# Pruebas de selección y pruebas para evaluar escuelas: nuevas consideraciones sobre su uso y la difusión de sus resultados

**Felipe Martínez Rizo***

**CUADERNO No. 21**


Instituto Nacional para la
Evaluación de la Educación

***Colección Cuadernos de Investigación***

ISSN 1665-9457

*Director General del Instituto Nacional para la Evaluación de la Educación

Este texto puede consultarse en: www.inee.edu.mx

**MÉXICO, MAYO, 2006**

# CONTENIDO

## INTRODUCCIÓN

En julio de 2005, el Consejo Técnico del Instituto Nacional para la Evaluación de la Educación (INEE) adoptó un documento, preparado por el Director General, en el que se establece el criterio del Instituto sobre la difusión de resultados de evaluaciones desagregados hasta el nivel de cada escuela en lo individual (Martínez Rizo, 2005).

El documento fue publicado como el N° 15 de la serie de *Cuadernos de Investigación* del INEE, y puede verse en la página Web del Instituto (web.inee.edu.mx). En él se reitera el compromiso del INEE con los principios de transparencia y rendición de cuentas; se precisa que los resultados deberán difundirse de manera que se tengan en cuenta las características de cada evaluación, sus alcances y límites, para evitar inducir a error a quienes los consulten.

Esta toma de posición se hizo necesaria por la persistente presencia, en algunos medios, de posturas críticas respecto al INEE que lo acusaban de negarse a difundir los resultados de su trabajo. El INEE, a su vez, según el punto de vista desarrollado ampliamente en el documento referido, consideraba que las críticas se debían al desconocimiento de las características técnicas de las evaluaciones a cargo del Instituto, que hacen inapropiada una difusión desagregada al nivel de escuela, a diferencia de otras evaluaciones cuyas características pueden dar sustento adecuado a tal tipo de difusión, que en ese caso es absolutamente apropiada.

En septiembre de 2005, el Centro Nacional de Evaluación para la Educación Superior difundió un texto, titulado *Resultados educativos: la secundaria (2002-2003)*, (Ceneval, 2005). El contenido de la obra es un conjunto de ordenamientos de escuelas secundarias, basados en los resultados de las pruebas de ingreso a la enseñanza media superior denominadas *Exani I*, que aplica anualmente el Ceneval.

Los ordenamientos comprenden algunos centenares de secundarias públicas y privadas de todo el país, y se hacen en varias formas: según el promedio de aciertos de los alumnos de cada escuela que presentaron el *Exani I*; según el número de alumnos dentro del uno por ciento de más alto desempeño del país; y según el puntaje más alto obtenido por un alumno o alumna de la escuela.

En enero de 2006 el Ceneval publicó un segundo volumen con título y contenido similares, con los resultados de las pruebas *Exani I* aplicadas en 2003-2004 (Ceneval, 2006).

La publicación de la primera de estas dos obras fue interpretada por algunas personas como una muestra de la apertura del Ceneval, en contraste con la supuesta cerrazón del INEE respecto a la difusión de resultados; pero el seguimiento del tema por parte de los medios de comunicación fue limitado. La difusión de la segunda obra, en cambio, fue seguida por una fuerte reacción crítica respecto al INEE en algunos medios de comunicación, al asociarse el tema con el cambio del titular del Ceneval, quien lo atribuyó a su decisión de difundir los ordenamientos de secundarias mencionados.

Ante esta situación, la Dirección General del INEE reiteró su postura de compromiso con una difusión de resultados responsable, que deberá ser a la vez transparente y congruente con las características de la evaluación de que se trate, al tiempo que expresaba su absoluto respeto a las decisiones de las autoridades del Ceneval, de las cuales el INEE es completamente ajeno.

La Dirección General procedió además a analizar las obras mencionadas, de lo que se derivó el texto que se presenta a continuación, enriquecido por los comentarios del Consejo Técnico, quien revisó una versión previa en su reunión de los días 17 y 18 de marzo de 2006.

Este análisis complementa el anterior que sirvió de base para fijar la postura del Instituto respecto a la difusión de resultados desagregados por escuela (cfr. supra), ya que ahora se considera un caso particular, el *Exani I*, el cual permite precisar los principios generales planteados en aquel. Por esta razón se difunde este nuevo texto, esperando que contribuya a enriquecer la cultura de evaluación que poco a poco va desarrollándose en el país.

## Propósitos y características de una prueba

El punto de partida del análisis siguiente es el principio general que se plantea en el documento sobre la difusión de resultados por escuela adoptado por el Consejo Técnico del INEE al que se ha hecho referencia:

*El INEE considera que la decisión de difundir o no resultados de evaluaciones desagregados para cada escuela debe tomarse con base en las características técnicas de las evaluaciones, que se derivan a su vez de los propósitos de cada evaluación. Cuando se pretende evaluar escuelas en lo individual, los instrumentos y las aplicaciones deben reunir características precisas que hacen posible el cumplimiento de ese propósito; si se pretende en cambio evaluar al sistema educativo como tal, los instrumentos y la aplicación deberán tener otras características que harían inadecuado el que los resultados se utilizaran para un propósito diferente* (Martínez Rizo, 2005: 5).

Para aplicar ese principio general al caso de los ordenamientos de secundarias del Ceneval, el análisis debe considerar las características particulares que tienen las pruebas cuyo propósito es la selección de alumnos para su ingreso al siguiente nivel de estudios,

en contraste con las que pretendan utilizarse para evaluar las escuelas de las que egresan determinados alumnos.

### *Pruebas de selección*

El propósito de una prueba de selección es, desde luego, ofrecer elementos a autoridades educativas e instituciones para tomar decisiones sobre el ingreso a planteles o programas de cierto nivel de aspirantes que hayan terminado el nivel anterior.

