Investigando la validez de evaluaciones a escuelas a partir de "pruebas con altas consecuencias" (High-Stakes Tests)

Daniel Koretz
Hui Leng Ng
Escuela de posgrados en Educación de Harvard

V Seminario Internacional de Investigación sobre Calidad de la Educación - ICFES

Octubre 30 de 2014

Contexto

- Hay un uso creciente de pruebas de logro y estimaciones de modelos de valor agregado (MVA) para evaluar maestros y escuelas
- La evidencia muestra que los puntajes en pruebas con altas consecuencias están frecuentemente sesgados (puntajes inflados)
- Investigaciones recientes, basadas en pruebas con bajas consecuencias y pruebas con altas consecuencias, cuestionan la consistencia de estimaciones de MVA

Un ejemplo de puntajes inflados: ganancias en matemáticas estandarizadas en Kentucky, 1992-1996

	KIRIS*	NAEP**
Grado 4º	0,61	0,17
Grado 8º	0,52	0,13

1. KIRIS: *Kentucky Instructional Results Information Service Program,* sistema de rendición de cuentas del Estado de Kentucky implementado desde 1992.
2. NAEP: *National Assessment of Educational Progress,* prueba estandarizada a nivel nacional que se realiza en los grados 4°, 8° y 12°.

Estudios recientes sobre la consistencia de calificaciones de MVA de maestro a partir de pruebas con consecuencias altas y bajas

- Papay (2011): correlaciones de rango de 0,15 a 0,58
- Corcoran et al. (2011): datos de Houston, 0,59 en matemáticas y 0,50 en lectura
- Mihaly et al. (2013): datos del proyecto *"*Midiendo la enseñanza efectiva*"*, correlaciones *entre componentes estables* desde 0,39 hasta 0,54

Consistencia vs. validez

- Estudios recientes que comparan estimaciones entre pruebas no exploran asuntos de validez más allá de consistencia y fiabilidad
  - Papay (2011): señaló el riesgo de la inflación pero no utilizó las consecuencias como un predictor
  - Corcoran et al. (2011): “Nuestros resultados no sugieren que una prueba sea superior a la otra para la construcción [de estimaciones de MVA]”
- Pero los estudios que muestran la inflación de puntajes sugieren que la investigación también debe enfocarse en otros aspectos de la validez

Metas de este estudio

- Examinar la consistencia entre estimaciones de MVA para *escuelas* utilizando pruebas con consecuencias altas y bajas
  - a través de grados, años, áreas y modelos
- Examinar evidencia adicional sobre validez
  - ¿Existe evidencia de que la inflación de puntajes contribuye en la inconsistencia?

Hipótesis

- Las calificaciones serán inconsistentes sin consideración del modelo, incluso a nivel de escuela en lugar del nivel de maestro
- Las diferencias entre puntajes de altas y bajas consecuencias serán predichas por variables correlacionadas con la preparación de la prueba (pertenencia a minorías, pobreza)
- Las calificaciones mostrarán demasiada consistencia dentro de pruebas y entre áreas
- La evidencia sobre validez será más negativa al nivel de escuela

Contexto del estudio: Houston

- De gran tamaño (203.000 estudiantes)
- Una minoría proporcionalmente alta (8% blancos no hispánicos)
- El 80% son estudiantes con desventajas económicas (almuerzo gratis o con tarifa reducida)
- Dos pruebas analizadas en grados 3-8 en años analizados
  - TAAS (*Texas Assessment of Academic Skills* - "Evaluación de habilidades académicas de Texas"): prueba estatal usada para rendición de cuentas
  - SAT-9 (*Scholastic AptitudeTest* – "Prueba de aptitud escolar"): prueba comercial de referencia normativa con consecuencias relativamente bajas

Datos

- Resultados: puntajes en grado 5° en lectura y matemáticas en 2000, 2001 (antes de *No Child Left Behind* – "Ningún niño dejado atrás")
- El 9.9% carecía al menos de 1 puntaje TAAS, el 0.3% carecía de puntaje SAT-9
  - Se crearon bases de datos específicas por prueba y agrupadas
- Solamente se incluyeron escuelas con datos de estudiantes válidos para ambas pruebas
- Conteos finales: 20.921 estudiantes con SAT-9, 17.992 conTAAS y 17.787 en la intersección, en 164 escuelas

Métodos

- Cálculo de estimaciones de MVA por escuela:
  - Modelos de coeficientes aleatorios de 2 niveles, 6 modelos de ajuste de covariables y 6 modelos de ganancia de puntajes que difieren en las covariables usadas
  - Las variables fueron centradas de acuerdo con la media general
- Estimación de las correlaciones de la discrepancia de TAAS-SAT:
  - Modelo de coeficientes aleatorios de 3 niveles con la diferencia de puntajes por estudiante como resultado
- Estimación de la consistencia de los rankings:
  - Correlaciones de rango de Spearman
  - Consistencia en la asignación a bandas de rendimiento

