Usando Pruebas para Evaluar: La experiencia de Estados Unidos

Daniel Koretz
Facultad de Educación de Harvard

3er Seminario Internacional del ICFES sobre Investigación en Educación

1 Noviembre 2012

Planteamiento del problema

- Creciente interés en el mundo por utilizar pruebas para evaluar y rendir cuentas
  - Para monitorear el desempeño de escuelas y sistemas
  - Para estimular mejoramiento
  - Para seleccionar y ordenar estudiantes

- Experiencia sustancial con *pruebas cuyos resultados generan consecuencias* en EEUU
  - Numerosos programas desde inicios de 1970s
  - Evaluaciones de impacto desde finales de 1980s

- Investigación (principalmente en EEUU) es limitada y presenta serios problemas

- Necesidad de construir sistemas que minimicen estos problemas

Qué sabemos y qué no sabemos acerca de las *pruebas cuyos resultados generan consecuencias*

- El efecto sobre el logro escolar no es claro
  - Débiles diseños de investigación, datos aún más débiles
  - Evidencia inconsistente sobre efectos modestos

- Efectos mixtos sobre práctica educativa
  - Algunas mejoras
  - Algunos efectos indeseables –preparación inapropiada para las pruebas, uso de trucos para obtener mejores resultados

- Puntajes pueden ser inflados significativamente (aumentar mucho más que el aprendizaje real)

