Documento de Trabajo N° 5

Funcionamiento diferencial del ítem en la evaluación educativa a nivel América Latina y el Caribe

Investigadora:
Pamela Woitschach
Director:
Luis Ortiz

2019

Este documento de trabajo es el informe de avance preliminar y no editado del proyecto de investigación PINV15-1150 *Funcionamiento diferencial del ítem en la evaluación educativa a nivel América Latina y el Caribe*, financiado por el CONACYT a través del Programa Pro-Ciencia con recursos del Fondo para la Excelencia de la Educación e Investigación (FEEI) del FONACIDE.

ÍNDICE

INTRODUCCIÓN

Los programas internacionales de evaluación de sistemas educativos (International Large-Scale Assessments in Education [ILSA]) tienen como finalidad conocer, describir y comparar los resultados de nivel de competencia tanto dentro y como entre los países participantes a través de la aplicación de pruebas de evaluación educativa estandarizada (Maddox, 2019). Una segunda finalidad de los programas internacionales de evaluación de sistemas educativos es la de analizar las variables y factores asociados al rendimiento (Ferández-Alonso, 2004). La evaluación estandarizada de sistemas educativos inicia hace más de cinco décadas con la creación de la Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA), organismo de cooperación internacional no gubernamental conformado por instituciones de investigación públicas y privadas de más de 60 países, desde entonces la IEA ha liderado el análisis educativo curricular a nivel mundial. En la actualidad los estudios de la IEA más conocidos como TIMSS (Trends in International Mathematics and Science Study) y PIRLS (Progress in International Reading Literacy Study), son desde hace más de dos décadas desarrollados por el Boston College (Mullis, Martin, Gonzalez, & Kennedy, 2003; Mullis, Martin, & Loveless, 2016). Como una alternativa esta vez orientada a la evaluación de sistemas educativos a través de las competencias de sus estudiantes, emerge el Programa Internacional para la Evaluación de Estudiantes (PISA) de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cuyo objetivo primordial es el de otorgar a los sistemas educativos participantes información válida, confiable y comparable que permita la mejora de sus políticas y resultados educativos.

Estos programas internacionales de evaluación de sistemas educativos tienen un gran impacto tanto mediático, académico como investigador. Por lo que, una vez publicados los resultados, los gabinetes de comunicación de los gobiernos y los organismos internacionales preparan notas de prensa que son consumidas ávidamente por los medios de comunicación. Así mismo, se realizan informes de resultados y análisis secundarios que son publicados en revistas especializadas (LLECE, 2014, 2016b; M. O. Martin & Mullis, 2013; Mullis, Martin, Foy, & Arora, 2012; OCDE, 2014; UNESCO, OEI, MEC, & DGEEC, 2013).

El rápido crecimiento de la cultura de evaluación a nivel mundial, con el objeto de la rendición de cuentas de los gobiernos, ha sido extendido no solo a países industrializados, sino que a aquellos en vías de crecimiento (Smith, 2014). Como ejemplo de este crecimiento, podemos observar las evaluaciones regionales como el Southern and *Eastern Africa Consortium for Monitoring Educational Quality*, que evalúa los conocimientos de los estudiantes de países de África (Hungi, 2011) y *The Southeast Asia Primary Learning Metrics* (SEA-PLM, 2016) programa de evaluación regional de la educación primaria en Asia. América Latina, no es se encuentra fuera de esta corriente y las evaluaciones educativas son lideradas desde el año 1996 por la UNESCO a través del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), bajo la coordinación técnica de la Oficina Regional de Educación de la UNESCO para América Latina y el Caribe (LLECE & UNESCO-OREALC, 2016; UNESCO-OREALC & LLECE, 2000, 2010).

El objetivo de esta investigación es el de identificar la presencia del funcionamiento diferencial del ítem de las pruebas TERCE aplicadas a 15 países de América Latina. Con el fin de ofrecer un panorama general de las evaluaciones en América Latina, la presente investigación centra su apartado teórico en una breve introducción en dos subapartados, el primero centrado en la metodología de los programas de evaluación con mayor trayectoria en la región latinoamericana, PISA de la OECD y los estudios de la UNESCO; y seguidamente, en el segundo apartado, orientado a ofrecer al lector un panorama teórico sobre la medición de la invarianza a través del análisis del funcionamiento diferencial del ítem.

1. PROGRAMA INTERNACIONAL PARA LA EVALUACIÓN DE ESTUDIANTES (PISA)

PISA evalúa desde el año 2000 en ciclos trianuales, las competencias y destrezas básicas de los estudiantes de 15 años en las áreas de matemáticas, ciencias naturales y comprensión lectora, siendo uno el área elegido como eje central de cada aplicación trianual (OECD, 2002). Los países participantes de PISA se clasifican como sistemas educativos provenientes de países miembros y no miembros de la OCDE; dado el impacto de sus resultados a escala mundial, se ha observado un incremento sustancial de la participación tanto a nivel de sistemas educativos como de sub-agrupaciones (alumnos, familias, profesores y directores). La muestra evaluada es seleccionada mediante un proceso de muestreo estratificado de conglomerados de dos etapas de selección (Martínez-Arias, 2006) estrategia que implica un trabajo colaborativo de los gobiernos participantes, quienes se constituyen en los agentes encargados de la puesta en marcha de la evaluación en sus respectivos territorios nacionales (Ferández-Alonso, 2004).

La construcción del instrumento de evaluación se inicia con la especificación de las competencias a ser evaluadas, este contenido a diferencia de la evaluación curricular es acordado por los diversos expertos internacionales quien se focalizan en las destrezas y competencias que los estudiantes deben poseer al término de la educación obligatoria. La evaluación es realizada con lápiz y papel en la mayoría de los países, aunque la progresiva inclusión de la evaluación computarizada fue iniciada en PISA 2012 para la competencia matemática, extendida a la evaluación de la competencia científica de PISA del 2015 y a la comprensión lectora en el entorno digital evaluada en PISA 2018.

Los formatos de respuesta a los ítems son de opción múltiple y respuesta construida codificados como respuestas de acierto y error (OECD, 2014), el proceso de construcción de las pruebas y análisis de los datos se realizan bajo técnicas analíticas vanguardistas y robustas que garantizan la calidad de la información obtenida. Para garantizar la efectividad en el uso del tiempo de aplicación y respuesta de la prueba, el programa PISA al igual que otros programas centrados en la evaluación educativa estandarizada, basan la construcción de las pruebas en la utilización de un diseño matricial de distribución de los ítems que permite, por un lado la aplicación de ítems que garanticen la cobertura del contenido a evaluar y por el otro, la disminución del tamaño del error muestral por medio de la disminución del tiempo de aplicación de la prueba y el aseguramiento de una tasa de participación amplia (OECD, 2002, 2005, 2009, 2012, 2014). Dado que las pruebas son aplicadas en diversas lenguas, PISA incluye dos versiones fuente de la prueba, una en inglés y otra en francés, además ha

desarrollado guías detalladas de traducción y adaptación del contenido del test producto del trabajo colaborativo de los gobiernos (Martínez-Arias, 2006).

Una prueba piloto de las pruebas es realizada en los países participantes y análisis sobre los niveles de dificultad, discriminación, opciones de respuesta, análisis de distractores, tasas de omisiones y funcionamiento diferencial entre países, es presentado como criterios de selección de los ítems que conformarán la prueba cognitiva final (OECD, 2012). Los análisis psicométricos realizados se basan en la Teoría de la Respuesta al Ítem (TRI), que permiten la comparabilidad y la construcción de las puntuaciones, así como el uso de diseños matriciales de distribución de los ítems. Pesos muestrales y pesos replicados son construidos, así como también la inclusión de cinco valores plausibles para la estimación de las puntuaciones. En cuanto al análisis de los factores asociados al rendimiento, modelos jerárquico-lineales son utilizados, respetando la naturaleza agrupada de los datos, permitiendo la inclusión de niveles de agregación, valores plausibles y pesos muestrales en el análisis.

Como resultado de PISA son publicadas cápsulas informativas para el público en general, informes de resultados por competencias e informes de factores asociados al conocimiento, estos últimos se encuentran orientados a los actores políticos y gestores educativos. A nivel técnico se presentan manuales metodológicos que detallan las características del proceso de evaluación, así como el análisis e interpretación de los resultados amplia (OECD, 2002, 2005, 2009, 2012, 2014).

Los resultados de logro en las competencias evaluadas son presentados en el formato conocido como tabla de ligas, que permiten la comparación de las puntuaciones entre los sistemas educativos participantes. Datos informativos tales como los porcentajes de cada país en los niveles de rendimiento, así como tamaños de los efectos, el rendimiento diferenciado ya sea por género o contexto socioeconómico, así como la influencia de las escuelas en el rendimiento se pueden observar con claridad en los informes publicados y que, a su vez, presentan tablas comparativas de forma a observar una comparativa de los resultados obtenidos a lo largo de las evaluaciones realizadas.

