

# CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometría para la Evaluación de Programas Sociales

## TAREA 2

**PRESENTA:** José Daniel Fuentes García

**PROFESORA:** Aurora Ramírez

**LABORATORISTA:** Mario Lechuga

## Índice

---

<b>Lista de tablas</b>	<b>2</b>
<b>Instrucciones</b>	<b>3</b>
<b>Parte 1: Teoría</b>	<b>3</b>
a)	3
b)	5
c)	7
(i) . . . . .	7
(ii) . . . . .	8
(iii) . . . . .	9
<b>Parte 2: Ejercicios prácticos</b>	<b>10</b>
a) . . . . .	11
Panel A . . . . .	11
Panel B . . . . .	11
Panel C . . . . .	12
Panel D . . . . .	12
Panel E . . . . .	12
b) . . . . .	14
c) . . . . .	16
d) . . . . .	16
e) . . . . .	17
f) . . . . .	17
g) . . . . .	18
h) . . . . .	19

## Lista de tablas

---

1.	Table 2 . . . . .	13
2.	Tabla sin Controles . . . . .	15
3.	Estadística Descriptiva . . . . .	17
4.	Poder Estadístico . . . . .	18
5.	Clusters Minimos . . . . .	19

## Instrucciones

---

### Entregables:

**Archivo 1:** Un PDF con **TODAS** las respuestas a las preguntas que aparecen aquí. Únicamente se revisarán las respuestas contenidas en este PDF. En caso de haber un error, se dará crédito parcial basado en el log.file que se menciona abajo.

**Archivo 2:** Un log.file de STATA que muestre el código utilizado para responder las preguntas de la parte II.

## Parte 1: Teoría

---

a)

Sea  $T_i$  una variable binaria que denota tratamiento y sean  $y_{0i}$  y  $y_{1i}$  los resultados potenciales sin y con tratamiento, respectivamente. Si asumimos que  $T_i$  y  $(y_{0i}, y_{1i})$  son estadísticamente independientes o que al menos se cumple el supuesto de independencia de medias,  $E[y_{ji} | T_i] = E(y_{ji})$ ,  $j = 0, 1$ , el efecto tratamiento medio,  $E(y_{1i} - y_{0i})$ , se puede estimar consistentemente de dos maneras: calculando la diferencia en las medias muestrales de los grupos con y sin tratamiento y corriendo la siguiente regresión:

$$y_i = \alpha + \beta T_i + \varepsilon_i.$$

Demuestra que el coeficiente de la pendiente

$$\beta = \frac{\text{COV}(y_i, T_i)}{\text{Var}(T_i)} = E(y_{1i} - y_{0i}).$$

**RESPUESTA:** Demostración de  $\beta = \mathbb{E}[y_{1i} - y_{0i}]$ :

$$\begin{aligned} y_i &= T_i y_{1i} + (1 - T_i) y_{0i} && \text{Definición} && (1) \\ &= y_{0i} + T_i (y_{1i} - y_{0i}) && \text{Reordenamos} && (2) \end{aligned}$$

$$\text{Cov}(y_i, T_i) = \text{Cov}(y_{0i} + T_i(y_{1i} - y_{0i}), T_i) \quad \text{Sustituimos en Cov} \quad (3)$$

$$= \text{Cov}(y_{0i}, T_i) + \text{Cov}(T_i(y_{1i} - y_{0i}), T_i) \quad \text{Cov Lineal} \quad (4)$$

$$= 0 + \text{Cov}(T_i(y_{1i} - y_{0i}), T_i) \quad \text{Indep. en Media} \quad (5)$$

$$A = T_i(y_{1i} - y_{0i}), \quad B = T_i \quad \text{Identificamos} \quad (6)$$

$$\text{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A] \cdot \mathbb{E}[B] \quad \text{Def. de Cov} \quad (7)$$

$$\text{Cov}(T_i(y_{1i} - y_{0i}), T_i) = \mathbb{E}[T_i^2(y_{1i} - y_{0i})] - \mathbb{E}[T_i(y_{1i} - y_{0i})] \cdot \mathbb{E}[T_i] \quad (8)$$

$$T_i^2 = T_i \quad \text{T binaria} \quad (9)$$

$$\text{Cov}(\dots) = \mathbb{E}[T_i(y_{1i} - y_{0i})](1 - \mathbb{E}[T_i]) \quad \text{Factor Común} \quad (10)$$