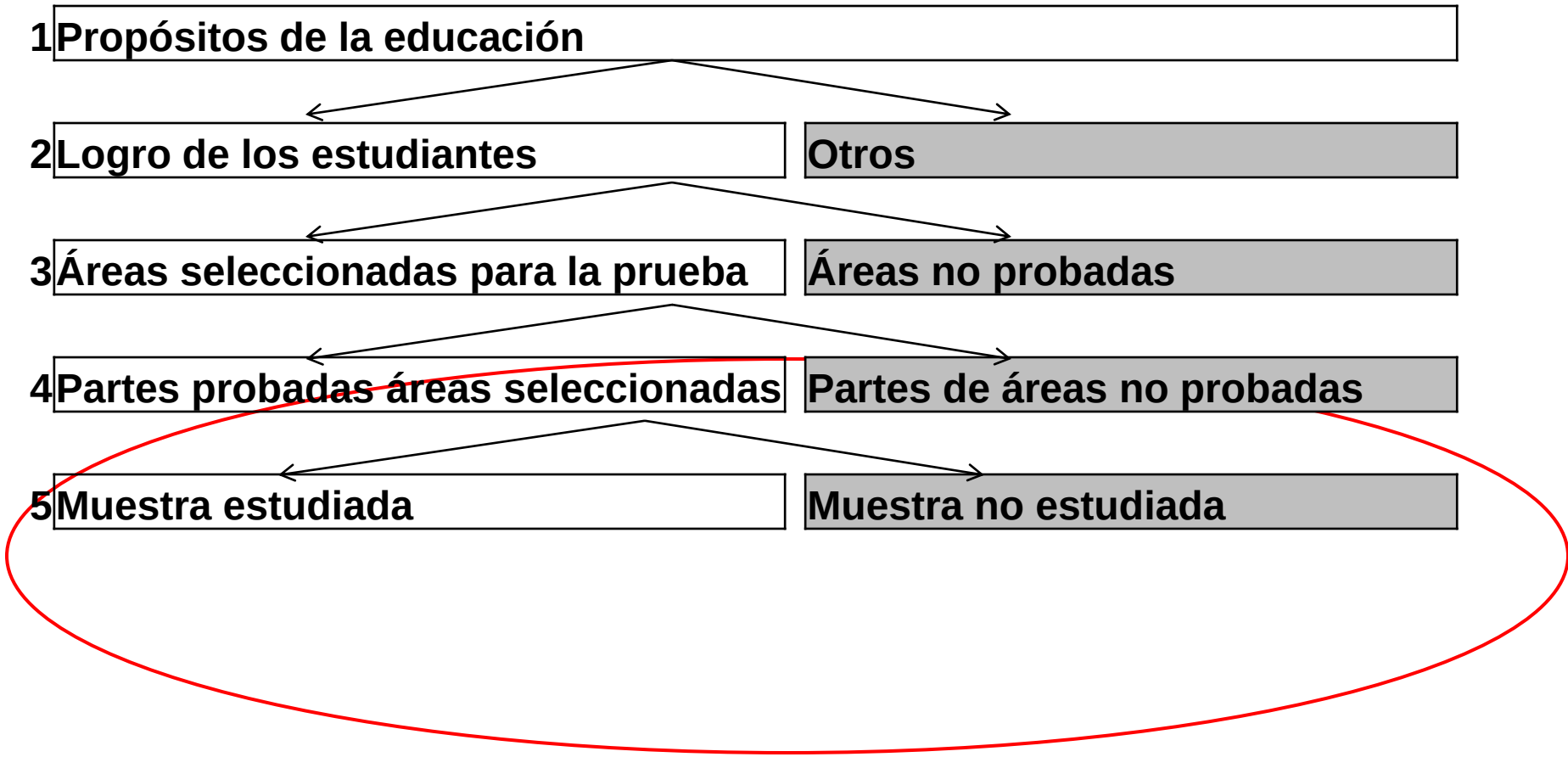
Tópicos

- El “principio de muestreo” de las pruebas
- Evidencia de inflación de puntajes
- Respuestas a *pruebas cuyos resultados generan consecuencias*: cómo se genera la inflación de puntajes
- Implicaciones para el desarrollo de nuevos programas de pruebas y evaluación

El “principio de muestreo” de las pruebas: Analogía con una encuesta electoral

- El 3 de Junio de 2010, una encuesta con 2.000 personas del Centro Nacional de Consultoría predijo 61.6% para Santos, 29.8% para Mockus
- Resultados finales: 69.1% para Santos, 27.5% para Mockus
- ¿Nos debería importar cómo votaron los 2.000 encuestados en particular?
- ¿Por qué es valiosa la información sobre esas 2.000 personas?

Muestreo para construir una prueba



¿Qué consecuencias tiene un muestreo incompleto?

- En todos los casos:
  - Evaluación educativa sistemáticamente incompleta
- Baja presión: efectos modestos
  - Error de medición (incertidumbre): fluctuaciones en puntajes
  - (Usualmente) diferencias modestas entre pruebas
- Alta presión (rendición cuentas): efectos muy grandes
  - Incentivos para concentrarse en la muestra utilizada, no en todo el área
  - Instrucción restringida, mala preparación para la prueba
  - Inflación de puntajes

