PERSPECTIVAS METODOLÓGICAS : ¿EVALUACIÓN CUANTITATIVA O CUALITATIVA?

Richard J. Shavelson

SK Partners & Universidad de Stanford

Overview

- "Cuidado con las dicotomías falsas": Los métodos de evaluación deben ser impulsados por preguntas de evaluación, no viceversa
- Tipos de preguntas de evaluación
  - Descriptivas
  - Causales
  - Mecanismos
- Buscar "lo que importa" que subyace en las evaluaciones de escuelas y de la enseñanza
  - Políticas y política
  - Métodos de evaluación
  - Diseño estadístico

Diseño de investigación/evaluación impulsado por preguntas

- ¿Qué está pasando?
- ¿Existe un efecto sistemático (causal)?
- ¿Cómo o por qué (mecanismo) está pasando?

- Investigación descriptiva y evaluación formativa
- Investigación impulsada por la teoría o la práctica y evaluación sumativa
- Investigación impulsada por la teoría o la práctica y evaluación formativa o sumativa

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press

Preguntas y métodos

Pregunta / Método	"¿Qué está pasando?" Descriptivo	"¿Existe un efecto sistemático?" Causal	"¿Cómo o por qué está pasando?" Mecanismos
Cualitativo	• Estudio de caso • Etnografía • Observación • Entrevista • Etc. Ejemplo: Holland & Eisenhart	• Etnografía • Múltiples estudios de casos • Etc. (?) Ejemplo: Holland & Eisenhart	• Estudio de caso • Etnografía • Observación • Entrevista • Etc. Ejemplo: Seguimiento a Experimento TN
Cuantitativo	• Encuesta de probabilidad (o resumen estadístico basado en cualitativo) • Estadísticas descriptivas • Comparaciones de estadísticas • Correlaciones • Regresiones • Etc. Ejemplo: NAEP	• Experimentos aleatorios • Cuasi-experimentos • Diseño causal de datos longitudinales • Discontinuidad de regresión • Intento por tratar • Etc. Ejemplo: Experimento de reducción de tamaño de grupo de Tennessee	• Encuesta • Experimentos aleatorios • Cuasi-experiments • Etc. Ejemplo: Seguimiento a Experimento TN

¿Qué está pasando?

"Si quieres saber qué está pasando, tienes que salir y ver qué está pasando" – Yogi Berra

- A menudo característico de la evaluación formativa
- Invita varios tipos de descripción:
  - Caracterizar a una población
  - Describir el alcance y la gravedad de un problema desde varios puntos de vista
  - Desarrollar una teoría o conjetura
  - Identificar cambios al paso del tiempo
- También puede incluir asociaciones entre variables, como las características de las escuelas (e.g., tamaño, ubicación, bases económicas) que se relacionan con (digamos) el ofrecimiento de instrucción musical y artística.

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press

¿Por qué tan pocas mujeres que empiezan carreras en CTIM terminan trabajando en esos campos?

- En su momento, varias explicaciones diferentes:
  - Las mujeres no estaban bien preparadas antes de llegar a la universidad;
  - Se les discriminaba en la universidad;
  - No querían competir con los hombres por los empleos.
- Primeros estudios de casos etnográficos de un pequeño grupo de mujeres de primer año en dos universidades públicas residenciales: la mitad en cada campus planeaba carreras tradicionales para mujeres
- Con base en un análisis de los datos etnográficos obtenidos de un año de observación y entrevistas abiertas con las participantes, se desarrollaron modelos para describir cómo participaron las 23 mujeres en la vida universitaria
- Compromiso con el trabajo académico era la principal razón para persistir y predijo el comportamiento real de todos los 23 casos.

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press

Evaluación Nacional del Progreso Educativo (NAEP)

- La simple recolección de datos no es por sí misma científica. Es la organización y análisis rigurosos de los datos para responder a preguntas claramente especificadas que forman la base de la descripción científica, no los datos mismos.
- NAEP encuesta y describe el desempeño de alumnos de 4º, 8º, 12º grados en una variedad de materias académicas, incluyendo matemáticas y lectoescritura, así como información sobre antecedentes.
- Existen métodos estadísticos y psicométricos modernos para resumir esta serie compleja de datos en informes sobre el aprovechamiento académico y su relación con otros factores. Esta combinación de rigurosa recolección de datos, análisis e informes es lo que distingue la descripción científica de la observación informal.

