Medir en educación

Recursos de evaluación del Consejo Australiano para la Investigación Educativa 1

Sineace

SERIE: DOCUMENTOS TÉCNICOS

MEDIR EN EDUCACIÓN
RECURSOS DE EVALUACIÓN DEL CONSEJO AUSTRALIANO PARA LA INVESTIGACIÓN EDUCATIVA 1

Sineace

Serie Documentos Técnicos, 19

Cuidado de la edición
Dirección de Evaluación y Gestión del Conocimiento
Verónica Alvarado Bonhote, Directora
Diana Zapata Pratto, Especialista en Gestión de Publicaciones

Traducción
Cecilia Torres Llosa

Maquetación
Ángel García Tapia

Se terminó de imprimir en noviembre de 2017 en:
EDITORIAL SUPER GRÁFICA E.I.R.L.
Calle Luisa Beausejour 2049
Cercado de Lima

Hecho el Depósito Legal en la Biblioteca Nacional del Perú N.° 2017-14515
ISBN N.° 978-612-4322-29-7

Tiraje: 500 ejemplares
Primera edición
Lima, noviembre de 2017

Calle Manuel Miota N.° 235 - San Antonio, Miraflores, Lima 18, Perú
Teléfonos: (+51 1) 637-1122, (51-1) 221-4826, (51-1) 221-4807 anexo 108
E-mail: sir@sineace.gob.pe / www.sineace.gob.pe

Título original: *Educational Measurement. Assessment Resource Kit*
Autor: Geoff N. Masters

ÍNDICE

PRESENTACIÓN

Entre los años 2009 y 2015, el SINEACE estuvo abocado a la elaboración de estándares de aprendizaje, con la finalidad de contribuir con el Ministerio de Educación y diversos actores a lograr mejores aprendizajes en los estudiantes de la educación básica regular. Durante dicho proceso, se conoció la experiencia australiana de evaluación de los aprendizajes realizada por el Consejo Australiano para la Investigación Educativa (ACER). Su trabajo inspiró la elaboración de los estándares de aprendizaje en forma de mapas de progreso, lo que significó un cambio de paradigma en la manera de enseñar, toda vez que no indican tanto qué debe enseñarse, sino qué debe aprender un estudiante.

La pertinencia para el Perú de la experiencia de ACER , motivó que el SINEACE considerara la conveniencia de poner al alcance de los docentes los folletos que comprendían el Kit de Recursos de Evaluación, que había sido un valioso aporte para la elaboración de estándares de aprendizaje. Para ello suscribimos un convenio con ACER, que permite poner al alcance de los lectores de habla castellana el primer número de la serie, que rescata la importancia de la evaluación en el contexto educativo y reflexiona acerca de la necesidad de objetividad en la medición del desempeño educativo.

Este número busca alcanzar una definición de la medición y aborda el tema de las variables educativas; que permitirán tomar las decisiones más apropiadas

para aplicar en cada estudiante según su desarrollo. Asimismo propone un método para construir medidas educativas que sean unidimensionales, objetivas y con un nivel de intervalo.

La evaluación es un recurso valioso en la educación, pues proporciona información acerca del proceso de enseñanza-aprendizaje, la cual debe ser valorada para la toma de decisiones de quienes intervienen en este. Permite a los profesores, además, expresar un juicio de valor sobre el desempeño de los alumnos, ya sea en general o sobre alguna faceta particular de estos.[1]

El ser humano siempre ha tenido la necesidad de medir. Para ello, ha utilizado conceptos como tamaño, peso, volumen o densidad, conceptos que luego fueron afinándose, lo que permitió separar una variable determinada, independientemente de las demás. Trasladar esta necesidad de medir a un contexto educativo resulta un paso obvio, pues este conocimiento es prioritario tanto para los estudiantes, como para sus padres y la comunidad en general. Para ello, es imprescindible usar medidas del progreso en el desempeño.

Las mediciones de desempeño educativo hacen falta para investigar formas de mejorar el aprendizaje de los estudiantes, y para efectuar el seguimiento al rendimiento de los estudiantes a través de los años, lo que es de suma utilidad, si se requiere diseñar nuevos programas o políticas en este sector.

Con el propósito de lograr objetividad, se ha buscado uniformizar las unidades de medida, lo que hizo necesario poner a disposición instrumentos de medición calibrados en una misma unidad. Con este mismo fin, la medición debe realizarse bajo condiciones controladas, y es necesario intentar que los intervalos entre conocimiento y conocimiento sean regulares y medibles. Esto cobra relevancia puesto que, mientras más objetiva sea la medición, mayor posibilidad habrá para hacer comparaciones entre aulas, secciones o grados.

---

[1] Gimeno Sacristán, José (1998). La evaluación en la enseñanza. En *Comprender y transformar la enseñanza* (pp. 334-397). Madrid: Editorial Morata.

Decidir la forma que tomará el reporte y el método de evaluación elegido también reviste importancia. Se presentan diferentes métodos y se muestran formas variadas de equivalencia con los modelos propuestos, pues se trata de estimar las dificultades de las tareas (proceso de *calibración*). Al mismo tiempo, es preciso que las variables sean estables, lo que hace recomendable preparar un banco de ítems para desarrollar y calibrar según se plantea en el mapa de progreso.

El SINEACE difunde este libro que, aunque basado en otra realidad, muestra que la evaluación educativa es universal y pertinente para diversas realidades y culturas. Se espera que la presente publicación contribuya a promover entre los actores educativos el debate acerca de la evaluación en tanto recurso valioso en el proceso de enseñanza-aprendizaje, así como para la elaboración de políticas que contribuyan al desarrollo del desempeño educativo.

Consejo Directivo Ad Hoc
SINEACE

GLOSARIO

Análisis de coincidencias

El análisis estadístico de qué tan bien se responde a una asignación (o a una persona) encajando las expectativas de un modelo de medida.

Banco de tareas

Una colección de los mejores productos calibrados en la misma variable de medida.

Calibración

El proceso de estimar las dificultades de las tareas evaluadas desde las respuestas de los estudiantes a ellas.

Configuración estándar

El proceso de fijar un desempeño estándar.

Crédito parcial

El puntaje de las respuestas en las asignaciones de los individuos en varias categorías ordenadas.

Criterio de referencia

El proceso de interpretar el desempeño evaluado de los individuos en términos de objetivos de aprendizaje especificados (algunas veces utilizado para describir la interpretación del desempeño de la evaluación en términos de un desempeño estándar especificado).

Dificultad

La ubicación en una variable de medición (un parámetro para ser estimado).

Equivalencia

Un proceso estadístico que convierte puntajes en diferentes evaluaciones para la misma escala, para compararlos directamente.

Errores de medida

Una indicación de lo desconocido asociado con el estimado de una de las habilidades del estudiante.

Escala de competencia descrita

Una escala/variable de medición descrita en términos de conocimiento, habilidades, comprensión, aptitudes o valores observados normalmente en varias ubicaciones de la escala.

Estándares de desempeño

Niveles de habilidad fijados como metas o requerimientos para propósitos particulares (normalmente operados como "puntuación límite").

Evaluación adaptada a computadora

Un proceso por el cual se seleccionan las tareas para ser administradas una a la vez desde un banco de tareas con base en el desempeño del estudiante en las tareas precedentes.

Funcionamiento diferencial

La observación de que una asignación es atípicamente más fácil o más difícil para un grupo de estudiantes que para el otro (por ejemplo, inusualmente difícil para las mujeres).

Habilidad

Término genérico para la ubicación de un individuo en una medición variable (es decir, un parámetro a ser estimado).

Inclinación del producto

La observación de que una asignación es atípicamente más fácil o más difícil para un grupo de estudiantes que para otro (por ejemplo, inusualmente difícil para las mujeres).

Logit

Una unidad de medida.

Mapa de progreso

(ver lo descrito en Escala de competencia).

Medida

El proceso de estimar la ubicación de los estudiantes (habilidades) en una variable medida desde sus respuestas a un conjunto de tareas.

Modelo Rasch

Un modelo de medida capaz de brindar mediciones objetivas en una unidad definida de medida.

Norma referencial

El proceso de interpretar el desempeño de un estudiante en términos de desarrollo de un grupo relevante (por ejemplo, estudiantes de la misma edad).

Objetividad

Una característica del sistema de medida que permite la medición comparada sin fijarse en las particularidades de las tareas utilizadas o el puntaje.

Producto

Una pregunta de prueba o tarea.

Puntaje dicotómico

Puntaje de las respuestas de los individuos solo en dos categorías (normalmente correcto/equivocado).

Puntuaciones

Una característica del sistema de medida que permite la medición comparada sin fijarse en las particularidades de las tareas utilizadas o el puntaje.

Objetividad

Calificación del desempeño del estudiante o respuestas realizadas en términos de un conjunto de categorías ordenadas (es decir, una escala de puntuación).

Resultados

Los resultados de aprendizaje (por ejemplo, conocimiento, habilidades, comprensión, actitudes, valores).

Unidad de medida

Una cantidad constante de una variable continua que puede ser repetida y contabilizada.

Unidimensionalidad

Un estado idealizado en el cual las respuestas de los estudiantes a un conjunto de tareas son gobernadas solo por la variable en la que dichas tareas son medidas.

Unión de productos

Asignaciones compartidas por dos o más pruebas, lo que permite que aquellas pruebas puedan ser equivalentes.

Variable

Algo que varía; en este contexto, una habilidad (actitud, etc.) que puede ser conceptualizada como una variable junto a un continuo.

MEDIR EN EDUCACIÓN
RECURSOS DE EVALUACIÓN DEL CONSEJO AUSTRALIANO PARA LA INVESTIGACIÓN EDUCATIVA 1

1. ¿QUÉ ES LA MEDICIÓN?

1.1. Conceptualización de las variables

En la vida, las ideas más poderosas son las ideas simples. Muchas áreas de emprendimiento, incluso la ciencia y la religión, implican la búsqueda de ideas unificadoras que ofrecen las explicaciones más sencillas para una amplia variedad de experiencias humanas.

En la historia de la humanidad, desde el principio nos vimos rodeados de objetos terriblemente complejos. Para dar sentido al mundo, nos resultó útil —y probablemente necesario— ignorar esta complejidad e inventar formas sencillas de pensar en los objetos a nuestro alrededor y de describirlos. Una estrategia útil consistía en enfocarnos en las formas en que los objetos se diferenciaban unos de otros.

Los conceptos *grande y pequeño* eran una distinción particularmente útil. El tamaño era una idea que nos ayudaba a ignorar la miríada de otras formas en que los objetos se diferenciaban —incluyendo el color, la forma y la textura— y a enfocarnos en una sola característica: su *tamaño*. La noción abstracta de tamaño era una idea poderosa, porque se podía usar para describir objetos tan distintos como los ríos, los animales, las rocas y los árboles.

A lo largo de muchos años de nuestra historia, el concepto de *tamaño* nos resultó, sin duda, de mucha ayuda. Pero a medida que hicimos observaciones más detalladas de los objetos y reflexionamos sobre la base de ellas, encontramos que era útil distinguir el tamaño del peso, incluso si ambos están estrechamente relacionados. Y a medida que lidiábamos con la experiencia de que los objetos más grandes no eran siempre más pesados, introdujimos conceptos más sofisticados, como densidad y gravedad específica.

Cada una de estas ideas nos ofrecía una forma de enfocarnos en una característica que diferenciaba a los objetos en un momento dado y consistía

en una herramienta para lidiar con una complejidad del mundo que de otra forma hubiese resultado inmanejable. El tamaño, el peso, la longitud, el volumen y la densidad fueron solo algunas de nuestras ideas para describir la forma en que los objetos varían; otras variables incluían la dureza, la temperatura, la inercia, la velocidad, la aceleración, la maleabilidad y el momento. A medida que nuestra comprensión mejoró y nuestras observaciones se tornaron más sofisticadas, encontramos útil inventar nuevas variables que se distinguían muy sutilmente de las que ya existían: por ejemplo, distinguir la masa del peso, la velocidad de la rapidez, la temperatura del calor.

Dimensión

[*di(s)* – separado *metiri* – medir]:

separado para la medición.

La ventaja de una variable residía en que nos permitía poner a un lado —al menos temporalmente— las formas muy complejas en que los objetos se diferenciaban y observar los objetos a través de un solo lente a la vez. Por ejemplo, los objetos se podían colocar en un solo orden de peso creciente, independientemente de las distintas formas, colores, áreas de superficie, volúmenes y temperaturas. El "lente" del peso nos permitía ver objetos desde solo una de infinitas dimensiones posibles.

A veces nos preguntábamos si habíamos inventado estas variables o si las habíamos descubierto nada más. ¿El concepto de

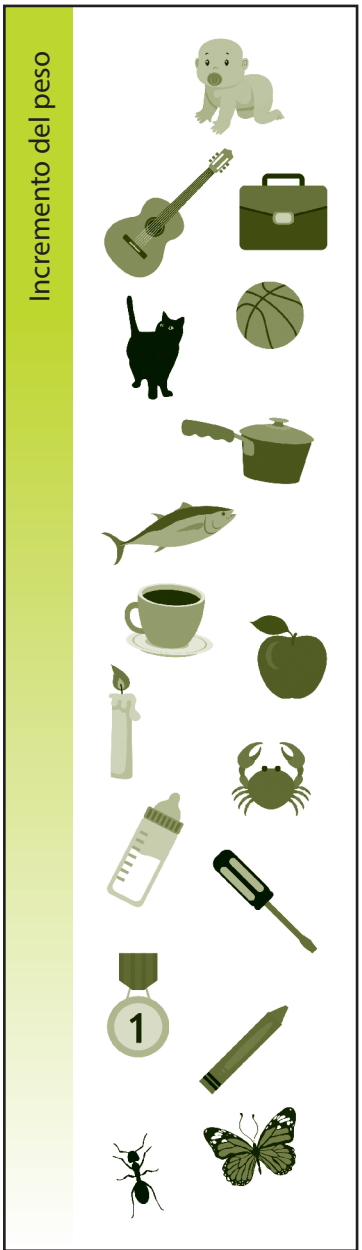
momento era una invención humana o es que "descubrimos" el momento? Ciertamente, fue una decisión humana el enfocar la atención en aspectos específicos de la variabilidad del mundo y el trabajar para aclarar y operacionalizar variables. El trabajo esmerado y relativamente reciente de Anders Celsius (1701-1744) y de Gabriel Fahrenheit para desarrollar una definición útil de temperatura da cuenta de esto. Por otro lado, las variables que desarrollamos pretendían representar diferencias *reales* entre los objetos. Finalmente, la pregunta acerca de si las variables fueron descubiertas o inventadas tenía un interés filosófico limitado: la pregunta importante sobre una variable era si resultaba útil en la práctica.

1.2. Variabilidad humana

Pero no solo los objetos inanimados mostraban una tremenda complejidad, también lo hacían las personas. Nuevamente, una estrategia para lidiar con esta complejidad era enfocarse en las formas particulares en que las personas variaban. Algunos humanos corrían más rápido que otros, algunos tenían más fuerza, otros eran mejores cazadores, bailarines más gráciles, guerreros superiores, artesanos más hábiles, maestros más sabios, consejeros más compasivos, comediantes más graciosos, mejores oradores. La lista de dimensiones

"Los conceptos físicos son las creaciones libres de la mente humana y no están —a pesar de que lo parezcan— determinados únicamente por el mundo externo". separado para la medición.

Albert Einstein. *La evolución de la Física, 1938.*



Se puede visualizar la variable peso como un continuo aumento de la pesadez.

en las cuales se podía comparar a los humanos era interminable y el lenguaje que desarrollamos para describir esta variabilidad era vasto e impresionante.

Al lidiar con la complejidad humana, nuestra decisión de enfocar un aspecto de la variabilidad a la vez fue, al menos, tan importante como lidiar con la complejidad de los objetos inanimados. Con el fin de seleccionar a la persona idónea para liderar una cacería, era deseable enfocarse en sus habilidades individuales y reconocer que el mejor cazador no era necesariamente el que mejor bailaba alrededor de la fogata o el que mejor contaba una historia al grupo. Hubo momentos en los que nada menos que nuestra existencia dependía de la claridad que tuviésemos de las fortalezas y debilidades relativas de nuestros compañeros humanos.

La decisión de prestar atención a un aspecto de esta variabilidad también resultó importante cuando se trataba de monitorear el desarrollo de destrezas, comprensiones, actitudes y valores en los jóvenes. Como adultos, buscábamos desarrollar diferentes tipos de habilidades en los niños, incluyendo habilidades para la caza, la danza, la lectura, la escritura, la narración de historias, la creación y el uso de armas y herramientas, la construcción de viviendas y la preparación de comidas. También buscábamos desarrollar el

conocimiento de los niños con respecto a la geografía local, la flora y la fauna, y su comprensión de las costumbres y los rituales tribales, las ceremonias religiosas y la historia oral. Para monitorear el progreso de los niños hasta que fueran adultos maduros, sabios y bien formados, solíamos encontrar conveniente enfocarnos en un aspecto de su desarrollo a la vez.

A veces nos preguntábamos si las variables que usábamos para lidiar con la complejidad del comportamiento humano eran reales, tal como la temperatura y el peso lo son. ¿Realmente diferían los niños en su capacidad para leer? ¿Las diferencias en estas habilidades eran "reales" en el mismo sentido que las diferencias entre objetos con respecto a la energía potencial o el momento?

Nuevamente, la pregunta importante era si una variable como la capacidad de lectura resultaba útil en la práctica. La experiencia común sugería que los niños sí se diferencian en sus habilidades para la lectura y que las capacidades de lectura de los individuos se desarrollan a lo largo del tiempo. ¿Pero la idea de la variable de una competencia lectora creciente lograba sustentarse en observaciones minuciosas del comportamiento de lectura? ¿Esta idea resultaba útil para comprender y promover el desarrollo de la lectura? Como con todas las variables, la pregunta más importante con respecto a las dimensiones de la variabilidad humana era si ayudaban o no a lidiar con las complejidades de la experiencia humana.

En resumen, la decisión de enfocar la atención en un aspecto de la variabilidad a la vez fue un progreso significativo en el manejo de la complejidad. La conceptualización de variables fue nuestro primer paso hacia la medición.

1.3. Invención de unidades

El segundo paso hacia la medición fue la invención de unidades que representaran montos iguales de la variable que se estaba midiendo. Se logró un progreso importante en el conteo de unidades en la relación con la más intangible de las variables: el tiempo.

Medidas bíblicas

Y Dios le dijo a Noé: "… Así es como lo harás: la longitud del arca 300 codos, su ancho 50 codos y su alto 30 codos".

Génesis 6:15

Efrón le respondió a Abraham: "Mi Señor, escúcheme; un pedazo de tierra que vale 400 siclos[1] de plata, ¿cuánto es eso?"… Y Abraham pesó la plata que Efrón había mencionado… 400 siclos de plata, de acuerdo con las pesas de uso común entre los comerciantes.

Génesis 23:15

El tiempo, a diferencia de otras variables como la longitud y el peso, no podía ser manipulado y era mucho más difícil de conceptualizar. Pero, increíblemente, el hombre se encontró viviendo dentro de un reloj gigante. Al inspeccionar cuidadosamente el tictac rítmico del mecanismo del reloj, el hombre aprendió a medir el tiempo contando unidades de este.

