



Education in the Knowledge Society

journal homepage <http://revistas.usal.es/index.php/eks/>

Large-Scale Assessment Tests and school effectiveness: A systematic literature review

Pruebas de evaluación a gran escala y eficacia escolar: una revisión sistemática de la literatura

Cristina Frade-Martínez^a (*), Adriana Gamazo^b, Susana Olmos-Migueláñez^c^a Instituto Universitario de Ciencias de la Educación (IUCE), Universidad de Salamanca, Salamanca, España<https://orcid.org/0000-0003-1781-0790> cristina.frade@usal.es^b Instituto Universitario de Ciencias de la Educación (IUCE), Universidad de Salamanca, Salamanca, España<https://orcid.org/0000-0002-7523-9484> adrianagamazo@usal.es^c Instituto Universitario de Ciencias de la Educación (IUCE), Universidad de Salamanca, Salamanca, España<https://orcid.org/0000-0002-0816-4179> solmos@usal.es

ARTICLE INFO

Keywords

educational assessment; school effectiveness; systematic literature review; academic achievement; secondary education.

Palabras clave

evaluación educativa; eficacia escolar; revisión sistemática de la literatura; rendimiento académico; educación secundaria.

ABSTRACT

Different international agencies have designed large-scale assessment tests (ILSA) for decades to measure student performance and establish comparisons between education systems. We present a systematic review of the literature that aims to learn about the evolution and characteristics of the secondary analyses of three ILSAs (PISA, TIMSS, and PIRLS) and identify possible improvements and future lines of research. We searched three repositories: Web of Science, Scopus, and ERIC, focused on secondary analyses of data from these tests developed since 2000, whose main objective was to analyze questions related to predictors of school performance and effectiveness. After applying different criteria, 63 articles were selected from the 470 identified. This analysis allowed us to identify the increase in interest in this topic over the last few years, the lack of studies in certain regions, and the prevalence of multilevel methodology. The review has also highlighted some limitations in the existing literature, such as the lack of consideration of the elements inherent to the design of large-scale evaluations, as well as those inherent to the data used for the analysis, such as the impossibility of establishing causal relationships or the lack of data at the classroom level. These results are used to make recommendations and detect less explored areas, such as curriculum-related or country-level variables.

RESUMEN

Desde hace décadas diferentes organismos internacionales han diseñado pruebas de evaluación internacional a gran escala (ILSA) dirigidas a medir el rendimiento de los estudiantes y tratando de establecer comparaciones entre los sistemas educativos. Se presenta una revisión sistemática de la literatura que pretende conocer la evolución y las características de los análisis secundarios de tres ILSA (PISA, TIMSS y PIRLS), así como identificar posibles mejoras y futuras líneas de investigación. Se realizaron búsquedas en tres repositorios: Web of Science, Scopus y ERIC, centrados en análisis secundarios de datos de estas pruebas desarrolladas desde el año 2000 cuyo objetivo principal era analizar cuestiones relacionadas con los predictores del rendimiento y la eficacia escolar. Tras aplicar diferentes criterios, se seleccionaron 63 artículos de los 470 identificados. Este análisis ha permitido identificar el aumento del interés por este tema en los últimos años, la falta de estudios en determinadas regiones o la prevalencia de la metodología multinivel. La revisión también ha puesto de manifiesto algunas limitaciones de la literatura existente, como la falta de consideración de los problemas inherentes al diseño de evaluaciones a gran escala, así como los relacionados con los datos utilizados para el análisis, como por ejemplo la imposibilidad de establecer relaciones causales o la falta de datos a nivel de aula. Estos resultados sirven de base para hacer algunas recomendaciones, así como para detectar áreas menos exploradas, como las variables relacionadas con el currículo o las variables a nivel de país.

(*) Autor de correspondencia / Corresponding author

1. Introducción

Diferentes organismos internacionales han desarrollado desde hace más de dos décadas pruebas de evaluación para conocer la situación de los sistemas educativos a nivel mundial que son conocidas bajo la denominación de “pruebas de evaluación a gran escala” o ILSA (por sus siglas en inglés, *International Large-Scale Assessments*). Estas ILSA evalúan las competencias de los estudiantes en un determinado momento de su escolarización, como el Programa para la Evaluación Internacional de los estudiantes (PISA), propuesto por la Organización para la Cooperación y el Desarrollo Económico (OCDE) con el objetivo de “evaluar de forma sistemática lo que los jóvenes saben y son capaces de hacer al finalizar su Educación Secundaria Obligatoria” (Instituto Nacional de Evaluación Educativa [INEE], Ministerio de Educación y Formación Profesional, 2023a) mediante la evaluación de sus competencias en matemáticas, ciencias y comprensión lectora. Otro ejemplo lo tenemos en el Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS) desarrollado por la Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA) y dirigido a estudiantes de 4º de Primaria y 2º de la ESO (Instituto Nacional de Evaluación Educativa [INEE], Ministerio de Educación y Formación Profesional, 2023b) o en el Estudio Internacional para el Progreso en Comprensión Lectora (PIRLS) también desarrollado por la IEA y dirigido a estudiantes de 4º de Primaria.

Podríamos decir, por lo tanto, que el objetivo que persiguen las ILSA es establecer comparaciones entre los países para así comprobar qué se hace bien en determinados contextos y tratar de replicar esas buenas prácticas en otros para mejorar la educación de los jóvenes.

La eficacia escolar ha sufrido un auge en su investigación desde finales del siglo pasado. En ese momento, se diferenciaron dos corrientes de estudio de este fenómeno. Por un lado, el movimiento *school effectiveness* que hacía hincapié en el estudio del funcionamiento de los centros educativos para tratar de determinar por qué unas escuelas funcionaban mejor que otras (sus estudiantes obtenían rendimientos académicos más altos) (Muñoz-Repiso Izaguirre et al., 2000). Es decir, se centraba en “conocer qué hace que una escuela sea eficaz” (Murillo, 2005, p. 3).

Por otro lado, surgió el movimiento *school improvement*, centrado en este caso en comprobar en qué consistían concretamente los procesos desarrollados por las escuelas con el fin de mejorar el rendimiento de sus estudiantes (Muñoz-Repiso Izaguirre et al., 2000).

Considerando ambas corrientes de investigación como complementarias, diferentes autores defendieron esta afirmación, dando como resultado un nuevo movimiento de investigación conocido como “Movimiento de mejora de la eficacia escolar” (Murillo, 2005, p. 3), movimiento en el que se enmarca y sobre el que se construye esta investigación.

Es en este momento en el que cabe reflexionar sobre cómo ha sido la investigación en eficacia escolar y su relación con las ILSA, aunque si bien es cierto que existen algunas revisiones sistemáticas de la literatura que abordan la relación de algunas variables contextuales con el rendimiento académico (Álvarez-Sotomayor & Martínez-Cousinou, 2020) o bien los usos de PISA en la investigación educativa (Gamazo, 2019), no consta la existencia de revisiones similares a la planteada en este artículo.

Esta revisión sistemática de la literatura (*Systematic Literature Review* o SLR) tendrá como fin conocer los estudios, sobre todo de carácter secundario, que se han realizado sobre tres grandes pruebas de evaluación internacional a gran escala (concretamente PISA, TIMSS y PIRLS) y su relación con la eficacia escolar. Del mismo modo, pretenderemos comprender cuáles han sido los objetivos de este tipo de estudios, las metodologías empleadas, así como sus propias limitaciones.

Por tanto, el objetivo que nos planteamos para la presente revisión sistemática de la literatura es conocer la evolución y características de los análisis secundarios de estas tres grandes ILSA (PISA, TIMSS y PIRLS), así como valorar las posibles mejoras y líneas futuras que podrían introducirse en las investigaciones en este campo de conocimiento.

2. Metodología

Para alcanzar el propósito descrito, se ha optado por combinar dos metodologías de revisión que nos han permitido extraer la información de forma precisa y planificada y que pueden ser, en caso de que así se desee, replicables. Estas dos metodologías fueron el *mapping* (o mapeo) (Kitchenham et al., 2011; Torres-Carrion et al., 2018) y la revisión sistemática de la literatura (García-Peñalvo, 2022; Kitchenham, 2004).

Con la primera de las metodologías se ha pretendido dar respuesta a una serie de preguntas de investigación de corte más genérico, “*they are intended to provide an overview of a topic area*” (Kitchenham et al., 2011, p. 3), concretado en seis cuestiones.

