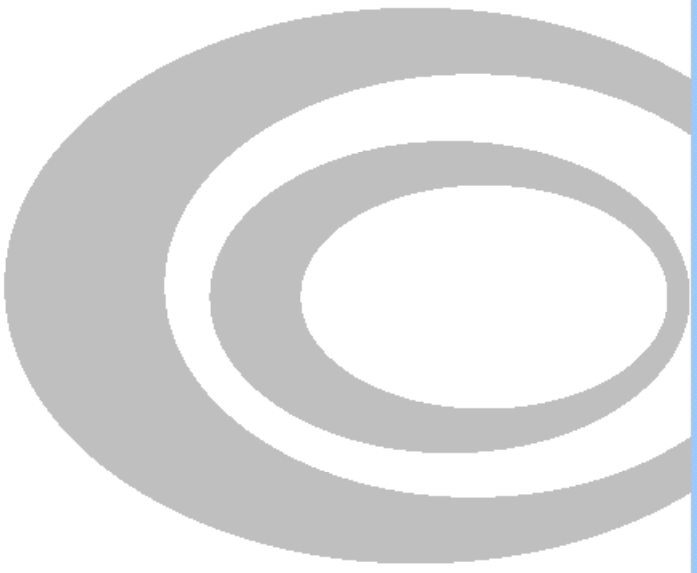


Programa de Promoción de la Reforma Educativa en América Latina y el Caribe

Grupo de Trabajo sobre Estándares y Evaluación

Sobre prueba y mediciones de aprendizaje: un breviario para la "alfabetización evaluativa"

Gerald W. Bracey

Grupo de Análisis para el Desarrollo

SOBRE PRUEBAS Y MEDICIONES DE APRENDIZAJE: UN BREVIARIO PARA LA "ALFABETIZACIÓN EVALUATIVA"

Gerald W. Bracey

Edición en español: Lima, setiembre 2002

Grupo de Trabajo sobre Estándares y Evaluación del PREAL y GRADE

Presentación

Gerald Bracey es un prolífico analista de las políticas educativas y defensor de las escuelas públicas de los Estados Unidos.

Este breve texto pretende dar la información mínima necesaria para quienes lean informes o artículos sobre resultados de pruebas de aprendizaje escolar comprendan mejor sus contenidos y cuenten con algunos criterios para interpretar correctamente su sentido.

Aunque la audiencia para quien escribe el autor es el público norteamericano, gran parte de su contenido resulta pertinente para los ciudadanos de América Latina, que empezamos recién a participar de la "cultura de la evaluación".

Agradecemos a Gerald Bracey la autorización brindada para traducir al español y difundir este texto, originalmente publicado en el año 2000 por *American Youth Policy Forum* (www.aypf.org).

Grupo de Trabajo sobre Estándares y Evaluación de PREAL y GRADE.
http://www.grade.org.pe/gtee-preal
Lima, setiembre del 2002

TABLA DE CONTENIDO

31. ¿Qué es un estándar?
32. ¿Qué es un estándar de contenido? ¿Qué es un estándar de desempeño?
33. ¿Qué es alineamiento?
34. ¿Qué es dar credenciales?

PARTE III **:** Algunas cuestiones sobre pruebas

1. ¿Por qué es el “enseñar para las pruebas” un problema en un contexto educacional pero no en un contexto atlético?
2. ¿Quién desarrolla pruebas?
3. ¿Qué agencias supervisan el uso debido de las pruebas?
4. ¿Por qué causan tantos problemas los coeficientes de correlación?
5. ¿Por qué no hay un promedio nacional con significado para el *SAT* o *ACT*?
6. ¿Por qué disminuye el puntaje promedio del *SAT*?
7. ¿Por qué fue "recentrado" el *SAT*?
8. ¿"Funcionan" el *SAT* y el *ACT*?
9. ¿Dependen demasiado las universidades del *SAT*?
10. ¿Por qué es necesaria la "alfabetización evaluativa"?

SOBRE PRUEBAS Y MEDICIONES DE APRENDIZAJE: UN BREVIARIO PARA LA "ALFABETIZACIÓN EVALUATIVA"

Gerald W. Bracey

INTRODUCCIÓN

La necesidad de una "alfabetización evaluativa"

Las pruebas educacionales empezaron a emerger gradualmente en la conciencia pública aproximadamente desde 1960. Hace cuarenta años, la gente no prestaba mucha atención a las pruebas. Pocos estados manejaban programas estatales de pruebas. La Evaluación Nacional de Progreso Educativo (NAEP) no existiría hasta algunos decenios después. Los puntajes del *SAT (Scholastic Aptitude Test)* no habían empezado su declive de dos decenios. Sólo los orientadores, los funcionarios de admisiones de las universidades y una minoría de estudiantes que querían ir a la universidad prestaban atención a esos puntajes del *SAT*; pocos más lo hacían. No había estudios internacionales que aplicaran pruebas a estudiantes de diferentes países. Solo Denver tenía una prueba de "competencia mínima" como requerimiento para graduarse de la secundaria.

Ahora, las pruebas están por doquier. Miles de estudiantes de la ciudad de Nueva York asistieron a clases de verano en un intento por elevar sus puntajes en las pruebas lo suficiente como para pasar al cuarto grado. Un número de colegios en Nueva York fueron descubiertos haciendo trampa en diversas formas, debido a la presión sobre los puntajes de las pruebas. Los expertos están discutiendo si la política de Chicago de retener a los estudiantes que no logran puntajes suficientemente altos es un éxito o un fracaso. Se critica al Consejo de Educación del Estado de Massachusetts por establecer un puntaje aprobatorio demasiado bajo en las pruebas estatales. El Consejo de Educación de Virginia está lidiando con cómo rebajar el puntaje de corte excesivamente alto de Virginia sin parecer que también está rebajando los estándares. Arizona reprobó al 89% de sus estudiantes en la primera aplicación de su nuevo programa de pruebas. Las pruebas están siendo ampliamente usadas - y mal usadas - para evaluar a estudiantes, profesores, directores y administradores.

Desgraciadamente, es fácil interpretar equivocadamente las pruebas. Algunas de las inferencias sobre resultados recientes de pruebas hechas por políticos, empleadores, los medios y el público en general no son válidas. Para evitar malas interpretaciones, es importante que los ciudadanos informados y los formuladores de políticas entiendan qué significa realmente la terminología de las pruebas. Este glosario espera proporcionar dicho conocimiento básico.

Este breviario está organizado en tres partes.

La parte I presenta algunas estadísticas esenciales para comprender conceptos sobre las pruebas y para hablar inteligentemente sobre ellas. Aquéllos que están familiarizados con la

terminología estadística pueden saltarse la Parte I e ir directamente a la discusión de la terminología actual de las pruebas. La parte II presenta algunos términos fundamentales. Tanto la Parte I como la II tienen que ver con el "qué": qué es una mediana, un puesto percentil, una prueba referida a normas, etc.

La Parte III amplía las partes I y II con discusiones sobre cuestiones importantes de las pruebas. Estas son más preguntas referidas a "quién" y a "por qué". En conjunto, estas tres partes tienen el potencial de elevar la comprensión pública sobre lo que es, con demasiada frecuencia, una fuente de travesuras políticas y de conflictos ásperos en el terreno educacional.

- Los editores

PARTE I

TERMINOLOGÍA ESTADÍSTICA ESENCIAL

1. ¿Qué es una media? ¿Qué es una mediana? ¿Qué es una moda?

Estas son las tres palabras con que la gente llama a algo "promedio". El término más común, tanto en lo que se refiere a pruebas como en la cultura general, es la media (promedio), que es, simplemente, la suma de todos los puntajes dividido por el número de puntajes. Si alguien tiene las medidas de once personas, para calcular la media (promedio), suma todas las once alturas y las divide por once.

La mediana, otra estadística común, es el punto por encima del cual yacen la mitad de los puntajes y por debajo del cual está la otra mitad. Si Ud. tiene las medidas de altura de once personas y las coloca en orden ascendente o descendente, el valor que tenga para el sexto puntaje es la mediana (cinco serán superiores a ella, cinco inferiores).

Las medias y las medianas pueden diferenciarse en cuanto a lo bien que ellas representen el "promedio", porque las medias están afectadas por los valores extremos y las medianas no. Las medianas sólo involucran el contar hasta la mitad de la distribución de lo que sea que se esté contando. Si se está promediando la fortuna de once personas y una de ellas es Bill Gates, el salario promedio se contará en billones, aun si las otras diez personas están viviendo por debajo del nivel de pobreza. Calculando la mediana, Bill es sólo un individuo más, y para encontrar la mediana sólo necesita Ud. encontrar la persona cuyo puntaje parte al grupo por la mitad.

La tercera estadística que también se considera como "promedio" se llama la moda. Es simplemente el puntaje más comúnmente ocurrente en una lista de puntajes. Suponga que se tiene los pesos de once personas. Si cuatro de ellas pesan 150 libras y no más de tres caen en cualquier otro peso, la moda es 150 libras. Las modas no aparecen con frecuencia en las discusiones sobre pruebas porque la media y la mediana suelen ser más descriptivas. En el ejemplo precedente de peso, por ejemplo, 150 libras sería la moda aun si fuese el más alto o el más bajo peso registrado.

Para ilustrar los diferentes promedios, considere esta lista como las fortunas de los residentes en Redmond, Washington (que, para nuestros objetivos, tiene solo 11 habitantes).

$ 10,000	$ 10,000	$ 20,000	$ 20,000
$ 20,000	$ 50,000	$ 60,000	$ 70,000
$ 75,000	$125,000	$ 70 billones	

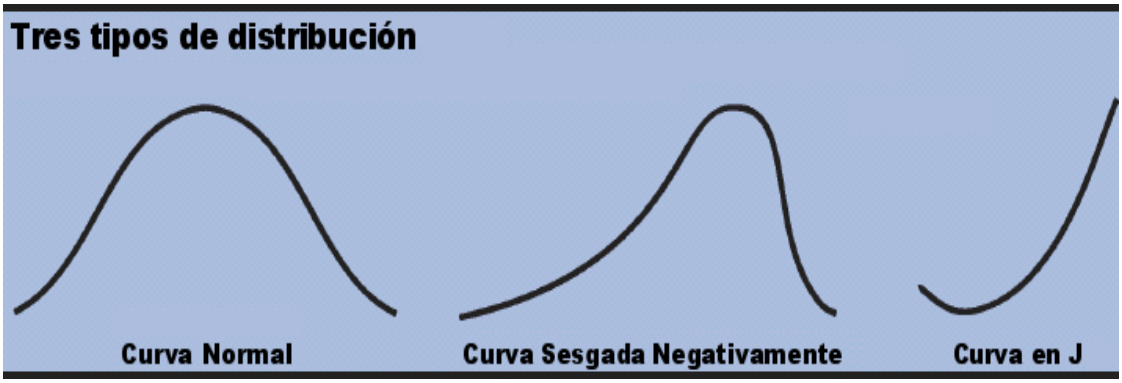
Fortuna media	=	$ 6.4 billones
Fortuna mediana	=	$ 50,000
Fortuna modal	=	$ 20,000

Los setenta billones constituían, aproximadamente la fortuna neta de Bill Gates en 1999. Cuando calculamos la media, esa riqueza es contabilizada y todos los habitantes parecen billonarios, con una fortuna promedio (media) de más de 6 billones.

Cuando calculamos la mediana, buscamos el puntaje que divide el grupo en mitad. En el ejemplo, es $ 50,000; cinco personas tienen más de $ 50 mil y cinco tienen menos. Los billones de Gates no importan, porque solamente estamos buscando el punto medio de la distribución.

En el Redmond de nuestro ejemplo, tres personas tienen fortunas equivalentes a $ 20,000, así que éste es el número más recurrente y es, por lo tanto, la moda.

Muchas distribuciones de estadísticas en educación caen en una curva acampanada, también llamada una "distribución normal". En una distribución normal de puntajes, la media, la mediana y la moda son idénticas.



Las modas se vuelven útiles cuando la forma de la distribución no es normal y tiene dos o más valores en los que se concentran los puntajes. Así, si Ud. diese una prueba y el puntaje más frecuente fuese 100, esa sería la moda, pero si hubiese otro grupo de puntajes, digamos que alrededor de 50, sería más descriptivo referirse a la distribución como "bimodal".

La curva en la izquierda es normal. La del medio es sesgada, con muchos puntajes amontonados en la parte alta. Esto podría suceder porque la prueba fue fácil para la gente que la rindió o porque la instrucción ha sido efectiva y la mayoría de la gente aprendió casi todo lo que necesitaban saber para la prueba.

Al construir una "prueba referida a normas" los que elaboran la prueba *imponen* una distribución normal de puntajes debido a la forma en la cual se seleccionan ítems o preguntas para la prueba. Cuando se trata de pruebas "referidas a criterios", una curva acampanada sería irrelevante. Usualmente, estamos buscando tomar una decisión de “sí” o “no” sobre la gente: ¿alcanzaron o no el criterio? O, queremos colocarlas en categorías como "básica", “proficiente" y "avanzada". El connotado educador Benjamín Bloom sostenía que en educación la existencia de una curva acampanada era un reconocimiento de fracaso: mostraría que la mayoría de la gente aprendió una cantidad promedio, unos pocos aprendieron mucho y otros pocos aprendieron poco. La meta de la educación, sostenía Bloom, debería ser una curva un poco con forma de una "j" inclinada, como la curva en la derecha (en el cuadro). Esto indicaría que la mayoría de la gente habría aprendido mucho y solo unas pocas aprendido poco.