La rotación regular de la Tierra sobre su propio eje marcaba cantidades iguales de tiempo y proveyó a los humanos una unidad básica de medida: el día. Al contar los días, estábamos en capacidad de reemplazar descripciones cualitativas de tiempo ("hace mucho tiempo") por descripciones cuantitativas ("hace cinco días"). Este era el segundo requerimiento para la medición: una unidad de medida, es decir, una cantidad fija de una variable que podría repetirse sin modificarse y, por ende, podía contarse. La invención de las unidades permitió responder a la pregunta *¿cuánto*? a partir de contar *cuántas* unidades.

La rotación regular de la Luna alrededor de la Tierra proveyó una unidad de tiempo mayor, la "luna" o el mes lunar. Y la rotación regular de la Tierra alrededor del Sol resultó en las estaciones y en una unidad todavía mayor: el año. El movimiento de estos cuerpos celestes

[1] Antigua unidad monetaria usada en el Oriente Próximo y en Mesopotamia.

nos aportó un instrumento para marcar cantidades iguales de tiempo y nos enseñó que las unidades se podían combinar para formar unidades mayores o subdividir para formar unidades aun más pequeñas (horas, minutos, segundos).

Las civilizaciones antiguas crearon formas de tabular sus mediciones del tiempo en calendarios grabados en piedra y usaron las sombras que se movían para inventar unidades más pequeñas que el día. Mediante la observación del movimiento rítmico del reloj gigante en que vivíamos, los humanos alcanzamos una sofisticación probablemente mayor en la medición del tiempo antes de desarrollar una sofisticación similar en la medición del variables más tangibles, como la longitud, el peso y la temperatura.

> Y José guardó granos en gran abundancia, como la arena del mar, hasta que dejó de medirla, porque no podía medirse.
>
> Génesis 41:49

La invención de unidades de medida también jugó un papel crucial en la comunicación exacta sobre las distancias. En la historia temprana del hombre, "un camino largo" se convertía en "una caminata de dos días", lo que permitía nuevamente responder a la pregunta *¿cuánto?* mediante el conteo de unidades. Para distancias más pequeñas, se contaban los pasos. Mil pasos eran una milla. Se definieron otras variables a partir de las partes del cuerpo: un pie, un cúbito (la longitud del antebrazo), una mano, o en términos de objetos que se

podían cargar y colocar de un extremo al otro: la cadena, el eslabón (1/100 de una cadena), la vara, la percha o el mástil, la yarda (un palo).

El uso reciente y continuo que hacemos de muchas de estas unidades nos recuerda que el dominio de la medición de longitudes es reciente. Lo mismo aplica para las unidades que usamos para medir otras variables (por ejemplo, piedras para medir el peso). Y otras unidades se inventaron tan recientemente que conocemos el nombre de sus inventores (por ejemplo, Celsius y Fahrenheit).

1.4. En busca de la objetividad

La invención de unidades como pasos, pies, palmos, cúbitos, cadenas, piedras, varas y mástiles, que podían repetirse sin modificarse, puso a disposición de los seres humanos instrumentos para la medición. Sin embargo, una pregunta importante al hacer mediciones consistía en si diferentes instrumentos daban medidas numéricamente equivalentes para el mismo objeto.

Manteniendo los estándares

Dado que Edward Masters, Thomas Draper, Henry Chesheire y Margaret Ball están ante este tribunal por tener y usar *strikes** ilegales, entonces la corte ordena en este día que John Stratford, *esquire***, y Thomas Corby, *esquire*, deben ser y se desea que sean quienes examinen si las personas mencionadas han hecho que sus *strikes* sean iguales al estándar provisto por el señor del feudo de Atherston y certifiquen sus procedimientos frente a esta corta en las siguientes Sesiones Generales de la Paz.

Warwickshire, Epiphany, 1673

**strike*: instrumento para la medición de granos.

***esquire*: título nobiliario.

Si dos instrumentos no proveían de medidas numéricamente equivalentes, entonces era posible que no estuvieran calibrados en la misma unidad. Una cosa era ponerse de acuerdo en el uso de un pie para medir longitudes, ¿pero de quién debía ser el pie? ¿Qué tal si mi piedra era más pesada que la tuya? ¿Y si tu cadena era más larga que la mía? Un requisito fundamental para tener mediciones útiles consistía en que las medidas resultantes fueran independientes del instrumento de medición y de la persona que lo usara: en otras palabras, debían ser objetivas.

Para lograr este tipo de objetividad era necesario establecer y compartir unidades de medida comunes o estándar. Por ejemplo, en 1790 se acordó medir la longitud con la unidad del metro, definida como una millonésima parte de la distancia desde el Polo Norte hasta la línea ecuatorial. Después del Tratado del Metro de 1875, un metro se redefinió como la longitud de una barra de platino e iridio que se encuentra en la Oficina Internacional de Pesos y Medidas cerca de París, y en 1983 un metro se definió como la distancia que viaja la luz en el vacío durante 1/299.792.458 de segundo. Todas las barras marcadas en metros y centímetros se calibraron teniendo en cuenta esta unidad estándar. Se fundaron oficinas de pesos y medidas para garantizar que los estándares se mantuvieran y que los instrumentos se calibraran con exactitud con respecto a las unidades estándar. Así, las medidas se podían comparar directamente de un instrumento a otro, requisito esencial si se quiere tener una comunicación precisa y si se trata de dirigir con éxito el comercio, la ciencia y la industria.

Si dos instrumentos no daban medidas numéricamente equivalentes, entonces una posibilidad más seria era que no estuviesen midiendo la misma variable. El indicio más sencillo de que se trataba de este problema se daba cuando dos instrumentos producían ordenamientos significativamente distintos para un grupo de objetos.

Por ejemplo, dos varas para medir —una calibrada en centímetros, la otra en pulgadas— daban medidas numéricas distintas para un mismo objeto. Pero cuando un número de objetos se medía tanto en pulgadas como en

centímetros, y luego se "ploteaban" los resultados usando centímetros y pulgadas, los puntos resultantes asemejaban una línea recta (de hecho, si no hay error de medición, se formaría una línea recta perfecta). En otras palabras, las dos varas proveían medidas de longitud consistentes.

Sin embargo, si al usar un instrumento el objeto A resultaba significativamente mayor que el objeto B, pero al usar un segundo instrumento el objeto B era significativamente mayor que el objeto A, esto sería muestra de una inconsistencia básica. ¿Qué podríamos concluir sobre la posición relativa de los objetos A y B con respecto a nuestra variable?

Un requisito fundamental en medición era que no debe importar qué instrumento se use o quién esté midiendo (es decir, el requisito de objetividad/imparcialidad). Solo si instrumentos diferentes proveían medidas consistentes era posible lograr este tipo de objetividad en nuestras medidas.

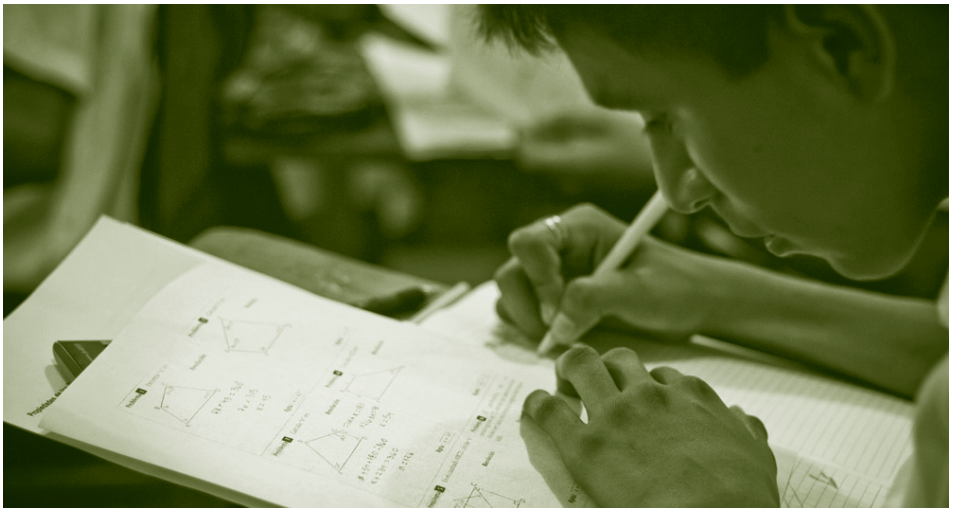
En resumen

La medición es una de las invenciones más poderosas y significativas de la humanidad.

La medición empieza con la decisión de prestar atención solo a una forma en la que los objetos o las personas difieren. Esta decisión de enfocarse únicamente en un aspecto de la variabilidad permite que estos objetos se conceptualicen de forma tal que tengan un único orden con respecto a una variable (dimensión). La conceptualización de una variable como un continuo de cantidades en aumento es el primer paso hacia la medición.

El segundo paso hacia la medición es la invención de una unidad. Una unidad es la cantidad de una variable que puede repetirse y contarse sin modificarse. Su uso asegura que diferencias numéricas iguales representen las mismas cantidades de esta variable.

El tercer y último paso es asegurar la objetividad de la medición. Las medidas son objetivas si no dependen del conocimiento del instrumento que se usa para obtenerlas o de la persona involucrada en el proceso de medición. El testeo de la objetividad consiste en comprobar si se obtienen medidas numéricas equivalentes usando diferentes instrumentos o si diferentes personas hacen la medición.



2. LA ASPIRACIÓN DE MEDIR VARIABLES EDUCATIVAS

En contextos educativos, es común separar y prestar atención al desarrollo del estudiante en un solo un aspecto a la vez.

Cuando un maestro busca determinar qué etapa ha alcanzado un estudiante en su aprendizaje para monitorear su progreso a lo largo del tiempo o para tomar decisiones sobre las experiencias de aprendizaje más apropiadas para cada individuo, entonces estas preguntas suelen abordarse en relación con un dominio de aprendizaje a la vez. Por ejemplo, es común evaluar el logro de un niño en razonamiento numérico por separado de muchas otras dimensiones en las que el niño debe estar avanzando (como la lectura, la escritura y el lenguaje oral), incluso si esos aspectos del desarrollo están relacionados.

La mayoría de variables educativas puede conceptualizarse como aspectos del aprendizaje en los que los estudiantes progresan a lo largo de los años. La lectura es un ejemplo. Esta empieza en la infancia temprana, pero continúa desarrollándose a lo largo de la primaria, dado que los niños desarrollan habilidades para extraer significados cada vez más sutiles de textos cada vez más complejos. Y para la mayoría de los niños, el desarrollo de la lectura no se detiene ahí: continúa hasta la secundaria.

Los maestros y los gestores educativos usan medidas del progreso de los estudiantes con diversos propósitos.

Se recurre a medidas de variables educativas cuando existe el deseo de asegurar que las plazas limitadas en programas educativos se ofrezcan a quienes más lo merezcan y estén en mejores condiciones de aprovecharlas. Por ejemplo, las plazas en las escuelas de medicina son limitadas debido al costo de proveer los programas de medicina y a la necesidad limitada de profesionales médicos en la comunidad. Las escuelas de medicina buscan asegurar que las plazas se ofrezcan a los estudiantes sobre la base de su probable éxito en la escuela y, de ser posible, en la medida en que los postulantes parezcan adecuados

para la subsiguiente práctica médica. Para asignar las plazas de forma justa, las escuelas de medicina se toman la molestia de identificar y medir atributos relevantes de los postulantes. Las universidades y escuelas que ofrecen becas sobre la base del mérito académico pasan por problemas similares para identificar y medir a los candidatos en dimensiones de logro apropiadas.

Las medidas de la competencia y el éxito escolar se determinan al culminar programas de educación o entrenamiento. ¿El estudiante ha alcanzado un nivel suficiente de comprensión y conocimiento al terminar el curso como para satisfacer los objetivos de tal curso? ¿El estudiante ha alcanzado suficientes niveles de competencia como para permitírsele ejercer (por ejemplo, como contador, abogado, pediatra, piloto)? Las decisiones de este tipo se toman identificando primero las áreas de conocimiento, habilidad y comprensión en las cuales se debe demostrar un mínimo nivel de competencia y midiendo los niveles de competencia o logro de los candidatos en dichas áreas.

También se requieren mediciones del desempeño educativo para investigar formas de mejorar el aprendizaje de los estudiantes: por ejemplo, para evaluar el impacto de iniciativas educativas específicas, para comparar la efectividad de diferentes maneras de estructurar y administrar la distribución educativa, y para identificar las estrategias de enseñanza más efectivas y las formas más económicas de elevar el rendimiento de las secciones de la población estudiantil con menor desempeño. La mayoría de la investigación educativa, incluyendo la evaluación de programas educativos, depende de medidas confiables de aspectos del aprendizaje de los estudiantes. Parte de la investigación más informativa consiste en hacer seguimiento al progreso de los estudiantes en una o más variables a lo largo de los años (es decir, estudios longitudinales del progreso).

La intención de separar y medir variables en educación se hace explícita en la construcción y uso de test educativos. La intención de obtener una única puntuación para cada estudiante, de modo que todos se puedan ordenar en una serie según esa puntuación, refleja la intención de medirlos mediante una

única variable y se conoce como la intención de unidimensionalidad. En un test de este tipo se pretende que una puntuación más alta represente una cantidad mayor de la variable que se mide y que las puntuaciones más bajas representen una cantidad menor. El uso de un test educativo para proveer un único orden de los estudiantes a lo largo de una variable educativa es idéntico al principio de la intención de ordenar objetos a lo largo de una única variable de mayor peso (ver página 18).

Se conoce como intención de unidimensionalidad a la intención de obtener una única puntuación para cada estudiante de modo que todos se puedan ubicar en un único orden de puntuaciones.

A veces, los test se construyen con la intención de proveer no una puntuación, sino varias puntuaciones. Por ejemplo, un test de razonamiento podría construirse con la intención de obtener para cada estudiante tanto una puntuación en razonamiento verbal como una puntuación en razonamiento numérico. Los test de este tipo se conocen como *test compuestos*. El conjunto de ítems de razonamiento verbal constituye un instrumento de medición; el conjunto de ítems de razonamiento numérico constituye otro. El hecho de que ambos se administren en la misma situación de evaluación responde simplemente a una decisión práctica.

No todos los conjuntos de preguntas se construyen con la intención de que estas conformen un instrumento de medición. Por ejemplo, algunos cuestionarios se elaboran

con la intención de reportar las respuestas a cada pregunta por separado, pero sin combinar las respuestas a las preguntas (por ejemplo, *¿cuántas horas en promedio al día pasas viendo televisión?, ¿qué tipo de libros o revistas prefieres leer?*). Se hacen preguntas de este tipo no porque exista la intención de proveer evidencia sobre una misma variable subyacente, sino porque existe un interés en conocer cómo responde una población de estudiantes a cada pregunta por separado. La mejor forma de comprobar si un conjunto de preguntas conforma un instrumento de medición es definir si el autor tiene la intención de combinar las respuestas para obtener una única puntuación para cada estudiante.

El desarrollo de todo instrumento de medición comienza con el concepto de variable. La tabla a continuación muestra algunos de los cientos de variables listadas en el *Mental Measurements Yearbook* (Anuario de Medidas Mentales), para las cuales se construyeron instrumentos de medición (test y cuestionarios).

Actitud hacia las matemáticas	Escritura narrativa
Adaptación a la escuela	Expresión artística
Adición de fracciones	Expresión escrita
Agresión	Expresividad
Altruismo	Extroversión
Ansiedad	Fuerza del yo
Asertividad	Función motriz fina
Atención	Función motriz gruesa
Autoconfianza	Habilidad lectora
Autocuidado personal	Habilidad para la aritmética
Caligrafía	Habilidad para la resolución de problemas
Cautela	Hiperactividad
Cognición de relaciones semánticas	Identificación de relaciones
Colaborar	Impulsividad
Competencia interpersonal	Lectura oral
Competitividad	Liderazgo potencial
Comportamiento asocial	Madurez escolar
Comprensión de diálogos	Madurez para el kindergarten
Comprensión de ventas	Memoria de oraciones
Comprensión del lenguaje	Orientación al logro
Comprensión matemática	Orientación hacia la meta

Comprensión oral	Pensamiento crítico
Comunicación de información	Pensamiento intuitivo
Conciencia	Percepción auditiva
Conocimiento del vocabulario	Percepción de objetos en el espacio
Conocimientos sobre nutrición	Precisión administrativa
Copiado de formas	Proeza física
Corrección	Programar en computación
Creatividad	Razonamiento
Decisión gerencial	Razonamiento abstracto
Deletreo	Razonamiento verbal
Depresión	Reconocimiento de letras
Destreza	Resiliencia emocional
Destreza manual	Satisfacción con la vida
Diferenciación táctil	Satisfacción corporal
Discriminación auditiva	Seguir direcciones
Discriminación visual	Sociabilidad
Disposición a la aventura	Tolerancia al estrés
División de decimales	Toma de decisiones
Empatía	Velocidad de tipeo

La intención que subyace a cada uno de estos instrumentos —y de muchos otros reportados en la literatura sobre medición educativa y psicológica— es reunir una serie de ítems capaces de proveer evidencia sobre la variable de interés y, luego, combinar las respuestas a estos para obtener medidas de la variable. Esta intención sugiere la pregunta si los ítems reunidos para medir una variable cualquiera funcionan de manera conjunta como para constituir un instrumento de medición útil.

2.1. ¿Intervalos iguales?

Cuando un estudiante responde un test, el resultado es una puntuación que pretende ser una medida de la variable que fue diseñado para medir. Estas puntuaciones proveen un único orden de personas que responden el test: desde aquel que obtuvo la puntuación más baja (la persona que responde la menor cantidad de ítems correctamente o que está de acuerdo con la menor cantidad de afirmaciones de un cuestionario) hasta el que obtuvo la puntuación

más alta. Dado que las puntuaciones ordenan a los estudiantes a lo largo de una variable, se les describe como si tuvieran propiedades "ordinales".

Es común asumir que las puntuaciones en un test también tienen propiedades de "intervalo", es decir, que las diferencias en las puntuaciones representan las mismas diferencias en la variable que se ha medido (por ejemplo, que la diferencia entre las puntuaciones 25 y 30 en un test de comprensión de lectura representa la misma diferencia en la habilidad para leer que la diferencia entre las puntuaciones 10 y 15). El intento de atribuir propiedades de intervalo a las puntuaciones es un intento por tratarlas como si fueran medidas similares a las de longitud en centímetros o de peso en kilogramos. Pero las puntuaciones no son conteos de una unidad de medida y, por lo tanto, no comparten las propiedades de intervalo de las medidas.

Las puntuaciones son conteos de los ítems respondidos correctamente y, por consiguiente, dependen de las particularidades de los ítems que se han contado. Una puntuación de 16 sobre 20 ítems fáciles no tiene el mismo significado que una puntuación de 16 sobre 20 ítems difíciles. En tal sentido, una puntuación es como el conteo de objetos. Un conteo de 16 papas no es una "medida" porque no se están contando unidades *iguales*. 16 papas pequeñas no representan la misma cantidad de papas que 16 papas grandes. Cuando compramos y vendemos papas, usamos y contamos una unidad (kilogramo o libra) que mantiene su significado para papas de distinto tamaño.

