Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Análisis metodológico del proyecto PISA como evaluación internacional

Methodological analysis of the PISA project as international assessment

Jornet Meliá, Jesús M.
Universidad de Valencia

Resumen

En este artículo el objetivo es realizar un análisis global del Proyecto PISA como programa de evaluación internacional. Asimismo, se analizan características metodológicas de los procesos métricos que sustentan los instrumentos que se utilizan en el proyecto: las pruebas de logro y los cuestionarios de contexto. Se identifican las fortalezas y debilidades del proyecto, en diversas áreas, desde el Diseño y validación del proyecto como programa de evaluación internacional, sus usos y modos de comunicación de resultados, hasta las características métricas de sus instrumentos. Se proponen alternativas para optimizar el uso del proyecto en general y en los países involucrados en particular. Se concluye que se trata de un proyecto que, por su calidad metodológica e impacto socio-político es valioso, si bien tiene diferentes aspectos que podrían mejorarse -tanto como programa de evaluación, como en sus fundamentos de medición educativas- para que pudiera aportar informaciones de mayor calidad y que pudieran orientar decisiones de mejora.

Fecha de recepción
2 Abril 2016

Fecha de aprobación
16 Junio 2016

Fecha de publicación
17 Junio 2016

Palabras clave:

PISA; análisis metodológico; evaluación del aprendizaje; cuestionarios de contexto; evaluación de sistemas educativos; educación, medición.

Abstract

In this paper the aim is to conduct a comprehensive analysis of PISA as international assessment program. In addition, methodological characteristics of the metrical processes underlying the instruments used in the project are also analyzed: achievement tests and questionnaires context. The strengths and weaknesses of the project in various areas are identified: from design and validation of the project as international assessment program, its uses and modes of communication of results to the metric characteristics of their instruments. They propose alternatives to optimize the use of the project in general and in particular the countries involved. It is concluded that this is a project that, for methodological quality and socio-political impact is valuable, although it has different aspects that could be improved -both as program evaluation, and educational measurements- fundamentals of measurement so that he could provide information higher quality and that could guide decisions for improvement.

Reception Date
2016 April 2

Approval Date
2016 June 16

Publication Date:
2016 June 17

Keywords:

PISA; methodological analysis; assessment of learning; background questionnaires; assessment of educational systems; education; measurement.

La evaluación de sistemas educativos se ha ido convirtiendo en una práctica habitual en los últimos años. Evaluaciones nacionales e internacionales, se han instaurado en el panorama socio-educativo aportando informaciones que han puesto de manifiesto la calidad educativa como un elemento central en las políticas gubernamentales.

Pese a que se vienen realizando a nivel internacional desde la década de los cincuenta

Autor de contacto / Corresponding author

Jornet-Meliá, Jesús M. Facultad de Filosofía y Ciencias de la Educación. Universidad de Valencia. Avda. Blasco Ibáñez, 30. 46010 – Valencia (España). jesus.m.jornet@uv.es

del siglo pasado, parece obvio que hasta la puesta en marcha del Proyecto PISA, su impacto social ha sido menor. A partir de él, no solo el interés político se ha focalizado en sus resultados, sino también el social, sobrepasando incluso a las informaciones ofrecidas por los profesionales implicados y derivadas de la investigación educativa.

En prácticamente todos los países iberoamericanos, su compromiso por incrementar la calidad educativa y sus esfuerzos para que la educación se convierta en un factor clave de desarrollo personal y social, de eliminación de desigualdades y prevención de la exclusión social, ha favorecido que el impacto de las evaluaciones internaciones fuera cada vez mayor en el s. XXI. En este marco de intereses y compromisos, el Proyecto PISA se ha constituido como un programa de evaluación que ha actuado como impulsor de la preocupación por la mejora educativa.

No obstante, el Proyecto PISA, si bien tiene un gran desarrollo metodológico y un elevado control, como todo estudio evaluativo tiene elementos a mejorar para que logre un verdadero impacto positivo en las sociedades y, en concreto, en los países iberoamericanos. En este trabajo, partimos de un trabajo anterior (Jornet, 2014) que realizamos en el que se analizaban las asignaturas pendientes de las evaluaciones a gran escala, por lo que en gran medida seguiremos el modo de trabajo allí planteado, con la finalidad de concretar mejor las líneas que sería interesante abordar, desde nuestro punto de vista, para la mejora del proyecto PISA.

Cuando analizamos cualquier programa de evaluación, en ocasiones, olvidamos que evaluar y medir son dos acciones diferentes. En la mayor parte de evaluaciones a gran escala, en especial en la actualidad, medir y evaluar se han presentado como una sola acción bajo el apelativo de Evaluación o incluso Diagnóstico (ver como ejemplo cualquier informe sobre evaluación de sistemas, sea nacional o internacional). Con la medición lo que pretendemos es recabar información acerca de la cantidad con que se da un fenómeno, mientras que con la evaluación, comparando la cantidad medida con criterios de valor bien especificados, aportamos juicios evaluativos. Es claro que la medición es instrumental para evaluar, pero no puede confundirse la medición con la evaluación. Aunque parezca una cuestión obvia, no lo debe ser tanto, cuando en la literatura y en la práctica evaluativa, encontramos muchas acciones que se valorizan en función de su calidad métrica, pero olvidan la validez y, más aún, los criterios y factores explicativos, que faciliten la comprensión del hecho evaluado y, por ende, representar una autentica evaluación que aporte informaciones para mejorar. Por ejemplo, en las evaluaciones diagnósticas establecidas en la legislación española se enfatiza un enfoque sistémico, si bien en la práctica lo que suele aportarse como evaluación son únicamente medidas de logro acerca del alumnado de un centro, zona, etc... Por desgracia, este mismo hecho se observa en la mayor parte de evaluaciones nacionales e internacionales, incluido el Proyecto PISA.

Por este motivo, analizaremos las cualidades del Proyecto PISA desde ambas perspectivas: como programa de evaluación y como proyecto de medición.

En el trabajo mencionado organizamos el análisis en tres núcleos de trabajo: a) las tipologías de evaluaciones a gran escala: características del Proyecto PISA, b) las asignaturas pendientes como planes de Evaluación, y c) las asignaturas pendientes de dichos planes desde la perspectiva de la Medición Educativa. En este caso, entendemos que esta estructura puede ayudarnos a organizar el análisis del Proyecto PISA. Por ello, la asumimos como referencia para el desarrollo de este artículo.

Por último, señalar que este artículo pretende, junto al de Martínez-Rizo (2016), ofrecer una visión global acerca del análisis del Proyecto PISA, a modo de presentación del conjunto de trabajos que se incluyen en este monográfico. En él participan autores de reconocido prestigio cuyas aportaciones recorren desde el impacto de PISA en el

contexto iberoamericano hasta aportaciones de tipo metodológico especializadas que pueden orientar procesos de mejora y optimización del proyecto.

El Proyecto PISA. Sus características como programa de evaluación a gran escala

No haremos mención a la historia y desarrollo del Proyecto PISA, pues es ampliamente conocida y existen referencias de gran calidad que lo describen, tanto desde la propia OCDE (OCDE, 1999), como realizadas por otros investigadores (Cordero, Crespo & Pedraja, 2013; Martínez, 2006).

Las características del Proyecto PISA, desde nuestro punto de vista, se podrían sintetizar en las siguientes:

- Se trata de un Proyecto cuya *Unidad de Análisis* (sobre la que pretende orientar la evaluación) es de carácter internacional, siendo su foco principal informar acerca del resultado diferencial de los sistemas educativos de diversos países, basado en el análisis de las competencias adquiridas por el alumnado de 15 años en tres áreas fundamentales (Matemáticas, Comprensión lectora y Ciencias) [1]
- El *Dominio Educativo* (DE, en lo sucesivo) o Universo de Medida, es la referencia de metas educativas que representan el logro de aprendizajes o competencias que constituyen el objetivo final de un sistema educativo. De este modo, nos referimos al DE como el conjunto de objetivos, actividades y tareas a que se refiere un programa educativo en general, o una materia en particular (Jornet & Suárez, 1989). En este sentido, el Proyecto PISA tuvo que afrontar desde sus inicios la dificultad de identificar un universo de medida que pudiera considerarse común para el conjunto de países evaluados. Al tratarse de un estudio internacional, era obvio que no podía tomar como referencia un diseño curricular específico, pues ello comprometería la validez de las pruebas. Por este motivo, tomó como referencia un constructo teórico de carácter educativo, para el cual (en sus tres competencias) debió realizarse un esfuerzo de definición que aparece plasmado en los marcos teóricos del Proyecto (OECD, 1999, 2003; 2006; 2009; 2012; 2015), y en el que se establece un planteamiento cognitivista en el que se enfatiza la idea de Competencia, como muestra de utilización del conocimiento adquirido hasta la edad objetivo del proyecto, para la resolución de problemas o uso de la información habitual en la vida cotidiana.
- La *población objeto*, en todos los casos, es una muestra estadísticamente representativa de los países que se integran para la realización de cada informe. Por tanto, se trata de un estudio muestral. No obstante, en cada país existe la posibilidad de ampliación de su muestra para informar acerca de algún estrato de interés local. Por ejemplo, en España, a través de las diferentes oleadas de PISA se han ido integrando como unidades de análisis independiente de España las Comunidades Autónomas (CCAA, en lo sucesivo), de forma que en el estudio se cuenta con muestras representativas, a nivel español y de algunas –o de la mayoría- de las CCAA, según la oleada (MEC, 2009; MECD, 2012). Ello permite realizar inferencias acerca de España y de las CCAA para poder representar mejor el funcionamiento diferencial de los logros educativos dentro del país.
- *La metodología que se utiliza para el Diseño de este tipo de pruebas, es la del muestreo matricial.* En este tipo de estrategia, se diseña un marco conceptual que define la competencia a evaluar, identificando los componentes fundamentales de las características de la competencia objeto de evaluación. Validado el marco conceptual por especialistas, constituye una convención que actuará como referencia para el diseño de los reactivos. A partir del mismo se diseñan los ítems, de forma que normalmente se genera un banco muy numeroso de ellos, pues la finalidad es poder representar de manera adecuada la competencia a evaluar, para asegurar la validez de constructo y de

contenido [2]. Es a partir de este banco de ítems desde donde se elaboran, mediante muestreo de matrices, cuadernillos que permiten "muestrear" el dominio de contenidos. Los cuadernillos así diseñados incluyen reactivos equivalentes en cuanto a dificultad y representan de forma sucinta el contenido, de manera que si se pudiera aplicar la totalidad de los cuadernillos se dispondría de una visión muy detallada acerca del DE evaluado. No obstante, la estrategia de muestreo de matrices tiene por objeto facilitar una logística de evaluación viable, pues no sería posible aplicar a todos los sujetos todos los reactivos. De este modo, aplicando una cantidad reducida de ellos, se puede realizar una inferencia –con cierto nivel de error- acerca del comportamiento que tendría el alumnado en el conjunto de ítems. Por ello, con este tipo de pruebas, el objetivo no es valorar individualmente al alumnado, dado que cada estudiante responde a un conjunto de cuadernillos diferentes, sino poder inferir un nivel de logro representativo de diferentes estratos muestrales. No obstante, todos los cuadernillos incluyen ítems de anclaje, que permiten establecer una estimación del logro del alumnado si se le administraran todos los reactivos, lo que permite estimar posibles valores a asignar a cada estudiante, denominados valores plausibles (OECD, 2002; 2005; 2009b; 2012b; 2014)-.