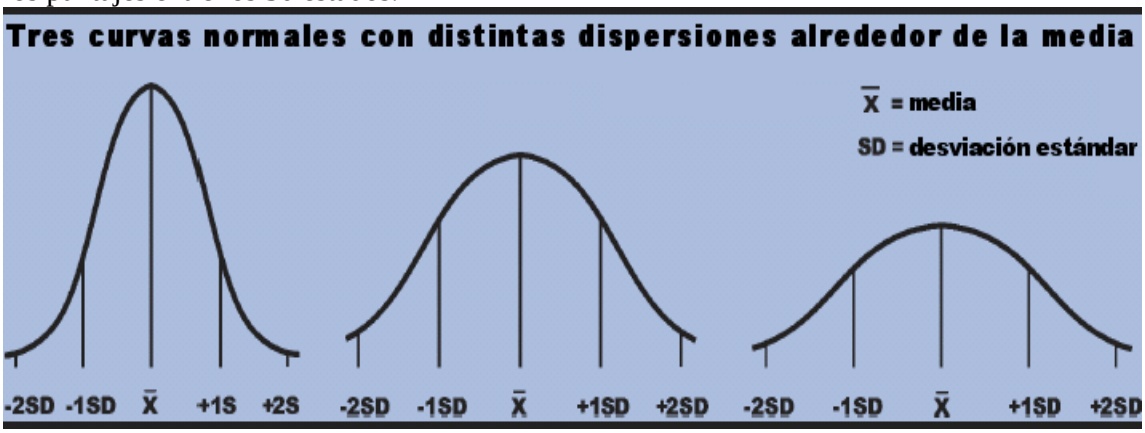
2. ¿Qué significa decir "Ninguna medida de tendencia central sin una medida de dispersión" y por qué habría de decir alguien esto?

La media, la mediana y la moda son todas medidas de promedio o lo que los estadísticos llaman "medidas de tendencia central". Necesitamos una medida de cómo los puntajes se distribuyen alrededor de este promedio. ¿Consiguen todos aproximadamente los mismos puntajes o están los puntajes distribuidos ampliamente? Una manera de reportar sobre la dispersión es el rango: la diferencia entre el puntaje más alto y el más bajo. El problema con el rango es que, como la media, puede estar afectado por puntajes extremos.

La medida de dispersión más común se llama "desviación estándar". En el mundo de la estadística, la diferencia entre el puntaje promedio y cualquier puntaje determinado se llama una "desviación". La desviación estándar nos dice cuán grandes son en promedio estas

desviaciones. Los estadísticos usan mucho la desviación estándar porque tiene propiedades matemáticas útiles e importantes, particularmente cuando los puntajes están distribuidos en una curva normal, de forma acampanada.

En el siguiente cuadro se muestran tres distribuciones diferentes y sus desviaciones estándar. Nótese que todas éstas son curvas acampanadas. Se diferencian en cuán esparcidos están los puntajes alrededor del promedio. A pesar de estas diferencias, algunas cosas son iguales. Por ejemplo, la distancia entre la media y la desviación estándar + 1 ó -1 siempre contiene 34% de los puntajes. Otro 14% caerá entre + ó - una y + ó - dos desviaciones estándar. Una persona que tiene un puntaje de una desviación estándar sobre la media siempre obtendrá puntajes en el percentil 84 - 34% de los puntajes estarán entre la media y una desviación estándar y luego habrá otro 50% que estará debajo de la media (Ver "Puntajes escalares" en la Parte II, Sección 7). Simplemente reportando los promedios a menudo se oculta importantes diferencias que podrían tener importantes implicancias para las políticas. Por ejemplo, en el Tercer Estudio Internacional de Ciencias y Matemáticas, los puntajes promedio en matemáticas y ciencias del 8° grado para los Estados Unidos estaban bastante cerca al promedio de las 41 naciones que participaron en el estudio. Como nación, parecíamos promedio. Sin embargo, los estados con más altos puntajes en los Estados Unidos sobrepasaban a casi todas las naciones, mientras que los estados con los más bajos puntajes sobrepasaban a solo tres de las 41 naciones. El promedio ocultaba cuánto variaban los puntajes entre los 50 estados.



3. ¿Qué es una distribución normal?

Para los estadísticos, una distribución "normal" de puntajes en las pruebas es la curva acampanada. No hay nada "mágico" acerca de las curvas acampanadas, a pesar del título de un famoso libro. Sucede, sin embargo, que muchas características humanas tales como tallas y pesos están distribuidas en la forma de una curva acampanada. Las calificaciones escolares y los puntajes de las pruebas han sido tradicionalmente expresadas en forma de curvas acampanadas.

4. ¿Qué es la significancia estadística?

Las pruebas de "significancia estadística" permiten a los investigadores juzgar si sus resultados son "reales" o podrían haber ocurrido por casualidad. Los investigadores educacionales suelen decir cosas como "la diferencia promedio entre los dos grupos fue significativa "al nivel de punto cero uno (.01)". ¿Qué es lo que pueden querer decir? Quieren decir que la diferencia entre los puntajes promedio de los dos grupos probablemente no

ocurrió por casualidad. Dicho más precisamente, las probabilidades de que **sí** ocurrió por casualidad son menores que una en cien. Esto se escribe como p<.01. La "p" es por "probabilidad"-- la probabilidad de que los resultados hubiesen podido ocurrir por casualidad.

5. ¿Por qué necesitamos pruebas de significancia estadística?

Porque usamos muestras, no poblaciones totales.

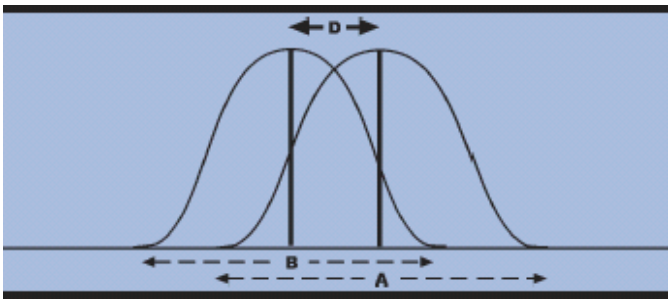
Tomemos el caso más simple, cuando comparamos sólo dos grupos. Digamos que a un grupo de estudiantes se le enseña a leer con enfoque de lenguaje integral y al otro fonéticamente. Al final del año aplicamos una prueba de lectura y encontramos que los dos grupos difieren. ¿Es probable o improbable que esa diferencia haya ocurrido por casualidad? Eso es lo que una prueba de significancia estadística nos dice.

Podría Ud. bien preguntarse, si los dos grupos en realidad tuvieron el mismo puntaje promedio, ¿por qué encontramos alguna diferencia en primer lugar? La respuesta es que estamos tratando con muestras, no poblaciones. Si se aplica la prueba a todos (la población total), cualquier diferencia que se encuentre es real, sin importar cuán grande o pequeña sea (presumiendo, por el momento, que no hay error de medición en la prueba). Pero cualquier muestra dada podría no ser representativa de la población. Esto es particularmente cierto en la investigación educacional que a menudo debe usar "muestras de conveniencia", es decir, los alumnos de colegios cercanos. Si Ud. comparase dos muestras diferentes, podría encontrar un resultado distinto. Si Ud. comparase fonética contra lenguaje integral en otro colegio podría Ud. obtener puntajes algo distintos y es poco probable que la diferencia fuera **exactamente** como la que encontró en la comparación anterior.

6. ¿Cómo se vincula la significancia estadística con la significancia práctica?

No están vinculadas. Los resultados de un experimento como nuestro ejemplo antes mencionado pueden ser altamente significativos estadísticamente, pero aun así no tener importancia práctica alguna. Por el contrario, resultados que no son estadísticamente significativos pueden ser tremendamente importantes en términos prácticos. Reitero, la significancia estadística es sólo una declaración de probabilidades. "¿Cuán probable es que las diferencias que observamos ocurrieran por casualidad?". Es importante tener en cuenta esto, porque muchos investigadores han sido entrenados en el uso de la significancia estadística y actúan como si la significancia estadística y la significancia práctica fuesen la misma cosa. Las posibilidades de encontrar un resultado estadísticamente significativo crecen conforme la muestra se agranda. Las pruebas estadísticas más comunes fueron diseñadas para muestras pequeñas, más o menos del tamaño de un salón de clase. Si las muestras son grandes, diferencias pequeñas se pueden convertir en significativas. A medida en que las muestras aumentan en tamaño tenemos más confianza en que estamos consiguiendo una muestra representativa, una muestra que representa con precisión la población total.

La decisión sobre la significancia práctica deber ser sopesada en otros términos. Por ejemplo, ¿podemos encontrar evidencia colateral de que los estudiantes a quienes se les enseña a leer con un enfoque de lenguaje integral difieren de los estudiantes a quienes se les enseña por fonética? ¿Informan los profesores que a los chicos a quienes se les enseña con un método o con el otro **les gusta** leer más? ¿Difieren los dos grupos en cuanto a cuánto leen los niños en casa? ¿Cuánto cuestan los dos programas? ¿Justifican los beneficios de uno u otro programa esos costos?



Tomemos un ejemplo. Supongamos que las dos distribuciones graficadas arriba representan los puntajes de estudiantes que han aprendido a leer con dos programas de instrucción diferentes. Sus puntajes promedio difieren por la cantidad D. Una prueba de significancia estadística nos dirá cuán probable fue que una D de ese tamaño pudiese haber ocurrido de casualidad si, realmente, en toda la población D era igual a O.

Y ahora, ¿qué? Bien, parece que deberíamos considerar A sobre B. Pero esa decisión no se puede basar únicamente en resultados estadísticos. El cálculo del "tamaño del efecto" (descrito en la siguiente sección) dará alguna idea de cuán grande es la diferencia en términos prácticos, pero eso por sí mismo no conducirá a una decisión. Necesitamos determinar con certeza que una prueba fue igualmente justa para ambos programas. En un estudio que comparó fonética contra lenguaje integral, los estudiantes de ambos programas obtuvieron un puntaje más o menos igual en una prueba estandarizada. Los estudiantes en el programa de fonética, sin embargo, tuvieron puntajes pobres en una prueba sobre carácter, trama y escenario -- aspectos de la lectura que son tratados con el enfoque de lenguaje integral pero no con la fonética.

Si pensamos que el resultado estadístico es válido, entonces podemos hacer preguntas como: ¿Cómo se sienten los profesores acerca de los dos programas? ¿Cómo se sienten los estudiantes? ¿Cuesta un programa mucho más que el otro? ¿Cuánta capacitación adicional del profesor se requiere para que los profesores se vuelvan competentes en los dos programas? ¿Los estudiantes de un programa pasan más tiempo leyendo voluntariamente que los estudiantes en el otro programa? Un "texto programado" usado para enseñar las nociones de B.F. Skinner acerca del aprendizaje fue utilizado en programas de psicología para estudiantes de pregrado en los años 60. Se pregonaba que garantizaba que los estudiantes dominarían los conceptos. Lo hicieron. Pero el formato del libro lo hizo, simultáneamente, difícil de leer y aburrido. Los estudiantes salieron odiando tanto a Skinner como a los textos programados.

7. ¿Qué es un coeficiente de correlación?

Los "coeficientes de correlación" muestran cómo se relacionan los cambios en una variable con cambios en otra. Un ejemplo usado varias veces en este documento es la correlación entre los puntajes *SAT* y las calificaciones en el primer año de la universidad. La gente que obtiene puntajes más altos en el *SAT* tiende a tener calificaciones más altas en su primer año de universidad. Esto es una indicación de una correlación positiva: conforme los puntajes de las pruebas suben, las calificaciones tienden también a subir. La palabra importante en la última oración es "tienden". No todas las personas que logran buen puntaje en el *SAT*

saldrán bien en la universidad. Si la relación entre los puntajes en las pruebas y las calificaciones fuese perfecta y positiva, entonces la correlación estaría en su valor más alto posible, + 1.00. Si la relación entre los puntajes en las pruebas y las calificaciones fuese perfecta y negativa, el coeficiente de correlación sería -1.00. Esto describiría una situación peculiar en la cual la gente con los puntajes más altos en las pruebas recibirían luego las calificaciones más bajas en la universidad.

Toda esta terminología estadística es importante cuando se lee e interpreta pruebas y puntajes de pruebas, el tema de la Parte II.

PARTE II

LA TERMINOLOGÍA DE LAS PRUEBAS

1. ¿Qué es lo estandarizado en una prueba estandarizada?

Virtualmente todo. Las preguntas son las mismas para todos los que rinden la prueba. Están en el mismo formato para todos los que la toman (usual, pero no exclusivamente, el formato de opción múltiple). Las instrucciones son las mismas para todos los estudiantes (se hacen algunas excepciones para estudiantes con ciertas discapacidades). Los límites de tiempo son los mismos (salvo excepciones similares). El modo de asignar puntajes es el mismo para todos los que toman la prueba, y no hay lugar para la interpretación.

La forma en que se informa a los padres o directiva del colegio sobre los puntajes es igual para todos los que rinden la prueba. Incluso los procedimientos para crear la prueba son bastantes estandarizados. Las estadísticas usadas para analizar las pruebas son estandarizadas.

Cuando hay la posibilidad de hacer interpretaciones de respuestas abiertas, como en algunas pruebas de cuociente intelectual (C.I.) administradas individualmente, los propios administradores están bastante estandarizados. Es decir, están entrenados en cómo aplicar la prueba, qué variaciones a la respuesta aceptar y cuáles rechazar (esto es especialmente importante cuando se aplican pruebas a niños pequeños, que son cualquier cosa menos estandarizados) y en cómo, en general, comportarse en el escenario de la prueba . No sería útil obtener un puntaje de C.I. que pase de 100 a 130, o caiga a 70, dependiendo de quién toma la prueba a un niño

2. ¿Qué es una norma? ¿Qué es una prueba referida a normas?

La norma es una mediana particular, la mediana de una prueba estandarizada referida a normas. A ella y a otras medianas también se las denomina como el 50^{avo} percentil. Cualquiera que sea el puntaje que divide a los que tomaron la prueba en dos grupos, con 50% de los puntajes por encima y 50% por debajo de ese puntaje, esa es la norma.

