和 Abdulkadiroğlu 等人（2021）展示了如何隔离集中分配算法中的随机变化，以及如何利用这种变化来估计因果效应。

著名的 Gale 和 Shapley（1962）延迟接受（DA）算法是最广泛用于学校分配的算法[18]。为了简述这一机制，考虑一组 N 个申请者（以 i 为索引）申请一组具有固定容量的 J 个学校（以 j 为索引）。申请人提交对学校偏好的排序列表，定义了一组部分偏好排序，用 $\succ_i$ 表示。申请人还被赋予每个学校的优先权（例如，有兄弟姐妹入学的申请人可能是最高优先权，其次是住在附近的申请人），用 ϕ_{ij} 表示 $\in \{1, .., P, \infty\}$，其中 $\phi_{ij} < \phi_{kj}$ 表示学校 j 优先考虑申请人 i 而不是申请人 k。申请人的类型定义为 $\theta_i = (\succ_i, \phi_i)$，其中 $\phi_i = (\phi_{i1}, \ldots, \phi_{iJ})$收集了申请人在所有学校的优先权；类型收集在 $\theta = (\theta_1, \ldots, \theta_N)$。由于优先权是粗略的（即有较少的比其他学生优先的类别），学生类型被进一步增加了一组同分决胜（tie-breaking）的数字 $g = (g_1, \ldots, g_N)$，$g_i \mid \theta \sim U(0, 1)$。每个学生的增强优先级由 $\phi_{ij} = \phi_{ij} + g_i$ 给定。DA 机制将输入（g，θ），并使用以下算法计算学生的分配情况:

- 第 0 步：每个申请人根据 $\succ_i$，申请她最喜欢的学校。每所学校根据增强的优先权 ϕ_{ij} 对这些申请者进行排名，并暂时录取排名最高的申请者，但以其容量为限。所有其他申请人都被拒绝。
- 步骤 k>0：每个在步骤 k-1 中被拒绝的申请人都申请她下一个最喜欢的学校。每个学校对这些新申请者和它在步骤 k-1 中录取的申请者进行排名（通过 ϕ_{ij}）。每所学校从这个申请池里暂时录取排

[17] 实行集中派位制度的城市包括巴尔的摩、波士顿、马萨诸塞州剑桥、新泽西州卡姆登、芝加哥、丹佛、印第安纳波利斯、明尼阿波利斯、纽瓦克、纽约市、新奥尔良、奥克兰、旧金山、西雅图、塔尔萨和华盛顿特区。集中派位在全球也很普遍，并在不断增长，截至 2020 年，有 51 个国家在小学或中学阶段采用这种做法（见 https://www.ccas-project.org/）。

[18] 对市场设计（market design）的经济研究涵盖了对 Gale-Shapley 等匹配工具的研究和使用。Abdulkadiroğlu 和 Andersson（2022）回顾了学校选择的市场设计方法。

名最高的申请者，直至录取完为止，拒绝其余的申请人。当没有新的申请人时，DA 就会终止，返回一组分配，$Z = (Z_i, \ldots, Z_N)$，其中 $Z_i = j$ 表示将学生 i 分配到学校 j[19]。

在一个较高的水平上，带有单一的同分决胜的 DA 机制产生了一个从能力、学生类型和抽签号码集到学校分配集的函数 $M(\bullet)$：$M(g, \theta) = Z$。和上面的简单抽签工具变量一样，集中分配的 Z_i 很可能受到 g 的随机变化的影响：在其他条件不变的情况下，g 较低的学生更有可能被分配到一个理想的学校。但与简单的抽签不同的是，现在有一个将这种随机性转化为分配的复杂过程，$M(\bullet, \theta)$，它取决于非随机学生类型的全部向量。具有较高优先权和/或某些偏好的学生更有可能被分配到某些学校，不管他们抽到的是哪种同分决胜。以申请者类型为条件，学校的录取是可忽略的。DA 的平等待遇属性确保具有相同 θ_i 的申请人具有相同的学校分配概率。然而，正如 Abdulkadiroğlu 等人（2017）所表明的，在规模大的地区，这种条件是不切实际的，因为有几乎和学生一样多的类型。在高维度的情况下，DA 在类型的条件下产生的有用变化很少。

Abdulkadiroğlu 等人（2017）对这个问题的解决方案利用了 Rosenbaum 和 Rubin（1983）的倾向得分，定义为分层 RCT 中处理的分层条件概率。在有抽签同分决胜的 DA 匹配中，干预由虚拟变量 $Z_{ij} = 1\{Z_i = j\}$表示。相关的倾向得分是一组概率 $p_{ij} \equiv E(Z_i \mid \theta)$，每个概率都是学生偏好和优先权的高维列表的标量函数。Rosenbaum 和 Rubin(1983)的倾向得分定理意味着，在一个以 θ 为条件进行随机处理的实验中，对 p_{ij} 的控制可以消除 θ 和潜在结果之间的关系所产生的任何 OVB。换句话说，p_{ij} 的集合可以用来确定集中分配所引起的风险组合。

4.2 理论和模拟的倾向得分

一般来说，DA 倾向得分是类型的一个未知函数。但 Abdulkadiroğlu 等人（2017）推导出 DA 倾向得分的大市场（large-market）近似值，鉴于学生偏好和优先权的数据，很容易计算出来。具体来说，Abdulkadiroğlu 等人（2017）推导了在连续经济中（continuum economy）集中分配倾向得分的等式，其中有大量的学生申请有限数量的学校。连续经济的分数很容易计算，与有限市场

[19] DA 分配是稳定的，因为没有一对匹配的学生和学校愿意交换分配（一个 "阻断对"）。当学生可以对所有的学校进行排名时，DA 也是策略预防的，因为学生从错误地报告他们的偏好中得不到任何好处。关于这些概念和相关概念的回顾，见 Roth 和 Sotomayor（1990）。

的分数非常接近，而且通常足够准确，支持使用 Z_{ij} 来估计因果效应。

大市场（large-market）近似法是通过定义特定学校的分界线，定义为每所学校的最后一个抽签号码。在连续经济中，分界线是非随机的，所以每个申请人的分配率仅由相关的分界线（由他或她的优先级决定）和分界线周围的同分决胜变化决定。申请人的分配类型是相互独立的，这是一个进一步的简化。由此产生的大市场 DA 倾向得分将每个学校的申请者分为三组：总是、从不和有条件地在学校就读的申请者，这取决于他们的席位优先权相对于学校的优先权截止点的位置[20]。

大市场分数的主要好处是降维：即使在有数千种类型的匹配中，大市场分数也是由（相对）粗略的学校分界线集决定的。大市场分数还有一个其他的好处，即区分集中分配风险的不同来源。例如，因果效应可以分别对有条件入学的申请人和总是入学的申请人进行估计。这些群体的因果效应的差异有时可以与学校选择的经济模型联系起来，如罗伊式的收益选择。Abdulkadiroğlu 等人（2017）表明，识别这些不同人群的公式适用于其他集中式机制，如随机独裁算法，并可以扩展到具有多个同分决胜的 DA（即不同学校的不同抽签号码）和立即接受机制（有时被称为 "波士顿机制"）。满足平等待遇（ETE）属性的更大类随机机制的倾向性得分可以通过多次重新抽出随机的同分决胜号码来模拟，并分别计算每个申请人被分配到一个特定学校的模拟份额。这些模拟的分数使我们不再需要大市场的近似值。

一些集中分配方案，如用于波士顿和纽约市考试学校和纽约市筛选学校的方案，采用了非抽签同分决胜的方法，如考试分数，而不是抽签号码，或与抽签号码一起。Abdulkadiroğlu 等人（2021 年）展示了如何将非随机筛选与抽签变化结合在一个统一的局部 *DA* 分数方法中。局部分数也是在一个大市场模型中得出的，有连续的申请人和一组连续分布的同分决胜。和以前一样，大市场模型允许将学生类型划分为总是、从不和有条件地被录取的学生，并产生简单、粗略的分配风险公式[21]。这个框架将 RD 式的识别策略推广到有多个处理和驱动变量的环境中。

[20] 对于从未有席位的申请者，$p_{ij} = 0$，而对于总是有座位的申请者，p_{ij} 等于 i 没有被分配到她喜欢的 j 学校的概率。对于有条件有席位的申请者，p_{ij} 是 i 通过 j 的分数线，但成绩不比 j 好的概率。Abdulkadiroğlu 等人（2017）展示了后两个概率，以及 p_{ij} 是如何由 i 的等级顺序列表中的学校组的大市场截止点决定的。

[21] 精确的定义和公式见 Abdulkadiroğlu 等人（2021）的 4.2 节。

Borusyak 和 Hull（2020）对 Abdulkadiroğlu 等人（2017）的方法进行了进一步的扩展，以在集中式分配系统中进行因果推断，他们展示了倾向得分解决方案如何能适用于任何变量 $Z_i = M_i\,(g, w)$，根据已知的公式 $Mi\,(\bullet)$(如 DA 机制)，将外生冲击 g(如随机同分决胜者)和非随机数据 w(如申请人的偏好和优先权)结合起来[22]。正如我们在下面进一步讨论的那样，控制 $\mu_i = E[M_i\,(g, w)\,|w]$，在其他部分固定的情况下，将 Z_i 平均到外生冲击上，这足以消除分配给不同 Z_i 值的个人比较中的选择偏差。这使倾向性评分方法普遍适用于多值或连续处理的环境。重要的是，即使 Z_i 不是由满足 ETE 属性的机制产生的，或者说，在 Z_i 表示集中式学校分配的情况之外，这一结果也是成立的。Borusyak 和 Hull(2020)通过反复抽出 g，并对抽出的 Z_i 进行平均，固定非随机的 w，来计算 μ_i。虽然可能需要计算，但这种模拟程序产生了一个从复杂（但已知）的分配方案中提取有用变化的一般方法。

DA 还产生了各种其他专门的工具变量，有更简单的倾向性得分。其中一个是关于申请人是否获得其第一选择学校的虚拟变量。这个第一选择的分配虚拟变量是以第一选择的偏好和优先风险组为条件随机分配的，这可以被简单地控制。另外，资格工具变量表明，g_i 是否优于提供席位的最差抽签号码（控制了排名的学校集）。虽然有效，但第一选择和资格工具变量可能会留下很多有用的分配变化（Narita，2016）。我们在下面说明这一现象[23]。