Una segunda razón por la cual las puntuaciones de test comunes no tienen las propiedades de las medidas es que están encuadradas por límites superiores e inferiores. No es posible tener una puntuación menor que cero o mayor que la máxima puntuación posible en ese test. El efecto de estos llamados efectos "piso" o "techo" es que diferencias iguales en las puntuaciones de los test no representan diferencias iguales en la variable medida. En un test de matemáticas de 30 ítems, una diferencia de un punto en los extremos del rango de puntuaciones (por ejemplo, la diferencia entre las puntuaciones 1 y 2 o entre

las puntuaciones 28 y 29) representa una diferencia mayor en el desempeño matemático que una diferencia de un punto cerca del centro del rango de puntuaciones (por ejemplo, la diferencia entre las puntuaciones 14 y 15).

A pesar de que las puntuaciones en un test no tienen propiedades de intervalo, suele ser común tratarlas (erróneamente) como si las tuvieran. Se asume que se trata de propiedades de intervalo cuando se usan puntuaciones numéricas en procedimientos estadísticos sencillos, como el cálculo de promedios o de desviaciones estándar, o en procedimientos estadísticos más sofisticados, como el análisis de regresión o de varianza. En estos procedimientos comunes, los usuarios de las puntuaciones las tratan como si tuvieran las propiedades de intervalo de las pulgadas, los kilogramos o las horas.

2.2. Objetividad

Todo constructor de test sabe que los ítems individuales no son importantes en sí mismos. Ningún ítem es indispensable: los ítems se construyen como oportunidades para recoger evidencia sobre la variable de interés y todo ítem podría reemplazarse por otro similar. Más importante que los ítems individuales de los test es la variable sobre la que estos pretenden proveer evidencia.

Un ítem particular desarrollado, por ejemplo, como parte de un test de cálculo no es significativo en sí mismo. En realidad, puede que los estudiantes no lo vuelvan a ver ni lo tengan que resolver nunca más. La pregunta importante a hacerse sobre un ítem de un test no es si resulta o no significativo en sí mismo, sino si constituye un vehículo útil para recoger evidencia sobre la variable que se quiere medir (en este caso, la habilidad para el cálculo).

Otra forma de decirlo es que, para sacar una conclusión sobre la habilidad de un estudiante en cálculo, no debería importar qué ítems en particular se le dan para que los resuelva. Cuando construimos un test, nuestra intención es que los resultados tengan una generalidad más allá de la especificidad de los ítems del test. Esta intención es idéntica a nuestra intención de que las

mediciones de la altura no dependan de los detalles del instrumento de medición (por ejemplo, si usamos una regla de metal, una de madera, una cinta métrica de construcción, una cinta métrica de costurera, etc.). Una intención fundamental de todas las medidas es que su significado debe relacionarse con alguna variable general, como estatura, temperatura, destreza motriz o empatía, y no deben estar condicionadas a la especificidad del instrumento utilizado para obtenerlas (¡solo imagina cuán inconvenientes serían las mediciones físicas si, cada vez que se reportaran, tuvieran que acompañarse de información sobre el instrumento que se utilizó para obtenerlas!).

Una intención fundamental de todas las medidas es que su significado debe estar relacionado con alguna variable general, y no deben estar condicionadas a la especificidad del instrumento utilizado para obtenerlas.

La intención de que las mediciones de variables educativas tengan un significado generalmente independiente del instrumento que se usó para obtenerlas es especialmente importante cuando existe la necesidad de comparar los resultados de test diferentes. Un maestro o una escuela que desea aplicar un test antes de empezar un curso (un *pretest*) y luego hacerlo una vez terminado el curso (un *postest*), para estimar su impacto, no querrá usar el mismo en ambas ocasiones. Una escuela de medicina que usa exámenes de admisión para seleccionar a los postulantes quizás desee comparar los resultados obtenidos con diferentes tipos de exámenes de admisión en distintas situaciones de evaluación. O un

sistema escolar que desea monitorear sus estándares a lo largo del tiempo, o el crecimiento durante la escolaridad, va a querer comparar los resultados de test usados en diferentes años o de test de diferente dificultad, diseñados para grados distintos (por ejemplo, los test de comprensión de lectura en 3.er, 4.to y 5.to grados).

En educación existen muchas situaciones en las que buscamos mediciones libres de la especificidad del instrumento utilizado para obtenerlas y que, por lo tanto, sean comparables de un instrumento al otro.

Cuando se miden variables educativas, también existe la intención de que las mediciones resultantes no dependan de las personas que midieron. Esta consideración es especialmente importante cuando las mediciones se basan en juicios sobre el trabajo de los estudiantes o en su desempeño. Para asegurar la objetividad de las mediciones que se basan en estos juicios, se suele proveer a los jueces de guías claras y de entrenamiento, así como darles ejemplos para ilustrar las puntuaciones (por ejemplo, muestras de los textos de los estudiantes o videos de la performance en un baile), usar varios jueces, establecer procedimientos para identificar y lidiar con las discrepancias, o hacer ajustes estadísticos para manejar diferencias sistemáticas en el rigor o la lenidad de los jueces.

A pesar de que, claramente, existe la intención de que las mediciones educativas tengan un significado libre de la especificidad de los test particulares, sus puntuaciones ordinarias (por ejemplo, el número de ítems resuelto correctamente) están completamente condicionadas a estos. Una puntuación de 29 en un test no tiene el mismo significado que una medición de 29 centímetros o 29 kilogramos. Para que la puntuación 29 tenga algún sentido, es necesario conocer el número total de ítems del test: ¿29 de 30 ítems? ¿29 de 40? ¿29 de 100? Incluso, saber que un estudiante obtuvo 29 de 40 no es muy útil. Tener éxito en 29 ítems fáciles no representa la misma habilidad que tener éxito en 29 ítems difíciles. Para poder comprender realmente el significado de una puntuación de 29 de 40 sería necesario considerar cada uno de los 40 ítems.

Un dilema antiguo en la evaluación educativa es que, mientras que los ítems en particular nunca son de interés en sí mismos sino que sirven únicamente como indicadores de una variable que sí interesa, el significado de una puntuación siempre está limitado a un conjunto de ítems en particular. Así como tenemos la intención de que la medición de la habilidad para escribir de un estudiante sea independiente de los jueces que la evalúan, también buscamos mediciones de variables —como el razonamiento numérico— que sean neutrales con respecto a los ítems particulares que se incluyeron en un test y que los trasciendan. La teoría moderna de la medición (descrita en la siguiente sección) resuelve justamente este dilema.

En resumen

En la educación, buscamos mediciones de una amplia gama de variables. Las mediciones confiables de variables educativas son esenciales para evaluar de forma exitosa la efectividad de los programas educativos, monitorear los estándares educativos a lo largo del tiempo, comparar los niveles de logro en distintos sistemas educativos, investigar las relaciones e influencias sobre el logro educativo, colocar becas y plazas en los cursos educativos, medir el crecimiento individual a lo largo del tiempo y tomar decisiones sobre el nivel que un individuo ha alcanzado en su aprendizaje. La medición educativa siempre empieza con la intención de estimar las posiciones de los estudiantes con respecto a una variable de interés.

En la educación asumimos que las puntuaciones en los test y cuestionarios tienen propiedades de un nivel de intervalo, cuando calculamos estadísticas simples como los promedios y las desviaciones estándar, y también cuando usamos procedimientos más sofisticados como los análisis de regresión. Sin embargo, las puntuaciones de test comunes, dado que no corresponden a conteos de una unidad en cantidades fijas, no tienen propiedades de intervalo.

Diferencias numéricas iguales no representan, en términos generales, diferencias iguales en la variable de interés.

En educación también tenemos la intención de que nuestras mediciones tengan una generalidad que vaya más allá de la especificidad de un conjunto de ítems y de las personas involucradas en el proceso de medición.

Los ítems de un test no son importantes en sí mismos: simplemente son oportunidades convenientes e intercambiables de recoger evidencia sobre la variable para cuya medición se ha diseñado. Sin embargo, el significado de una puntuación numérica está condicionado a los conjuntos específicos de ítems. Cada test tiene un único conjunto de puntuaciones numéricas, equivalente a las varas de medición calibradas de acuerdo con sus propias unidades de longitud.

El siguiente artículo describe un modelo de medición que puede usarse para lo siguiente:

- Determinar el grado en que un conjunto de ítems funcionan juntos para proveer mediciones de una única variable.
- Definir una unidad de medición para la construcción de medidas de variables educativas que tengan un nivel de intervalo.
- Construir medidas numéricas que tengan un significado independiente del conjunto particular de ítems que se usó.

Cuando la medición no alcanza

Los test referidos a una norma

Las puntaciones de test ordinarias (conteo de ítems respondidos correctamente) no comparten las propiedades de medidas tales como las longitudes en centímetros o los pesos en kilogramos. Una puntación en un test, por ejemplo 28 de 40, no tiene un significado general como 28 cm, porque está condicionada a un conjunto particular de ítems.

En un intento por proveer de significados a puntuaciones de un test que vayan más allá de las especificidades de un instrumento particular, una práctica común consiste en referir las puntuaciones a una población determinada de estudiantes. Por ejemplo, si el 65% de los estudiantes escoceses de 4.to grado tuvieron una puntuación menor a 28 en un test particular, entonces la puntuación de 28 en ese test se volvería a expresar como el percentil 65 de estudiantes escoceses de 4.to grado. Si bien este método de interpretación de puntuaciones de un test es útil, no provee una escala de medición con propiedades de intervalo. Es equivalente a marcar una vara para medir las estaturas de los niños no con unidades constantes, como las pulgadas o los centímetros, sino con percentiles. El percentil 60 sería la marca que separa el 60% de los estudiantes más bajos del 40% de los estudiantes más altos; el percentil 90 separaría al 90% más bajo de los estudiantes del 10% más alto, y así sucesivamente. Por supuesto, el percentil 60 de los estudiantes de 4.to grado tendría una posición distinta en la vara para medir que el percentil 60 de los estudiantes de cualquier otro grado.

Siempre es posible localizar percentiles en una vara para medir que esté calibrada en pulgadas o centímetros, pero los percentiles no sustituyen una unidad bien definida y, por lo tanto, no proveen una base para la medición (contar unidades).

En la página 77 se encuentra una discusión más detallada sobre las interpretaciones de puntuaciones referidas a las normas.

Los test referidos a un criterio

Un segundo intento de abordar las limitaciones de las puntuaciones condicionadas a un test es referir los resultados a ámbitos muy bien definidos del aprendizaje, para intentar concluir de forma absoluta si un individuo domina o no cada ámbito. Un ejemplo de tal ámbito sería "restar números de dos dígitos". ¿El estudiante domina la sustracción de números de dos dígitos o no? En un test referido a un criterio, el conjunto de ítems se redacta para cada ámbito y se considera que un estudiante domina el ámbito si responde al 80% de los ítems de manera correcta.

Si bien este método de interpretación de puntuaciones de un test es en apariencia atractiva, en la práctica tiene una serie de defectos. Incluso si los ámbitos están tan bien definidos como "restar números de dos dígitos", el éxito del estudiante depende del conjunto particular de ítems que se aplique. Los ítems de sustracción escritos de forma vertical son más difíciles que los ítems escritos de forma horizontal. Los ítems de sustracción que requieren reagrupar son más difíciles que aquellos que no. Los ítems de sustracción que implican ceros son notoriamente más difíciles. Que un estudiante responda el 80% de ítems de forma correcta depende de los ítems que se apliquen. En un intento por eliminar este tipo de ambigüedades, un enfoque común es definir los ámbitos de desempeño de forma más precisa (por ejemplo, restar números de dos dígitos cuando están escritos de forma vertical y cuando no es necesario reagrupar). Pero la consecuencia de este enfoque es la fragmentación y atomización del currículo escolar, de modo que se tienen listas cada vez más largas de habilidades cada vez más triviales.

En la práctica, un test referido a un criterio es como una vara de medir con una sola marca (el criterio). Los estudiantes cumplen con ese criterio o no. Estas varas de medir específicas para cada ámbito, con sus resultados "sí/no", ofrecen una base muy limitada para monitorear el crecimiento individual en áreas significativas del aprendizaje a lo largo del tiempo.

3. UN MODELO PARA MEDIR

El artículo anterior identificó varias razones por las cuales las puntuaciones comunes en un test (el conteo de ítems respondidos correctamente) no tienen las mismas propiedades que medidas tales como las longitudes en centímetros o las temperaturas en grados Celsius.

- A pesar de que la intención de la mayoría de test desarrollados es generar una única puntuación para cada estudiante (en otras palabras, construir una única dimensión a lo largo de la cual se pueda ordenar a los estudiantes de menos a más), muchos de los test desarrollados no están acompañados de una verificación explícita de la validez de resumir las respuestas a los ítems en una única medida.

- A pesar de que en la mayoría de análisis estadísticos el uso de puntuaciones en un test implica que tiene niveles de intervalo apropiados —dado que las puntuaciones no corresponden a unidades de medida—, por lo general, no tienen una escala de intervalo.

- A pesar de que nuestro interés en las evaluaciones educativas siempre se dirige a una variable subyacente —y no a un conjunto específico de ítems—, las puntuaciones comunes en un test (por ejemplo, 28 de 40) siempre están condicionadas a un test en particular y, por ende, no tienen un significado neutral e independiente del instrumento (como 28 centímetros o 28 kilogramos).

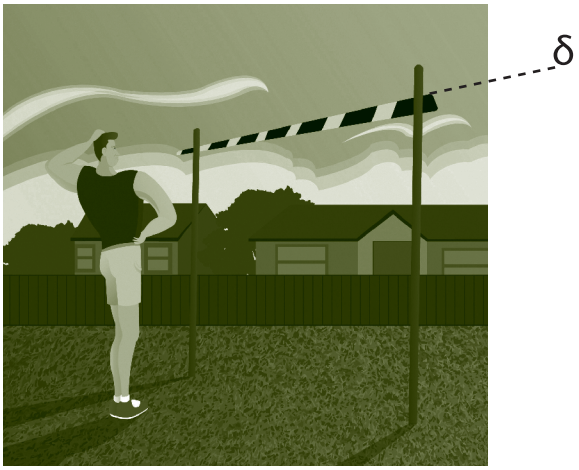
En resumen, las puntuaciones numéricas *correctas* no tienen las propiedades de las medidas.

Este artículo describe un método para construir *medidas* educativas. Estas *medidas* —si se pueden construir— comparten las propiedades descritas en las páginas 26 a 36. En otras palabras, son las siguientes:

- Estimados de posiciones en una única variable (unidimensionales).

- Expresadas en una unidad de medición constante (con un nivel de intervalo).
- Libres de las particularidades del instrumento usado (objetivas).

No es fácil conseguir medidas con estas propiedades. El método aquí descrito requiere datos (observaciones) que satisfagan un conjunto de requerimientos muy demandantes. A pesar de que las medidas educativas deben construirse a partir de las respuestas a los ítems de un test, no todos los conjuntos de ítems cumplen con estos requerimientos ni son capaces de producir medidas unidimensionales, en un nivel de intervalo, objetivas.



¿Cómo se relacionan la dificultad de la tarea (δ) y la habilidad de la persona (β)?)

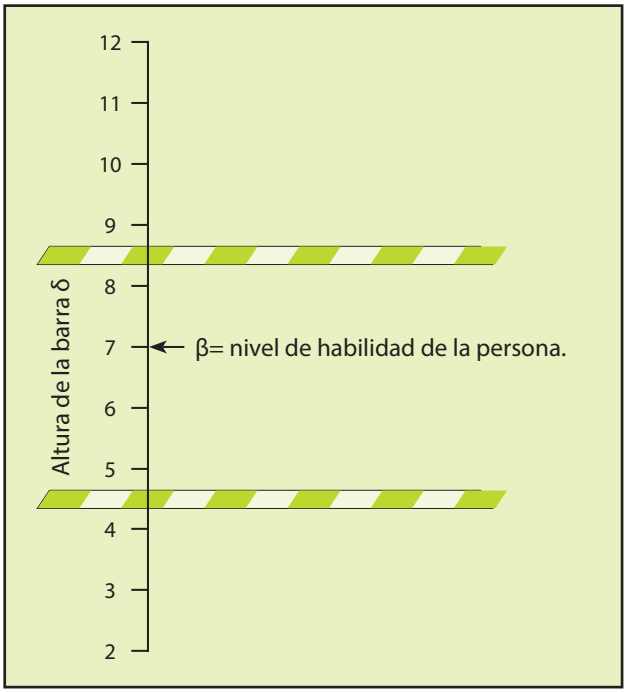
3.1. Una variable

El modelo de medición que se describe aquí empieza con la intención de enfocarse en un único aspecto de la variabilidad (es decir, una variable) y estimar la posición de los individuos en esa variable.

Supongamos, por ejemplo, que la variable de interés sea la habilidad para salto alto. Podríamos plantear como hipótesis que los individuos varían en

la habilidad para el salto y que es posible obtener estimados útiles de esta para el salto alto, a partir de la observación del desempeño en algunas tareas relevantes. Que esta idea se sostenga en la práctica dependerá de la medida en que el desempeño en las tareas que usemos sea consistente con la proposición de que los individuos varían a lo largo de una única dimensión de la habilidad para el salto alto.

La noción de una variable de salto alto se representa en la imagen siguiente. En dicha imagen uno se imagina que la habilidad para el salto alto aumenta a medida que subimos en la página. La altura de la barra determina la dificultad de la tarea y se representa mediante la letra griega delta (δ). A lo largo de este continuo de dificultad en aumento, también se pueden trazar las habilidades de los individuos para el salto alto (β). Se ha marcado la habilidad imaginada β = 7 de un individuo.



Una escala de la dificultad en aumento que muestra el nivel de habilidad de la persona (β).

Se pueden hacer varias observaciones a partir de esta imagen.

En primer lugar, se pueden conceptualizar las habilidades de los individuos para el salto alto (β) y las dificultades (δ) del salto alto como posiciones en un mismo continuo. Las tareas más fáciles y los individuos de menor habilidad se ubicarán en la parte inferior del continuo; las tareas más difíciles y los individuos de mayor habilidad se ubicarán en la parte superior.

En segundo lugar, la habilidad para el salto alto de este individuo se ha etiquetado con la letra griega β (beta), a fin de reflejar el hecho de que nunca podremos conocer con exactitud la habilidad de esta persona: solo podemos imaginarla y luego estimarla a partir de las observaciones de su desempeño. A mayor el número de intentos de salto alto de esta persona, más información tendremos y más podremos confiar en nuestro estimado.

En tercer lugar, la variable de salto alto se ha marcado (optimistamente) en lo que parecen ser unidades iguales. Para desarrollar *medidas* de la habilidad para el salto alto, necesitamos una unidad constante de dicha habilidad.

3.2. Planificación de las observaciones

Para medir individuos en una variable, es necesario recoger evidencia que sea relevante para ella. En el caso de la habilidad para el salto alto, es probable que la evidencia en forma de observaciones sobre el éxito o fracaso en un conjunto de tareas sea una base apropiada para estimar las habilidades del individuo. En el caso de otras variables, la evidencia más apropiada podría recogerse mediante tareas de papel y lápiz, o evaluando portafolios de trabajo, proyectos completos o productos tecnológicos o de arte.

Para estimar la posición de una persona con respecto a una variable, no suele bastar con observar su desempeño en (o su respuesta a) una única tarea. El éxito o fracaso en una sola tarea de salto alto o una sola pregunta de un test de lectura provee información muy limitada sobre la habilidad de un

individuo. Las medidas confiables demandan múltiples observaciones.