Los editores de pruebas se refieren a la mediana de sus pruebas como “la norma nacional”. Si la prueba ha sido debidamente construida, el estudiante promedio en la nación sacaría un puntaje equivalente a la norma nacional.

A diferencia de la temperatura interna del cuerpo, no hay nada evaluativo sobre la norma de una prueba. Noventa y ocho punto seis grados Fahrenheit (98.6 F) es la norma para la temperatura del cuerpo. Es un indicador de salud y las desviaciones de esta norma son malas. La norma en los puntajes de pruebas, sin embargo, simplemente señala un lugar en *el medio* de la distribución de puntajes (pese a lo cual, algunos administradores colocan a estudiantes en clases para niños retrasados, o en programas de niños dotados y talentosos, basándose únicamente, en las relaciones de los estudiantes con esta norma).

Una vez que la norma ha sido determinada, todos los otros puntajes son descritos con referencia a esta norma -- de allí el término "prueba referida a normas". Las pruebas de Habilidades Básicas de Iowa y otras pruebas comerciales de logros, el *SAT* y las pruebas de *C.I.* son todos ejemplos de pruebas referidas a normas.

La idea de establecer normas nacionales de este modo incomoda a algunas personas porque, por definición, la mitad de todas las personas que rinden la prueba tienen que estar debajo del promedio. Ellas alegan que podría dañar a algunos niños el pensar que están debajo del promedio cuando realmente lo están haciendo bastante bien.

¿Cómo podría estar yéndole bastante bien a alguien y aun así estar debajo del promedio? Porque una prueba referida a normas no dice nada sobre cuan bien se está desempeñando cualquier persona. Si una persona obtiene un puntaje en el percentil 75 en dicha prueba, se sabrá que tuvo un desempeño mejor que el de 75% de las otras personas que rindieron la prueba. Eso es todo. Quizás *todos* los que tomaron la prueba lo hicieron muy mal. Sucede simplemente que esa persona se desempeñó mejor que el 75% del grupo. Por otra parte, si su puntaje está en el percentil 30 de los que rinden el Examen de Graduados (*Graduate Record Examination- GRE*) está "debajo del promedio" pero aun en un grupo bastante selecto. Si se preocupó por tomar el GRE, es probable que complete cuatro años o más de universidad, algo que logra sólo una cuarta parte de todos los adultos en los EEUU, y solo el 50% de aquéllos que hoy en día empiezan su vida universitaria.
Es importante recordar que *los puntajes de una prueba referida a normas son siempre relativos, nunca absolutos*[1]. Si Ud. visita África y compara su talla con un grupo de Watusis, es muy probable que Ud. esté debajo del promedio; si visita a pigmeos y realiza las mismas mediciones, Ud. podría estar en el percentil 99. Su talla absoluta nunca cambió, pero la naturaleza del grupo referencial sí.

Además, más o menos cada cinco años, los editores de las pruebas comerciales re – normalizan sus pruebas de logros. La currícula cambia para reflejar cambios en los conocimientos o cambios de énfasis pedagógicos. Las pruebas antiguas podrían no medir los contenidos de la nueva currícula. Así que los editores de pruebas deben re – normalizar sus pruebas con cierta frecuencia para mantenerlas actualizadas. Hay evidencia abrumadora de que el logro educacional ha fluctuado hacia arriba y hacia abajo en los últimos 40 años, así que el "percentil 50" refleja diferentes cantidades de logros en diferentes momentos.

Para escaparse del relativismo de las pruebas referidas a normas se ha buscado desarrollar pruebas que tengan "puntajes referidos a criterios".

[1] Hasta 1996 el SAT fue una excepción a esta regla. Su "norma" se estableció en 1941 y permaneció fija hasta que el Consejo Universitario (*College Board*) "recentró" el SAT en 1996. "Recentrar" equivale a establecer una nueva norma, que es lo que hacen los productores de pruebas comerciales aproximadamente cada 5 años.

3. ¿Qué es una prueba referida a criterios?

En teoría, para cualquier tarea, podemos imaginar su logro en un continuo que va desde una carencia total de habilidad hasta una excelencia evidente. Cualquier nivel de logro a lo largo de ese continuo puede ser referido a criterios específicos de desempeño. Por ejemplo, si la habilidad fuese patinaje en el hielo, el continuo podría ir desde "No poder pararse sobre el hielo "hasta" aterrizar tras un triple giro". El béisbol profesional usa un sistema referido a criterios. Las ligas mayores representan "excelencia notoria" mientras que los diversos niveles de equipos rurales representan puntos diferentes de logro en un continuo. Se puede entrenar a jueces para que estén de acuerdo casi unánimemente sobre la calidad de una actuación.

Desgraciadamente, los dominios educacionales casi nunca son ni remotamente tan específicos como aquéllos que se encuentran en el atletismo. Los "criterios" de las pruebas referidas a criterios generalmente se limitan a establecer un puntaje de corte en alguna prueba. Muchas pruebas actuales que se consideran referidas a criterios estarían mejor denominadas como "referidas a contenidos". Así, en el Programa Estándares de Aprendizaje de la Comunidad de Virginia, se describió ciertos contenidos que los estudiantes deberían esforzarse por aprender. Después se desarrollaron pruebas para medir cuán bien los estudiantes habían dominado el material especificado en los estándares.

Estas pruebas tienen puntajes de corte, puntajes que determinan si un estudiante pasa o reprueba. A este puntaje de corte se le refiere frecuentemente como el "criterio". Debido a ello estas pruebas son a menudo llamadas pruebas referidas a criterios, pero la frase no está usada en el sentido original descrito en el primer párrafo de esta sección. El "criterio" es simplemente lograr un puntaje por encima del puntaje de corte designado, a fin de graduarse de secundaria. Si el puntaje de corte es, digamos, 70, lo único que importa es obtener un 70 o algo mejor. La decisión de aprobar o reprobar se basa en el puntaje y nada más. Una verdadera prueba referida a criterios, en cambio, tendría criterios asociados a los puntajes por encima de 70 y con los puntajes más bajos también.

En casi todos los estados, la prueba para obtener una licencia para conducir es parcialmente una prueba referida a contenidos con un "criterio" y también una verdadera prueba referida a criterios. La prueba de lápiz y papel cubre contenidos específicos y los postulantes deben obtener un cierto número de ítems correctos para pasar. Además hay una prueba al volante con verdaderos criterios. El postulante por ejemplo, debe parquear su auto en paralelo a cierta distancia del sardinel y sin tumbar los postes que representan otros autos.

4. ¿Cómo se desarrollan las pruebas referidas a normas y las referidas a criterios?

Los procedimientos para las dos pruebas son bastante diferentes. En las pruebas referidas a normas los productores de pruebas examinan los materiales curriculares producidos por diversos editores de libros de texto y de cuadernos de trabajo. Luego los redactores de ítems construyen ítems para medir las habilidades y temas más comúnmente reflejados en estos libros. Estos ítems son luego evaluados por paneles de expertos respecto a su "validez de contenido". Validez de contenido es un índice de si una prueba mide o no lo que dice medir (considerado con mayor detalle en la sección sobre validez de pruebas). Una prueba que reclama ser una medida de habilidades de lectura pero que consiste sólo de ítems de vocabulario no tendría alta validez de contenido.

Después, los ítems deben ser probados o piloteados para ver si se "comportan" debidamente. El comportamiento adecuado de un ítem es un concepto estadístico. Si muchas personas logran responder correctamente el ítem o muchas se equivocan, el ítem no se comporta debidamente. La mayoría de ítems incluidos en las pruebas referidas a normas son aquéllos a los que entre 30% y 70 % de los estudiantes responden correctamente en las pruebas piloto. El diseñador de pruebas también eliminará preguntas en las que personas con puntajes generales altos se equivocan y en las que personas con puntajes generales bajos lo hacen bien. La teoría es que cuando eso ocurre, hay algo extraño en el ítem.

Los constructores de pruebas escogen ítems que obtienen entre 30 y 70% de respuestas correctas debido a la manera como suelen generalmente usarse las pruebas referidas a normas. Se usan para hacer predicciones diferenciales (por ejemplo, quién va tener éxito en la universidad) o para asignar recompensas diferencialmente (por ejemplo, a quién se admite en programas para niños dotados y talentosos). Si todos responden los ítems de manera correcta o si todos tienen los ítems errados, todos tendrían el mismo puntaje, y no sería posible hacer predicciones diferenciales. Recuerde que el uso principal de las pruebas referidas a normas es hacer dichas predicciones.

Para pruebas referidas a normas, el vocabulario debe limitarse a palabras que se espera todos conozcan, excepto, claro está, en una prueba de vocabulario. Términos que fueron tomados de áreas especializadas como el arte o la música, por ejemplo, resultarían novedosos para muchos estudiantes, que entonces podrían escoger una respuesta equivocada (o una correcta) por la razón equivocada. Una prueba hecha por un maestro, por el contrario, puede incorporar palabras que han sido recientemente usadas en la su enseñanza, sean o no familiares a la mayoría de las personas.
Como una pequeña disgresión observamos que elaborar una prueba con "palabras que se espera todos conozcan" no es tan simple como uno podría inicialmente pensar. En una nación políglota como los Estados Unidos, diferentes subculturas usan diferentes palabras. Cuando la palabra "regata" apareció en algunas ediciones del *SAT* se de*sat*ó una pequeña guerra; la gente argüía que los estudiantes de familias de bajos ingresos tenían mucho menores probabilidades de haberse encontrado antes con "regata" o palabras similares que reflejan actividades sólo de los acaudalados.

El proceso de desarrollar una prueba referida a criterios es bastante distinto. Para casi todas dichas pruebas se especifica una serie de objetivos o quizás incluso un currículo íntegro, y el fin de la prueba es determinar cuán bien han dominado los estudiantes los objetivos o el currículo. Como en las pruebas hechas por maestros de aula, una prueba referida a criterios puede contener palabras que son inusuales o raras en el habla diaria y la lectura, siempre que aparezcan en el currículo y siempre que los estudiantes hayan tenido una oportunidad de aprenderlos.

Con una prueba referida a criterios no estamos muy interesados en diferenciar a los estudiantes por medio de sus puntajes. Sin duda la meta de algunas de dichas pruebas, como aquéllas que se requieren para obtener la licencia de conducir, es que todos logren una nota aprobatoria. Cuando las pruebas referidas a criterios diferencian entre los estudiantes usualmente es para colocarlos en categorías tales como básico, proficientes y avanzados -- más que para ordenar a los estudiantes según niveles de percentil.

Históricamente, la mayoría de las pruebas usadas en los Estados Unidos han sido pruebas referidas a normas: pruebas estandarizadas de logros, el *SAT* y el *ACT*, la prueba de C.I. etc. Las pruebas estatales desarrolladas recientemente son referidas a criterios, en el sentido de tener un puntaje de corte.
Tanto las pruebas referidas a normas como las referidas a criterios deben ser evaluadas en términos de dos cualidades técnicas, confiabilidad y validez, consideradas a continuación.

5. ¿Qué es la confiabilidad en una prueba?

En las evaluaciones, la confiabilidad es una medida de consistencia. Es decir, si un grupo de personas rindió una prueba en dos ocasiones diferentes, ellas deberían obtener puntajes bastante parecidos las dos veces (asumiendo que no quede memoria de la primera ocasión para la segunda). Si la gente obtuviera un puntaje alto en la primera oportunidad y bajo en la segunda, no tendríamos base alguna para interpretar lo que la prueba significa.

Inicialmente, la manera más común de determinar la confiabilidad era hacer que una persona rindiera la misma prueba dos veces o tomar formas alternativas de una misma prueba. Los puntajes de las dos administraciones de la prueba estarían correlacionados. Generalmente, uno esperaría que la correlación entre las dos administraciones llegase a .85 o más, acercándose al máximo posible que es + 1.00 (ver "¿Qué es un coeficiente de correlación?" para una explicación de qué valores puede tomar).
Dar pruebas dos veces a la gente es a menudo inconveniente. También está el problema del tiempo: si la segunda administración es muy cercana a la primera, la memoria de la primera podría afectar la segunda. Si el intervalo entre las pruebas es muy largo, muchas cosas pueden cambiar en la estructura cognitiva de una persona, lo que podría disminuir la correlación. Una alternativa para probar la confiabilidad se llama "confiabilidad partida". Esto significa tratar a cada mitad de la prueba como una prueba independiente y correlacionar las dos mitades. Usualmente, las preguntas impares resultan estar correlacionadas con las pares.

6. ¿Qué es la validez de una prueba?

La confiabilidad es el *sine qua non* de una prueba: si no es confiable, hay que deshacerse de ella. Sin embargo, una prueba puede ser confiable sin ser válida. Si un tirador al blanco dispara 10 veces y siempre da en la misma posición del tablero pero a un pie de distancia del blanco, podríamos decir que el tirador es confiable (da en el mismo lugar cada vez) pero no válido, ya que el objetivo es el blanco.

La validez es algo más complicado que la confiabilidad.

Hay varios términos que pueden ser usados junto con la palabra validez, contenido, constructo, criterio, consecuencial y aparente. Una prueba tiene validez de contenido si mide lo que dice estar midiendo. Esto requiere que la gente analice el contenido de la prueba en relación con lo que se supone que la prueba mide. Esto podría requerir, en el caso de pruebas referidas a criterios, contrastar una prueba con los contenidos de un syllabus.

La validez relacionada a criterios, también llamada validez predictiva, ocurre si una prueba predice algo que estamos interesados en predecir. El *SAT* fue desarrollado para predecir las calificaciones en el primer año en la universidad. Para ver si lo hace, correlacionamos los puntajes en la prueba con las calificaciones. Si la prueba tiene validez predictiva, aquéllos

con puntaje alto en ella también tienen la tendencia a sacar mejores calificaciones que aquéllos con puntaje bajo.