La SLR, en cambio, se ha empleado con el fin de comprender en mayor profundidad en qué se centraban estos estudios, qué metodologías de investigación eran las más empleadas en el desarrollo de este tipo de análisis secundarios de las pruebas de evaluación a gran escala, así como su evolución, limitaciones, etc. Aspectos que fueron concretados en otras seis preguntas de investigación que definiremos a continuación.

Para el desarrollo de esta SLR se han seguido las directrices de la Declaración PRISMA 2020 (Page et al., 2021), suponiendo, de este modo, la realización de un proceso sistematizado de búsqueda, recopilación y análisis de la información.

Por tanto, se han conjugado dos metodologías complementarias con el fin de obtener la mayor información posible sobre los estudios desarrollados en las últimas décadas.

2.1. Preguntas de investigación

Como decíamos previamente, se han planteado distintas preguntas de investigación según la metodología a aplicar y a partir de las cuales hemos podido establecer la cadena de búsqueda apropiada para así dar respuesta a esas preguntas. Diferenciaremos a continuación entre preguntas de investigación del *mapping* (M) y preguntas de investigación de la SLR (PI).

Mapping

Para comprender la trayectoria que ha seguido y sigue esta temática, hemos iniciado la revisión de la literatura empleando la metodología de mapeo de la literatura. Con anterioridad a la selección de las bases de datos de las que obtener los documentos a analizar, se plantearon las siguientes cuestiones a resolver:

- M1: ¿Cuántos estudios se han publicado en revistas científicas relacionando alguna prueba estandarizada (PISA, TIMSS o PIRLS) con la eficacia escolar?
- M2: ¿Desde cuándo se han contabilizado las publicaciones y hasta cuándo? ¿Cuándo ha experimentado su mayor auge la investigación en esta temática?
- M3: ¿Quiénes son los autores de referencia en este campo?
- M4: ¿Qué países realizan más estudios de este tipo?
- M5: ¿En qué revistas científicas se publican más estudios relacionados con las pruebas a gran escala y la eficacia escolar?
- M6: ¿Cuáles son las metodologías más empleadas para el análisis de datos secundario de las pruebas internacionales de evaluación a gran escala?

SLR

Con la SLR se ha pretendido dar respuesta a las siguientes preguntas de investigación:

- PI1: ¿Cuáles son las variables predictoras más utilizadas en la investigación sobre rendimiento y eficacia escolar?
- PI2: ¿Se ha centrado la investigación basada en los resultados de las pruebas a gran escala en el rendimiento académico o en la eficacia escolar?
- PI3: ¿Los autores tienen en cuenta la estructura jerárquica de los datos de las pruebas a gran escala (estudiante, aula, escuela, región, país...) en sus análisis?
- PI4: ¿Se trabaja con los valores plausibles y los pesos muestrales de las diferentes jerarquías (alumnos y escuelas) en la estimación de parámetros?
- PI5: ¿Se detectan muchas limitaciones a este tipo de estudios? ¿Cuáles son las más comunes?
- PI6: Generalmente, ¿cuáles son las implicaciones que tienen en la sociedad los estudios basados en pruebas de evaluación a gran escala?

2.2. Cadena de búsqueda

Para alcanzar el objetivo planteado en esta revisión sistemática de la literatura, y tras la realización de una primera metabúsqueda en la que se probaron múltiples alternativas para la selección de la cadena de búsqueda en las diferentes bases de datos, se ha determinado que las bases de datos más apropiadas para realizar la búsqueda de información y en función de nuestra temática (ILSA y su relación con la eficacia escolar) eran Web of Science, Scopus y ERIC, esta última por su marcado carácter educativo. Además, se ha limitado la búsqueda a artículos publicados en revistas científicas como tipo de archivo y a partir del año 2000, ya que fue cuando se llevó a cabo la primera edición de las pruebas PISA.

Cuando nos referimos a pruebas estandarizadas o pruebas a gran escala estamos hablando, en este caso, de PISA, TIMMS y PIRLS. Es por este motivo por el que hemos empleado las siguientes cadenas de búsqueda para cada una de las bases de datos elegidas:

- WoS→ (PISA OR TIMMS OR PIRLS) (Topic) and "school effectiveness" (Topic) → 35 resultados (31 artículos).
- Scopus → TITLE-ABS-KEY (pisa OR timms OR pirls) AND "school effectiveness" → 440 resultados (384 artículos).
- ERIC → abstract:(PISA or TIMMS or PIRLS) AND "school effectiveness" → 69 resultados (55 artículos).

2.3. Fases

Diferenciamos una serie de fases a desarrollar consecutivamente y que han consistido en:

1. Eliminar duplicados.
2. Aplicar los criterios de inclusión y exclusión seleccionados (que se detallarán a continuación) revisando título, resumen y palabras clave de los artículos.
3. Aplicar los criterios de calidad (que se detallarán a continuación) al texto completo.
4. Extraer los datos de los artículos incluidos en la selección final para dar respuesta a las preguntas de investigación tanto del mapping como de la propia SLR.

En la primera fase se eliminaron los duplicados generados tras realizar la búsqueda en las tres bases de datos seleccionadas mediante las cadenas de búsqueda informadas previamente.

A continuación, se determinaron una serie de criterios de elegibilidad (inclusión y exclusión) de los artículos que fueron aplicados a la revisión del título, el resumen y las palabras clave del artículo. Estos criterios de inclusión y exclusión (CI y CE respectivamente) fueron:

- CI1: La publicación está relacionada con la temática investigada (eficacia escolar, rendimiento o pruebas a gran escala).
- CI2: El artículo está en castellano o inglés.
- CI3: Se utilizan los datos de las pruebas internacionales a gran escala como PISA, TIMMS o PIRLS.
- CI4: Se hace referencia a los factores predictores del rendimiento académico.
- CI5: Se trata de artículos empíricos sobre la eficacia escolar.
- CI6: La variable criterio del estudio es el rendimiento o la eficacia.
- CI7: La metodología empleada es un análisis multinivel.
- CE1: La temática no tiene relación con nuestro tema de investigación.
- CE2: El artículo está en algún idioma diferente del castellano o el inglés.
- CE3: Los datos en los que se basa el artículo no pertenecen a pruebas estandarizadas internacionales.
- CE4: Se trata de un artículo meramente teórico.

En este caso, se ha considerado necesario que los artículos cumplieran con todos los criterios de inclusión para ser seleccionado y proceder a la revisión de su texto completo aplicando los criterios de calidad. En el caso de los criterios de exclusión, bastaría con que los artículos cumplieran con alguno de ellos para ser eliminados del proceso de revisión.

Como mencionábamos, con el fin de ultimar la selección de artículos para su revisión final, se redactaron una serie de criterios de calidad en forma de rúbrica. De este modo, se le otorgó una puntuación de 0 en caso de que no se cumpliera, 0,5 si se cumplía parcialmente y 1 en caso de que ese criterio se cumpliera en su totalidad.

En función de las puntuaciones obtenidas, aquellos artículos con una puntuación inferior a 8.5 fueron excluidos de la selección final (al considerar que el no alcanzar esta puntuación mínima suponía no poder ser considerado como de calidad para esta investigación), siendo el resultado final de nuestra SLR aquellos artículos que superaron esta puntuación de corte. Estos criterios de calidad han sido una adaptación de los aplicados por Gamazo (2019) (ver Tabla 1).

