La Comparabilidad de los Resultados de las Pruebas Nacionales de Lectura y Matemáticas, 1998-2003

Felipe Martínez Rizo

CUADERNO No. 10

Instituto Nacional para la Evaluación de la Educación

Colección Cuadernos de Investigación

ISSN 1665-9457

La Comparabilidad de los Resultados de las Pruebas Nacionales de Lectura y Matemáticas, 1998-2003

Felipe Martínez Rizo*

CUADERNO No. 10

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

MÉXICO, JUNIO, 2004

Contenido

▲ Introducción 3

▲ Asegurando la comparabilidad: el desarrollo de técnicas de equiparación 4

▲ Los diseños de equiparación 8

▲ Los métodos de equiparación 9

▲ La anomalía de lectura de las pruebas del NAEP de 1986 11

▲ Los criterios para hacer buenas equiparaciones 12

▲ El comportamiento de los resultados de las pruebas de estándares 14

▲ Las pruebas de estándares y los criterios de calidad de las equiparaciones 17

▲ Conclusiones 25

▲ Referencias 27

▲ Anexos 28
—A. Condiciones que conducen a una equiparación satisfactoria 28
—B. Estándares AERA-APA-NCME relativos a equiparación 29
—C. Puntajes en Pruebas Nacionales, según Gaviria 31

Introducción

> *La equiparación requiere que se tengan en cuenta muchos puntos prácticos. La manera en que se manejen esos puntos puede tener consecuencias muy importantes sobre la calidad de la equiparación. El proceso de construcción de la prueba que se siga, y la forma en que se diseñe el estudio de equiparación, son puntos cruciales para que pueda hacerse una equiparación adecuada. Si hay problemas con la construcción de la prueba o con los datos, entonces ninguna cantidad de manipulaciones estadísticas puede llevar a resultados adecuados de la equiparación.* (Kolen y Brennan, 1995: 292)

Las personas interesadas en educación esperan que los sistemas de evaluación informen, entre otros aspectos, sobre la evolución en el tiempo de la calidad educativa: saber si la calidad avanza o retrocede es, en efecto, tan importante para la toma de decisiones como saber si, en un momento dado, esa calidad es buena, regular o mala. Para saber si la calidad mejora, es indispensable contar con evaluaciones estrictamente comparables; evaluar el nivel de aprendizaje que alcanzan los alumnos de un sistema o subsistema educativo grande demanda instrumentos de medición, los cuales puedan aplicarse a todos los alumnos, o a muestras representativas de ellos. ¿Cómo asegurar la comparabilidad de los resultados de las evaluaciones que utilizan acercamientos a la evaluación en gran escala? El asunto es, a la vez, importante y complejo.

A partir de las décadas de 1970 y 1980, la Secretaría de Educación Pública desarrolló instrumentos de medición del rendimiento escolar, con diversos propósitos, como seleccionar alumnos para el ingreso a la educación secundaria y normal. A partir de 1998, la SEP comenzó la aplicación de pruebas de lectura y matemáticas a muestras representativas, tanto nacional como estatalmente, de los alumnos de primaria y secundaria de todo el país. Esos instrumentos, conocidos con el nombre de *pruebas de estándares*, se desarrollaron con el propósito de contar con mediciones que permitan conocer el estado del sistema y su evolución en el tiempo.

El Instituto Nacional para la Evaluación de la Educación (INEE) asumió, a partir de junio de 2003, la responsabilidad de dichas pruebas, por ser las más congruentes con la misión del nuevo organismo. En esa fecha el Instituto recibió las últimas pruebas aplicadas por la SEP en el mes de mayo de ese mismo año, para proceder a su análisis; recibió también los resultados de las pruebas aplicadas de 1998 a 2002. Como parte de su *Primer Informe Anual*, el INEE dio a conocer, desde noviembre de 2003, los resultados preliminares de las pruebas aplicadas en mayo del mismo año, así como varios estudios retrospectivos de las aplicaciones de años anteriores (cfr. INEE, 2003). Desde entonces esos resultados están a disposición del público en la página web del Instituto. Los resultados definitivos, que difieren muy poco de los preliminares, están igualmente en la página de Internet del INEE.

En el *Primer Informe Anual* se analizan los resultados de 2003, pero se advierte con claridad que no son estrictamente comparables con los de años anteriores (Cfr. INEE, 2003: 69). En el apartado relativo a los posibles avances o retrocesos en la calidad educativa se reitera la observación, y se indica que por esa razón sólo se hacen comparaciones en el tiempo con otros indicadores, más confiables, y con los resultados de las pruebas TIMSS de 1995 y 2000. (INEE, 2003: 140)

La segunda parte del *Primer Informe Anual* contiene, además, un diagnóstico global sobre la evaluación educativa hecha en nuestro país hasta 2002, incluyendo tanto las pruebas de aprendizaje como estadísticas e indicadores, evaluaciones de escuelas e internacionales. En ese lugar se mencionan con claridad las fallas de las pruebas nacionales, incluyendo lo relativo a las equiparaciones. (INEE, 2003: 45-46).

Que los resultados de las pruebas no sean estrictamente comparables es, desde luego, lamentable. Como se sabe, la importancia de que las evaluaciones informen sobre los avances o retrocesos de la calidad educativa en el tiempo es clara. Por ello es importante sustentar, de la manera más sólida posible, la opinión mencionada sobre la no comparabilidad de los resultados en cuestión. Ese es el propósito de este trabajo.

Asegurando la comparabilidad: el desarrollo de técnicas de equiparación

Supongamos que se diseña una prueba del rendimiento escolar en matemáticas, se aplica a una muestra representativa de alumnos de 6° grado y se encuentra que los alumnos obtuvieron, en promedio, una calificación de, digamos, 500 puntos.

Si un año después se diseña una nueva prueba de matemáticas y se aplica a una nueva muestra, igualmente representativa, de los alumnos de 6° grado del mismo lugar, y se obtiene un resultado promedio de 505 puntos, este hecho puede ser debido, al menos, a dos explicaciones distintas: una es que el nivel promedio de los alumnos de 6° grado ha

mejorado ligeramente de un año al otro; otra explicación es que la prueba del segundo año resultó ligeramente más sencilla que la del año anterior. Como se verá más adelante hay, desde luego, muchos otros factores que pueden afectar la comparabilidad de los resultados de pruebas aplicadas en momentos diferentes.

Una forma sencilla de evitar la dificultad consistiría en aplicar en el segundo año la misma prueba aplicada un año antes. En ese caso no habría duda de que el nivel de dificultad del instrumento sería idéntico. Esta solución, sin embargo, presenta otro problema: si se trata de la misma prueba, es posible que los alumnos a quienes se aplique en el segundo año resulten beneficiados, porque ellos mismos, o sus maestros, hayan tenido alguna información sobre las preguntas que contiene el instrumento y, por lo tanto, sus resultados se vean favorecidos, sin que haya una mejora real del nivel.

Para evitar ese riesgo, los sistemas de evaluación en gran escala que aplican pruebas periódicamente deben construir instrumentos que sean, a la vez, diferentes y estrictamente comparables. Lo anterior es fácil de decir, pero difícil de lograr. No basta, en efecto, que quienes elaboran las preguntas que formarán la segunda prueba procuren hacerlas con grados de dificultad similares a las que integraban la primera. Es necesario que se cumplan muchos otros requisitos técnicos.

El desarrollo de la psicometría puede entenderse como un largo esfuerzo para elaborar instrumentos de medición cuyos resultados, además de válidos y confiables, sean comparables entre sí, aunque se obtengan en distintos lugares y momentos. A eso apuntó, en la primera mitad del siglo XX, la construcción de preguntas de opción múltiple, las técnicas de escalamiento y estandarización y, en general, la Teoría Clásica de los Tests (TCT); en la misma dirección apuntan los aportes de la segunda mitad del siglo pasado, como la Teoría de Respuesta al Ítem (TRI), entre otros.

El uso cada vez más extendido de instrumentos estandarizados de medición mostró que, para tener resultados estrictamente comparables, no era suficiente que las pruebas se desarrollaran con base en las mismas especificaciones y que los resultados se transformaran a escalas similares. Pese a ello puede haber diferencias significativas en el grado de dificultad promedio de dos pruebas, por lo cual comenzaron a buscarse técnicas específicas para asegurar la comparabilidad. Así comenzó a hacerlo, desde 1941, el *College Board*, para la prueba de ingreso a educación superior llamada *Scholastic Aptitude Test* (SAT), y desde 1959 el *American College Testing Program* para su propia prueba (ACT). (Cfr. Arce Ferrer y Backhoff, s/f)

Las técnicas necesarias para contar con instrumentos de evaluación cuyos resultados sean estrictamente comparables, se designan con la palabra inglesa *equating*, que en español suele traducirse como *equiparación*. El hecho de que sólo recientemente se cuente con una palabra especial para designar estos procesos, y que todavía no pueda considerarse terminología estándar, debería bastar para mostrar la novedad del tema.