Determinar si una prueba tiene o no *suficiente* validez predictiva para justificar su continuación es asunto de juicio o análisis de costo beneficio. Poquísimas universidades requerirían el *SAT* si tuviesen que pagarlo (ahora los estudiantes pagan los costos). Las predicciones a partir de las calificaciones en las escuelas secundarias y el puesto en la clase serían suficientemente altas. El *SAT* añade poca precisión a las predicciones y costaría millones de dólares a las universidades si éstas, en vez de los postulantes, cargasen con el costo.

La validez de constructo es un concepto más abstracto. Es un poco como la validez de contenidos porque estamos tratando de determinar si una prueba mide lo que dice que mide, pero esta vez no estamos interesados en los contenidos, como aritmética o historia, sino en constructos psicológicos tales como inteligencia, ansiedad o auto-estima. La validez de constructo es mayormente de interés para otros profesionales que están trabajando en el campo del constructo. Ellos tratarían de determinar si una nueva prueba de, digamos, ansiedad, arroja mejor información para propósitos de tratamiento o si encaja mejor con otros constructos en el campo.

La validez consecuencial se refiere a las consecuencias de una prueba y si aprobamos o no esas consecuencias. También se refiere a inferencias hechas a partir de la prueba. Por ejemplo, una vez que se conoce una prueba, los profesores a menudo usan más tiempo enseñando el material que está en la prueba que el material que no lo está. ¿Es eso una cosa buena? La respuesta depende de cómo juzgamos lo que está siendo enfatizado y lo que se deja de lado. Pudiese ser que la prueba esté haciendo un buen trabajo al enfocar la atención de los profesores en material importante, pero pudiese ser que la prueba esté causando que los profesores dejen de lado otro material igualmente importante y que estrechen mucho su enseñanza. Numerosos estados han desarrollado pruebas para determinar si los estudiantes han dominado ciertos contenidos y habilidades. En la primera administración de estas pruebas, muchos estudiantes fallaron. Algunos infirieron que los profesores no estaban enseñando el material adecuado o no estaban enseñando bien. Otros dedujeron que los puntajes de corte para las pruebas estaban demasiado altos, y algunos decían que las pruebas simplemente eran malas. Todas éstas eran consecuencias de usar la prueba.

Los investigadores no se han puesto de acuerdo sobre la importancia de la "validez aparente". La validez aparente tiene que ver con la percepción que sobre la prueba tiene el que la rinde. Si el contenido de la prueba le parece inapropiado o irrelevante, la cooperación del que rinde la prueba con la prueba con el instrumento de medición se pone en riesgo, posiblemente también perturbando las otras clases de validez.

7. ¿Qué es un percentil? ¿Una equivalencia a un grado? ¿Un puntaje escalar? ¿Una estanina?

Estos términos son todos medidas que se usan para informar sobre los resultados de las pruebas. Los dos primeros son las más comunes, mientras que estanina es rara vez ya usada. Quiere decir " de estándar nueve" y fue una manera de transformar los percentiles a nueve categorías. Esto fue importante en el momento en que se inventó, porque los datos eran procesados en computadoras por medio de tarjetas perforadas de 80 columnas y había que

ahorrar espacio en las tarjetas. Condensando los 99 puestos percentiles en 9 estaninas, los resultados de las pruebas ocupaban sólo una columna.

Los percentiles, las equivalencias a grados, y las equivalencias de curvas normales son pertinentes solamente a las pruebas referidas a normas. Los puntajes escalares se usan tanto para pruebas referidas a normas como para las referidas a criterios.

Percentiles. Los percentiles proveen información en términos de cómo se desempeñó determinado niño, aula, colegio o distrito en relación con otros niños, aulas, colegios o distritos. Un estudiante en el primer percentil es sobrepasado por todos, un estudiante en el 99° percentil sobrepasa a todos y un estudiante en el 50^{avo} percentil está en el promedio nacional.

Es importante notar que los percentiles son posiciones relativas, no puntajes. A partir sólo de puestos no se puede decir nada sobre el desempeño. Cuando los ocho velocistas finalistas corren los 100 metros en las Olimpiadas, alguien *tiene* que quedar último. Esta persona sigue siendo el 8° ser humano más rápido en el planeta ese día. Los percentiles usualmente se reportan con referencia a algún grupo nacionalmente representativo, pero pueden ser adaptados a "normas locales".

Las grandes ciudades a menudo se comparan a sí mismas con otras grandes ciudades, a fin de evitar aparecer en categorizaciones o “rankings” nacionales que incluyen puntajes de los suburbios. Los suburbios casi nunca se comparan a sí mismos con otros suburbios, porque se ven mejor cuando se les compara con muestras nacionales que incluyen a estudiantes de las grandes ciudades y de áreas rurales pobres.

Equivalencias a grados. Las equivalencias a grados también clasifican a los estudiantes con referencia al desempeño del estudiante promedio. Un equivalente a grado de 3.6 se asignaría al estudiante que recibió un puntaje igual al promedio obtenido en una prueba rendida en el sexto mes de clases por alumnos del tercer grado. Si un estudiante en el cuarto mes del cuarto grado recibe un equivalente al grado de 4.4 en una prueba, se dice que esa estudiante está "en el nivel del grado". Esta manera de conceptualizar el nivel de un grado crea mucha confusión.

Algunas veces los periódicos generan escándalos informando que la mitad de los estudiantes en algún colegio "no están leyendo a nivel de grado". No es un escándalo. Hemos definido "nivel de grado" como el puntaje del estudiante promedio. Entonces, a nivel nacional, la mitad de todos los estudiantes están *siempre* por debajo del nivel de grado. Por definición.

No tenemos que definir el nivel de grado de esta manera. Podríamos dar al nivel de grado una interpretación referida a criterios y esperar que todos los niños lo lograsen, pero ésta no es la manera como usualmente se le define.

El concepto de nivel de grado también crea confusión cuando los estudiantes sacan un puntaje por encima o por debajo del grado que están cursando. Los padres de los alumnos de cuarto grado cuyos hijos están leyendo, digamos, a un nivel de séptimo grado se preguntarán porqué su niño no está en el séptimo grado, por lo menos en lectura. Pero un niño de cuarto grado que recibe un equivalente de grado siete en una prueba no está leyendo como un niño de séptimo grado. Se trata del equivalente de calificación que el niño promedio del séptimo

grado obtendría *leyendo material de cuarto grado*. Es poco probable -- aunque no imposible -- que un niño de cuarto grado que lee en un nivel de “séptimo grado” pudiese realmente arreglársela con material de lectura para el séptimo año escolar.

Un "**puntaje escalar**" es difícil de explicar sin entrar en mucho detalle estadístico. Conceptualmente, se trata de convertir los puntajes crudos en una métrica en términos de la desviación estándar (Por favor, ver el cuadro que muestra la desviación estándar en la sección ¿? de este texto). Supongamos que una prueba tenía 100 ítems y otra sólo 50. Un puntaje crudo de 40 probablemente significaría cosas muy diferentes en las dos pruebas. Pero las dos pruebas pueden convertirse a una escala en términos de sus desviaciones estándar.

Convertir a puntajes escalares los puntajes crudos produce una escala con un promedio de 0.0 y una desviación estándar de 1.0 (el puntaje promedio menos el puntaje promedio = 0, y 0 dividido por cualquier cifra es cero). Los estadísticos decidieron tempranamente que tal escala no se veía muy bien. Sucede que se puede añadir una constante a todos los puntajes escolares o multiplicarlos todos por una constante sin cambiar la relación entre ellos. La distribución de puntajes escalares tiene un promedio de 0.0 y una desviación estándar de 1.0. Si multiplicamos todos los puntajes escalares por 100 y añadimos 15, obtenemos la escala común del C.I. (una media de 100 y una desviación estándar de 15). Si los multiplicamos por 100 y añadimos 500 obtenemos la escala del *SAT* (una media de 500 y una desviación estándar de 100).

Los puntajes escalares también permiten una comparación normativa de puntajes a través de diferentes escalas: un puntaje de C.I. (*IQ*) de 115 es el "mismo" que un puntaje verbal de *SAT* de 600, porque ambos están a una desviación estándar por encima del promedio, que es 100 y 500 respectivamente. Una persona con un puntaje de C.I. (*IQ*) de 115 y un puntaje verbal de *SAT* de 600 está en el 84° percentil en ambas pruebas. Si los puntajes en las dos pruebas hubiesen sido diferentes, querríamos explorar si las pruebas estaban o no midiendo constructos diferentes. (En este caso, los constructos están altamente correlacionados. Cuando se inventó el *SAT* en 1926 se le mencionaba como una prueba de inteligencia). Los puntajes escalares sólo pueden ser interpretados significativamente si los puntajes caen en una curva normal o de forma acampanada, o de manera muy parecida a ella.

8. ¿Qué son las preguntas de opción múltiple?

Son preguntas en las que uno lee algún material, luego escoge una respuesta de una lista de respuestas pre seleccionadas, usualmente cuatro o cinco. Inventadas en 1914 por Frederick J. Kelly de la Universidad de Kansas, las preguntas de opción múltiple hicieron posible aplicar pruebas masivamente a los reclutas militares en la Guerra Mundial I, para evaluar sus talentos y habilidades diferenciales. Todos podían tomar la prueba a la misma hora, en el mismo formato, en un corto período de tiempo, y las respuestas podían ser calificadas rápidamente y a bajo costo.

Estas cualidades explican porque aún hoy en día frecuentemente se prefieren las pruebas de opción múltiple. Desde la I^{a} Guerra Mundial, los principales cambios en la tecnología de la opción múltiple han sido desarrollos en la tecnología empleada para la calificación. Las computadoras escudriñan y califican miles de hojas de respuestas en una hora. La principal desventaja de las preguntas de opción múltiple es que usualmente someten a prueba muestras pequeñas de conocimientos, fuera de contexto. Se pueden elaborar pruebas de opción

múltiple que miden razonamientos de un orden más elevado, pero uno raramente las ve excepto en escuelas de postgrado.

9. ¿Qué miden las pruebas de opción múltiple?

Esta pregunta vaga podría tener una variedad de respuestas, pero tal como se usa acá se refiere al "nivel" de conocimientos o habilidades que miden las diversas pruebas. Mucha gente se opone a las pruebas de opción múltiple sosteniendo que ellas sólo pueden medir trocitos de información o conocimientos descontextualizados. Otros sostienen que las pruebas de opción múltiple pueden medir el razonamiento y pensamientos de orden más elevado tan bien como cualquier otra clase de prueba. La resolución de esta disputa parecería estar en la palabra "pueden". Las preguntas de opción múltiple *pueden* medir todo tipo de habilidades analíticas, pero rara vez *lo hacen*.

El uso de pruebas de opción múltiple para evaluar el pensamiento de orden más elevado se ve principalmente en escuelas de postgrado y no en las pruebas usadas en las escuelas primarias o secundarias o en cualquier otra parte en gran escala. Esas pruebas podrían, por ejemplo, describir un experimento en psicología. La exposición del experimento podría ocupar una página entera o más -- un texto bastante más largo que lo visto en otras pruebas. Las preguntas entonces pedirían a los estudiantes que saquen conclusiones acerca de lo que demuestran los resultados del experimento. Una prueba completa podría consistir de sólo dos o tres pasajes como ése, con cuatro a ocho preguntas elaboradas alrededor de cada uno.

En contraste, la mayoría de las pruebas usadas en las escuelas requieren que los estudiantes contesten muchas preguntas en un periodo corto de tiempo. Los estudiantes que se detienen a pensar sobre una pregunta están en problemas: no terminarán la prueba. Por cierto, una recomendación que da el Consejo Universitario (*College Board*) a los estudiantes que practican para el *SAT* es "avancen". El *SAT*, más complejo que la mayoría de las pruebas de logros, contiene preguntas como las siguientes:

Costillas: Pulmón

a) cráneo: cerebro
b) apéndice: órgano
c) calcetín: pie
d) esqueleto: cuerpo
e) pelo: cuero cabelludo

Si el producto de cinco números enteros es negativo, ¿a lo más cuántos de los cinco pueden ser negativos?

a) 1
b) 2
c) 3
d) 4
e) 5

En la primera pregunta, la persona debe evaluar cada respuesta alternativa para ver cuál describe una relación lo más parecida a aquélla en la primera parte del ítem. El pensamiento analógico es importante en el mundo real, pero rara vez dicho pensamiento está tan constreñido por el formato de un ítem. Ni tampoco tiene lugar en un lapso de tiempo tan corto.

En la segunda pregunta, los estudiantes sólo necesitan recordar que el producto de dos números enteros negativos es positivo y que el producto de un positivo y un negativo es negativo. Por lo tanto, cualquier grupo de números enteros que contiene un número impar de números enteros pueden ser todos negativos (aun: - 1 x- 1 x - 1 x - 1 = + 1; multiplicándolo por otro - 1 = - 1 y así, sucesivamente). Aquéllos que abogan por la evaluación del desempeño o evaluación "auténtica" generalmente están tratando de superar los límites de las pruebas de opción múltiple.

10. ¿Qué es la evaluación "auténtica"?

La evaluación auténtica es un intento por medir el desempeño directamente, en un escenario de "la vida real". En una prueba de opción múltiple es usualmente imposible saber por qué un estudiante escoge una determinada respuesta. ¿Memoria repetitiva? ¿Acierto afortunado? ¿Acierto después de eliminar dos respuestas erróneas? ¿Una línea de pensamiento bien razonada hasta la respuesta correcta? ¿Una línea de pensamiento bien razonada hasta una respuesta errónea?

En pruebas de aritmética a veces se puede recoger algunas pistas, porque las respuestas incorrectas pueden ser diseñadas para reflejar ciertos tipos particulares de errores. Este tipo de uso diagnóstico de las pruebas es bastante difícil en otras áreas temáticas. Por estas y otras razones, algunas personas se han interesado en elaborar pruebas que evalúan el desempeño en escenarios más "auténticos".