Tabla 1. Criterios de calidad de las publicaciones

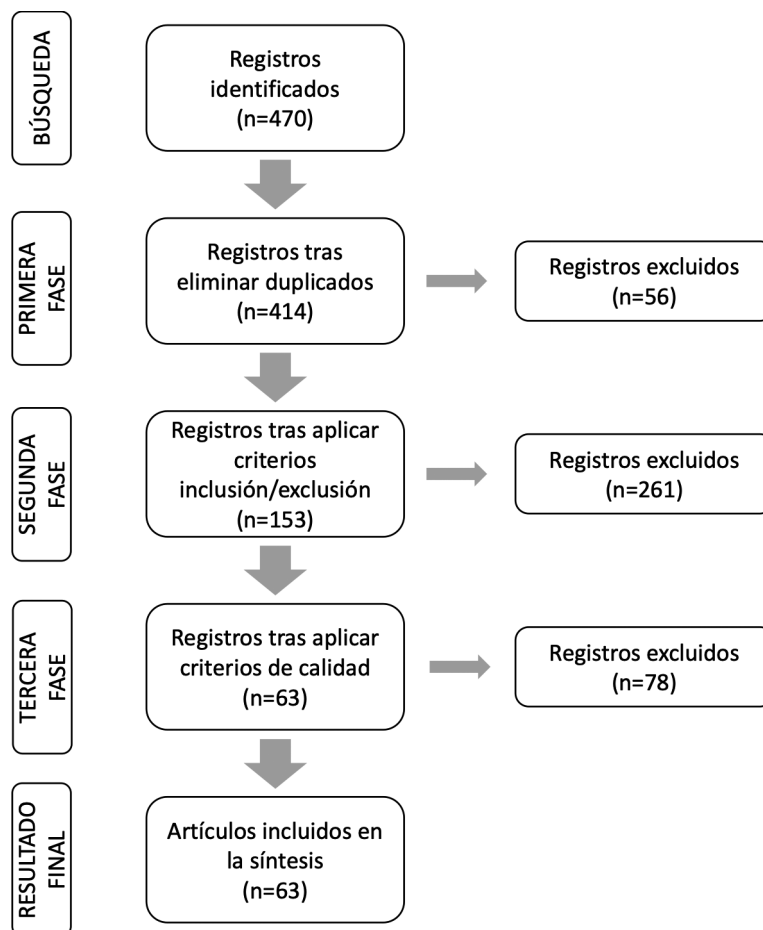
Pregunta	No	Parcial	Si
1. ¿Existen preguntas de investigación cualitativas y cuantitativas claras (u objetivos)?	No hay objetivos/preguntas de investigación claros, ni se intenta explicarlos en el texto.	Los objetivos/preguntas no se especifican como tales, pero se deducen claramente del texto	Los objetivos/preguntas están claramente identificados o etiquetados como tales en el texto
2. ¿Permiten los datos recogidos al autor o autores abordar la pregunta de investigación (objetivo)?	Los datos recogidos son insuficientes para abordar los objetivos/preguntas	Los datos recogidos permiten responder parcialmente a los objetivos/preguntas	Los datos recogidos permiten al autor o autores responder con éxito a todos los objetivos/preguntas
3. ¿Utiliza una muestra completa de PISA o de las otras pruebas de evaluación a gran escala incluidas en este análisis (es decir, un muestreo realizado por expertos)?	Utiliza una submuestra ad hoc y no está claro si la muestra es representativa de la población subestudiada	Utiliza una submuestra ad hoc, pero su composición y representatividad se explican claramente	Utiliza una muestra realizada originalmente por los expertos de PISA
4. ¿Son adecuadas las mediciones (origen claro, o validez conocida, o instrumento estándar) y están claramente definidas?	El origen de las variables no es claro/válido/fiable, y/o las variables no están claramente definidas	El origen de las variables es claro, válido y fiable, pero su definición carece de claridad	El origen de las variables es claro, válido y fiable y están claramente definidas
5. ¿Se explican claramente los métodos?	El método no se explica claramente en el texto	El método se explica un poco en el texto	El método se explica claramente en el texto
6. ¿Son los métodos utilizados adecuados para responder a las preguntas/objetivos?	Los métodos no arrojan los resultados necesarios para responder a las preguntas de la investigación	Los métodos permiten al autor o autores responder a las preguntas de la investigación, pero no son los mejores métodos para hacerlo	Los métodos utilizados son los más adecuados para responder a las preguntas de investigación planteadas por los autores
7. ¿Se responden todas las preguntas de investigación?	Al menos dos preguntas de investigación no tienen respuesta	Se responde a todas las preguntas de la investigación menos a una	Se responde a todas las preguntas de la investigación
8. ¿Se basan las conclusiones en los datos?	Las conclusiones no se basan en los datos presentados en el documento	Las conclusiones se basan en cierta medida en los resultados obtenidos, pero el autor o autores hacen afirmaciones que no se basan en los datos	Todas las conclusiones se basan en los resultados presentados en el documento
9. ¿Se explica el uso de valores plausibles?	No se menciona el uso de valores plausibles / No hay uso de valores plausibles / Hay un uso incorrecto de valores plausibles	El uso de valores plausibles se menciona, pero no se explica claramente	El uso de valores plausibles es adecuado y está correctamente explicado
10. ¿En qué consiste el uso de las ponderaciones muestrales explicadas?	No se menciona el uso de las ponderaciones muestrales / No hay uso de las ponderaciones muestrales / Hay un uso incorrecto de las ponderaciones muestrales	Se menciona el uso de pesos de muestreo, pero no se explica claramente	El uso de las ponderaciones muestrales es adecuado y está correctamente explicado

La búsqueda en las bases de datos seleccionadas dio como resultado un total de 470 artículos. Eliminando los artículos duplicados nos quedamos finalmente en esta primera fase con 414 artículos (24 de ERIC, 8 de WoS y 382 de Scopus).

En la siguiente fase se procedió a la aplicación de los criterios de inclusión y exclusión, tras la cual pasamos de 414 artículos a 153.

Al texto completo de estos 153 artículos se le aplicaron los criterios de calidad (Tabla 1), estableciendo el punto de corte para su aceptación en la obtención de, al menos, 8.5 puntos, como se mencionó previamente. Esto supuso que los resultados finales de esta SLR nos dejaran un total de 63 artículos, habiendo excluido 78 artículos. Podemos ver en la Figura 1 los resultados de este procedimiento.

Figura 1. Diagrama de flujo PRISMA sobre el proceso de búsqueda



Nota. Adaptado de Page et al., 2021.

Los 63 artículos restantes fueron sometidos a una revisión en profundidad seguida de un proceso de vaciado de datos para así poder dar respuesta a las preguntas de investigación del mapping y de la propia SLR.

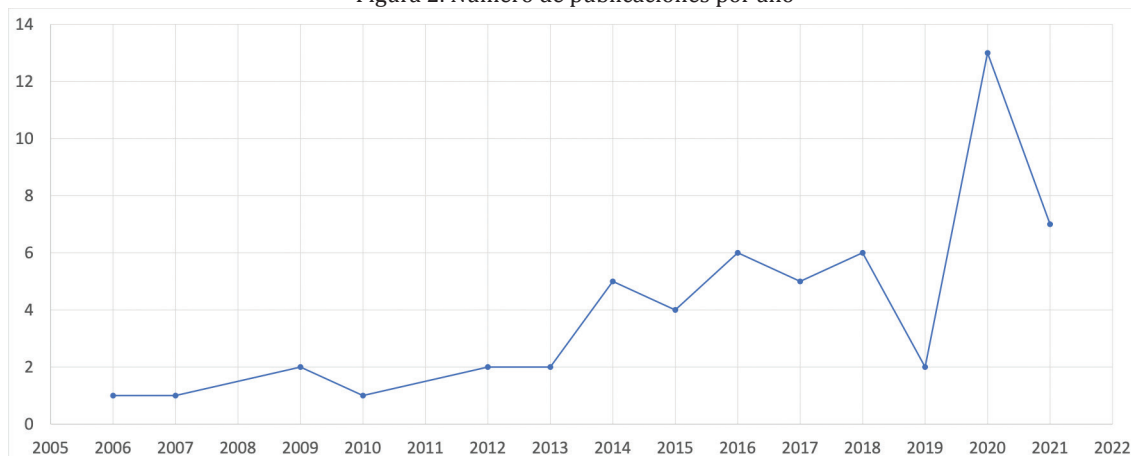
3. Resultados

Comenzaremos respondiendo a las preguntas de investigación planteadas para el estudio de *mapping*. Como respuesta a las dos primeras cuestiones: resultaron seleccionados un total de 63 artículos publicados en revistas científicas; concretamente entre los años 2006 y 2021, a pesar de haber incluido en la búsqueda inicial en las bases de datos todas aquellas publicaciones desde el año 2000.

En la Figura 2 podemos ver la evolución de los artículos seleccionados, concentrándose un gran número de publicaciones en el año 2020, con 14 artículos publicados en ese año (M2).

Con algunas excepciones (donde destacan las pocas publicaciones en el año 2019), observamos que, desde que se inició la investigación en esta temática, ha habido un interés creciente en obtener información en este campo mediante la explotación de los datos empleando análisis secundarios de los resultados en las pruebas internacionales de evaluación a gran escala.