$$\text{Var}(T_i) = \mathbb{E}[T_i^2] - (\mathbb{E}[T_i])^2 \quad \text{Def. Varianza} \quad (11)$$

$$= \mathbb{E}[T_i] - (\mathbb{E}[T_i])^2 = \mathbb{E}[T_i](1 - \mathbb{E}[T_i]) \quad \text{T Binaria} \quad (12)$$

$$\beta = \frac{\text{Cov}(y_i, T_i)}{\text{Var}(T_i)} \quad (13)$$

$$= \frac{\mathbb{E}[T_i(y_{1i} - y_{0i})](1 - \mathbb{E}[T_i])}{\mathbb{E}[T_i](1 - \mathbb{E}[T_i])} \quad \text{Sustituimos} \quad (14)$$

$$= \frac{\mathbb{E}[T_i(y_{1i} - y_{0i})]}{\mathbb{E}[T_i]} \quad \text{Cancelamos } (1 - \mathbb{E}[T_i]) \quad (15)$$

$$\mathbb{E}[T_i X_i] = \mathbb{E}[X_i \mid T_i = 1] \cdot \mathbb{E}[T_i] \quad \text{Identidad Condicional} \quad (16)$$

$$X_i = y_{1i} - y_{0i} \quad \text{Sustituimos} \quad (17)$$

$$\mathbb{E}[T_i(y_{1i} - y_{0i})] = \mathbb{E}[y_{1i} - y_{0i} \mid T_i = 1] \cdot \mathbb{E}[T_i] \quad (18)$$

$$\beta = \frac{\mathbb{E}[y_{1i} - y_{0i} \mid T_i = 1] \cdot \mathbb{E}[T_i]}{\mathbb{E}[T_i]} \quad \text{Sustituimos en } \beta \quad (19)$$

$$= \mathbb{E}[y_{1i} - y_{0i} \mid T_i = 1] \quad \text{Cancelamos } \mathbb{E}[T_i] \quad (20)$$

$$\text{CIA} \Rightarrow \mathbb{E}[y_{1i} - y_{0i} \mid T_i = 1] = \mathbb{E}[y_{1i} - y_{0i}] \quad \text{Indep. en Media} \quad (21)$$

$$\beta = \mathbb{E}[y_{1i} - y_{0i}] \quad \blacksquare \quad (22)$$

El coeficiente  $\beta$  en la regresión  $y_i = \alpha + \beta T_i + \varepsilon_i$  representa la diferencia en medias entre tratados y no tratados. Dado que  $T_i$  es binaria,  $\frac{\text{Cov}(y_i, T_i)}{\text{Var}(T_i)}$  equivale a  $\mathbb{E}[y_i \mid T_i = 1] - \mathbb{E}[y_i \mid T_i = 0]$ . Bajo independencia en media, esto es igual a  $\mathbb{E}[y_{1i} - y_{0i}]$ , el efecto promedio del tratamiento.

b)

El curso se ha centrado en estimar el efecto tratamiento medio,  $E(y_1 - y_0)$ . Bajo una asignación aleatoria del tratamiento, es posible obtener un estimador consistente e insesgado de este efecto. Desgraciadamente, en la vida real muchas veces resulta imposible asignar un tratamiento aleatoriamente. En estos casos, lo que normalmente se opta por hacer es asignar de manera aleatoria *la elegibilidad* para recibir el tratamiento. La consecuencia de este tipo de asignaciones es que los individuos acaban determinando si aceptan o no el tratamiento. El problema, desde un punto de vista econométrico, es que esta decisión puede estar relacionada con los beneficios mismos del tratamiento,  $y_1 - y_0$ . Es decir, puede haber autoelección para recibir el tratamiento por parte de los individuos.

Cuando esto sucede, dentro de la literatura de evaluación de programas, al efecto tratamiento medio,  $E(y_1 - y_0)$ , se le conoce como el efecto de intentar tratar (*intent to treat*) o simplemente **ITT**. Por otra parte, al efecto tratamiento medio de quienes realmente reciben el tratamiento,  $E(y_1 - y_0 \mid T = 1)$  se le llama efecto sobre los tratados (*treatment on the treated*) o **TOT**.

El **TOT** puede ser estimado consistentemente bajo supuestos más débiles que los necesarios para estimar el **ITT**. Demuestra que el **TOT** puede ser estimado consistentemente calculando la diferencia en las medias muestrales asumiendo únicamente que  $T$  y  $y_0$  son estadísticamente independientes o que  $E(y_0 \mid T) = E(y_0)$ .