Conjuntos de covariables usados en los modelos

1. Ninguno

2. Sólo antecendentes a nivel de estudiante

3. Antecedentes a nivel de estudiante y escuelas

4. Logro anterior en grado 3º a nivel de estudiante solamente

5. Logro anterior en grado 3º y antecendentes a nivel de estudiante

6. Logro anterior en grado 3º y antecendentes a nivel de estudiante y escuela

Métodos, continuación

- Construcción de las bandas de rendimiento:
  1. Tres bandas, cortes en ± 1 DE posterior
  2. Quintiles
  3. 5 bandas, asimétricas, con proporciones desiguales: modeladas siguiendo ligeramente el sistema de la ciudad de Nueva York
- Múltiples rasgos-múltiples métodos- MRMM (*Multitrait-Multimethod*):
  - Las medidas de rasgos fueron correlaciones de Spearman intra áreas, inter pruebas
  - Las medidas de métodos fueron correlaciones intra pruebas, inter áreas
  - Las correlaciones de rasgos deben exceder las correlaciones de métodos

Modelos de efectos aleatorios de 2 niveles para calificación de escuelas

Modelos de ajuste de covariables:

$$Y\downarrow isy = \mu + \alpha Y\downarrow is(y-1) + \mathbf{B}\downarrow \mathbf{is}\ \boldsymbol{\beta} + \mathbf{PA}\downarrow \mathbf{is}\ \boldsymbol{\pi} + \mathbf{S}\downarrow \mathbf{s}\ \boldsymbol{\gamma} + \psi\downarrow s + \varepsilon\downarrow is$$

Modelo de ganancia de puntajes:

$$[Y\downarrow isy - Y\downarrow is(y-1)] = \mu + \mathbf{B}\downarrow \mathbf{is}\ \boldsymbol{\beta} + \mathbf{PA}\downarrow \mathbf{is}\ \boldsymbol{\pi} + \mathbf{S}\downarrow \mathbf{s}\ \boldsymbol{\gamma} + \psi\downarrow s + \varepsilon\downarrow is$$

$\mathbf{B}\downarrow \mathbf{is}$ vector de variables de antecedentes, estudiante *i* en escuela *s*

$\mathbf{PA}\downarrow \mathbf{is}$ logro previo (grado 3º), matemáticas+ lectura

$\mathbf{S}\downarrow \mathbf{s}$ medias de variables de estudiantes por escuela

Modelo de efectos aleatorios de 3 niveles para la estimación de la correlación de diferencias de puntajes

$$\left(TAAS_{isy} - SAT9_{isy}\right) = a_{000} + a_{100}NONWHITE_{isy} + a_{200}ECONDIS_{isy}$$
$$+ a_{010}SM\_NONWHITE_{sy} + a_{020}SM\_ECONDIS_{sy}$$
$$+ a'_{300}(\mathbf{X}_{isy}) + \left(e_{isy} + u_{0sy} + v_{00y}\right)$$

$\mathbf{X}_{isy}$ Vector de variables de control: puntaje en TAAS en el año anterior y cuadrática, género, competencia limitada en inglés, discapacidad

NONWHITE: No blanco
ECONDIS: Desventajas económicas

Modelos para estimación de correlaciones dentro de escuelas agregadas para nivel de estudiante MRMM

Intercepto aleatorio de 2 niveles, modelos de pendiente e intercepto aleatorios, centrados en media de escuela

WS(R): $$R_{is}^{\text{TAAS}} = g_{0,\text{R}} + g_{1,\text{R}} R_{is}^{\text{SAT-9}} + \left(u_{s,\text{R}} + v_{s,\text{R}} R_{is}^{\text{SAT-9}} + e_{is}^{\text{R}} \right)$$

WS(M): $$M_{is}^{\text{TAAS}} = g_{0,\text{M}} + g_{1,\text{M}} M_{is}^{\text{SAT-9}} + \left(u_{s,\text{M}} + v_{s,\text{M}} M_{is}^{\text{SAT-9}} + e_{is}^{\text{M}} \right)$$

WT(TAAS): $$M_{is}^{\text{TAAS}} = g_{0,\text{TAAS}} + g_{1,\text{TAAS}} R_{is}^{\text{TAAS}} + \left(u_{s,\text{TAAS}} + v_{s,\text{TAAS}} R_{is}^{\text{TAAS}} + e_{is}^{\text{TAAS}} \right)$$

