La escala de calificación del TERCE

Al igual que otras evaluaciones nacionales e internacionales, el TERCE (2013) reporta los resultados de sus pruebas de aprendizaje a través de dos mecanismos centrales. El primero consiste en una escala de puntaje, y el segundo en la distribución de estudiantes según niveles de desempeño.

Antecedentes SERCE

La escala de puntaje utilizada en el Segundo Estudio Regional Comparativo y Explicativo (SERCE, 2006) tuvo una media de 500 puntos y una desviación estándar de 100. En dicha ocasión, el cálculo de puntajes se realizó utilizando la Teoría de Respuesta a Ítem además de otros criterios específicos que se describen más abajo.

El TERCE, como parte del ciclo de crecimiento cualquier estudio de gran escala, se construyó innovando en aspectos metodológicos del proceso de construcción de las pruebas y de la determinación de puntajes, lo que sumado a las diferencias en los criterios de elegibilidad de los examinados, determinó que la comparación con SERCE tuviera que hacerse como un proceso distinto y paralelo al de la estimación de puntajes del TERCE.

Así, llegado el TERCE surgió el desafío de identificar un mecanismo para facilitar la comparabilidad de los resultados obtenidos por las entidades participantes en ambas mediciones.

En relación con lo anterior, vale la pena destacar dos puntos. En primer lugar, la fijación de una escala en un valor promedio determinado con un respectivo valor de desviación estándar no tiene ninguna implicancia en los resultados u ordenamiento de los países. El proceso de "escalamiento" consiste en transformar los *logits* que arrojan los modelos de medición—que tienen promedio cero y desviación estándar igual a uno—en un valor arbitrario que sea más adecuado para la comprensión del público en general[1]. Así, la fórmula general para la transformación de la escala de logits a la escala deseada se hace de la siguiente forma:

$$Puntaje = Promedio\ arbitrario\ de\ la\ escala + (logit\ de\ cada\ estudiante * desviación\ estándar\ arbitraria)$$

[1] Los logits, una vez ajustado el modelo IRT, no necesariamente tienen un promedio cero y desviación estándar uno, por lo que antes de aplicar la transformación indicada en la fórmula, se los estandariza para que el promedio sea cero y la desviación estándar uno.

Como se desprende de la fórmula anterior, esta transformación lineal respeta el ordenamiento de los datos, por lo que no puede afectar la posición relativa de estudiantes ni países.

En segundo lugar, los criterios de inclusión/exclusión de las evaluaciones (*ver nota 5*) son claves para la rigurosidad de éstas. Los avances científicos en los ámbitos de la medición y el muestreo han permitido la **utilización de criterios más inclusivos** para maximizar el uso de la información en pos de obtener estimaciones más precisas del logro de los estudiantes. Es probable que un cambio en los criterios de exclusión afecte marginalmente la posición relativa de los estudiantes o algún dato agregado, sin embargo, cuando se mejoran estos criterios se obtienen estimaciones más fidedignas del logro.

Entregas de resultados del TERCE

Considerando la riqueza de información que entrega el TERCE, así como también los diferentes focos de los análisis de resultados, se planteó una estrategia continua de difusión de los hallazgos del estudio, cuyo hito inicial tuvo lugar en diciembre de 2014 con la primera entrega de resultados. Dicha entrega tuvo como foco la comparación entre los resultados de SERCE y TERCE, por lo que se emplearon criterios de selección de casos y metodologías de análisis que garantizaran tal comparabilidad, tal como se explica en detalle más adelante. Adicionalmente se comparó la distribución de niveles de desempeño, lo que se basó en la misma metodología empleada en SERCE para identificar puntos de corte.

En cuanto a la segunda etapa de difusión, que comenzó en julio de 2015, su objetivo es entregar una estimación de puntajes basada en una nueva metodología (valores plausibles y criterios de exclusión actualizados) e identificar el rol e importancia relativa de diversos factores asociados. La escala en que estos puntajes se reportan tiene un promedio de 700 puntos con una desviación estándar de 100.

Procedimientos para posibilitar la comparabilidad SERCE-TERCE

Para realizar la comparación de puntuaciones entre SERCE y TERCE fue necesario recurrir a una metodología que hiciera posible dicho ejercicio. Las metodologías comúnmente utilizadas por los programas internacionales de medición para este tipo de comparaciones se basan en el empleo de bloques de preguntas comunes entre ambas mediciones. Estas preguntas comunes se denominan *bloques ancla* y sirven para asegurar que la base de

comparación sea compartida, lo que permite, adicionalmente, situar las preguntas de los bloques no comunes en una escala comparable de puntuaciones.

Para obtener estimaciones de puntajes comparables, se comenzó por establecer una base de datos de TERCE que empleara los mismos criterios de exclusión que se habían establecido en el SERCE, y luego se utilizó el mismo software de análisis métrico (*Winsteps*). Para asegurar que las puntuaciones de TERCE se expresaran en una escala comparable con SERCE, se instruyó al programa para que dejara constantes los parámetros de los ítems establecidos en el SERCE (incluyendo los de los ítems de anclaje). En estas condiciones el programa estimó los parámetros de los nuevos ítems de TERCE (los que no pertenecen a los bloques de anclaje) y con esa información se calcularon los puntajes de los estudiantes participantes en TERCE.

