UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE PSICOLOGÍA
Departamento de Psicobiología y Metodología de las Ciencias del Comportamiento

TESIS DOCTORAL

Evaluaciones educativas a gran escala en Latinoamérica:TERCE

Large scale educational assessments in Latin America:TERCE

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Pamela Raquel Woitschach Mendoza

Directores

Rosario Martínez-Arias
Rubén Fernández-Alonso
José Muñiz-Fernández

Madrid
Ed. electrónica 2019

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE PSICOLOGÍA

Departamento de Psicobiología y Metodología en Ciencias del Comportamiento

TESIS DOCTORAL

EVALUACIONES EDUCATIVAS A GRAN ESCALA EN LATINOAMÉRICA: TERCE

LARGE SCALE EDUCATIONAL ASSESSMENTS IN LATIN AMERICA: TERCE

MEMORIA PARA OPTAR AL GRADO DE DOCTOR PRESENTADA POR

Pamela Raquel Woitschach Mendoza

Directores

Rosario Martínez-Arias

Rubén Fernández-Alonso

José Muñiz-Fernández

Madrid, 2018

En honor a mis padres Juan Alberto y Luciana Aurora

AGRADECIMIENTOS

Cada etapa de este recorrido estuvo marcada y escenificada por profesionales y personas de bien, que en suma caracterizan a la comunidad académica y científica de las áreas de psicología y educación de Paraguay, España y Canadá; quienes me han brindado oportunidades y aprendizajes que han permitido mi pleno desarrollo intelectual. A mis padres en Paraguay, mi profundo agradecimiento por su ejemplo de tenacidad, amor y entrega. Así también, mi agradecimiento al Programa de Becas en el Exterior "Don Juan Carlos Antonio López" del Gobierno de la Republica del Paraguay, por valorar mi perfil investigador y sobre todo creer en mi potencial.

España me ha brindado lo mejor de su comunidad científica, los conocimientos adquiridos y el apoyo incondicional recibido me han orientado y permitido el desarrollo de este proyecto, que representa el punto culminante de mi iniciación en la carrera científica. He sido privilegiada al contar con tutores, pioneros y renombrados científicos en lo que a investigación en psicometría y educación se refiere. A Rosario Martínez-Arias, Rubén Fernández-Alonso y José Muñiz-Fernández, mi eterno agradecimiento por dejarme formar parte de sus investigaciones. Mi gratitud por el apoyo y las oportunidades de capacitación brindadas al grupo de investigación en Psicometría de la Facultad de Psicología de la Universidad de Oviedo, al Departamento de Ciencias de la Educación de la Universidad de Oviedo y al Servicio de Evaluación Educativa de la Dirección General de Ordenación Académica e Innovación Educativa del Gobierno del Principado de Asturias.

Este viaje intelectual no habría sido igual, sin la privilegiada oportunidad de formar parte del Measurement, Evaluation and Research Methodology Program (MERM), del Department of Educational and Counselling Psychology, and Special Education (ECPS) de la University of British Columbia en Vancouver Canadá. A mi tutor,

el renombrado teórico y matemático Dr. Bruno Zumbo, mi eterno agradecimiento por esas profundas horas de diálogo y trabajo científico-académico. Así también a la Dra. Yan Liu profesora asistente del ECPS, por sus magistrales clases de modelos jerárquico-lineales cambio y crecimiento.

No puedo olvidar en estos párrafos de agradecimiento a las instituciones y organizaciones que han reconocido mi trabajo. En primer lugar, *The International Test Comission (ITC)* por el reconocimiento como *Joven Investigador* año 2016. En la misma línea, mi eterno agradecimiento a la *Asociación Interuniversitaria de Investigación Pedagógica (AIDIPE)* por el premio de investigación recibido en el *I Encuentro de Doctorados e Investigadores Nóveles 2017*. Por último, mi agradecimiento a la *Sociedad Científica en Psicología y Educación* con el *Premio ACIPE 2018* en reconocimiento al mejor artículo publicado en la *Revista de Psicología y Educación*.

Finalmente, mi reconocimiento a todos los expertos de las diferentes universidades que he visitado y con los que he publicado, así como a los alumnos a nivel mundial con los cuales he compartido mis conocimientos. Me han guiado tanto como motivado en este camino y cada una de las temáticas estudiadas en esta Tesis Doctoral han sido validadas por su expertis y valiosa opinión. No solo han sabido ser una fuente de apoyo constante, sino que también han sido una fuente inagotable de conocimiento que me ha permitido disfrutar de este viaje intelectual. Sin duda todos ustedes han marcado mi carrera profesional y mis anhelos personales.

Pamela Woitschach

MacKenzie Mount, Revelstoke British Columbia

Canadá, marzo del 2018

MOTIVACIÓN PERSONAL

Años atrás cuando me encontraba estudiando las materias de Psicometría y Validez del Máster de Metodología de Investigación en Ciencias del Comportamiento y la Salud de la Universidad Complutense de Madrid, recordaba una y otra vez una imagen de mi adolescencia en la educación secundaria de una institución educativa pública de Paraguay.

> Era un día de primavera y se sentía en el ambiente aquel clima de América del Sur, tan cálido y abrazador como su gente. La «escuelita» se denominaba al edificio de la institución educativa donde realizaba mis estudios haciendo honor a su carencia de infraestructura. En aquel ansiado día de la prueba de competencia regional de matemática, dado el tamaño de las aulas, nuestros asientos fueron ubicados en el patio de la institución. Bajo el sol radiante compartíamos el espacio con quienes alegres disfrutaban de un acalorado juego de balón pie. Al tiempo que me dispuse a leer las preguntas de la prueba, mi mente se abstrajo por completo, percatándome solo de que el equipo al que aprendí a alentar al tiempo que esperaba recibir la prueba, estaba ganando…

Cada día que analizo las bases de datos de esta Tesis Doctoral y de otras investigaciones, me pregunto en qué contexto está inmerso el estudiante 53,765, del clúster 2305 y del estrato 14; y cómo sus características personales y las circunstancias espacio temporales del proceso de evaluación conforman y condicionan sus respuestas a la prueba. Las evaluaciones educativas a gran escala, sus implicaciones y consecuencias, en un proceso que aglomera información de miles de estudiantes, docentes, directivos y familias; y que tiene un indudable impacto en la política educativa de los países participantes, es el foco central de este trabajo, que deseo pueda llegar a ser de una lectura fructífera para todos, como fructífero fue su desarrollo.

Un fracaso no siempre es un error; puede ser simplemente lo mejor que uno puede hacer dadas las circunstancias. El verdadero error es dejar de intentarlo. B. F. Skinner

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTADO ALFABETICO DE SIGLAS Y ACRÓNIMOS

AERA: American Educational Research Association

ANOVA: Analysis of variance

APA: American Psychological Association

AwC: Alignment-within-CFA

BBR: Replicación repetida balanceada

BID: Banco Interamericano de Desarrollo

BSEM: Bayesian structural equation modeling

CFA: Confirmatory factor analysis

CLF: Component loss function

DIF: Differential item functioning

EM: Método iterativo

EXANI: Examen Nacional de Ingreso

GII: Gender inequality index

HDI: Human development index

HGLMM: Hierarchical generalized linear mixed model

HLM: Hierarchical linear model

HOP: Índice de posesiones en el hogar

ICC: Intraclass correlation coefficient

IEA: International Association for the Evaluation of Educational Achievement

IRT: Item response theory

ISEC: Índice socioeconómico y cultural

LLECE: Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación

LMM: Linear mixed models

MACS: Means and covariance structure analysis

MG-CFA: Multiple-group confirmatory factor analysis

MI: Measurement invariance

ML: Maximum likelihood estimation

MLR: Maximum likelihood estimation with robust standard errors

NAEP: National Assessment of Educational Progress

NCME: National Council on Measurement in Education

OCDE: Organización para la Cooperación y el Desarrollo Económicos

OECD: Organisation for Economic Co-operation and Development

OREALC: Oficina Regional de Educación para América Latina y el Caribe

OTL: Oportunidad de aprendizaje

PQL: Penalized quasi-likelihood

PERCE: Primer Estudio Regional Comparativo y Explicativo

PIRLS: Progress in International Reading Literacy Study

PISA: Programme for International Student Assessment

SBA: Índice de servicios básicos de la vivienda

SCA: Índice de servicios de comunicación de la vivienda

SACMEQ: Southern and Eastern Africa Consortium for Monitoring Educational Quality

SEA-PLM: Southeast Asia Primary Learning Metrics

SEC: Nivel socioeconómico y cultural de la familia

SERCE: Segundo Estudio Regional Comparativo y Explicativo

SIMCE: Sistema de Medición de la Calidad de la Educación

TALIS: Teaching and Learning International Survey

TCT: Teoría clásica de los test

TERCE: Tercer Estudio Regional Comparativo y Explicativo

TIMMS: Trends in International Mathematics and Science Study

TRI: Teoría de la respuesta al ítem

UNDP: United Nations Development Programme

UNESCO: United Nations Educational, Scientific and Cultural Organization

RESUMEN

La presente Tesis Doctoral «Evaluaciones Educativas a gran escala en Latinoamérica: Tercer Estudio Regional Comparativo y Explicativo (TERCE)» analiza la equidad de las evaluaciones educativas estandarizadas del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación en América Latina y el Caribe (LLECE). El objetivo es conformar un análisis metodológico de la equidad del proceso de evaluación desde una mirada *in vivo* antes que *in vitro,* en este análisis de la equidad confluirán dos tradiciones. Por un lado, la aproximación educativa donde la equidad es entendida como la distribución de los conocimientos y oportunidades sociales y escolares y, por otro lado, la tradición psicométrica que busca aportar evidencias de la validez de las medidas, que garanticen el uso e interpretación de los resultados de las pruebas en la *evaluación educativa estandarizada*. En apariencia pueden parecer dos tradiciones totalmente independientes. Sin embargo, tomados en el análisis metodológico de la evaluación desde el punto de vista del modelo ecológico tanto el *contexto* en donde se desarrolla la evaluación, como el *proceso de validación* son aspectos simbióticamente conectados.

El trabajo está conformado por cinco estudios. El primero tiene como objetivo establecer la invarianza y analizar la viabilidad de la comparación de las medidas de la prueba de logro educativo entre países y culturas *(Capítulo I)*. En un segundo momento, una triada de estudios desde una perspectiva educativa intentan determinar el efecto global de los centros educativos *(Capítulo II);* estimar el impacto de la segregación escolar *(Capítulo III)*; y conocer la influencia de las oportunidades de aprendizaje (OTL) otorgadas por el contexto escolar y los sistemas educativos en el rendimiento de los estudiantes *(Capítulo VI)*. Finalmente, el *Capítulo V* tiene el objetivo de analizar las causas subyacentes al funcionamiento diferencial del ítem (DIF) desde una visión ecológica del proceso de evaluación.

La muestra empleada en todos los estudios corresponde al alumnado escolarizado en 6º grado de Educación Primaria de 15 países de América Latina participantes del TERCE del 2013, que en suma representan a más de 9 millones de estudiantes. Los análisis se realizaron utilizando como variable dependiente las distintas pruebas cognitivas y contaron con la incorporación de una metodología robusta que incluye el uso de cinco valores plausibles, variables de estratificación, incluyendo los pesos muestrales y pesos replicados. En el *Capítulo I* se utiliza la técnica del *Alignment* para la determinación de los parámetros de invarianza aproximada, mientras que en los *Capítulos II, III y IV* se emplean modelos jerárquico-lineales de dos y tres niveles de agregación. Por último, el estudio del DIF se realizó mediante una regresión logística multinivel para variables con distribución Bernoulli.

El primer estudio destaca la presencia de parámetros no invariantes en prácticamente el 50% de los ítems de cada cuadernillo de ciencias naturales, advirtiéndose que en esos casos los parámetros no son equivalentes entre los países. Las variabilidades están concentradas fundamentalmente en Chile, Brasil, Ecuador, Guatemala y Nuevo León (México). También se resalta que la mayoría de los parámetros no-invariantes corresponden a ítems que evalúan los procesos cognitivos de comprensión y reconocimiento de conceptos. El formato y nivel de puntuación de los ítems también aparece vinculado a la invarianza ya que todos los ítems de crédito parcial presentan parámetros no-invariantes. En cuanto a las comparaciones entre países se observa la presencia de un alto nivel de disparidad en la forma en cómo los estudiantes de cada país comprenden y responden al ítem, por lo que el uso de rankings de logro educativo resultados puede no ser adecuado.

Adentrados en la triada de estudios sobre la equidad educativa, el *Capítulo II* analiza el efecto de los centros educativos y observa que, en los sistemas educativos

latinoamericanos el coeficiente de correlación intraclase (ICC), entendido como el porcentaje de varianza entre centros de un modelo multinivel sin predictores, gira en torno al 40%, valor que se ha mantenido desde las evaluaciones realizadas en 1997 y 2006. El ICC se ve reducido en un 60% al incluir variables de tipo socioeconómico. Por otra parte, el efecto neto de los centros escolares en América Latina es de 13% en la materia de lectura, 23% en matemáticas y 25% en ciencias naturales.

El *Capítulo III* señala que los estudiantes de bajo nivel socioeconómico que asisten a centros educativos con aulas de conformación heterogénea presentan un bajo resultado educativo, a diferencia de los estudiantes de nivel socioeconómico alto que asisten a centros educativos heterogéneos. Una vez el modelo es controlado por variables sociodemográficas, se observa que a medida que los centros presentan mayores dispersiones en el índice socioeconómico tienden también a presentar resultados más bajos. Un claro aporte de este estudio es la certeza de que los estudiantes de nivel socioeconómico alto de la muestra no se ven penalizados por asistir a agrupaciones heterogéneas.

En el *Capítulo IV*, donde se analiza el efecto de las OTL, se destaca que el 40% de las diferencias en los resultados entre los centros educativos se explican por factores de contexto o entrada. El modelo ajustado por variables de contexto predice diferencias del orden de 0.92 desviaciones típicas entre el alumnado de los centros con mayor y menor nivel socioeconómico, y las variables OTL (cuaderno de anotaciones, asistencia docente, clima y practicas docentes en el aula) explican cerca de un 10% de las diferencias entre los centros y un 2% de las diferencias entre el alumnado.

Finalmente, en el *Capítulo V* se destaca que el 32% de los ítems del cuadernillo uno de ciencias naturales, presenta evidencias de DIF entre las agrupaciones de sexo masculino y femenino. Aun cuando se analiza el DIF en dos o tres niveles de agregación,

su presencia es significativa y favorece al sexo masculino. A nivel de sistema educativo las variables explicativas son: el Índice de inequidad de género (GII) y el Índice de desarrollo humano (HDI). En consecuencia, una vez controlado el DIF por el HDI el estudiante del género femenino tiene cuatro veces más posibilidades de dar una respuesta correcta que los estudiantes del género masculino.

Palabras clave: Evaluación internacional, equidad, invarianza, eficacia escolar, marco ecológico de respuesta al ítem, validez, DIF, modelos jerárquico-lineales.

SUMMARY

This doctoral thesis "Large Scale Educational Assessments in Latin America: The Third Comparative Regional Explanatory Study (TERCE)" examines the equity of standardised educational evaluations from the Latin American Laboratory for the Evaluation of Educational Quality in Latin America and the Caribbean (LLECE). The objective is perform a methodological analysis of the equity of the evaluation process in the real world rather than theoretically, bringing together two traditions in the analysis. One is the educational approach in which equity is understood as the fair distribution of knowledge and social and school opportunities, the other is the psychometric tradition which seeks to provide evidence of validity for measurement which backs the use and interpretation of results of tests in *standardised educational evaluation*. They may appear to be two completely independent traditions, however when viewed through a methodological analysis of evaluation and the lens of an environmental model, both the *context* in which the evaluation is performed, and the *validation process* are symbiotically connected.

This work comprises five studies. The objective of the first is to establish the invariance of testing measures for educational achievement between different countries and cultures and to analyse whether comparing them would be viable (*Chapter I*). Secondly, a trio of studies from an educational perspective attempt to determine the overall effects of schools (*Chapter II*); estimate the impact of school segregation (*Chapter III*); and understand the influence of the opportunity to learn (*Chapter IV*). Finally, the objective of *Chapter V* is to analyse the underlying causes of differential item functioning (DIF) from an environmental view of the evaluation process.

The sample used in all of the studies were the students in the 6th year of primary education in the 15 Latin American countries participating in TERCE 2013, more than

nine million students. The analyses were performed using the various cognitive tests as the dependent variable and used a robust methodology which included the use of five plausible values, stratification variables, including sample weights and replicate weights. In *Chapter I* the *Alignment* technique is used to determine the approximate parameters of invariance, while in *Chapters II, III, and IV* two-and three-level hierarchical linear models are used. Finally, the DIF study was performed using a multilevel logistical regression model for variables with a Bernoulli distribution.

The first study highlights the presence of non-invariant parameters in almost 50% of the items in each science test booklet, indicating that in those cases the parameters are not equivalent between countries. The variability is mainly concentrated in Chile, Brazil, Ecuador, Guatemala and Nuevo León (Mexico). It also indicates that the majority of the non-invariant parameters are in items evaluating the cognitive processes of comprehension and concept recognition. The item formats and scores also seem linked to the invariance as all of the partial credit items present non-invariant parameters. There is a great disparity in the comparisons between countries in the way the students in each country understand and respond to items, which means that the use of rankings of educational achievement may not be appropriate.

Chapter II examines the effect of schools and notes that in Latin American education systems the intraclass correlation coefficient (ICC), which is the percentage of variance between schools in a multilevel model without predictors, is around 40%. This is a value that has not significantly changed since the evaluations performed in 1997 and 2006. The ICC falls by 60% when socioeconomic variables are included. The net effect of schools in Latin America is 13% in reading, 23% in mathematics, and 25% in sciences.

Chapter III indicates that students from lower socioeconomic levels who attend schools with classes with a diverse student make-up exhibit lower educational

achievement, in contrast to students from higher socioeconomic levels who attend diverse schools. Once the model controls for sociodemographic variables, the extent to which schools with a wider range of sociodemographic variables tend to also exhibit lower results is clear. One clear contribution of this study is the certainty that students from higher socioeconomic levels in the sample are not penalised for attending diverse classes.

Chapter IV, which examines the effect of Opportunity to Learn (OTL) variables , highlights that 40% of the differences in results between schools is explained by factors related to context or admission. The model, adjusted for context variables, predicts differences in the order of 0.92 standard deviations between students at schools with high or low socioeconomic levels, and that the OTL variables (notebooks, teacher attendance, teaching climate and practice in the classroom) explain almost 10% of the differences between schools and 2% of the differences between students.

Finally, *Chapter V* highlights that 32% of the items in science test booklet one exhibit evidence of DIF between groups of boys and girls. DIF is significant even when looking at classroom and school level, and boys score higher than girls. The explanatory variables at the level of education system are: The Gender Inequality Index (GII), and the Human Development Index (HDI). Consequently, once the DIF is controlled for HDI, girls are four times as likely to give a correct answer than boys.

Key words: International evaluation, equality, invariance, school effectiveness, ecological item response framework, validity, DIF, hierarchical-linear models.

PREFACIO

La educación se erige como un derecho humano fundamental, estipulado en el artículo 26 de la Declaración Universal de los Derechos Humanos de 1948, por lo que gobiernos de todos los pueblos y naciones del mundo se han abocado a su monitoreo con el fin de garantizar su cobertura, calidad y equidad. Como una consecuencia lógica de este interés, la evaluación educativa estandarizada inicia cinco décadas atrás con la implementación a nivel mundial de los primeros estudios centrados en la evaluación comparada de los conocimientos académicos y el diagnóstico situacional de los sistemas educativos[1] (Tian & Sun, 2018).

Las *evaluaciones educativas estandarizadas* se sustentan en una infraestructura metodológica robusta y coherente, que permite el uso de pruebas estandarizadas, con la finalidad de acumular datos y evidencias objetivas que permitan la toma de decisiones y la mejora de los sistemas educativos. Fernández-Alonso (2004), concluye que dada la extrema complejidad del proceso de evaluación y el alcance de sus resultados, las evaluaciones educativas estandarizadas son el fruto de un compromiso político que supone un esfuerzo logístico coordinado de todos los sistemas educativos participantes

En virtud de ese compromiso, los estados miembros de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), designan al Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) como un agente clave para el monitoreo y seguimiento del Marco de Acción de la Agenda de Educación 2030 y del Objetivo de Desarrollo Sostenible N° 4 (UNESCO, 2016), que busca garantizar una educación inclusiva y equitativa de calidad en la región. El LLECE bajo la coordinación técnica de la Oficina Regional de Educación de la

[1] En adelante nos referiremos a las evaluaciones diagnósticas y de comparación de sistemas educativos con el término general de «*evaluaciones educativas estandarizadas*».

UNESCO para América Latina y el Caribe (OREALC-UNESCO Santiago), tiene como objetivo dar cumplimiento al eje central propuesto por la UNESCO, por lo que establece en 1996 la evaluación de sistemas educativos de América Latina y el Caribe (UNESCO-OREALC & LLECE, 2000). Siendo el *Tercer Estudio Regional Comparativo y Explicativo* (TERCE) del año 2013 la tercera versión del estudio. La riqueza metodológica de esta colección de evaluaciones educativas radica por sobre todo en que es diseñada y aplicada únicamente a países de Latinoamérica, 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 visión propuesta por Martínez-Arias (2006) subraya que las evaluaciones educativas estandarizadas de carácter internacional mantienen inmersa la idea latente de considerar al mundo como un *laboratorio educativo global*, donde el contexto ya sea de carácter político o educativo produce diversos resultados. En consonancia con ese enfoque y basado en el Modelo Ecológico de Bronfenbrenner (1979) la UNESCO construye sobre los resultados del TERCE las recomendaciones para las políticas públicas de mejora de la educación en América Latina y el Caribe. El modelo ecológico de Bronfenbrenner al igual que el modelo de Contexto, Entrada, Proceso y Producto (CIPP) propuesto por Stufflebeam a finales de los 60`, son utilizados en la vasta mayoría de las evaluaciones educativas. La adopción de estos modelos se realiza con el fin de enriquecer y profundizar el análisis de la mejora educativa desde una perspectiva holística que aprecia la estructura anidada de los sistemas educativos y las interrelaciones entre los diferentes niveles (UNESCO-OREALC, 2016a).

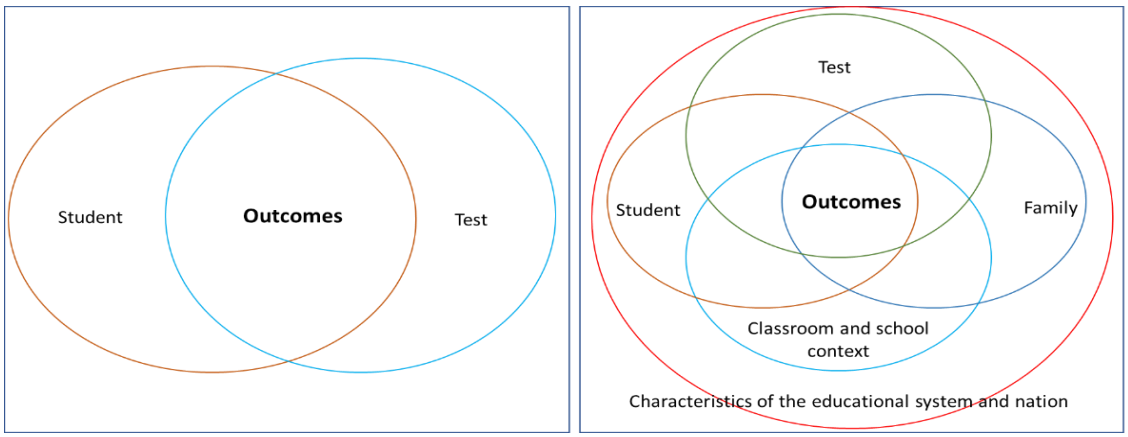
Esta Tesis Doctoral se enmarca en un programa de investigación que tiene como objetivo apreciar que las evaluaciones educativas se desarrollan en ambientes naturales donde la mediación «contextual» determina las oportunidades de aprendizaje a las que el

estudiante es expuesto y por ende los resultados de la prueba; y se encuentra implícita no solo en la construcción de *inferencias, decisiones y recomendaciones sobre los resultados obtenidos,* sino que también en el *proceso de validación de las medidas obtenidas en la prueba*. Si bien desde una perspectiva métrica, en apariencia estamos posicionados entre dos aspectos al parecer totalmente independientes, tomados éstos en el análisis metodológico desde el punto de vista ecológico de Bronfenbrenner (1979), como es propuesto por Chen y Zumbo (2017), tanto el *contexto* como el proceso de *validación* son aspectos simbióticamente conectados.

Adentrados desde una perspectiva psicométrica, *The Standards for Educational and Psychological Testing* del 2014 menciona que el propósito de las pruebas utilizadas en las evaluaciones educativas estandarizadas va más allá del análisis a nivel individual o de aula, teniendo el propósito no solo de informar sobre el proceso de enseñanza y el aprendizaje a como curricular, sino también realizar inferencias sobre los resultados del proceso enseñanza aprendizaje (AERA, APA, & NCME, 2014, p. 184). Sin embargo, a escala mundial los usos actuales de los resultados obtenidos de las pruebas en las evaluaciones educativas estandarizadas incluso van más allá de solo esos usos previstos. En efecto los resultados de las pruebas en las evaluaciones educativas estandarizadas son generalmente utilizados para la comparación entre países y culturas, teniendo también implicancias en la supervisión, intervención, innovación o modificación en todos los niveles de la política educativa (Por ejemplo, a nivel del estudiante, centro y sistema educativo) (Lietz, Cresswell, Rust & Adams, 2017). Por lo tanto, dado el alto impacto del uso legítimo de la prueba en las evaluaciones educativas estandarizadas, la relevancia de la validez de las medidas no puede ser subestimada (Sireci, 2015).

Si bien el uso de pruebas estandarizadas nace de una tradición psicométrica aplicada a la evaluación experimental en psicología, esta visión tradicional del análisis

que se observa en la Figura 1 de la izquierda destaca que el resultado de la prueba (*outcomes*) se halla en suma determinado por las características del estudiante (*student*) y las características de la prueba (*test*). La expansión del uso de pruebas estandarizadas en las evaluaciones educativas ha motivado un cambio radical en esta visión, ya que, a diferencia de la perspectiva tradicional, las evaluaciones de sistemas educativos presentan una visión globalizada donde los educadores y las instituciones tienen un papel preponderante en los resultados obtenidos.



Nota: Elaboración propia basada en Ruhe y Zumbo, 2009. Nota: Elaboración propia basada en Zumbo et al. (2015)[2]

Figura 1. Modelo de evaluación en Psicología y el Modelo Ecológico de respuesta al ítem-test propuesto por Zumbo et al. (2015).

El trabajo de Zumbo y Gelin (2005) se constituye como el precursor del modelo ecológico de respuesta al item-test de Zumbo et al. (2015) que a su vez se sustenta en el modelo ecologico propuesto por Bronfenbrenner en 1979. En un paso mas allá en el desarrollo del modelo Chen y Zumbo (2017), proponen el uso del marco de referencia ecológico del proceso de respuesta y rendimiento en la prueba como una evidencia de validez. Este giro en la perspectiva se puede observar en la Figura 1 (derecha), donde el resultado de la evaluación *(outcomes)* es determinado en principio por las características

[2] Cabe resaltar que si bien el grafico presenta diagramas de tamaño similar, no se espera que el impacto de cada nivel de agregación representado en el diagrama sea de exclusiva igual dimensión.

de la prueba y del estudiante, así como, por las características de la familia, del aula y la escuela *(classroom and school)* que se presentan como oportunidades de aprendizaje a las que el estudiante se encuentra expuesto y que también determinan la experiencia de aprendizaje, la respuesta al ítem y por ende el resultado de la prueba *(outcomes).* Por último, se observa el impacto de las características externas a la escuela que provienen del sistema educativo de los países y naciones en donde el estudiante se encuentra inmerso.

Si pensáramos por un momento en una analogía que escenifique la diferencia entre ambas perspectivas de análisis pensaríamos por ejemplo en una obra de arte y en como la forma artística entre el fondo (contexto) y la forma (respuesta al ítem y resultado de una prueba) es organizada. Por un lado, la perspectiva tradicional que si bien no omite la existencia de un contexto especifico de donde el evaluado proviene y forma parte (fondo), se centra en las características de la prueba y las características del estudiante (forma). Mientras que, desde la perspectiva ecológica propuesta por Zumbo, et al. (2015), el contexto ya sea familiar, social, escolar o político en donde el estudiante se encuentra inmerso se convierten en elementos que contribuyen a entender el proceso de respuesta al ítem y por ende el resultado de la prueba.

Este interés en la confluencia entre el proceso de validación y el contexto en donde la evaluación se desarrolla si bien no es novel para el ámbito educativo tampoco lo es para la psicometría, dado que ya fue introducido por Cronbach & Meehl en 1955, seguidos por Cronbach en 1982 y Messick 1989. Este último propone una visión unificada y apuesta por la incorporación de las consecuencias sociales del uso de los test al concepto de validez. Su principal aporte sería la visión de la validez como un juicio de evaluación integral del grado en que la evidencia empírica y teórica fundamenta de forma idónea y apropiada las interpretaciones y acciones basadas en los puntajes derivados de los test y

otros modos de evaluación (Messick, 1989, p. 5). Esto se ve a su vez reflejado en las subsecuentes publicaciones de *Standards*, donde la validez es vista como un concepto holístico e integrado que incluye evidencias sobre el contenido de la prueba, los procesos de respuesta, la estructura interna, la relación con otras variables y las consecuencias sociales del proceso de evaluación (AERA, APA, & NCME, 1999; AERA et al., 2014).

Retomando la idea de párrafos anteriores, donde se resaltaba que la UNESCO construye sobre los resultados del TERCE las recomendaciones para las políticas públicas de mejora de la educación desde el Modelo Ecológico de Bronfenbrenner (1979) y del Modelo CIPP; y que ambos modelos se utilizan no solo para el análisis del logro educativo, sino que también para la construcción de ítems de los cuestionarios y el análisis de factores asociados. Por lo que la construcción de *inferencias, decisiones y recomendaciones sobre los resultados obtenidos,* así como el *proceso de validación de las medidas obtenidas en la prueba* no pueden ser exentos del análisis desde una visión ecológica que permita la comprensión profunda de todos los factores que lo determinan.

Por lo que el programa de investigación que sustenta esta Tesis Doctoral se encuentra alineada a ofrecer evidencias de validez con el modelo ecológico propuesto por Chen y Zumbo (2017) y con la idea holística y articulada de validez propuesta por Messick (1989) que ofrecen una visión amplia del proceso de respuesta al ítem como evidencia de validez. El objetivo es conformar un análisis metodológico de la equidad del proceso de evaluación donde confluirán dos tradiciones, por un lado, la aproximación educativa donde la equidad es entendida como la distribución de los conocimientos y oportunidades sociales y escolares y, por otro lado, la tradición psicométrica que busca aportar evidencias de la validez de las medidas, que garanticen el uso e interpretación de los resultados de las pruebas educativas. Siendo su fin último el de enriquecer y profundizar el análisis desde

una perspectiva holística que aprecia la estructura anidada de los sistemas educativos y las interrelaciones entre los diferentes niveles (UNESCO-OREALC, 2016a).

INTRODUCCIÓN

Esta Tesis Doctoral analiza desde una visión ecológica el grado de justicia y equidad de las evaluaciones educativas estandarizadas del LLECE en América Latina y el Caribe. Tal y como postulan Zumbo and Hubley (2016), el objetivo es conformar un análisis metodológico de la equidad de la evaluación desde una mirada más *in vivo* antes que *in vitro*. En consecuencia, dos son los aspectos para profundizar; por un lado, la equidad del contexto educativo y por el otro el impacto de este en la validez de las medidas en las pruebas de evaluación educativa estandarizada.

El Programa para la Evaluación Internacional de Alumnos (PISA) organizado por la Organización para la Cooperación y el Desarrollo Económico (OECD) define a la *equidad* como el suministro de oportunidades de alta calidad en educación, para todos los estudiantes, independientemente de su género, antecedentes familiares o características socioeconómicas (OECD, 2013, 2016a). Siendo la justicia (fairness) un componente principal de la equidad, que implica el aseguramiento de que las circunstancias personales y sociales por ejemplo el género, la condición socioeconómica y el origen étnico, no sean un obstáculo para lograr el potencial educativo. Por otro lado, desde una perspectiva posicionada en la prueba, el capítulo III de la última versión de los *Standards* denominado «*fairness in testing*» refiere que, el objetivo primordial es el de proteger a los que responden la prueba de las amenazas que afectan las interpretaciones justas y válidas de los puntajes de la prueba. Pudiendo estas amenazas provenir de cuatro fuentes: el contenido del test, el contexto en donde se desarrolla la evaluación, el formato de respuesta al test y las oportunidades de aprendizaje a las que son expuestos los evaluados.

En congruencia con lo anterior esta Tesis Doctoral presenta cinco estudios que pretenden abarcar el análisis de la equidad desde una doble perspectiva. Por un lado, el

estudio de las amenazas que afectan a la justicia y validez en el uso e interpretación de los resultados de las pruebas en las evaluaciones educativas estandarizadas y, por otro, la equidad en el acceso a las oportunidades de aprendizaje en los sistemas educativos de Latinoamérica.

A diferencia de gran parte de la investigación en psicología o de la investigación experimental, los estudios educativos se desarrollan en ambientes naturales y por lo que el proceso de validación de las medidas no puede separarse de contexto social y escolar en el que están inmersos los estudiantes que responden a las pruebas. Por ello, el primer objetivo de esta Tesis Doctoral es indagar sobre el grado de pertinencia en la comparabilidad del constructo entre los países participantes *(Capítulo I).* Se realizó un estudio sobre la invarianza de las medidas de todos los cuadernillos de la prueba de ciencias naturales utilizando una técnica psicométrica novedosa, el *Alignment* para la evaluación de la invarianza aproximada (Asparouhov & Muthén, 2014), dicha técnica nos ha permitido no solo el uso de variables de estratificación, pesos muestrales y clústeres de agrupación, sino que también ha permitido la comparación accesible y eficiente de las dieciséis naciones participantes de forma simultánea. Los resultados del análisis invarianza pretenden determinar la variabilidad o invariabilidad entre la forma en como los países ven, analizan y responden al ítem de la prueba de ciencias naturales.

Los estudios que conforman los *Capítulos II, III, IV* intentan, desde una perspectiva más educativa y pedagógica, analizar el contexto social y educativo en el cual se aplican las pruebas. Estos tres estudios abarcan tópicos relevantes para el análisis de la equidad de los sistemas educativos: el impacto de los factores socioeconómicos y el rol de los centros educativos, el impacto de la segregación escolar en los resultados escolares y efecto de las oportunidades de aprendizaje sobre el desempeño académico. En esta triada

de estudios se emplean modelos jerárquico-lineales de dos y tres niveles de agregación, que a su vez incluyeron el uso de los pesos muestrales y valores plausibles con el fin de asegurar la calidad metodológica del análisis.

El *Capítulo V* presenta el análisis del funcionamiento diferencial del ítem que incluye la aplicación de la novel propuesta de la ecología del proceso de respuesta al ítem y rendimiento en la prueba propuesto en sus inicios por Zumbo y Gelin (2005) y que corresponde a la tercera generación del estudio del funcionamiento diferencial del ítem desarrollado más tarde por Zumbo et al. (2015) y Chen y Zumbo (2017). En la misma el análisis DIF es concebido, no es sólo como una técnica indispensable para el análisis de la validez interna, sino también como un aspecto clave para determinar el grado de justicia y equidad en el uso de las pruebas y como una herramienta que nos acerca a una explicación profunda ya sea de carácter cognitivo o social del proceso de respuesta y el rendimiento en una prueba. El análisis DIF es realizado mediante modelos jerárquico-lineales de dos y tres niveles de agregación que incluyen a su vez los pesos muestrales a nivel del estudiante y el centro educativo. La estrategia analítica se ha desarrollado con la inclusión gradual de variables explicativas por cada nivel de agregación, en la misma línea del novedoso estudio realizado por Chen y Zumbo (2017).

Los cinco estudios analizan la equidad de las evaluaciones educativas de esta Tesis Doctoral emplean datos provenientes del programa de evaluación TERCE aplicado en América Latina y se fundamentan en las materias educativas que a criterio de Baker (2014) son las materias más utilizadas para la comparación internacional, denominadas también como inteligencia académica. Por ello esta introducción se organiza en tres apartados. El primero denominado, *los antecedentes de las evaluaciones educativas estandarizadas,* que hace un breve recorrido sobre el nacimiento y consolidación de estas evaluaciones a

nivel mundial y especialmente sobre su desarrollo en el contexto latinoamericano. El segundo apartado sintetiza las evidencias encontradas por estos programas de evaluación en relación con los factores asociados al logro académico, mientras que el último punto está reservado a mostrar la metodología de las evaluaciones *del LLECE para el conjunto de países de Latinoamérica*.

Antecedentes de las Evaluaciones Educativas Estandarizadas a Escala Mundial

Como se acaba de mencionar este apartado recrea el desarrollo histórico de las *evaluaciones educativas estandarizadas*. Si bien es importante tener en cuenta que cada uno de los programas de evaluación que se mencionarán a continuación podrían ser un tópico de estudio intensivo, ahora simplemente se intentará dar una visión lo más sintética y global posible, que permita introducir al lector la tradición de estudios en el que se encuadran las cinco investigaciones que conforman la Tesis Doctoral. Siguiendo el desarrollo histórico de los hechos se mencionarán primeramente los programas de evaluación educativa de carácter internacional, para finalizar enumerando con detalle las evaluaciones educativas estandarizadas que operan en los países de América Latina y el Caribe.

Las evaluaciones educativas estandarizadas arrancan hace más de cinco décadas coincidiendo con la creación de *The International Association for the Evaluation of Educational Achievement* (IEA), organismo de cooperación internacional no gubernamental conformado por instituciones de investigación públicas y privadas de más de 60 países. En 1959 la IEA inició el estudio piloto de doce países, el cual concluyó que las comparaciones entre países eran posibles, aunque presentaban serias dificultades que

obligaban a cuidar los mecanismos de su organización. Desde entonces la IEA ha impulsado evaluaciones en diez áreas curriculares y estudios de otros temas específicos como el clima de aula, el uso de los recursos informáticos para el aprendizaje, la formación docente o la educación infantil. 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 a los estudios de la IEA surge a principios de siglo el estudio PISA de la OECD, el cual abandona la evaluación curricular para centrarse en la valoración de las competencias y destrezas básicas de los estudiantes de 15 años.

Como consecuencia de la globalización, la cultura de la evaluación educativa estandarizada con el objetivo de rendición de cuentas de los gobiernos se ha extendido a todos los rincones del mundo y ha dejado de ser un ejercicio exclusivo de países industrializados (Smith, 2014). Ejemplos destacados a escala regional son 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) que es el programa de evaluación regional de la educación primaria en Asia. En América Latina las evaluaciones educativas son lideradas desde el año 1996 por la UNESCO a través del LLECE y bajo la coordinación técnica de la OREALC (LLECE & UNESCO-OREALC, 2016; UNESCO-OREALC & LLECE, 2000, 2010).

Dentro de las evaluaciones nacionales en las américas es necesario destacar el National *Assessment Educational Project* realizado en los Estados Unidos de Norteamérica (Beaton et al., 2011), que es, sin duda, el programa nacional con mayor

tradicción y más influyente en el plano del desarrollo de la métrica aplicada a la evaluación de los sistemas educativos. Por su parte en la región latinoamericana la evaluación educativa ha experimentado un notable desarrollo a partir de la década de los 90 y en la actualidad la gran mayoría de los países disponen de programas y unidades técnicas para su ejecución, A continuación, se reseñan los programas nacionales que operan en la región organizándolos en cuatro grupos según el momento en que son instaurados por los gobiernos y ministerios de educación nacionales. El primer grupo corresponde a aquellos países que disponían de programas de evaluación antes del año 1990 cuando comienzan las reformas educativas en la región; el segundo grupo estaría constituido por los países que iniciaron la evaluación educativa en torno a 1990; un tercero grupo serían las incorporaciones desde 1995 y el último estaría conformado por los países que instauraron estos programas a principios de este siglo. Una síntesis de las evaluaciones tanto nacionales, regionales e internacionales se recoge en el Anexo-Tabla 1 de la Tesis Doctoral.

Chile, Costa Rica y México son los países pioneros de la evaluación educativa en la región. En Chile el interés por la evaluación del sistema educativo se inicia en la década del 60 y a finales de los 80’ se consolida con la implementación del Sistema de la Calidad de la Educación (Ministerio de Educación Unidad de Currículum y Evaluación, 2004). El recorrido de México es similar: las evaluaciones comienzan a ejecutarse dentro del Programa para la Modernización Educativa en la década de los 80 y unos más tarde se afianzan con la creación del Sistema Nacional de Evaluación Educacional (Secretaría de Educación Pública & Organización de Estados Iberoamericanos para la Educación la Ciencia y la Cultura [OEI], 1994). A su vez Costa Rica aplica entre 1981 y 1986 las primeras pruebas nacionales diagnósticas (OEI, 1997).

Brasil, Colombia y Argentina conforman el grupo de países que inician la evaluación de su sistema educativo en torno a 1990. Brasil inicia la aplicación de pruebas educativas a nivel nacional en 1990 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2007). Por su parte, en 1991 se instaura el Sistema Nacional de Evaluación de la Calidad del Ministerio de Educación de Colombia a través del Instituto Colombiano para el Fomento de la Educación Superior (Fundación Centro de Estudios en Políticas Públicas & Fundación Konrad Adenauer, 2005). Finalmente, en 1993 Argentina crea el Sistema Nacional de Evaluación de la calidad educativa, el cual inicia al mismo tiempo con evaluaciones educativas anuales provinciales de finales de cada etapa educativa (Dirección Nacional de Información y Evaluación de la Calidad Educativa, 2003).

El tercer grupo está compuesto por *Paraguay, Uruguay, Perú, Ecuador Guatemala y República Dominicana,* quienes inician el proceso de evaluación nacional entre los años 1995 y finales de la década del noventa. Paraguay instaura en 1995 el Sistema Nacional de Evaluación del Proceso Educativo y la primera evaluación se realiza en el año 1996. Mientras que Uruguay instala en el año 1996 la evaluación censal de aprendizajes de la educación primaria (Administración Nacional de la Educación Pública, 2003). Al mismo tiempo en 1996 Perú aplica las primeras evaluaciones nacionales a través del Ministerio de Educación (Ministerio de Educación del Perú & OEI, 1994). A pesar de que Ecuador inicia su interés por la investigación educativa en la década del 80, no es hasta 1996 cuando instala el Sistema Integral de Evaluación Educativa (Ministerio de Educación y Cultura del Ecuador & OEI, 1994). A mediados de la década del 90 Guatemala crea el Programa Nacional de Evaluación del Rendimiento Escolar (Centro de Investigaciones Económicas Nacionales, 2002) y al mismo tiempo Honduras consolida la Unidad de

Medición de la Calidad Educativa en 1997 con la evaluación del rendimiento académico de carácter censal (Fundación para la Educación Ricardo Ernesto Maduro Andreu, 2017). Mientras que República Dominicana inicia en 1996 el Sistema Nacional de Pruebas que miden el aprendizaje de los estudiantes (Programa de Promoción de la Reforma Educativa en América Latina y el Caribe [PREAL], 2006).

Los últimos países en incorporarse a la evaluación del sistema educativo nacional son *Panamá y Nicaragua*. Panamá inicia las evaluaciones de competencias en el 2005, mientras que Nicaragua inaugura las evaluaciones estandarizadas en el año 2002 y materializa en el 2006 la creación del Consejo Nacional de Evaluación y Acreditación del Sistema Educativo Nacional (UNESCO, 2010). Finalmente, República Dominicana aplica en el 2011 la evaluación diagnóstica primer ciclo de básica y en el 2013 la evaluación diagnostica primer ciclo de la educación media.

En definitiva, la evaluación de los sistemas educativos con el objetivo de acumular evidencias que orienten las decisiones políticas ha experimentado un notable desarrollo en las últimas décadas, proliferando los programas de evaluación internacional y creándose organismos nacionales y consorcios internacionales dedicados a la evaluación educativa (Tian & Sun, 2018). Además, en las comparaciones internacionales los factores culturales, contextuales adquieren un papel decisivo en la validez de las puntuaciones obtenidas y la posibilidad de que esas medidas puedan ser generalizadas y comparables ente los países y regiones (Ercikan, Roth, & Asil, 2015). La región latinoamericana se constituye en un escenario peculiar donde las características del proceso educativo (mayor desigualdad en las oportunidades de aprendizaje) y la presencia de rasgos propios del contexto (desequilibrios socioeconómicos, pueblos originarios, etc.) juegan un rol preponderante en el logro educativo

Efectos Escolares y Factores Asociados al Logro Académico en las Evaluaciones Educativas Estandarizadas

La aparición y aumento de los programas educativos a nivel mundial ha permitido la acumulación de evidencias científico-técnicas sobre la realidad educativa. Este apartado tiene como objetivo adentrarnos a una perspectiva más orientada a la educación. Es una revisión sobre las principales evidencias sobre la efectividad de los sistemas educativos y el análisis de los factores que se asocian al progreso en materia de educación.

En 1964 el Congreso de los Estados Unidos promulga la ley "Civil Rights Act" en un contexto histórico caracterizado por la doctrina Separated but equals. El gobierno de Estados Unidos de Norteamérica solicita la elaboración del estudio denominado *Equality of Educational Opportunity* (Coleman et al., 1966), más tarde conocido como el Informe de Coleman, cuya finalidad última era auditar el efecto de la inversión educativa realizada ininterrumpidamente desde el final de la segunda guerra mundial. El estudio fue llevado a cabo por el National Center for Education Statistics of the U.S. Office of Education y estuvo compuesto por 60.000 profesores y 4.000 escuelas distribuidas a lo largo de los Estados Unidos.

Dos son los resultados obtenidos en el Informe Coleman que se encuentran estrechamente relacionados a los puntos analizados en esta Tesis Doctoral. El primero, trata sobre el poder de las instituciones educativas para minimizar el efecto del contexto. El estudio, influenciado por las teorías sociológicas del momento (Bordieu & Passeron, 1996), demostró que la escuela, así como los recursos educativos que se utilizaban tenían una escasa influencia en la determinación de los logros de los alumnos en comparación con sus diferencias de origen. Los resultados señalaron que los resultados estaban determinados por las características de las familias y las diferencias en el acceso a recursos

educativos del alumnado. No obstante, los datos también señalaron que los estudiantes en situación de desventaja social se verían beneficiados por participar en aulas con estudiantes de niveles socioeconómicos más altos, mientras que los estudiantes de clase socioeconómica alta, presentarían un rendimiento más bajo al que cabría esperar dado su nivel socioeconómico.

