**Propósitos y Representaciones**
ISSN 2307-7999
e-ISSN 2310-4635

Dic. 2019, Vol. 7, N° SPE, e347
Número especial: Evaluación psicológica: Perspectivas y desafíos de la psicología educativa
http://dx.doi.org/10.20511/pyr2019.v7nSPE.347

**ARTÍCULOS DE INVESTIGACIÓN**

# Calidad de los sistemas educativos: Modelos de evaluación

## Quality of Education Systems: Evaluation Models

**Rubén Fernández-Alonso***
Consejería de Educación del Gobierno del Principado de Asturias, Oviedo, España
Universidad de Oviedo, Oviedo, España
ORCID: http://orcid.org/0000-0002-7011-0630

**José Muñiz**
Universidad de Oviedo, Oviedo, España
ORCID: http://orcid.org/0000-0002-2652-5361

**Recibido** 01-06-19 **Revisado** 13-08-19 **Aprobado** 29-10-19 **En línea** 07-11-19

***Correspondencia**

Email: fernandezaruben@uniovi.es

**Citar como:**

Fernández-Alonso, R., & Muñiz, J. (2019). Calidad de los sistemas educativos: Modelos de evaluación. Propósitos y Representaciones, 7(SPE), e347. doi: http://dx.doi.org/10.20511/pyr2019.v7nSPE.347

## Resumen

El grupo de Investigación Psicometría de la Universidad de Oviedo participó en diferentes ponencias y talleres organizados en el marco del II Congreso Internacional de Evaluación Psicológica celebrado en noviembre de 2018 en la Universidad San Ignacio De Loyola (Lima, Perú). El presente trabajo recoge parte de aquellas aportaciones. En concreto el objetivo de este escrito es presentar los modelos matemáticos y procedimientos metodológicos disponibles para el análisis de los datos en las evaluaciones de sistemas educativos. El diseño, ejecución y diseminación de resultados de un programa de evaluación de sistema educativo es una tarea compleja que supone un desafío en diferentes ámbitos, entre los que destaca el análisis de datos. Estos programas tienen dos grandes finalidades: conocer y describir el nivel de conocimientos y competencias de la población de estudiantes e identificar y analizar los factores de contexto y proceso asociados a los resultados educativos. Para cumplir ambas finalidades la evaluación de sistemas educativos se ha dotado de soluciones metodológicas singulares y específicas. En este escrito se presentan tres de ellas. Dos están orientadas a expresar los resultados del aprendizaje: valores plausibles y métodos de punto de corte, mientras que la última está centrada en analizar la relación entre los factores escolares y los resultados.

**Palabras clave:** Evaluación de sistemas educativos; Teoría de Respuesta al ítem; Valores Plausibles; Modelos jerárquico-lineales; Eficacia escolar.

## Summary

The Psychometric Research Group of the University of Oviedo participated in different presentations and workshops organized within the framework of the II International Congress of Psychological Evaluation held in November 2018 at the San Ignacio De Loyola University (Lima, Peru). This work gathers part of those contributions. Specifically, the aim of this paper is to present the mathematical models and methodological procedures available for the analysis of data in the evaluation of educational systems. The design, execution and dissemination of the results of an evaluation program of the education system is a complex task that poses a challenge in different areas, among which data analysis stands out. These programs have two main purposes: to know and describe the level of knowledge and skills of the student population and to identify and analyze the context and process factors associated with educational outcomes. In order to fulfil both purposes, the evaluation of education systems has been provided with unique and specific methodological solutions. Three of them are presented in this paper. Two are aimed at expressing learning outcomes: plausible values and cut-off methods, while the last focuses on analyzing the relationship between school factors and outcomes.

**Keywords:** Evaluation of Education Systems; Item Response Theory; Plausible Values; Hierarchical-linear Models; School Effectiveness.

## Introducción

La evaluación de sistemas educativos cumple en 2019 sesenta años. En junio de 1959, bajo el auspicio de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), se inició el *Twelve-Contry Study*, un programa de cooperación transnacional que exploró la posibilidad de realizar equiparaciones internacionales rigurosas del rendimiento académico y que se considera el primer ensayo mundial de evaluación del desempeño escolar. Como corresponde a una etapa de gestación el *Estudio Piloto 12-países* señaló las limitaciones y desafíos a los que se enfrentaba la evaluación de sistemas educativos (traducción y adaptación cultural de las pruebas, logística de la aplicación, comparabilidad de resultados…), pero también apuntó que, bajo ciertas condiciones, la comparación era posible (Foshay, Thorndike, Hotyat, Pidgeon & Walker, 1962).

Desde entonces, la evaluación de sistemas educativos ha crecido y se ha generalizado. La primera evaluación nacional, *The National Assessment of Educational Progress* (NAEP), fue organizada en 1969 por el Departamento de Educación de los Estados Unidos de Norteamérica. En Latinoamérica la mayoría de los países comenzaron a evaluar sus sistemas educativos en la década de los 90, si bien en algunos casos (v. g., Chile, México o Costa Rica) el inicio es anterior (Woitschach, 2018). En la actualidad se desarrollan varios estudios de alcance mundial: *Trends in International Mathematics and Science Study* (TIMSS), *Programme for International Student Assessment* (PISA), *Progress in International Reading Literacy Study* (PIRLS), *International Civic and Citizenship Education Study* (ICCS), e *International Computer and Information Literacy Study* (ICILS). También se cuenta con evaluaciones internacionales de ámbito regional. En Latinoamérica y el Caribe destacan los estudios del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), si bien es posible citar ejemplos en cada continente. En África, *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SAMEQ) y *Programme for the Analysis of Education Systems* (PASEC); en Asia, *Southeast Asia Primary Learning Metrics* (SEA-PLM); y en Oceania, *Pacific Islands Literacy and Numeracy Assessment* (PILNA).

Pese a su diversidad todos ellos persiguen dos finalidades: conocer y describir el nivel de conocimientos y competencias de la población de estudiantes, ya sea en un momento puntual o a lo largo de la escolarización; e identificar y analizar los factores de contexto y proceso asociados a los resultados educativos (Fernández-Alonso, 2004). El objetivo de este escrito es presentar las principales soluciones metodológicas y analíticas para ese doble propósito. Por ello, el trabajo se organiza en dos apartados: en el primero se mostrarán los modelos matemáticos y métodos destinados a estimar y describir los resultados del aprendizaje y en el segundo se recrearán los modelos para analizar los factores asociados al desempeño escolar.

## Calidad de los sistemas educativos: Modelos de evaluación

La evaluación del sistema de educativo tiene dos modos de informar de los resultados del alumnado. Por un lado, las puntuaciones promedio (*scale scores*) agregadas a nivel poblacional, de estratos u otras variables de interés (demográficas, tipo de centro, etc.). Son puntuaciones sintéticas que permiten comparaciones entre grupos y, por ello, suelen tener impacto mediático. Los resultados también se presentan como niveles de rendimiento (*achievement levels*), que son estándares de desempeño que describen los conocimientos y competencias de la población.

### Los resultados expresados como puntuaciones promedio

En sus inicios los programas de evaluación de sistemas educativos expresaban los resultados cognitivos del alumnado empleando los fundamentos de la Teoría Clásica del Test (Foshay et al., 1962). Sin embargo, para mantener una adecuada validez de contenido, estas evaluaciones manejan un gran número de ítems lo que obliga a distribuirlos en diferentes modelos de cuadernillos siguiendo los principios del diseño experimental (Adams & Wu, 2002; Allen, Carlson & Zelenak, 1999; Allen, Donoghue & Schoeps, 2001; Beaton, 1987; Fernández-Alonso & Muñiz, 2011; Frey, Hartig & Rupp, 2009; Mullis, Martin, Kennedy, Trong & Sainsbury 2009; Olson, Martin & Mullis, 2008). En el momento de la aplicación de la prueba cada estudiante sólo responde a un modelo de cuadernillo, es decir, se enfrenta a una submuestra del banco completo, con el agravante de que los cuadernillos distan mucho de ser perfectamente paralelos (Lord, 1962). En este contexto los modelos clásicos son inapropiados para informar de los resultados de los estudiantes (Muñiz, 1997, 2018). Estas limitaciones en la equiparación y comparabilidad de los resultados no fueron solventadas hasta el último cuarto del siglo XX, cuando el programa NAEP empleó por primera vez modelos matemáticos derivados de la Teoría de Respuesta a los Ítems (TRI, Beaton, 1987; Bock, Mislevy & Woodson, 1982; Messick, Beaton & Lord, 1983) que, desde entonces, es la aproximación dominante para expresar los resultados cognitivos en las evaluaciones de sistemas educativos.