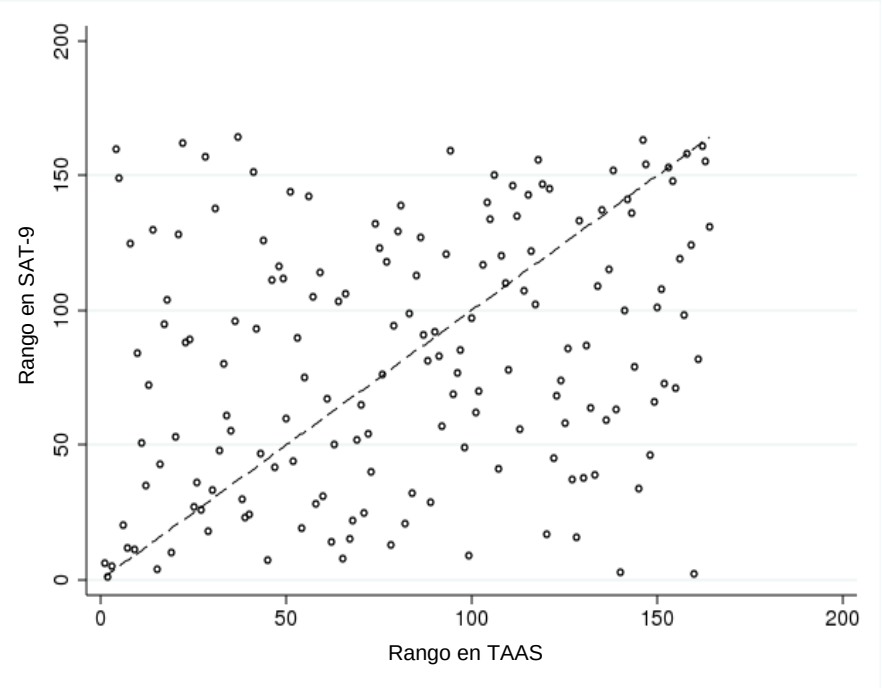
WS(R):	intra escuela (lectura)
WS (M):	intra escuela (matemáticas)
WT(TAAS):	intra prueba (prueba TAAS)

Conclusiones principales

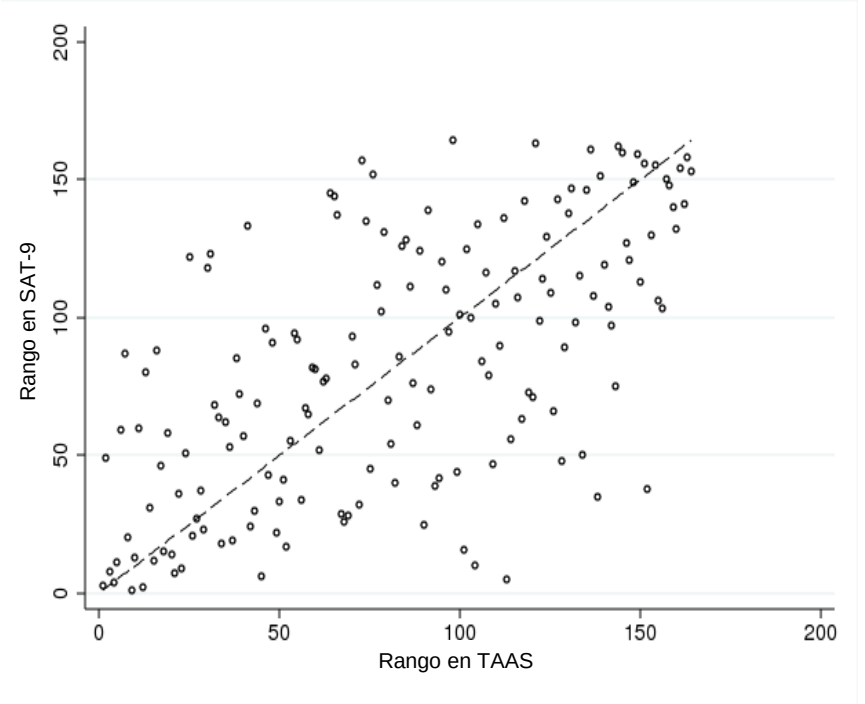
- Se obtuvieron resultados similares a través de modelos y años
- Las estimaciones de MVA fueron inconsistentes entre pruebas a pesar del enfoque en escuelas en lugar de maestros
  - Correlaciones de Spearman mediocres o débiles
  - Débil concordancia en la clasificación en bandas de rendimiento
- Evidencia adicional de validez arroja dudas sobre las calificaciones de altas consecuencias:
  - MRMM mostró un método más fuerte que el efecto de rasgos para TAAS al nivel de escuela
  - En otros estudios la diferencia de puntaje por estudiante es predicha por factores asociados con la preparación para la prueba e inflación de puntajes