- *En todas las oleadas del Proyecto PISA, desde el año 2000, se han incluido sistemas de Cuestionarios de Contexto*, para recoger información acerca de variables y/o indicadores de entrada, proceso y contexto. Se toman como fuentes de información: el alumnado, profesorado, equipos directivos, y familias. A través de las diferentes oleadas, se han ido perfeccionando los cuestionarios de contexto y han ido incluyendo diversos indicadores complejos o compuestos por diferentes reactivos.

En síntesis y, siguiendo la tipología mencionada, el Proyecto PISA se enmarca en la categoría de Evaluaciones Internacionales, referidas a constructos teóricos, de carácter muestral.

Líneas de investigación a considerar para la mejora del Proyecto PISA

Para organizar esta revisión seguimos el esquema que se presenta en la tabla 1. Se identifican siete líneas de mejora acerca del Proyecto como Programa de Evaluación y tres referidas a sus Instrumentos de Medición.

1. El Proyecto PISA como Programa de Evaluación.

En cuanto a PISA como Programa de Evaluación, nos centraremos en tres elementos que nos parece de interés comentar-: a) el diseño y validez del programa, b) el uso de sus resultados, y c) la comunicación de los mismos.

Cualquier programa de evaluación debe nacer de un concepto concreto de calidad educativa, desde el que se especifican los criterios que sirven para estructurar el modo en que se emitirá el juicio de valor que representa la evaluación.

En el Proyecto PISA se declara como objetivo de la evaluación comprobar el nivel de competencias adquiridas por el alumnado de 15 años (al finalizar la educación obligatoria) para su desarrollo en la sociedad (OCDE, 1999).

Tabla 1. Síntesis de líneas de revisión metodológica sobre el Proyecto PISA

Líneas de análisis metodológico	Evaluación	Diseño y validación de la Evaluación (4)
		Usos de la Evaluación (2)
		Comunicación de Resultados (1)
	Medición	Validación de Pruebas de Logro (2)
		Diseño y Validación de Cuestionarios de Contexto (1)

Sin embargo, la declaración manifestada acerca del objetivo de la evaluación implica un *concepto de calidad para la evaluación basado en el concepto de excelencia*, es decir, se toman como referencia los logros del alumnado y se da por supuesto que a mayor nivel medio de los puntajes obtenidos por los sujetos de cada país, mejor es el sistema educativo del mismo. El Proyecto PISA nace de la pretensión de recabar indicadores de logro educativo para completar el marco de indicadores que trabaja habitualmente la OCDE (MECD, 2014; 2015). Por ello, no extraña que implique esa inferencia respecto al concepto de calidad subyacente de los sistemas educativos. Las legislaciones educativas de cualquier país suelen ser más ambiciosas, enfatizando como objetivo del sistema educativo la formación integral de la persona y sus consecuencias para el desarrollo humano y la transformación social. Es obvio que, si representamos el producto de un sistema educativo a partir de los logros mostrados en pruebas estandarizadas externas en tres competencias, no se puede dar respuesta a si un sistema cumple con los objetivos psico-socio-educativos para los que cada país da respuesta con su legislación a las necesidades de desarrollo personal y social.

Aspectos de índole psico-socio-afectiva que representen mejor la formación y desarrollo humano no están presentes. De igual modo, al no focalizar la evaluación sobre los procesos internos que se desarrollan en cada país, ni en las condiciones en que opera la educación (variables de entrada), ni en sus contextos, difícilmente se puede disponer de una imagen realista de calidad de los sistemas. Siguiendo a De la Orden (1997; 2007), en el modo en que describe los efectos que se consideran en las evaluaciones educativas, podríamos señalar que PISA aporta información acerca de la eficacia en cuanto al logro de resultados en las competencias evaluadas. Otros aspectos, tales como la eficiencia (grado en que se optimizan los recursos y medios para el logro de objetivos) o la funcionalidad (grado en que el sistema responde a las necesidades sociales específicas de cada país), no están implicados en el diseño del proyecto. Aún siendo conscientes de que simplificamos el análisis, el Proyecto responde a una sola cuestión: ¿Qué niveles de logro en las competencias medidas obtienen los países participantes? De manera subsidiaria, se han ido integrando análisis que involucran a más variables, como el nivel socio-económico y cultural (NSEC, en lo sucesivo) –como referencia habitual para interpretar si el logro es superior al esperado por su NSEC o no-.

Asimismo, y a partir de las aportaciones que permite el análisis estadístico de datos se introdujo el análisis acerca de la equidad del sistema. En este caso, el análisis se basa en la comparación de la varianza intra-centros e inter-centros y sus relaciones con el NSEC de las familias de los centros incluidos en el estudio. Desde nuestro punto de vista, si bien se informa de excelencia y equidad, con un planteamiento simplificado es difícil establecer la coherencia entre ambos objetivos de evaluación. Tampoco es posible que informe con el diseño actual del impacto y relevancia que tiene la educación en un país.

Como efecto de la situación descrita encontramos que, en muchas ocasiones, se interpretan mal los resultados por parte de los usuarios (políticos y administradores de la educación). *Este hecho conlleva que se*

realicen inferencias erróneas derivadas del planteamiento del proyecto como estudio evaluativo -ver estudios sobre Validez Consecuencial en Mehrens (1997), Popham, (1997) o Martínez-Rizo et al (2015)-.

De este modo, el Proyecto PISA está más cercano a un análisis basado en Indicadores que a un estudio evaluativo propiamente dicho, pues se trata de un análisis de logros guiados desde el concepto de excelencia, y no de un análisis acerca de cómo se producen los procesos educativos en los países y el impacto y relevancia que puede tener la educación en cada país (Jornet, Sánchez-Delgado & Perales, 2015). Planteamientos que la propia OCDE valora como recomendables para que se realice la educación, como los enfoques inclusivos, no se valoran a partir de PISA. En definitiva, el proyecto claramente es un estudio comparativo, diferencial, en cuanto a resultados, pero no se trata de un análisis basado en metodología comparada, por lo que es muy difícil extraer lecciones de innovación educativa.

¿Podríamos mejorar esta situación? Se pueden abordar dos líneas para dotar de mayor utilidad al proyecto. En primer lugar, se podría *ampliar el DE* del mismo incluyendo en el concepto de resultado educativo a áreas de tipo psico-socio-afectivo y de integración social. Por otra parte, hay que considerar que la calidad de un sistema educativo no puede representarse únicamente por los resultados del alumnado. Para ello deberían analizarse variables de entrada, proceso y contexto de los sistemas de cada país, y enfocar el análisis de la funcionalidad del sistema educativo en cada caso. ¿Este enfoque se debería realizar por la propia OCDE? Obviamente, se requerirían muchos más medios y podría resultar inviable, pero si se quiere aprovechar la información del proyecto, deberían ser las instituciones nacionales quienes asumieran esa labor. No obstante, para dar coherencia a los análisis sería conveniente que desde el staff de PISA se establecieran los lineamientos para la realización de este tipo de estudios. Un enfoque sistémico, holista, podría satisfacer de mejor manera las necesidades de información para orientar las decisiones a nivel de macro-sistema.

Otro elemento a considerar es la distancia que existe entre la definición de los constructos trabajados en el Proyecto PISA y las metas educativas que están presentes en los diseños curriculares de los sistemas educativos de cada país. La verdad es que no podemos responder en la actualidad a una pregunta básica desde el punto de vista de la validez de la evaluación: *¿En qué grado representa lo que se evalúa en PISA el currículum de cada país?* Es posible que sea más representativo de unos diseños curriculares que de otros. Esta carencia de información no permite contextualizar la interpretación de los puntajes nacionales y, de hecho, podría constituir un sesgo evaluativo si se diera el caso de que existieran distancias diferentes entre el DE medido en PISA y el DE de cada diseño curricular nacional. Podríamos señalar que si diferenciamos entre DE origen de las pruebas (marcos conceptuales de las competencias medidas en el proyecto), DE implementado en cada país (según sus propios diseños curriculares y circunstancias) y DE medido por las pruebas, existe un vacío en la continuidad lógica que nos ayude a analizar la validez del estudio: nos falta el análisis del DE/Origen y DE/Prueba respecto del DE/Nacional. El análisis, en este caso, no sería muy costoso de realizar, pues se podría sustentar en comités de expertos y podría ser una garantía adicional que apoyara una interpretación más certera de los resultados que aporta el proyecto –ver figura 1-.

Figura 1. Representación gráfica de los Dominios Educativos de referencia entre los que se debe establecer la coherencia para el logro de la Validez del Proyecto

Por último, en relación al Diseño del Proyecto PISA como evaluación internacional, nos parece necesario insistir sobre la necesidad de *enfatizar el análisis de las variables de entrada, proceso y contexto y su relación con los resultados*. Desde un planteamiento sistémico, "...la equidad, por ejemplo, se informaría a partir de las variables de procesos –sin descuidar la influencia del contexto, los efectos en los resultados, ni las relaciones entre las de entrada y los procesos-, pero se orientaría esencialmente hacia el análisis de si realmente se crean oportunidades para la igualdad en el sistema (procesos) y no sólo en los resultados analizados con una única referencia de entrada (el nivel socioeconómico y cultural familiar). Por ello, debería basarse en una orientación que permitiera acercarse a un enfoque de evaluación de necesidades (en el sentido en que lo propone Tejedor, 1995)" (Jornet, 2014, p. 119).

Aunque se ha ido realizando un esfuerzo considerable, incluyendo análisis secundarios de carácter investigador con datos del Proyecto, tanto desde la OCDE como por parte de algunos institutos nacionales de evaluación, y aportaciones como los estudios particulares que se han difundido como *PISA In Focus*, el esfuerzo creemos que debería ser mayor.