Tópicos

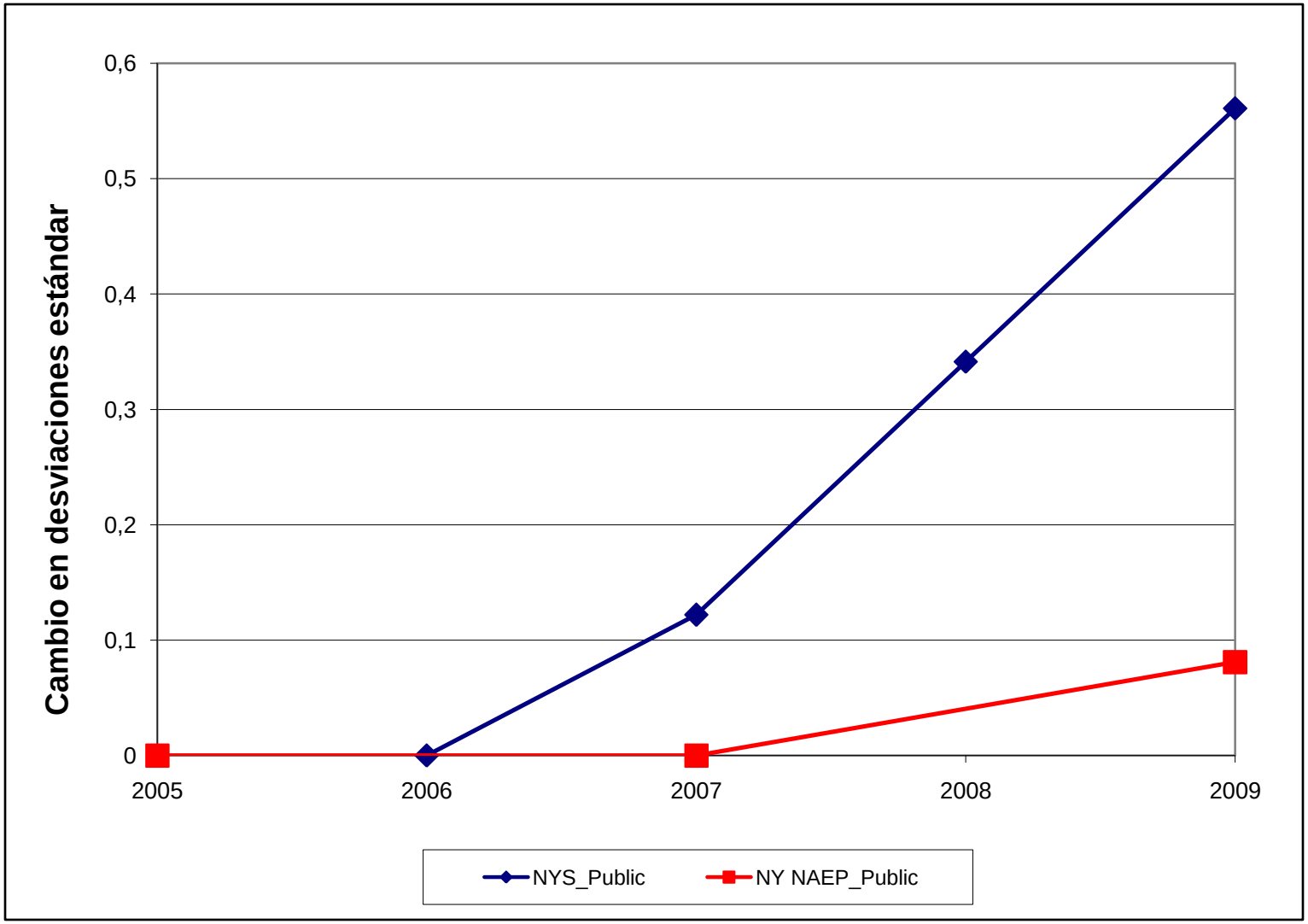
- El “principio de muestreo” de las pruebas
- Evidencia de inflación de puntajes
- Respuestas a *pruebas cuyos resultados generan consecuencias*: cómo se genera la inflación de puntajes
- Implicaciones para el desarrollo de nuevos programas de pruebas y evaluación

Lógica de estudios sobre inflación de puntajes

- Los puntajes **sólo** tienen significado si sirven para generalizar al área
  - Una encuesta es útil solo si sus resultados son generalizables a todo el electorado

- Si las ganancias son generalizables al área, éstas **tienen** que ser generalizables a otras pruebas dentro de la misma área
  - Si una encuesta es precisa, otras encuestas mostrarán resultados similares

