40° Reunión de Coordinadores Nacionales del LLECE

UNESCO Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

LLECE Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación

Plan de análisis del piloto ERCE 2019
Pruebas de logro

Jorge Manzi
MIDE UC

PERÚ Ministerio de Educación

EL PERÚ PRIMERO

Objetivos de la presentación

1. Presentar las actividades que componen el análisis del comportamiento empírico de los ítems de las pruebas de logro, en fase piloto para el ERCE, orientados a describir los logros de aprendizaje de los estudiantes de 3º y 6º grado en Latinoamérica y el Caribe.
2. Presentar las propiedades a estudiar, la metodología con que se revisarán y los criterios para considerar que el comportamiento de ítems y escalas es aceptable en las pruebas de logro de aprendizaje del ERCE en su fase piloto.

Propósitos del plan de análisis para ERCE piloto 2019:

➢Asegurar que el comportamiento de los ítems es apropiado para su uso en la aplicación definitiva, en base a los tres pilares fundamentales de la medición, establecidos por los Estándares de AERA, APA & NCME (2014), como son:

- Confiabilidad
- Validez
- Imparcialidad de la medición

➢Dado que la prueba ERCE tiene también el objetivo de mostrar cuál es el avance en los aprendizajes de los estudiantes de la región respecto del TERCE, se incluye el estudio de propiedades asociadas a asegurar la calidad de la futura equiparación de ambas pruebas.

Recepción y preparación de bases de datos

- UNESCO será el responsable del aseguramiento de la calidad de las bases de datos.
- MIDE UC será el responsable de la re-estructuración de las bases de datos para realizar los análisis.
- MIDE UC realizará una última verificación de las bases de datos recibidas, cautelando principalmente la existencia de valores dentro de rango en todas las variables, así como la ausencia de información duplicada al interior de cada país. En caso que en esta fase se identifiquen problemas de digitación o a nivel de registros, MIDE UC deberá devolver a UNESCO las bases de datos para su ajuste y corrección (lo que puede conllevar consecuencias para el calendario de análisis).

Metodología

- Tanto las propiedades a nivel de instrumento como las propiedades a nivel de los ítems que lo componen serán analizadas en base a dos enfoques, la Teoría Clásica de Tests y la Teoría de Respuesta al ítem.
- En todas las pruebas, al igual que en TERCE y SERCE, se ajustarán modelos IRT de 1 parámetro, pues este tipo de modelo logra llevar a una misma escala las estimaciones de habilidad de las personas y las estimaciones de dificultad de los ítems.
  - En el caso de pruebas compuestas exclusivamente por ítems respuesta cerrada y que, por tanto, son puntuados de forma dicotómica, ajustará el modelo Rasch (Rasch, 1960).
  - En el caso de pruebas que combinan ítems de respuesta cerrada y abierta, se ajustará un modelo de créditos parciales (Masters, 1982).

Análisis de los ítems

- El análisis de las propiedades de cada ítem permite describir en qué medida su funcionamiento se encuentra alineado con el comportamiento esperado en función del modelo de medición.
- En este análisis se tendrá información y criterios de selección para distintas propiedades de los ítems, las que deben ser analizadas en conjunto con el contenido de los ítems para decidir cuáles son los mejores candidatos para conformar la prueba definitiva ERCE.
- Los análisis se realizarán por cuadernillo y de manera global ocupando la información de todas las formas del instrumento.
- Los análisis se realizarán con la base de datos consolidada que combina a todos los países, lo que se complementará con análisis por país.

Análisis de los ítems

Propiedad	TCT	IRT
➢Omisión	X	
➢Dificultad	X	X
➢Capacidad discriminativa	X	
➢Comportamiento distractores	X	
➢Ajuste al modelo		X

Teoría Clásica de Tests: Omisión

Criterios:

- Omisión hasta un 20% se acepta el ítem.
- Entre 21% y 30% se analizará el patrón de omisión (el ítem se acepta si la omisión es mayor para los examinados con menor rendimiento en el cuadernillo).
- Sobre un 30% el ítem se excluye.