Los modelos TRI son funciones logísticas que estiman la competencia o habilidad del alumnado en la variable evaluada en función de sus respuestas a un conjunto de ítems y de los parámetros o propiedades métricas de dichos ítems. Su formulación matemática es la siguiente (Mazzeo, 2018):

(1) $$p(u_p; \beta|\Theta)$$

Donde, $\Theta \equiv (\Theta_1, \Theta_2, \dots, \Theta_m)$ es el vector de competencia o habilidad del estudiante $p$ condicionado por su vector o patrón de respuestas a los ítems de la prueba $u_p \equiv (u_{p1}, u_{p2}, \dots, u_{pn})'$ y por el vector de los parámetros de los ítems $\beta = (\beta_1, \beta_2, \dots, \beta_i)'$. El número de parámetros de los ítems determina el modelo TRI empleado en cada estudio. Por ejemplo, LLECE, SACMEQ, PILNA, ICCS y PISA (en este caso hasta el año 2012) combinan el modelo de Rasch para los ítems dicotómicos y el modelo de crédito parcial para los ítems politómicos (Adams & Wu, 2002; Hungi, 2011; Martin & Kelly, 1997; Pacific Community, 2016; Schulz, Carstens, Losito & Fraillon, 2018; UNESCO-Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC], 2016a). A partir del año 2015 PISA combina el modelo de Birnbaum (2-parámetros) para los ítems binarios y el modelo de crédito parcial generalizado de Muraki para ítems con tres o más categorías (Organisation for Economic Cooperation and Development [OECD], 2017). Por su parte, NAEP, TIMSS y PIRLS conjugan tres modelos según el formato de los ítems: 3-parámetros para ítems de elección múltiple, 2-parámetros para ítems abiertos binarios y el modelo de Muraki para ítems politómicos (Martin, Mullis & Hooper, 2016, 2017; National Center for Education Statistics [NCES], 2018).

Los modelos TRI presentan indudables ventajas sobre la aproximación clásica (Muñiz, 2018). En contraprestación son menos intuitivos ya que la escala de puntuaciones ($\Theta$) está indeterminada, se mueve entre extremos infinitos. Para solventar la indeterminación los resultados se ofrecen en puntuaciones transformadas. La más conocida expresa los resultados en una escala normal con media 500 puntos y desviación típica 100 [*N(500,100)*] (Hungi, 2011; Martin et al., 2016, 2017; OCDE, 2017), si bien otros valores son posibles (NCES, 2018; Pacific Community, 2016; UNESCO-OREALC, 2016a; UNESCO-OREALC, & LLECE, 2016a).

Tradicionalmente, la evaluación psicoeducativa calcula la función (1) empleando estimadores puntuales de máxima verosimilitud ponderada o procedimientos bayesianos (Muñiz, 1997, 2018). Sin embargo, en la evaluación de sistemas educativos las puntuaciones individuales no tienen interés. Mazzeo (2018) denomina estos estudios como *group-score assessement* para resaltar que su objetivo es estimar y comparar parámetros poblacionales (v. g., promedios por país) y no evaluar desempeños individuales como ocurre en la mayoría de la investigación educativa y psicológica. Además, está demostrado que los estimadores puntuales presentan sesgos al recuperar los parámetros poblacionales (Beaton, 1987; Mislevy, Beaton, Kaplan & Sheehan, 1992; von Davier, Gonzalez & Mislevy, 2009). Por ello, la evaluación de sistemas educativos ha desarrollado un procedimiento singular y específico para reportar los resultados cognitivos: los valores plausibles (VP).

Un VP puede definirse como *una muestra aleatoria tomada de una función de densidad multivariante a posteriori que contiene la distribución de las probabilidades que tiene un estudiante de obtener una puntuación en la materia evaluada en función de sus respuestas a un banco parametrizado de ítems y de sus características sociodemográficas y personales*. Matemáticamente el modelo se expresa del siguiente modo (Mazzeo, 2018):

(2) $$f(\Theta|u_p, x_p) \propto p(u_p; \hat{\beta}|\Theta)\emptyset(\Theta; \Gamma' x_p, \Sigma)$$

El término $f(\Theta|u_p, x_p)$ representa la función de densidad a posteriori del nivel de competencia del estudiante $(\theta)$ condicionado por sus respuestas a los ítems $(u_p)$ y sus características sociodemográficas y personales $(x_p)$. Esta función de densidad recoge la distribución de puntuaciones probables del estudiante y es el producto de dos distribuciones de probabilidad. Por un lado, un modelo TRI visto en (1) que estima el nivel de competencia del alumnado condicionado por sus respuestas a unos ítems de parámetros conocidos $[p(u_p; \hat{\beta}|\Theta)]$ y, por otro, de un modelo de estructura poblacional $[\emptyset(\Theta; \Gamma' x_p, \Sigma)]$ donde $\Gamma'$ es la matriz de coeficientes de regresión de las variables socio-demográficas poblacionales sobre los resultados y $\Sigma$ es la matriz de varianza-covarianza de las características de la población. Por tanto, el segundo término del producto es una función de densidad continua que estima la probabilidad de que un estudiante tenga un determinado nivel de competencia condicionada por sus características sociodemográficas y personales, el efecto que a nivel poblacional tienen esas características sobre el desempeño y la relación que existe entre las variables empleadas para definir dichas características. Por características sociodemográficas se entienden variables como género, edad y nivel socioeconómico y cultural del estudiante, resultados promedio de la escuela, así como otros factores extraídos del análisis de componentes principales de las respuestas a los cuestionarios de contexto (Martin et al., 2016, 2017; Mazzeo, 2018; OECD, 2017; NCES, 2018; UNESCO-OREALC, 2016a).

El procedimiento de estimación de los PV se desarrolla en dos fases. La primera es similar al ajuste de un banco de ítems ordinario: se selecciona un número casos sin ponderar igual para todos los grupos (v. g., misma *N* de estudiantes para todos los países o estratos) que funciona como muestra de calibración. A partir del vector de respuestas, se calculan los parámetros de los ítems empleando algún procedimiento de estimación puntual. En la segunda fase se trabaja con todos los casos y sus correspondientes pesos muestrales para estimar la función de densidad a posteriori de la que se extraerán aleatoriamente los VP. En esta fase la matriz incluye el vector de las respuestas a los ítems de los estudiantes y toda la información sobre sus características sociodemográficas y de los factores extraídos del análisis de componentes principales. Los parámetros de los ítems de la primera fase se fijan como información previa para todos los grupos y las variables sociodemográficas funcionan como covariables en un modelo de regresión múltiple. La estimación de la función de densidad multivariada se hace por separado para grupo (país o estrato muestral) de modo que los parámetros de los ítems se mantienen constantes en todos los países, y son complementados por el efecto específico de las covariables sobre las puntuaciones en cada país o estrato. La descripción de la lógica y los fundamentos de los VP puede consultarse en Mazzeo (2018), NCES (2018) y von Davier et al. (2009) y el detalle para su ejecución en Wu, Adams, Wilson y Haldane (2007).

De la función de densidad estimada para cada estudiante, tal y como se acaba de describir, se toman aleatoriamente un número determinado de VP, entre 5 y 20, que son puntuaciones probables del estudiante (OECD, 2017; Martin et al., 2016, 2017; NCES, 2018). La figura 1 muestra las funciones de densidad de dos estudiantes que respondieron a los mismos ítems y cuyas características sociodemográficas y personales son similares. El estudiante 2 acertó más ítems que el estudiante 1 y por ello su función de densidad se ubica más a la derecha en la escala *N(500,100)*. No obstante, los valores probables de cada estudiante presentan un rango muy amplio. En este ejemplo se extraen aleatoriamente 5 VP para cada estudiante. Nótese que, en general, los valores probables del estudiante 2 son más altos que los del estudiante 1. Sin embargo, el VP-2 del estudiante 2 (en torno a 480 puntos) es inferior al VP-4 del estudiante 1 (en torno a 520). Es por esta razón que los PV, al contrario que los estimadores TRI puntuales, no pueden emplearse para reportar resultados individuales y sólo se usan en la evaluación de sistemas educativos para describir parámetros poblacionales (Mazzeo, 2018; NCES, 2018).