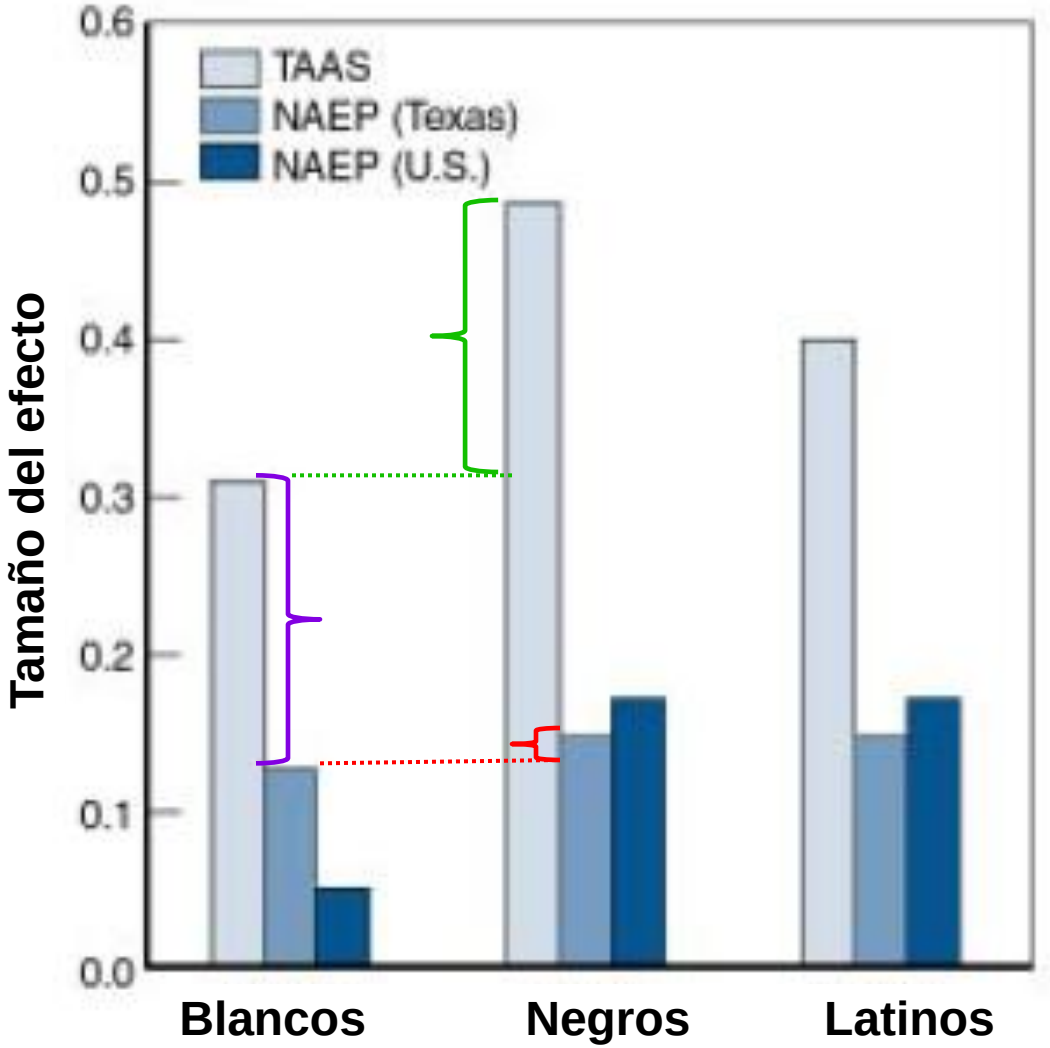
Tendencia puntajes Grado 8° Matemáticas N.York



Cambios en lectura, Grado 4 KIRIS y NAEP, 1992-1994

	KIRIS	NAEP
Ganancia puntajes escala	18,8	-1
Ganancia estandarizada	0,76	-0,03

Una mirada al "milagro de Texas" (Klein, et al. 2000)



Tópicos

- El “principio de muestreo” de las pruebas
- Evidencia de inflación de puntajes
- Respuestas a *pruebas cuyos resultados generan consecuencias*: cómo se genera la inflación de puntajes
- Implicaciones para el desarrollo de nuevos programas de pruebas y evaluación

Buena vs mala preparación para una prueba

- Buena: entrega a estudiantes conocimientos y destrezas que pueden aplicar en cualquier parte
  - En estudios posteriores
  - En empleos posteriores
  - En consecuencia, en otras pruebas
- Mala: genera ganancias específicas a la prueba que no son generalizables mas allá de la prueba

Caminos para aumentar los puntajes

Enseñar más

Trabajar más duro

Trabajar más efectivamente

Reasignación

Entrenamiento

Trampas

Cambiar a quien toma la prueba

Reasignación

- Cambiar los recursos de enseñanza para que coincidan con las pruebas
  - Dentro de un área
  - Entre áreas
- Reasignar logro
- Dentro de un área puede generar cambios significativos ó inflación
  - Hay inflación si el material que recibe menos énfasis es también importante para la inferencia

Oportunidades de la reasignación

Patrones recurrentes y predecibles en la prueba:

- Enfasis recurrente
  - Parte del contenido evaluado aparece más en las pruebas que otros contenidos
- Omisiones recurrentes de la prueba

Algebra 1

7.1	
7.2	2003S #17 (o)
7.3	
7.4	2003S #38 (m) 2002F #37 (m) 2000S #36 (m)
7.5	
7.6	
7.7	

Fuente: Quincy MA High School Dept de Matemáticas.