La medición requiere entonces que las observaciones se hagan bajo condiciones controladas. Puede que la idea de que los individuos varían en su habilidad para el salto alto haya surgido de observaciones casuales de personas que saltan sobre troncos, rocas, setos o vallas. Pero para comparar (y medir) la habilidad para el salto alto no les pediríamos a algunos individuos que salten una valla, a otros que salten un seto y a unos terceros que salten una soga. Más bien, estandarizaríamos las condiciones de observación para minimizar la influencia de factores irrelevantes con respecto a la variable de interés. Lo mismo rige para toda medición. Por ejemplo, cuando medimos la estatura de los niños, los medimos en una situación controlada y artificial: sin zapatos, con la barbilla hacia arriba y pegados a la pared.

La medición requiere de observaciones en condiciones controladas.

3.3. Registros de observaciones

Una vez que se ha decidido cuál es el método de evaluación que se usará, se requiere tomar una decisión sobre las observaciones o los juicios que se registrarán. Existen varias posibilidades. Una es registrar los *ratings* del desempeño de los individuos o de su trabajo. En la evaluación del desempeño en áreas tales como la gimnasia, la oratoria, el buceo

o la música instrumental, y también en la evaluación de los textos escritos por los estudiantes y de los productos de su trabajo en tecnología y arte, se suelen usar puntajes de jueces. Una segunda posibilidad es usar un sistema de puntuaciones de crédito parcial para identificar a los estudiantes que dieron respuestas parcialmente correctas o que fueron parcialmente exitosos al resolver un problema. Una tercera posibilidad es usar *puntuaciones dicotómicas* para registrar el éxito o fracaso en una tarea. Por ejemplo, las respuestas de los estudiantes a las preguntas de un test se suelen registrar como correctas o incorrectas. Los intentos de los individuos de pasar la barra de salto alto también se registran de forma dicotómica (pasó/la tocó).

Se suelen tabular los registros de las observaciones. La siguiente tabla muestra cómo se podría tabular un conjunto de registros de salto alto para N personas. Los resultados de cada persona se registran en una fila de la tabla. Se asume que los individuos tienen diferentes habilidades (β_1, β_2, β_3, …, β_n). Cada columna corresponde a una altura particular de la barra (con la dificultad δ) y el resultado de cada intento se registra ya sea con 1 (éxito) o con 0 (fracaso).

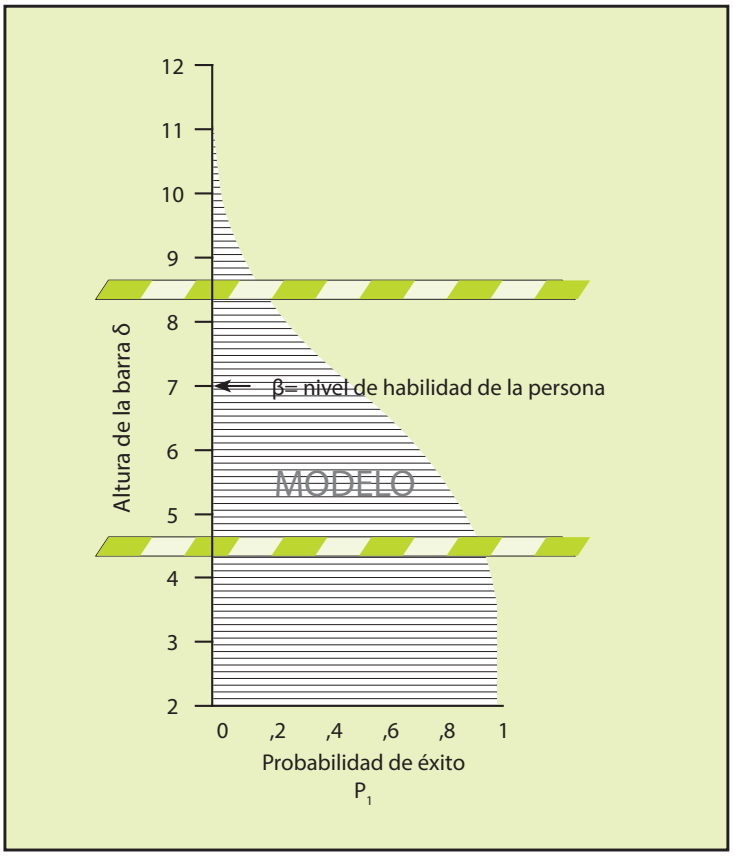
Tareas

	δ_1	δ_2	δ_3	δ_4	δ_5	...	δ_L
β_1	1	1	1	1	0		0
β_2	1	1	0	1	0		0
β_3	1	1	1	1	0		0
β_4	1	1	0	0	0		0
β_5	1	0	1	0	1		0
...							
β_N	1	0	1	0	1		0

3.4. Un modelo de medición

El modelo de medición desarrollado por el matemático danés Georg Rasch (1901-1980) provee una base para estimar la habilidad β de una persona a partir de la fila de desempeños registrados para ella. El modelo propone una relación matemática entre la habilidad β de una persona, la dificultad δ de la tarea que se intenta resolver y la probabilidad P_1 de que la persona tenga éxito en esa tarea.

En la siguiente imagen se muestra esta relación matemática.



Para una persona con habilidad β, la probabilidad de éxito modelada P1 disminuye con el aumento de la dificultad de la tarea δ.

Esta imagen muestra cómo en el modelo Rasch la probabilidad de éxito P_1 de una persona disminuye a medida que aumenta la dificultad de la tarea. Se muestra la relación para una persona de habilidad $\beta = 7$. Mientras más difícil sea la tarea (es decir, mientras más alta esté la barra), menor será la probabilidad de éxito de la persona.

Cuando el resultado del intento de una persona en una tarea se registra con 1 para el éxito y 0 para el fracaso, la probabilidad de éxito P_1 y la probabilidad de fracaso P_0 suman uno ($P_1 + P_0 = 1$).

Expresado matemáticamente, el modelo Rasch provee la probabilidad P_1 de que una persona de habilidad β tenga éxito en una tarea de dificultad δ de la siguiente manera:

$$P_1 = \exp (\beta - \delta) / 1 + \exp (\beta - \delta)$$

Nótese que la probabilidad de éxito P_1 depende de $\beta - \delta$ (en otras palabras, cuán distante está la barra del nivel de habilidad de la persona). Cuando la barra se ajusta a la habilidad de la persona, entonces $\beta - \delta = 0$, y

$$P_1 = \exp (0) / 1 + \exp (0) = 0{,}5$$

3.5. Una unidad de medición

El modelo Rasch se puede reorganizar como:

$$\beta - \delta = 1n (P_1 / P_0)$$

Donde la unidad en que β y δ se expresa se llama "logit". Cuando el modelo Rasch se usa para construir medidas de habilidad, la variable de medición resultante se calibra en logits.

La tabla de la derecha muestra la probabilidad modelada P_1 de que una persona con una habilidad $\beta = 7$ tenga éxito en tareas con dificultades δ en un rango de 2 a 12 logits.

δ	P_1
12,0	,007
11,8	,008
11,6	,010
11,4	,012
11,2	,015
11,0	,018
10,8	,022
10,6	,027
10,4	,032
10,2	,039
10,0	,047
9.8	,057
9,6	,069
9,4	,083
9,2	,100
9,0	,119
8,8	,142
8,6	,168
8,4	,198
8,2	,231
8,0	,269
7,8	,310
7,6	,354
7,4	,401
7,2	,450
7,0	,500
6,8	,550
6,6	,599
6,4	,646
6,2	,690
6,0	,731
5,8	,769
5,6	,802
5,4	,832
5,2	,858
5,0	,881
4,8	,900
4,6	,917
4,4	,931
4,2	,943
4,0	,953
3,8	,961
3,6	,968
3,4	,973
3,2	,978
3,0	,982
2,8	,985
2,6	,988
2,4	,990
2,2	,992
2,0	,993

3.6. La clave para la objetividad

Una intención fundamental en toda medición es que las medidas de la variable sean independientes de los detalles del instrumento particular que se usó para obtenerlas. En la medición educativa, nuestro interés siempre está puesto en la variable (es decir, el constructo) que se vaya a medir, y no en un ítem o conjunto de ítems en particular. Cada test es, simplemente, una muestra conveniente de muchos ítems que podrían usarse para recoger evidencia sobre esa variable.

En el nivel más elemental, esta intención implica que, si tuviéramos que considerar a dos personas A y B con las supuestas habilidades β_A y β_B en la variable de interés, entonces nuestro estimado de la diferencia $\beta_A - \beta_B$ entre ambas no debería depender de los ítems particulares del test que usamos para calcular esta diferencia. Si en un conjunto de ítems se estimó que la persona A sería, digamos, 1 logit más hábil que la persona B, entonces debería calcularse que la persona A sea 1 logit más hábil que la persona B en cualquier otro conjunto de ítems que midan esa variable (dentro de los márgenes de error de medición).

En nuestra analogía del salto alto, esperaríamos que nuestro estimado de las habilidades relativas de salto alto de personas A y B tenga un significado generalizable, es decir, que su significado no se limite a las pocas observaciones que hicimos o a las alturas particulares en las que colocamos la barra. Solo si nuestros estimados de las habilidades relativas de los individuos son generalizables a tareas más allá de las usadas para obtenerlos, tendremos alguna esperanza de estar construyendo "medidas" de esa variable.

Cuando dos personas A y B intentan lograr la misma tarea dicotómica, existen cuatro resultados posibles: ambas personas tienen éxito (✓✓); la persona A tiene éxito pero la persona B fracasa (✓✗): la persona A fracasa pero B tiene éxito (✗✓); ambas personas fracasan (✗✗). Solo dos de estos resultados (✓✗ y ✗✓) contienen información sobre las habilidades *relativas* de las personas A y B.

Si la probabilidad de que la persona A tenga éxito en la tarea es de P_A y la probabilidad de que la persona B tenga éxitos es de P_B, entonces las probabilidades de estos cuatro posibles resultados se dan según las siguientes probabilidades combinadas (ver nota en la página 51):

Ambas personas tienen éxito $P_{✓✓} = P_A \times P_B$

La persona A tiene éxito pero B fracasa $P_{✓✗} = P_A \times (1 - P_B)$

La persona A fracasa pero B tiene éxito $P_{✗✓} = (1 - P_A) \times P_B$

Ambas personas fracasan $P_{✗✗} = (1 - P_A) \times (1 - P_B)$

La probabilidad condicional de que la persona A tenga éxito y B fracase, si una persona tiene éxito y otra fracasa, es:

$$P_{✓✗}/(P_{✓✗}+P_{✗✓}) = P_A x(1-P_B)/[P_A x(1-P_B)+(1-P_A)x\ P_B]$$

Y la probabilidad condicional de que la persona A fracase y B tenga éxito, si una persona tiene éxito y otra fracasa, es:

$$P_{\times\checkmark}/(P_{\checkmark\times}+P_{\times\checkmark})=(1-P_A)xP_B/[P_Ax(1-P_B)+(1-P_A)xP_B]$$

A partir de estas dos ecuaciones se da lo siguiente:

$$P_{\checkmark\times} / P_{\times\checkmark} = \exp(\beta_A - \delta) / \exp(\beta_B - \delta) = \exp(\beta_A - \beta_B)$$

En otras palabras:

$$\beta_A - \beta_B = 1n\ (P_{\checkmark\times} / P_{\times\checkmark})$$

Las implicancias de esta característica del modelo Rasch se muestran en el gráfico de la siguiente página.

En dicha imagen están marcadas las habilidades para el salto alto ($\beta_A = 7$ logits; $\beta_B = 6$ logits) de las personas A y B. El gráfico muestra cómo las probabilidades Rasch de que ambas personas A y B tengan éxito (✓✓), A fracase y B tenga éxito (✗✓), A tenga éxito y B fracase (✓✗), y ambas personas fracasen (✗✗) varían según la dificultad δ de la tarea.

Lo importante en esta imagen es que la ratio del ancho de la zona gris claro (P✓✗) y el ancho de la zona gris oscuro (P✗✓) es *constante* para todos los valores de δ (por ejemplo ,24/,09 = ,36/,13 = ,11/,04).

El significado de esta característica del modelo Rasch es que no resulta necesario conocer o estimar la altura de la barra (dificultad de la tarea) para estimar las habilidades relativas de las personas A y B.

Si estas personas tuvieran múltiples intentos para pasar la barra a *cualquier* altura, se podría obtener un estimado de sus habilidades relativas a partir del

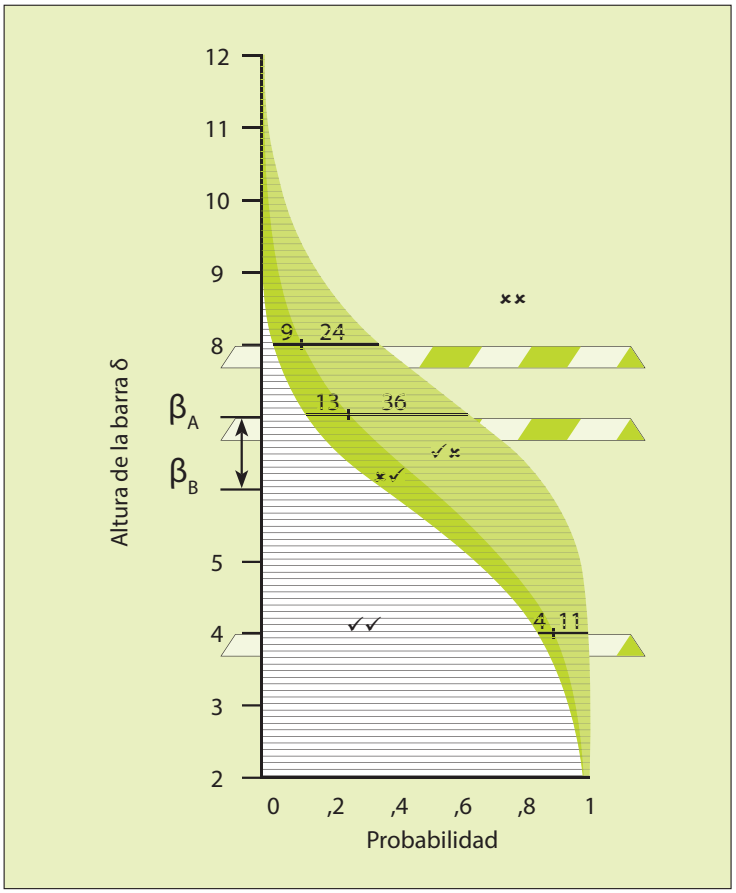
número de veces en las que A tuvo éxito y B fracasó (N✓✗) y del número de veces en las que A fracasó y B tuvo éxito (N✗✓). Todo lo que se requiere es calcular las proporciones:

$P✓✗ = N✓✗ / (N✓✗ + N✗✓)$

$P✗✓ = N✗✓ / (N✓✗ + N✗✓)$

Y sustituir en:

$b_A - b_B = 1n (p_{✓✗} / p_{✗✓}) = 1n (N_{✓✗} / N_{✗✓})$



Donde $b_A - b_B$ es un estimado de la diferencia $\beta_A - \beta_B$. En otras palabras, si un conjunto de datos de salto alto es conforme con el modelo Rasch, entonces la diferencia ($\beta_A - \beta_B$) entre las personas A y B puede estimarse (en logits) poniendo la barra a cualquier altura y contando simplemente los resultados ✓✗ y ✗✓.

Nota sobre probabilidades

Si una lanza una moneda, existen dos resultados posibles: cara (H) y sello (T). Si la moneda no ha sido alterada, entonces la probabilidad para cada resultado es de 50:50. En otras palabras, las probabilidades son:

P(H) = 0,5

P(T) = 0,5

Probabilidad combinada

Si se lanzan dos monedas inalteradas y los resultados de estos lanzamientos son independientes el uno del otro, entonces se dan cuatro posibles resultados igual de probables: HH, HT, TH, y TT.

Las probabilidades son:

P(HH) = P(H) x P(H) = 0,5 x 0,5 = 0,25

P(HT) = P(H) x P(T) = 0,5 x 0,5 = 0,25

P(TH) = P(T) x P(H) = 0,5 x 0,5 = 0,25

P(TT) = P(T) x P(T) = 0,5 x 0,5 = 0,25

Probabilidad condicional

Si nos dicen que un determinado lanzamiento de dos monedas resultó en *impares* (HT o TH) en lugar de *pares* (TT o HH), la probabilidad de que el resultado sea HT y no TH es:

P(HT) / (P(HT) + P(TH)) = 0,25 / (0,25 + 0,25) = 0,5

3.7. Un ejemplo

Para ilustrar esta característica fundamental del modelo Rasch, ahora consideraremos los resultados de un ejercicio hipotético de salto alto.

Supongamos que, para estimar las habilidades relativas para el salto alto de las personas A y B, colocamos la barra a tres alturas diferentes (i, ii e iii) y registramos los resultados de las dos personas en 100 intentos para cada altura:

Altura de la barra	A y B pasan ✓✓	A pasa B falla ✓✗	B pasa A falla ✗✓	A y B fallan ✗✗	Total de intentos
iii (difícil)	1	11	4	84	100
ii	14	36	13	37	100
i (fácil)	64	24	9	3	100

Solo las partes sombreadas de esta tabla contienen información sobre las habilidades relativas de las personas A y B. Ahora podemos usar la ecuación final que aparece en la página 50 para estimar la diferencia entre las habilidades para el salto alto de las personas A y B:

$b_A - b_B = 1n (N_{✓✗} / N_{✗✓})$

Esta diferencia podría estimarse usando por separado los resultados de las personas en las tres alturas:

En función de los intentos en la altura i:

$b_A - b_B = 1n (24 / 9) = 0{,}98$ logits

En función de los intentos en la altura ii:

$b_A - b_B = 1n (36/ 13) = 1{,}02$ logits

En función de los intentos en la altura iii:

$b_A - b_B = 1n\ (11 / 4) = 1{,}01$ logits

3.8. Analizando el ajuste

Nótese que, dado que los "resultados" de la tabla de la página 52 se ajustan muy bien al modelo Rasch, los tres estimados son casi idénticos. Los intentos en las alturas i, ii y iii llevan a la conclusión de que la habilidad de la persona A para el salto alto es alrededor de un logit mayor que la habilidad de la persona B.

La comparación de estos tres estimados provee un test sencillo del ajuste de esta data al modelo Rasch. Si, en un experimento real de salto alto de este tipo, las tres alturas no llevaran a estimados similares de la diferencia entre las personas A y B, entonces habría evidencia de que la data no se comporta conforme al modelo.

3.9. Comparaciones objetivas

Nótese que, en el ejemplo anterior, no fue necesario conocer las alturas i, ii y iii de la barra (en pulgadas, centímetros o logits) para estimar las habilidades relativas de las personas A y B. Cuando la data se ajusta al modelo Rasch, es posible comparar habilidades sin saber —o incluso tener que estimar— las dificultades de la tarea. Esta es una característica única del modelo Rasch.