Desde sus inicios PISA se ha caracterizado por el incremento en el número de países participantes, los países participantes desde el año 2000 al 2018 representan el 80% de la economía mundial (Smith, 2014). Este incremento en la participación de países es un claro ejemplo de la relevancia de la evaluación estandarizada como herramienta de mejora de las políticas educativas. Como consecuencia del incremento en la participación de sistemas educativos no-miembros de la OCDE y como una herramienta de monitoreo de los objetivos del desarrollo sostenible promovidos por la Asamblea General de las Naciones Unidas en 2015, se lanza en el 2015 el plan piloto de PISA para el Desarrollo (PISA-D). Este proyecto piloto de seis años de duración cuyo objetivo principal es la accesibilidad de las herramientas utilizadas en PISA para la evaluación de países de ingreso medio y bajo, es implementado inicialmente con la participación de países como: Bután, Camboya, Ecuador, Guatemala, Honduras, Paraguay, Senegal y Zambia. La participación de Paraguay en las evaluaciones PISA, se da por primera vez en el 2018 y como bien lo indica Addey (2019) la brecha lingüística que se observa entre los organizadores de la evaluación y los gestores de la

evaluación a nivel nacional es aún bastante notoria, lo que dificulta la inclusión efectiva del país en particular y los países en vías de desarrollo en general, en el programa de evaluación.

2. TERCER ESTUDIO REGIONAL COMPARATIVO Y EXPLICATIVO TERCE DE LA UNESCO

Como una alternativa más efectiva a nivel regional, se observa la triada de estudios realizados hasta la fecha por la UNESCO en conjunto con el LLECE. La evaluación de la UNESCO es caracterizada desde sus inicios por la participación íntegra de países de la región latinoamericana, región que en suma se caracteriza por los niveles más altos de desigualdad social e inequidad en materia educativa del mundo (UNESCO-OREALC, 2016a, p. 89). La UNESCO a través del LLECE y bajo la coordinación técnica de la OREALC Santiago, analiza los conocimientos y factores asociados de los países de América Latina desde la década del 90. El Primer Estudio Internacional Comparativo (PERCE) del año 1997 contó con la participación de trece países (UNESCO-OREALC & LLECE, 2000), para el año 2006 el Segundo Estudio Regional Comparativo y Explicativo (SERCE) contó con la participación de dieciséis países (UNESCO-OREALC & LLECE, 2010). La última versión del estudio en el año 2013 (TERCE) incluyó además del estado de Nuevo León a quince países de la región (LLECE, 2014). Actualmente, el LLECE se encuentra coordinando el Cuarto Estudio Regional Comparativo y Explicativo (ERCE) a ser aplicado en el año 2019.

El objetivo primordial de las evaluaciones de la UNESCO es el de determinar el desempeño escolar de los países basados en un currículo académico en común. Análisis posteriores se orientan a conocer la relación entre el desempeño escolar y los factores asociados al aprendizaje; siendo estos analizados desde la información recabada del estudiante y su familia, las escuelas y los sistemas educativos (UNESCO-OREALC & LLECE, 2016). Los resultados del programa de evaluación son publicados en formato de informes de logro cognitivo e informes de factores asociados, que se encuentran orientados a la población escolar y de sistemas educativos. Así como reportes técnicos orientados al público con conocimiento estadísticos y psicométricos de cada estado participante. El LLECE, así como su nombre bien lo refiere no solo es un laboratorio de medición de la calidad educativa regional, sino que es también en un laboratorio de capacitación para todos los sistemas de evaluación educativa nacional de los países participantes.

Dado que la evaluación TERCE es de carácter curricular, para el desarrollo de los instrumentos y los marcos de evaluación, un análisis curricular conjunto de todos los países es desarrollado con el fin de determinar los contenidos comunes a ser evaluados. Una vez obtenida esta información tablas de doble entrada (dominios y ejes temáticos) son construidas, siguiendo en la misma línea de las evaluaciones PISA de la OECD. El diseño muestral de TERCE es aleatorio sistemático por conglomerados bietápico (escuela-aula) y los estratos analizados son escuelas urbanas públicas, urbanas privadas y rurales. La inclusión de estados nacionales se da desde la prueba SERCE, donde estados como el de Nuevo León (México) y Goias (Brasil) participaron, pero solo el estado de Nuevo León se mantuvo en la evaluación TERCE del 2013.

Dado el amplio contenido a ser evaluado en las pruebas de logro cognitivo, TERCE al igual que PISA, utiliza un diseño matricial. En este caso de bloques incompletos, donde los ítems que constituyen la prueba son de opción múltiple y preguntas abiertas codificadas en formato binario y de crédito parcial. Una vez construidas las pruebas de logro cognitivo, se realizan aplicaciones previas a muestras específicas de los países participantes con el fin de afinar los análisis psicométricos y la selección de los ítems que conformarán la prueba final. Los análisis psicométricos se centran en la teoría clásica de las pruebas (TCT) y en la teoría de la respuesta al ítem (TRI). La particularidad de las pruebas de la UNESCO en América Latina es que son aplicadas desde la década de los 90`, por lo que es posible la comparabilidad en series longitudinales. Claro ejemplo de ello es la inclusión de ítems de anclaje de SERCE en la prueba TERCE.

La estimación de las puntuaciones incluye el uso de cinco valores plausibles para el cálculo de los resultados de logro cognitivo. En PERCE la escala estaba anclada a una media de 250 y una desviación típica de 50 puntos, mientras que en SERCE esto fue modificado a una escala con anclada en una media de 500 y una desviación típica de 100 puntos. Finalmente, en TERCE la escala estuvo ubicada en una media de 700 y una desviación típica de 100. Una característica clave de las evaluaciones educativas a gran escala es el formato de presentación de los resultados por países en lo que es denominado tabla de ligas. Este formato de presentación de resultados ordena a los países en forma de ranking de acuerdo a la puntuación obtenida en las materias evaluadas. Esta primera pieza de información en cada reporte de logro cognitivo permite no solo el ordenamiento de los países, sino que también una comparación global del rendimiento obtenido. Pero la primera característica intrínseca al uso de estas tablas es (1) la suposición de que todos los países han sido evaluados bajo las mismas condiciones de prueba y (2) que las puntuaciones no se ven afectadas por características sociales y de contexto, por lo que la comparación es posible.

La multiculturalidad juega un papel relevante en los programas de evaluación educativa. TERCE tiene como principal finalidad la comparabilidad de los resultados obtenidos, hecho que se constata en la presentación de los resultados de los países en formato de tabla de ligas. Es de suma importancia garantizar la equivalencia de las pruebas, puesto que el factor multicultural juega un papel decisivo en la validez de los puntajes y en la posibilidad de que estos puntajes sean comparables entre países, regiones y culturas (Ercikan, Roth, & Asil, 2015). La principal estrategia para garantizar la equivalencia y la comparabilidad de las pruebas entre agrupaciones se realiza a través del estudio del funcionamiento diferencial del ítem (DIF).

3. FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM EN LAS PRUEBAS EDUCATIVAS A GRAN ESCALA

Asociaciones avocadas a la creación de normas y criterios para la construcción, uso de las pruebas y prácticas de evaluación en las áreas de psicología y educación han puesto el foco de atención en lo que a buenas prácticas, diversidad cultural y justicia en el uso de las pruebas se refiere. En esta línea *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA] y The National Council on Measurement in Education [NCME], 2014) y *The International Test Commission* [ITC] han publicado guías orientadas a público especializado (técnicos y

psicómetras), así como materiales orientados a los usuarios de test, con el fin de proveer un soporte que permita asegurar la justicia en el uso y la comparabilidad de los resultados de los test en contextos de diversidad cultural y lingüística.

El análisis del DIF en las pruebas de evaluación educativa es uno de los pasos previos para la conformación de los ítems que formarán parte de la prueba final. En este sentido evaluaciones educativas a gran escala como PISA, TIMMS, PIRLS y las evaluaciones de la UNESCO han incursionado en el análisis del funcionamiento diferencial del ítem como criterio de análisis de ítems y como información relevante a ser entregada a los países participantes (Joncas & Foy, 2012).

El DIF ocurre cuando diversos grupos de examinandos de igual capacidad en el criterio evaluado tienen en promedio, respuestas sistemáticamente diferentes a un ítem (AERA, APA, & NCME, 2014, p. 16). En otras palabras, un ítem presenta un funcionamiento diferencial cuando dos grupos que tienen el mismo nivel de habilidad rinden de modo diferente en un ítem por factores ajenos al propio constructo que se mide (Abad, Olea, Ponsoda, & García, 2011). A decir del LLECE éste es un indicador de equidad, pues permite conocer el comportamiento de los ítems teniendo en cuenta las características culturales que puedan influir en su funcionamiento, con el fin de garantizar que la calificación final en el estudio se realice a partir de ítems que no presenten DIF (ORELAC-UNESCO-LLECE, 2010, p. 231).

Tres generaciones del estudio del DIF son analizadas por Zumbo (Zumbo, 2007) quien puntualiza que la primera generación el DIF en aquel entonces conocido como sesgo de los ítems, estaba centrado en el análisis de las agrupaciones de sexo o raza denominados como grupo focal y de referencia. En la segunda generación el término de sesgo fue modificado por funcionamiento diferencial del ítem y los estudiosos se centraban en el desarrollo de técnicas más sofisticadas para la evaluación. En tanto que, la Tercera Generación del DIF viene de la mano de una visión más integradora que busca no solo su identificación, sino que más bien apuesta por el acercamiento a una explicación de las causas que subyacen a la presencia del DIF; como un fenómeno en donde la caracterización del contexto denominado por el autor como "testing situation" mueve progresivamente el análisis del DIF a modelos multinivel.

