

Aleatorização e Experimentos

IBM0288 - 2026.2

Prof. Raphael Gouvea

0.1 Para reflexão

Frontier Knowledge in College and Student Success*

Barbara Biasi[†] Song Ma[‡]

February 26, 2026

Abstract

We study whether exposure to frontier knowledge in college helps students succeed. Applying text analysis to 459,410 course syllabi and 107 million academic publications, we develop a measure of “frontier knowledge proximity” and show that course content varies substantially, even within the same institution, driven primarily by instructor heterogeneity. Linking syllabi to individual student records and exploiting course updates unobserved to students at enrollment, we find that frontier exposure helps students graduate on time, raises GPA, leads students to attend graduate school, and raises earnings. The distributional pattern, however, is asymmetric: frontier knowledge is most equalizing at the degree-completion margin, where disadvantaged students benefit most, but its returns to graduate school attendance and earnings rise with students’ pre-college resources. A mediation analysis confirms that frontier exposure operates through improved attainment for the disadvantaged, while generating additional labor-market premia for the advantaged—revealing complementarity between frontier content and pre-college resources.

0.2 Aula passada: resultados potenciais

Para cada unidade i em uma população, definimos dois resultados potenciais:

- Y_{i1} : resultado para a unidade i se ela recebesse o tratamento
- Y_{i0} : resultado para a unidade i se ela **não** recebesse o tratamento (controle)

...

Seja D_i uma variável binária que indica o status de tratamento da unidade i :

- $D_i = 1$: a unidade i recebeu o tratamento
- $D_i = 0$: a unidade i **não** recebeu o tratamento (controle)

! Resultado observado

O resultado observado para a unidade i , denotado por Y_i , depende dos resultados potenciais e do status de tratamento: $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$. Observamos Y_{i1} se $D_i = 1$ e Y_{i0} se $D_i = 0$.

0.3 Aula passada: efeitos tratamento

- O efeito causal do tratamento para a unidade i é definido como: $\tau_i = Y_{i1} - Y_{i0}$

...

! O efeito tratamento individual é não observável

Não é possível observar τ_i diretamente, pois apenas um dos resultados potenciais é observado para cada unidade.

- O efeito tratamento médio é definido como:

$$ATE = E[Y_1 - Y_0] = E[\tau_i]$$

- O efeito tratamento médio sobre os tratados é definido como:

$$ATT = E[Y_1 - Y_0 \mid D = 1]$$

- O efeito tratamento médio sobre os **não** tratados é definido como:

$$ATU = E[Y_1 - Y_0 \mid D = 0]$$

0.4 Na saúde e na doença

- Os EUA gastam uma parcela elevada do PIB em saúde, mas têm indicadores piores que outros países que possuem um sistema universal.
- O Medicare atende idosos e o Medicaid atende alguns grupos de baixa renda, mas a maioria dos trabalhadores permanece sem seguro, por escolha ou por falta de oferta a preços acessíveis.
- Em 2010, o governo Obama sancionou o Affordable Care Act (ACA) para tentar melhorar a situação.
- O ACA exige que americanos comprem plano de saúde sob pena de multa fiscal.
- Democratas e Republicanos têm visões distintas do ACA desde o início.

...

  Pergunta Empírica

Qual o impacto de ter plano de saúde sobre as condições de saúde da população?

0.5 Por que não comparar as médias de dois grupos?

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)

NHIS: *National Health Interview Survey*. Dados de 2009.

Health Index: variável categórica que vai de 1 a 5, codificada pela resposta do entrevistado à pergunta: "Você diria que seu estado geral de saúde é excelente, muito bom, bom, regular ou ruim?"

0.6 Um exemplo ilustrativo

- Dois alunos estrangeiros chegam ao MIT e têm que escolher entre pagar por plano de saúde ou não.