Entrenamiento

- Se concentra en preparar en detalles sin importancia sustantiva de la prueba
  - Detalles menores del contenido, sin importancia
  - Detalles sobre cómo se presenta el material en la prueba
- Incluye trucos para tomar la prueba (por ejemplo, proceso de eliminación, vínculos)
- Puede inflar los puntajes o simplemente perder tiempo

Oportunidades del entrenamiento

De nuevo, patrones recurrentes en la prueba

- Detalles menores del contenido recurrentes (énfasis y omisión)
- Formas de presentación recurrentes
  - Formatos de ítem
  - Otros aspectos de la presentación
- Demandas recurrentes de respuesta (por ejemplo, cómo se califica el trabajo)

2008 item, New York Grado 7
Prueba de Matemáticas

¿Cuál herramienta es más apropiada para medir la masa de una porción de queso?

a. Regla

b. Termómetro

c. Taza para medir

d. Báscula

2009 item, New York Grado 7
Prueba de Matemáticas

¿Cuál herramienta sería la más apropiada para que Natasha encuentre la masa de una sandía?

a. Escala

b. Regla de pulgadas

c. Metro de madera

d. Taza de medir

Item de Grado 8 MCAS

Eva tiene cuatro grupos de pitillos. Las medidas de los pitillos están abajo. ¿Cuál grupo de pitillos no puede ser utilizado para formar un triángulo?

A. Grupo 1: 4 cm, 4 cm, 7 cm
B. Grupo 2: 2 cm, 3 cm, 8 cm
C. Grupo 3: 3 cm, 4 cm, 5 cm
D. Grupo 4: 5 cm, 12 cm, 13 cm

Ejemplo de entrenamiento (¿trampa?)

"La pregunta en la hoja de revisión para … (en el) examen … dice en parte:

'La cantidad promedio que cada miembro de la banda debe conseguir es función del número de miembros de la banda, b, bajo la regla f(b)=12000/b.'

La pregunta en la prueba real dice en parte:

'La cantidad promedio que cada porrista debe pagar es función del número de porristas, n, bajo la regla f(n)=420/n'"

Strauss, V., *The Washington Post*, Julio 10, 2001, p. A09

Entrenamiento: basado en una característica incidental de ítems de la prueba

Siempre que usted tenga un triángulo recto – triángulo con un ángulo de 90 grados – puede utilizar el Teorema de Pitágoras… la suma de los cuadrados de los lados del triángulo (los lados próximos al ángulo recto) será igual al cuadrado de la hipotenusa (el lado opuesto al ángulo recto)…

Dos de los más frecuentes cocientes que se ajustan al Teorema de Pitagoras son 3:4:5 y 5:12:13. Dado que estos son cocientes, cualquier múltiplo de estos números también funcionará, tales como 6:8:10, y 30:40:50.

Princeton Review, *Cracking The MCAS Grade 10 Mathematics*

Tópicos

- El “principio de muestreo” de las pruebas
- Evidencia de inflación de puntajes
- Respuestas a *pruebas cuyos resultados generan consecuencias*: cómo se genera la inflación de puntajes
- Implicaciones para el desarrollo de nuevos programas de pruebas y evaluación

Qué sabemos

- Usar solamente una prueba para evaluar y rendir cuentas no es adecuado
  - Pruebas omiten muchos resultados importantes
  - Pruebas con consecuencias para el evaluado generan efectos mixtos en la práctica
  - Pruebas con consecuencias para el evaluado producen ganancias infladas en puntajes

- Inflación de puntajes socava la evaluación de dos maneras:
  - La mejora total es exagerada
  - La efectividad relativa (por ejemplo de las escuelas) es estimada incorrectamente

¿Y los modelos de valor agregado? (MVA)?

