

Resultados del estudio de validez, de las pruebas ENLACE y Excale

Eduardo Backhoff Escudero
México, D.F. 12 de febrero de 2014.



Presentación



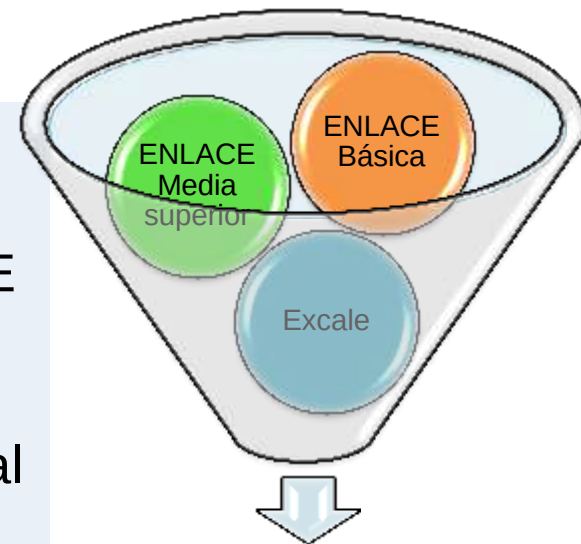
¿Para qué un estudio de validez?

- ✓ La continuación o suspensión de una política educativa debe obedecer a una **evaluación sólida** sobre su naturaleza e impacto.
- ✓ Toda recomendación que se emita a los diseñadores de las políticas educativas debe estar **sustentada en evidencias**.

Por ello:

Se optó por realizar un **estudio de carácter técnico** de **ENLACE** y **Excale**, que incluyera recomendaciones tanto para las autoridades del INEE como de la SEP.

- ✓ Con el fin de que las nuevas pruebas de aprendizaje que formen parte del Sistema Nacional de Evaluación Educativa contribuyan al propósito de ***mejorar la calidad de la educación mexicana por la calidad de su diseño y aplicación, así como por el uso que se haga de sus resultados.***



Planea



¿En qué consistió?



- ✓ **Convenio con la UAA** (finales de 2013) para coordinar el estudio como agente externo.
- ✓ **Participación de 10 especialistas** nacionales y de otros países para evaluar las evidencias de validez.
- ✓ **Comprendió cuatro etapas:**
 - ✓ Diseño del estudio y metodología
 - ✓ Estudio e informe de ENLACE Básica
 - ✓ Estudio e informe de Excale (Básica)
 - ✓ Estudio e informe de ENLACE Media Superior
 - ✓ Informe integral con conclusiones y recomendaciones
- ✓ **El informe final se entregó al INEE en diciembre de 2014.**