- Kuzdar (cazaque) vem de um país com sistema universal e tem receio de ficar muito doente com o frio de MA.
- Maria (chilena) já está acostumada com o frio e decide usar o dinheiro do plano de saúde para outras coisas.
- A tabela abaixo mostra os **resultados potenciais** de ambos e suas escolhas.

	Kuzdar	Maria
Resultado Potencial sem plano: Y_{i0}	3	5
Resultado Potencial com plano: Y_{i1}	4	5
Tratamento: D_i		
Condição de saúde observada		
Efeito tratamento		

	Kuzdar	Maria
Resultado Potencial sem plano: Y_{i0}	3	5
Resultado Potencial com plano: Y_{i1}	4	5
Tratamento: D_i	1	0
Condição de saúde observada		
Efeito tratamento		

	Kuzdar	Maria
Resultado Potencial sem plano: Y_{i0}	3	5
Resultado Potencial com plano: Y_{i1}	4	5
Tratamento: D_i	1	0
Condição de saúde observada	4	5

	Kuzdar	Maria
Efeito tratamento		
	Kuzdar	Maria
Resultado Potencial sem plano: Y_{i0}	3	5
Resultado Potencial com plano: Y_{i1}	4	5
Tratamento: D_i	1	0
Condição de saúde observada	4	5
Efeito tratamento	1	0

0.7 Visualizando o viés de seleção

Comparação entre variáveis de interesse:

$$Y_K - Y_M = Y_{K,1} - Y_{M,0}$$

...

Mostrando viés de seleção:

$$\begin{aligned} Y_K - Y_M &= Y_{K,1} - Y_{M,0} \\ &= \underbrace{Y_{K,1} - Y_{K,0}}_{=1} + \underbrace{Y_{K,0} - Y_{M,0}}_{=-2} \end{aligned}$$

$$\begin{aligned} Y_K - Y_M &= Y_{K,1} - Y_{M,0} \\ &= \underbrace{Y_{K,1} - Y_{K,0}}_{\text{efeito tratamento}=1} + \underbrace{Y_{K,0} - Y_{M,0}}_{\text{viés de seleção}=-2} \end{aligned}$$

0.8 Utilizar médias elimina o viés de seleção?

$$\begin{aligned} Avg_n[Y_{i1} - Y_{i0}] &= \frac{1}{n} \sum_{i=1}^n Y_{i1} - \frac{1}{n} \sum_{i=1}^n Y_{i0} \\ &= Avg_n[Y_i | D_i = 1] - Avg_n[Y_i | D_i = 0] \\ &= Avg_n[Y_{i1} | D_i = 1] - Avg_n[Y_{i0} | D_i = 0] \\ &\neq Avg_n[Y_{i1} - Y_{i0}] = ATE \end{aligned}$$

0.9 Hipótese: efeito causal constante e homogêneo

$$\begin{aligned} Y_{i1} &= Y_{i0} + \kappa \\ Y_{i1} - Y_{i0} &= \kappa \end{aligned}$$

...

- κ é tanto o efeito individual quanto o efeito médio!

...

- Como a comparação de médias se relaciona com κ ?

0.10 Diferença Simples de Médias

$$\begin{aligned}
 DSM &= Avg_n[Y_{i1} \mid D_i = 1] - Avg_n[Y_{i0} \mid D_i = 0] \\
 &= Avg_n[Y_{i0} + \kappa \mid D_i = 1] - Avg_n[Y_{i0} \mid D_i = 0] \\
 &= \underbrace{\kappa}_{\text{Efeito causal médio}} + \underbrace{Avg_n[Y_{i0} \mid D_i = 1] - Avg_n[Y_{i0} \mid D_i = 0]}_{\text{Viés de seleção}}
 \end{aligned}$$

! Diferença Simples de Médias

Utilizando a diferença simples de médias como estimador do efeito tratamento, obtemos como resultado **efeito causal médio** + **viés de seleção**.