4.3 用分数控制进行估算

一旦计算出来，集中分配的倾向性得分就可以用各种方式来估计学校效能。例如，Abadie（2003）提出了 LATEs 和相关参数的估计方法，通过工具变量倾向性得分进行逆加权。Rosenbaum 和 Rubin(1983)提出的倾向得分匹配，是获得当地因果效应的等权平均数的另一种选择。一个简单的选择是在线性 IV 回归中调整分配倾向得分。对于一个给定的学校 j，考虑 IV 的 second 和第一阶段：

$$Y_i = \beta D_i + X_i^{'}\mu + \eta_i, \tag{19}$$

[22] Borusyak 和 Hull（2020）也讨论了局部的解决方案，其中 g 被视为用户指定带宽内的随机性。请注意，与激励性的 DA 机制不同，对于一般的 Borusyak 和 Hull（2020）解决方案，g 中的外生冲击不需要定义在相同的观察 "水平"。

[23] 集中分配机制中第一选择 IV 的例子包括 Deming（2011）、Abdulkadiroğlu 等（2013）、Deming 等（2014）和 Hastings 等（2009）。资格 IV 的例子包括 Dobbie 和 Fryer（2014），Lucas 和 Mbiti（2014），以及 Pop-Eleches 和 Urquiola（2013）。第一选择 IV 也被用于分散的分配机制（Abdulkadiroğlu 等，2011；Cullen 等，2006；Dobbie 和 Fryer，2011；Hoxby 等，2009）。

$$D_i = \pi Z_{ij} + X_i'\psi + \upsilon_i, \tag{20}$$

其中，D_i 表示在学校 j 的入学，X_i 是前置控制的向量，$z_i = Z_{ij} - p_{ij}$ 是一个"重新中心化"的 offer 工具变量，从 offer 指标 Z_{ij} 中减去倾向得分 p_{ij}。Borusyak 和 Hull（2020）表明这种情况如何识别条件型 IV 系数的加权平均值。特别地，IV 估计值由以下方式给出：

$$\beta_{IV} = \int w_{IV}(t)\beta_{IV}(t)dF_\theta(t), \tag{21}$$

其中，$F_\theta(\cdot)$给出了类型 θ_i 的分布并且

$$\beta_{IV}(t) = \frac{E[Y_i \mid Z_{ij}=1, \theta_i = t] - E[Y_i \mid Z_{ij}=0, \theta_i = t]}{E[D_i \mid Z_{ij}=1, \theta_i = t] - E[D_i \mid Z_{ij}=0, \theta_i = t]}$$

是类型 $\theta_i = t$ 的学生的 Wald 估计值。类似于等式(8)中的加权函数。等式(21)中的权重 $w_{IV}(t)$积分为 1，并与 $\theta_i = t$ 类型的学生份额、条件类型分配方差 $Var(Z_{ij} \mid \theta i = t)$ 和条件第一阶段 $E[D_i \mid Z_{ij} = 1, \theta_i = t] - E[D_i \mid Z_{ij} = 0, \theta_i = t]$ 成比例。因此，等式(19)中的 IV 系数 β_{IV} 等于分配微弱地增加入学时$\beta_{IV}(t)$的凸平均值，并且可以解释为假设 2A 和 2C 的条件型类似情况成立时 LATE 的加权平均值。

等式(19)和(20)可以看作是对高维类型向量 θ_i 的风险组控制 IV 情况在等式(6)和(7)的概括，该向量通过倾向得分 p_{ij} 而不是通过一组风险组指标进行线性调整。注意到当 pij 包括在控制向量 X_i 中，并且用未经调整的 offer Z_{ij} 取代重新中心化的工具变量z_i时，估计值没有变化，这就加强了这种联系。这一观察来自于 Frisch-Waugh-Lovell 定理：将 Z_{ij} 回归到 p_{ij} 和 X_{ij} 的残差是$z_i = Z_{ij} - p_{ij}$，因为 $p_{ij} = E[Z_{ij} \mid \theta_i]$ 以 1 的系数预测 Z_{ij}，并且 X_i 中所有其他前定变量在控制 p_{ij} 时与 Z_{ij} 无关。因此我们通过将 Z_{ij} 重新中心化在 p_{ij} 上，或通过控制 p_{ij} 的线性函数，或 2SLS 估计中 p_{ij} 的任何更灵活的函数（有或没有其他的前置协变量 X_i），可以得到与等式(21)中所示相同的 IV 估计值[24]。

与等式（8）一样，对 p_{ij} 的线性调整有可能通过有效地汇总集中分配中的所有条件性随机变化而产生对学校 j 的有效性的精确估计。权重 $w_{IV}(t)$舍弃了那些无论是否随机同分决胜都总是或从不分配给 j 的类型，而对分配和不分配的可能性相同的类型给予更多的权重。通过增加预测剩余结果变化的前定控制，可以进一步提高精

[24] Frisch-Waugh-Lovell 定理的另一个应用表明，对风险组控制模型（6）-（7）的 2SLS 估计等同于控制经验倾向得分，该得分被计算为每个风险组的平均 offer 率。当非参数估计每个风险组的分数不可行时，对分配机制产生的理论分数的控制使估计可行。

确度。因此，控制 p_{ij} ，而不是重新中心化 Z_{ij} ，在实践中可以导致更小的标准误差。Abdulkadiroğlu 等人（2017）更进一步，控制倾向得分的每个值的指标，以及学生的人口统计学和滞后的成绩测量。

集中的学校 offer 和倾向性得分可以作为实证研究学校部门（如特许学校）的有效性和分配到具有不同特征的学校（如具有高地区评级或具有某些同行特征的学校）的效果的基础。形式上，用 C_j 表示学校 j 的特征，让 $z(i)$ 和 $d(i)$ 表示学生 i 的分配和入学学校的指标。考虑工具变量 $C_{z(i)} = \sum_j C_j Z_{ij}$ 和处理变量 $C_{d(i)} = \sum_j C_j D_{ij}$ ，分别测量分配学校和入学学校的特征。例如，C_j 表示特许学校，$C_{z(i)}$ 是被分配到特许学校的一个指标，而 $C_{d(i)}$ 表示特许学校的入学。Borusyak 和 Hull(2020)的特征延伸到 IV 估计，用 $C_{z(i)}$ 作为 $C_{d(i)}$ 的工具变量，同时控制 $\sum_j C_j p_{ij}$ ，这样形成了按分配倾向得分加权的学校特征的平均值[25]。对于检查特许学校效能，这意味着控制分配到特许学校的总风险。我们在下一节再来讨论其他学校特征的 IVs。

表 6 通过对丹佛公立学校的特许效应的分析，说明了集中分配倾向得分所发挥的作用（这里使用的样本来自 Abdulkadiroğlu 等（2017））。第（1）列显示了一个精确的 0.42σ 的上特许学校对数学考试成绩的影响，通过 2SLS 估计，特许学校 offer 工具变量灵活地控制模拟的特许倾向得分和其他基线人口统计[26]。其余各列显示了类似但不太精确的估计值，这些估计值来自于较粗糙的工具变量：第（2）列中的第一选择特许分配指标和第（3）列中的任何特许分配资格指标，两者都控制了适当的基于偏好的风险组。这些替代策略摒弃了机制产生的特许分配中的一些随机变化，因此产生了不太精确的估计，反映在第二阶段的标准误差中。表 6 倒数第二行显示，第一选择和资格方法的样本量需要增加 1.6 和 3.5 倍，才能与风险控制的集中分配 IV 策略的精度相匹配。

[25] 如前所述，该规范确定的系数与 IV 程序相同，该程序用重新中心化的 $C_j (Z_{ij} - p_{ij})$。前置变量可以包括在任一回归中以提高精确度。

[26] 样本包括 2011-2012 和 2012-2013 学年的 4-10 年级的申请人。详情见 Abdulkadiroğlu 等人（2017）的表 6 和表 9。

5 单个学校的 VAM

对于许多高风险的决策来说，了解一个广泛的学校部门（如特许学校与传统公立学校）的有效性是不够的。家长们希望知道哪些学校能够为他们的孩子提供最多的学习机会。政策制定者在决定是否关闭、重组或扩大他们地区的学校时，同样也会依赖个别学校效能的衡量（Rockoff 和 Turner，2010；Abdulkadiroğlu 等，2016；Cohodes 等，2021）。这种对学校效能数据的需求反映在最近公开可用的衡量的激增上。

例如，2015 年的《每个学生成功法案》授权美国所有州采用小学和中学的问责制度，包括对学生平均成绩和成长的公开衡量。同时，《美国新闻与世界报道》和 GreatSchools.org 等私营公司制作了大量受欢迎的学校评级，这些评级通常在 Zillow 和 Redfin 等房地产网站上占据显著位置。这种评级似乎影响了家庭对居住地的选择，以及对子女入学的选择（Bergman 和 Hill，2018；Hasan 和 Kumar，2019）。

几乎所有的公立和商业学校的表现衡量标准都来自于观察比较：通常是一个学校注册学生的平均考试分数水平或增长，有时会根据观察到的学生人口结构的差异进行调整。这种水平和增长的衡量标准非常类似于观察性的增值模型（VAMs），长期以来，人们一直在考虑和争论教师和学校排名的问题（如 Kane 和 Staiger，2008；Rothstein，2010；Chetty 等，2014a；Deming，2014）[27]。作为这种 VAMs 基础的选择可观察因素假设，反映了一种不同于上述抽签和基于不连续的识别策略的消除选择偏差的方法。然而，从 Angrist 等人（2016b，2017）开始的最新文献显示，这种准实验性的变化可以被纳入学校的 VAM 议程，产生更可靠的单个学校效能的估计，同时努力解决获得这种精细的因果估计的某些基本问题。

在本节中，我们首先概述了观察性学校 VAMs 的基本逻辑和通常应用于增值估计的经验贝叶斯方法。接下来，我们讨论如何用学校抽签来检验观察性 VAMs 的关键识别假设。然后，我们描述了当发现识别假设被违反时，可以通过部分纠正选择偏差来进一步利用这种变化来改进观察模型。

[27] 对学校效能的观察研究至少可以追溯到科尔曼（1966 年）的报告，该报告在横断面回归中著名地表明，相对于家庭背景的贡献，学生成绩的方差中归因于教育投入的部分很小。可以说，随后的长期观察和准实验文献描绘了一幅更加细致的画面。