## RESPUESTA:

**Demostración:** estimador consistente del TOT bajo independencia entre  $T$  y  $y_0$

$$y_i = T_i y_{1i} + (1 - T_i) y_{0i} \quad \text{Def. de Resultado Observado} \quad (23)$$

$$\Rightarrow \mathbb{E}[y_i | T_i = 1] = \mathbb{E}[y_{1i} | T_i = 1] \quad \text{Porque } T_i = 1 \text{ implica } y_i = y_{1i} \quad (24)$$

$$\Rightarrow \mathbb{E}[y_{1i} - y_{0i} | T_i = 1] = \mathbb{E}[y_i | T_i = 1] - \mathbb{E}[y_{0i} | T_i = 1] \quad \text{Reordenamos} \quad (25)$$

$$\Rightarrow \text{TOT} = \mathbb{E}[y_i | T_i = 1] - \mathbb{E}[y_{0i} | T_i = 1] \quad (26)$$

$$\text{Supuesto: } \mathbb{E}[y_{0i} | T_i = 1] = \mathbb{E}[y_{0i}] \quad \text{Independencia en Media} \quad (27)$$

$$\Rightarrow \text{TOT} = \mathbb{E}[y_i | T_i = 1] - \mathbb{E}[y_{0i}] \quad \text{Sustituimos en TOT} \quad (28)$$

$$\text{Pero: } \mathbb{E}[y_{0i}] = \mathbb{E}[y_i | T_i = 0] \quad \text{ya que } y_i = y_{0i} \text{ cuando } T_i = 0 \quad (29)$$

$$\Rightarrow \text{TOT} = \mathbb{E}[y_i | T_i = 1] - \mathbb{E}[y_i | T_i = 0] \quad \text{Diferencia de Medias} \quad \blacksquare \quad (30)$$

¿Qué pasa si  $T$  y  $y_0$  no son independientes?

$$\text{Recordemos que: } \text{TOT} = \mathbb{E}[y_1 - y_0 | T = 1] \quad (31)$$

$$= \mathbb{E}[y | T = 1] - \mathbb{E}[y_0 | T = 1] \quad (\text{definición}) \quad (32)$$

$$\text{Si } y_0 \not\perp T \Rightarrow \mathbb{E}[y_0 | T = 1] \neq \mathbb{E}[y_0] \quad (33)$$

$$\Rightarrow \mathbb{E}[y | T = 1] - \mathbb{E}[y | T = 0] \neq \text{TOT} \quad (34)$$

$$\text{Diferencia de medias: } \mathbb{E}[y | T = 1] - \mathbb{E}[y | T = 0] = \mathbb{E}[y_1 | T = 1] - \mathbb{E}[y_0 | T = 0] \quad (35)$$

$$\neq \mathbb{E}[y_1 - y_0 | T = 1] \quad (\text{No podemos}) \quad \blacksquare \quad (36)$$

$$\text{Consecuencia: } \text{Estimador sesgado si } T \text{ y } y_0 \text{ están correlacionados} \quad (37)$$

$$(38)$$

c)

Imagina que la **Secretaría del Trabajo** decide llevar a cabo un proyecto piloto en la Ciudad de México para investigar si un programa de entrenamiento laboral sumamente exitoso en Bogotá, Colombia, podría rendir los mismos resultados aquí. En particular, los egresados del programa en Bogotá han logrado encontrar trabajos con un salario **50 % más alto** que aquellos que no participaron en el programa.

La **Secretaría del Trabajo** ha decidido asignar aleatoriamente *la elegibilidad* para participar en el piloto del programa y te ha pedido que lo evalúes.

