Research Design meets Market Design: Theory and Applications

Parag A. Pathak, MIT
Econometric Society Summer School

August 2022

Research Design meets Market Design

- Several lectures have focused on the theoretical and *allocative* properties of matching mechanisms
    - e.g., how many obtained a desired choice under different systems?
- But resource allocation algorithms also generate data that can be used to study the *productive* dimensions of assignment
    - e.g., does a medical resident obtain better training at their first choice program?

Research Design meets Market Design

- Several lectures have focused on the theoretical and *allocative* properties of matching mechanisms
  - e.g., how many obtained a desired choice under different systems?

- But resource allocation algorithms also generate data that can be used to study the *productive* dimensions of assignment
  - e.g., does a medical resident obtain better training at their first choice program?

- Today I'll talk about a growing literature studying how to use allocation algorithms as "experiment generators" based on two papers:
  1) Abdulkadiroğlu, Angrist, Pathak, Narita (2017): "Research Design meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica*, 85(5): 1373-1432.
  2) Abdulkadiroğlu, Angrist, Pathak, Narita (2022): "Breaking Ties: Regression Discontinuity Design meets Market Design." *Econometrica*, 90(1): 117-151.

DA with lottery tie-breaking

- DA and other allocation mechanisms satisfy the **equal treatment of equals** property: applicants with the same preferences and priorities (or "type") have the same probability distribution over assignments
- Embedded in DA, therefore, is a stratified randomized trial. How best to use these designs?

DA with lottery tie-breaking

- DA and other allocation mechanisms satisfy the **equal treatment of equals** property: applicants with the same preferences and priorities (or "type") have the same probability distribution over assignments
- Embedded in DA, therefore, is a stratified randomized trial. How best to use these designs?
- Two applications
  1) Study of charter and other sectors in the Denver
  2) Study of long-term effects of Pre-K in Boston
- Both studies use algorithm-generated seat offers to compute causal effects as part of an IV strategy, which isolates the random part of offers, eliminating confounding generated by differences in student type

Defining the Problem

Research Design: Extracting Ignorable Assignment

- Let $D_i(s)$ indicate whether student i is offered a seat at school s
  - Applicants are characterized by prefs and priorities, their type, θ
  - Type affects assignment and is correlated w/outcomes, hence a powerful source of omitted variables bias (OVB)

Research Design: Extracting Ignorable Assignment

- Let $D_i(s)$ indicate whether student i is offered a seat at school s
  - Applicants are characterized by prefs and priorities, their type, θ
  - Type affects assignment and is correlated w/outcomes, hence a powerful source of omitted variables bias (OVB)
- DA induces a stratified RCT
  - Let W_i be any r.v. independent of lottery numbers

$$Pr[D_i(s) = 1 | W_i, \theta_i = \theta] = Pr[D_i(s) = 1 | \theta_i = \theta] \tag{1}$$

  - W_i includes potential outcomes and student characteristics like sibling and free lunch status
  - Conditioning on type therefore eliminates any OVB in comparisons by offer status

Research Design: Extracting Ignorable Assignment

- Let $D_i(s)$ indicate whether student i is offered a seat at school s
  - Applicants are characterized by prefs and priorities, their type, θ
  - Type affects assignment and is correlated w/outcomes, hence a powerful source of omitted variables bias (OVB)
- DA induces a stratified RCT
  - Let W_i be any r.v. independent of lottery numbers

$$Pr[D_i(s) = 1 | W_i, \theta_i = \theta] = Pr[D_i(s) = 1 | \theta_i = \theta] \tag{1}$$

  - W_i includes potential outcomes and student characteristics like sibling and free lunch status
  - Conditioning on type therefore eliminates any OVB in comparisons by offer status
- But full type conditioning is **impractical**: it eliminates many students and schools from statistical analyses
  - Denver's 5,000 charter applicants include 4,300 types

Propensity Score

We condition instead on the **propensity score**, the probability of assignment to school s for a given type:

$$p_s(\theta) = Pr[D_i(s) = 1|\theta_i = \theta]$$

Theorem (Rosenbaum & Rubin 1983)

Conditional independence property (1) *implies that for any* W_i *that is independent of lottery numbers,*

$$P[D_i(s) = 1|W_i, p_s(\theta_i)] = P[D_i(s) = 1|p_s(\theta_i)] = p_s(\theta_i)$$

Propensity Score

We condition instead on the **propensity score**, the probability of assignment to school s for a given type:

$$p_s(\theta) = Pr[D_i(s) = 1 | \theta_i = \theta]$$

Theorem (Rosenbaum & Rubin 1983)

Conditional independence property (1) *implies that for any* W_i *that is independent of lottery numbers,*

$$P[D_i(s) = 1 | W_i, p_s(\theta_i)] = P[D_i(s) = 1 | p_s(\theta_i)] = p_s(\theta_i)$$

- Why is this useful?
  - The score is much coarser than θ: many types share a score
  - The score identifies the maximal set of applicants for whom we have a randomized school-assignment experiment
  - The score reveals the experimental design embedded in DA: we know (and will show) its structure; particularly valuable in non-lottery case

Example 1: The Score Pools Types

- Five students $\{1, 2, 3, 4, 5\}$; three schools $\{a, b, c\}$, each with one seat
  - student preferences

$$1 : a \succ b$$
$$2 : a \succ b$$
$$3 : a$$
$$4 : c \succ a$$
$$5 : c$$

  - school priorities
    - 2 has priority at b
    - 5 has priority at c

Example 1: The Score Pools Types

- Five students $\{1, 2, 3, 4, 5\}$; three schools $\{a, b, c\}$, each with one seat
  - student preferences

$$1 : a \succ b$$
$$2 : a \succ b$$
$$3 : a$$
$$4 : c \succ a$$
$$5 : c$$

  - school priorities
    - 2 has priority at b
    - 5 has priority at c
- Types are unique, ruling out research with full-type conditioning

Example 1: The Score Pools Types

- Five students $\{1, 2, 3, 4, 5\}$; three schools $\{a, b, c\}$, each with one seat
  - student preferences

$$
\begin{aligned}
1 &: a \succ b \\
2 &: a \succ b \\
3 &: a \\
4 &: c \succ a \\
5 &: c
\end{aligned}
$$

  - school priorities
    - 2 has priority at b
    - 5 has priority at c
- Types are unique, ruling out research with full-type conditioning
- The score pools: DA assigns students 1, 2, 3, and 4 to school a each with probability 0.25
  - 5 beats 4 at c by virtue of priority; this leaves 1, 2, 3, and 4 all applying to a in the second round and no one advantaged there