- Los MVA son en cierta medida la mejor manera de medir resultados
- Los MVA implican temas adicionales, por ejemplo:
  - Grandes cantidades de ruido aleatorio
  - Resultados inestables entre pruebas
  - Dificultad infiriendo los efectos verdaderos de maestros y escuelas
- Los MVA no manejan el problema de inflación de puntajes

¿Qué no sabemos?

- No hemos identificado los mejores tipos de evaluación basada en pruebas de logro ni de rendición de cuentas
  - ¿Cuáles programas maximizan mejoras reales?
  - ¿Cuáles programas minimizan trucos, mala preparación para pruebas e inflación de puntajes?
- Motivo: investigación y evaluación insuficiente
  - Investigación muestra razones para preocuparse
  - Investigación no ha evaluado aún los mejores diseños

Sugerencias

- Llevar a cabo monitoreo y evaluación en forma continua
  - Evaluar el sistema de evaluación, no solo la educación
- Probar nuevos diseños de pruebas
- Probar nuevos diseños de sistemas de rendición de cuentas más grandes

Necesidad de sistemas de monitoreo y evaluación

- Necesidad de monitorear:
  - Respuestas de comportamiento de educadores
  - Otras formas de trucos
  - Inflación de puntajes
- Necesidad de investigar sobre variaciones en efectos, por ejemplo:
  - Variaciones entre tipos de escuelas
  - Variaciones entre tipos de estudiantes

Necesidad de experimentar con nuevos diseños de pruebas

Para estimar mejor las ganancias reales y mejorar los incentivos

- Maximizar la cobertura
- Minimizar la repetición innecesaria de:
  - Contenidos
  - Estilos de presentación
  - Demanda de tareas y calificación
- Construir pruebas “con auditoría”
  - En programas de pruebas basadas en muestreo
  - Con items insertados (“evaluaciones auto monitoreadas”)