Teniendo presente tal propósito, es importante señalar que los alumnos de rendimiento muy alto o muy bajo pueden ser detectados con cualquier tipo de prueba, pero no sucede lo mismo con los alumnos de rendimiento medio, que generalmente forman la mayoría de los sustentantes de una prueba de selección cualquiera. Por ello este tipo de pruebas tienden a evitar el uso de ítems demasiado fáciles o demasiado difíciles, que serán respondidos por casi todos los sustentantes o sólo por muy pocos, privilegiando, más bien, preguntas de dificultad media, porque son las que ayudarán más a distinguir a los alumnos de rendimiento medio. Por ello, en forma congruente con su propósito las pruebas de selección suelen tener las siguientes características:

- *Se diseñan con referencia a una norma estadística*, ya que se busca comparar a los sustentantes entre sí, para elaborar ordenamientos según una distribución normal, que pueda usarse para aceptar a unos con preferencia a otros. La construcción en referencia a la norma estadística maximiza la posibilidad de distinguir sustentantes de niveles de competencia medios, que son los que plantean problemas al momento de la selección.
- *Su contenido es principalmente de habilidades*, con base en las investigaciones que han mostrado el valor predictivo de ellas. El rendimiento en habilidades, además, suele tener una distribución normal. En caso de incluir conocimientos, estas pruebas tratarán también de incluir preguntas de dificultad media, dado su propósito de selección, y no darán importancia al grado en que los contenidos sean representativos del conjunto del currículo enseñado en el nivel anterior.
- *Cuidan la comparabilidad de resultados de aspirantes a un mismo programa* usando la misma forma, o formas bien equiparadas, para todos los que están en tal situación.
- *El error estándar de medición y el intervalo de confianza tienen importancia sólo a nivel individual*, ya que deben compararse resultados de unos sustentantes con otros. Los casos de diferencias no significativas estadísticamente, o *empates técnicos*, son relevantes sólo cuando ocurren alrededor de los puntos de corte, que determinan quiénes acceden a los estudios deseados y quiénes quedan fuera.

- *Cuidan que presenten la prueba todos los aspirantes a ingresar al mismo programa*, y no importa que lo hagan todos los egresados de una misma escuela.

***Pruebas para evaluar escuelas***

En contraste con lo anterior, si el propósito de una evaluación es llegar a juicios sobre la calidad de la escuela de origen de los sustentantes, las pruebas de rendimiento que eventualmente se utilicen deberán tener otras características. El diseño que deberán tener las pruebas que tengan este segundo propósito se distinguirá por los siguientes rasgos:

- *Deberá tener un referente criterial* preciso, que no es otro que la competencia mostrada por los sustentantes en relación con los objetivos de aprendizaje del currículo utilizado en las escuelas que se pretende evaluar. Como no se pretende distinguir entre alumnos de rendimiento medio para propósitos de selección, no importa que muchos alumnos obtengan resultados elevados o, por el contrario, que sea el caso de muy pocos; o sea que no se obtenga una distribución normal.
- *Privilegiará el contenido de conocimientos*, buscando que los que incluya la prueba cubran el espectro del currículo, con cuidado especial de la *sensibilidad a la instrucción* de las preguntas de la prueba; se minimizará, en cambio, el componente de habilidades, ya que la distribución de éstas entre los alumnos no se debe tanto a la escuela como a los antecedentes personales y familiares de cada alumno. En otras palabras, se buscará incluir los conocimientos que se enseñan en la escuela, sean fáciles o difíciles, tratando de evitar los que pueden adquirirse más bien fuera de las aulas.
- *Cuidará la comparabilidad de resultados de alumnos de todas las escuelas de origen que se quiera comparar*, utilizando la misma forma para todos ellos, o empleando técnicas rigurosas de equiparación.
- *Deberá manejar el error estándar de medición y el intervalo de confianza a nivel de escuela*, ya que las comparaciones deben hacerse en este nivel.
- *Será fundamental que los alumnos de cada escuela a evaluar que presenten la prueba sean la totalidad o una muestra estrictamente representativa del total*.

## TABLA 1. CARACTERÍSTICAS DE LAS PRUEBAS DE SELECCIÓN Y DE LAS QUE SE USEN PARA EVALUAR ESCUELAS

| Aspecto | Pruebas de selección | Pruebas para evaluar escuelas |
|---|---|---|
| Referente | Normativo, estadístico; *distribución normal* | Criterial, curricular; no importa la distribución estadística |
| Contenidos | Habilidades; si hay contenidos no importa sensibilidad a instrucción o representatividad curricular, pero sí dificultad media | Contenidos, con cuidado de su sensibilidad a la instrucción y de la representatividad curricular |
| Comparabilidad | A cuidar entre aspirantes al mismo destino | A cuidar entre sustentantes de todas las escuelas de origen a comparar |
| Error de medición, intervalo confianza | A nivel de alumno | A nivel de escuela |
| Sustentantes | Todos los que aspiran al mismo destino | Todos los de las escuelas a evaluar o muestras representativas |

La tabla 1 muestra sintéticamente que las pruebas para valorar la calidad de las escuelas de origen de determinados alumnos deberán tener características diferentes y, en general, opuestas a las que deben satisfacer las pruebas que sirven para seleccionar los mejores aspirantes para el nivel educativo siguiente.

### *La validez de las pruebas*

Otra forma de abordar la cuestión es considerando la validez de las pruebas, que es una de las cualidades básicas de cualquier instrumento de medición, y suele definirse como la propiedad de un instrumento que consiste en que mida efectivamente lo que se pretende medir. Precisando la definición anterior, debe recordarse que, en sentido estricto, *la validez no es una cualidad del instrumento en sí mismo*, con abstracción del uso que se haga de los resultados de su aplicación, *sino de las inferencias que se hagan a partir de tales resultados*.