Example 2: Further Pooling in Large Markets

- Four students $\{1,2,3,4\}$; three schools $\{a,b,c\}$, each with one seat and no priorities
  - student preferences

$$
\begin{aligned}
1 &: c \\
2 &: c \succ b \succ a \\
3 &: b \succ a \\
4 &: a
\end{aligned}
$$

Types are again unique

Example 2: Further Pooling in Large Markets

- Four students $\{1, 2, 3, 4\}$; three schools $\{a, b, c\}$, each with one seat and no priorities
  - student preferences

$$\begin{aligned} &1 : c \\ &2 : c \succ b \succ a \\ &3 : b \succ a \\ &4 : a \end{aligned}$$

  Types are again unique

- There are $4! = 24$ possible assignments. Enumerating these, we find
  - $p_a(1) = 0$, since 1 doesn't rank a
  - $p_a(2) = 2/24 = 1/12$
  - $p_a(3) = 1/24$
  - $p_a(4) = 1 - p_a(1) - p_a(2) = 21/24$
- No pooling

Understanding Example 2

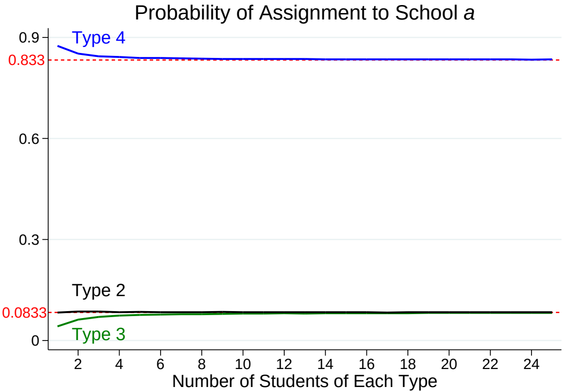
- The probability of assignment to *a* is determined by
  - Failure to be seated at schools ranked more highly than *a*
  - Success in the competition for *a* conditional on this failure

Understanding Example 2

- The probability of assignment to a is determined by
  - Failure to be seated at schools ranked more highly than a
  - Success in the competition for a conditional on this failure
- Type 2 is seated at a when:
  - Schools he's ranked ahead of a (schools b and c) are filled by others
  - He also beats type 4 in competition for a seat at a
    - This happens for two realizations of the form $(s,t,2,4)$ for $s,t=1,3$
- Type 3 is seated at a when:
  - Schools he's ranked ahead of a (school b) are filled by another and he beats type 4 at a
    - This happens only when the lottery order is $(1,2,3,4)$

The Large-Market P-Score

- An $n-$*scaled* version of Example 2:
  - n each of types 1-4 apply to 3 schools, each with n seats
  - Enumeration with large n is a chore, but repeating lottery draws reveals a common score for types 2 and 3 for n more than a few:



Deriving the DA Score

Score Computation

- $p_s(\theta)$ is generated by a permutation distribution, a relative frequency generated by all possible lottery realizations

Score Computation

- $p_s(\theta)$ is generated by a permutation distribution, a relative frequency generated by all possible lottery realizations
- That is 26,000! lotteries for DPS....
  - The LLN tells us it's enough to sample lotteries. But since covariates are discrete, the resulting empirical $\hat{p}_s(\theta)$ has as many points of support as does θ
  - Sim scores are a black box; sample-based $\hat{p}_s(\theta)$ must be smoothed

Score Computation

- $p_s(\theta)$ is generated by a permutation distribution, a relative frequency generated by all possible lottery realizations
- That is 26,000! lotteries for DPS....
  - The LLN tells us it's enough to sample lotteries. But since covariates are discrete, the resulting empirical $\hat{p}_s(\theta)$ has as many points of support as does θ
  - Sim scores are a black box; sample-based $\hat{p}_s(\theta)$ must be smoothed
- Is there a formula?
  - Except for special cases, $p_s(\theta)$ as no closed form

Score Computation

- $p_s(\theta)$ is generated by a permutation distribution, a relative frequency generated by all possible lottery realizations
- That is 26,000! lotteries for DPS....
  - The LLN tells us it's enough to sample lotteries. But since covariates are discrete, the resulting empirical $\hat{p}_s(\theta)$ has as many points of support as does θ
  - Sim scores are a black box; sample-based $\hat{p}_s(\theta)$ must be smoothed
- Is there a formula?
  - Except for special cases, $p_s(\theta)$ as no closed form
- The large market continuum model provides the formula we need

Score Computation

- $p_s(\theta)$ is generated by a permutation distribution, a relative frequency generated by all possible lottery realizations
- That is 26,000! lotteries for DPS....
  - The LLN tells us it's enough to sample lotteries. But since covariates are discrete, the resulting empirical $\hat{p}_s(\theta)$ has as many points of support as does θ
  - Sim scores are a black box; sample-based $\hat{p}_s(\theta)$ must be smoothed
- Is there a formula?
  - Except for special cases, $p_s(\theta)$ as no closed form
- The large market continuum model provides the formula we need
  - The DA score for a continuum market approximates the score as a function of a few easily-computed sufficient statistics
  - The DA score is automatically coarse: no simulation, smoothing or rounding required
  - The DA score reveals the nature of the stratified trial embedded in DA: which schools have random assignment and why

DA Formalities

- I students with preferences $\succ_i$ and priorities for school s given by $\rho_{is} \in \{1, ..., K, \infty\}$
- Student i's type is $\theta_i = (\succ_i, \rho_i)$, where ρ_i is the vector of i's ρ_{is}
- $s = 1, ..., S$ schools, with capacity vector $\mathrm{q} = (q_1, ..., q_S)$
  - In the continuum (large market), $I = [0, 1]$ and q_s is the proportion of I that can be seated at s
- Student i's *lottery number*, r_i, is i.i.d. uniform $[0, 1]$
- Student i's *rank* at school s is

$$\pi_{is} = \rho_{is} + r_i$$

DA Formalities

- I students with preferences $\succ_i$ and priorities for school s given by $\rho_{is} \in \{1, ..., K, \infty\}$
- Student i's type is $\theta_i = (\succ_i, \rho_i)$, where ρ_i is the vector of i's ρ_{is}
- $s = 1, ..., S$ schools, with capacity vector $\mathrm{q} = (q_1, ..., q_S)$
    - In the continuum (large market), $I = [0, 1]$ and q_s is the proportion of I that can be seated at s
- Student i's *lottery number*, r_i, is i.i.d. uniform $[0, 1]$
- Student i's *rank* at school s is

$$\pi_{is} = \rho_{is} + r_i$$

- DA assignment is determined by a *vector of cutoffs,* c_s: applicants to s with $\pi_{is} \leq c_s$ and $\pi_{i\tilde{s}} > c_{\tilde{s}}$ $\forall \tilde{s}$ they prefer to s, are seated at s
    - Lottery numbers matter for assignment to s only in the *marginal priority group*