A raíz y, en ocasiones, como reacción a las conclusiones del informe de Coleman, se desarrollaron diferentes estudios que tratan de replicar sus resultados. Plowden en 1967, Weber (1971) y alcanzándose una etapa de afianzamiento en el estudio del efecto de las escuelas con Rutter, Mortimore, Ouston, y Moughan (1979) quienes oficializan la línea de investigación en eficacia escolar con el estudio *Fifteen thousand hours: secondary schools and their effect on children*. Los avances teóricos y metodológicos de la eficacia escolar son evidentes desde las pioneras propuestas de principios de la década del 80' (Edmonds, 1982; Lezotte, 1989), presentando un abanico conceptual e ideológico amplio. De esta manera, aparecen estudios desde una vision productiva de la educacion (Hopkins, 1990; Rodríguez, 1991), seguidas por propuestas más integradoras (Fernández, Trevignani, & Silva, 2003; Martinic & Pardo, 2003; Murillo, 2003b), hasta desembocar en el concepto de la eficacia escolar más orientado al componente de equidad (Creemers & Kyriakides, 2010; Lezotte & McKee, 2011; Murillo & Duk, 2011). Murillo ofrece una definición de escuela de éxito que está muy relacionada con el enfoque que se mantiene en la presente Tesis Doctoral y que señala que la escuela eficaz presenta tres características relevantes: el valor añadido, la equidad y el desarrollo integral de los alumnos, por tanto:

> Se entiende que una escuela eficaz es eficaz si consigue un desarrollo integral de todos y cada uno de sus alumnos mayor de lo que sería esperable teniendo en cuenta su

rendimiento previo y la situación social, económica y cultural de las familias (Murillo, 2003c, p.2).

Dos son las temáticas de interés dentro de la investigación sobre el movimiento de eficacia escolar. Por un lado, el foco de atención se encuentra en determinar cuánto influye la escuela sobre el rendimiento de los alumnos; es decir, estimar la magnitud de los efectos escolares y analizar sus propiedades científicas. Y, por otro, cuáles son las características que generan esas diferencias entre escuelas, es decir, identificar los factores de aula, escuela y contexto que hacen que una escuela sea eficaz (Murillo, 2003a, 2003b; Teddlie & Reynolds, 2000; Teddlie, Reynolds, & Sammons, 2000).

En general los estudios educativos han encontrado que las fuentes de variabilidad de los resultados educativos operan en diferentes niveles de agregación y, por ello, proceden tanto de las características propias del evaluado y su contexto familiar, como de la agrupación escolar y de las características del sistema educativo en el que se encuentra inmerso. Entre los estudios realizados en el ámbito latinoamericano que han aportado sólidas evidencias sobre los factores asociados al logro educativo se encuentran la colección de estudios del LLECE (UNESCO-OREALC & LLECE, 2000, 2010, 2016a); la síntesis de investigación sobre Latinoamérica del *International Handbook of School Effectiveness and Improvement* (Towsend, 2007); y la colección de estudios sobre eficacia escolar promovida por el Centro de Investigación y Documentación de España en conjunto con los países miembros del Convenio Andrés Bello (Murillo et al., 2006). A continuación, se revisan los factores asociados a los resultados educativos en las evaluaciones Latinoamericanas.

La literatura científica coincide en destacar que, *a nivel del estudiante y su familia*, los antecedentes escolares, las prácticas educativas en el hogar y los recursos económicos son

los principales factores que impactan en el rendimiento a nivel del estudiante (Banco Interamericano de Desarrollo [BID], 2017; OECD, 2016a; Scheerens, 2016; UNESCO-OREALC & LLECE, 2016a). En la región latinoamericana a diferencia de otros contextos educativos, emergen variables relacionadas de forma negativa al logro educativo como el hecho de ser niña, pertenecer a una población indígena y/o rural, presentar bajos niveles socioeconómicos y poseer un trabajo infantil remunerado (UNESCO-OREALC, 2016a; UNESCO-OREALC & LLECE, 2016a). En el sentido contrario la asistencia regular a clases, el interés por la lectura, el apoyo de los padres, las buenas relaciones con sus pares y las expectativas de los docentes y familia sobre el logro académico son variables asociadas de forma positiva al logro académico en las *evaluaciones educativas estandarizadas* de la región (Castro-Aristizabal, Castillo-Caicedo, & Mendoza-Parra, 2016; Suárez-Enciso, Elías, & Zarza, 2016).

En cuanto al rol de los centros educativos, las variables relevantes son el tipo de población al que atienden, los recursos e infraestructura de las instituciones educativas, las características y las practicas pedagógicas de los docentes se destacan como principales factores de eficacia escolar (UNESCO-OREALC & LLECE, 2000, 2010, 2016a). La región latinoamericana se caracteriza por un nivel de segregación elevado (Murillo & Martínez-Garrido, 2017b), donde se observa una amplia brecha entre el logro educativo entre agrupaciones como las indígenas y no indígenas (Webb, Canales, & Becerra, 2017; Woitschach, 2016); y las agrupaciones de migrantes y no migrantes (Ting & Ronald, 2017). Otras variables asociadas a la segregación escolar provienen del contexto sociodemográfico, como ser la diferencia de logro educativo entre instituciones educativas de carácter público o privado y las instituciones educativas de zonas urbanas o rurales

(Arcidiácono et al., 2014; Balarín, 2016; Castro-Aristizabal & Giménez, 2017; Murillo, 2016; Murillo & Martínez-Garrido, 2017a, 2017b).

Los resultados académicos demuestran que los centros donde el profesorado presenta bajos niveles de ausencia laboral, donde existen mayores recursos y acceso a recursos tecnológicos presentan un mayor nivel de logro educativo (UNESCO-OREALC & LLECE, 2016a). Así también los centros educativos donde existe liderazgo educativo y gobernanza, el rendimiento académico de los estudiantes es mayor (Hernández-Castilla, Murillo, & Martínez-Garrido, 2013; OECD, 2016b; Sans-Martín, Guàrdia-Olmos, & Triadó-Ivern, 2016).

Desde una visión más amplia a nivel de sistema educativo, América Latina se caracteriza por ser un contexto social de escasos recursos educativos, si bien el progreso en materia de salud, nutrición y equidad de género se encuentra estrechamente ligado al progreso en educación (UNESCO, 2016); esa relevancia se ve poco reflejada en el nivel de inversión en materia educativa, que en promedio ronda entre el 4,5% al 5,2% del producto interno bruto (BID, 2017). Este bajo nivel de inversión sumado al contexto social en donde la educación se desarrolla impacta de forma directa en el logro educativo y en la calidad docente. En promedio, aproximadamente un tercio de los alumnos de nivel primario de los países de la región no parece haber adquirido aprendizajes básicos y se observa que el acceso a las oportunidades de aprendizaje es limitado (UNESCO-OREALC, 2016a), mientras que el profesorado desarrolla su trabajo en condiciones laborales deficientes, caracterizado por las escasas oportunidades de desarrollo profesional (UNESCO-OREALC, 2013; UNESCO-OREALC & LLECE, 2016a).

Metodología en las Evaluaciones Educativas Estandarizadas del LLECE

Posicionados desde el contexto educativo y socioeconómico presentado en los párrafos precedentes y dado que la principal finalidad de las evaluaciones educativas es la de acumular datos y evidencias objetivas que permitan la toma de decisiones y la mejora de los sistemas educativos (Fernández-Alonso, 2004), los programas de evaluación educativa estandarizada, han incursionado en la mejora y promoción de avances relacionados con la metodología del proceso de evaluación, la sistematización de los resultados y los análisis secundarios con el fin de garantizar la credibilidad de los resultados obtenidos.

En lo que a metodología del proceso de evaluación se refiere el NAEP con el fin de proveer un sustento robusto al proceso de evaluación educativa ha introducido importantes avances técnicos en los campos del testeo y la evaluación, la validez de contenido, la especificación de los factores asociados, de las propiedades métricas de los estimadores de resultados del alumnado, el escalamiento de respuestas, los diseños muestrales complejos y los diseños de evaluación matricial, el test equating, la adaptación cultural de las pruebas, y las estrategias de comunicación de resultados a las diferentes audiencias. Avances técnicos que le han permitido ser un pionero en la evaluación nacional y convertirse en un referente de evaluación educativa a nivel mundial (Beaton et al., 2011; Harmon et al., 1997; Messick, Beaton, & Lord, 1983; Mislevy, 1992; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992).

A su vez PISA, ha sistematizado los avances metodológicos producto de su colección de estudios en una serie de reportes técnicos (OECD, 2002, 2005, 2009, 2012, 2014) que al igual de la serie de reportes técnicos de PIRLS y TIMMS (Joncas & Foy, 2012) conforman el acervo histórico del desarrollo metodológico de las *evaluaciones educativas estandarizadas* a escala mundial.

El objetivo de este apartado es el de mostrar un recorrido sobre los principales avances metodológicos que los estudios del LLECE ha incluido desde sus inicios. Ya que estos avances nos permitirán (esperemos) analizar el concepto de equidad del proceso de evaluación que se desarrolla en un contexto caracterizado por la desigualdad de oportunidades educativas (UNESCO-OREALC, 2016a). Tres preguntas son abordadas en los primeros párrafos de este apartado ¿Cuál es la finalidad de las evaluaciones del LLECE?, ¿Es este programa realmente exitoso en la región? y por último ¿Cuál es la evolución metodológica por la que ha atravesado?

La UNESCO enfatiza que el objetivo fundamental de esta colección de estudios es el de determinar el nivel de desempeño escolar desagregado por materias evaluadas, además de conocer cuál es la relación entre el desempeño escolar y los factores asociados ya sean estos vinculados a las características de los estudiantes, sus familias, las escuelas o el sistema educativo (UNESCO-OREALC & LLECE, 2016b). Sus estudios tienen el propósito de entregar información útil para formular y ejecutar políticas educativas en todos los países de la región (UNESCO-OREALC & LLECE, 2000, p. 7). A diferencia de PISA y en consonancia con los estudios de los estudios liderados por la IEA, los estudios del LLECE se encuentran orientados al currículo educativo del conjunto de países participantes. Donde la educación es vista como un concepto multidimensional y el logro académico es un elemento imprescindible, pero que no se constituye como el único indicador de éxito del sistema educativo (UNESCO-OREALC & LLECE, 2016b, p. 3).

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. Países como *Argentina, Brasil, Chile, Colombia, Costa Rica, México, Paraguay, Perú y República Dominicana* han sido constantes en su participación

en la triada de estudios iniciados en 1997. 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 ya que Bolivia, Honduras y Venezuela desistieron, pero Ecuador, Salvador, Guatemala, Nicaragua, Panamá y Uruguay además del estado nacional de Nuevo León (México) se sumaron (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 y solo Cuba y San Salvador se ausentaron (LLECE, 2014). Actualmente, el LLECE coordina la organización del Cuarto Estudio Regional Comparativo y Explicativo (ERCE) a ser aplicado en el año 2019 y que contará con la participación de diecisiete países entre los que se destaca el regreso de Cuba, Bolivia, El Salvador y Venezuela.

Dada la alta participación de los países de la región este programa tiene un gran impacto tanto mediático, académico como investigador. Cuando se hacen públicos los resultados, los gabinetes de comunicación de los ministerios nacionales y los organismos internacionales preparan notas de prensa que son consumidas ávidamente por los medios de comunicación. Así mismo, se realizan *informes de resultados* (LLECE & UNESCO-OREALC, 2016; OREALC-UNESCO, 2014; OREALC-UNESO & LLECE, 2008; UNESCO-OREALC, 2013, 2014, 2016a; UNESCO-OREALC & LLECE, 2000, 2010, 2016a, 2016b), *reportes técnicos* (UNESCO-OREALC-LLECE, 2010; UNESCO-OREALC, 2016b; UNESCO-OREALC & LLECE, 2001) y *análisis secundarios* que son publicados en revistas especializadas (Murillo & Martínez-Garrido, 2016).

Hasta aquí hemos respondido a las dos primeras preguntas planteadas en este apartado ¿Cuál es la finalidad de las evaluaciones del LLECE?, ¿Es este programa

realmente exitoso en la región? En general podemos destacar de que el éxito de las evaluaciones educativas a nivel regional radica en que éstas iniciativas sirven de base a las políticas y programas educativos para la rendición de cuentas (UNESCO-OREALC, 2016a). Son una medida precisa del desempeño escolar que se encuentra contextualizado a países de América Latina y el Caribe. 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.

En este segundo momento nos centraremos en lo que al proceso metodológico se refiere con el análisis conjunto de la triada de reportes técnicos producto de las evaluaciones realizadas por el LLECE en 1997 (UNESCO-OREALC & LLECE, 2001), en el 2006 (UNESCO-OREALC-LLECE, 2010) y la evaluación del 2013 donde observaremos una etapa de madurez metodológica de la región (UNESCO-OREALC, 2016b). Este recorrido comprende desde la instrumentación técnica, el proceso de evaluación y puntuaciones asignadas, así como el reporte de los resultados.

Las evaluaciones educativas del LLECE han sido de carácter curricular desde sus inicios, por lo que, en cuanto a *los instrumentos y el desarrollo de los marcos de evaluación*, un análisis curricular conjunto de todos los países participantes es realizado con anterioridad con el objetivo de determinar los aspectos comunes y los contenidos a evaluar en las pruebas. Los estudios del LLECE al igual que otras pruebas de evaluación educativa, utilizan tablas de especificaciones de contenido de doble entrada que contienen dominios de contenido y ejes temáticos (UNESCO-OREALC, 2016b). Un profundo análisis curricular es realizado para la determinación de los contenidos comunes de las mallas curriculares de cada país participante.

La prueba presenta un diseño matricial de bloques incompletos y los ítems que constituyen la prueba son de opción múltiple y de preguntas abiertas codificadas en formato binario y de crédito parcial. Previa a la aplicación final de la prueba, se realizan pruebas pilotos, que son aplicadas a muestras representativas provenientes de todos los países participantes, fruto de ese análisis es la selección de ítems para la aplicación final.

Entre los análisis realizados tanto para las pruebas pilotos, así como el análisis final de las pruebas del estudio se focalizan en la teoría clásica de los test (TCT) y en la teoría de la respuesta al ítem (TRI). Teniendo en cuenta la participación de Brasil, se realiza el proceso de traducción de las pruebas cognitivas al idioma portugués. Aunque si bien los reportes técnicos no presentan una información detallada sobre el proceso de traducción, se detalla que las pruebas son enviadas al Brasil para su traducción al portugués y cotejadas posteriormente por los expertos de la UNESCO (UNESCO-OREALC, 2016b, p. 63). La ausencia de esta información clave limita la realización de otros análisis psicométricos, tales como el funcionamiento diferencial de los ítems o la comparabilidad entre las versiones idiomáticas.

En cuanto al sesgo y funcionamiento diferencial del ítem-escala. El análisis del funcionamiento diferencial de los ítems es incluido en las versiones del SERCE y el TERCE, bajo la técnica de Mantel-Haenzel y analizando las variables de país y género. Si bien el informe técnico de TERCE indica que se incluirá el análisis del DIF desde la TRI, los resultados no son presentados en el informe ya que la presencia de DIF no es un criterio de eliminación de ítem y se considera información complementaria para los países.

La *estimación de las puntuaciones* es realizada mediante la TRI y valores plausibles son utilizados tanto para el cálculo de los resultados de logro cognitivo como para el análisis de factores asociados. Las escalas de habilidad de competencia han variado desde la primera aplicación del estudio. 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 de las principales utilidades de la existencia de una colección de medidas es la posibilidad de comparación de series longitudinales. En este caso el LLECE ha incorporado ítems de anclaje de SERCE en TERCE con el objetivo de comparar con los resultados de ambas pruebas. Si bien el reporte técnico de TERCE afirma que el contenido de las pruebas es el mismo, dado que los instrumentos provienen del análisis curricular y de una tabla de especificaciones construida con procesos idénticos, es importante destacar que el currículo de los países presenta variaciones propias de las reformas y programas educativos instaurados entre el 2009 y el 2013, por lo que el marco de contenido no exclusivamente el mismo. La comparabilidad SERCE-TERCE es realizada mediante el scaling y el equiating dentro del marco de la TRI optando por una calibración concurrente.

El *diseño muestral* del estudio de 1997 fue clasificado como bietápico aleatorio, mientras que para SERCE fue aleatorio estratificado por conglomerados de una selección (nivel de escuela) y en el caso de TERCE aleatorio sistemático por conglomerados bietápico (escuela y aula). En PERCE los estratos de selección inicialmente fueron sobre las características demográficas (mega ciudad, urbano, rural) y el seguidamente por el tipo de administración de los centros (público, privado). Mientras que en SERCE y TERCE

los estratos de selección son urbano público, urbano privado y rural. La participación de estados nacionales se dio a partir del SERCE, evaluación que incluyo a Nuevo León y Goias (Brasil), mientras que TERCE conto por segunda vez consecutiva con la participación de Nuevo León (México). Debido a que la muestra no es auto ponderada, pesos muestrales son aplicados. Estos pesos muestrales tienen dos aplicaciones importantes. La primera es que permiten reconstruir el tamaño poblacional del país o estrato estudiado. Y una segunda aplicación importante de estos pesos, es que permiten emplear métodos de replicación o re-muestreo.

Las principales diferencias en el diseño muestral entre SERCE y TERCE se refieren a las variables y la tasa de exclusión, al cálculo del tamaño muestral en referencia a los valores mínimos de participación y a la no exclusión de alumnos con lengua materna no español. Entre las variables explicitas a diferencia de SERCE, el TERCE no incluyó una muestra a nivel de aula; lo que a su vez se presenta como una desventaja al momento del análisis de los factores asociados al rendimiento académico. Otro aspecto de suma importancia fue la incorporación del sobre-muestreo como consecuencia de la infra participación observada en algunos países en PERCE y SERCE, ya que el sobre-muestreo en estos casos fue opcional para los países participantes.

Los cuestionarios de contexto son construidos para la evaluación de características socio-contextuales de los estudiantes, sus familias, los profesores y directores de las instituciones escolares. Estos cuestionarios en su mayoría presentan valores psicométricos aptos para su aplicación y se constituyen en una herramienta eficaz para la generación de resultados ajustados a las características propias del contexto. En cuanto a la presentación de los resultados, informes de logro cognitivo y factores asociados son publicados por la organización, en años posteriores a cada evaluación.

Finalmente, volvemos a considerar la idea de que TERCE se basa en el uso de pruebas estandarizadas, con la finalidad de obtener información válida y objetiva para el diagnóstico e intervención, en todos los niveles del sistema educativo. Por lo que en consecuencia y dada la relevancia del estudio en América Latina, el punto focal de esta investigación incluye y analiza desde una perspectiva ecológica en cinco estudios la equidad de las evaluaciones educativas estandarizadas de la UNESCO en América Latina y el Caribe. Los estudios de esta Tesis Doctoral responden a objetivos de investigación específicos y se fundamentan en las materias educativas que a criterio de Baker (2014) son las materias más utilizadas para la comparación internacional. El estudio se erige como un análisis metodológico del proceso de evaluación desde una mirada *in vivo*, antes que *in vitro* (Zumbo & Hubley, 2016).

PREGUNTA, OBJETIVOS E HIPÓTESIS DE LA TESIS DOCTORAL

La Tesis Doctoral se encuentra organizada en cinco estudios que conforman capítulos sobre la evidencia de equidad de las evaluaciones estandarizadas lideradas por el LLECE. Dos son los aspectos para profundizar; por un lado, la equidad del contexto educativo y por el otro el impacto de este en la validez de las medidas en las pruebas de evaluación educativa estandarizada. Conjugados en un cuestionamiento clave sobre:

¿Cuáles son las amenazas del contexto socioeducativo que impactan en la justicia en el uso y la interpretación de los resultados obtenidos en las pruebas de evaluación educativa estandarizada TERCE?

Tomando en cuenta la principal finalidad del programa de evaluación educativa del LLECE se analiza la posibilidad de comparación de las medidas entre las dieciséis naciones. Seguidamente se identifican y analizan los factores asociados a la variabilidad de los resultados obtenidos en las pruebas educativas en cada nivel del contexto educativo (estudiante, centro y sistema educativo). Por último en el estudio final de esta Tesis Doctoral *(Capítulo V)*, en la misma línea de la teoría ecológica utilizada tanto para la construcción de las recomendaciones para las políticas públicas del LLECE (UNESCO-OREALC, 2016a), como para los estudios de la Tercera Generación del DIF de Chen y Zumbo (2017), se analiza el proceso de respuesta a la prueba desde una perspectiva innovadora que toma en cuenta la realidad socio-económica en donde las evaluaciones educativas estandarizadas del LLECE son desarrolladas.

Por consiguiente, los objetivos específicos de esta Tesis Doctoral son:

a) Establecer la presencia de invarianza de las medidas de la prueba de logro educativo de TERCE y analizar la viabilidad de la comparación entre países.
b) Determinar el efecto global de los centros escolares en el rendimiento de los estudiantes en los 15 países participantes en TERCE.
c) Especificar el impacto de la segregación escolar en el logro académico de los estudiantes en las pruebas del TERCE.
d) Conocer la influencia de las oportunidades de aprendizaje otorgadas por el contexto escolar y los sistemas educativos en la región latinoamericana en el logro académico de los estudiantes.
e) Determinar la presencia/ausencia de DIF y analizar las causas subyacentes al proceso de respuesta al ítem desde una perspectiva ecológica.

Derivando en las siguientes hipótesis de investigación,

a) La presencia de variabilidad del constructo a través de los países participantes deriva en la imposibilidad de comparación entre países.
b) Por lo que en consecuencia se espera determinar las variables que se asocian con la variabilidad de las puntuaciones obtenidas. Donde el contexto socioeconómico de las familias y los recursos de las instituciones serán los principales determinantes del logro académico.
c) En cuanto a la conformación del aula dado el contexto socioeconómico, las aulas heterogéneas permitirán el aprendizaje equitativo mientas que las aulas homogéneas favorecerán el mejor rendimiento de los estudiantes de niveles socioeconómicos más altos.
d) Por otro lado, los recursos de los centros educativos favorecerán al alumnado, que obtendrá puntuaciones en rendimiento académico aun controlado el contexto socioeconómico de donde provienen.
e) Las variables relacionadas a las características individuales, así como las escolares y el sistema educativo en donde la prueba se desarrolla serán los principales determinantes y explicarán el funcionamiento diferencial de los ítems de la prueba.

CAPITULO I. Measurement Invariance of the Academic Performance for the Sixteen Nations of the UNESCO Assessment Program

ABSTRACT

In the context of international assessments, the comparability of scores between countries assumes that the measures are equivalent. UNESCO's Third Regional Comparative and Explanatory Study (TERCE) program reports on the results for mathematics, science, and reading for 15 Latin American countries and the State of Nuevo León in Mexico. A standard reporting practice is to rank order the countries according to their performance levels in each of these three subjects. An implicit assumption in this ranking is that the measures are sufficiently invariant to allow an un-confounded interpretation. The study's objective is to investigate the use of a relatively newly developed psychometric method—the alignment method (Asparouhov & Muthén, 2014)—for the analysis of the measurement invariance for the TERCE and to determine the comparability of the scores obtained in the assessment. The analysis was carried out with 82 items of the Science test applied to 61,921 students. The alignment method was used for the item pool of the test under the MLR estimation strategy to assess the approximate measurement invariance. The data analyses were performed with the Mplus 8 program. The preliminary results indicate that the alignment method based on a configural model automates the process of invariance measurement. In summary, the research shows the effectiveness of the use of the technique for the detection of invariance in complex samples, providing evidence of non-invariant items that may affect the validity of interpretations in cross-cultural comparisons.

Key words: Invariance, alignment, standardized educative evaluation, complex survey, UNESCO.

INTRODUCTION

Around the globe, researchers and decision makers in education are increasingly using standardized tests in order to make cross-cultural comparisons of achievements in education. A clear example of this practice is the Educational Trend Studies of the International Association for the Evaluation of Educational Achievement (IEA), which have been published over the past two decades. TIMMS the latest version, evaluates mathematics and sciences in 64 countries (Martin, Mullis, & Hooper, 2016), whereas PIRLS evaluates reading comprehension in 61 countries (Mullis, Martin, Foy, & Hooper, 2017). Similarly, the Program for International Student Assessment (PISA), in its latest version, analyzes student knowledge in mathematics, science, and reading in 72 countries (OECD, 2016).

At the regional level in Latin America since 1997, UNESCO has been assessing the knowledge of elementary education students using Third Comparative and Explanatory Study of Education (TERCE), which focused on the evaluation of mathematics, reading, and science in 15 countries (LLECE, 2014). As a partner to the international studies, the localized national studies aim to establish monitoring of internal education policies as well as to allow for a local comparison of the specific needs of each education system.

Diverse cultural contexts characterize international assessments as the multicultural factor plays a decisive role in the validity of the scores and in the possibility that these scores are comparable between countries or regions (Ercikan, Roth, & Asil, 2015). Due to the high impact of the results of such studies, in order to adequately guarantee the efficiency of the evaluation and the credibility of the data, the educational

evaluation programs have promoted advances related to the systematization of procedures and secondary analysis.

In the educational achievement reports, the results of each country are presented in what is commonly known as the league table where the countries appear as a ranked list in order to easily show performance differences between countries. As the interpretation and use of these results signficantly impact the educational policy decisions at regional and international levels, the importance of the validity of the measures cannot be overestimated (Sireci, 2015). Due to that high stakes of the tests results on educational programs and society in general, there is much interest in detecting measure variability in order to minimize the bias of the items (AERA, APA, & NCME, 2014; Hambleton, Merenda, & Spielberger, 2005; International Test Commission [ITC], 2018, 2017).

Rutkowski and Svetina (2014) have reported that only TALIS 2008 and PISA 2012 performed invariance analyzes for the students' context questionnaire, and that just TALIS has published the results in technical reports (OECD, 2010) or working documents (Desa, 2014). Although researchers has been considerable effort expended to guarantee the equivalence of the measures, there has been less marked interest in the invariance analysis of international assessments (Byrne & van de Vijver, 2017; Johnson et al., 1994; OECD, 2014; UNESCO-OREALC, 2016). The lack of engagement with the invariance analysis is likely due to the factors involved with complex data survey, which include the matrix design of items, a large sample design with stratifications, weights, and clustering.

Measurement invariance

Among the antecedents of the invariance measurement, Jöreskog's (1971) seminal work with the analysis of the similarities and differences of the factorial structures between

groupings is notable. Sörbom (1974), as the pioneer in the specification of the estimation of the means and latent factors, is also significant, as he used a strategy based on structural equation models in Lisrel. Building on this early work, Byrne, Shavelson, and Muthén (1989) then introduced the distinction between the measurement of partial and incomplete invariance. Years later, Meredith (1993) went one step beyond measurement and took on the in-depth analysis of the consequences of MI on the validity of the tests. More recently, Zumbo (2013) has indicated that the paradox of the study of invariance is, perhaps from a mathematical perspective, a trivial concept. From a historical view of psychometrics, however, invariance is one of the essential properties of the theory of item response theory (IRT). For more than three decades, this theory has been the basis of the construction of the tests and the adjustment of the items in the educative assessments, allowing not only for the matrix design of the items but also for the establishment of performance trends.

The measurement of invariance under the technique of multiple-group confirmatory factor analysis can be applied to the comparison between countries or between cultural groupings of the same country and either or focusing on characteristics such as language, belonging or not belonging to a specific sub-group (such as one based on race), migration conditions, gender, and socioeconomic status. A study of invariance using the multiple-group confirmatory factor analysis (MG-CFA) under the MACS strategy for the mathematics test from 21 countries in TIMMS 1999 has shown a pattern of invariance (Wu, Li, & Zumbo, 2007). Another study looked at 55 countries with the PISA 2009 reading test using MG-CFA and included characteristics such as culture, economic development of the country, and language; it found that the reading test is not invariant among the 55 economies (Asil & Brown, 2016). Elosua and Mujika-Lizaso (2013) evaluated the linguistic versions of the PISA 2009 reading test, which was applied

in Spain in four different language versions (Spanish, Basque, Catalan and Galician). They found evidence of metric equivalence between the four linguistic versions. In the same way, Segeritz and Anand-Pant (2013) evaluated the invariance of the PISA 2003 students' questionnaires applied in Germany, and they found evidence for comparability of measures between cultural groups (immigrants and non-immigrants).

Among the main difficulties of effectively utilizing the CFA technique is having to deal with models that have a high magnitude of restrictions, which require modification in the indexes in order to achieve an acceptable fit to establish a comparison between several groups (Asparouhov & Muthén, 2014; Byrne & van de Vijver, 2017; Muthén & Asparouhov, 2014). This challenge is in addition to the difficulty of managing lost values due to the matrix design of the tests (Asil & Brown, 2016). A historical review of the main techniques has demonstrated the advantages and difficulties in evaluating invariance in international comparisons (van de Schoot, Schmidt, de Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015). With the aim of solving the difficulties of traditional techniques, Asparouhov and Muthén (2014) presented a new psychometric method, "The Alignment." This procedure represents a new generation of methodology for scaling latent variable framework (Munck, Barber, & Torney-Purta, 2017), shows a substantial improvement in the detection of approximate invariance between groups, and allows for the use of complex data survey.

The research goal of this study is to incorporate this novel alignment method for the analysis of invariance in large-scale assessments, which implies the use of several indicators and complex survey design (e.g., stratification, sample weights, items matrix design and several clusters). Due to the objectives the research goal, the theoretical introduction will focus on the specification of the method and the type of estimation.

Additionally, the discussion will highlight their primary applications in both simulation and real data studies. For a more in-depth detail of the alignment's algorithm and procedure, articles such as those by Asparouhov and Muthén (2014) and Muthén and Asparouhov (2014), which incorporate a technical vision, or by Byrne and van de Vijver (2017), which uses a more pedagogical perspective, can guide both specialized and novel users. Finally, the study presents an empirical study with 61,921 students belonging to 2,955 schools in 15 countries of Latin America.

The Alignment Method

The alignment method is an exploratory tool that allows the estimations of factor means and variances, while its results offer information about the approximate invariance measurement (Asparouhov & Muthén, 2014). It is based on the idea of starting the invariance analysis from a configural model that does not show invariance; from there it starts to search for the most substantial number of invariant parameters that are possible leaving the mean and covariance of the factors to vary between groups. This criterion of rotation is similar to that used in the exploratory factor analysis (Muthén & Asparouhov, 2014). Once the alignment is completed, a detailed analysis can be made to determine which parameters of measurement are approximately invariant and which are not. Even though the alignment began with an exploratory optics, Marsh et al. (2017) introduces the alignment-within-CFA (AwC) as a confirmatory perspective. Given these adjustments, the alignment method demonstrates a greater flexibility in the face of traditional confirmatory methods.

Alignment optimization procedure.

The alignment optimization procedure began with the estimation of the configural model, where factor mean is $\alpha_g = 0$ and factor variance is $\psi_g = 1$ for each group *g*. As well, the intercept and parameters are estimated without restrictions while discovering the most optimal measurement invariance pattern (Asparouhov & Muthén, 2014). This analysis work uses a non-identifying model that recognizes the parameters by choosing a certain rotation, which results in a simple and interpretable factor loading pattern. This process is possible by incorporating a simplicity function *(F)* similar to the exploratory approach. The function *F* accumulates the total non-invariance measurement. The simplicity function implies that for each pair of groups, every intercept and loading parameter that the alignment method adds to the total loss function difference between the parameters scaled through the component loss function (CLF) *f*.

After, the total loss function *F* is minimized in a solution where there are few large non-invariant measurement parameters and many approximate invariant measurements instead of many medium-sized non-invariant measurement parameters (Muthén & Asparouhov, 2014). The alignment method has two optimization procedures, namely FIXED and FREE. The FIXED alignment optimization assumes that $\alpha_1 = 0$, while the FREE alignment optimization estimates α_1 as an additional parameter. The optimization procedure is available with Maximum Likelihood or Bayesian estimation. Asparouhov and Muthén (2014) have showed the advantages of the Bayesian alignment estimation over the Maximum-Likelihood to be: the model flexibility, the improvement of the model fit by using BSEM with small residual covariances, and the facility of the results interpretation. Taking into account the alignment research papers published since 2013 to 2018, the most common estimation methods reported are ML or MLR. Based on both the

data characteristics as well as on the recommendations of Byrne and van de Vijver (2017), this study focuses on MLR estimation. The MLR estimation uses the standard error calculations with the Huber-Whit sandwich estimator (Muthén & Asparouhov, 2014), which allows work with complex samples that include stratified designs (e.g., countries and regions), clusters (e.g., schools), and sample weights.

Alignment ad-hoc testing procedure

Although the first purpose of the alignment is to produce a comparison between factor means and variances, the method produces an output of the degree of measurement invariance. After the model is estimated, the invariance analysis is performed on all parameters, specifying one parameter at a time and comparing between two groups. These comparisons are made again and again between the groups to create a set where each parameter is evaluated against the mean of the set of invariants (Byrne & van de Vijver, 2017). If this parameter, for each group, is significantly different from the mean, then it is called a non-invariant parameter. The author controls the type I error with an algorithm that controls the value of alpha in .001. The program produces for each parameter of the model the groups that are invariant paired according to the differences in the p-value (Asparouhov & Muthén, 2014). Complete information about which parameters are invariant and non-invariant are showed in the results. The author's indications based on simulation data is to follow a cut point of 25% limit of non-invariance for accurate alignment results (Muthén & Asparouhov, 2014). Additionally, a measurement of R^2 for each measurement parameter provides parameter variation between groups in the configuration model that is explained by the variation in the mean and variance of the factor between the groups.

Alignment method applications

Kim, Cao, Wang, and Nguyen (2017) have studied the alignment method in contrast with other approaches to measurement invariance. Their research concluded that both Bayesian and alignment optimization are recommended for approximate invariance or non-invariant item detections. In particular, they remarked that the power of invariant and non-invariant groups detection decreases when the sample is small, and that it fails when the group sample is large. Based on their simulation study and method comparisons, Kim et al. explained the alignment method is useful when the research objectives are to establish approximate invariance and the factor means comparison are of focal interest. Following studies with simulated data, Flake and McCoach (2017) used polytomous indicators under conditions of partial measurement invariance with MLR estimation. Their work has shown that the alignment method has an excellent recovery of the exact parameters and produced estimates with little bias (and that bias can be increased when the pattern of non-invariance is higher). Additionally, they mentioned the method does not always flag the non-invariant items when they are in the medium or small conditions, and that it shows a better performance for the intercepts than for the loadings. Considering the information provided by Flake and McCoach (2017), in this study, R^2 works better for some items than for others and is influenced by the group size.

The study of Munck, Barber, & Torney-Purta (2017) compared attitudes toward immigrants in 92 groups organized by gender and country. The results has demonstrated that the use of the alignment with MG-CFA makes it feasible and comprehensible to assess measurement invariance in large data sets with an automatized process. Taking into consideration research results from the alignment theory, the performance for the attitudinal test (tradition and conformity) in 26 countries with 49.980 subjects shows a

percentage of non-invariance between 16% for factor loadings, 50% for intercepts, and an overall of 33% (Muthén & Asparouhov, 2014; Muthén & Asparouhov, 2017). Based on PISA 2009 with 53 countries for students' questionnaires, the percentage of non-invariance allowing was 21% factor loadings, 63% intercepts, and an overall of 42%.

Byrne and van de Vijver (2017) performed the invariance analysis using alignment method on a family values scale under an ML estimation with 18 items and 27 countries, reporting a satisfactory performance of the alignment method. The non-invariant finding was, as a rule of thumb, less than 25%. Additionally, the study shows R2 values for invariant items >.31. Based on the total contribution (loadings and intercepts), the authors found a substantially different value for two of the entirely invariant parameters (where one of them shows a more considerable total contribution). Notwithstanding this difference, the researchers argued that the results can be associated with the size of the (smallest) groups for which significance is not easy to achieve.

Lomazzi (2018) compared the use of the MG-CFA and the alignment procedure with ML estimation for a gender role scale. The alignment results shows an acceptable degree of non-invariance in 35 of the 59 countries and an effective performance of the alignment method. Where the percentage of non-invariance was decreasing (1st at 51%, 2nd at 39%, 3rd at 27%, and finally 21%) by deleting from the model the non-invariant groups.

Until now the results of the use of alignment method based on complex cognitive tests have been unexplored. In order to address this gap in the literature, the present study asks the following questions:

(a) What is the feasibility of using the alignment method with complex samples, and is it reasonable to assume that measures of academic achievement are comparable between different educational contexts?

(b) Is it possible to use alignment as a method for the detection of invariance in large-scale educational evaluation tests?

(c) Are the scores of the academic achievement of the students participating in the TERCE tests comparable among the participating countries?

(d) If the test shows non-invariant items, what can be said about them?

METHOD

Sample

This study used a sample from the science test portion of the TERCE conducted in 2013, which included 15 countries plus a national state, consisting of 61,921 students (μ_{age} = 12.47, SD = 0.96) and an average of 2,813 participating schools. The TERCE objective was to evaluate the knowledge of 6th-grade students. The sample design has been stratified by conglomerates, with random and systematic selection in two stages. In these designs, the sampling units (i.e., schools, classrooms, and students) are selected in two or more stages and these said sample units do not have the same probability of being chosen. Table 1 illustrates a brief description of the sample distribution.

Table 1
Sample Distributions from Booklets

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5	Booklet 6
Argentina (1)	631	619	617	611	614	732
Brazil (2)	506	491	511	490	497	826
Chile (3)	840	846	837	828	834	824
Colombia (4)	718	725	731	718	722	804
Costa Rica (5)	590	579	588	585	590	836
Dominican Republic (6)	612	624	601	600	621	838
Ecuador (7)	807	798	798	805	805	833
Guatemala (8)	677	684	685	669	677	823
Honduras (9)	655	643	654	651	645	827
Mexico (10)	600	600	602	607	608	840
Nicaragua (11)	634	626	610	623	628	833
Panama (12)	605	587	586	577	603	839
Paraguay (13)	539	542	541	538	540	814
Peru (14)	807	805	806	794	796	821
Uruguay (15)	471	406	470	476	470	821
Nuevo Leon (16)	705	697	706	704	698	822
Student's sample	10397	10326	10343	10276	10348	10231
Cluster's sample	2823	2816	2803	2812	2818	2811

Source TERCE Technical Report (UNESCO-OREALC, 2016, p.26)

Description of the 6th-grade science test

The science test evaluates three cognitive processes (recognition of information and concepts; understanding and application of concepts; and scientific thinking and problem solving), and five domains of knowledge (health, living beings, environment, the earth and the solar system, and matter and energy). The items were composed of multiple response options and constructed responses, and the final data set has the responses coded as dichotomous and partial credit items (UNESCO-OREALC, 2016). The definition of the domain is the reference to the operational definition of the content. Educational evaluations use the insertion of tables of double entry, where is indicated the content or

areas for the domain evaluated and the operations or cognitive processes employed for the resolution of the test problems.

Table 2
Matrix Designs of TERCE Items

Domain	Process			Total	%
	Recognition of objects and elements	Understanding and applications of concepts	Scientific thinking and problem resolving		
Health	5	7	7	19	21%
Living beings	9	10	6	25	27%
Ambient	4	15	3	22	24%
The earth and the solar systems	3	6	4	13	14%
Matter and energy	3	6	4	13	14%
Total	24	44	24	92	100%
%	26%	48%	26%	100%	---

Source TERCE Results Report. Learning achievements (UNESCO-OREALC & LLECE, 2016, p. 82)

Moreover, the science test was composed of 92 items that were distributed in six blocks or clusters between 15 and 16 items each. These blocks were distributed in six different booklet models by an incomplete block design. This way of organizing the items allows the results to be generalized to the entire population, although the students respond only to a small part of the item bank (Fernández-Alonso & Muñiz, 2011). Each booklet was made up of two blocks or clusters of items between $\sum$=26 and 30 items, and each cluster appeared twice throughout the collection of booklets (once at the beginning of the booklet and a second time in the second position of the booklet). This rotation of the cluster presentation order allowed for controlling the effect of the order of the positions. Finally, the test contained two anchoring blocks with the 2009 SERCE test, which allowed for trend studies (UNESCO-OREALC, 2016).

Table 3
Matrix Design of TERCE Science Test

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Total, Items	Deleted item	Final item/database TERCE
Booklet 1	1	2					-	0	30
Booklet 2			1	2			-	3	27
Booklet 3	2				1		-	4	27
Booklet 4			2			1	-	5	26
Booklet 5		1				2	-	3	28
Booklet 6				1	2		-	5	26
Total	15	15	15	15	16	16	92	10	82

Source TERCE Technical Report (UNESCO-OREALC, 2016, p.150)

Variables of interest

All the variables included in the analysis are from UNESCO data collection and are available online for researchers who are interested in educational data sets (UNESCO-OREALC, 2016). The variables extracted from the 6th-grade science test are as follows:

(a) Six booklets corresponding to the science test: A final sample provided by TERCE of 82 dichotomous items (n = 79) and partial credit (n = 3).

(b) Stratification variable: stratification country1, classification of the 15 countries of Latin America (Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, and Uruguay) and a national state (State of Nuevo León, Mexico).

(c) Cluster variable: ID_cluster refers to the classification of the participating schools.

(d) Weight variable: The swgc is the senatorial weight of each country provided for TERCE.

Analysis

Based on the research goals, the analysis will start in consecutive order of procedure. First, individual databases were created for each booklet applied (n = 6) following the distribution of items of the test matrix. It is important to note that all the alignment analysis was carried out separately for every booklet (B1, B2, B3, B5, B6). A total of 82 items were applied in TERCE science test, where 79 items were binary (0 incorrect and 1 correct) and 3 were partial credit items where total credit means the correct answer (2), partial credit: partial correct answer (1), no credit: incorrect answer (0). All items (binary and partial credit[3]) were modeled at the same time in every booklet in a one-dimensional model.

The study of the booklets dimensionality has been omitted since TERCE assures compliance with that assumption (UNESCO-OREALC, 2016, p. 247). In the second stage, the alignment analysis was the focus. The booklets' analysis was performed using the MLR estimation since the stratification, cluster, and sample weight data were included in the analysis under the mixture complex calculation strategy (Muthén & Muthén, 2017). Initially the alignment analysis was specified FREE and based on the result outputs FIXED in booklet 2 (FIXED in group 11) and booklet 3 (FIXED in group 14)[4].

3 The results of the additional analysis performed for every type of item codifications (partial credit: fully correct and fully incorrect model) are partially included in this article.

[4] Mplus syntax used in this study is presented in the Appendix A.

RESULTS

Although the science test was constituted by 82 items, each of the booklets consisted of an average of 30 items. It is also important to note that, given the matrix design, each of the items was applied in two different booklets with a different position. With this information in mind, for a more accessible comprehension of the primary analysis, the collection of booklets will always be considered in a joint comparison, following the same pattern of alignment optimization results (invariance/non-invariance pattern and factor mean comparisons).

The alignment approach began with the configural model, where factor means is $\alpha_g = 0$, and factor variance is $\psi_g = 1$ for each group, and every factor loading, and every item intercept were freely estimated. For each booklet, the analysis began with the FREE alignment optimization. Based in a Mplus warning message, the reference group with factor mean value close to 0.0 is specified using the FIXED alignment option for booklet 2 (group 11) and booklet 3 (group 14).

Also, the output results show the pattern of invariant and non-invariant parameters for every intercepts/threshold and factor loading, showing in bold parenthesis the non-invariant groups. Based on the amount of information from the Mplus outputs, for this study, only the non-invariant parameters are reported. Taking into account the global percentage of non-invariant parameters in Table 4, all the percentages are below the cut point (25%) proposed by Muthén and Asparouhov (2014). Booklet 2 possesses the lowest percentage (2,31%) and booklets 1 and 5 have the highest values (3.52% - 3,77%).

Following these results, 15 is the average of non-invariant items across all the booklets. Among the 16 groups compared, there is some level of variability presented.

The findings indicate booklets 2, 3, 4, and 6 have non-invariant parameters for factor loading, while booklets 1 and 5 have shown less than 1% of non-invariant parameters. According to other studies, the percentage of non-invariant parameters is concentrated in the intercepts than in the loading factors (Byrne & van de Vijver, 2017). This fact corresponds to Muthén (2013), whom points out that the invariance must be focused on item difficulty rather than on the factor loadings.

Following with the non-invariant results and examining the average per countries group in Table 5, it is observed that all the groups are non-invariant in at least one intercept. Chile has the highest amount of non-invariant, followed by Brazil with 19 non-invariant intercepts and Costa Rica, Guatemala, and Nuevo Leon with 12 non-invariant intercepts. On the other extreme side, Paraguay and Colombia are non- invariant in 3 or 4 non-invariant intercepts and are followed by the rest of non-invariant countries in 6 and 12 parameters. Finally, Uruguay has a non-invariant intercept only in booklet 5.