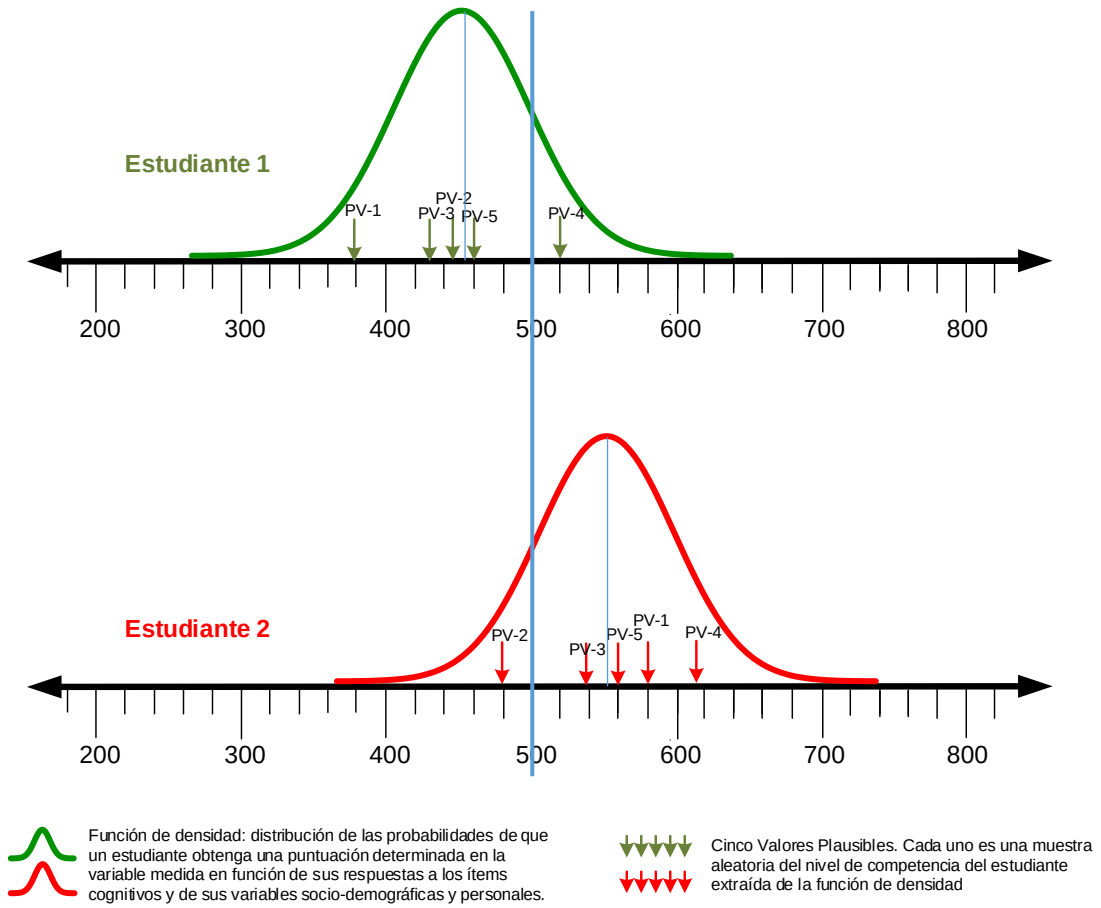


*Figura 1*. La lógica de asignación de puntuaciones: función de densidad a posteriori $[f(\theta|u_p, x_p)]$ y selección de valores plausibles de dos estudiantes.

Finalmente la puntuación de cada grupo (país, jurisdicción, estrato…) se expresa como el promedio de los VP. Sin embargo, el error típico de la media no se calcula como la razón entre la desviación típica y la raíz cuadrada del número de casos ya que los estudiantes no se seleccionan por un procedimiento aleatorio simple, sino mediante un muestreo por conglomerados en dos o más etapas. Ello obliga a emplear métodos de resmuestreo para calcular los errores típicos de los estimadores. El detalle de dicho cálculo puede consultarse en OECD (2009).

**Los resultados expresados como escalas cualitativas: los niveles de desempeño**

Las escalas de puntuación sintéticas que se acaban de describir resumen parámetros poblacionales y permiten ubicar y comparar desempeños grupales. Sin embargo, la puntuación numérica no informa de los conocimientos y competencias efectivamente alcanzadas por el alumnado (Fernández-Alonso, 2004). Para responder a este tipo de preguntas se emplea una metodología, denominada niveles de desempeño, cuya finalidad es establecer puntos de corte en la escala continua y analizar los conocimientos, destrezas y habilidades que demuestra el grupo estudiantes que supera un determinado nivel o punto de corte (Kelly, Mullis & Martin, 2000; Martin et al., 2016, 2017; UNESCO-OREALC, 2016a).

Es un procedimiento arbitrario, pero al tiempo muy práctico y eficiente. Su lógica es similar al uso que la industria textil hace de las medidas antropométricas (OECD, 2017; Servicio de Evaluación Educativa del Principado de Asturias, 2018a). Los rasgos físicos, igual que los resultados en una prueba cognitiva, son muy variables y se expresan en escalas numéricas y continuas. Por ejemplo, el ancho de la cadera de los hombres oscila normalmente entre 65 y 150

centímetros, es decir, en un rango de 85 centímetros. Sin embargo, los productores textiles colapsan o agrupan este rango en unas pocas categorías: talla S (entre 78 y 85 cm.); talla M (entre 86 y 94 cm.) y así sucesivamente. Salvando las distancias, determinar puntos de corte sigue la misma idea: arbitrar unos límites o intervalos en una escala continua para agrupar las puntuaciones en unos pocos niveles de desempeño. El procedimiento se ejecuta en dos grandes fases (Servicio de Evaluación Educativa del Principado de Asturias, 2018a, 2018b):

- Determinar puntos de corte en la escala de resultados para establecer grupos de desempeño o niveles de rendimiento y asignar ítems a dichos niveles.
- Elaborar descripciones que resuman las competencias del alumnado en cada uno de los niveles de desempeño.

***Determinar puntos de corte y asignar ítems a niveles de rendimiento.*** Existen diferentes procedimientos para establecer puntos de corte (Muñiz, 2018). En sus primeras ediciones TIMSS señalaba los puntos de corte *a priori* sobre la escala de percentiles (Kelly et al., 2000). En la actualidad, tanto TIMSS como PIRLS fijan cuatro cortes en la escala *N(500,100)*: 400, 475, 550 y 625 puntos, creando otros tantos grupos con el alumnado que obtiene una puntuación de ±5 puntos sobre dichas marcas (Martin y Mullis, 2012; Martin et al., 2016, 2017). Así, por ejemplo, el grupo Nivel Bajo está compuesto por los estudiantes que lograron entre 395 y 405 puntos (figura 2).

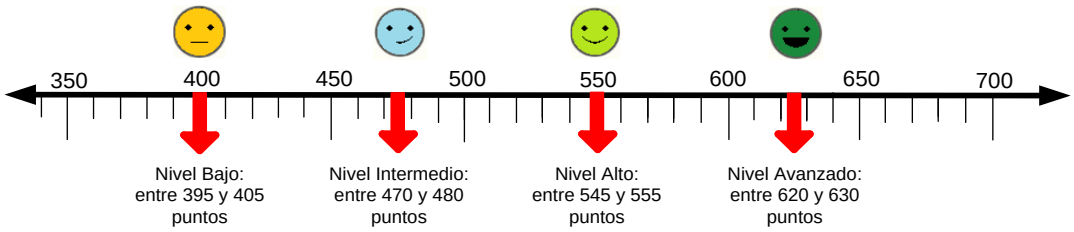


*Figura 2*. Puntos de corte y grupos de nivel de rendimiento en TIMSS y PIRLS
Fuente: Servicio de Evaluación Educativa del Principado de Asturias

A continuación, se compara el porcentaje de acierto de los cuatro grupos en cada uno de los ítems. Existen varios criterios para asignar los ítems a los niveles de rendimiento, si bien todos ellos se basan en el mismo principio: un ítem se asignará a un nivel de rendimiento cuando la mayoría de los estudiantes de dicho nivel (como mínimo el 60%) responden correctamente al ítem al tiempo que la mayoría de los estudiantes del nivel inmediatamente inferior lo falla (menos del 50% de aciertos). Por ejemplo, un ítem se asignará al Nivel Alto si es acertado por al menos el 60% de los estudiantes de dicho grupo y más de la mitad de los estudiantes del Nivel Intermedio lo falla.

Tomar unas puntuaciones *a priori* en la escala *N(500,100)* no es el único modo de establecer puntos de corte. Otros estudios como PISA, LLECE o PILNA determinan sus puntos de corte mediante el consenso de un grupo de expertos sobre las características de los ítems (Hungi et al., 2010; OECD, 2017; UNESCO-OREALC, 2016a). En este caso un panel de expertos trabaja con los ítems ordenados por su nivel de dificultad y colegiadamente acuerdan señalar los puntos de corte. En este punto las marcas más críticas son límites superior e inferior. Determinar el límite inferior supone identificar los ítems que preguntan por aspectos básicos y elementales, de modo que el alumnado que no responda acertadamente a los mismos pasará a formar parte del grupo de menor competencia. Para establecer el límite superior se deben aislar los ítems con mayor complejidad, aquellos que sólo podrán resolver los estudiantes avanzados o excelentes. Delimitadas las marcas extremas el rango de puntuación comprendido entre ambas se divide equitativamente en tantos puntos como grupos sean necesarios.

La figura 3 ejemplifica el procedimiento para establecer 6 niveles de rendimiento con un test de 20 ítems (Servicio de Evaluación Educativa del Principado de Asturias, 2018a). La parte central del gráfico muestra conjuntamente la distribución de resultados [*N(500,100)*] y los ítems ordenados por su dificultad. Como la puntuación del alumnado y la dificultad de los ítems están en la misma escala es posible predecir, por ejemplo, que el alumnado que obtenga 600 puntos tiene altas probabilidades de acertar los 15 ítems cuya dificultad está por debajo de esta marca (todos los comprendidos entre el ítem 05 y el ítem 18). Igualmente, también es más probable que ese alumnado falle los 5 ítems cuyo grado de dificultad está por encima de 600 puntos (ít09, 07, 20, 17 y 19).

En este ejemplo el grupo de expertos acordó que los ítems básicos eran el 12, 13 y 18. De los tres, el gráfico señala que el ítem 13 es el más difícil y, por tanto, el límite inferior se ubica inmediatamente por encima de la dificultad de dicho ítem, en este caso 360 puntos. Se dirá entonces que la probabilidad de acertar un ítem básico por parte de los estudiantes que obtienen menos de 360 puntos es inferior al puro azar ($p < 0,50$). En el caso del límite superior el panel de expertos consensuó que sólo el alumnado más competente acertará los ítems 17 y 19. Entre ambos, el ít17 es más fácil y, por tanto, el límite superior se establece por debajo del nivel de dificultad de este ítem (680 puntos). De este modo se predice que un estudiante que obtenga más de 680 puntos tendrá una probabilidad superior al azar ($p > 0.50$) de acertar un ítem muy complejo. Por tanto, 680 puntos marca la diferencia entre el alumnado avanzado o excelente y el resto de estudiantes evaluados. Como en este ejemplo se buscaba establecer seis niveles de rendimiento, el espacio entre las marcas inferior y superior se divide en cuatro partes de 80 puntos cada una. Los ítems que caen dentro de cada cuadrante son asignados a su respectivo nivel de rendimiento.