- Y_{i0} é tudo aquilo que afeta a saúde da pessoa para além do status de participação no plano de saúde.

0.11 Exemplo 1: Plano de saúde nos EUA

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	-.01 (.01)	.15	.17	-.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	-.47 (.05)	3.49	3.93	-.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

- Painei B na tabela mostra que os dois grupos são distintos em várias dimensões observáveis:

- anos de estudo
- tamanho da família
- emprego
- renda
- **Pergunta:** Se o único problema fossem as variáveis observáveis, como poderíamos obter o efeito causal médio, ou seja, sem o viés de seleção?

1 Como a aleatorização elimina o viés de seleção?

1.1 Experimento aleatório

Passo a passo:

1. Obtenha uma amostra aleatória da população de interesse.
2. Divida a amostra em dois grupos, **tratamento** e **controle**, sendo que a seleção dos indivíduos em cada grupo é feita de modo aleatório.
3. Introduza o tratamento (um programa, política, estratégia, etc.) e observe os resultados.

...

💡 Lei dos Grandes Números

Explique intuitivamente o que se espera dos dois grupos gerados pela atribuição aleatória do tratamento, ou seja, como eles se comparam antes e depois do tratamento?

1.2 Efeito causal médio

! Experimento Aleatório

Como os grupos de controle e tratamento vêm da mesma população, eles são similares em **todas** as dimensões observáveis (outras variáveis da amostra) e **não observáveis** (!), incluindo o valor esperado dos **resultados potenciais**!

O que acontece no modelo de efeito constante?

...

$$\begin{aligned}
 &= E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
 &= E[Y_{i1} \mid D_i = 1] - E[Y_{i0} \mid D_i = 0] \\
 &= E[Y_{i0} + \kappa \mid D_i = 1] - E[Y_{i0} \mid D_i = 0] \\
 &= \kappa + \underbrace{E[Y_{i0} \mid D_i = 1] - E[Y_{i0} \mid D_i = 0]}_{=0} \\
 &= \kappa
 \end{aligned}$$

1.3 Estimador do ATE

- Como os grupos são comparáveis, qualquer diferença observada nos resultados após o tratamento deve ser atribuída ao tratamento.
- Estimador: a diferença na média do resultado entre o grupo de tratamento e o grupo de controle.
- Estimador não viesado para ATE.

1.4 Inferência Estatística

- A diferença observada é estatisticamente significativa ou poderia ser apenas devido ao acaso da amostragem?
- Teste t de diferença de médias para duas amostras independentes.
- Hipóteses:
 - H_0 : Não há efeito do tratamento; $ATE = 0$.
 - H_A : Há um efeito do tratamento; $ATE \neq 0$.

1.5 Estimador do ATE: regressão MQO

- Podemos estimar o ATE em um experimento por meio de uma regressão MQO: $Y_i = \beta_0 + \kappa \times D_i + \epsilon_i$

Interpretação dos coeficientes:

...

- $\beta_0 = E[Y_i \mid D_i = 0]$: a média do grupo de controle.

• $\kappa = ATE$: a diferença na média do resultado entre o grupo de tratamento e o grupo de controle.

...

- Equivalência: estimador via MQO e diferença de médias são iguais

• Diferença: MQO permite inclusão de controles

Viés e eficiência

O estimador do ATE em um experimento aleatório é não viesado. A inclusão de covariáveis não é necessária para eliminar o viés de seleção, mas pode reduzir o desvio-padrão e aumentar a eficiência do estimador.

1.6 Ameaças à validade interna

Validade Interna

Uma análise estatística tem validade interna se a inferência estatística sobre efeitos causais é válida para a população estudada.

- Falhas no processo de aleatorização: participantes não sorteados conseguem “dar um jeitinho” para entrar no grupo de tratamento

• Falhas em seguir o protocolo de tratamento: *compliance* parcial

▸ Possível focar nas pessoas que seguiram o protocolo

▸ Estima-se o efeito *intent-to-treat*: efeito da **atribuição aleatória do tratamento** e não efeito do tratamento em si.