Otro problema que afecta al proyecto PISA, y que es común a otros estudios evaluativos a gran escala, es la selección de sujetos. Indudablemente uno de los puntos fuertes de PISA es que siguen un proceso de muestreo y recogida de información minuciosa y atenta a los factores de representatividad y calidad de la información. Sin embargo, también es cierto que las muestras se establecen para ser representativas de un país. En caso de que dentro de un país alguna de sus regiones desee ser analizada adicionalmente como unidad independiente del país, también puede ampliar su muestra. Con todo, cuando se realizan análisis secundarios que se refieren a otros estratos muestrales hay que recordar que se trabaja ya con grupos, no con muestras; lo que implica que las inferencias que se realicen se ven afectadas por un incremento del error muestral.

Superar este hecho es difícil, pues es muy costoso, en términos de incremento del número de casos a incluir, lograr muestras estadísticamente representativas en un mayor número de estratos. No obstante, es conveniente que en los estudios que se den con situaciones en las que no se trabaje con muestras estadísticamente representativas, que se indique expresamente que se trabaja con grupos, pero no con muestras, por lo que las inferencias a realizar quedan condicionadas por un mayor nivel de error. Adicionalmente, sería conveniente que los países participantes,

si quieren utilizar el proyecto PISA para responder a algunas cuestiones particulares, se involucren aumentando la muestra para los estratos necesarios. De hecho, gran parte de las informaciones que se aportan en estudios diferenciales nacionales utilizando variables diferenciadoras de estratos inferiores al poblacional, apoyados en los datos del proyecto, se realizan con grupos -excepto en algunos casos-. Si se realiza así, es preciso que, tal como hemos señalado, se relativicen las interpretaciones de cualquier análisis diferencial realizado con grupos y no con muestras.

Un tercer elemento a considerar es que *el alumnado encuestado es en todo caso "normativo"*, es decir, no se incluyen personas con discapacidades o dificultades o trastornos de aprendizaje identificados; lo que implica que se trabaja sin tener en cuenta a los grupos vulnerables sobre los que la educación debería mostrar especialmente su potencial de cambio [3]. Entendemos que la vulnerabilidad personal va más allá de la socioeconómica, que normalmente es la que se considera para referirse a "grupos vulnerables". Analizar la capacidad del sistema educativo como elemento de inclusión requiere del estudio de los procesos que se ponen en marcha para atender a este tipo de alumnado y, en todo caso, también de los resultados que se obtienen con los programas de refuerzo y las estructuras –medios personales y materiales- que se ponen en marcha para ayudarles a superar sus dificultades (Booth & Ainscow, 2000).

Por último, y como cuarto elemento a considerar en cuanto al Diseño y Validación del Proyecto PISA, presentamos algunas *consideraciones adicionales en relación a los análisis diferenciales y la valoración del cambio educativo*.

La sistematización del Proyecto PISA en cuanto a sus calendarios de aplicación entendemos que inicialmente es adecuada. Cualquier estudio periódico debe basarse en un calendario bien estructurado en el que se establezcan los elementos a evaluar. Sin embargo, este hecho que, como hemos señalado es positivo, parece no entenderse bien cuando los interesados en la evaluación reciben los resultados. Tan sólo unos ejemplos.

Al tratarse de un estudio internacional, lógicamente, no se estructuran sus periodos de aplicación vinculándolos a programas de innovación educativa que puedan haberse dado en los países participantes. Evaluar si la educación mejora o no en relación a un cambio de legislación en un país requiere de un proceso de evaluación específico que considere medidas pre-post y de seguimiento, considerando que es la Evaluación de Programas la que debe responder a este tipo de preguntas. Un proyecto como PISA puede aportar información indirecta, pero siempre será parcial, no sólo por lo comentado anteriormente, sino también porque su calendario de aplicación es independiente de lo que ocurra al interior de cada país con sus cambios legislativos. Además, el cambio educativo debería evaluarse cuando una ley se haya implantado en su totalidad, y de manera particular, según vayan produciéndose hechos o cambios en el modo de trabajo interno en escuelas y sistema que puedan tener un significado específico. Esos significados serían, en todo caso, los que deberían orientar el foco de la evaluación y, con seguridad, no siempre las respuestas están en los resultados escolares.

Por otra parte, los estudios de seguimiento que se realizan a partir del proyecto son necesariamente trasversales (o transeccionales en términos de Hernández, Fernández & Baptista, 2010). Cuando se comparan los niveles de resultados obtenidos en diferentes periodos de tiempo, es preciso considerar que se trata –por supuesto de sujetos diferentes, pues siempre se analiza a los que tienen 15 años-, pero también los cambios en los centros muestreados en cada oleada pueden implicar efectos que no se controlan, aunque el muestreo se realice con las mismas garantías de representatividad y grado de confianza. De hecho se comparan cohortes diferentes de centros dentro de cada país.

En ocasiones hemos observado cómo se presentan estos datos por instituciones nacionales como si se tratara de estudios longitudinales, sin cuestionar tan siquiera la equivalencia real de las pruebas, ni el error de la diferencia que pueda darse entre pruebas de una y otra oleada. Al trabajar con grandes muestras, el resultado más observado es el de la estabilidad a través de los años. Es obvio que falta clarificación acerca del tipo de análisis y la forma de aportar la información que se da entre diferentes oleadas de PISA. En cualquier caso, las transformaciones sociales en general y, los cambios producidos por la educación, se observan a medio-largo plazo. En periodos de tres años es difícil constatar cambios significativos en los niveles de logro. La implementación de cambios en los modos de trabajo en el interior de cada país en general, y de las escuelas en particular, requiere tiempo. Téngase en cuenta que se somete a evaluación siempre a alumnado de 15 años, por lo que los cambios a observar desde una a otra oleada de PISA deberían haberse dado en lo trabajado, al menos desde los 12 a los 15 años. Si tenemos en cuenta el énfasis diferencial que se da en el análisis de cada competencia, se debería poder observar cambios desde la aplicación realizada en una primera oleada y la tercera, en que se volviese a evaluar con el mismo énfasis cada competencia. Ello supone que desde el momento en que se aplica PISA en una primera ocasión, inmediatamente se debería estar introduciendo cambios para que el recorrido en la enseñanza obligatoria se diera de manera completa –o casi completa en cada país participante- hasta los 15 años. En cualquier caso, desde la recogida de información hasta la publicación de informes también se da un tiempo de latencia a considerar, de prácticamente un año. Como hemos señalado, los cambios socio-educativos y, más en concreto los relativos a las prácticas docentes, requieren de mayor tiempo, pues debe partirse de comprender e identificar las debilidades del sistema, arbitrar soluciones para superarlas (cambios organizacionales escolares, en metodología didáctica, formación del profesorado...), e implementarlas. Si se tiene éxito con los cambios introducidos, y los podemos observar cuando una generación completa haya sido educada en ese marco de innovación, posiblemente necesitaremos al menos entre 15 y 20 años. Por ello, aunque hemos señalado que el calendario parece adecuado, quizás sería necesario ampliar los plazos entre oleadas, al menos a cada 5 años. Y dejar esos tiempos intermedios para explotar mejor la información a nivel nacional, arbitrar medidas de mejora, etc.

Por otra parte, *la diversidad socio-cultural que se da en el mundo entendemos que es una riqueza y no un problema.* Los efectos de la globalización siempre tienen dos caras, una positiva y la otra negativa. La tensión entre la valorización de la cultura local y la internacionalización es un debate abierto, pues requiere de un análisis muy profundo que permita encontrar el equilibrio entre respeto y valoración de la cultura de un país y lo que a nivel internacional se asume como referencia. La realidad es que, tal como señalan diversos autores la definición de competencias medidas en el proyecto están orientadas a un tipo de sociedad en la que los ingenieros y tecnólogos constituyan finalmente la élite que permita evolucionar económicamente a un país en un mismo modelo de desarrollo (Martínez-Rizo, 2016). Las necesidades de formación para desarrollarse personalmente de manera adecuada en un espacio geopolítico y social determinado, no tienen por qué ser los mismos en todos los países. De ahí la necesidad de que se reconsideren las competencias a evaluar y se analicen en su definición respecto a las características locales, que no son sólo las que se definen en un diseño curricular nacional. Podemos aprender unos de otros, pero es necesario reconocer el potencial que existe en cada país, según sus circunstancias, para aportar elementos de mejora social, de inclusión, de convivencia y participación. No existe un único modelo deseable. Por ello, identificamos un reto importante para el proyecto PISA en particular y para cualquier proyecto de evaluación internacional: promover acciones de reconocimiento de la diversidad. Aunque pueda parecer anecdótico,

un ejemplo nos puede ayudar a entender mejor esta perspectiva. Se evalúan aspectos de ciencias acerca del modo en que se pueden entender los cambios climáticos; sin embargo, un agricultor o un pastor de un país del "tercer mundo", cuya formación se ha basado en vivir toda su vida en el campo y, en todo caso, escuchar a sus mayores, puede leer y predecir los cambios simplemente mirando el tipo de nubes que hay en el cielo (su altura, colores, formas...), observando la dirección del viento y el olor de la tierra y las plantas. Y, en muchas ocasiones, comprobamos que sus predicciones son más certeras que las aportadas por institutos meteorológicos basados en gran cantidad de información analizada a partir de modelos estocásticos. Los niños y adolescentes de los países desarrollados probablemente ni localizarían las estrellas o constelaciones más importantes. El mundo es diverso, y el conocimiento también. Un programa de evaluación que no esté abierto a integrar la diversidad, y se plantee a partir de las opiniones de un conjunto de especialistas de un número reducido de países desarrollados [4], además de tener fuertes carencias en cuanto a la validez de la evaluación, constituyen un elemento de guía hacia una globalización del conocimiento que lleve a olvidar la diversidad socio-cultural. Como modelo de evaluación es necesario tenerlo en cuenta, pues *un programa que olvide la diversidad corre graves riesgos en caer en la falta de justicia y equidad en la evaluación.*

Por último, también *es frecuente observar que no se generaliza analizar el tamaño del efecto* (Borges & Sánchez-Bruno, 2004; Frías, Pascual & García, 2000; Ledesma, Macbeth & Cortada De Kohan, 2008). En muchas ocasiones pequeñas diferencias se comentan como comprobadas, cuando es cuestionable que desde un punto de vista sustantivo realmente signifiquen algo.