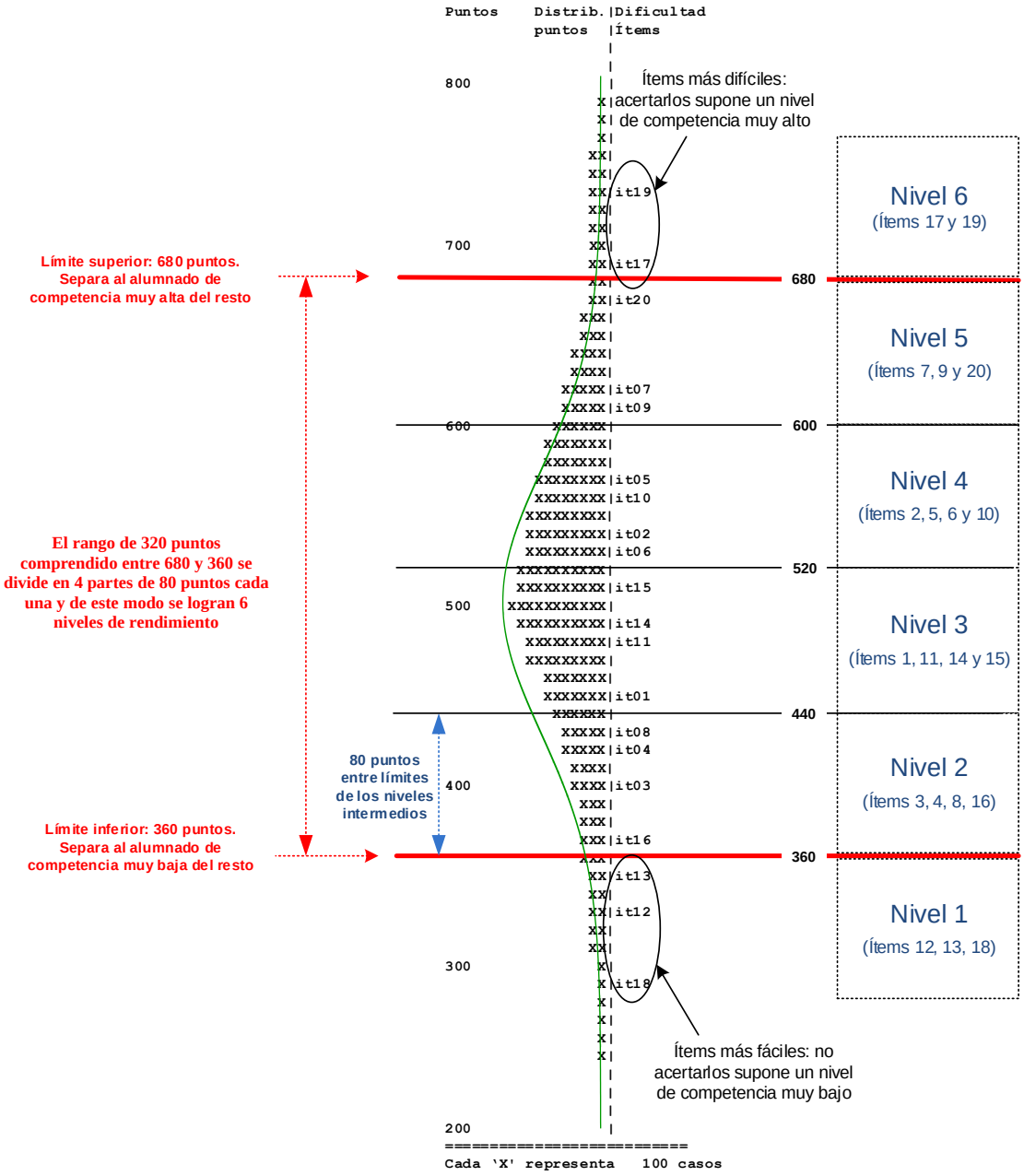


*Figura 3*. La lógica de la asignación de ítems a niveles de rendimiento en los métodos de acuerdo entre expertos
Fuente: Servicio de Evaluación Educativa del Principado de Asturias

***Elaborar descripciones que resuman las competencias.*** Con los ítems ordenados por su nivel de dificultad o asignados a los niveles de rendimiento un panel de expertos en la materia y curso evaluado analiza su contenido, encargándose de tres tareas (Nungi et al., 2010; OECD, 2017; UNESCO-OREALC, 2016a):

- Elaborar pequeñas descripciones que definan las destrezas y habilidades específicas necesarias para responder correctamente a cada uno de los ítems. Generalmente estas descripciones son muy puntuales y concretas puesto que se refieren a los conocimientos y procesos cognitivos que se ponen en juego en el momento de responder a ítems determinados.

- El conjunto de ítems de cada nivel conforma el abanico de competencias, conocimientos y destrezas del alumnado de dicho nivel. Por ello, la segunda tarea consiste en redactar una descripción general que resuma y caracterice cada uno de los niveles de desempeño.
- Seleccionar un grupo de ítems que ejemplifiquen las competencias propias de cada grupo o nivel de desempeño. Estos ítems serán liberados y publicados en los correspondientes informes de resultados.

Las descripciones contenidas en las escalas de competencia descritas tienen cuatro características (Servicio de Evaluación Educativa del Principado de Asturias, 2018a):

- Son jerárquicas e inclusivas: el modelo predice que el alumnado de un determinado nivel tiene altas probabilidades de responder acertadamente a los ítems de niveles inferiores.
- Son probabilísticas, no deterministas: no puede concluirse que todo el alumnado de un nivel responderá correctamente a los mismos ítems o que, por pertenecer al mismo nivel, su competencia real en la materia sea idéntica.
- Son empíricas: las descripciones contenidas en los niveles de rendimiento provienen primariamente de respuestas a ítems concretos y, por tanto, representan logros efectivos del alumnado.
- Tienen potencial para orientar la práctica educativa: por el modo en que se construyen los niveles tienen potencial para predecir los próximos aprendizajes que está en condiciones de abordar con garantías el estudiante.

## ¿Cómo se identifican y analizan los factores asociados a los resultados educativos?

La evaluación del sistema educativo tiene que ofrecer orientaciones de política educativa para la mejora escolar (Fernández-Alonso, 2004; UNESCO-OREALC, 2016b). Ello supone identificar y estudiar los elementos vinculados a los resultados educativos. Para recoger la información necesaria con la que sintetizar estos factores se aplican cuestionarios de contexto para alumnos, familias, profesorado, directivos escolares y, en ocasiones, autoridades educativas de los países participantes. Esta información permite construir variables simples e índices complejos. Las primeras reflejan hechos observables o comprobables documentalmente (v. g., género, edad, etc.) y se generan mediante recodificaciones y cálculos aritméticos. Los índices complejos resumen hechos no observables o variables latentes (v. g., actitudes y creencias personales, clima de aula, liderazgo pedagógico, etc.) y se construyen mediante análisis factoriales confirmatorios o modelos TRI (Adams & Wu, 2002; OECD, 2017; UNESCO-OREALC, 2016a).

Previo a la construcción de los índices y con el fin de organizar el análisis se desarrollan marcos teóricos basados en la aproximación sistémica (Adams & Wu, 2002; Mullis et al., 2002, 2009; Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas del Principado de Asturias, 2011). La figura 4 ejemplifica un marco teórico que funciona como una matriz de especificaciones de doble entrada para seleccionar y ubicar las variables e índices considerados en el análisis.

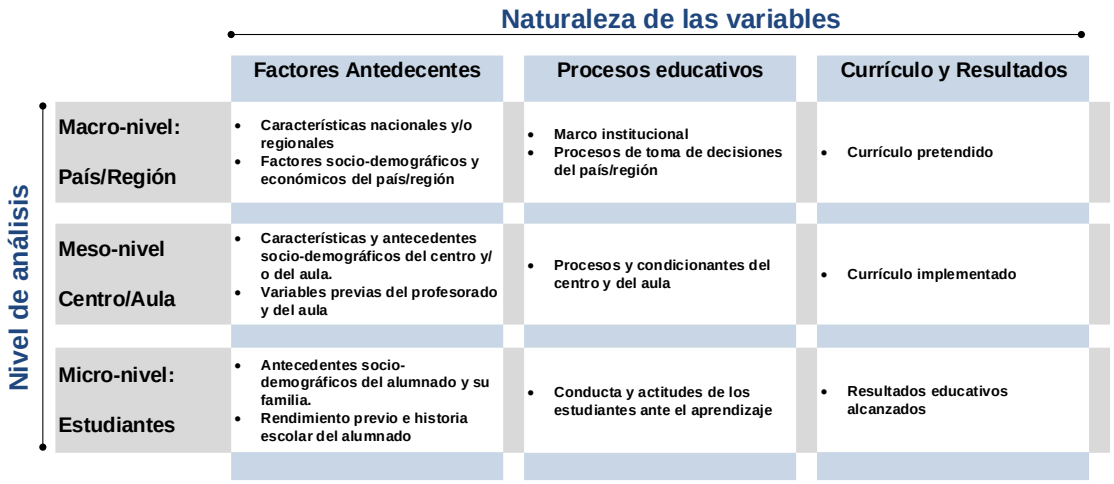


*Figura 4*. Marco teórico para un estudio de factores asociados

El eje de coordenadas distingue tres tipos de variables según su naturaleza: factores antecedentes y de contexto sociodemográfico que, por definición, son estables y poco permeables a la acción educativa; factores de proceso, que por su carácter moldeable tienen mayor potencial y capacidad de mejora escolar; y resultados educativos, entendidos en sentido amplio ya que incluyen resultados cognitivos, afectivos y otros productos deseables como la satisfacción de los usuarios con el servicio educativo (Muñoz-Repiso et al., 1995; Murillo, 2003). El segundo eje de la tabla señala que los datos presentan una estructura jerárquica o multinivel (Scheerens & Bosker, 1997; Scheerens, 2016): los estudiantes (micro-nivel o Nivel 1) se escolarizan en aulas, éstas conforman centros (meso-nivel o Nivel 2) y éstos se ubican en áreas geográficas dentro de un mismo sistema educativo (macro-nivel o Nivel 3).