Figura 2. Número de publicaciones por año



La primera publicación sobre el tema fue la de Ruiz de Miguel y Castro Morera (2006) en la que las autoras estudiaron los factores de la eficacia escolar en el área de matemáticas en una serie de países que habían participado en PISA 2003; mientras que la publicación más reciente es la de Pomianowicz (2023) referida a las diferencias en el rendimiento académico en comprensión lectora entre los inmigrantes de primera y segunda generación de los diferentes países participantes en PISA 2018.

Centrándonos en los autores de referencia en este campo de investigación (M3), según el número de publicaciones en este periodo, destaca Martínez-Abad de entre los 133 autores diferentes del periodo comprendido entre 2006 y 2021 en los 63 artículos analizados. En este caso, ha trabajado en colaboración con Gamazo en tres de las cuatro publicaciones que han sido seleccionadas. En dos de estos artículos destaca el empleo de la metodología de minería de datos o Data mining para investigar sobre los factores asociados al rendimiento académico y a la eficacia escolar (Gamazo & Martínez-Abad, 2020; Martínez-Abad et al., 2020); mientras que en el tercero se utiliza el análisis multinivel para identificar, de nuevo, factores relacionados con la eficacia escolar (Gamazo et al., 2018).

Del mismo modo, debemos indicar que Van Damme, Van Den Noortgate y Gielen han trabajado juntos en las tres publicaciones seleccionadas, realizando estudios sobre:

1. Las diferencias de rendimiento en lectura en las distintas regiones belgas con datos de PISA 2009 (Ning et al., 2016).
2. La influencia del clima disciplinario en el rendimiento en lectura en diferentes países que participaron en PISA 2009 (Ning et al., 2015).
3. Los factores asociados al rendimiento académico en matemáticas en una serie de países que participaron en PISA 2012 (Liu et al., 2014).

Otra de las autoras con mayor número de publicaciones seleccionadas en esta SLR ha sido Ruiz de Miguel, con estudios centrados en el análisis de los factores asociados al rendimiento académico en las diferentes áreas estudiadas en PISA en alguna de sus ediciones (Blanco Blanco et al., 2014; Ruiz de Miguel, 2009) o también mediante el estudio de la eficacia escolar en el área de las matemáticas (Ruiz de Miguel & Castro Morera, 2006).

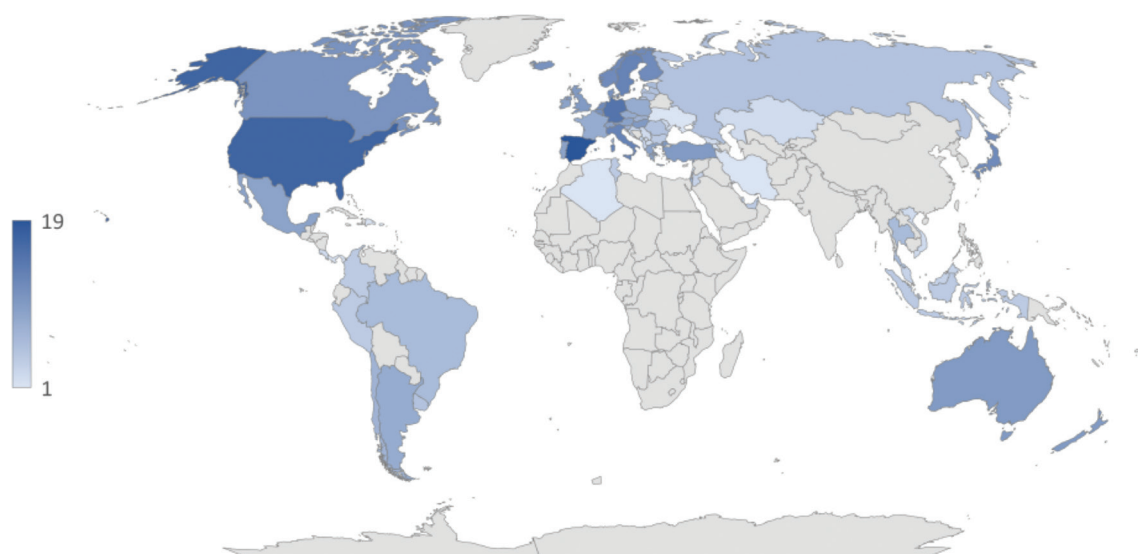
En la Tabla 2 observamos, de manera individual, los autores con mayor número de publicaciones sobre el tema que nos concierne:

Tabla 2. Número de publicaciones por autor

Autor	Número de publicaciones
Martínez-Abad, F.	4
Gamazo, A.	3
Van Damme, J.	3
Van Den Noortgate, W.	3
Gielen, S.	3
Ruiz de Miguel, C.	3

La Figura 3 nos indica el número de publicaciones en función del país al que se refieren los datos de las investigaciones (M4), es decir, en cuántas publicaciones se han incluido los datos de ese país (tanto individual como conjuntamente) siempre y cuando hayan aparecido en al menos dos de las publicaciones seleccionadas.

Figura 3. Número de publicaciones por país



Los países que han participado en un mayor número de estudios de este tipo han sido España (20 artículos), Estados Unidos (17), Suecia (16) y Alemania (15).

De los 63 artículos, en 32 se realizan análisis tratando de establecer comparaciones entre diferentes países o incluso entre regiones de un mismo país a partir de datos de alguna de las pruebas estandarizadas a gran escala seleccionadas (PISA, TIMSS o PIRLS). Los 31 artículos restantes se centraron en un único país o en analizar los datos relativos a una región concreta de algún país. Podemos ver en la siguiente tabla (Tabla 3) cuales han sido los países sobre los que se han concentrado un mayor de número de estudios a nivel individual: España con 8 artículos y Estados Unidos y Argentina con 4 respectivamente.

Tabla 3. Estudios individuales por país y referencias bibliográficas

País	Referencia
España	(Blanco Blanco, 2014)
	(Burger & Walk, 2016)
	(Doncel Abad & Cabrera Álvarez, 2020)
	(Gamazo et al., 2018)
	(Gil-Flores & García-Gómez, 2017)
	(Gómez-Fernández & Mediavilla, 2021)
	(Martínez-Abad, 2019)
	(Martínez-Abad et al., 2020)
Estados Unidos	(Li, 2016)
	(Liu, 2021)
	(Wu, Gao, et al., 2020)
	(Wu, Shen, et al., 2020)
Argentina	(Formichella & Krüger, 2017)
	(Ibañez Martín & Formichella, 2017)
	(Krüger, 2013)
	(Quiroz et al., 2018)

Para responder a la quinta pregunta de investigación formulada para el mapping, hemos analizado las revistas científicas en las que han sido publicados los diferentes artículos seleccionados, habiendo una gran diversidad de revistas (45). Entre ellas destacan Revista de Educación y School Effectiveness and School Improvement con cinco publicaciones sobre el tema estudiado cada una de ellas (M5).

En la Tabla 4 observamos el número de artículos publicado en cada revista con más de una publicación, de acuerdo con el criterio seguido en la respuesta a la pregunta M4, así como el índice de impacto y el cuartil en el que se encuentra dicha revista en la actualidad de acuerdo con SCIMAGO (SJR) y con Web of Science (JIF – SSCI o JIF - ESCI).

Tabla 4. Número de publicaciones por fuente

Fuente	Número de pub.	Scimago		Web of Science	
		Impacto	Cuartil	Impacto	Cuartil
Revista de Educación	5	0,345	Q3	-	-
School Effectiveness and School Improvement	5	1,017	Q1	3,000	Q2
Education Policy Analysis Archives	3	0,41	Q2	0,4*	Q4**
Compare	2	0,719	Q1	1,8	Q3
Educational Studies	2	0,605	Q2	1,7	Q3
Frontiers in Education	2	0,661	Q2	2,3*	Q2**
Frontiers in Psychology	2	0,891	Q2	3,8	Q1
International Journal of Educational Research	2	1,017	Q1	3,2	Q2
International Journal of Science Education	2	1,003	Q1	2,3	Q3
Large-Scale Assessments in Education	2	0,996	Q1	3,1*	Q1**
Scandinavian Journal of Educational Research	2	0,82	Q1	1,9	Q3

Nota. Estos índices y cuartiles han sido consultados y actualizados a fecha 4 de julio de 2023 y se refieren a los rankings del año 2022.