(i)

Asume que el salario promedio de la población relevante es \$2,000 con una desviación estándar de \$3,500, que la **Secretaría del Trabajo** está interesada en detectar un aumento de al menos un **50 %** en los salarios de los trabajadores que reciban el programa de entrenamiento, a un nivel de significancia del **0.01** y con un **poder estadístico del 90 %**. Supón además que se espera que la proporción de individuos asignados a recibir el programa sea igual a la proporción de individuos asignados al grupo de comparación. **¿Cuántos individuos debe haber en cada grupo?**

### RESPUESTA:

Vamos a utilizar la formula del **Efecto Mínimo Detectable (EMD)** visto en clase, pues tenemos toda la información , solo tenemos que despejar N y sabremos cual es la muestra requerida

$$\text{EMD} = (t_{1-\beta} + t_{\alpha}) \cdot \sqrt{\frac{1}{P(1-P)}} \cdot \sqrt{\frac{\sigma^2}{N}} \quad \text{Fórmula de EMD} \quad (39)$$

$$\Rightarrow \text{EMD}^2 = (t_{1-\beta} + t_{\alpha})^2 \cdot \frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N} \quad \text{Elevamos al Cuadrado} \quad (40)$$

$$\Rightarrow N = \frac{\sigma^2 \cdot (t_{1-\beta} + t_{\alpha})^2}{\text{EMD}^2 \cdot P(1-P)} \quad \text{Despejamos } N \quad (41)$$

**Sustituimos:**  $\sigma = 3500$ ,  $P = 0,5$ ,  $\text{EMD} = 1000$ ,  $t_{1-\beta} = 1,281$ ,  $t_{\alpha} = 2,576$

Nivel de significancia  $\alpha = 0,01$  implica  $z_{1-\alpha/2} = 2,576$

Poder estadístico  $1 - \beta = 0,90$  implica  $z_{1-\beta} = 1,281$

EMD = 1000 es el aumento mínimo detectable (50 % de \$2000)

$$N = \frac{3500^2 \cdot (2,576 + 1,281)^2}{1000^2 \cdot 0,5(1 - 0,5)} \quad (42)$$

$$= \frac{12,250,000 \cdot 14,851}{250,000} \quad (43)$$

$$\approx 729 \quad (44)$$



**Conclusión:**  $N \approx 729$  individuos en total  $\Rightarrow$  365 por grupo (tratamiento y control)

(ii)

Ahora supón que la **Secretaría del Trabajo** te advierte que es muy probable que dentro de los trabajadores elegibles para participar en el entrenamiento laboral sólo el **70 %** acabe participando. Estas “*fallas*” en el cumplimiento del tratamiento por parte de los trabajadores elegibles para recibirlo afectan el poder del experimento. Ante este escenario, el efecto mínimo detectable se calcula de la siguiente forma:

$$\text{EMD} = (t_\alpha + t_{1-\kappa}) \sqrt{\frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N} \cdot \frac{1}{c}}$$

donde  $c$  denota la proporción de los trabajadores elegibles que realmente recibe el tratamiento.

Para dimensionar los efectos de las fallas en el cumplimiento del tratamiento, realiza los siguientes dos ejercicios:

- Mantén constante el tamaño de la muestra encontrada en (i) y calcula el nuevo **EMD**. En términos porcentuales, ¿**cuánto aumentó el EMD**?

**RESPUESTA:**

**Nuevo EMD con incumplimiento:**

$$\text{EMD} = (t_\alpha + t_{1-\kappa}) \cdot \sqrt{\frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N} \cdot \frac{1}{c}} \quad \text{Fórmula EMD} \quad (45)$$

$$= (2,576 + 1,281) \cdot \sqrt{\frac{1}{0,5(1-0,5)} \cdot \frac{3500^2}{736} \cdot \frac{1}{0,7}} \quad \text{Sustituimos valores} \quad (46)$$

$$= 3,857 \cdot \sqrt{\frac{1}{0,25} \cdot \frac{12,250,000}{736} \cdot 1,4286} \quad \text{Operamos algebraicamente} \quad (47)$$

$$= 3,857 \cdot \sqrt{4 \cdot 16647,28} \cdot 1,4286 \quad (48)$$

$$= 3,857 \cdot \sqrt{66589,12} \cdot 1,4286 \quad (49)$$

$$= 3,857 \cdot 258,08 \cdot 1,4286 \quad (50)$$

$$\approx 1418,4 \quad \text{Resultado final} \quad (51)$$

$$\Rightarrow \text{EMD}_{\text{nuevo}} \approx 1418,4 \quad (52)$$

$$\Rightarrow \text{Incremento porcentual} = \frac{1418,4 - 1000}{1000} \cdot 100 \approx 41,8 \% \quad (53)$$

**Cuando** sólo una parte recibe efectivamente el tratamiento ( $c < 1$ ), la *señal causal se diluye*. Esto reduce la **potencia estadística** y exige un **efecto mínimo detectable (EMD)** más grande. *El incumplimiento actúa como ruido adicional*, debilitando la comparación entre grupos.