En el sistema educativo muchas veces la cuestión de selección no se plantea en el tránsito de los alumnos de un grado o nivel a otro; suele ser el caso del tránsito de un grado al siguiente en el mismo nivel, o de un nivel a otro cuando ambos se ofrecen en una misma escuela. Así, los alumnos que terminan la primaria o la secundaria en una escuela que ofrece la secundaria o la preparatoria, respectivamente, generalmente transitan automáticamente de un nivel a otro sin que medie proceso alguno de selección.

En otros casos, en especial cuando el número de solicitantes rebasa con mucho la capacidad de la institución en cuestión, es necesario algún tipo de selección. Para decidir a

cuáles aspirantes admitir pueden emplearse criterios muy distintos, como la capacidad de pago, el género, la pertenencia a cierto grupo étnico o religioso, o simplemente la cercanía del domicilio y la escuela o el hecho de que un hermano o hermana del aspirante hayan ingresado con anterioridad.

Si prevalecen principios de equidad o igualdad de oportunidades, sin embargo, es frecuente que se prefiera emplear como criterio para la admisión el de la mayor o menor *capacidad* de los aspirantes, en el entendido que de ella dependerá la probabilidad de éxito en los estudios que se desea comenzar. En esta situación, si un aspirante muestra tener más capacidad que otro, deberá ser preferido, independientemente del nivel general de sus respectivas escuelas de procedencia, ya que se trata de una decisión relativa a cada persona en lo individual: sería el caso de un aspirante que proviene de una escuela excelente, pero él o ella muestran menor capacidad que otro u otra que proviene de una escuela mediocre. *Se trata de medir la capacidad individual de los aspirantes, y no la calidad institucional de las escuelas de origen*.

Por esa razón los contenidos de las pruebas de selección privilegian las habilidades, además de tener referente normativo y las demás características mencionadas. Lógicamente, estas pruebas dan importancia a la *validez predictiva*, reflejada en índices satisfactorios de correlación entre el puntaje de los sustentantes en la prueba y sus calificaciones en los estudios a los que acceden.

En el caso de pruebas cuyos resultados se usarán para evaluar escuelas, en cambio, lo que se pretende medir es la calidad institucional de las escuelas en cuanto tales. Ésta deberá reflejarse en los resultados de los egresados, pero con múltiples salvedades, debido principalmente a la influencia del entorno en los resultados.

En efecto, el nivel promedio de conocimientos que muestren en conjunto los egresados de cierta escuela se debe en parte a ella (por la calidad de sus maestros, el cumplimiento de planes y programas, las metodologías de enseñanza y aprendizaje, la organización y el clima escolar, entre otros elementos), pero también a factores que no controla, en especial el nivel de los alumnos que recibe y la calidad de los apoyos que tienen en el hogar y el entorno social en el que viven. Es posible que una escuela haya hecho un mejor trabajo con alumnos de entorno muy desfavorable, en comparación con otra que recibe alumnos privilegiados, pese a lo cual, la media de los resultados de los egresados de la segunda puede seguir siendo superior.

En esta circunstancia, la validez deberá cuidarse prestando atención a indicadores de otras dimensiones de la calidad de una escuela, que los resultados de sus egresados no necesariamente reflejan; es el caso de la selectividad de cada escuela al ingreso, de la proporción de alumnos de primer ingreso que llegan al final del nivel en la misma institución, o bien, la proporción de los alumnos que terminan el nivel en una escuela habiéndolo comenzado en otra, aspectos que no se atienden en el *Exani I*.

Los altos resultados de algunas escuelas pueden deberse más al hecho de que seleccionan el ingreso de alumnos de gran capacidad, y no necesariamente a lo que la escuela misma aporte. Para sustentar bien este tipo de juicios deberían controlarse cuidadosamente otras variables, de preferencia con diseños longitudinales que permitan hacer medidas del valor agregado por la escuela a los resultados finales de los alumnos, distinguiéndolos del valor aportado por el medio familiar y social de cada alumno y el atribuible a su habilidad individual.

Cuando se pretende apreciar la calidad de la escuela de origen, la validez predictiva no es importante, ya que no se busca medir la capacidad individual de los alumnos para predecir el éxito que podrán tener en estudios ulteriores, sino valorar la calidad de las escuelas de origen.

## EL CASO DEL *EXANI I*

Las pruebas *Exani I* buscan apoyar la toma de decisiones sobre el ingreso a planteles de enseñanza media superior de alumnos que están terminando la secundaria o lo han hecho previamente; son pruebas de selección.

En forma consistente con lo anterior, se trata de pruebas construidas con referencia a la norma estadística; su contenido combina habilidades (32 preguntas de 128) con conocimientos (96 preguntas, 12 para cada una de ocho áreas curriculares) sin cuidar la representatividad curricular ni la sensibilidad a la instrucción; se trata de aplicar la misma forma a todos los aspirantes de un mismo plantel; no preocupa la cantidad o proporción de egresados de cierta escuela que presenten la prueba, pero sí que lo hagan todos los aspirantes; el tema de resultados iguales o no distinguibles solamente interesa en casos individuales.

Las pruebas *Exani I* son adecuadas para su propósito de selección, ya que permiten ordenar a los sustentantes de manera razonablemente confiable, según su probabilidad de cursar con éxito el nivel educativo siguiente, especialmente si el puntaje obtenido en la prueba por los sustentantes se utiliza junto con otros elementos, como su promedio de calificaciones en el nivel anterior.

Correlativamente, las pruebas *Exani I* no tienen los rasgos necesarios para sustentar juicios sobre la calidad de las escuelas de origen de los sustentantes, las secundarias. El argumento se precisa señalando las limitaciones de los ordenamientos de secundarias basados en el *Exani I*.