Para continuar ilustrando que los estimados Rasch de habilidad son independientes de las dificultades de las tareas, nótese que podríamos haber estimado las habilidades relativas de las personas A y B a partir de los resultados de cualquiera de los 200 intentos:

i y ii	bA - bB	= 1n ((11 + 36) / (4 + 13))	= 1n (47/17)
			1,02 logits

ii y iii bA - bB = 1n ((36 + 24) / (13 + 9)) = 1n (60/22)
1,00 logits

i y iii bA - bB = 1n ((11 + 24) / (4 + 9)) = 1n (35/13)
0,99 logits

O, el mejor estimado posible, de los resultados de todos los 300 intentos

bA - bB = 1n ((11 + 36 + 24) / (4 + 13 + 9))
= 1n (71 + 26) = 1,00 logits

Dado que es posible simplemente sumar las columnas del centro de la tabla de la página 52 de esta forma, sin prestar atención a las dificultades de las tareas, no existe una razón para que las personas A y B tengan que hacer más de un intento para cada altura. Las dos personas podrían hacer un intento cada una para L alturas diferentes y registrar los resultados en una tabla como esta:

Medida de altura	A y B pasan ✓✓	A pasa B falla ✓✗	B pasa A falla ✗✓	A y B fallan ✗✗	Total de intentos
1	0	1	0	0	1
2	0	1	0	0	1
3	1	0	0	0	1
...					1
L	0	0	1	0	1

Nuevamente, las habilidades relativas de las personas A y B podrían estimarse sumando las columnas del centro de la tabla para obtener $N_{✓✗}$ y $N_{✗✓}$ y luego sustituir en:

$b_A - b_B = 1n (N_{✓✗} / N_{✗✓})$

Si el desempeño de las personas A y B fuera consistente con el desempeño en la página 52, entonces los totales de las dos columnas del centro de la

tabla tendrían un ratio de aproximadamente 2,7:1, lo que llevaría a una diferencia estimada de alrededor de 1,0 logits, sin importar las alturas que intentaron.

En el caso de arriba, el ajuste de la data al modelo puede probarse al comparar el estimado b_A - b_B, que se basa en los intentos en las primeras L/2 alturas con los estimados basados en los intentos de las segundas L/2 alturas, o comparando el estimado basado en todas las alturas pares con aquel basado en todas las alturas impares. Estos cuatro estimados serán muy similares si las observaciones registradas se ajustaran al modelo Rasch.

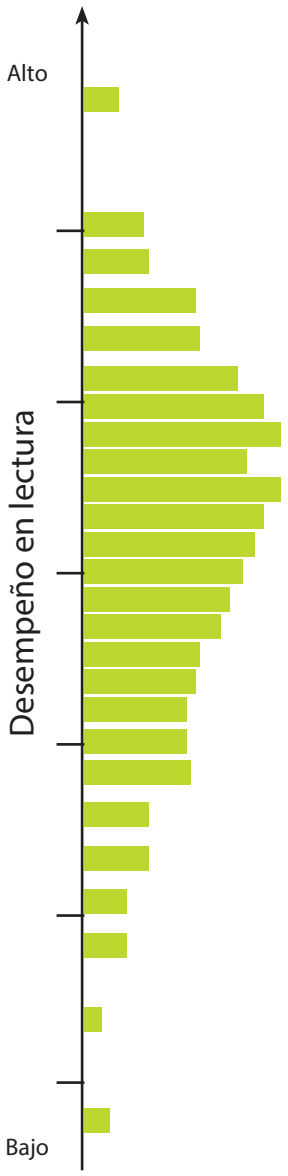
3.10. Aplicación a data de test

El procedimiento que se acaba de aplicar a data de salto alto también se puede aplicar a un conjunto de data de test. En un test, cada evaluado tiene un solo intento por ítem. Si ese intento se registra, ya sea como correcto (✓) o como incorrecto (✗), entonces podría resumirse el desempeño de las personas A y B en un test de la longitud L según la siguiente tabla:

Número de Item	A y B correcto ✓✓	A correcto B incorrecto ✓✗	B correcto A incorrecto ✗✓	A y B incorrectos ✗✗	Total de intentos
1	0	1	0	0	1
2	0	1	0	0	1
3	1	0	0	0	1
...					1
L	0	0	1	0	1

Las habilidades relativas de las personas A y B se estiman sumando las dos columnas centrales de la tabla, de modo que se obtienen $N_{✓✗}$ y $N_{✗✓}$ y se sustituyen en:

$$b_A - b_B = 1n\ (N_{✓✗} / N_{✗✓})$$



Distribución de las posiciones estimadas de los estudiantes con respecto a una variable de desempeño en lectura.

Se podría evaluar el ajuste de esta data al modelo Rasch comparando los estimados obtenidos de diferentes subconjuntos de ítems (por ejemplo, ítems pares, ítems impares; primera mitad del test, segunda mitad del test).

3.11. Estimación de la posición en una variable

Si se considera de esta manera a todas las parejas de estudiantes que dan el test, y se recogen todos estos pares estimados, entonces es posible estimar las posiciones de todos estudiantes que dan el test con respecto a la variable. Este procedimiento se conoce como el método "por pares" de estimar las habilidades de los estudiantes en una variable.[1]

El resultado de aplicar el procedimiento de estimación por pares a un conjunto de data de test es un estimado de la posición de cada estudiante con respecto a la variable para cuya medición se ha diseñado el test. El diagrama de esta página muestra la distribución de las posiciones estimadas de los estudiantes de 3.er año con respecto a una variable de logro

[1] Choppin, Bruce. (1976). "Desarrollos recientes en el banco de ítems". En D. N. de Gruijter & L. J. van der Kamp, eds., *Advances in Psychological and Educational Measurement* (pp. 97-110). Londres: Wiley.

creciente en lectura. Estas posiciones estimadas se trazan en una escala de nivel de intervalo en *logits*.

En resumen

El modelo Rasch descrito en este artículo especifica los requisitos que debe cumplir un conjunto de data de test, si se quiere que provea *medidas* que: (i) sean estimados de las posiciones de los individuos en una única variable/dimensión; (ii) se expresen en una unidad constante de medición, y (iii) estén libres de las particularidades del instrumento que se usó para obtenerlas.

No es fácil obtener medidas con estas características. A pesar de que se pueden construir a partir de las respuestas a ítems de test, no todos los conjuntos de ítems cumplen estos requisitos y son capaces de producir medidas unidimensionales, de nivel intervalo y objetivas.

La clave de las medidas objetivas reside en el hecho de que, cuando dos personas A y B hacen un intento en un ítem, bajo el modelo Rasch, el ratio $P_{✓✗} / P_{✗✓}$ se define solo en términos de las habilidades relativas de estas dos personas:

$$\beta_A - \beta_B = 1n(P_{✓✗} / P_{✗✓})$$

Donde $P_{✓✗}$ es la probabilidad modelada de que la persona A tenga éxito y la persona B fracase en el ítem, y $P_{✗✓}$ es la probabilidad de que A falle y B tenga éxito. Es esta característica del modelo la que hace posible que se den medidas que estén "libres" de las particularidades de los ítems usados para obtenerlas.

Cuando se aplica el modelo Rasch, este provee una medida para cada estudiante en un continuo marcado por intervalos iguales llamados *logits*.

4. TRAZANDO VARIABLES

Una característica fundamental de las medidas es que indican posiciones en variables generales. En otras palabras, tienen significados que no se limitan a —y pueden generalizarse más allá de— los instrumentos específicos usados para obtenerlas. Por ejemplo, las medidas de longitud en centímetros indican posiciones en la variable general "longitud" y tienen significados que no dependen de los detalles del instrumento usado (por ejemplo, una regla de madera, una wincha, un calibrador, una cinta métrica).

Las medidas educativas también están pensadas para indicar posiciones en variables generales. Por ejemplo, las medidas de la habilidad para la lectura están pensadas para indicar posiciones en la variable general "habilidad de lectura" y para tener significados que no se limiten a un pasaje del texto en particular o a las preguntas específicas del test usado para obtenerlas.

Los constructores de test saben que las preguntas individuales no significan nada en sí mismas: simplemente son oportunidades para recoger muestras del comportamiento con el propósito de estimar posiciones en la variable general de interés.

Cuando se trata de *interpretar* medidas educativas, es importante ver más allá de las especificidades del instrumento hacia las generalidades de la variable de medición subyacente. A continuación, nos abocaremos a este tema.

4.1. Distinguir una variable

En nuestra discusión sobre la medición de la habilidad para el salto alto, notamos que se podía conceptualizar las dificultades de las tareas de salto alto y las habilidades de los individuos como posiciones en una misma variable. En los eventos reales de salto alto, se determina la dificultad de una tarea mediante la altura de la barra desde el suelo. Cuando la barra se coloca a alturas

crecientes, estas tareas de dificultad creciente definen los niveles crecientes de la habilidad para el salto alto.

Pero una alternativa para medir la altura de la barra desde el suelo consistiría en estimar la dificultad de cada tarea de salto alto a partir del registro de los éxitos de la persona en esa tarea. Si un grupo de personas intentara el mismo conjunto de tareas, entonces se estimaría como más fácil la altura que la mayoría del grupo pasó y como más difícil la altura que la minoría pasó. Usando el modelo de medición de la página 45, se estimaría la dificultad de cada tarea (en logits) a partir de los registros disponibles de los saltos.

Este proceso de estimación de dificultades de un conjunto de tareas se conoce como "calibración". Para ilustrarlo es conveniente empezar por considerar los intentos de un individuo en dos tareas de salto alto Y y Z con dificultades δ_Y y δ_Z. Si la persona tuvo un intento para cada altura, entonces hay cuatro resultados posibles: la persona tiene éxito en ambos (✓✓); tiene éxito en Y pero falla en Z (✓✗); tiene éxito en Z pero falla en Y (✗✓), y falla en ambos (✗✗).

Solo dos de estos cuatro resultados posibles (✓✗ y ✗✓) son útiles para estimar las dificultades relativas de ambas tareas.

Si se siguen pasos paralelos a los descritos en la página 49 —que incluyen calcular la probabilidad condicional de que la persona tenga éxito en una tarea con la condición de que tenga éxito en una pero falle en la otra—, entonces la distancia entre las tareas Y y Z en la variable es:

$$\delta_Z \text{ y } \delta_Y = 1n\ (P_{✓✗} / P_{✗✓})$$

Si esta persona hace intentos en las tareas Y y Z en varias ocasiones, y en cada ocasión se registra si el resultado es ✗✗, ✓✗, ✗✓ o ✓✓, entonces se puede estimar la distancia entre las tareas Y y Z:

$$d_Z \text{ y } d_Y = 1n\ (n_{✓✗} / n_{✗✓})$$

Donde $d_Z - d_Y$ es un estimado de $\delta_Z - \delta_Y$, $n_{✓✗}$ es el número de veces en que la persona tiene éxito en Y pero falla en Z, y $n_{✓✗}$ es el número de veces en que la persona tiene éxito en Z pero falla en Y.

En este caso, la observación importante es que este estimado no depende de la habilidad de la persona. La distancia entre las tareas Y y Z puede estimarse contando los resultados $_{✓✗}$ y $_{✗✓}$ en el caso de *cualquier* persona. Y cuando la data se ajusta al modelo, los estimados obtenidos de esta forma a partir del desempeño de diferentes individuos son estadísticamente equivalentes.

De esta observación se puede concluir que, para estimar la distancia entre las tareas Y y Z, no es necesario pedir a los individuos que hagan más de un intento en cada tarea. Para cualquier grupo de personas, todo lo que se requiere es que se mantenga un registro de la cantidad de resultados ✓✗ y ✗✓ para el grupo. Luego, la distancia se puede estimar como:

$$d_Z - d_Y = 1n\ (N_{✓✗} / N_{✗✓})$$

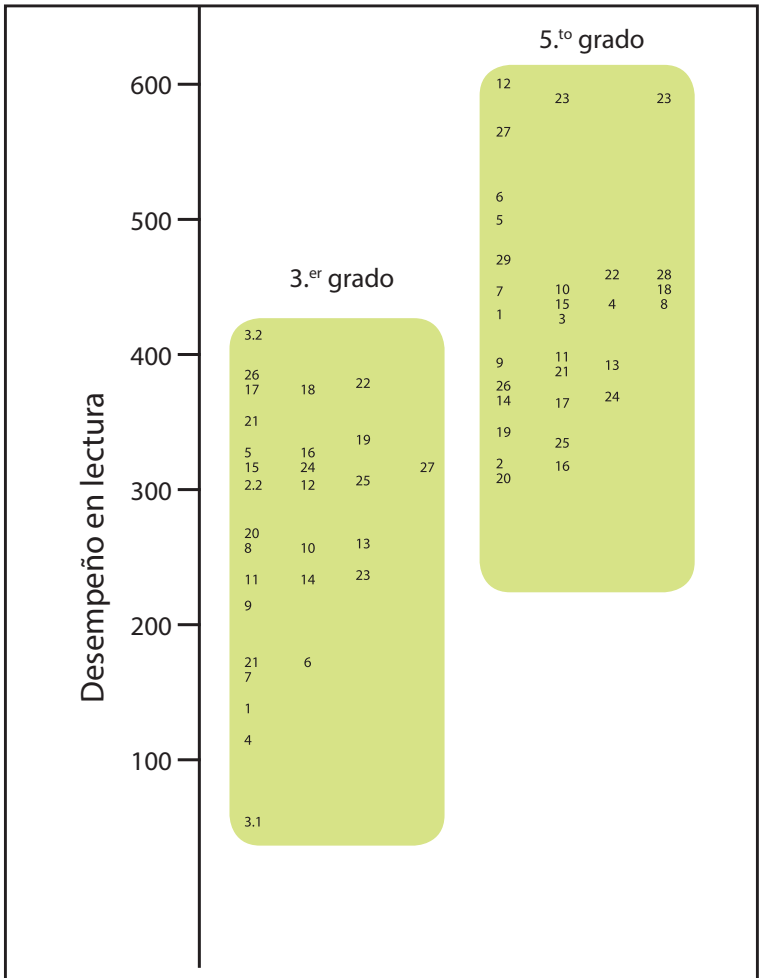
Donde $N_{✓✗}$ es el número de personas que tiene éxito en Y pero falla en Z, y $N_{✗✓}$ es el número de personas que tiene éxito en Z pero falla en Y.

Se puede repetir este proceso para todos los pares posibles de tareas, y se puede recoger las distancias estimadas entre las tareas para calibrar todas ellas a lo largo de la misma variable.

Cuando se aplica este proceso a registros del desempeño de estudiantes en un conjunto de ítems de test, se obtiene un estimado (en logits) de la dificultad de cada ítem, lo que permite calibrar todos los ítems a lo largo de la variable en la que se está midiendo a los estudiantes.

El diagrama siguiente muestra dos conjuntos de ítems calibrados en una variable de desempeño creciente en la lectura. Los ítems mostrados aquí se

aplicaron en dos test: uno para estudiantes de 3.er grado y uno para estudiantes de 5.to grado. El ítem más sencillo (numerado 3.1 en el test del año 3) está en la parte inferior del diagrama; el más difícil (numerado 13 en el test del año 5) está en la parte superior. A partir de este diagrama queda claro que el test del 5.to grado fue más difícil que el de 3.er grado en términos generales, a pesar de que hubo muchos ítems en el rango 300 a 400 en ambos (los números en la escala vertical son logits múltiples).



Ítems de lectura calibrados a lo largo de la variable de desempeño creciente en la lectura

"La publicación de 1675 de John Ogilby de los mapas de las rutas principales para salir de Londres fue la primera de su tipo en toda Europa. Los caminos se ilustraron como líneas paralelas hacia la parte superior de la página, sin importar qué dirección tenían en la tierra. Se incluyeron dibujos de los instrumentos usados para su construcción: ruedas de carretera para medir las distancias, cuadrantes y cadenas de topografía. También se señalaron los lugares de atracción a lo largo de los caminos".

John Ure

Cuando una cantidad de ítems del test se calibró de esta forma a lo largo de la variable, las posiciones de los ítems individuales revelaron la variable subyacente. Cada ítem es un ejemplo de la variable en la región en que está calibrado. Por ejemplo, el ítem 3.1 es un ítem de lectura relativamente sencillo que requiere un nivel bajo de habilidad para la lectura. Este ítem requiere que los niños de 3.er grado observen la carátula de un libro apropiado para su edad e identifiquen los elementos claves de la historia, a partir de título del libro y de la ilustración. Es probable que los niños con habilidades muy bajas para la lectura (menor a 100 en esta escala) no puedan completar tareas de este tipo. El ítem 3.1 es el único de ambos test que ilustra este nivel incipiente en el desarrollo de la lectura.

En el otro extremo, los ítems de lectura más difíciles en estos dos test son los ítems 12, 23 y 27 en el test de 5.to grado. Para responderlos de forma correcta, los estudiantes deben interpretar la expresión "por último, pero no por menos" ("por último, pero no menos importante"), inferir el significado del lenguaje figurativo y demostrar una comprensión de la conexión entre contenido y forma de un fragmento de un texto. Estos son ejemplos de niveles de habilidad para la lectura relativamente altos en la región 600 de esta escala.

Un análisis de lo que los estudiantes deben hacer para dar la respuesta correcta a cada ítem de estos dos test señala el inicio de un mapa del desarrollo de la lectura entre 3.er y 5.to grado de primaria. El mapa de desempeño en lectura de la página 65 resume las habilidades evaluadas por la mayoría de los ítems de ambos test. Una versión más detallada de este mapa incluiría ejemplos de los ítems del test para ilustrar las posiciones a lo largo del mapa. Y se obtendría una comprensión aún más rica si se añadieran a esta imagen otros ítems calibrados y se investigaran las características típicas y las demandas de los ítems en varias ubicaciones del continuo.

4.2. Análisis de la estabilidad

Cuando se mapean variables de esta manera, una pregunta importante es si la ubicación de los ítems a lo largo de la variable son estables para todos los estudiantes que participaron de la medición.

Se puede medir y comparar a los individuos de forma significativa solo si la variable en sí es estable.

En el nivel más elemental, podemos preguntar si se obtiene la misma diferencia estimada $d_Z - d_Y$ en las dificultades de los dos ítems Y y Z de las respuestas de diferentes grupos de estudiantes (por ejemplo, entre estudiantes de sexo femenino o masculino, de rendimiento alto o bajo, con número par o impar). Este test se puede realizar al contar los estudiantes en cada grupo con Y correcta y Z incorrecta ($N_{✓✗}$) y los estudiantes con Y incorrecta y Z correcta ($N_{✗✓}$) y luego estimar la diferencia de la siguiente manera:

$$d_Z - d_Y = 1n\ (N_{✓✗} / N_{✗✓})$$

Si las observaciones se ajustan al modelo Rasch, esta diferencia es la misma (estadísticamente equivalente) para diferentes subgrupos de estudiantes. Dicho de forma más general, si un instrumento es estable en su funcionamiento

Si una vara midiera distinto una alfombra, una imagen o un pedazo de papel, entonces la vara como instrumento de medición estaría dañada con respecto a su confiabilidad. La función de un instrumento de medición debe ser independiente del objeto de medición dentro del rango de objetos para los cuales se supone que ha sido pensando el instrumento.