5.1 估计观察到的 VAMs

传统的学校增值估计的出发点是一个恒定效应模型，其思路与前面考虑的一样，现在扩展到允许每个学校有一个独特的因果效应。考虑一个有多所学校的环境，以 $j = 1, \ldots J$，每所学校都有自己的因果"增值"（value-added）参数 β_j，衡量其相对于被遗漏的学校对成绩结果 Y_i 的影响:

$$Y_i = \mu + \sum_{j=1}^{J} \beta_j D_{ij} + \varepsilon_i, \tag{22}$$

如前所述，$D_{ij} \in \{0, 1\}$ 指的是学生 i 在学校 j 的入学，ε_i 捕捉其他决定成绩的因素[28]。

为了估计增值参数 β_j，政策制定者和实践者可以使用回归控制和结果差异化策略的组合。将未观察到的 ε_i 回归到观察到的协变量 X_i，其中也可能包括学生的人口统计学和滞后的考试分数，产生一个增强的因果模型:

$$Y_i = \sum_{j=1}^{J} \beta_j D_{ij} + X_i' \mu + \eta_i, \tag{23}$$

其中（也和以前一样），在过渡到带有协变量的模型时，我们将符号 μ 重新用于连接到 X_i 的系数向量，它被定义为包括一个常数。VAM 估计所依据的关键的可观察因素选择假设是：对于每一个 j，$E[D_{ij}\eta_i] = 0$。换句话说，学生能力中未被 X_i 解释的部分必须与学校入学不相关。在可观察因素的选择下，OLS 回归可以恢复等式(23)的参数。

这个框架嵌套了几种类型的学校质量测量方法。最简单的水平测量法有效地将 X_i 只等于一个常数，从而将学校质量作为入学学生的平均成绩水平来衡量。如果学校招收的学生具有系统性的未经调整的能力 ε_i，那么水平模型的估计值就会有偏差，因为学校不是随机分配的。更为复杂的 VAMs 通过在 X_i 中加入这些控制因素来说明人口统计学和滞后考试成绩的可观察差异。在这种情况下，可观察因素的选择假设要求在具有相同特征和过去成绩的学生中没有系统地选择学校入学。

调整滞后成绩的另一种方法是收益模型，即首先对同期和过去的考试成绩进行差分，以消除时间上的不可控因素。如果我们让Δ

[28] 实际上，等式（22）来自一个可加分离的潜在结果模型，$Y_i(j) = \mu + \beta_j + \varepsilon_i$，这意味着 因果效应 $Y_i(j) - Y_i(k) = \beta_j - \beta_k$ 在不同的学生中是不变的。

$Y_i = Y_{ig} - Y_{i(g-1)}$ 表示学生成绩相对于较早年级的变化，那么收益模型就可以得到：

$$\Delta Y_i = \mu + \sum_{j=1}^{J} \beta_j D_{ij} + \Delta_{\varepsilon_i}, \tag{24}$$

其中，$\Delta\varepsilon_i = \varepsilon_i - Y_{i(g-1)}$。这里的相关识别假设要求潜在的结果趋势在各学校之间是平行的，即 $E[\Delta\varepsilon_i \mid D_{ij}= 1] = E[\Delta\varepsilon_i \mid D_{ik} = 1]$ 对于所有 $j \neq k$，因此，ΔY_i 对 D_{ij} 的线性回归可以得到参数 β_j。等式(23)和等式(24)中的回归调整和结果差分可以通过在收益回归中增加额外的人口统计控制来进一步结合。

5.2 经验贝叶斯方法

正常情况下的 EB 收缩率

对(23)和(24)这样的 VAM 模型进行 OLS 估计，可以得到一组针对学校的增值估计值 β_j[29]。一个关键的问题是，我们将在下一小节中讨论，这些观察性 VAM 程序所依据的识别性假设是否成立，从而使 β_j 对等式(22)中的因果参数 β_j，给出一个无偏的估计。暂且不说这个，另一个也许同样重要的问题是，即使这些识别假设成立，β_j 的估计是否精确到对决策有用。与第 3 节中讨论的部门范围内的效果估计相比，单一学校的 VAM 估计可能涉及大量的抽样误差，特别是对于小型或新的学校，学生的观察值相对较少。

经验贝叶斯（EB）分析提供了一种策略来调节个体估计值 β_j 的抽样方差。EB 方法将学校 VAM 参数 β_j 视为学校质量分布的抽样，通常假设为正态分布。从估计的 β_j's 中得出代表该分布的超参数的 EB 估计。这个估计的分布产生了对各个学校质量的后验预测。EB 方法使用全套的 VAM 估计值来减少单个学校质量估计值的抽样方差，接受后验估计值的偏差作为交换（Morris，1983；Raudenbush 和 Bryk，1986；Efron，2012）。

为了说明问题，假设 β_j 是 β_j 的无偏和正态分布的估计值，已知方差等于其平方标准误差 s_j^2。这种正态性假设可以被看作是每所学校学生人数增加时的渐进式近似。接下来，假设潜在参数 β_j 是从学校群体中定义的分布 G_β 中随机抽取的。暂时假设 G_β 是一个正态分

[29] 在一些应用中，研究人员通过首先将 Y_i 回归到 X_i，然后计算所得残差的学校平均数来估计增值。如果学校的入学率独立于控制变量 X_i，这种方法在大样本中产生的估计值与 OLS 估计值（23）相同，但如果入学率与控制因素相关，一般会产生渐进的不同估计值。由于纳入 X_i 通常是出于对选择偏差的考虑，因此在可能的情况下，使用不强加这种独立性假设的 OLS 估计似乎更可取。

布，并且与各学校的抽样方差 s_j^2 无关。这得出以下层次模型：

$$\beta_j \mid \beta_j, s_j^2 \sim N(\beta_j, s_j^2), \tag{25}$$

$$\beta_j \mid s_j^2 \sim N(\mu_\beta, \sigma_\beta^2). \tag{26}$$

这个模型有两个超参数，μ_β 和 σ_β^2[30]。这些超参数的矩估计方法由以下等式给出：

$$\mu_\beta = \frac{1}{J}\sum_{j=1}^{J}\beta_j, \tag{27}$$

$$\sigma_\beta^2 = \frac{1}{J}\sum_{j=1}^{J}[(\beta_j - \mu_\beta)^2 - s_j^2]. \tag{28}$$

这里的方差估计式在等式(28)中减去了 s_j^2，作为偏差校正：β_j's 的未校正样本方差被抽样方差膨胀了[31]。最大似然法应用于学校层面的模型(25)和(26)，或者从模型(23)中残差的分布假设出发，对个人结果进行全面的参数化说明，为简单的矩量法提供了更多的选择。

EB 估计的最后一步是构建每个学校质量的后验值。给定的模型是(25)和(26)，β_j 的后验分布由 $\beta_j \mid \beta_j, s_j^2 \sim N(\beta_j^*, V_j^*)$ 给出，其中

$$\beta_j^* = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\right)\beta_j + \left(\frac{s_j^2}{\sigma_\beta^2 + s_j^2}\right)\mu_\beta, \tag{29}$$

并且 $V_j^* = \frac{s_j^2\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}$。等式(29)表明，后验平均数 β^* 是无偏估计 β_j 和先验平均数 μ_β 的加权平均。当 β_j 的抽样方差 s_j^2 接近 0 时，其权重也接近 1。通过将嘈杂的无偏估计值按照其抽样误差的比例向先验均值缩减，后验均值减少了方差，对于估计值比较嘈杂的学校，缩减的幅度更大。经验贝叶斯后验均值 β_j^* 将估计的超参数 μ_β 和 σ_β^2 加入(29)。

何时缩减？

缩小的 EB 后验平均数是否应该优于噪声较大但无偏的估计值 β_j，取决于分析者的目标。要看到这一点，请注意，以学校 j 的真实增值为条件，两个估计值的均方误差（MSE）为[32]：

---

[30] 鉴于(22)中的模型，平均数 μ_β 捕获相对于一个被遗漏的类别的平均学校质量。

[31] Kline 等人（2020）概述了对方差分量进行偏差校正估计的一般方法。

[32] 后验均值的 MSE 公式忽略了超参数的估计误差。参见 Morris（1983）和 Armstrong 等人（2020）对纳入先验分布中的估计误差方法的讨论。

$$E[(\beta_j-\beta_j)^2\mid\beta_j,s_j^2]=s_j^2,$$
$$E[(\beta_j^*-\beta_j)^2\mid\beta_j,s_j^2]=(\frac{\sigma_\beta^2}{\sigma_\beta^2+s_j^2})^2 s_j^2+(\frac{s_j^2}{\sigma_\beta^2+s_j^2})^2(\beta_j-\mu_\beta)^2 \tag{30}$$

如果我们只对一个特定的学校感兴趣，这个条件 MSE 公式显示，并不清楚两个估计值中哪个更好；缩减减少了方差，但如果学校与平均水平有很大差别，可能会导致大量偏差（如等式(30)中第二项所反映的）。另一方面，如果我们对评估许多个学校感兴趣，MSE 的相关概念是对 β_j 的分布进行整合：

$$E[(\beta_j-\beta_j)^2\mid s_j^2]=s_j^2,$$
$$E[(\beta_j^*-\beta_j)^2\mid s_j^2]=(\frac{\sigma_\beta^2}{\sigma_\beta^2+s_j^2})^2 s_j^2. \tag{31}$$

这个等式表明在无条件的情况下，后验均值的 MSE 明显低于无偏估计 β_j[33]。事实上，根据条件平均函数的标准属性，后验均值在该模型下的所有函数（β_j, s_j^2）中具有最小的 MSE。因此，当我们想要一个在所有学校中平均表现良好的估计值时，EB 方法是有用的。

一旦计算出 EB 后验分布，它们就可以用于几个目的[34]。首先，如等式(31)所示，EB 缩减可以得到一组具有较小的平均 MSE 的学校估计值。第二，缩减可以纠正将学校增值作为回归因子的模型中的测量误差。将无偏但有噪声的估计值 β_j 放在回归的右边，会由于经典的测量误差而导致向零的衰减偏差；后验平均值会引入非经典的测量误差来纠正这一点，所以用右边的 β_j^* 回归得到的系数与用真实的 β_j 相同。这种修正与传统的变量误差回归密切相关，后者同样使用估计的信噪比(signal-to-noise ratio)修正测量误差，但通常对所有观测值使用一个共同的收缩因子（Draper 和 Smith，1998）[35]。

第三，EB 后验分布可用于对个别学校进行决策。为此，使用后验的正确特征将取决于决策者的损失函数，一般来说，使用后验

---

[33] 这一观察与经典的 James 和 Stein（1961）的结果密切相关，当在平方损失下估计三个或更多的参数时，OLS 是不允许的，并被基于收缩的估计式所支配。

[34] 研究者有时会报告后验平均估计值的方差，作为增值分散性的总结。这样的程序可能会产生误导性的结果，因为相对于潜在参数的基本分布而言，后验均值的分布在结构上是分散的（形式上，$\mathrm{Var}(\beta_j^*)<\sigma^2<\mathrm{Var}(\beta_j)$）。为了理解增加值的方差，超参数估计 σ_β^2 更有用。