Illustrating Cutoffs and Marginal Priorities

Rank	Priority	Lottery No.	Offer	
1.13	1	.13	1	
1.99	1	.99	1	
2.05	2	.05	1	
2.35	2	.35	1	c_s=2.35
2.57	2	.57	0	
2.61	2	.61	0	
3.12	3	.12	0	
3.32	3	.32	0	

- Marginal priority, denoted ρ_s, is the integer part of c_s; here, $\rho_s = 2$
- The *lottery cutoff, denoted* τ_s, is the decimal part of c_s; here, $\tau_s = .35$

Assignment Outcomes: Partitioning Types

- Let Θ_s denote the set of types who rank s
  - $B_{\theta s}$ denotes the set of schools that type θ prefers to s

Assignment Outcomes: Partitioning Types

- Let Θ_s denote the set of types who rank s
  - $B_{\theta s}$ denotes the set of schools that type θ prefers to s
- This set is partitioned by:
  - Θ_s^n, defined by $\rho_{\theta s} > \rho_s$
    - These **never-seated** applicants have worse than marginal priority at s
    - *No one* in this group is seated at s

Assignment Outcomes: Partitioning Types

- Let Θ_s denote the set of types who rank s
  - $B_{\theta s}$ denotes the set of schools that type θ prefers to s
- This set is partitioned by:
  - Θ_s^n, defined by $\rho_{\theta s} > \rho_s$
    - These **never-seated** applicants have worse than marginal priority at s
    - *No one* in this group is seated at s
  - Θ_s^a, defined by $\rho_{\theta s} < \rho_s$
    - These **always-seated** applicants clear marginal priority at s
    - *Everyone* in this group is seated at s when not seated at a school in $B_{\theta s}$

Assignment Outcomes: Partitioning Types

- Let Θ_s denote the set of types who rank s
  - $B_{\theta s}$ denotes the set of schools that type θ prefers to s
- This set is partitioned by:
  - Θ_s^n, defined by $\rho_{\theta s} > \rho_s$
    - These **never-seated** applicants have worse than marginal priority at s
    - *No one* in this group is seated at s
  - Θ_s^a, defined by $\rho_{\theta s} < \rho_s$
    - These **always-seated** applicants clear marginal priority at s
    - *Everyone* in this group is seated at s when not seated at a school in $B_{\theta s}$
  - Θ_s^c, defined by $\rho_{\theta s} = \rho_s$
    - These **conditionally seated** applicants have marginal priority at s
    - Members of this group are seated at s when not seated at a school in $B_{\theta s}$ *and* they clear the lottery cutoff at s

Assignment Risk: Most Informative Disqualification

- Define

$$\text{MID}_{\theta s} = \begin{cases} 0 & \text{if } \rho_{\theta \tilde{s}} > \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \\ 1 & \text{if } \rho_{\theta \tilde{s}} < \rho_{\tilde{s}} \text{ for some } \tilde{s} \in B_{\theta s} \\ \max\{\tau_{\tilde{s}} \mid \rho_{\theta \tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\} & \text{if } \rho_{\theta \tilde{s}} \geq \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \end{cases}$$

Assignment Risk: Most Informative Disqualification

- Define

$$\mathrm{MID}_{\theta s} = \begin{cases} 0 & \text{if } \rho_{\theta\tilde{s}} > \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \\ 1 & \text{if } \rho_{\theta\tilde{s}} < \rho_{\tilde{s}} \text{ for some } \tilde{s} \in B_{\theta s} \\ \max\{\tau_{\tilde{s}} \mid \rho_{\theta\tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\} & \text{if } \rho_{\theta\tilde{s}} \geq \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \end{cases}$$

- MID tells us how the lottery number distribution for applicants to s is truncated by qualification at more preferred schools

Assignment Risk: Most Informative Disqualification

- Define

$$\text{MID}_{\theta s} = \begin{cases} 0 & \text{if } \rho_{\theta\tilde{s}} > \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \\ 1 & \text{if } \rho_{\theta\tilde{s}} < \rho_{\tilde{s}} \text{ for some } \tilde{s} \in B_{\theta s} \\ \max\{\tau_{\tilde{s}} \mid \rho_{\theta\tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\} & \text{if } \rho_{\theta\tilde{s}} \geq \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \end{cases}$$

- MID tells us how the lottery number distribution for applicants to s is truncated by qualification at more preferred schools
  - MID is 0 when priority status is worse-than-marginal at all higher ranked schools (**no truncation**)

Assignment Risk: Most Informative Disqualification

- Define

$$\mathrm{MID}_{\theta s} = \begin{cases} 0 & \text{if } \rho_{\theta\tilde{s}} > \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \\ 1 & \text{if } \rho_{\theta\tilde{s}} < \rho_{\tilde{s}} \text{ for some } \tilde{s} \in B_{\theta s} \\ \max\{\tau_{\tilde{s}} \mid \rho_{\theta\tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\} & \text{if } \rho_{\theta\tilde{s}} \geq \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \end{cases}$$

- MID tells us how the lottery number distribution for applicants to s is truncated by qualification at more preferred schools
  - MID is 0 when priority status is worse-than-marginal at all higher ranked schools (**no truncation**)
  - MID is 1 if $B_{\theta s}$ includes a school where θ is seated with certainty (**complete truncation**)

Assignment Risk: Most Informative Disqualification

- Define

$$\text{MID}_{\theta s} = \begin{cases} 0 & \text{if } \rho_{\theta\tilde{s}} > \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \\ 1 & \text{if } \rho_{\theta\tilde{s}} < \rho_{\tilde{s}} \text{ for some } \tilde{s} \in B_{\theta s} \\ \max\{\tau_{\tilde{s}} \mid \rho_{\theta\tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\} & \text{if } \rho_{\theta\tilde{s}} \geq \rho_{\tilde{s}} \text{ for all } \tilde{s} \in B_{\theta s} \end{cases}$$