En este sentido *¿podríamos mejorar esta situación?* Entendemos que en este caso, pueden darse soluciones técnicas que mejoren los análisis a través del tiempo. No nos referimos a que si siempre fueran los mismos centros los que actúen como referencia, fuera más realista el resultado, pues probablemente sabiendo que van a ser sujetos de análisis, acabarían orientando la enseñanza en función de qué se evalúa en el proyecto y cómo se evalúa –tipología de instrumentos-, con lo que los resultados se falsearían en la práctica. Es más válida una distribución muestral aleatoria en cada oleada, controlando la equivalencia de características de los centros que se incluyen en cada periodo, tal como se realiza, pero incrementando el control en la selección de los centros participantes, orientándose a la homogeneidad de la tipología de centros analizados entre oleadas. Entendemos que la solución, en todo caso, se puede dar aumentando la cultura de evaluación y definiendo mejor la utilidad del proyecto, sobre lo que posteriormente volveremos. En esta línea de propuestas se podrían: a) informar de manera específica de los niveles de equivalencia real estudiados entre las pruebas de las diferentes oleadas (y en concreto del error de la diferencia), b) informar del tamaño del efecto, d) informar del grado de homogeneidad en la tipología de centros analizados en cada oleada, d) explicar de la manera más explícita posible cuáles son los límites de interpretación de los resultados obtenidos, y e) integrar un planteamiento de complementariedad metodológica (Bericat, 1998), apoyando a los países para que desde sus institutos nacionales, puedan realizar estudios sobre buenas prácticas locales con enfoque de carácter cualitativo, basados en estudios de casos. En esta labor, tanto la OCDE –como responsable del proyecto- como los Institutos Nacionales tienen un trabajo indudable a realizar y que debería hacerse de manera conjunta, estableciendo lineamientos acerca de cómo entender las informaciones relativas al cambio educativo que aporta el proyecto y el modo en que desarrollar este tipo de estudios.

En cuanto a los Usos del Proyecto PISA como programa de evaluación, resaltamos tres aspectos que pueden afectar al mejor aprovechamiento de sus resultados.

En primer lugar, señalar que inicialmente y, posiblemente derivado de su origen como

estudio de indicadores, su carácter principal *es Descriptivo, es decir, describe los niveles de logro alcanzados por el alumnado de cada país, pero no se enfatiza el estudio de los factores que puedan explicarlos*. La carencia de estudios explicativos (si bien se ha ido paliando progresivamente a través de la historia del proyecto), hace difícil que se pueda comprender por qué se obtienen determinados niveles y no otros. La escasez de estudios explicativos limita sin duda su utilidad, al no favorecer la comprensión de los resultados. Ciertamente, tanto desde la OCDE, como desde los institutos nacionales y por parte de investigadores educativos independientes, se han ido realizando estudios explicativos. Las aportaciones de los estudios multinivel (Hox, 2002; Murillo, 2008; Andréu, 2011; Murillo & Hernández-Castilla, 2011a; 2011b) han ido ampliando la utilidad de este tipo de evaluaciones en general y del proyecto PISA en particular. No obstante, la escasa calidad de las informaciones que aportan nos debería poner de manifiesto que los esfuerzos no son suficientes. Se utilizan técnicas muy sofisticadas para llegar a identificar informaciones ya comprobadas desde hace muchos años. Hemos avanzado poco desde lo comprobado por Coleman et al (1966) acerca de que es el nivel socioeconómico y cultural familiar el mejor predictor del logro académico. Casi todos los estudios posteriores realizados con proyectos evaluativos, incluido PISA, aportan informaciones parecidas (Gaviria, Martínez & Castro, 2004; Lizasoain & Joaristi, 2010; Murillo & Hernández-Castilla, 2011). La sofisticación de los análisis actuales y la mejora en los cuestionarios de contexto (que ha sido evidente en el proyecto PISA) han permitido ir identificando otros factores asociados. Sin embargo, la utilidad de la información no mejora lo aportado desde otros enfoques de carácter investigador (Murillo, 2003, 2007).

Sin duda, el proyecto *PISA está diseñado para satisfacer una visión macro-analítica y no meso, ni micro-analítica.* Sin embargo, los colectivos implicados en la educación (desde los legisladores y políticos hasta el profesorado) esperan encontrar un mayor número de respuestas en el proyecto para la mejora de la organización educativa en general y de la práctica docente en particular (Jornet, García-García & González-Such, 2014). Como en casos anteriores, nos volvemos a preguntar *¿podemos adoptar alguna estrategia adicional para mejorar esta situación?*

Lógicamente el esfuerzo que se realiza para llevar a cabo el proyecto PISA y el impacto socio-político que tiene requiere plantearse modos para incrementar su utilidad. Entre ellos, comentaremos algunas opciones.

A partir de diversos estudios se pone de manifiesto que, desde el punto de vista macro-analítico, más allá de la influencia del nivel socioeconómico y cultural de los países, deben existir otras variables relativas a la estructura socio-económica de los mismos que puedan estar relacionadas con el valor social objetivo y subjetivo que se da a la educación (Jornet, Perales & Sánchez-Delgado, 2011; Sancho-Álvarez, Jornet & González-Such, 2016). Probablemente completar el proyecto con un nivel de análisis macro-analítico más comprehensivo, sistémico, podría maximizar la utilidad para orientar decisiones políticas, pero ello requiere de un modelo que trabaje de manera integral los logros con sus contextos mediatos e inmediatos.

Por otra parte, extraer conclusiones para la práctica educativa es difícil, pues el proyecto no está diseñado para ello, aunque se aporten algunos indicios (Carabaña, 2015). Sin embargo, cualquier evaluación, en términos de De la Orden (2012), tiene un carácter optimizante. En este sentido, el esfuerzo entendemos que debe llevarse a cabo desde los institutos nacionales, no por parte de la OCDE. Un buen ejemplo, lo tenemos en el trabajo realizado por el INEE de México con sus trabajos sobre PISA para docentes (ver por ejemplo, INEE, 2005). “Leer los resultados del proyecto” con intención de aprender a mejorar (en colaboración entre especialistas en medición-evaluación y docentes) conlleva claros beneficios. Tampoco podemos esperar que un proyecto de estas características nos

aporte todo. Orientar la práctica docente requiere de otros planteamientos evaluativos. Probablemente con la línea emergente de PISA para centros se pueda valorar si es una buena referencia para la mejora, aunque desde nuestro punto de vista esa línea no debería ser prioritaria en el desarrollo del proyecto, pues entra como un modelo más en el "mercado de las evaluaciones de acreditación o certificación" (junto al European Foundation for Quality Management Excellence Model –EFQM-, Normas ISO...), cuando tampoco está diseñado para ello. De momento, tan sólo es un inicio, pero estimamos que, siendo loable, deberían ser los institutos nacionales los que, como desarrollo de su cultura de evaluación, establecieran el nexo necesario entre los resultados de los informes y los usuarios potenciales de cada proyecto evaluativo y, en este caso, también.

Por último, simplemente poner de manifiesto que *el Proyecto PISA ha publicado sus bases de datos desde sus orígenes. Potencialmente existe una gran oportunidad de investigación a partir de ellas.* Son complejas, sin duda, y se requiere de una formación elevada para utilizarlas apropiadamente. Sin embargo, estimamos que los investigadores evaluativos, y educativos en general, han aprovechado poco estas oportunidades *¿Podemos mejorar esta situación?* Indudablemente, sí. Sería conveniente que el número de investigadores que profundizara en las informaciones disponibles del Proyecto se incrementara. No podemos, ni debemos, como investigadores esperar a que nos aporte todo su potencial el proyecto sin hacer nada. Los investigadores de cada país son los que mejor posición tienen –al conocer más directamente la cultura, los intereses sociales, educativos y problemas de su país- para plantear hipótesis y objetivos de estudio. Debemos implicarnos más para maximizar los beneficios que los estudios evaluativos en general, y los derivados del proyecto PISA en particular, pueden aportarnos. De hecho, desde el proyecto PISA se aporta documentación y alternativas de formación para que los investigadores aprendamos las estructuras de las bases y cómo manejar algunas particularidades de los datos, como por ejemplo, acerca de los valores plausibles.

Finalmente, en relación al Diseño y Validez de PISA, como proyecto de Evaluación, señalar que *la comunicación de resultados es posiblemente una de las áreas en que hace falta una revisión más profunda.* En evaluación siempre hay que considerar de manera diferencial a los potenciales usuarios o audiencias implicadas (Green, 1988; Weiss, 1984). La estructura del informe es la de cualquier informe de investigación, dirigida a técnicos, no a los posibles usuarios de una evaluación de estas características. Por ello, no extraña el hecho de las deficientes interpretaciones que se realizan en muchas ocasiones acerca de los resultados del proyecto. Indudablemente se trata de un informe muy cuidado técnicamente y, a través de los años, ha ido incluyendo aportaciones e innovaciones. Incluso usos gráficos que se han ido popularizando entre los investigadores y que no eran de uso frecuente. Con todo, incluso, los gráficos parecen entenderlos los investigadores, pero no siempre los usuarios profesionales de la educación. En un trabajo reciente se pudo comprobar que profesionales (con nivel de grado y máster) no comprendían algunos de los gráficos más frecuentes de uso en este proyecto (García-Bellido, 2015) ¿Es un problema del Proyecto PISA? No lo genera, por supuesto; pero indica que no tiene en cuenta el nivel de comprensión de aquéllos a quienes se dirige.

En cualquier caso, el modo en que transmite sus resultados es equívoco. En especial, destacar el uso de rankings. Como práctica evaluativa para la comunicación de resultados ha sido rechazada por muchos especialistas en evaluación (Martínez-Rizo, 2015; Martínez-Rizo, 2016; Ravela, 2002, 2003...) ¿Por qué? Hay varias razones. Comentaremos algunas de ellas.

La mera ordenación de los países por un único criterio conduce a errores de apreciación acerca de los niveles mostrados por cada uno

de ellos. Pequeñas diferencias (de uno o dos puntos) pueden ser en muchos casos obviadas por quien reciben el informe, fijándose únicamente en la posición en que se sitúa el país objetivo de su análisis. Adicionalmente, un ranking nunca está totalmente contextualizado. En el Proyecto PISA se han intentado contextualizar los niveles de logro en relación con el NSEC de los países, pero ello no es suficiente. Las características internas de los países son las que determinan si sus resultados indican un mayor o menor grado de funcionalidad para sus circunstancias particulares, si el sistema responde adecuadamente a las expectativas sociales en la medida de sus posibilidades. No responde pues preguntas básicas en educación: ¿Se han conseguido las metas que nuestro sistema se ha planteado? ¿El sistema aporta más o menos de lo que puede conseguirse en nuestro país? ¿Qué aporta el sistema educativo –identificado con un nivel de logro en una posición internacional determinada- para el desarrollo personal y social de los ciudadanos de nuestro país? Serían preguntas que nos interesaría que respondieran los estudios evaluativos de sistemas educativos. En un mundo diverso y complejo, definido por las desigualdades en oportunidades de desarrollo, una información aportada como ranking tan sólo aporta visiones erróneas que avivan los debates políticos mal informados y que no contribuyen a orientar procesos de mejora. La competitividad, que no la competencia, es lo que en todo caso estimula una forma de transmisión de resultados de este tipo. En competiciones deportivas en las que lo que importa es ganar es lícito el ranking, y da igual ganar por uno o cinco puntos, por un segundo o por décimas de segundo. La valoración social no puede ser tratada con una simplificación que conduce a errores evidentes. Tenemos demasiadas pruebas de ello y que han dado origen a señalar como un factor de calidad de los estudios evaluativos la Validez Consecuencial y que resulta una carencia de la gran mayoría de proyectos evaluativos a gran escala y del proyecto PISA también (sobre la que se trabaja en este monográfico y existen diversos estudios al respecto, como por ejemplo: Taut & Palacios, 2016).