*revistas indexadas en el Emerging Sources Citation Index. ** revistas en las que el cuartil se basa en JCI en lugar de JIF.

A continuación, empleamos la clasificación elaborada por Gamazo (2019) basada en la propuesta por Liou y Hung (2015), donde se diferencian en tres grupos las metodologías de análisis para así poder dar respuesta a la última pregunta planteada (M6), a saber:

- Avanzado: correlación canónica, análisis factorial, análisis de discriminación, análisis clúster, SEM (*Structural Equation Modeling*), regresión logística, modelos multinivel, TRI (Teoría de Respuesta al Ítem), BD/ML (*Big Data y Machine Learning*) y otros.
- Intermedio: regresión múltiple y comparación múltiple post hoc.
- Básico: análisis descriptivos, correlacionales, ANOVA, chi-cuadrado y prueba t de Student.

Siguiendo esta clasificación, observamos en la Tabla 5 cuáles han sido las metodologías empleadas en el análisis de los datos procedentes de las pruebas de evaluación a gran escala.

Tabla 5. Número de publicaciones según la metodología utilizada

Nivel	Metodología	n
Avanzado	Correlación canónica	1
	Análisis factorial	4
	Análisis de discriminación	-
	Análisis clúster	3
	SEM	6
	Regresión logística	3
	Modelos multinivel	55
	TRI	3
	BD/ML	-
	Otros	7
Intermedio	Regresión múltiple	6
	Comparación múltiple post hoc	-
Básico	Descriptivo	24
	Correlacional	6
	ANOVA	6
	Chi-cuadrado	-
	Prueba t	1

Como vemos, en casi todas las investigaciones incluidas en esta SLR se desarrollan modelos multinivel (55 casos), siendo la metodología más empleada. En 16 ocasiones, esta es la única metodología utilizada por los autores, pero lo más común es combinarla con otras metodologías para el análisis de datos. De entre ellas, destaca la complementariedad metodológica que combina el análisis descriptivo de las variables que serán posteriormente incluidas en los modelos multinivel con el propio análisis multinivel (Caponera & Losito, 2016; Doncel Abad & Cabrera Álvarez, 2020; Laukaityte & Rolfsman, 2020; Sortkær & Reimer, 2018) entre otros.

A continuación, procederemos a dar respuesta a las preguntas de investigación planteadas para la SLR.

La primera pregunta de investigación se refiere a las variables predictoras más utilizadas en la investigación sobre rendimiento y eficacia escolar, esto nos servirá para comprender cuál ha sido la evolución de la investigación sobre las pruebas a gran escala, ya que comprobaremos si estas variables han ido cambiando a lo largo del tiempo en función del año de publicación (PI1). La mayor parte de los estudios emplean metodologías en las que se tiene en cuenta la estructura anidada de los datos, es decir, diferencian entre las variables a nivel de alumno y de escuela y, en algunos casos, también a nivel de país (cuando se trata de estudios comparativos entre países, evidentemente) o de aula (sobre todo en el caso de los estudios basados en PIRLS), aplicando mayoritariamente una metodología de análisis multinivel. Por lo tanto, para dar respuesta a esta primera pregunta, y partiendo de la base de que en las pruebas internacionales de evaluación a gran escala se tienen en cuenta una gran cantidad de variables diferentes, presentaremos aquellas que han sido más frecuentemente utilizadas como predictoras del rendimiento académico de los alumnos, diferenciando entre el nivel de estudiante y de escuela.

Nos referiremos en primer lugar a las variables predictoras a nivel de alumno que han sido históricamente más utilizadas y que son: el género, el nivel socioeconómico y cultural, la lengua hablada en el hogar, el índice de inmigración y el de repetición (Gamazo et al., 2018; Krüger, 2013; Martínez-Abad et al., 2020; Wu, Shen, et al., 2020).

En cuanto a las variables a nivel de escuela más empleadas en estos estudios, destacan: el nivel socioeconómico medio de la escuela, su localización (urbana/rural), la titularidad y la escasez tanto de profesorado como de recursos educativos en el centro, entre otras (Gómez-Fernández & Mediavilla, 2021; Pomianowicz, 2023; Sousa et al., 2012; van Hek et al., 2018; Wu, Gao, et al., 2020).

En los casos en los que se han realizado comparaciones entre países considerando el país de procedencia como un tercer nivel (además del de escuela y del de estudiante), las variables con las que han trabajado los diferentes autores han sido el Producto Interior Bruto (PIB) del país, el PIB per cápita, el gasto en educación o el

índice Gini (Gamazo & Martínez-Abad, 2020; Gil-Flores & García-Gómez, 2017; Lee et al., 2021; Liu et al., 2014; Ruiz de Miguel, 2009; Ruiz de Miguel & Castro Morera, 2006; Tan & Liu, 2018).

Reflexionando sobre la propia evolución en el tiempo de estos estudios secundarios de PISA y partiendo de la respuesta de la segunda pregunta de investigación del mapping, con los años ha habido una serie de variables que se han mantenido como el género o el nivel socioeconómico de los alumnos y el tamaño o la titularidad de la escuela, y otras que han ido ganando una mayor relevancia, como las referidas al liderazgo de los equipos directivos de los centros o el clima disciplinario.

Centrándonos ahora en la variable criterio de los estudios (PI2), concretamente diferenciando entre si esta es el rendimiento académico o la eficacia escolar. Con apenas una excepción, en 62 de los 63 artículos la variable de estudio es el rendimiento académico en alguna de las tres competencias básicas estudiadas en las ILSA. En la mayor parte de los casos basados en PISA, el rendimiento académico sobre el que se trabaja coincide con la competencia principal estudiada en la correspondiente edición de esta prueba.

En la tercera pregunta de la SLR nos referíamos a la ya mencionada antes estructura jerárquica de los datos (PI3). De las 63 investigaciones seleccionadas, 60 tienen en cuenta esta estructura en sus análisis, diferenciando mayoritariamente el nivel de escuela y de alumno. Algunos estudios, como por ejemplo el de Lee et al. (2021) o el de Tan y Liu (2018), incluyen también el nivel de país con el fin de realizar comparaciones entre países en función de una serie de variables, siendo dos de las más comunes el producto interior bruto (PIB) o el gasto medio en educación.

Además, algunos estudios basados en datos de TIMSS y de PIRLS, incluyen también el nivel de aula (Bokhove et al., 2019; Grilli et al., 2016) al disponer de esta información, teniendo en cuenta variables como el número de horas de matemáticas o comprensión lectora a la semana, el nivel socioeconómico y cultural medio de la propia clase, el tiempo medio invertido en la realización de deberes de matemáticas, etc.

Por lo tanto, en relación con esta tercera pregunta de investigación, podemos concluir que mayoritariamente los autores tienen en cuenta la estructura jerárquica de los datos de las ILSA.

De los 63 artículos, 53 se basan en datos de PISA en alguna de sus ediciones. De entre ellos, y como respuesta a la cuarta pregunta planteada para esta SLR, en 48 se tienen en cuenta los valores plausibles siguiendo las recomendaciones de los diferentes organismos internacionales (OECD e IEA). En tres investigaciones (Ertem, 2021; Sousa et al., 2012; Tan & Dimmock, 2022) no se tienen en cuenta estos valores plausibles, mientras que en otras dos (Agasisti et al., 2021; Lee et al., 2021), estos valores se emplearon, pero en un caso sin tener en cuenta cada uno de los valores plausibles sino la media de los valores (Lee et al., 2021) lo que supone una pérdida de variabilidad y, en otro caso, empleando el primero de los valores plausibles (Agasisti et al., 2021).

En cuanto al empleo de los pesos muestrales, en 27 de las 53 publicaciones basadas en PISA se utilizan ambos: a nivel de estudiantes y centros. En otras 12 investigaciones, los autores aplican únicamente el peso muestral a nivel de alumnos, frente a 14 artículos en los que no se incluyen los pesos muestrales.

Si combinamos el uso de los valores plausibles con el de los pesos muestrales utilizados en estudios basados en PISA, obtendríamos los siguientes resultados (ver Tabla 6).