- MID tells us how the lottery number distribution for applicants to s is truncated by qualification at more preferred schools
  - MID is 0 when priority status is worse-than-marginal at all higher ranked schools (**no truncation**)
  - MID is 1 if $B_{\theta s}$ includes a school where θ is seated with certainty (**complete truncation**)
  - For those who are marginal or worse at all schools they prefer to s, and marginal somewhere, MID is the most forgiving cutoff in the set of schools at which they're marginal
    - Applicants who clear $\max\{\tau_{\tilde{s}} \mid \rho_{\theta\tilde{s}} = \rho_{\tilde{s}}, \tilde{s} \in B_{\theta s}\}$ are seated in $B_{\theta s}$, and so not at risk for a seat at s

The DA Propensity Score

Theorem

In a continuum economy, $\Pr[D_i(s)=1|\theta_i=\theta]=\varphi_s(\theta)\equiv$

$$\begin{cases} 0 & \text{if } \theta\in\Theta_s^n \\ (1-MID_{\theta s}) & \text{if } \theta\in\Theta_s^a \\ (1-MID_{\theta s})\times\max\left\{0,\dfrac{\tau_s-MID_{\theta s}}{1-MID_{\theta s}}\right\} & \text{if } \theta\in\Theta_s^c \end{cases}$$

where we set $\varphi_s(\theta)=0$ *when* $MID_{\theta s}=1$ *and* $\theta\in\Theta_s^c$

The DA Propensity Score

Theorem

In a continuum economy, $\Pr[D_i(s) = 1 | \theta_i = \theta] = \varphi_s(\theta) \equiv$

$$\begin{cases} 0 & \text{if } \theta \in \Theta_s^n \\ (1 - MID_{\theta s}) & \text{if } \theta \in \Theta_s^a \\ (1 - MID_{\theta s}) \times \max\left\{0, \dfrac{\tau_s - MID_{\theta s}}{1 - MID_{\theta s}}\right\} & \text{if } \theta \in \Theta_s^c \end{cases}$$

where we set $\varphi_s(\theta) = 0$ *when* $MID_{\theta s} = 1$ *and* $\theta \in \Theta_s^c$

- $\mathrm{MID}_{\theta s}$, τ_s, and Θ are population quantities, fixed in the continuum
    - Our second theorem shows that the sample analog of $\varphi_s(\theta)$ converges uniformly to the finite market score as market size grows

DA Econometrics

- Estimating the DA score described by Theorem 1
  - **Formula**: assign students by priority status to Θ_s^n, Θ_s^a, or Θ_s^c as realized in the match; plug empirical τ_s and $\mathrm{MID}_{\theta s}$ into $\varphi_s(\theta)$

DA Econometrics

- Estimating the DA score described by Theorem 1
  - **Formula**: assign students by priority status to Θ_s^n, Θ_s^a, or Θ_s^c as realized in the match; plug empirical τ_s and $\mathrm{MID}_{\theta s}$ into $\varphi_s(\theta)$
  - **Frequency**: use empirical offer rates in cells with constant $\varphi_s(\theta)$

DA Econometrics

- Estimating the DA score described by Theorem 1
  - **Formula**: assign students by priority status to Θ_s^n, Θ_s^a, or Θ_s^c as realized in the match; plug empirical τ_s and $\text{MID}_{\theta s}$ into $\varphi_s(\theta)$
  - **Frequency**: use empirical offer rates in cells with constant $\varphi_s(\theta)$
- Simulating the finite-market score
  - We do this using 1,000,000 lottery draws, running DA for each, and computing the empirical assignment rate for each type
    - This converges (as # of draws increase) to the finite-market score
    - No dimension reduction in our data: 1148 score values for 1523 applicants with non-degenerate charter risk. The simulated score requires smoothing

DA Econometrics

- Estimating the DA score described by Theorem 1
  - **Formula**: assign students by priority status to Θ_s^n, Θ_s^a, or Θ_s^c as realized in the match; plug empirical τ_s and $\mathrm{MID}_{\theta s}$ into $\varphi_s(\theta)$
  - **Frequency**: use empirical offer rates in cells with constant $\varphi_s(\theta)$
- Simulating the finite-market score
  - We do this using 1,000,000 lottery draws, running DA for each, and computing the empirical assignment rate for each type
    - This converges (as # of draws increase) to the finite-market score
    - No dimension reduction in our data: 1148 score values for 1523 applicants with non-degenerate charter risk. The simulated score requires smoothing
  - By contrast with simulation, Theorem 1 reveals *why* we have random assignment at one school or another.

Denver's Charter Sector: Empirical Strategy

- DPS has a large charter sector, part of the SchoolChoice match.
- Impact evaluation for the charter sector
  - An any-charter offer dummy, D_i, is the sum of all individual charter offers (our instrument)
  - The *any-charter p-score* (our key control) is the sum of the scores for each charter that type θ ranks
  - C_i indicates any-charter enrollment (our "endogenous" variable)

Denver's Charter Sector: Empirical Strategy

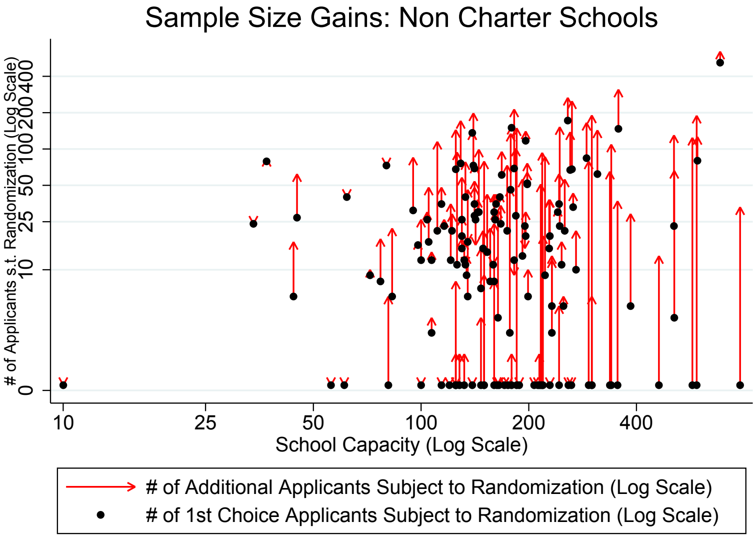
- DPS has a large charter sector, part of the SchoolChoice match.
- Impact evaluation for the charter sector
  - An any-charter offer dummy, D_i, is the sum of all individual charter offers (our instrument)
  - The *any-charter p-score* (our key control) is the sum of the scores for each charter that type θ ranks
  - C_i indicates any-charter enrollment (our "endogenous" variable)
- 2SLS First and Second stages

$$C_i = \sum_x \gamma(x) d_i(x) + \delta D_i + \nu_i$$

$$Y_i = \sum_x \alpha(x) d_i(x) + \beta C_i + \varepsilon_i$$

  - $d_i(x)$: dummies for propensity score values (cells), indexed by x
  - $\gamma(x)$ and $\alpha(x)$: associated "score effects"