Rangos en TAAS y SAT-9, correlaciones de rango mínimo y máximo (de 48 correlaciones a través de modelos, años y áreas)

r=0,27; lectura, 2000



r=0,63; matemáticas, 2000



Concordancia observada y por azar entre clasificaciones basadas en TAAS y SAT-9

		±1 DE		Quintiles		Desigual	
Lectura	Observada	64	79	24	40	31	48
	Por azar	57	63	20	20	28	28
	Coeficiente k de Cohen	0,21	0,34	0,05	0,27	0,09	0,26
Matemáticas	Observada	68	80	30	40	38	47
	Por azar	55	63	20	20	28	28
	Coeficiente k de Cohen	0,22	0,48	0,5	0,27	0,15	0,28

Matriz de múltiples rasgos y múltiples métodos para dos medidas de ansiedad y depresión

		Auto reporte		Observación	
		Ansiedad	Depresión	Ansiedad	Depresión
Auto rep.	Ansiedad	Fiabilidad	Método	Rasgo	Ninguno
Auto rep.	Depresión	Método	Fiabilidad	Ninguno	Rasgo
Observ.	Ansiedad	Rasgo	Ninguno	Fiabilidad.	Método
Observ.	Depresión	Ninguno	Rasgo	Método	Fiabilidad

Fuente: Nunnally and Bernstein, 1994

Análisis con multiples métodos y multiples rasgos, nivel de escuela

Las correlaciones intra áreas (WS) inter pruebas (rasgo), deben exceder las correlaciones intra pruebas (WT) inter áreas (método)

	WS(R)	WS(M)	WT(TAAS)	WT(SAT9)
Covariables de estudiante y escuela	0,30	0,61	0,72	0,55
No covariables	0,43	0,58	0,69	0,69

WS(R):	intra escuela (lectura)
WS (M):	intra escuela (matemáticas)
WT(TAAS):	intra prueba (prueba TAAS)
WT(SAT):	intra prueba (prueba SAT)

Análisis de múltiples métodos y múltiples rasgos, nivel de estudiante (intra escuelas)

	WS(R)	WS(M)	WT(TAAS)
Correlación	0,70	0,73	0,61
DE de las pendientes intra escuelas	0,17	0,28	0,15

WS(R):	intra escuela (lectura)
WS (M):	intra escuela (matemáticas)
WT(TAAS):	intra prueba (prueba TAAS)

Regresión multinivel predictora de la diferencia TAAS-SAT9

	1	2
Fixed Effects		
NONWHITE ($\hat{a}_{100}$)	0.147***	0.146***
ECONDIS ($\hat{a}_{200}$)	0.101***	0.095***
SM_NONWHITE ($\hat{a}_{010}$)	0.470***	
SM_ECONDIS ($\hat{a}_{020}$)		0.407***
Intercept	-0.098***	-0.098***
TAAS_M(Y-1)	-0.011	-0.011
TAAS_M(Y-1)_sq	-0.101***	-0.101***
FEMALE	0.052***	0.052***
LEP	0.078***	0.077***
SPECED	0.079	0.079
DISABLED	-0.037	-0.036
Random Effects		
Between-year ($\hat{s}_v^2$)	0.000	0.000*
Between-school, within-year ($\hat{s}_u^2$)	0.064***	0.060***
Within-school & within-year ($\hat{s}_e^2$)	0.376***	0.376***

Key: * p <.05; ** p < .01; *** p < .001. All variables were grand-mean-centered.

Conclusiones

- Las calificaciones son inconsistentes entre pruebas
  - Esto representa varianza entre escuelas, no sólo entre maestros
- Parte de esta inconsistencia representa un sesgo (puntajes inflados) en calificaciones basadas en pruebas de altas consecuencias
- Por lo tanto, las calificaciones no son solamente incorrectas aleatoriamente, sino que también son sistemáticamente incorrectas. Esto genera:
  - Errores en las recompensas y sanciones para las escuelas
  - Errores en las investigaciones que evalúan la efectividad de programas
  - En el contexto de EEUU, una ilusion de mayor equidad

Próximos pasos

- Es necesario replicar el estudio en otros contextos, pruebas y sistemas de rendición de cuentas
- Es necesario tomar medidas para avanzar en la minimización de preparaciones inapropiadas para pruebas y puntajes inflados
- Es necesario incluir la investigación de evidencia de validez, así como la consistencia y la fiabilidad, en la evaluación de los sistemas de rendición de cuentas basada en pruebas.

Diapositivas adicionales

Referencias

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). Teacher effectiveness on high- and low-stakes tests. Working paper. Consultado el 7 de julio de 2014, dehttps://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. Working paper, Measuring Effective Teaching Project. http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193

Referencias, continuación

Nunnally, J. M., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.