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press

¿Existe un efecto sistemático?

- Los diseños de evaluaciones que intentan identificar efectos sistemáticos tienen en su origen una intención de establecer una relación de causa y efecto
- La labor causal se construye tanto sobre la teoría como sobre los estudios descriptivos
- La búsqueda de efectos causales no se puede llevar a cabo en un vacío: idealmente, una fuerte base teórica, así como extensa información descriptiva existen para proporcionar el fundamento para entender las relaciones causales
- Para la *evaluación sumativa*, por ende, un programa debe de haber pasado por un periodo de desarrollo y estar en una situación consistente de operación (por ejemplo, > = 3 años) antes de hacer pruebas de efectos causales

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press

¿Reducir el tamaño del grupo mejora el aprovechamiento?

- Estudio de Tamaño de Grupo de Tennessee—Experimento aleatorio
- Los maestros de primaria de todo el estado se dividieron en forma aleatoria en 3 condiciones (en escuelas de tamaño adecuado)
  1. Un maestro en un grupo normal (22-26 alumnos)
  2. Un maestro y un ayudante de maestro en un grupo normal (22-26 alumnos/2)
  3. Un maestro en un grupo pequeño (13-17 alumnos)
- Se identificó un efecto causal favorable sólo en la condición 3, especialmente para alumnos pertenecientes a minorías
- Fue la base para un reforma escolar, pero ¡TN no la llevó a cabo!

Fuente: Shavelson & Towne (2002). *Scientific Research in Education.* National Academy Press.

¿Cómo o por qué está pasando?

- Búsqueda de mecanismos causales
- Múltiples métodos aplicables
- Tamaño de grupo de Tennessee
  - Comprendieron que había un efecto
  - No pudieron precisar los mecanismos que le dieron origen
  - Siguieron múltiples años de investigación

Lo que importa: la política, la medición & el diseño

- **La política importa**: Cualquiera que sea el objeto de evaluación:
  - Está integrada en múltiples contextos
  - Y cuando una política a gran escala está en juego, la política importa muchísimo
  - Ignora la política a tu cuenta y riesgo
- **La medición importa**: Cualquiera que sea el objetivo (constructo) que interesa, distintas formas de medición pueden producir resultados diferentes: hay que alinear confiabilidad, validez y utilidad con el propósito planeado
- **El diseño importa**: Las diferentes formas como se diseñan las mediciones para abordar el tema de evaluación—y sus premisas subyacentes—pueden producir resultados muy distintos

La política & la evaluación importan: Sistema de Evaluación del Aprendizaje de California (CLAS)

Reforma sistémica que alinea resultados de aprendizaje de los alumnos con plan de estudios (indagatorio) con evaluaciones alternativas (desempeño)

Unidad de evaluación

- Enfoque sobre evaluación experimedntal e innovadora para evaluación sumativa
- Muestra matriz de resultados de prueba de opciones múltiples en anteriores CAP en mientras
- Recabar información adicional sobre evaluación formativa integrada en los salones de clase
- Con moderación combinar información
- En última instancia (10 años) usar la evaluación externa como una "auditoría" de la información recabada e integrada en el plan de estudios

Gobernador

- Promesa de campaña: El estado recabará y proporcionará resultados individuales de las preubas de los alumnos e informará a los padres de familia (votantes)
- Deseaba una prueba común de opciones mútliples para cada alumno, sin muestreo matriz
- Deseaba experimento con evaluación alternativa, siempre que lo permitiera el tiempo y el dinero
- Se molestó por las protestas de facciones políticas (e.g., la derecha religiosa)
- Afirmó que CLAS se equivocó; despido del jefe de la unidad de evaluació; el Estado volvió a un sistema de evaluación empleado dos generaciones antes

Lo que importa: premisas subyacentes del diseño: valor agregado

- Al usar el valor agregado para evaluar a maestros, colegios y universidades
  - La política importa: ¡exigencia de responsabilidad!
  - La evaluación importa: vs. "¡cualquier número sirve!"
  - Las premisas subyacentes del diseño de valor agregado importan: ¡dudosas afirmaciones causales!
- Mediciones de valor agregado (MVA)
  - Un instrumento de política muy delicado
  - Cuando se ejerce de manera inadecuada, probablemente haga más mal que bien