  - Se prestará atención a la comparación de la omisión entre bloques

Teoría Clásica del Test: Dificultad y discriminación

Criterio para la dificultad:

- Dificultad aceptable entre 20% (0,2) y 80% (0,8). Ítems fuera de este rango se aceptan si sus dificultades IRT se encuentran entre [-3 ,+3].

Criterios para la discriminación:

- Se incorporarán en la prueba ítems con correlaciones biseriales sobre 0,3.
- Se incluirán si fuera necesario para asegurar la cobertura de la tabla de especificaciones ítems con correlaciones biseriales algo inferiores, en tanto esto no perjudique la confiabilidad.

Teoría clásica: Distractores

Criterios:

- Se revisará la distribución de las respuestas correctas entre distractores para evaluar su distribución, prefiriéndose ítems cuyos distractores presenten distribuciones relativamente homogéneas.
- Se considerará especialmente problemático el caso de ítems donde al menos uno de sus distractores presente una correlación biserial positiva (similar o mayor a la de la respuesta correcta) con el puntaje total.

Teoría de Respuesta al Ítem: Ajuste/INFIT y OUTFIT

Criterio para INFIT y OUTFIT:

- Se aceptarán valores de INFIT y OUTFIT el rango [0.7 , 1.3], ampliable al rango [0.5 , 1.5], en caso que la eliminación del ítem afecte negativamente la validez de contenido de la prueba y su comportamiento sea adecuado en las propiedades directamente relacionadas con la confiabilidad (dificultad y discriminación).

Análisis de los ítems

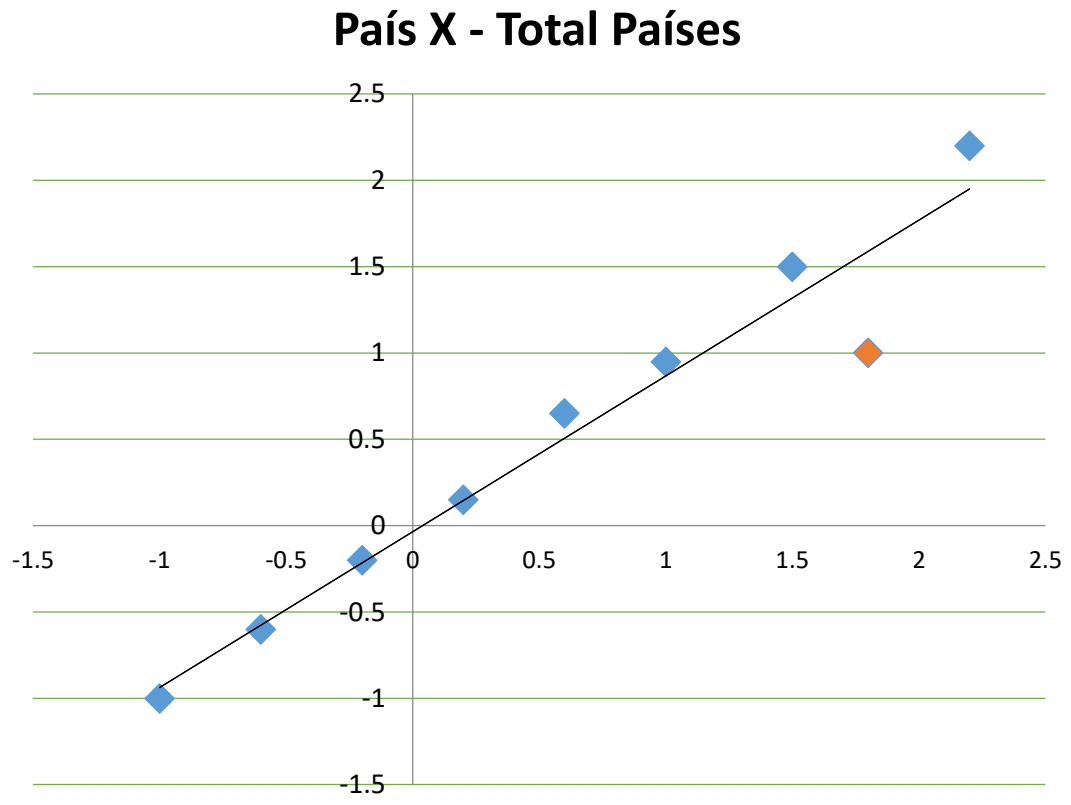
Propiedad	TCT	IRT
➢Omisión	20%*	-
➢Dificultad	20% (0,2) y 80% (0,8)	[-3 ,+3]
➢Capacidad discriminativa	0.3 **	
➢Comportamiento distractores	Biseriales nulas o negativas	-
➢Ajuste al modelo	-	[0.7 , 1.3]]***