Las estrategias para la detección del DIF tanto uniforme como no uniforme, se basan en estrategias centradas en el análisis de varianza, técnicas derivadas de la TRI y las estrategias de análisis multinivel. Entre las principales técnicas se encuentran el Método Delta, el Índice de Estandarización, Mantel-Haenszel, la Regresión Logística y por último las estrategias de detección del DIF mediante el análisis multinivel.

El método Delta se basa en la comparación de las puntuaciones de los índices de dificultad convertidos a puntuaciones Z en un gráfico bivariado. Por otro lado, el Índice de Estandarización cuantifica las diferencias entre las proporciones de acierto de los grupos de referencia y focal para cada una de las categorías en las que se divide la prueba, ofreciendo un indicador global para cada una de esas diferencias para el ítem (Muñiz, 2003, p. 252). La técnica de Mantel-Haenszel es, debido a su facilidad de implementación y el hecho de contar

con una prueba de significación asociado, una técnica ampliamente utilizada; sin embargo, al momento de detectar el DIF no uniforme, no es una técnica muy adecuada. (Abad et al., 2011). Por último, la Regresión Logística resulta efectiva para la detección tanto del DIF uniforme como no uniforme, aunque la técnica presenta la desventaja de ser calculada sobre la puntuación total de la prueba, siendo probablemente esta puntuación no fiable.

Los métodos derivados de la TRI presentan mayor solidez teórica y complejidad estadística para la detección de DIF, debido a la invarianza de los parámetros y la comparación de puntuaciones o patrones de respuesta en relación con el nivel de habilidad del sujeto. Las técnicas se basan en la comparación de las CCI de los distintos grupos; o bien en comparar los parámetros de los ítems y probar la hipótesis nula de igualdad de CCI (Navas, 1994).

Estudios comparativos realizados como por ejemplo en Reino Unido y España, analizan los ítems liberados en las pruebas PISA con métodos Mantel-Haenszel, Regresión Logística y Medias Estandarizadas y destacan que en veintiséis ítems analizados dos de ellos presentan DIF (Elousa, 2006). En otro estudio realizado con ítems provenientes de PIRLS y utilizando como técnica del model-based recursive partitioning (MBRP), se identificó que el lenguaje materno en comparación con la lengua que habla el niño en la escuela, o la lengua que hablan en la casa; es la variable que tiene mayor incidencia en la presencia de DIF uniforme en las pruebas (Holmes, Finch, & French, 2016). Por otro lado, la exploración del DIF en las pruebas de evaluación educativa de matemáticas en estudiantes de 11 años en Inglaterra, a través de la técnica de regresión logística, determinó la presencia de DIF por género en 36 de los 40 ítems, 18 a favor de las niñas y 18 a favor de los niños (Ong, Williams, & Lamprianou, 2015).

Los métodos multinivel se caracterizan por su facilidad de aplicación y la inclusión de variables en todos los niveles de agrupación. El modelo mixto lineal generalizado (GLMM) o el modelo mixto lineal generalizado jerárquico (HGLMM) pertenece a la familia de modelos generales de efectos mixtos. Diversas investigaciones han denotado la eficacia de esta técnica de análisis para la detección del DIF (Balluerka, Gorostiaga, Gomez-Benito, & Hidalgo, 2010; Balluerka, Plewis, Gorostiaga, & Padilla, 2014; Swanson, Clauser, Case, Nungester, & Featherman, 2002, van den Noortgate, & de Boeck, 2005), dado que los modelos multinivel permiten la inclusión de variabilidad proveniente de los distintos niveles propios del contexto educativo y se erigen como una de las técnicas modernas mas eficaces para la detección del DIF en pruebas de evaluacion educativa (Chen & Zumbo, 2017).

Dada la inclusión de Paraguay en las evaluaciones educativas a gran escala como una de las prioridades del Plan Nacional de Desarrollo 2030, esta investigación tiene el objetivo de analizar el DIF en las pruebas de lectura y ciencias naturales de TERCE, dado que TERCE a diferencia de PISA, es el programa de evaluación educativa a gran escala que cuenta con la participación de más del 80% de países de América Latina y el Caribe. Es de vital importancia conocer el comportamiento de los ítems en poblaciones caracterizadas por la multiculturalidad, como es el caso de Paraguay y los países de América Latina y el Caribe. La presente investigación se encuentra enmarcada en el eje de educación para la formación, eje transversal de la tecnología educativa. Alineados en el Plan de Desarrollo 2030 de la República del Paraguay y focalizados en el eje 1 de reducción de la pobreza y desarrollo

social de la estrategia 1.2. Las evaluaciones educativas dentro del Plan de Desarrollo de la Nación tienen como finalidad la producción de informes de investigación que sirvan de base objetiva y válida para la toma de decisiones sobre las políticas educativas a nivel nacional.

El objetivo de este trabajo es el análisis de los ítems liberados de la prueba de lectura y ciencias naturales de TERCE aplicada a Países de América Latina y el Caribe. Los objetivos del estudio son:

1 Identificar los países participantes de la evaluación TERCE en América Latina y el Caribe.
2 Establecer las características de cada país participante en TERCE.
3 Determinar la presencia de funcionamiento diferencial del ítem en las pruebas de lectura y ciencias naturales.
4 Explorar por medio de las técnicas de análisis jerárquico los condicionantes de logro en las pruebas de lectura y ciencias naturales.
5 Comparar la eficacia de los métodos de evaluación del funcionamiento diferencial del ítem.

4. MÉTODO

4.1. Muestra

La muestra del estudio TERCE está compuesta por el alumnado matriculado en 3er. y 6to. Grado de la educación primaria de 15 países de América Latina y el Estado de Nuevo León (México). La muestra de TERCE es seleccionada mediante un muestro bietápico por conglomerados y estratificado (Joncas & Foy, 2012; OECD, 2009), obteniéndose una muestra por encima de los 60.000 estudiantes y 2000 escuelas que representan a casi 9 millones de estudiantes de América Latina y el Caribe. En este estudio en particular, debido a que nuestro principal objetivo es trabajar con los ítems de cada prueba de logro cognitivo, el marco muestral se orienta a los estudiantes que respondieron a los ítems incluidos en este análisis del funcionamiento diferencial. Recordemos que TERCE basa su aplicación de pruebas en un diseño matricial, donde los estudiantes responden a un número limitado de ítems. Para este efecto la muestra de esta investigación está compuesta por 12.415 estudiantes evaluados en lectura y 13.498 estudiantes evaluados en ciencias naturales, pertenecientes a 2802 centros educativos de quince países de América Latina y el Caribe.

4.2. Diseño Muestral en TERCE e Implicaciones para el Análisis

Como se acaba de apuntar TERCE comparte con otros programas de evaluación de sistemas educativos el empleo de diseños muestrales estratificados, aleatorios y sistemáticos. En estos diseños las unidades muestrales (centros, aulas, estudiantes…) se seleccionan en dos o más etapas y dichas unidades muestrales no tienen la misma probabilidad de ser elegidos. Esto significa que, dentro de un mismo país o estrato, todos los estudiantes no representan exactamente igual al conjunto de la población o, dicho de otro modo, que algunos estudiantes, a la hora de representar a la población total de su país, son más importantes que otros. Para graduar su importancia o representatividad a cada estudiante se le asigna un peso muestral cuyo tamaño o valor es inversamente proporcional a la probabilidad que cada estudiante tiene de ser elegido (Martin, Mullis, & Foy, 2015; Martin, Mullis, & Hooper, 2016; OECD, 2014).

El software de análisis clásico (por ejemplo, SPSS) asume que los casos se seleccionan según un muestreo aleatorio simple, por lo que sus resultados tienden a infra estimar los errores típicos de los estadísticos lo que aumenta exponencialmente la probabilidad de cometer un error de Tipo I, es decir, de encontrar falsos positivos al rechazar la hipótesis nula, siendo ésta verdadera. Por su parte el software de análisis multinivel (por ejemplo, HLM), aunque es más adecuado ya que reconoce la estructura anidada de los datos, también se basa en el principio de que los datos se obtienen de un muestreo aleatorio y en este caso los resultados tienden a sobreestimar la varianza muestral y, por tanto, a aumentar la probabilidad de cometer un error de tipo II, esto es, rechazar la hipótesis alternativa, siendo ésta verdadera (OECD, 2009).

5. INSTRUMENTOS

En este estudio en particular, se utilizaron los datos provenientes de las pruebas de logro cognitivo en lectura y ciencias naturales, así como la información proveniente de los cuestionarios de contexto del estudiante, la familia, los directores y profesores participantes del estudio TERCE aplicado en el año 2013.

5.1 Pruebas de logro cognitivo

En general cada prueba de logro cognitivo de TERCE sigue un diseño matricial, por lo que cada una de ellas se organizan en seis bloques o clúster de entre 16 y 17 ítems cada uno. Esos bloques a su vez de distribuyeron en seis modelos de cuadernillos diferentes mediante un diseño de bloques incompletos. Esta forma de organizar los ítems de la prueba permite que los resultados se puedan generalizar a la población completa, aunque los estudiantes respondan solo a una pequeña parte del banco de ítems. (Fernández-Alonso & Muñiz, 2011).