### *El referente de las pruebas*

La construcción con referencia a la norma estadística hace que las pruebas del *Exani I* busquen incluir preguntas de nivel de dificultad medio, evitando las demasiado fáciles y las

demasiado difíciles. Ahora bien, si las secundarias enseñan bien ciertos contenidos curriculares básicos, de manera que casi todos los alumnos los manejan correctamente, y las preguntas al respecto no se incluyen, se perderá de vista una parte fundamental del desempeño de esas escuelas, y los juicios que se hagan respecto a ellas tendrán un sesgo importante. Algo análogo, en sentido contrario, debe decirse sobre la posible exclusión de preguntas sobre temas importantes pero difíciles, por el hecho de que no ayudan a distinguir a los aspirantes a ingresar al siguiente nivel, perdiendo de vista lo que pueden decir sobre el desempeño de las escuelas de origen.

### *El contenido y la cuestión de la validez*

El mayor o menor desarrollo de habilidades verbales y numéricas generales debe tanto o más a las características individuales de los sustentantes y a su medio familiar y social, que a los esfuerzos de la escuela a la que asistieron. Lo mismo ocurre con unos contenidos curriculares, más que con otros. Contenidos informativos, especialmente del campo de las ciencias sociales, pueden aprenderse más en casa, en la calle o en la televisión, mientras que otros, como los relativos a las matemáticas, deben más a la escuela, por lo general.

Por ello para valorar la calidad de las escuelas de origen de los sustentantes de una prueba es importante considerar la cobertura de sus contenidos respecto al currículo, y su sensibilidad a la instrucción. Lo anterior es particularmente necesario cuando el currículo es definido por una instancia externa a la escuela, como es el caso del Sistema Educativo Mexicano. El peso que tienen las habilidades generales en el E*xani I*, la escasa cobertura curricular de sus contenidos, y el no tener en cuenta su posible sensibilidad a la instrucción, hacen que esas pruebas no sean adecuadas para el propósito de evaluar las escuelas de origen de los sustentantes.

### *La comparabilidad*

Si bien en la mayoría de los casos los aspirantes a ingresar a la misma institución presentan el *Exani I* utilizando una misma forma del examen, –lo que no plantea problemas en lo que hace a la comparabilidad de los resultados– no puede decirse lo mismo en cuanto a la comparabilidad estricta de los resultados obtenidos por el conjunto de sustentantes de todo un ciclo escolar, que son los que se utilizan para preparar los ordenamientos objeto de este análisis. En este segundo caso sí se está en la situación de utilización de formas diferentes del examen, cuya equivalencia debería garantizarse y documentarse.

En el volumen citado relativo al ciclo 2003-2004 se reconoce que para contar con *información estadística confiable y válida que permita efectuar comparaciones en diversos niveles*

*de agregación es conveniente aplicar a los datos originales procedimientos técnicos que eliminen las eventuales –e inevitables– diferencias en dificultad entre versiones distintas del examen* (Ceneval, 2005: 37).

Se informa que "*se ha tenido cuidado de que las versiones que se aplican tengan, en general, un nivel de dificultad semejante y, aunque se reconoce la posibilidad de que la calificación global pueda estar influida por diferencias inevitables entre versiones, se cuenta con estudios estadísticos que permiten afirmar que los errores debidos a este factor son relativamente pequeños[…] Se ha eliminado, o al menos reducido hasta donde se ha juzgado suficiente, el efecto de las diferencias entre versiones, de modo que los resultados así ajustados (o igualados, para usar la terminología técnica) son comparables para fines estadísticos"* (2005: 38).

El texto de 2003-2004 indica que se perfeccionaron los esfuerzos en este sentido, avanzando en relación a lo hecho en la ocasión anterior. Esta referencia alude al apéndice contenido en la tercera parte del volumen de 2002-2003 (Ceneval, 2005: 369-377).

Sin embargo, el contenido de ese apéndice no presenta evidencias de que se hayan utilizado procedimientos rigurosos de equiparación. El análisis que se hace en el sentido de que variables como el género de los alumnos, su edad o el régimen de la escuela a la que asisten tienen mayor peso en los resultados que la versión del examen utilizada no soluciona el problema de la comparabilidad estricta de los resultados de una u otra versión.

El manejo de reactivos comunes (ancla) puede permitir una buena equiparación, pero no se ofrecen evidencias de que se haya hecho lo necesario para ello, y hay otros problemas que afectan la comparabilidad de los resultados obtenidos con versiones diferentes del *Exani I* (como el que las pruebas se apliquen en momentos diferentes del ciclo escolar).

Al respecto puede consultarse el Cuaderno de Investigación N° 10 del INEE para ver un resumen de las muchas exigencias técnicas que deben cumplirse para que pueda asegurarse la comparabilidad de los resultados obtenidos con versiones diferentes de una misma prueba (Martínez Rizo, 2004).

### *El manejo del error de medición*

Un problema más en torno del *Exani I* –si se le quiere usar para evaluar las escuelas de origen de los sustentantes– es que no se da información sobre el error de medición y el intervalo de confianza de los resultados a nivel de plantel, lo que es fundamental para valorar el nivel de significatividad estadística de las diferencias.

Cuando se ordena un número grande de escuelas (2 mil 290 y 4 mil 863 en este caso) es probable que la diferencia que separa los resultados de decenas o centenas de planteles sea demasiado pequeña para ser significativa, por lo que los ordenamientos son engañosos. En este caso es seguro que no hay diferencia entre los planteles que ocupan los diez primeros lugares de los ordenamientos difundidos, por lo que el manejo mediático del *top*

*ten* de escuelas carece de sentido. Es probable incluso que la diferencia entre la escuela del primer puesto y la del lugar cien o la que separa el lugar cien del 250, por ejemplo, tampoco sean significativas; la difusión de este tipo de ordenamientos lleva al lector no especializado a hacer interpretaciones incorrectas.

En el mismo sentido debe reflexionarse sobre los cambios de los ordenamientos de un año a otro, que pueden variar mucho sin reflejar cambios reales en la calidad de la escuela de que se trate. El hecho de que un plantel suba o baje cien lugares en el *ranking* no es significativo, si la diferencia que los separa no lo es.