En ese país, que masificó antes que ningún otro el acceso a la educación superior, la necesidad de racionalizar los procedimientos de ingreso a las universidades propició la creación del *College Board*, en noviembre del año 1900. Hasta 1925 las pruebas que se utilizaban consistían de pocas preguntas de las que hoy se llaman de tipo ensayo; en 1926 se aplicó por primera vez la prueba denominada *Scholastic Aptitude Test* que, desde 1937, se componía únicamente de preguntas de opción múltiple. (Martínez Rizo, 2002: 356-357)

Sin embargo, como no se contaba entonces con técnicas de equiparación, *la constatación de que el porcentaje de sustentantes que obtenía puntajes aprobatorios variaba de un año a otro en forma considerable, hizo pensar que lo que cambiaba en realidad era el grado de dificultad de la prueba, y no el nivel promedio de los sustentantes, el cual se podía suponer más estable. Por ello, a principios de los años treinta se decidió establecer una proporción fija de aprobados, ajustando las puntuaciones del conjunto de sustentantes de cada año en consecuencia*. (Donlon, 1984, en Martínez Rizo, 2002: 357).

Como se mencionó, desde 1941 el *College Board* comenzó a cuidar sistemáticamente la comparabilidad de los resultados anuales del SAT, con procedimientos elementales de equiparación. El uso de procedimientos basados en la Teoría de Respuesta al Ítem comenzó hasta 1982. (Cfr. Arce Ferrer y Backhoff, s/f; Martínez Rizo, 2002: 357-359).

En lo que seguramente fue el primer debate público al respecto, la importancia de contar con resultados estrictamente equiparables de este tipo de evaluaciones, se evidenció en los Estados Unidos a principios de la década de 1980. Desde mediados de los años sesenta, los resultados obtenidos en promedio por los sustentantes del SAT comenzaron a ser ligeramente inferiores, cada año, a los del anterior. La tendencia a la baja continuó hasta mediados de la década siguiente, con la consiguiente alarma de educadores y ciudadanos.

La duda sobre la comparabilidad de los resultados hizo que se formara una comisión de especialistas que analizaran la cuestión (*Advisory Panel on the Scholastic Aptitude Test Score Decline*). La conclusión del estudio fue que las pruebas sí eran comparables; el descenso en los resultados debía atribuirse, probablemente, a los cambios en la composición social de los aspirantes, que cada año representaban un porcentaje creciente del grupo de edad de referencia, y cada vez incluían proporciones crecientes de aspirantes provenientes de grupos socialmente desfavorecidos.

A parte de los Estados Unidos, la psicometría se desarrolla sobre todo en otros países de habla inglesa, especialmente el Reino Unido, Canadá y Australia. Es interesante observar que, pese a que cuenta también con mediciones estandarizadas del rendimiento escolar desde la década de 1940, el tema de la comparabilidad de los resultados en Inglaterra es tema de discusiones inconclusas al día de hoy.

En un trabajo reciente sobre las evaluaciones de los alumnos de educación obligatoria en Inglaterra, Tarsh (2004) señala que *una cuestión sobre la que se discute de manera persis-*

tente, es la que pregunta en qué medida las tendencias del rendimiento de los alumnos que registran las estadísticas son genuinas, o en qué medida reflejan, por ejemplo, el impacto de cambios en los contenidos curriculares, en los métodos de evaluación… o en el que los maestros enfoquen su enseñanza a preparar a los alumnos para las pruebas. Tarsh añade: *ha habido varios estudios de los estándares a lo largo del tiempo, pero ninguno ha sido concluyente para acabar con la discusión*. (2004: 15)

Otra manera de apreciar el desarrollo reciente de las técnicas de equiparación es revisando su tratamiento en la literatura especializada. Antes de los años ochenta sólo puede identificarse un texto (de Angoff, 1971) sobre *equating* que tuviera cierto impacto, según los especialistas. A partir de esa fecha, en cambio, los estudios sobre el tema se multiplican.

Desde mediados del siglo XX varias asociaciones especializadas en evaluación educativa y psicológica, publicaron documentos con carácter normativo para los profesionales del campo. El primero de esos textos data de 1954 : *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, de la American Psychological Association (APA).

En 1955 apareció el documento *Technical Recommendations for Achievement Tests*, de la American Educational Research Association (AERA) y el National Council on Measurement in Education (NCME). En 1966 apareció un tercer documento, que remplazó a los dos anteriores, con el nombre *Standards for Educational and Psychological Tests and Manuals*, preparado por las tres organizaciones mencionadas. En 1974 se publicó una nueva versión de ese texto; en 1985 una más, y en 1999 la más reciente, siempre por la AERA, la APA y el NCME.

El tema de la equiparación de resultados de los tests no se encuentra en los textos de 1954, 1955, 1966 ni 1974. Aparece por primera vez en el documento de 1985, y de nuevo en el de 1999. En la versión de 1985, se dedican al tema de la equiparación cuatro normas o estándares: del 4.6 al 4.9 (AERA-APA-NCME, 1985: 34). En la versión más reciente, de 1999, el número de estándares relativos al tema de equiparación se duplica, llegando a ocho: del 4.10 al 4.17 (AERA-APA-NCME 1999: 57-59). En anexo se presentan los estándares relativos a equiparación de la versión más reciente del documento de AERA-APA-NCME. Su revisión bastará para mostrar la complejidad de la cuestión. En 2000 el CENEVAL publicó un documento con estándares de calidad para instrumentos de evaluación, que retoma varios de AERA-APA-NCME e incluye otros adecuados a la situación de México (Martínez Rizo *et al.*, 2000).

Si se tiene conciencia de lo reciente del desarrollo de técnicas de equiparación en el país en que surgieron, no debería sorprender que en los países de habla hispana, que no tienen una tradición psicométrica comparable, esas técnicas aún no se utilicen o hayan comenzado a hacerlo sólo en los años más recientes. El Sistema de Medición de la Calidad Educativa (SIMCE) de Chile, el más reconocido de su tipo en América Latina, que data de 1988, sólo

comenzó a emplear las técnicas modernas de equiparación en 2000, tras la introducción de los modelos psicométricos avanzados de la Teoría de Respuesta al Ítem (Comisión para el Desarrollo y Uso del SIMCE, 2003). Fuera de Chile y Brasil (SAEB), al parecer ningún otro país de Hispanoamérica cuenta hasta ahora con series de resultados estrictamente comparables en escala nacional.

Como dice María José Navas Ara:

> *…el tema de la equiparación ha pasado del reducto constituido por organismos dedicados a la evaluación y especialistas en el campo de la medida a la arena pública, al escrutinio directo de educadores, legisladores y del público en general. Este sistema generalizado de tests sometido al escrutinio público conduce a que los tests apenas si pueden ser reutilizados, se "queman" muy rápidamente y solamente se pueden volver a aplicar algunos de sus ítems, pero no la prueba como tal. Por consiguiente, se asiste a una necesidad creciente de construir formas alternativas de un mismo test cada vez con mayor frecuencia. La equiparación de puntuaciones… se está convirtiendo en absolutamente fundamental y necesaria en las instituciones y organismos dedicados a la construcción de tests. Angoff va todavía más lejos al considerar la equiparación como el tema fundamental de la psicometría*. (1996: 298).

Los diseños de equiparación

Hay básicamente dos tipos de diseño, según sea factible o no aplicar en un mismo momento, o un momento cercano, las diversas formas de la prueba que se quiere equiparar a grupos de sustentantes, los cuales reúnan características precisas.

- *Diseños con un grupo o grupos equivalentes, sin pruebas ni ítems ancla*
  Cuando es posible aplicar en forma simultánea o cercana las formas a equiparar, no es necesario que dichas formas tengan elementos comunes. La condición necesaria y suficiente para hacer la equiparación consiste en que las formas se apliquen a grupos estadísticamente equivalentes. Esto puede conseguirse de dos maneras: con dos grupos aleatorios (*random groups*); o con un solo grupo que se subdivide en dos, a quienes se aplican en diverso orden las formas a equiparar (*single group with counterbalancing*).

- *Diseños con grupos no equivalentes y pruebas o ítems ancla*
  Cuando no es posible aplicar simultáneamente las formas a equiparar, como ocurre normalmente en el caso de un sistema de evaluación que, por razones de seguridad, debe aplicar una nueva prueba cada año, la equiparación debe hacerse

mediante el uso de una prueba o conjunto de ítems comunes a las dos formas. Esos elementos comunes se designan con las expresiones de *prueba-ancla* o *ítems-ancla*, y el uso de una u otros da lugar a las variantes de este enfoque.

Los métodos de equiparación

Independientemente del diseño elegido, la comparación de las puntuaciones obtenidas en una u otra forma de una prueba, para efectos de equiparación, puede hacerse mediante diversos procedimientos estadísticos que, de manera simplificada, pueden agruparse en tres clases:

- Métodos lineales, cuando se consideran equivalentes las puntuaciones originales que den lugar a puntuaciones equiparadas estandarizadas iguales. Este método implica que la forma de las distribuciones sea igual, que los momentos de tercero (sesgo) y cuarto orden (apuntamiento) sean iguales.
- Métodos equipercentilares, cuando se consideran equivalentes puntuaciones originales cuyos percentiles sean iguales. Este método es menos preciso, pero no requiere que haya igualdad en la forma de las distribuciones.
- Métodos basados en la *Teoría de la Respuesta al Ítem*, si se consideran *puntuaciones verdaderas equiparadas las que corresponden al mismo nivel estimado en la característica evaluada por los tests*. (Navas Ara, 1996: 302).