El análisis de factores asociados debe ser coherente con el marco teórico, cuestión que se refleja, tanto en el tipo de modelos matemáticos empleados, como en la estrategia de ajuste y comparación de dichos modelos. Ambos aspectos son desarrollados a continuación.

**Modelos matemáticos en el análisis de factores asociados.**

En una estructura multinivel los estudiantes que comparten agrupaciones jerárquicas de orden superior (v. g., grupos-aula) tienden a ser más parecidos entre sí y sus desempeños más homogéneos que los de quienes no comparten dichas agrupaciones. En este contexto, no se puede mantener el supuesto de independencia de las observaciones que sustenta las soluciones analíticas del modelo general lineal. De hecho, los modelos de regresión múltiple clásicos presentan importantes limitaciones para analizar datos anidados, ya que infraestiman los errores de medida cuando no contemplan las estructuras jerárquicas de orden superior, o bien destruyen las diferencias internas de los grupos cuando eliminan las estructuras jerárquicas de orden inferior (Openshaw, 1982; Robinson, 1950).

Hace tres décadas se publicaron los primeros trabajos con modelos multinivel (Paterson & Goldstein, 1991), también denominados jerárquico lineales (Raudenbush & Bryk, 2002) o de coeficientes aleatorios (Longford, 1993). En conjunto conforman una familia de modelos matemáticos desarrollados específicamente para analizar datos de naturaleza compleja. En la actualidad su uso está generalizado porque son muy versátiles; pueden implementarse sobre variables criterio medidas en cualquier escala: continuas, ordinales, discretas o binarias; y permiten modular los datos en diseños longitudinales o de crecimiento, medidas repetidas, estudios experimentales con grupo

de control y estructuras de clasificación cruzada, por citar las aplicaciones más recurrentes en la investigación educativa (Hox, 1998; Raudenbush & Bryk, 2002).

En la evaluación de sistemas educativos los datos responden a la estructura de un diseño anidado, en el que los casos (estudiantes, Nivel 1) se agrupan en unidades de información más amplias (aulas, centros…, Nivel 2) y éstas a su vez en estructuras de orden superior (estratos poblacionales, regiones, países…, Nivel 3). En este tipo de diseño los modelos multinivel más empleados son el análisis de varianza de un factor de efectos aleatorios, análisis de regresión con medias como resultados, análisis de covarianza de un factor de efectos aleatorios, análisis de regresión con coeficientes aleatorios, y análisis de regresión con medias y pendientes como resultados (Gaviria Soto & Castro Morera, 2005; Pardo, Ruiz & San Martín, 2007; Raudenbush & Bryk, 2002).

Un modelo jerárquico-lineal puede entenderse como un modelo de regresión clásico con regresores a diferentes niveles y, por ello, el modelo de regresión clásico es un buen inicio para entender la lógica del análisis jerárquico-lineal (Gaviria Soto & Castro Morera, 2005). Supongamos que se quiere predecir el resultado de un estudiante en una prueba en función de su puntuación en un índice socioeconómico y cultural (ISEC). El modelo de regresión simple señala que la puntuación verdadera del estudiante ($y_i$) será un modelo aditivo de tres términos. Dos de ellos fijos y comunes para todos los casos: el intercepto ($\beta_0$) que es la puntuación esperada para los estudiantes cuyo ISEC ($X_i$) es igual a la media ($X_i - \bar{X}_{\cdot} = 0$); y la pendiente ($\beta_1$) que es la ganancia (o pérdida) esperada en el resultado por cada unidad que aumenta (o disminuye) el ISEC del estudiante ($X_i - \bar{X}_{\cdot} \neq 0$). El tercer término es el error de estimación ($\varepsilon_i$) que se supone aleatorio (Raudenbush & Bryk, 2002).

(3) $$y_i = \beta_0 + \beta_1(X_i - \bar{X}_{\cdot}) + \varepsilon_i$$

Ahora bien, un diseño jerárquico asume que los interceptos y pendientes no son fijos, sino que cada escuela tiene los suyos. Los interceptos varían porque los centros obtienen promedios de puntuación diferentes para estudiantes de igual ISEC, y las pendientes también varían porque las diferencias entre estudiantes de ISEC bajo y alto son más grandes en unos centros que en otros. Estas variaciones en pendientes e interceptos obligan especificar un modelo multinivel que introduce nuevos términos en la parte aleatoria de la ecuación. Para este ejemplo, donde sólo se incluye el ISEC de cada estudiante (medida de nivel 1) y no hay predictores en el nivel 2 se especifica un modelo de regresión con coeficientes (interceptos y pendientes) aleatorios, que matemáticamente se define como (Raudenbush & Bryk, 2002):

(4)
Nivel 1: $$y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{\cdot j}) + \varepsilon_{ij}$$
Nivel 2: $$\beta_{0j} = \gamma_{00} + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

Y que en su manera compacta queda del siguiente modo:

(5) $$y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\cdot j}) + [\mu_{0j} + \mu_{1j}(X_{ij} - \bar{X}_{\cdot j}) + \varepsilon_{ij}]$$

En este caso la predicción del resultado del estudiante *i* en la escuela *j* ($y_{ij}$) tiene una parte fija y otra aleatoria, esta última contenida en el corchete. Igual que en el modelo clásico la parte fija contiene dos términos, $\gamma_{00}$ y $\gamma_{10}(X_{ij} - \bar{X}_{\cdot j})$, cuyo significado es

similar al de $\beta_0$ y $\beta_1(X_{ij} - \bar{X}_{.j})$ en la ecuación (3). Sin embargo, ahora la parte aleatoria es más compleja. Además del error de estimación asociado al estudiante ($\varepsilon_{ij}$), se incluyen dos nuevos términos de variación: una asociada al hecho de que los centros tienen interceptos diferentes ($\mu_{0j}$) y una segunda variación porque el efecto del ISEC sobre el rendimiento $[\mu_{1j}(X_{ij} - \bar{X}_{.j})]$ es distinto en cada escuela.

La figura 5 representa gráficamente los términos de estas ecuaciones con un ejemplo ficticio pero muy plausible. El eje de ordenadas recoge las puntuaciones en la prueba en la escala *N(500,100)* y el eje de coordenadas las puntuaciones en el ISEC en una escala *N(0,1)*. La línea verde central es la recta de regresión que resume el efecto del ISEC sobre los resultados en el conjunto de la población (todos los centros y estudiantes): el intercepto general ($\gamma_{00}$) es igual a 500 puntos y la recta tiene una pendiente ($\gamma_{10}$) de 31º grados, que según la escala de este gráfico supone que por cada unidad que aumenta el ISEC se predicen 15 puntos de ganancia en el resultado de la prueba.

En la figura también se han seleccionado dos escuelas que tienen interceptos y pendientes distintas y se ha señalado un caso, que llamaremos el estudiante 7 de la escuela 2 ($y_{72}$), que logró 565 puntos en la prueba y cuyo ISEC es igual a 1 punto. Según la ecuación (3) el estudiante $y_{72}$ habría obtenido 50 puntos por encima del valor esperado en función de su ISEC individual, ya que:

$$y_{72} = \beta_0 + \beta_1(X_i - \bar{X}_{.}) + \varepsilon_i =$$
$$565 = 500 + 15(1 - 0) + \varepsilon_i =$$
$$\varepsilon_i = 565 - 515 = 50$$