Table 4

Non-invariant Parameters by Booklets

Booklet 1			Booklet 2			Booklet 3		
Item	Item intercept	Factor Loading	Item	Item intercept	Factor Loading	Item	Item intercept	Factor Loading
IT1_1	Chile, Dominican Republic		IT 2_1	Brazil		IT 3_1	Dominican Republic	
IT1_3	Chile, Dominican Republic, Guatemala, Nicaragua, Nuevo Leon	Paraguay	IT 2_2	Mexico, Nuevo Leon		IT 3_2	Guatemala	
IT 1_4	Colombia, Costa Rica		IT 2_3	Dominican Republic		IT 3_3	Dominican Republic, Nicaragua	
IT 1_6	Honduras		IT 2_4	Panama		IT3_12	Peru	
IT 1_10	Brazil		IT 2_11	Chile		IT 3_15_1	Costa Rica, Ecuador, Guatemala, Nuevo Leon	
IT 1_11	Panamá		IT 2_12	Costa Rica, Honduras, Nicaragua		IT 3_15_2	Costa Rica, Guatemala, Nuevo Leon	
IT 1_15_1	Costa Rica, Ecuador, Honduras	Chile, Ecuador	IT 2_13	Chile		IT3_16	Brazil, Dominican Republic	
IT 1_15_2	Colombia, Costa Rica, Dominican Republic, Guatemala, Mexico, Nicaragua		IT 2_14	Chile		IT 3_18	Brazil	
IT 1_16	Chile, Dominican Republic		IT 2_15	Chile		IT 3_19	Guatemala, Panama	
IT 1_17	Brazil, México		IT 2_17	Brazil		IT 3_20	Dominican Republic, Mexico	
IT_1_18		Costa Rica	IT 2_18	Brazil, Chile, Peru		IT 3_21	Argentina	
IT 1_24	Chile		IT 2_20	Honduras, Paraguay		IT 3_22	Chile	
IT 1_30_1	Argentina, Brazil, Ecuador		IT 2_23	Paraguay		IT 3_24	Dominican Republic	
IT_1_30_2	Brazil, Ecuador		IT 2_26	Ecuador		IT 3_25	Costa Rica, Guatemala, Panama	
						IT 3_27	Chile, Nuevo Leon	
$\sum_{\text{test items}}$:30			$\sum_{\text{test items}}$:27			$\sum_{\text{test items}}$:27		
Noninvariant	6.05%	0.83%		4.62%			6.02%	
Average	3.52%			2.31%			3.06%	

Table 4

Non-invariant Parameters (continued)

Booklet 4			Booklet 5			Booklet 6		
Item	Item intercept	Factor Loading	Item	Item intercept	Factor Loading	Item	Item intercept	Factor Loading
IT4_1	Brazil, Honduras		IT 5_1	Chile, Dominican Republic		IT 6_3	Nuevo Leon	
IT4_2	Colombia, Mexico, Nuevo Leon		IT 5_2		Nicaragua	IT 6_5	Brazil, Chile, Peru, Nuevo Leon	
IT 4_3	Brazil		IT 5_3		Costa Rica	IT 6_7	Argentina	
IT 4_4	Chile		IT 5_8	Brazil, Nuevo Leon		IT 6_9	Nuevo Leon	
IT 4_6	Brazil		IT_5_9	Uruguay		IT 6_10	Chile, Paraguay	
IT 4_11	Chile		IT_5_13	Brazil		IT 6_12	Nuevo León	
IT 4_12	Nicaragua		IT 5_14	Brazil		IT 6_13	Ecuador	
IT 4_13	Chile		IT 5_15_1	Argentina, Brazil, Costa Rica, Ecuador	Ecuador	IT 6_15	Argentina, Chile, Peru	
IT 4_14	Argentina, Peru		IT5__15_2	Brazil, Costa Rica, Ecuador		IT 6_19	Dominican Republic	
IT 4_15	Peru		IT 5_16	México		IT 6_20	Honduras	
IT 4_17	Guatemala		IT 5_17	Chile, Peru		IT 6_21	Chile	
IT 4_20	Brazil		IT 5_18	Guatemala		IT 6_24	Nuevo Leon	
IT 4_23	Chile, Costa Rica		IT 5_19	Nicaragua		IT 6_25	Guatemala, Paraguay	
IT 4_24	Chile		IT 5_22	Chile		IT 6_26	Costa Rica	
IT 4_25	Chile		IT 5_23	Guatemala, Panama				
IT 4_26_1	Argentina, Chile, Ecuador, Panama		IT 5_25	Chile, Costa Rica				
			IT 5_26	Argentina, Brazil, Honduras				
			IT 5_27	Honduras				
			IT 5_28_1	Ecuador, Guatemala	Ecuador			
$\sum_{\text{test items}}$: 26			$\sum_{\text{test items}}$: 28			$\sum_{\text{test items}}$: 26		
Noninvariant	5.55%			6.66%	0.66%		5.06%	
Average	2.83%			3.77%			2.52%	

Table 5

Distribution of Non-Invariant Parameters by Countries and Booklets

Groups	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5	Booklet 6	Total
Chile (3)	4	5	2	7	4	4	26
Brazil (2)	5	3	2	3	6	1	20
Costa Rica (5)	3	1	3	1	3	1	12
Guatemala (8)	2	0	5	1	3	1	12
Nuevo Leon (16)	1	1	3	1	1	5	12
Dominican Rep. (6)	4	1	5	0	1	1	12
Ecuador (7)	3	1	1	1	3	1	10
Argentina (1)	1	0	1	2	2	2	8
Honduras (9)	2	2	0	1	2	1	8
Nicaragua (11)	2	1	1	1	1	0	6
Peru (14)	0	1	1	2	1	2	7
Mexico (10)	2	1	1	1	1	0	6
Panama (12)	1	1	2	1	1	0	6
Paraguay (13)	0	2	0	0	0	2	4
Colombia (4)	2	0	0	1	0	0	3
Uruguay (15)	0	0	0	0	1	0	1
$\sum_{\text{Non-invariant intercepts}}$	32	20	27	23	30	21	153

The method aims to compare factor mean and variances across groups. The alignment solution has a few significant non-invariant parameters and many invariant parameters rather than many medium-sized non-invariant parameters, which allows for a meaningful comparison. Factor means comparisons for every booklet obtained in the alignment optimization is showed in Table 6. This factor means values were correlated with the average of the five TERCE plausible values for each participating country. The results of the correlation show a high degree of concordance between the score obtained in the test and the factor means provided by the alignment method (B1: .98, B2: .98, B3: .94, B4: .97, B5: .97, B6: .96).

This concordance between the data provided by the alignment and the average of the five plausible values of each group allows us to take a step beyond in the analysis.

Consequently, the analysis of the difference pattern between the means of the factors of each group is presented from the highest to the lowest value. In broad terms, the versatility of the results extracted from the output of the Mplus, since the mode of presentation allows the researchers to obtain an overview of the distribution of the means of the factors for all the booklets of the science test. First, it is highlighted that in 4 of the 6 booklets the first five positions are occupied by the same groups: Chile, Costa Rica, Mexico, Nuevo Leon, Colombia, and Uruguay. The lower end of the scores also houses almost the same collection of groups with some small variations: Paraguay, Nicaragua, Dominican Republic, Honduras, and Guatemala.

Booklet 1 shows that the first group of countries (Chile, Costa Rica, Mexico, and Nuevo Leon) have an average higher than at least nine of the remaining groupings. This group is followed by a second set of countries, Colombia, Uruguay, and Peru, where apparently the test has a different structure to the higher groups and the groups of lower value in the factor (Dominican Republic, Nicaragua, Paraguay, Honduras, Panama, and Guatemala). Booklet 2 shows practically the same pattern observed in booklet 1 with a group of countries with similar means factor in the central positions of the distribution (which show the same pattern of differences with the last five countries of the lower end). As for booklet 3, once again, Chile, Costa Rica, Mexico, Nuevo Leon, and Colombia form a group of similar factors means with precisely the same eight countries with lower means. Booklet 4 has in its upper-end Chile, Nuevo Leon, and Mexico, which make scores higher than the other nine countries. The same pattern is seen in booklet 5 for Nuevo Leon, Costa Rica, Chile, and Mexico with a factor mean different from the other 10 countries. Finally, in booklet 6, Costa Rica and Chile replicate the same pattern of nine countries with a significant small mean.

Table 6

Factor Mean Comparisons across Booklets

Book 1				**Book 2**			**Book 3**		
Ranking	G	FM	G with significantly smaller FM	G	FM	G with significantly smaller FM	G	FM	G with significantly smaller FM
1	Chile (3)	1.079	14-1-2-7-8-9-12-6-11-13	Chile (3)	1.198	7-2-8-1-14-11-12-9-6-13	Chile (3)	0.947	1-7-14-11-12-6-8-13-9
2	Costa Rica (5)	0.774	14-2-7-8-9-12-6-11-13	Nv. Leon (16)	1.068	2-8-1-14-11-12-9-6-13	Costa Rica (5)	0.595	14-11-12-6-8-13-9
3	Mexico (10)	0.696	14-1-2-7-8-9-12-6-11-13	Costa Rica (5)	0.936	2-8-1-14-11-12-9-6-13	Mexico (10)	0.515	14-11-12-6-8-13-9
4	Nv. Leon (16)	0.687	14-1-7-8-9-12-6-11-13	Colombia (4)	0.806	1-11-12-9-6-13	Nv. Leon (16)	0.485	14-11-12-6-8-13-9
5	Colombia (4)	0.546	9-12-6-11-13	Uruguay (15)	0.764	14-11-12-9-6-13	Colombia (4)	0.471	14-11-12-6-8-13-9
6	Uruguay (15)	0.495	9-12-6-11-13	Mexico (10)	0.742	1-14-11-12-9-6-13	Uruguay (15)	0.360	12-6-8-13-9
7	Peru (14)	0.322	9-12-6-11-13	Ecuador (7)	0.534	11-12-9-6-13	Brazil (2)	0.335	14-12-6-8-13-9
8	Argentina (1)	0.286	12-6-11-13	Brazil (2)	0.406	12-9-6-13	Argentina (1)	0.329	11-12-6-8-13-9
9	Brazil (2)	0.249	11-13	Guatemala (8)	0.392	11-12-9-6-13	Ecuador (7)	0.266	8-13-9
10	Ecuador (7)	0.188	11-13	Argentina (1)	0.318	12-9-6-13	Peru (14)	0.000	13-9
11	Guatemala (8)	0.011		Peru (14)	0.313	11-12-9-6-13	Nicaragua (11)	-0.087	1
12	Honduras (9)	-0.144		Nicaragua (11)	0.000	13	Panama (12)	-0.221	
13	Panama (12)	-0.144		Panama (12)	-0.109		Republican D. (6)	-0.221	
14	Republican D. (6)	-0.256		Honduras (9)	-0.132		Guatemala (8)	-0.352	
15	Nicaragua (11)	-0.299		Republican D. (6)	-0.277		Paraguay (13)	-0.472	
16	Paraguay (6)	-0.305		Paraguay (13)	-0.474		Honduras (9)	-0.498	
Score	**Correlation: .98**			**Correlation: .98**			**Correlation: .94**		
Book 4				**Book 5**			**Book 6**		
1	Chile (3)	0.729	1-14-2-12-11-8-9-13-6	Nv. Leon (16)	0.734	14-7-1-2-9-12-8-11-6-13	Costa Rica (5)	0.682	1-14-7-8-9-11-12-6-13
2	Nv. Leon (16)	0.649	1-14-2-12-11-8-9-13-6	Costa Rica (5)	0.728	14-7-1-2-9-12-8-11-6-13	Chile (3)	0.657	1-14-7-8-9-11-12-6-13
3	Mexico (10)	0.519	1-14-2-12-11-8-9-13-6	Chile (3)	0.651	14-7-1-2-9-12-8-11-6-13	Colombia (4)	0.533	8-9-11-12-6-13
4	Costa Rica (5)	0.459	11-8-9-13-6	Mexico (10)	0.587	14-7-1-2-9-12-8-11-6-13	Brazil (2)	0.526	7-8-9-11-12-6-13
5	Uruguay (15)	0.375	13-6	Colombia (4)	0.508	12-8-11-6-13	Nv. Leon (16)	0.510	14-7-8-9-11-12-6-13
6	Colombia (4)	0.290	13-6	Uruguay (15)	0.392	12-8-11-6-13	México (10)	0.411	8-9-11-12-6-13
7	Ecuador (7)	0.223	13-6	Peru (14)	0.284	12-8-11-6-13	Uruguay (15)	0.411	8-9-11-12-6-13
8	Argentina (1)	0.134	13-6	Ecuador (7)	0.181	11-6-13	Argentina (1)	0.224	9-12-6-13
9	Peru (14)	0.070	13-6	Argentina (1)	0.175	11-6-13	Peru (14)	0.159	12-6-13
10	Brazil (2)	0.002		Brazil (2)	0.153	11-6-13	Ecuador (7)	0.119	12-6-13
11	Panama (12)	-0.054		Honduras (9)	0.063	6-13	Guatemala (8)	-0.039	13
12	Nicaragua (11)	-0.102		Panama (12)	-0.125		Honduras (9)	-0.103	
13	Guatemala (8)	-0.107		Guatemala (8)	-0.136		Nicaragua (11)	-0.146	
14	Honduras (9)	-0.223		Nicaragua (11)	-0.330		Panama (12)	-0.236	
15	Paraguay (13)	-0.428		Republican D. (6)	-0.387		Republican D. (6)	-0.413	
16	Republican D. (6)	-0.459		Paraguay (13)	-0.415		Paraguay (13)	-0.451	
Score	**Correlation: .97**			**Correlation: .97**			**Correlation: .96**		

The comparison between groups is one of the objectives of this type of international evaluation. Knowing which items are entirely invariant across groups, therefore, is a useful tool. Figure 1 presents information about how the items are distributed in full invariant and non-invariant. The top of the figure includes a line detailing the total number of items in every booklet for a more accessible understanding of the results, pointing to the non-invariant items that represent between 40% - 64% of the science test. The full invariant items have the characteristic of being equivalent in the 16 participating nations of the TERCE study for intercepts and factor loading, so they are clearly useful for making comparisons between education systems across countries.

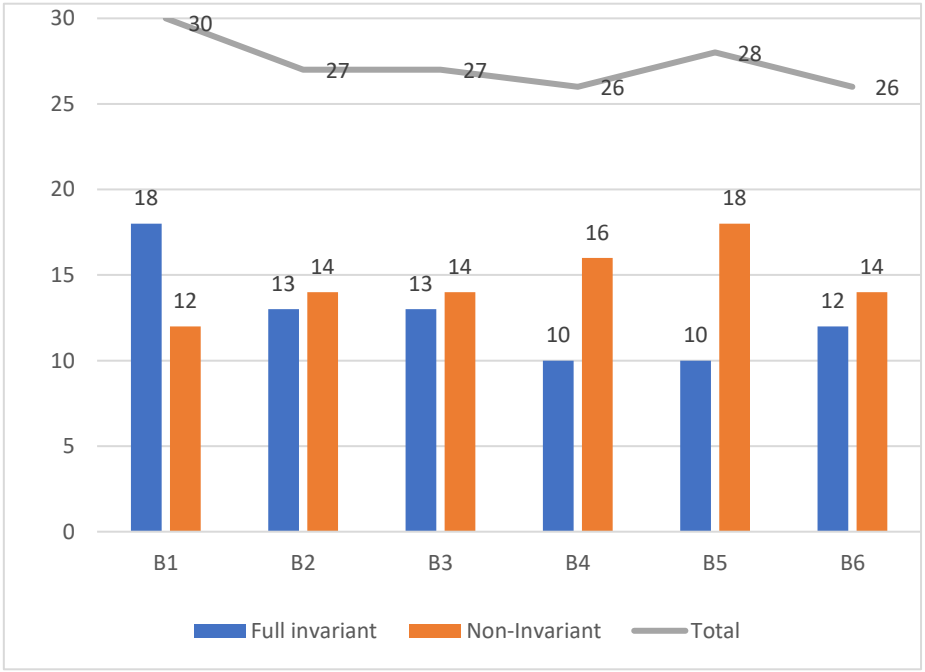


Figure 1. Distribution of the non-invariant and full invariant items across booklets. Full invariant: invariant parameter (factor loading & item intercept), non-invariant: non-invariant parameter (factor loading OR item intercept), total: total items per booklet, B1 to 6: where B indicates booklets and their number.

The techinal 8 output from the alignment results gives a complete information about the contribution made from each variable to the final simplicity function. Bear in mind the research goals this study will not include the fit function contribution results due to the extensive information within every booklet.

Considering the information of the total fit function contribution that agglomerates the fit function contribution of the loading and intercept factor (which gives information about the degree of contribution to the final simplicity function), lower values can be understood as an indication this item exhibited the least amount of non-invariance. The expected values for the items wholly invariant or non-invariant vary, following the same pattern as observed in Byrne and van de Vijver (2017), where the completely invariant items do not always exhibit the pattern of showing less contribution. Likewise, R^2 results show similarity to those reported by Flake and McCoach (2017), even when being completely invariant items have a R^2 value of 0.000. Having discussed these results in personal communication with the authors, they have detailed that this can occur due to different characteristics (Asparouhov, personal communication, February 6, 2018).[5]

In light of the information illustrated in the methodological section (Table 2), as well as taking into account the distributions of the items using the cognitive process in Table 5, the distributions were 49% of the non-invariant items applying the process of understanding and concepts applications. The lower levels of the distributions are recognition of objects and elements (30%) and scientific thinking and problem solving

[5] It is true that R2 can be close to zero even for invariant items even though that is somewhat unusual. The best way to understand this is to compute the R2 by hand, see formula (13) and (14) https://www.statmodel.com/download/webnotes/webnote18.pdf This can happen for example if the power was not sufficient to establish the non-invariance (such as small sample size or many missing values for that item or unusually large SE due to empty cells in bivariate tables). Or it can happened if the average aligned loading is close to 0. (Asparouhov, personal communication, February 6, 2018)

(21%). Additionally, in Figure 2 he observed and expected range can be appreciated, where Chi-Square test was applied to discover the pattern of distributions ($\chi_2^2 = 2.75$).

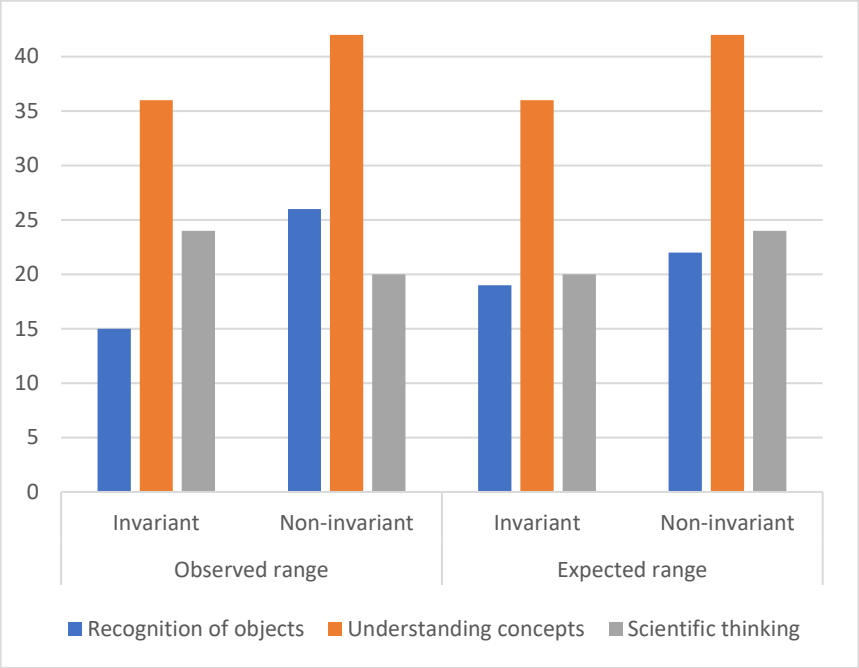


Figure 2. Chi-Square test was applied to discover the pattern of distributions between the cognitive process and the invariance pattern.

DISCUSSION AND CONCLUSION

The research aim was to effectively apply the alignment method to international educational assessments, which are characterized by large sample sizes that follow a hierarchical pattern and include several groups. Considering the first research question, the study began by investigating the possibility of using the alignment as a method for the detection of invariance in large-scale educational evaluation tests. The alignment menthod, as a way to compare factor means and variances, has revealed a model with the presence of a few non-invariant items and a higher number of invariant items. This demonstrates an improved process for measuring approximate invariance, which allows the use of characteristics from a complex survey data. Considering the versatility of the method, the computation time, and the information offered, the alignment method is a useful alternative when it is necessary to manage multiple groups. By handling a large number of items and considering the different types of indicators, the study has also discovered a helpful tool for modelling partial credit and binary items.

Even though the technique shows a positive performance in regard to the management of large datasets, there are still several questions related to the results values, especially in terms of the pattern of the fit function contribution by the indicator and R^2 values. The same situation was observed in the research material, indicating that one possible explanation of the differences can be associated with the sample size (Byrne & van de Vijver, 2017). The same pattern as Flake and McCoach's (2017) results were shown here, where R^2 does not correspond precisely with the expected value for invariant items, which is also probably related to the sizes of the groups.

Among the leading results it is significant to note that in at least 50% of the items some level of variability is observed. This result indicates that at least one group (i.e., country) is non-invariant in those parameters. The classification by countries indicates that this variability is concentrated in Chile, Brazil, Costa Rica, Ecuador, Guatemala, and Nuevo Leon. In terms

of the mean factors calculated by alignment optimization, a high level of correlation is observed between the average of the five plausible values that were used to create the ranking of TERCE results. The presence of 50% variability reveals that the ranking is built based on values that are obtained from non-invariant indicators, which means that comparison is not entirely accurate.

It is possible to take into account the characteristics of the items by condensing the information into three essential attributes of the large-scale educational evaluation tests: the type of cognitive processes used, the matrix design of the item, and the type of item codification applied. In broad terms, the results show the non-invariant items are focused on: first, understanding and application of concepts; second, recognizing objects; and, third, demonstrating scientific thinking. The non-invariant items are equally distributed in every booklet, avoiding the presence of position patterns. Finally, after performing a comparison between the type of codification used and based on the gradual increment of the partial credit items in educational assessments, the study revealed the pattern of an invariant and non-invariant item between every model. The principal results are that the non-invariance is presented in all types of codification (i.e., partial credit, fully correct, and fully incorrect).

The research on the measurement of invariance had three decades of practice without significant advances until the introduction of the alignment, which represents a new generation of methodology for scaling latent variable framework (Munck et al., 2017). The alignment method, as this study has shown, provides a substantial improvement in the detection of approximate invariance between groups, and allows the use of sophisticated data survey. Given its relative novelty, it is essential to remark on the alignment method's advantages and limitations classified here by the computational speed, the type of estimates, the number of indicators, groups and the sample size needed for the calculation, the types of parameters to be estimated, the invariance pattern, and the type of data allowed.

Based on the study, the average for the speed of calculation between all booklets was not more than 360 seconds. This speed implies a remarkably effective method, especially when taking into account the size of the sample and groups in comparison. Another advantage of the alignment method, especially for such large group comparisons, is the ability to start from the configurational model without requiring the scalar and metric invariance.

Regarding the number of indicators, this study was available to perform the analysis for all the indicators (i.e., binary, partial credit) at the same time, which provided a handy tool to achieve the matrix design of the items. In the same line Muthén & Asparouhov (2017) indicates a satisfactory performance of the technique from three indicators for each factor, regarding the number of groups. Based on the large size of the groups, the study focused only on 16 groups. Other researchers noted that there is a dependence between the size of the groups and how definites are the non-invariance pattern (although the alignment has shown satisfactory performance in situations where the invariance pattern is well known). The parameter biases can increase when there is an increase in the degree of non-invariance, a decrease in group sample size, and an increase in the numbers of groups (Flake & McCoach, 2017; Kim et al., 2017; Muthén & Asparouhov, 2014).

Another current limitation, especially for educational and psychological uses, is the numbers of factors allowed in every analysis. At the time of this research, it is possible to work with only one factor at a time and cases with multiple factors are aligned one factor at the time. Additionally, another current limitation is the use of covariates and a full structural equation model (Asparouhov & Muthén, 2014). Considering samples from educational assessments, where there is usually the presence of missing data, the analysis can be done only for each booklet separately (due to the fact of the matrix design implies the presence of missing data by design).

Among the leading results, it is highlighted that in at least 50% of the items some variability is observed. Due to the high stakes of the test uses for policy decisions (Ercikan et al., 2015), this percentage indicates that the ranking is based on values that are obtained from non-invariant indicators, which means their comparison would be not accurate when taking into account any characteristics that may be useful to psychometricians. As the non-invariant items mostly refer to cognitive processes of understanding and association of concepts, variability could be related to the differentiation of content and concept between each group. Finally, whichever form of coding is used with only an exception in one item, the results shows the presence of non-invariant parameters.

The interest in detecting measure variability to minimize the bias of the items comes from the high stakes of the tests results on educational programs and society in general (AERA et al., 2014; Hambleton et al., 2005; ITC, 2018, 2017). Although the interpretation and use of results have significant impacts on educational policy decisions at regional and international levels, it is important to keep in mind that "tests are imperfect measures of constructs because they either leave out something that should be included . . . alternatively else [they] include something that should be left out or both" (Messick, 1993 p. 34). The focuses, then, should not be on the assessments themselves or on the test applications, but it should be more about how the results are interpreted (Sireci, 2015). In other words, the major concern is with the implications of the test uses and their personal and social consequences (Zumbo & Hubley, 2016).

Appendix A

Mplus scripts

```
DATA: FILE = PC6_61921_BOOK5_28_W.dat;
FORMAT IS 1f2.0 28f1.0 1f4.0 1f5.2;
VARIABLE: NAMES =country1 IT5_1-IT5_28 ID_cluster swgc;
Categorical= IT5_1-IT5_28;
USEVARIABLES = IT5_1-IT5_28;
```

Data set input, information about the data format and names. In user variables we only include the items we are using for these analysis

```
STRATIFICATION =country1;
WEIGHT= swgc;
CLASSES= c1(16);
CLUSTER= ID_cluster;
```

The complex survey variables. The stratification are 16 nations, we use senatorial weights to and school's groups.

```
KNOWNCLASS = c1(country1= 1-16);
ANALYSIS: TYPE= COMPLEX MIXTURE;
ESTIMATOR = MLR;
PROCESSORS = 2;
ALIGNMENT = FREE;
ALGORITHM=INTEGRATION;
MODEL: %OVERALL%
f BY IT5_1-IT5_28;
OUTPUT: TECH1 TECH8 ALIGN;
PLOT: TYPE = PLOT2;
```

The type of analysis MIXTURE COMPLEX and KNOWCLASS are for real data. The estimations MLR is default based on the data complex survey.

PROCESSORS: option to speed up computations

Alignment: FREE was used for every book. And based in program warnings we FIXED in booklets 2 (g:11) and 3 (g:14).

AI: is used with ML estimation to indicate the optimization method to use

MODEL: 1 factor by all items per booklet

PLOT2: Sample proportions and estimated probabilities, Item characteristic curves, Information curves and Measurement parameter plots

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standars for Educational and Psychological Testing*. Washington: AERA.

Asil, M., & Brown, G. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing, 16*(1), 71-93. doi:10.1080/15305058.2015.1064431

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495-508. doi:10.1080/10705511.2014.919210

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological bulletin, 105*(3), 456-466. doi:10.1037/0033-2909.105.3.456

Byrne, B., & van de Vijver, F. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29*(4), 539-551. doi:10.7334/psicothema2017.178

Desa, D. (2014). *Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses*. doi: 10.1787/19939019

Elosua, P., & Mujika-Lizaso, J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension test in Spain. *Psicothema, 25*(3), 390-395. doi:10.7334/psicothema2013.46

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.

Fernández-Alonso, R., & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de

las Competencias Básicas [Design notebooks for the evaluation of basic skills]. *Aula Abierta, 39*(2), 3-34.

Flake, J., & McCoach, B. (2017). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* 1-15. doi:10.1080/10705511.2017.1374187

Hambleton, P., Merenda, F., & Spielberger, C. (2005). *Adapting educational and psychological test for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

International Test Commission [ITC]. (2018). *ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations*. International Test Comission.

International Test Commission [ITC]. (2017). ITC guidelines for translating and adapting tests (Second Edition), *International Journal of Testing, 2*(18), 101-134. doi: 10.1080/15305058.2017.1398166

Johnson, E. G., Nancy, L. A., Bourque, M. L., Bowker, D. W., Caldwell, N. W., Donoghue, J. R., . . . Kline, D. L. (1994). *The NAEP 1992 technical report*. Retrieved from http://files.eric.ed.gov/fulltext/ED376191.pdf

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426. doi:10.1007/bf02291366

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544. doi:10.1080/10705511.2017.1304822

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [LLECE]. (2014). *Primera entrega de resultados Tercer Estudio Regional Comparativo y Explicativo Terce* [First delivery of results Third Regional Comparative and Explanatory Study Terce]. Santiago de Chile: UNESCO.

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, data, analyses, 12*(1). doi:10.12758/mda.2017.09

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2017). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*. doi:10.1037/met0000113-10.1037/met0000113.supp

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015.* Retrieved from http://timss.bc.edu/publications/timss/2015-methods.html#

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi:10.1007/bf02294825

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). Phoenix, AZ: Oryx Press.

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Retrieved from http://timssandpirls.bc.edu/pirls2016/international-results/

Munck, I., Barber, C., & Torney-Purta, J. (2017). Measurement Invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The Alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*. doi:10.1177/0049124117729691

Muthén, B. (2013). Late-Breaking News: Some Exciting New Methods (AVI video). Available from https://www.statmodel.com/UConnM3.shtml

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*(978). doi:10.3389/fpsyg.2014.00978

Muthén, B., & Asparouhov, T. (2017). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research.* 1-28. doi:10.1177/0049124117701488

Muthén, L.K., & Muthén, B.O. (2017). *Mplus user´s guide* (8th ed.). Los Angeles, CA:Muthén & Muthén.

Organisation for Economic Co-operation and Development [OECD]. (2014). *PISA 2012 technical report*. Retrieved from https://goo.gl/V2z9oT

Organisation for Economic Co-operation and Development [OECD]. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. Paris: OECD Publishing.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57. doi:10.1177/0013164413498257

Segeritz, M., & Anand-Pant, H. (2013). Do they feel the same way about math?: Testing measurement invariance of the PISA "students' approaches to learning" instrument across immigrant groups within Germany. *Educational and Psychological Measurement*, *73*(4), 601-630. doi:10.1177/0013164413481802

Sireci, S. (2015). Beyond ranking of nations: Innovative research on PISA. *Teachers College Record*, *117*(1), 1-8.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical & Statistical Psychology*, *27*(2), 229-239. doi:10.1111/j.2044-8317.1974.tb00543.x

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & La Oficina Regional de Educación para América Latina y el Caribe

[UNESCO-OREALC]. (2016). *Reporte técnico del Tercer Estudio Regional Comparativo y Explicativo* [Technical report of the Third Regional Comparative and Explanatory Study]. Santiago de Chile: UNESCO.

van de Schoot, R., Schmidt, P., de Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6*(1064). doi:10.3389/fpsyg.2015.01064

Wu, A., Li, Z., & Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multigroup confirmatory factor analysis: A demostration with TIMSS data. *Practical Assesment, Research & Evaluation, 12*(3), 1-26.

Zumbo, B.D. (2013). On matters of invariance in latent variable models: Reflections on the concept, and its relations in classical and item response theory. In Paolo Guiudici, Salvatore Ingrassia, and Maurizio Vichi (Eds.), *Statistical Models for Data Analysis* (pp.399-408). New York: Springer.

Zumbo, B.D., & Hubley, A. M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice, 23*(2), 299-303. doi:10.1080/0969594X.2016.1141169

CAPÍTULO II. Influencia de los Centros Escolares sobre el Rendimiento Académico en Latinoamérica

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

RESUMEN

La estimación del efecto de los centros es la piedra angular en el estudio de la eficacia escolar. En Latinoamérica, las conclusiones sobre el efecto de los centros no siempre son consistentes. Esta investigación analiza la influencia de las características de los centros sobre el rendimiento académico en Latinoamérica. La muestra es de 63.750 estudiantes, de 2955 escuelas de quince países latinoamericanos. La edad media del alumnado es de 12,4 años (DT = 0,94). El 69,4% asiste a un centro público, el 50,3% son mujeres y el 20,4% ha repetido al menos un curso. Se emplea el análisis jerárquico-lineal de dos y tres niveles, las variables de ajuste utilizadas son los antecedentes socioeconómicos del alumno y variables escolares como la dependencia, ruralidad, infraestructura y el nivel socioeconómico de los centros. Los resultados muestran que el efecto bruto de los centros es del 33% para Ciencias Naturales y 28% para Matemáticas y Lectura. El impacto de las variables socioeconómicas reduce esta variabilidad entre en 35% y el 68%. El efecto neto de los centros fue de 25% en Ciencias Naturales, 23% en Matemática y 13% en Lectura. Se discuten estos resultados y se analizan posibles implicaciones para las políticas educativas en Latinoamericana.

Palabras clave: Latinoamérica, Equidad, educación primaria, rendimiento académico, modelos jerárquicos lineales.

INTRODUCCIÓN

La iniciativa *Educación Para Todos* de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, 2000) persigue entre sus objetivos mitigar la desigualdad en materia de educación y suprimir las discriminaciones en las posibilidades de aprendizaje de grupos desfavorecidos. La investigación sobre la eficacia escolar aborda esta problemática, contribuyendo con un amplio volumen de hallazgos (Scheerens, 2016; Scheerens, Witziers, y Steen, 2013; Teddlie y Reynolds, 2000; Towsend, 2007). De acuerdo con Teddlie y Reynolds (2000) los estudios sobre eficacia escolar tienen dos grandes propósitos, por un lado estimar los efectos de la escuela, ya sean brutos o bajo alguna definición de eficacia relativa, es decir, una vez eliminado el influjo de las variables de contexto o antecedentes al proceso educativo, y por otro, analizar los factores asociados, es decir, los procesos y características que permiten que las escuelas promuevan el aprendizaje del alumnado más allá de lo que sería esperable en función de las variables de contexto. El presente trabajo se encuadra dentro de la primera finalidad, y pretende conocer los efectos escolares entendido como el aporte de los centros educativos en el rendimiento de sus alumnos, estimando las diferencias brutas entre los centros y el efecto de éstos una vez descontadas las variables que describen el contexto escolar en América Latina.

El estudio del efecto bruto del centro educativo, que operativamente se define como el porcentaje de las diferencias imputadas al factor en un ANOVA de efectos aleatorios o como la correlación intraclase de un modelo multinivel sin predictores, es la línea primigenia y más antigua de la eficacia escolar, y cuenta con una tradición de cinco décadas (Murillo, 2005). En América Latina la investigación sobre los efectos escolares se inició hace dos décadas, por lo que en la actualidad ya se dispone de un amplio conjunto de trabajos (Blanco-Bosco, 2009; Murillo, 2003, 2007, 2008). Sin duda, los estudios latinoamericanos con mayor alcance son los realizados por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación

(LLECE), que, hasta la fecha, ha realizado tres ediciones del *Estudio Regional Comparativo y Explicativo*: la primera (denominada *PERCE*) en 1997, la segunda (*SERCE*) en 2006 y la tercera (*TERCE*) en 2013. Se trata de una evaluación internacional, donde han participado una veintena de países de la región y cuya finalidad es analizar los logros del aprendizaje del alumnado y los factores asociados a este logro (UNESCO-OREALC y LLECE, 2000, 2010, 2016b).

Sin embargo, pese al corpus de investigación acumulado (o precisamente por ello) las conclusiones sobre el efecto bruto del centro no siempre son consistentes, llegando incluso a parecer contradictorias, debido al escaso consenso en la forma de medición de la misma, antes que en su conceptualización teórica (Murillo, 2005). Por ejemplo, los análisis secundarios realizados con datos de las comparaciones internacionales de la región generalmente informan de efectos brutos similares a los publicados en el estudio original (Miralles, Castejón, Pérez, y Gilar, 2012). Sin embargo, no faltan ocasiones donde los autores comunican efectos superiores (Castro-Aristizabal, Castillo-Caicedo, y Mendoza-Parra, 2016; Doneschi, 2017) o inferiores (Cervini, 2012; Murillo y Román, 2011) a los publicados originalmente. La diferencia en los datos probablemente radica en el hecho de que los estudios difieren en aspectos básicos, como la materia evaluada, la etapa educativa y la edad del alumnado, las características de los centros (titularidad y zona geográfica), los diseños muestrales empleados para seleccionar al alumnado participante y el número de niveles de agregación empleados en el análisis multinivel.

Existen bastantes datos que indican que el efecto bruto del centro es mayor en las materias científico-matemáticas que en las sociolingüísticas (Blanco-Bosco, 2011; Casas, Gamboa, y Piñeros, 2002; Cervini, 2002a, 2002b, 2012; Scheerens, 2016; UNESCO-OREALC y LLECE, 2001, 2010; Zorrilla, 2009). Sin embargo, Carvallo (2006) y Cervini, Dari y Quiroz (2016) señalan lo contrario, y otros autores informan de diferencias muy pequeñas (Murillo y Román, 2011), cuando no efectos similares en ambas materias (Cervini, 2006). Incluso dentro

de una misma materia se estiman efectos diferentes en función de sus ejes de especificación. Peña-Suárez, Campillo-Álvarez, Santarén-Rosell y Muñiz (2012), con datos de PISA 2006 y empleando como variable dependiente los resultados en las competencias científicas, encuentran que el efecto bruto del centro es mayor cuando la variable a predecir son las capacidades científicas (identificar cuestiones científicas; explicar fenómenos científicamente; y usar pruebas científicas) que cuando son las actitudes hacia la ciencia (interés por la ciencia; y apoyo a la investigación científica.

Igualmente, los datos disponibles no presentan una pauta clara que relacione el tamaño del efecto bruto del centro y la edad o etapa educativa del estudiante. Una proporción de la investigación señala que a medida que aumenta la edad disminuye ligeramente el tamaño del efecto (Scheerens et al., 2013). Sin embargo, Cervini (2010) encuentra mayor efecto en educación secundaria que en primaria, y no faltan estudios que informan de que el efecto es similar en ambas etapas (Cervini, 2009) y que, por tanto, no encuentran relación entre la edad del alumnado y el tamaño del efecto bruto del centro (Murillo y Román, 2011). Las variaciones a la hora de estimar el efecto bruto del centro también se observan en función de algunas características de la escuela como la titularidad o la zona geográfica. Piñeros y Rodríguez Pinzón (1998) encuentran que el coeficiente de correlación intraclase es mayor en los centros privados que en los públicos, aunque la revisión de Casas et al. (2002) no permite confirmar la afirmación anterior. Por su parte, Reimers (2002) señala que el efecto bruto en las escuelas rurales e indígenas es entre dos y tres veces mayor que en las escuelas urbanas.

Por último, el tamaño del efecto bruto del centro también parece depender de características metodológicas del estudio, tales como el diseño muestral empleado, el modo en que se define o instrumentaliza la variable dependiente y el número de niveles incluidos en el análisis jerárquico-lineal. Carvallo (2006), trabajando con edades similares, señala un efecto menor en estudios censales que en estudios muestrales, y Cervini (2004a, 2004b), después de

excluir los centros más pequeños de su estudio, encuentra una disminución del efecto bruto, probablemente por eliminar parte de las variaciones vinculadas a las escuelas rurales y remotas. Por su parte, Casas et al. (2002), en un estudio longitudinal con datos colombianos, encuentran una disminución del tamaño del efecto asociada a un cambio en el tipo de examen, lo que indica que una modificación en la especificación de la variable dependiente puede afectar al poder discriminatorio de las pruebas a nivel de centro. Finalmente, la evidencia sobre las variaciones del efecto bruto del centro en función del número de niveles de los modelos jerárquico-lineales parece bien documentada. Cervini (2006, 2010, 2012), comparando el tamaño del efecto en modelos de entre dos y cuatro niveles, encuentra que a medida que aumenta el número de niveles disminuye el tamaño del efecto del centro como consecuencia lógica de la distribución de la varianza entre-centros en niveles de agregación superiores (país, provincia) o inferiores (aula), y Scheerens (2016), en una revisión que compara estudios que incluyen datos del nivel estudiante, profesor/aula y escuela, señala que la omisión de niveles intermedios (profesor/aula) conduce a una sobreestimación del efecto del siguiente nivel superior, como podría ser el caso de las escuelas.

En general se ha asumido que los sistemas educativos más equitativos tienden a mostrar efectos brutos entre centros más pequeños (OCDE, 2002). Sin embargo, esto no es del todo cierto, un sistema educativo podría presentar diferencias absolutas pequeñas entre centros, pero si todas las variaciones estuvieran determinadas por el contexto sociológico y demográfico la capacidad de mejora de los centros sería nula o muy limitada. El caso opuesto sería aquel sistema educativo que mostrara grandes diferencias entre sus centros, pero donde la influencia de los antecedentes sociológicos sobre los resultados fuera muy débil, ya que en este caso las prácticas educativas de los centros tendrían mayor capacidad compensatoria. Por ello, otra forma tradicional de estimar la inequidad de los sistemas educativos es calcular el grado de asociación entre los desempeños escolares y las características del contexto educativo,

entendiendo que la relación entre los antecedentes socioeconómicos y culturales del alumnado y su rendimiento académico es una medida de la equidad con la que un sistema educativo distribuye las oportunidades de aprendizaje y del potencial compensatorio de cada escuela (OECD, 2010). Para su estimación se especifica un modelo, generalmente jerárquico-lineal, donde las variables de contexto se incluyen en diferentes niveles, y que permite estimar los efectos escolares netos, que operativamente se definen como la proporción de varianza que queda por explicar una vez controlados los factores antecedentes (Rodríguez-Jiménez y Murillo, 2011).

Existen abundantes datos que apoyan la relevancia de los factores sociodemográficos para explicar las diferencias en los resultados de los centros (OECD, 2002, 2004, 2007, 2010, 2013, 2016; Sirin, 2005). Sin embargo, al igual que ocurriera al hablar del efecto bruto, la estimación de la potencia explicativa de las variables de contexto varía de unos estudios a otros. Por ejemplo, SERCE, una vez detraídas las variables de ajuste, estimó los siguientes efectos netos de centro: 44% en Ciencias Naturales; 37% en Matemáticas; y 31% en Lectura (UNESCO-OREALC y LLECE, 2010). Sin embargo, Murillo y Román (2011) consideran muy elevados estos valores y reanalizan los datos encontrando que el efecto neto del centro oscila entre el 17% (Lectura) y el 22% (Matemática). Por su parte, Cervini (2016), con la base de datos TERCE y un modelo jerárquico-lineal de tres niveles, aún rebaja más este efecto, situándolo entre el 11% en Matemáticas y el 7% en Lectura. Las razones de esta disparidad en las estimaciones parecen ser tres: la variable dependiente empleada, el número y la naturaleza de las variables independientes y el nivel o niveles de análisis en que son consideradas estas últimas variables.

En general los datos disponibles indican que los factores demográficos, económicos y culturales tienen mayor impacto cuando la variable criterio es la lengua vernácula que cuando se trata de una materia científico-matemática (Cervini, 2002b, 2009, 2012; Cervini et al., 2016).

Sin embargo, este resultado no siempre se replica en estudios latinoamericanos ya que Cervini (2006, 2010) y Murillo y Román (2011) encuentran efectos similares en Lenguaje y Matemáticas, y Zorrilla (2009) encuentra que el efecto es mayor en Matemáticas. En todo caso, parece que cuando la variable dependiente es una lengua extranjera (por ejemplo, inglés en los países hispanohablantes) el efecto del contexto social y demográfico sobre los resultados es mayor que en cualquier otra materia (Consejería de Educación y Ciencia del Principado de Asturias, 2007).

Una segunda fuente variación en la proporción de varianza explicada por los modelos de eficacia relativa es el número y la naturaleza de las variables antecedentes. En general, la variable con mayor fuerza suele ser un índice construido, generalmente tipificado, que resume el nivel socioeconómico y cultural de la familia del alumnado o, en su defecto, variables que describen el nivel de estudios y profesiones de los progenitores (Peña Suárez, Fernández-Alonso, y Muñiz, 2009; Sirin, 2005). Doneschi (2017) en un estudio de dos niveles que compara la inclusión progresiva de variables de contexto en cada nivel, señala que el índice socioeconómico es la variable con mayor capacidad para dar cuenta de las diferencias en los resultados, mientras que otros estudios informan que el nivel socioeconómico explica aproximadamente la mitad de la varianza entre centros (Casas et al., 2002; Rodríguez-Jiménez y Murillo, 2011). También han mostrado su relevancia algunas variables dicotómicas, como el género, la escolarización temprana, la lengua materna o la condición de emigrante, y otras medidas en escala ordinal o continua, como el número de libros en el hogar, las posesiones materiales y las características de la vivienda (Murillo y Román, 2011; UNESCO-OREALC y LLECE, 2000). Igualmente, en los estudios Latinoamericanos parecen importantes dos variables más, que generalmente no se consideran en la investigación con países desarrollados: ser indígena y compatibilizar trabajo y estudios (UNESCO-OREALC y LLECE, 2016a). Mención especial merece el tratamiento de la repetición escolar, ya que es una variable con

importantes efectos sobre el rendimiento académico incluso una vez descontadas el resto de las variables de contexto ya mencionadas (Gobierno del Principado de Asturias, 2016). Evidentemente en los diseños de investigación descriptivos *expost facto* la repetición es una variable antecedente, sin embargo, también es una variable modificable por los centros, ya que es una decisión que mayoritariamente se toma en los claustros escolares. Por tanto, considerarla o no como variable de contexto afectará a la proporción de varianza explicada por los modelos de ajuste.

Finalmente, el potencial de los factores de contexto para aquilatar las diferencias entre centros depende del nivel de análisis en el que se incluyen dichos factores. Cuando las variables se introducen únicamente a nivel de alumnado tienen menos incidencia, que cuando se manejan como promedios de aula o centro (Cervini, 2002a, 2002b, 2004b, 2006, 2012; Cervini et al., 2016). Esta limitación de los antecedentes sociológicos para explicar la varianza intra-centros está relacionada con el hecho de que la historia escolar, el rendimiento previo y algunas variables psicológicas, como la capacidad, la motivación, las expectativas o los estilos de aprendizaje, tienen más efecto sobre los resultados individuales que las variables de contexto económico y cultural (Fernández Alonso, Suárez-Álvarez, y Muñiz, 2015, 2016; Rodríguez, Piñeiro, Regueiro, Estevez, y Val, 2017; Suárez-Álvarez, Fernández-Alonso, y Muñiz, 2014). No obstante, los modelos que dan cuenta de un mayor porcentaje de varianza son aquellos que consideran conjuntamente las variables de ajuste en dos o más niveles de análisis, lo que muestra la superioridad de los modelos jerárquico-lineales para explicar los resultados educativos (Fernández-Alonso, Álvarez-Díaz, Suarez-Alvarez, y Muñiz, 2017). En general la reducción de la varianza es mayor en el nivel de centro (o aula) que en el individual (Ferrão y Fernandes, 2001; Peña-Suárez et al., 2012), si bien aquí también se advierten notables diferencias en la potencia de los modelos, en algunos casos la varianza explicada está por debajo del 50% (Cervini, 2002a, 2002b, 2012; Murillo y Román, 2011), en otros se sitúa en

torno al 60% (Cervini, 2010; Rodríguez-Jiménez y Murillo, 2011), y en las soluciones más potentes los factores antecedentes dan cuenta del 80% (Cervini et al., 2016) y hasta del 90% de las diferencias entre los centros (Doneschi, 2017).

En este contexto el objetivo general del presente estudio es analizar los efectos de los centros educativos sobre el rendimiento académico en Latinoamérica, empleando para ello los datos obtenidos en la evaluación TERCE. Este objetivo general se concreta en cuatro objetivos específicos: a) ¿Cuál es el efecto global de los centros escolares en el rendimiento académico de los estudiantes de Latinoamérica?, b) ¿Qué porcentaje de la varianza total y de la varianza entre los centros está asociada al nivel socioeconómico y cultural de los estudiantes? c) ¿Qué incremento experimenta la proporción de la varianza explicada cuando se incluyen otras variables de ajuste?, y d) ¿Cuál es el efecto neto de los centros escolares latinoamericanos sobre el rendimiento académico cuando se controlan las variables de ajuste?

MÉTODO

Participantes

La población objetivo se definió como el alumnado matriculado en 6º de educación primaria en el curso 2013 en los 15 países participantes y el estado de Nuevo León (México). Dentro de cada país la muestra fue seleccionada siguiendo un diseño bietápico y estratificado propio de este tipo de comparaciones internacionales (Joncas y Foy, 2012). En la primera etapa los centros participantes se seleccionaron con una probabilidad proporcional a su tamaño, y en la segunda etapa se eligió un grupo aula del centro mediante un muestreo aleatorio simple. En el presente estudio se excluyó al alumnado sin información en las pruebas de Ciencias Naturales, Matemáticas y Lectura por lo que la base final quedó compuesta por 61938 estudiantes participantes en la prueba de Ciencias Naturales, 63750 en la prueba de Matemáticas y 60949 participantes en la prueba de Lectura que asisten a 2955, 2934 y 2954 escuelas respectivamente,

los cuales representan a una población de prácticamente 9 millones de estudiantes. La tabla 1 recoge el número de estudiantes y centros considerados en cada materia. La media de edad del alumnado es de 12,4 años (DT = 0,94). El 69,4% asiste a un centro público; el 50,3% son mujeres; y el 20,4% ha repetido al menos un curso en el momento de la aplicación de la prueba.