• Atrito: se o motivo não for correlacionado com o tratamento, não gera viés

• Efeitos experimentais: solução pode ser o procedimento *duplo-cego*

• Amostras pequenas: não geram viés, mas imprecisão

1.7 Ameaças à validade externa

Validade Externa

Uma análise estatística tem validade externa se a inferência e as conclusões podem ser generalizadas da população e dos contextos estudados para outras populações e contextos.

- Amostras não representativas: população estudada e população de interesse devem ser suficientemente similares

• Intervenção não representativa: um programa de escala pequena e supermonitorado pode ser diferente do programa efetivamente implementado

• Efeitos de equilíbrio geral: um programa de escala pequena e temporário pode alterar o contexto econômico quando implementado em larga escala e de modo permanente

2 Aplicações

2.1 Aplicação: Oregon Health Plan (OHP)

- Contexto: debate sobre a expansão do Medicaid para os não segurados

- Intervenção:
 - 2008: Oregon criou uma **loteria pública** para oferecer chance de inscrição no Medicaid
 - ~75.000 inscritos → ~30.000 vencedores (tratamento) vs. ~45.000 não sorteados (controle)
- Critérios de elegibilidade: 19–64 anos, pobres, sem seguro há ≥6 meses, poucos ativos financeiros

2.2 Principais Resultados

	Oregon		Portland area	
	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
Outcome				
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Figura 1: Efeitos do OHP sobre cobertura e uso de serviços médicos.

Outcome	Oregon		Portland area	
	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	-.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	-.011 (.005)
Any medical debt?			.568	-.032 (.010)
Sample size	23,741		12,229	

Figura 2: Efeitos do OHP sobre indicadores de saúde e finanças.

2.3 Aplicação: Auditorias CGU

Ferraz e Finan: QJE 2008

- Tema: corrupção política e mecanismos de accountability.
- Questão central: quando eleitores têm acesso a informações sobre corrupção, isso afeta suas escolhas eleitorais?
- Contexto do Brasil:

- Programa federal de auditoria aleatória em municípios (CGU).
- Municípios sorteados para auditorias → publicação dos resultados na mídia.

2.4 Desenho Experimental

- Experimento ideal: auditar municípios, medir corrupção e liberar a informação para um grupo aleatório de eleitores.
- Comparação: resultados eleitorais em municípios **com informação divulgada** vs. **sem informação** → efeito causal da transparência.
- Problema: experimento direto seria **antiético e politicamente inviável**.
- Alternativa dos autores: explorar o sorteio aleatório do programa de auditoria da CGU e o **timing das eleições**.

2.5 Timing das eleições

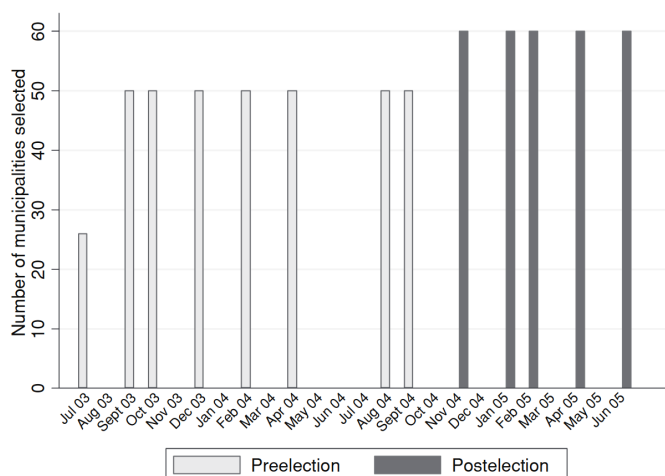


FIGURE II
Timing of the Release of the Audits

Notes. Figure shows the dates for the release of the audit reports for every municipality that was audited in the first thirteen lotteries. The lighter bars denote the 376 municipalities that were audited before the elections. The darker bars denote 300 municipalities audited after the elections. The figure was calculated based on data from the CGU.