¿Qué es la
validez de una
prueba

=

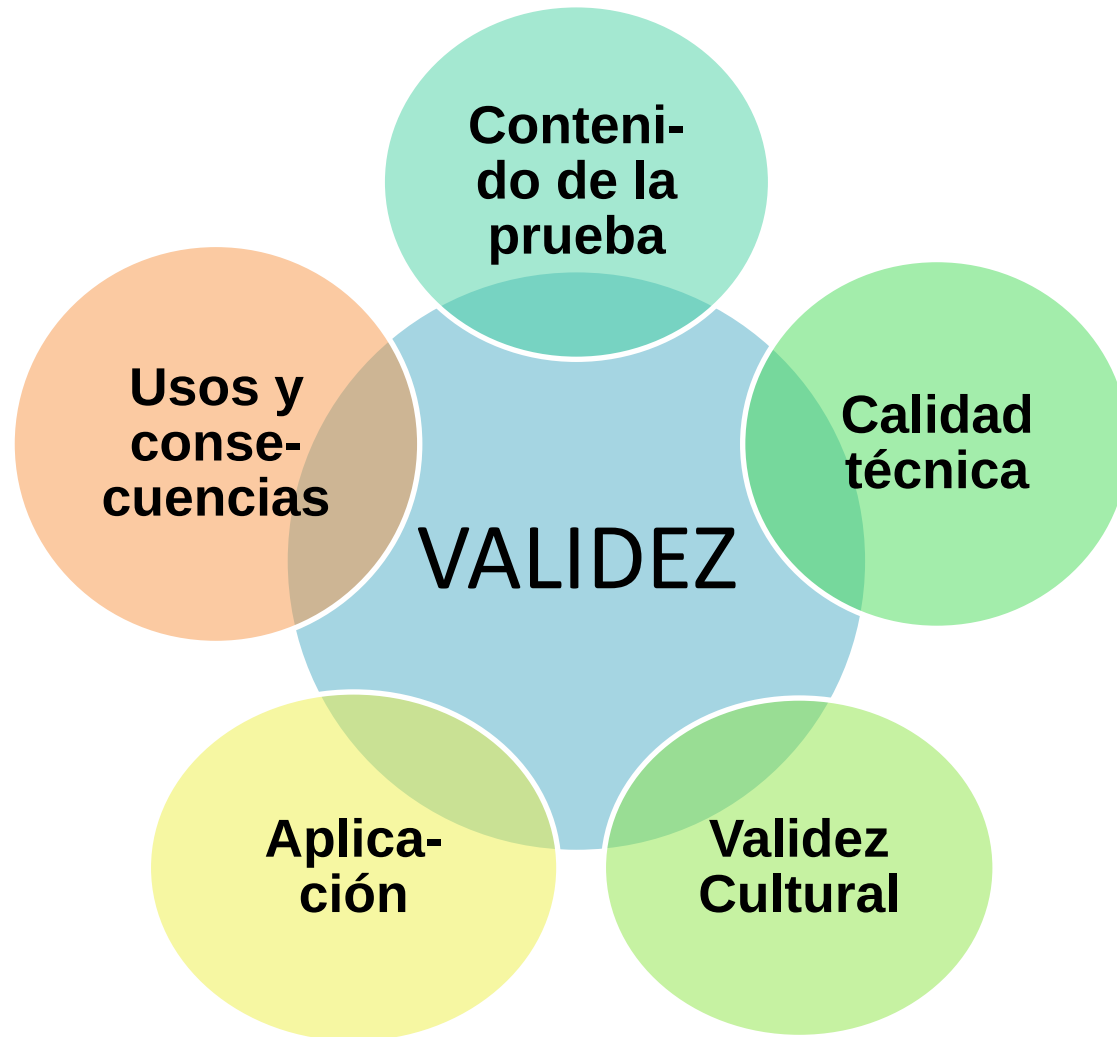
En términos generales, una prueba es válida si mide lo que pretende medir.

Por ejemplo:

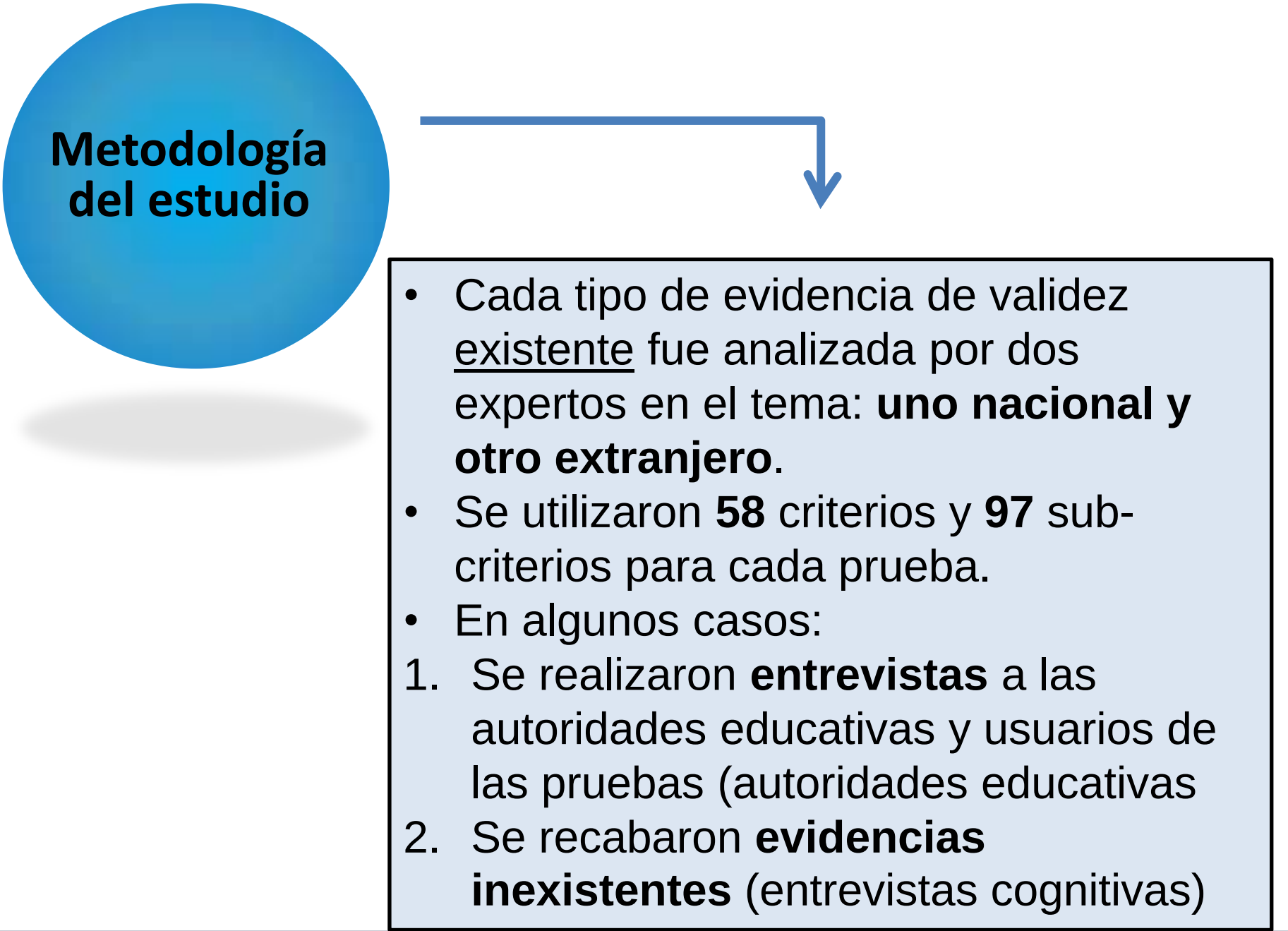
- Una prueba de inteligencia
- Una prueba de comprensión de lectura de 3er grado