[35] 注意由于左边的经典测量误差不会导致偏差，所以把 β_j^* 而不是 β_j 放在回归的左边会导致偏差而不是纠正偏差。

平均值以外的函数可能是最佳的。例如，一个地区政策制定者可能对关闭所有低于质量阈值$\bar{\beta}$的学校感兴趣，并认为每个学校被错误关闭的成本相同，在这种情况下，根据β_j小于$\bar{\beta}$的后验概率进行决策是最理想的。（由参数化正态/正态模型中的$\Phi(\frac{\bar{\beta}-\beta_j^*}{\sqrt{V_j^*}})$给出）。最近的研究是由 Gu 和 Koenker（2020）讨论了用于排名和选尾决策（tail selection decisions）的 EB 方法。

EB 扩展

通过增加预测学校质量的协变量和允许更灵活的先验分布形式，这里描述的简单 EB 框架可以得到有益的扩展。$\beta_j | s_j^2, C_j \sim N(C_j'\mu, \sigma_\beta^2)$给出学校特征$C_j$的列表，如学校部门，在某些情况下很有用。由此得出的 EB 后验均值将β_j向估计的线性指数$C_j'\mu$缩小（可从β_j对 C_j 的回归中估计）而不是一个常数收缩。直接允许 σ_β^2 取决于学校的特征是很简单的。

本着同样的精神，等式(26)的正态性可能被G_β的一个更一般的模型所取代。无论如何，正态性下得出的线性收缩估计值具有理想的特性。特别是，β_j^*与β_j对β_j的线性回归的拟合值相吻合，所以它可以被解释为对增值的最佳线性预测，而不管混合分布如何。尽管如此，估计一个更灵活的 G_β，以获得学校质量分布形状的更完整的情况，并形成改进的后验，可能是有意义的。为实现此目的的方法包括 Kiefer 和 Wolfowitz（1956）非参数最大似然估计式（NPMLE；也见 Robbins（1956）和 Koenker 和 Mizera（2014））和 Efron（2016）提出的指数族反卷积估计式。Gilraine 等人（2020）将 NPMLE 方法应用于教师增值分布的估计[36]。

在其他一些情况下，允许β_j和s_j^2相关是有用的。例如，可能是较新或较小的学校效果较差，在这种情况下，学生较多的学校（并且因此s_j^2较小）会倾向于有更高的β_j。这里的一个简单策略是将s_j^2作为 G_β的条件分布中的一个协变量；NPMLE 估计式也可以用来估计β_j和s_j^2的无限双变量分布。另一种方法是在估计先验分布之前，对β_j's 进行方差稳定变换，使其产生近似恒定的抽样方差（例如，

[36] 见 Kline 和 Walters(2021)和 Kline 等人(2021)关于非参数 EB 方法在教育之外的其他最新应用。

见 Brown（2008））。由于到目前为止使用 EB 方法的学校质量实证文献通常依赖于正态性和独立性假设，放宽这些假设的实际意义尚不清楚。

5.3 用抽签测试 VAM 的有效性

当然，一个精确但有很大偏差的学校增值后验值的估计，很可能和一个极其嘈杂的估计一样有问题。我们接下来考虑使用准实验性的学校分配变化来检验观察性 VAMs 的选择偏差。观察性 VAM 的关键假设是观察变量的选择：在一组包含控制变量 X_i 的条件下，学校的入学率与随机分配的一样好。这一假设可以借助于以下形式的回归来检验：

$$Y_i = \sum_j \alpha_j D_{ij} + X_i'\mu + \nu_i \tag{32}$$

其中 X_i 是因果模型(23)中的控制向量。这个回归模型与(23)的不同之处在于，它被定义为回归，因此可能有不同于因果模型的参数。因此，这个模型中的误差项 ν_i，与潜在结果的随机部分 η_i 不同。

在可观察因素的选择下，因果模型(23)和回归(32)的参数是一致的，因此，对于所有学校 j 来说，$\alpha_j = \beta_j$。此外，回归的残差与残差能力是一致的：$\nu_i = \eta_i$。反过来，学生的能力应该与学校提供的任何随机性无关，理由与假设 1A 相同。因此，对可观察因素进行选择的结果是OLS残差和（重新中心化的）分配指标Z_{if}的正交性：

$$E[(Z_{il} - p_{il})\nu_i] = 0, \tag{33}$$

对于可能的多所学校 l=1,...,L。在这里，我们使用分配的倾向性得分 p_{if}，以解决提供的任何非随机性，如第 4 节的集中分配方案，但同样的逻辑可以适用于分散的抽签，通过调整风险组固定效应，如第 2-3 节。

对等式(33)的检验是通过将回归模型(32)的残差对一组工具变量进行回归得到的，Z_{if} - p_{if}。检验拒绝是观察到的 VAM 系数的选择偏差的症状：即 $\nu_i \neq \eta_i$，这样有 $\alpha_j \neq \beta_j$，对于某些或所有学校 j。Angrist 等人（2016b）展示了这个程序如何可以被视为 L 正交限制的拉格朗日乘数（LM）检验，它规定了 VAM 有效性和条件独立的学校分配的共同原假设。

对等式(33)的综合检验可以分解为两个概念上不同的检验：一个是捕捉观察到的 VAM 系数 α_j 平均预测学校有效性 β_j 的程度，一个是捕捉任何选择偏差 $b_j \equiv \alpha_j - \beta_j$ 在学校之间的变化。形式上，基于同质化 ν_i 的假设，考虑构建一个测试统计量：

$$T = \frac{(Y-Y)' P_Z (Y-Y)}{\sigma_v^2}$$

其中，Y是成绩结果的N×1向量，Y是估计等式(32)的回归拟合值的N×1向量，$\sigma_v^2 = (Y-Y)'(Y-Y)/N$估计了$\nu_i$的方差，$P_Z$是重新中心化的提供$Z_{if}$ - p_{if}的投影矩阵[37]。这是对VAM残差$\hat{\nu}_i = Y - Y$对Z_{if} - p_{if}回归显著性的联合检验：

$$\nu_i = \psi_0 + \sum \psi_l (Z_{il} - p_{il}) + u_i,$$

其中，$\psi_1, \ldots, \psi_L = 0$的空值被用来计算残差。Angrist等人（2016b）表明这个测试统计量可以改写为两项之和：

$$T = \frac{(\varphi - 1)^2}{\sigma_v^2 (Y' P_Z Y)^{-1}} + \frac{(Y - \varphi Y)' P_Z (Y - \varphi Y)}{\sigma_v^2} \tag{34}$$

其中$\varphi = (Y' P_Z Y)^{-1} Y' P_Z Y$是2SLS系数估计值，它在$Y_i$的等式中使用所有$Z_{if}$ - p_{if}作为Y_i'的工具变量。由于这些重新中心化的offer在设计上与学生人口统计学或滞后考试分数不相关，φ可以被视为估计第二阶段的等式

$$Y_i = \tau_0 + \varphi \alpha_{d(i)} + X_i' \tau + \eta_i, \tag{35}$$

其中$\alpha_d(i) = \sum_j \alpha_j D_{ij}$表示$i$入学学校的观察性VAM系数。

第二阶段的参数φ有时被称为"预测系数"，当等式(35)的版本用准实验变量拟合时，被用来评估观察性质量指标α_j的平均预测有效性——无论是对教师、学校，还是最近对教育之外的事物（例如Chetty等，2014a；Daming，2014；Chetty和Hendren，2018；Abaluck等，2021）。典型的原假设是φ= 1，这意味着$\alpha_d(i)$的一个单位的增加会转化为Y_i的一个单位的增加。与这种预测关系的偏差表明"预测偏差"。因此，Angrist等人（2016b）的测试统计量分解(34)的第一项是没有预测偏差的空值的Wald统计量。第二项是Sargan（1958）统计量，用于对等式（35）的2SLS估计中的过度识别限制进行LM检验。直观地说，这一项检查在每个抽签准实验中，VAM系数是否具有同等的预测性。VAM有效性T的综合检验检查了L个限制条件(33)，将预测偏差的单一检验与L-1个额外的限制条件结合起来，这些限制条件来自于各种特定学校的招生抽签。

表7使用Angrist等人（2017）分析的学生和抽签样本说明了波士顿中学的这些测试。第一行报告了来自2SLS模型的预测系数，

[37] 实际上，$P_Z = Z(Z'Z)^{-1}Z'$，其中Z是一个N×L矩阵，堆叠重新中心化分配的观测值Z_{il} - p_{il}。

该模型将 VAM 预测值$\alpha_d(i)$与特许学校的抽签 offer 和第一选择的集中分配相联系，控制必要的风险组固定效应和额外的基线协变量。第(1)列显示了未受控制的 VAM 的结果，它只是比较了调整了年份效应的各学校六年级的平均数学成绩。第（2）列的滞后分数模型进一步调整了学生的人口统计学和滞后（五年级）的成绩，而第（3）列的收益模型用成绩和滞后成绩的差异来代替结果，保留人口统计学的协变量。

检验结果显示，调整了滞后成绩的 VAM 模型的偏差要比未受控制的情况小得多。未控制模型的 2SLS 预测系数估计值只有 0.40，表明在比较各学校未经调整的成绩水平时存在很大的预测偏差。相比之下，调整了滞后成绩产生的预测系数为 0.86，与 1 只有微小的统计学差异（p=0.07）。收益模型的相应估计值等于 0.95，在常规水平上不能拒绝收益模型中没有预测偏差的原假设（p = 0.55）。这种最小的预测偏差反映了关于教师和学校增值的文献中的一个共同发现：调整过去成绩的方法消除了教室和学校之间比较的大部分选择偏差。因此，平均而言，来自控制良好的 VAM 的估计值提供了对学生考试成绩的因果效应的合理可靠的估计（Chetty 等，2014a；Angrist 等，2017）。

尽管这种预测偏差很小，但表 7 的最后一行显示，对所有限制的综合检验对这三个模型都有决定性的拒绝。拒绝的原因主要是 2SLS 过度识别限制的失败：即使收益模型平均来说很好地预测了学生的结果，但对于一些 L 学校特定的抽签来说，这种预测的有效性很差。第（2）和（3）列中更复杂的 VAMs 的测试结果在图 3 中描述，它给出了 2SLS 估计的 "视觉 IV "表示。具体来说，对于 Angrist 等人（2017）的每一个学校抽签，我们将分配对六年级数学分数的简约式效应 Y_i 与分配对估计的观察增值$\alpha_d(i)$的第一阶段效应做对比 。通过这些点的加权最佳拟合线的斜率与表中的预测系数估计值φ相对应，我们绘制了基准的 45 度线作为参考。正如所看到的，收益模型使这两条线更加接近，反映出预测偏差最小。但有几个抽签点离这两条线很远，说明各学校的选择偏差不为零（在 10%的水平上，有颜色的点在统计上明显远离这条线）。

