

¿Importa la Calidad del Profesor?

Evidencia de Asignación Aleatoria de Estudiantes a Profesores

Carrell & West (2010), *Journal of Political Economy*

Martín Fuentes Santelices

The Economic Club

2026



Índice

- 1 Motivación
- 2 Contexto Empírico
- 3 Modelo Empírico
- 4 Resultados
- 5 Conclusiones
- 6 Literatura Reciente

Motivación

- ¿Cómo se mide la “calidad” de un profesor?
 - Primaria/secundaria: pruebas estandarizadas (*value-added*).
 - Educación superior: evaluaciones subjetivas de los estudiantes.
- Problema central: los profesores pueden “*enseñar para el examen*” — suben notas o evaluaciones sin generar aprendizaje real.
- Pregunta de investigación: **¿el profesor que mejora el desempeño contemporáneo también mejora el aprendizaje de largo plazo?**
- Reto empírico habitual: los estudiantes se *autoseleccionan* hacia ciertos profesores $\Rightarrow$ sesgo de selección.

Contexto: U.S. Air Force Academy (USAFA)

- Panel de **10,534 estudiantes** (2000–2007), 2,820 secciones, 421 profesores.
- Currículo **núcleo obligatorio** (~ 30 cursos): mismo programa, mismos exámenes, mismo período de evaluación.
- Cálculo I es requisito de 8 cursos posteriores (Cálculo II, Física, Ingeniería, Astronáutica, etc.).
- *Todos los alumnos deben completar el currículo obligatorio, que equivale a 8 semestres.*

Diseño de Identificación

- La oficina de registro asigna estudiantes a **secciones de forma aleatoria** (estratificada por género y atletas).
- Los autores verifican aleatoriedad con **10,000 remuestreos** por sección: distribución de p -valores $\approx$ uniforme.
- Regresión de p -valores sobre características del profesor (rango, experiencia, grado terminal): **ningún coeficiente significativo** de 36 estimados.
- Los exámenes de matemáticas se califican en “fiestas de calificación” — cada profesor califica una sola pregunta para *todos* los estudiantes $\Rightarrow$ elimina sesgo de calificación por profesor.

Modelo de Valor Agregado

Especificación

- Modelo de efectos aleatorios de dos ecuaciones (curso inicial y curso posterior), con controles de estudiante, sección y semestre-curso:

$$Y_{it}^1 = X_{it}^1\beta + \gamma_t^1 + \lambda_{j_1}^1 + \lambda_{j_2}^1 + v_{s_1}^1 + v_{s_2}^1 + \varepsilon_{it}^1$$

$$Y_{it}^2 = X_{it}^2\beta + \gamma_t^2 + \lambda_{j_2}^2 + \lambda_{j_1}^2 + v_{s_2}^2 + v_{s_1}^2 + \varepsilon_{it}^2$$

- γ_t : efecto fijo de **semestre-curso** t .
- $\lambda_{j_1}^1$: valor agregado del profesor inicial **en su propio curso** (aprendizaje contemporáneo).
- $\lambda_{j_1}^2$: valor agregado del **mismo** profesor inicial **en el curso posterior** (aprendizaje profundo).
- Efectos aleatorios de sección v y error idiosincrático ε controlan choques comunes de aula.

Descomponiendo la Covarianza

- ¿Qué parte del efecto contemporáneo *persiste* hacia el curso posterior?

$$\lambda_{j1}^2 = \rho \lambda_{j1}^1 + \phi_{j1}^2$$

- ρ : fracción del valor agregado contemporáneo que persiste.
- ϕ_{j1}^2 : aprendizaje “profundo” **no reflejado** en la nota del curso inicial (p.ej. comprensión conceptual).
- Estimación: $\hat{\rho} \approx -0.14$, **no significativo** $\Rightarrow$ sin evidencia de persistencia lineal positiva del valor agregado contemporáneo hacia el aprendizaje posterior.
- *Nota: los autores no reportan un SE directo para ρ (el test recae sobre $Cov(\lambda^1, \lambda^2)$); como chequeo de robustez usan el método 2SLS de Jacob et al. (2010) y obtienen el mismo resultado (negativo y no significativo).*

¿Qué es exactamente el “Aprendizaje Profundo”?

- $\phi_{j_1}^2$: lo que **queda** en $\lambda_{j_1}^2$ después de restar la parte que persiste linealmente ($\rho \lambda_{j_1}^1$).
- **Profundo**: comprensión transferible a problemas nuevos.
- **Superficial**: memorizar procedimientos para el examen.
- Ejemplo: aplicar la regla de la cadena sin entender por qué funciona basta para $\lambda_{j_1}^1$ alto, pero no ayuda en ramos avanzados de Ingeniería ($\lambda_{j_1}^2$).