2.6 Probabilidade de reeleição

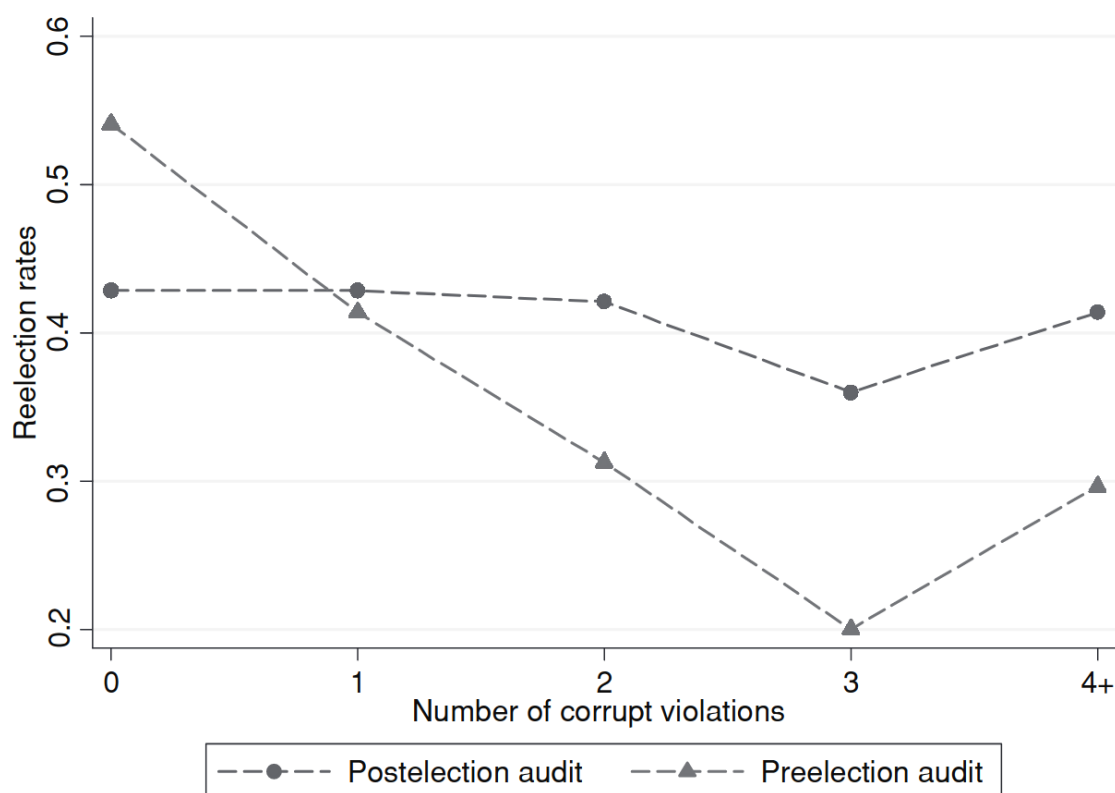


FIGURE III

Relationship between Reelection Rates and Corruption Levels

Notes. Figure shows the unadjusted relationship between the proportion of first-term mayors who were reelected in the 2004 elections and the number of corrupt incidents reported in the audit reports for municipalities audited before and after the elections. The points represented by circles are calculated for the municipalities that audited after the elections. The points represented by triangles are calculated for the municipalities audited before the elections. The figure was calculated for our entire sample of 373 municipalities based on data from Brazil's Electoral Commission and the CGU audit reports.