Gains Over First Choice



STRIVE DA Score Anatomy

Table 2: DA Score anatomy

				DA Score = 0			DA Score in (0,1)		DA Score = 1
				Θ_s^n	Θ_s^c	Θ_s^a	Θ_s^c	Θ_s^a	Θ_s^a
				$(\rho_{\theta s} > \rho_s)$	$(\rho_{\theta s} = \rho_s)$	$(\rho_{\theta s} < \rho_s)$	$(\rho_{\theta s} = \rho_s)$	$(\rho_{\theta s} < \rho_s)$	$(\rho_{\theta s} < \rho_s)$
	Eligible applicants	Capacity	Offers	$0 \leq MID \leq 1$	$MID \geq \tau_s$	$MID = 1$	$MID < \tau_s$	$0 < MID < 1$	$MID = 0$
STRIVE Prep School	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GVR	324	147	112	0	0	159	0	116	49
Lake	274	147	126	0	0	132	0	26	116
Highland	244	147	112	0	0	121	0	21	102
Montbello	188	147	37	0	0	128	0	31	29
Federal	574	138	138	78	284	3	171	1	37
Westwood	494	141	141	53	181	4	238	0	18

Notes: This table shows how formula scores are determined for STRIVE school seats in grade 6 (all 6th grade seats at these schools are assigned in a single bucket; ineligible applicants, who have a score of zero, are omitted). Column 3 records offers made to these applicants. Columns 4-6 show the number of applicants in partitions with a score of zero. Likewise, columns 7 and 8 show the number of applicants subject to random assignment. Column 9 shows the number of applicants with certain offers.

2SLS and Semiparametric Alternatives

Table 6: Comparison of 2SLS and semiparametric estimates of charter effects

	Frequency (saturated)		Formula (saturated)		Simulation (hundredths)	
	2SLS	Semiparametric	2SLS	Semiparametric	2SLS	Semiparametric
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.496***	0.443***	0.524***	0.486***	0.543***	0.474**
	(0.076)	(0.105)	(0.071)	(0.105)	(0.075)	(0.212)
	{0.071}		{0.076}		{0.079}	
Reading	0.127*	0.106	0.120*	0.118	0.106	0.127
	(0.065)	(0.107)	(0.073)	(0.115)	(0.069)	(0.173)
	{0.065}		{0.069}		{0.071}	
Writing	0.325***	0.326***	0.356***	0.364***	0.324***	0.305**
	(0.079)	(0.102)	(0.082)	(0.113)	(0.079)	(0.145)
	{0.077}		{0.080}		{0.080}	
N	1,102	1,093	1,083	1,081	1,137	1,137

Notes: This table compares 2SLS and semiparametric estimates of charter attendance effects on the 2012-13 TCAP scores of Denver 4th-10th graders. The instrument is an any-charter offer dummy. The semiparametric estimator is described in Section 3.5. In addition to score variables, 2SLS estimates include controls for grade tested, gender, origin school charter status, race, gifted status, bilingual status, subsidized price lunch eligibility, special education, limited English proficient status, and baseline test scores. Semiparametric models use these same variables as controls when computing the score weighting function. Standard errors in parentheses are from a Bayesian bootstrap. Conventional robust standard errors for 2SLS estimates are reported in braces.
*significant at 10%; **significant at 5%; ***significant at 1%

The semiparametric scheme uses a score-weighted Abadie (2003)-style estimand; 2SLS estimates are close, but more precise

Alternative IV Results

Table 8: Other IV strategies

	Charter attendance effect				
	Offer instrument with DA score (frequency) controls (saturated)	First choice charter offer with risk set controls	Qualification instrument with risk set controls	Sample size increase for equivalent gain (col 2 vs col 1)	Sample size increase for equivalent gain (col 3 vs col 1)
	(1)	(2)	(3)	(4)	(5)
		A. First stage estimates			
First stage for charter offers	1.000	0.774***	0.476***		
	--	(0.026)	(0.024)		
First stage for charter enrollment	0.410***	0.323***	0.178***		
	(0.031)	(0.035)	(0.027)		
		B. 2SLS estimates			
Math	0.496***	0.596***	0.409***		
	(0.071)	(0.102)	(0.149)	2.0	4.4
Reading	0.127**	0.227**	0.229		
	(0.065)	(0.102)	(0.144)	2.5	4.9
Writing	0.325***	0.333***	0.505***		
	(0.077)	(0.119)	(0.162)	2.4	4.5
N (students)	1,102	1,125	1,969		
N (schools)	30	15	24		

Non-lottery tie-breaking

A Local Selection Story

- Consider two applicants to Boston Latin School
- Randomness comes from applicant i's *tie-breaker*, R_i
  - Unlike lottery case, we do not know this distribution $F_R^i(r)$
- Among such applicants, BLS offers go to all qualified: $R_i < \tau_s$, where R_i is the tie-breaker and τ_s the cutoff

A Local Selection Story

- Consider two applicants to Boston Latin School
- Randomness comes from applicant i's *tie-breaker*, R_i
  - Unlike lottery case, we do not know this distribution $F_R^i(r)$
- Among such applicants, BLS offers go to all qualified: $R_i < \tau_s$, where R_i is the tie-breaker and τ_s the cutoff
- Define tie-breaker intervals around cutoff:

$$t_{is}(\delta) = \begin{cases} n & \text{if } R_i > \tau_s + \delta & \textit{never seated} \\ a & \text{if } R_i < \tau_s - \delta & \textit{always seated} \\ c & \text{if } R_i \in [\tau_s - \delta, \tau_s + \delta] & \textit{conditionally seated} \end{cases}$$

A Local Selection Story

- Consider two applicants to Boston Latin School
- Randomness comes from applicant i's *tie-breaker*, R_i
  - Unlike lottery case, we do not know this distribution $F_R^i(r)$
- Among such applicants, BLS offers go to all qualified: $R_i < \tau_s$, where R_i is the tie-breaker and τ_s the cutoff
- Define tie-breaker intervals around cutoff:

$$t_{is}(\delta) = \begin{cases} n & \text{if } R_i > \tau_s + \delta & \textit{never seated} \\ a & \text{if } R_i < \tau_s - \delta & \textit{always seated} \\ c & \text{if } R_i \in [\tau_s - \delta, \tau_s + \delta] & \textit{conditionally seated} \end{cases}$$

Lemma (Simple Local Score)

For any pre-treatment variable, W_i, and applicant type, θ:

$$\underbrace{\lim_{\delta \to 0} E[1(R_i \le \tau_s) | \theta_i = \theta, t_{is}(\delta) = t, W_i]}_{\textit{Local Propensity Score}} = \psi_s(\theta, t) = \begin{cases} 0 & \textit{if } t = n \\ 1 & \textit{if } t = a \\ 0.5 & \textit{if } t = c \end{cases}$$

No OVB!