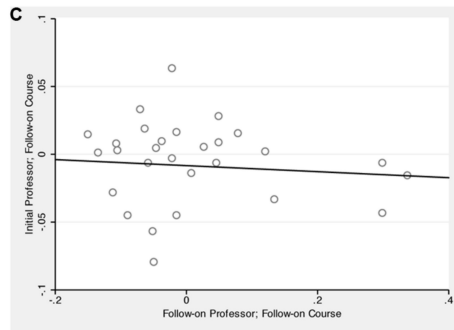
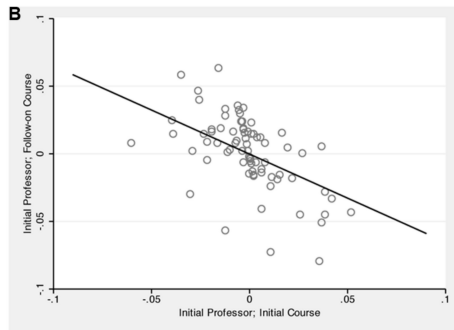
Profesor	λ^1 (Cálculo I)	$\rho \lambda^1$ (persiste)	ϕ (deep learning)	λ^2 (efecto futuro)
Rayo (Entrenador)	+0.50	−0.07 (perjudica)	+0.10 (poco profundo)	+0.03 (casi nulo)
Joaquín (Conceptual)	−0.50	+0.07 (beneficia)	+0.50 (muy profundo)	+0.57 (altísimo)

Tabla 4: Varianzas y Covarianzas del Modelo

Parámetro	Interpretación	Estimación	SD
$\text{Var}(\lambda_{j1}^1)$	Valor agregado, curso inicial	0.0028	0.052
$\text{Var}(\lambda_{j1}^2)$	Valor agregado, curso posterior	0.0025	0.050
$\text{Cov}(\lambda_{j1}^1, \lambda_{j1}^2)$	Persistencia entre cursos	-0.0004	n.s.
$\text{Var}(\lambda_{j2}^1)$	Placebo: prof. futuro $\rightarrow$ nota pasada	0.0000	—
$\text{Var}(\lambda_{j2}^2)$	Valor agregado, prof. seguimiento	0.0186	0.136
$\text{Var}(v_{s1}^1)$	Choque de sección, curso inicial	0.0255	—
$\text{Var}(v_{s1}^2)$	Persistencia del choque de sección	0.0248	—
$\text{Var}(v_{s2}^1)$	Placebo: sección futura $\rightarrow$ nota pasada	0.0000	—
$\text{Var}(v_{s2}^2)$	Choque de sección, curso posterior	0.0100	—

- La desviación estándar de las notas brutas de los alumnos es de **12 puntos porcentuales**.
- I.e: para λ_{j1}^1 , si un profesor está 1 SD por encima de la media, sus alumnos sacan $0.052 \times 12 = +0.6$ puntos.
- El efecto de valor agregado es más grande para cursos futuros (λ_{j2}^2). En cursos avanzados, el profesor importa mucho más, +1.6 puntos. *Podría atribuirse a menor n ?*

Visualizando la Covarianza: Paneles B y C (Fig. 1 del paper)



- **B:** Mismo profesor, dos cursos. Pendiente $\approx \hat{\rho} = -0.14$: más valor agregado hoy, menos aprendizaje profundo.
- **C – X:** $\hat{\lambda}_{j2}^2$ (ese profesor, en su curso de seguimiento, contemporáneo), **Y:** $\hat{\lambda}_{j1}^2$ (ese *mismo* profesor, aprendizaje profundo que generó cuando dictaba el curso inicial). Subconjunto que enseña ambos cursos. *Relación plana/ruidosa: el aprendizaje profundo que generas en un rol no predice tu desempeño contemporáneo en el otro rol.*

Características Observables del Profesor

Regresión sobre el profesor del curso inicial (j_1)

Característica del profesor j_1	$\hat{\lambda}_{j_1}^1$ (contemporáneo)	$\hat{\lambda}_{j_1}^2$ (posterior)
Profesor asociado/titular	-0.69*	+0.70*
Grado terminal (doctorado)	-0.28	+0.38
Más de 3 años de experiencia	-0.79***	+0.66**

- Profesores **más jóvenes y menos calificados**: mejores notas *inmediatas*.
- Profesores **más experimentados y calificados**: mejor desempeño en cursos *posteriores*.