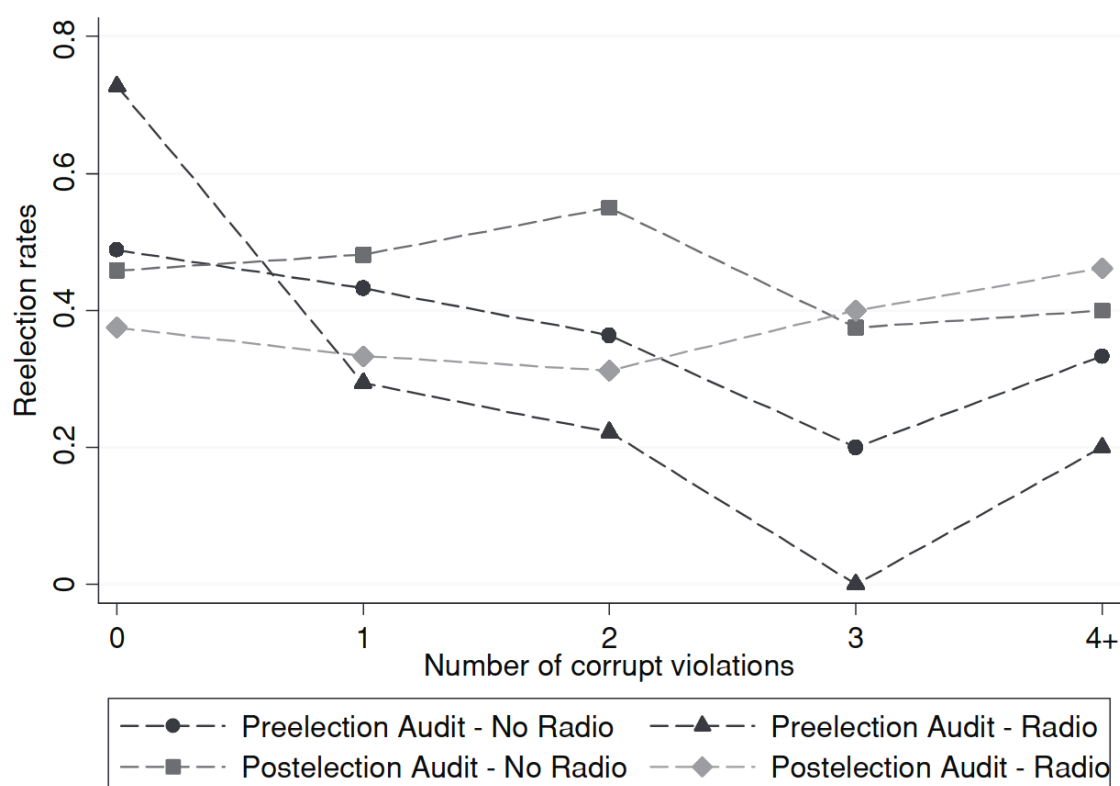


FIGURE IV

Relationship between Reelection Rates and Corruption Levels

Notes. Figure shows the unadjusted relationship between the proportion of first-term mayors who were reelected in the 2004 elections and the number of corrupt incidents reported in the audit reports for municipalities audited before and after the elections and the existence of local radio. The points represented by circles are calculated for the municipalities that were audited before the elections and do not have a local AM radio station. The points represented by triangles are calculated for the municipalities that were audited before the elections and have a local AM radio station. The points represented by squares are calculated for the municipalities that were audited after the elections and do not have a local AM radio station. The points represented by diamonds are calculated for the municipalities that were audited after the elections and had a local AM radio station. The figure was calculated for our entire sample of 373 municipalities based on data from Brazil's Electoral Commission, the CGU audit reports, and IBGE.

2.7 Resultados Principais

- Prefeitos candidatos à reeleição **acusados de corrupção** tiveram **redução** na probabilidade de reeleição.
- Impacto **maior quando a mídia local (rádio)** amplificava os resultados.
- Efeito **mais forte nos casos de corrupção mais graves**.
- Informação pública → accountability eleitoral efetiva.