- Local score-conditioning kills OVB from type and non-random tie-breakers

Corollary (local conditional independence)

$$\lim_{\delta \to 0} E[1(R_i \leq \tau_s) | \theta_i = \theta, t_{is}(\delta) = t, W_i = w, \underbrace{\psi_s(\theta, t) = p}_{score}] = p.$$

- An assignment-based framing of non-parametric RD identification
- This result presumes continuously differentiable tie-breaker distributions, but ignores outcomes

No OVB!

- Local score-conditioning kills OVB from type and non-random tie-breakers

Corollary (local conditional independence)

$$\lim_{\delta \to 0} E[1(R_i \leq \tau_s) | \theta_i = \theta, t_{is}(\delta) = t, W_i = w, \underbrace{\psi_s(\theta, t) = p}_{score}] = p.$$

- An assignment-based framing of non-parametric RD identification
- This result presumes continuously differentiable tie-breaker distributions, but ignores outcomes
- A conditional independence (CI) building block for more complex matches to come

Local Conditional Independence: Fact or Fantasy?

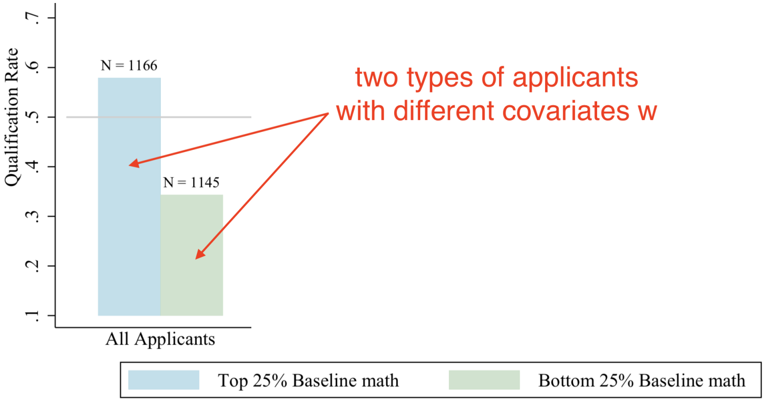
For NYC's elite Townsend Harris school, we compute

$$\Pr[R_i \leq \tau_s | t_{is}(\delta) = c, W_i = w, \psi_s(t) = 0.5]$$

Local Conditional Independence: Fact or Fantasy?

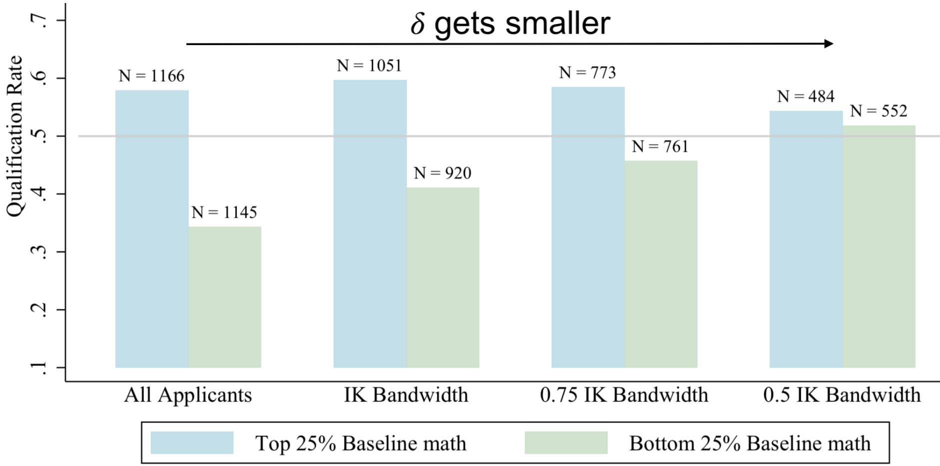
For NYC's elite Townsend Harris school, we compute

$$\Pr[R_i \leq \tau_s | t_{is}(\delta) = c, W_i = w, \psi_s(t) = 0.5]$$



Shrinking bandwidth

Qualification rates converge to 0.5 as $\delta \to 0$ in our data:



Most Informative Disqualification for Serial Dictatorship

- In a serial dictatorship, we need only be concerned with the easiest cutoff to clear at schools preferred to s
- Recall *most informative disqualification* (MID):

$$MID_{\theta s} = \begin{cases} 0 & \text{if } B_{\theta s} = \emptyset \\ \max\{\tau_b \mid b \in B_{\theta s}\} & \text{otherwise} \end{cases}$$

Anyone clearing $MID_{\theta s}$ does better than s

Most Informative Disqualification for Serial Dictatorship

- In a serial dictatorship, we need only be concerned with the easiest cutoff to clear at schools preferred to s
- Recall *most informative disqualification* (MID):

$$MID_{\theta s} = \begin{cases} 0 & \text{if } B_{\theta s} = \emptyset \\ \max\{\tau_b \mid b \in B_{\theta s}\} & \text{otherwise} \end{cases}$$

  Anyone clearing $MID_{\theta s}$ does better than s

- The *global* SD p-score is therefore the probability that

$$MID_{\theta s} < R_i \leq \tau_s,$$

  that is,

$$p_s(\theta) = E[D_i(s)|\theta_i = \theta] = \max\{0, F_R(\tau_s|\theta) - F_R(MID_{\theta s}|\theta)\},$$

  where $D_i(s)$ indicates i is offered a seat at s

Screened Serial Dictatorship Goes Local

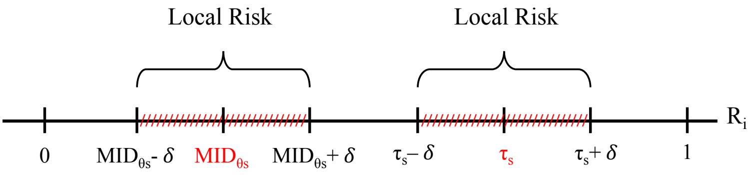
- Mission not accomplished: with non-lottery tie-breaking,
  - $F_R(\cdot|\theta)$ is an unknown and unrestricted fn of θ
  - $p_s(\theta)$ leaves OVB from R_i uncontrolled