5.4 抽签的偏差纠正

结合 OLS 和 IV 的估计

鉴于表 7 中的测试拒绝情况，一个自然的问题是，准实验性录取变化是否可以用来减少观察性 VAM 中明显的选择偏差。在基于

抽签的学校质量估计和观察性 VAM 估计都可用的情况下，在使用有偏的（但精确的）观察性估计和无偏的（但有噪声的）抽签估计之间存在着权衡。接下来我们通过扩展第 5.2 节的经验贝叶斯框架，描述这种权衡的潜在解决方案。

假设我们有一组来自等式（23）的可能有偏的观察性 OLS VAM 估计值，满足：

$$\alpha_j \mid \beta_j, b_j, s_{j,\alpha}^2 \sim N(\beta_j + b_j, s_{j,\alpha}^2). \tag{36}$$

如前所述，参数 β_j 给出了学校 j 的真实质量，$s_{j,\alpha}^2$ 是 OLS 估计 α_j 的平方标准误差，而正态性假设是每个学校有很多学生时的渐近近似值。然而现在，估计值可能有偏差——由学校特定的参数 b_j 表示。拒绝第 5.3 节中基于抽签的综合检验表明，对于一些或所有学校来说，$b_j \neq 0$。

假设除了 α_j，我们还有每个学校的准实验 VAM 估计值 β_j。例如，这些估计值可能来自于等式(23)中的 D_{ij} 指标与风险调整后的 offer 工具变量 $Z_{ij} - p_{ij}$，如第四节所述。抽签估计值被假定为真实 VAM 参数的一致和渐近正态的估计值：

$$\beta_j \mid b_j, s_{j,\beta}^2 \sim N(\beta_j, s_{j,\beta}^2) \tag{37}$$

我们通常期望 $s_{j,\beta}^2$ 大于 $s_{j,\alpha}^2$ 因为 IV 比 OLS 更不精确[38]。最后，我们扩展了层次模型(26)，允许学校质量的双变量分布和学校间的选择偏差。让 $\Theta_j \equiv (\beta_j + b_j, \beta_j)'$ 表示学校 j 的观察性 VAM 和因果参数的 2×1 向量：

$$\Theta_j \mid S_j \sim N(\mu_\Theta, \Sigma_\Theta). \tag{38}$$

矩阵 Σ_Θ 描述了因果有效性和选择偏差在各学校之间的联合分布。矩阵 S_j 对角线上有抽样方差 $s_{j,\alpha}^2$ 和 $s_{j,\beta}^2$，在对角线外有 OLS 和 IV 估计的抽样协方差。当 $b_j = 0$ 和 OLS 残差 ν_i 是同方差时，这个协方差等于 $s_{j,\alpha}^2$（Hausman，1978）。

在(36)、(37)和(38)描述的模型下，Θ_j 的后验均值在观察到的估计值 $\Theta_j = (\alpha_j, \beta_j)'$ 的基础上给出，即：

$$\Theta^* \equiv E[\Theta_j \mid \Theta_j, S_j] = (\Sigma_\Theta^{-1} + S_j^{-1})^{-1} S_j^{-1} \Theta_j + (\Sigma_\Theta^{-1} + S_j^{-1})^{-1} \Sigma_\Theta^{-1} \mu\Theta. \tag{39}$$

Θ^*的第二个元素是β_j的后验均值，同时使用学校质量的 OLS 和 IV 估计，这是两个估计值和先验均值的线性组合[39]。

[38] 请注意，在高斯-马尔科夫模型的经典假设下，这将得到保证。

[39] 见 Angrist 等人（2017）对特殊情况下两个估计值的权重的表达。Chetty 和 Hendren（2018）使用类似的方法来计算邻里效应的 EB

按照第 5.2 节的 EB 方法，等式(39)中的 "混合"增值后验均值可以通过估计超参数 μ_Θ 和 Σ_Θ 来近似，其依据是观察到的 OLS 和 IV 估计的联合分布及其抽样方差和协方差，然后将这些超参数估计插入等式(39)。通过上述 Hausman（1978）对 S_j 的逻辑，可以证明，当观察到的 VAM 接近无偏（$Var(b_j) \approx 0$），并且误差是同方差时，不需要对 IV 估计进行加权。我们回到传统的 EB 缩减公式(29)，应用于 OLS α_j 的估计。在另一个极端，当选择偏差严重到使观察的 VAM 估计值无用时（即 $Var(\mathrm{b_j}) \to \infty$），不再对 α_j 加权，我们对准实验估计值 β_j 采用常规 EB 收缩公式。在中间情况下，混合后验在两套增值估计之间进行了最佳的偏差和方差折衷。

招生不足的偏差校正：IV VAM

在实践中，基于抽签的估计 β_j 可能不是每所学校都有。例如，在集中分配系统中，通常有一些学校在分配上缺乏准实验性的变化。这是一个招生不足的问题，形式化为 $Z_{ij} - p_{ij} = 0$，对于某个学校 j 的所有学生 i。被分配到这种学校的申请人（$Z_{ij} = 1$）从未面临任何不被分配的风险（$p_{ij} = 1$），也许是因为学校面临的需求少，所以不需要基于抽签的配给，而所有其他学生从未被分配（$Z_{ij} = p_{ij} = 0$）。由于所有学生的 $Z_{ij} - p_{ij} = 0$，这所学校的分配不能作为学校招生的工具变量。因此，鉴于申请不足，我们的工具变量比等式(23)中的内生变量少，不能使用 2SLS 来估计 β_j's 或应用混合后验公式(39)。

Angrist 等人（2021 年）提出的工具变量增值模型（IV VAM）方法，提供了一个解决招生不足问题的办法[40]。 IV VAM 避开了对 $\beta_j's$ 的识别不足问题，用一个模型将增值与一组较低维度的学校特征联系起来，通过 IV 进行估计。这种方法从一个假设的学校层面的增值 β_j 对 K×1 的学校特征向量 M_j 的预测开始：

$$\beta_j = M_j'\varphi + \nu_j. \tag{40}$$

等式(40)是一个学校之间的模型，通常用于学校效应的分层线性模型（Raudenbush 和 Bryk，1986）。特征 M_j 可能包括 OLS 增值系数 α_j，以及其他学校属性，如部门或人口统计学。系数向量 φ 捕捉到这些预测因素和因果学校质量之间的系统关系。残差 ν_j，定义为 $E[M_j\nu_j]=0$，反映了没有被 M_j 解释的学校质量变化。超参数 $\sigma_\nu^2 \equiv Var(\nu_j)$ 总结了这种残差变化的程度。

将学校层面的预测(40)插入学生层面的因果模型(22)，可以得

[40] IV VAM 方法简化并概括了 Angrist 等人（2017）所探讨的带有申请不足的混合估计的参数化方法。

到：

$$Y_i = \tau_0 + M'_{d(i)}\varphi + \varepsilon_i + \nu_{d(i)}, \tag{41}$$

其中，$M_{d(i)} = \sum_j M_j D_{ij}$ 是学生 i 的入学学校的特征向量。IV VAM 的第一步是通过 2SLS 来估计等式(41)，将 L 个超额申请学校的风险调整后的 offer 向量作为 $M_{d(i)}$ 的工具变量，$Z_i = (Z_{i1} - p_{i1}, ..., Z_{iL} - p_{iL})$。

这个 2SLS 程序概括了第 5.3 节中介绍的测试方法。当 $M_{d(i)}$ 仅由 OLS 学校增值估计值 $\alpha_{d(i)}$，2SLS 估计值 φ 检查基础 OLS 模型的预测偏差，而伴随的过度识别检验评估了不同抽签的 OLS 预测值的变化。在更普遍的情况下，2SLS 预测系数描述了学校特征 M_j 和增值之间的关系，而过度识别检验统计量可用于构建 σ_v^2 的估计值（如果 M_j 包括无偏的 OLS 系数，该估计值应该为零）。IV VAM 估计程序的机制及其与 VAM 情况测试的联系在 Angrist 等人（2021）中有进一步的详述[41]。

IV VAM 的第二步是在所有可用的信息下，包括学校特征 M_j 和可用的抽签准实验的估计值，构建单个学校质量的最小均方误差预测。让 ρ 表示 Y_i 对风险调整后的 offer 向量 $\tilde{Z}_i$ 进行回归的 $L \times 1$ 系数向量，v_ρ 表示 ρ 的抽样协方差矩阵。J×1 的 IV VAM 后验向量为：

$$\beta^* = \Pi'(\Pi\Pi' + V_\rho / \sigma_v^2)^{-1}\rho + [I_J - \Pi'(\Pi\Pi' + V_\rho / \sigma_v^2)^{-1}\Pi]M\varphi. \tag{42}$$

这里 M 是一个 $J \times K$ 的矩阵，收集所有学校的特征 M_j，IJ 是 $J \times J$ 的识别矩阵，Π 是一个 $L \times J$ 的矩阵，由 J 个学校入学指标 D_{ij} 对 $\tilde{Z}_i$ 的回归得出的第一阶段系数。在一个特殊情况下，如果没有招生不足（L = J），M_j 等于 OLS 增值系数，这个等式就会坍缩为双变量收缩等式(39)[42]。在招生不足的情况下，IV VAM 后验结合了每所学校的增值预测 $M_j\varphi$ 和简约式的 offer 效应 ρ，通过第一阶段的矩阵 Π 说明 offer 的合规性。一个 EB 实施方案将 Π 的 OLS 估计值和 V_ρ，以及从 IV VAM 第一阶段得到的 φ 和 σ_v^2 的 2SLS 估计值一起加入(42)。

5.5 风险控制的增值模型（RC VAM）。

使用集中分配变化来估计学校增值的另一种策略是 Angrist 等人（2021）的风险控制增值模型（RC VAM）。这种方法的出发点是，

[41] 在 $Var(v_j) \neq 0$ 的情况下，(41)的 2SLS 估计所依据的排他性要求学校 offer 与残余的学校质量 $v_{d(i)}$ 不相关，这不是由 offer 和潜在结果的独立性保证的。正如 Angrist 等人（2021）所详述的，这使得 IV VAM 成为 Kolesár 等人（2015）的 "许多无效工具 "框架的一个特例。