Los tres métodos tienen múltiples variantes, y pueden utilizarse con uno u otro de los diseños de equiparación anteriormente mencionados. Pueden usarse o no técnicas de suavizado (*smoothing*) antes o después de hacer la equiparación (presuavizado o postsuavizado).

Cuando se usa un diseño con grupos no equivalentes, que utiliza una prueba o un conjunto de ítems-ancla, se puede simular un diseño de un solo grupo, utilizando *las puntuaciones en el test de anclaje V para estimar la media y la varianza de las puntuaciones que se habrían obtenido en los tests X e Y de haber aplicado las dos formas del test a todos los sujetos*. Se puede también equiparar sólo con base en la información disponible, sin hacer la estimación anterior. En el primer caso se habla de *métodos basados en el grupo sintético*, y en el segundo de *métodos basados en los datos*. (Navas Ara, 1996: 309).

Los métodos lineales de grupo sintético incluyen el de Tucker y el de Levine, este último basado en puntuaciones observadas o en puntuaciones verdaderas, y con variantes para pruebas de igual o diferente confiabilidad. Existe también un método de estimación de frecuencias. Los métodos lineales basados en los datos incluyen el de Lord, el de doble equiparación, los de anclaje predictor y de anclaje predicho, cuando se maneja un test ancla externo; cuando se usa uno interno, se tienen los métodos de Lord, Thurstone y Swineford y Fan.

En cuanto a los métodos equipercentilares, en el caso de grupo sintético hay también un método de estimación de frecuencias. En cuanto a los métodos basados en los datos se emplea el método directo de Angoff, o encadenado, y los de Braun y Holland.

Los métodos anteriores tienen en común que se basan en la Teoría Clásica de las Pruebas. Los modelos de medición más avanzados, desarrollados a partir de mediados del siglo XX por autores como Lord y Rasch, permiten acercamientos a la equiparación que, en principio pueden dar resultados más confiables.

Esos modelos, conocidos genéricamente como de rasgos latentes o de la Teoría de Respuesta al Ítem (TRI), permiten estimar parámetros de los ítems que no varían en las diversas poblaciones, por lo tanto es posible estimar, en una misma escala, la habilidad de sujetos que respondan diversos subconjuntos de ítems. Por ello es posible, en principio, hacer mejores equiparaciones. En la práctica, sin embargo, los diferentes modelos que aplican la TRI (de uno, dos o tres parámetros) pueden dar resultados más o menos precisos en diversas circunstancias, lo que también ocurre con los métodos basados en la Teoría Clásica de las Pruebas. Se distinguen métodos basados en los momentos, en la curva característica del ítem y otros.

No hay un método absolutamente superior a los otros: cada uno puede ser más o menos adecuado en determinadas circunstancias. En una obra de referencia básica sobre el tema, Kolen y Brennan (1995) señalan, por ejemplo, que unos métodos son más apropiados cuando se deben manejar muestras muy pequeñas, otros con muestras medianas y unos más con muestras grandes; la mejor o peor estandarización de las condiciones de aplicación, hace también preferibles unos métodos respecto a otros.

Referente a métodos de equiparación basados en la TRI, los autores citados consideran que los modelos de un parámetro (Rasch) son más adecuados con muestras pequeñas, en tanto que con muestras grandes son preferibles los modelos de tres parámetros. Kolen y Brennan señalan también que, para seleccionar un método, es importante saber si interesa más la precisión de los resultados alrededor de la media o lejos de ella: los métodos equipercentilares y los basados en modelos TRI, tres parámetros, son preferibles si importa tener resultados precisos a lo largo de toda la escala, mientras que los métodos basados en modelos TRI de un parámetro (Rasch) dan resultados precisos cerca de la media. (Kolen y Brennan 1995: 270-271).

La enumeración anterior, basada en Navas Ara (1996) y Kolen y Brennan (1995) no es exhaustiva, pero basta para dar una idea de la diversidad de diseños y métodos que se han desarrollado para diversas situaciones, en donde se requiere equiparar formas de una prueba para asegurar que los puntajes derivados sean equivalentes. La enumeración nos da una idea de la complejidad técnica del tema. La información siguiente lo confirma.

La anomalía de lectura de las pruebas del NAEP de 1986

La experiencia más antigua de evaluación del sistema educativo de un país es la del *National Assessment of Education Progress* (NAEP), creado en Estados Unidos a finales de la década de 1960. El NAEP desarrolló la serie de resultados comparables más importante de cualquier sistema educativo. Además del SAT y el ACT, las pruebas del NAEP han dado lugar a algunos de los desarrollos más importantes en lo concerniente a técnicas de equiparación, por los especialistas del *Educational Testing Service*, dada precisamente la necesidad de que se pudiera apreciar el avance o retroceso de los niveles de aprendizaje alcanzados por los alumnos estadunidenses.

Los resultados preliminares de las evaluaciones del NAEP correspondientes al año 1986, en particular las relativas a lectura, mostraron un comportamiento que, a la vista de educadores y estudiosos parecía extraño; los resultados *mostraban un decremento sorprendentemente grande respecto a los de 1984 en el rendimiento de los alumnos de 17 años y, en menor grado, en el de los de 9 años… Se consideró extremadamente improbable que cambios de tal magnitud en la competencia lectora hubieran ocurrido en sólo dos años, sin que la comunidad educativa tuviera conciencia de ello* (Zwick, en Kolen y Brennan, 1995: 21-22).

Dado el tiempo que toman los cambios educativos, en efecto, no es de esperarse que el nivel promedio de un conjunto muy grande, como el del sistema educativo norteamericano, se modifique significativamente de un año a otro. Por ello el ETS decidió realizar una serie de estudios para revisar los resultados; se verificó en particular la calidad de las equiparaciones.

El resultado de la revisión fueron las diferencias en las aplicaciones de 1986 respecto a las de 1984, en lo relativo al manejo de los ítems comunes a las dos formas de la prueba. En 1986, además de secciones para evaluar la comprensión lectora, la prueba incluía matemáticas y/o ciencias para los alumnos de nueve años, y computación e historia y/o literatura para los de 17. Los ítems-ancla aparecían en distinto orden y el tiempo disponible para responder fue diferente. El estudio concluyó que esos cambios explicaban la mayor parte de la diferencia de los resultados, más que el nivel de competencia de los alumnos. (Kolen y Brennan, 1995: 22).

De manera similar a lo que había ocurrido con las pruebas del SAT en los años treinta del siglo XX, la *anomalía de lectura del NAEP de 1986* lleva a la misma conclusión: dada la fuerte inercia de un sistema grande de educación, no es posible que ocurran cambios significativos, tanto en sentido ascendente como descendente, en lapsos cortos.

En una escuela y, más aún, en un grupo, pueden darse cambios bruscos: el cambio de un maestro experimentado y dedicado por otro sin experiencia o negligente, puede hacer que los resultados de los alumnos de un grupo cambien significativamente. Pero en la escala de un sistema grande no es esperable que tales cambios ocurran.

De las consideraciones anteriores se concluye un principio importante:

Si los resultados de las mediciones muestran variaciones inconsistentes a lo largo del tiempo, habrá que revisar con cuidado la calidad de instrumentos y aplicaciones, para ver si los cambios deben atribuirse a modificaciones reales en el nivel promedio de rendimiento de los alumnos evaluados, o bien a diferencias en la dificultad promedio de las pruebas.

Los criterios para hacer buenas equiparaciones

El análisis de la *anomalía de lectura del NAEP de 1986* muestra que para hacer una buena equiparación no basta utilizar diseños de investigación y métodos estadísticos adecuados, sino que deben cuidarse muchos otros aspectos.

La definición estándar de los criterios de una buena equiparación fue formulada en 1980 por Lord, retomada por Angoff y aceptada desde entonces por los especialistas. Según esta definición una buena equiparación debe satisfacer cuatro requisitos:

- *Mismo constructo*: las pruebas deben medir la misma característica, sea ésta una aptitud, una destreza, un rasgo latente u otra.
- *Invarianza de la población*: la elección de la subpoblación que se utilice para calcular la función de equiparación de los puntajes de las pruebas no afecta los resultados de la transformación.
- *Simetría*: el proceso es reversible, es decir la distribución condicional de frecuencias después de la transformación es igual si se hace sobre una u otra de las formas que se equiparan.
- *Equidad*: las puntuaciones de las pruebas, después de la equiparación, deben ser totalmente intercambiables, es decir, que sea cual sea el nivel de habilidad de los sujetos, debe ser igual aplicar una u otra prueba.

En su sentido más estricto, el cuarto requisito implicaría que, para cualquier grupo de sujetos, de cualquier nivel de habilidad, fueran iguales no sólo las medias de las puntuaciones obtenidas en las dos pruebas equiparadas, sino también la desviación estándar y las medidas de sesgo y apuntamiento, así como el error estándar de medida. Para que eso ocurra, las dos pruebas deben tener una confiabilidad perfecta, o bien ser perfectamente paralelas, lo que se reconoce como imposible. Dos pruebas que sean totalmente equivalentes deberían dar como resultado exactamente el mismo ordenamiento de los sustentantes.