Además, los ordenamientos simples, se realizan sin informar del nivel del error de medida que podría explicar que determinadas diferencias no significaran nada, por lo que en sí mismos constituyen una falacia ¿Algún experto podría explicar a partir de los datos del ranking si una diferencia entre 498 y 490 puntos es debida a que unos tienen adquirida tal o cual competencia y los otros no? ¿No es necesario para mejorar, cuando evaluamos, aportar las claves para comprender las diferencias sustantivas entre los hechos o fenómenos evaluativos?

¿Se podría mejorar esta situación? Las poblaciones parecen acostumbradas a los rankings y, sin duda, es una tentación para el evaluador aportar sus resultados de este modo. Se trata nuevamente de trabajar por incrementar la cultura de evaluación. Podrían aportarse diversas soluciones alternativas o complementarias. Alguna de ellas ya se ha probado en algún instituto nacional (por ejemplo por el INEE de México o el de España).

En primer lugar, sería necesario identificar los planos de análisis -macro, meso y micro analíticos- en los que se pueda comprender si determinadas prácticas políticas en la organización de la educación o en la administración escolar o de aula pueden estar relacionadas o no con los niveles de logro ¿Sería una tarea a abordar por la OCDE? No necesariamente. En este sentido la OCDE podría asumir un rol como el de director de una orquesta, en el que sin entrar a tocar directamente un instrumento armonice las informaciones para que cada país entienda mejor qué le aportan sus resultados y por dónde podría mejorar; es decir, establecer lineamientos de uso para cada plano de análisis. Distribuir responsabilidades para asumir la mejora es esencial cuando trabajamos en un enfoque global de estudio internacional.

Una segunda cuestión es que si se aportan informaciones en forma de ranking, se

especifiquen si existen diferencias estadísticamente significativas entre países. Es complejo tratar esta información con el número de países participantes, pero si se establecieran estándares realmente criteriales (Jornet & González-Such, 2009) para la interpretación de los niveles de logro, en gran medida, esta opción minimizaría los efectos negativos de la observación de rankings o de la agrupación de países por estándares normativos.

Por otra parte, otra acción que se podría arbitrar desde la OCDE, es la identificación de grupos de países, a partir de variables estructurales, tales como por ejemplo: tamaño y dispersión de la población, estructura socioeconómica de las regiones geopolíticas, tipología de la oferta de centros de enseñanza, etc. Este tipo de agrupaciones se podría abordar desde análisis de conglomerados (por ejemplo, de k-medias). Posteriormente, se trataría de analizar para los grupos resultados de logro (promedio de cada grupo y si existen diferencias entre los grupos). De ese modo se podría contextualizar mejor los resultados obtenidos.

En esta misma línea, se podría abordar el análisis de conglomerados por características socio-culturales (idioma –o idiomas existentes en países-, niveles de inmigración, movilidad poblacional interna…), o bien, conglomerados de países en función de indicadores de nivel de desarrollo económico, o por regiones geopolíticas.

En definitiva, se trataría de obviar los rankings e integrar la información de los países en categorías y contextualizar las categorías por variables de contexto, no sólo por su NSEC.

Es claro que con la información que se aporta es muy difícil, por no decir imposible, realizar recomendaciones de mejora, cuando deberíamos entender que una evaluación que no aporte este tipo de informaciones difícilmente es útil.

En conjunto, se trata de establecer lineamientos de interpretación que orienten a los usuarios hacia qué se puede afirmar en función de los resultados de un país y que no es lícito afirmar.

2. El Proyecto PISA: algunas notas sobre sus características métricas

La revisión acerca de las características del proyecto PISA en sus aspectos métricos, la organizaremos en tres grandes núcleos; dos de ellos referidos a las pruebas de logro y otro en relación a los cuestionarios de contexto.

En primer lugar, respecto a las pruebas de logro académico, nos referiremos a problemas relacionados con el Diseño de Contenidos de las pruebas que, en gran medida, han estado ya comentados en cuanto a la Validez del proyecto como Programa de Evaluación. En este caso, nos centraremos específicamente en los elementos que afectan a la validez de los instrumentos de medida propiamente dichos.

En primer lugar advertimos una carencia importante de información fundamental: falta de estudios sobre alineación de las pruebas respecto al constructo de las que se originan, y respecto a la alineación con los estándares interpretativos. Faltan por tanto evidencias empíricas acerca de la validez de constructo y contenido. Los estudios de alineación de pruebas respecto al DE origen y a los estándares de interpretación son fundamentales para asegurar la validez de la interpretación de las puntuaciones de las pruebas ¿Qué significa obtener una puntuación de 523? ¿Qué tipo de competencias se tienen adquiridas y cuáles no? Difícilmente se pueden responder estas preguntas y, cuando se intentan responder, se realizan de una manera muy primaria, refiriéndose únicamente como ejemplos a qué ítems no se han sabido responder en un país u otro, lo que habitualmente sirve tan sólo para ofrecer titulares a la prensa. Los estudios de alineación, si además se realizan, tomando como referencia una taxonomía cognitiva pueden aportar las bases necesarias para responder este tipo de cuestiones que, en definitiva, son las que expresan si los resultados de una prueba son válidos y, en consecuencia útiles *¿Se puede mejorar esta situación?* Existe suficiente metodología en el

ámbito de los estudios de validación de pruebas como para poder ofrecer respuestas más certeras, basadas en comités de especialistas que analicen los niveles de alineación (Martínez-Rizo, 2015; Rothman, 2004; Webb, 1997; Webb, Herman & Webb, 2007). No se trata de estudios complejos, pero sí minuciosos que no sólo redundan en una mejor interpretación de las puntuaciones, sino también ofrecen las evidencias necesarias para mejorar la validez las pruebas.

En esta misma línea de revisión, hay que señalar otra carencia: la falta de estudios de Validez Cultural. Un proyecto internacional implica necesariamente riesgos de que se den factores de invalidez por diferencias culturales y lingüísticas. Existen estudios con diversos proyectos internacionales que ponen de manifiesto la existencia de funcionamiento diferencial de ítems (DIF, por sus siglas en inglés) o incluso sesgo, que implicaría favorecer la observación de niveles de logro diferenciados vinculados a una deficiente elaboración de los reactivos. En estudios nacionales, si se da una diversidad cultural y/o lingüística en el país objeto de estudio, también se dan. Como en el caso anterior, se pueden arbitrar soluciones ante este tipo de problemáticas. En primer lugar, los análisis de reactivos por medio de comités de juicio siguiendo metodologías como las descritas por Solano-Flores, Contreras-Niño y Backhoff, (2006), Solano-Flores (2009, 2013) o Basterra (2011). No es esperable que en un estudio que en la actualidad ya llega a más de 80 países, las pruebas funcionen con las mismas garantías de validez, ni tan siquiera con las mismas propiedades métricas.

Cualquier prueba estandarizada desarrollada en un contexto socio-educativo determinado requiere de un estudio de adaptación y validación para su aplicación en un contexto diferente (Hambleton, 2005; Solano-Flores, 2008). Los estudios piloto que se realizan en las instituciones participantes sería conveniente que se informasen de forma integrada en un estudio específico acerca de la validez y propiedades métricas de las pruebas. Ello garantizaría la interpretabilidad del estudio y, en su caso, ayudaría a relativizar los usos e interpretaciones de las puntuaciones, dando mayores garantías de calidad a los usuarios del proyecto.

En este sentido, la colaboración inter-institucional con la OCDE sería fundamental para poder disponer de tales informaciones.

Una aportación indudable de la OCDE a través de PISA es la cantidad de informes de investigación que ha generado (ver: http://www.oecd-ilibrary.org/education/oecd-education-working-papers_19939019?page=1). Sin embargo, respecto a DIF y Sesgo resulta llamativo que en un estudio internacional, existiendo metodología adecuada para tal propósito (ver por ejemplo, Camilli, & Shepard, 1994; González-Montesinos & Jornet, 2010), no se encuentre en los informes del Proyecto, estudios de estas características. Ante tal carencia, sería un ámbito de interés a realizar por investigadores independientes. Si bien, en este caso, entendemos que la responsabilidad debería recaer en los constructores de las pruebas y, en su caso, con la colaboración de los institutos nacionales.

Por otra parte, otro factor relevante a mencionar es que el Proyecto PISA se basa en un diseño de reactivos específico, propio, característico. Se trata de una evaluación estandarizada, con "personalidad propia" (Ruiz-Primo & Li, 2015; Ruiz-Primo, Li & Minstrell, 2014; Shavelson, et al., 2002). Los avances que pueden haberse dado con sus desarrollos en las formas de abordar los diseños de ítems, por el contrario pueden constituir un factor no controlado de posible DIF y/o sesgo, al estar alejados del modo en que habitualmente se evalúe al alumnado en las aulas. Sobre este tema, sin duda, el análisis debería ser mayor que el que realizaremos en estas líneas. Somos conscientes de que para que se pueda considerar que una competencia está adquirida el alumnado debería poder resolver cualquier problemática con ella relacionada, presentada en cualquier formato evaluativo, aunque sea muy diferente del que habitualmente está acostumbrado a

enfrentarse. Sin embargo, de todos es conocido que las personas tienden a estudiar tal cual son evaluados. La forma de evaluación condiciona la expresión del nivel de logro, pues pone de manifiesto el modo en que se ha enseñado. Las soluciones al respecto plantean también preguntas a responder.

Dado que el Proyecto libera reactivos de cada oleada, ¿el profesorado debería formarse en diseñar reactivos similares para que a través de toda la educación obligatoria, hasta los 15 años, el alumnado también fuera evaluado con este tipo de formatos de ítems? ¿Entre los recursos de aprendizaje se debería incluir someter al alumnado a diferentes formatos evaluativos, incluido el formato PISA o el de otros proyectos evaluativos? Desde nuestro punto de vista, que el alumnado sea evaluado de diferentes maneras podría ser una solución que beneficiara la generalización del aprendizaje, pero es tan sólo una opinión, pues condicionaría la labor de los docentes que, en definitiva, son los que mejor conocen cómo orientar la enseñanza más si se trata de realizarla con un carácter individualizado, personalizado. Quedaría, pues, ahí planteado un debate para la Comunidad Educativa.