En el volumen correspondiente a 2002-2003 el Ceneval indica que *los listados que incluyen nombres sólo mencionan las escuelas cuyos resultados están en los niveles de desempeño más altos del país, es decir, todas las escuelas cuyo nombre aparece en este libro ocupan una posición destacada en el ámbito nacional, sin importar en qué lugar aparezcan* (op. cit. 2005: 13).

Estas palabras son un reconocimiento de lo que se señala en este análisis, en el sentido de que las diferencias que hay entre las primeras cien escuelas, o grupos similares, no son significativas; sin embargo, el hecho de presentarlas en forma de ordenamiento induce al lector a una interpretación en sentido contrario. Por otra parte, la decisión de no dar nombres de todas las escuelas, inclusive las de resultados más bajos, es congruente con las limitaciones de los ordenamientos, pero resulta inconsistente con la decisión de difundir los resultados de la parte alta de la distribución escuela por escuela, sin advertir al lector sobre los problemas técnicos mencionados.

Aunque lo que sigue no resuelve, desde luego, los otros problemas que se están analizando, el relativo a las diferencias significativas –o no– entre las escuelas que se comparan en los ordenamientos podría reducirse si se calcularan el error de medición y el intervalo de confianza correspondiente, y si en lugar de ordenamientos con posiciones singulares (primera, segunda, y así sucesivamente) se manejaran grupos de escuelas cuyo desempeño no sea estadísticamente diferente, lo cual no hacen las obras aquí analizadas.

Debe recordarse que las mejores prácticas de evaluación en el nivel internacional incluyen la obligación de reportar resultados con información sobre el error de medición u otras medidas del margen de error o de confianza de los resultados. En este sentido el primero de los volúmenes analizados incluye un párrafo sorprendente:

*En todos los casos* [de los ordenamientos] *se muestra el resultado numérico, trátese de valores máximos o de promedios, y cada lector puede juzgar si las diferencias entre una y otra posición son suficientemente importantes*… (Ceneval, 2005: 377).

*En realidad es imposible que el lector, incluso si se trata de un especialista, pueda juzgar la importancia de las diferencias que separan a las escuelas, puesto que no se proporciona la información necesaria para ello, lo que hace pensar que no se estimó. Los lectores no especializados, por su parte, tenderán a interpretar, erróneamente, todas las diferencias como si fueran significativas, como han hecho los medios que hablan del* top ten *de las secundarias del país, o de determinada entidad federativa.*

### *El número de sustentantes*

Este último punto constituye una de las mayores debilidades del uso de resultados del *Exani I* como base para valorar la calidad de las escuelas de origen de los sustentantes. En efecto: es claro que los sustentantes no constituyen la totalidad de los alumnos de las escuelas que se pretende evaluar, y que tampoco constituyen una muestra representativa del total.

Para comprender mejor lo que sigue conviene recordar nociones elementales de muestreo.

### De poblaciones y muestras

El sentido común tiende a pensar que la representatividad de una muestra dependerá única o principalmente de su tamaño y del volumen de la población de la que se saque.

En realidad el tamaño del universo sólo influye tratándose de poblaciones pequeñas; con poblaciones grandes casi no importa. En cambio hay otros factores importantes para definir la representatividad, que son la mayor o menor homogeneidad de la población, el margen de precisión y la probabilidad de error aceptables y, sobre todo, la forma en que se obtenga la muestra, en concreto si es estrictamente aleatoria o no.

Para precisar estas ideas, la tabla 2 ofrece un ejemplo del tamaño que deberá tener una muestra para que se puedan hacer inferencias con respecto a la población con una precisión de uno por ciento y probabilidad de 95 por ciento.

El otro factor que puede influir en el tamaño de la muestra se mantiene constante: en todos los casos se supone una población de la misma homogeneidad, con una desviación estándar de cinco unidades en el ejemplo imaginario que se maneja.

**TABLA 2. TAMAÑOS DE MUESTRA PARA DIFERENTES TAMAÑOS DE POBLACIÓN**

| Tamaño de la población | Tamaño de la muestra | % de la muestra respecto a la población |
|---|---|---|
| 100 | 95 | 95.000 |
| 1,000 | 645 | 64.500 |
| 10,000 | 1,537 | 15.370 |
| 100,000 | 1,783 | 01.780 |
| 1,000,000 | 1,812 | 00.180 |
| 10,000,000 | 1,815 | 00.018 |

Cuando la población es pequeña –cien sujetos– se necesita casi toda –95– para tener una muestra con la precisión y probabilidad deseadas. Con una población de mil sujetos basta algo más de la mitad del total para una muestra de ese nivel de precisión: 645 casos. Para una población de diez mil sujetos se requiere una muestra de mil 537, poco más del 15 por ciento del total. Con poblaciones de cien mil sujetos o más, el tamaño absoluto de la muestra ya casi no cambia: mil 783 personas; para una población de un millón se necesitan mil 812 sujetos en la muestra, y sólo mil 815 para diez millones. Obviamente la proporción que representa la muestra sobre el total es cada vez menor, como muestra la tercera columna de la tabla anterior.

En todos los casos, se supone que se trata de muestras estrictamente aleatorias. Este punto es fundamental. La naturaleza aleatoria de una muestra no se obtiene simplemente escogiendo a los integrantes de la muestra en forma no intencional. Es posible que, sin pretenderlo, estén presentes factores de sesgo difíciles de apreciar a simple vista.

La aleatoriedad estricta implica que cada miembro de la población tenga una probabilidad conocida, y en principio igual, de ser incluido en la muestra, lo que supone que se cuente con un listado completo de los integrantes de la población –el marco muestral– y que se utilice un procedimiento que asegure igualdad de probabilidades de ser elegido a todos ellos. Lo anterior no se cumple con procedimientos no rigurosos como, por ejemplo, incluir en una encuesta de opinión a las personas que el entrevistador se encuentre accidentalmente al caminar por la calle; pues aunque no se tenga la intención expresa de incluir o excluir a ciertas personas, el lugar mismo donde se hagan las entrevistas puede representar un fuerte sesgo. Piénsese en la diferencia que representará buscar entrevistados en una calle de una colonia popular o en un barrio residencial exclusivo. No debe confundirse muestreo aleatorio con muestreo carente de rigor.