Thurstone (1928: 547)[1]

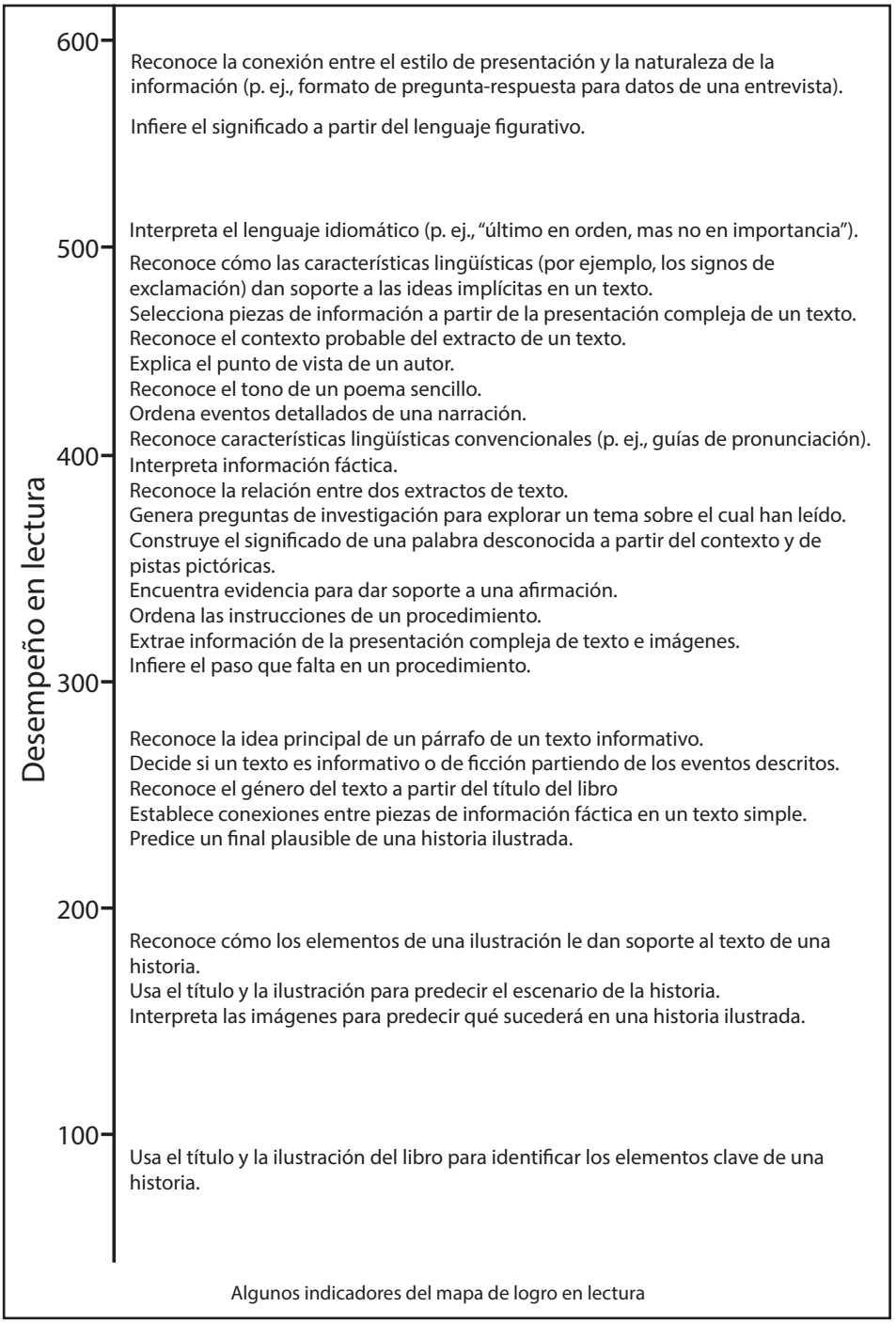
para todos los estudiantes con los cuales se usará, entonces se obtendrán estimados estadísticamente equivalentes de la diferencia $d_j - d_i$ para cada par de ítems (i, j), a partir de las respuestas de diferentes subgrupos de estudiantes.

El análisis estadístico de la estabilidad de las dificultades de los ítems para diferentes subgrupos de estudiantes se conoce como análisis del "funcionamiento diferencial de los ítems" (*dif*, por sus siglas en inglés).

El gráfico de la página 67 muestra de forma pictórica los resultados de un análisis *dif*. Este gráfico se construyó calibrando los ítems de un test estatal de lectura para primaria (y para alumnos de género masculino y femenino por separado) y trazando luego estos dos conjuntos de estimados de la dificultad. El ítem más fácil en este conjunto (tanto para hombres como para mujeres) se ubica en la parte posterior izquierda del gráfico; el ítem más difícil está en la parte superior derecha. La línea sin sombra muestra la región de equivalencia estadística según el modelo (región de confianza del 95%).

Dos ítems (ver flechas) se ubican sobre la línea. En relación con los otros ítems de este

[1] Thurstone, L. L. (1928). "Attitudes can be measured". *American Journal of Sociology* 33, 529-554.



Algunos indicadores del mapa de logro en lectura

test, estos dos ítems son significativamente más difíciles tanto para las mujeres como para los hombres. Una pregunta que se podría hacer sobre estos dos ítems es si su contenido coloca a las mujeres en desventaja. Un tercer ítem está justo por debajo de la línea y, en relación con los otros ítems del test, es más difícil para los chicos que para las chicas.

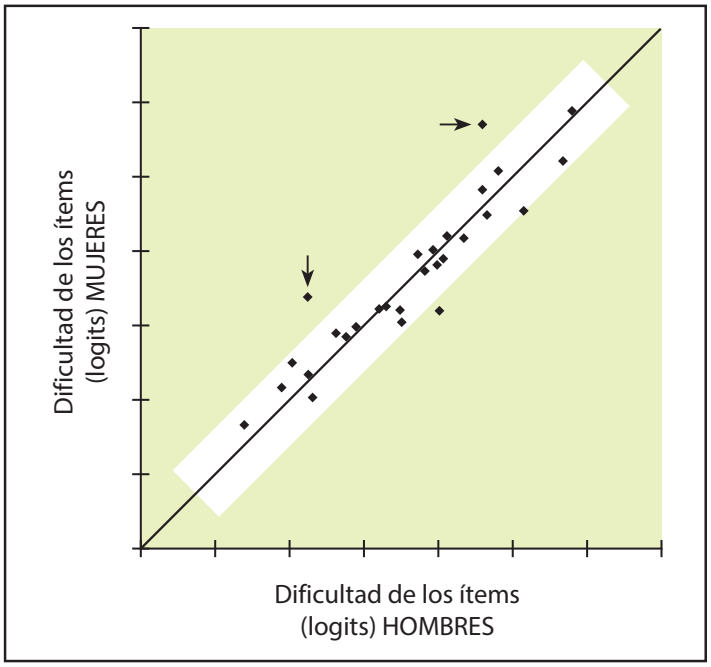
Cuando se usa de forma rutinaria en la construcción de test, el análisis *dif* proporciona una base para identificar ítems que podrían perder la neutralidad para ciertos grupos de estudiantes. Solo si los ítems mantienen su dificultad relativa en todas las poblaciones de estudiantes con las que se supone que se usarán (es decir, son "neutros"), se puede decir que proveen la estabilidad requerida de un instrumento de medición.

4.3. Bancos de ítems

El mapa de la página 65 muestra los ítems de lectura de dos test calibrados en un continuo de desempeño creciente en lectura. Otros ítems de lectura podrían desarrollarse y calibrarse de acuerdo con este continuo, si las respuestas a estos ítems también se ajustaran al modelo Rasch. En teoría, no hay un límite en la cantidad de ítems que podrían calibrarse para una variable, y mientras mayor sea la cantidad de ítems calibrados, más rica será la descripción e ilustración de esa variable.

Aquí nos referimos como un "banco de ítems" a una colección de ítems calibrados. Algunos autores usan el término "banco de ítems" para cualquier tipo de colección de preguntas de evaluación. Nosotros seguimos la convención propuesta por Bruce Choppin de reservar el término banco para referirnos a un conjunto de ítems calibrados juntos para una variable de medición común. Solo conforman un banco si los ítems han sido calibrados en conjunto para la misma variable.

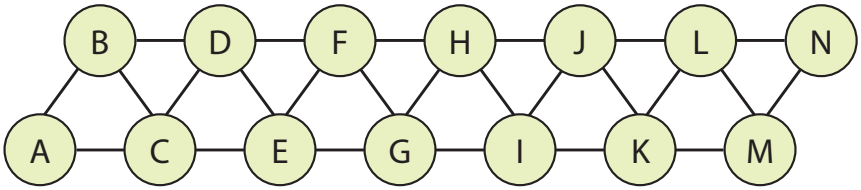
Un banco de ítems se construye calibrando conjuntamente ítems de diferentes test o mediante estudios "equiparativos" en los que los estudiantes resuelven



ítems de más de un test. Los ítems de las páginas 61 y 65 se calibraron para la misma variable usando el hecho de que algunos de los ítems de 3.er grado también fueron incluidos en el test de 5.to grado. Estos ítems comunes proveyeron el *"link"* necesario para una calibración conjunta de ambos test. En el proceso de calibración, el conjunto total de los ítems del 3.er y 5.to grado se trata como un solo test en el que solo algunos ítems (los comunes) fueron aplicados a todos los estudiantes.

Se podría calibrar más ítems para esta variable de lectura si se añaden algunos de los ítems de la página 61 en nuevos test a medida que estos se desarrollan. Los ítems del banco que ya están calibrados proporcionarían el *link* requerido para añadir más ítems a la variable (en la práctica, esto se haría estimando por separado las dificultades de todos los ítems en un nuevo test, y luego adaptando cada una de las dificultades del ítem en función de la cantidad requerida para que la dificultad promedio de los ítems comunes sea igual al promedio del banco). Este proceso se conoce como *equiparación de ítems comunes.*

Un proceso alternativo para calibrar un gran número de ítems respecto de la misma variable consiste en pedir a grupos de estudiantes que respondan a más de un test. Un grupo de estudiantes que toma dos test provee el *link* necesario para calibrar ambos con respecto a la misma variable. En el proceso de calibración, los dos test son tratados como un gran test. Este proceso se conoce como *equiparación de personas comunes* y se usó para equiparar las catorce formas del test TORCH (Test de Comprensión de Lectura, por sus siglas en inglés). El test más fácil (la forma A) se diseñó para estudiantes de 3.er grado; el test más difícil (la forma N), para estudiantes de 10.mo grado. Los test se ordenaron según la dificultad prevista y se les pidió a todos los estudiantes del estudio de equiparación que intentaran responder dos conjuntos de dificultad similar. En el diagrama de abajo, cada línea que conecta los dos test representa a un grupo de estudiantes que intenta resolver esos test.



El establecer *links* entre los catorce test de comprensión de lectura de esta forma permitió calibrar los ítems a lo largo de un continuo de habilidad creciente para la lectura. Los profesores que usan TORCH eligen un test apropiado para las habilidades de lectura reales de los estudiantes. Dado que todos los test están calibrados a lo largo de la misma variable, el desempeño en un test se puede comparar directamente con el desempeño en cualquier otro, y se puede monitorear el crecimiento en la lectura a lo largo del tiempo.

Los bancos de ítems varían en tamaño desde varias docenas hasta muchos miles. Una vez que se ha construido un banco, se puede usar como fuente de ítems calibrados para la construcción de nuevas formas de test. Cualquier combinación de los ítems calibrados seleccionados de un banco es capaz de proporcionar medidas de los estudiantes con respecto a la variable del banco. Cuando las respuestas de los estudiantes son acordes al modelo Rasch, estas

medidas son directamente comparables con medidas basadas en cualquier otra selección de ítems del banco.

Las ventajas de un banco de ítems incluyen el hecho de que no es necesario aplicar exactamente los mismos ítems a todos los estudiantes. Un conjunto de ítems relativamente sencillos puede seleccionarse y aplicarse a los estudiantes con niveles de logro relativamente bajos; un conjunto de ítems más difíciles puede aplicarse a estudiantes más hábiles, y el resultado de estos dos test se puede comparar directamente. Este tipo de medidas de los estudiantes son "objetivas" en el sentido descrito en las páginas 32-35, y su significado no depende del conocimiento de los ítems particulares que se usaron para obtenerlas.

4.4. Test adaptativo computarizado

Cuando los ítems se extraen de un banco de ítems calibrados y las respuestas de los estudiantes se condicen con el modelo Rasch, entonces es posible comparar directamente el desempeño de los estudiantes que respondieron a selecciones diferentes de ítems. En el test adaptativo computarizado, los ítems se presentan de uno en uno en una pantalla. Después de que un estudiante ha intentado responder un ítem, la habilidad del estudiante (β) se reestima sobre la base a su desempeño en ese ítem y en todos los anteriores. Se rastrea el banco de ítems automáticamente en busca de aquel cuya dificultad estimada se acerque lo más posible a la nueva estimación de la habilidad del estudiante. Se aplica este ítem y el proceso continúa. El test suele terminar cuando se alcanza un nivel específico de confianza con respecto a la habilidad del estudiante.

Un test adaptativo computarizado se adapta ítem a ítem a la persona que está dando el test y consta de ítems empatados con el nivel de habilidad de los estudiantes individuales. En un test adaptativo computarizado no existe una razón por la cual dos estudiantes tendrían que responder a un ítem común. Y, dado que todos los ítems están calibrados y han sido extraídos del

mismo banco, los resultados de los estudiantes en el test son directamente comparables, sin importar qué ítems respondieron. La ventaja de un test adaptativo computarizado es que contiene muy pocos ítems —o quizá ninguno— que sean inapropiadamente fáciles o difíciles para los estudiantes individuales.

En resumen

Cuando se calibran los ítems según una variable, estos empiezan a dar un significado a dicha variable. Son indicadores de observaciones típicas en una posición específica de la variable. Si se les toma en cuenta en su conjunto, los ítems de test o cuestionarios calibrados amplían la comprensión de la naturaleza del proceso típico de desarrollo o progreso. Forman un "mapa" para observar y monitorear el proceso de los estudiantes. A mayor número de ítems calibrados para una variable, más riqueza en la descripción e ilustración de la variable.

La clave en la calibración objetiva de ítems de test o cuestionarios se encuentra en el hecho de que, según el modelo Rasch, cuando un individuo responde dos ítems Y y Z, el ratio $P_{✓✗}$ / $P_{✗✓}$ está determinado solo por las dificultades relativas de ambos ítems:

$$\delta_Z - \delta_Y = 1n\ (P_{✓✗} / P_{✗✓})$$

Donde $P_{✓✗}$ es la probabilidad modelada de que el individuo tenga éxito en Y pero falle en Z, y $P_{✗✓}$ es la probabilidad de que el individuo falle en Y pero tenga éxito en Z. Cuando la data se condice con el modelo Rasch, es posible estimar las dificultades relativas de cualquier par de ítems Y y Z, a partir de un registro simple del desempeño del estudiante en ambos ítems, sin importar qué estudiantes estuvieron involucrados.

En la práctica, resulta esencial que se coloquen los *checks* en la medida en que las observaciones se condigan con el modelo. Solo si los estimados de

la dificultad de los ítems son estables para toda la población de estudiantes con la que se usarán, se puede medir y comparar a todos los estudiantes de la población en esa misma variable. Los *checks* en el funcionamiento diferencial del ítem indican la medida en que la variable mantiene su significado para un subgrupo particular de una población de estudiantes.

Cuando una gran cantidad de ítems se calibra con respecto a una variable, estos conforman un banco de ítems. Una ventaja de este es que permite seleccionar ítems y combinarlos en test diferentes y comparar directamente el desempeño de los estudiantes en ellos. Un test adaptativo computarizado se extrae de un banco de ítems calibrados para construir test a la medida del desempeño ítem por ítem de un individuo.

5. REPORTANDO MEDIDAS

La medición educativa empieza con la intención de estimar la ubicación de los estudiantes con respecto a alguna variable de interés. En educación nos interesamos en distintos aspectos del desarrollo del estudiante, incluyendo la habilidad para la lectura, la alfabetización científica, el respeto por otras culturas, la competencia matemática, el amor por el aprendizaje, el razonamiento lógico, la pericia en el manejo de la tecnología y las habilidades interpersonales. Cada intento por medir es un intento por conocer el nivel actual de logro de los estudiantes en algún aspecto de su desarrollo.

Se diseñan los instrumentos de medición —test y cuestionarios— para ofrecer observaciones que se pueden usar para estimar los niveles de logro. Pero los instrumentos de medición particulares nunca son importantes en sí mismos: cualquier ítem del test se puede reemplazar por uno o varios ítems igual de apropiados, y cualquier test se puede reemplazar por una selección alternativa de ítems.

En educación tenemos la intención de que nuestras medidas tengan una generalidad que vaya más allá del conjunto particular de ítems que se usó para obtenerlas. Nos interesa conocer el desempeño de un estudiante en una selección particular de ítems solo en la medida en que indica la posición del estudiante en la variable subyacente para cuya medición se diseñó el test. Esta intención es común a todo tipo de medición. Por ejemplo, cuando usamos una balanza, contamos con que la que usemos en particular sea irrelevante con respecto al resultado, y que la medida de nuestro peso sea expresada en una métrica que no sea peculiar de ese instrumento.

Si se van a comparar las medidas entre instrumentos de forma significativa (por ejemplo, dos test de lectura), y se van a usar para medir el cambio y monitorear el crecimiento, entonces se deben reportar en escalas de medición que no estén atadas a un instrumento en particular.

5.1. Interpretando medidas

En la medición educativa, nuestro interés principal al interpretar y reportar el logro de los estudiantes se suele centrar en el conocimiento, las habilidades, la comprensión, las actitudes o los valores que los estudiantes han adquirido. Puede que también estemos interesados en comparar los niveles del logro de unos estudiantes con los de otros (por ejemplo, estudiantes de la misma edad o el mismo grado, estudiantes en otros estados o países) o en conocer cuánto han progresado los estudiantes desde un momento anterior. Pero para muchas finalidades —sobre todo para la instrucción— nuestro interés principal es saber cómo están progresando los estudiantes con relación a algún continuo del desarrollo del conocimiento, habilidades, comprensión, actitudes o valores.

Para interpretar medidas educativas de esta manera, es necesario primero dar un significado sustantivo a la variable que se está midiendo, distinguiendo los tipos de observaciones que se suelen hacer en distintas ubicaciones a lo largo de la variable. En las páginas 75 y 76 se muestra un ejemplo de este tipo de mapeo. Los números en el lado izquierdo de la página indican niveles crecientes en la habilidad para la lectura tal como los define el Marco Lexile para la Lectura *(Lexile Framework for Reading)*.[1] Los títulos literarios son ejemplos de libros en diferentes niveles de dificultad para la lectura. El libro más fácil es *Ronald Morgan va a batear* (200 lexiles); el más difícil es *Jonathan Livingston Seagull* (900 lexiles). Un estudiante con un nivel de, por decir, 880, debería estar en condiciones de leer un texto en ese nivel (por ejemplo *El pony colorado*) con un nivel de comprensión del 75%. En el lado derecho de la página se encuentran ejemplos de textos de nivel creciente en este continuo.

El mapa de progreso de las páginas 75 y 76 permite interpretar las habilidades para la lectura de los estudiantes (medidas en la escala Lexile) en relación con

[1] Stenner, Jackson. (1996). *Measuring reading comprehension with the lexile framework.* Artículo presentado en la Fourth North American Conference on Adolescent/Adult Literacy, Washington (EE.UU.).

los tipos de textos que es probable que puedan leer y comprender, y sugerir libros que podrían ser apropiados para niveles específicos de habilidad para la lectura (el que un libro sea apropiado o no para un estudiante en particular depende, claro está, no solo de su nivel de dificultad, sino también de sus contenidos y el nivel de interés para ese estudiante).