Tabla 1

Distribución de la muestra por materia evaluada

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

Procedimiento

El estudio es implementado por el LLECE en cooperación con las coordinaciones de los sistemas educativos de los países participantes, quienes otorgan los permisos de acceso al personal experto y externo al centro encargado de la aplicación de las pruebas, conservando la confidencialidad de los participantes (UNESCO-OREALC, 2016b). La aplicación se desarrolló en dos jornadas, la primera dedicada a Lectura y Escritura y la segunda a Matemáticas y Ciencias. La evaluación de cada materia ocupó entre 45 y 60 minutos, con un descanso de 30 minutos. El cuestionario de contexto del alumnado, de 45 minutos de duración, se aplicó al final de la segunda jornada, tras un receso de 15 minutos. El primer día se entregaron los cuestionarios para el centro, profesorado y familias, y se recogieron al final de la segunda jornada.

Instrumentos

Se emplearon pruebas cognitivas de conocimientos en las tres materias mencionadas, y las puntuaciones en dichas pruebas se emplean como variables criterio en este estudio. Por su parte, las variables de ajuste se extraen de las respuestas a los cuestionarios de contexto ya reseñados. Lo instrumentos, creados por expertos en varios campos de estudio contratados

LLECE y validados en un estudio piloto, cumplen con los criterios psicométricos comunes a todo procedimiento de evaluación (Martínez-Arias, 2010).

Test de rendimiento académico

La prueba de Ciencias constaba de 92 ítems de elección múltiple y abiertos de respuesta corta agrupados en seis bloques diseñados para cubrir una tabla de especificaciones bivariado que evaluaba tres procesos cognitivos (*Reconocimiento de información y conceptos; Comprensión y aplicación de conceptos; y Pensamiento científico y resolución de problemas*), y cinco dominios de conocimientos (*Salud; Seres vivos; Ambiente; La tierra y el sistema solar; y Materia y energía*). La prueba de Matemáticas contenía 98 ítems de los dos formatos mencionados y distribuidos en seis bloques de 16 ó 17 ítems cada uno. Las especificaciones de contenido se organizaron en una matriz de doble entrada: tres procesos cognitivos (*Reconocimiento de objetos y elementos; Solución de problemas simples; y Solución de problemas complejos*), y cinco dominios de conocimientos (*Numérico; Geométrico; Medición; Estadístico; y Variación*) (UNESCO-OREALC, 2016b). Finalmente, la prueba de Lectura constaba de 96 ítems de elección múltiple y respuesta construida asignados a seis clústers de 16 ítems cada uno, y distribuidos sobre una matriz de especificaciones que evaluaba tres procesos cognitivos *(Recuperación literal de información; Realización de inferencias; y Lectura crítica)* y dos dominios de conocimientos *(Comprensión de textos; y Dominio metalingüístico y teórico)* (UNESCO-OREALC, 2016b).

Dado que los estudiantes no pueden responder al conjunto de ítems en el tiempo de evaluación disponible, los bloques de ítems se distribuyeron en seis modelos de cuadernillos por cada materia siguiendo un diseño matricial de bloques incompletos parcialmente balanceado (Fernández-Alonso y Muñiz, 2011) y cada estudiante respondió a un cuadernillo que contenía entre 31 y 33 ítems. El promedio del coeficiente alfa de Cronbach de los

cuadernillos fue de ,72 en Ciencias Naturales; ,80 en Matemáticas; y ,85 en Lectura. La puntuación fue calculada mediante la metodología de valores plausibles, por ser la más eficiente para recuperar los parámetros poblaciones en las evaluaciones de sistemas educativos (Mislevy, Beaton, Kaplan, y Sheehan, 1992; OECD, 2009). Las puntuaciones individuales fueron expresadas en una escala con media 700 puntos y desviación típica 100 (UNESCO-OREALC, 2016b).

Variables de control o ajuste

Para describir las características sociológicas del alumnado se han considerado seis variables, cinco de ellas dicotómicas: *Género* (1 = ser mujer); *Indígena* (1 = pertenecer a una etnia indígena); *Repetidor* (1 = haber repetido algún curso durante la escolaridad); *Trabajar* (1= el estudiante trabaja y recibe una remuneración por esa actividad); y *Cuaderno* (1= el estudiante posee cuaderno para anotaciones). La sexta variable es una estimación del *Nivel socioeconómico y cultural del alumnado* (ISEC), un índice estandarizado por TERCE y compuesto por 17 ítems que recababan información sobre el nivel educativo de los progenitores, tipo de trabajo que realizan, ingresos familiares, disponibilidad de material de lectura del hogar y bienes y servicios del barrio donde se ubica la vivienda. Los valores del alfa de Cronbach del índice oscilan entre ,8 y ,9 según el país (UNESCO-OREALC, 2016b).

Dentro de las características del contexto social y demográfico de la escuela se consideraron cinco variables, dos de ellas dicotómicas: *Titularidad* (0 = centro público; 1 = centro privado) y *Ruralidad* (1 = centro rural). Los recursos de la escuela se estimaron mediante el *Nivel de Infraestructura del centro,* un índice estandarizado elaborado con 19 ítems del cuestionario para las direcciones que evaluaba las instalaciones, equipamientos y servicios de escuela. Los valores del alfa de Cronbach del índice oscilan entre ,7 y ,9 según el país

(UNESCO-OREALC, 2016b). Las dos últimas variables fueron el *Nivel socioeconómico y cultural del centro y del país*, calculados como el promedio del ISEC por centro y por país.

Análisis de datos

Se emplearon modelos jerárquico-lineales, que previenen contra errores de inferencia estadística como el sesgo de segregación o la heterogeneidad de la regresión, y permiten separar los efectos de la escuela de los factores propios de los estudiantes (Martínez-Arias, Gaviria-Soto, y Castro-Morera, 2009). Inicialmente, para cada competencia y país (incluida la muestra completa) se ajustó una secuencia de tres modelos de 2 niveles (alumnado y escuela): Modelo I, sin variables predictoras; Modelo II, que incluye el ISEC en ambos niveles (estudiante y centro); y Modelo III, que añade al modelo anterior el resto de las variables de contexto (*Nivel estudiante:* género, indígena, repetidor, trabajar, cuaderno; *Nivel centro:* titularidad, ruralidad, nivel de infraestructura). En total se especificaron 153 ajustes (3 modelos x 3 competencias x 17 muestras). Con el Modelo I se estimó el coeficiente de correlación intraclase (ICC) para la determinación del efecto bruto de los centros, y con los Modelos II y III se calculó la proporción de varianza explicada, entendida como la reducción de varianza (total y por nivel) que experimenta el Modelo I al incluir las variables de control (Efecto ISEC) y el efecto neto del centro, entendido como el porcentaje de varianza contenida en el nivel de centros de los Modelos II y III.

Adicionalmente, empleando la muestra completa, los modelos I, II y III se especificaron para un análisis jerárquico-lineal de 3 niveles (alumnado, escuela y país) lo que dio lugar a 9 ajustes más (3 modelos x 3 competencias). En las tablas de resultados estos modelos se identificarán como TERCE 3-N, y permiten estimar los efectos ya mencionados, pero segregando la varianza al nivel de país. El análisis se realizó con HLM 7.01 (Raudenbush, Bryk, Cheong, Congdon, y du Toit, 2011), empleando el método de estimación de Máxima

Verosimilitud y tomando como variable dependiente los cinco valores plausibles de cada competencia. En los modelos ajustados con la base completa los casos fueron ponderados empleando los pesos senatoriales ofrecidos por TERCE, de modo que, en cada país, la suma de los pesos equivalía a 5000 estudiantes escolarizados en 200 escuelas. En los modelos ajustados para cada país individualmente considerado, se emplearon los pesos muestrales de modo que la suma de estos pesos equivalía a la matrícula total de estudiantes de 6º de educación primaria y al número total de centros de educación primaria de cada país (UNESCO-OREALC, 2016b).

El rango de casos perdidos en las variables osciló entre el 2% y el 12%, y para su recuperación se empleó una estrategia en dos pasos. Inicialmente los casos incompletos fueron imputados por la media del sujeto y, a continuación, los datos totalmente perdidos fueron recuperados mediante el método iterativo (método EM con variables auxiliares) que implementa el módulo Missing Value Análisis del programa SPSS 15. Fernández-Alonso, Suárez-Álvarez y Muñiz (2012) encontraron que esta estrategia de dos pasos es la que mejor recupera los datos poblacionales cuando el mecanismo de la pérdida no es aleatorio y el porcentaje de datos perdidos es similar al registrado en TERCE.

RESULTADOS

La Tabla 2 muestra la distribución de la varianza en los niveles alumnado y centro para cada país y materia evaluada, así como el porcentaje de la varianza total contenida en el nivel de centro, expresada como el coeficiente de correlación intraclase (ICC). El porcentaje de varianza total atribuido al nivel de país, que no está recogida en el modelo TERCE 3-N, fue del 13% Ciencias Naturales, 26% en Matemáticas y 17% en Lectura. En todo caso el modelo TERCE 3-N muestra una importante reducción del efecto bruto del centro cuando se compara con el modelo de dos niveles (TERCE 2-N), si el análisis no hubiese considerado el nivel de país el ICC sería entre 10 y 23 puntos porcentuales mayor. Para el conjunto de la muestra Ciencias es

la materia con mayor ICC (33%), mientras que Lectura y Matemáticas presentan 5 puntos porcentuales menos.

Tabla 2

Modelo I Efecto bruto. Distribución de la varianza por nivel de análisis y efecto bruto del centro expresado como porcentaje de varianza entre centros

	Ciencias Naturales			Matemática			Lectura		
	Varianza Nivel 1	Varianza Nivel 2	Efecto bruto	Varianza Nivel 1	Varianza Nivel 2	Efecto bruto	Varianza Nivel 1	Varianza Nivel 2	Efecto bruto
TERCE (2-N)	5943	4443	43%	5119	5428	51%	5787	4782	45%
TERCE (3-N)	5903	3560	33%	4926	3005	28%	5822	3019	28%
Argentina	6256	3421	35%	4847	2793	37%	6541	3931	38%
Brasil	5991	3323	36%	4062	3480	46%	6409	2980	32%
Chile	9569	3412	26%	8226	2399	23%	8087	2060	20%
Colombia	6149	2767	31%	4201	2744	40%	5401	3807	41%
Costa Rica	6102	1852	23%	4561	1456	24%	5255	1586	23%
Ecuador	5878	3778	39%	4164	3140	43%	4954	3268	40%
Guatemala	4270	2909	41%	3553	2764	44%	4279	2643	38%
Honduras	4589	3131	41%	3326	2815	46%	4018	2343	37%
México	5638	2445	30%	7715	2810	27%	6365	3358	35%
Nicaragua	3674	2546	41%	2361	1630	41%	4315	2350	35%
Nuevo León	6065	2227	27%	8979	2109	19%	7031	2227	24%
Panamá	5473	3558	39%	3166	2220	41%	5481	3340	38%
Paraguay	4927	4143	46%	3601	4055	53%	5116	3420	40%
Perú	4696	4195	47%	5302	6943	57%	4678	6601	59%
Rep. Dominicana	4368	1171	21%	2670	612	19%	4357	1439	25%
Uruguay	8827	2927	25%	8301	2074	20%	8213	2811	25%

Los datos del Modelo I señalan que en todos los países el efecto bruto de la escuela es significativo, aunque no es menos cierto que se advierten importantes variaciones en la magnitud del ICC de cada país. Perú presenta el mayor porcentaje de variación entre centros, llegando a ser incluso superior a la varianza dentro de los centros en Matemáticas (57%) y Lectura (59%), seguido de Paraguay donde el coeficiente de correlación intraclase supera el

45% en dos de las tres áreas. En Argentina, Ecuador, Guatemala, Honduras, Nicaragua y Panamá los ICC de todas las materias están en rango ,35 -, 45. En el extremo contrario, es decir, los países con un tamaño del efecto del centro menor son Chile, Costa Rica, República Dominicana, Uruguay y el estado de Nuevo León, donde el ICC no alcanza el 30% en ninguna materia.

Por otro lado, las estimaciones del efecto bruto de los centros son bastante estables. La correlación de los ICC de los países es de ,85 entre Matemáticas y Lectura; ,82 entre Ciencias y Lectura; y ,95 entre Ciencias y Matemáticas, lo que indica que, dentro de cada país, las diferencias entre sus centros tienden a ser similares independientemente de la materia considerada.

Tabla 3

Porcentaje de varianza total y por niveles explicada por el Modelo II (ISEC en ambos niveles) y el Modelo III (todas las variables de ajuste)

	Ciencias Naturales			Matemática			Lectura		
	Nivel 1	Nivel 2	Total	Nivel 1	Nivel 2	Total	Nivel 1	Nivel 2	Total
Dos Niveles (TERCE 2-N)									
Modelo II	1%	50%	22%	1%	48%	25%	2%	69%	32%
Modelo III	3%	52%	24%	2%	52%	28%	4%	72%	35%
Tres Niveles (TERCE 3-N)									
Modelo II	2%	36%	20%	1%	35%	22%	1%	65%	29%
Modelo III	4%	39%	22%	2%	41%	29%	4%	68%	33%

Todos los porcentajes de varianza son significativos al nivel 0,001 (bilateral)

La tabla 3 muestra el porcentaje de varianza total y por nivel explicada por el Modelo II (ISEC del alumnado y del centro) y el Modelo III (con todas las variables de control) para el conjunto de países. Lectura y Matemáticas presentan una asociación entre resultados y variables de ajuste similar, explicando en torno al 30% de la varianza total, mientras que en el caso de Ciencias este porcentaje se ubica sobre el 20%. La fuerza explicativa del ISEC es

mucho mayor en el nivel 2, ya que en el conjunto del estudio (Modelo II, TERCE 3-N) el ISEC da cuenta del 65% de las diferencias intercentros en Lectura y en torno al 35% en Ciencias Naturales y Matemáticas, mientras que los porcentajes de varianza explicada en el nivel individual no superan el 2%. Cuando se incluyen el resto de las variables de contexto (Modelo III, TERCE 3-N) la ganancia explicativa es más discreta en Ciencias, donde la varianza total explicada se incrementa un 2%, que en Lectura y Matemáticas (4% y 7% de incremento respectivamente). Por su parte, los incrementos de la varianza explicada en el nivel 2 están entre el 3% y el 6%, y entre el 1% y el 3% en el nivel 1.

Tabla 4

Porcentaje de varianza total explicada por el Modelo II (ISEC en ambos niveles) y el Modelo III (todas las variables de ajuste) país

	Modelo II			Modelo III		
	Ciencias Naturales	Matemática	Lectura	Ciencias Naturales	Matemática	Lectura
Argentina	20%	16%	25%	27%	20%	32%
Brasil	26%	30%	26%	30%	31%	28%
Chile	14%	12%	13%	18%	14%	16%
Colombia	14%	24%	30%	18%	27%	34%
Costa Rica	16%	15%	21%	20%	20%	26%
Ecuador	13%	11%	32%	16%	15%	36%
Guatemala	32%	28%	32%	34%	35%	35%
Honduras	6%	2%	17%	14%	12%	22%
México	18%	17%	28%	22%	22%	32%
Nicaragua	8%	6%	20%	10%	12%	22%
Nuevo León	19%	13%	21%	20%	17%	25%
Panamá	26%	23%	31%	29%	27%	33%
Paraguay	11%	6%	31%	13%	11%	35%
Perú	32%	33%	47%	34%	38%	51%
Rep. Dominicana	19%	16%	19%	21%	19%	22%
Uruguay	19%	17%	27%	27%	21%	32%

Nota: Todos los porcentajes de varianza son significativos al nivel 0,001 (bilateral)

La tabla 4 muestra el porcentaje de varianza total explicada por los Modelos II y III en cada país. Perú es el país donde las variables antecedentes tienen mayor efecto sobre los resultados, estimándose que entre el 35% y el 50% de la varianza total se explica por factores de contexto. Guatemala presenta valores estables en torno al 35% y en Brasil y Panamá los factores de contexto explican en torno a un tercio de las diferencias totales en todas las materias. En el extremo contrario, Chile, Costa Rica, Nicaragua, Honduras, República Dominicana y Nuevo León son los países donde el efecto de los antecedentes no supera el promedio del modelo TERCE 3-N en las tres competencias. Mención especial merecen Ecuador y Paraguay, donde los factores de contextos explican el 35% de la varianza en Lectura y un porcentaje mucho más discreto en las competencias científico-matemáticas. En todo caso, la correlación entre el porcentaje de varianza total explicada por país en Ciencias y Matemáticas es más alta (,89), que la encontrada cuando se compara con la varianza en Lectura (la correlación Lectura-Matemáticas es ,64 y Lectura-Ciencias ,54), lo que parece señalar que los países donde los factores de ajuste explican más diferencias en una materia tienden a presentar mayor capacidad explicativa en el resto.

Tabla 5

Efecto Neto. Porcentaje de varianza entre centros sin explicar en los modelos II y III

	Modelo II			Modelo III		
	Ciencias Naturales	Matemática	Lectura	Ciencias Naturales	Matemática	Lectura
TERCE (2-N)	28%	36%	21%	27%	34%	20%
TERCE (3-N)	26%	24%	14%	25%	23%	13%
Argentina	19%	24%	17%	17%	21%	11%
Brasil	17%	23%	11%	15%	22%	10%
Chile	15%	12%	10%	12%	11%	8%
Colombia	21%	21%	18%	19%	19%	15%
Costa Rica	10%	13%	7%	10%	12%	5%
Ecuador	31%	38%	15%	30%	36%	12%
Guatemala	13%	22%	11%	12%	17%	9%
Honduras	37%	45%	24%	34%	40%	22%
México	16%	13%	11%	15%	11%	9%
Nicaragua	37%	38%	19%	36%	35%	17%
Nuevo León	10%	7%	7%	9%	6%	6%
Panamá	19%	24%	13%	18%	21%	12%
Paraguay	39%	50%	15%	40%	49%	12%
Perú	23%	36%	23%	22%	32%	19%
Rep. Dominicana	8%	5%	10%	7%	5%	9%
Uruguay	10%	4%	2%	7%	3%	2%

La tabla 5 muestra el efecto del centro, esto es, el porcentaje de varianza entre centros sin explicar en los Modelos II y III después de controlar las variables antecedentes de cada modelo. El ajuste TERCE 3-N señala que, una vez detraído el influjo de todas las variables de contexto, quedaría por explicar el 25% de las diferencias entre los centros en Ciencias Naturales; el 23% en Matemáticas y el 13% en Lectura. Nótese que el modelo de dos niveles con datos completos (TERCE N-2) sobrevalora indebidamente el efecto neto, especialmente en Matemáticas y Lectura. La varianza sin explicar dentro de los países se ajusta al patrón general: en todos los casos (salvo República Dominicana) el efecto neto es mayor en

Matemáticas que en Lectura. No obstante, se advierten importantes diferencias entre los países en el porcentaje de varianza por explicar: en República Dominicana, Uruguay y Nuevo León el margen de mejora de los centros una vez descontado las variables de control está por debajo del 10% de la varianza total. En el extremo contrario Ecuador, Honduras, Nicaragua, Paraguay y Perú son los países que presentan mayor margen de mejora en el conjunto de las tres materias. En todo caso, los datos de los países de nuevo vuelven a estar altamente correlacionados, especialmente en Ciencias y Matemáticas (r = ,96), lo que indica que los países que presentan efectos netos de centro más altos en una materia tienden a presentar efectos más altos en las otras dos.

DISCUSIÓN Y CONCLUSIONES

Para responder a la primera pregunta del estudio se ha estimado el efecto bruto del centro en cada materia expresado como el ICC de un modelo jerárquico-lineal sin predictores, entendiéndose que un efecto bruto mayor se asocia a mayores diferencias entre los centros de un mismo país. Los datos señalan que no se han producido avances en la región en las dos últimas décadas. Los ICC ofrecido por PERCE en el año 1997 (UNESCO-OREALC y LLECE, 2001) y el calculado para TERCE 2-N en el Modelo I (tabla 2) apenas han variado: en Lectura es similar en ambas ediciones y en Matemáticas se observa un descenso de 4 puntos porcentuales en el último estudio. Además, se observa que los ICC de los nueve países que participaron en los dos estudios han experimentado un crecimiento generalizado que se sitúa entre 2 y 22 puntos porcentuales según el país y la materia considerada. Al comparar los ICC de los países con datos en ambos estudios también se advierte que la correlación del efecto bruto del centro es de ,78 en Matemáticas y de ,81 en Lectura, es decir, que los países que presentaban mayores diferencias entre sus centros en la primera evaluación tienden a seguir mostrando más diferencias dos décadas después. Para interpretar coherentemente estas magnitudes debe tenerse en cuenta que el estudio PISA señala que los ICC de los países

Latinoamericanos son superiores al promedio de los países desarrollados. Por ejemplo, son entre 2 y 3 mayores que el efecto bruto de España, y entre 4 y 5 veces más que los ICC de los sistemas educativos con menor efecto bruto (OECD, 2002, 2004, 2007, 2010, 2013, 2016). Estos datos son coherentes con las estimaciones de Scheerens (2000), que concluye que en los países desarrollados el ICC está en torno al 10-15%, mientras que en los países en vías de desarrollo el rango se eleva hasta el 30-40%. Por tanto, la primera conclusión es que, desde que se dispone de datos comparables para la región, las diferencias brutas entre los centros en América Latina se mantienen estables, cuando no aumentan, y que aquellos países con mayores diferencias entre sus centros parece que no han logrado reverter la situación durante este tiempo.

Se ha apuntado que un efecto bruto alto entre centros no tiene porqué significar *per se* una mayor inequidad educativa, ya que el valor del ICC pudiera reflejar las decisiones políticas sobre la ordenación académica de los países (OECD, 2007). Por ello, el estudio se plantea dos preguntas referidas al efecto que las variables de contexto tienen sobre los resultados, ya que en un modelo sistémico estas variables no pueden ser manipuladas por los centros para mejorar sus rendimientos. Por tanto, en la medida en que las variables de ajuste expliquen más varianza el margen de mejora de los centros estará más limitado. Para el conjunto de los países, una vez se introduce el ISEC (Modelo II), la varianza entre centros del Modelo I se reduce en un 65% en Lectura y un 35% en Matemáticas y Ciencias, estimación coherente con los datos reportados en otras investigaciones (Cervini, 2012; Cervini et al., 2016; Murillo y Román, 2011), que señalan que el efecto de ISEC es mayor en Lectura que en las materias científico-matemáticas. Según los datos de PISA la fuerza de la relación entre nivel socioeconómico y cultural y los resultados educativos de los países de la región es similar al promedio de la OCDE (OECD, 2002, 2004, 2007, 2010, 2013). En el Modelo III se introducen el resto de las variables de ajuste y, en términos generales, se aprecia una pequeña ganancia explicativa en las diferencias, lo que

es coherente con los resultados de Doneschi (2017). En general, el ISEC ha mostrado buena capacidad para dar cuenta de la varianza de resultados lo que parece indicar que en TERCE se han superado las limitaciones que presentaba el ISEC construido en la segunda edición (SERCE), y que era la razón por la cual algunos autores señalaron que el índice socioeconómico del segundo estudio tenía menos impacto del esperado en los resultados de la prueba (Cervini, 2012; Murillo y Román, 2011). No obstante, se han observado tres casos (Nicaragua, Honduras y Paraguay) donde el ISEC presentaba un potencial débil o muy débil para explicar la varianza (en especial varianza entre los centros) en Matemáticas, si bien al introducir la colección completa de variables de ajuste todos ellos experimentan ganancias significativas, lo que demuestra la relevancia del uso de índices consistentes que resuman las características sociodemográficas de las familias para introducirlos en diferentes niveles de los modelos jerárquico-lineales (Peña Suárez et al., 2009).

La última pregunta del estudio pretendía estimar el efecto neto del centro, esto es el porcentaje de varianza entre centros que queda sin explicar en el Modelo III una vez se introducen todas las variables de ajuste. Este efecto neto se interpreta como el margen de mejora de los centros, es decir, un porcentaje más alto señala una mayor capacidad de la escuela para mejorar sus resultados más allá de los condicionantes sociológicos y compensar las desigualdades de acceso y permanencia mediante sus prácticas educativas. Los valores estimados del presente estudio (13% en Lectura y 23% en Matemáticas) son coincidentes con los datos de Murillo y Román (2011), y algo más altos que los estimados por Cervini (2012, 2016), probablemente porque la eliminación de escuelas pequeñas y remotas operado por este último haya comprimido las diferencias entre el centro. Los efectos netos estimados también son consistentes con los ofrecidos por LLECE en el segundo estudio regional (UNESCO-OREALC y LLECE, 2010), ya que la correlación entre los efectos netos del centro por país de SERCE y del presente estudio son: ,81 en Lectura y ,82 en Matemáticas y Ciencias Naturales,

lo que señala que en la última década el grado en que las variables antecedentes determinan los resultados de los centros se ha mantenido estable en la región.

Los resultados encontrados ayudan a comprender los fenómenos educativos en América Latina y tienen importantes implicaciones para la toma de decisiones políticas en la región más desigual del mundo en materia educativa (UNESCO-OREALC, 2016a, p. 89) en especial en los procesos de escolarización, los criterios de asignación, ordenación académica y distribución de recursos. Una lectura conjunta de los datos permite establecer al menos dos perfiles de países en función de las diferencias entre centros y del efecto de las variables de ajuste sobre sus resultados. El primer grupo estaría formado por Chile, Costa Rica, Nuevo León, República Dominicana y Uruguay, países que presentan un efecto bruto del centro relativamente pequeño, con ICC de hasta ,27 y un moderado impacto de las variables de ajuste sobre los resultados de entre ,14 a ,32. Es el grupo de países donde los resultados de los centros parecen ser más homogéneos y menos determinados por los antecedentes escolares, pero como contrapunto son los países donde el efecto neto del centro es menor (en general menos del 10%). El segundo grupo está formado por Perú, Panamá, Brasil y Guatemala, que son países donde las diferencias brutas entre sus centros son mayores (el ICC se sitúa mayoritariamente en torno a ,40) y donde el conjunto de variables de contexto explica mayor porcentaje de la varianza total inicial (en torno a un 30%). En estos países los centros muestran mayor heterogeneidad en sus resultados y, según los datos aquí presentados parecen ser los sistemas educativos más desiguales.

El estudio tiene algunas limitaciones. En primer lugar, las medidas de desigualdad presentadas –efecto bruto, porcentaje de varianza explicada por factores de contexto y efecto neto del centro- son clásicas, pero no agotan las posibilidades de estudio de las desigualdades en educación. Sería posible, por ejemplo, construir indicadores de segregación que analicen cómo afecta al rendimiento el agrupamiento del alumnado en los centros en función de su origen social (OECD, 2016; Robert, 2007). La limitación más significativa en el diseño de la

investigación es la falta de una medida del logro previo, que permita detraer varianza fundamentalmente del nivel intra-centros. Intentando aliviar esto se ha utilizado la variable de repetición como un indicador de la historia escolar con dificultades de aprendizaje o de permanencia en el sistema educativo, sin embargo, no hay duda de que una medida de rendimiento previo es el mejor predictor del rendimiento del alumnado individualmente considerado (Fernández-Alonso et al., 2015, 2017). Las futuras líneas de investigación se orientan al estudio de los residuales de centros a la luz de los modelos ajustados y el estudio de los factores de proceso educativo que pudieran dar cuenta del porcentaje de varianza no explicado en los modelos aquí presentados, tales como los estilos parentales (Osorio y González-Cámara, 2016), o las actitudes de los profesores (Álvarez-Martino et al., 2016; Cunha et al., 2015), por citar sólo dos de las muchas posibles. Se ha advertido que, en el modelo que predice el resultado en Matemáticas, tres países presentaban alteraciones muy importantes en el valor del coeficiente de regresión del ISEC dependiendo del uso o no de los pesos muestrales, lo que parece demandar análisis adiccionales de corte metodológico para estudiar cómo los pesos muestrales influyen sobre la relación entre el ISEC y los resultados dentro de los países. Finalmente, existen evidencias que señalan que las diferencias entre los centros de un país pueden ser el resultado de las tradiciones y las circunstancias socio históricas asociadas a la propia construcción y estructura del sistema educativo (Foces Gil, 2015), por lo que cabría la posibilidad de analizar si la organización y ordenación de las enseñanzas de los países están vinculadas a las diferencias en el desempeño de los centros educativos.

REFERENCIAS

Álvarez-Martino, E., Álvarez, M., Castro, P., Campo, M. A., y González-Mesa, C. (2016). Teacher's perception of disruptive behaviour in the classrooms. *Psicothema*, 28(2), 174-180. doi: 10.7334/psicothema2015.215

Blanco-Bosco, E. (2009). La desigualdad de resultados educativos. Aportes a la teoría desde la investigación sobre eficacia escolar. *Revista Mexicana de Investigación Educativa, 14*(43), 1019-1049.

Blanco-Bosco, E. (2011). *Los límites de la escuela: educación, desigualdad y aprendizajes en México*. México D.F.: El Colegio de México, Centro de Estudios Sociológicos.

Carvallo, M. (2006). Factores que afectan el desempeño de los alumnos mexicanos en edad de educación secundaria: un estudio dentro de la corriente de eficacia escolar. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 4*(3), 30-53.

Casas, A., Gamboa, L. F., y Piñeros, L. J. (2002). *El efecto escuela en Colombia, 1999-2000*. Bogotá: Universidad del Rosario.

Castro-Aristizabal, G., Castillo-Caicedo, M., y Mendoza-Parra, J. (2016). *Principales determinantes en la adquisición de competencias en América Latina: Un análisis multinivel a partir de los resultados en PISA 2012*. Cali: Pontificia Universidad Javeriana. doi:10.2139/ssrn.2744657

Cervini, R. (2002a). Desigualdades en el logro académico y reproducción cultural en Argentina. *Revista Mexicana de Investigación Educativa, 7*(16), 445-500.

Cervini, R. (2002b). Desigualdades socioculturales en el aprendizaje de matemática y lengua de la educación secundaria en Argentina: Un modelo de tres niveles. *Revista Electrónica de Investigación y Evaluación Educativa, 8*(2), 135-158.

Cervini, R. (2004a). Influencia de los factores institucionales sobre el logro en matemática de los estudiantes en el último año de la educación media de Argentina- Un modelo de tres niveles. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 2*(1), 1-24.

Cervini, R. (2004b). Nivel y variación de la equidad en la educación media de Argentina. *Revista Iberoamericana de Educación, 34*(4), 1-18.

Cervini, R. (2006). Los efectos de la escuela y del aula sobre el logro en matemáticas y en lengua de la educación secundaria. Un modelo multinivel. *Perfiles Educativos, 27*(112), 68-97.

Cervini, R. (2009). Comparando la inequidad en los logros escolares de la educación primaria y secundaria en Argentina: Un estudio multinivel. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 7*(1), 5-21.

Cervini, R. (2010). El efecto escuela en la educacion primaria y secundaria: El caso de Argentina. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 8*(1), 7-25.

Cervini, R. (2012). El efecto escuela en países de América Latina: Reanalizando los datos del SERCE. *Archivos Analíticos de Políticas Educativas, 20*(39), 1-25.

Cervini, R., Dari, N., y Quiroz, S. (2016). Las Determinaciones Socioeconómicas sobre la distribución de los Aprendizajes Escolares. Los Datos del TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 61-79. doi:10.15366/reice2016.14.4.003

Consejería de Educación y Ciencia del Principado de Asturias (2007). *Evaluación de Diagnóstico Asturias 2007. Procesos y resultados*. Oviedo: Servicio de Evaluación y Calidad. Recuperado de https://goo.gl/uZT7Ck

Cunha, J., Rosario, P., Macedo, L., Nunes, A. R., Fuentes, S., Pinto, R., y Suárez, N. (2015). Parents' conceptions of their homework involvement in elementary school. *Psicothema*, 27(2), 159-165. doi: 10.7334/psicothema2014.210

Doneschi, A. (2017). *Desigualdad de aprendizajes en Uruguay:determinantes de los resultados de PISA 2012.* (Tesis de maestría inédita), Universidad de la República, Montevideo. Recuperado de https://goo.gl/YYnmWX

Fernández-Alonso, R., Álvarez-Díaz, M., Suarez-Alvarez, J., & Muñiz, J. (2017). Students' achievement and homework assignment strategies. *Frontiers in Psychology, 8*, 286. doi:10.3389/fpsyg.2017.00286

Fernández-Alonso, R., y Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las Competencias Básicas. *Aula Abierta, 39*(2), 3-34.

Fernández-Alonso, R., Suárez-Álvarez, J., y Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas. *Psicothema, 24*(1), 167-175.

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2015). Adolescents' homework performance in mathematics and science: Personal factors and teaching practices. *Journal of Educational Psychology, 107*(4), 1075-1085. doi:10.1037/edu0000032

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2016). Homework and performance in mathematics: the role of the teacher, the family and the student´s background. *Revista de Psicodidáctica, 21*(1), 5-23. doi:10.1387/RevPsicodidact.13939

Ferrão, M. E., y Fernandes, C. (2001). A escola brasileira faz diferença? Uma investigaçao dos efeitos da esco la na proficiência em Matemática dos alunos da 4ª série. en C. Franco (Ed.), *Promoçao, ciclos e avaliaçao educacional* (pp.155-172). Curitiba: Artmed.

Foces, J. A. (2015). PISA, IDE e IPE: Evidencia empírica de las desigualdades educativas entre las regiones españolas. *Revista de Psicología y Educación, 10*(1), 173-192.

Gobierno del Principado de Asturias. (2016). La repetición escolar: hechos y creencias. *Informes de Evaluación, 2*. Recuperado de https://goo.gl/MKsGVn

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.

Martínez-Arias, R. (2010). La evaluación del desempeño. *Papeles del Psicólogo, 31*(1), 85-96.

Martínez-Arias, R., Gaviria-Soto, J. L., y Castro-Morera, M. (2009). Concepto y evolución de los modelos de valor añadido en educación. *Revista de Educación* 348, 15-45.

Miralles, M. J., Castejón, L., Pérez, A., y Gilar, R. (2012). El análisis de los efectos de la escuela sobre el rendimiento académico en matemáticas: Un análisis multinivel con datos de PISA 2003. *Revista de Psicología y Educación, 7*(1), 83-110.

Mislevy, R. J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characterstics from sparse matrix samples of item responses. *Journal of Education Measurement, 29*(2), 131-161. doi:10.1111/j.1745-3984.1992.tb00371.x

Murillo, F. J. (2003). *La investigación sobre Eficacia Escolar en Iberoamerica. Revisión Internacional sobre el Estado del Arte*. Bogotá: CAB/CIDE.

Murillo, F. J. (2005). *La investigación sobre eficacia escolar*. Barcelona: Octaedro.

Murillo, F. J. (2007). School Effectiveness Research in Latin America. en T. Townsend (Ed.), *International Handbook of School Effectiveness and Improvement* (pp.75-92). Dordrecth: Springer.

Murillo, F. J. (2008). Enfoque, situación y desafíos de la investigación sobre eficacia escolar en Latinoamérica en *Eficacia escolar y factores asociados en América Latina y el Caribe* (pp.17-48). Santiago de Chile: OREALC-UNESCO y LLECE.

Murillo, F. J., & Román, M. (2011). ¿La escuela o la cuna? Evidencias sobre su aportación al rendimiento de los estudiantes de América Latina. Estudio multinivel sobre la estimación de los efectos escolares. *Revista de currículum y formación del profesorado, 15*(3), 27-50.

OCDE. (2002). *Conocimientos y aptitudes para la vida. Primeros resultados del Programa Internacional de Evaluación de estudiantes (PISA) 2000 de la OCDE*. México: Aula XXI Santillana.

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. (2007). *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*. París: OECD Publishing.

OECD. (2009). *PISA Data Analysis Manual SPSS®* (Second ed.). París: OECD Publishing.

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

OECD. (2013). *PISA 2012 Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed* doi:/10.1787/9789264201132-en

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education* PISA (Ed.) doi:/10.1787/9789264266490-en

Osorio, A. y González-Cámara, M. (2016). Testing alleged superiority of the indulgent parenting style among Spanish adolescents. *Psicothema*, 28(4), 414-420. doi: 10.7334/psicothema2015.314

Peña-Suárez, E., Campillo-Álvarez, Á., Santarén-Rosell, M., y Muñiz, J. (2012). El papel de los centros escolares en la adquisición de la competencia científica. *Revista Iberoamericana de Psicología y Salud, 3*(1), 75-87.

Peña-Suárez, E., Fernández-Alonso, R., y Muñiz, J. (2009). Estimación del valor añadido de los centros educativos. *Aula Abierta, 37*(1), 3-18.

Piñeros, L. J., y Rodríguez, A. (1998). *Los insumos escolares en la Educación Secundaria y su efecto sobre el rendimiento académico de los estudiantes: un estudio en Colombia*. Washington, D.C.: Banco Mundial.

Raudenbush, S., Bryk, A., Cheong, Y. K., Congdon, R., & du Toit, M. (2011). *HLM 7 Hierarchical Linear and Nonlinear Modeling*. United States of America: SSI Scientific Software International, Inc.

Reimers, F. (2002). *Distintas escuelas, diferentes oportunidades. Los retos para la igualdad de oportunidades en Latinoamérica*. Madrid: La Muralla.

Robert, P. (2007). *The influence of educational segregation on educational achievement* (Vol. RSCAS 2007/29): European University Institute.

Rodríguez-Jiménez, O., y Murillo, F. J. (2011). Estimación del efecto escuela para Colombia. *Revista Internacional de Investigación en Educación, 3*(6), 299-316.

Rodríguez, S., Piñeiro, I., Regueiro, B., Estevez, I., & Val, C. (2017). Estrategias cognitivas, etapa educativa y rendimiento académico. *Revista de Psicología y Educación, 12*(1), 19-34.

Scheerens, J. (2016). *Educational Effectiveness and Ineffectiveness a Critical Review of the Knowledge Base*: Springer Netherlands. doi: 10.1007/978-94-017-7459-8_1

Scheerens, J., Witziers, B., & Steen, R. (2013). A Meta-analysis of School Effectiveness. *Revista de Educación,* 361, 619-645.

Sirin, S. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research, 75*(3), 417-453. doi:10.3102/00346543075003417

Suárez-Álvarez, J., Fernández-Alonso, R., & Muñiz, J. (2014). Self-concept,motivation, expectations, and socioeconomic level as predictors of academic performance inmathematics. *Learning and Individual Differences, 30,* 118-123. doi:10.1016/j.lindif.2013.10.019

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research*. London: Falmer Press.

Towsend, T. (2007). *International handbook of school effectiveness and improvement*. Dordrecth: Springer.

UNESCO-OREALC. (2016a). *Recomendaciones de Políticas Educativas en América Latina en base al TERCE*. Santiago de Chile: UNESCO.

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

UNESCO-OREALC, y 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, y LLECE. (2001). *Informe Técnico. Primer estudio internacional comparativo sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la educación básica*. Santiago de Chile: UNESCO.

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

UNESCO-OREALC, y LLECE. (2016a). *Informe de resultados del Tercer Estudio Regional Comparativo y Explicativo. Factores Asociados*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2016b). *Informe de resultados del Tercer Estudio Regional Comparativo y Explicativo. Logros de aprendizaje*. Santiago de Chile: UNESCO.

UNESCO. (2000). *Marco de Acción de Dakar. Educación para Todos cumplir nuestros compromisos comunes.* Recuperado de https://goo.gl/BEao3p

Zorrilla, M. (2009). ¿Cuál es la aportación de la escuela secundaria mexicana en el rendimiento de los alumnos en Matemáticas y Español? *Revista electrónica de investigación educativa, 11*(2), 1-29.

CAPÍTULO III. Estructura Sociocultural de los Centros Educativos y Desempeño Académico en Latinoamérica

RESUMEN

El Tercer estudio regional comparativo y explicativo de la UNESCO analiza y compara los resultados académicos en matemáticas, ciencias naturales y lectura en 15 países de América Latina más el Estado de Nuevo León México. La presente investigación analiza el impacto del agrupamiento heterogéneo u homogéneo en el rendimiento en matemáticas. La muestra está conformada por 61.673 estudiantes que representan a más de nueve millones de estudiantes de toda la región. Para el análisis de los datos inicialmente se calcularon estadísticos descriptivos y correlaciones de Pearson, para luego realizar análisis de regresión lineal múltiple, se utilizó el software SPSS y el módulo de replicates bajo la estimación BBR, teniendo en cuenta la naturaleza de los datos de la base TERCE. Los principales resultados evidencian que las agrupaciones heterogéneas benefician a los estudiantes de SEC bajo, sin perjudicar a los estudiantes de SEC alto. Con los datos de la muestra completa de TERCE las agrupaciones heterogéneas permiten reducir a la mitad la diferencia de resultados entre las clases sociales. Una medida practica que los sistemas educativos podrían aplicar con el fin de disminuir las diferencias en los resultados de aprendizaje, controlando el efecto de las clases sociales en la educación.

Palabras Clave: Segregación escolar, efecto del nivel socioeconómico, eficacia escolar, TERCE

INTRODUCCIÓN

Cinco décadas después de la promulgación de la ley ¨Civil Rights Act of 1964" en Estados Unidos la desigualdad de oportunidades educativas sigue siendo un hecho transcultural, tal y como demuestran estudios anglosajones (Palardy, Rumberger, & Butler, 2015), centroeuropeos (Felouzis, 2005; Gülseli & de Valk, 2012; Riedel, Schneider, Schuchart, & Weishaupt, 2010), asiáticos (Ting & Ronald, 2017) e hispanoamericanos (UNESCO-OREALC & LLECE, 2016a). El estudio sobre la desigualdad educativa se inicia como reacción a las conclusiones del informe *Equality of Educational Opportunity* (Coleman et al., 1966), y es una línea de investigación influenciada por las teorías educativas de ascendencia sociológica que señalan que el éxito educativo está determinado en gran parte por el capital cultural y los saberes específicos y lingüísticos del grupo dominante y que, por tanto, la educación reproduce o perpetúa las desigualdades de una determinada sociedad (Bordieu & Passeron, 1996).

Si bien la desigualdad es un campo basto es posible organizar los trabajos en función de tres finalidades básicas: describir la desigualdad en la distribución de los estudiantes en las escuelas en función de sus características sociológicas, analizar el impacto que tienen los antecedentes sociológicos en el rendimiento escolar; comparar los efectos que distintas agrupaciones del alumnado (homogéneas o heterogéneas) tienen en los resultados educativos.

La línea de estudio más clásica tiene como finalidad documentar la existencia de desigualdades entre los centros por el efecto de la composición de su matrícula (Cervini, Dari, & Quiroz, 2016; Martínez-Garrido, 2015; Murillo & Román, 2011; Scheerens, 2000; Woitschach, Fernández-Alonso, Martínez-Arias, & Muñiz-Fernández, 2017). Las medidas de segregación han evolucionado en las últimas décadas, Murillo (2016) en un trabajo de corte metodológico, compara la eficacia de cinco medidas de estimación de la segregación en América Latina señalando que entre el 45% y el 60% de estudiantes del grupo minoritario

deberían cambiar de escuela para que exista una distribución homogénea en las mismas. Por su parte, Jaume (2013) ofrece un dato similar (47%) para Argentina y la UNESCO-OREALC y LLECE (2016a) lo eleva hasta el 55% para el conjunto de Latinoamérica. La proliferación de estudios en esta línea permite afirmar que dentro de los sistemas educativos existen altos niveles de segregación económica (Castro-Aristizabal & Giménez, 2017; Chiu, 2015; Murillo, 2016; Murillo & Martínez-Garrido, 2017a), racial-étnica (Murillo & Martínez-Garrido, 2017a) y lingüística (Nava-Gómez & Pérez-Cervantes, 2015; Ortíz-Sandoval, 2012).

Una segunda línea de investigación está compuesta por los estudios que muestran el impacto de los antecedentes sociodemográficos en los resultados educativos. En ese contexto se destaca el volumen y consistencia de los resultados las investigaciones centradas en analizar el efecto del nivel socioeconómico y cultural (Sirin, 2005; White, 1982). Los programas de evaluación internacional como PISA, TIMSS y PIRLS han señalado que las características de las familias, los recursos materiales en el hogar y la composición de la matrícula del centro impactan en mayor o menor medida en los resultados de todos los países participantes (Benavides, León, & Etesse, 2014; Chiu, 2015; Martin, Mullis, Foy, & Hooper, 2016; Mullis, Martin, Foy, & Hooper, 2016; Mullis, Martin, Foy, & Hooper, 2017; OECD, 2016; Phillips, Larsen, & Hausman, 2015).

De igual modo, las comparaciones en la región latinoamericana (PERCE, SERCE y TERCE) han demostrado la marcada influencia del capital cultural y económico de las familias en el rendimiento de los estudiantes en todas las materias evaluadas (UNESCO-OREALC & LLECE, 2000, 2010, 2016), hecho que también se confirma en evaluaciones nacionales (Treviño, Valenzuela, & Villalobos, 2014). El contexto latinoamericano, se caracteriza por una amplia diversidad lingüística y cultural, son destacables los trabajos que demuestran que los estudiantes indígenas presentan menor rendimiento incluso después de controlar otras variables

de contexto (Cueto & Secada, 2003; UNESCO-OREALC, 2016a; Webb, Canales, & Becerra, 2017), fenómeno relacionado con el uso de una lengua indígena (Suárez-Enciso, Elías, & Zarza, 2016; UNESCO-OREALC & LLECE, 2016; Woitschach, 2016) y las condiciones de vulnerabilidad de las familias que en su mayoría residen en zonas rurales (UNESCO-OREALC & LLECE, 2016). Igualmente, existen abundantes evidencias que analizan los efectos del acceso a la educación privada que está vinculada a la desigualdad de origen económico, la proliferación de oferta educativa privada ha aumentado gradualmente los niveles de segregación en las escuelas (Arcidiácono et al., 2014; Castro-Aristizabal, Castillo-Caicedo, & Mendoza-Parra, 2016; Castro-Aristizabal & Giménez, 2017; Gasparini, Jaume, Serio, & Vazquez, 2011; Murillo & Martínez-Garrido, 2017b), aunque está forma de segregación también puede observarse en escuelas públicas, asociada en este caso a la segregación residencial (Balarín, 2016), o a las políticas de asistencia a los estudiantes en situación de vulnerabilidad en los sistemas educativos de la región (Elacqua, 2012).

Finalmente, dentro de la línea que analiza el impacto de las variables sociológicas en el rendimiento es necesario mencionar los trabajos que estudian la asociación entre la clase social y la conducta docente. Las conclusiones parecen afirmar, que el efecto de los docentes es alto en aulas heterogéneas (Sanders, Wright, & Horn, 1997). Agirdag, van Avermaet, y van Houtte (2013) señalan que los docentes tienden a presentar expectativas bajas cuando los grupos-aula están conformados por estudiantes no nativos y de clase trabajadora, produciendo un efecto Pigmalión, dicho efecto ha sido ampliamente replicado en diferentes culturas (Chang, 2011; OECD, 2013; Rosenthal & Jacobson, 1968; Treviño, 2003).