*En términos técnicos, la validez de un instrumento se refiere al **grado en que distintas evidencias respaldan las interpretaciones y usos que se hacen de los resultados de una prueba.***

El estudio consistió en buscar **evidencias de 5 áreas de la validez**



Metodología del estudio



- Cada tipo de evidencia de validez existente fue analizada por dos expertos en el tema: **uno nacional y otro extranjero**.
- Se utilizaron **58** criterios y **97** sub-criterios para cada prueba.
- En algunos casos:
 1. Se realizaron **entrevistas** a las autoridades educativas y usuarios de las pruebas (autoridades educativas
 2. Se recabaron **evidencias inexistentes** (entrevistas cognitivas)



- ✓ El informe final del Comité de validez comprende una descripción detallada para cada una de las pruebas, según los 58 criterios y 97 sub-criterios evaluados por el comité.
- ✓ Algunos de los resultados más destacados en cada área evaluada:

- 1. Alineación a los referentes** (Currículo)
- 2. Aspectos psicométricos** (Estándares internacionales)
- 3. Atención a la diversidad** (Ausencia de Sesgo Cultural)
- 4. Aplicaciones** (Control de la información en todas sus etapas)
- 5. Usos y consecuencias** (Alineación con sus propósitos)

Alineación a los referentes (1)

¿Las pruebas estuvieron alineadas al referente (currículo nacional/Marco curricular Común)?

¿Las pruebas atienden adecuadamente los campos formativos, asignaturas y niveles de demanda cognitiva considerados en los planes y programas de estudio?

1. Diferencias importantes entre las tres pruebas. En **ENLACE** (Básica y Media Superior [MS]) debido a su diseño censal, se evalúa un **dominio reducido del currículo**.
2. En **Excale**, por ser muestral y con un diseño matricial, evalúa un **dominio curricular más amplio**.

Alineación a los referentes (2)

3. Notables diferencias en la calidad de la documentación de cada prueba, sobretodo en lo que se refiere a las evidencias de que cada reactivo se alinea a lo que se supone que se desea evaluar.
4. *En Excale, las evidencias son extensas y detalladas; en ENLACE Básica no fue homogénea entre años, asignaturas; y en ENLACE MS, las especificaciones de los reactivos fueron pobres.*
5. En cuanto a la elaboración y validación de reactivos: las tres pruebas siguen lo que recomienda la literatura, pero hubo diferencias en cuanto a calidad, especificidad y suficiencia. Se requiere mejorar.