* Se admitirá hasta 30% si la omisión es mayor para los examinados con menor rendimiento
** Puede admitirse número inferior en tanto no perjudique la confiabilidad.
***Ampliable al rango [0.5 , 1.5], en tanto las demás propiedades estén OK.

Interacción ítem-país

Para evaluar la posible interacción ítem-país, se utilizará la estrategia de TIMSS 2015 (Martin, Mullis & Hooper, 2016).

- Se busca examinar si la dificultad de los ítems es invariante entre países.
- Se conceptualiza como interacción entre el ítem y el país aquellos ítems donde el ordenamiento de las dificultades de los ítems varía considerablemente para un país respecto del comportamiento general del test (observado en el total de países)



Interacción ítem-país

- En este análisis se detectarán los ítems que presenten grandes diferencias respecto del comportamiento global.
- Se buscará establecer si estas diferencias en un determinado ítem son concordantes con indicadores de cobertura curricular del contenido evaluado en dicho ítem a través de los países.
- Se considerará como un criterio de exclusión de un ítem, el hecho que presente una alta interacción a través de los países, especialmente cuando dicha interacción sea consistente con la información sobre cobertura curricular en los países involucrados.
  - En prueba definitiva se volverá a verificar empíricamente esta propiedad.

Análisis de funcionamiento diferencial de ítems (DIF) según género.

- El sesgo según género se conceptualizará como el de ítems donde, para un mismo nivel de habilidad, examinados de distinto género muestren diferencias sustantivas en su probabilidad de acierto.
- Para detectar ítems que presenten funcionamiento diferencial de género, se seguirá lo propuesto por ICCS (Schulz, Ainley & Fraillon, 2011) e ICILS (Fraillon, Schulz, Friedman, Ainley & Gebhardt, 2015). En dicha propuesta se considera sospechoso de sesgo a los ítems que muestran distancias absolutas mayores a 0.6 logit entre hombres y mujeres.

Comportamiento de ítems ancla para análisis de tendencias

Re-análisis de propiedades del test y de ítems con los países que participan en ambas mediciones

Para que la pruebas ERCE puedan entregar información acerca de cuánto han variado los logros de aprendizaje de los estudiantes, la prueba ERCE se pondrá en una misma escala con la prueba TERCE.

- La equiparación de las pruebas ERCE y TERCE se realizará mediante una calibración conjunta de ambas mediciones, procedimiento que requiere estabilidad en el comportamiento psicométrico de los ítems ancla.
- Se revisará el comportamiento específico de los ítems seleccionados para anclaje, a partir de los mismos indicadores antes generados en el análisis de ítems, poniendo foco en aquellas propiedades de los ítems que inciden en el adecuado procedimiento de equiparación entre las pruebas ERCE y TERCE.

Análisis de los ítems que componen los bloques anclas

El análisis de los ítems que componen los bloques ancla se enfocará en cuatro propiedades que son fundamentales para una correcta equiparación mediante una calibración concurrente:

1. Se analizará que las dificultades de los ítems se mantengan consistentes entre ambas mediciones, mediante un análisis de regresión de sus parámetros de dificultad en TCT e IRT.
2. Se verificará que los ítems ancla muestren altas correlaciones ítem-test, supuesto de los modelos IRT a utilizar, o, al menos, que sean equivalentes a las de la prueba TERCE.
3. Se verificará que en todos los ítems ancla el ajuste al modelo sea adecuado.
4. Se verificará que los ítems ancla no muestren comportamiento diferencial según género, ni interacción ítem-país.

MUCHAS GRACIAS