La tercera línea de investigación es la más novedosa y nace del intento de abordar el estudio de la segregación desde una visión más pragmática que busca ofrecer evidencias para que los sistemas educativos pueden manejar la desigualdad dentro de la exigencia de las

escuelas inclusivas (UNESCO-OREALC, 2016a; UNESCO-OREALC & LLECE, 2016). Son estudios que ponen su acento lo que ocurre dentro de las aulas y de los centros cuando estos escolarizan grupos heterogéneos de estudiantes en función de sus características sociológicas.

Como estudio seminal el informe el Informe Coleman (1966, p. 22), cita en sus resultados que los estudiantes que provienen de familias con características como estabilidad económica y educación de los padres, pero que asisten a escuelas con grupos desfavorecidos presentan un rendimiento más bajo al que tendrían si fueran a escuelas con niños de sus mismas características socioeconómicas. Por otro lado, si estudiantes con características de desventaja social asisten a escuelas con niños de un nivel socioeducativo mayor, los niños en desventaja presentan un incremento en su rendimiento. Mientras que Róbert (2007) señala que los alumnos de alto nivel socioeconómico no se ven afectados por la integración con estudiantes de bajo nivel socioeconómico, aunque tampoco encuentra que los estudiantes de bajo estatus socioeconómico y cultural se beneficien de la integración en grupos de mayor nivel sociológico. Oakes, Ormseth, Bell, & Camp, 1990 destacan que una de las consecuencias de las agrupaciones es que los estudiantes con alto nivel SEC tienen mayor acceso a las oportunidades y calidad en el aprendizaje, comparados con alumnos de bajo SEC y bajo nivel de habilidad. Denotando la necesidad de proveer altos niveles de calidad en educación a los niños minoritarios en las escuelas donde ellos se encuentran (Diem & Boorks, 2013).

En cuanto al efecto de la agrupación sobre el aprendizaje la revisión teórica realizada por Slavin (1990) destaca la existencia entre un leve y hasta un nulo efecto de la agrupación homogénea. Por el otro lado, el metaanálisis de Kulik y Kulik (1992) destaca la ganancia académica de las agrupaciones por habilidad. Mientras que efectos negativos de las agrupaciones por habilidad se observan en escuelas de Gran Bretaña (Boaler, Wiliam, & Brown, 1998).

De entre las consecuencias negativas de las agrupaciones homogéneas, Oakes (2005) apunta que no existe un beneficio claro y que los grupos minoritarios se ven afectados académica y emocionalmente al agruparse. Mientras que, Kulik y Kulik (1992) apuntan que las agrupaciones no generan un efecto devastador en la autoestima de los estudiantes y que en todo caso se observa un leve efecto positivo en las agrupaciones de alto nivel de habilidad y un leve efecto negativo en las agrupaciones de bajo nivel de habilidad.

En cuanto a la percepción de los estudiantes sobre las ventajas o desventajas del agrupamiento por habilidades ya sea homogéneo u heterogéneo, un estudio realizado con 44 estudiantes compara las percepciones en el ámbito académico y social. Indicando que los estudiantes perciben la existencia de un beneficio de estas agrupaciones, así como la existencia de las desventajas sociales al estar agrupados de forma homogénea. Mientras que las percepciones de la agrupación heterogénea son negativas en cuanto al ámbito académico y positivas en lo que se refiere a la socialización y diversidad entre pares (Adams-Byers, Whitsell, & Moon, 2004).

El presente trabajo desde una mirada innovadora para América Latina se encuadraría dentro de esta tercera línea de investigación sobre desigualdades. Planteándose dos objetivos específicos que pretenden dar cuenta del comportamiento del tipo de agrupamiento en las aulas de 16 naciones de Latinoamérica.

1. Analizar el impacto de la segregación escolar en el logro académico de los estudiantes en las pruebas de matemáticas de TERCE
2. Determinar qué tipo de agrupación del alumnado (homogénea o heterogénea) influye de forma positiva en el logro académico del alumnado.

MÉTODO

Participantes

El Tercer Estudio Regional Comparativo y Explicativo (TERCE) tiene como finalidad evaluar la calidad de la educación primaria en América Latina y el Caribe es organizado por el Laboratorio Latinoamericano de la Calidad de la Educación (LLECE) y coordinado por la Oficina Regional de Educación para América Latina y el Caribe. La muestra de esta investigación está compuesta por 61.673 estudiantes y 2.957 centros educativos de educación primaria obligatoria matriculados en el año académico 2013 en 15 países más el estado de Nuevo León (México) participantes en el TERCE y que representan a una población de nueve millones de estudiantes de 6º curso de América Latina. Dentro de cada país la muestra fue seleccionada siguiendo un diseño bietápico por conglomerados y estratificado propio de este tipo de comparaciones internacionales (Joncas & Foy, 2012; UNESCO-OREALC, 2016b).

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).

Estos pesos muestrales tienen dos aplicaciones importantes. La primera es que permiten reconstruir el tamaño poblacional del país o estrato estudiado. Y una segunda aplicación importante de estos pesos, es que permiten emplear métodos de replicación o re-muestreo. Los métodos de replicación, que se emplean en la evaluación de sistemas educativos desde hace un cuarto de siglo (Mislevy, Beaton, Kaplan, & Sheehan, 1992), se basan en la idea de que los estimadores poblacionales (media, desviación típica…) y sus correspondientes errores típicos serían más precisos si en los estudios se pudiera contar con varias muestras de una misma población en vez de que con una única muestra. Esta segunda aplicación de los pesos es sumamente importante a la hora establecer las conclusiones de un estudio ya que los procedimientos de replicación son más adecuados para estimar las varianzas muestrales y los errores típicos de los estimadores poblacionales que los métodos de análisis clásico. 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).

El estudio TERCE dispone de dos tipos de pesos: los pesos replicados y pesos senatoriales. Los pesos replicados, son transformaciones del peso original cuya suma, como se acaba de apuntar, representa el número total de estudiantes de 6º curso de cada país. Por su parte, el peso senatorial es una transformación del peso original para que la suma de pesos por país sea exactamente igual en todos ellos. Lo que permite estimar las estadísticas

internacionales haciendo que todos los países aporten el mismo peso, sin perjudicar a los países pequeños, que de otra forma contribuirían en menor medida que los grandes al cálculo de los estadísticos internacionales.

Materiales y Procedimiento

El proceso de evaluación fue desarrollado bajo los estándares éticos de la UNESCO y contó con la participación y seguimiento de los gobiernos de cada país participante. Un mayor detalle sobre el estudio se puede encontrar en el estudio realizado con la muestra TERCE de Woitschach et al. (2017) así como en el reporte técnico del LLECE (UNESCO-OREALC, 2016b)

La prueba de matemáticas se desarrolló a partir de una tabla de especificaciones organizada en cinco dominios y tres procesos cognitivos y constaba de 98 ítems (UNESCO-OREALC, 2016b), en su mayoría de elección múltiple, agrupados en seis bloques. La puntuación de cada estudiante fue calculada mediante la metodología de valores plausibles que es la más eficiente para recuperar los parámetros poblaciones en las evaluaciones de sistemas educativos (OECD, 2009).

Todas las variables independientes se extrajeron con información proveniente de las respuestas ofrecidas por el alumnado, sus familias, el profesorado y las direcciones a sus correspondientes cuestionarios de contexto. Las variables independientes son de dos tipos: variables interés y variables de control o ajuste.

Cinco variables que describen las características familiares.

(a) El nivel socioeconómico y cultural de la familia (SEC) está construido por TERCE y se trata de una variable ordinal de cuatro niveles, siendo 1 = SEC bajo y 4 = SEC alto. Para este estudio se han elaborado tres variables continuas, construidas a partir de las respuestas del alumnado y sus familias mediante un análisis factorial de máxima verosimilitud y

expresadas en puntos típicos, N (0,1). *Índice de servicios básicos de la vivienda* (SBA) que resume la disponibilidad o no en el hogar de los siguientes servicios: agua, electricidad, alcantarillado y recogida de basura. *Índice de servicios de comunicación de la vivienda* (SCA), que resume la disponibilidad en el hogar de medios de comunicación como teléfono, TV por cable o Internet. *Índice de posesiones en el hogar* (HOP), que resume la disponibilidad en el hogar de diversos recursos materiales, como electrodomésticos o automóviles. Las alfas de Cronbach de los tres índices construidos son: SBA = 0,55; SCA = 0,46; HOP = 0,87. La pregunta sobre el *Número de libros en el hogar* fue recodificada a los valores medios del intervalo y sus valores posibles son: 0, 7,15, 25, 40 y 80. Finalmente, la variable *Recibe ayuda del gobierno*, es dicotómica y 1 señala a las familias que son beneficiarias de alguna ayuda estatal.

(b) Estudios familiares: Para resumir la información se construyeron cuatro variables dicotómicas que reflejan el nivel de estudios más alto alcanzado por cualquiera de los dos progenitores: *Estudios básicos* (variable de referencia), *postobligatorios*, *superiores*, y *sin información* sobre estudios familiares.

(c) Estructura familiar: Para resumir dicha información se construyeron cinco variables dicotómicas que reflejan otras tantas estructuras familiares: *Nuclear* (variable de referencia), *monoparental*, *extensa*, *otra estructura*, y *sin información* sobre la estructura familiar.

(d) Cinco variables para describir el perfil del alumnado: Todas ellas dicotómicas (0/1) donde 1 significa: ser *mujer*, ser *indígena*, hablar en casa una *lengua distinta* a la empleada en la evaluación, tener que *trabajar* (además de estudiar) y *haber repetido* durante la escolaridad.

(e) Siete variables que describen las características del centro: Que fueron calculadas como los promedios por centro de los índices SEC, SBA, SAC y HOP, y como el porcentaje de

familias del centro que reciben ayuda del gobierno y que tienen estudios básicos y superiores.

Las variables de interés en este estudio están relacionadas con la influencia que la composición de la matrícula de las escuelas (centros) tienen en logro académico. Para estudiar este hecho se han elaborado tres variables:

Variables de interés.

(a) Nivel de segregación del centro (sd_centro): Expresa la desviación típica del índice SEC del centro.

(b) Interacción 1 (SECQ1 x sd_centro): Esta variable estima el rendimiento de los estudiantes de SEC bajo que se escolarizan en agrupaciones heterogéneas. Esta variable busca estimar el efecto de la interacción entre el nivel SEC del alumnado y el nivel de segregación del centro. Se creó una variable dicotómica, denominada SECQ1 donde 1 = estudiante cuyo SEC se encuentra en el cuartil inferior y 0 = cualquier otro SEC, y que fue multiplicada por el nivel de segregación del centro. Al margen de los valores 0, una puntuación baja señala que los estudiantes se escolarizan en agrupaciones homogéneas en cuanto a SEC y valores altos indicarán agrupaciones heterogéneas.

(c) Interacción 1 (SECQ4 x sd_centro): Esta variable estima el rendimiento de los estudiantes de SEC alto que se escolarizan en agrupaciones heterogéneas. Para calcular el término de la interacción se creó una variable dicotómica, denominada SECQ4 donde 1 = estudiante cuyo SEC se encuentra en el cuartil superior y 0 = cualquier otro SEC, y que fue multiplicada por el nivel de segregación del centro. Al margen de los valores 0, una puntuación baja señala que los estudiantes se escolarizan en agrupaciones homogéneas en cuanto a SEC y valores altos indicarán agrupaciones heterogéneas.

Análisis de los Datos

A continuación, se calcularon estadísticos descriptivos y correlaciones Pearson y posteriormente se compararon dos modelos de regresión múltiple. El primero sólo con las variables de interés y el segundo incluyendo también todas las variables de ajuste o control, ambos modelos se ajustaron para la muestra completa y para cada país. Todos los análisis se realizaron con SPSS, pero en vez de emplear el módulo original, se implementaron las macros del módulo *Replicates* desarrollado por ACER para el estudio PISA (OECD, 2009), ya que las estimaciones resultantes son más eficientes que las ofrecidas por el módulo original de SPSS o por el software de análisis multinivel al adecuarse mejor al diseño muestral empleado por TERCE. Para la estimación se eligió la opción de replicación repetida balanceada (BBR) para 100 pesos replicados con la variante de Fay (Fay, 1989; Rust, 1985). Para confirmar que dicha opción era la correcta, previamente se calcularon las medias en matemáticas, y sus correspondientes errores típicos, de todos los países, encontrándose resultados exactamente iguales a los reportados en el informe TERCE. Esta evidencia parece confirmar la idoneidad de la opción elegida.

En los modelos con la muestra completa se emplearon los pesos senatoriales por lo que los países tienen una aportación paritaria al estimador internacional. En los análisis segregados por país se emplearon las réplicas de los pesos originales, de modo que los datos reproduzcan adecuadamente el tamaño poblacional. El rango de casos perdidos en las variables oscila entre el 5% y el 12%, y para su recuperación se empleó una estrategia en dos pasos. Inicialmente los casos incompletos fueron imputados con la media del sujeto y, a continuación, los datos totalmente perdidos fueron recuperados mediante el método iterativo (método EM con variables auxiliares) que implementa el módulo Missing Value Análisis de SPSS (Fernández-Alonso, Suárez-Álvarez, & Muñiz, 2012).

RESULTADOS

La tabla 1 muestra los coeficientes de regresión y sus correspondientes errores típicos de dos modelos de regresión.

Tabla 1

Coeficientes de regresión de los modelos 1 y 2

		Modelo 1		Modelo 2	
		Stat	SE	stat	SE
	@INTERCEPT	711,2	4,7	618,0	4,2
	Efecto segregación (sd_SEC)	-8,8	6,4	-54,8	3,7
	Inter: SEC_Bajo * Baja_Segregación (SECQ1xSD)	-65,1	2,4	18,2	3,2
	Inter: SEC_Alto * Baja_Segregación SECQ4xSD	85,7	2,8	-1,5	3,3
Características familiares	Índice SEC (Nivel Socioeconómico y cultural)			18,2	1,7
	Índice SBA (SS. Básicos: agua, luz...)			-18,5	2,1
	Índice SCO (SS. Comunicación: TV cable, Net.)			-3,4	1,6
	Índice HOP (Posesiones en el Hogar)			2,9	1,4
	Número de libros en el hogar			1,4	1,9
	Recibir ayudas del gobierno			-22,1	2,0
Estudios familiares	Estudios Postobligatorios			-11,0	1,1
	Estudios Superiores			-14,5	2,2
	Sin información sobre estudios			2,5	1,7
Estructura familiar	Familia Monoparental			-1,9	0,6
	Familia Extensa			-15,2	1,2
	Otra Estructura Familiar			0,3	0,0
	Sin información sobre estructura familiar			-29,9	4,1
Perfil del estudiante	Ser indígena / nativo			-25,7	1,4
	Ser niña			6,6	0,6
	Hablar otra lengua distinta en casa			-5,5	0,6
	Haber repetido			-11,7	1,0
	El niño/a trabaja			-5,1	2,3
Variables de centro	Promedio SEC			-40,1	5,6
	Promedio SBA			3,2	3,8
	Promedio SCO			26,1	3,9
	Promedio HOP			-25,8	4,1
	Proporción Familias con estudios superiores			47,1	3,3
	Proporción de familias que reciben ayudas			-55,7	5,7
	Proporción de estudiantes que trabajan			-6,2	18,8

En el primer modelo, que sólo incluye las variables de interés, indica que el efecto de la segregación no es significativo, aunque los estudiantes de SEC bajo que se escolarizan en centros con agrupaciones heterogéneas e inclusivas tienen a presentar peores resultados que aquellos que se escolarizan en centros donde el alumnado procede única o mayoritariamente

de los estratos sociales bajos, mientras que los estudiantes de SEC alto que se escolarizan en grupos heterogéneos presentan mejores resultados. En el segundo modelo se incluyen todas las variables de ajuste y los resultados presentan importantes modificaciones. En primer lugar, el efecto de la segregación es significativo indicando que a medida que los centros presentan mayores dispersiones en el índice SEC tienden al presentar resultados más bajos, siendo el efecto importante por cada punto que aumenta la desviación típica del SEC del centro, se predicen 55 puntos de caída en matemáticas. Sin embargo, una vez se controlan las variables de ajuste los resultados indican que los estudiantes de SEC bajo se benefician por asistir a centros con agrupaciones inclusivas, mientras que los estudiantes de SEC alto no se ven perjudicados por asistir a agrupaciones heterogéneas ya que efecto de esta interacción no prácticamente nulo y, por supuesto, no significativo.

La Figura 1 representa el beneficio en términos de equidad educativa de las agrupaciones heterogéneas. En grupos altamente homogéneos el modelo predice que la diferencia de puntuación entre un estudiante de SEC bajo y otro de SEC alto está en torno a 55 puntos, es decir, más de media desviación típica. Sin embargo, esta distancia se reduce a la mitad se (unos 21 puntos) cuando estos estudiantes se escolarizan en grupos heterogéneos.

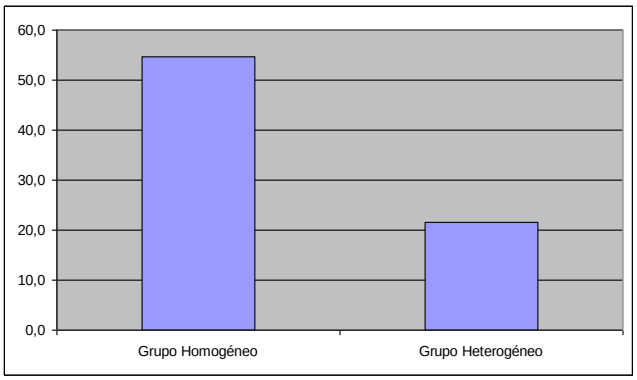


Figura 1. Diferencia de puntuación en matemáticas para los estudiantes de SEC Alto y Bajo en función del tipo de agrupamiento del centro. Todos los países de TERCE

Los datos parecen indicar que los agrupamientos heterogéneos permiten reducir la diferencia en los resultados en matemáticas entre las clases sociales. La siguiente cuestión es si esta afirmación se puede extender a todos los países de TERCE. En la siguiente tabla (Tabla 2) se muestran los resultados del modelo dos segregados por países. Por motivos de parsimonia se prescinde de presentar los coeficientes de regresión de las variables de ajuste.

Tabla 2

Coeficientes de regresión del modelo 2 segregado por países

	Efecto segregación (sd_SEC)		Inter: SEC_Bajo * Baja_Segregación (SECQ1xSD)		Inter: SEC_Alto * Baja_Segregación SECQ4xSD	
	Stat	SE	Stat	SE	Stat	SE
Total	**-54,8**	**3,7**	**18,2**	**3,2**	-1,5	3,3
Argentina	***-45,5***	***16,5***	***31,1***	***12,7***	***-20,0***	***8,9***
Brasil	-12,3	23,1	13,5	12,8	***-26,0***	***14,5***
Chile	1,8	19,9	19,0	12,3	12,3	9,4
Colombia	***-49,4***	***13,4***	14,5	13,4	1,6	16,3
Costa rica	***-30,8***	***13,6***	13,6	10,4	-0,3	9,5
Dominicana	-3,0	12,4	11,8	6,5	3,3	8,1
Ecuador	***-23,8***	***11,0***	***24,5***	***11,0***	-12,8	12,0
Guatemala	-14,2	12,5	8,0	8,8	-3,1	12,7
Honduras	***-35,1***	***13,9***	4,0	11,7	8,4	14,1
México	-15,2	12,5	15,8	12,5	-2,3	11,0
Nicaragua	-20,8	14,7	-1,4	9,4	-3,1	10,3
Panamá	***-23,4***	***12,7***	***29,0***	***11,8***	-13,8	12,1
Paraguay	***-52,3***	***10,3***	***26,8***	***10,3***	-4,1	10,0
Perú	9,3	12,3	5,5	9,7	5,5	9,4
Uruguay	10,6	29,4	33,2	23,4	-22,4	22,1
Nuevo León (México)	-19,8	12,2	***24,5***	***12,2***	0,5	9,8

Nota: En negrita valores p<.05.

Los resultados indican que los países presentan efectos diferentes a la hora de tratar las diferencias sociales. En la mayoría de los países se encuentra que el efecto que las agrupaciones heterogéneas presentan coeficientes negativos. Sin embargo, sólo en seis casos el efecto es significativo (Argentina, Colombia, Costa Rica, Ecuador, Honduras y Paraguay), al que se puede unir Panamá, donde el efecto es marginalmente significativo (alfa > 0,1). En el extremo contrario se encontrarían Chile, Perú y Uruguay que incluso presentan efectos positivos, aunque ninguno de ellos se muestra estadísticamente significativo.

Por otro lado, el efecto positivo de las agrupaciones heterogéneas para los estudiantes de SEC bajo también aparece en la mayoría de los países. Sin embargo, sólo en cinco casos (Argentina, Ecuador, Panamá, Paraguay y Nuevo León) este efecto parece libre de la duda estadística. En todo caso, debe notarse que los errores de estimación son importantes, por lo que, aunque los efectos netos en países como Brasil, Chile, Colombia, Costa Rica, República Dominicana, México y Uruguay son importantes no alcancen el nivel de significación estadística.

Finalmente, también es mayoritaria la conclusión de que los estudiantes de SEC alto no se ven penalizados por asistir a agrupaciones heterogéneas. El efecto de la interacción SEC alto y baja segregación es cercano a cero en Colombia, Costa Rica, República Dominicana, Guatemala, Honduras, México Nicaragua, Paraguay, Perú y Nuevo León. Sólo en dos casos (Argentina y Brasil) se encuentra un efecto negativo y significativo, mientras que en Uruguay el efecto no es significativo, peso a que el coeficiente de regresión es importante.

DISCUSIÓN Y CONCLUSIONES

La desigualdad de oportunidades educativas es un hecho transcultural latente. América Latina no solo presenta el desafío de integración dentro de la escuela, sino que además trabaja arduamente en la disminución de la brecha de acceso a la educación en las comunidades más desfavorecidas. Los resultados de esta investigación ofrecen evidencias de apoyo para las políticas de inclusión educativa, desde una mirada intra-escuela. El acervo científico de América Latina se encuentra orientado a conocer cuáles son los factores que atentan sobre el proceso de enseñanza aprendizaje, y es bastante escasa la evidencia sobre qué hacer con lo que hoy en día se encuentra dentro de las aulas y probablemente de muy difícil modificación a corto plazo.

La segregación escolar impacta en el logro académico de los estudiantes analizados en esta investigación y se encuentra en la misma línea de estudios realizados en la región. Los estudiantes agrupados en escuelas homogéneas tienden a presentar resultados más bajos en la misma línea de Oakes (2005) y Oakes et al. (1990). Si bien este estudio no se ha focalizado en el análisis de variables como la autoestima, por lo que este efecto, a diferencia del estudio de Oakes, se refiere exclusivamente al rendimiento escolar. El efecto de las agrupaciones heterogéneas en esta investigación al contrario del estudio de Slavin (1990), constata que los estudiantes no se ven penalizados por la inclusión, pero si se existe un efecto negativo en las agrupaciones homogéneas, en este caso para las agrupaciones de SEC bajo. Siendo nuestros resultados congruentes con los de Robert (2007), en el aspecto de que los estudiantes de SEC alto no se ven perjudicados por la inclusión, encontrando diferencias en comparación al estudio antes mencionado, en que los estudiantes de SEC bajo, si se ven benefician de la integración. Este beneficio, tal vez está relacionado al acceso de educación de calidad, que caracteriza a los ambientes donde los estudiantes de SEC alto asisten.

La evidencia que sustenta el panorama de inequidad en la educación en América Latina es amplia y contundente (Arcidiácono et al., 2014; Benavides et al., 2014; Gasparini et al., 2011; Jaume, 2013; Murillo, 2016; Murillo & Martínez-Garrido, 2017a, 2017b; Nava-Gómez & Pérez-Cervantes, 2015; Treviño et al., 2014; UNESCO-OREALC, 2016a; Webb et al., 2017; Woitschach et al., 2017). Diversas investigaciones se han realizado con el propósito de estimar el nivel de segregación en los distintos países de la Región (Arcidiácono et al., 2014; Benavides et al., 2014; Gasparini et al., 2011; Jaume, 2013; Murillo, 2016; Murillo & Martínez-Garrido, 2017a, 2017b; Nava-Gómez & Pérez-Cervantes, 2015; Treviño et al., 2014; Webb et al., 2017). Estudios de evaluación educativa a gran escala han formulado recomendaciones a los sistemas educativos basados en la composición de las escuelas de la región, denotando la ausencia de un análisis específico del impacto de la conformación de las aulas en el rendimiento académico

de los estudiantes (UNESCO-OREALC, 2016a). Por lo que este estudio se propuso ir un paso más allá e identificar el impacto de las agrupaciones homogéneas y heterogéneas en el rendimiento académico de los estudiantes.

Los resultados, tanto del conjunto de la muestra TERCE, como de la mayoría de los países indican que las aulas como mayor dispersión tienen a presentar resultados más bajos una vez controladas las variables de ajuste. Este resultado es coherente con las evidencias que muestran que desde el punto de vista de la eficacia pura los grupos-aula homogéneos son superiores a los heterogéneos (Kulik & Kulik, 1992). Sin embargo, desde el punto de vista del sistema educativo la eficacia absoluta tiene menos relevancia que la distribución de los resultados. Un sistema educativo difícilmente será de calidad si la distribución del capital cultural no es equitativa (Diem & Boorks, 2013). En ese sentido y en concordancia con otros estudios (Oakes, 2005; Oakes et al., 1990), los datos indican que las agrupaciones heterogéneas benefician a los estudiantes de SEC bajo, sin perjudicar a los estudiantes de SEC alto (Róbert, 2007; Slavin, 1990). Con los datos de la muestra completa de TERCE las agrupaciones heterogéneas permiten reducir a la mitad la diferencia de resultados entre las clases sociales.

REFERENCIAS

Adams-Byers, J., Whitsell, S. S., & Moon, S. M. (2004). Gifted students' perceptions of the academic and social/emotional effects of homogeneous and heterogeneous grouping. *Gifted Child Quarterly, 48*(1), 7-20. doi:10.1177/001698620404800102

Agirdag, O., van Avermaet, P., & van Houtte, M. (2013). School segregation and math achievement: A mixed-method study on the role of self-fulfilling prophecies. *Teachers College Record, 115*(3), 1-50.

Arcidiácono, M., Cruces, G., Gasparini, L., Jaume, D., Serio, M., & Vázquez, E. (2014). *La segregación escolar público-privada en América Latina*. Santiago de Chile: Naciones Unidas.

Balarín, M. (2016). La privatización por defecto y el surgimiento de las escuelas privadas de bajo costo en el Perú. ¿Cuáles son sus consecuencias? *Revista de la Asociación de Sociología de la Educación, 9*(2), 181-196.

Benavides, M., León, J., & Etesse, M. (2014). *Desigualdades educativas y segregación en el sistema educativo peruano. Una mirada comparativa de las pruebas PISA 2000 y 2009* (Vol. Avances de Investigación 15). Lima: Grupo de Análisis para el Desarrollo (GRADE).

Boaler, J., Wiliam, D., & Brown, M. (1998). *Students' experiences of ability grouping disaffection, polarisation and the construction of failure*. Paper presented at the Annual Conference of British Educational Research Association, Northern Ireland. Recuperado de: http://www.leeds.ac.uk/educol/documents/000000789.htm

Bordieu, P., & Passeron, J. C. (1996). *La Reproducción elementos para una teoría del sistema de enseñanza*. México D.F.: Fontamara S.A.

Castro-Aristizabal, G., Castillo-Caicedo, M., & Mendoza-Parra, J. (2016). Principales determinantes en la adquisición de competencias en América Latina: Un análisis multinivel a partir de los resultados en PISA 2012 *Documentos de trabajo Facultad de Ciencias Económicas y Administrativas* (22), 4-31. doi:10.2139/ssrn.2744657

Castro-Aristizabal, G., & Giménez, G. (2017). ¿Por qué los estudiantes de colegios públicos y privados de Costa Rica obtienen distintos resultados académicos? *Revista de la Facultad Latinoamericana de Ciencias Sociales, 25*(49), 195-223. doi:10.18504/pl2549-009-2017

Cervini, R., Dari, N., & Quiroz, S. (2016). Las Determinaciones Socioeconómicas sobre la distribución de los aprendizajes escolares. Los datos del TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 61-79. doi:10.15366/reice2016.14.4.003

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. 2 volumes. Washington, D.C.: Office of Education, U. S. Department of Health, Education, and Welfare, U. S. Government Printing Office. OE-38001; Superintendant of Documents Catalog No. FS 5.238:-38001.

Cueto, S., & Secada, W. (2003). Eficacia escolar en escuelas bilingues en Puno, Perú. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(1), 1-23.

Chang, J. (2011). A case study of the pygmalion effect: Teacher expectations and student achievement. *Canadian Center of Science and Education - International Education Studies, 4*(1), 198-201.

Chiu, M. (2015). Family inequality, school inequalities, and mathematics achievement in 65 countries: microeconomic mechanisms of rent seeking and diminishing marginal returns. *Teachers College Record, 117*(1), 1-32.

Diem, S., & Boorks, J. (2013). Integration was a solution, but integration does not address quality education": A conversation about school desegregation with Dr. Michael A. Middleton. *Teachers College Record, 115*(11), 1-11.

Elacqua, G. (2012). The impact of school choice and public policy on segregation: Evidence from Chile. *International Journal of Educational Development, 32*, 444-453. doi:https://doi.org/10.1016/j.ijedudev.2011.08.003

Fay, R.E. (1989). Theory and application of weighting for variance calculation *JSM Proceedings Survey Research Methods Section* (pp. 212-217). Alexandria, VA: American Statistical Association.

Felouzis, G. (2005). Ethnic segregation and its effects in middle school in France. *Revue française de sociologie, 46,* 3-35. doi:10.3917/rfs.465.0003

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas. *Psicothema, 24*(1), 167-175.

Gasparini, L., Jaume, D., Serio, M., & Vazquez, E. (2011). *La segregación escolar en Argentina* (Vol. Documento de Trabajo Nº 123). Buenos Aires: Centro de estudios distributivos, laborales y sociales.

Gülseli, B., & de Valk, H. (2012). Navigating the school system in Sweden, Belgium, Austria and Germany: School segregation and second generation school trajectories. *Ethnicities, 12*(6), 776-799. doi:10.1177/1468796812450857

Jaume, D. (2013). *Un estudio sobre el incremento de la segregación escolar en Argentina* (Vol. Documento de Trabajo No. 143). Buenos Aires: Centro de estudios distributivos, laborales y sociales.

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.

Kulik, J., & Kulik, C.L. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36*(2), 73-77. doi:10.1177/001698629203600204

Martin, M.O., Mullis, I.V.S., & Foy, P. (2015). Assessment design for PIRLS, PIRLS literacy, and ePIRLS in 2016. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 Assesment*

Framework. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Martin, M.O., Mullis, I.V.S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Recuperado de Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Martin, M.O., Mullis, I.V.S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015* Recuperado de Chestnut Hill, MA: http://timss.bc.edu/publications/timss/2015-methods.html#

Martínez-Garrido, C. (2015). *Investigación sobre enseñanza eficaz. Un estudio multinivel para Iberoamerica.* (Tesis doctoral inédita), Universidad Autónoma de Madrid.

Mislevy, R. J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characterstics from sparse matrix samples of item responses. *Journal of Education Measurement, 29*(2), 131-161. doi:10.1111/j.1745-3984.1992.tb00371.x

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Recuperado de Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Recuperado de Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/international-results/

Murillo, F.J. (2016). Midiendo la segregación escolar en América Latina. Un análsis metodológico utilizando el TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 33-60. doi:10.15366/reice2016.14.4.002

Murillo, F.J., & Martínez-Garrido, C. (2017a). Estimación de la magnitid de la segregación escolar en América Latina. *Magis Revista Internacional de Investigación en Educación, 9*(19), 11-30. doi:10.1590/es0101-73302017167714

Murillo, F.J., & Martínez-Garrido, C. (2017b). Segregación social en las escuelas públicas y privadas en América Latina. *Educação & Sociedade*, 1-24. doi:10.1590/ES0101-73302017167714

Murillo, F.J., & Román, M. (2011). ¿La escuela o la cuna? Evidencias sobre su aportación al rendimiento de los estudiantes de América Latina. Estudio multinivel sobre la estimación de los efectos escolares. *Revista de currículum y formación del profesorado, 15*(3), 27-50.

Nava-Gómez, G., & Pérez-Cervantes, V. (2015). Educación intercultural ¿inclusión o segregación?: Una mirada antropológica y lingüistica a la educación indígena en el estado de México. *Compendio Investigativo de Academia Journals Celaya*, 3927-3934.

Oakes, J. (2005). *Keeping Track: how schools structure inequality* (2d ed. ed.). New Haven: Yale University Press.

Oakes, J., Ormseth, T., Bell, R., & Camp, P. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: RAND Corporation.

Organisation for Economic Co-operation and Development [OECD]. (2009). *PISA Data analysis manual SPSS*® (Second ed.). París: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2013). *PISA 2012 Results: What makes schools successful (Volume IV): Resources, policies and practices*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2014). *PISA 2012 Technical report*. Recuperado de https://www.oecd.org/pisa/pisaproducts/PISA%202012%20Technical%20Report_Chapter%209.pdf

Organisation for Economic Co-operation and Development [OECD]. (2016). *PISA 2015 Results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC]. (2016a). *Recomendaciones de políticas educativas en América Latina en base al TERCE*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC]. (2016b). *Reporte técnico Tercer Estudio Regional Comparativo y Explicativo. TERCE*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [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.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-

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

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2016). *Informe de resultados del Tercer Estudio Regional Comparativo y Explicativo. Factores Asociados*. Santiago de Chile: UNESCO.

Ortíz-Sandoval, L. (2012). Bilingüismo y educación: La diferenciación social de la lengua escolar. *América Latina hoy, 60*, 139-150.

Palardy, G., Rumberger, R., & Butler, T. (2015). The effect of high school socioeconomic, racial, and linguistic segregation on academic performance and school behaviors. *Teachers College Record*, *117*(12), 1-52.

Phillips, K.J.R., Larsen, E.S., & Hausman, C. (2015). School choice & social stratification: How intra-district transfers shift the racial/ethnic and economic composition of schools. *Social Science Research, 51*, 30-50. doi:10.1016/j.ssresearch.2014.12.005

Riedel, A., Schneider, K., Schuchart, C., & Weishaupt, H. (2010). School choice in German primary schools: How binding are schools districts? *Journal for Educational Research Online*, *2*(1), 94-120.

Róbert, P. (2007). *The influence of educational segregation on educational achievement* (Vol. RSCAS 2007/29): European University Institute.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils intellectual development*. New York: Holt, Rinehart and Winston.

Rust, K.F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381-397.

Sanders, W.L., Wright, S. P., & Horn, S.P. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67. doi:10.1023/a:1007999204543

Scheerens, J. (2000). *Improving school effecttiveness*. París: UNESCO: International Institute for Educational Planning.

Sirin, S. (2005). Socioeconomic status and academic achievement: A Meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453. doi:10.3102/00346543075003417

Slavin, R.E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*(3), 471-499. doi:10.2307/1170761

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

Ting, L., & Ronald, L. (2017). School segregation policy and its educational ramifications for internal migrant children in urban China. *Asian Journal of Social Sience Studies, 2*(2), 1-10. doi:10.20849/ajsss.v2i2.158

Treviño, E. (2003). Expectativas de los docentes en aulas con estudiantes indígenas en Bolivia, México y Perú. *Revista Latinoamericana de Estudios Educativos (México), 33*(2), 83-118.

Treviño, E., Valenzuela, J.P., & Villalobos, C. (2014). ¿Se agrupa o se segrega al interior de los establecimientos escolares chilenos? Segregación académica y socioeconómica al

interior de las escuelas. Análisis de su magnitud, principales factores explicativos y efectos. *Centro de Investigación Avanzada en Educación Universidad de Chile*(11), 1-12.

Webb, A., Canales, A., & Becerra, R. (2017). Capítulo IX Las desigualdades invisibilizadas: población indígena y segregación escolar. In I. Irarrázaval, E. Piña, & M. Letelier (Eds.), *propuestas para Chile Concurso de Políticas Públicas 2016* (pp. 279-305). Santiago: C.I.P. - Pontificia Universidad Católica de Chile.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological bulletin, 91*(3), 461-481. doi:10.1037/0033-2909.91.3.461

Woitschach, P. (2016). *Effect of the indigenous language on educational assessments test.* Poster presented to the 10th International Test Commission [ITC] - Improving Policy and Practice: Opportunities and Challenges in an International Context, Vancouver, Canadá.

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

CAPÍTULO IV. Análisis de la Oportunidad de Aprendizaje en el Estudio TERCE de la UNESCO

Fernández, S.; Woitschach, P.; Álvarez-Díaz, M. y Fernández-Alonso, R. (2018). Análisis de la Oportunidad de Aprendizaje en el estudio TERCE de la UNESCO. *Revista de Investigación Educativa*, *36*(2), 509-528. doi: http://dx.doi.org/10.6018/rie.36.2.307831

RESUMEN

El objetivo de este estudio es analizar el efecto que la Oportunidad de Aprendizaje (OTL) tiene en el desempeño de las escuelas de América Latina, considerando los resultados obtenidos en la prueba de Ciencias Naturales del Tercer Estudio Regional Comparativo y Explicativo (TERCE) de las Naciones Unidas para la Educación la Ciencia y la Cultura (UNESCO). El estudio TERCE se realizó en el año 2013 en quince países de América Latina más el Estado de Nuevo León (México). La muestra estuvo compuesta por 61.937 estudiantes, con edad media de 12.42 años (DT=0.94). El 49.6% son mujeres, el 69.4% asiste a un centro público, el 65,8% a un centro urbano y el 18.1% ha repetido un curso académico. Se ajustaron cuatro modelos jerárquicos lineales de intercepto aleatorios de tres niveles (alumno, escuela y país), empleando el programa HLM 7.01. Entre los principales resultados se observa que, una vez descontadas las variables socioeconómicas y el historial de repetición académica, los factores asociados al clima, a las prácticas de aula, y al desarrollo docente, en tanto que variables de OTL, presentan un margen de mejora destacable en el funcionamiento de las escuelas.

Palabras clave: Oportunidad de aprendizaje; eficacia docente; ciencias naturales; modelos jerárquicos lineales.

INTRODUCCIÓN

El concepto de Oportunidad de aprendizaje (OTL), cuyo origen está en el modelo de Carroll (1963) que lo definió como la cantidad de tiempo determinado para aprender, ha sido ampliamente tratado en la investigación educativa por ser considerado un factor clave a la hora de explicar las diferencias de rendimiento del alumnado y su estudio ha modificado en los últimos años la visión sobre los determinantes del rendimiento educativo (McDonnell, 1995). El metanálisis de Scheerens (2017), que incluye estudios realizados entre 1997 y 2009, reporta un tamaño del efecto (*d*) de las variables OTL de 0.44, superior al efecto de factores como la implicación familiar o el liderazgo escolar. Lamain, Scheerens, & Noort (2017) recopilan 51 estudios primarios realizados en las dos últimas décadas en todas las regiones mundiales y contabilizan que el tamaño del efecto de las variables OTL es positivo y significativo en 84 de los 192 efectos analizados, es decir, en el 44% de los casos recopilados.

Las evaluaciones internacionales (v. g., el Estudio Internacional de Tendencias en Matemáticas y Ciencias –TIMSS o el Programa para la Evaluación Internacional de Alumnos –PISA) han mostrado el efecto de las variables OTL sobre el rendimiento. Angell, Kjærnsli y Lie (2006), con datos de TIMSS, destacan que el grado de cobertura de los contenidos curriculares explica entre el 25% y el 50% de la varianza entre los países. Klieme (2016) estima que la medida OTL en TIMSS 2011 incrementa el poder explicativo de las diferencias entre países de entre un 3 y un 6%, y en el caso de PISA 2012 la varianza entre países pasa del 85% al 96% al incluir la OTL como predictor. Mo, Singh y Chang (2013) reanalizando datos de TIMSS 2003 encuentran una asociación positiva entre el resultado en Ciencias y las variables OTL. Por otra parte, Luyten (2017) concluye que el efecto de variables OTL sobre el rendimiento científico-matemático es positivo y alto en los 22 países de PISA 2012 analizados, y relativamente moderado en TIMSS 2011, resultado tal vez ligado a la validez de la medida OTL. Como segunda derivada de las comparaciones internacionales, las variables OTL, además de

explicar diferencias en el rendimiento, pueden ayudar a desarrollar modelos de enseñanza-aprendizaje más adaptados a los curricula nacionales, sin necesidad de remarcar puntajes diferenciales entre realidades socioculturales diversas (Floden, 2002).

Por su parte, los estudios de eficacia escolar en Latinoamérica han señalado que, el efecto positivo de la OTL sobre los resultados escolares se mantiene incluso después de controlar las variables de contexto (Cueto, Guerrero, León, Zapata & Freire 2014; Cueto, León, Ramírez, Guerrero, 2008; Fernández, 2004; Martínez-Garrido, & Murillo, 2016; Murillo, 2007; Murillo, & Hernández-Castilla, 2011; UNESCO-OREALC, & LLECE, 2000, 2010; Velez, Schiefelbein, & Valenzuela, 1994).

Hasta el momento el término OTL se ha tratado como un constructo unidimensional, cuando en realidad está lejos de tener una acepción unívoca. Durante las últimas décadas el significado original (tiempo de aprendizaje) se ha complejizado y ahora se considera un concepto poliédrico y multimensional, que exige el uso de diversos métodos de recogida de información y el manejo de diferentes variables. Para su estimación los estudios con grandes muestras emplean mayoritariamente métodos indirectos (encuestas al alumnado y profesorado), aunque también son posibles procedimientos directos (observación de aula o revisión de las tareas y cuadernos del alumnado). Igualmente las variables empleadas son diversas, si bien las más frecuentes son: currículum impartido y tasas de cobertura del contenido; tiempo de aprendizaje; calidad de la instrucción, de las prácticas de enseñanza y de los recursos; y gestión, disciplina, clima de trabajo y relaciones personales en el aula (Cueto et al., 2008, 2014; OECD, 2013). En lo que se resta se revisirán estudios previos que han manejado variables OTL similares a las disponibles en la base de datos del *Tercer Estudio Regional Comparativo y Explicativo* (TERCE), que será la matriz empleada en el presente análisis. Se trata de variables vinculadas al tiempo de aprendizaje, las características del proceso instructivo, los recursos disponibles y el ambiente de trabajo del aula.

Como ya se señaló, el tiempo de aprendizaje es un elemento canónico del constructo OTL (Fernández, Fernández-Alonso, Arias, Fernández-Raigoso, & Burguera, 2016). Stallings y Knight (2014) estiman que, en un aula óptimamente gestionada al menos el 85% del tiempo se dedica a actividades de enseñanza y aprendizaje y el 15% a tareas administrativas, logísticas y de gestión. Bruns y Luque (2014) encuentran que los países iberoamericanos quedan 20 puntos porcentuales por debajo de ese 85%, lo que se traduce en un día menos de instrucción por semana. Observaciones estandarizadas en el aula han mostrado que la pérdida significativa del tiempo de instrucción es un fenómeno importante en Latinoamérica, siendo el absentismo docente una de las principales causas, ya que la mitad del tiempo perdido se debe a ausencias del profesorado, impuntualidad al inicio o finalización de las clases o a la realización durante las mismas de actividades distintas de las instructivas (Abadzi, 2009; Chaudhury, Hammer, Kremer, Muralidharan, & Rogers, 2006). Los datos anteriores pueden incluso considerarse conservadores, al menos comparados con la estimación de Wolff, Schiefelbein y Valenzuela (1994) que señalaban que en la región se perdía un tercio del calendario lectivo por huelgas, asuetos y ausencias docentes. En todo caso, los resultados disponibles muestran que el tiempo sobre la tarea está positivamente asociado con el rendimiento en pruebas estandarizadas en diferentes contextos (rural o urbano, bilingues o monolingues), edades y materias (Cueto & Secada, 2003; Hernández-Castilla, Murillo, & Martínez-Garrido, 2014; Schuh Moore, DeStefano, & Adelman, 2012). Los estudios del Laboratorio Latinoamericano de la Calidad Educativa (LLECE) señalan que la asistencia y puntualidad docente aumentan el promedio de los estudiantes entre 6 y 34 puntos, según el país y la materia, encontrándose que este efecto significativo se reproduce en todos los países, aún después de descontar el efecto del nivel socioeconómico del alumnado, y similares resultados también se encontraron con muestras españolas (Servicio de Evaluación Educativa del Principado de Asturias, 2017). No obstante, los resultados no son unánimes. Así, Vélez et al. (1994), después de revisar 18 estudios

latinoamericanos sobre el tema, contabilizaron 60 efectos distintos que describen la relación absentismo docente-rendimiento académico de los cuales sólo 18 confirmaban la hipótesis de que a mayor absentismo peores resultados, mientras que en 34 casos la relación, si bien positiva, no era significativa. Carvallo (2006), con datos del estudio del Examen Nacional de Ingreso (EXANI) de México, también señala un resultado atípico, no pudiendo asociar la impuntualidad docente con descensos en los rendimientos escolares.

Un segundo elemento del constructo OTL es el clima de aula, entendido como la atmósfera de orden, buenas relaciones y ambiente de trabajo orientado a la consecución de los objetivos educativos (Scheerens, 2016). La información disponible sobre la relación entre clima de aula y resultados educativos ofrece un panorama consistente y coherente. Marzano, Marzano y Pickering (2003) estiman que el orden de aula tiene un efecto moderado (d = 0.52) sobre el rendimiento académico, si bien con los datos de Korpershoek, Harms, de Boer, van Kuijk, y Doolaard (2016) este efecto se puede calificar de pequeño aunque significativo (g = 0.17). A medio camino entre ambas estimaciones se encuentran Durlak, Weissberg, Dymnicki, Taylor y Schellinger (2011), que obtienen un efecto entre pequeño y moderado (g = 0.27). Los datos de PISA señalan que las diferencias en Matemáticas según la percepción discente del orden de aula son del 40% de la desviación típica para el conjunto de la OCDE, confirmándose esta relación, en mayor o menor medida, en todos los países (Gil Flores, 2014; Gobierno del Principado de Asturias, 2011). En el contexto latinomericano las tres evaluaciones del LLECE apuntan en la misma dirección: en el primer estudio regional (PERCE) el clima de aula predecía diferencias en Matemáticas y Lectura en torno a una desviación típica, mientras que en el segundo (SERCE) las ganancias esperadas eran más moderadas (entre 10 y 60 puntos según la materia y el país). Finalmente, en TERCE el rango de las diferencias en función del clima de aula osciló entre 6 y 18 puntos (UNESCO-OREALC & LLECE, 2016a). De igual modo, existe un buen número de investigaciones que señalan que las aulas con buenas relaciones personales y un clima de trabajo

estimulante, cordial, afectivo y seguro tienden a presentar mejores resultados educativos (Carvallo, 2006; López, 2006; Ramos Ramírez, 2013; Román, 2010; Torres-Fernández, 2008). En esta misma línea argumental se ha encontrado que en las escuelas de bajo rendimiento el clima de aula se caracteriza por la tensión, la escasa participación y las malas relaciones entre docentes y discentes (Hernández-Castilla et al., 2014). Finalmente, algunos datos apuntan a que en las aulas con mejor clima de trabajo los estudiantes también presentan ganancias en factores socioafectivos (motivación, autoconcepto) y mayor satisfacción hacia la escuela (Martínez-Garrido, 2015).