Un hecho que podría mejorar la comprensión de las puntuaciones observadas a partir del proyecto sería si se analizasen los modos de evaluación a que han estado sometidos los estudiantes que forman parte de cada oleada, y se investigase si la distancia entre las formas en que han sido evaluados tiene relación con las puntuaciones obtenidas. Ello podría constituir una evidencia adicional de validez. De hecho estudios sobre sensibilidad instruccional ponen de manifiesto que las distancias entre formatos evaluativos pueden ser un factor explicativo de las puntuaciones obtenidas (Ruiz-Primo & Li, 2015). No obstante ello implicaría que se establecieran protocolos de evaluación acerca del modo en que el profesorado evalúa al alumnado en cada país. No se trata de una tarea simple, resulta sin duda compleja y no sería fácilmente aprehensible desde un mero estudio externo basado en cuestionarios dirigidos al profesorado o al alumnado. Con todo hay experiencias satisfactorias en que se puede obtener una representación bastante adecuada acerca de las prácticas evaluativas que realizan los docentes (Martínez-Rizo, 2012). Es claro que este tipo de análisis debería liderarse desde el proyecto PISA (estableciendo lineamientos), pero su realización sería necesario que la asumieran los institutos nacionales, dado que debería abordarse desde la complementariedad metodológica (cuantitativa-cualitativa: Bericat, 1998).

Por último, señalar la necesidad de que se informe del grado de implementación del currículum y su distancia con las pruebas en cada país. Este aspecto ya lo mencionamos genéricamente al analizar los factores intervinientes en el Diseño y Validez del Proyecto PISA como programa de evaluación. En este caso, cuanto menos, una comprobación a partir de las informaciones que puede ofrecer el profesorado a partir de los cuestionarios de contexto, pero referido específicamente a los reactivos que se incluyen en las pruebas, podría aportar una información de gran utilidad para comprender mejor las puntuaciones obtenidas.

El segundo núcleo de análisis sobre el que nos referiremos es el relativo al estudio de propiedades métricas de la pruebas del proyecto PISA. No vamos a detenernos en este aspecto pues es, probablemente de los mejor cuidados del proyecto. Sin embargo, fue objeto de controversia en un momento dado, en relación al modo de utilización de la Teoría de Respuesta al Ítem (IRT, por sus siglas en inglés). Las dificultades fueron superadas y hoy se asume como uno de los puntos fuertes del proyecto PISA. Con todo, es conveniente recordar la necesidad de seguir protocolos de adaptación de pruebas cuando se trata de pruebas internacionales, siguiendo las recomendaciones y protocolos que se aportan en trabajos como el de Hambleton, Merenda y Spielberg (2005) o, más recientemente en el trabajo de Muñiz, Elosúa y Hambleton (2013).

La técnica de muestreo matricial para el diseño y desarrollo de pruebas es compleja y, de hecho, hay pocas personas realmente especializadas en esta metodología. Un

problema que estimamos que no está clarificado es la equivalencia real de los cuadernillos que se diseñan para cada oleada y entre oleadas. Entendemos que esta información sería clave para poder asumir si el proceso final utilizado realmente se ajusta a las propiedades métricas deseables y mantiene indicios de validación de contenido entre cuadernillos. A ello, hay que añadir que las pruebas se traducen/adaptan a diferentes idiomas de diversos países, con diferentes contextos socio-educativos, currícula, etc... Incluso, sería necesario constatar si las distribuciones de cuadernillos entre países son equivalentes o no, simplemente con un análisis frecuencial de los cuadernillos que finalmente han respondido en cada país. Este tipo de estudios sería necesario incluirlo como garantía de que los procesos de muestreo matricial cumplen con su finalidad adecuadamente.

Por último, en relación a los Cuestionarios de Contexto, desde la primera oleada de PISA hasta la actualidad se ha observado un incremento importante en su calidad y en el uso de sus indicadores, sean simples o complejos. Las características de los cuestionarios no difieren mucho de los que habitualmente se han ido utilizando en otros estudios evaluativos, sean nacionales o internacionales. Como puntos fuertes de los mismos se podría señalar que se ha realizado un esfuerzo considerable para integrar lo que denominamos indicadores complejos, y en identificar indicadores simples (reactivos) que tienen capacidad diferencial respecto a los niveles de logro. Como en tantos otros sistemas de cuestionarios de contexto, las debilidades devienen de su diseño inicial. No se identifica una teoría clara que sirva como orientación acerca de cuáles pueden ser los factores explicativos que es necesario medir. Es claro que parcialmente se apoya en hallazgos de investigación educativa acerca de algunos factores que puedan explicar el logro, pero no se observa una teoría sistémica que dé sentido al conjunto de elementos que se consideran en los cuestionarios.

Adicionalmente, no se aportan informaciones acerca de propiedades métricas de los mismos –o de las subescalas que se integran- ni de evidencias de validación. Probablemente, como instrumentos de medida, que también lo son como las pruebas de logro, reciben menos atención en el conjunto del proyecto. No se da la invarianza entre países, al contrario de lo que se da en las pruebas de logro, lo cual es lógico.

La Educación es, sin duda, una manifestación política y cultural. Por lo tanto, es expresión directa del modo psico-socio-cultural como se entiende el rol de la educación en cada país. Desde nuestro punto vista, sería deseable que se abriera, dentro del proyecto, el establecimiento de lineamientos para la configuración final de los cuestionarios de contexto, pero que se diera libertad para determinar en cada país un ámbito amplio de muestreo de informaciones a partir de ellos. En otros sistemas de evaluación se identifican indicadores de manera diferencial por países para representar mejor las características e informaciones que interesan en cada país. Por ejemplo, en el Portafolios de Laeken –aunque se trate de un modelo de evaluaciones basado en indicadores no comparable con PISA-, se introduce una estrategia que podría ser de utilidad en el proyecto que nos ocupa. Se diferencian entre: Indicadores Primarios, Secundarios y Terciarios. Los primeros y los segundos serían de obligada aplicación en cada país, recabando información sustantiva, común entre los países. Los primarios serían los que aportaran información clave, los secundarios ayudarían a matizar a los anteriores, incluyendo pequeñas adaptaciones socio-culturales de cada país. Los terciarios serían de libre disposición de cada país, de forma que se pudieran muestrear informaciones que en los momentos de desarrollo socio-educativo de cada país pudieran resultar de interés. Somos conscientes que desde el proyecto PISA se pueden incluir hasta tres reactivos en cada país, diferentes del conjunto común. No obstante, esa proporción nos parece escasa, pues no permite avanzar en el conocimiento específico

de problemáticas que pudieran ser de interés local.

La utilidad final de los cuestionarios de contexto se basa en su capacidad explicativa del logro y en su calidad para describir los elementos de entrada, procesuales y de contexto que deben servir para comprender mejor los resultados obtenidos, se establezcan o no análisis que relacionen las informaciones contextuales con las de resultados. En este sentido, entendemos que constituye una debilidad, pero que es mejorable, si se adoptan medidas de mejora de su validez y de control de sus propiedades métricas (ver revisión metodológica para el desarrollo de cuestionarios de contexto en Jornet, López-González & Tourón, 2012).

Conclusiones

Cuando se analiza un proyecto de evaluación que ha conseguido un alto nivel de impacto social y político, sus detractores están atentos a encontrar razones en trabajos técnicos que les ayuden a hacerlo desaparecer. Sus partidarios esperan lo contrario: encontrar motivos fundamentados para afirmar su valor y su permanencia.

En este caso, creemos que el proyecto PISA es mejorable, pero precisamente por el impacto socio-político que ha tenido, presenta un valor que no han conseguido otros proyectos, sean nacionales o internacionales: *situar a la educación en el centro de la preocupación social y política*. Tan sólo por eso, ya merece la pena trabajar para optimizar sus modos de hacer y sus aportaciones. Se trata, sin duda, de una oportunidad para que las sociedades centren su atención en cómo mejorar sus sistemas educativos y, a ser posible, sus modos de organización y prácticas docentes.

Más allá de esta apreciación y posicionamiento general, creemos que es necesario poner de manifiesto las carencias o lagunas para que pueda mejorarse. Se trata, sin duda, de un proyecto muy ambicioso, complejo y con una alta sofisticación técnica. Como proyecto orientado desde una institución de gran potencia, como es la OCDE, entendemos que tiene todas las posibilidades de pervivir y de asentarse en el panorama internacional; y entendemos que esto es positivo De hecho, es conveniente participar en él y que se mantenga. Si bien, estimamos que debería revalorizarse en función de la capacidad que muestre para superar temas importantes:

- Problemas en cuanto a su validez, como programa de evaluación, mejorando la validez respecto las características curriculares, socio-económicas y culturales de cada país.
- Incrementar la calidad técnica de las pruebas de logro, trabajando también para mejorar su validez, alineación curricular y a estándares criteriales, y en la elaboración de reactivos.
- Asumir un enfoque diferente en cuanto al desarrollo de los cuestionarios de contexto, dando más rol a las instituciones nacionales.
- Acercarse a proporcionar informaciones que permitan abordar Estudios Comparados, no sólo ser un estudio comparativo diferencial respecto a niveles de logro.
- Establecer planos de análisis que permitan un mayor nivel de comprensión de sus resultados para los usuarios del proyecto, lo que mejoraría su validez consecuencial.
- Trabajar en profundidad los modos en que se aporta la información evaluativa, con el fin de que esté realmente clara y sea comprensible para los usuarios.
- Finalmente, no olvidar que cualquier proyecto evaluativo implica juicios de valor y, por tanto, se trata de una acción que debe ajustarse en todo su recorrido metodológico a asegurar posiciones éticas, que afirmen la justicia y la equidad. De ahí la necesidad de los estudios de DIF y sesgo.

Líneas de desarrollo que se han ido adoptando, como por ejemplo “PISA para centros”, o sus “aplicaciones para ordenador”, entendemos que no son precisamente las prioritarias, ni van a aportar las mejoras que se necesitan para que el proyecto como programa de evaluación sea más válido.

Sólo si mejoramos la calidad del proyecto (que ya de por sí es elevada), podremos llegar a disponer de informaciones comprensibles y, en definitiva, aumentar su credibilidad y utilidad, como dos factores clave para la validez de los programas de evaluación.