Una vez obtenidos los resultados TERCE equiparados en escala SERCE, se procedió a calcular el error estándar de cada país, y de la región, usando el método de Linealización de Taylor (mismo método reportado en SERCE). Dicha información, más los errores que se habían estimado en el SERCE y los errores de linking (estimados a partir de los ítems comunes), fueron empleados como insumos para determinar en qué países se había producido un cambio estadísticamente significativo de sus puntuaciones medias entre 2006 y 2013, y en qué dirección se producían estos cambios.

Avances en medición de aprendizajes y modificaciones implementadas en TERCE

La evaluación de aprendizajes ha experimentado importantes cambios durante la última década, perfeccionándose sus técnicas de recolección de datos, de medición, y también sus análisis. Siguiendo estas innovaciones, fue necesario que el TERCE adaptara sus procedimientos y mejorara sus criterios metodológicos para así obtener mayor precisión en sus cálculos, y consecuentemente entregar datos de mejor calidad a los países.

Uno de los cambios implementados en TERCE fue el cálculo de puntuaciones con el método de **valores plausibles**, los cuales posteriormente hacen posible la estimación de una puntuación media a nivel nacional, metodología mejorada para aproximar los puntajes de los países[2]. Además, siguiendo las prácticas de otros programas

[2] Esta decisión fue tomada en consenso, mediante el Acuerdo 8 de la XXXII Reunión de CNs Quito, 7-8 abril 2014, en base a Recomendación 3.1 de la Reunión de Conformación del CTAN en Montevideo, 10-11 septiembre 2011.

internacionales que emplean esta metodología, se incluyeron **covariables**[3] para estimar los valores plausibles, lo que marca otra diferencia con el SERCE[4] que estimó puntajes IRT en forma directa. Los valores plausibles son la base para los reportes de resultados de las pruebas de aprendizaje y los análisis de factores asociados. Sin perjuicio de lo anterior, la estimación directa de los puntajes IRT igualmente se produjo, y forma parte, de las base de datos del TERCE.

También se implementaron modificaciones relacionadas con los **criterios de exclusión**[5] que utiliza la prueba, asunto que finalmente determina el tamaño de la muestra efectiva con la cual posteriormente se realizan los análisis. Es conveniente aclarar que los criterios de exclusión del SERCE eran más restrictivos, lo que determinó que un mayor número de estudiantes no fueran considerados como participantes de la prueba, y por lo tanto no contaban con un puntaje. Al aplicar los criterios del SERCE, se excluyó el 29,36% del total de participantes de la muestra total del TERCE, a nivel regional. Esta situación varía dependiendo del país y la prueba aplicada, llegando incluso en ciertos casos a superar el 45%. Los criterios de exclusión de TERCE, por su parte, fueron establecidos siguiendo la tendencia en las pruebas internacionales, que consideran como válido a un examinado que alcanza a responder unas pocas preguntas, aunque deje el resto de la prueba en blanco.

La nueva escala del TERCE

En resumen, el TERCE incorporó dos elementos principales para mejorar la precisión de las estimaciones de logro: (1) cambios en los criterios de exclusión que mejoran la alineación del TERCE con los criterios de exclusión empleados en otras pruebas internacionales, y (2) mejoras el procedimiento de estimación de puntajes utilizando la técnica de valores plausibles. De estos dos elementos, el cambio de criterios de exclusión es el que podría afectar marginalmente la posición relativa de los países generando, por ejemplo, que un

[3] Las covariables utilizadas fueron: índice de actividades recreativas en el hogar; índice de prácticas de organización del aula; índice de asistencia y puntualidad docente; género del estudiante; número de adultos en el hogar; número de niños en el hogar; ¿vives con tu madre y padre?; ¿tienes libro de lenguaje?; ¿tienes libro de matemática?; ¿tienes cuaderno para tomar nota en clases?; ¿tienes en tu sala libros para leer?; ¿haces tareas de la escuela en tu casa?; ¿te preguntan si hiciste las tareas de la escuela?; ¿trabajas?.

[4] Los datos de la prueba TERCE se basan en el uso de versiones generalizadas del modelo de Rasch para respuestas binarias, en el uso de técnicas basadas en la teoría de la imputación de valores faltantes de Rubin (1987), denominados como valores plausibles, y en el uso de análisis estadísticos ponderados sobre la base de valores plausibles, para la obtención de inferencias poblacionales por país.