En la práctica, la equiparación será mejor si la confiabilidad de las pruebas de que se trate es lo más similar que se pueda, y si al menos las medias de las puntuaciones son iguales. (Cfr. Petersen, Kolen y Hoover, 1993: 242; Martínez Arias, 1995: 637-638; Navas Ara, 1996: 300-301).

En el caso que nos interesa, de equiparación horizontal del rendimiento escolar, con formas que se aplican año tras año a sujetos del mismo grado, también es deseable que

la dificultad de las formas, estimada con base en el piloteo previo de los ítems, sea lo más similar posible.

Si se quieren comparar los resultados de alumnos de grados distintos en un mismo año, o para el seguimiento longitudinal de una cohorte a lo largo de años y grados, el tema se vuelve todavía más complejo. En este caso se habla de equiparación vertical, o de escalamiento para lograr comparabilidad. En este caso, desde luego, se supone de entrada que la dificultad de las formas será diferente. (Cfr. Navas Ara, 1996: 299-300 y 338-341).

Cada una de las numerosas variantes de los diseños y métodos de equiparación tiene exigencias particulares, pero pueden identificarse requisitos comunes a todas. Los anexos A y B, presentan dos formulaciones recientes de los requisitos para una buena equiparación:

- Una consiste en los estándares relativos a equiparación de la versión más reciente de la obra *Estándares para pruebas educativas y psicológicas*, preparada por las tres principales organizaciones americanas relacionadas. (AERA-APA-NCME. 1999).
- La otra es un listado más detallado, hecho por dos de los más destacados especialistas del tema (Kolen y Brennan, 1995: 284).

A partir de lo expuesto se elaboró la siguiente lista de criterios para una buena equiparación. Como puede apreciarse, algunos criterios son de tipo general, necesarios para cualquier instrumento de medición de buena calidad, aunque no se usaran para hacer comparaciones en el tiempo; otros criterios se refieren más precisamente a las exigencias que implica el propósito de mantener series históricas comparables.

Debe añadirse que el cumplimiento perfecto de todos los criterios constituye un ideal difícil de alcanzar. La calidad de las equiparaciones no es una variable dicotómica, la cual sólo acepte los valores extremos de buena o mala. Hay, sin duda, grados, que el listado permite precisar, para apreciar en qué medida una equiparación se acerca o se aleja de lo deseable. Los criterios se definen para lograr una buena equiparación:

1. Los contenidos de la prueba y sus especificaciones están bien definidos, y no ha habido cambios importantes en los referentes curriculares u otros.
2. En función de los propósitos que se persigan, se ha precisado también si las pruebas a utilizar se harán con referencia a la norma estadística, o a un criterio.
3. Se ha escogido el diseño metodológico y el método estadístico que se usarán para las equiparaciones, en función de las comparaciones y análisis que se quiera hacer con los puntajes transformados.
4. La extensión de la prueba es adecuada: al menos treinta ítems y de preferencia más, considerando cada dimensión o dominio a evaluar como una prueba diferente.
5. Las formas a equiparar son de la misma extensión.
6. Las formas tienen índices semejantes de confiabilidad. Para equiparaciones horizontales la dificultad es también similar.

7. Los ítems a utilizar se pilotearon y se cuenta con estadísticos de cada uno.
8. Se definieron los ítems-ancla de manera que sean representativos de la prueba en conjunto, tanto en contenido como en características estadísticas.
9. Los ítems-ancla son al menos 20 por ciento del total de ítems que tenga la prueba, o al menos treinta, en el caso de pruebas largas, en donde el 20 por ciento sería mayor. Se toma en cuenta si la prueba es unidimensional o no.
10. Los ítems-ancla son idénticos, en todas sus partes, en las formas a equiparar.
11. Cada ítem-ancla se sitúa aproximadamente en el mismo lugar en las dos formas.
12. Las características formales de las pruebas (tipografía, formato, etcétera) son adecuadas, y se mantienen iguales en las aplicaciones sucesivas.
13. Los procedimientos de muestreo son adecuados, asegurando la representatividad de los alumnos a quienes se aplican las pruebas con respecto a la población y, en su caso, a subconjuntos que se desea evaluar. Las muestras de sujetos a las que se aplican las formas antigua y nueva son comparables.
14. Hay procedimientos estándar de aplicación y medidas de seguridad, los cuales evitan que algunos sustentantes conozcan previamente la prueba o algunos ítems, asegurando la igualdad de condiciones de los sustentantes y la ausencia de sesgos. Los procedimientos se mantienen idénticos a lo largo del tiempo.
15. La aplicación de las formas tiene lugar en momentos similares del año.
16. Los procedimientos de lectura de los resultados aseguran la calidad de las bases de datos que se utilizarán en los análisis.
17. Las técnicas de calibración, escalamiento, estimación de estadísticos y cálculo de resultados son adecuados, según el propósito y las características de la evaluación, los instrumentos y las muestras de alumnos a los que se aplican.

A continuación se analizan las pruebas de estándares nacionales, en lo relativo a la comparabilidad de sus resultados en el tiempo, con base en los criterios anteriores.

El comportamiento de los resultados de las pruebas de estándares

La Tabla 1 resume las puntuaciones promedio de lectura y matemáticas obtenidas por los alumnos de los grados evaluados durante 1998 a 2003. Se presentan sólo resultados nacionales, sin desglosarlos según los servicios educativos que forman los estratos de la muestra. Aunque en 1998 hubo aplicación en secundaria, no se realizó en 1999, y en 2000 la prueba fue totalmente nueva; por ello no se ofrecen datos de 1998 para este nivel.

Un primer análisis permite apreciar si los resultados de las pruebas, a lo largo del periodo en cuestión, presentan un patrón consistente, o si muestran cambios difíciles de explicar, que deberían llevar a revisar la comparabilidad de los resultados.

En 2003, la SEP y el INEE solicitaron a un experto externo que hiciera un análisis sobre la comparabilidad de los resultados de las pruebas de estándares de lectura y matemáticas. Ese análisis llegó a conclusiones prácticamente idénticas a las de este trabajo, más allá de diferencias menores de las cifras exactas. En el Anexo C, puede verse la tabla de donde partió ese estudio. (Gaviria Soto, 2003a y 2003b).

TABLA 1
Puntajes promedio en Pruebas Nacionales

	1998	1999	2000	2001	2002	2003
Primaria						
Lectura						
2°	432.5	---	428.5	---	375.0	---
3°	418.5	427.6	---	419.4	---	456.3
4°	430.4	---	470.1	---	428.2	---
5°	453.0	459.6	---	470.5	---	479.8
6°	470.1	481.5	495.2	500.0	489.5	466.3
Matemáticas						
2°	453.8	---	428.1	---	375.8	---
3°	424.1	422.6	---	414.2	---	472.6
4°	429.1	---	465.1	---	424.1	---
5°	459.6	461.0	---	472.9	---	402.8
6°	478.2	486.3	494.7	507.0	486.0	402.0
Secundaria						
Lectura						
1°	---	---	500.6	494.9	492.9	487.6
2°	---	---	496.9	528.7	519.8	516.3
3°	---	---	514.8	---	548.1	546.6
Matemáticas						
1°	---	---	510.2	505.6	509.5	419.2
2°	---	---	507.6	517.6	520.2	433.8
3°	---	---	521.6	---	533.9	448.3

Una rápida revisión de la tabla anterior muestra variaciones importantes en los puntajes de los diversos años y grados; la manera como se presentan esos cambios, así como sus dimensiones, resultan inverosímiles, si se piensa que se refieren a un sistema educativo que tiene en cada grado algo más de dos millones de alumnos.

Leyendo los renglones de la tabla podemos ver los cambios en las puntuaciones promedio obtenidas año con año por los alumnos del mismo grado. En cuanto a lectura, los resultados de 2002 en 2° de primaria son 53 puntos inferiores a los de 2000 (375 vs. 428.5),

en tanto que los de 3° de primaria en 2003 son 37 puntos superiores a los de 2001 (456.3 vs. 419.4). En 4° parece haber otra caída muy fuerte en 2002, respecto a 2000 (más de 40 puntos) pero ese año se presentaba una aparente mejora de la misma importancia respecto a 1998 (430.4-470.1-428.2).

En 5° grado los resultados son más estables; en 6° parece haber mejora de 30 puntos de 1998 a 2001, y luego un retroceso de 11 puntos en 2002 y 25 más en 2003 (de 470.1 a 500 y luego a 489.5 y a 466.3). En matemáticas ocurre algo similar: 53 puntos menos en 2° grado entre 2000 y 2002 (428.1 vs. 375.8); 58 puntos más en 3° entre 2001 y 2003 (414.2 vs. 472.6); subida espectacular de 1998 a 2000 y caída similar de 2000 a 2002 en 4° grado (429.1-465.1-424.1); otra fuerte caída en 5° grado entre 2001 y 2003 (472.9 vs. 402.8); y caída fuerte de 84 puntoa entre 2002 y 2003 (486 vs. 402.0).