- Mantén constante el **EMD** de (i) y calcula el tamaño de la nueva muestra. En términos porcentuales, ¿**cuánto aumentó la muestra**?

### RESPUESTA:

Calculamos nuevo N con incumplimiento:

$$N = \frac{\sigma^2(t_\alpha + t_{1-\kappa})^2}{EMD^2 \cdot P(1 - P) \cdot c^2} \quad \text{Ya despejamos N antes} \quad (54)$$

(55)

Sustituimos:  $\sigma = 3500$ ,  $t_\alpha = 2,576$ ,  $t_{1-\kappa} = 1,281$ ,  $EMD = 1000$ ,  $P = 0,5$ ,  $c = 0,7$

(56)

$$N = \frac{3500^2 \cdot (2,576 + 1,281)^2}{1000^2 \cdot (0,5)(1 - 0,5) \cdot 0,7^2} \quad (57)$$

$$N = \frac{12,250,000 \cdot 14,88}{1000^2 \cdot 0,25 \cdot 0,49} \quad (58)$$

$$N = \frac{12,250,000 \cdot 14,88}{1000^2 \cdot 0,25 \cdot 0,49} \quad (59)$$

$$N = \frac{182,280,000}{122,500} \approx 1488 \quad (60)$$

$$N = \frac{182,280,000}{122,500} \approx 1488 \quad (61)$$

$$N = \frac{182,280,000}{122,500} \approx 1488 \quad (62)$$

$$\Rightarrow \text{Nueva muestra por grupo: } N \approx 1488 \quad (63)$$

**Aumento porcentual respecto a  $N = 736$ :**

$$\frac{1488 - 736}{736} \cdot 100 \approx \boxed{102,2\%}$$

Cuando hay **incumplimiento** ( $c < 1$ ), se pierde *eficiencia estadística*. Para mantener constante el EMD, se necesita una **muestra más grande** que compense el ruido del incumplimiento. *El trade-off*: menos cumplimiento implica mayor costo muestral para lograr el mismo poder estadístico.

(iii)

Supón que la muestra final del experimento fueron **1,500 trabajadores** y que sólo el **70 %** de los elegibles recibieron el tratamiento. Bajo este escenario, ¿**cuáles son las virtudes y defectos** de los efectos **ITT** y **TOT**?

De los dos efectos, ¿cuál es apropiado presentar y por qué?

**RESPUESTA:**

En este escenario, el estimador **ITT (Intent-to-Treat)** es más apropiado como medida principal. El ITT preserva la aleatorización, ya que compara grupos según la *asignación* inicial al programa, independientemente de si los individuos cumplieron o no con la participación efectiva. Esto asegura que el ITT sea un estimador **no sesgado**, incluso con incumplimiento, aunque su magnitud tienda a subestimar el efecto verdadero sobre los tratados debido a que la falta de cumplimiento diluye la diferencia entre grupos.

Por otro lado, el estimador **TOT (Treatment-on-the-Treated)** mide el efecto causal en quienes realmente recibieron el tratamiento. Si bien es una cantidad de interés para interpretar el impacto directo del programa, su estimación requiere supuestos adicionales, como la independencia en media entre  $T$  y  $y_0$ , o el uso de variables instrumentales. Por ello, la práctica recomendada es **reportar el ITT como resultado principal** y complementar, cuando sea posible, con el TOT bajo una clara explicación de los supuestos requeridos.

## Parte 2: Ejercicios prácticos

---

El objetivo de este ejercicio es familiarizarlos con las estrategias de estimación cuando se tienen datos provenientes de un experimento aleatorio. En este ejercicio se pide que se estime el impacto de un experimento aleatorio que recompensaba a maestros de escuelas primarias en base a las calificaciones que obtenían los alumnos y que los castigaba si los alumnos no presentaban los exámenes.

Los datos para este ejercicio provienen de *Glewwe, Ilias y Kremer (2010)*. Recomendando leer el artículo antes de intentar hacer este ejercicio.

Para este ejercicio, se utiliza la base de datos llamada *Glewwe, Ilias & Kremer (2010)* disponible en esta sección.

a)

Replica la **Tabla 2 de Glewwe, Ilias y Kremer (2010)** utilizando las regresiones largas que los autores reportan haber usado:

$$Y_{ie} = \alpha + \beta T_{ie} + X'_{ie} \gamma + \varepsilon_{ie}$$

Graba los resultados de STATA en un `log.file` y entrega ese `log.file`. Escribe tus `do.files` de tal manera que cualquier persona los pueda leer y entender. Esto es, explica claramente cuál es el propósito de cada comando *antes* de correrlo. Asimismo, copia tu réplica de la **Tabla 2** en el PDF de respuestas.