Cualquier variable de medición en educación puede trazarse en un mapa e ilustrarse con ejemplos del tipo de habilidades, respuestas y comportamientos que caracterizan los niveles de desarrollo a lo largo de esa variable. Las descripciones e ilustraciones otorgan un significado sustantivo a una variable y esclarecen la naturaleza del crecimiento en el área que se está midiendo. Los mapas de progreso de este tipo también se conocen como "escalas descritas de competencia" y el proceso de interpretar el nivel de lo logro de los estudiantes con respecto a estos mapas como "hacer referencias estándar".

En la construcción de variables de medición, es común definir niveles de logro amplios y describir e ilustrar observaciones típicas para cada nivel. La escala de competencia en Cívica en la página 80 se construyó a partir de un análisis del desempeño de estudiantes estadounidenses en la Evaluación Nacional del Progreso en Educación (National Assessment of Educational Progress). La escala numérica en el lado izquierdo de la página está dividida en cuatro niveles amplios, y los tipos de conocimiento y comprensión típicos de los estudiantes se han descrito para cada nivel.[2] Tal como en el ejemplo Lexile, estos niveles descritos proveen un marco de referencia para interpretar las medidas de logro.

5.2. Estándares de desempeño

Así como proveen marcos de referencia para reportar y describir los niveles actuales de logro de los estudiantes y grafican el progreso a lo largo del

[2] Anderson, Lee et ál. (1990). *The Civics Report Card: Trends in Achievement from 1976 to 1988 at Ages 13 and 17*. Princeton: Education Testing Service.

Títulos de obras literarias	Ejemplo de texto
990 *Jonathan Livingston Seagull* **980** *Paternidad* **960** *Las aventuras de Tom Sawyer* **960** *Instrucciones del juego Pictionary* **920** *Matar a un ruiseñor* **920** *El León, la bruja y el armario* **900**	Discutí la cuestión en todas sus formas, política y científicamente, y aquí presento un extracto de un artículo trabajado con cuidado que publiqué en el número del 30 de abril. Decía lo siguiente: "Después de examinar una a una las distintas hipótesis, rechazando cualquier otra sugerencia, se torna necesario admitir la existencia de un animal marino de enorme poder. Las grandes profundidades del océano son completamente desconocidas para nosotros. Los sondeos no pueden alcanzarlas…".
890 *Stuart Little* **880** *El pony colorado* **870** *Un sabor a moras* **830** *Sounder* **810** *La señora Frisby y las ratas de NIMH* **810** *Johnny Appleseed* **800**	Era más alto que una gran hoja de guadaña y de un lavanda muy pálido sobre el agua azul oscuro. Se barría hacia atrás y cuando el pez justo nadaba bajo la superficie el viejo pudo ver su enorme masa y las franjas moradas que lo atravesaban. Su aleta dorsal miraba hacia abajo y sus pectorales enormes estaban extendidos. En este círculo el viejo pudo ver el ojo del pez y los dos peces succionadores grises que nadaban a su alrededor. A veces se pegaban a él. A veces se precipitaban lejos.
780 *Manual del Boy Scout* **780** *La casa de la pradera* **770** *Un grillo en Time Square* **730** *Harriet, la espía* **710** *Desaparecida* **700** *Donde crece el helecho rojo* **700**	Templeton, claro está, se sentía miserable con respecto a la pérdida de su huevo amado. Pero no podía resistir alardear. "Vale la pena guardar cosas", decía con voz arisca. "Una rata nunca sabe cuándo algo le podría resultar útil. Nunca boto nada". "Bueno", le dijo una de las ovejas, "todo este asunto es bueno para Charlotte, ¿pero qué hay del resto de nosotros? El olor es insufrible. ¿Quién querría vivir en un granero perfumado de huevos podridos?". "No te preocupes, te acostumbrarás," dijo Templeton.
690 *Cómo comer gusanos fritos* **670** *Fiebre de chocolate* **650** *En los bancos de Plum Creek* **640** *Hardy Boys Submarine Caper* **620** *Jack y Jill* **610** *Flossie y el zorro* **600**	No sabía cómo el mundo ha sido simplificado para los reyes. Para ellos, todos los hombres son súbditos. "Acércate, para que te pueda ver mejor", dijo el rey que se sentía sumamente orgulloso de por fin ser el rey de alguien. El principito miró hacia todos lados buscando un sitio donde sentarse, pero todo el planeta estaba copado y bloqueado por el magnífico manto del rey. Así que se quedó y, ya que estaba cansado, bostezó. "Es contrario a la etiqueta bostezar en presencia del rey", dijo el monarca.
580 *El niño que pagaba el pato* **560** *Algo sobre el campo de juego* **560** *El rescate de Madeline* **550** *Los chicos del vagón de carga* **540** *Sarah, sencilla y alta* **530** *Hay un chico en el baño de las chicas* **500**	"Aar." Encyclopedia respondió tras un momento. Siempre esperaba un momento. Quería ser útil. Pero le daba miedo que a la gente no le gustase si respondía sus preguntas muy rápido y sonaba muy inteligente. Su papá le hacía más preguntas que cualquier otra persona. El Sr. Brown era el jefe de la policía de Idaville. El pueblo tenía cuatro bancos, tres cines y una pequeña liga. Tenía el número usual de grifos de gasolina, iglesias, escuelas, tiendas y casas cómodas en calles sombreadas.

490 *Comandante Sapo* **480** *George, el curioso* **450** *Un cocodrilo debajo de mi cama* **450** *Sophie y Gussie* **440** *Algo extraño está sucediendo* **430** *Lejanía* **400**	El sábado siguiente mi mamá me llevó a la autopista para tomar el bus a Nueva York. Era la primera vez que iba sola y mi mamá estaba nerviosa. "Escucha, Margaret: no te sientes al lado dTe ningún hombre. Siéntate o sola o elige alguna señora simpática. Y trata de sentarte adelante. Si no hay aire acondicionado, abre tu ventana Y cuando llegues, pídele a una señora que te indique el camino para bajar las escaleras. La abuela se encontrará contigo en la oficina de información". "Ya sé, ya sé". Lo habíamos conversado tres docenas de veces…
380 *Historias de 4to grado* **370** *¿Dónde está el gato?* **350** *Conejito* **320** *Las aventuras del Sr. Rana* **310** *Quique duerme fuera de casa* **300** *Mog, el gato olvidadizo* **300**	Cuando Oso llegó a casa, sacó toda la plata de su alcancía. Luego se fue al centro de la ciudad… y le compró un sombrero hermoso a la luna. Esa noche puso el sombrero en la cima de un árbol para que la luna lo encontrase. Luego esperó y vio cómo la luna se deslizaba lentamente por el árbol hasta llegar a las ramas y se probaba el sombrero. "¡Hurra!" gritó Oso. "Le queda perfecto". Durante la noche, mientras el oso dormía, el sombrero se cayó del árbol. En la mañana Oso encontró el sombrero en la puerta de su casa. "¡Así que la luna también me consiguió un sombrero!", exclamó Oso.
290 *El cocodrilo de Zack* **270** *Bingo, el mejor perro del mundo* **260** *Un Pez, dos peces, pez rojo, pez azul* **220** *Tren de carga* **210** *Sapo y Sepo, un año entero* **200** *Ronald Morgan va a batear* **200**	En el gran cuarto verde había un teléfono y un globo rojo. Y una foto de la vaca saltando sobre la luna. Y había tres ositos sentados en unas sillas. Y dos gatitos y un par de mitones. Y una casita de juguete y un ratón joven. Y un peine y un cepillo y un tazón lleno de papilla. Y una señora mayor muy tranquila que susurraba "cállate". Buenas noches, cuarto. Buenas noches, luna. Buenas noches, vaca saltando sobre la luna. Buenas noches, luz y globo rojo. Buenas noches, osos. Buenas noches, sillas. Buenas noches, gatitos.

tiempo, variables de medición como las ilustradas en las páginas 75-76 y 80 también se usan para definir expectativas o metas con respecto al desempeño de los estudiantes. Por ejemplo, el mapa de la página 75-76 se puede usar para identificar el nivel de habilidad para la lectura que resulta razonable esperar de todos los estudiantes al terminar el 4.to grado. El mapa de progreso del desarrollo del conocimiento de Cívica que se muestra en la página 80 puede resultar útil al pensar sobre el nivel de conocimiento de Cívica que se debe establecer como meta para todos los estudiantes de 8.vo grado. Las expectativas o las metas con respecto al logro de los estudiantes también se conocen como "estándares de desempeño".

Un estándar de desempeño identifica los tipos de habilidades o de comprensión esperados en los estudiantes y, operacionalmente, asume la forma de un puntaje

mínimo que se debe lograr para considerar que un estudiante ha alcanzado el estándar (por ejemplo, 270 en la escala Lexile; 320 en la escala de Cívica). Los puntajes mínimos, también conocidos como puntajes de corte, se determinan mediante un proceso de "definición de estándares" en el cual se decide, ítem por ítem, qué desempeño es probable en un estudiante que apenas cumple el estándar. Por ejemplo, para definir el puntaje para pasar una evaluación de último año de la carrera de Odontología, los expertos en este campo pueden decidir cuál es la probabilidad de que un dentista mínimamente competente responda cada uno de los ítems de la evaluación.

Cuando se establece un estándar de desempeño, existe un interés especial no solo en conocer dónde se ubican los estudiantes con respecto al continuo de logro subyacente, sino también en conocer dónde se ubican con respecto a un punto determinado (puntaje de corte) de ese continuo. En algunos contextos, como por ejemplo en las evaluaciones de fin de carrera, esta pregunta puede ser de interés fundamental al interpretar los resultados.

5.3. Reportando el crecimiento

Una variable de medición también ofrece un marco de referencia para monitorear y reportar el crecimiento a lo largo del tiempo. Al medir en varias ocasiones el nivel de logro de un individuo con respecto a una variable, es posible hacer seguimiento al desarrollo de ese individuo a lo largo del tiempo, graficar su trayectoria de crecimiento, y evaluar la mejora entre una ocasión y la siguiente. La interpretación del nivel del logro actual de un estudiante que hace referencia al logro de ese estudiante en una ocasión previa también se conoce como "referencia ipsativa".

En estudios longitudinales del logro de los estudiantes, se sigue a los mismos individuos a lo largo de varios años. En estos estudios se suele recoger no solo medidas de logro, sino también información sobre la historia y las experiencias educativas de los estudiantes, su contexto familiar y actividades extracurriculares relevantes. Se hace el intento de comprender los factores

que influyen sobre el aprendizaje a lo largo de varios años de escolaridad. Los estudios longitudinales dependen de la posibilidad de generar y comparar medidas de logro con respecto a la(s) misma(s) variable(s) durante un rango amplio de edad.

También es posible medir y comparar los logros de un *grupo* de estudiantes en diferentes ocasiones. Cuando se sigue el progreso de un grupo, se pueden hacer preguntas con respecto a las tasas de crecimiento promedio o típicas. El *Third International Mathematics and Science Study* (TIMSS), por ejemplo, mide el logro en matemáticas y ciencias de estudiantes de 4.to grado de cada país participante y, cuatro años más tarde, mide el logro de los estudiantes de 8.vo grado en esos países. De esta manera fue posible medir, en cada país, el crecimiento promedio o típico en el logro en matemáticas y ciencias a lo largo de cuatro años.

5.4. Comparando logros

También se pueden interpretar las medidas del logro de los estudiantes al compararlas con los logros de otros estudiantes. El proceso de comparar las medidas de un estudiante con las medidas de otro se conoce como "hacer referencias normativas".

Una medida se interpreta de forma normativa cada vez que se compara con el desempeño de otros. Observaciones como que un estudiante ha obtenido el puntaje más alto de su clase, ha tenido un desempeño en el 10% superior de los estudiantes del Estado, tiene una edad de lectura de 6.2 y muestra un logro en el percentil 85 de su grupo etario en el país, son ejemplos de interpretaciones del logro en referencia a la norma.

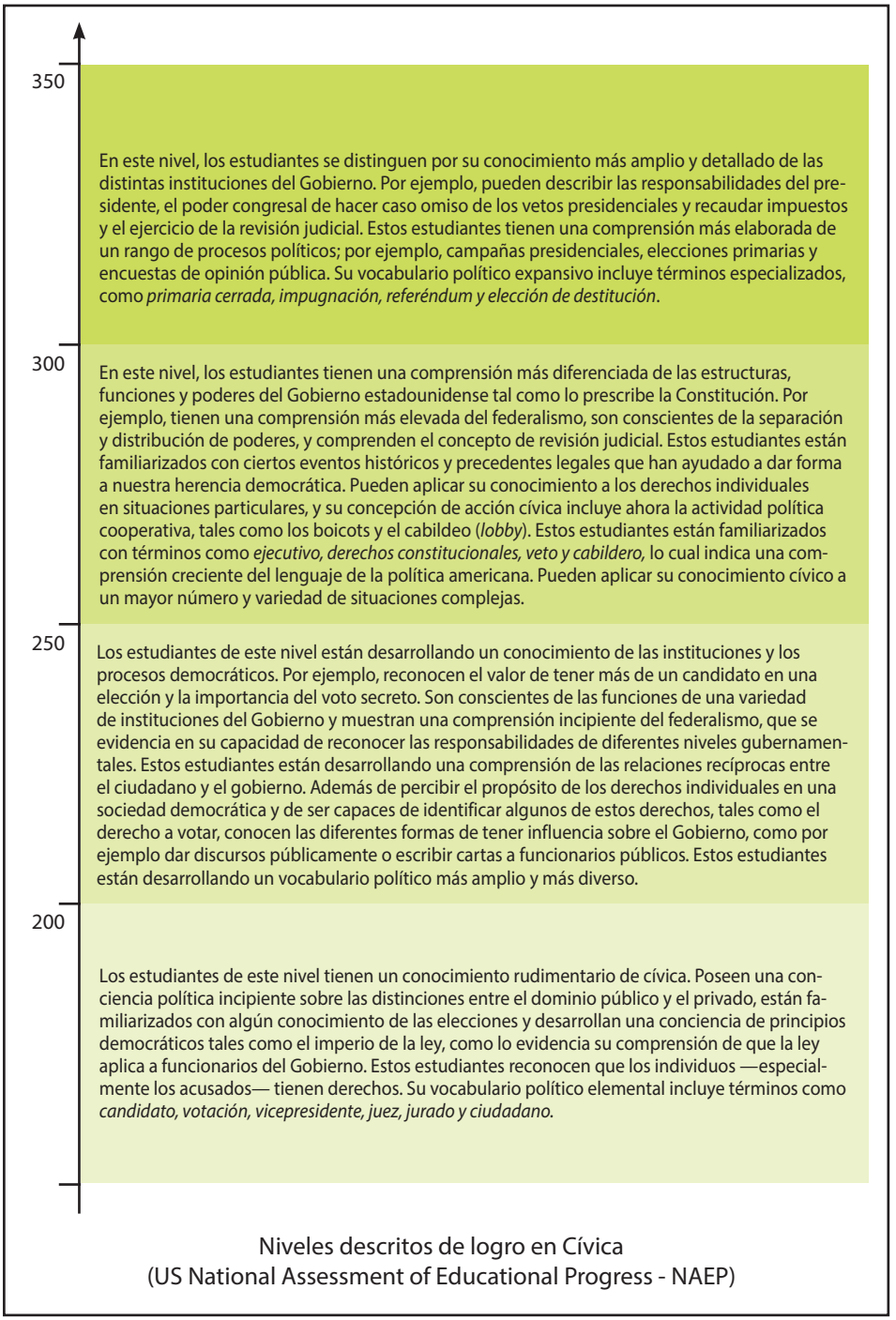
Si se interpreta el logro de un estudiante en comparación con los logros de otros estudiantes, entonces es importante aclarar cuál es la naturaleza del grupo con el que se compara. ¿Se trata de todos los estudiantes de diez años del estado o provincia? ¿De todos los estudiantes de diez años del país? ¿De todos los

estudiantes de 5.to grado del estado o provincia? ¿Todos los estudiantes de 5.to grado del país?

Algunos programas de evaluación miden los logros de todos los estudiantes de un sistema educativo en algunos grados en particular. Estos programas de evaluación de "cohorte completo" o de "población" permiten comparar el desempeño de los estudiantes —o el desempeño promedio de una clase o escuela— con los logros de todos los estudiantes de ese mismo grado en todo el sistema educativo.

Pero las medidas no siempre están disponibles para todos los estudiantes de la población que se desea comparar, lo cual requiere que se haga inferencias para una población a partir de una muestra de estudiantes seleccionada con cuidado. La extracción y evaluación de muestras de estudiantes es una práctica común en los estudios internacionales de logro tales como el Programa Internacional de Evaluación de Estudiantes (PISA, por sus siglas en inglés) y el Tercer Estudio Internacional de Matemáticas y Ciencias (TIMSS, por sus siglas en inglés), y en estudios nacionales de logro como la Evaluación Nacional del Progreso en el Aprendizaje (NAEP, por sus siglas en inglés). Al medir los logros de muestras representativas de estudiantes, estos programas proveen información sobre los logros de cohortes nacionales de estudiantes. Un estado o una escuela que elige usar el material de evaluación de estos programas podrá luego comparar el desempeño individual o grupal con normas nacionales e internacionales.

Un ejemplo de un estudio nacional de este tipo fue el *Australian National School English Literacy Survey* (Estudio Nacional Australiano de Alfabetización Escolar), que midió los logros en alfabetización de muestras nacionales cuidadosamente seleccionadas de estudiantes de 3.er y 5.to años. Los logros de lectura medidos para estas muestras nacionales se muestran en la página 82. Cada una de las barras de este gráfico corresponde a un puntaje en el test de lectura de 3.er y 5.to años, y se muestra el porcentaje aproximado de los estudiantes para cada barra.