Tabla 6. Uso de los valores plausibles (PV) y de los pesos muestrales en la investigación basada en datos de PISA

PV		PM	
		Sí	24
Sí	48	No	12
		Students	12
No	3	Sí	1
		No	2
Más o menos	2	Sí	2

Nota. “PV” se refiere a valores plausibles; “PM” a los pesos muestrales; “Más o menos” dentro de “PV” indica que no se han utilizado los valores plausibles según las recomendaciones de la IEA o la OCDE; “El de los estudiantes” dentro de “PM” hace referencia a que únicamente se utiliza el peso muestral del nivel de estudiantes.

Centrándonos ahora en las 10 publicaciones restantes construidas sobre los datos obtenidos de TIMSS, de PIRLS o de la base de datos que combinaba ambas pruebas en su edición de 2011 (Grilli et al., 2016), destacamos que en todas ellas se tienen en cuenta el uso de los valores plausibles y de los pesos muestrales siguiendo las indicaciones de los organismos internacionales pertinentes.

No en todos los estudios seleccionados los autores han indicado las limitaciones que han encontrado o que se han derivado de sus investigaciones. Aun así, aquellos que sí han hecho referencia a este aspecto destacan mayoritariamente una gran limitación ligada a las pruebas a gran escala y es su carácter transversal, lo que lleva implícito que los resultados de estas investigaciones no puedan ser considerados como afirmaciones causales, sugiriendo la necesidad de la realización estudios longitudinales (Bokhove et al., 2019; Costa & Araújo, 2018; Li, 2016; Wu, Gao, et al., 2020; Wu, Shen, et al., 2020). Evidentemente, en el caso de las ILSA, este tipo de estudios son imposibles de realizar, puesto que se dirigen a una población objeto concreta (estudiantes de 15 años en PISA, por ejemplo) que en cada edición es representada por alumnos diferentes. Algunos autores indican que el no disponer de información sobre el rendimiento previo de los alumnos supone también una gran limitación, ya que no permite realizar comparaciones en función del capital cultural de las familias (Danhier, 2018; Huang & Sebastian, 2015; Marteleto & Andrade, 2014; Myrberg et al., 2019). Del mismo modo, el hecho de que en las diferentes ediciones de algunas pruebas haya escalas e indicadores que varían en función de las tendencias socioeducativas del momento, dificulta la realización de estudios longitudinales. En palabras de Martínez-Abad et al. existe una “falta de sistematicidad en la perspectiva de las evaluaciones PISA” (2020, p. 8).

De la mano de la última limitación indicada, está el posible sesgo que se deriva de la propia extracción de datos de los cuestionarios de contexto, ya que estos datos se recogen mediante autoinformes de los propios profesores, de los alumnos y de los directores de los centros (Danhier, 2018; Wu, Gao, et al., 2020; Wu, Shen, et al., 2020).

También se han indicado otras limitaciones como puede ser el compromiso metodológico que supone el uso de la media de los valores plausibles (Lee et al., 2021), en lugar de utilizarlos todos como recomiendan los organismos internacionales; que no se dispone de datos a nivel de aula (Burns et al., 2020; He & Fischer, 2020; Wu, Gao, et al., 2020), lo cual podría proporcionar una información muy relevante sobre las prácticas del profesorado; el alto número de valores perdidos en ciertas variables o en ciertos países (Guo et al., 2018) que, junto con los sesgos culturales de las propias ILSA (Lenkeit, 2012), derivan en la imposibilidad de la generalización de las conclusiones obtenidas en este tipo de estudios comparativos.

En ocasiones, algunos autores han indicado como limitación tanto a sus estudios como a las ILSA la deseabilidad social a la hora de responder a los cuestionarios por parte de los estudiantes y su “*tendency to give definite answers and use the scales’ upper ends*” (Lopes et al., 2022, p. 35).

Como norma general, las implicaciones de estos estudios derivan en recomendaciones adaptadas para cada país que tratan de paliar las deficiencias resaltadas en la investigación en relación con los factores asociados al rendimiento académico.

Por ejemplo, en el caso de Turquía, Yetişir y Bati (2021) señalan la necesidad de elaborar reformas educativas a largo plazo en el país para así tratar de mejorar la calidad de las prácticas docentes, puesto que en su estudio se encontraron con que el rendimiento educativo de los alumnos turcos podría mejorar paliando dos grandes deficiencias de su sistema educativo: la escasez de recursos educativos y de profesorado en los centros escolares.

En otros estudios, los autores aluden al hecho de que sería interesante explorar lo que realmente sucede en determinados contextos, como podrían ser las escuelas donde los alumnos rinden por encima de lo esperado (escuelas de alta eficacia), mediante la recolección de datos de carácter cualitativo que permitan profundizar en estos estudios gracias a la aplicación de métodos mixtos en la investigación (Burns et al., 2020; Ertem, 2021; Wu, Shen, et al., 2020).

4. Discusión y conclusiones

La revisión que hemos presentado supone una profundización en una línea de investigación cada vez más trabajada en la actualidad como es el análisis secundario de las ILSA y, concretamente, su relación con la eficacia escolar. El propósito de esta investigación era doble, conocer la situación actual de la investigación y sus perspectivas de futuro; es decir, de dónde venimos y hacia dónde vamos, preguntas a las que hemos tratado de dar respuesta.

Como veremos más adelante, una de las principales críticas que se hace a este tipo de estudios es la imposibilidad de establecer relaciones causales entre las variables predictoras y el rendimiento. Por este motivo, es

necesario asegurar unas garantías estadísticas que permitan, al menos, establecer inferencias sobre una base empírica robusta. En el caso de las ILSA, estas garantías estadísticas se centran en una adecuada representatividad de la muestra (uso de pesos muestrales) y una correcta representación de la variabilidad de los datos de rendimiento (uso de los valores plausibles), así como tener en cuenta la estructura jerárquica de los datos. Sin embargo, estas recomendaciones, que pueden encontrarse en todos los manuales técnicos de las ILSA¹, no se encuentran en todos los estudios recogidos en esta revisión. De las 63 publicaciones referenciadas, 34 utilizan de manera adecuada estos mecanismos estadísticos en conjunto, lo cual apunta a una necesidad de mejorar el nivel de rigor metodológico existente en la literatura científica basada en datos de evaluaciones a gran escala. Cabe destacar que la garantía más extendida es la consideración de la estructura jerárquica de los datos, con casi un 100% de investigaciones que la cumplen.

Esto, sumado a las propias limitaciones de las ILSA como que se trate de pruebas que no tienen en cuenta las diferentes culturas o la idiosincrasia de los países por el hecho de ser estandarizadas, el hecho de que promuevan la competitividad entre los países o que se trate de estudios transversales que no permiten valorar la evolución de los sistemas educativos, como indican en algunos de los estudios revisados, da como resultado que debemos cuestionarnos si, por un lado, esta es la manera más adecuada de diagnosticar la situación de los sistemas educativos a nivel mundial y si, de ser así, las comparaciones que se hacen entre países son válidas, ya que como bien sabemos existen muchísimas diferencias sociales, culturales y económicas entre ellos. De este modo, se plantea la necesidad de ahondar en esta línea de investigación y tratar de conocer si estas diferencias entre países e incluso entre territorios de un mismo país son significativas, siempre orientándolo hacia la mejora, ya no de la eficacia escolar de los centros o del rendimiento educativo de los alumnos, sino de la propia educación, tratando de lograr el pleno desarrollo personal y la participación de los individuos en la sociedad.

Si nos centramos en las perspectivas de futuro que tendría esta línea de investigación, y según los resultados obtenidos en la SLR, podríamos plantear diferentes posibilidades como la exploración de la relación que pudieran tener con la eficacia escolar otras variables de proceso menos estudiadas hasta el momento y que podrían presentar también efectos significativos en ella además de poder ser modificadas a nivel de país. Algunas de estas variables son las relacionadas con el currículo educativo y con el uso de TIC en la educación de los estudiantes. Además, también sería interesante que en las ILSA se dispusiera de la información necesaria para agrupar a los estudiantes en función de la clase en la que se encuentran facilitando así la generación de variables que permitan comparar los resultados educativos a nivel de clase y no solo según el centro educativo. Del mismo modo, se podría aplicar esta lógica de tal manera que pudiésemos comparar hasta qué punto la forma en la que los profesores imparten las clases influye en el rendimiento educativo de los estudiantes, lo que nos aportaría información sobre cuáles son las mejores prácticas de aula que ayuden a mejorar el aprendizaje de los estudiantes.