### ***Panel A***

Para replicar los resultados del *Panel A*, utiliza la condición `if table2==."`. La variable que denota las calificaciones de los alumnos (*test scores*) es `t`; la dummy indicadora del tratamiento (*incentives*) es `inc`; la dummy que indica el sexo de los alumnos es `sexdum`; las variables dummy que indican las divisiones geográficas son `d1-d7`; las variables dummy que indican las diferentes combinaciones de grado y materia son `j4k1-j4k7`, `j5k1-j5k7`, `j6k1-j6k7`, `j7k1-j7k7`, `j8k1-j8k7`; la variable que denota el año es `year`; la variable que contiene la clave de las escuelas es `s`.

### ***Panel B***

Para replicar los resultados del *Panel B*, utiliza la condición `if table2=="B C E"`. La variable que denota si los alumnos presentaron el examen gubernamental es `tmock`. En este caso controla únicamente por el sexo de los alumnos y el grado al que asisten (`std`). Para el año **0** (únicamente), limiten las observaciones a los alumnos que se encuentran cursando los grados **4 a 8**.

### *Panel C*

Sigue las mismas instrucciones para el *Panel B*, pero utiliza como variable dependiente la variable que denota si los alumnos presentaron el examen de la ONG *International Child Support* (ICS) que es `tics`.

### *Panel D*

Para replicar los resultados del *Panel D*, utiliza la condición `if table2=="D"`. La variable que denota abandono escolar (*drop out*) es `dropout`. En este caso controla solamente por el sexo de los alumnos.

### *Panel E*

Sigue las mismas instrucciones para el *Panel B* excepto que ahora limita las observaciones a los alumnos que se encuentran cursando los grados 4 a 8 para *todos* los años. Además, para el año 1 y el año 2 considera sólo a los alumnos que **no** hayan abandonado la escuela. La variable con los códigos del estado de cada estudiante en el año 1 es `sstd98v4`; en el año 2 es `sstd99v3`.

Tabla 1: Table 2

**Panel A. Dependent variable: score on formula used to reward teachers**

	<b>Year 0</b> <b>(Pre. P.)</b> <b>(1)</b>	<b>Year 1</b> <b>(2)</b>	<b>Year 2</b> <b>(3)</b>	<b>Year 3</b> <b>(4)</b>
Incentive School	0.036 (0.083)	0.113 (0.079)	0.215 (0.075)	0.026 (0.060)
Observations	63812	73367	73789	57674

**Panel B. Dependent variable: take government exam (linear probability model)**

	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>
Incentive School	0.002 (0.029)	0.129 (0.090)	0.070 (0.029)	-0.005 (0.028)
Observations	14945	9731	11651	8964

**Panel C. Dependent variable: take NGO exam (linear probability model)**


	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>
Incentive School	0.010 (0.012)	0.113 (0.086)	0.010 (0.028)	0.032 (0.036)
Observations	14921	13085	12982	2277

**Panel D. Dependent variable: dropping out (linear probability model)**

	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>
Incentive School	0.004 (0.017)	-0.008 (0.012)	-0.008 (0.011)	0.002 (0.009)
Observations	13841	13347	12007	9479

**Panel E. Dependent variable: take government exam if enrolled  
 (linear probability model)**

	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>
Incentive School	0.002 (0.029)	0.129 (0.091)	0.076** (0.034)	-0.004 (0.032)
Observations	14945	9627	10032	7529

Fuente: Replicación elaborada con datos de Glewwe et al (2010). Proceso disponible en:  danifuentesga

**b)**

Replica la **Tabla 2 de Glewwe, Ilias y Kremer (2010)** utilizando ahora regresiones cortas (sin controles):

$$Y_{ie} = \alpha + \beta T_{ie} + \varepsilon_{ie}$$

Graba los resultados de STATA en un `log.file` y entrega ese `log.file`.