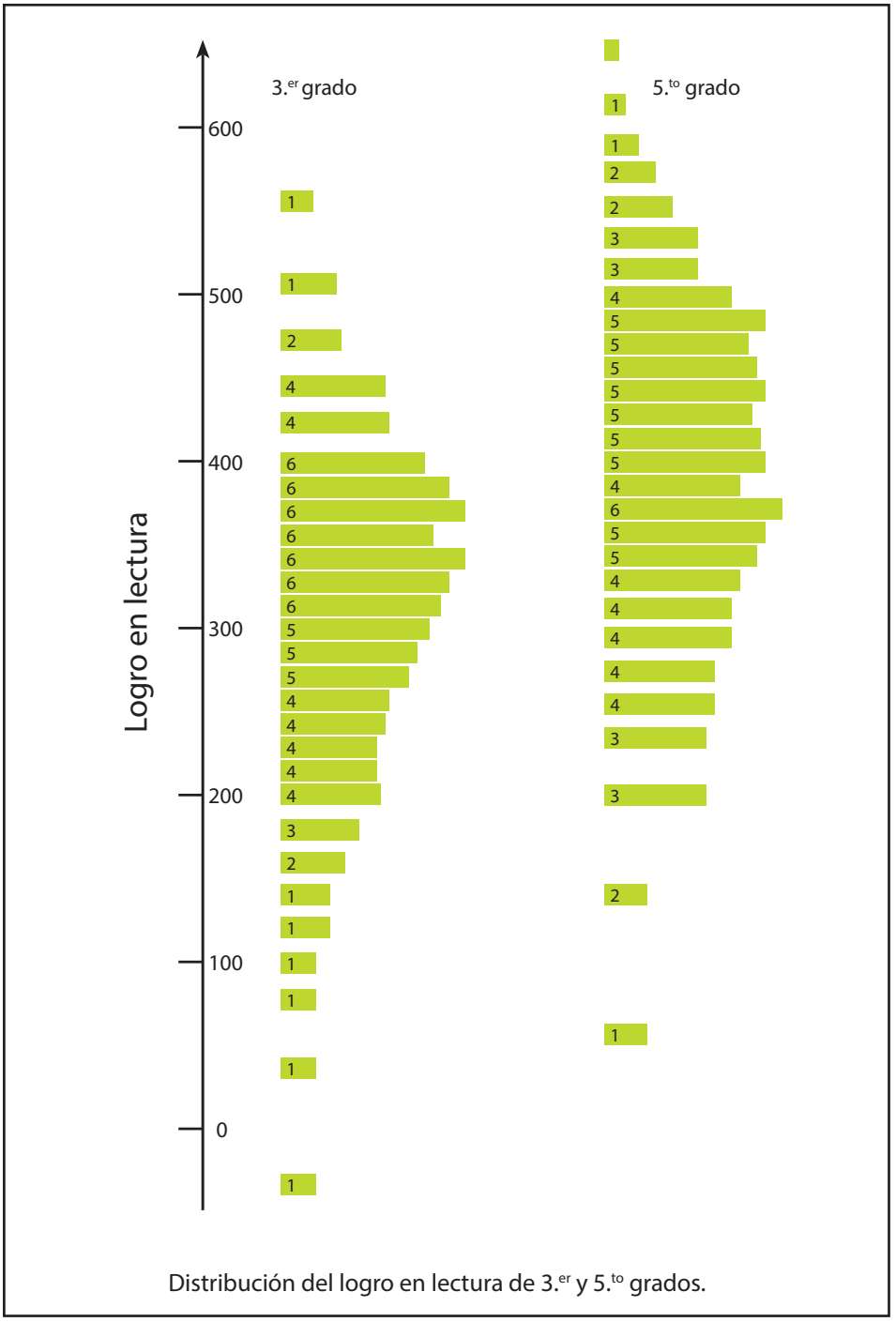
Niveles descritos de logro en Cívica
(US National Assessment of Educational Progress - NAEP)

La variable de lectura en la que se midió a los estudiantes también aparece en la página 65. Al referirse a esta página, puede interpretarse el nivel de logro en lectura medido para un estudiante en términos del tipo de conductas de lectura que es probable en él. Al referirse a la página 65, es posible interpretar esa misma medida en términos de los logros de lectura de los estudiantes australianos de 3.er y 5.to grados. Por ejemplo, en la página 65 se observa que, en el caso de un estudiante con una medida de lectura de 200 en esta escala, resultaría típico que estuviese en condiciones de usar combinaciones de imágenes y textos para manifestar algo de comprensión (por ejemplo, usar el título de un libro y la ilustración de la carátula para identificar los elementos clave de la historia, interpretar una imagen para predecir qué sucederá a continuación en una historia, usar el título y la ilustración para predecir el escenario de la historia o reconocer cómo los elementos de una ilustración dan soporte al texto en una historia). En la página 65 se observa que el 89% de los estudiantes de 3.er grado y el 97% de los de 5.to grado tuvieron un desempeño por encima de este nivel.

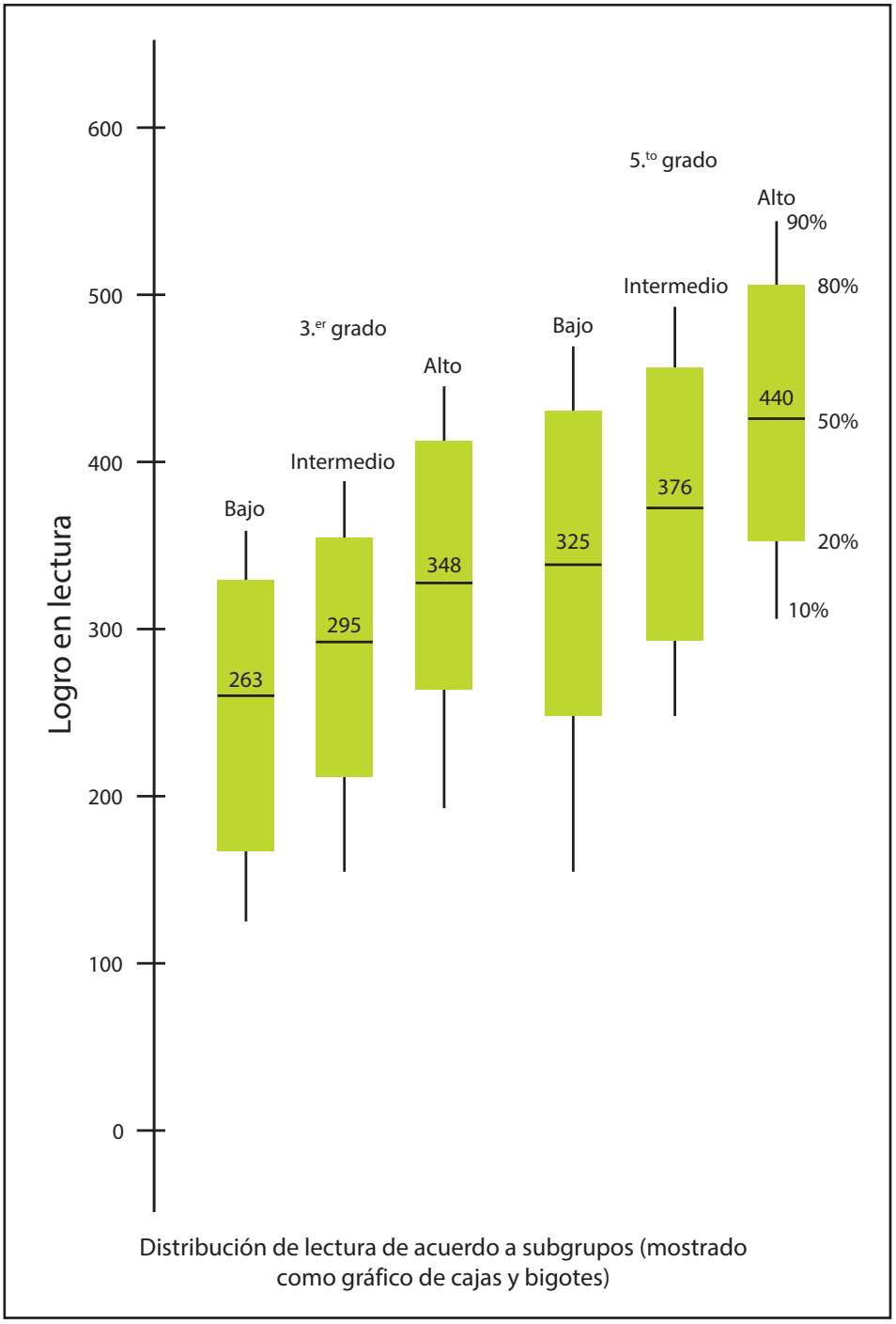
Los editores de las evaluaciones comerciales suelen proporcionar normas de evaluación que permiten a los usuarios comparar el desempeño en un test con el desempeño de estudiantes de esa misma edad o de ese mismo grado. Las normas de los test nos presentan el porcentaje de estudiantes de una muestra normativa que logra cada uno de los puntajes del test.

5.5. Comparando subgrupos

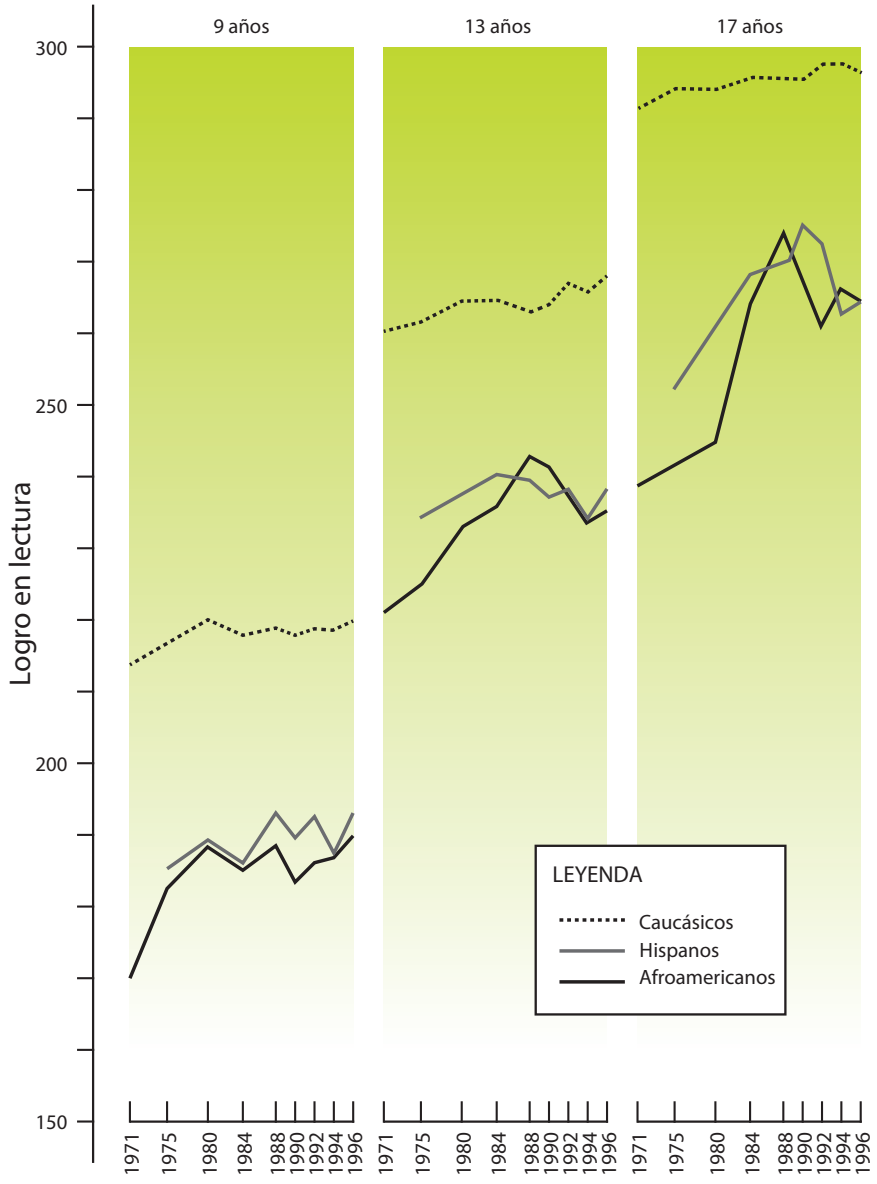
Cuando se mide una cantidad lo suficientemente grande de estudiantes con respecto a la misma variable, es posible comparar y reportar el desempeño de subgrupos de estudiantes en dicha variable. En la página 83 se ilustra la comparación de subgrupos de estudiantes, donde se ha agrupado a los estudiantes de 3.er y 5.to grado según su estatus socioeconómico en función de la ocupación de sus padres. Se construyeron cinco categorías socioeconómicas para cada nivel del año: se muestran la más baja (trabajos manuales y no capacitados), la intermedia y la más alta (trabajos profesionales y gerenciales).



Distribución del logro en lectura de 3.er y 5.to grados.



Distribución de lectura de acuerdo a subgrupos (mostrado como gráfico de cajas y bigotes)



Tendencias en el logro en lectura en los Estados Unidos 1971-1996

Los diagramas de caja y bigotes se construyeron para mostrar la mediana, el 60% y el 80% central de los estudiantes de cada subgrupo.

En los seis diagramas de caja y bigotes se puede observar que la diferencia entre los logros de lectura entre los grupos socioeconómicos más altos y más bajos es mayor para el 5.to grado que para el 3.er grado. También se puede ver que, en promedio, los estudiantes de 3.er año en el grupo socioeconómico más alto tienen niveles de lectura más altos que los estudiantes de 5.to grado del grupo más bajo. En programas de evaluación estatales y en estudios nacionales e internacionales se suelen identificar subgrupos de estudiantes y comparar y reportar sus logros.

5.6. Monitorear tendencias a lo largo del tiempo

La construcción y conservación de variables de medición resulta esencial en los intentos por monitorear tendencias en el logro educativo a lo largo del tiempo. Dado que no suele ser deseable aplicar el mismo test a los mismos estudiantes en ocasiones diferentes, o incluso administrar el mismo test a diferentes grupos de estudiantes año tras año, los intentos por monitorear tendencias a lo largo del tiempo dependen de la calibración de distintos test para una variable común.

Un esfuerzo nacional mayor por monitorear tendencias en el logro educativo a lo largo del tiempo ha sido el NAEP. En la página 84 se muestran resultados seleccionados del NAEP Lectura. Los logros de lectura promedio en las muestras de estudiantes caucásicos, afroamericanos e hispanos de 9, 13 y 17 años durante el período de 1971 a 1996. En estos veinticinco años, todos los test de lectura se calibraron según la misma variable de lectura, lo que permitió comparar los niveles de lectura de estos estudiantes, así como graficar y analizar las tendencias en el logro de la lectura en este período.

En el gráfico de la página 84 se puede ver que los estudiantes caucásicos tuvieron niveles de logro significativamente mayores que los estudiantes

afroamericanos o hispanos a lo largo de un período de veinticinco años en el caso de las tres edades. Sin embargo, mientras que durante este período no hubo un cambio significativo en los niveles de lectura de los estudiantes caucásicos, los niveles de lectura promedio de los estudiantes afroamericanos e hispanos crecieron significativamente entre los años 1971 y 1996. Esto rige para las tres edades. Aun así, una mirada más detallada muestra que, si bien hubo mejoras en el logro de lectura de los estudiantes afroamericanos e hispanos entre 1971 y fines de la década de 1980, parece no ser así durante la década de 1990.

Muchos programas de evaluación de gran escala, incluyendo algunos estudios nacionales e internacionales, proporcionan a los responsables de las políticas y a administradores públicos información sobre tendencias en el logro educativo. Un prerrequisito para estudiar tendencias es la construcción cuidadosa de la variable de medición cuyo crecimiento y decrecimiento se monitoreará.

En resumen

La medición educativa es un proceso que estima las ubicaciones de los estudiantes con respecto a alguna variable de interés. Una vez que se han tomado las medidas de una variable educativa, se puede hacer una serie de preguntas sobre la medida de cualquier estudiante.

¿Qué clase de conocimientos, habilidades, comprensión, actitudes o valores indica la medida? En otras palabras, ¿dónde se ubica el estudiante con respecto a la variable de interés y qué se puede concluir sobre su nivel de logro actual? Esta pregunta se puede responder al referirse a observaciones típicas en distintas ubicaciones a lo largo de la variable (por ejemplo, los tipos de tarea que es probable que el estudiante pueda completar y los tipos de respuestas que es probable que dé).

¿El estudiante muestra un desempeño acorde con el nivel esperado para los

estudiantes de su edad o grado? En otras palabras, ¿dónde se ubica el estudiante en relación con un estándar de desempeño previamente especificado? Esta pregunta se puede responder al establecer una relación entre las medidas del estudiante y el nivel de logro esperado o elegido como meta para esa edad o ese grado, y al decidir si se encuentra significativamente por encima o por debajo de ese nivel.

¿Cuál ha sido el progreso del estudiante desde la última evaluación? En otras palabras, ¿qué crecimiento se ha dado? Esta pregunta se puede responder midiendo el logro de un estudiante en varias ocasiones y monitoreando la mejora a lo largo del tiempo.

¿Qué sucede en la comparación entre el logro de un estudiante y el logro de otros estudiantes de la misma edad o el mismo grado? En otras palabras, ¿cómo está con respecto a normas de edad o grado? Esta pregunta se puede responder mediante la relación entre las medidas del estudiante y la distribución de medidas para el grupo norma (referencia).

Se pueden hacer preguntas similares para grupos enteros de estudiantes. Por ejemplo: ¿qué tipos de textos puede leer y comprender un niño de seis años promedio? ¿Qué porcentaje de los estudiantes de 5.^{to} grado alcanzan los estándares esperados de desempeño? ¿Cuáles son las tasas típicas en el desarrollo de la psicomotricidad fina en niños de tres años? ¿Qué sucede en la comparación entre el desempeño matemático nacional y los *benchmarks* internacionales? Las respuestas a preguntas de este tipo son esenciales para una toma de decisiones informada en educación y dependen de la disponibilidad de mediciones confiables basadas en variables construidas cuidadosamente.

¿Qué es la medición?

La medición es la ubicación de objetos respecto de las variables según la experiencia. Empezamos la medición con la idea de una variable. Esta idea se puede visualizar como una línea que apunta en la dirección hacia donde hay "más". Damos un significado explícito a una variable al especificar los tipos de preguntas (observaciones, ítems de un test) con los que esperamos definirlo. Testeamos la validez de estas preguntas al exponerlas a la experiencia y descubrir si hay condiciones útiles bajo las cuales dibujen una línea. Hacemos que el significado de la variable sea operacional al estimar (calibrar) las posiciones relativas de las preguntas válidas sobre esta línea y etiquetamos la línea en consecuencia. Si podemos inventar preguntas que mantengan su ubicación sobre la línea en un rango útil de aplicaciones, entonces tendremos una variable con una definición operacional a lo largo de la cual se puede intentar hacer mediciones objetivas.

Hacemos mediciones en esta variable al aplicar una selección apropiada de preguntas a un objeto (persona) que queremos medir, observando qué pasa y estimando del patrón de reacciones (respuestas) la ubicación probable del objeto entre las preguntas ordenadas. Evaluamos la validez de esta medición al comparar el patrón de respuesta observado con su expectativa estimada para ver si se puede aceptar el patrón como un resultado plausible del sistema de medición que hemos definido.

Benjamin D. Wright

Universidad de Chicago

30 de marzo de 1979

Medir en educación
Recursos de evaluación del Consejo Australiano para la Investigación Educativa 1

Durante la elaboración de estándares de aprendizaje en que estuvo abocado el SINEACE entre los años 2009-2015, se conoció la experiencia australiana de evaluación de los aprendizajes realizada por el Consejo Australiano para la Investigación Educativa (ACER). El SINEACE suscribió un convenio con el ACER para traducir los folletos que comprenden el Kit de Recursos de Evaluación, que fueron un valioso material durante el proceso; esto permite poner al alcance de los lectores de habla castellana el primer número de la serie, el cual rescata la importancia de la evaluación en el contexto educativo.

Si bien este libro está basado en otra realidad, muestra que la evaluación educativa es universal y pertinente para diversas realidades y culturas: proporciona información acerca del proceso de enseñanza-aprendizaje; permite investigar formas de mejorar el aprendizaje de los estudiantes; asimismo, permite hacer seguimiento del rendimiento de los estudiantes a través de los años, y diseñar programas que contribuyan al desarrollo del sector.

SERIE
DOCUMENTOS TÉCNICOS

Sineace
SISTEMA NACIONAL DE EVALUACIÓN, ACREDITACIÓN Y CERTIFICACIÓN DE LA CALIDAD EDUCATIVA
PERÚ Ministerio de Educación

ISBN: 978-612-4322-29-7
9 786124 322297