En secundaria, en lectura hay caídas insignificantes de 2002 a 2003: 492.9 vs. 487.6 puntos en 1°; 519.8 vs. 516.3 en 2°; y 548.1 vs. 546.6 en 3°. En matemáticas las caídas son muy fuertes, acercándose a los 100 puntos: 509.5 vs. 419.2 en 1°; 520.2 vs. 433.8 en 2°; y 533.9 vs. 448.3 en 3°.

Si se analizan los puntajes de la Tabla 1 en sentido vertical aparecen otras variaciones extrañas, si se considera que las pruebas de primaria estén también equiparadas verticalmente, y que en secundaria se aplica la misma prueba para los tres grados. Normalmente debería esperarse que los puntajes de un grado superior fueran siempre mayores a los de los grados inferiores, eso ocurre en la mayoría de los casos, pero no en todos: en 1998 los resultados de 3° de primaria son inferiores a los de 2° tanto en lectura como en matemáticas; en 2003 en lectura los resultados de 6° grado son algo inferiores a los de 5°. En secundaria se presenta este tipo de anomalías en 2000, con resultados inferiores en 2° grado respecto a 1°.

Se puede hacer una lectura más compleja de los datos de la tabla, en diagonal: dado que la muestra se aplicó en las mismas escuelas todos los años, aunque no con los mismos alumnos, se puede intentar una lectura longitudinal: los grupos que un año están en cierto grado escolar, un año después (en la columna siguiente a la derecha) se encuentran también en el grado siguiente, un renglón por debajo. La lectura en diagonal permite seguir una misma cohorte de alumnos a lo largo del tiempo.

En el caso de primaria, el análisis en diagonal muestra un patrón general de avances, pero de muy distinta importancia, y con algunos retrocesos. En lectura, entre el 4° grado de 2000 y el 5° de 2001 hay sólo 0.4 puntos de aparente progreso (470.1 vs. 470.5), mientras que de 2° en 2002 a 3° en 2003 se pasa de 375 a 456.3. Pero de 2° grado en 2000 a 3° en 2001 se retrocede de 428.5 a 419.4. En matemáticas se observan aparentes retrocesos de 2° a 3° en 1998 y 1999 (453.8 vs. 422.6) o de 4° en 2002 a 5° en 2003 (424.1 vs. 402.8). En secundaria se presentan descensos muy fuertes entre 2002 y 2003, tanto de 1° a 2° (509.5 vs. 433.8) como de 2° a 3° (520.2 vs. 448.3).

Al igual que en el caso del NAEP, estos cambios, tanto en sentido ascendente como descendente, no pueden atribuirse sin más a modificaciones reales del nivel promedio de competencia de los alumnos de un sistema educativo como el mexicano que, en primaria y secundaria, cuenta con más de veinte millones de alumnos, en aproximadamente 130 mil escuelas. La inercia de un conjunto de tales dimensiones es, sin duda, muy grande, por lo que no son esperables cambios bruscos importantes.

Si se considera que los casos de descensos fuertes son verosímiles, porque se cree hay razones para pensar que se está produciendo un deterioro grave de la calidad educativa, los ascensos que también se presentan en los resultados de las pruebas en algunos años y grados resultan tanto más inexplicables.

Las inconsistencias de los resultados deben llevarnos a revisar las características de instrumentos y aplicaciones de las pruebas nacionales de lectura y matemáticas, entre 1998 y 2003, para detectar si atienden adecuadamente los requisitos de calidad de las equiparaciones enlistadas antes.

Las pruebas de estándares y los criterios de calidad de las equiparaciones

De los prerrequisitos básicos se cumple solamente el cuarto; en cuanto a los otros cinco hay problemas significativos, que cuestionan la comparabilidad, como se explica a continuación:

1. No hubieron cambios importantes en los referentes de las pruebas de 1998 a 2003; los cambios de 2000 en el currículo de primaria fueron menores. Sin embargo, la definición de los dominios de contenido y el mapeo para tener especificaciones adecuadas no fueron suficientemente precisos en relación con el currículo.

2. Este punto es importante para las pruebas de estándares, las cuales teóricamente se conciben como criteriales, pues lo que se pretende con ellas no es simplemente clasificar a los sustentantes comparándolos entre sí, sino saber cuántos logran los objetivos curriculares en diversos niveles de competencia, más o menos satisfactorios. La interpretación más interesante de los resultados, es la que se hace en términos de qué porcentaje de alumnos se sitúa, año con año, en los cuatro niveles de competencia que se manejan.

Lo que se constata al revisar pruebas y resultados es que ni la planeación de las primeras, ni el análisis de los segundos, se ha hecho de manera consistente con los principios aplicables a las pruebas criteriales. Los ítems que conforman la prueba no se distribuyen sistemáticamente a lo largo de la escala de dificultad, ya que no se cuenta con estadísticos derivados del piloteo previo. Los puntos de corte de los niveles de competencia tampoco se definen con métodos apropiados para pruebas criteriales.

Por lo anterior, no es posible comparar los porcentajes de alumnos que se sitúan en uno, u otro, de los niveles de competencia a lo largo de los años en que se han aplicado las pruebas.

3. La elección del diseño metodológico (equiparación para grupos equivalentes mediante ítems-ancla) es adecuada, pero el método estadístico escogido no es el más conveniente, dadas las circunstancias de las pruebas nacionales, las cuales se aplican a muestras muy grandes y en las que se requiere comparar resultados a todo lo largo de la escala de puntuaciones.

En lugar de un método basado en el modelo de Rasch, sería más apropiado usar un método equipercentilar (no lineal) o uno basado en el modelo TRI de tres parámetros. El estudio de Magriñá (2003), relativo sólo a las pruebas de 3°, 5° y 6° de primaria de 1999, muestra que, como era de esperar, el método basado en el modelo de Rasch distorsiona los valores extremos.

Es discutible cuál es el método estadístico más apropiado para hacer la equiparación de las pruebas de que se trata, por lo cual, convendría emplear varios para poder comparar sus resultados, lo que no se hizo hasta 2003.

4. Para analizar el cumplimiento del cuarto criterio es necesario considerar primero lo relativo al número de dimensiones que miden las pruebas de estándares.

En 1996, cuando comenzaron a planearse las pruebas, se juzgó que las secciones de competencia lectora y matemáticas eran dimensiones o dominios diferentes; sin embargo, desde las primeras aplicaciones, y hasta 2002, las dos secciones se consideraron como pertenecientes a una sola dimensión. Un análisis hecho como parte del análisis de las pruebas, mostró que tal idea carece de sustento, y que las dos secciones constituyen, en efecto, dimensiones claramente diferentes. (Gaviria Soto, 2003b).

El error tuvo serias consecuencias, tanto para este criterio relativo a la extensión de las pruebas, como en cuanto al criterio del número de ítems comunes y el de los procedimientos para escalar las puntuaciones, que se analizarán más adelante.

En efecto: si las secciones de las pruebas pertenecen a una misma dimensión, la extensión total es claramente suficiente para que puedan hacerse equiparaciones adecuadas, ya que se rebasa siempre el número mínimo de treinta ítems.

En cambio, si cada sección constituye una prueba diferente, dado que mide una dimensión diferente, la extensión de cada una no es tan grande, y en algunos casos podría ser insuficiente. Es importante añadir que, debido a la ausencia de pilotaje previo (*cfr. infra*), al hacer el análisis de los resultados de las aplicaciones debieron descartarse algunos ítems, por detectarse un comportamiento estadístico anómalo. El número de ítems utilizado efectivamente para el análisis de los resultados es, pues, inferior al total contenido en los cuadernillos de las pruebas. En la Tabla 2 se presenta el número de ítems incluidos efectivamente en el análisis, después de descartar aquellos que tuvieron un comportamiento estadístico anómalo.

Como puede apreciarse en la tabla, con excepción de las pruebas de lectura en 1998, todas las demás incluyeron al menos treinta ítems que, en principio, podrían ser suficientes para hacer equiparaciones.

TABLA 2
Número de ítems de las Pruebas Nacionales

	1998	1999	2000	2001	2002	2003
Primaria						
Lectura						
2°	14	---	42	---	38	---
3°	15	37	---	38	---	58
4°	15	---	39	---	48	---
5°	17	38	---	44	---	52
6°	17	38	43	43	63	58
Matemáticas						
2°	30	---	32	---	37	---
3°	30	43	---	43	---	46
4°	34	---	48	---	46	---
5°	33	45	---	60	---	48
6°	33	41	48	58	44	46
Ciencias N.						
6°	--	--	--	--	--	--
Secundaria						
Lectura	17	---	62 (3°)	77	67	62
Matemáticas	33	---	42 (3°)	40	52	48
Ciencias N.	---	---	---	---	---	12

5. Esta tabla muestra también que el quinto criterio no se cumple: la extensión de las pruebas, en efecto, varía mucho de una aplicación a otra, aún considerando solamente las pruebas de 1999 en adelante. Inevitablemente los cambios afectan de manera importante la calidad de las equiparaciones.