El caso de los grupos de voluntarios, es decir, que participan por decisión propia, es un ejemplo claro de muestra no aleatoria, que puede tener un fuerte sesgo de autoselección. Es probable, en efecto, que la mayor o menor disposición a participar voluntariamente en cierta actividad se deba a la influencia de factores que inciden en los resultados de manera sistemática.

Esto ocurre frecuentemente en pruebas de selección que se aplican a sustentantes voluntarios. En algunas ocasiones, como en escuelas privadas que ofrecen el siguiente ciclo de estudios, los alumnos no tienen necesidad de presentar la prueba de admisión, lo que elimina a un grupo de características que son seguramente diferentes al promedio de la población, probablemente con un mayor nivel socioeconómico y mejor rendimiento académico. En sentido opuesto, en el caso de escuelas públicas que no ofrecen el nivel siguiente, cuyos alumnos deben concursar si desean ser aceptados en otro plantel, es probable que los egresados de menor rendimiento, frecuentemente de nivel socioeconómico bajo, opten por no presentar la prueba, dejar de estudiar y entrar al mercado de trabajo.

Es importante señalar que todas las escuelas secundarias del Sistema Educativo Nacional tienen una población de alumnos de tercer grado relativamente pequeña, en lo que se refiere a la obtención de muestras. Las escuelas más grandes de este nivel rara vez tendrán más de trescientos alumnos en tercer grado, suponiendo la existencia de seis grupos con cincuenta alumnos cada uno. Unas cuantas en todo el país rebasarán ligeramente tal cifra, en tanto que la gran mayoría tendrá una matrícula mucho menor, en la mayoría de los casos de menos de cien personas, y en muchos otros, muy inferior, con pocas decenas de estudiantes.

Si se tiene presente lo anterior, la tabla 2 contiene una lección importante: inclusive con una muestra estrictamente aleatoria, la proporción de alumnos necesaria para hacer inferencias con la precisión y la probabilidad de error que se manejan es muy alta, generalmente superiores al noventa por ciento de la población. Esto implica que aún diferencias pequeñas entre la población y la muestra, menores al diez por ciento, implicarán márgenes de error mayores. El problema se agrava mucho si las muestras utilizadas no son aleatorias, sino fruto de la autoselección, como es el caso en casi todas las aplicaciones del *Exani I*. La selección es una de las amenazas más importantes de la validez de este tipo de estudios.

Debe añadirse que los supuestos de homogeneidad, precisión y probabilidad que se manejan en el ejemplo anterior no son particularmente exigentes, sino frecuentes en la investigación social.

**Los sustentantes del Exani I**

De 30 mil 337 secundarias registradas en el ciclo escolar 2003-2004 en México, según los datos del Ceneval, los sustentantes del *Exani I* reportaron haber estudiado en 16 mil 072 secundarias (2006: 30).

*La proporción de secundarias* con egresados que presentaron la prueba en ese ciclo escolar varía fuertemente por entidad federativa: desde 93.3 por ciento en el Distrito Federal hasta 11.4 por ciento en Zacatecas. *La proporción de alumnos de secundaria* que presentaron el examen en el ciclo de referencia va de 98.5 por ciento, nuevamente en el Distrito Federal, hasta sólo 0.3 por ciento en Nuevo León. Sólo en otras dos entidades la cifra es mayor a ochenta por ciento (Aguascalientes y Quintana Roo) y en una más (estado de México) es de 76.4 por ciento; en la entidad siguiente (Tabasco) sólo el 48.7 por ciento de los alumnos de secundaria presentó el examen (Ceneval, 2006: 31).

Dado el carácter voluntario que tiene la presentación del *Exani I* en casi todos los lugares, las secundarias no representadas son principalmente de dos tipos: escuelas privadas cuyos alumnos seguirán en el mismo plantel, o en otro semejante, sin necesidad de examen de ingreso alguno, o secundarias con alumnado poco numeroso y de origen humilde, que no pretenden seguir estudios de nivel medio superior. Por ello los resultados del *Exani I* no

pueden usarse para comparar entidades federativas; asimismo, la presencia de muchas secundarias privadas de algunas entidades en los puestos más altos de los ordenamientos del no puede interpretarse en el sentido de un mejor nivel de la entidad, ya que los sustentantes y su escuela de origen no son la totalidad ni una muestra representativa del total.

Por lo que toca a *la cantidad de alumnos de una misma secundaria* que presentaron el *Exani I* en 2003-2004, el Ceneval informa que, de las 16 mil 072 escuelas de origen, el grupo con menos de diez sustentantes estuvo formado por 8 mil 919 planteles, con 28,665 sustentantes en conjunto, lo que arroja en promedio 3.2 en cada una de las escuelas de este grupo. En otras 4 mil 863 secundarias presentaron el *Exani I* de diez a 69 alumnos, haciendo un total de 134 mil 038 sustentantes, para un promedio de 27.6 por escuela. En las 2 mil 290 escuelas restantes más de setenta alumnos presentaron el examen, con un total de 394 mil 755 alumnos y un promedio de 172.4 (2005: 30).

No hay datos sobre el alumnado de cada secundaria, por lo que incluso una cifra alta puede ser una fracción del total. En el volumen relativo a 2002-2003 se reconoce expresamente lo anterior, pese a lo cual se presentan los resultados en la forma que se está comentando:

*Dado que los exámenes se aplican a solicitud de los planteles en los cuales ingresan o aspiran a ingresar esos egresados, y no en las escuelas secundarias de donde egresan, la proporción que los examinados representan de los egresados no es la misma en todos los casos. En pocas palabras, el Ceneval no tiene datos de todas las escuelas secundarias del país, y de algunas de ellas tiene muy pocos* (Ceneval, 2005: 12).