La palabra "auténtica" no es un término muy bien elegido, especialmente porque implica que cualquier otra clase de evaluación es "no auténtica". Quizás un mejor término sería evaluación "directa". Todas las tareas a las que se les ha llamado "auténticas" son una forma de evaluación directa. Sus defensores sostienen que no podemos determinar cuán bien alguien conoce una técnica o un cuerpo de conocimientos a menos que tengamos una oportunidad para observar directamente su desempeño. No podemos decir mucho sobre las habilidades de redacción de alguien usando pruebas de opción múltiple. En dichas pruebas, una persona lee unas cuantas oraciones en las cuales se han subrayado algunas partes. La persona luego selecciona una de entre cuatro o cinco opciones sobre cuál de las partes subrayadas contiene una palabra mal escrita, un error gramatical o sintaxis incorrecta. Para conocer las habilidades de redacción de los estudiantes, debemos observarlos desempeñarse – ¡ellos tienen que escribir! Debemos hacer que los estudiantes se desempeñen haciendo ellos mismos “el trabajo de edición”.

Más allá de esto, los promotores de las pruebas “auténticas” sostienen que la evaluación debería reflejar algún problema complejo de la vida misma. Dichas evaluaciones requieren necesariamente de mucho tiempo y, por lo tanto, son caras. Por ello, las pruebas de opción múltiple son utilizadas con más frecuencia, especialmente para propósitos de responsabilización a gran escala. En las aulas u otros escenarios de instrucción, las evaluaciones auténticas y directas son usualmente preferibles.

11. ¿Qué son las pruebas de desempeño?

Las pruebas de desempeño están estrechamente relacionadas a la evaluación auténtica. Podemos decir que toda evaluación auténtica involucra desempeño, pero podría haber algunas actuaciones o desempeños triviales que no satisfacen los requisitos como para ser

considerados evaluación auténtica. Por ejemplo, evaluar la redacción de los estudiantes involucra un desempeño: ellos tienen que escribir. Pero ¿sobre qué escriben? Si la tarea es trivial o banal, no está realizándose una evaluación auténtica.

La autenticidad también puede ser desbaratada al momento de la calificación. Por ejemplo, cuando los estudiantes sí escriben, alguien tiene que calificar sus composiciones. Si los textos son parte de un programa estatal de pruebas no serán calificados por profesores locales sino por alguna organización especializada en calificar muestras de redacción. Una evaluación estatal genera muchas composiciones y, como consecuencia, los calificadores tienen que evaluar los textos muy rápidamente, tanto como uno cada diez segundos. Tal rapidez impide una atención muy cuidadosa a la composición. Las composiciones son evaluadas según alguna fórmula y la creatividad auténtica bien podría ser castigada. Así, tanto la enseñanza como la evaluación resultan maltratadas.

12. ¿Qué son los portafolios?

Los portafolios son una variedad de la evaluación de desempeño, por lo general colecciones de varias clases de productos que se exhiben. Algunos distritos también usan portafolios de matemáticas, que pueden ser colecciones de actividades para resolver problemas y ejemplos de cómo los estudiantes resolvieron el problema de las matemáticas. En portafolios de ciencias, se recogen los resultados de experimentos u otras investigaciones. Los portafolios de redacción son los más comunes, sin embargo, y son considerados análogos a los portafolios de los artistas, colecciones que muestran una variedad de muestras de escritura que van desde exposiciones hasta la narrativa y la poesía. Los programas de transición de la escuela al trabajo frecuentemente involucran portafolios para demostrar el verdadero trabajo que se ha realizado.

13. ¿Qué es una prueba de "altas implicancias"

Una prueba de altas implicancias es una que da lugar a alguna forma de castigo para aquéllos que obtienen puntajes bajos o alguna forma de recompensa para aquéllos con puntajes altos, u, ocasionalmente, ambos. Por muchos años, las pruebas de altas implicancias más comunes fueron el *SAT* y el *ACT*. Los estudiantes con puntajes altos tenían una mejor oportunidad de ser admitidos en universidades selectivas. Esto es cierto aún, aunque después de que pasó la expansión demográfica, muchas universidades pasaron de *seleccionar* estudiantes a *reclutarlos*, para poder mantener o aumentar el número de sus programas y profesores. Las pruebas de CI (*IQ*) fueron también algunas veces de altas implicancias, dando lugar a que niños fueran colocados en programas de dotados y talentosos por un lado, ó en programas de nivelación ó de educación especial en el otro.

El problema con las pruebas de altas implicancias es que ellas hacen que la gente preste demasiada atención a elevar los puntajes, en detrimento de una educación más integral. Cuando mucho depende de los resultados, los maestros enseñarán técnicas para tomar pruebas, alinearán estrechamente su enseñanza con la prueba, y aún harán trampa ocasionalmente para verse bien, para hacer que su director o vecindario sean bien vistos o para mantener sus puestos de trabajo.

14. ¿Qué es una prueba de C.I. (*IQ*.)?

Una prueba de C.I., o Cociente de Inteligencia, es una prueba que mide algunas habilidades de pensamiento que están muy relacionadas con la escolaridad pero no son específicas a ella. Fueron desarrolladas inicialmente en Francia para determinar cuáles niños no recibían

beneficios de los programas escolares regulares. Cuando fueron trasladadas a este país, se consideraba que las pruebas de C.I. (*IQ*) medían "g" o un factor mental general que determinaba buena parte de las habilidades de pensamiento de una persona. El factor "g", a su vez, era considerado como una entidad controlada por un solo gen. El rol de la genética en la determinación de la inteligencia aún se debate mucho, como se apreció en la controversia sobre el libro *La Curva Acampanada*.

La teoría genética de la inteligencia fue rebatida por aquéllos que creen que la experiencia y el ambiente son más importantes, dando lugar desde los años 20 al debate "naturaleza - crianza". Hoy en día se reconoce como ingenua esa polarización, pero se sigue debatiendo sobre la dimensión e importancia del rol que juegan los genes versus el ambiente. Algunos arguyen que los genes son responsables de tan poco como 20% de la inteligencia y otros sostienen que determinan el 80%.

Desde el comienzo, no todos suscribieron la teoría del factor "g". Otra teoría extendida alega que la inteligencia se compone de un número de habilidades específicas. En 1983 Howard Gardner propuso una teoría de siete inteligencias separadas que goza la aceptación de muchos educadores. Desde entonces él ha añadido dos inteligencias adicionales.

Las pruebas de C.I. (*IQ*) consisten de habilidades específicas, tales como repetir una cantidad de dígitos escogidos al azar, o usar cubos o cuadros para copiar un dibujo mostrado por quien administra las pruebas. La Stanford - Binet, una de las pruebas de C.I. (*IQ*), más populares, usa 15 subpruebas agrupadas en cuatro categorías; razonamiento verbal, razonamiento cuantitativo, razonamiento abstracto/visual y memoria de corto plazo.

Las pruebas de C.I. (*IQ*) han sido criticadas por tener un sesgo cultural. Esta es un área muy cargada, demasiado compleja para resolverla acá. Basta decir que los niños de familias más acomodadas tienen más posibilidad de tener experiencias de vida que contribuyen a un mayor rendimiento en las pruebas de C.I. Por ejemplo, un estudio encontró que los padres de clase media hablaban con sus hijos cuatro veces más que los padres que vivían en la pobreza. También tienen más temprano acceso a libros.

Pruebas como la Stanford-Binet o las de C.I. de Wechsler deben ser administradas individualmente por administradores altamente entrenados. Se han desarrollado pruebas de C.I. administradas en grupo, pero éstas no suelen llamarse pruebas de C.I. Son usualmente dadas en la escuela, en conjunción con pruebas de logros. Los puntajes en estas pruebas se usan para "predecir" los puntajes de pruebas de logro académicos del estudiante. Las pruebas de C.I. son generalmente consideradas como una forma de prueba de "habilidad", en contraste con las pruebas de logros. Esto no es algo conceptualmente sólido, como se anota en la próxima sección.

15. ¿Cuál es la diferencia entre una prueba de habilidad o aptitud y una prueba de logros?

Principalmente, las pruebas de logro están conectadas más directamente a lo que se enseña en la escuela que las pruebas de habilidad. Casi toda la gente piensa que éstas son diferentes "tipos" de pruebas, y la distinción ha causado muchos problemas. Casi todos piensan en la habilidad en términos de "potencial". Los estudiantes que sacan puntajes más bajos en una prueba de logros que en una prueba de habilidad son a menudo catalogados como de bajos

méritos, es decir que no responden a su habilidad. La etiqueta "de altos méritos" se aplica a gente cuyos puntajes en las pruebas de logros son más altos que sus puntajes en pruebas de habilidad.

Sin embargo, una sola prueba nunca puede medir "potencial". Todo lo que puede medir es lo que los estudiantes saben y pueden hacer en un punto único, en el momento que rinden la prueba. Cuando se mira de cerca, las distinciones entre logros y habilidad se tornan conceptualmente borrosas; las pruebas con diferentes nombres no necesariamente miden cosas diferentes. Casi todo lo que podemos decir con validez es que el conocimiento y las habilidades que se miden en las pruebas de logro se parecen a los tipos de cosas que se enseñan en las escuelas, mientras que las habilidades que se miden en las pruebas de habilidad no parecen de base escolar y dependen menos de conocimientos específicos. Por ejemplo, algunas pruebas de habilidad tienen ítems de analogías. El pensamiento analógico es importante para tener éxito en el colegio, y probablemente también en la vida, pero los maestros casi nunca enseñan explícitamente a los niños a trabajar con analogías.

Algunas pruebas de habilidad también incluyen items perceptuales. Estos a menudo toman la forma de una serie de cuatro o cinco figuras geométricas. La tarea del estudiante es seleccionar cuál de otro grupo de figuras sería el siguiente en la serie. Los estudiantes que son buenos en este tipo de ítem a menudo desarrollan buenas habilidades en juegos de video o juegos perceptuales como ajedrez. Los estudiantes con puntajes altos en estas pruebas "no verbales" y bajos en las partes verbales y cuantitativas de las pruebas de habilidad tienen dificultades en el colegio. Estos niños son perceptualmente orientados, pero las escuelas tienen mucho que ver con símbolos: números y letras.

Algunos han sostenido que las pruebas de habilidad *predicen* los logros futuros y que las pruebas de logros *resumen* los logros pasados. Esto es solamente una convención que describe usos corrientes. Las pruebas de logros pueden ser usadas para predecir logros futuros. Cuando hacemos tales predicciones, estamos correlacionando los puntajes de las pruebas con algo en el futuro, como el éxito en la universidad. Cualquier prueba puede ser usada para hacer tales predicciones.

A decir verdad, las correlaciones involucradas pueden ser calculadas para dos variables cualesquiera. Podríamos usar altura o peso o densidad de cejas para predecir logros futuros -- todo lo que necesitamos hacer es meter dentro de la ecuación las diferentes alturas, pesos o número de pelos de la gente y sus notas de universidad. El que dichas predicciones den o no resultados significativos y estadísticamente importantes es otro asunto. Podríamos encontrar que la densidad de las cejas no predijo nada, en cuyo caso tendríamos que cesar de usarla como pronosticadora y, aun cuando sí predijera las calificaciones, no está claro que las políticas de admisiones tendrían entonces que ser cambiadas para tomar esto en cuenta. Típicamente, las pruebas de logros dadas en la secundaria predicen calificaciones en la universidad tan bien como la prueba de "aptitud" o "habilidad" más frecuentemente usada, la *SAT*.

16. ¿Qué son *ITBS, ITED, TAP, STANFORD-9, METRO, CTBS y TERRA NOVA*?

Excepto la *ITED*, todas éstas son pruebas de logros muy difundidas y producidas comercialmente y referidas a normas: las pruebas Iowa de Habilidades Básicas, las pruebas de Iowa de Desarrollo Educacional, Pruebas de Logro y Proficiencia; la 9ª versión de las Pruebas de Logros de Stanford, las Pruebas de Logros Metropolitanas, las Pruebas Integrales

de Habilidades Básicas; y una nueva versión de las Pruebas Comprensivas de Habilidades Básicas con un nombre caprichoso, Terra Nova.

Una "batería" completa, como a menudo se les llama, ofrece pruebas de lectura, matemáticas, lenguaje, vocabulario, ciencias y estudios sociales. Las últimas dos no se aplican tan a menudo como las primeras tres, debido a la gran variedad de programas curriculares de ciencias y estudios sociales que se dan en las escuelas. A menos que las currícula de ciencias y estudios sociales hayan sido específicamente alineadas con las pruebas, las pruebas podrían no reflejar lo que se está enseñando en un determinado grado.

La *ITED*, para los grados 9-12, no se usa en muchos lugares porque es bastante más difícil que las otras. Contiene pasajes largos de lectura, requiere que los estudiantes resuelvan problemas de matemáticas de pasos múltiples y que analicen experimentos simulados de ciencias. Casi todos los estados y distritos aplican la más fácil *TAP*.

17. ¿Qué es una prueba de competencia mínima?

Como fue originalmente concebida, una prueba de competencia mínima era una garantía de que los alumnos del último año de secundaria estaban saliendo del colegio "mínimamente competentes".

En los años '70, como ahora, la gente se preocupaba de que los estudiantes estuviesen siendo "promovidos socialmente" en base al tiempo pasado en la escuela y dejaban el colegio sin tener un mínimo nivel de habilidades. Pronto se vio, sin embargo, que el nivel mínimo no podía ser especificado a través de medios técnicos. Siempre hubo cierta arbitrariedad al establecer qué habilidades serían medidas y cuál sería el puntaje de corte.

Las pruebas de competencia mínima se volvieron muy populares; en un momento existían en alguna forma en 35 estados. Han sido reemplazadas más recientemente por lo que se conoce generalmente como "el movimiento de estándares" que demanda "estándares altos", "expectativas altas" y "un currículum desafiante para todos los estudiantes" -- algo más que un mínimo. Los puntajes de corte fueron usualmente establecidos para que suficiente cantidad de estudiantes fallasen inicialmente como para satisfacer a aquéllos que habían reclamado las pruebas en primer lugar, pero para que en el momento de la graduación virtualmente todos hubiesen aprobado. Una decisión judicial que emana de la época de las pruebas de competencia mínima que podría aparecer otra vez, sostenía que para que un estado retuviese diplomas basándose en una prueba, tenía que probar que los niños habían sido realmente provistos de oportunidades para aprender el material de la prueba (Debra P. vs. Turlington 1981).