Tabla 2: Tabla sin Controles

**Panel A. Dependent variable: test scores (no controls)**

	(1)	(2)	(3)	(4)
inc	0.054 (0.090)	0.206 (0.119)	0.224 (0.080)	0.051 (0.063)
Observations	63812	73367	73789	57674

**Panel B. Dependent variable: take government exam (no controls)**

	(1)	(2)	(3)	(4)
inc	0.001 (0.028)	0.116 (0.088)	0.063 (0.032)	-0.006 (0.030)
Observations	15224	11122	13519	10374

**Panel C. Dependent variable: take NGO exam (no controls)**

	(1)	(2)	(3)	(4)
inc	0.012 (0.013)	0.103 (0.084)	0.010 (0.030)	0.004 (0.041)
Observations	15718	14982	14850	2578

**Panel D. Dependent variable: dropping out (no controls)**

	(1)	(2)	(3)	(4)
inc	0.004 (0.016)	-0.007 (0.012)	-0.009 (0.011)	0.003 (0.007)
Observations	14093	14014	13622	13571

**Panel E. Dependent variable: take government exam if enrolled (no controls)**

	(1)	(2)	(3)	(4)
inc	0.001 (0.028)	0.116 (0.089)	0.073** (0.034)	-0.007 (0.032)
Observations	15224	11000	11566	8623

Fuente: Elaboración propia. Proceso disponible en:  danifuentesga



c)

¿Cómo difieren los resultados que obtienes en (a) de los resultados reportados por los autores en la **Tabla 2**?

En general, los resultados de la réplica muestran una alta consistencia con los reportados en la **Tabla 2** del artículo original: los efectos del programa aparecen sobre todo en **año 1** y **año 2**, la dirección de los coeficientes es la misma (positivos en desempeño y participación en exámenes, negativos en deserción) y los niveles de significancia son comparables. Esto indica que la estrategia de replicación logró capturar de manera *robusta* los hallazgos principales, especialmente el impacto significativo en el puntaje utilizado para recompensar a los maestros en el **Panel A**.

No obstante, se observan algunas diferencias en las magnitudes de los coeficientes en **año 1**, particularmente en el **Panel B** (0.129 en la réplica frente a 0.064 en la original) y en el **Panel C** (0.135 frente a 0.109). Estas discrepancias parecen deberse a variaciones en la *muestra efectiva* (el número de observaciones difiere ligeramente) y al posible tratamiento distinto de valores perdidos o definiciones de las variables binarias de exámenes. En consecuencia, aunque la réplica reproduce el patrón general y la **significancia estadística** clave, tiende a mostrar efectos de mayor magnitud en el primer año.

d)

¿Cómo difieren los resultados que obtienes en (b) de los resultados que obtienes en (a)?

La **short regression** difiere de la **long regression** principalmente en que incluye un número reducido de controles, lo cual se refleja en ligeras variaciones tanto en la magnitud de los coeficientes como en los errores estándar. En términos generales, los efectos estimados del programa mantienen la misma *dirección* y *significancia estadística*, lo que confirma la robustez de los resultados.

Sin embargo, en la mayoría de los paneles los coeficientes de la short regression tienden a

ser de mayor magnitud que en la long regression, lo que sugiere que la inclusión de controles adicionales en el modelo largo atenúa parcialmente los efectos del tratamiento. Esta diferencia es esperable, ya que la long regression captura mejor la heterogeneidad observable, mientras que la short regression ofrece una estimación más sesgada por omisión de variables.

e)

Utiliza el año **0** como línea basal (información de los individuos antes de que inicie el tratamiento). Calcula y reporta en una tabla la desviación estándar y el número de individuos en las escuelas con tratamiento y el número de individuos en las escuelas de comparación para las calificaciones de los alumnos (**t**), la dummy indicadora de si los alumnos presentaron el examen gubernamental (**tmock**), la dummy indicadora de si los alumnos presentaron el examen de la ONG **tics**) y la tasa de abandono (**dropout**).

Antes de obtener las estadísticas descriptivas, asegúrate de estar seleccionando las muestras adecuadas (por ejemplo, para calcular la media y la desviación estándar de las calificaciones de los alumnos es necesario utilizar la condición `if (table2==."& year==0)`).