6. Por lo que se refiere a la confiabilidad de las formas de las pruebas, la Tabla 3 presenta la situación de los instrumentos aplicados en primaria y la Tabla 4 la de secundaria. Para cada aplicación de las pruebas se presenta el índice de confiabilidad alfa (α) de Cronbach. Se presenta también, en cada caso, el promedio del porcentaje de aciertos que tuvieron los ítems de cada prueba.

Como se observó, el propósito de las equiparaciones no es otro sino el de asegurar la comparabilidad de los puntajes obtenidos con instrumentos cuya dificultad puede variar,

aunque haya tratado de igualarse en el proceso de construcción. Por ello el que la dificultad sea igual no puede considerarse como un requisito previo; sin embargo, tratándose de equiparaciones horizontales, el piloteo previo de ítems permite construir formas lo más parecidas que se pueda en cuanto a su dificultad; al no haberse contado con estadísticos derivados del piloteo de los ítems, la construcción de las formas no pudo cuidar este punto. Por ello se considera útil ofrecer también un índice de dificultad estimado *a posteriori*.

TABLA 3
Dificultad y confiabilidad de las pruebas de estándares, 1998-2003. Primaria

	1998	1999	2000	2001	2002	2003
Primaria						
Lectura						
2°						
Cronbach	0.712	---	0.837	---	0.800	---
% aciertos	0.622	---	0.502	---	0.436	---
3°						
Cronbach	0.635	0.741	---	0.716	---	0.863
% aciertos	0.434	0.396	---	0.379	---	43.9%
4°						
Cronbach	0.649	---	0.798	---	0.835	---
% aciertos	0.464	---	0.437	---	0.441	---
5°						
Cronbach	0.631	0.822	---	0.771	---	0.848
% aciertos	0.435	0.452	---	0.441	---	45.6%
6°						
Cronbach	0.631	0.828	0.840	0.848	0.871	0.858
% aciertos	0.464	0.501	0.475	0.504	0.449	42.6%
Matemáticas						
2°						
Cronbach	0.857	---	0.853	---	0.811	---
% aciertos	0.544	---	0.458	---	0.450	---
3°						
Cronbach	0.805	0.855	---	0.851	---	0.795
% aciertos	0.429	0.439	---	0.427	---	31.5%
4°						
Cronbach	0.738	---	0.886	---	0.608	---
% aciertos	0.391	---	0.492	---	0.346	---
5°						
Cronbach	0.683	0.711	---	0.789	---	0.638
% aciertos	0.377	0.369	---	0.389	---	31.6%
6°						
Cronbach	0.557	0.682	0.785	0.798	0.716	0.784
% aciertos	0.343	0.332	0.393	0.423	0.352	29.7%

La Tabla 3 permite apreciar que la confiabilidad de las pruebas varía en las aplicaciones de los diversos años, lo cual dificulta la realización de buenas equiparaciones. Llama además la atención que, contra lo que suele ocurrir, la confiabilidad de las pruebas de matemáticas es, en muchos casos, inferior a la de lectura, lo cual tiene que ver, desde luego, con la longitud de las pruebas.

TABLA 4

Dificultad y confiabilidad de las pruebas de estándares, 1998-2003. Secundaria

	1998	1999	2000	2001	2002	2003
Secundaria						
Lectura						
1°						
Cronbach	0.647	---	0.859	0.890	0.863	0.868
% aciertos	0.532	---	0.472	0.478	0.450	48.8%
2°						
Cronbach	0.657	---	0.896	0.900	0.878	0.886
% aciertos	0.595	---	0.540	0.540	0.506	54.4%
3°						
Cronbach	---	---	0.864	---	0.890	0.890
% aciertos	---	---	0.494	---	0.561	60.1%
Matemáticas						
1°						
Cronbach	0.610	---	0.533	0.382	0.507	0.540
% aciertos	0.378	---	0.334	0.299	0.332	33.7%
2°						
Cronbach	0.657	---	0.484	0.503	0.578	0.623
% aciertos	0.418	---	0.279	0.317	0.352	36.4%
3°						
Cronbach	---	---	0.502	---	0.644	0.683
% aciertos	---	---	0.290	---	0.378	39.2%

En el caso de las pruebas de secundaria, la confiabilidad de las pruebas de lectura es bastante homogénea y, en general, adecuada. La de las pruebas de matemáticas, por el contrario, es desigual y, en todos los casos, es muy baja para que pueda considerarse aceptable.

En lo referente a la dificultad, estimada *a posteriori* con el promedio del porcentaje de respuestas correctas de todos los ítems obtenido por los alumnos, debe tenerse en cuenta que las pruebas aplicadas cada año en los tres grados de secundaria han sido las mismas, por lo que las diferencias en su dificultad son atribuibles a diferencias de habilidad de los sujetos evaluados.

Tanto en el caso de secundaria como en el de primaria, se debe tomar en cuenta que en 2003 las puntuaciones originales, crudas, se estandarizaron con escalas separadas para la prueba de lectura y para la de matemáticas; de 1998 a 2002, en cambio, se uso una sola escala para las dos pruebas.

7. El séptimo prerrequisito, fundamental desde el punto de vista estadístico, no se cumplió: los ítems no se pilotearon antes de su aplicación, por lo que no se contó con los estadísticos correspondientes antes de armar las pruebas.

8. Dadas las fallas relativas a los criterios 1 (ausencia de especificaciones claras) y 6 (ausencia de piloteo previo), es claro que los ítems-ancla no se definieron de manera que fueran representativos de la prueba en conjunto.

9. Dado que estamos ante un caso de equiparación con grupos no equivalentes, el cual usa ítems comunes, es importante lo relativo al número de tales ítems empleado en las equiparaciones. Debe recordarse que no se deben considerar las dos partes del instrumento aplicado cada año como una sola prueba, sino que la parte de lectura y la de matemáticas constituyen una prueba cada una, ya que miden dimensiones o dominios diferentes.

La Tabla 5 resume la información al respecto. En cada casilla se anota primero el número de ítems nuevo y luego, entre paréntesis, el de ítems comunes con formas previas. Obviamente las primeras pruebas (1998) no pudieron tener ítems comunes con pruebas previas, por lo cual, entre paréntesis se pone un guión.

Las cifras de la Tabla 5 se refieren al total de ítems incluidos en las pruebas de los diversos años, tanto los que se mantuvieron en el momento de los análisis, como los descartados por su comportamiento estadístico anómalo; por ello las cifras totales no coinciden con las de la Tabla 2, que sólo considera los ítems que permanecieron para los análisis.

El total de ítems comunes de una prueba, respecto a las de años anteriores, puede estar formado por cantidades diferentes de ítems de uno u otro año, lo cual añade complejidad al tema; por ello en las cifras de ítems comunes que se ponen entre paréntesis, se indica con un superíndice el año o años a los cuales corresponden los ítems comunes.

TABLA 5
Ítems comunes en las Pruebas Nacionales

	1998	1999	2000	2001	2002	2003
Primaria						
Lectura						
2°	15 (-)	---	43 (0)	---	37 (0)	---
3°	17 (-)	42 (0)	---	0 (41^{99})	---	47 (6^{99})
4°	17 (-)	---	42 (0)	---	47 (6^{00})	---
5°	18 (-)	38 (7^{98})	---	0 (44^{99})	---	43 (8^{98} 5^{99})
6°	18 (-)	38 (7^{98})	0(38^{99}+7^{98})	0(36^{99}+7^{98})	56 (14^{99})	58 (0^{02})
Matemáticas						
2°	33 (-)	---	20 (12^{98})	---	30 (6^{98} 5^{00})	---
3°	32 (-)	36 (12^{98})	---	34 (12^{98})	---	34 (5^{98})
4°	36 (-)	---	48 (0)	---	40 (3^{98} 5^{00})	---
5°	37 (-)	28 (21^{98})	---	12 (49^{99})	---	23 (6^{98} 1^{99})
6°	37 (-)	29 (19^{98})	14 (34^{99})	13(47^{vv})	41 (8^{vv})	47 (1^{02})
3° secundaria						
Lectura	17 (-)	---	65 (0)	---	68 (6^{00})	0 (74^{02})
Matemáticas	33 (-)	---	48 (0)	---	59 (1^{00})	1 (59^{02})

Puede apreciarse que el número de ítems comunes varía mucho. En primaria se tiene, en un extremo, pruebas prácticamente idénticas en 3° y 5° de lectura en 1999 y 2001, así como en 6° en 1999, 2000 y 2001.

En el otro extremo, en lectura las pruebas de 2° en 2000 fueron completamente diferentes respecto a las de 1998, al igual que las de 2002 respecto a 2000. Lo mismo ocurre con las pruebas de 4° en 2000 y con las de 2003, tanto en lectura como en matemáticas; en el segundo caso hubo sólo un ítem común con respecto a la prueba de 2002. En lectura, además de 58 ítems nuevos se incluyeron aparentemente ocho ítems comunes; una revisión más detallada permitió detectar, sin embargo, que se trató de otros tantos ítems abiertos (de respuesta construida, en vez de opción múltiple) los cuales fueron comunes con la forma de 2002, pero que no se calificaron ni se utilizaron en ninguno de los dos años, por lo cual se trató de una prueba completamente diferente, sin ítem-ancla alguno. Por ello en la casilla correspondiente se anota un cero con superíndice 02.