Aspectos técnicos (1)

¿Cómo fue la calidad de los reactivos que integran las pruebas?, ¿son equivalentes las versiones de las pruebas?, ¿qué tan confiables son?, ¿se utilizaron técnicas psicométricas modernas y apropiadas?, ¿están bien documentadas las evidencias técnicas?

1. Las pruebas presentan manuales técnicos, pero la documentación que proporciona cada una difiere en la calidad y cantidad de la información proporcionada.
2. Se detectaron insuficiencias en ENLACE debido a la premura de los tiempos de entrega de resultados.

Aspectos técnicos (2)

3. No se reportan de manera sistemática estudios/análisis que exploren que los reactivos que integran las pruebas no favorecen o desfavorecen a ciertos grupos sociales (género, etnia, región, etc.) conocidos como estudios DIF.
4. ENLACE Básica llevó a cabo análisis para identificar la copia, pero sin aparentes consecuencias administrativas.
5. Todas las pruebas cuentan con un sistema de Banco de reactivos pero no se presenta en todos los casos documentación detallada de ellos.
6. No se presentan evidencias de validez y confiabilidad de los niveles de logro y puntos de corte

Atención a la diversidad (1)

¿Las pruebas tomaron en cuenta las diferencias culturales, lingüísticas y socioeconómicas que hay entre los estudiantes mexicanos?

1. No se encontró documentación que determine que las calificaciones obtenidas con el mismo conjunto de ítems son confiables y válidas para grupos de estudiantes de distinto grupo étnico, lingüístico, y nivel socioeconómico.
2. Las especificaciones de ENLACE no fueron suficientes para precisar características gráficas, textuales y contextuales.

Atención a la diversidad (2)

2. No hubo evidencia de que se revisen aspectos lingüísticos y fuentes de sesgo en ENLACE.
3. En un microanálisis, en las tres pruebas se encontraron problemas lingüísticos y de contenido.
4. No existen mecanismos formales y sistemáticos para eliminar reactivos con sesgo, ni de corrección de sesgo.
5. Solo Excale hizo estudios pilotos con poblaciones indígenas, para prevenir sesgos, pero no es suficiente
6. Ninguna prueba hizo estudios de validez de invarianza.

Aplicaciones (1)

¿Cómo fueron aplicadas las pruebas? ¿Tuvieron controles efectivos para evitar su filtración? ¿Cómo se cuidó y documentó cada etapa del proceso, desde la impresión hasta el levantamiento de la información?

- Existen procedimientos para asegurar la calidad de la aplicación y de control a nivel aula, escuela y entidad.
- Se levantan reportes firmadas por aplicador, coordinador, director y, en algunos casos, padres de familia.
- Sin embargo se percibe que se requieren controles de calidad más estrictos en fases cruciales que involucren verificaciones aleatorias por terceras personas (ENLACE Básica)

Aplicaciones (2)

- Existe calidad en el diseño de los instrumentos, en la lectura de resultados y su calificación.
- No existen controles de calidad estrictos en las fases de aplicación donde se involucre a personal externo (caso de ENLACE Básica).
- No se percibe que se hayan tomado medidas para desmotivar irregularidades, como en los casos escuelas con patrones de copia (ENLACE Básica)
- Debe haber estrategia para proteger integridad de la información de las pruebas, considerar formas de fraude, y valorar la extensión de los cuestionarios de contexto.

Usos y consecuencias (1)

¿Cómo se difundieron los resultados de las pruebas?

¿Qué uso se dio a los resultados?

¿Cuáles consecuencias han traído consigo su uso en el sistema educativo?

1. ENLACE enuncia un gran número de usos que se describen con diferente especificidad, pero el diseño de las pruebas no se alinea a todos los usos enunciados.
2. Las 3 pruebas no parten de un modelo que precise los usos previstos; los recursos necesarios y disponibles; la diseminación de resultados y capacitación de usuarios; consecuencias esperadas y formas de determinar si los usos previstos se dan realmente.