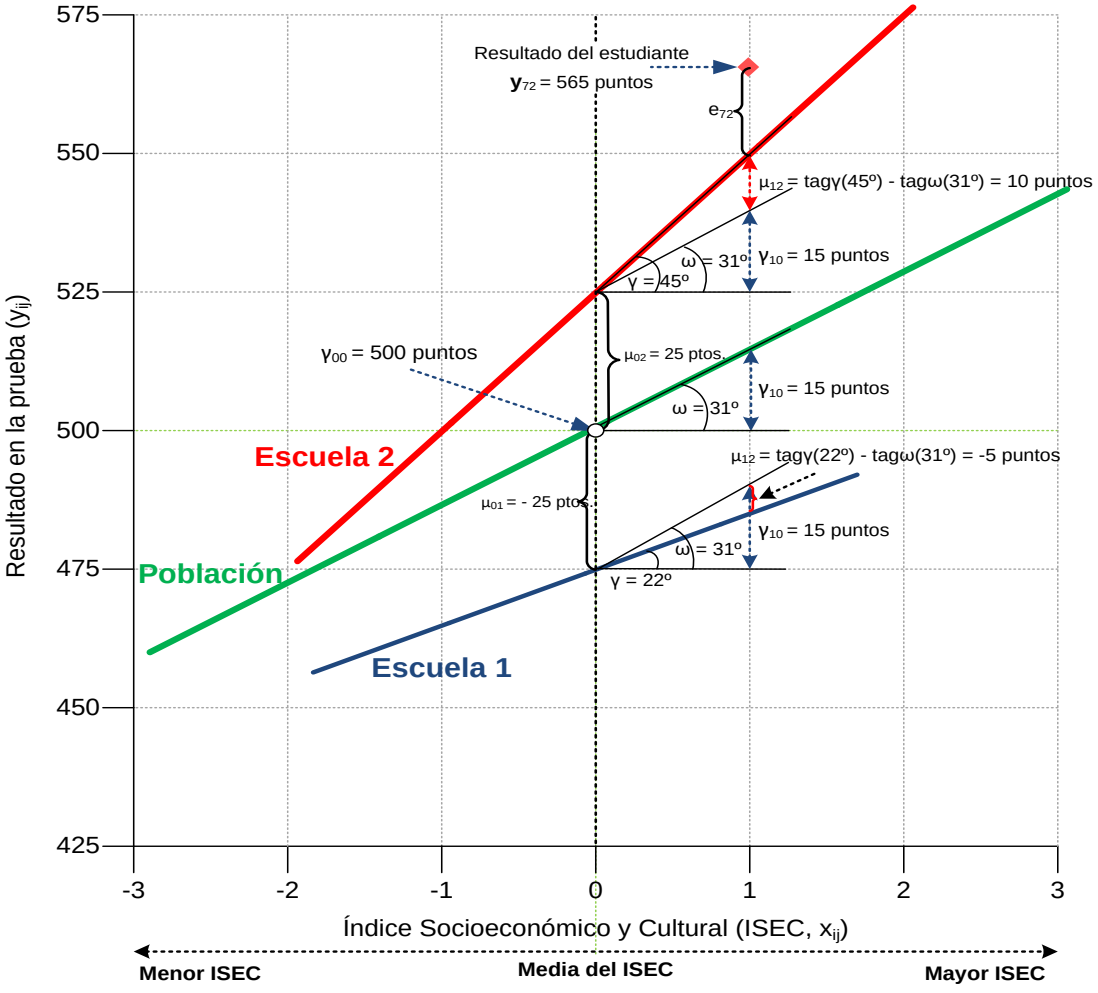

*Figura 5*. Representación en el plano de los términos de una ecuación multinivel

Sin embargo, el estudiante $y_{72}$ asiste a la Escuela 2 donde el intercepto y la pendiente que resumen la relación entre los resultados y el ISEC son diferentes al parámetro general y también al de otras escuelas. Por ello, el resultado del estudiante $y_{72}$ puede ser explicado no sólo por su ISEC, sino también por la escuela a la que asiste. En primer lugar, el intercepto de la Escuela 2 se sitúa en 525 puntos, es decir, el estudiante $y_{72}$ se escolariza en un centro cuyo promedio supera en 25 puntos la media poblacional ($\mu_{02} = 25$). Nótese que en el caso de la Escuela 1 la situación es la contraria: $\mu_{01} = -25$. Además, en la Escuela 2 la pendiente de regresión ($\mu_{12}$) tiene una inclinación de 45º, es decir, la ganancia en el resultado por cada unidad que aumenta el ISEC es mayor que los 15 puntos predichos por el modelo poblacional ($\gamma_{10}$). En concreto, para los valores escalares de este gráfico: $\mu_{12} = tag(45º) - tag(31º) = (1 \times 25) - (0{,}6 \times 25) = 10$. Por tanto, sustitudendo los valores de (5):

$$y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + [\mu_{0j} + \mu_{1j}(X_{ij} - \bar{X}_{.j}) + \varepsilon_{ij}] =$$
$$565 = 500 + 15(1 - 0) + [25 + 10(1 - 0) + \varepsilon_{ij}] =$$
$$\varepsilon_{ij} = 565 - (515 + 25 + 10) = 15$$

El ejemplo muestra que de los 50 puntos de error que el modelo de regresión simple imputa al estudiante, 25 puntos se explican porque éste asiste a una escuela con buenos resultados y otros 10 puntos adicionales porque en su escuela el efecto del ISEC sobre los resultados es mayor que el estimado para el conjunto de la población, por lo que finalmente el error relacionado con el estudiante queda reducido a 15 puntos. Esto ilustra una de las ventajas de los modelos jerárquico-lineales frente a la regresión clásica: permiten identificar y descomponer la varianza de resultados en diferentes niveles: individual, centro/aula y sistema educativo. En general los análisis señalan que las estructuras de orden superior acumulan menos varianza que los niveles inferiores, lo que es totalmente compatible con la realidad educativa: el desempeño escolar tiene un importante componente de motivación y esfuerzo individual, por lo que es esperable que los factores individuales expliquen gran parte de las diferencias. De igual modo es muy plausible que las variables de aula (clima ordenado, metodología docente, etc.) incidan mucho más en los resultados del alumnado que los factores de sistema educativo cuyo efecto sobre los resultados es siempre más indirecto (Woitschach, Fernández-Alonso, Martínez-Arias & Muñiz, 2017).

Por otro lado, el uso de los modelos jerárquico-lineales no sólo está recomendado por relacionar datos provenientes de patrones complejos de variabilidad. También ocurre que muchos fenómenos educativos son de naturaleza multinivel, es decir, una misma medida puede tener significados y presentar efectos distintos según el nivel de análisis en que sea considerada. Los deberes o tareas escolares en el hogar son un ejemplo de este tipo de variables (Trautwein, 2007). Supóngase preguntas del tipo: *con qué frecuencia haces tus deberes* o *cuánto tiempo te lleva hacerlos*. Analizadas a nivel individual las medidas reflejan el hábito de trabajo o dedicación del estudiante. Sin embargo, si las respuestas se promedian por aula la medida tiene un significado distinto, ya que describe la política de deberes escolares del profesorado, es decir, la frecuencia o la cantidad de deberes asignados. Además, los efectos sobre el rendimiento son diferentes según el nivel de análisis: en general se ha encontrado que el efecto del tiempo de deberes a nivel individual es negativo o, en el mejor de los casos, no significativo, mientras que la frecuencia o el tamaño de los deberes tienden a estar positiva y significativamente asociados a los resultados (Fernández-Alonso, Álvarez-Díaz, Suárez-Álvarez & Muñiz,

2017; Fernández-Alonso, Suárez-Álvarez & Muñiz, 2015, 2016; Fernández-Alonso et al., 2019; Trautwein, 2007).

**Ajuste y comparación de modelos para analizar factores asociados.**

La estrategia de ajuste de un análisis jerárquico-lineal comienza especificando modelos muy sencillos a los que se van añadiendo variables, mientras se mantienen aquellos factores significativos en los modelos previos. Ello permite comparar el incremento del porcentaje de varianza explicada por los modelos sucesivos y la mejora que experimentan los parámetros de ajuste con la introducción de nuevas variables. La especificación de los modelos debe ser coherente con el marco teórico del estudio (ver figura 4) y la estrategia es muy flexible ya permite establecer diversos modelos dependiendo de los objetivos del estudio y de las variables de interés. No obstante, en los análisis de factores asociados hay tres modelos básicos que, de una u otra manera, suelen estar incluidos en todos los estudios.

***Modelo nulo.*** La estrategia multinivel comienza ajustado un modelo sin predictores. Se trata, por tanto, de un análisis de varianza de un factor de efectos aleatorios que se conoce como modelo nulo o vacío y cubre tres finalidades: estima la magnitud de la varianza total y cómo ésta se distribuye entre los diferentes niveles de agregación; sirve de base para comparar el ajuste y mejora de la capacidad explicativa del resto de modelos; y permite estimar el efecto de centro, es decir, la proporción de las diferencias en el resultado que son imputables a la acción educativa de las escuelas. La estimación del efecto del centro es la línea primigenia y más antigua de la eficacia escolar (Scheerens, 2016; Scheerens & Bosker, 1997; Scheerens, Witziers & Steen, 2013; Teddlie & Reynolds, 2000; Towsend, 2007) y en Latinoamérica la investigación sobre los efectos escolares tiene más de dos décadas por lo que se dispone de un amplio conjunto de trabajos (entre otros, Casas, Gamboa & Piñeros, 2002; Cervini, 2012; Cervini, Dari & Quiroz, 2016; Murillo, 2003, Murillo & Román, 2011; UNESCO-OREALC & LLECE, 2000, 2010, 2016b). El efecto del centro señala, fundamentalmente, el porcentaje de variaciones relacionadas con diferencias en la calidad y oferta formativa de los centros. En general se asume que en los sistemas educativos más equitativos el tamaño de este efecto es más pequeño dado que las diferencias en el resultado de los centros tienden a ser más menores (Woitschach et al., 2017).