Screened Serial Dictatorship Goes Local

- Mission not accomplished: with non-lottery tie-breaking,
  - $F_R(\cdot|\theta)$ is an unknown and unrestricted fn of θ
  - $p_s(\theta)$ leaves OVB from R_i uncontrolled
- Local risk fixes this

Screened Serial Dictatorship Goes Local

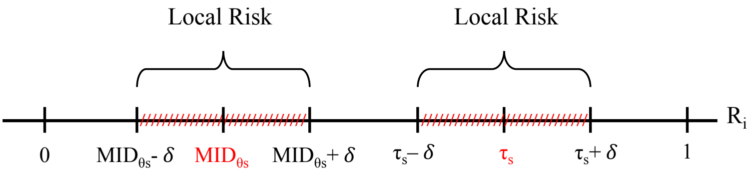
- Mission not accomplished: with non-lottery tie-breaking,
  - $F_R(\cdot|\theta)$ is an unknown and unrestricted fn of θ
  - $p_s(\theta)$ leaves OVB from R_i uncontrolled
- Local risk fixes this
- Local risk comes from tie-breakers near two cutoffs for each type



- "in-bandwidth" applicants have assignment risk of .5

Screened Serial Dictatorship Goes Local

- Mission not accomplished: with non-lottery tie-breaking,
  - $F_R(\cdot|\theta)$ is an unknown and unrestricted fn of θ
  - $p_s(\theta)$ leaves OVB from R_i uncontrolled
- Local risk fixes this
- Local risk comes from tie-breakers near two cutoffs for each type



- "in-bandwidth" applicants have assignment risk of .5
- Below $MID_{\theta s}$ and above τ_s, risk is 0; offers are certain in-between
  - Risk is non-zero only when $MID_{\theta s} < \tau_s$

Local Risk at King College Prep

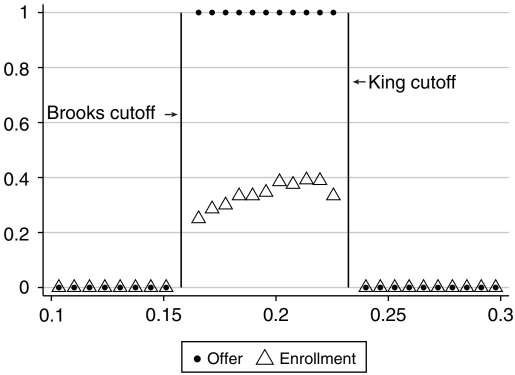


FIGURE 1. OFFERS AND ENROLLMENT AT KING

Note: Offers and enrollment for 373 applicants to King with MID given by the Brooks cutoff.

Doing DA

- Adds features (and notation!), but the building blocks of the lottery-case inform the non-lottery case
- We deploy the building blocks of the lottery-case in service of the non-lottery case
- As with the simple serial dictatorship, we identify local risk by taking limits of the tie-breaker distribution; don't need to know what that distribution is as long as its smooth

Doing DA

- Adds features (and notation!), but the building blocks of the lottery-case inform the non-lottery case
- We deploy the building blocks of the lottery-case in service of the non-lottery case
- As with the simple serial dictatorship, we identify local risk by taking limits of the tie-breaker distribution; don't need to know what that distribution is as long as its smooth
- NYC school assignment: a situation with both lottery and non lottery-tie breaking
- Centralized assignment generates many sources of variation:
  - ✓ Within school lottery-based acceptance/rejection or within school non-lottery-based acceptance/rejection
  - ✓ Higher ranked lottery-based rejection
  - ✓ Higher ranked non-lottery based rejection (as in the serial dictatorship example)

Long-Term Effects of Universal Preschool

The US Preschool Landscape

- Early childhood programs are viewed as a promising tool to improve outcomes and address socioeconomic disparities
  - Evidence that early-life deficits generate persistent negative impacts (Knudsen et al., 2006; Almond et al., 2018)
  - High estimated rates of return for historical early-childhood interventions (Heckman et al., 2010; Hendren and Sprung-Keyser, 2020)
- Recent policy efforts center on expanding public preschool
  - Public preschool in 44 states; substantial programs in 24 of 40 largest cities (NIEER, 2019)
  - Share of 4-year-olds in state-funded preschool increased from 14% in 2002 to 34% in 2019 (NIEER 2003, 2019)
  - Universal preschool for 3- and 4-year-olds included in recent $3.5 trillion budget proposal from President Biden/Senate Democrats

Preschool Evidence

- Randomized evaluations of small-scale demonstration programs
  - Perry, Abecedarian Projects: resource-intensive interventions in 1960s/70s, $N \approx 100$
  - Large impacts on short-term test scores/behavior and long-term economic outcomes (Schweinhart et al., 2011; Ramey et al., 2012; Heckman et al., 2013; Garcia et al., 2020)
- Observational studies of Head Start
  - Sibling comparisons (Currie and Thomas, 1995; Garces et al., 2002; Deming, 2009; Miller et al., 2021)
  - Discontinuities in program rules, rollout designs (Ludwig and Miller, 2007; Carneiro and Ginja, 2014; Bailey et al., 2020)
  - "Sleeper effects:" Test score impacts fade out, but effects re-emerge for adult outcomes – suggests non-cognitive channel
- Recent randomized evaluations of large-scale programs
  - Head Start Impact Study (HSIS; Puma et al., 2010), Tennessee Voluntary Prekindergarten Program (TN-VPK; Lipsey et al., 2018)
  - Immediate test score gains that fade out quickly. Ineffective programs, or sleeper effects?

Classifying Preschool Evidence

Randomized design	Long-term outcomes	Large-scale program	Examples
Yes	Yes	No	Perry, Abecedarian
No	Yes	Yes	Siblings, RDs, rollouts
Yes	No	Yes	HSIS, TN-VPK
Yes	Yes	Yes	?