Algunas premisas clave en la medición del valor agregado

- Las mediciones de valor agregado pretenden proporcionar estimados causales de los efectos de la universidad en el aprendizaje de los alumnos; se quedan cortas
- Se conocen muy bien las premisas para hacer inferencias causales de los datos observados (e.g., Holland, 1986; Reardon & Raudenbush, 2009)

- *Susceptibilidad de manipulación*: En teoría, los estudiantes podrían quedar expuestos a cualquier tratamiento (i.e., asistir a cualquier universidad).
- *No interferencia entre unidades*: El resultado de un estudiante depende sólo de su asignación a un tratamiento dado (e.g., no existen efectos de pares).
- *La premisa métrica*: Los resultados de las pruebas están en una escala de intervalos.
- *Homogeneidad*: El efecto causal no varía como función de alguna característica de los estudiantes.
- *Tratamiento sumamente ignorable*: La asignación a un tratamiento es esencialmente aleatoria después de condicionar sobre variables de control.
- *Forma funcional*: La forma funcional (típicamente lineal) empleada como control de las características del estudiante es la correcta.

Algunas decisiones clave en la medición del valor agregado

- ¿Cuál es el tratamiento & comparado con qué?
  - Si la universidad A es el tratamiento, ¿cuál es el control o la comparación?
  - ¿Cuánto dura el tratamiento (e.g., 3, 4, 5, 6, + años)?
  - ¿Qué tratamiento nos interesa?
    - ¿Enseñanza-aprendizaje ajustando por *efectos de contexto*?
    - ¿Enseñanza-aprendizaje con contexto de pares?
- ¿Cuál es la unidad de comparación?
  - ¿Institución o universidad o carrera (asumir mismo tratamiento para todos)?
  - Elección práctica entre precisión de definición de tratamiento y tamaño suficiente de muestra para estimación
  - Los estudiantes cambian de carrera/universidad: ¿A qué tratamiento se atribuyen los efectos?

Algunas decisiones claves en la medición de valor agregado (Continúa)

- ¿Qué se debe medir como resultados?
  - ¿Habilidades genéricas (e.g., pensamiento crítico, resolución de problemas) en general o de una carrera? ¿Conocimientos específicos de una materia y resolución de problemas?
  - ¿Como debe medirse?
    - Respuesta elegida (opción múltiple)
    - Respuesta elaborada (ensayo de argumentación con justificación)
    - Etc.
  - ¿Qué tan válidas son las mediciones cuando se traducen para evaluaciones entre países?
- ¿Qué covariantes deben usarse para hacer ajustes con el fin de dar cuenta del sesgo de selección?
  - Covariantes individuales: resultados paralelos anteriores a la prueba con resultados de la prueba
  - Múltiples covariantes: cognitivos, afectivos, biográficos (e.g., SES)
  - Efectos de contexto institucional : puntuación promedio anterior a la prueba, SES promedio
- ¿Cómo lidiar con "clasificacion" de estudiantes (por habilidad y otros elementos)? ¡La elección de asistir a una universidad "no es casual!"

¿Todas estas preocupaciones importan?: ¡Datos de Colombia!

- ¡Sí!
- Datos (>64,000 estudiantes, 168 IHE (instituciones de educación superior) y 19 Grupos de Referencia, como ingeniería, derecho y educación) del singular sistema de evaluación de universidades de Colombia
  - Todos los alumnos del último año de preparatoria toman el examen de ingreso a la universidad: SABER 11: lengua, matemáticas, química, y ciencias sociales)
  - Todos los graduados de la universidad hacen un examen de egreso : SABER PRO: razonamiento cuantitativo (RC), lectura crítica (LC), redacción, e inglés, además de exámenes sobre materias específicas
- Enfoque sobre habilidades genéricas de RC y LC