2.8 Aplicação em Finanças

Choi, Laibson e Madrian: RFS 2010

- Tema: racionalidade de investidores.

- Questão central: por que investidores não escolhem sempre o fundo mais barato?
- Contexto: fundos de índice S&P 500 oferecem praticamente o mesmo retorno, mas cobram **taxas (fees) diferentes**.
- Hipótese existente: taxas refletem heterogeneidade em outros serviços oferecidos em conjunto com os fundos.
- Estratégia dos autores: conduzir **experimentos controlados** em um contexto em que esses outros serviços são inexistentes.

2.9 Participantes e Tarefa

- **Participantes:**

- Estudantes de MBA (Wharton, 2005)
- Estudantes universitários (Harvard, 2005)
- Funcionários de Harvard (staff, 2007)
- **Tarefa principal:** alocar US\$ 10.000 hipotéticos entre **4 fundos reais do S&P 500**.

• **Incentivos:**

- Pagamento inicial modesto + pagamento adicional baseado no **retorno futuro do portfólio escolhido**.
- Para o staff, horizonte de retorno de **1 mês**; para os estudantes, **1 ano**.

• **Documentos fornecidos:** lâmina dos fundos e “choice sheet” para registrar as alocações.

• **Objetivo:** observar se participantes escolhem **fundos de menor taxa (fee)** quando outras características são idênticas.

2.10 Condições de Informação

- Aleatorização em 4 condições principais:

+ **Prospectus-only:** apenas lâmina + choice sheet.

1. **Fees treatment:** lâmina + choice sheet + planilha mostrando **fees detalhados e impacto no portfólio**.
2. **Returns treatment:** lâmina + choice sheet + planilha com **retornos passados anualizados**.
3. **FAQ treatment (staff only):** lâmina + choice sheet + respostas a perguntas básicas sobre fundos e S&P 500.

. . .

- Objetivo do desenho: testar como **informação sobre custos, retornos passados ou conhecimento geral** afeta a alocação dos participantes.

- Controle experimental: todos os grupos tiveram tempo livre para decidir, sem poder consultar outros participantes; os mínimos de investimento reais foram respeitados.