Gray-Lobe, Pathak, Walters (2021)

- This paper studies the long-term impacts of Boston's universal public preschool program
- Research design leverages randomization built into Boston's school assignment mechanism
- More than 4,000 four-year-olds randomized to seats between 1997 and 2003
- Key outcomes: college attendance and college type
- Also look at medium-term impacts on test scores, grade progression, disciplinary outcomes

Public Preschool in Boston: Program Features I

- Key features of public preschool in Boston:
  - Open to all children, regardless of income
  - Programs housed in public school facilities (elementary schools, early learning centers, K-2 schools)
  - Certified teachers with BA or MA, same licensing requirements and pay scale as K-12
  - Costs roughly $13-14K for full-day (today's $), roughly half for 1/2-day
  - Slots allocated through Boston's centralized school assignment system. Oversubscribed $\Longrightarrow$ rationing by lottery
- Alternative preschool options:
  - Lack of data on other preschool options during our study period (work in progress)
  - In recent cohorts a large share of lottery losers attend other (majority private) preschools (Weiland et al., 2019)

Public Preschool in Boston: Program Features II

- We study cohorts applying 1997-2003, a period of transition for BPS preschools
  - 1997: BPS partially phased-out K1 (4-year-olds) in favor of full-day K2 (5-year-olds)
  - 1999: mix of full-day and half-day programs serving around 900 K1 students per year
  - Programs heterogeneous, quality potentially uneven (Marshall et al., 2006; Jan, 2007)
- Continued changes after our study period
  - 2005: Mayor Menino promises universal pre-K for 4-year-olds. BPS creates Department of Early Childhood, expands staffing
  - Program subsequently grows to about 2,500 students per year
  - Efforts to standarize programs and improve quality (Sachs and Weiland, 2010)
  - NIEER (2020) gives Boston 8 out of 10 benchmarks, tied for 6th out of 40 city-wide programs (unclear relevance to our study period)

Table 1: Descriptive statistics and covariate balance

	Average characteristics		Offer differentials	
	All applicants	Randomized applicants	No controls	Risk controls
	(1)	(2)	(3)	(4)
		Panel A. Applicant demographics		
Black	0.432	0.407	-0.011	-0.015
			(0.011)	(0.017)
White	0.166	0.149	-0.012	-0.023*
			(0.008)	(0.012)
Hispanic	0.291	0.344	0.036***	0.020
			(0.011)	(0.015)
Female	0.495	0.488	0.011	0.060***
			(0.011)	(0.020)
Age (enrollment year)	4.569	4.580	-0.025	-0.031
			(0.017)	(0.031)
Bilingual Spanish	0.108	0.187	0.044***	0.004
			(0.008)	(0.005)
Sample size	8786	4215	8786	4215

Table 2: Attrition

	Non-offered followup rate (1)	Offer differential (2)
Name submitted to NSC	0.987	0.008**
		(0.003)
		4215
Ever observed in SIMS	0.910	0.028***
		(0.010)
		4215
Any MCAS score	0.845	0.038***
		(0.013)
		4215
Number of MCAS scores	9.053	0.518***
		(0.184)
		4215

Table 3: Effects of preschool attendance on on-time college enrollment

	Non-offered mean (1)	First stage (2)	Reduced form (3)	2SLS estimate (4)
Any college	0.459	0.645***	0.054***	0.083***
		(0.015)	(0.019)	(0.030)
	2668		4175	

Table 3: Effects of preschool attendance on on-time college enrollment

	Non-offered mean (1)	First stage (2)	Reduced form (3)	2SLS estimate (4)
Any college	0.459	0.645***	0.054***	0.083***
		(0.015)	(0.019)	(0.030)
	2668		4175	
Two-year college	0.097		0.018	0.028
			(0.012)	(0.019)
	2668		4175	
Four-year college	0.363		0.036*	0.055*
			(0.019)	(0.029)
	2668		4175	
Massachusetts college	0.329		0.055***	0.085***
			(0.019)	(0.029)
	2668		4175	
Public college	0.260		0.024	0.038
			(0.018)	(0.027)
	2668		4175	
Private college	0.199		0.029*	0.045*
			(0.016)	(0.025)
	2668		4175	

Table 6: Effects of preschool attendance on MCAS test scores

	ELA scores		Math scores	
	Non-offered mean	2SLS	Non-offered mean	2SLS
	(1)	(2)	(3)	(4)
Grade 3	-0.424	-0.048	-0.400	0.024
		(0.068)		(0.094)
	2025	3241	677	1092
Grade 4	-0.340	-0.025	-0.302	-0.063
		(0.067)		(0.066)
	2020	3219	2022	3226
Grade 5	-0.366	0.071	-0.276	0.022
		(0.080)		(0.076)
	1316	2056	1319	2059
Grade 6	-0.311	0.027	-0.221	-0.023
		(0.072)		(0.067)
	1690	2625	1948	3113
Grade 7	-0.203	0.049	-0.180	-0.003
		(0.064)		(0.064)
	1948	3109	1950	3114
Grade 8	-0.194	-0.009	-0.157	0.024
		(0.065)		(0.063)
	1936	3087	1939	3093
Grade 10	-0.158	0.066	-0.0958	-0.031
		(0.062)		(0.064)
	1801	2852	1785	2847
All grades (stacked)	-0.283	0.029	-0.215	0.005
		(0.056)		(0.057)
Number of students	2279	3615	2249	3569
Number of scores	12736	20189	11640	18544

Table 7: Effects of preschool attendance on disciplinary outcomes

	Non-offered mean	2SLS
	(1)	(2)
Ever suspended	0.166	-0.021
		(0.023)
	2099	3335
Number of suspensions	0.663	-0.241*
		(0.141)
	2099	3335
Ever truant	0.654	0.027
		(0.029)
	2099	3335
Times truant	28.05	-4.408
		(3.557)
	2099	3335
Days absent	16.55	-1.617
		(1.262)
	2099	3335
Juvenile incarceration	0.007	-0.008*
		(0.005)
	2099	3335
Disciplinary index	0.000	0.167***
		(0.063)
	2099	3335

Wrapping Up

- Algorithms used to allocate resources generate experiments
- So far, we've focused on the class of DA-based assignment algorithms, which includes covers many matching algorithms in the field, but more remains to be done
  - ✓ TTC, Linear programming, etc
  - ✓ Dynamic assignment systems (e.g., waitlists)
- Other directions to explore.... large market sequence, randomization/permutation inference for non-lottery case, etc...
- Several ongoing applications, including estimation of single school effects or value-added models or work focused on counterfactual questions