[5] Para que un examinado sea incluido en la base de datos para el TERCE, el estudiante debe estar presente en la sesión donde se aplica la prueba respectiva y deben contar al menos con tres respuestas marcadas. En el caso de la comparación TERCE-SERCE se usaron los criterios de exclusión del SERCE, es decir, se excluyeron cuando había omisión en las dos últimas preguntas del segundo bloque y cuando el INFIT o OUTFIT eran menor a 0,7 o superior a 1,3, respectivamente.

país que en diciembre se ubicó por sobre el promedio regional, en abril se ubique por debajo este mismo parámetro. Cabe señalar que la implementación de actualizaciones metodológicas es una práctica común en los estudios de gran escala ya que, tal como se señaló anteriormente, permiten ir mejorando la calidad de las estimaciones.

Por tales motivos, los resultados presentados para el TERCE no fueron idénticos entre diciembre 2014 y julio 2015. Mientras que en diciembre se aplicaron los criterios del SERCE a los datos del TERCE, en julio aplicamos los nuevos criterios de definidos para el TERCE lo cual sumado al cambio en la metodología de estimación de los puntajes, produjo leves cambios en los puntajes medios de los países. En consideración de lo anterior se decidió cambiar la escala de puntajes en la cual se presentan los resultados, y así evitar que las variaciones que se produjeron entre los puntajes TERCE de diciembre 2014 y julio 2015 resulten confusas. La nueva escala, que como se explicó anteriormente, está centrada en 700 puntos, no tiene incidencia alguna en la posición relativa de los países.

Considerando lo anterior, el CTAN[6] recomendó que los puntajes fuesen presentados en una escala diferente de la de SERCE (media de 500 puntos). Los países, por su parte, atendiendo a los desafíos comunicacionales que implicaba comunicar la modificación del puntaje entre las dos entregas de resultados acordaron un cambio de escala en la XXXIII Reunión de Coordinadores de LLECE realizada el 23 y 24 de octubre en Monterrey, México. Sumado a lo anterior, los representantes de los países presentes en Monterrey acordaron que esta nueva escala fuese utilizada y aplicada a todos los informes que se presentaron a partir del 2015, y a posibles estudios futuros del LLECE.

Por este motivo, y siguiendo las recomendaciones del CTAN y los acuerdos de los coordinadores nacionales del LLECE, se decidió construir una nueva escala que tiene una media centrada en 700 puntos y una desviación estándar de 100 puntos. Cabe recordar que lo que modifica las puntuaciones medias de los países es el cambio en otros criterios técnicos utilizados para el cálculo del puntaje y no en la escala en sí misma, tal como se explica en el siguiente apartado.

[6] La III Reunión del Consejo Técnico Consultivo de Alto Nivel (CTAN), se realizó en Ciudad de México el 6 y 7de Octubre 2014. Los participantes de esta reunión fueron: Felipe Martínez Rizo, Eugenio González, Wolfram Schulz y Martín Carnoy (video-conferencia). Como integrantes de la Sub-comisión de la Asamblea de Coordinadores Nacionales asistieron Margarita Zorrilla (México) y Harvey Sánchez (Ecuador).

Demostración que la utilización de una nueva escala no posee incidencia en la distribución de puntajes

Las escalas de calificaciones son transformaciones lineales de las puntuaciones brutas, es decir de la habilidad estimada con un modelo IRT. La media y la desviación estándar de las calificaciones transformadas son distintas de la habilidad IRT calculada inicialmente. Lo anterior significa que si se altera la escala de medida, ello no implica alteración alguna en la distribución de los resultados. En el caso del TERCE si la distribución de la calificación en bruto es simétrica, también lo será la distribución de las calificaciones transformadas o estandarizadas.

Como se mencionó anteriormente, la escala con la que se presenta la información no afecta la distribución ni el ordenamiento de los resultados. Para demostrar aquello, a continuación se plantea un ejercicio simple en el que es posible comparar los resultados por país utilizando la puntuación en bruto y luego el puntaje estándar. Como era de esperar, el resultado que se obtiene –en términos de ordenamiento- es idéntico, independiente de si se utiliza un criterio o el otro. Finalmente, los elementos que podrían generar variaciones en los resultados de los países obedecen a cambios en otros criterios de cálculo, que de ahora en adelante serán diferentes a los utilizados en SERCE (2006).

Tabla 1: Resultados Lectura Tercero: Habilidad IRT versus el puntaje estandarizado, ordenamiento por países.

País	Habilidad IRT, puntaje en bruto
CHL	1.0035
CRI	0.5302
NLE	0.3288
URU	0.2544
PER	0.1810
MEX	0.1773
COL	0.1369
BRA	0.1128
ARG	-0.0046
ECU	-0.0324
HON	-0.2201
GTM	-0.2401
PAN	-0.3465
NIC	-0.4910
PAR	-0.5355
REP	-0.9193

País	Puntaje estandarizado
CHL	789.50
CRI	748.36
NLE	730.85
URU	724.39
PER	718.01
MEX	717.68
COL	714.17
BRA	712.08
ARG	701.88
ECU	699.46
HON	683.15
GTM	681.41
PAN	672.17
NIC	659.61
PAR	655.74
REP	622.38

Fuente: Resultados de TERCE, a presentarse en julio2015