Cada cuadernillo está formado por dos bloques o clúster de ítems, y cada clúster aparece dos veces a lo largo de toda la colección de cuadernillos, una vez al inicio del cuadernillo y una segunda vez en la segunda parte del cuadernillo. Esta rotación del orden de presentación del clúster permite controlar el efecto del orden de las posiciones. Finalmente, cada prueba contiene dos bloques de anclaje con la prueba SERCE, permitiendo estudios de tendencia (LLECE, 2016b). Los ítems fueron ajustados empleando el programa Winsteps (Linacre, 2005) y utilizando el modelo de un parámetro de Rasch. La puntuación de cada estudiante para cada materia fue calculada mediante la metodología de valores plausibles que es la más eficiente para recuperar los parámetros poblaciones en las evaluaciones de sistemas educativos (Mislevy, Beaton, Kaplan & Sheehan, 1992; OECD, 2009; von Davier, Gonzalez & Mislevy, 2009). En TERCE las puntuaciones individuales fueron estimadas conjugando las respuestas de los estudiantes a los ítems con la información proveniente de diferentes covariables que funcionan como factores de imputación, y fueron expresadas en una escala con media 700 puntos y desviación típica 100 (UNESCO/OREALC, 2016).

5.2. Prueba de Lectura

Los estudiantes respondieron a la prueba de lectura, construida en base a una tabla de especificaciones que contiene dos dominios (comprensión de textos y dominio metalingüístico) y tres procesos cognitivos (literal, inferencial y crítico). La prueba de lectura

contó en total con 96 ítems, agrupados en seis bloques. Los bloques a su vez se distribuyeron en seis cuadernillos de aplicación, siguiendo un diseño matricial. Cada cuadernillo está compuesto por un total de 30 a 32 ítems.

5.3. Prueba de Ciencias Naturales

Los estudiantes respondieron a la prueba de ciencias naturales, la misma ha sido construida en base a una tabla de especificaciones que contiene cinco dominios (salud, seres vivos, ambiente, la tierra-sistema solar, y materia-energía) y tres procesos cognitivos (reconocimiento de información y conceptos, comprensión y aplicación de conceptos y pensamiento científico y resolución de problemas) (UNESCO-OREALC, 2016) y constaba de 92 ítems, en su mayoría de elección múltiple, agrupados en seis bloques. Cada cuadernillo consta de dos bloques que contienen en total entre 31 y 33 ítems.

5.4. Cuestionarios de contexto del estudiante, la familia y la escuela

Los cuestionarios de contexto del estudiante, la familia, los directores y el profesorado fueron construidos teniendo en cuenta las principales características de la región, por lo que un extenso análisis teórico fue realizado por los organizadores, recabando los principales cuestionamientos de cada sistema educativo alineados con las temáticas educativas y sociales más analizadas. Un detallado resumen de los constructos analizados se puede observar en el reporte técnico de TERCE (UNESCO-OREALC, 2016).

Esta breve introducción sobre el tipo de muestreo y los instrumentos utilizados nos permitirá orientar al lector en la organización de las bases de datos utilizadas en este estudio en particular. Cinco bases de datos provenientes de TERCE fueron utilizadas para este estudio, entre ellas:

(1) Base de datos de logro cognitivo en lectura: Que incluye los datos del estudiante con sus respectivos valores plausibles, pesos senatoriales y respuesta a los ítems de la prueba de lectura. Muestra de 12.415 estudiantes.
(2) Base de datos de logro cognitivo en ciencias naturales: Que incluye los datos del estudiante con sus respectivos valores plausibles, pesos senatoriales y respuesta a los ítems de la prueba de ciencias naturales. Muestra de 13.498 estudiantes.
(3) Base de datos del cuestionario de contexto del estudiante: Que alberga la información contextual del estudiante, como ser sus datos familiares, tipo de escuela y contexto social al que pertenece.
(4) Base de datos del director de la Institución Educativa: Esta base incluye información sobre el tipo de escuela al que accede el estudiante, pesos senatoriales del nivel de escuela y sus características sociales, de infraestructura y pedagógicas. Muestra de 2.802 escuelas.
(5) Base de datos del país: Esta base de datos ad hoc, que contiene información obtenida vía agregación de variables de niveles inferiores como ser: la familia y/o la escuela proveniente de las bases de datos TERCE. Muestra de 15 países de América Latina y el Caribe.

5.5. Variables y estrategia de análisis de datos

Para cada uno de los objetivos se realizaron diferentes análisis de datos y se utilizaron variables específicas, las cuales fueron incluidas en el análisis teniendo en cuenta el objetivo

del estudio. Esta sección está organizada en dos apartados que incluyen la información sobre las variables y análisis estadísticos aplicados. El apartado uno contiene la información para el cumplimiento de los objetivos 1, 3 y 5. Mientras que el apartado dos contiene la información para los objetivos 2 y 4.

5.5.1. Apartado 1

Objetivos: (1) Identificar los países participantes de la evaluación TERCE en América Latina y el Caribe y (3) Determinar la presencia de funcionamiento diferencial del ítem en las pruebas de lectura y ciencias naturales y (5) Comparar la eficacia de los métodos de evaluación del funcionamiento diferencial del ítem.
Variable dependiente: 30 ítems en formato binario provenientes del cuadernillo uno de la prueba de lectura. Además de otros 28 ítems en formato binario provenientes del cuadernillo uno de la prueba de ciencias naturales. En cada uno de los ítems de ambos cuadernillos, la codificación utilizada indica que 0 representa una respuesta incorrecta y 1 representa una respuesta correcta.

5.5.2. Variables de ajuste

(a) Habilidad: Entendida como la media de los cinco valores plausibles. Para una mejor comparación esta variable fue estandarizada a la región con una distribución normal de media 0 y desviación típica 1. Se construyó la variable de habilidad para la materia de lectura, así como para la materia de ciencias naturales.
(b) Género del estudiante: Donde 0 es la codificación para las mujeres y 1 representa a los hombres.

Análisis de datos: Se trabajó con dos bases de datos, una para logro cognitivo del estudiante en lectura, y otra para ciencias naturales. Utilizando el software IBM SPSS se observaron las características descriptivas de la muestra, para lo que se estimaron sobre las bases ponderadas la media, la desviación típica y la varianza de las variables de logro cognitivo en lectura y ciencias naturales, género, ubicación geográfica de la escuela, tipo de institución educativa y lengua del estudiante. Para la identificación del DIF se utilizó un modelo de regresión logística lineal de distribución Bernoulli mediante el software HLM (Raudenbush et al., 2011). El objetivo del análisis fue identificar la presencia de DIF en cada uno de los ítems de las pruebas de lectura y ciencias naturales.

5.5.3. Apartado 2

Objetivos: (2) Establecer las características de cada país, (4) Explorar por medio de las técnicas de análisis jerárquico los condicionantes de logro en las pruebas.
Variable dependiente: Cinco valores plausibles de la prueba de lectura y ciencias naturales de TERCE.
Variables de ajuste: Cinco de ellas son dicotómicas: Género (1 = ser mujer); condición de Indígena (1 = pertenecer a una etnia indígena); condición de Repetición (1 = haber repetido algún curso durante la escolaridad); Trabajo remunerado (1= el estudiante trabaja y recibe una remuneración por esa actividad); y Conexión a Internet (1 = el estudiante dispone de conexión a Internet en el hogar). La última variable es una estimación del Nivel Socioeconómico y Cultural del alumnado (SEC), que es un índice estandarizado construido

por TERCE y compuesto por 17 ítems que recogen información sobre el nivel educativo de los padres, el tipo de trabajo que realizan, el rango de ingresos familiares, así como información sobre los bienes y servicios del barrio en el que se ubica la vivienda, y la disponibilidad de material de lectura del hogar. Los valores del alfa de Cronbach de este índice oscilan entre .8 y .9 según el país (UNESCO-OREALC, 2016).

Dentro de las características del contexto social y demográfico de la escuela se han considerado cuatro variables, dos de ellas dicotómicas: Titularidad (1 = centro privado) y Ruralidad del centro (1 = centro rural). El volumen de recursos de la escuela se estimó mediante el Nivel de Infraestructura de la escuela, que es un índice estandarizado elaborado con información de 10 ítems del cuestionario del director referidos al tipo de instalaciones, equipamientos y servicios con los que cuenta la escuela. Los valores del alfa de Cronbach de este índice oscilan entre .7 y .9 según el país (UNESCO-OREALC, 2016). La cuarta variable es el Nivel Socioeconómico y Cultural de la escuela, estimado como el promedio por centro del SEC del alumnado escolarizado en el mismo. Finalmente, se ha considerado una variable de ajuste a nivel de país, en este caso una estimación del nivel de riqueza, medido a través del Producto Interno Bruto Per Cápita del año 2013 (UNESCO-OREALC & LLECE, 2016a). Análisis de datos: Inicialmente se calcularon los estadísticos descriptivos de todas las variables (media, desviación típica y varianza) y posteriormente se ajustaron dos modelos jerárquico-lineales de interceptos aleatorios y segregados en tres niveles (alumno, escuela y país). Se realizó un análisis para cada materia evaluada y se siguió una estrategia de modelización donde el primer modelo sin predictores (modelo nulo) fue utilizado para conocer la distribución de la varianza en cada nivel. Mientras que el segundo modelo incluyó a las variables de ajuste con el fin de conocer algunos condicionantes de logro cognitivo. En el ajuste se empleó el método de estimación de máxima verosimilitud con errores típicos robustos usando el Programa HLM 7.01 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). En todos los análisis se utilizaron los pesos senatoriales proveídos por TERCE, los cuales están diseñados para que todos los países, independientemente del tamaño de su población, contribuyan por igual en el análisis de resultados (OREALC-UNESCO, 2016).