Tabla 3: Estadística Descriptiva

	Calificaciones (t)			Examen gobierno (tmock)			Examen ONG (tics)			Abandono escolar (dropout)		
	Media	SD	N	Media	SD	N	Media	SD	N	Media	SD	N
Control	0.00	1.00	33614	0.80	0.40	10472	0.81	0.39	10737	0.13	0.34	7382
Tratamiento	0.05	1.03	30198	0.80	0.40	9507	0.83	0.38	9768	0.13	0.34	6711
Total	0.02	1.02	63812	0.80	0.40	19979	0.82	0.38	20505	0.13	0.34	14093

Fuente:Elaboración propia. Proceso disponible en:  danifuentesga

f)


Utiliza el comando **sampsi** de **STATA** para calcular el poder estadístico que tienen los autores para identificar los efectos del año **2** que contraste en el inciso (b) para las califi-

caciones de los alumnos (`t`), la dummy indicadora de si los alumnos presentaron el examen gubernamental (`tmock`), la dummy indicadora de si los alumnos presentaron el examen de la ONG (`tics`) y la tasa de abandono (`dropout`).

**Nota:** Lee con atención la descripción del comando `sampsi` en el *help* de STATA. Habiendo hecho esto, normaliza a 0 el #1 (`#1=0`); utiliza la `beta` que encontraste en el inciso (b) para cada una de las variables como #2, respectivamente; y utiliza las siguientes opciones: `sd(#)`, `alpha(#)`, `n1(#)`, `n2(#)` y `onesided`. Utiliza las desviaciones estándar y los números de individuos en las escuelas de comparación y con tratamiento que calculaste en el inciso (e) para sustituirlas en `sd(#)`, `n1(#)` y `n2(#)`, respectivamente. Como nivel de significancia (`alpha(#)`) utiliza el mínimo entre el *valor-p* de cada una de las `betas` o 0.05. Graba los resultados de STATA en un `log.file` y entrega ese `log.file`. Interpreta los resultados.

Tabla 4: Poder Estadístico

Panel	$\beta$ (Year=2)	N total	Poder
Calificaciones ( <code>t</code> )	0.224	63,812	1.000
Examen gobierno ( <code>tmock</code> )	0.063	19,979	1.000
Examen ONG ( <code>tics</code> )	0.010	20,505	0.585
Abandono escolar ( <code>dropout</code> )	-0.009	14,093	0.470

Fuente: Replicación elaborada con datos de Glewwe et al (2010). Proceso disponible en:  `danifuentesga`

**g)**


La aleatorización reportada por los autores aparentemente no se realizó a nivel alumno. La aleatorización se llevó a cabo a **nivel escuela**. Por lo tanto, los cálculos de poder se deben ajustar para tomar esto en consideración. Repite los cálculos que realizaste en el inciso (f), pero en lugar de definir el tamaño de muestra (`n1(#)`, `n2(#)`) y obtener el poder, define un nivel de poder de 0.8 (`power(0.8)`) y obtén el **tamaño de muestra necesario** para

alcanzar dicho poder. Después, utiliza el comando `sampclus` para realizar el ajuste al tamaño de muestra requerido para lograr un poder del 80 % dado que la aleatorización se llevó a cabo a nivel escuela.

**Nota:** Lee con atención la descripción del comando `sampclus` en el *help* de STATA. Habiendo hecho esto, define el número de escuelas en el estudio como el número de clusters (grupos). Para obtener el coeficiente de correlación intragrupal (*intraclass correlation*) utiliza el comando `loneaway` y la condición `if year==0`. Usa como `response_var` la variable con la cual vas a realizar el cálculo de poder. Usa como `group_var` la variable `s`. Graba los resultados de STATA en un `log.file` y entrega ese `log.file`.

Tabla 5: Clusters Minimios

Outcome	n (sin cluster)	n (con cluster)	Mín. clusters
t	449	17,676	135
tmock	499	9,226	71
tics	18,808	116,986	894
dropout	26,490	282,304	2,155

Nota: SD se redondeo a dos decimales. Fuente: Elaboración propia. Proceso disponible en:  danifuentesga

**h)**

Dado lo encontrado en los incisos (a)–(g), ¿crees en los resultados reportados por los autores en la Tabla 2? ¿Por qué?

**RESPUESTA:**

En general, **sí creo en los resultados reportados en la Tabla 2** para las variables de *calificaciones (t)* y *examen de gobierno (tmock)*, porque los efectos replicados son **robustos**, **significativos** y el **poder estadístico es alto**.

**Sin embargo**, *no confiaría demasiado en las conclusiones para tics y dropout*, ya que la réplica muestra **bajo poder estadístico** y los **tamaños de muestra requeridos** para alcanzar un poder del 80 % serían **enormes** bajo el diseño real de clusters.