Para finalizar, cabría destacar las propias limitaciones que presenta la realización de un estudio de este tipo ligadas a la metodología empleada. Por un lado, las SLR se refieren a una línea muy específica dentro de un campo de investigación. Es decir, no se incluye todo lo estudiado sobre el tema, si no tan solo lo que hemos delimitado según nuestra cadena de búsqueda (pruebas de evaluación a gran escala y eficacia escolar, en este caso), lo que podría introducir un sesgo en la propia investigación, concretamente nuestro propio sesgo como investigadoras, puesto que somos quienes hemos delimitado este tema de estudio. Por otro lado, cualquier revisión de la literatura debe asumir el sesgo de publicación propio del sistema de producción científica actual, en el cual los resultados significativos o positivos se encuentran sobrerrepresentados, por lo que la estimación de la relevancia de las variables que puede hacerse a través de una revisión sistemática debe siempre tener en cuenta este hecho para tomar perspectiva sobre la adecuación de dichas conclusiones a la realidad educativa.

En cualquier caso, lo que ha quedado claro tras la realización de este estudio es la necesidad de explotación de las pruebas de evaluación a gran escala para conocer cuáles son los factores relacionados con la eficacia escolar y, en el caso concreto de PISA, la realización de estudios que se adecúen a las recomendaciones de la OCDE.

Financiación

Este trabajo de investigación es parte del proyecto de I+D+i financiado por el MCIN/AEI/10.13039/501100011033 titulado “EFI-SEIS. Escuelas eficaces para la mejora del Sistema” y con referencia PGC2018-099174-B-I00, y por FEDER: Una manera de hacer Europa. Del mismo modo, la investigación se realiza bajo una Ayuda para contratos predoctorales con referencia PRE2019-087412 financiada por MCIN/AEI/10.13039/501100011033 y por FSE invierte en tu futuro.

1. <https://www.oecd.org/pisa/data/pisa2018technicalreport/>

Contribución de los autores

Cristina Frade-Martínez: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Visualización, Redacción-borrador original.

Adriana Gamazo: Conceptualización, Curación de datos, Metodología, Supervisión, Redacción-revisión y edición.

Susana Olmos-Migueláñez: Conceptualización, Investigación, Metodología, Supervisión, Validación, Redacción-revisión y edición.

Referencias

- Agasisti, T., Avvisati, F., Borgonovi, F., & Longobardi, S. (2021). What School Factors are Associated with the Success of Socio-Economically Disadvantaged Students? An Empirical Investigation Using PISA Data. *Social Indicators Research*, 157(2), 749-781. <https://doi.org/10.1007/s11205-021-02668-w>
- Álvarez-Sotomayor, A., & Martínez-Cousinou, G. (2020). Inmigración, lengua y rendimiento académico en España. Una revisión sistemática de la literatura. *Revista Internacional de Sociología*, 78(3), Article 3. <https://doi.org/10.3989/ris.2020.78.3.19.083>
- Blanco Blanco, A., López Martín, E., & Ruiz de Miguel, C. (2014). Aportaciones de los modelos jerárquico-lineales multivariados a la investigación educativa sobre el rendimiento. Un ejemplo con datos del alumnado español en PISA 2009. *Revista de educación*, 365, 122-149. <https://doi.org/10.4438/1988-592X-RE-2014-365-267>
- Bokhove, C., Miyazaki, M., Komatsu, K., Chino, K., Leung, A., & Mok, I. A. C. (2019). The Role of "Opportunity to Learn" in the Geometry Curriculum: A Multilevel Comparison of Six Countries. *Frontiers in Education*, 4, 63. <https://doi.org/10.3389/feduc.2019.00063>
- Burger, K., & Walk, M. (2016). Can children break the cycle of disadvantage? Structure and agency in the transmission of education across generations. *Social Psychology of Education*, 19(4), 695-713. <https://doi.org/10.1007/s11218-016-9361-y>
- Burns, E. C., Martin, A. J., & Collie, R. J. (2020). Supporting and thwarting interpersonal dynamics and student achievement: A multi-level examination of PISA 2015. *International Journal of Research & Method in Education*, 43(4), 364-378. <https://doi.org/10.1080/1743727X.2020.1757639>
- Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: Evidence from TIMSS. *Large-scale Assessments in Education*, 4(1). <https://doi.org/10.1186/s40536-016-0030-6>
- Costa, P., & Araújo, L. (2018). Skilled Students and Effective Schools: Reading Achievement in Denmark, Sweden, and France. *Scandinavian Journal of Educational Research*, 62(6), 850-864. <https://doi.org/10.1080/00313831.2017.1307274>
- Danhier, J. (2018). How Big is the Handicap for Disadvantaged Pupils in Segregated Schooling Systems? *British Journal of Educational Studies*, 66(3), 341-364. <https://doi.org/10.1080/00071005.2017.1322682>
- Doncel Abad, D., & Cabrera Álvarez, P. (2020). Comunidades Autónomas bilingües, identidades y desempeño educativo según PISA 2015. *Revista de educación*, 387, 163-188. <https://doi.org/10.4438/1988-592X-RE-2020-387-443>
- Ertem, H. Y. (2021). Examination of Turkey's PISA 2018 reading literacy scores within student- level and school-level variables. *Participatory Educational Research*, 1, 248-264. <https://doi.org/10.17275/per.21.14.8.1>
- Formichella, M. M., & Krüger, N. S. (2017). Reconociendo el carácter multifacético de la educación: Los determinantes de los logros cognitivos y no-cognitivos en la escuela media argentina. *El Trimestre Económico*, 84(333), 165-191.
- Gamazo, A. (2019). *Factores asociados al rendimiento y a la eficacia escolar: Un estudio basado en métodos mixtos a partir de PISA 2015*. <https://doi.org/10.14201/gredos.140406>
- Gamazo, A., & Martínez-Abad, F. (2020). An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. *Frontiers in Psychology*, 11, 575167. <https://doi.org/10.3389/fpsyg.2020.575167>
- Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., & Rodríguez-Conde, M. J. (2018). Evaluación de factores relacionados con la eficacia escolar en PISA 2015. Un análisis multinivel. *Revista de Educación*, 379, 56-84. <https://doi.org/10.4438/1988-592X-RE-2017-379-369>
- García-Peñalvo, F. J. (2022). Developing robust state-of-the-art reports: Systematic Literature Reviews. *Education in the Knowledge Society*, 23, Article e28600. <https://doi.org/10.14201/eks.28600>