6. RESULTADOS

Objetivos

6.1 Identificación de países participantes Evaluación TERCE América Latina y Caribe

Tabla 1

Datos de la distribución media de la muestra y la población

	Tamaño de la muestra	Tamaño de la población
Argentina	3.639	760.311
Brasil	2.983	2.043.907
Chile	5.044	262.569
Colombia	4.308	1.046.752
Costa Rica	3.520	105.218
República Dominicana	3.661	184.352
Ecuador	4.818	416.114
Guatemala	4.056	227.627
Honduras	3.880	170.860
México	3.618	2.599.591
Nicaragua	3.726	115.937
Panamá	3.413	66.069
Paraguay	3.222	118.744
Perú	4.789	609.457
Uruguay	2.799	52.096
Nuevo León (México)	4.197	115.783
Total	61.673	8.895.387

Tabla 2

Distribución de la muestra por materia evaluada

	Ciencias Naturales	Matemática	Lectura
Estudiantes	61938	63750	60949
Escuelas	2955	2934	2954

TERCE evalúa a gran número de naciones de América Latina y el Caribe, en la tabla 1 es posible observar la presencia de 15 países más el estado de Nuevo León (México), evaluados en las materias de ciencias naturales, matemática, lectura y escritura. Si bien en este estudio estamos centrados en la muestra perteneciente al 6to grado de la educación obligatoria y las materias de Lectura y Ciencias Naturales (tabla 2), TERCE también ha evaluado a estudiantes de 3er. grado. Los datos observados en las tablas 1 (col.1) y 2 se refieren a datos sin la inclusión de pesos muestrales, es importante recordar que TERCE utiliza pesos muestrales

que tienen el objetivo de representar a la población total de cada país participante (Tabla 1, Col.2). El estudio TERCE si bien aplica la evaluación a aproximadamente una media de 60.000 estudiantes (tabla 1, col.1) que asisten a cerca de 3.000 escuelas, estos representan a un total de casi nueve millones (tabla 1, col. 2) de estudiantes de la educación primaria (obligatoria) de las 16 naciones participantes.

6.2. Características de países participantes en la evaluación TERCE.

Tabla 3

Descriptivos del promedio de logro cognitivo de TERCE desagregado por materias

	Ciencias Naturales				*Lectura*			
País	Media	Desviación estándar	Varianza	Muestra	Media	Desviación estándar	Varianza	Muestra
Argentina	699,61	85,39	7291,66	874,39	705,75	91,59	8387,88	820,43
Brasil	698,46	77,27	5969,92	854,92	722,39	88,16	7772,98	834,52
Chile	779,01	96,73	9356,74	845,25	776,75	91,89	8443,06	807,96
Colombia	734,58	82,82	6858,60	815,74	735,26	81,17	6588,14	834,61
Costa Rica	757,13	71,05	5048,35	837,94	757,68	74,57	5560,60	793,81
República Dominicana	641,20	63,43	4023,81	833,41	634,71	68,09	4636,11	851,28
Ecuador	708,39	85,92	7381,96	821,64	684,76	86,82	7537,41	806,23
Guatemala	683,96	73,52	5405,01	824,71	672,59	81,57	6653,63	851,00
Honduras	668,76	78,27	6126,88	856,04	661,99	77,34	5980,89	865,97
México	733,90	80,22	6434,74	821,40	735,60	91,67	8402,86	807,54
Nicaragua	659,72	64,96	4219,18	876,08	665,26	75,53	5705,23	822,91
Panamá	670,05	83,51	6973,13	856,76	673,07	89,02	7924,85	818,48
Paraguay	649,99	84,14	7080,25	834,90	654,13	90,74	8233,00	845,77
Perú	707,04	83,85	7030,75	833,62	702,05	91,93	8451,83	831,60
Uruguay	722,91	106,14	11265,58	869,82	729,13	103,93	10801,36	822,48

TERCE mantiene una escala de logro cognitivo con media de 700 puntos y 100 puntos de desviación típica, por lo que al observar la tabla 3 la distribución de los países se encuentra organizado en torno a los 700 puntos. Los datos presentados pertenecen a la media de los cinco valores plausible que ofrece la evaluación TERCE ponderado con el peso senatorial, que permite la representación de la población total de cada país. Claramente se pueden observar los países que lideran el ranking de TERCE en el siguiente orden: Costa Rica, Nuevo León, México, Colombia, México y Chile, recordemos que estas naciones son las que mayor experiencia en la participación en pruebas de evaluación educativa presentan. En la cola de las puntuaciones de logro en ciencias naturales, se encuentran República Dominicana, Paraguay y Panamá.

Los resultados de logro cognitivo a su vez han sido segregados por el género (tabla 4), el tipo de escuela (tabla 5), la ubicación geográfica (tabla 6) y la lengua materna (tabla 7). Es importante destacar que no existen grandes diferencias en relación al género del estudiante, pero el hecho de pertenecer a una escuela pública, de zona rural y hablar una lengua materna indígena presenta diferencias de hasta una desviación típica en el rendimiento del

estudiantado, lo que puede llegar a representar hasta un grado de diferencia entre los estudiantes que pertenecen a una escuela urbana, privada que hablan la lengua castellana.

Tabla 4

Promedio del logro cognitivo en TERCE desagregado por género

	Ciencias Naturales				*Lectura*			
	Media	*Desviación estándar*	*Varianza*	*Muestra*	*Media*	*Desviación estándar*	*Varianza*	*Muestra*
Femenino	705,79	88,33	7801,35	6723	701,73	95.21	9064.13	6106
Masculino	701,63	92,06	8474,81	6774	698,56	94.23	8879.98	6308

Tabla 5

Promedio del logro cognitivo en TERCE desagregado por tipo de escuela

	Ciencias Naturales				*Lectura*			
	Media	*Desviación estándar*	*Varianza*	*Muestra*	*Media*	*Desviación estándar*	*Varianza*	*Muestra*
Pública	690,74	84,84	7197,54	11010	684,86	89,16	7498,77	10113
Privada	761,08	91,08	8294,83	2488	767,2	89,17	7950,87	2301

Tabla 6

Promedio del logro cognitivo en TERCE desagregado por ubicación geográfica

	Ciencias Naturales				*Lectura*			
	Media	*Desviación estándar*	*Varianza*	*Muestra*	*Media*	*Desviación estándar*	*Varianza*	*Muestra*
Urbana	711,13	88,09	7759,8	9933	719,32	92,12	8486,5	9013
Rural	666,28	85,5	7309,83	3564	649,24	81,83	6696,61	3401

Tabla 7

Promedio del logro cognitivo en TERCE desagregado por la lengua materna

	Ciencias Naturales			
	Media	*Desviación estándar*	*Varianza*	*Muestra*
Lengua materna de la prueba	708,79	89,84	8071,88	11649
Lengua materna indígena o distinta a la lengua de la prueba	665,02	91,16	8309,34	962
	Lectura			
	Media	*Desviación estándar*	*Varianza*	*Muestra*
Lengua materna de la prueba	704,03	93,92	8820,5	11424
Lengua materna indígena o distinta a la lengua de la prueba	655,03	92,31	8521,47	991

6.3. Funcionamiento diferencial del ítem en las pruebas de ciencias naturales.

Para identificar la presencia del funcionamiento diferencial del ítem se procedió a utilizar un modelo de regresión logística de distribución Bernoulli, dado que los datos provienen de una estructura anidada que incluye el efecto de las escuelas y los sistemas educativos el uso de modelos jerárquico-lineales se hace necesario, ya que estos modelos respetan la estructura anidada de los datos en educación. A continuación, en la tabla 8 se pueden observar los coeficientes del modelo que incluye en el nivel 1 las variables de clasificación (género) y el nivel de habilidad en la variable de ciencias naturales. Recordemos que la variable dependiente en este punto es los ítems binarios del cuadernillo número 1 de la prueba de lectura y la prueba de ciencias naturales.

Iniciando nuestro análisis desde el lado derecho de la tabla 8 la primera columna se refiere al número que identifica a cada ítem de los cuales 28 son ítem binarios y 2 son ítems de crédito parcial. Dada la técnica utilizada, este estudio procedió a analizar los ítems de formato binario. La columna número dos indica el valor del coeficiente de regresión para el intercepto de la variable donde valores positivos indican que las mujeres son más propensas a presentar una respuesta favorable en el ítem, mientras que valores negativos se refieren a la propensión en el caso de los varones a responder positivamente en ese ítem. Entre los principales resultados se observa la presencia de DIF en nueve de los 28 ítems de la prueba de ciencias naturales, de los cuales cinco favorecen a los hombres y cuatro de los ítems favorecen a las mujeres.