Usos y consecuencias (2)

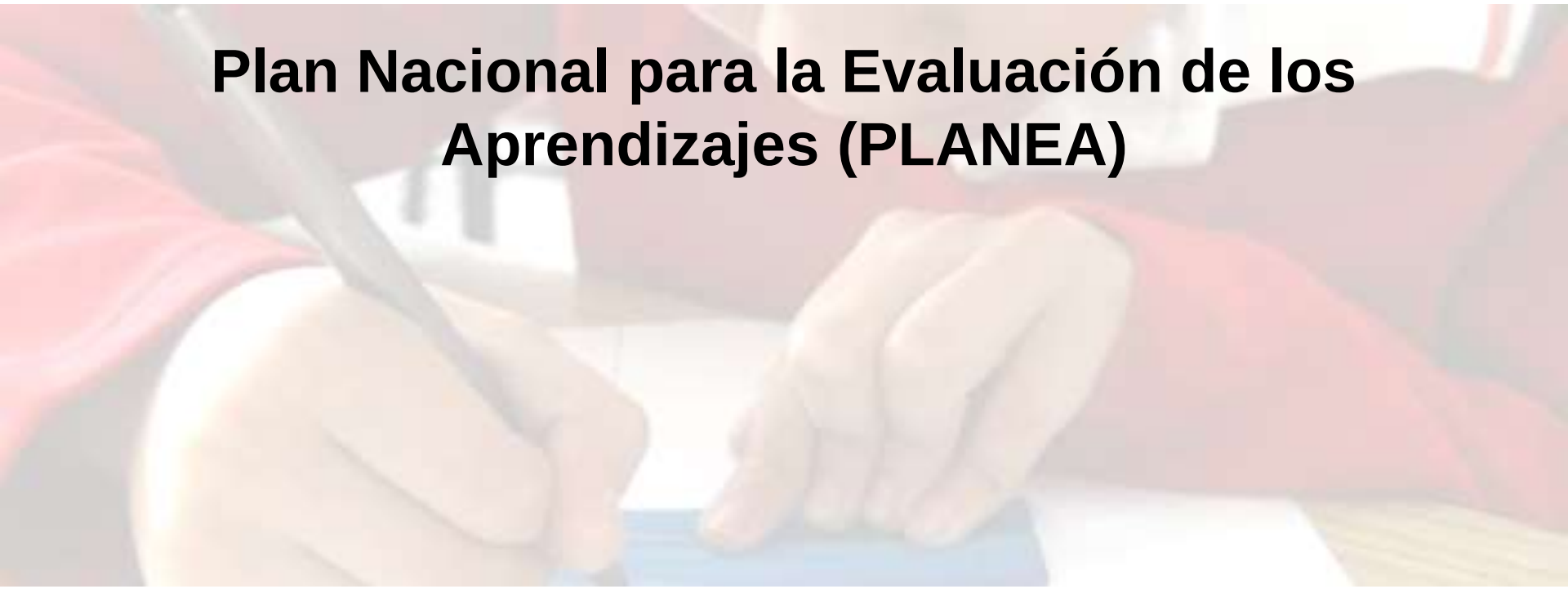
- En el caso de Excale, los resultados no son oportunos.
- Los resultados deben llegar en tiempo y forma que propicie usos propuestos, considerando la diferente necesidad y capacidad de cada grupo de usuarios.
- No se aprecia coordinación entre SEP/INEE y los estados, para fomentar la capacidad de interpretar los resultados de las evaluaciones.
- Los resultados de Excale perdieron visibilidad por la atención que generó ENLACE.
- Los planes de operación deben incluir procesos de capacitación para interpretación y uso de resultados *(por ejemplo: maestros, periodistas, tomadores de decisión)*.

Usos y consecuencias (3)

- No hay mecanismos suficientes para facilitar el acceso a bases de datos para investigadores, ni apoyo para usuarios de bases, sólo hay casos aislados de uso de bases.
- No hay esfuerzo sistemático para ver si se dan usos-consecuencias previstas o imprevistas, positivas o negativas. *(Existen estadísticas gruesas de acceso a portales, pero no hay diagnóstico de tipos de usuarios y usos de la información.)*
- No hay mecanismos de reacción cuando se detectan usos de los resultados que son injustificados y perniciosos.
- En ENLACE se observa una infación de resultados a lo largo del tiempo.

Finalmente, el Comité de validez emitió una serie de recomendaciones en cada área que fue evaluada, las cuales fueron tomadas en cuenta para el diseño del:

Plan Nacional para la Evaluación de los Aprendizajes (PLANEA)



Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement. (1999). *Standards for educational and psychological testing*. Washington: Authors.

Por su atención, gracias.



Instituto Nacional para la
Evaluación de la Educación

México

Av. Barranca del Muerto 341
Col. San José Insurgentes
Deleg. Benito Juárez, CP. 03900
México, D.F. Tel. 5482 0900

www.inee.edu.mx



INEE MEXICO



@INEEMX



INEE MEXICO



+ INEE