En síntesis, y tal como señalamos en líneas anteriores, desde una perspectiva conceptual, la diversidad socio-cultural es un bien de la humanidad y no un problema. Los estudios internacionales deberían comprometerse en conciliar las perspectivas locales con las referencias internacionales, y que estas últimas no supongan un peligro acerca de una progresiva disminución de la diversidad. En el marco de la globalización, la educación también está inmersa. Y no puede constituir un elemento que diluya las características propias de los países, sino que coadyuve en su mejora. Entendemos que *el Proyecto PISA es en sí mismo una oportunidad para el diálogo internacional.* De ahí que estimamos necesario que en la mejora del proyecto se concilien en mayor grado las características que se orientan desde la OCDE con las que puedan resultar funcionales para cada país. Y animamos a los institutos y agencias nacionales a colaborar para que el dialogo entre todas las instancias permita encontrar soluciones que ayuden a superar algunos de los problemas aquí mencionados. El rol de establecimiento de lineamientos por parte de la OCDE y la colaboración activa por parte de las instituciones nacionales, sin duda, podría aportar mejoras. Un ejemplo al respecto ya lo tenemos: el Grupo Iberoamericano de PISA (GIP). En esta línea de colaboración y apertura es muy interesante el trabajo que aporta Andreas Schleicher en este monográfico, acerca del futuro de PISA. Se lo agradecemos. En él nos presenta propuestas de apertura hacia constructos psico-socio-afectivos y la disposición de colaboración del staff de PISA con los países participantes dentro de un modelo que concibe la educación en un mundo cada vez más interconectado, aspectos –en gran medida en línea con los comentados en este trabajo- y que sin duda pueden coadyuvar a incrementar la utilidad del proyecto.

Referencias

Andréu, J. (2011). El análisis multinivel: una revisión actualizada en el ámbito sociológico. *Metodología de Encuestas*, Vol. 13, 161-176.2 ISSN: 1575-7803

Backhoff, E., Bouzas, R., Contreras, C., Hernández, P. & García, P. (2007). *Factores escolares y aprendizaje en México. El caso de la educación básica*. México: INEE.

Basterra, M. R. (2011). Cognition, culture, language, and assessment. En M. R. Basterra, E. Trumbull & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 72-95). New York: Routledge.

Bericat, E. (1998). *La Integración de los métodos cuantitativos y cualitativos en la investigación social. Significado y medida*. Barcelona, Ariel Sociología.

Booth, T. & Ainscow, M. (2000). *Index for Inclusion*. Bristol: CSIE.

Borges, A. & Sánchez-Bruno, A. (2004). Algunas consideraciones metodológicas relevantes para la investigación aplicada. *Revista Electrónica de Metodología Aplicada*, *9*(1), pp. 1-11.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage.

Carabaña, J. (2015). *La inutilidad de PISA para las escuelas*. Madrid: La Catarata.

Coleman, J. S.; Campbell, E.; Hobson, C.; McPartland, J.; Mood, A.; Weinfeld, F. & York., R. (1966). *Equality of educational opportunity*. Washington: Government Printing Office.

Cordero, J. M., Crespo, E. & Pedraja, F. (2013). Rendimiento educativo y determinantes según PISA: una revisión de la literatura en España. *Revista de Educación*, 362, 273-297. DOI: http://dx.doi.org/10.4438/1988-592X-RE-2011-362-161

De la Orden, A. (2007). Evaluación de la calidad de la educación. Un modelo sistémico como base para la construcción de un sistema de indicadores. En INEE, *Conceptos,*

metodologías y experiencias para la construcción de indicadores educativos (pp. 6-21). México: Instituto Nacional para la Evaluación de la Educación (INEE).

De la Orden, A. (2012). La función general de la evaluación y la optimización educativa. Ponencia invitada en el *I Foro Iberoamericano de Evaluación Educativa*. México: Ensenada, UABC-IIDE (5-7 Noviembre). Recuperado de http://uee.uabc.mx/uee/eventos/primerForoRIEE/ponencias/6.pdf

De la Orden, A. (Dir.) (1997). Desarrollo y validación de un modelo de calidad universitaria como base para su evaluación. *RELIEVE*, *3*(1), art.2. DOI: http://dx.doi.org/10.7203/relieve.3.1.6334

De la Orden, A., & Jornet, J.M. (2012). La utilidad de las evaluaciones de sistemas educativos: la consideración del contexto. *Bordón*, *64* (2), 69-88.

Ercikan, K., & Solano-Flores, G. (2016). Assessment and sociocultural context: A bidirectional relationship. En G. T. L. Brown & L. Harris (Eds.), *Human Factors and Social Conditions of Assessment*. New York: Routledge.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, *27*, 275-285.

Frías, M.D., Pascual, J., & García, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema*, 12(2), 236-240.

García-Bellido, Rosario (2015). *Diseño y validación de un instrumento para evaluar la competencia "Aprender a aprender" en profesionales de la educación.* Tesis Doctoral. Universitat de València. Recuperado de http://roderic.uv.es/handle/10550/43599

Gaviria, J. L., Martínez, R., & Castro, M. (2004). Un Estudio Multinivel Sobre los Factores de Eficacia Escolar en Países en Desarrollo: El Caso de los Recursos en Brasil. *Policy Analysis Archives*, *12*(20). DOI: http://dx.doi.org/10.14507/epaa.v12n20.2004

González-Montesinos, M.J. & Jornet, J.M. (2010). *Modelo para detección de funcionamiento diferencial de reactivos (DIF) en pruebas INEE. Informe Técnico 2010*. México: INEE, Dirección General de Pruebas, Documento interno.

Green, J. (1988). Stakeholder Participation and Utilization in Program Evaluation. *Evaluation Review*, April (12), 91-116. DOI: http://dx.doi.org/10.1177/0193841X8801200201

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la investigación*. México: Editorial Mc Graw Hill.

Hox, J. (2002). *Multilevel Analysis: Techniques and applications*. London: Lawrence Erlbaum Associates.

INEE -Instituto Nacional para la Evaluación de la Educación de México- (2005). *PISA para docentes: La evaluación como oportunidad de aprendizaje.* México D.F.: INEE. Recuperado de http://www.inee.edu.mx/index.php/84-publicaciones/materiales-para-docentes-capitulos/455-pisa-para-docentes-la-evaluacion-como-oportunidad-de-aprendizaje

Jornet, J. M. & Suárez, J. M. (1989). Conceptualización del dominio educativo desde una perspectiva integradora en Evaluación Referida al Criterio (ERC). *Bordón*, 41, 237-275.

Jornet, J. M. (2014). Asignaturas pendientes en las evaluaciones a gran escala. En M. C. Cardona, y E. Chiner. (Eds.). *Investigación educativa en escenarios diversos, plurales y globales*. (pp. 115 – 128). Madrid: EOS.

Jornet, J. M., Perales, M. J. & Sánchez-Delgado, P. (2011). El Valor Social de la Educación:

Entre la Subjetividad y la Objetividad. Consideraciones Teórico-Metodológicas para su Evaluación. *Revista Iberoamericana de Evaluación Educativa*, *4*(1), 51-77. Recuperado de http://www.rinace.net/riee/numeros/vol4-num1/art3.pdf

Jornet, J.M. & González-Such, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, *16*, 103-123. Recuperado de http://dadun.unav.edu/bitstream/10171/9172/1/16%20Estudios%20Ee.pdf

Jornet, J.M., García-García, M. & González-Such, J (Eds.)-(2014). *La evaluación de sistemas educativos. Informaciones de interés para los colectivos implicados*. Universitat de València –PUV-.

Jornet, J.M., López-González, E., & Tourón, J. (Coords.) (2012). Los cuestionarios de contexto en la evaluación de sistemas educativos. *Bordón*, *64*(2). (Monográfico).

Jornet, J.M., Sánchez-Delgado, P. & Perales, M. J. (2015) *La evaluación del impacto y la relevancia de la educación en la sociedad*. Universitat de València -PUV-.

Ledesma, R., Macbeth, G. & Cortada de Kohan, N. (2008). El tamaño del efecto: una revisión conceptual y aplicaciones de la vista con sistema de estadística. *Revista Latinoamericana de psicología, 40*(3), 425-439.

Lizasoain, L. & Joaristi, L. (2010). Estudio Diferencial del Rendimiento Académico en Lengua Española de Estudiantes de Educación Secundaria de Baja California (México). *Revista Iberoamericana de Evaluación Educativa, 3*(3), pp. 115-134. Recuperado de http://www.rinace.net/riee/numeros/vol3-num3/art6.pdf

Martínez Rizo, F. (2009) Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *REDIE*, *11*(2), 1.18. Recuperado de http://redie.uabc.mx/redie/article/download/231/388

Martínez Rizo, F. (2012). *La evaluación en el aula: promesas y desafíos de la evaluación formativa*. México: Universidad Autónoma de Aguascalientes.

Martínez-Rizo, F. (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA. *RELIEVE*, *22*(1). DOI: http://dx.doi.org/10.7203/relieve.22.1.8244

Martinez-Rizo, F. (Coord.) (2015). *Las pruebas ENLACE y Excale. Un estudio de validación. Cuaderno de Investigación No. 40*. México. DF: Instituto Nacional para la Evaluación de la Educación. Recuperado de http://publicaciones.inee.edu.mx/buscadorPub//P1/C/148/P1C148_01E01.pdf

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

MECD. (2014). *Panorama de la educación. Indicadores de la OCDE 2014*. Madrid: MECD. Recuperado de http://www.mecd.gob.es/dctm/inee/indicadores-educativos/panorama2014/panorama2014web.pdf?documentId=0901e72b81b20622

MECD. (2015). *Panorama de la educación. Indicadores de la OCDE 2015*. Madrid: MECD. Recuperado de http://www.mecd.gob.es/dctm/inee/internacional/panorama-de-la-educacion-2015.-informe-espanol.pdf?documentId=0901e72b81ee9fa3

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16-18.

Ministerio de Educación, Cultura y Deporte. (2012). *PISA 2012. Programa para la Evaluación Internacional de los alumnos. Informe español*. Madrid: MECD. http://www.mecd.gob.es/dctm/inee/internacional/pisa2012/pisa2012lineavolumeni.pdf?documentId=0901e72b81786310

Ministerio de Educación (2009). PISA 2009. *Programa para la evaluación internacional de los alumnos. Informe español*. Madrid:

Ministerio de Educación. http://www.mecd.gob.es/dctm/ievaluacion/internacional/pisa-2009-con-escudo.pdf?documentId=0901e72b808ee4fd

Muñiz, J., Elosua, P. & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, *25*(2), pp. 151-157. DOI: http://dx.doi.org/10.7334/psicothema2013.24

Murillo, F.J. (2003). Una panorámica de la investigación iberoamericana sobre eficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *1*(1), 1-14.