[42] 关于这种等价关系的细节，见 Angrist 等人（2021）的附录。

像第 4 节中的 DA 机制这样的分配系统产生了关于学生偏好和优先事项的丰富数据，这可能说明了学校之间的大部分非随机排序。RC VAM 通过在回归(32)中的控制向量 X_i 中加入分配倾向分数 p_{ij} 的函数，使用这些数据来加强传统 VAM 估计中的可观察因素选择假设。换句话说，RC VAM 不是使用集中的分配信息来产生学校分配的工具，而是使用这个信息来构建新的控制变量，以帮助缓和观察性模型中的选择偏差[43]。

带有风险控制的 RC VAM 模型所依据的选择可观察变量假设与 Dale 和 Krueger（2002，2014）以及 Mountjoy 和 Hickman（2020）的大学质量研究中所援引的假设相呼应。这些研究控制了学生的大学申请组合和录取 offer，以估计在特定大学就读的回报，并假设录取决定是随机的。同样，RC VAM 要求在分配风险和其他可观察因素的条件下，不遵守集中分配录取 offer 的情况要像随机一样。这种联系在下面 Angrist 等人（2021）的结果中得到了正式体现：

$$\varepsilon_i \perp D_i \,|\, (p_i, X_i, Z_i) \Rightarrow \varepsilon_i \perp D_i \,|\, (p_i, X_i), \tag{43}$$

其中，$D_i \in \{1, \ldots, J\}$ 是学生 i 就读的学校，$Z_i \in \{1, \ldots, J\}$ 是 i 的学校分配，$p_i = (p_{i1}, \ldots, p_{iJ})$ 是所有学校的倾向得分向量，ε_i 是模型(22)中的学生能力项。这一结果表明，如果在具有相同的分配风险、协变量和录取通知书的学生中，学校的录取情况与能力无关。那么仅以风险和协变量为条件，录取情况也与能力无关——因为录取通知书是以风险为条件的随机性，一旦我们以分配倾向得分为条件，就没有必要控制录取情况。因此，对倾向得分的了解使我们能够利用录取通知书的条件随机性来检验 RC VAM 对可观察因素的选择假设（等式(43)的右侧），而不是直接控制 offer 变量，即使是通过 Dale 和 Krueger 式的随机不遵守录取 offer 的假设（等式(43)的左侧）来激励 RC VAM 策略。

对纽约市初中和高中的 RC VAM 识别假设的测试出现在图 4 中。结果显示，使用 Angrist 等人（2021）的样本和与图 3 相同的可视化 IV 测试程序，纽约市初中和高中质量的 RC VAM 估计值几乎是无偏的[44]。 两个面板的第一列再次显示，未控制的 VAM，有效地比较了成绩水平，严重地未能预测集中式学校分配的简约式效应。值

[43] Abdulkadiroğlu 等人（2020 年）开发了一种相关的控制函数方法，该方法控制了从适合于学生等级排序的偏好列表的随机效用离散选择模型中得出的偏好。

[44] 由于纽约市有许多录取工具变量，Angrist 等人（2021 年）将它们依据学校的观察性 VAM 估计分为 20 个组别。Angrist 等人（2021 年）也发现，在丹佛中学的样本中，RC VAM 几乎是无偏的。

得注意的是，第二列显示，通过增加分配风险控制，大部分的选择偏差被消除了；当传统的 VAM 控制（人口统计学和滞后的考试分数）被进一步增加时，预测系数与 1 难以区分，所有的点都紧密地聚集在 45 度线上。初中和高中 RC VAMs 的综合检验 p 值分别为 0.21 和 0.84。对于高中来说，似乎没有遗漏变量偏差，这一点尤其令人印象深刻，因为纽约市的结果（SAT 分数）来自于与滞后分数控制不同的测试，因此可能更容易出现遗漏变量偏差（Chetty 等人（2014b）在不同背景下提出的观点）。这些结果表明，利用集中分配系统的分配风险信息是减轻学校 VAMs 中选择偏差的一个有希望的策略。

6 结论：学校质量测量的下一步是什么？

美国学区越来越多地使用集中分配系统，为应用本章所概述的方法提供了新的机会。政策制定者和家长对学校有效性的可靠衡量标准的日益增长的需求，也同样会推动这种分析。最后，我们对这项工作可能采取的新方向进行了简要介绍。

这里回顾的大部分研究都集中在基于成绩的学校质量衡量上。近年来，人们对学校对学生成绩以外的结果的影响越来越感兴趣。这一方向的早期努力包括探索学校对非认知结果的影响，如缺勤、停学和社会情感发展；以及长期的结果，如教育程度、犯罪、政治参与、就业和收入（Deming，2011；Deming et al.，2014；Angrist 等，2016a；Dobbie 和 Fryer，2015；Abdulkadiroğlu 等，2020；Dobbie 和 Fryer，2020；Jackson 等人，2020；Beuermann 等人，2021；Cohodes 和 Feigenbaum，2021）。基于成绩的增值衡量和其他衡量之间的联系构成了未来研究的一个重要领域。更长期的结果也提出了新的计量经济学问题，因为传统的基于成绩的 VAMs 所使用的滞后控制策略对于像收入这样的东西是不可用的。因此，在研究长期效果时，基于抽签的方法可能特别重要。

第 5 节中讨论的增值模型提出了每个学校对所有学生都有一个单一的因果效应。在实践中，学校的增值可能是异质性的——例如，一所学校对那些在不同年份就读的学生，或对具有不同准备水平的学生的影响可能是不同的。而城市特许学校似乎对基线考试分数低的学生特别有利（Angrist 等，2012 年；Frandsen 和 Lefgren，2021 年；Chabrier 等，2016 年）。这类匹配效应带来了新的方法论挑战。特别是 Goldsmith-Pinkham 等（2022）最近的工作强调了在有许多干预和异质效应的情况下基于回归的方法的问题；匹配效应也使基于抽签的 VAM 有效性测试的解释复杂化，因为除了选择偏差外，这些测试对 LATEs 和其他干预效果参数之间的差异很敏感。这表明，除了线性 2SLS 模型外，更灵活的非参数模型也可以发挥作用。

第三个研究前沿是将这里讨论的学校质量测量工具与集中分配产生的偏好数据相结合，努力了解学校选择和学校效能之间的相互作用。关于家庭是否根据对学生成绩的因果影响来选择学校的证据不一：（Rothstein，2006；Abdulkadiroğlu 等，2020）认为学校的选择更多地取决于同龄人的成绩而不是学校的因果效应。然而，其他研究表明，逻辑障碍和看似复杂或不透明的选择系统阻碍了一些家庭为他们的孩子选择最有效的学校（Walters，2018；Bergman 等，

2020；Kapor 等，2020）。对这些问题的更好理解应该有双重回报，即改善学校问责措施和教育成果。

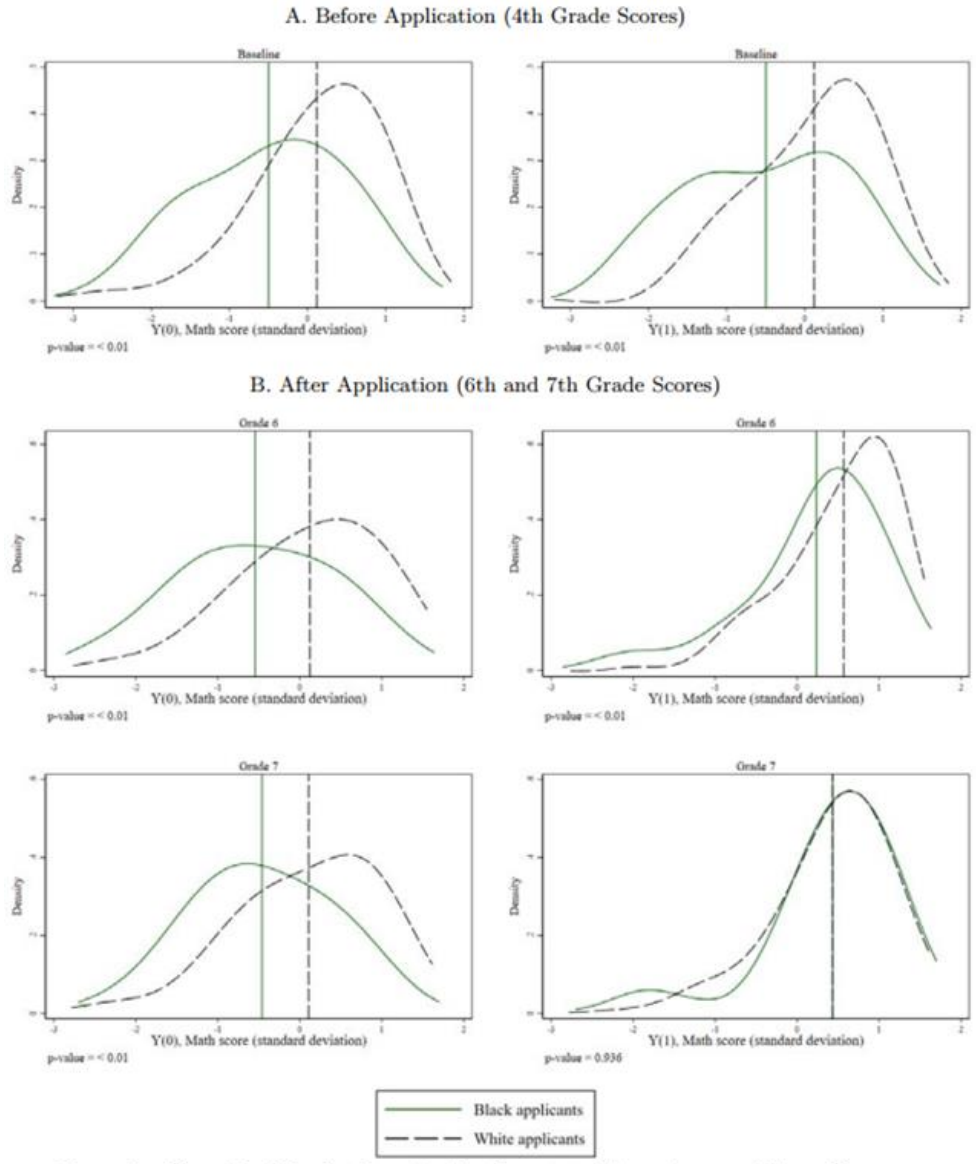


Figure 1: Complier Distributions for Applicants to Massachusetts Urban Charters

图 1：马萨诸塞州城市特许学校机构申请者的分布情况

注：本图绘制了城市特许学校申请者样本中黑人和白人抽签者的未干预（Yi (0)）和干预（Yi (1)）潜在结果的估计分布图。分布的估计如第 3.2 节所述，使用脚注 10 中描述的经验法则带宽。垂直线显示的是按种族划分的平均潜在结果。测试每个小组中白人和黑人分布平等的 p 值来自于加权 bootstrap 程序，使用估计的 complier CDF 中最大的绝对黑人-白人距离作为测试统计。每个 bootstrap 迭代中的 CDFs 是通过 2SLS 估计等式(11)来计算的，其中 $g(X_i, Y_i)=1\{Y_i \leq y\}$，对于一个网格中的点 y，用 *iid* 指数加权观察。这里使用的样本来自 Angrist 等人（2013）的研究。