18. ¿Qué son las pruebas de colocación avanzada? (*Advanced Placement –AP*)

Las pruebas de colocación avanzada las toman estudiantes de secundaria para ganar crédito universitario. Desde su origen en 1900, el Consejo de Universidades ha tratado de "impulsar" la enseñanza por medio de las evaluaciones. Las pruebas de colocación avanzada (AP) son la culminación de un esfuerzo por proveer a los estudiantes de secundaria con instrucción de alta calidad en áreas de estudios escolares construidas alrededor de un determinado currículo y que culminan con pruebas basadas en ese currículo. Normalmente, aunque solo aproximadamente la mitad de las escuelas secundarias de EEUU ofrece cursos de colocación

avanzada, más de un millón de estudiantes toma las pruebas *AP*, cada año, cifra diez veces mayor que hace 20 años.

El mayor incentivo para tomar las pruebas es el crédito universitario. Los calificadores entrenados evalúan las pruebas *AP* en una escala de cinco puntos, y muchas universidades conceden crédito por puntajes de tres o más. Ya que las pruebas cuestan mucho menos que los cursos universitarios, los estudiantes que salen airosos logran un inicio acelerado en la universidad y a la vez ahorran dinero. NO es necesario tomar un curso *AP* antes de rendir la prueba *AP*. Muchos colegios secundarios ofrecen cursos "avanzados" o "acelerados" o de "excelencia", que cumplen casi lo mismo sin adherirse estrictamente al syllabus del *AP*.

Un incentivo secundario para tomar las pruebas *AP* es la admisión a la universidad. Los funcionarios de admisiones han favorecido a los estudiantes que toman las pruebas *AP*, particularmente a los que toman más de una. Un incentivo para los padres es el dinero. Los cursos *AP* se dan gratis en las escuelas públicas y las pruebas cuestan $ 75, considerablemente menos que los cobros por los mismos cursos y prueba en la universidad.

19. ¿Qué es el bachillerato internacional?

El bachillerato internacional (BI) es un programa riguroso de estudios que se originó en Suiza. Los exámenes BI algunas veces son comparados con los de *AP*, pero hay una diferencia. Los estudiantes pueden tomar una prueba de *AP* sin haber tomado un curso *AP*. Para ser elegible para un examen BI en cambio, los estudiantes deben estar matriculados en un colegio que haya sido acreditado por medio del proceso de acreditación muy riguroso del BI y estar tomando el curso en el cual desean ser examinados. En el sistema BI, el examen determina 75% de la calificación en el curso. Mientras el número de pruebas BI dados en los EEUU se ha triplicado en la última década, aún es pequeño en comparación a los *AP*, ya que sólo se rinden 14,000 pruebas anualmente.

20. ¿Qué es la evaluación nacional de progreso educativo?

La Evaluación Nacional de Progreso Educativo, (*National Assessment of Educational Progress - NAEP*) empezó como un estudio nacional sobre lo que los estudiantes y adultos jóvenes saben y pueden hacer en las áreas de lectura, matemáticas y ciencias. Desde sus inicios a finales de los años sesenta, el *NAEP* ha añadido a sus evaluaciones historia, geografía, redacción y, más recientemente, arte y cívica. La idea original del *NAEP* fue simplemente establecer lo que sabe y no sabe una muestra de personas. Sus creadores lo vieron mucho como una encuesta de salud que pudiera determinar la incidencia de diversas enfermedades. Sin conocer la frecuencia de, por ejemplo, la tuberculosis, sería difícil saber cuánto esfuerzo sería necesidad par erradicarla.

En 1982, el Servicio Educacional de Pruebas (*Educational Testing Service – ETS*) ganó el contrato federal para administrar el *NAEP* (el dinero es parte del Presupuesto del Departamento de Educación de los EEUU) y lo apodó "La Libreta de Notas de la Nación". Esto puede ser solo parcialmente cierto, porque el *NAEP* no está alineado con ningún currículo en particular. Los estudiantes que aprenden matemáticas con el programa "Matemáticas Conectada" bien podrían tener diferentes puntajes *NAEP* que los de un distrito en que aprenden con “Saxon Math”. Así pues, uno no puede especificar que un currículo sea "mejor" que el otro. Cuando se propuso, mucha gente y organizaciones temieron que *NAEP*

llevaría a un currículum nacional estándar y al control federal de la educación. Como consecuencia, *NAEP* se ubicó en una agencia de política mantenida por el estado, la Comisión de Educación de los Estados con base en Denver, y fue prohibida de informar datos en ningún conjunto más pequeño que una "región". En 1988 una nueva legislación federal permitió brindar información a nivel del estado y ahora cerca de 40 estados participan en evaluaciones de *NAEP* estado por estado.

21. ¿Qué es el Consejo de Directores de la Evaluación Nacional (*National Assessment Governing Board – NAGB)?*

En los años ochenta se formó un Consejo de Directores de la Evaluación Nacional (*NAGB*) para dar lineamientos de política para la conducción de *NAEP*. La *NAGB* se propuso cambiar el *NAEP* para que en lugar de ser una evaluación de cómo son las cosas, fuera un programa sobre cómo deberían ser. Es decir, el *NAGB* hizo pasar al *NAEP* de ser descriptivo a ser prescriptivo.

Para hacer esto, la *NAGB* estableció "niveles de proficiencia" para cada una de las pruebas, denominando a los desempeños ya sea como "básico" "proficiente" o "avanzado" (es posible obtener puntajes "menos que básico", pero éste realmente no es un nivel como los otros). Estos niveles de proficiencia han sido criticados por estudios conducidos por la Oficina de Contabilidad General, el Centro para Investigación en Evaluación, Estándares, y Pruebas para Estudiantes (CRESST) y algunos eminentes psicometristas de los EEUU. En la primavera de 1999, el Concejo Nacional de Investigación declaró que los niveles de proficiencia eran "fundamentalmente defectuosos" y deberían ser reemplazados.

Los niveles de proficiencia no ofrecen una perspectiva sobre el desempeño estudiantil que resulte corroborado por otros indicadores. Por ejemplo, en las más recientes evaluaciones *NAEP* de matemáticas y ciencias, pocos niños de cuarto grado lograron "proficiencia" y virtualmente ninguno de los de cuarto grado logró el nivel "avanzado". Sin embargo, estos mismos estudiantes de cuarto grado obtuvieron puntajes por encima del promedio en matemáticas, comparados con estudiantes de 26 naciones, y fueron terceros en el mundo en ciencias. Además los puntajes del *NAEP* en matemáticas y ciencias han mejorado desde 1977, el primer año para el cual se recogieron datos de tendencias de largo plazo.

22. ¿Qué es el Tercer Estudio Internacional de Matemáticas y Ciencias (*TIMSS*)?

TIMSS es un tercer intento por comparar los logros en matemáticas y ciencias de distintas naciones. Las comparaciones internacionales se han vuelto un barómetro sobre cómo les está yendo a las escuelas por todo el mundo. El Tercer Estudio Internacional de Matemáticas y Ciencias (*TIMSS*) es, a la fecha, el estudio más grande, más reciente y mejor controlado sobre el tema.

Sin embargo, tiene problemas. Por lo pronto, la confiabilidad de las pruebas no es impresionante, algo que ha sido casi pasado por alto. *TIMSS* administró pruebas a estudiantes de 26 países en el 4° grado, en 41 países en el 8° grado y entre 16 y 21 países, dependiendo de la prueba, en el año final de escuelas secundarias. Se llama Año Final porque, en muchos casos, no corresponde al 12° grado en los Estados Unidos.

El *TIMSS* ha generado un estereotipo muy difundido pero falso: cuanto más tiempo están los estudiantes americanos en la escuela, más retrasados quedan con respecto a sus pares extranjeros. El estereotipo se deriva del hecho de que los estudiantes americanos obtienen puntajes muy altos tanto en matemáticas como en ciencias en el nivel de 4° grado, puntajes promedio en el 8° grado y casi los más bajos en el estudio del Año Final. La declinación del 4° al 8° grado es probablemente real, pero la caída mayor entre el 8° y el 12° probablemente no lo es.

Uno de los descubrimientos del componente de análisis curricular del *TIMSS* fue que los educadores americanos consideran los grados sexto a octavo como la culminación de la escuela primaria, mientras que la mayoría de las otras naciones industrializadas los ven como el comienzo de la secundaria y del estudio académico mas intenso. La consecuencia es que el 7° y 8° grados en muchos otros países incluyen el estudio del álgebra y la geometría, mientras que sólo alrededor del 15% de los niños americanos reciben instrucción en álgebra en 8° grado. El resto recibe un repaso de temas anteriormente cubiertos. En parte, este repaso es necesario, debido a otra cosa descubierta en el análisis curricular de TIMSS: los libros de texto americanos son como tres veces más gruesos que los de otras naciones. Los maestros en otros países enseñan menos temas e invierten más tiempo en cada uno. Los maestros americanos tratan de enseñar todo lo que está en el texto. Esto da una cobertura que a menudo es breve y superficial.

Los resultados del Año Final, que indican un desempeño pobre de los estudiantes americanos, son ostensiblemente equívocos. Solo cinco naciones cumplieron con los criterios establecidos por el mismo estudio para considerar válidos los datos. Por otra parte, los sistemas educativos de la mayoría de otros países no son comparables a los de los Estados Unidos después del 8° grado (ni en muchos casos, entre ellos mismos). En otras naciones, los estudiantes entran a programas focalizados; algunos cursan estudios intensivos de matemáticas y ciencias, otros entran a programas técnicos o vocacionales, otros reciben formación en las artes y humanidades. La duración de estos programas varía, pero los estudiantes de las otras naciones tenían como promedio un año más de edad que los estudiantes norteamericanos y algunos tenían la misma edad que los estadounidenses que están en el último año de la *universidad.*

También hay diferencias culturales entre los países que producen grandes diferencias en los puntajes de pruebas. En casi todas las otras naciones, los estudiantes son estudiantes, no estudiantes y trabajadores. Pero 55% de los estudiantes norteamericanos en el estudio indicó que trabajaba más de 21 horas por semana. La investigación sobre la relación entre el trabajo y el desempeño en el colegio encuentra que trabajar hasta 20 horas a la semana está asociado con un mejor desempeño, pero más allá de eso, trabajar tiene un impacto perjudicial en la escolaridad: los estudiantes no duermen suficientemente, flojean en las tareas para hacer en casa y se saltan las comidas, especialmente el desayuno.

Los estudiantes estadounidenses que no trabajaban mucho tuvieron puntajes similares al promedio internacional, al igual que los del 8° grado. Aquellos que trabajaban 21 - 35 horas por semana (28%) estaban bastante por debajo del promedio y los que trabajaban más de 35 horas por semana (27%) salían fuera del cuadro. Cuando uno selecciona a subgrupos de estudiantes Americanos que más se parecen a sus pares extranjeros en otras dimensiones, ellos también obtienen puntajes promedio, como los obtenían los del octavo grado.

23. ¿Qué es "*How in the World do Students Read*"?

Este es el titulo de un libro publicado en 1992 que resume un estudio internacional de lectura conducido por la misma organización que realizó el TIMSS. Es virtualmente desconocido. Los niños norteamericanos de 10 años y los de 14 años sólo fueron superados por estudiantes de una nación, Finlandia. Hubo 27 países que participaron en la prueba para los de menor edad y 31 en la de mayores.
A los estudiantes americanos les ha ido bien siempre en comparaciones internacionales de lectura. Esto es probablemente debido al esfuerzo concertado que hacen los maestros de primaria para enseñar lectura y la menor cantidad de tiempo que dedican a las matemáticas y ciencias.

24. ¿Qué es el Consejo de Universidades?

El Consejo de Universidades (*College Board*) empezó en 1900 como un grupo pequeño de universidades del noreste de los EEUU. Por muchos años se llamó el Consejo de Pruebas de Admisión Universitaria y su objetivo inicial era darle coherencia a la currícula de las escuelas secundarias. Las universidades habían encontrado que estudiantes con calificaciones escolares parecidas a menudo tenían experiencias muy distintas en términos de la sofisticación y el rigor de los cursos que habían tomado. El Consejo pensó que podría eliminar la confusión desarrollando exámenes en diversas materias. A partir de estos exámenes, las escuelas secundarias podrían determinar qué era lo que las universidades valoraban y, de acuerdo a eso, cambiar su currícula. Impresionado con los procedimientos para hacer pruebas desarrolladas por los militares durante la Primera Guerra Mundial, el Consejo decidió desarrollar una sola prueba para predecir el éxito en la universidad. En 1926, introdujo la Prueba de Aptitud Académica (*Scholastic Aptitude Test*) a la que casi siempre se hace referencia con sus iniciales, *SAT*. La mayoría de las actividades del Consejo están orientadas a apoyar algún aspecto de las 3,300 instituciones que constituyen su membresía.

Hasta ahora tiene como función principal facilitar la transición de la secundaria a la Universidad. Establecida legalmente como una corporación sin fines de lucro, en el otoño de 1999 el Consejo anunció su primer emprendimiento comercial, una página web que ofrecerá tutoría a bajo costo para el *SAT* y los cursos *AP* e información sobre ayuda financiera. El Consejo de Universidades podría en el futuro dar cursos de *AP* por medios electrónicos.

25. ¿Qué es el Servicio Educacional de Pruebas - ETS (*Educational Testing-Service)*?