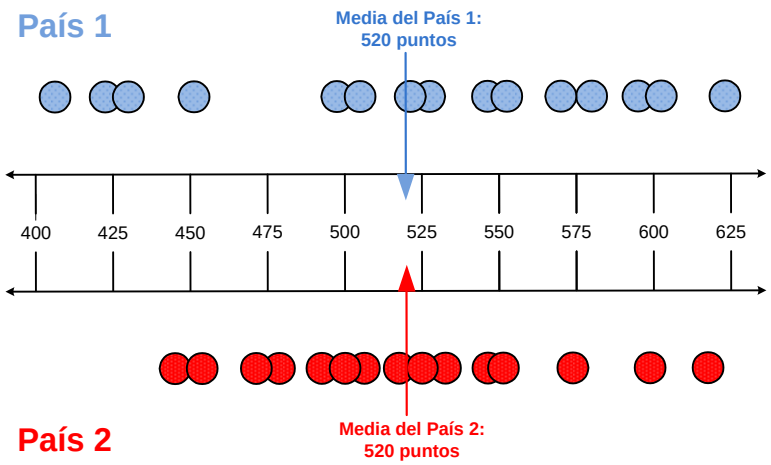


*Figura 6*. La lógica del efecto del centro
Fuente: Servicio de Evaluación Educativa del Principado de Asturias

La figura 6 muestra los resultados simulados de dos países compuestos por 15 centros (cada uno representado por un círculo). La posición del círculo señala el promedio de cada centro en la escala *N(500,100)*. La media de ambos países es idéntica (520 puntos), pero en el País 2 las diferencias entre los centros, entendida como la varianza en las medias de los centros es mucho menor (aproximadamente un 50% más pequeña). Por tanto, el tamaño del efecto del centro en el País 2, es decir, las diferencias o desigualdades entre sus escuelas son más pequeñas y se concluye que, con relación al País 1, sus resultados son más equitativos.

***Modelo de ajuste o modelos background.*** En el segundo modelo se incluyen como predictores la información disponible sobre los antecedentes escolares y variables de contexto sociodemográfico y escolar del alumnado. Las variables más empleadas son el índice que resume el nivel socioeconómico y cultural del alumnado o, en su defecto, variables como estudios y profesiones de los progenitores, número de libros en el hogar, posesiones materiales o características de la vivienda (Palardy, Rumberger & Butler, 2015; Peña Suárez, Fernández-Alonso & Muñiz, 2009; Sirin, 2005). Otras variables muy usadas en los modelos de ajuste son el género, la lengua materna o la condición de emigrante y en los estudios latinoamericanos también parecen importantes, ser indígena y compatibilizar trabajo y estudios, que generalmente no se consideran en la investigación con países desarrollados (UNESCO-OREALC & LLECE, 2016b). Por su parte, las variables relativas a los antecedentes escolares con mayor efecto sobre los resultados son, por este orden, el rendimiento previo, la repetición escolar y la escolarización temprana.

Los modelos de ***background*** pueden especificarse con predictores en un único nivel (v. g, regresiones con medias como resultados o regresiones con coeficientes aleatorios). Sin embargo, la potencia explicativa aumenta al incluir factores de ajuste en todos los niveles empleando análisis de covarianza de un factor o análisis de regresión con medias y pendientes como resultados. Por ello, es altamente recomendable que los modelos de ***background*** incluyan variables y factores en todos los niveles jerárquicos del análisis (Scheerens, 2016).

Los resultados del modelo de variables *background* pueden interpretarse en términos de inequidad educativa: cuando mayor sea el porcentaje de varianza explicada por los predictores incluidos en este modelo, mayor será la determinación de los resultados por

factores antecedentes y, por tanto, mayor el nivel de inequidad. El gráfico 7 compara el "efecto del nivel socioeconómico y cultural del alumnado (ISEC) sobre la puntuación del alumnado" en dos países con idéntico resultado en la prueba (500 puntos) y nivel socioeconómico y cultural (ISEC = 0 puntos). En el País 1 por cada punto que aumenta el ISEC se predicen 25 puntos de ganancia en la prueba, mientras que en País 2 se predicen 10 puntos de ganancia en la prueba por cada punto que aumenta el ISEC. Por tanto, se concluye que el País 2 parece más equitativo ya que los resultados de su alumnado están menos determinados por los antecedentes socioeconómicos y culturales que en el caso del País 1.

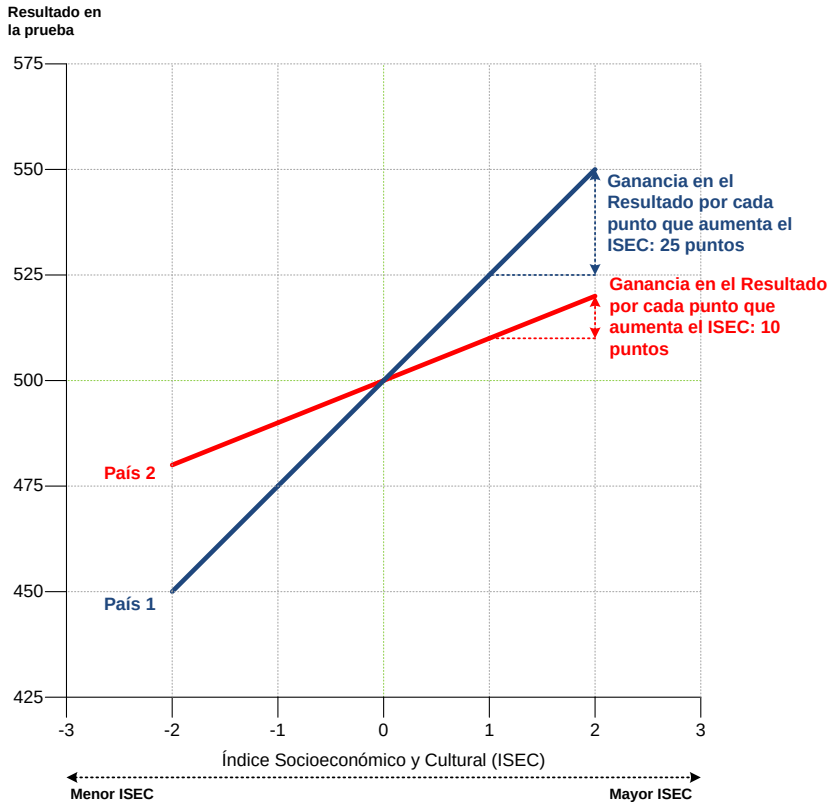


*Figura 7*. La lógica del efecto de los antecedentes sociológicos en los resultados
Fuente: Servicio de Evaluación Educativa del Principado de Asturias

***Modelos de procesos educativos y variables de interés.*** Una vez que se dispone de un modelo con las variables antecedentes, el siguiente paso en la estrategia es añadir al mismo las variables y factores que describen los procesos educativos. Como ya se apuntó, el modelo de background descuenta la varianza imputable a factores antecedentes y demográficos. Por tanto, la varianza que explican los modelos de proceso puede interpretarse como un efecto neto y no contaminado. En otras palabras, los procesos que aparezcan estadísticamente significativos lo serán después de descontar o neutralizar el efecto de los antecedentes y, por tanto, puede descartarse que los resultados

del modelo estén afectados por hipótesis alternativas relativas a las características de contexto social y demográfico.

Las especificaciones de los modelos de proceso pueden ser muy variadas. La estrategia más analítica es introducir una a una todas las variables de interés sobre el modelo de ajuste (UNESCO-OREALC & LLECE, 2016b). Es la más detallada y en general la que más posibilidades tiene de mostrar significaciones estadísticas en los factores asociados. También es posible especificar modelos que incluyen todas las variables de proceso de un mismo nivel de análisis. Por ejemplo, introducir todos los factores de proceso medidos en el nivel de estudiante y de esta forma explorar el efecto conjunto sobre los resultados de variables como los hábitos de lectura, las actitudes, motivación, expectativas académicas del alumnado, etc., y estimar qué porcentaje de la varianza es explicada por las variables personales del alumnado.

Otra solución, probablemente más sustantiva desde el punto de vista teórico, consiste en especificar un modelo para estudiar procesos concretos incluyendo variables medidas en diferentes niveles de análisis. Un ejemplo posible sería el estudio de los procesos de aula (Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas del Principado de Asturias, 2011). En este caso, el modelo incluye variables medidas a nivel individual (por ejemplo, la valoración de la labor docente por parte del alumnado), y otras medidas a nivel de aula (clima de trabajo, tiempo efectivamente dedicado al aprendizaje, perfiles de metodología docente, etc.). Con este tipo de modelos es posible analizar la interacción entre variables de distinto nivel, respondiendo a preguntas de investigación muy interesantes como, por ejemplo, estudiar la influencia de una determinada metodología de docente (variable de aula) sobre en el aprendizaje de alumnado con diferentes niveles de comprensión (variable individual), pudiendo identificar metodologías de enseñanza-aprendizaje que beneficien sobremanera al alumnado con mayores problemas de comprensión.

En general el último modelo suele incluir todas las variables de proceso en todos sus niveles (UNESCO-OREALC & LLECE, 2000, 2010). Ello permite estimar la capacidad predictiva del modelo completo y comparar los efectos del conjunto de variables y factores analizados. Aquellas variables que siguen manteniendo su significación estadística en el modelo final pueden considerarse los factores más relevantes sobre los que apoyar las conclusiones del estudio y orientar las políticas educativas para la mejora del sistema educativo.

## Conclusiones

La evaluación de sistemas educativos supone un desafío en diferentes ámbitos: disposición y logística de los recursos; especificación de marcos teóricos; análisis de datos y comunicación de resultados. En relación con el análisis de datos el principal reto es responder a las dos finalidades de estos estudios: expresar los resultados y competencias de la población escolar; e identificar y estudiar los factores asociados a los resultados educativos que permitan orientar las decisiones políticas para la mejora de los sistemas educativos.