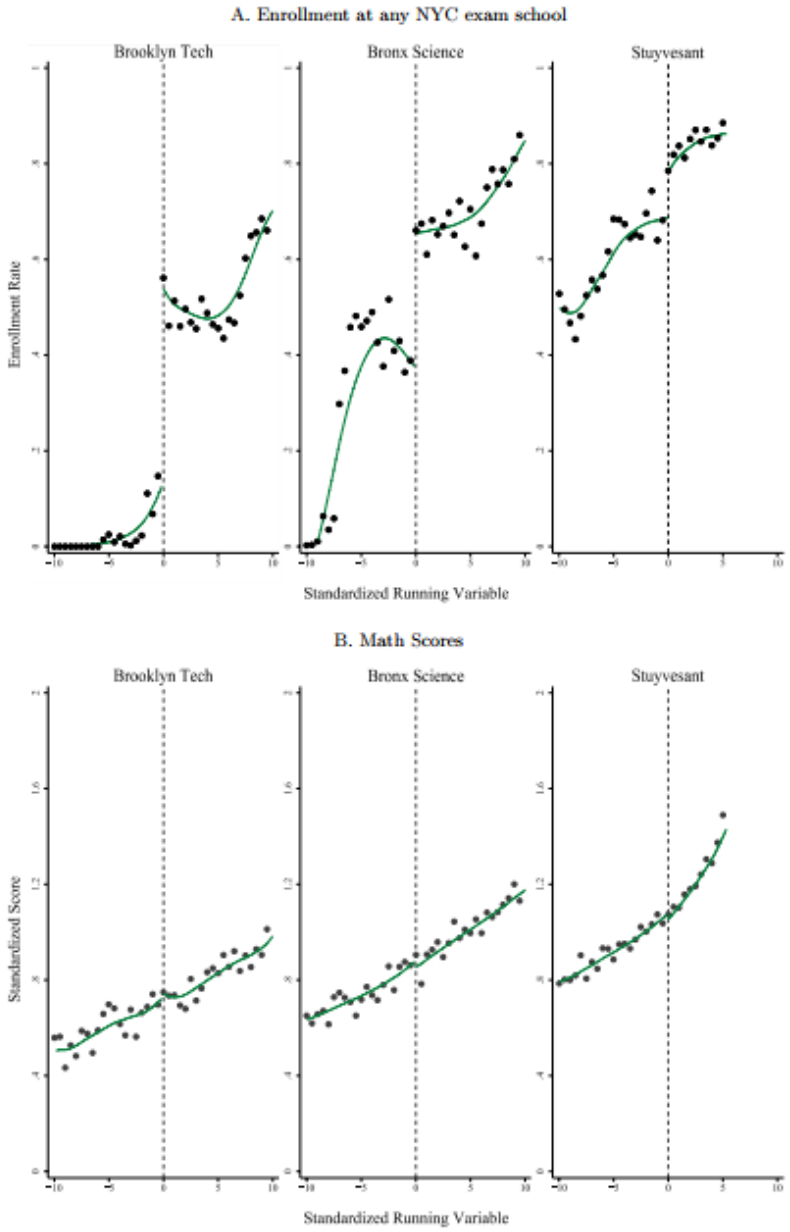


Figure 2: First Stage and Reduced Form for NYC Exam School Admissions RD

图 2：纽约市考试学校招生 RD 的第一阶段和简约式

注：A 组显示了纽约市考试学校的申请者的入学率，这三所考试学校是布鲁克林科技大学、布朗克斯科学大学和斯蒂文森大学，是学生的录取分数线与每所学校的录取分数线的距离的函数。B 组显示了相应的对 Regents 数学标准化考试分数的简约式影响。该图展示了 Abdulkadiroğlu 等人（2014）中应用的录取 RD 方法。

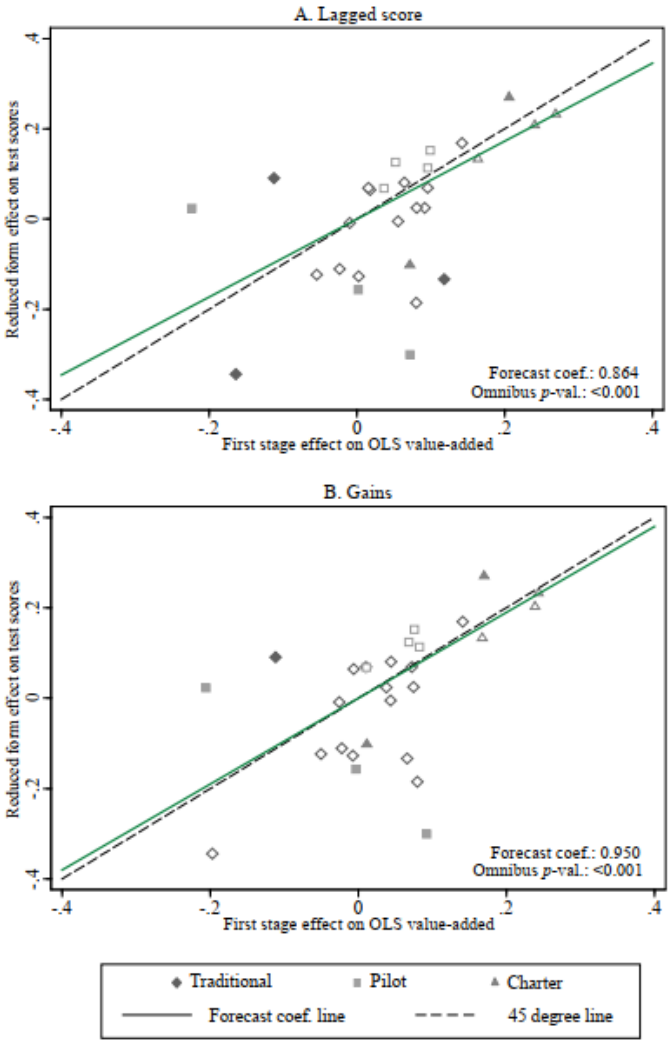


Figure 3: Visual IV Tests for VAM Bias

图 3：VAM 偏差的视觉 IV 测试

注：本图显示了 28 个初中入学抽签的简约式估计值与增值的第一阶段。结果是标准化的六年级数学考试成绩。学校被归类为属于特许、试点和传统公共部门。填充的标记表示在10%的水平上，简约式和第一阶段的估计值有明显的不同。实线的斜率等于表 7 中的预测系数，而虚线表示 45 度线。这里使用的样本来自 Angrist 等人（2017）。

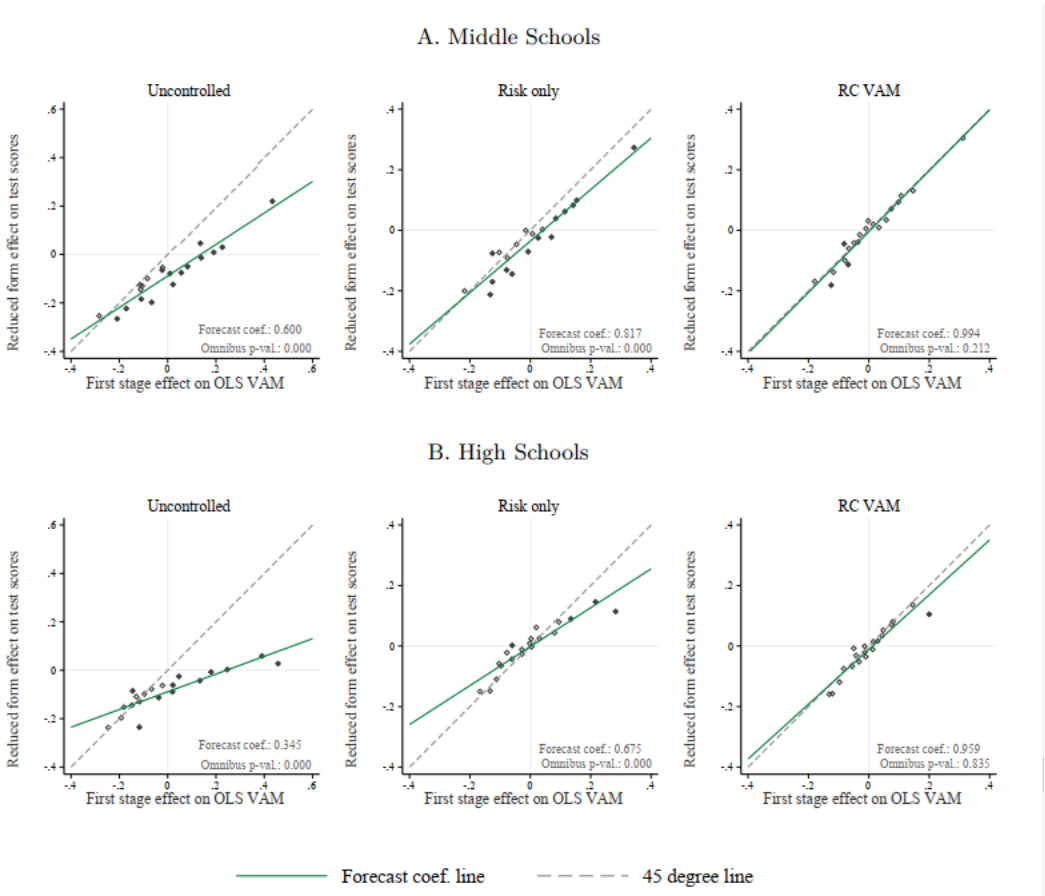


Figure 4: Visual IV Tests for VAM Bias

图 4：VAM 偏差的视觉 IV 测试

注：该图描绘了纽约市初中和高中样本的 20 个学校分配指标中的每一个的简约式的估计值与增值的第一阶段。结果是初中的六年级数学纽约州测试分数，和高中的 SAT 数学分数。分配是按估计的常规 VAM 的分位数来分的。填充的标记表示在 10%的水平上有显著差异的简约式和第一阶段的估计。实线的斜率等于 Angrist 等人（2021）表 2 中的预测系数，而虚线表示 45 度线。这里使用的样本来自 Angrist 等人（2021）。