Debe recordarse que en ningún caso se trata de una muestra seleccionada aleatoriamente, sino que, salvo en los pocos lugares en donde el examen es obligatorio, por lo que se trata de un caso evidente de autoselección. Como se ha apuntado, la no presentación del examen puede deberse a causas contrastantes, susceptibles de sesgo sociocultural: alumnos de nivel socioeconómico alto, quienes planean estudiar el bachillerato en escuelas que no requieren se haga examen de admisión, o alumnos de nivel bajo que no piensan seguir estudiando.

Como se mostró en el punto anterior, inclusive si se tratara de muestras aleatorias, proporciones muy altas de sustentantes pueden ser insuficientes para tener una precisión adecuada en el caso de poblaciones pequeñas, como son todos los casos de las escuelas secundarias. Tratándose de grupos autoseleccionados de sustentantes el problema es aún mayor. No hay fundamento para utilizar los resultados de las aplicaciones del *Exani I* para valorar la calidad de las escuelas de origen de los sustentantes, aunque sí para su propósito de selección.

Las obras que se analizan comprenden pasajes en los que se reconocen en principio las limitaciones de los datos para hacer inferencias sobre las escuelas de origen de los sustentantes, pese a lo cual se hacen.

*En el ámbito mundial, las evaluaciones externas utilizadas y estudiadas más ampliamente descansan en pruebas o exámenes estandarizados, cuidadosamente elaborados, a fin de asegurar la uniformidad de su contenido y dificultad… Estos exámenes son de varios tipos. Podemos distinguir los de aplicación por muestreo y los de aplicación por demanda… Ambas evaluaciones ofrecen ventajas y desventajas. Así, desde el punto de vista de la información que proporcionan, las del primer tipo permiten conocer las características o atributos generales de la población muestreada, pero no las particularidades de los individuos que la conforman. Con las del segundo tipo, por el contrario, se conocen los atributos de los individuos evaluados y los del grupo específico que ellos conforman, pero no los que caracterizan a la población de la que forman parte* (Ceneval, 2005: 8).

*[…]el estudio no considera las secundarias con menos de diez sustentantes y separa a las restantes en dos grupos: las escuelas con entre diez y 69 sustentantes, y aquellas con setenta o más. Si bien esta agrupación no indica cuán representativa de la institución es la cantidad de sustentantes que de ella provienen, si indica la mayor o menor presencia de las escuelas en el estudio*. (Ceneval, 2005: 31)

El hecho que pese a estas declaraciones se hagan y difundan los ordenamientos que constituyen el centro de las dos publicaciones que se analizan parece reflejar un desconocimiento o, al menos, una falta de respeto de los principios básicos del muestreo. Así lo confirma el pasaje siguiente, en el que se comparan los grupos de sustentantes del *Exani I* con las muestras del estudio de PISA aplicado en 2003. El tamaño comparativamente menor de dicha muestra, y la pequeña proporción que representa de la población de la que forma parte, se aducen como justificación del uso de los resultados de los sustentantes del Exani I para hacer inferencias relativas a sus escuelas de origen y a las entidades federativas en que se ubican.

*Los datos de las siguientes diez entidades no se definieron mediante el muestreo; el objetivo del Exani I no es efectuar un estudio del rendimiento educativo de los egresados de tercero de secundaria, pero esta circunstancia no resta validez a los datos. Admítase nuevamente el caso de la evaluación de PISA: la muestra de México en 2003 fue de 30 mil jóvenes de 15 años de un universo de más de 2 millones de jóvenes de esa edad y de un millón 270 mil que estaban matriculados en algún curso; es decir, la muestra es de 2.5% de la población de 15 años que se encontraba estudiando en el momento de efectuarse la aplicación. El porcentaje de sustentantes* [del Exani I] *en el Distrito Federal es de 98.5% y de Morelos (la entidad de la que se ofrecen detalles con el porcentaje más bajo de cobertura) de 34.2% de la población en el último año de secundaria* (Ceneval, 2006: 151-152).

Contra lo que se afirma en el párrafo anterior, y con base en lo que se ha explicado en este trabajo, debe concluirse que las características del *Exani I* y de los grupos de sustentantes a los que se aplica sí restan validez no a los datos mismos, pero sí a las inferencias que se hacen a partir de ellos respecto a las escuelas de origen de los alumnos.

## Conclusión

Sin desconocer que los trabajos del Ceneval a que se refiere este análisis tienen elementos valiosos, el análisis realizado de las dos publicaciones lleva a la conclusión de que las comparaciones de entidades federativas y subsistemas y, muy especialmente, los ordenamientos de secundarias que se presentan, carecen de sustento sólido.

Dichos ordenamientos constituyen el punto más destacado de las dos obras en cuestión, como puede apreciarse tanto por la proporción de las obras que se les dedica, como por la atención que se les ha prestado; los comentarios al respecto del Ceneval confirman tal impresión, pues se han concentrado en defender tales ordenamientos.

El análisis hecho muestra, en cambio, que esa forma de usar los resultados no es apropiada e induce a error al lector, porque las características de las pruebas y de los conjuntos de personas a las que se aplicaron no permiten hacer los ordenamientos en cuestión de manera adecuada.

La revisión del uso que se hace de diferentes tipos de pruebas en lugares con mayor tradición en evaluación de aprendizajes aporta elementos valiosos para el análisis aquí realizado. Las pruebas del SAT (*Scholastic Aptitude Test*) que, con un diseño basado en preguntas de opción múltiple se aplican desde la década de 1930 en Estados Unidos, son pruebas de selección como el *Exani I* y se usan también para comparar una dimensión particular de la calidad de las instituciones *de destino* de los sustentantes, concretamente el nivel medio de los alumnos de nuevo ingreso de universidades prestigiadas, más o menos exigentes y selectivas. Como se trata de pruebas bien equiparadas, sobre todo desde los años ochenta, permiten apreciar también el nivel general de las sucesivas cohortes anuales de egresados de *high school*; no permiten, en cambio, y no se usan para ello, valorar la calidad de las escuelas de origen de los sustentantes.