Un tercer elemento que configura el constructo OTL son las prácticas de enseñanza, que incluyen el modo de trabajar los contenidos de aprendizaje, la metodología didáctica, las explicaciones, el tipo, orginización y orientación de las tareas de enseñanza y la evaluación de los aprendizajes. Se ha encontrado una relación positiva entre los resultados educativos y ciertas tareas docentes como establecer objetivos de aprendizaje claros, plantear actividades de alto nivel de exigencia cognitiva, controlar el trabajo y los deberes en el hogar y organizar una evaluación variada y justa (Gobierno del Principado de Asturias, 2011), y análisis complementarios sugieren que el alumnado con niveles de comprensión más bajos se beneficia especialmente cuando asiste a un aula en la que la claridad de las explicaciones del profesorado es valorada positivamente por la mayoría del alumnado, en tanto que el alumnado con mayor nivel de comprensión se beneficia cuando se le enfrenta a actividades de aprendizaje con alta exigencia cognitiva y que demandan reflexión sobre el propio aprendizaje, lo cual señala la necesidad de establecer prácticas de enseñanza diferenciadas y de claro nivel reflexivo e investigador (Fernández et al., 2016). La investigación latinoamericana apunta a conclusiones similares. Si bien los primeros estudios del LLECE no encontraron resultados concluyentes (UNESCO-OREALC & LLECE, 2000, 2010), en cambio TERCE estima que las prácticas docentes predicen ganancias de entre 9 y 36 puntos en las pruebas cognitivas (UNESCO-

OREALC & LLECE, 2016a). Igualmente, Martínez-Garrido (2015) y Román (2010) encuentran efectos positivos de la metodología docente sobre los resultados. En conjunto, los datos asocian ciertas características docentes a buenos resultados, tales como clases bien preparadas, enseñanza estructurada, objetivos claros, colaboración docente, actividades variadas y participativas, estrategias de aprendizaje activas y uso frecuente de la evaluación y seguimiento del progreso del alumnado (Murillo & Román, 2009; Torres-Fernández, 2008; Velez, et al., 1994). Por el contrario, las escuelas tienden a presentar rendimientos más bajos cuando la metodología docente se basa en estrategias de memorización y reproducción de contenidos (Carvallo, 2006; Hernández-Castilla et al., 2014).

El último componente del concepto OTL son los recursos materiales disponibles. Los resultados de la investigación parecen dependientes del contexto en el que se desarrollan los estudios. Así, Gaviria, Martínez-Arias y Castro (2004) han señalado que en los países desarrollados, donde el gasto educativo es razonablemente alto y las diferencias en la calidad y cantidad de los recursos de las escuelas son relativamente bajas, las evidencias sobre la relación recursos-resultados tiende a ser débil, pero cuando estos estudios se realizan en países en vías de desarrollo el efecto de los recursos sobre el rendimiento es más nítido. En el contexto latinoamericano, Velez et al. (1994), después de analizar varios informes de investigación del último cuarto del siglo XX, encontraron una relación positiva entre la disponibilidad de libros de texto y materiales de lectura y el rendimiento en 13 de las 17 investigaciones revisadas, y Gaviria et al. (2004) concluyen que los recursos de la escuela impactan en los resultados educativos, haciendo mención especial a la existencia o no de libros de texto. Por su parte, Murillo y Román (2009), analizando los datos de SERCE, indican que la existencia de recursos didácticos y las instalaciones académicas constituyen el segundo factor escolar con mayor efecto sobre el desempeño de los estudiantes. En una línea similar los datos de TERCE señalan que el simple hecho de disponer de un cuaderno individual de trabajo como recurso de uso habitual en

el aula, se relaciona con un mejor rendimiento en la mitad de la muestra de tercer grado, y que el 17.6% de los estudiantes de este nivel no cuenta con el material de estudio apropiado, así como el 28.9% de los estudiantes del sexto grado, lo que refleja que, a pesar de que el nivel de pobreza ha disminuido en los últimos años, aún se carece de recursos básicos en el aula (UNESCO-OREALC & LLECE, 2016a).

MÉTODO

La estructura jerárquica de la información de las evaluaciones educativas internacionales y la búsqueda de modelos que ajusten las variables contextuales de cada nivel a la variabilidad individual de rendimiento escolar reclama el uso de una metodología que permita analizar la interacción entre factores individuales y sociales.

La finalidad del presente trabajo es analizar el impacto que tiene la oportunidad de aprendizaje (OTL) en los resultados del alumnado latinoamericano de 6º grado de educación en Ciencias Naturales. Este objetivo general se concreta en las siguientes cuestiones:

1. ¿Qué impacto tiene sobre los resultados en Ciencias Naturales la asistencia y metodología docente, el clima de trabajo en el aula y la disponibilidad por parte del alumnado de un recurso básico como el cuaderno de trabajo?
2. Una vez controladas las variables antecedentes del contexto educativo, ¿se sigue manteniendo este hipotético impacto de las variables relacionadas con la oportunidad de aprendizaje?
3. ¿Qué porcentaje de varianza de los resultados educativos explican las variables asociadas a la oportunidad de aprendizaje?

Población y Muestra

La población se definió como el alumnado matriculado en 6º curso de enseñanza obligatoria en el curso 2013 en los 15 países participantes y en el estado de Nuevo León (México). En cada

país la muestra fue seleccionada siguiendo un diseño bietápico por conglomerados y estratificado propio de las evaluaciones internacionales (Joncas & Foy, 2012; OECD, 2009). En la primera etapa los centros (unidades primarias de la muestra) fueron seleccionados con una probabilidad proporcional a su tamaño, y en la segunda etapa se seleccionó un grupo-aula completo de cada centro, obteniéndose una muestra por encima de los 67.000 estudiantes. Del presente estudio se excluyó al alumnado sin información en la prueba de Ciencias Naturales, por lo que la base final quedó compuesta por 61.637 estudiantes escolarizados en 2.955 centros educativos, que representan a una población de prácticamente 9 millones de estudiantes escolarizados en 6º curso en la región. La tabla 1 recoge el número de estudiantes participantes y la población total a la que representan por país.

La media de edad del alumnado es de 12,42 años y la desviación típica 0,94. El 69,4% asiste a un centro público y el 65,8% a un centro urbano; el 49,6% son mujeres y el 81,9% está escolarizado en el curso correspondiente a su edad en tanto que el 18,1% restante ha repetido al menos un curso en el momento de la aplicación de la prueba.

Tabla 1

Datos 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

Instrumentos

En el estudio se emplearon dos tipos de instrumentos: (a) pruebas de conocimientos escolares, a partir de las cuales se construye la variable dependiente del estudio; (b) cuestionarios de contexto para el alumnado, sus familias, el profesorado y las direcciones de los centros, de donde se extraen todas las variables de interés y de ajuste del presente estudio, salvo la relativa al nivel de riqueza de los países. Las pruebas fueron aplicadas en el programa de evaluación TERCE de las Naciones Unidas para la Educación la Ciencia y la Cultura (UNESCO), cuyas bases de datos son de acceso libre para su uso en investigación (UNESCO-OREALC, 2016).

Rendimiento en Ciencias Naturales

El alumnado respondió a una batería de pruebas que evaluaban Lectura, Matemáticas y Ciencias, si bien en el presente estudio se decidió usar los resultados de Ciencias como variable dependiente. La prueba de Ciencias se desarrolló a partir de una tabla de especificaciones organizada en cinco dominios y tres procesos cognitivos (UNESCO-OREALC, 2016) y constaba de 92 ítems, en su mayoría de elección múltiple, agrupados en seis bloques (cuatro bloques de ítems nuevos y dos bloques con ítems de anclaje provenientes de SERCE). Los ítems se distribuyeron en seis cuadernillos siguiendo un diseño matricial (Fernández-Alonso, & Muñiz, 2011), y cada estudiante respondió a un cuadernillo que contenía entre 31 y 33 ítems a resolver en unos 60 minutos de evaluación. Los ítems fueron ajustados al modelo Rasch empleando el programa Winsteps (Linacre, 2005). La puntuación de cada estudiante 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 se estimaron conjugando las respuestas de los estudiantes a los ítems con información proveniente de diferentes covariables que funcionan como factores de imputación,

y fueron expresadas en una escala con media de 700 puntos y desviación típica 100 (UNESCO-OREALC & LLECE, 2016b).

Variables de ajuste

Cuando la variable dependiente es el rendimiento escolar conviene incluir variables de ajuste que eviten sobreestimar los efectos de las variables de interés (Fernández-Alonso, Álvarez-Díaz, Suárez-Álvarez, & Muñiz, 2017). Dentro de los datos disponibles en la matriz de TERCE se han elegido seis variables de control que permiten describir las características sociológicas del alumnado. Se trata de variables tradicionalmente relevantes en la predicción de rendimiento académico (Liu & Whitford, 2011; UNESCO-OREALC & LLECE, 2016a; Woitschach, Fernández-Alonso, Martínez-Arias & Muñiz, 2017). 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).

Variables de interés (OTL)

En total se han manejado cuatro variables OTL, dos de ellas a nivel de alumnado: *Cuaderno* (1= el estudiante tiene cuaderno o libreta para tomar notas en clase) y un índice estandarizado que estima la *Asistencia y Puntualidad Docente,* y que fue construido con las respuestas del alumnado a tres ítems de elección múltiple: el profesorado falta a clases, llega tarde y se va temprano, y cuyas opciones de respuesta eran: 1 = "Nunca o casi nunca"; 2 = "A veces"; 3 = "Siempre o casi siempre". El valor del alfa de Cronbach oscila entre .4 y .7, según el país (UNESCO-OREALC, 2016).

Se han manejado además dos variables, expresadas como índices estandarizados a nivel de centro: el índice *Prácticas para el Desarrollo del Aprendizaje*, elaborado a partir de la respuesta del alumnado a 13 ítems que se refieren a la frecuencia con la que ocurren ciertos hechos en la clase (por ejemplo, el profesorado nos motiva a seguir estudiando, tienen las clases preparadas, etc.). Las opciones de respuesta eran: 1 = "Nunca o casi nunca"; 2 = "A veces"; 3 = "Siempre o casi siempre". El valor del alfa de Cronbach de este índice oscila entre .6 y .9 según el país (OREALC-UNESCO, 2016). Por su parte, el índice *Clima de Aula* se construyó con las respuestas a seis ítems destinados a valorar el nivel de atención del alumnado en el aula, el

respeto entre el alumnado, el número de interrupciones durante las clases o la frecuencia de comportamientos agresivos. El alfa de Cronbach para este índice osciló entre .7 y .9 (OREALC-UNESCO, 2016, p. 303).

Procedimiento de recogida y análisis de datos

La aplicación de las pruebas cognitivas fue realizada por personal experto y externo al centro y se desarrolló en dos jornadas, la primera dedicada a Lectura y Escritura y la segunda a Matemáticas y Ciencias Naturales. La evaluación de cada materia ocupó entre 45 y 60 minutos, con un descanso de 30 minutos; la aplicación de los cuestionarios de contexto del alumnado tuvo una duración de 45 minutos, tras un receso de 15 minutos. El primer día se entregaron los cuestionarios para el centro, profesorado y familias, que se recogieron al final de la segunda jornada. Se tuvieron en cuenta los estándares éticos de la UNESCO, y las familias de los estudiantes seleccionados recibieron una comunicación a través de las direcciones escolares sobre su participación en el estudio.

En el análisis de datos inicialmente se calcularon los estadísticos descriptivos de todas las variables y las correlaciones de Pearson. A continuación, se ajustaron cuatro modelos jerárquico-lineales de intercepto aleatorios y segmentados en tres niveles: alumno, centro y país. La estrategia de modelización fue la siguiente: inicialmente, se ajustó un modelo nulo sin predictores, para comprobar la distribución de la varianza en cada nivel. El segundo modelo incluyó sólo las variables de ajuste, el tercero sólo las variables OTL y el último consideró conjuntamente las variables de los modelos previos. 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). En TERCE la suma total de los pesos de cada país equivale a 5000 estudiantes (peso del nivel 1) escolarizados en 200 escuelas (peso del nivel 2).

El rango de casos perdidos en las variables osciló entre el 2% y el 12%, y para su recuperación se empleó una estrategia en dos pasos. Inicialmente los casos incompletos fueron imputados con la media del sujeto y, a continuación, los datos totalmente perdidos se recuperaron mediante el método iterativo EM con variables auxiliares que implementa el módulo Missing Value Análisis de SPSS 22. Fernández-Alonso, Suárez-Álvarez y Muñiz (2012) encontraron que esta estrategia en dos pasos es la que mejor recupera los datos poblacionales en estudios con tipos de pérdida (no aleatoria) y porcentajes de datos faltantes similares a los registrados en TERCE.

RESULTADOS

La Tabla 2 muestra los estadísticos descriptivos y correlaciones entre las variables. En general, el sentido de las asociaciones es el esperado: las aulas con mejor clima tienden también a presentar mejores puntuaciones en las prácticas educativas (r = .23) y el profesorado con mejores valores en el índice de asistencia docente, también presenta mejores puntuaciones en el índice de prácticas educativas (r = .15). Igualmente, las correlaciones entre las variables independientes y el rendimiento en Ciencias Naturales presentan, en general, la magnitud y dirección esperadas.

Tabla 2

Estadísticos descriptivos y coeficientes de correlación Pearson entre las variables

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Mujer	-															
2	Indígena	-.011**	-														
3	Trabaja	-.120**	.047**	-													
4	Ha repetido	-.075**	.062**	.084**	-												
5	SEC Alumno	.025**	-.095**	-.038**	-.127**	-											
6	Internet	-.019**	-.134**	-.045**	-.155**	.266**	-										
7	Asistencia docente	.022**	-.067**	-.040**	-.090**	.106**	.113**	-									
8	Tener cuaderno	.064**	-.038**	-.032**	-.113**	.116**	.092**	.086**	-								
9	SEC Escuela	-.063**	-.221**	-.042**	-.144**	.296**	.565**	.206**	.127**	-							
10	Escuela privada	.008	-.075**	-.034*	-.146**	.231**	.330**	.113**	.093**	.507**	-						
11	Escuela rural	.069**	.178**	.015	.088**	-.180**	-.480**	-.161**	-.072**	-.646**	-.360**	-					
12	Infraestructura	-.063**	-.161**	-.031*	-.109**	.182**	.481**	.154**	.064**	.769**	.453**	-.604**	-				
13	Prácticas docentes	.003	-.022	-.047**	.035*	.010	-.061**	.147**	.027	-.115**	-.010	.140**	-.170**	-			
14	Clima de aula	.092**	.071**	-.029	-.034*	.013	-.130**	.007	.002	-.182**	.045**	.232**	-.184**	.227**	-		
15	PIB x cápita 2013	.063	-.453**	.172	-.509**	.269	.190	-.523**	.377*	.297	-.312*	-.232	.354*	-.146	-.388*	-	
16	Rendimiento	-.001	-.097**	-.070**	-.212**	.207**	.288**	.269**	.161**	.470**	.287**	-.289**	.427**	.061**	.004	.032	-
	Media	0.496	0.074	0.044	0.201	0.144	0.536	0.008	0.743	0.073	0.146	0.430	0.044	-0.030	-0.011	14,96	703.0
	Desviación Est.	0.500	0.262	0.206	0.401	0.985	0.498	0.943	0.436	1.005	0.353	0.495	0.988	0.994	0.978	4,71	91.5

**p < .01; *p< .05

La Tabla 3 presenta los resultados obtenidos en el ajuste multinivel, con las variables organizadas por cada nivel de análisis. El modelo nulo indica que la mayor parte de la varianza (54%), se sitúa dentro de los centros, en tanto que, prácticamente uno de cada tres puntos de la varianza se ubica en el nivel 2. Finalmente, el 14% de las diferencias se encuentran entre los países.

El modelo I, que incluye todas las variables de ajuste, explica el 20% de la diferencia total y una parte importante de las diferencias entre los centros. Todas las variables de ajuste a nivel de centro son significativas y operan en la dirección esperada. Destaca, en todo caso, el efecto del SEC de centro, ya que por cada punto que aumenta se predice ganancias cercanas al 35% de la desviación típica en la escala de Ciencias. En el nivel individual la variable con más impacto es la repetición de curso: al alumnado repetidor se le predice una pérdida cercana a un cuarto de desviación típica en la escala de Ciencias.

El modelo II, sólo con variables OTL, explica el 5% de la varianza total y casi el 10% de las diferencias entre centros. De nuevo todas las variables funcionan en la dirección esperada. Cabe destacar que en este modelo el índice de prácticas docentes parece tener más fuerza explicativa que el clima de aula.

El Modelo III, que incluye todas las variables, explica prácticamente el 25% de la varianza total y casi la mitad de las diferencias entre los centros. Todas las variables de interés continúan siendo significativas pese a estar controladas por las variables de ajuste y, de hecho, el efecto de las variables de interés apenas pierde fuerza explicativa. Por su parte la significación de las variables de ajuste también funciona en el sentido esperado.

Tabla 3

Modelos Multinivel para predecir el efecto de la oportunidad de aprendizaje en Ciencias Naturales (estudio TERCE)

	Modelo Nulo β	**Modelo I Contexto β(SE)**	**Modelo II: OTL β(SE)**	**Modelo III: Contexto+OTL β(SE)**
Intercepto	696.27 (9.66)***	636.61 (23.14)***	687.79 (9.57)***	640.15 (20.01)***
Nivel 1 (Alumnado)				
Mujer	-	-1.96 (1.96)	-	-3.02 (2.01)
Indígena	-	-0.01 (3.58)	-	1.93 (3.88)
Trabaja	-	-10.73 (6.46)	-	-9.33 (5.66)
Ha repetido	-	-22.51 (5.53)***	-	-20.92 (2.81)***
SEC Alumno	-	5.95 (2.01)**	-	5.66 (1.55)***
Internet/casa	-	8.31 (3.44)*	-	7.48 (2.95)*
Asistencia docente	-	-	12.40 (1.72)***	11.21 (1.12)***
Tener cuaderno	-	-	13.10 (2.97)***	11.04 (2.55)***
Nivel 2 (Centro)				
SEC Escuela	-	34.94 (4.82)***	-	32.93 (3.90)***
Escuela privada	-	14.30 (6.14)*	-	10.09 (5.60)†
Escuela rural	-	21.84 (11.43)†	-	18.28 (5.47)***
Infraestructura	-	7.81 (3.55)**	-	9.42 (3.12)**
Prácticas docentes	-	-	8.23 (3.69)*	10.28 (2.14)***
Clima de aula	-	-	6.14 (2.35)**	6.22 (2.23)**
Nivel 3 (País)				
PIB x Cápita 2013 (mil. $)	-	0.79 (0.77)	-	0.79 (1.12)
Distribución de la varianza				
Dentro del centro	5903.34	5783.13	5786.84	5682.59
Entre los Centros	3560.36	2196.31	3225.10	1999.65
Entre los Países	1465.71	745	1414.25	645.71
Total	10929.41	8724.44	10426.19	8327.95
Porcentaje de varianza explicada				
Dentro del Centro	-	2%	2%	4%
Entre los Centros	-	38%	9%	44%
Entre los Países	-	49%	4%	56%
Total	-	20%	5%	24%

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

Dado que HLM no ofrece coeficientes estandarizados, en ocasiones no es fácil comparar el impacto relativo de las variables expresadas en escalas y métricas diferentes (variables dicotómicas vs. continuas). El gráfico siguiente intenta paliar esta limitación mostrando el rango de ganancia de las variables de interés en función de los coeficientes

estimados en el modelo final. En el caso de la variable dicotómica, los puntos de ganancia sobre el intercepto indican la diferencia entre dos grupos (tener cuaderno o no). En las variables continuas, la longitud de la barra señala la distancia entre los percentiles 10 y 90 de la distribución de frecuencias de cada variable. Los resultados indican que, una vez descontado el efecto de los antecedentes sociodemográficos del alumnado, las familias y el centro, las variables relativas a la OTL presentan un importante margen de mejora. Así, la asistencia a clase del docente y las prácticas de aula predicen ganancias en torno a un 25% de la desviación típica de la escala, en tanto que el disponer de cuaderno y de un clima de trabajo ordenado en el aula suponen, respectivamente, un 16% y 11% de ganancia adicional.

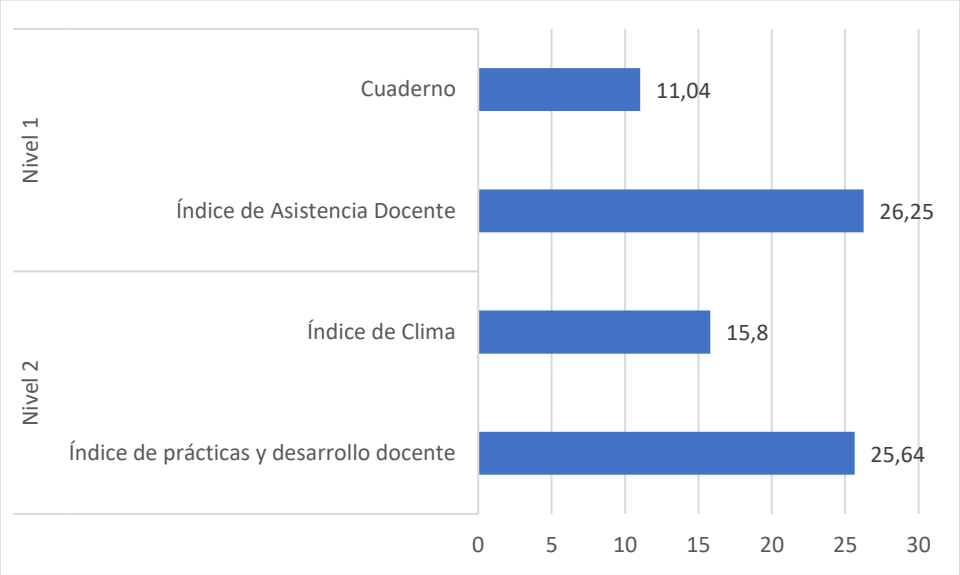


Figura 1. Estimación del efecto de las variables del Modelo del contexto y OTL en el rendimiento en Ciencias Naturales.

DISCUSIÓN Y CONCLUSIONES

Iberoamérica es una de las regiones con mayor desigualdad y donde el tamaño del efecto bruto del centro es bastante alto (Woitschach, et al, 2017), por lo que se espera que las variables de contexto tengan un fuerte impacto en los resultados educativos. Los análisis

realizados con esta muestra representativa de millones de escolares latinoamericanos así parecen confirmarlo. El modelo I mostró que aproximadamente un 40% las diferencias en los resultados de los centros se explican por estos factores de contexto o entrada. Los datos señalan un efecto sobre los resultados en Ciencias vinculados al nivel socioeconómico de las familias, la disponibilidad de internet en el hogar y la repetición de curso. Por su parte, el SEC del centro está fuertemente vinculado a los resultados, ya que el modelo predice diferencias del orden de 0.92 desviaciones típicas entre el alumnado de los centros con mayor y menor nivel SEC. Esto supone prácticamente un nivel de rendimiento completo en la escala TERCE (UNESCO-OREALC, & LLECE, 2016b). Igualmente, otras características del centro, como titularidad, ruralidad e infraestructuras también están asociadas con los resultados en Ciencias.

En relación con el primer objetivo del estudio el modelo II señala que las variables OTL explican cerca de un 10% de las diferencias entre los centros y un 2% de las diferencias entre el alumnado. Son valores más pequeños que los reportados por Angell, Kjærnsli, & Lie (2006), pero se encuentran en la línea de los estudios que destacan el poder explicativo de las variables OTL (Lamain et al., 2017; Luyten, 2017; Mo et al., 2013; Scheerens, 2017).

En todo caso, y respondiendo al segundo objetivo, el modelo III muestra que, aún después de controlar los factores de entrada, las variables de interés mantienen su significación estadística, lo que permite concluir que, una vez descontadas las características sociodemográficas, los factores asociados a la OTL mantienen un margen de mejora importante en el funcionamiento de los centros. Además, se advierte que las variables de interés más destacadas son las vinculadas al profesorado, dato que es coherente con los resultados de estudios previos realizados en diferentes países de la región (Cueto et al., 2008, 2014; Martínez-Garrido, & Murillo, 2016; Murillo, 2007; UNESCO-OREALC,

& LLECE, 2000, 2010; Velez et al., 1994). Así, el coeficiente del índice de Prácticas y desarrollo docente predice que las aulas en las que el profesorado explica con paciencia, anima, felicita, motiva y pregunta con regularidad sobre qué entendió el alumnado, obtendrán en torno a 26 puntos más que las aulas en las que el profesorado puntúa más bajo en este índice. Estos resultados son compatibles con lo señalado por Martínez-Garrido (2015); Torres-Fernández, 2008; UNESCO-OREALC y LLECE (2016a). Por su parte, el índice de Asistencia y puntualidad docente predice ganancias similares puesto que el alumnado cuyo profesorado asiste con regularidad presenta un efecto positivo de 26 puntos en el rendimiento. Estos datos están en línea con las evidencias encontradas por Cueto y Secada (2003); DeStefano, Friedlander, Adelman, & Schuh Moore (2010); Fernández et al. (2016); Hernández-Castilla et al. (2014); Schuh Moore et al. (2012), que señalan la importancia del tiempo efectivo de aula.

Entre las variables de OTL más relevantes se encuentra el Clima de aula que permite aumentar 6 puntos el rendimiento de los estudiantes por cada punto de incremento en el índice, llegando a conseguir las aulas con mejor clima 16 puntos más en su rendimiento. De nuevo los resultados son consistentes con la evidencia disponible sobre la importancia de la atmósfera de trabajo en el aula (Murillo & Román, 2009; Ramos Ramírez, 2013; Román, 2010; Torres-Fernández, 2008; UNESCO-OREALC & LLECE, 2016a).

Finalmente, disponer de un simple cuaderno de clase, parece una buena aproximación para una estimación general de los recursos disponibles, puesto que se predicen ganancias en torno al 11% de la desviación típica después de descontar el efecto de las variables de contexto. Estos datos son coherentes con la evidencia previa que señala que los recursos escolares impactan en los resultados, al menos en los estudios realizados

en países en vías de desarrollo (Gaviria et al., 2004; Murillo & Román, 2009; UNESCO-OREALC & LLECE, 2016a; Velez et al., 1994).

Para dar cuenta del tercer objetivo del estudio se comparan los porcentajes de varianza explicados por los modelos I y III, ya que en este último se introducen las variables de interés, una vez que en primero sólo se consideraron las variables de contexto. En este caso, la ganancia en el porcentaje de varianza entre centros es de 6 puntos porcentuales (38% en el Modelo I frente al 44% en el modelo III), lo que puede parecer un efecto modesto, pero que es coincidente con los datos observados en TIMSS (Luyten, 2017). En todo caso la comparación de la varianza explicada por los modelos de ajuste y final permite extraer dos conclusiones generales: en primer lugar, la ecuación que analice la relación OTL-Rendimiento debe incluirse varibles de ajuste para protegerse de la sobreestimación del efecto de las variables de interés (Fernández-Alonso, Suárez-Álvarez & Muñiz, 2016). La segunda lectura tiene que ver con el tamaño del efecto encontrado. Los datos señalan una pequeña ganancia entre los modelos I y III, que unido al hecho de que las variables de proceso suelen tener mayor impacto en las materias científico-matemáticas que en las áreas de lecto-escritura, donde los resultados parecen estar más influenciados por el contexto familiar del estudiante (Woitschach et al., 2017), podría lleva a concluir que el impacto de las variables OTL es muy modesto. Sin embargo, no conviene olvidar que pequeños efectos sostenidos a lo largo del tiempo pueden marcar diferencias importantes (Prentice, & Miller, 1992), lo que ratifica la relevancia de los resultados encontrados.

Finalmente, el estudio tiene algunas limitaciones que conviene tener presente a la hora de interpretar los datos. Las variables OTL manejadas se extrajeron de encuentas de opinión, método que tiene sus limitaciones ya que las respuestas de los informantes pueden estar sometidas a los sesgos propios de la percepción sujetiva y la deseabilidad social (Husen, 1967; Luken, 2017). Además, en los datos disponibles faltan variables

clásicamente vinculadas al constructo OTL, como la cantidad de programa impartido o el tiempo efectivo de aprendizaje, así como registros y observaciónes sobre la calidad del proceso instructivo, más allá de las atribuciones realizadas por estudiantes y profesorado (Kurtz, 2013; Elliot, 2015). Tampoco se dispone de una medida de los conocimientos previos del alumnado (más allá de la condición de repetición incluída en este estudio), lo que hubiera permitido controlar el efecto del rendimiento previo del alumnado y de las diferencias en el curriculum en función de esos niveles previos (Floden, 2002), y que probablemente sea la razón del pequeño porcentaje de varianza explicada en el nivel 1. Por último, señalar que los modelos empleados son correlacionales por lo que en un futuro sería necesario emplear modelos causales que permitan delimitar la relevancia predictiva de los diferentes componentes del indicador OTL.

REFERENCIAS

Abadzi, H. (2009). Instructional time loss in developing countries: concepts, measurement, and implications. *World Bank Research Observer 24*(2), 267-290. doi;10.1093/wbro/lkp008

Angell, C., Kjærnsli, M., & Lie, S. (2006). Curricular effects in patterns of student responses to TIMSS science items. En S. J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science* (pp. 277–290). London: Routledge.

Bruns & Luque (2014). *Great Teachers How to raise student learning in Latin America and the Caribbean*. Washington DC: International bank for recosntruction and development/ The World Bank. doi:10.1596/ 978-1-4648-0151-8.

Carvallo, M. (2006). Factores que afectan el desempeño de los alumnos mexicanos en edad de educación secundaria: un estudio dentro de la corriente de eficacia

escolar. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 4*(3), 30-53.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 723-733.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives 20*(1), 91-116.

Cueto, S., Guerrero, G., Leon, J., Zapata, M. & Freire, S. (2014) The relationship between socioeconomic status at age one, opportunities to learn and achievement in mathematics in fourth grade in Peru, *Oxford Review of Education, 40*(1): 50-72, doi: 10.1080/03054985.2013.873525

Cueto, S., León, J., Ramírez, C., & Guerrero, G. (2008): Oportunidades de aprendizaje y rendimiento escolar en matemática y lenguaje: resumen de tres estudios en Perú. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *6*(1), 29-41

Cueto, S., & Secada, W. (2003). Eficacia escolar en escuelas bilingues en Puno, Perú. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(1), 1-23. Recuperado de: https://goo.gl/2qrMaa

DeStefano, J., Friedlander, E., Adelman, E., & Schuh Moore, A. (2010). *Using opportunity to learn and early grade reading fluency to measure school effectiveness: school quality in Nepal*. USAID/ EQUIP2, Washington DC: FHI 360.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: a meta-

analysis of school based universal interventions. *Child Development, 82*, 405-432. doi:10.1111/j.1467-8624.2010.01564.x

Elliott, S. N. (2015). Measuring opportunity to learn and achievement growth: Key research issues with implications for the effective education of all students. *Remedial and Special Education, 36 (1)*, 58-64.

Fernández, T. (2004). De las escuelas eficaces a las reformas educativas de segunda generación. *Estudios Sociológicos, 22*(65), 377-408.

Fernández, S., Fernández-Alonso, R., Arias, J. M, Fernández-Raigoso, M., & Burguera, J. L. (2016). Oportunidad de aprendizaje y eficacia docente. Análisis exploratorio de factores asociados. *Bordón, 68*(4), 49-65. doi.org/10.13042/Bordon.2016.38075

Fernández-Alonso, R., Álvarez-Díaz, M., Suárez-Álvarez, J., & Muñiz J. (2017). Students' achievement and homework assignment strategies. *Frontiers in Psychology, 8*, 286. doi: 10.3389/fpsyg.2017.00286

Fernández-Alonso, R. & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las Competencias Básicas. *Aula Abierta, 39*(2), 3-34.

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas. *Psicothema, 24*(1), 167-175.

Fernández-Alonso, R., Suarez-Álvarez, J., & Muñiz, J. (2016). Homework and performance in mathematics: the role of the teacher, the family and the student´s background. *Revista de Psicodidáctica, 21*(1), 5-23. doi:10.1387/RevPsicodidact.13939

Floden, R. E. (2002). The measurement of opportunity to learn. En A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement.* Washington: NRC.

Gaviria, J. L., Martínez-Arias, R., & Castro, M. (2004). Un estudio multinivel sobre los factores de eficacia escolar en países en desarrollo: El caso de los recursos en Brasil. *Education Policy Analysis Archives, 12*(20). Recuperado de: https://goo.gl/BhZRZL

Gil Flores, J. (2014). Factores asociados a la brecha regional del rendimiento español en la evaluación PISA. *Revista de Investigación Educativa, 32*(2), 393-410. doi: 10.6018/rie.32.2.192441

Gobierno del Principado de Asturias (2011). *Evaluación de Diagnóstico Asturias 2010*. Oviedo: Consejería de Educación y Ciencia. Recuperado de: https://goo.gl/KAv8hN

Hernández-Castilla, R., Murillo, F. J., & Martínez-Garrido, C. (2014). Factores de ineficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 12*(1), 103-118. Recuperado de: https://goo.gl/C4B8ri

Husen, T. (1967*). International study of achievement in mathematics: a comparison of twelve countries*. (Vol. III). Nueva York: Wiley & Sons.

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. En 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.

Klieme, E. (2016). TIMSS 2015 and PISA 2015 How are they related on the country level? *Deutsches Institut für Internationale Pädagogische Forschung*. Recuperado de: https://goo.gl/1cS1RS

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research, 86,* 643-680. doi: 10.3102/0034654315626799

Kurtz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19(1),* 159-184

Lamain, M., Scheerens, J., & Noort, P. (2017). Review and "vote count" analysis of OTL-effect studies. En J. Scheerens (Ed.), *Opportunity to learn, curriculum alignment and test preparation a research review* (pp. 55-101). Switzerland: Springer International Publishing.

Linacre, J. M. (2005). *A User's Guide to WINSTEPS/MINISTEP: Rasch-Model* Computer Programs (Version 3.55). Chicago: MESA Press.

López, M. (2006). Todo el que llega aquí se contagia: el éxito escolar. En F. J. Murillo (Ed.), *Estudios sobre eficacia escolar en Iberoamérica. 15 buenas investigaciones* (pp. 261-286). Bogotá: Convenio Andrés Bello.

Liu, X., & Whitford, M. (2011). Opportunities-to-learn at home: Profiles of students with and without reaching science proficiency. *Journal of Science Education and Technology, 20*(4), 375-387. doi:10.1007/s10956-010-9259-y

Luyten, H. (2017). Predictive power of OTL measures in TIMSS and PISA. In J. Scheerens (Ed.), *Opportunity to learn, curriculum alignment and test preparation a research review* (pp. 103-119). Switzerland: Springer. doi: 10.1007/978-3-319-43110-9_5

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305-322.

Martínez-Garrido, C. (2015). *Investigación sobre enseñanza eficaz. Un estudio multinivel para Iberoamerica*. (Tesis inédita de doctorado). Universidad Autónoma de Madrid, España.

Martínez-Garrido, C., & Murillo, F. J. (2016). Investigación Iberoamericana sobre enseñanza eficaz. *Revista Mexicana de Investigación Educativa*, *21*(69), 471-499.

Marzano, R. J., Marzano, J. S., & Pickering, D. J. (2003). *Classroom management that works. Research-based strategies for every teacher*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD). doi: https://goo.gl/ngMUYR

Mislevy, R. J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characterstics from sparse matrix samples of item responses. *Journal of Education Measurement, 29*(2), 131-161. doi:10.1111/j.1745-3984.1992.tb00371.x

Mo, Y., Singh, K. & Chang, M. (2013) Opportunity to learn and student engagement: a HLM study on eighth grade science achievement. *Educational Research for Policy and Practice 12*(1), 3-19.doi:10.1007/s10671-011-9126-5

Murillo, F. J. (2007). School effectiveness research in Latin America, En T. Townsend (ed.), *International handbook of school effectiveness and improvement*, Nueva York: Springer, 75-92.

Murillo, F. J., & Hernández-Castilla, R. (2011). Factores escolares asociados al desarrollo socioafectivo en Iberoamérica. *Revista Electrónica de Investigación y Evaluación Educativa*, *17*(2). 1-23.

Murillo, F. J., & Román, M. (2009). Mejorar el desempeño de los estudiantes de América Latina. *Revista Mexicana de Investigación Educativa, 14*(41), 451-484.

OECD. (2009). *PISA data analysis manual SPSS®* (2ª edition). París: OECD Publishing.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* París: OECD Publishing. doi:10.1787/9789264190511-en

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160-164. doi:/10.1037/0033-2909.112.1.160

Ramos-Ramírez, G. (2013). *La investigación sobre eficacia escolar en el Salvador. Estudio retrospectivo y prospectivo.* (Tesis doctoral inédita), Universidad Autónoma de Madrid, Madrid.

Raudenbush, S., Bryk, A., Cheong, Y. K., Congdon, R., & du Toit, M. (2011). *HLM 7 hierarchical linear and nonlinear modeling*. Chicago: SSI Scientific Software International, Inc.

Román, M. (2010). Investigación latinoamericana sobre enseñanza eficaz. *Revista Educación y Ciudad*, 19, 81-96.

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness a critical review of the knowledge base*: Netherlands: Springer doi: 10.1007/978-94-017-7459-8_1

Scheerens, J. (2017). *Opportunity to learn, curriculum alignment and test preparation a research review* (pp. 23-53). Switzerland: Springer International Publishing.

Scheerens, J. (2017). Meta-analyses and descriptions of illustrative studies. En J. Scheerens (Ed.), *Opportunity to learn, curriculum alignment and test preparation. A Research Review* (pp. 23-53). Switzerland: Springer International Publishing.

Schuh Moore, A., DeStefano, J., & Adelman, E. (2012). *Opportunity to Learn as a measure of school effectiveness in Guatemala, Honduras, Ethiopia, and Nepal.* USAID/ EQUIP2, Washington DC: FHI 360

Servicio de Evaluación Educativa del Principado de Asturias (2017). ¿Cuánto importa el orden del aula en los resultados educativos? *Informes de Evaluación, 7*. doi: 10.13140/RG.2.2.21639.29608

Stallings, J., Knight, S. & Markham, D. (2014). *Using the stallings observation system to investigate time on task in four countries.* Washington, DC: World Bank.

Torres-Fernández, P. (2008). La investigación Iberoamericana de eficacia escolar ¿Qué nos dejó a los cubanos? *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 6*(4), 81-97.

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.

UNESCO-OREALC, & LLECE. (2016a). *Informe de resultados del tercer estudio regional comparativo y explicativo. Factores asociados*. Santiago de Chile: UNESCO.

UNESCO-OREALC, & LLECE. (2016b). *Informe de resultados del tercer estudio regional comparativo y explicativo. Logros de aprendizaje*. Santiago de Chile: UNESCO.

Velez, E., Schiefelbein, E., & Valenzuela, J. (1994). Factores que afectan al rendimiento académico en la Educación Primaria. Revisión de la Literatura en América Latina y el Caribe. *Revista Latinoamericana de Innovaciones Educativas, 17*, 29-53.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? En M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assesments 2*, 9-36. Princeton, NJ: Educational Testing Service.

Wolff, L., Schiefelbein, E., & Valenzuela, J. (1994). *Mejoramiento de la calidad de la educación primaria en América Latina y el Caribe*. Washington, D.C: Banco Mundial.

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

CAPÍTULO V. An Ecological View of Measurement: Focused on Multilevel Model Explanation of Gender Differential Item Functioning

ABSTRACT

UNESCO's Third Regional Comparative and Explanatory Study (TERCE) program reports on the results for 15 Latin American countries and the State of Nuevo León in Mexico. The Third Regional Comparative and Explanatory Study of UNESCO analyses and compares the academic results in mathematics, sciences, and reading in 15 countries of Latin America. The aims were to investigate ecological explanations of gender differential item functioning. Validity is the foundation of a testing procedure, and the process of validating is key to the overall success of the educative assessment as a whole. This study deals specifically with the position of an ecological point of view which includes and situates the person, process, context, and time of the testing situation. These descriptions pinpointed specific incidents of how and what variables at the individual, school, or country level can give a deep understanding of the response process in Latin America countries. The present study analysed the item of the science test applied in 2013 to 6th grade students and the data pool consists of 9,689 students from in 2,663 schools. A progressive inclusion of the variance distribution in different Bernoulli logistic regression models has been carried out. The estimation used was the penalized quasi-likelihood (PQL) for a Bernoulli distribution of the outcome. In summary, the main results have shown the presence of DIF in 32% of the sciences test booklet one, and that the main sources of gender DIF were related to the human development level of the countries participants in TERCE 2013.

Key words: educative standardized evaluation, DIF, validity, hierarchical linear model, UNESCO.

INTRODUCTION

It is fundamental that there is robust evidence of validity that supports test score interpretations and uses in educational assessments. The greater the impact of test score social consequences, the higher the level of validity evidence is required to support the interpretations and uses. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) states that the major purposes for educational testing are to inform decisions about test takers, as well as to make inferences about their results and the teaching-learning process. The current uses of educational testing results, however, go beyond those purposes, especially in terms of their global significance. In addition to comparisons between countries, the results from these international educative assessments are mostly used for supervision, intervention, innovation or changes in all levels of educational policies.

Throughout the last century, the conceptualization of validity and validation have evolved through the theories and the strategies to discover and support the inferences, and through the policy implications of the evaluation process. The last version of the Standards refers to validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test. Meanwhile, validation is defined as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use (AERA et al., 2014, p. 11). According to AERA et al. (2014), validity is viewed as a holistic or integrated concept that includes evidence from the test content, the response processes, the internal structure, the relations among and with other variables, and the social consequence of testing. In conjunction, these sources of validity evidence are synthesized on three different sets of standard procedures such as establishing intended uses and

interpretations, the uses regarding samples and setting used in validation, and finally the specific forms of validity evidence.

Recently, Gomez Benito, Sireci, Padilla, Hidalgo, and Benitez (2018) proposed a conceptual strategy that transforms differential item functioning (hereafter referred to as DIF) in an integrated validation study for all sources of evidence (instead of only evidence of validity in the internal structure). Although the Gomez Benito et al. pointed out their DIF validation studies proposal can be extended to educational testing, the mixed methods framework proposed did not address the complete scenario of the testing situation factors.

The Standards states that DIF occurs when diverse groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item (AERA et al., 2014, p. 16). Positioned from Zumbo's descriptions of the third generation, DIF is an integrated and ecological view of testing procedures in which the person does not exist as an isolated unit, and DIF analysis is focused on the sources of contextual and holistic explanations than on individuals. Beginning as early as 2005, Zumbo and Gelin (2005) recognized the intrinsic value of the contextual contribution to the overall response process. This early work is the precursor to the ecology of the item responding framework, which in educational assessments can include items and test characteristics, individual, classroom or school characteristics, and country factors. More recently, evidence of the impact of country characteristics can be seen in Chen and Zumbo (2017), using two-level logistic regression model with PISA data set.

Up until now, the evidence of DIF from a holistic point of view that is based on multilevel analysis includes the information of the students at the individual level and item characteristics at the nested level (Balluerka, Gorostiaga, Gomez-Benito, & Hidalgo,

2010; Balluerka, Plewis, Gorostiaga, & Padilla, 2014; Swanson, Clauser, Case, Nungester, & Featherman, 2002). Given that DIF usually occurs in the context of observational rather than experimental studies, especially in educational assessments, the practice of including contextual information can address not only the sources of DIF evidence but also move towards an ecological, and even a more scientific, explanation of the item response process. Multilevel regression models can therefore expand the knowledge of DIF causes, specifying a DIF parameter that varies randomly over items and testing hypotheses on sources of DIF shared by the school and country bundles. Thus, the objective of this research is to identify the underlying explanations of gender differential item functioning in international assessments using multilevel regression models.

Generalized Linear Mixed Model

Generalized linear mixed model (GLMM) or hierarchical generalized linear mixed model (HGLMM) belongs to a general family of mixed effects models, which can be used for continuous, binary, ordinal, categorical, nominal, categorical, and variable dependent, it is including both random and fixed effect in the analysis. When the variable of interest is binary, where usually zero means an incorrect answer and one is equal to a correct answer, the distribution must be considered from a binomial viewing. Given the predicted value of the outcome, the level 1 random effect can take on only one of two values, and therefore cannot be normally distributed. Thus, the level 1 random effect cannot have homogeneous variance. Instead, the variance of this random effect depends on the predicted value as specified below (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011, p. 104).

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right).$$

In other words, η_{ij} is the log of the odds of success. Thus, if the probability of success, ϕ_{ij}, is 0.5, the odds of success is 1.0 and the log-odds or logit is zero. When the probability of success is less than 0.5, the odds are less than one and the logit is negative; when the probability is greater than 0.5, the odds are greater than unity and the logit is positive. Thus, while ϕ_{ij} is constrained to be in the interval (0,1), η_{ij} can take on any real value. The level 1 model can be expressed by the next equation (1):

$$\text{logit}\left[\text{Prob}\left(Y_{ijk}=1\right)\right]=\pi_{0jk}+\pi_{1jk}*\text{ability}+\pi_{2jk}*\text{grouping},\ (1)$$

where Y_{ijk} is the binary response/probability of success for a test taker i, from the school j and country k. The level - 1 intercept expresses π_{0jk} as a function of random intercept at level - 2 β_{00k} plus the level - 1 residual error term r_{0jk} and the random intercept at level - 3 γ_{000} plus the level -2 residual error term $u_{00k.}$ The level -1 intercept is a function of the grand mean units at level - 2 and level - 3. If the clustered structure is omitted or not taken into account, then the data may lead to misleading results and incorrect conclusions. The linear mixed regression model allows a random intercept (i.e., each cluster has a different intercept), and a random slope (i.e., each cluster has a different slope).