表 1：马萨诸塞州城市特许学校抽签的平衡性和减员情况

Table 1: Balance and Attrition for Massachusetts Urban Charter Lotteries

	Means		
	MA Urban (1)	Urban Applicant (2)	Balance Coefficient (3)
A. Balance			
Female	0.484	0.498	0.002 (0.017)
Black	0.201	0.479	-0.011 (0.015)
Hispanic	0.321	0.244	0.026 (0.014)
Asian	0.072	0.017	0.001 (0.005)
White	0.375	0.204	-0.007 (0.012)
Special education	0.200	0.176	-0.005 (0.013)
English language learner	0.161	0.103	0.003 (0.010)
Subsidized lunch status	0.688	0.688	0.008 (0.015)
Baseline math score	-0.416	-0.336	-0.020 (0.033)
Baseline English score	-0.454	-0.359	0.002 (0.035)
Joint p-value			0.694
B. Attrition			
Has outcome score	0.702	0.801	0.012 (0.010)
Observations	234,793	6,038	6,038

注：本表第(1)和(2)列分别报告了马萨诸塞州城市学区学生和城市特许抽签申请人的基线特征和成绩测试分数的指标的平均值。第(3)列报告了协变量对特许 offer 虚拟变量的回归系数，控制了抽签风险组指标。括号中报告了稳健的标准误差。联合 P 值来自于对假设的检验，即抽签 offer 在所有基线特征上是平衡的。这里使用的样本来自 Angrist 等人（2013）。

表 2：马萨诸塞州城市特许学校的 2SLS 估计值

Table 2: 2SLS Estimates for Massachusetts Urban Charter Schools

	Treatment Variable					
	Attendance Indicator				Years Attended	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)
Math score effects	0.329 (0.020)	0.454 (0.039)	0.407 (0.019)	0.579 (0.038)	0.236 (0.009)	0.314 (0.020)
First stage	0.567 (0.015)		0.551 (0.015)		1.017 (0.030)	
Non-charter outcome mean	-0.320		-0.268		-0.268	
Number of Applicants	4,281		4,590		4,590	
Sample Coverage	Application Year		All Years		All Years	
Sample Size	4,281		11,458		11,458	

注：本表报告了波士顿特许中学就读对马萨诸塞州城市特许学校申请者数学考试成绩的影响的 OLS 和 2SLS 估计值以及第一阶段估计值。第(1)-(2)列将干预定义为在申请后的学年进入特许学校的指标，并将样本限制在该申请年的考试成绩。第(3)-(4)列使用相同的干预定义，但汇集了 4 至 7 年级的抽签后测试分数。第(5)-(6)列使用全样本，但将干预定义为按结果年级在特许学校度过的年数。所有模型都控制了抽签风险组指标和学生性别、种族、特殊教育、英语学习者、补贴午餐状态、以及年级和年份指标。第(1)-(2)列括号内报告了稳健的标准误差。标准误差在第(3)-(6)列中按学生据类。这里使用的样本来自 Angrist 等人（2013）。

表 3：马萨诸塞州城市特许学校的抽签 compliers 的特征

Table 3: Characteristics of Lottery Compliers at Massachusetts Urban Charter Schools

	Compliers				
	Untreated (1)	Treated (2)	Pooled (3)	Always-takers (4)	Never-takers (5)
Female	0.506	0.510	0.508	0.539	0.463
	(0.023)	(0.021)	(0.016)	(0.024)	(0.017)
Black	0.401	0.380	0.390	0.623	0.490
	(0.022)	(0.021)	(0.016)	(0.023)	(0.017)
Hispanic	0.250	0.300	0.275	0.183	0.228
	(0.02)	(0.018)	(0.013)	(0.019)	(0.014)
Asian	0.022	0.024	0.023	0.004	0.024
	(0.007)	(0.005)	(0.004)	(0.003)	(0.005)
White	0.229	0.216	0.223	0.154	0.215
	(0.018)	(0.016)	(0.012)	(0.016)	(0.014)
Special education	0.190	0.181	0.186	0.158	0.177
	(0.018)	(0.016)	(0.012)	(0.018)	(0.013)
English language learner	0.143	0.148	0.145	0.054	0.088
	(0.015)	(0.013)	(0.010)	(0.011)	(0.010)
Subsidized lunch	0.689	0.705	0.697	0.698	0.666
	(0.021)	(0.019)	(0.014)	(0.022)	(0.016)
Baseline math score	-0.274	-0.312	-0.293	-0.394	-0.301
	(0.047)	(0.041)	(0.032)	(0.045)	(0.036)
Baseline English score	-0.352	-0.349	-0.350	-0.362	-0.299
	(0.050)	(0.043)	(0.033)	(0.046)	(0.038)
Share of sample			0.546	0.197	0.257

注：本表报告了马萨诸塞州城市特许学校抽签申请者中，compliers、always-takers 和 never-takers 的平均基线特征估计值。如第 3.2 节所述，均值是从控制抽签风险组指标的 2SLS 和 OLS 回归中计算出来的。括号内为稳健的标准误差。这里使用的样本来自 Angrist 等人（2013）。

表 4：波士顿特许学校的反事实的命运

Table 4: Counterfactual School Destinies for Boston Charter Compliers

	Target sector					
	Proven providers		Expansion charters		Other charters	
Destiny	Z = 0 (1)	Z = 1 (2)	Z = 0 (3)	Z = 1 (4)	Z = 0 (5)	Z = 1 (6)
Proven providers		1.000	-0.052 (0.038)		0.000 (0.024)	
Expansion charters	0.269 (0.046)			1.000	0.231 (0.034)	
Other charters	0.008 (0.026)		0.047 (0.026)			1.000
Traditional publics	0.528 (0.058)		0.694 (0.058)		0.529 (0.042)	
Pilots	0.180 (0.041)		0.174 (0.041)		0.118 (0.023)	

注：本表报告了在波士顿特许学校抽签中，未被处理（*Z=0*）和被处理（*Z=1*）的 compliers 在特定的后备学校类型中的比例。本表报告了波士顿特许学校抽签的申请人中，未被处理（Z=0）和被处理（Z=1）的人在特定后备学校类型中的比例。左边标明的目的地是招收接受干预和未接受干预的 compliers 的部门。括号内为稳健的标准误差。这里使用的抽签样本来自 Cohodes 等人（2021）。

表 5：波士顿特许学校的多部门 2SLS 估计值

Table 5: Multi-sector 2SLS Estimates for Boston Charter Schools

	Before charter expansion			After charter expansion			
		Estimates			Estimates		
	Non-charter mean (1)	Proven providers (2)	Other charters (3)	Non-charter mean (4)	Proven providers (5)	Expansion charters (6)	Other charters (7)
Math Score	0.117	0.320	0.183	-0.074	0.365	0.326	0.193
		(0.037)	(0.026)		(0.070)	(0.074)	(0.055)
First stage							
Immediate offer		1.304	1.554		0.795	0.659	0.930
		(0.067)	(0.047)		(0.054)	(0.046)	(0.052)
Waitlist offer		1.027	0.984		0.400	0.348	0.853
		(0.050)	(0.061)		(0.048)	(0.041)	(0.071)

注：本表报告了特许抽签对特许学校入学年限的第一阶段影响，以及波士顿多种类型的特许学校入学对数学考试成绩的 2SLS 估计。样本中叠加了五至八年级的抽签后考试成绩。内生变量是在不同类型的特许学校（扩张前的成熟提供者、扩张前的其他特许学校、扩张后的成熟提供者、扩张学校和扩张后的其他特许学校）中度过的年限的计数。工具变量是每个学校类型的即时和等待名单上的抽签 offer 虚拟变量。对于在抽签当天获得席位的申请者来说，立即 offer 等于 1。候补名单上的申请者获得了候补名单上的席位，则等于 1。控制因素包括抽签风险组，以及性别、种族、民族、女性-少数族裔交互、特殊教育、英语学习者、补贴午餐状态、以及年级和年份指标。在括号内报告了标准误差，按学生聚类。这里使用的样本来自 Cohodes 等人（2021）。

表 6：关于丹佛特许效应的替代 IV 策略

Table 6: Alternative IV Strategies for Denver Charter Effects

	Instruments		
	Offer	First Choice	Qualification
	(1)	(2)	(3)
Math score	0.417	0.515	0.379
	(0.050)	(0.064)	(0.092)
First stage	0.443	0.347	0.457
	(0.024)	(0.022)	(0.021)
Risk controls	DA Score	first-choice risk sets	preference risk sets
Equivalent sample increase vs. column (1)		1.64	3.46
Observations	2,099	2,222	3,502

注：本表报告了对丹佛市采用集中分配策略的学生就读特许学校的影响的 IV 估计。第一行比较了特许学校就读对数学成绩影响的替代 2SLS 估计。第二行报告了相应的第一阶段的估计。第(1)列以集中分配到特许学校作为特许入学的工具变量，控制模拟特许分配倾向得分的百分比和其他基线协变量。第(2)列以特许学校的第一选择分配为工具，控制第一选择的固定效应和其他基线协变量。第(3)列用特许学校资格工具来衡量特许学校的入学率，控制偏好的固定效应和其他基线协变量。第四行报告了为达到相当于使用任何特许学校资格工具变量的精确收益而需要增加的样本量。括号中报告了稳健的标准误差。这里使用的样本与 Abdulkadiroğlu 等人（2017）一样。

表 7：波士顿学校的 VAM 偏差测试

Table 7: VAM Bias Tests for Boston Schools

	Value-added model		
	Uncontrolled (1)	Lagged score (2)	Gains (3)
Forecast coefficient	0.396 (0.056)	0.864 (0.075)	0.950 (0.084)
p-values:			
Forecast bias	<0.001	0.071	0.554
Overidentification	<0.001	0.003	0.006
Omnibus	<0.001	<0.001	<0.001

注：本表报告了使用波士顿六年级学生的数学成绩对传统的VAMs进行偏差测试的结果。未控制的模型只包括测试年的指标作为控制变量。滞后分数 VAM 包括基线数学和 ELA 分数的立方多项式，以及申请年份、性别、种族、午餐补贴、特殊教育、英语学习者状态、以及基线缺勤和停课的指标。收益 VAM 放弃了滞后的分数控制，并而使用基线的分数增长作为结果。预测系数来自于测试分数对传统 VAMs 的拟合值的 IV 回归，用抽签 offer 指标的拟合值来作为工具变量。IV 模型是通过渐进有效的 GMM 程序估计的，并控制了分配层的固定效应、人口统计学变量和滞后分数。预测偏差检验检查预测系数是否等于 1，而过度识别检验检查 IV 模型的过度识别限制。综合检验结合了预测偏差和过度识别限制。括号内报告了标准误差。这里使用的样本来自 Angrist 等人（2017）。