2.11 Principais Resultados

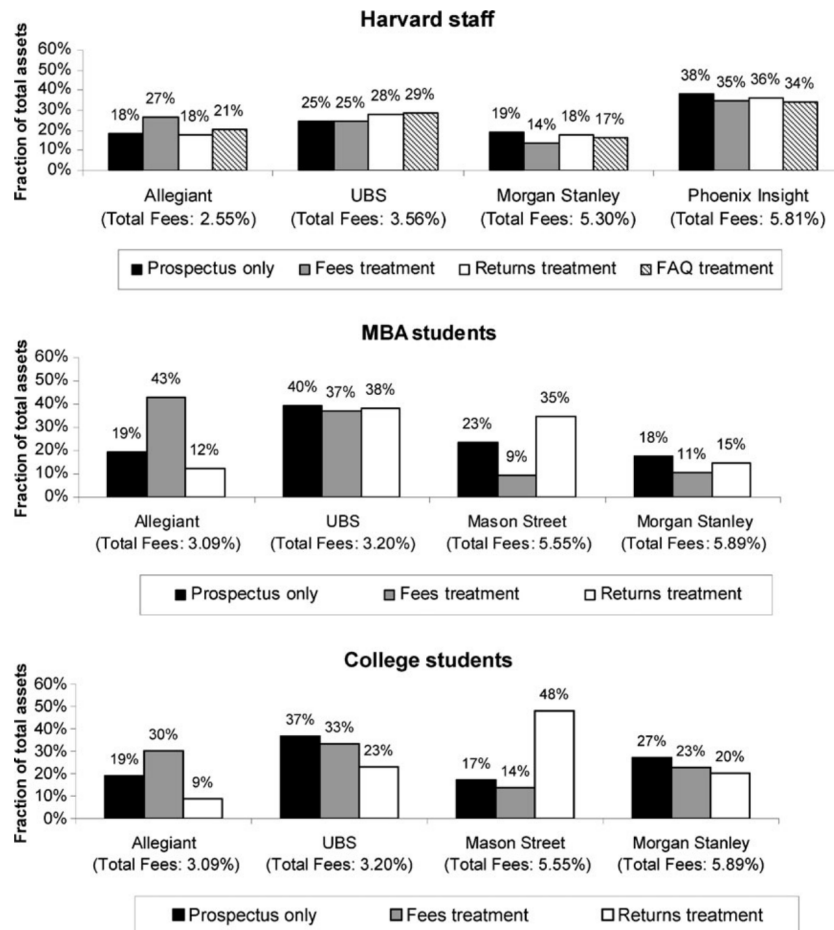


Figure 1
Average fund allocations by subject type and experimental condition

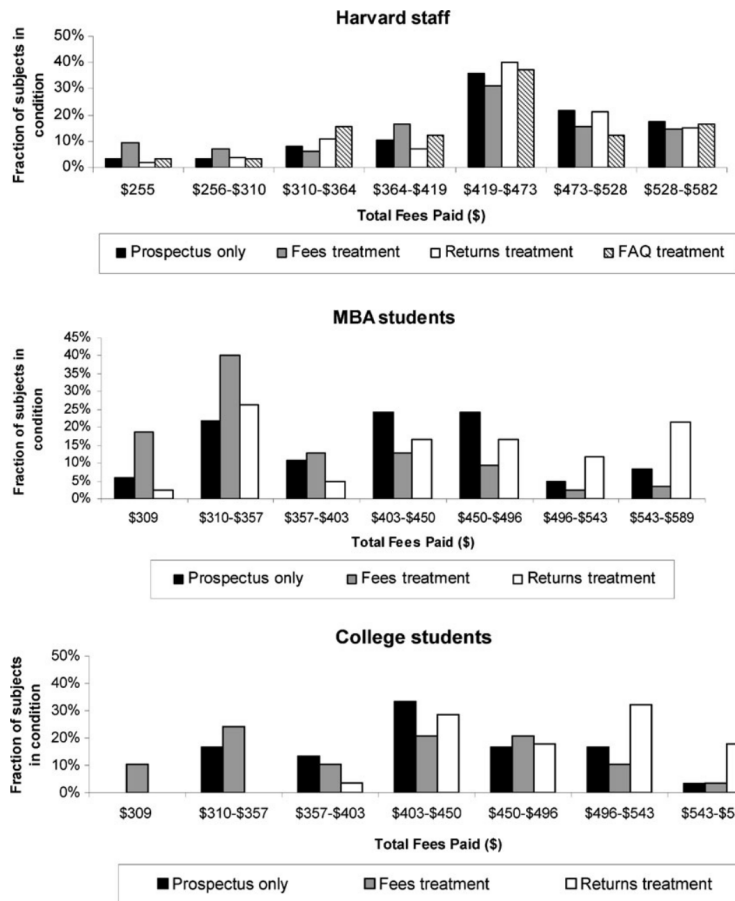


Figure 2
Histogram of fees paid by subject type and experimental condition

2.12 Resultados principais

- Participantes alocam US\$ 10.000 entre quatro fundos S&P 500 e são recompensados pelo retorno subsequente
- Os participantes falham em **minimizar taxas**
- Rejeitamos a hipótese de escolha por **serviços extras não ligados ao portfólio**
- Custos de busca por informações importam, mas, mesmo eliminados, as taxas não são minimizadas
- Participantes dão peso elevado a **retornos anualizados desde a criação do fundo**
- **Taxas pagas diminuem com conhecimento financeiro**
- Participantes que escolhem fundos caros **percebem que estão cometendo um erro**