Prob (IT1_19$_{ijk}$=1|π_{jk}) = ϕ_{ijk}

log [ϕ_{ijk}/ (1 - ϕ_{ijk})] = η_{ijk}

$\eta_{ijk} = \pi_{0jk} + \pi_{1jk}*(\text{ability}_{jk}) + \pi_{2jk}*(\text{grouping}_{ijk})$

Level-2 model

$\pi_{0jk} = \beta_{00k} + r_{0jk}$

$\pi_{1jk} = \beta_{10k} + r_{1jk}$

$\pi_{2jk} = \beta_{20k} + r_{2jk}$

Level-3 model

$\beta_{00k} = \gamma_{000} + u_{00k}$

$\beta_{10k} = \gamma_{100} + u_{10k}$

$\beta_{20k} = \gamma_{200} + u_{20k}$

Combined equation

$$\eta_{ijk} = \gamma_{000} + \gamma_{100}*\text{ability} + \gamma_{200}*\text{grouping}_{ijk} + r_{0jk} + r_{1jk}*\text{ability}_{ijk} + r_{2jk}*\text{grouping}_{ijk} + u_{00k} + u_{10k}*\text{ability}_{ijk} + u_{20k}*\text{grouping}_{ijk} \quad (2)$$

Above is the mixed model in which the first right side of the equation is the fixed effect and the left second part of the equation is the random term (Equation 2). Random effects are represented as random variables in an LMM; therefore, a random effect has a distribution with an error term, which allows one to generalize the results to a population with a defined probability distribution. β_{00k} are the means of the level 1 regression coefficients, r_{2jk} are random variables that represent unexplained variability across schools, γ_{000} are the means of the level 1 regression coefficients and u_{20k} are the random variables that represent unexplained variability across countries. Random intercepts represent random deviations for a given cluster or subject from the overall fixed intercept. Random slopes represent random deviation for a given cluster or a subject from the overall fixed effects (slopes). Random effects are random values associated with random factors, contain measurement errors, and vary from sample to sample.

METHOD

Sample

The data pool used was from the science test which consists of 9,692 students (49.3% students identified as a girl and 50.7% identified as a boy) and 2,663 schools participating in the Third Regional Comparative and Explanatory Study (TERCE) conducted in 2013 in 15 countries in Latin America. The objective of TERCE was to evaluate the knowledge of 6th-grade students. The sample design has been stratified by conglomerates, with a random and systematic selection in two stages. In these designs, the sampling units (schools, classrooms and students) are selected in two or more stages, and these sample units do not have the same probability of being chosen.

Table 1
Students and Schools Sample Distributions

Countries	Student sample	School sample
Argentina	631	194
Brazil	506	122
Chile	840	187
Colombia	718	149
Costa Rica	590	185
Dominican Republic	612	164
Ecuador	807	182
Guatemala	677	173
Honduras	655	193
Mexico	600	161
Nicaragua	634	171
Panama	605	181
Paraguay	539	182
Peru	807	261
Uruguay	471	158
Total	9692	2663

Source TERCE Technical Report (UNESCO-OREALC, 2016, p.26)

Measurement

Description of the 6th-grade science test

TERCE evaluates three cognitive processes (recognition of information and concepts, understanding and application of concepts, and scientific thinking and problem solving), and five domains of knowledge (health, living beings, environment, the earth and the solar system, and matter and energy). The items were composed of multiple response options and constructed responses; the final data set has the responses coded as binary items (UNESCO-OREALC, 2016). Moreover, the science test is composed of 92 items that were distributed in six blocks or clusters. These blocks were distributed in six different booklet models by an incomplete block design. Each booklet is made up of two blocks or clusters of items between 26 and 30 items totally, and each cluster appears twice throughout the collection of booklets, once at the beginning and once at the second position of the booklet. Not-missing values were reported as TERCE provided complete data sets

The variable focus of the analysis was extracted from booklet number one of the 6th-grade science test as follows:[6]

(a) Dependent variable: Item 19 from the science booklet. It is important to denote that TERCE items were presented to the students in a multiple-choice format, but that information is not available to researchers due to TERCE that recoded responses in binary format in the open access dataset. In the current item 19, the coding 0 represents an incorrect answer and 1 represent a correct answer (mean 0.54 and SD=.498). The distribution of the sample by gender is as follows.

[6] The variables included in the analysis are from UNESCO data collection are available online for researchers who are interested in educational data sets (UNESCO-OREALC, 2016).

Table 2
Gender Distributions on item 19

		Gender		
		Girls	Boys	Total
IT1_19	0	2314	2116	4430
	1	2469	2793	5262
Total		4783	4909	9692

The predictors included at the student level were extracted from TERCE student data set.

(a) Sciences ability: Mean of five plausible values for every student in the science test. Thus, the complete data set presents a mean of 703.7038 (SD=90.24). For a better comparison, that variable was standardized to the region in a normal distribution with mean 0 and standard deviation 1.

(b) Gender recoded as a 0 for girls and 1 for boys.

The predictors included at the school level were extracted from TERCE school principal and family data set.

(a) School SES: Index of socioeconomic and cultural status standardized to the region, which is a continuous variable with a mean of 0.28 (SD=1.05), with a minimum value of -2.41 and maximum of 3.27. The index includes information from 17 items about mother education and house services, resources, and infrastructure (Alpha de Cronbach ± .08 between countries).

(b) School physical resources: Index of the school infrastructure standardized to the region, which is a continuous variable with mean 0.29 (SD=1.03), with a minimum of -2.37 and a maximum of 2.86. The index includes information from 19 items about services, resources, and school physical infrastructure (Alpha de Cronbach ± .07 between countries).

The predictors included at the country level were extracted from the Human Development Report 2013 of the United Nations Development Programme (UNDP, 2013).

(a) Gender inequality index (GII): GII is an index measuring gender disparity. It ranges from 0, which indicates that women and men perform equally, to 1, which indicates that women have the poorest opportunities in all measured dimensions.

(a) Human development index (HDI): HDI is a composite index of life expectancy, education, and average income. It ranges from 0 to 1. A nation scores higher on HDI when its population has a longer life expectancy at birth, longer period of education, and higher average income.

Analysis

Given the multilevel nature of the TERCE data, a gradual inclusion of the variance distribution in different Bernoulli logistic regression models has been carried out. First, we processed the analysis including a two-level model (student-school; student-country), next we tested a three-level model analysis including the student, school, and country information. PQL was the type of estimation applied, which involves the use of a standard HLM model with the introduction of appropriate weighting at level 1. However, after this standard HLM analysis has converged, the linearized dependent variable and the weights must be recomputed. Then, the standard HLM analysis is recomputed. This iterative process of analyses and re-computing weights and linearized dependent variable continues until estimates converge (Raudenbush et al., 2011). All variables in two level models were centered at the group mean (Enders & Tofighi, 2007) and in the case of three level models all variables were centered following Brincks et al.'s (2017) strategy; which implies the use of grand mean centered in order to preserve two sources of variability: within-country, between -school variability and between country variability. The study

included senatorial weights from students and school in all the analysis carried out (UNESCO-OREALC, 2016).

Based on the research goals, the analysis carries out a consecutive order of steps. First, we identify gender DIF using two- and three-level binary (Bernoulli) logistic regression models for every item of the booklet. Equation one was used including only level 1 predictors (ability and gender). There were two goals to be accomplished in this step: the first was to identify the gender DIF in the country average, and the second was to discover significant variability in the random gender slope, which exemplifies not only the presence of gender DIF but also the variability across countries.

The second step in the analysis was to run a Bernoulli logistic regression model, treating the data in two- and three-level hierarchical modelling. Each analysis included in level 1 the same variables as equation 1 and controlling their effects by adding different predictors in each subsequent level. All the variables at student level were left constant in all models, and each predictor at level 2 and 3 was included separately based on the complexity of the model and to avoid the collinearity (considering the sample size at a higher level of fifteen countries). Given this same information, Browne and Draper (2006) were able to obtain unbiased variance components with REML estimation with only 6 units at the highest level for a simple model.

1. Two-level Bernoulli logistic regression models including the student level at level-1 and school grouping at level-2.

Predictors included at level-1: Ability in sciences and gender. Both variables have been centred around the group mean.

Predictors included at the random slope of gender in level-2: School SES and school infrastructure index.

2. Two level Bernoulli logistic regression models including the student level at level-1 and country grouping at level-2.

Predictors included at level 1: Ability in sciences and gender. Both variables have been centred around the group mean.

Predictors included at the random slope of gender in level-2: Gender inequality index and human development index.

3. Three level Bernoulli logistic regression models including the student level at level-1 and school grouping at level-2, and 15 countries at level-3.

Predictors included at level-1: Ability in sciences and gender. Both variables have been centred around the grand mean.

Predictors included at the random slope of gender in level-2: School SES and school infrastructure index.

Predictors included at the random slope of gender in level-3: Gender inequality index and human development index.

RESULTS

An exhaustive analysis of every item in the booklet one was carried out. Nine items were flagged with significant ($p<.05$) coefficients for gender DIF in science booklet number one – which corresponds to 32% of this booklet. Given that DIF distribution in those nine items, girls are more likely to endorse a correct answer in four items, and boys in five of the items. Taking into account our research goals, the item that presented a significant variability in the random slope of gender was considered as a second criterion and selected for the following analysis. Four of nine items with DIF were flagged with a

significant coefficient in gender DIF as well as a significant variability between countries. In broad terms, our first approach has shown the presence of gender DIF in at least 32% of the binary items in booklet number one. Moreover, in consideration of DIF notably, that presence is homogeneous between countries in around five of the items, regardless of some variations between countries in four items flagged with DIF.

The next step further analyzed the association between the presence of gender DIF and other predictors. Consequently, for the following steps, the item number 19 was included in all the models, taking into consideration the complexity of the models and the methodological goal of this research. Firstly, from the perspective of a non-nested structure, a Chi-Square test was applied to discover the association between the responses' distribution of item 19 and gender, showing a significant association between those variables ($\chi^2 = 27.166, \mathrm{p} < .000$). Secondly, taking into account a nested structure of the data, different models were performed. Even though all the models will be explained in the subsequent pages, a brief description of our model zero (gender DIF) for all the levels analyzed is presented in Table 3.

Table 3
Summary of the DIF Results Presented by Multilevel Models

		Fixed Effect Gender (gamma γ_{20})		Random Effect Gender				
Item 19		Coefficients	odds ratio		SD	Variance component	Chi-square (df)	p-value
Two-level student/country	M0 Gender-DIF	0.311****	1.365	u_2	0.214	0.04592	41.137 (14)	<0.001
Two-level student/school	M0 Gender-DIF	0.288	1.334	u_2	0.123	0.01526	1442.365 (1618)	>0.500
				r_2	1.677	2.81446	1425.808 (1606)	>0.500
Three-level student/school/country	M0 Gender-DIF	0.199****	1.220	u_{20}	0.327	0.10704	39.14620 (14)	<0.001

Note. Gender was codified by 0 for girls and 1 for boys. **p<.05, ***p<.01, ****p<.001, SD: Standard deviation, df: degrees of freedom.

The model zero (M0) is based in equation 1, and it has the aim of detecting not only if an average gender DIF effect exists, but also if this DIF effect has shown

variability across groups (schools or countries). Comparing a holistic visualization of the gender DIF coefficients in all models (column 3 of Table 3), we were able to detect a positive coefficient. Based on our gender codification in the data set, girls equal zero and boys equal one. This result shows that even though when omitting or including variability across levels, boys are more like to endorse (answer correctly) on item number 19 than girls. It is important to note that the results with model zero (M0) are not controlling for contextual variables. That result, or phenomenon is variant across countries but is constant across schools (column 9, table 3).

Progressively, we drew in more of the nested structure information in our analysis. The next step included the variation of school level. We analyzed two-level Bernoulli logistic regression models including the student variables at level 1 (ability and gender) and controlling the random slope of gender by school characteristics (school SES and school infrastructure) at level two. Each variable was included separately in the analysis in contemplation of the estimation complexity and to avoid the multicollinearity due to the high correlation (r=.783, n=2663, p=.000). With the intention of discovering predictors that can explain the relationship between items responses and gender in different levels, we included variables that characterized the school profile. Table 4 displays all the models analyzed; it clearly shows the presence of DIF favouring boys in all the models (column 2) but not a significant variability between schools, which represent a similar profile of DIF across schools (column 10).

The coefficients of all the variables included at the school level can be seen in Table 5, column two. The coefficients of gender DIF are significant and positive in all the models. Given our variable codification, the intercept of the model is zero for urban school in M5 and zero for public schools in M6, but both of those variables are a non-significant predictor of gender slope (p=.584 and p=.733). In the same line, school climate

(M2) and teacher strategy (M3) present a negative coefficient as well as non-significant values (p=.935 and p=.365). However, in Table 5, two variables associated with school and family resources were positive but not significant predictors of gender slope at the school level (columns 4-5 and 10-11). As a result, none of the variables (such as school climate, type of professor strategy used, and rural or private school) are significant predictors of the relationship between gender and item responses.

Focusing on our principal purpose—the impact of country predictors—the two-level Bernoulli logistic regression model was run. Student characteristics were included at level 1 (ability and gender), while random gender slope was controlled by country characteristic (GII and HDI). Taking into consideration that the correlation between GII and HDI is -.703, which implies a high correlation between those two indexes, every variable in the model was included separately.

Similar to the results in Table 3, the results on Table 6 shows that gender DIF for item 19 is favouring boys even after it is controlled by the country gender inequality index. It is important to observe, however, the gender DIF switches to favouring girls when controlled by the country level of human development (column 2). It is noteworthy that girls are four times more likely to endorse that item correctly when the country increases the amount in their human development index (Table 7).

Table 4

Two level Models: Student and School

	Fixed Effect Gender (gamma γ_{20})					Random Effect gender u_2			
Item 19	Coefficients	p-value	odds ratio	Confidence Interval	Who is more likely to endorse	Standard deviation	Variance component	Chi-square (df)	p-value
M0 Gender_DIF	0.288502	<0.001	1.334	(1.166-1.528)	boys	0.12355	0.01526	1442.365 (1618)	>0.500
M1_SCH_SES	0.266529	<0.001	1.305	(1.129-1.510)	boys	0.13506	0.01824	1440.718 (1617)	>0.500
M2_CLIMATE	0.277729	<0.001	1.320	(1.151-1.514)	boys	0.12426	0.01544	1440.048 (1617)	>0.500
M3_STRATEGY	0.274064	<0.001	1.315	(1.144-1.513)	boys	0.12589	0.01585	1442.621 (1617)	>0.500
M4_INFRASTR	0.245505	<0.001	1.278	(1.105-1.479)	boys	0.13948	0.01945	1439.905 (1617)	>0.500
M5_RURAL	0.290627	<0.001	1.337	(1.129-1.584)	boys	0.12389	0.01535	1442.414 (1617)	>0.500
M6_TYPE_SCH	0.290875	<0.001	1.337	(1.134-1.577)	boys	0.12278	0.01507	1442.435 (1617)	>0.500

Note. Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom.

Table 5

Two level Models: Student and School

	Fixed Effect Gender (gamma γ_{20})		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Item 19	Coefficients Gender (gamma γ_{20})	p-value	SCH_SES (gamma γ21)	p-value	SCH_CLIMATE (gamma γ21)	p-value	TEACH_STRAT (gamma γ21)	p-value	SCHO_INFAEST (gamma γ21)	p-value	RURAL (gamma γ21)	p-value	TYPE_SCH (gamma γ21)	p-value
M0 Gender_DIF	0.288502	<0.001	-	-	-	-	-	-	-	-	-	-	-	-
M1_SCH.SES	0.266529	<0.001	0.060	0.378	-	-	-	-	-	-	-	-	-	-
M2_CLIMATE	0.277729	<0.001	-	-	-0.073	0.289	-	-	-	-	-	-	-	-
M3_STRATEGY	0.274064	<0.001	-	-	-	-	-0.075	0.388	-	-	-	-	-	-
M4_INFRASTR	0.245505	<0.001	-	-	-	-	-	-	0.110	0.102	-	-	-	-
M5_RURAL	0.290627	<0.001	-	-	-	-	-	-	-	-	-0.007	0.962	-	-
M6_TYPE_SCH	0.290875	<0.001	-	-	-	-	-	-	-	-	-	-	-0.009	0.950

Note. Gender was codified by 0 for girls and 1 for boys.

Table 6

Two level Models: Student and Country

	Fixed Effect Gender (gamma γ_{20})					Random Effect gender u2			
Item 19	Coefficients	p-value	odds ratio	Confidence Interval	Who is more likely to endorse	Standard deviation	Variance component	Chi-square (df)	p-value
M0 Gender-DIF	0.311543	0.002	1.365531	(1.144-1.630)	boys	0.21430	0.04592	41.13742 (14)	>0.001
M1 DIF controlled by GII	1.437591	0.009	4.210539	(1.541-11.506)	boys	0.10718	0.01149	20.83019 (13)	0.076
M2 DIF controlled by HDI	-1.896979	0.022	0.150021	(0.031-0.721)	girls	0.09330	0.00871	19.29655 (13)	0.114

Note. Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom.

Table 7

Two level Models: Student and Country

	Fixed Effect Gender (gamma γ_{20})			Model 1 controlled by GII		Model 2 controlled by HDI	
Item 19	Coefficients Gender (gamma γ_{20})	p-value	Who is more likely to endorse	GII (gamma γ_{21})	p-value	HDI (gamma γ_{21})	p-value
M0 Gender-DIF	0.311543	0.002	boys	-	-	-	-
M1 Gender inequality index	1.437591	0.009	boys	-2.819408	0.026	-	-
M2 Human development index	-1.896979	0.022	girls	-	-	2.984254	0.010

Note. Gender was codified by 0 for girls and 1 for boys.

Considering the strength of the multilevel approach, we carried out an analysis that allowed for the insertion of the variability between schools and countries. For that purpose, different models were performed including the three-level Bernoulli logistic regression models. Four different models were analyzed. For all the models, the variables at level 1 were constant (ability and gender) and the slope of gender was controlled by one variable separately at each time in every level 2 and 3 (Table 8). In the first model (M1) implemented (column 2-3), gender DIF was controlled by the school's socioeconomic status at level 2 and country index of gender inequality (GII) at level 3.

In addition to the variables at level 1 (ability and gender), the second model (M2) included the school SES at level 2 and the index of human development at the country level. In the third model (M3), the variation in gender coefficient was controlled by the school infrastructure index (level 2) and the gender inequality index at the country level (level 3). In the last model (M4), the coefficient of gender was controlled by the school infrastructure index (level 2) and the human development index at the country level (level 3). After controlling for level 2 and level 3 variables, the principal result is that the coefficients of gender DIF are significant in all models. Notwithstanding, the inclusion of the country human development index switches the sign of gender coefficient. Hence, this results in favouring girls over boys (Table 8).

Table 8

Three Level Models: Student, School and Country

	(M1)		(M2)		(M3)		(M4)	
Fixed effects	Coefficients	p-value	Coefficients	p-value	Coefficients	p-value	Coefficients	p-value
Gender γ_{200}	1.198166	0.344	-3.067581	0.148	1.095772	0.392	-1.111502	0.093
SCH_SES γ_{210}	0.132051	0.387	0.080799	0.596	-	-	-	-
SCH_INFRA γ_{210}	-	-	-	-	0.199960	0.180	0.066322	0.185
GII γ_{201}	-2.546304	0.407	-	-	-2.321548	0.456	-	-
HDI γ_{201}	-	-	4.400968	0.124	-	-	1.829206	0.048
Random effects	Variance component	p-value	Variance component	p-value	Variance component	p-value	Variance component	p-value
Gender r_2	2.82194	>0.500	2.82088	>0.500	2.85952	>0.500	0.06129	>0.500
Gender u_{20}	0.07641	0.004	0.04202	0.039	0.06211	0.011	0.10653	0.225

Note. Gender was codified by 0 for girls and 1 for boys.

After running a series of analysis including variables at both the school and country level, as well as bearing in mind the previous non-remarkable results for two-level analysis, we decided to allow the inclusion of the natural variability for the school level. As well, due to the model complexity we omit the inclusion of predictors at level 2 (Table 9). The data is then presented in a holistic visualization, which includes the variability or the impact of the student characteristics, school's effects, and country properties. We found that, even after controlling for different conditions, gender uniform DIF is still present. However, the association between gender and item responses changes to favouring girls when the variable human development index is included at the country level (Table 9 and 10, column 2). Considering the negative relationship between gender DIF and GII, this result implies that a medium size probability of gender DIF is associated with lower inequality. Taking into consideration the relationship between gender and items response controlled by HDI, girls are four times more likely to give a correct answer than boys. That relationship suggests that with higher levels in HDI, it is more likely to favour girls than boys in most of the nations participating in TERCE.

Table 9

Three Level Models: Student, School and Country

	Fixed Effect Gender (gamma γ_{200})					Random Effect Gender			
Item 19	Coefficients	p-value	odds ratio	Confidence Interval		Standard deviation	Variance component	Chi-square (df)	p-value
M0 Gender-DIF	0.199494	0.234	1.220785	(0.865-1.722)	r_2	1.67764	2.81446	1425.80876 (1606)	>0.500
					u_{20}	0.32717	0.10704	39.14620 (14)	>0.001
M1 DIF controlled by GII	1.715366	0.167	5.558710	(0.442-69.851)	r_2	1.67919	2.81969	1425.87216 (1606)	>0.500
					u_{20}	0.29750	0.08851	33.54459 (13)	0.002
M2 DIF controlled by HDI	-3.544203	0.095	0.028892	(0.000-2.043)	r_2	1.67939	2.82036	1426.13816 (1606)	>0.500
					u_{20}	0.19896	0.03959	22.71597(13)	0.045

Note. Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom.

Table 10

Three Level Models: Student, School and Country

	Fixed Effect Gender (gamma γ_{20})			M1 controlled by GII		M2 controlled by HDI	
Item 19	Coefficients Gender (gamma γ_{20})	p-value	Who is more likely to endorse	GII (gamma γ_{201})	p-value	HDI (gamma γ_{21})	p-value
M0 Gender-DIF	0.199494	0.234	boys	-	-	-	-
M1 Gender inequality index	1.715366	0.167	boys	-3.785609	0.215	-	-
M2 Human development index	-3.544203	0.095	girls	-	-	5.063227	0.075

Note. Gender was codified by 0 for girls and 1 for boys.

It is important to keep in mind that, although they certainly recognize that contextual effects are worthy of consideration, conventional validation practices do not pay much attention to contextual effects as part of validation. That is, although conventional validation practice would not disagree with the generic role of context in assessment, it does not pay much attention to it. Conventional validation practices place the contextual effects in the background while individual differences between test takers are in the foreground (Zumbo & Forer, 2011). This is particularly important given the well-known large education inequality in Latin America that are related to contextual factors (UNESCO-OREALC, 2016a).

This research has aimed to provide a holistic explanation about why DIF was occurring in educative assessments in Latin America contexts. The validity of the inferences one makes from test scores is bounded by place, time, and use of the score resulting from a measurement operation (Zumbo, 2007). In our case, DIF was explained by various factors from an ecological view standpoint, including the information of the schools' and countries' characteristics. Even though TERCE states they performed a gender DIF analysis, their report indicates that no item has shown to be a significant gender DIF. The results obtained for TERCE are not available for methodological analysis. Additionally, the technical report informs that gender DIF is not a criterion of item elimination (UNESCO-OREALC, 2016b, p. 252). However, the notorious absence of significant gender DIF results can be explained not only by the technique used (in this case, Mantel-Haenszel analysis) but also due to the omission of the information from the nested structure of the data.

The data reveals to us, in a holistic visualization of the results, that even if the model includes or omits the variability or the impact of the students' characteristics, schools' effects and countries' properties, that gender DIF is still present. However, the association between gender and item responses changes to favoring girls when the human development index is

included at the country level. A further dilemma arises for the particular process of DIF validity studies as the nested nature of the data cannot be underestimated and test takers have to be viewed in their complete life circumstances. A compounding variable in testing is the fact that although a great deal of the work is done in isolation, it is nevertheless influenced by contextual factors, such as the class environment, the school resources, country politics, and socioeconomic reality.

Validity is the foundation of a testing procedure, and the process of validating is key to the overall success of the educative assessment as a whole. This study deals specifically with the position of an ecological point of view which includes and situates the person, process, context, and time of the testing situation. These descriptions pinpointed specific incidents of how and what variables at the individual, school, or country level can give a deep understanding of the response process in Latin America countries.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

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

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. Multivariate Behavioral Research, 52(2), 149-163. doi:10.1080/00273171.2016.1256753

Browne, W., & Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, 1(3), 473-514. doi:10.1080/00273171.2016.1256753

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 Publishing.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. Psychological Methods, 12(2), 121-138. doi:10.1037/1082-989X.12.2.121

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

Raudenbush, S., Bryk, A., Cheong, Y. K., Congdon, R., & du Toit, M. (2011). HLM 7 hierarchical linear and nonlinear modeling. Lincolnwood,IL: SSI Scientific Software International.

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

United Nations Development Programme [UNDP]. (2013). Human development report 2013. The rise of the South: Human progress in a diverse world. New York: UNDP.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & La Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2016a). Recomendaciones de políticas educativas en América Latina en base al TERCE [Recommendations for educational policies in Latin America based on TERCE]. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & La Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2016b). Reporte técnico tercer estudio regional comparativo y explicativo. TERCE [Technical report third comparative and explanatory regional study. TERCE]. Santiago de Chile: UNESCO.

Zumbo, B.D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), Handbook of Statistics, Vol. 26: Psychometrics (pp.45-79). The Netherlands: Elsevier Science B.V.

Zumbo, B. D., & Forer, B. (2011). Testing and Measurement from a Multilevel View: Psychometrics and Validation. In James A. Bovaird, Kurt F. Geisinger, & Chad W. Buckendahl (Editors). *High Stakes Testing in Education - Science and Practice in K-12 Settings*, (pp.177-190). American Psychological Association Press, Washington, D.C..

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.

DISCUSIÓN Y CONCLUSIÓN GENERAL DE LA TESIS DOCTORAL

Seis son los marcos de acción del Foro Mundial sobre Educación, que reafirman el compromiso de los países de todas las regiones del mundo con la mejora de la educación. Esta iniciativa denominada *"Educación Para Todos"* de la UNESCO (2000), persigue entre sus objetivos velar por una educación primaria gratuita obligatoria y de calidad (Objetivo II), así como mitigar la desigualdad en materia de educación y suprimir las discriminaciones en las posibilidades de aprendizaje de grupos desfavorecidos (Objetivo V). Esta mejora de todos los aspectos de la educación debe estar basada en resultados educativos que puedan ser reconocidos y mesurables, especialmente en áreas educativas primordiales (Objetivo VI). En América Latina, la colección de estudios liderados por el LLECE desde 1997 se erigen como una herramienta clave que ofrece evidencias objetivas que sustentan la mejora en las políticas educativas de las naciones de la región.

La cultura de la evaluación educativa como herramienta para el diagnóstico, rendición de cuentas y mejora de la política educativa, es la herramienta más poderosa de la política educativa de los últimos veinte años (Volante, 2007, p. 4) y ha dejado de ser un ejercicio exclusivo de las regiones más desarrolladas de mundo, para convertirse en una herramienta de mejora en el 75% de los países en vías de desarrollo (Benavot & Tanner, 2007). El aumento de la tasa de participación no solo se ha visto en el número de países que concurren, sino también en el tamaño de la muestra de estudiantes, escuelas y sub-estados nacionales, este fenómeno. De la mano del crecimiento de las evaluaciones educativas estandarizadas a nivel internacional y regional, se observa también el incremento de la cantidad de pruebas (áreas y materias) aplicadas a nivel nacional.

No exentos del crecimiento y mejora observados a escala mundial, los estudios organizados por el LLECE han incursionando en el aumento de la participación de países, la

implementación de módulos nacionales de evaluación, ampliación áreas evaluadas, inclusión de comparabilidad entre evaluaciones y, finalmente en la mejora del proceso metodológico de la evaluación que garantizan la calidad de sus resultados. El contexto latinoamericano que a diferencia de los países anglosajones presenta características relacionadas a la escasa inversión en educación y el alto nivel de inequidad social. Esta realidad condiciona el proceso y por ende los resultados de logro educativo, por lo que en consonancia con esa realidad, la UNESCO construye sobre los resultados del TERCE las recomendaciones para las políticas públicas de mejora de la educación en América Latina desde el Modelo Ecológico (Bronfenbrenner, 1979) y el modelo CIPP (Stufflebeam, 1983) con el fin de enriquecer y profundizar el análisis de la mejora educativa desde una perspectiva holística que aprecia las características del contexto social de la región (UNESCO-OREALC, 2016a).

La Tesis Doctoral fundamentó sus objetivos y resultados en un programa de investigación que tiene el objetivo de comprender que la mediación «contextual» determina no solo las oportunidades de aprendizaje a las que el estudiante es expuesto, sino que también la forma en como los estudiantes comprenden y responden al ítem-prueba. El objetivo de conformar un análisis metodológico de la equidad del proceso de evaluación desde una mirada in vivo antes que in vitro, mediante el desafío titánico de unir en el análisis dos tradiciones de análisis. Por un lado, la aproximación educativa donde la equidad es entendida como la distribución de los conocimientos y oportunidades sociales y escolares y, por otro lado, la tradición psicométrica que busca aportar evidencias de la validez de las medidas, que garanticen el uso e interpretación de los resultados de las pruebas en la *evaluación educativa estandarizada*.

Es importante tener en cuenta que, aunque ciertamente los métricos reconocen que los efectos contextuales son dignos de consideración, las prácticas de validación convencionales no se encuentran centradas en los efectos del contexto como parte del proceso

de validación. Es decir, aunque la práctica de validación convencional no estaría en desacuerdo con el rol genérico del contexto en la evaluación, pero no es el efecto del contexto el foco central de análisis. Las prácticas de validación convencionales colocan los efectos contextuales en segundo plano, mientras que las diferencias individuales entre los examinados se encuentran en primer plano (Zumbo & Forer, 2011), diferenciación observada en la Figura 1. del prefacio de esta Tesis Doctoral. El creciente interés en garantizar la justicia en las pruebas que aglomeran en sus estudios a países con diversidad lingüística y cultural, puede ser observado con la constante creación de guías y manuales técnicos que buscan sistematizar y avalar el proceso, entre ellas se destacan *The Standards for Educational and Psychological Testing* (AERA et al., 2014), *The ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally diverse populations* (ITC; 2018) y las diferentes publicaciones técnicas resultado de cada evaluación educativa a escala mundial como la colección de manuales tecnicos de PISA (OECD, 2002, 2005, 2009, 2012, 2014), PIRLS, TIMMS (Joncas & Foy, 2012)por su puesto los del LLECE (UNESCO-OREALC, 2016).

Con el fin de responder a las interrogantes planteadas en la investigación se emplearon diversas las estrategias analíticas, que tal y como sustentan los *Standards* (AERA et al., 2014) el análisis no puede ser realizar atendiendo exclusivamente a las características personales del evaluado. A diferencia de gran parte de la investigación en psicología o de la investigación experimental, los estudios educativos se desarrollan en ambientes naturales y por lo que el proceso de validación de las medidas no puede separarse de contexto social y escolar en el que están inmersos los estudiantes que responden a las pruebas, ya que las pruebas de diagnóstico de los sistemas educativos centran la responsabilidad del logro académico del estudiante en actores tales como el profesorado, el centro y el sistema educativo. Por lo que de esta forma es crucial analizar el proceso de la evaluación educativa estandarizada desde una visión holística del contexto, tal vez mejor denominado como una mirada ecológica del proceso de evaluación.

Teniendo en cuenta que el fin primero de las evaluaciones educativas estandarizadas en este caso de escala regional, es la de comparar a los países participantes, este recorrido analítico se ha iniciado con la visión puesta en la comparabilidad de las medidas. Los instrumentos de medida aplicados en TERCE, están construidos en base a una tabla de especificaciones de doble entrada que identifica bloques de contenido y procesos cognitivos, y que provienen de un análisis en conjunto de la malla curricular de todos los países participantes. Si bien el análisis psicométrico presentado por la organización se destaca por su meticulosidad, implicando a teorías psicométricas como la TCT y la TRI, tanto para la selección de ítems como para la construcción del puntaje de los estudiantes relega un aspecto primordial como el análisis de la invarianza del constructo entre los países participantes. Este aspecto es de crucial importancia dado que una de las primeras informaciones que resalta en los informes de resultado es la tabla con las puntuaciones promedio de los países, presentadas en comparativa global. A primera vista lo que el lector intuye es que el constructo ya sea de lectura, ciencias naturales o matemática, es equivalente para cada uno de los países participantes de la evaluación. Otra de las principales ideas que el lector no especializado aprecia es la omisión de cómo el contexto social determina el logro cognitivo de los estudiantes.

El primer estudio ha utilizado una novedosa herramienta de análisis psicométrico, que permitió no solo la comparación simultánea de varias agrupaciones, sino también la inclusión de características como los niveles de estratificación, los clústeres de agrupación y los pesos muestrales. Como punto de partida, los resultados de este primer capítulo de análisis nos enfrentaron a la realidad de la existencia de parámetros no invariantes en la prueba de ciencias naturales. Dado que las medidas obtenidas por las agrupaciones no son equivalentes la comparación regional sobre el constructo evaluado puede no ser adecuada. El análisis realizado a los seis cuadernillos de la prueba de ciencias naturales nos mostró que en general al menos el 50% de los ítems que constituyen un cuadernillo presentan un parámetro invariante ya sea

este de intercepto o de carga del factor. Además, la clasificación de países de acuerdo con el peso de la carga del factor presenta una correlación alta con la media de los cinco valores plausibles obtenida a nivel país en la prueba de ciencias naturales. Recordemos en este caso que la ordenación de los países fue realizada por los organizadores teniendo en cuenta los pesos muestrales y la media de los cinco valores plausibles de cada área evaluada.

La evidencia observada sobre la variabilidad del constructo entre las agrupaciones nos orientó a identificar las causas probables que podrían impactar en la presencia de parámetros no-invariantes. En un primer momento, nos centramos en el tipo de codificación utilizado. Aunque los ítems de las pruebas cognitivas de TERCE son presentadas a los estudiantes en formato de opción múltiple y respuesta abierta, las bases de datos provistas al público en general presentan respuestas sin valores perdidos codificadas en formatos binario y de crédito parcial. En un ejercicio metodológico, se constataron tres tipos de codificación de las respuestas abiertas de cada cuadernillo[7]. En un segundo momento, se constató la relevancia del tipo de proceso cognitivo utilizado para responder al ítem y su relación con la presencia o no de no de parámetros no-invariantes.

De ambos análisis se concluye que no existe una relación significativa entre el tipo de proceso cognitivo utilizado, tipo de codificación de os ítems y la presencia de parámetros invariantes o no-invariantes. En cuanto al tipo de codificación del ítem, es importante destacar que todos los ítems de crédito parcial en todas las formas de codificación son ítems con parámetros no invariantes, lo cual esté tal vez relacionado con la forma en que los estudiantes procesan y responden al estímulo o tal vez con el tipo de calificación y codificación del corrector. En ese sentido, se destaca que la mayoría los parámetros no-invariantes se refieren a ítems que requieren del uso de procesos cognitivos de *entendimiento y comprensión de*

[7] Las codificaciones utilizadas y los parámetros no invariantes se presentan en el capítulo Anexo Tabla 2.

conceptos; por lo que la diferencia es observada en cómo los estudiantes procesan y comprenden el contenido de los ítems.

En un segundo momento la Tesis Doctoral, se centró en identificar los factores asociados a los resultados obtenidos, dado que los pesos de carga obtenidos en el factor en el primer estudio (Capítulo I) presentaron un alto nivel de correlación con la puntuación obtenida por los países en el ranking de distribución de conocimientos. Los tres estudios que conforman los capítulos *II, III y IV* se refieren al análisis de las características a nivel de aula, centro, sistema educativo y su efecto en el logro cognitivo en ciencias naturales, matemática y lectura.

Desde una visión educativa y pedagógica el capítulo II tuvo como objetivo identificar el efecto de los centros educativos en los resultados educativos. Si bien es tema de investigación con una amplia tradición en países europeos y anglosajones, los datos observados en América Latina presentaban cierta inconsistencia. Los resultados de la Tesis Doctoral confirman que el efecto neto de los centros escolares de América Latina participantes en el TERCE es superior al estimado para los países desarrollados. Por su parte el impacto de las características socioeconómicas de los estudiantes y sus familias oscila entre el 35% y el 68%; y este efecto tiende a ser mayor en las materias lingüísticas que en las científico-matemáticas. Dado que estos datos son uno de los principales aportes de la Tesis Doctoral, el análisis fue realizado con todas las materias evaluadas en TERCE y su impacto en la comunidad científica ha sido oficialmente reconocido[8].

Hasta aquí en la Tesis Doctoral se ha identificado la existencia de diferencias entre países en la conceptualización del constructo evaluado en las pruebas educativas aplicadas en TERCE, dado que este está estrechamente ligado a los resultados de logro cognitivo, hemos

[8] Distinción otorgada por la Sociedad Científica en Psicología y Educación con el Premio ACIPE 2018 en reconocimiento al mejor artículo publicado en la Revista de Psicología y Educación.

determinado cual es el efecto de contexto socioeconómico de las familias, los centros educativos y el sistema educativo de cada país en los resultados de la prueba.

En un paso más allá en la determinación de las variables que afectan al resultado académico y profundizando en lo que a la realidad del centro educativo se refiere. El capítulo *III Fairness in Testing de los Standards*, describe el impacto de las oportunidades de aprendizaje en el proceso de evaluación, e indica que en ocasiones los recursos escolares a los que tienen acceso los estudiantes en desventaja social se erigen como fuentes de inequidad, que pueden afectar al resultado de una prueba de evaluación educativa (AERA et al., 2014, p. 56). En consecuencia, el *capítulo III* de la Tesis Doctoral se orientó por tanto al análisis de las oportunidades de aprendizaje que ofrecen tanto las familias, como las instituciones y los sistemas educativos y cómo éstas impactan en el logro de los estudiantes.

Los resultados observados se encuentran también en la misma línea del estudio de Coleman de hace más de cincuenta años atrás. Los estudiantes de menor nivel socioeconómico tienen acceso a una menor infraestructura que los de un nivel socioeconómico más elevado, observándose un menor acceso a laboratorios, libros y cuadernos etc. El resultado es claro, en América Latina, la educación se encuentra en proceso de asegurar un acceso global al derecho de la educación para los niños con menores recursos económicos, la región latinoamericana se encuentra en una tesitura de frenar el impacto de las características de la sociedad, mientras que América del Norte y los países de Europa, se encuentran más bien orientados a garantizar una educación de calidad que cumpla con los más altos estándares de calidad a escala mundial.

El último estudio de esta Tesis Doctoral tiene como objetivo la fusión de las características del contexto educativo en donde se desarrolla la prueba, en lo que se denomina ecología del proceso de respuesta al ítem y rendimiento en la prueba. Este capítulo tuvo el fin de fundamentar y proveer de una explicación profunda al funcionamiento diferencial del ítem,

que nos orientó hacia la causa de las diferencias entre países. Para este efecto nos hemos sustentado en la explicación del DIF desde una visión ecológica del proceso de respuesta propuesto por Zumbo et al. (2015) como la tercera generación del estudio del DIF. Para este efecto teniendo en cuenta que se trata de una teoría novel y dada las características metodológicas del estudio, se utilizó el cuadernillo uno de la prueba de ciencias naturales. Encontrándose que más del 30% de la prueba presenta DIF basado en el sexo de los estudiantes y que este puede ser explicado por características del sistema educativo.

Las implicancias de los resultados de esta Tesis Doctoral abarcan no solo desde al campo educativo, sino que también se extienden al campo metodológico. Desde la perspectiva educativa los resultados del *Capítulo II, III y IV* se erigen como uno de los contados artículos científicos del acervo teórico sobre el análisis de los efectos escolares, que conjuga el uso de una metodología de análisis estadístico robusta que incluye a todas las naciones participantes, así como el análisis de todas las materias evaluadas en 6to grado por TERCE. El análisis de la conformación de las aulas en América latina, a más de darnos una medida sobre el nivel de segregación escolar, nos orienta al uso de sus resultados como una estrategia de conformación de aulas con el fin de proveer un acceso equitativo a las oportunidades de aprendizaje de las agrupaciones más desfavorecidas. En la misma línea del estudio de las oportunidades de aprendizaje el *Capítulo IV* se centra en las características de OTL medidas por TERCE y destaca que variables como el clima de aula, las prácticas pedagógicas y el acceso a materiales educativos determinan en gran medida el rendimiento de los estudiantes y se erigen como elementos de fundamentales de la oportunidad de aprendizaje.

Ambos estudios de corte metodológico fueron realizados bajo el uso de estrategias de analíticas novedosas que permiten la inclusión de todas las variables que grafican el escenario educativo y el aporte de sus resultados pueden ser aplicados tanto a nivel metodológico, como a nivel de política educativa. Y se erigen como evidencias de la validez de las medidas del

estudio TERCE, tanto desde la evaluación del constructo de las pruebas como del análisis del proceso de respuesta a la prueba. Por lo que, en este sentido, la finalidad o el foco de atención de estudios a futuro se centra en el aspecto de carácter ético del uso de las pruebas en las evaluaciones educativas estandarizadas (Padilla, Gómez , Hidalgo, & Muñiz, 2006).

El constante crecimiento de las evaluaciones educativas y el impacto de sus resultados en la política educativa de los países, a criterio de esta tesis se encuentra en un estadio orientado a la explicación de los fenómenos observados, convirtiéndose en una de las herramientas explicativas más poderosa para las políticas educativas y los sistemas de rendición de cuentas de los gobiernos (Lewis & Lingard, 2015). Bien, lo citaba Karp (1988) más de treinta años atrás, si bien el rol del uso de pruebas es comúnmente percibido como una herramienta útil, la interpretación de los resultados debe ser vista desde una perspectiva holística del ser humano, y no solo como la suma de partes. El argumento de que el todo es la suma de las partes es relevante en especial, en este caso. Es decir, el científico del comportamiento supone que el individuo está funcionando como un sistema interactivo, por lo que la interpretación de las puntuaciones de la prueba e inferencias realizadas sobre la misma, deben tener en cuenta la propensión y las predisposiciones del individuo para interactuar con su entorno.

Esta Tesis Doctoral fue desarrollada exclusivamente con datos proveídos por la UNESCO, por lo que entre las principales limitaciones del estudio se destaca la ausencia de las respuestas primarias del alumnado, dado que la codificación de los ítems de las pruebas cognitivas proveídos por la organización, son recodificaciones en formato binario sin la presencia de valores perdidos. Es importante en este punto destacar que los ítems se presentaban a los estudiantes en formato de opción múltiple y respuesta construida, y que por lo tanto se desconoce el porcentaje y el tipo de pérdida de valores perdidos; y por consiguiente el tipo de imputación de los valores utilizado. Por otro lado, la ausencia de información relacionada al proceso de traducción de las pruebas ha decaído en la imposibilidad de

desarrollar estudios que nos acerquen a una explicación profunda sobre el comportamiento de las pruebas en las versiones del castellano y el portugués.

La limitación más significativa en el diseño de la investigación es la falta de una medida del logro previo, así como la ausencia de información a nivel de aula característica observada, por ejemplo, en el SERCE. Si bien, los estudios de la UNESCO nunca han analizado variables de rendimiento previo, otras investigaciones han concluido que no hay duda de que una medida de rendimiento previo es el mejor predictor del rendimiento del alumnado individualmente considerado (Fernández-Alonso et al., 2015, 2017). Las futuras líneas de investigación se orientan al estudio de los residuales de centros a la luz de los modelos ajustados y el estudio de los factores de proceso educativo que pudieran dar cuenta del porcentaje de varianza no explicado en los modelos aquí presentados, tales como los estilos parentales (Osorio y González-Cámara, 2016), o las actitudes de los profesores (Álvarez-Martino et al., 2016; Cunha et al., 2015), por citar sólo dos de las muchas posibles. En cuanto a los estudios de corte psicométrico, el paso a futuro es la réplica de las puntuaciones de los estudiantes, esta vez, con los ítems que han se han caracterizado por ser invariantes entre las agrupaciones y carentes de funcionamiento diferencial del ítem, lo que podría dar una visión comparativa del rendimiento de las agrupaciones, basados en medidas ausentes de sesgo.

Finalmente, esta Tesis Doctoral destaca que, si bien el estudio TERCE se desarrolla desde una metodología robusta que compromete a todas las naciones participantes en el proceso, el contexto nacional tiene una implicancia elevada en el logro educativo de los estudiantes, e impacta no solo en el acceso a oportunidades de aprendizaje, sino que también en la forma en como los estudiantes perciben y responden a la prueba de evaluación. Por lo que la inclusión de las características del contexto social y educativo del alumnado, en especial en pruebas que tienen el objetivo de la evaluación y comparación multicultural, debe formar parte del análisis no solo en la construcción de las *inferencias, decisiones y recomendaciones sobre*

los resultados obtenidos, sino que también en todos los pasos del *proceso de validación de las medidas obtenidas en la prueba.* Lograr una mayor equidad en la educación no solo es un imperativo de justicia social, sino también es una forma de utilizar los recursos de manera más efectiva, aumentar la oferta de habilidades que impulsan el crecimiento económico y promover la cohesión social (OECD, 2016a, p. 4).

Recopilatorio de las principales evidencias

A continuación, se enumeran las principales conclusiones obtenidas del conjunto de capítulos incluidos en esta Tesis Doctoral:

Capítulo I

- El uso de la técnica del alignment es un método novedoso, eficiente y de alto grado de utilidad para la comparación de gran cantidad de agrupaciones.
- El alignment ha presentado un desempeño óptimo para el análisis de las 16 naciones en simultáneo y la inclusión de variables de estratificación, clústeres y pesos muestrales.
- Se observa la presencia de parámetros no invariantes en al menos el 50% de los ítems de cada cuadernillo de la prueba de ciencias naturales.
- La ausencia de equivalencia (parámetros no invariantes) se concentra en el intercepto del ítem antes que en la carga de los factores.
- Las variabilidades están concentradas en países tales como Chile, Brasil, Ecuador, Guatemala y Nuevo León (México).
- El tipo de proceso cognitivo utilizado para resolver el ítem de la prueba de ciencias naturales no presenta relación significativa con la presencia de parámetros no invariantes, aunque si bien la mayoría de los parámetros no invariantes corresponden a ítems de comprensión y reconocimiento de conceptos.

- El 100% de los ítems de crédito parcial presentan parámetros no invariantes en el intercepto del ítem.
- Si bien los ítems de crédito parcial fueron a su vez recodificados en formato binario ya sea totalmente correcto o totalmente incorrecto, en todos los tipos de recodificación se encontraron parámetros no invariantes. Esto debe ser una llamada de atención a la inclusión de este tipo de ítems, o bien una mejora en el proceso de corrección de estos.
- Existe una correlación que oscila entre .94 a .98 entre la media de los cinco valores plausibles y la carga factorial resultante del análisis del alignment en la prueba de ciencias naturales.
- Las comparaciones entre países no son precisas debido a la presencia de un alto nivel de disparidad en la forma en cómo los estudiantes de cada país comprenden y responden al ítem en la prueba de ciencias naturales.

Capítulo II

- El ICC de los centros en América Latina en la prueba TERCE del 2013 se mueve en torno al 40%. Dicho indicador se ha mantenido estable desde las evaluaciones PERCE y SERCE.
- El ICC de los nueve países participantes en PERCE y SERCE han experimentado un crecimiento que se sitúa entre 2 y 22 puntos porcentuales según país y materia.
- La correlación entre el efecto bruto obtenido en SERCE y TERCE oscila entre .78 y .81.
- El valor del ICC se reduce en un 60% al incluir variables de tipo socioeconómico. El impacto es mayor en lectura, que en las materias científico-matemáticas.
- El ISEC ofrecido por TERCE presenta una adecuada capacidad para explicar la varianza de resultados, superando las limitaciones del ISEC ofrecido en SERCE.