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

Murillo, F.J. (2008). Los modelos multinivel como herramienta para la investigación educativa. *Magis, Revista Internacional de Investigación en Educación*, 1, 45-62.

Murillo, F.J. & Hernández-Castilla, R. (2011a). Factores escolares asociados al desarrollo socio-afectivo en Iberoamérica. *RELIEVE*, 17(2). DOI: http://dx.doi.org/10.7203/relieve.17.2.4007

Murillo, F. J., & Hernández-Castilla, R. (2011b). Efectos escolares de factores socio-afectivos. Un estudio Multinivel para Iberoamérica. *Revista de Investigación Educativa*, 29(2), 407-42. Recuperado de http://revistas.um.es/rie/article/view/111811

OECD. (1999). *Measuring student knowledge and skills. A new Framework for assessment.* Paris: OECD. Recuperado de http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33693997.pdf

OECD. (2002). PISA 2000 Technical Report. Paris: OECD. Recuperado de https://www.oecd.org/pisa/pisaproducts/33688233.pdf

OECD. (2003). *The PISA 2003. Assessment Framework. Mathematics, Reading, Science and Problem Solving knowledge and skills.* Paris: OECD. Recuperado de http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33694881.pdf

OECD. (2005). *PISA 2003. Technical Report.* Paris: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/pisa2003technicalreport.htm

OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy. A Framework for PISA 2006.* Paris: OECD. Recuperado de http://www.oecdilibrary.org/docserver/download/9806031e.pdf?expires=1465551043&id=id&accname=guest&checksum=02E5A7F7B73F1CCFA0DA6E8336A2F3D8

OECD. (2009). *PISA 2009. Assessment Framework. Key competencies in reading, mathematics and science.* Paris: OECD. Recuperado de https://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD. (2009b). *PISA 2006. Technical Report.* Paris: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/42025182.pdf

OECD. (2012). Pisa 2012. *Assessment and analytical Framework. Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD. https://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf

OECD. (2012b). *PISA 2009. Technical Report.* París: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/pisa2009technicalreport.htm

OECD. (2014). *PISA 2012. Technical Report.* París: OECD Publishing. Recuperado de https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

OECD. (2015). PISA 2015. *Assessment and Analytical Framework. Science, Reading, Mathematic and Financial Literacy*. Paris. OECD. http://www.keepeek.com/Digital-Asset-Management/oecd/education/pisa-2015-assessment-and-analytical-framework_9789264255425-en#page201

Popham, W. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9-13.

Ravela, P. (2002). *¿Cómo presentan sus resultados los sistemas nacionales de evaluación educativa en América Latina?* Documento de Trabajo No. 2. Santiago de Chile: PREAL. Recuperado de http://www.preal.org/docs-trabajo/ravela_n22.pdf

Ravela, P. (2003). *¿Cómo aparecen los resultados de las evaluaciones educativas en la prensa?* Grupo de Trabajo sobre Estándares y Evaluación del PREAL. Recuperado de http://www.preal.cl/GTEE/pdf/prensa.pdf

Rothman, R. (2004). Benchmarking and alignment of state standards and assessment. En S. Fuhrman and R. Elmore (Eds.), *Redesigning accountability systems for education (pp. 96-114)*. New York: Teachers College Press.

Ruiz-Primo, M.A. (2006). A Multi-Method and Multi-Source Approach for Studying Fidelity of Implementation. CSE Report 677. SEAL, Stanford University/CRESST.

Ruiz-Primo, M. A., & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA Science items. *Teachers College Record*, *117*(1), 1-36.

Ruiz-Primo, M. A., Li, M., & Minstrell, J. (2014). Building a framework for developing and evaluating contextualized ítems in science assessment (DECISA). Proposal submitted to the DRL–CORE R7D Program to National Science Foundation. Washington, DC: National Science Foundation.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 269-393. DOI: 10.1002/tea.10027

Ruiz-Primo, A., Jornet, J.M. & Backhoff, E. (2006). *Acerca de la Validez de los exámenes de la calidad y el logro educativos (Excale).* México: Instituto Nacional de Evaluación Educativa. Recuperado de http://www.uv.es/gem/gemhistorico/publicaciones/Acerca_de_la_Validez_de_los_examenes_de_la_calidad_y_el_logro_educativos_Excale.pdf

Ruiz-Primo, M.A.; Li, M.; Wills, K.; Giamellaro, M.; Ming-Chih, L.; Mason, H., & Sand, D. (2012). Developing and Evaluating Instructionally Sensitive Assessments in Science. *Journal of research in science teaching*. *49*(6), pp. 691–712.

Sancho-Álvarez, C., Jornet, J. & González-Such, J. (2016). El constructo Valor Social Subjetivo de la Educación: validación cruzada entre profesorado de escuela y universidad. *Revista de Investigación Educativa*, 34(2). DOI: http://dx.doi.org/10.6018/rie.34.2.226131

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2002). *Evaluating new approaches to assessing learning*. CSE Technical Report 604. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California, Los Angeles.

Solano-Flores, G. & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2-3), pp. 245-263. DOI: http://dx.doi.org/10.1080/13803611.2013.767632

Solano-Flores, G. (2008, July). A conceptual framework for examining the assessment capacity of countries in an era of globalization, accountability, and international test comparisons. Paper given at the *6th Conference of the International Test Commission*. Liverpool, UK.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. En M. R. Basterra, E. Trumbull, & G. Solano-Flores, *Cultural validity in assessment* (pp. 3-21). New York: Routledge.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational*

Measurement: Issues and Practice, 28(2), pp.9-18.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2006). Test translation and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons. *REDIE: Electronic Journal of Educational Research, 8*(2). Recuperado de http://redie.uabc.mx/vol8no2/contents-solano2.html

Tejedor, F. (1995). Perspectiva metodológica del diagnóstico y evaluación de necesidades en el ámbito educativo. Metodología en el diagnóstico y evaluación en los procesos de intervención educativa. *Actas del V Seminario de Modelos de Investigación Educativa.* Murcia: AIDIPE.

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Science Education.* Madison, Wisconsin: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N.M., Herman, J. & Webb, N.L. (2007). Alignment of mathematics state-level standards and assessment: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17-29.

Weiss C.H. (1984). Toward the future of stakeholders approaches in evaluation. En R.F. Conner, D.G. Altman & C. Jackson (Eds), *Evaluation studies. Review Annual, 9,* pp. 255-268. Beverly Hills, California: Sage Publishers.

Notas

[1] Aunque no en todos los países la enseñanza obligatoria concluye a los 15 años, se adoptó esa convención por ser la más representativa.

[2] Utilizado en la mayor parte de pruebas a gran escala de tipo muestral: TIMSS, PIRLS (ambas de la IEA), Pruebas diagnósticas estatales de España, EXCALE del INEE de México, por ejemplo.

[3] Existe evidencia de que este tipo de casos se eliminan de la muestra en el momento de recoger información. El esquema comparativo de niveles de logro conlleva que en muchos países se evite que este tipo de alumnado sea tenido en cuenta, pues se aspira a mostrar el mejor nivel.

[4] Véase en el Anexo I la síntesis de especialistas que han participado en el desarrollo de los marcos teóricos de cada competencia. La diversidad de especialistas –sin duda muy bien formados- de qué –y cuántos- países provienen. De ahí la necesidad de las aportaciones del GIP, como inicio de considerar las características propias iberoamericanas.

Agradecimientos

Este trabajo se ha realizado en el marco del proyecto I+D+I Sistema educativo y cohesión social: diseño de un modelo de evaluación de necesidades (SECS/EVALNEC). Ref. EDU2012-37437, financiado por el Ministerio de Economía y Competitividad de España.

Anexo 1

Países a los que pertenecen los especialistas responsables de los marcos conceptuales de las pruebas

2000	2003	2006	2009	2012	2015
Matemáticas					
Holanda, Italia, Irlanda, España, Dinamarca, Corea, USA, Austria	Holanda, Japón, Alemania, 2 USA, Rep. eslovaca, Irlanda, Polonia, Corea,	Holanda, USA, Alemania, Japón, Polonia, Dinamarca,	Holanda, Alemania, USA, Polonia, Dinamarca, Japón	2 Australia, Alemania, 3 USA, Japón, Polonia, Dinamarca, Reino Unido	

	Dinamarca, España,				
Lectura					
2 USA, Reino Unido, Canadá, Holanda, Bélgica, Finlandia, Francia, Alemania, Japón	2USA, Reino Unido, Canadá, Holanda, Bélgica, Finlandia, Francia	Holanda, 2 USA, Reino Unido, Canadá, Bélgica, Finlandia, Francia	2 USA, Japón, Holanda, Reino Unido, Bélgica, Corea, Francia, Alemania, España		
Ciencias					
Reino Unido, Australia, Suiza, Corea, Noruega, Alemania, 2 USA	Reino Unido, Australia, Suiza, Noruega, Alemania, 2 USA, Corea,	USA, Polonia, Australia, Rep. eslovaca, Noruega, 2 Francia, Italia, Reino Unido, Japón, Alemania	Australia, Noruega, Japón, 3 Alemania, 2 Francia, Holanda, China, 2 USA, Finlandia,		2 Reino Unido, USA, Alemania, Sudáfrica, Francia, Australia, Singapur
Resolución de problemas					
	2 USA, Hungría, Reino Unido, 2 Holanda, Alemania, Grecia			Alemania, Hungría, 3 USA, Alemania, Luxemburgo, Singapur	
Expertos técnicos					
			5 USA, Australia, Francia, Bélgica, Francia, 2 Holanda,	Alemania, México, Singapur, 3 USA, 2 Holanda,	4 USA, 2 Alemania, Chile, Japón, Noruega, Chipre, Holanda
				Literatura financiera	**Literatura científica**
				2 USA, Francia, Nueva Zelanda, Australia, Rep. Checa, Canadá, Reino Unido	3 USA, Dinamarca, Holanda, Italia, Japón, China

Fuente: Extraído de OECD 2000; 2003; 2006; 2009; 2012; 2015.

Nota: cuando aparece un nº junto a un país indica el número de especialistas de dicho país que han intervenido en el diseño del marco conceptual en cada caso.

Autores / Authors

To know more / Saber más

Jornet-Meliá, Jesús M. (jesus.m.jornet@uv.es).
Catedrático de Evaluación y Medición Educativa del departamento de Métodos de Investigación y Diagnóstico en Educación de la Universidad de Valencia (España). Es uno de los co-directores de esta sección monográfica sobre "Evaluaciones internacionales: PISA". Su dirección postal es: Facultad de Filosofía y Ciències de la Educación. Universidad de Valencia. Avda. Blasco Ibáñez, 30. 46010-Valencia (España).

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]