Necesidad de experimentar con nuevos diseños de sistemas de evaluación

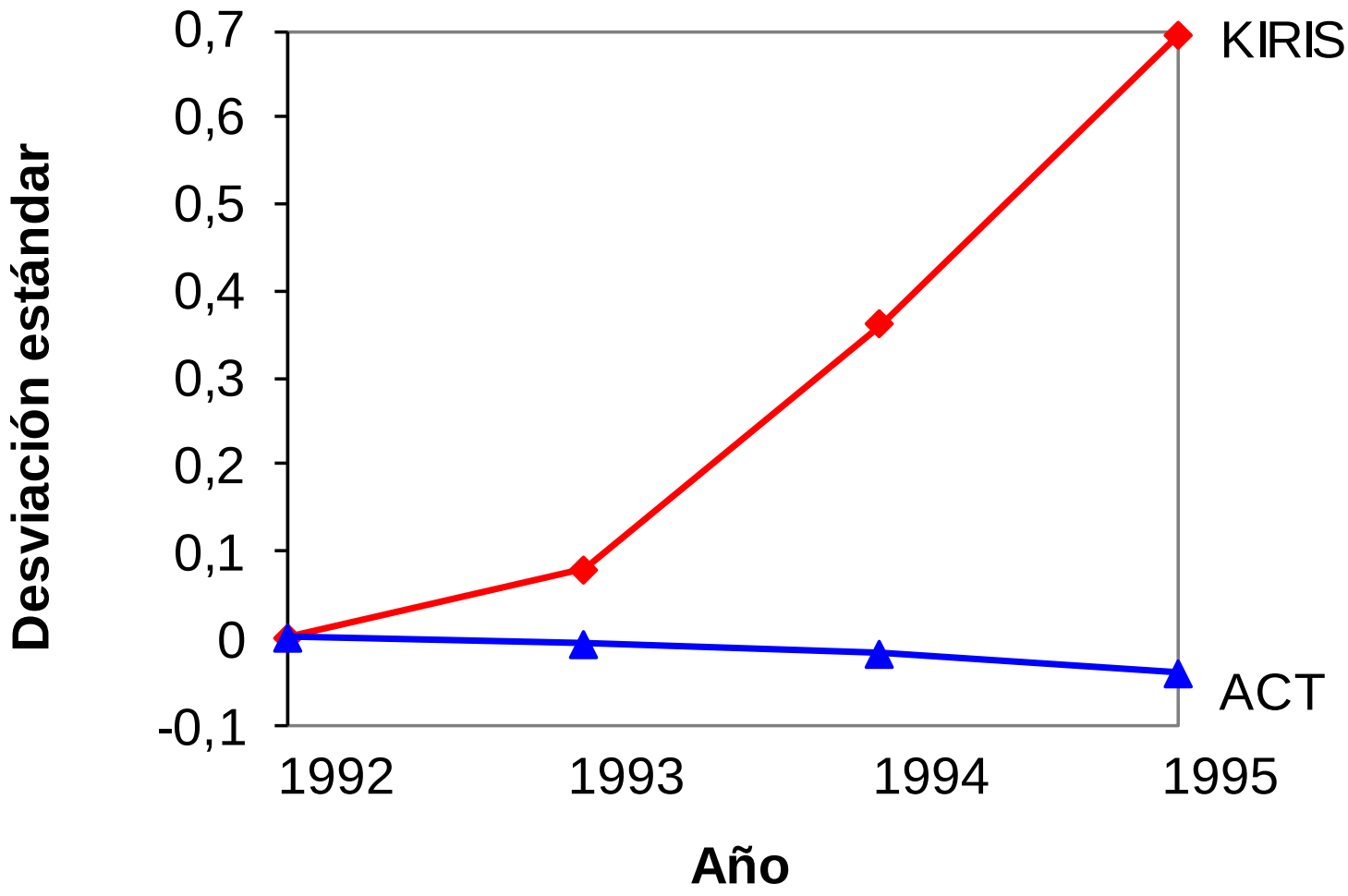
- Necesidad de encontrar maneras para que otros propósitos *cuenten*
  - Incluyendo destrezas de mayor nivel que son difíciles de valorar con una prueba impuesta externamente
- Necesidad de explorar el uso de múltiples medidas
  - Medidas objetivas adicionales
  - Medidas subjetivas
- Necesidad de monitorear trucos, considerar rendición de cuentas “dinámica”

Siguientes pasos: Cuatro puntos claves

- Sea cauteloso: aproveche los problemas evidenciados en la experiencia estadounidense
- Pruebe con un enfoque más amplio: no solamente puntajes en las pruebas
- Monitoree y evalúe el sistema rutinariamente y esté preparado para modificar los programas de pruebas y evaluación
- Participe en investigación y desarrollo anticipatorios