Tabla 8

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Ciencias Naturales

Step 1 Detección del DIF uniforme									
	DIF de género					Efecto aleatorio			
Ítem	Coeficientes	p-value	odds ratio	(IC)	A quien favorece	DS	Componente de varianza	Chi-cuadrado (gl)	p-value
1	-0,040039	0,413	0,96075	(0.868-1.064)	*	0.05410	0,00293	15.69473 (14)	0,332
2	0,144211	0,017	1,155128	(1.031-1.295)	mujeres	0.11482	0,01318	21.91023 (14)	0,080
3	-0,044113	0,396	0,956846	(0.859-1.066)	*	0,03335	0,00111	4.57866(14)	0,500
4	-0,016294	0,749	0,983838	(0.884-1.095)	*	0,09763	0,00953	18.75172 (14)	0,174
5	0,187361	0,001	1,206063	(1.092-1.332)	mujeres	0,07021	0,00493	18.32716 (14)	0,192
6	0,007366	0,894	1,007393	(0.897-1.132)	*	0,08054	0,00649	9.70090 (14)	0,500
7	-0,087034	0,129	0,916646	(0.817-1.029)	*	0,06923	0,00479	19.29364 (14)	0,154
8	-0,017677	0,708	0,982478	(0.890-1.085)	*	0,03256	0,00106	8.30249 (14)	0,500

9	-0,159336	0,004	0,85271	(0.771-0.943)	varones	0,05722	0,00327	15.88112 (14)	0,320
10	-0,119958	0,064	0,886958	(0.780-1.008)	varones	0,1499	0,02247	25.14333 (14)	0,033
11	0,048287	0,383	1,049472	(0.935-1.177)	*	0,10339	0,01069	17.66822 (14)	0,222
12	0,094311	0,159	1,098901	(0.959-1.259)	*	0,16726	0,02798	26.11881 (14)	0,025
13	-0,005795	0,919	0,994222	(0.882-1.121)	*	0,11299	0,01277	15.23681 (14)	0,362
14	0,060572	0,327	1,062444	(0.935-1.208)	*	0,14922	0,02227	26.29020 (14)	0,024

Nota. (IC) indica el intervalo de confianza, (DS) desviación estándar, (gl) grados de libertad.

Tabla 8 *Continuación*

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Ciencias Naturales

	Step 1 Detección del DIF uniforme								
	DIF de género						Efecto aleatorio		
Ítem	Coeficientes	p-value	odds ratio	(IC)	A quien favorece	DS	Componente de varianza	Chi-square (gl)	p-value
15	*	*	*	*	*	*	*	*	*
16	-0,265109	<0.001	0,767122	(0.675-0.872)	*	0,11438	0,01308	17.70124 (14)	0,220
17	0,012768	0,801	1,01285	(0.910-1.127)	*	0,07883	0,00621	11.69148 (14)	0,500
18	0,046997	0,3	1,048118	(0.954-1.151)	*	0,02846	0,00081	5.77771 (14)	0,500
19	-0,264493	<0.001	0,767595	(0.671-0.878)	varones	0,17325	0,03001	31.31477 (14)	0,005
20	0,021256	0,679	1,021483	(0.917-1.138)	*	0,0928	0,00861	20.10097 (14)	0,127
21	-0,075383	0,157	0,927388	(0.832-1.033)	*	0,09012	0,00812	13.90594 (14)	0,500
22	-0,445878	<0.001	0,640262	(0.566-0.724)	varones	0,135	0,01822	23.68940 (14)	0,050
23	-0,006196	0,907	0,993823	(0.889-1.111)	*	0,10488	0,011	20.31065 (14)	0,120
24	0,281632	<0.001	1,325291	(1.189-1.477)	mujeres	0,09803	0,00961	19.52041 (14)	0,146
25	0,060705	0,206	1,062586	(0.963-1.172)	*	0,05653	0,0032	9.75794 (14)	0,500
26	-0,253945	<0.001	0,775734	(0.699-0.861)	varones	0,0367	0,00135	10.06399 (14)	0,500
27	0,299354	<0.001	1,348987	(1.210-1.503)	mujeres	0,08671	0,00752	9.16958 (14)	0,500
28	0,003798	0,936	1,003806	(0.909-1.109)	*	0,0469	0,0022	13.30477 (14)	0,500

29	-0,082645	0,149	0,920678	(0.820-1.034)	*	0,10458	0,01094	19.47147 (14)	0,147
30	*	*	*	*	*	*	*	*	*

Nota. (IC) indica el intervalo de confianza, (DS) desviación estándar, (gl) grados de libertad.

Recordemos que valores de significación inferiores a .05 son observados como valores significativos, por lo que el p. value <.05 es observado en los ítems 2, 5, 9, 10 (tabla 8) y los ítems 19, 22, 24, 26 y 27 (tabla 8 continuación). Las columnas del 7 al 10 indican los valores del efecto aleatorio, es decir, la variabilidad observada entre las agrupaciones. En general, se observa que la presencia del DIF varía entre los países participantes, dado que se observan valores significativos y no significativos en los resultados.

Tabla 9

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Lectura

	Step 1 Detección del DIF uniforme								
	DIF de género					Efecto aleatorio			
Ítem	Coeficientes	p-value	odds ratio	(IC)	A quien favorece	DS	Componente de varianza	Chi-cuadrado (gl)	p-value
1	-0,018660	0,891	0,981513	(0.800-1.204)	*	0.251181	0,06341	30.37266 (14)	0,07
2	-0.154494	0.296	0.856849	(0.631-1.163)	*	0.25310	0.06406	28.42417 (14)	0.013
3	0.182367	0.343	1.200054	(0.806-1.787)	*	0.48495	0.23518	56.40754 (14)	<0.001
4	-0.111043	0.506	0.894900	(0.631-1.269)	*	0.40696	0.16562	47.88655 (14)	<0.001
5	0.207217	0.172	1.230249	(0.903-1.676)	*	0.08366	0.00700	15.88963 (14)	0.320
6	0.021466	0.889	1.021698	(0.740-1.411)	*	0.28392	0.08061	28.15201 (14)	0.014
7	-0.113581	0.371	0.892632	(0.686-1.162)	*	0.15170	0.02301	12.85637 (14)	>.500
8	-0.131319	0.363	0.876938	(0.650-1.183)	*	0.10645	0.01133	11.04753 (14)	>.500
9	-0.041696	0.792	0.959161	(0.688-1.337)	*	0.11833	0.01400	11.77313 (14)	>.500
10	-0.065852	0.685	0.936270	(0.666-1.317)	*	0.28884	0.08343	25.28568 (14)	0.032
11	0.046136	0.712	1.047217	(0.806-1.361)	*	0.08990	0.00808	8.71317 (14)	>.500
12	0.125642	0.523	1.133877	(0.751-1.711)	*	0.56235	0.31624	80.06078 (14)	<0.001
13	0.023708	0.865	1.023991	(0.763-1.373)	*	0.25358	0.06430	27.13465 (14)	0.018
14	0.036620	0.792	1.037299	(0.774-1.390)	*	0.17739	0.03147	22.80794 (14)	0.063

Nota. (IC) indica el intervalo de confianza, (DS) desviación estándar, (gl) grados de libertad.

Tabla 9 *Continuación*

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Lectura

	Step 1 Detección del DIF uniforme DIF de género						Efecto aleatorio		
Ítem	Coeficientes	p-value	odds ratio	(IC)	A quien favorece	DS	Componente de varianza	Chi-cuadrado (gl)	p-value
15	0.102063	0.566	1.107453	(0.763-1.607)	*	0.46243	0.21384	58.70687 (14)	<0.001
16	0.137086	0.595	0.928068	(0.692-1.245)	*	0.26654	0.07104	29.83032 (14)	0.008
17	-0.252693	0.058	0.776706	(0.598-1.009)	*	0.14934	0.02230	17.63643 (14)	0.223
18	0.154533	0.365	1.167113	(0.819-1.663)	*	0.39936	0.15949	46.25983 (14)	<0.001
19	-0.065773	0.641	0.936344	(0.696-1.259)	*	0.08625	0.00744	14.70776 (14)	0.398
20	-0.185319	0.164	0.830839	(0.634-1.089)	*	0.11149	0.01243	14.66737 (14)	0.401
21	-0.057185	0.777	0.944419	(0.617-1.446)	*	0.59826	0.35792	83.62219 (14)	<0.001
22	-0.147958	0.317	0.862467	(0.635-1.171)	*	0.24878	0.06189	27.70283 (14)	0.016
23	-0.027519	0.864	0.972857	(0.693-1.366)	*	0.36510	0.13330	41.55360 (14)	<0.001
24	-0.186284	0.219	0.830038	(0.608-1.132)	*	0.33217	0.11034	43.18511 (14)	<0.001
25	-0.012322	0.924	0.987754	(0.751-1.299)	*	0.17821	0.03176	22.32642 (14)	0.072
26	-0.281065	0.037*	0.754979	(0.581-0.981)	varones	0.04688	0.00220	9.72312 (14)	>.500
27	0.035213	0.808	1.035841	(0.764-1.405)	*	0.24008	0.05764	27.24426 (14)	0.018
28	-0.153205	0.271	0.857954	(0.644-1.143)	*	0.19766	0.03907	23.57138 (14)	0.051
29	0.03588	0.848	1.036540	(0.699-1.537)	*	0.51072	0.26083	66.79154 (14)	<0.001
30	-0.019536	0.899	0.980654	(0.708-1.358)	*	0.37266	0.13887	45.73350 (14)	<0.001

Nota. (IC) indica el intervalo de confianza, (DS) desviación estándar, (gl) grados de libertad.
En cuanto a los resultados de la prueba de lectura, se observa que el ítem 26 del cuadernillo uno, presenta funcionamiento diferencial del ítem y este favorece a los estudiantes del sexo masculino.