En secundaria, como se mencionó, tras la aplicación de 1998 no hubo aplicación para el año siguiente, y las pruebas de 2000 fueron totalmente nuevas. En 2001 tampoco hubo aplicación en tercer grado. En 2002 se repitieron, con muy pocos cambios, las pruebas de 2001. En 2003, por el contrario, las pruebas fueron casi nuevas. Hay, pues, variación e inconsistencia considerables en el número de ítems-ancla de una aplicación a otra de las pruebas de estándares de 1998 a 2003. Dado el modelo de equiparación utilizado, este criterio

es fundamental y no se puede considerar que se cumpla. La calidad de las equiparaciones se ve inevitablemente afectada.

10. Este criterio sí se cumple: los ítems-ancla se manejaron en forma idéntica en las formas en que se emplearon.

11. Dados los cambios en el número de los ítems comunes a las diversas formas de las pruebas, y en la extensión de las mismas, es claro que la ubicación de los ítems-ancla en las formas no pudo cuidarse de manera sistemática.

12. Las características formales de las pruebas (tipografía, formato, etcétera) fueron casi iguales en los diversos años, aunque hubo diferencias menores.

13. Las muestras utilizadas en las pruebas de estándares son buenas, de modo que son representativas del conjunto del alumnado del sistema educativo y de los diversos tipos o modalidades de los servicios educativos (escuelas públicas y privadas, etcétera). Como la muestra de escuelas ha sido la misma desde 1998 hasta 2003, en principio habría fundamento para juzgar que los alumnos a quienes se aplicaron las pruebas fueron comparables. Sin embargo, como se indica en el criterio 15, las pruebas se aplicaron en unas ocasiones al fin del ciclo escolar y en otras al inicio del mismo; en estos casos, las pruebas destinadas a un grado se aplicaron a los alumnos del grado siguiente, considerándolos comparables. Este supuesto es inadecuado; la deserción que ocurre al transitar de un grado a otro no es aleatoria; esto afecta de manera especial a los alumnos con menor rendimiento. En esos casos parece claro que no se cumple con el criterio de comparabilidad de los sujetos evaluados.

14. Los procedimientos de administración de la prueba sí aseguraron la igualdad de condiciones de los sustentantes en cada aplicación; las medidas de seguridad fueron adecuadas también, por lo que no hay razones para temer la presencia de sesgos significativos. Los procedimientos no han cambiado.

15. La aplicación de las formas no tuvo lugar en momentos similares en todos los años cuando aplicaron: en 1998, 1999 y 2001, lo cual sucedió a inicios del año escolar, en octubre y noviembre; en cambio en 2000, 2002 y 2003 se aplicaron al final del ciclo escolar, en mayo-junio.

16. La precisión de los procedimientos ópticos de lectura de resultados es muy alta, por lo cual no hay razones para dudar de la calidad de las bases de datos que contienen los resultados de las pruebas aplicadas. Los cuestionarios de contexto asociados a las pruebas sí pueden verse afectados por la ausencia de controles de calidad sistemáticos de los procedimientos manuales de captura de datos y por la falta de informes al respecto, pero esas fallas no afectan las puntuaciones de las pruebas de logro.

17. En lo referente a las técnicas empleadas en diversos puntos del proceso psicométrico, sí se detectan fallas, las cuales afectan las equiparaciones:

- La ya mencionada falta de estadísticos de los ítems, por la ausencia de piloteo antes de la aplicación operativa, impidió que pudiera controlarse en forma adecuada la proporción de ítems de diverso grado de dificultad, en función de los propósitos de las pruebas.
- Como se explicó, hasta 2002, el escalamiento de las puntuaciones de los alumnos se hizo considerando que las dos materias (lectura y matemáticas) constituían una sola prueba, lo cual no es correcto, como muestra un análisis de dimensiones. En 2003 se calcularon dos escalas separadas, una para lectura y otra para matemáticas. Obviamente, esto impide comparar los puntajes de 2003 con los de años previos.
- La no difusión pública de resultados de las pruebas de estándares hasta 2002, impide saber con certeza si se hicieron análisis que utilizaran ponderadores para el cálculo de medias, desviaciones estándar y errores estándar de las puntuaciones, tanto para el total de la muestra como para los estratos de escuelas públicas y privadas; urbanas y rurales; cursos comunitarios y escuelas indígenas; secundarias generales, técnicas y telesecundarias. Los análisis de los resultados de 2003 sí usaron esos ponderadores. Las comparaciones con resultados de años anteriores pueden verse afectadas.

Conclusiones

Como resultado de la revisión presentada, se debe afirmar que las pruebas de lectura y matemáticas aplicadas de 1998 a 2003 no cumplen la mayoría de los criterios necesarios para que sus resultados puedan compararse, incluyendo varios de los más importantes para tal propósito.

Las fallas son de tal consideración, que pueden explicar los cambios bruscos e inconsistentes que se encuentran en los resultados de las pruebas. Las diferencias de 2002 a 2003, en especial, además del nulo anclaje entre las formas respectivas, se explican sin duda por la modificación del criterio para el escalamiento de los puntajes crudos obtenidos por los alumnos. Hasta 2002, tras de calibrar en una sola escala los ítems, se estimaron las puntuaciones de los sujetos en lectura, matemáticas y globalmente. En 2003 se utilizaron escalas diferentes para los ítems de lectura y matemáticas.

El no haber seguido procedimientos apropiados para la planeación de una prueba criterial, ni para el análisis de sus resultados, hace especialmente inadecuado comparar porcentajes de alumnos que se sitúen, de año en año, en los diversos niveles de competencia definidos.

Las altas y bajas que se observan en las puntuaciones obtenidas en diversos años, grados y áreas curriculares, así como en la proporción de alumnos en diversos niveles de com-

petencia no pueden atribuirse, con base en esos datos, a cambios en los niveles de competencia reales de los sustentantes. La conclusión es fuerte, pero la evidencia al respecto parece incuestionable.

Esto no quiere decir que no haya elementos valiosos en los resultados de las pruebas revisadas. Dada la calidad de las muestras utilizadas, los resultados de cada aplicación permiten hacer comparaciones entre entidades de la república y entre las modalidades de servicio que constituyen los estratos de la muestra, como ya lo ha llevado a cabo el INEE, a partir del *Primer Informe Anual*, en noviembre de 2003.

Esas comparaciones deben tener en cuenta el valor estadísticamente significativo o no de las diferencias que se encuentren; cuidando lo anterior, es posible hacer algunas comparaciones a lo largo del tiempo, no en cuanto al nivel absoluto de los resultados obtenidos en años sucesivos, pero sí de la posición relativa de los subsistemas educativos considerados, sin perder de vista el efecto de regresión a la media.

Debe añadirse algo importante, recordando la cita de Kolen y Brennan que se incluye como epígrafe de este trabajo: las deficiencias detectadas son de tal naturaleza, que hacen imposible corregir retroactivamente las equiparaciones referidas:

Si hay problemas con la construcción de la prueba o con los datos, entonces ninguna cantidad de manipulaciones estadísticas puede llevar a resultados adecuados de la equiparación. (1995: 292).

Con base en lo anterior, las fallas detectadas no pueden corregirse mediante nuevos cálculos de las escalas utilizadas o calibraciones *a posteriori*. Es posible una nueva aplicación de alguno de los instrumentos utilizados en el pasado, con lo que podrían hacerse algunas comparaciones.

Una última lección se impone: *para observar el movimiento la referencia debe permanecer inmóvil* (Gaviria Soto, 2003a). Para que el sistema educativo nacional pueda contar con resultados de evaluaciones que informen de manera válida, confiable y consistente sobre el nivel de aprendizaje alcanzado por los alumnos, y sobre su avance o retroceso real, necesitamos referentes firmes, cuya construcción implique necesariamente el respeto a las exigencias técnicas correspondientes.

La complejidad de las metodologías psicométricas contemporáneas hace que esfuerzos considerables, bien intencionados pero técnicamente deficientes, no puedan ofrecer resultados de calidad suficiente para que sean útiles. La experiencia de los países que han avanzado más en este terreno muestra también que el desarrollo de mediciones comparables implica periodos considerables.

El INEE está convencido de la importancia de aportar tal tipo de información, y está haciendo lo necesario para ello. Sólo necesita que se le dé el tiempo indispensable para una tarea delicada y compleja.

Referencias

AERA-APA-NCME (1985). *Standards for Educational and Psychological Testing*. Washington. American Psychological Association.

AERA-APA-NCME (1999). *Standards for Educational and Psychological Testing*. Washington. American Educational Research Association.

ANGOFF, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, Educational Testing Service.

ANGOFF, W. H. (1971). Scales, norms and equivalent scores. En R. L. THORDNIKE, Ed. *Educational Measurement*. Washington. American Council on Education.

ARCE FERRER, ÁLVARO J. y EDUARDO BACKHOFF ESCUDERO (s/f). *Efecto del método de igualación en la equivalencia de las versiones de exámenes nacionales*. Mecanograma.

COMISIÓN PARA EL DESARROLLO Y USO DEL SIMCE (2003). *Evaluación de aprendizajes para una educación de calidad*. Santiago de Chile. Ministerio de Educación.