Evaluaciones Estudiantiles

Regresión sobre el profesor del curso inicial (j_1)

Pregunta de evaluación (profesor j_1)	$\hat{\lambda}_{j_1}^1$ contemp.	$\hat{\lambda}_{j_1}^2$ posterior
Claridad y organización	+0.51***	-0.46**
Calidad de preguntas planteadas	+0.70***	-0.59**
Conocimiento de la materia	+0.56**	-0.44*
Evaluación general del curso	+0.49**	-0.39*
Cantidad aprendida	+0.59**	-0.47*
Efectividad facilitando aprendizaje	+0.54***	-0.45**

- Las 6 preguntas: **positivas** con $\hat{\lambda}_{j_1}^1$, **negativas** con $\hat{\lambda}_{j_1}^2$ — patrón sistemático.
- Correlación de rangos: profesor peor rankeado en $\hat{\lambda}_{j_1}^2 \rightarrow 6^{\circ}-7^{\circ}$ mejor en evaluaciones y en $\hat{\lambda}_{j_1}^1$.
- Los estudiantes premian notas altas “hoy”, no aprendizaje duradero.*

Conclusiones

Hallazgos Principales

- Los profesores sí importan — afectan tanto el desempeño contemporáneo como el de cursos posteriores.
- Pero **ser buen profesor “hoy” no implica serlo “después”** — no hay evidencia de persistencia lineal positiva del valor agregado, y las características docentes y evaluaciones apuntan en direcciones opuestas para ambos horizontes.
- Profesores experimentados y con doctorado: peor evaluación inmediata, mejor aprendizaje duradero.
- Las evaluaciones estudiantiles **premian lo contrario** de lo que predice el aprendizaje de largo plazo.

¿Por Qué Ocurre Esto? Tres Explicaciones

1. **Amplitud curricular.** Profesores menos experimentados se apegan más estrictamente al currículo evaluado; los más experimentados lo *amplían* y generan comprensión más profunda — a costa de la nota inmediata.
 2. **Hábitos de estudio transferidos.** Si el curso inicial “enseña para el examen”, el estudiante puede desarrollar hábitos de estudio superficiales que *traslada* (mal) al curso posterior.
 3. **Esfuerzo compensatorio (más cínica).** Estudiantes de profesores de *bajo* valor agregado inicial podrían **aumentar su esfuerzo** después, para “compensar” la nota más baja de lo esperado.
- *Los propios autores explicitan que no pueden distinguir entre estas tres explicaciones con sus datos.*

Implicaciones de Política

- Usar evaluaciones estudiantiles o notas contemporáneas para **tenure y promoción** puede incentivar el mal tipo de docencia.
- Los sistemas de pago por desempeño (*merit pay*) basados en valor agregado de corto plazo enfrentan el mismo riesgo.
- Recomendación de los autores: usar **medidas de aprendizaje de largo plazo** (cursos de seguimiento) cuando sea posible.
- Analogía con primaria/secundaria: el sesgo de “enseñar para el examen” documentado por Rothstein (2010) parece extenderse a la educación superior.

Literatura Reciente (I): 2014–2016

- **Braga, Paccagnella & Pellizzari (2014)**, *Economics of Education Review*
 - Replican el diseño de Carrell & West: asignación aleatoria de estudiantes a profesores.
 - Confirman que la calidad docente importa y correlaciona **negativamente** con las evaluaciones estudiantiles — quienes mejor preparan para cursos futuros reciben peores evaluaciones.
 - Aporte: teoría de “**utilidad realizada**” — los estudiantes evalúan notas y esfuerzo inmediato, no aprendizaje de largo plazo.
- **Rodríguez & Rubio (2016)**, *International Review of Economics Education*.
 - Misma metodología de valor agregado; asignación aleatoria por apellido.
 - Relación positiva pero **moderada** entre publicar en revistas top y calidad docente; sin relación con investigación ni actividades docentes.
 - Consideran las evaluaciones estudiantiles “sospechosas y discutibles” — los alumnos no distinguirían satisfacción de aprendizaje.

Literatura Reciente (II): 2020–2026

- **Stroebe (2020)**, *Basic and Applied Social Psychology*
 - Síntesis teórica y empírica de **97 estudios** multisectoriales.
 - Las evaluaciones estudiantiles (SET) están sesgadas por atractivo físico, género y disciplina — no miden efectividad docente.
 - Estudiantes premian lenidad al calificar $\Rightarrow$ profesores reducen exigencia $\Rightarrow$ inflación de notas y menos aprendizaje.
 - Conclusión textual: “*SETs are unrelated to teaching effectiveness.*”
- **Kursad, Dibek & Erdogan (2026)**, *Studies in Educational Evaluation*
 - Meta-análisis multinivel, **57 estudios**, 241 correlaciones.
 - Relación modesta pero significativa entre SET y rendimiento ($r = 0.30$) — señal detectable, pero limitada.
 - Heterogeneidad muy alta ($I^2 = 85.35\%$): la relación depende fuertemente del contexto.
- *Las evaluaciones estudiantiles capturan, en el mejor de los casos, una señal débil y contexto-dependiente de aprendizaje real.*

Gracias!



The Economic Club

Muchas gracias!