- Gil-Flores, J. G., & García-Gómez, S. (2017). Importancia de la actuación docente frente a la política educativa regional en la explicación del rendimiento en PISA. *Revista de Educación*, 378, 52-77. <https://doi.org/10.4438/1988-592X-RE-2017-378-361>
- Gómez-Fernández, N., & Mediavilla, M. (2021). Exploring the relationship between Information and Communication Technologies (ICT) and academic performance: A multilevel analysis for Spain. *Socio-Economic Planning Sciences*, 77, 101009. <https://doi.org/10.1016/j.seps.2021.101009>
- Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS Combined Data: Multivariate Multilevel Modelling of Student Achievement. *The Annals of Applied Statistics*, 10(4), 2405-2426. <https://doi.org/10.1214/16-AOAS988>
- Guo, S., Li, L., & Zhang, D. (2018). A multilevel analysis of the effects of disciplinary climate strength on student reading performance. *Asia Pacific Education Review*, 19, 1-15. <https://doi.org/10.1007/s12564-018-9516-y>
- He, J., & Fischer, J. (2020). Differential associations of school practices with achievement and sense of belonging of immigrant and non-immigrant students. *Journal of Applied Developmental Psychology*, 66, 101089. <https://doi.org/10.1016/j.appdev.2019.101089>
- Huang, H., & Sebastian, J. (2015). The role of schools in bridging within-school achievement gaps based on socioeconomic status: A cross-national comparative study. *Compare: A Journal of Comparative and International Education*, 45(4), 501-525. <https://doi.org/10.1080/03057925.2014.905103>
- Ibañez Martín, M. M., & Formichella, M. M. (2017). Logros educativos: ¿Es relevante el género de los estudiantes? *Education Policy Analysis Archives*, 25(3), 1-32. <https://doi.org/10.14507/epaa.25.2520>
- Instituto Nacional de Evaluación Educativa [INEE], Ministerio de Educación y Formación Profesional. (2023a). *PISA*. Instituto Nacional de Evaluación Educativa. <https://d66z.short.gy/XIVqb0>
- Instituto Nacional de Evaluación Educativa [INEE], Ministerio de Educación y Formación Profesional. (2023b). *TIMSS*. Instituto Nacional de Evaluación Educativa. <https://d66z.short.gy/MozmE8>
- Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele, UK, Keele Univ.*, 33.
- Kitchenham, B., Budgen, D., & Brereton, P. (2011). Using mapping studies as the basis for further research—A participant-observer case study. *Information & Software Technology*, 53, 638-651. <https://doi.org/10.1016/j.infsof.2010.12.011>
- Krüger, N. S. (2013). Segregación social y desigualdad de logros educativos en Argentina. *Education Policy Analysis Archives*, 21(86), 1-30. <https://doi.org/10.14507/epaa.v21n86.2013>
- Laukaityte, I., & Rolfsman, E. (2020). Low, medium, and high-performing schools in the Nordic countries. Student performance at PISA Mathematics 2003-2012. *Education Inquiry*, 11(3), 276-295. <https://doi.org/10.1080/20004508.2020.1721256>
- Lee, M., Ryoo, J. H., & Walker, A. (2021). School Principals' Time Use for Interaction with Individual Students: Macro Contexts, Organizational Conditions, and Student Outcomes. *American Journal of Education*, 127(2), 303-344. <https://doi.org/10.1086/712174>
- Lenkeit, J. (2012). How effective are educational systems? A value-added approach to study trends in PIRLS. *Journal of Educational Research Online*, 4(2), 143-173.
- Li, H. (2016). How is formative assessment related to students' reading achievement? Findings from PISA 2009. *Assessment in Education: Principles, Policy & Practice*, 23(4), 473-494. <https://doi.org/10.1080/0969594X.2016.1139543>
- Liou, P.-Y., & Hung, Y.-C. (2015). Statistical techniques utilized in analyzing PISA and TIMSS data in science education from 1996 to 2013: A methodological review. *International Journal of Science and Mathematics Education*, 13(6), 1449-1468. <https://doi.org/10.1007/s10763-014-9558-5>
- Liu, H., Damme, J., Gielen, S., & Van den Noortgate, W. (2014). School processes mediate school compositional effects: Model specification and estimation. *British Educational Research Journal*, 41. <https://doi.org/10.1002/berj.3147>
- Liu, Y. (2021). Distributed Leadership Practices and Student Science Performance through the Four-Path Model: Examining Failure in Underprivileged Schools. *Journal of Educational Administration*, 59(4), 472-492. <https://doi.org/10.1108/JEA-07-2020-0159>
- Lopes, J., Oliveira, C., & Costa, P. (2022). Determinantes escolares y de los estudiantes en el rendimiento lector: Un análisis multinivel con estudiantes portugueses. *Revista de Psicodidáctica*, 27(1), 29-37. <https://doi.org/10.1016/j.psicod.2021.05.001>
- Marteletto, L., & Andrade, F. (2014). The Educational Achievement of Brazilian Adolescents: Cultural Capital and the Interaction between Families and Schools. *Sociology of Education*, 87(1), 16-35. <https://doi.org/10.1177/0038040713494223>

- Martínez-Abad, F. (2019). Identification of Factors Associated with School Effectiveness with Data Mining Techniques: Testing a New Approach. *Frontiers in Psychology*, 10, 2583. <https://doi.org/10.3389/fpsyg.2019.02583>
- Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M.-J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies in Educational Evaluation*, 66, 100875. <https://doi.org/10.1016/j.stueduc.2020.100875>
- Muñoz-Repiso Izaguirre, M., Murillo, F. J., Barrio, R., Brioso, M. J., M^a Lourdes Hernández, & Pérez-Albo, M. J. (2000). *La Mejora de la Eficacia Escolar: Un estudio de casos*. Centro de Investigación y Documentación Educativa.
- Murillo, F. J. (2005). La investigación en eficacia escolar y mejora de la escuela como motor para el incremento de la calidad educativa en Iberoamérica. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 3(2), Article 2.
- Myrberg, E., Johansson, S., & Rosén, M. (2019). The Relation between Teacher Specialization and Student Reading Achievement. *Scandinavian Journal of Educational Research*, 63(5), 744-758. <https://doi.org/10.1080/00313831.2018.1434826>
- Ning, B., Van Damme, J., Van Den Noortgate, W., Gielen, S., Bellens, K., Dupriez, V., & Dumay, X. (2016). Regional inequality in reading performance: an exploration in Belgium. *School Effectiveness and School Improvement*, 27(4), 642-668. <https://doi.org/10.1080/09243453.2016.1202842>
- Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X., & Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement: a cross-country comparative study. *School Effectiveness and School Improvement*, 26(4), 586-611. <https://doi.org/10.1080/09243453.2015.1025796>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, Article n71. <https://doi.org/10.1136/bmj.n71>
- Pomianowicz, K. (2023). Educational achievement disparities between second-generation and non-immigrant students: Do school characteristics account for tracking effects? *European Educational Research Journal*, 22(3), 297-324. <https://doi.org/10.1177/14749041211039929>
- Quiroz, S. S., Dari, N. L., & Cervini, R. A. (2018). Nivel Socioeconómico y Brecha entre Educación Secundaria Pública y Privada en Argentina. Los Datos de PISA 2015. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 16(4), 79-97. <https://doi.org/10.15366/reice2018.16.4.005>
- Ruiz de Miguel, C. (2009). Las escuelas eficaces: Un estudio multinivel de factores explicativos del rendimiento escolar en el área de matemáticas. *Revista de educación*, 348, 355-376.
- Ruiz de Miguel, C., & Castro Morera, M. (2006). Un Estudio Multinivel Basado en PISA 2003: Factores de Eficacia Escolar en el área de Matemáticas. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, 14, 1-26.
- Sortkær, B., & Reimer, D. (2018). Classroom disciplinary climate of schools and gender – evidence from the Nordic countries. *School Effectiveness and School Improvement*, 29(4), 511-528. <https://doi.org/10.1080/09243453.2018.1460382>
- Sousa, S., Park, E. J., & Armor, D. J. (2012). Comparing Effects of Family and School Factors on Cross-national Academic Achievement using the 2009 and 2006 PISA Surveys. *Journal of Comparative Policy Analysis: Research and Practice*, 14(5), 449-468. <https://doi.org/10.1080/13876988.2012.726535>
- Tan, C. Y., & Dimmock, C. (2022). The relationships among between-class ability grouping, teaching practices, and mathematics achievement: A large-scale empirical analysis. *Educational Studies*, 48(4), 471-489. <https://doi.org/10.1080/03055698.2020.1780416>
- Tan, C. Y., & Liu, D. (2018). What is the influence of cultural capital on student reading achievement in Confucian as compared to non-Confucian heritage societies? *Compare: A Journal of Comparative and International Education*, 48(6), 896-914. <https://doi.org/10.1080/03057925.2017.1369392>
- Torres-Carrión, P. V., González-González, C. S., Aciar, S., & Rodríguez-Morales, G. (2018). Methodology for systematic literature review applied to engineering and education. In *2018 IEEE Global Engineering Education Conference (EDUCON) (Santa Cruz de Tenerife, Spain - 17-20 April 2018)* (pp. 1364-1373). IEEE. <https://doi.org/10.1109/EDUCON.2018.8363388>
- van Hek, M., Kraaykamp, G., & Pelzer, B. (2018). Do schools affect girls' and boys' reading performance differently? A multilevel study on the gendered effects of school resources and school practices. *School Effectiveness and School Improvement*, 29(1), 1-21. <https://doi.org/10.1080/09243453.2017.1382540>

- Wu, H., Gao, X., & Shen, J. (2020). Principal leadership effects on student achievement: A multilevel analysis using Programme for International Student Assessment 2015 data. *Educational Studies*, 46(3), 316-336. <https://doi.org/10.1080/03055698.2019.1584853>
- Wu, H., Shen, J., Zhang, Y., & Zheng, Y. (2020). Examining the effect of principal leadership on student science achievement. *International Journal of Science Education*, 42(6), 1017-1039. <https://doi.org/10.1080/09500693.2020.1747664>
- Yetişir, M. İ., & Bati, K. (2021). The Effect of School and Student-Related Factors on PISA 2015 Science Performance in Turkey. *International Journal of Psychology and Educational Studies*, 8(2), 2. <https://doi.org/10.52380/ijpes.2021.8.2.433>