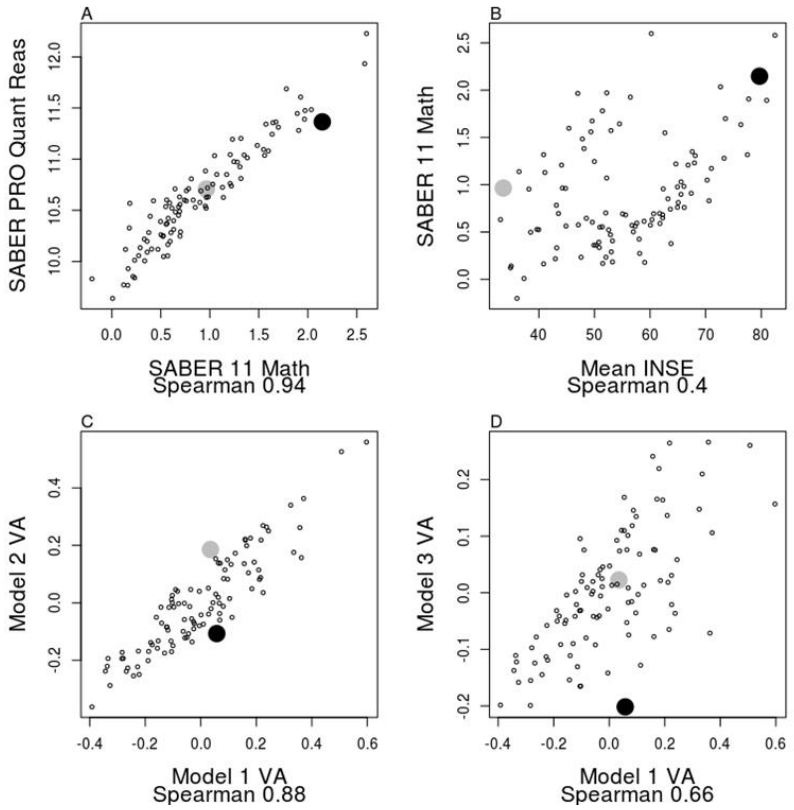
Estimadción de modelos de valor agregado

- Modelo jerárquico de 2 niveles de efectos mixtos
  - 1. Estudiante dentro de grupo de referencia
  - 2. Grupo de referencia
- Covariantes:
  - Nivel individual
    - Vector SABER 11 de 4 calificaciones debido a temas de confiabilidad
    - SES (INES)
  - Nivel de grupo de referencia
    - Media SABER 11 *o*
    - Media INSE

- **Modelo 1**: Ningún efecto de contexto: i.e., ninguna media SABER 11 ni INSE
- **Modelo 2**: Contexto con media INSE
- **Modelo 3**: Contexto con media SABER 11

Mediciones VA: ¡Instrumentos Delicados!

- Impacto en escuelas de ingeniería
  - Punto negro: Escuela de "matrícula de alta calidad"
  - Punto gris: Escuela de "matrícula de calidad promedio"



Generalizaciones de los hallazgos

- Exámenes de materias SABER PRO en derecho y educación
  - Estimados VA no son sensibles a variación en medición de resultados genéricos v. materias específicas
  - Mayores diferencias entre universidades (ICCs) con resultados de materias específicas que con resultados genéricos
- Evaluación AHELO de Habilidades Genéricas
  - Estimados VA con equivalente AHELO a los encontrados con exámenes SABER PRO
  - Menores diferencias entre universidades (ICCs) en resultados de habilidades genéricas AHELO que en resultados SABER PRO

MVA y evaluación de maestros

Como medicion única o como medición decisiva de "calidad docente" o "efectividad docente", no conviene la MVA

- Los estimados MVA han resultado ser inestables entre modelos estadísticos, años, y grupos que un maestro enseña
- Múltiples factores impactan los resultados de los avances de aprendizaje de los estudiantes dentro de las escuelas, y no pueden desentrañarse adecuadamente
  - Maestro actual + maestros anteriores + maestros que enseñan materias distintas
  - Condiciones escolares (e.g., pares, liderazgo, apoyo docente, calidad curricular, tutorías y otros apoyos estudiantiles, tamaño de grupo)
  - Condiciones extraescolares (e.g., barrios, capital social)
- Múltiples factores impactan incluso más los avances de aprendizaje de los estudiantes entre escuelas

Gracias