DONLON, THOMAS F. ed. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York. College Entrance Examination Board.

GAVIRIA SOTO, JOSÉ LUIS (2003). *Informe de validación de los resultados de las pruebas de Estándares Nacionales, México 1998-2002*. Mecanograma.

GAVIRIA SOTO, JOSÉ LUIS (2003). *Calibración y equiparación de las pruebas de estándares nacionales*. Mecanograma.

INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN (2003). *La Calidad de la Educación Básica en México. Primer Informe Anual 2003*. México. INEE.

KOLEN, MICHAEL J. y ROBERT L. BRENNAN (1995). *Test Equating. Methods and Practices*. New York. Springer-Verlag.

LORD, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale. Lawrence Erlbaum.

MAGRIÑÁ, ANTONIO (2003). *Evaluación inicial de los procesos de calibración y equiparación de los proyectos de las pruebas de Estándares Nacionales*. Mecanograma.

MARTÍNEZ ARIAS, ROSARIO (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid. Síntesis. Cap. 20. Puntuaciones de los tests. Transformaciones. Normas. Procedimientos de Equiparación.

MARTÍNEZ RIZO, FELIPE (2002). Revisión de literatura sobre evaluación de sistemas educativos. En Varios. *La calidad de la educación en México. Perspectivas, análisis y evaluación*. México. Miguel Ángel Porrúa, pp. 305-427.

MARTÍNEZ RIZO, FELIPE *et al.* (2000). *Estándares de calidad para instrumentos de evaluación*. México. Ceneval.

NAVAS ARA, MARÍA JOSÉ (1996). Equiparación de puntuaciones. En JOSÉ MUÑIZ Coord. *Psicometría*. Madrid. Editorial Universitas. Págs. 293-369.

PETERSEN, NANCY S., M. J. KOLEN y H. D. HOOVER (1993). Scaling, Norming and Equating. En ROBERT L. LINN Ed. *Educational Measurement*. 3rd Ed. Phoenix. American Council on Education-The Oryx Press. Págs. 221-262.

TARSH, JASON (2004). Country Highlight: Pupil Assessment in England. *OECD/INES/Network A-Review of Assessment Activities*. February/March 2004, pp. 11-17.

Anexo A

Condiciones que conducen a una equiparación satisfactoria

A. Requisitos generales:

1. Se han especificado con toda claridad los propósitos de la equiparación, el grado de precisión que se requiere y si se quiere tener puntuaciones comparables a lo largo de períodos prolongados.
2. Se ha seleccionado adecuadamente el modelo para la recolección de datos, el diseño de la equiparación, el método estadístico a utilizar y los procedimientos para seleccionar los resultados, teniendo en cuenta el contexto en el cual tendrá lugar el estudio de equiparación.
3. Se han seguido procedimientos adecuados de control de calidad.

B. Desarrollo de la prueba: aspectos comunes a todo diseño de equiparación:

1. Los contenidos de la prueba y las especificaciones estadísticas están bien definidos y son estables a lo largo del tiempo.
2. Gracias a aplicaciones piloto u operativas previas, se cuenta con estadísticas de todos o casi todos los ítems a utilizar para construir la forma para equiparar.
3. La prueba es razonablemente larga (al menos treinta ítems, y de preferencia más).

C. Desarrollo de la prueba: aspectos propios de los diseños con anclas:

1. Cada conjunto de ítems comunes es representativo de la prueba en conjunto, tanto en contenido como en características estadísticas.
2. Cada conjunto de ítems comunes es suficientemente extenso (al menos 20 por ciento de la prueba para pruebas de cuarenta ítems o más; al menos treinta ítems para pruebas largas).
3. Cada item común se sitúa aproximadamente en el mismo lugar en las dos formas a equiparar. Los troncos de los ítems comunes, los materiales de estímulo y los distractores son idénticos en las formas. Se controlan otros posibles efectos del contexto en el nivel de ítem.
4. Se utiliza el principio de doble relación (*double linking*). La forma antigua se administró en el lapso de un año, y en un momento del año similar a aquél en el cual se aplica la nueva forma.

D. Grupos de examinados:

1. En el caso de modelos con grupos equivalentes, los grupos examinados son representativos de los examinados en la aplicación operacional de la prueba; son estables en el tiempo; y son relativamente grandes.
2. En el caso de modelos de equiparación con pruebas o ítems-ancla, con grupos no equivalentes, los grupos que tomen las formas antigua y nueva no son muy diferentes.

E. Administración de las pruebas:

1. Las pruebas y los ítems se guardan de manera segura.
2. Las pruebas son administradas bajo condiciones estandarizadas cuidadosamente controladas, que se mantienen idénticas en las sucesivas aplicaciones.

F. Situación de los dominios y la enseñanza:

1. El currículo, los materiales de enseñanza y/o el campo de estudio son estables.

Fuente: Kolen y Brennan, 1995: 284; adaptada de Brennan y Kolen, 1987b.

Anexo B

Estándares AERA-APA-NCME relativos a equiparación

Estándar 4.10. Cualquier afirmación de que las puntuaciones obtenidas en diferentes formas de una prueba pueden utilizarse de manera intercambiable deberá ofrecer una clara fundamentación teórica (*rationale*) y evidencia empírica que la soporte…

Estándar 4.11. Cuando las afirmaciones sobre la equivalencia de las puntuaciones de forma a forma de una prueba se basen en procedimientos de equiparación, deberá ofrecerse información técnica detallada, sobre el método mediante el cual se establecieron las funciones de equiparación u otras relaciones (*linkages*), así como sobre la precisión de dichas funciones.

Estándar 4.12. En estudios de equiparación que se basen en la equivalencia estadística de grupos de examinados a quienes se apliquen diversas formas de una prueba, deberán describirse con detalle los métodos que garanticen tal equivalencia.

Estándar 4.13. En estudios de equiparación que utilicen diseños de prueba-ancla, deberán presentarse las características de la prueba-ancla y su semejanza respecto a las formas a ser equiparadas, incluyendo tanto las especificaciones de los contenidos, como las relaciones determinadas empíricamente entre las puntuaciones obtenidas. Si se usan ítems-ancla, como se hace en algunos estudios de equiparación basados tanto en la Teoría Clásica de las Pruebas, como en la Teoría de Respuesta al Ítem, deberán presentarse la representatividad

y las características psicométricas de los ítems-ancla.

Estándar 4.14. Cuando se usen conversiones de puntuaciones o procedimientos de comparación para relacionar resultados obtenidos con pruebas o formas de una prueba que no sean estrictamente paralelos, deberán describirse claramente la forma de construcción, la interpretación buscada y las limitaciones de esas conversiones o comparaciones.

Estándar 4.15. Cuando se preparen formas adicionales de una prueba tomando un subconjunto de ítems de una forma existente, o reordenando sus ítems, y existan razones sólidas para pensar que las puntuaciones obtenidas mediante esas formas puedan verse influenciadas por los efectos del contexto de los ítems, deberá ofrecerse información de que no se produce una distorsión indebida de las normas para las diferentes versiones, o para las relaciones (*linkages*) de las puntuaciones de unas formas y otras.

Estándar 4.16. Si se cambian las especificaciones de una prueba de una versión a otra, los cambios deberán identificarse en el manual de la prueba, con la advertencia en el sentido de que las puntuaciones de ambas versiones pueden no ser estrictamente equivalentes. Cuando ocurran cambios substanciales en las especificaciones de una prueba, los puntajes deberán reportarse en una escala diferente, o bien deberá advertirse claramente a los usuarios que las nuevas puntuaciones no son directamente comparables con las obtenidas mediante versiones anteriores de la prueba.

Estándar 4.17. Los programas de pruebas que pretendan mantener una escala común a lo largo del tiempo, deberán realizar verificaciones periódicas de la estabilidad de la escala en que se reportan las puntuaciones obtenidas.

Fuente: AERA-APA-NCME, 1999: 57-59

Anexo C

Puntajes en Pruebas Nacionales, según Gaviria

	1998	1999	2000	2001	2002	2003
Primaria						
Lectura						
2°	419.3	---	427.5	---	377.5	---
3°	417.4	432.1	---	425.9	---	456.3
4°	437.2	---	475.0	---	437.6	---
5°	456.8	466.7	---	479.7	---	479.8
6°	474.5	488.4	500.9	505.0	501.8	466.3
Matemáticas						
2°	436.8	---	422.5	---	378.0	---
3°	423.9	431.7	---	424.2	---	472.6
4°	434.0	---	472.9	---	428.5	---
5°	464.9	467.2	---	480.7	---	402.8
6°	480.0	490.7	499.8	511.2	493.3	402.0
Secundaria						
Lectura						
1°	---	---	507.9	500.9	504.1	487.6
2°	---	---	506.5	535.8	533.3	516.3
3°	---	---	523.9	---	563.1	546.6
Matemáticas						
1°	---	---	515.1	508.3	513.8	419.2
2°	---	---	511.4	520.9	525.1	433.8
3°	---	---	526.1	---	540.1	448.3

Fuente: Gaviria Soto, José Luis, 2003b.