ETS es una gran organización de pruebas y de investigación sobre pruebas que tiene su matriz en Lawrence Township, New Jersey, cerca de Princeton. Se desprendió del Consejo de Universidades en 1947. Sus productos más conocidos son la Prueba de Evaluación Académica (nacida como Prueba de Aptitud Académica) y desde 1982, la Evaluación Nacional de Progreso Educativo (*NAEP*). También desarrolla y administra pruebas de admisión a programas de leyes y medicina y pruebas para ser usadas en los negocios y la industria.

26. ¿Qué es el *SAT*?

El Consejo de Universidades desarrolló el *SAT* en 1926. Hasta 1994, las siglas representaban al "*Scholastic Aptitude Test*" (Prueba de Aptitud Académica). El primer *SAT* contenía tanto

preguntas de opción múltiple como de ensayo. Cuando el estallido de la Segunda Guerra Mundial impidió la administración de la parte ensayística, el Consejo decidió usar sólo la sección de opción múltiple para todas las administraciones.

Cuando el *ETS* cambió el nombre del *SAT* en 1994 a Prueba de Evaluación Académica, también empezó a referirse a la prueba como una prueba de "razonamiento", pero poco es lo que se cambió, excepto un mayor énfasis en "lectura crítica" y la supresión de la sección sobre antónimos.

El nuevo *SAT* tiene 138 ítems y se dispone de 180 minutos para responder la prueba, así que no es posible mucho razonamiento "profundo" sobre cada pregunta. El *ETS* convierte los puntajes crudos en puntajes escalares, de manera que la media es 500 y la desviación estándar es 100, produciendo una escala que va de 200 a 800.

Actualmente, cada año aproximadamente 1,200 000 alumnos del último año de secundaria toman el *SAT*. Cuando se incluye a alumnos del penúltimo y antepenúltimo año que también la rinden, el *ETS* administra alrededor de 2,000,000 de *SAT*s al año.

27. ¿Qué es el *PSAT*?

El *PSAT* es la "Prueba Preliminar de Evaluación Académica". Es una versión abreviada del *SAT* que contiene preguntas viejas del *SAT*. Algunas veces la toman alumnos del 10° grado como práctica. También es el criterio único con el cual los estudiantes pueden buscar las Becas Nacionales al Mérito. Este último uso es problemático, porque los varones sacan mejores puntajes en el *PSAT* y el *SAT* que las niñas. Como consecuencia, los varones obtienen hasta dos tercios de las becas. *ETS* añadió una prueba de redacción al *PSAT* y, debido a que las niñas se desempeñan mejor que los niños en esta prueba, el diferencial en la distribución de las becas se ha reducido en más o menos 50%.

28. ¿Qué es la Corporación Nacional de Becas al Mérito?

La Corporación Nacional de Becas al Mérito es una organización independiente sin fines de lucro de Evanston, Illinois, que administra dos programas de becas: el Programa Nacional de Becas al Mérito y el Programa Nacional de Becas por Logros. La Corporación utiliza el *PSAT* para seleccionar a los posibles becarios.

Cada año, cerca de 35,000 estudiantes con los puntajes más altos de *PSAT* reciben "Cartas de Reconocimiento", mientras que otros 15,000 son designados como semi- finalistas. Se les pide llenar solicitudes de beca y eventualmente cerca de 6,500 las reciben.

29. ¿Qué es el *ACT*?

Estas siglas indican tanto una serie de pruebas de admisión a las universidades como la organización que las elabora, el *American College Testing Program* localizado en Iowa City, Iowa. Mientras que los constructores del *SAT* querían identificar a estudiantes académicamente dotados y traerlos a las universidades costeras del Este, los elaboradores del *ACT* estaban más interesados en proveer tanta información académica como tutoría para *todos* los estudiantes que estarían asistiendo a las universidades estatales, especialmente las

universidades públicas del Medio Oeste. Cerca de 900,000 estudiantes del último año toman actualmente la batería de pruebas de *ACT*. La mayoría de universidades ahora aceptan tanto el *SAT* como el *ACT* para efecto de admisiones.

30. ¿Qué es *FAIRTEST*?

"*FairTest*" es el nombre más conocido de lo que formalmente es el Centro Nacional para la Evaluación Justa y Abierta de Cambridge, Massachusetts. *FairTest* empezó principalmente como una organización anti - *ETS* con su atención focalizada sobre el *SAT*. Desde que se fundó, ha ampliado su ámbito para interesarse en asuntos de equidad de género y equidad étnica y con cuestiones vinculadas al "movimiento por los estándares".

31. ¿Qué es un estándar?

La palabra admite muchas definiciones. Puede ser un estandarte o algo que registra una magnitud, como una barra de platino que establece los estándares para longitud. O puede ser algo ordinario o familiar, como una calidad estándar de carne o el equipamiento estándar en un auto. En la esfera de la educación, sin embargo, estándar se usa usualmente en referencia a un grado o nivel de exigencia, excelencia o logro (definido en el *American Heritage Dictionary*).

El "movimiento por los estándares" no es una organización o esfuerzo formal, sino que ha brotado de una preocupación tanto porque los estudiantes americanos no están aprendiendo lo suficiente como porque lo que están aprendiendo no es de calidad o rigor suficientemente alto. El que esto sea cierto o no es materia de considerable debate.

32. ¿Qué es un estándar de contenido? ¿Qué es un estándar de desempeño?

Los estándares de contenido especifican **qué**, los estándares de desempeño **cuánto**. Desde que el Consejo Nacional de Maestros de Matemáticas publicó sus estándares curriculares en 1989, casi todos los estándares han sido estándares de contenido, explicitando lo que los elaboradores de estándares pensaban que los estudiantes deberían saber o, por lo menos, estar expuestos a ello. Las pruebas que han sido elaboradas en torno a estos estándares de contenido, con sus puntajes de corte para la aprobación, pueden ser considerados estándares de desempeño. Los niveles de proficiencia del *NAEP* discutidos anteriormente fueron intentos para establecer los estándares de desempeño en las diversas evaluaciones *NAEP*.

33. ¿Qué es el alineamiento?

El alineamiento se refiere al grado en que un currículo está alineado con una prueba y viceversa. Es importante que una prueba esté alineada con un currículo. De otro modo, la prueba mediría cosas que no han sido enseñadas. Por otra parte, alinear un currículo con una prueba tiene sus problemas, porque la prueba abarca solo una pequeña parte de cualquier currículo. El alineamiento podría reducir el currículo.

En la evaluación de programas educacionales, es importante tener la prueba alineada con los objetivos del programa. Sin alineamiento, un programa efectivo podría parecer no serlo. En una evaluación del programa remedial "Éxito Para Todos", por ejemplo, las metas de la prueba (las Pruebas Integrales de Habilidades Básicas en este caso) no se emparejaban

completamente con los objetivos del programa educativo. Esto puede haber atenuado el impacto aparente del programa.

34. ¿Qué son las acreditaciones o certificaciones?

Acreditar es usar pruebas para otorgar o denegar credenciales o licencias para profesiones específicas. Un número de estados usa pruebas para acreditar o certificar que los maestros saben lo suficiente como para entrar a un aula. El uso de pruebas para este objetivo ha sido debatido acaloradamente a lo largo de los años. Algunos sostienen que gran parte de la enseñanza involucra una serie de habilidades no relacionadas al conocimiento de contenidos específicos y que estas habilidades no pueden ser medidas con pruebas de lápiz y papel. Otros afirman que todos los maestros necesitan algún nivel mínimo de conocimientos, independientemente de cualesquiera habilidades para la enseñanza que ellos posean. También hay pruebas para certificar a los abogados, médicos, contadores, y muchos otros profesionales. Estas pruebas también son desarrolladas por alguna de las editoras privadas de pruebas, generalmente en coordinación con la organización profesional que supervisa la profesión, tal como la Asociación Médica Americana, la Asociación Americana de Abogados, etc.

PARTE III

Algunas Cuestiones sobre las Pruebas

1. ¿Por qué es enseñar para una prueba un problema en círculos educacionales pero no en círculos atléticos?

Hace como 75 años, un educador observó que los entrenadores de tenis "enseñaban para la prueba". Es decir, instruían a sus alumnos en precisamente aquellas cosas que necesitarían para tener éxito en su deporte: cómo servir, cómo lanzar la bola, cómo volear, cómo llegar a la red. Esto es enseñar para la prueba, y es una práctica ampliamente aceptada. Sin duda, creeríamos loco a cualquier entrenador si hiciera otra cosa.

¿Por qué, entonces, es enseñar para la prueba un problema en educación? La respuesta es que el entrenamiento de tenis o de fútbol incorpora todos los aspectos del deporte, mientras que el entrenamiento para una prueba específica en educación usualmente no lo hace. El entrenamiento para el fútbol podría salir mal si el equipo opositor estableciera nuevas jugadas pero, en este caso, la "prueba" también se vuelve un instrumento de enseñanza: los jugadores aprenderán algo mientras lidian con las nuevas jugadas del opositor.

El currículo de, digamos, matemáticas puede ser considerado como un círculo grande que incorpora todo el campo. La prueba es una serie de círculos más pequeños que muestrean partes del grande. Siempre y cuando los maestros estén trabajando en todo el campo, la prueba es una representación válida de lo que está pasando, así como una veta de mineral representa el yacimiento más grande. Pero si uno se concentra sólo en la parte del campo cubierto por la prueba, la educación sufre. Teóricamente, las pruebas podrían cubrir todo un dominio, pero tomarían muchas horas y muchas horas y muchos dólares para administrarse.

Las pruebas de logro usadas comúnmente en las escuelas suelen tener sólo 25-40 ítems para cubrir una asignatura. En algunas evaluaciones de desempeño, nos acercamos a un sistema como el deporte. Podemos enseñar aspectos de redacción y luego hacer que los estudiantes escriban y observar cuán bien han aprendido esos aspectos. Sin embargo, como vimos en la sección sobre pruebas de desempeño, esta práctica puede ser desvirtuada si las muestras de redacción estudiantil son calificadas rápidamente usando una fórmula que se concentra en unos pocos elementos e ignora o hasta castiga la creatividad.

2. ¿Quién desarrolla pruebas?

Casi todas las pruebas en los EEUU son desarrolladas por editoriales comerciales tales como *CTB Mc Graw - Hill, Riverside*, *Harcourt, Educational Measurement* o por organizaciones privadas sin fines de lucro tales como el *ACT* o *ETS*. Unas pocas firmas son especializadas: *Measurement Incorporated* de Carolina del Norte califica muestras de redacción; *National Computer Systems* de Iowa se especializa en la calificación masiva de hojas de respuestas, *National Evaluation Systems* de Massachusetts se especializa en pruebas de profesores; y *Advanced Systems* en New Hampshire se especializa en pruebas de desarrollo hechas por encargo.

En años recientes, cada vez más, el desarrollo de pruebas ha tenido lugar a nivel de estados. A iniciativa de un gobernador, la legislatura o el Consejo de educación de un estado, se ha diseñado un programa de pruebas específicamente para un determinado estado. Así, existe el *Texas Assesssment of Academic Skills*, las pruebas de *Virginia Standard of Learning, the Massachusetts Comprehensive Assessment System*, y otras.

En algunos casos, como el de Virginia, las pruebas han sido derivadas de un marco curricular determinado. Estas pruebas son inicialmente desarrolladas por las firmas privadas de pruebas según especificaciones dadas por los estados. Las pruebas son luego revisadas por maestros y supervisores y por profesores universitarios en los diversos estados. Algunos estados tales como Virginia y Carolina del Norte han hecho contratos con investigadores universitarios para determinar si las pruebas responden a requerimientos técnicos en relación a su confiabilidad y validez.

3. ¿Qué agencias supervisan el debido uso de las pruebas?

Casi no hay regulación alguna de la industria de las pruebas. La Asociación Americana de Investigación Educacional, la Asociación Psicológica Americana, y el Consejo Nacional sobre Medición en Educación desarrollaron y adoptaron conjuntamente unos *Estándares para el Uso de Pruebas*, pero poca gente, salvo quienes realizan investigaciones con las pruebas, presta atención a estos estándares. Cuando las pruebas han sido mal utilizadas, como recientemente en Chicago y California, ni los editores de pruebas ni alguna de las tres organizaciones antes mencionadas han elevado objeciones públicas a las violaciones.

Tanto en Chicago como en California, se usa solamente los puntajes de pruebas para determinar si los niños han de ser o no promovidos o retenidos en un grado. Esto viola, por lo menos, dos estándares: que una prueba sólo no debería ser usada para tomar decisiones sobre personas, y que una prueba diseñada para un propósito no debería ser arbitrariamente aplicada a otro propósito. Las pruebas usadas en Chicago y California son pruebas de logros referidas a normas, que no fueron diseñadas para decisiones de promoción o retención, ni

están técnicamente a la altura de la tarea. Son lo suficientemente precisas como para decir que un niño está por encima o por debajo del promedio, pero no lo suficientemente precisas como para decir que un determinado niño debería quedarse otro año en el mismo grado.

Diversos educadores han reclamado algún tipo de "agencia de vigilancia" para monitorear a los que hacen las pruebas. A George Manaus del *Boston College* le gustaría ver una "Administración Federal para las Pruebas". Larry Cuban de *Stanford University* también arguye que el advenimiento de pruebas de altas implicancias aumenta la necesidad de una agencia supervisora ya que es fácil distorsionar el verdadero sentido de las cifras de las pruebas.

4. ¿Por qué causan tantos problemas los coeficientes de correlación?

El coeficiente de correlación es origen de muchos malentendidos porque los cerebros humanos parecen estar condicionados para inferir causalidad a partir una mera correlación. Sin embargo, con solamente un coeficiente de correlación, *no* podemos inferir causalidad. Las dos variables podrían estar causalmente relacionadas, ambas podrían estar siendo afectadas por una tercera variable, o la correlación podría simplemente ocurrir por algún artificio. Hay, por ejemplo, una correlación entre los puntajes del *SAT* y las calificaciones del primer año de universidad, pero no podemos decir que el *SAT causó* las calificaciones en la universidad.