6.4. Exploración, vía modelos multinivel, de los condicionantes de logro en las pruebas.

Los resultados obtenidos del modelo jerárquico lineal nulo identificaron el nivel de variabilidad que se observa en los diferentes niveles evaluados (tabla 9, col. 2), una vez incluida las variables explicativas en el modelo I de contexto de las escuelas (tabla 9, col. 3)

se pudo constatar una reducción de la variabilidad del 2% a nivel del estudiante, 48% a nivel de las escuelas y del 49% a nivel de los sistemas educativos, lo que indica que las variables tanto a nivel de las escuelas y el sistema educativo han explicado casi el 50% de la varianza del modelo nulo. Entre las principales variables explicativas se encuentran el hecho de haber repetido, ser mujer y tener un trabajo remunerado. Mientras que variables que favorecen el logro en ciencias naturales se encuentran a nivel de las escuelas. Escuelas privadas urbanas con familias de niveles socioeconómico alto, presentan un mejor rendimiento en la prueba de ciencias naturales a diferencia de las escuelas públicas, de zonas rurales.

Tabla 10

Modelos Multinivel de condicionantes de logro Cognitivo (TERCE)

	Ciencias Naturales		*Lectura*	
	Modelo Nulo	*Modelo I Contexto*	*Modelo Nulo*	*Modelo I Contexto*
	β	**β(SE)**	**β**	**β(SE)**
Intercepto	696.27 (9.66) ***	636.61 (23.14) ***	695.11 (11.07) ***	671.41 (9.29) ***
Nivel 1 (Alumnado)				
Ser del sexo femenino	-	-1.96 (1.96)	-	6.94 (1.59)***
Pertenecer a una etnia indígena	-	-0.01 (3.58)	-	-4.90 (2.79)
Trabajo infantil remunerado	-	-10.73 (6.46)	-	-5.74 (5.21)
Haber repetido	-	-22.51 (5.53)***	-	-20.30 (2.15)***
SEC Alumno	-	5.95 (2.01)**	-	18.38 (1.39)***
Nivel 2 (Centro)				
SEC Escuela	-	34.94 (4.82)***	-	26.49 (3.23)***
Escuela privada		14.30 (6.14)*	-	10.84 (3.55)**
Escuela rural	-	-21.84 (11.43)†	-	-1.22 (3.34)
Infraestructura	-	7.81 (3.55)**	-	8.88 (2.57)***
Nivel 3 (País)				
PIB x Cápita 2013 (mil. $)	-	0.79 (0.77)	-	0.65 (0.75)
Distribución de la varianza				
Dentro del centro	5903.34	5783.13	5821.932	5572.44
Entre los Centros	3560.36	2196.31	3018.548	957.47
Entre los Países	1465.71	745	1857.70	585.34
Total	10929.41	8724.44	*10698.17*	*7115.25*
Porcentaje de varianza explicada				
Dentro del Centro		2%	-	4%
Entre los Centros		38%	-	68%
Entre los Países		49%	-	68%
Total		20%	-	33%

Nota: †p < .10; *p < .05; **p < .01; ***p < .001

6.5. Comparación de eficacia de métodos de evaluación del funcionamiento diferencial del ítem.

Un exhaustivo análisis teórico del reporte técnico de TERCE se llevó a cabo con el fin de identificar si el programa de evaluación desarrolla análisis de DIF y cuáles son las técnicas utilizadas. Los principales resultados observados se refieren a que (1) el DIF es realizado con la estrategia de análisis de Mantel-Haenszel, técnica descripta en la introducción teórica. (2) el DIF se realiza por agrupaciones como el género y el país y (3) el análisis del DIF no es una técnica psicométrica decisiva para la inclusión de un ítem dentro de la prueba cognitiva final (UNESCO-OREALC, 2016, p. 252). Por lo que en este estudio se procedió a utilizar una técnica de detección del DIF vanguardista que permite la inclusión de la variabilidad natural del contexto educativo, obteniendo mejores resultados que la técnica del Mantel-Haenszel.

Tabla 11

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Ciencias Naturales TERCE

Ítem No.	Presencia/Ausencia	Tipo de DIF	A quien favorece	Tipo de técnica de detección	% de detección Método I	% de detección Método Mantel-Haenszel
1	Ausencia de DIF	*	*			
2	Presencia de DIF	Uniforme	mujeres	métodos multinivel		
3	Ausencia de DIF	*	*			
4	Ausencia de DIF	*	*			
5	Presencia de DIF	Uniforme	mujeres	métodos multinivel		
6	Ausencia de DIF	*	*			
7	Ausencia de DIF	*	*			
8	Ausencia de DIF	*	*			
9	Presencia de DIF	Uniforme	varones	métodos multinivel		
10	Presencia de DIF	Uniforme	varones	métodos multinivel	32%	0%
11	Ausencia de DIF	*	*			
12	Ausencia de DIF	*	*			
13	Ausencia de DIF	*	*			
14	Ausencia de DIF	*	*			
15		Ítem de crédito parcial				
16	Ausencia de DIF	*				
17	Ausencia de DIF	*				
18	Ausencia de DIF	*				
19	Presencia de DIF	Uniforme	varones	métodos multinivel		
20	Ausencia de DIF	*				

21	Ausencia de DIF	*		
22	Presencia de DIF	Uniforme	varones	métodos multinivel
23	Ausencia de DIF	*		
24	Presencia de DIF	Uniforme	mujeres	métodos multinivel
25	Ausencia de DIF	*		
26	Presencia de DIF	Uniforme	varones	métodos multinivel
27	Presencia de DIF	Uniforme	mujeres	métodos multinivel
28	Ausencia de DIF	*	*	*
29	Ausencia de DIF	*	*	*
30	Ítem de crédito parcial			

Tabla 12 continuación

Distribución de los ítems de acuerdo a la presencia de DIF – Prueba de Lectura TERCE

Ítem No.	Presencia/Ausencia	Tipo de DIF	A quien favorece	Tipo de técnica de detección	% de detección	% de detección Método Mantel-Haenszel
1	Ausencia de DIF	*	*	*		
2	Ausencia de DIF	*	*	*		
3	Ausencia de DIF	*	*	*		
4	Ausencia de DIF	*	*	*		
5	Ausencia de DIF	*	*	*		
6	Ausencia de DIF	*	*	*		
7	Ausencia de DIF	*	*	*		
8	Ausencia de DIF	*	*	*		
9	Ausencia de DIF	*	*	*		
10	Ausencia de DIF	*	*	*		
11	Ausencia de DIF	*	*	*	1%	0%
12	Ausencia de DIF	*	*	*		
13	Ausencia de DIF	*	*	*		
14	Ausencia de DIF	*	*	*		
15	Ausencia de DIF	*	*	*		
16	Ausencia de DIF	*	*	*		
17	Ausencia de DIF	*	*	*		
18	Ausencia de DIF	*	*	*		
19	Ausencia de DIF	*	*	*		
20	Ausencia de DIF	*	*	*		
21	Ausencia de DIF	*	*	*		
22	Ausencia de DIF	*	*	*		

23	Ausencia de DIF	*	*	*
24	Ausencia de DIF	*	*	*
25	Ausencia de DIF	*	*	*
26	Presencia de DIF	Uniforme	varones	métodos multinivel
27	Ausencia de DIF	*	*	*
28	Ausencia de DIF	*	*	*
29	Ausencia de DIF	*	*	*
30	Ausencia de DIF	*	*	*

CONCLUSIÓN

Al hacer públicos los resultados de las evaluaciones educativas estandarizadas, se parte del supuesto de que las variables y factores que son tenidos en cuenta para el análisis presentan buenos indicadores en cuanto a su calidad métrica. Diversas son las implicaciones de los resultados obtenidos por las evaluaciones educativas, recientes estudios publicados, nos dan un panorama sobre la educación en América Latina y el esfuerzo colaborativo de los países de la región para el desarrollo de una educación de calidad (Woitschach, Fernández-Alonso, Martínez-Arias y Muñiz, 2017). Todos los informes y trabajos citados en la literatura sobre el estudio TERCE, se basan en la idea de que las variables de resultados individuales que funcionan como variable dependiente, están construidas de forma consistente con los estándares de calidad internacionales (AERA, APA, & NCME, 2014); en general los programas publican informes técnicos donde muestran las propiedades métricas de los ítems (Martin & Mullis, 2012; M. O. Martin, Mullis, & Hooper, 2016; OECD, 2014; ORELAC/UNESCO-LLECE, 2010). Si bien el Informe de Resultados de TERCE, indica que los ítems de las pruebas pasaron por un proceso de adaptaciones nacionales para dar cuenta de su adecuación lingüística y de contenido a los distintos países participantes en el estudio (LLECE, 2016b, p. 71). Los valores obtenidos tanto en el proceso de adecuación lingüística como en el proceso de evaluación de la invarianza de las medidas en relación al DIF de genero carecen de notoriedad en los informes técnicos publicados.