Para cumplir el primer objetivo se han desarrollado dos procedimientos singulares. Las puntuaciones se expresan como valores plausibles y no como estimadores puntuales. Adicionalmente, las puntuaciones numéricas se traducen a descripciones de competencias mediante métodos de puntos de corte que establecen niveles de rendimiento. Por su parte, el análisis de factores asociados es inseparable de un marco teórico robusto, que asume la existencia

de factores de diversa naturaleza, íntimamente relacionados y manteniendo relaciones jerárquicas. En coherencia con este marco teórico en el análisis de factores asociados se emplean modelos jerárquico-lineales que deben ser ajustados de acuerdo con una estrategia que descomponga la varianza en los diferentes niveles de agregación y controle y descuente la parte de las variaciones debidas a los factores antecedentes. Sólo de esa manera se puede estimar un efecto neto de los procesos educativos y no contaminado, y orientar las políticas educativas a la mejora del sistema.

## Referencias


Adams, R. J., & Wu, M. L. (2002). *Technical report for the OECD Programme for International Student Assessment*. Paris: OECD Publications.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 Technical Report*. Washington, DC: U.S. Department of Education / NCES.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report*. Washington, DC: U.S. Department of Education / NCES

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 Technical Report*. Princeton, NJ: NAEP / Educational Testing Service.

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The Next Stage in Educational Assessment. *Educational Researcher, 11*(3), 4-16. doi: https://doi.org/10.3102/0013189X011003004

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

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., & 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: http://dx.doi.org/10.15366/reice2016.14.4.003

Fernández-Alonso, R. (2004). *Evaluación del rendimiento matemático* (tesis doctoral). Universidad de Oviedo. Recuperado de http://hdl.handle.net/10651/16615

Fernández-Alonso, R., Álvarez-Díaz, M., Suárez-Álvarez, J., & Muñiz, J. (2017). Students' achievement and homework assignment strategies. *Frontiers in Psychology, 8*, 286. doi: http://dx.doi.org/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. (2015). Adolescents' homework performance in mathematics and science: Personal factors and teaching practices. *Journal of Educational Psychology, 107*(4), 1075-1085. doi: http://dx.doi.org/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: http://dx.doi.org/10.1387/RevPsicodidact.13939

Fernández-Alonso, R., Woitschach, P., Álvarez-Díaz, M., González-López, A. M., Cuesta, M., & Muñiz, J. (2019). Homework and academic achievement in Latin America: A multilevel approach. *Frontiers in psychology, 10*, 95. https://doi.org/10.3389/fpsyg.2019.00095

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53. doi: https://doi.org/10.1111/j.1745-3992.2009.00154.x

Gaviria Soto, J. L., & Castro Morera, M. (2005). *Modelos jerárquicos lineales*. Madrid: La Muralla.

Hox, J. J. (1998). Multinivel modeling: When and Why. En I. Balderjahn, R. Mathar y M. Schader (Eds.). *Classification, data analysis and data highways* (pp 147-154). New York: Springer.

Hungi, N. (2011). Accounting for Variations in the Quality of Primary School Education, *SACMEQ Working Paper, 7*. Recuperado de http://www.sacmeq.org/sites/default/files/sacmeq/publications/07_multivariate_final.pdf

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Capelle, F., Paviot, L., & Vellien, J. (2010). SACMEQ III Project Results: Pupil Achievement levels in Reading and Mathematics. *SAMEQ Working Document, 1*. Recuperado de http://www.sacmeq.org/sites/default/files/sacmeq/reports/sacmeq-iii/working-documents/wd01_sacmeq_iii_results_pupil_achievement.pdf

Kelly, D. L., Mullis, I.V. S., & Martin, M. O. (2000). *Profiles of Student Achievement in Mathematics at the TIMSS International Benchmarks: U.S. Performance and Standards in an International Context*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22,* 259-267. doi: https://doi.org/10.1177/001316446202200202

Martin, M. O., & Kelly, D. L. (1997). *TIMSS technical report: Vol. II. Implementation and analysis: Primary and middle school years*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Mullis, I. V. S (Eds.) (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de http://timssandpirls.bc.edu/methods/pdf/TP11_Context_Q_ Scales.pdf

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.) (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.) (2017). *Methods and Procedures in PIRLS 2016*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Recuperado de https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Mazzeo, J. (2018). Large-scale group-score assessments. En W. J. van der Linden (Ed.) *Handbook of Item Response Theory. Vol. 3: Applications* (pp. 297-311). Boca Raton, FL: CRC Press.

Messick, S., Beaton, A. E., & Lord, F. (1983). *A new design for a new era*. Princeton. NJ: Educational Testing Service.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29*(2): 133-161. Recuperado de http://www.jstor.org/stable/1434599

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gonzalez, E. J., Chorostowski, S. J., & O'Connor, K. M. (2002). *TIMSS assessment framewoks and specificacions 2003* (2 ed.). Chestnut Hill, MA: Boston College.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide

Muñiz, J. (2018). *Introducción a la Psicometría*. Madrid: Pirámide

Muñoz-Repiso, M., Cerdán, J., Murillo, F. J., Calzón, J., Castro, M., Egido, I., García, R., & Lucio-Villegas, M. (1995). *Calidad de la educación y eficacia de las escuelas. Estudio sobre la gestión de los recursos educativos*. Madrid: Ministerio de Educación y Ciencia.

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

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.

National Center for Education Statistics (2018). *NAEP Technical Handbook: Methods and Procedures*. U.S. Department of Education: Washington, DC. Recuperado de https://nces.ed.gov/nationsreportcard/tdw/

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Openshaw, S. (1984). Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A: Economy and Space, 16*(1), 17-31. doi: https://doi.org/10.1068/a160017

Organisation for Economic Co-operation and Development [OECD]. (2009). *PISA Data Analysis Manual: SPSS® Users*, 2nd Ed. Paris: OECD Publishing. doi: http://dx.doi.org/10.1787/9789264056275-en

Organisation for Economic Co-operation and Development [OECD]. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. Recuperado de http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pacific Community (2016). *Pacific Islands Literacy and Numeracy Assessment (PILNA 2015). Regional Report*. Fiji Islands: Educational Quality Assessment Program.

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.

Pardo, A., Ruiz, M. Á., & San Martín, R. (2007). Cómo ajustar e interpretar modelos multinivel con SPSS. *Psicothema, 19*(2), 308-321.

Paterson, L., & Goldstein, H. (1991) New statistical methods for analysing social structures: An introduction to multilevel models. *British Educational Research Journal, 17*(4), 387-393. Recuperado de http://links.jstor.org/sici?sici=0141-1926%281991%2917%3A4%3C387%3ANSMFAS%3E2.0.CO%3B2-9

Peña Suárez, E., Fernández Alonso, R., & Muñiz Fernández, J. (2009). Estimación del valor añadido de los centros educativos. *Aula abierta, 37*(1), 3-18. Recuperado de http://digibuo.uniovi.es/dspace/bitstream/10651/7869/1/AulaAbierta.2009.37.1.3-18.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2ª ed.)*. Thousand Oaks, CA: Sage.

Robinson, A. H. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review, 15*, 351-357

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness. A critical review of the knowledge base*. Dordrecht, The Netherlands: Springer.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier.

Scheerens, J., Witziers, B., & Steen, R. (2013). A meta-analysis of school effectiveness studies. *Revista de Educación, 361*, 619-645. doi: https://doi.org/10.4438/1988-592X-RE-2013-361-235

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018). *ICCS 2016 technical report.* Amsterdam: The International Association for the Evaluation of Educational Achievement.

Servicio de Evaluación Educativa del Principado de Asturias. (2018a) ¿Cómo se describen los resultados del aprendizaje en las evaluaciones del sistema educativo? *Informes de Evaluación, 14*. Recuperado de https://www.educastur.es/documents/10531/879356/2018_11_informe_evaluacion_N14_V03_r/997be880-9627-4456-9ebb-e5465c20a4df

Servicio de Evaluación Educativa del Principado de Asturias. (2018b). *Evaluación de Diagnóstico Asturias 2018. Niveles de rendimiento 6º de EP*. Oviedo, España: Consejería de Educación y Cultura.

Servicio de Ordenación Académica, Formación del Profesorado y Tecnologías Educativas del Principado de Asturias (2011). *Evaluación de Diagnóstico Asturias 2010*. Oviedo, España: Consejería de Educación y Ciencia.

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

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

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

Trautwein, U. (2007). The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction, 17*, 372-388. doi: https://doi.org/10.1016/j.learninstruc.2007.02.009

UNESCO-OREALC. (2016a). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago de Chile: UNESCO. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000247123

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

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

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series, Volume 2*, 9−36.

Woitschach, P. (2018). *Evaluaciones educativas a gran escala en Latinoamérica: TERCE* (tesis doctoral). Universidad Complutense, Madrid.

Woitschach, P., Fernández-Alonso, R., Martínez-Arias, R., & Muñiz, J. (2017). Influencia de los Centros Escolares sobre el Rendimiento Académico en Latinoamérica. *Revista de Psicología y Educación, 12*(2), 138-154. doi: https://doi.org/10.23923/rpye2017.12.152

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest 2.0: generalised item response modelling software*. Camberwell, Victoria: Australian Council for Educational Research.