5. ¿Por qué no hay un promedio nacional para el *SAT* o el *ACT*?

Hacia fines de agosto cada año, el Consejo de Universidades y el *ACT* dan a conocer los resultados más recientes del "promedio nacional" en el *SAT* y el *ACT*. Mucho se ha argumentado sobre estos números, desde que en 1977 un informe analizó las causas de lo que era, en esos momentos, un declive de 14 años en el puntaje promedio *SAT*.

El "promedio nacional" no es significativo por una serie de razones. Primero, los estudiantes que rinden la prueba son un grupo auto seleccionado y, cada año, una proporción más y más grande de todos los alumnos del último año la vienen rindiendo. Hace treinta años, alrededor del 30% de todas las promociones tomó el *SAT;* hoy la cifra es alrededor del 43%

El porcentaje creciente de estudiantes que rinden el *SAT* y el *ACT* representa una selección cada vez más a fondo en la reserva de talentos. Además, las características demográficas de quienes toman el *SAT* han estado cambiando, especialmente desde los años 60. El *SAT* fue estandarizado en un grupo pequeño de estudiantes blancos que vivían mayormente en el Noreste y que planeaban asistir a las universidades de elite.

Empezando en los 60, sin embargo, a medida en que las universidades se abrieron a mujeres y minorías, más de estos dos grupos han tomado el *SAT*. Además, más y más estudiantes de familias de bajos ingresos y estudiantes con promedios de calificaciones en la secundaria no destacados han aspirado a una universidad y han tomado la prueba. Bajo estas circunstancias, asombra poco que el promedio del *SAT* haya caído. Por una variedad de razones discutibles, las mujeres y las minorías (excepto los estudiantes asiáticos en la sección de matemáticas) no sacan tan buenos puntajes en el *SAT* como los hombres.

6. ¿Por qué decayó el puntaje promedio del *SAT*?

La demografía de quienes han estado tomando las pruebas ha estado cambiando con el tiempo y todos los cambios están asociados con puntajes de pruebas más bajos -- más mujeres, más minorías, mas estudiantes de familias de bajos ingresos, más estudiantes con promedios bajos de notas escolares. Un estudio encontró que el puntaje promedio del *SAT* hubiese subido entre 1975 y 1990 si una variable -- los puestos en el orden de mérito de los estudiantes de secundaria -- hubiera permanecido igual. Pero más y más estudiantes en el 40% más bajo de los puestos en secundaria tomaron el *SAT*.

Cuando el Consejo de Universidades armó un panel en 1976 para estudiar la caída del puntaje promedio del *SAT*, el panel concluyó que fueron muchos los factores que causaron la caída. Uno de los documentos de base para el panel simplemente presentaba una lista de las diversas hipótesis que habían sido propuestas para explicar la caída: había 74 de ellas!

EL distinguido panel llamó al periodo de declive una "década de distracción". Durante este periodo, el país había sido impactado por los asesinatos de John F. Kennedy Jr., Robert F. Kennedy Jr., Martin Luther King, Jr. y Malcom X. Había soportado una guerra impopular y protestas contra ella. Había sufrido el escándalo Watergate. Virtualmente todas las áreas urbanas habían experimentado serios disturbios. Durante la década, los periódicos podían mostrar casi una "injuria - del - día": la policía golpeando a los manifestantes en la Convención Democrática Nacional de 1968, una mujer joven llorando sobre el cuerpo de una amiga en Kent State University, etc. Las drogas recreativas se habían vuelto populares y la televisión era omnipresente. Poco asombra que la gente estuviese prestando menos atención al análisis de las partes de la oración y a la factorización de las ecuaciones. Otros indicadores de logros en este periodo cayeron junto con el *SAT*.

7. ¿Por qué se "recentró" el *SAT*?

El Consejo de Universidades tomó esta acción en 1996 para hacer que un puntaje de 500 una vez más reflejar el puntaje promedio de gente que solicitaba entrar a la universidad. Como se da cuenta en la sección "¿Por qué no hay un promedio nacional significativo para el *SAT*?", el grupo con el cual se estableció el estándar en 1941 era una elite. Específicamente, eran 10,654 estudiantes que vivían en el Noreste. Noventa y ocho por ciento era blanco, 61% era de sexo masculino, y 41% había asistido a secundarias privadas de preparación para la universidad. Esto difícilmente representaba al cuerpo de estudiantes que estaba tomando la prueba en 1996, el año en que se recentró. Ese año, más de 1,000,000 de estudiantes se amontonaban angustiados los sábados en las mañanas para tomar el *SAT*: 21% de ellos era de alguna minoría, 52% era mujer, y 83% había estudiado en las escuelas públicas. La "reserva" de los tomadores de pruebas se había ampliado sustancialmente y se había democratizado bastante. Sin embargo, el puntaje escalar de 500 había sido asignado al puntaje promedio verbal y de matemáticas de la elite con la cual se estableció el estándar.

En 1941 un puntaje escalar de 500 representaba el puntaje promedio de los que planeaban asistir a la universidad, por lo menos en el Noreste. En 1996, representaba el puntaje promedio de nadie. Los estudiantes que recibían un 464, en 1996, por ejemplo, podrían creer que estaban "debajo del promedio" porque, después de todo, 500 era el "promedio". Pero era el promedio sólo para ese grupo inicial con el cual se fijó el estándar en 1941. Así, en 1996,

el Consejo de Universidades decidió hacer que 500 representara una vez más el puntaje promedio de todos los que tomaban el *SAT*.

La acción del Consejo fue controversial, porque parecía que los puntajes se elevaron sin una buena razón o, por lo menos, por ninguna razón relacionada con cómo los estudiantes estaban desempeñándose realmente. "La mayor dosis de Prozak educacional en la historia", se dijo en son de broma. A la gente también le preocupaba que se perdiese la información sobre tendencias. Sin embargo, el *ETS* provee escalas que traducen de atrás para adelante, y viceversa, las escalas viejas y nuevas. La gente puede seguir tendencias con cualquier escala que prefiera. El recentrar cumplió con el objetivo del Consejo: hacer que 500 representase de nuevo el verdadero puntaje promedio de todos los que toman el *SAT*.

8. ¿"Funcionan" el *SAT* y el *ACT*?

La respuesta depende en parte de la perspectiva de cada uno y en parte de cómo uno define "funcionar". La función de ambas pruebas es predecir las notas de los alumnos en el primer año de la universidad. Ambas pruebas hacen esto, pero difícilmente sus predicciones son perfectas. La correlación típica entre los puntajes de las pruebas y las calificaciones de los alumnos de primer año es aproximadamente + 0.45. Esto significa que la prueba da cuenta de más o menos 20% de lo que se incluye en las calificaciones. Otros factores dan cuenta de alrededor del 80% (ver "¿Qué es un "Coeficiente de Correlación"?).

Esta correlación de 0.45 se reduce en universidades altamente selectivas porque los puntajes *SAT* en esos colegios son menos diferenciados. Cuanto más se parecen las personas unas con otras, con menos éxito podemos hacer predicciones diferenciales sobre su desempeño. Supongamos, por ejemplo, que Ud. quisiera predecir el efecto del peso corporal en el éxito como un jugador de defensa. Si todos los que van a jugar pesaran 275 libras, no podría hacer Ud. predicciones porque todos tendrían el mismo "puntaje", 275. Conforme los puntajes se tornan más y más diferenciados, es posible hacer mejores predicciones.

En las universidades no selectivas, la abrumadora mayoría de universidades norteamericanas, las calificaciones escolares combinadas con el puesto en la promoción predicen el éxito en el primer año universitario tan bien como los puntajes en las pruebas de admisión. En las universidades selectivas las pruebas usualmente predicen mejor que las calificaciones, porque las calificaciones están agrupadas aún más estrechamente que los puntajes en las pruebas.

9. ¿Confían demasiado las universidades en el *SAT*?

Probablemente no. La cultura popular cree que las pruebas funcionan y que puntajes bajos en el *SAT* anulan la oportunidad de un estudiante de ser admitido a una universidad selectiva. *USA Today* recientemente sacó una tira cómica mostrando a una mamá que leía a su niño en la cama. La madre decía: "y el pequeño chanchito con el más alto puntaje en matemáticas y lenguaje vivió feliz para siempre. Los otros dos fueron tragados por el lobo". En *Ninguno de los de arriba*, David Owen declaró que "gente que no recuerda el número de sus zapatos se acuerda de cuánto obtuvo en el *SAT*". A decir verdad, las universidades usan muchos factores para tomar decisiones sobre admisiones y obtienen información de cosas como portafolios, cintas de video e historias personales. Uno de los mitos alrededor de las admisiones a las universidades es que todos los postulantes están compitiendo con todos los

otros postulantes. La verdad es que las universidades selectivas admiten por categorías. Quieren "cerebros" ciertamente, pero también quieren "el sueño americano" y "legados" (hijos de ex - alumnos). Ellos hacen ajustes para admitir al "talento especial". Esto incluye no sólo a atletas, sino a muchos aspirantes a las bellas artes y las artes escénicas, que tienden a no desempeñarse bien en las pruebas de lápiz y papel. "La conciencia social" también ha sido una categoría desde los años '60, pero está en decadencia puesto que las cortes han dictaminado en contra de al menos algunos programas de acción afirmativa. Por último, los decanos de admisión prefieren "clientes que pagan" - aquéllos que pueden pagar los $ 20,000.00 mayores costos anuales, sin ayuda financiera de la universidad.

Como evidencia de que las universidades no usan sólo los puntajes *SAT*, puede considerarse los ingresantes al primer año en la universidad Brown, una de las universidades más selectivas de la nación. En 1988, la Brown podría haber llenado los salones de primer año con solo estudiantes con puntajes entre 750 y 800 en la prueba verbal del *SAT*. En realidad, admitieron estudiantes con puntajes entre 350 y 800. Sólo un tercio de los postulantes con puntajes entre 750 y 800 fueron admitidos. Mirando la cantidad de estudiantes admitidos que obtuvieron puestos altos en la secundaria, Brown parecía estar más interesada en el puesto en la promoción escolar que en los puntajes de las pruebas.

También hay evidencia de que las universidades selectivas no necesitan el *SAT* o el *ACT*. Hace algunos años las universidades Bates y Bowdoin hicieron el *SAT* opcional para las admisiones, pero aun así, lo requerían para ubicación y consejería. Los estudiantes que presentaron puntajes en el *SAT* con sus solicitudes tenían puntajes cerca de 150 puntos más altos que aquéllos que no los entregaron. Pero no tenían promedios escolares más altos. La universidad notó que se volvía más diversa geográficamente y étnicamente, así como con respecto a posibles especializaciones. Los profesores quedaron más contentos con el carácter de las clases que resultaron de haber hecho opcional el *SAT*.

Finalmente, una persona que a menudo da conferencias a funcionarios de admisiones declara que siempre pide que levanten las manos aquéllos que continuarían usando el *SAT* si las universidades, no los estudiantes, tuviesen que pagarlo. Dice que aún le falta ver un solo brazo levantarse.

¿Por qué es necesaria la "alfabetización en evaluación"?

Espero que el lector que haya leído atentamente las páginas previas haya salido mejor informado pero no abrumado. Las pruebas son emprendimientos mucho más complejos que lo que se presenta usualmente en los medios. Los medios no examinan los resultados anunciados de las pruebas. Suelen más bien aceptarlos mayormente sin crítica. Cuando los resultados de la *NAEP* de Cívica fueron publicados en noviembre 1999, sólo *Education Week* hizo notar que los niveles de proficiencia del *NAEP* son defectuosos. Todos los demás presentaron la noticia como estadísticamente correcta.

Y las pruebas siguen llegando. Conforme este breviario se iba terminando, estos ítems aparecían en las noticias:

- La Prueba Nacional propuesta por la Administración Clinton había sido desarrollada y estaba lista para ser probada en el terreno, esperando solamente fondos del Congreso. Los fondos no estaban llegando.

- Arizona publicó los resultados de su primer programa de pruebas de estado y 89% de los estudiantes fallaron. Es interesante que el superintendente de instrucción pública del estado rindió las pruebas y aprobó por las justas.
- Las evaluaciones de colegios *charter* en Michigan y Ohio concluyeron que era muy pronto para decir si con la nueva gestión mejoraban los puntajes en las pruebas.
- John Stossel presentó un segmento del programa 20/20 de ABC alabando a las escuelas católicas por sacar puntajes más altos que las escuelas públicas - y a costos mucho más bajos. Como muchos informes similares, éste no tomó en cuenta los salarios bajos, el costo de los edificios y los subsidios provistos por la iglesia.
- Una denuncia contra un maestro de Chicago que publicó algunos ítems de pruebas en un diario pasó a juicio.
- El candidato presidencial republicano, el gobernador George W. Bush, usó las mejoras de puntajes en las pruebas de Texas como un elemento saltante de su campaña. Los críticos inmediatamente cuestionaron la validez de estas supuestas ganancias.
- California completó su desarrollo de un Índice de Desempeño Académico basado en pruebas, para evaluar sus escuelas.
- La mayoría de los estudiantes reprobó una nueva prueba de estudios sociales de Michigan.
- La Asociación Nacional para el Progreso de la Gente de Color anunció que auspiciará cursos preparatorios *SAT-ACT* para estudiantes de minorías.
- El superintendente de escuelas de Kansas City, Missouri, anunció un programa para elevar puntajes con el fin de restablecer la acreditación del estado al distrito escolar.
- Interrogado sobre por que se había sido cancelado el viaje anual para observar ballenas, el superintendente de escuelas de Palo Alto del Este, California, contestó: "A los estudiantes no se les toma pruebas sobre observación de ballenas, así es que no van a ir a observar ballenas".

Podrá llegar un momento para la educación en que las pruebas y los puntajes en las pruebas pierden tanta prominencia, pero ese momento no es ahora. En vista de la omnipresencia de pruebas y evaluaciones, "la alfabetización en evaluación" parece ser una necesidad perentoria.