Los resultados de este estudio se encuentran en consonancia con resultados obtenidos en el reporte técnico de TERCE, en lo que a variables descriptivas se refiere. Latinoamérica es una de las regiones con mayor desigualdad y donde el tamaño del efecto bruto del centro es bastante alto (Woitschach, Fernández-Alonso, Martínez-Arias, & Muñiz Fernández, 2017) por lo que se espera que las variables de contexto tengan un fuerte impacto en los resultados educativos. Los análisis realizados en este estudio presentan una validez comparativa ya que los datos representan a casi nueve millones de estudiante de 16 naciones de América Latina y el Caribe. Los estudios de la UNESCO no solo se han caracterizado por la estabilidad de sus aplicaciones, ya que son realizados desde la década del 90´, sino que además se caracterizan por ser una fuente de información contextualizada al ámbito latinoamericano, que a efectos metodológicos y de construcción de políticas públicas, son más acertados que los datos que otros programas de evaluación educativa ofrecen en un pequeño muestreo de países latinoamericanos participantes. El impacto de las variables sociodemográficas impacta no solo en el acceso a las oportunidades de aprendizaje, sino que también en la forma en

como los estudiantes responden a la prueba. América Latina en este sentido, es de las regiones con mayores niveles de desigualdad social y educativa (UNESCO-OREALC, 2013).

Tomando en cuenta este panorama y el continuo avance de la cultura de la evaluación que no solo es una tarea propia de países industrializados, sino que también ha llegado a ser una herramienta de uso habitual para la construcción de políticas educativas en países en vías de desarrollo (Smith, 2014), requiere a su vez del soporte metodológico que asegure la validez de las medidas obtenidas y para el aseguramiento de las medidas obtenidas, rigurosos análisis psicométricos deben ser realizados. El programa de evaluación TERCE sigue un riguroso sistema metodológico para la construcción de las pruebas de evaluación cognitiva, pero a su vez, se observa la ausencia de los resultados del análisis del funcionamiento diferencial del ítem de acuerdo a variables sociodemográficas como ser el género del estudiante.

Los primeros objetivos de este estudio se orientaron a identificar los datos descriptivos de las variables de rendimiento académico segregados por países, zonas geográficas, género, tipo de escuela y lengua materna. Los resultados una vez más indicaron la fuerza explicativa de variables sociodemográficas como el tipo de escuela, la zona geográfica y la lengua materna. Por lo que en la muestra latinoamericana el hecho de asistir a una escuela pública, en una zona rural y hablar una lengua distinta al castellano representa un impacto negativo de media desviación típica. Estos resultados se encuentran en la misma línea de datos obtenidos en las evaluaciones PERCE y SERCE (UNESCO-OREALC & LLECE, 2000, 2010; Woitschach, Fernández-Alonso, Martínez-Arias, & Muñiz Fernández, 2017), así como en estudios como los realizados por PISA (OECD, 2002, 2004, 2010).

En cuanto a la presencia de variabilidad en las respuestas a los ítems del cuadernillo número uno de ciencias naturales y de lectura, donde se ha observado la presencia de funcionamiento diferencial del ítem 10 de los ítems, lo que indica un alto grado de variabilidad en la forma en como los estudiantes responden a la prueba de acuerdo al género al que pertenecen. Si bien no se ha observado una pauta fuerte de beneficio a un solo tipo de género, la presencia de variabilidad es una amenaza al momento de interpretar las puntuaciones, dado que el objetivo primordial de estas evaluaciones es el de la comparabilidad, por lo que la invarianza de las medidas es una característica requerida.

La estrategia utilizada para la detección del DIF se corresponde con una de las más vanguardistas técnicas que permite la inclusión de variabilidad proveniente de diversos niveles de agregación y ha sido utilizada en diversos estudios de investigación (Chen & Zumbo, 2017). La eficacia de la técnica ha permitido la identificación de la presencia de DIF en al menos el 32% de los ítems de la prueba de ciencias y el 1% de los ítems de la prueba de lectura, lo que podría tener un impacto tanto en la construcción como en la interpretación de las puntuaciones.

En cuanto a las variables que explican la variabilidad de las puntuaciones, los resultados siguen la línea de las investigaciones desarrolladas en países latinoamericanos, que destacan que aspectos contextuales tienen un alto impacto en el rendimiento académico (Woitschach et al., 2017). Los sistemas educativos de América Latina podrían estar más concentrados en el control del impacto de variables sociodemográficas, antes que en ofrecer

una educación de calidad como es el caso de países industrializados. En América Latina variables como el hecho de tener un trabajo infantil remunerado es una realidad que impacta negativamente en el logro académico, en la misma línea que el hecho de pertenecer a una escuela rural y a una etnia indígena, resultados que se encuentran en consonancia con informes realizados en la región (Suárez-Enciso, Elías, & Zarza, 2016).

Finalmente, es importante resaltar que las limitaciones del estudio se encuentran vinculados a la disponibilidad de variables explicativas. Dado que los análisis son realizados con las bases de datos proveídas por la organización. Los estudios a futuro se orientan a obtener evidencias explicativas del funcionamiento diferencial del ítem observado. Con el fin de aportar evidencias que permitan la mejora de la evaluación educativa y el sustento de la construcción de políticas públicas en educación para países de América Latina y el Caribe.

BIBLIOGRAFÍA

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en Ciencias Sociales y de la Salud*. Madrid: Editorial Sïntesis.

Addey, C. (2019). Researching inside the international testing machine: PISA parties, midnight emails and red shoes. In B. Maddox (Eds.), *International large-scale assessments in education. Insider research perspectives* (13-29). Great Britain: Bloomsbury Academic.

AERA, APA, & NCME. (2014). *Standars for Educational and Psychological Testing*. Washington: AERA.

Balluerka, N., Gorostiaga, A., Gomez-Benito, J., & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, *22*(4), 1018-1025.

Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J.L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology*, *10*(2), 71-79. doi:10.1027/1614-2241/a000076

Chen, M.Y., & Zumbo, B.D. (2017). Ecological Framework of Item Responding as Validity Evidence: An application of Multilevel DIF Modeling using PISA Data. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 53-68). New York, NY: Springer International

Elousa, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y Comprensión *Relieve*, *12*(2).

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teachers College Record*, *117*(1), 1-28.

Ferández-Alonso, R. (2004). *Evaluación del rendimiento matemático.* (Tesis Doctoral), Universidad de Oviedo, Oviedo, Asturias.

Gomez Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benitez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109. doi:10.7334/psicothema2017.183

Holmes, W., Finch, M. E. H., & French, B. F. (2016). Recursive Partitioning to Identify Potential Causes of Differential Item Functioning in Cross-National Data. *International Journal of Testing, 16*, 21-53. doi:10.1080/15305058.2015.1039644

Joncas, M., & Foy, P. (2012). Sample Design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS and PIRLS International Study Centre, Boston College.

LLECE. (2014). *Primera entrega de resultados Tercer Estudio Regional Comparativo y Explicativo Terce*. Retrieved from Santiago de Chile:

Maddox, B. (2019). *International large-scale assessments in education. Insider research perspectives*. Great Britain: Bloomsbury Academic.

Maddox, B., Zumbo, B. D., Tay-Lim, B., & Qu, D. I. (2015). An Anthropologist Among the Psychometricians: Assessment Event, Ethnography, and Differential Item Functioning in the Mongolian Gobi. *International Journal of Testing, 15*, 291-309. doi:10.1080/15305058.2015.1017103

Martínez-Arias, R. (2006). La metodología de los estudios PISA. *Revista de Educación*, 111-129.

Muñiz, J. (2003). *Teoría Clásica de los Test*: Ediciones Pirámide.

Navas, M. J. (1994). Teoría Clásica de los Test versus Teoría de la Respuesta al Ítem. *Psicológica, 15*, 175-208.

OECD. (2002). *Results from PISA 2000. Reading for Change: Performance and Engagement across Countries*. París: OECD Publishing.

OECD. (2004). *Learning for Tomorrow's World: First Results from PISA 2003*: OECD Publishing.

OECD. (2005). *PISA 2003 techincal report*. Paris: OECD Publishing.

OECD. (2009). *PISA 2006 techincal report*. Paris: OECD Publishing.

OECD. (2010). *PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes (Volume II)* doi:10.1787/9789264091504-en

OECD. (2012). *PISA 2009 techincal report*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring Crossing Differential Item Functioning by Gender in Mathematics Assessment. *International Journal of Testing, 15*(337-355).

Smith, W. (2014). *The global expansion of the testing culture: National tetsing policies and reconstruction of education.* (Doctor of Philosophy), The Pennsylvania State University.Retrievedfrom https://etda.libraries.psu.edu/files/final_submissions/10282

Suárez-Enciso, S., Elías, R., & Zarza, D. (2016). Factores asociados al Rendimiento Académico de Estudiantes de Paraguay: Un análisis de los Resultados del TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 113-133. doi:10.15366/reice2016.14.4.006

Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75. doi:10.3102/10769986027001053

UNESCO-OREALC. (2013). *Situación educativa de América Latina y el Caribe: Hacia la educación de calidad para todos al 2015*. Santiago de Chile: UNESCO.

UNESCO-OREALC. (2016). *Reporte Técnico Tercer Estudio Regional Comparativo y Explicativo. TERCE*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2000). *Primer estudio internacional comparativo sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la educación básica. Segundo Informe*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2010). *SERCE. Factores asociados al logro cognitivo de los estudiantes de América Latina y el Caribe*. Santiago de Chile: UNESCO.

van den Noortgate, W., & de Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443-464.

Woitschach, P., Fernández-Alonso, R., Martínez-Arias, R., & Muñiz Fernández, J. (2017). Influencia de los centros escolares sobre el rendimiento académico en Latinoamérica. R*evista de Psicología y Educación, 12*(2), 138-154. doi:10.23923/rpye2017.12.152

Zumbo, B. D. (2007). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 45-79). The Netherlands: Elsevier Science.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, *5*, 1-23.