El desarrollo creciente de sistemas de evaluación basados en pruebas en todos los estados de la Unión Americana, como resultado de la ley conocida con el nombre *No Child Left Behind*, ha llevado en algunos casos a que se utilicen con propósitos de evaluación de las escuelas de origen pruebas diseñadas con otro propósito. Las reflexiones respecto a esta situación de un destacado estudioso de estos temas, quien analiza los riesgos que trae consigo ese fenómeno, pueden ser ilustrativas para el público mexicano (Cfr. Popham, 2004 y 2003).

Para terminar este trabajo, es necesario referirse a otra afirmación digna de atención, la cual está contenida en el primero de los dos volúmenes analizados:

*Las publicaciones de resultados que hemos editado y que seguiremos editando son clara muestra de que entre los riesgos de equívocos y malas interpretaciones debidos a sesgos, información incompleta o no equiparable, y la ausencia de información, el Ceneval ha optado por lo primero* (Ceneval, 2005: 13).

En contra de esta opinión, la postura del INEE considera que presentar información que ha sido construida erróneamente es peor que no dar información, ya que induce a error en la interpretación de los usuarios. A lo anterior hay que añadir que la disyuntiva entre no dar información y ofrecer una que induzca a error es falsa, ya que hay otra posibilidad, que el INEE considera la única aceptable: procesar correctamente la información, teniendo en cuenta sus propias características y los principios técnicos aplicables, y difundir los resultados precisando sus alcances y límites, de manera que se maximice la probabilidad de interpretaciones y usos correctos.

## REFERENCIAS

Ceneval (2006). *Resultados educativos: la secundaria (2003-2004)*. México. Autor.

Ceneval (2005). *Resultados educativos: la secundaria (2002-2003)*. México. Autor.

MARTÍNEZ RIZO, FELIPE (2005). Sobre la difusión de resultados por escuela. *Cuadernos de Investigación*, N° 15. México. Instituto Nacional para la Evaluación de la Educación.

MARTÍNEZ RIZO, FELIPE (2004). La comparabilidad de los resultados de las pruebas nacionales de español y matemáticas, 1998-2003. *Cuadernos de Investigación*, N° 10. México. Instituto Nacional para la Evaluación de la Educación.

POPHAM, W. JAMES (2004). *America's "Failing" Schools. How Parents and Teachers Can Cope With No Child Left Behind*. New York-London, RoutledgeFalmer.

POPHAM, W. JAMES (2003). *Test Better, Teach Better. The Instructional Role of Assessment*. Alexandria, Virginia. Association for Supervision and Curriculum Development

## *Títulos de esta Colección:*

1.- *Los resultados de las pruebas de PISA*
**Martínez Rizo, Felipe.**

2.- *Factores externos e internos a las escuelas que influyen en el logro académico de los estudiantes de nivel primaria en México, 1998-2002*
**Muñoz I, Carlos *et al*.**

3.- *Contextualización sociocultural de las escuelas de la muestra de estándares nacionales (1998-2002).*
_____ *Determinantes sociales y organizacionales del aprendizaje en la Educación Primaria de México. Un análisis de tres niveles*.
_____ *Perfil de las escuelas primarias eficaces de México (2001)*
**Fernández, Tabaré.**

4.- *Tercer Estudio Internacional de Matemáticas y Ciencias Naturales (TIMSS): resultados de México en 1995 y 2000*
**Backhoff, Eduardo y G. Solano.**

5.- *Estudio Sobre las Desigualdades Educativas en México: la Incidencia de la Escuela en el Desempeño Académico de los Alumnos y el rol de los Docentes*
**Treviño, Ernesto y G. Treviño.**

6.- *Evaluación inicial de los procesos de calibración y equiparación de las pruebas del proyecto de estándares nacionales.*
**Magriñá, Antonio.**

7.- *Factores asociados al aprendizaje del lenguaje y las matemáticas en 13 estados de México*.
**Cervini, Rubén.**

8.- *Acciones de Evaluación en las Instituciones Públicas de Educación Media Superior*
**Antonio, Rocío.**

9.- *El proyecto PISA: su aplicación en México*
**Vidal, Rafael, *et. al*.**

10.- *La Comparabilidad de los Resultados de las Evaluaciones. Importancia y Dificultad de la Equiparación*
**Martínez Rizo, Felipe.**

11.- *Marginación y rezago educativo en México.*
**Ávila, José Luis.**

12.- *Pruebas y rendición de cuentas*
**Martínez Rizo, Felipe.**

13.- *Panorama Educativo 2004. La edición 2004 de* Education at a Glance de la OCDE
**Martínez Rizo, Felipe.**

14.- *El Diseño de Sistemas de Indicadores Educativos: Consideraciones Teórico-Metodológicas*
**Martínez Rizo, Felipe.**

15.- *La telesecundaria mexicana: desarrollo y problemática actual*
**Martínez Rizo, Felipe.**

16.- *Sobre la difusión de resultados por escuela*
**Martínez Rizo, Felipe.**

17.- *Exámenes de la Calidad y el Logro Educativos (Excale):nueva generación de pruebas nacionales*
**Backhoff Escudero, Eduardo.**

18.- La Educación Mexicana en *Education at a Glance 2005*
**Martínez Rizo, Felipe.**

19.- *Metodología para evaluar la calidad de las traducciones de las pruebas internacionales: el caso mexicano de TIMSS-1995*
**Guillermo Solano-Flores, Eduardo Backhoff Escudero y Luis Ángel Contreras-Niño**

20.- *Acerca de la Validez de los Exámenes de la Calidad y el Logro Educativos (Excale)*
**María Araceli Ruiz-Primo, Jesús M. Jornet Meliá y Eduardo Backhoff Escudero**