Diapositivas adicionales

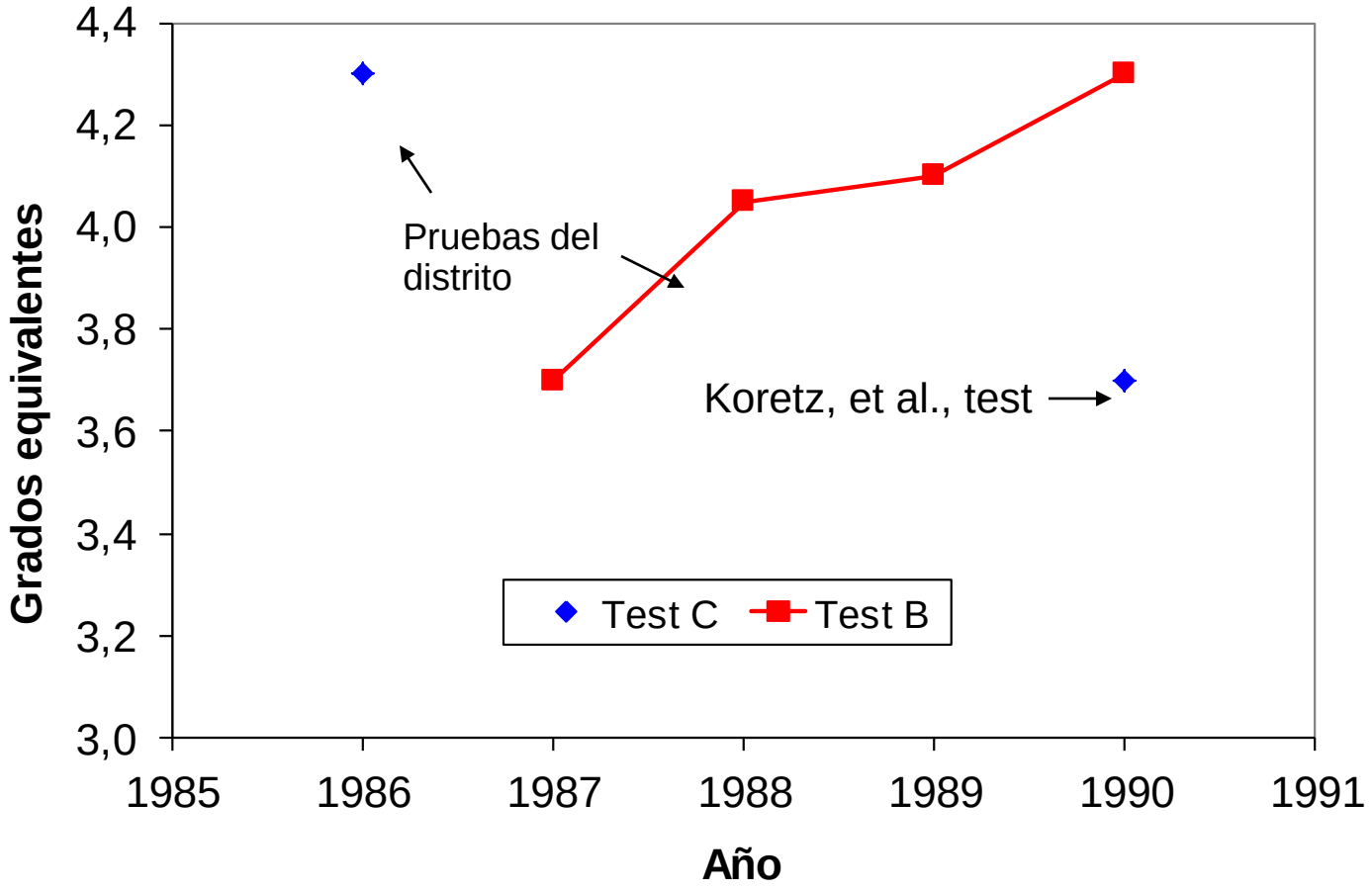
Tendencias en matemáticas, KIRIS y ACT



Ganancias estandarizadas en matemáticas Kentucky, 1992-1996

	KIRIS	NAEP
Grado 4°	0,61	0,17
Grado 8°	0,52	0,13

Desempeño en pruebas con y sin entrenamiento



Fuente: Adaptado de Koretz, Linn, Dunbar y Shepard (1991)

Ejemplos de tres listas de palabras

A	B	C
Siliculoso	Baño	Irresponsable
Vilipendiar	Viaje	Menosprecio
Epimisio	Alfombra	Minúsculo

Nuevas muestras de tres listas de palabras

A	B	C
Siliculoso	Baño	Irresponsable/ Parsimonioso
Vilipendio	Viaje	Menosprecio
Epimisio	Alfombra	Minúsculo

“Ley de Campbell” (1975)

“A mayor uso de cualquier indicador social cuantitativo para la toma de decisiones sociales, mayor es su exposición a presiones corruptas, y mayor su vulnerabilidad a ser distorsionado y a corromper el proceso social que esté busca monitorear.”

Donald T. Campbell, (1975). “Evaluando el impacto del cambio social planificado” In G. M. Lyons (Ed.), *Social Research And Public Policies : The Dartmouth/OECD Conference*.

Ejemplos de la Ley de Campbell

- Aerolíneas, estadísticas de tiempos
- Tiempos de entrega del Servicio Postal de Virginia Occidental
- “Tarjetas de Reporte” de Cardiología en New York

Para más ejemplos véase:

http://www.performanceincentives.org/data/files/directory/ConferencePapersNews/Rothstein.pdf