- El efecto neto de los centros escolares en América Latina es del 13% en lectura, 23% en matemáticas y 25% en ciencias naturales. Y el promedio de la correlación entre los resultados obtenidos en SERCE es de .82.
- Se observan dos grupos de países. Por un lado, países con diferencias entre centros y efecto de variables contextuales relativamente pequeñas y que representan a los países más homogéneos (equitativos): Chile, Costa Rica, Nuevo León, República Dominicana y Uruguay. Por el otro lado países donde el efecto del centro es mayor, así como el impacto de variables sociodemográficas: Perú, Panamá, Brasil, Guatemala.

Capítulo III

- En el modelo analizado sin la inclusión de variables de control, los estudiantes de bajo nivel socioeconómico que asisten a centros educativos heterogéneos e inclusivos presentan peores resultados, que los que asisten a agrupaciones homogéneas. Mientras que los de nivel socioeconómico alto escolarizados en centros educativos heterogéneos presentan mejores resultados en la prueba de matemáticas.
- Una vez controladas las variables sociodemográficas el efecto de segregación es significativo lo que parece indicar que a medida que los centros presentan mayores dispersiones en el índice socioeconómico tienden a mostrar resultados más bajos en la prueba de matemáticas.
- En grupos altamente homogéneos el modelo predice que la diferencia de puntuación entre un estudiante de SEC bajo y otro de SEC alto está en torno a 55 puntos, es decir, más de media desviación típica. Sin embargo, esta distancia se reduce a la mitad se (unos 21 puntos) cuando estos estudiantes se escolarizan en grupos heterogéneos.
- Una vez controlado el modelo por variables contextuales los estudiantes de SEC bajo se benefician por asistir a agrupaciones inclusivas (heterogéneas).

- Mientras que los estudiantes de nivel socioeconómico alto no se ven penalizados por asistir a agrupaciones heterogéneas.
- En cuanto a los resultados desagregados por países, no se observa un patrón claro, denotando que los países presentan efectos diferentes a la hora de tratar las diferencias sociales.

Capítulo IV

- Las aulas con mejor clima, con menores ausencias del profesorado y con puntuaciones más altas en el índice de prácticas educativas presentan una correlación que varía entre .23 y .15 con el rendimiento en la prueba de ciencias naturales.
- Teniendo en cuenta los resultados del análisis multinivel, el modelo nulo indica que la varianza se distribuye en 54% a nivel de centro y el 14% a nivel de países.
- El modelo que incluye variables de ajuste reduce el porcentaje de varianza del modelo nulo en un 20%, entre las variables con mayor poder explicativo se encuentra el SEC.
- El modelo que incluye variables de OTL explica un 5% de la varianza total y casi el 10% de las diferencias entre centros.
- Mientras que el modelo que incluye variables de ajuste y variables OTL en conjunto, explica el 25% de la varianza total y el 50% de las diferencias entre centros.
- El 40% de las diferencias en los resultados entre los centros educativos se explican por factores de contexto o entrada.
- El SEC del centro está fuertemente vinculado a los resultados. El modelo predice diferencias del orden de 0.92 desviaciones típicas entre el alumnado de los centros con mayor y menor nivel SEC.
- En América Latina, disponer de un simple cuaderno de clase, parece una buena aproximación para una estimación general de los recursos disponibles, puesto que se

predicen ganancias en torno al 11% de la desviación típica después de descontar el efecto de las variables de contexto.

Capítulo V

- El 32% (9 ítems) de los ítems del cuadernillo uno de la prueba de ciencias naturales, presenta evidencias de funcionamiento diferencial teniendo en cuenta la variable de género.
- Los estudiantes del sexo masculino tienen a dar una respuesta favorable en cinco ítems y los estudiantes del sexo femenino en cuatro de los nueve ítems que presentan DIF.
- La distribución del tipo de respuesta (acierto/error) en el ítem 19 y el sexo de la persona que lo responde presenta una relación estadísticamente significativa.
- Aun cuando se analiza el DIF en dos o tres niveles de agregación, su presencia es significativa.
- A nivel escolar no se encuentran variables significativas que expliquen la presencia de funcionamiento diferencial basado en el género del estudiante.
- A nivel de sistema educativo las variables explicativas de la presencia de funcionamiento diferencial basado en el género del estudiante son: el Índice de Inequidad de Género (GII) y el Índice de Desarrollo Humano (HDI).
- Una vez controlado el funcionamiento diferencial del ítem por el HDI, las mujeres tienen cuatro veces más posibilidades de dar una respuesta correcta que los hombres.
- Cuando se introducen variables en los niveles del centro educativo y sistema educativo solo el HDI es una variable explicativa significativa.
- En todos los modelos analizados la presencia de DIF favorece a los estudiantes del sexo masculino, pero una vez controlado el coeficiente por el HDI todos los ítems favorecen a las estudiantes del sexo femenino.

DISCUSSION AND GENERAL CONCLUSION OF THE DOCTORAL THESIS

There are six action frameworks from the World Education Forum that reaffirm the commitment of countries all over the world to improving education. Among the objectives of the UNESCO (2002) initiative "*Education for All*" are the provision of quality, free compulsory primary education for all (Objective II), along with reducing inequality in education and eliminating discrimination in learning possibilities for disadvantaged groups (Objective V). This improvement in all aspects of education must be based on results which can be recognized and measured, especially in primary education (OBJECTIVE VI). In Latin America research led by LLECE since 1997 has been a key tool offering objective evidence underpinning improvement in education policy in the region's countries.

The culture of educational evaluation for diagnosis, accountability, and improvement of educational policies is the most powerful tool in education policy in the last twenty years (Volante, 2007, p. 4) and is no longer something which is exclusive to the most developed areas of the world and is now an improvement tool in 75% of developing nations (Benavot & Tanner, 2007). The increase in participation is not only seen in the number of countries using it, but also in the sample sizes of students, schools, and sub-national states. The growth of standardised educational evaluations at an international level has gone hand in hand with an increase in the amount of tests (in terms of subjects tested) at a national level.

Research organized by LLECE has not been immune from that global growth and improvement, it has increased the number of participating countries, implemented national evaluation modules, widened the areas evaluated, added comparability between evaluations, and has improved the methodological process of evaluation to ensure the quality of the results. One characteristic distinguishing the Latin American environment from English-speaking countries is the relatively low investment in education and the high levels of social inequality. This reality affects the process, and therefore the results of educational attainment. This is why

UNESCO, confronted with this reality, used the results of TERCE to produce recommendations for public policies to improve education in Latin America from the Ecological Model (Bronfenbrenner, 1979) and the CIPP model (Stufflebeam, 1983), the aim of which was to achieve a deeper and richer analysis of educational improvement from a holistic perspective that considered the social characteristics of the region (UNESCO-OREALC, 2016a).

The basis of the objectives and results of this Doctoral Thesis was a research program the objective of which was to understand that "contextual" measurement determines not only the opportunities to lean that students are exposed to, but also the way the students understand and respond to test items. The goal was to create a methodological analysis of equality in the evaluation process from an in vivo rather than in vitro point of view, through the enormous challenge of uniting two analytical traditions. One, the educational approach in which equity is the fair distribution of knowledge and social and school opportunities, and the other, the psychometric tradition which seeks to provide evidence of validity for measurements which provide backing for the use of the test results in *standardised educational assessment*.

It is important to remember that, while it is true that metrics recognise that contextual effects are worthy of consideration, conventional validation practices do not focus on the effects of context as part of the validation process. Although conventional validation practice would not disagree with the generic role of context in evaluation, the effect of that context is not the main focus of analysis. Conventional validation practice puts contextual effects in the background, while individual differences between those tested are in the foreground (Zumbo & Forer, 2011), something which is seen in Figure 1. in the preface to this thesis. The growing interest in ensuring fairness in tests in studies in with linguistic and cultural diversity can be seen in the constant production of guides and technical manuals which seek to standardize the process. They include *The Standards for Educational and Psychological Testing* (AERA et al.,

2014), *The ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally diverse populations* (ITC; 2018) and the various technical publications for each global educational assessment, such as the technical manuals for PISA (OECD, 2002, 2005, 2009, 2012, 2014), PIRLS, TIMMS (Joncas & Foy, 2012) and of course LLECE (UNESCO-OREALC, 2016).

Various analytical strategies have been used to respond to the questions raised in this research. As the *Standards* (AERA et al., 2014) state, analysis cannot be performed by only looking at the individual characteristics of the person being evaluated. In contrast to much psychology and experimental research, educational studies are performed in natural surroundings, which is why validation of those measurements cannot be separated from the social or school contexts of students who respond to those tests. Diagnostic tests of educational systems place the responsibility for students' academic achievement on actors such as the teachers, schools and the education system itself. This is why it is crucial to examine the standardised educational evaluation process from a holistic contextual point of view, maybe better described as an ecological view of the evaluation process.

Because the primary aim of standardised educational evaluations is to compare participating countries, this part of the analysis began by looking at the comparability of the measures. The measuring instruments used in TERCE are constructed based on a double entry table of specifications which identifies blocks of content and cognitive processes, and which comes from a combined analysis of the curricula of all the participating counties. While the psychometric analysis presented by the organisation is meticulous, including psychometric theories such as TCT and TRU, both the item selection and the construction of the scoring for students downgrade an important basic aspect which is the construct invariance between the participating countries. This is a critically important aspect given that one of the first pieces of information in the reports of results is the table of average scores by countries, presented in

overall comparison. At first sight the reader may think that the construct, whether reading, science or mathematics, is the same for each participating country. Another significant idea that the reader may not appreciate is the omission of how social context determines students' cognitive attainment.

The first study used a novel psychometric analysis tool which allowed not only the simultaneous comparison of various groupings but also the inclusion of characteristics such as stratification levels, group clusters and sample weights. As a starting point, the results of this first chapter of analysis revealed the existence of non-invariant parameters in the science test. As the scores from the groups are not equivalent, regional comparison of that construct may be inappropriate. The analysis of the six science test booklets showed that generally at least 50% of test booklet items had an invariant parameter, albeit the intercept or loading factor. In addition, classification of countries in line with the loading factor weights exhibited a strong correlation with the mean of the five plausible values at country level in the science test. It should be remembered that the country ordering was performed by the organisers, considering sample weights and the mean of the five plausible values in each area being evaluated.

The evidence of construct variability between groups pointed towards the need to identify probable causes that might have an impact on the presence of non-invariable parameters. The first focus was on the type of coding used. Although the TERCE cognitive test items are presented to the students in multiple choice and open answer formats, the databases open the public generally give responses without missing values, coded in binary and partial credit formats. In a methodological exercise, three types of coding in open answers in each test booklet were checked[9]. Following that the importance of each type of cognitive process used

[9] The coding used, and the non-invariant parameters are presented in Annex Table 2

to respond to the item was verified along with its relationship to the presence or absence of non-invariant parameters.

The conclusion from those two analyses was that there was no significant relationship between the type of cognitive process used, the type of item coding and the presence of invariant or non-invariant parameters. It is important to highlight that all of the partial credit items in all coding types were items with non-invariant parameters, which may be related to the way students process and respond to the stimulus or may be with the type of scoring and coding by the marker. The majority of the non-invariant parameters are in items that require the use of the cognitive processes of *understanding and comprehension of concepts*; consequently, the difference is seen in how students process and understand the item content.

The second stage of the doctoral thesis focused on identifying the factors associated with the results from the first part, as the loading weights in the factor in the first study (Chapter I) demonstrated a strong correlation with the countries' scores in the knowledge distribution ranking. The three studies making up chapters *II, III* y *IV* are about the analysis of characteristics at class, school and education system level and their effect on cognitive achievement in science, mathematics and reading.

The objective of chapter II from an educational and pedagogical perspective was to identify the effect of schools on educational results. While this is something with a long tradition in European and English-speaking countries, the data from Latin America is rather inconsistent. The results of this doctoral thesis confirm that the net effect of Latin American schools participating in TERCE is higher than that estimated for developing countries. The effect of socioeconomic characteristics of students and their families ranges between 35% and 68% and tends to be higher in language related subjects than in science or mathematics. Given that this data is one of the major contributions of this doctoral thesis, the analysis was

performed on all of the TERCE evaluation subjects and its impact on the scientific community has been officially recognised[10].

At this point in the doctoral thesis, the existence of differences between countries has been identified in the conceptualization of the construct evaluated in the TERCE educational tests. As this is closely linked to results of cognitive achievement, the effect the socioeconomic context of families, schools and the education system in each country on test results was determined.

Going beyond the determination of the variables which affect academic results and going deeper into the reality of the schools, the *Standards* (AERA et al., 2014) in chapter *III Fairness in Testing* , describes the impact of opportunities to learn in the evaluation process, and indicates that on occasion the school resources that socioeconomically disadvantaged students have access to can be sources of inequality that can affect the results of educational evaluations (AERA et al., 2014, p. 56). Consequently, chapter *III* focuses on the analysis of opportunities to learn that are offered by families, schools and education systems, and how that impacts on student achievement.

The results seen are in line with research carried out by Coleman more than fifty years previously. Students from lower socioeconomic levels have access to poorer infrastructure than students from higher socioeconomic levels, less access to laboratories, books, notebooks etc. The result is clear; in Latin America education is in the process of ensuring overall access to the right to an education for children from poorer economic backgrounds. Latin America is reducing the impact of societal characteristics, while in North America and Europe education

[10] Distinction granted by the Scientific Society in Psychology and Education with the ACIPE 2018 prize in recognition of the best article published in the Journal of Psychology and Education.

is oriented towards ensuring high quality education which complies with the highest standards on a global scale.

The objective of the final study in this doctoral thesis was to integrate the contextual characteristics of where the test was performed, in what is called the ecology of the item response process, and test performance. The goal of this chapter was to thoroughly examine and provide a deep explanation for differential item functioning, leading towards the cause of differences between countries. This involved the explanation of DIF from ecological response process models proposed by Zumbo et al. (2015) as the third generation of DIF study. Bearing in mind that this involved a novel theory and given the methodological characteristics of the study, science test booklet one was used. More than 30% of the test exhibited DIF based on student gender which may be explained by characteristics of the education system.

The results of this doctoral thesis have implications not only in the field of education, but also in methodology. From an educational perspective, the results of *Chapters II, III* and *IV* forms one of the few scientific articles of a theoretical legacy on the analysis of the effects of schools which combines a robust statistical analytical methodology including all participating nations as well as an analysis of all subjects assessed in 6^{th} grade by TERCE. Analysis of classroom makeup in Latin America, in addition to providing a measure of school segregation, also hints towards the use of the results as a strategy for classroom composition with the aim of providing equal access to opportunities to learn for the most disadvantaged groups. *Chapter IV* similarly focuses on the OTL factors measured by TERCE and indicates that variables such as classroom climate, teaching practices, and access to educational materials largely determine student performance and are the fundamental elements of the opportunity to learn.

Both methodological studies were performed using novel analytical strategies which allowed the inclusion of all of the variables which describe the educational stage. The contribution of those results may be in their application at both the methodological and educational policy level. They stand as evidence of the validity of TERCE measures, in the evaluation of the test construct and the analysis of the test response process. This is why the aim or focus of attention of future studies must be on the ethical usage of the tests in standardised educational assessments (Padilla, Gómez , Hidalgo, & Muñiz, 2006).

The continued growth of educational assessments and the impact of their results on national educational policies, is becoming one of the most powerful explanatory tools for educational policies and systems of government accountability (Lewis & Lingard, 2015). As Karp (1988) stated more than thirty years ago, while the use of tests is commonly seen as a useful tool, interpretation should be from a holistic human perspective, not as just the sum of its parts. The behavioral scientist supposes that an individual function like an interactive system, so the interpretation of test scores and subsequent inferences must take into account an individual's tendency and predisposition to interact with their environment.

This doctoral thesis was produced using only data provided by UNESCO, meaning that one of the main limitations of the study is the absence of the original student responses. The item coding of the cognitive tests provided by UNESCO are binary format re-codings without missing values. It is important to note that the students were presented with items in multiple choice and constructed response formats, and therefore the percentage and type of missing values is unknown, along with the type of value imputation used. In addition, a lack of information about the translation process resulted in the impossibility of performing studies that would have given a deeper understanding of the behaviour of the tests in Spanish and Portuguese versions.

The most significant limitation in the research design is the lack of a measure of previous achievement, as well as the absence of information at the level of characteristic classroom, such as that seen in SERCE. Although UNESCO studies have never examined variables of previous achievement, other research has concluded that there is no doubt that measures of previous achievement are the best predictor of individual student performance (Fernández-Alonso et al., 2015, 2017). Future research could look at the school residuals in light of the adjusted models and the study of factors in the educational process that might account for the percentage of variance not explained by the models presented here, such as parental styles (Osorio y González-Cámara, 2016), and teacher attitudes (Álvarez-Martino et al., 2016; Cunha et al., 2015), to name two of many possibilities. A future step for psychometric studies would be the replication of student scores, this time with the items that have been identified as invariant between groups and lacking differential item functioning, which might provide a comparative view of group performance based on unbiased measures.

Finally, this doctoral thesis highlights that, while TERCE was developed from a robust methodology which involved all of the participating countries in the process, the national context has significant impact on student educational performance, not only in opportunities to learn, but also in the way that students view and respond to the assessment. For this reason, characteristics of student social and educational contexts, particularly in tests aimed at multicultural assessment and comparison, must form part of the analysis, not just in the construction of the *inferences, decisions and recommendations based on the results*, but also in all parts of the *validation process of the test measures*. Achieving better educational equity is not just an imperative of social justice, it is also a way of more effectively using resources and improving the skills available that will drive economic growth and encourage social cohesion (OECD, 2016a, p. 4).

Summary of the main evidence

The main conclusions of each chapter in this doctoral thesis are given below:

Chapter I

- The alignment technique is a novel, efficient and extremely useful method for comparing large numbers of groups.
- Alignment demonstrated excellent performance in the analysis of the 16 countries simultaneously, and the inclusion of variables of stratification, clusters and sample weights.
- Non-invariant parameters were seen in at least 50% of items in each test booklet for science.
- The absence of equivalence (non-invariant parameters) were mainly in the item intercept rather than in factor loadings.
- The variability was mainly seen in countries such as Chile, Brazil, Ecuador, Guatemala and Nuevo León (Mexico).
- The type of cognitive process used in responding to the test items in science does not exhibit a significant relationship with the presence of non-invariant parameters, most of the non-invariant parameters correspond to items related to comprehension and understanding concepts.
- All (100%) of the partial credit items exhibit non-invariant parameters in the item intercept.
- Although the partial credit items were recoded into a binary format, totally correct or totally incorrect, non-invariant parameters were found in all of the recoding types. This should be a call to look at the inclusion of these item types, or at least an improvement

in the marking process.

- There is a correlation ranging between .94 and .98 between the mean of the five plausible values and the factorial loading resulting from the alignment analysis of the science test.
- Comparisons between countries are not accurate owing to the great disparity in the way students in each country understand and respond to each test item in the science test.

Chapter II

- The ICC for Latin American schools in the 2013 TERCE test is around 40%. This indicator has remained largely unchanged since the PERCE and SERCE evaluations.
- The ICC in the nine participating countries in PERCE and SERCE has increased by between 2 and 22 percentage points depending on country and subject.
- The correlation between the net effect found in SERCE and TERCE is between .78 and .81.
- The ICC diminishes by 60% when socioeconomic variables are included. The effect is greater in reading than in science or mathematics.
- The ISEC offered by TERCE shows adequate capacity to explain the variance of the results, overcoming the limitations of the ISEC in SERCE.
- The net effect of schools in Latin America is 13% in reading, 23% in mathematics and 25% in science. The mean correlation between the results in SERCE is .82.
- Two groups of countries were identified. One with countries with relatively small differences between schools and relatively small effects of contextual variables, representing more homogeneous (more equitable) countries: Chile, Costa Rica, Nuevo Léon, The Dominican Republic and Uruguay. The other group made up of countries in

which the effect of the school is larger, as is the effect of sociodemographic variables: Peru, Panama, Brazil and Guatemala.

Chapter III

- In the model examined without control variables, students from low socioeconomic backgrounds attending inclusive schools with diverse populations exhibited worse results than students in more homogeneous groups. At the same time, students from higher socioeconomic levels attending diverse schools exhibited better results in mathematics.

- Once controlled for sociodemographic variables, the effect of segregation is significant which seems to indicate the extent to which schools with more diversity in socioeconomic indices tend to demonstrate lower results in the mathematics test.

- In very homogeneous groups the model predicts a difference in scores between a student from a lower socioeconomic group and a student from a higher socioeconomic group of around 55 points, more than half a standard deviation. However, this difference is halved (to 21 points) when the students are in more heterogeneous groups.

- Once contextual variables are controlled for, students from lower socioeconomic levels benefit from attending inclusive (heterogeneous) groups.

- Students from high socioeconomic levels do not seem to be penalized by attending heterogeneous groups.

- In the disaggregate data there is no clear pattern, indicating that countries demonstrate different effects when addressing social differences.

Chapter IV

- Classrooms with better climates, with fewer teacher absences and higher scores in

educational practice exhibit a correlation between .23 and .15 with performance in the science test.

- Considering the results of the multilevel analysis, the null model indicates that the variance is distributed 54% at school level and 14% at country level.
- The model including adjustment variables reduces the percentage of variance in the null model to 20%. SEC (socioeconomic class) is amongst the variables with the most explanatory power.
- The model which includes OTL variables explains 5% of the total variance and almost 10% of the differences between schools.
- The model which includes adjustment variables combined with OTL variables explains 25% of the total variance and 50% of the differences between schools.
- Factors related to context or admission explain 40% of the differences in results between schools.
- School SEC is closely linked to results. The model predicts differences in the order of 0.92 standard deviations between schools with a high SEC and schools with a low SEC.
- In Latin America, provision of a simple class exercise book seems to be a good approximation for the general estimation of available resources, as it predicts increases of around 11% of a standard deviation after controlling for the effects of context variables.

Chapter V

- Almost a third (32%; 9 items) of the items in booklet one of the science test showed evidence of differential item functioning based on gender.
- Of the nine items exhibiting gender-related DIF boys tended to be more likely to answer

correctly in five items and girls in four.

- The distribution of the type of response (correct/incorrect) in item 19 exhibited a statistically significant relationship with the gender of the respondent.
- The relationship is still significant even when the DIF is analysed a two or three group levels.
- No significant variables at school level were found which would explain the presence of gender-related differential item functioning.
- At the level of education system, the explanatory variables for the presence of gender-related differential item functioning are: The Gender Inequality Index (GII) and the Human Development Index (HDI).
- Once differential item functioning by HDI is controlled for, girls are four times as likely to answer correctly than boys.
- When school level and education system level variables are added, HDI alone is a significant explanatory variable.
- In all the models examined, DIF means boys score higher, but once the coefficient is controlled for HDI girls score higher in all of the items.

REFERENCIAS

Administración Nacional de la Educación Pública [ANEP]. (2003). *La evaluación de las ciencias en 6to. año de educación primaria: Aportes para la la elaboración de una agenda*. Montevideo: Programa de evaluación de aprendizajes.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (1977). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Arcidiácono, M., Cruces, G., Gasparini, L., Jaume, D., Serio, M., & Vázquez, E. (2014). *La segregación escolar público-privada en América Latina* (Documento de Trabajo, No. 167). La Plata: Centro de Estudios Distributivos, Laborales y Sociales [CEDLAS].

Asparouhov, T., & Muthén, B.O.(2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. doi:10.1080/10705511.2014.919210

Baker, D. (2014). *The schooled society: The educational transformation of global culture*. Palo Alto, CA: Stanford University Press.

Balarín, M. (2016). La privatización por defecto y el surgimiento de las escuelas privadas de bajo costo en el Perú. ¿Cuáles son sus consecuencias? *Revista de la Asociación de Sociología de la Educación*, *9*(2), 181-196.

Banco Interamericano de Desarrollo [BID]. (2017). *Aprender mejor: Políticas públicas para el desarrollo de habilidades*: Banco Interamericano de Desarrollo.

Beaton, A.E., Rogers, A.M., Gonzalez, E., Hanly, M.B., Kolstad, A., Rust, K.F., Sikali, E., & Jia, Y. (2011). *The NAEP Primer*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Benavot, A., & Tanner, E. (2007). *The growth of national learning assessments in the world, 1995-2006*. Background paper for the education for all global monitoring report 2008 Education for all by 2015: Will we make it. Recuperado de http://unesdoc.unesco.org/images/0015/001555/155507e.pdf

Bordieu, P., & Passeron, J.C. (1996). *La Reproducción. Elementos para una teoría del sistema de enseñanza*. México D.F.: Fontamara S.A.

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Castro-Aristizabal, G., Castillo-Caicedo, M., & Mendoza-Parra, J. (2016). Principales determinantes en la adquisición de competencias en América Latina: Un análisis multinivel a partir de los resultados en PISA 2012 *Documentos de trabajo Facultad de Ciencias Económicas y Administrativas* (22), 4-31. doi:10.2139/ssrn.2744657

Castro-Aristizabal, G., & Giménez, G. (2017). ¿Por qué los estudiantes de colegios públicos y privados de Costa Rica obtienen distintos resultados académicos? *Revista de la Facultad Latinoamericana de Ciencias Sociales, 25*(49), 195-223. doi:10.18504/pl2549-009-2017

Centro de Investigaciones Económicas Nacionales [CIEN]. (2002). *Informe de progreso educativo Guatemala 2002*. Guatemala: CIEN.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. (1966). *Equality of educational opportunity*. Washington, D.C.: U. S. Government.

Creemers, B.P.M., & Kyriakides, L. (2010). Using the dynamic model to develop an evidence-based and theory-driven approach to school improvement. *Irish Educational Studies, 29*(1), 5-23. doi:10.1080/03323310903522669

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.

Cronbach, L.J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*, 281-302.

Dirección Nacional de Información y Evaluación de la Calidad Educativa. (2003). *La evaluación de la calidad educativa en Argentina - Experiencias provinciales*. Buenos Aires: Ministerio de Educación Ciencia y Tecnología.

Ebel, R.L. (1961). Must all tests be valid? *American Psychologist, 16*(10), 640-647. doi:10.1037/h0045478

Edmonds, R. (1982). Programs of school improvement. An overview. *Educational Ledearship, 40*(3), 4-11.

Fernández-Alonso, R. (2004). *Evaluación del rendimiento matemático.* (Tesis doctoral inédita Universidad de Oviedo, Asturias, España).

Fernández-Aguerre, T., Trevignani, V., & Silva, C. (2003). *Las escuelas eficaces en Honduras*. Tegucigalpa: PNUD.

Fundación Centro de Estudios en Políticas Públicas, & Fundación Konrad Adenauer. (2005). *Informe del sistema educativo de Colombia*. Buenos Aires y Rio de Janeiro: CEPP-FKA.

Fundación para la Educación Ricardo Ernesto Maduro Andreu. (2017). *Informe de progreso educativo: Honduras*. Honduras: FEREMA.

Harmon, M., Smith, T.A., Martin, M.O., Kelly, D.L., Beaton, A.E., Mullis, I.V.S., Gonzales, E.J., & Orpwood, G. (1997). *Performance assessment in IEA´s third international mathematics and science study*. Boston College, MA: Center for the Study of Testing, Evaluation, and Educational Policy.

Hernández-Castilla, R., Murillo, F.J., & Martínez-Garrido, C. (2013). Factores de ineficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 12*(1), 103-118.

Hopkins, D. (1990). The international improvement project (ISIP) and effective schooling: Towards a synthesis. *School Organisation, 10*(2&3), 179-194.

Hungi, N. (2011). *Accounting for variations in the quality of primary school education*. SACMEQ Working Paper, Number 7. Recuperado de http://www.sacmeq.org/sites/default/files/sacmeq/publications/07_multivariate_final.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2007). *SAEB 2005 Primeiros resultados. Médias de desempenho do SAEB/2005 em perspectiva comparada*. Brasilia: Ministério da Educação - Brazil.

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.

Karp, B., T., (1988). *Selecting high school counsellors: The development of a behaviour-based interview pattern* (Master´s thesis, Department of Educational Policy and Administrative Studies, University of Calgary).

Klitgaard, R., & Hall, G. (1975). Are there unusually effective schools? *The Journal of Human Resources, 10*(1), 90-106. doi:10.2307/145121

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [LLECE]. (2014). *Primera entrega de resultados tercer estudio regional comparativo y explicativo Terce*. Santiago de Chile: UNESCO.

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación & Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe [LLECE, & UNESCO-OREALC]. (2016). Informe de resultados del tercer estudio regional comparativo y explicativo (TERCE). *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 9-32. doi:10.15366/reice2016.14.4.001

Lezotte, L. (1989). School improvement based on the effective school research. *International Journal of Educational Research, 13*(7), 815-825.

Lezotte, L., & McKee, K. (2011). *What effective schools do? Re-envisioning the correlates*. Bloomington: IN: Solution Tree Press.

Lewis, S., & Lingard, B. (2015). The multiple effects of international large-scale assessment on education policy and research, *Discourse: Studies in the Cultural Politics of Education*, *36*(5), 621-637. doi: 10.1080/01596306.2015.1039765

Lietz, P. , Cresswell, J. C., Rust, K. F. and Adams, R. J. (2017*). Implementation of Large-Scale Education Assessments. In Implementation of Large-Scale Education Assessments* (eds P. Lietz, J. C. Cresswell, K. F. Rust and R. J. Adams). doi:10.1002/9781118762462.ch1

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

Martinic, S., & Pardo, M. (2003). Aportes de la investigación educativa iberoamericana para el análisis de la eficacia escolar. In F. J. Murillo (Ed.), *La investigación sobre la eficacia escolar en Iberoamérica. Revisión Internacional sobre el estado del arte* (pp. 93-125). Bogotá: Convenio Andrés Bello.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). New York: Macmillan.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment*. Princeton, NJ: ETS. Recuperado de https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1993.tb01562.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1-3), 35-44. doi:10.1023/a:1006964925094

Messick, S., Beaton, A., & Lord, F. (1983). *National assessment of educational progress reconsidered: A new design for a new era* (NAEP-83-01). Princeton, NJ: National Assessment of Educational Progress.

Ministerio de Educación del Perú, & Organización de Estados Iberoamericanos para la Educación la Ciencia y la Cultura [OEI]. (1994). *Sistema educativo nacional del Perú.* Madrid: OEI.

Ministerio de Educación Unidad de Currículum y Evaluación. (2004). *Sistema de medición de la calidad de la educación. Informe de resultados*. Santiago de Chile: UCE-SIMCE.

Ministerio de Educación y Cultura del Ecuador, & Organización de Estados Iberoamericanos para la Educación la Ciencia y la Cultura [OEI]. (1994). *Sistema Educativo Nacional del Ecuador*. Madrid: OEI.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS.

Mislevy, R.J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characterstics from sparse matrix samples of item responses. *Journal of Education Measurement, 29*(2), 131-161. doi:10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R.J., Johnson, E.G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Kennedy, A. (2003). *PIRLS 2001 international report: IEA´s study of reading literacy achievement in primary school in 35 countries*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., & Loveless, T. (2016). *20 years of TIMSS international trends in mathematics and science achievement, curriculum, and instruction*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murillo, F.J. (2003a). El movimiento de investigación de eficacia escolar. In F. J. Murillo (Ed.), *La investigación sobre eficacia escolar en Iberoamérica. Revisión Internacional del estado del arte* (pp. 53-92). Bogotá: Convenio Andrés Bello.

Murillo, F.J. (2003b). Una panorámica de la investigación iberoamericana sobre eficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(1), 1-14.

Murillo, F.J. (2016). Midiendo la segregación escolar en América Latina. Un análsis metodológico utilizando el TERCE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 14*(4), 33-60. doi:10.15366/reice2016.14.4.002

Murillo, F.J., Cano, F., Castejón, J. L., Concha, C., Delprato, M. A., Ferräo, M. E., . . . Sancho, A. (2006). *Estudios sobre eficacia escolar en Iberoamérica. 15 Buenas Investigaciones*. Bogotá: Convenio Andrés Bello.

Murillo, F.J., & Duk, C. (2011). ¿Escuelas eficaces versus escuela inclusivas? *Revista Latinoamericana de Educación Inclusiva, 5*(1), 11-12.

Murillo, F.J., & Martínez-Garrido, C. (2016). *La educación en América Latina y el Caribe. Aportes del TERCE y sus Reanálisis*. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, *14*(4), 5-8.

Murillo, F.J., & Martínez-Garrido, C. (2017a). Estimación de la magnitud de la segregación escolar en América Latina. *Magis Revista Internacional de Investigación en Educación, 9*(19), 11-30. doi:10.1590/es0101-73302017167714

Murillo, F.J., & Martínez-Garrido, C. (2017b). Segregación social en las escuelas públicas y privadas en América Latina. *Educação & Sociedade, 38*(140), 727-750. doi:10.1590/ES0101-73302017167714

Oficina Regional de Educación para América Latina y el Caribe & Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [OREALC-UNESCO]. (2014). *El liderazgo escolar en América Latina y el Caribe. Un estado del arte con base en ocho sistemas escolares de la región*. Recuperado de: http://unesdoc.unesco.org/images/0023/002327/232799s.pdf

Oficina Regional de Educación para América Latina y el Caribe, Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [OREALC-UNESCO & LLECE]. (2008). *Eficacia escolar y factores asociados en América Latina y el Caribe*. Santiago de Chile: OREALC-UNESCO & LLECE.

Organisation for Economic Co-operation and Development [OECD]. (2002). *PISA 2000 Technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2005). *PISA 2003 Techincal report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2009). *PISA 2006 Techincal report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2012). *PISA 2009 Techincal report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2013). *PISA 2012 Results: Excellence through equity (Volume II): Giving every student the chance to succeed*. doi:/10.1787/9789264201132-en

Organisation for Economic Co-operation and Development [OECD]. (2014). *PISA 2012 Technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2016a). *PISA 2015 Results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2016b). *PISA 2015 Results (Volume II): Policies and practices for successful schools*. doi.org/10.1787/9789264267510-en

Organización de Estados Iberoamericanos para la Educación la Ciencia y la Cultura [OEI]. (1997). *Sistema educativo nacional de Costa Rica*. Madrid: OEI.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2010). *Compendio de los manuales del SERCE*. Recuperado de: http://unesdoc.unesco.org/images/0019/001919/191940s.pdf

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2013). *Situación educativa de América Latina y el Caribe: Hacia la educación de calidad para todos al 2015*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2014). *Revisión general 2015 de la educación para todos América Latina y el Caribe*. Recuperado de: http://unesdoc.unesco.org/images/0023/002327/232701s.pdf

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & La Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2016a). *Recomendaciones de políticas educativas en América Latina en base al TERCE*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & La Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC]. (2016b). *Reporte técnico tercer estudio regional comparativo y explicativo. TERCE*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [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.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2001). *Informe Técnico. Primer estudio internacional comparativo sobre lenguaje, matemática y factores asociados, para alumnos del tercer y cuarto grado de la educación básica*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2010). *SERCE. Factores asociados al logro cognitivo de los estudiantes de América Latina y el Caribe*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2016a). *Informe de resultados del tercer estudio regional comparativo y explicativo. Factores Asociados*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Oficina Regional de Educación para América Latina y el Caribe & Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación [UNESCO-OREALC-LLECE]. (2016b). *Informe de resultados del tercer estudio regional comparativo y explicativo. Logros de aprendizaje*. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO]. (2000). *Marco de acción de Dakar. Educación para todos cumplir nuestros*

compromisos comunes. Recuperado de http://unesdoc.unesco.org/images/0012/001211/121147s.pdf

Padilla, J. L., Gómez , J., Hidalgo, M. D., & Muñiz, J. (2006). La evaluación de las consecuencias del uso de los test en la teoría de la validez. *Psicothema, 18*(2), 307-312.

Plowden, B. (Ed). (1967). *Children and their primary schools: A report of the control advisory council for education (England)* (Vol. 1). London: Her Majesty's Stationery Office.

Programa de Promoción de la Reforma Educativa en América Latina y el Caribe [PREAL]. (2006). *Informe de progreso educativo República Dominicana*. Santo Domingo: PREAL-EDUCA.

Rodríguez, G. (1991). *Investigación evaluativa en torno a los factores de eficacia escolar de los centros públicos de EGB.* (Tesis doctoral inedita, Facultad de Filosofía y Ciencias de la Educación UNED, Madrid, España).

Ruhe, V., & Zumbo, B. (2009). *Evaluation in distance education and e-learning*. New York: The Guilford Press.

Rutter, M., Moughan, B., Mortimore, P., & Ouston, J., (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.

Sans-Martín, A., Guàrdia-Olmos, J., & Triadó-Ivern, X. (2016). El liderazgo educativo en Europa: Una aproximación transcultural. *Revista de Educación, Enero-Marzo*(371), 78-99. doi:10.4438/1988-592X-RE-2015-371-309

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness a critical review of the knowledge base*. Dordrecht: Springer. doi: 10.1007/978-94-017-7459-8

Secretaría de Educación Pública, & Organización de Estados Iberoamericanos para la Educación la Ciencia y la Cultura [OEI]. (1994). *Sistema educativo nacional de México*. Madrid: OEI.

Smith, W. (2014). *The global expansion of the testing culture: National tetsing policies and reconstruction of education.* (Doctoral dissertation, Pennsylvania State University, United States). Recuperado de https://etda.libraries.psu.edu/files/final_submissions/10282

Southeast Asia Primary Learning Metrics [SEA-PLM]. (2016). *Southeast Asia primary learning metrics. Global citizenship domain assessment framework*. Melbourne: Australian Council for Educational Research [ACER]. Recuperado de https://drive.google.com/file/d/0Bw2j_6JUc3aWZ2NXNURSRW8yZjJnanhTX0wzSzNqd0V0SmVz/view

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

Stufflebeam, D.L. (1983). The CIPP model for program evaluation. In Madaus, G.F., Seriven, M., S., Stufflebeam, D.L. (Eds.), *Evaluation models. Viewoints on educational and human services evaluation*, (pp.117-141). doi: 10.1007/978-94-009-6669-7_7

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp. 55-133). London: Falmer Press.

Tian, H., Sun, Z. (2018) *Academic achievement assessment.* doi: 10.1007/978-3-662-56198-0_4

Ting, L., & Ronald, L. (2017). School segregation policy and its educational ramifications for internal migrant children in urban China. *Asian Journal of social science studies, 2*(2), 1-10. doi:10.20849/ajsss.v2i2.158

Towsend, T. (Ed) (2007). *International handbook of school effectiveness and improvement*. Dordrecth: Springer.

United Nations Educational Scientific and Cultural Organization [UNESCO]. (2016). *Education for people and planet*. Recuperado de http://unesdoc.unesco.org/images/0024/002457/245752e.pdf

United Nations Educational Scientific and Cultural Organization [UNESCO]. (2010). *World data on education* (Vol. VII Ed. 2010/11). Buenos Aires: UNESCO.

Volante, L. (2007). *Evaluating test-based accountability systems: An international perspective*. Paper presented at the Association for Educational Assessment - Europe, Stockholm, Sweden. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.9825&rep=rep1&type=pdf

Webb, A., Canales, A., & Becerra, R. (2017). Capítulo IX las desigualdades invisibilizadas: Población indígena y segregación escolar. En I. Irarrázaval, E. Piña, & M. Letelier

(Eds.), P*ropuestas para Chile concurso de políticas públicas 2016* (pp. 279-305). Santiago de Chile: Pontificia Universidad Católica de Chile.

Weber, G. (1971). *Inner-city children can be taught to read: Four successful schools.* Washington, DC:Council for Basic Education.

Woitschach, P. (2016). *Effect of the indigenous language on educational assessments test.* Poster presented to the 10th International Test Commission [ITC] - Improving Policy and Practice: Opportunities and Challenges in an International Context, Vancouver, Canada.

Zumbo, B.D., & Forer, B. (2011). Testing and Measurement from a Multilevel View: Psychometrics and Validation. In James A. Bovaird, Kurt F. Geisinger, & Chad W. Buckendahl (Editors). *High Stakes Testing in Education - Science and Practice in K-12 Settings*, (pp.177-190). American Psychological Association Press, Washington, D.C..

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.

Zumbo, B. D., & Hubley, A. M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice, 23*(2), 299-303. doi:10.1080/0969594X.2016.1141169

Zumbo, B.D., Liu, Y., Wu, A., Shear, B., Olvera-Astivia, O., & Ark, T. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136-151. doi:10.1080/15434303.2014.972559

ANEXOS

Facultad de Psicología
Universidad Complutense de Madrid
Vicedecanato de Investigación, Doctorado e Infraestructuras

Autorización para utilizar una publicación en la tesis

Programa de Doctorado en Psicología – RD 99/2011

Doctorando/a: Pamela Raquel Woitschach Mendoza

Título de tesis: EVALUACIONES EDUCATIVAS A GRAN ESCALA EN LATINOAMÉRICA: TERCE

Director 1 y tutor: Rosario Martinez-Arias

Director 2: Rubén Fernández-Alonso

Director 3: José Muñiz-Fernández

Los abajo firmantes, en calidad de coautores del trabajo

Análisis de la Oportunidad de Aprendizaje en el estudio TERCE de la UNESCO publicado en la Revista de Investigación Educativa, 36(2), 509-528. (i) autorizan que este trabajo sea presentado por el doctorando/a para su tesis doctoral, (ii) declaran que el trabajo no ha sido presentado en ninguna otra tesis doctoral en la que los firmantes estén involucrados y (iii) se comprometen a no utilizarlo en ninguna otra tesis doctoral en la que estén involucrados.

Deberán firmar todos los co-autores del trabajo, consignando debajo su nombre completo y fecha de firma.

RUBEN FERNANDEZ ALONSO
Oviedo, 7 de VI de 2018

MARCOS ÁLVAREZ DÍAZ
Oviedo, 7 de junio de 2018

PdI SAMUEL FERNANDEZ
Dpt CC. EDUCACION
UNIVERSIDAD DE OVIEDO

Tabla 1.

Distribución de las principales evaluaciones nacionales e internacional en países de América Latina

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Argentina	T-N	N	LL-N	N	N	P-N	PIR-N	N	T-N		N	P-LL	N		P	N		P	LL-N
Brazil	N		LL-N		N	P	N		P-N		N	P-LL	N		P-N		N	P	LL
Chile			LL-N	N	I-T-N	P-N	N	N	T-N	N	N	P-LL-N	N	N	P-I-N	N	T-N	P-N	LL
Colombia	T		LL-N		I	N	PIR		N			P-LL-N	T		P-I-N		PIR	P-N	LL
Costa Rica			LL	N								LL		N	P-N		N	P	LL-N
Dominican Republic			LL									LL			I			N	LL-N
Ecuador		N	N			N						LL	N	N	N	N			LL
Guatemala		N	N	N	N	N					N	LL-N	N	N	I-N	N	N	N	LL-N
Honduras			LL-N	N	N	N	N	N	N	N			N	N	N	N	PIR-T-N	N	LL-N
Mexico	T		LL			P			P		N	P-LL-N	N	N	P-I-N	N	N	P-N	LL-N
Nicaragua								N				LL-N							LL
Panama					N						N	LL-N		N	P				LL-N
Paraguay		N	LL			N				N		LL-N			I	N			LL
Peru		N	LL		N	P	N			N		LL-N		N	P-N	N	N	P-N	LL
Uruguay		N			N			N	P		N	P-LL			P-N	N	N	P-N	LL-N

Nota: T- TIMMS, *N- National, LL- Llece, P-PISA, PIR-PIRLS, I-Cived.*

Tabla 2.

Distribución de parámetros no invariantes según tipo de codificación en ítems de crédito parcial

Booklet 1 invariant and non-invariant Intercepts/Thresholds		
Fully correct 0=0; 1and2=1	IT1_15$1	**(1)** 2 3 4 5 6 **(7)** 8 **(9)** 10 11 12 **(13)** 14 15 16
	IT1_30$1	1 **(2)** 3 4 5 6 **(7)** 8 9 10 11 12 13 14 15 **(16)**
Fully incorrect 0and1=0; 2=1	IT1_15$1	1 2 **(3) (4) (5) (6)** 7 **(8)** 9 **(10) (11)** 12 13 14 15 **(16)**
	IT1_30$1	1 2 3 4 5 6 (7) 8 9 10 11 12 13 14 15 16
Partial Credit 0=0; 1=1; 2=1	IT1_15$1	1 2 3 4 **(5)** 6 **(7)** 8 **(9)** 10 11 12 13 14 15 16
	IT1_15$2	1 2 3 **(4) (5) (6)** 7 **(8)** 9 **(10) (11)** 12 13 14 15 16
	IT1_30$1	**(1) (2)** 3 4 5 6 **(7)** 8 9 10 11 12 13 14 15 16
	IT1_30$2	1 **(2)** 3 4 5 6 **(7)** 8 9 10 11 12 13 14 15 16
Booklet 3 invariant and non-invariant Intercepts/Thresholds		
Fully correct 0=0; 1and2=1	IT2_15$1	1 2 **(3)** 4 5 6 7 8 9 10 11 12 13 14 15 16
Fully incorrect 0and1=0; 2=1	IT3_15$1	1 2 3 4 **(5)** 6 7 **(8)** 9 **(10)** 11 12 13 14 15 **(16)**
Partial Credit 0=0; 1=1; 2=1	IT3_15$1	1 2 3 4 **(5)** 6 **(7) (8)** 9 10 11 12 13 14 15 **(16)**
	IT3_15$2	1 2 3 4 **(5)** 6 7 **(8)** 9 10 11 12 13 14 15 **(16)**
Booklet 4 invariant and non-invariant Intercepts/Thresholds		
Fully correct 0=0; 1and2=1	IT4_26$1	**(1)** 2 3 4 5 6 **(7)** 8 9 10 11 12 13 14 15 16
Fully incorrect 0and1=0; 2=1	IT4_26$1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Partial Credit 0=0; 1=1; 2=1	IT4_26$1	**(1)** 2 **(3)** 4 5 6 **(7)** 8 9 10 11 **(12)** 13 14 15 16
	IT4_26$2	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Booklet 5 invariant and non-invariant Intercepts/Thresholds		
Fully correct 0=0; 1and2=1	IT5_15$1	1 **(2)** 3 4 **(5)** 6 **(7)** 8 9 10 11 12 13 14 15 16
	IT5_28$1	1 2 3 4 5 6 7 **(8)** 9 10 11 12 13 14 15 16
Fully incorrect 0and1=0; 2=1	IT5_15$1	1 **(2)** 3 4 **(5)** 6 **(7)** 8 9 10 11 12 13 14 15 16
	IT5_28$1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Partial Credit 0=0; 1=1; 2=1	IT5_15$1	**(1) (2)** 3 4 **(5)** 6 **(7)** 8 9 10 11 12 13 14 15 16
	IT5_15$2	1 **(2)** 3 4 **(5)** 6 **(7)** 8 9 10 11 12 13 14 15 16
	IT5_28$1	1 2 3 4 5 6 **(7) (8)** 9 10 11 12 13 14 15 16
	IT5_28$2	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Nota: Entre paréntesis parámetros no invariantes, la numeración corresponde a los países participantes.