Martínez-Rizo, Felipe (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de Méxito y el Grupo Iberoamericano de PISA. *RELIEVE*, *22* (1), art. M0. DOI: http://dx.doi.org/10.7203/relieve.22.1.8244

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA

Impact of large-scale tests in contexts of weak technical tradition: Experience of Mexico and the Iberoamerican Group for PISA

Martínez-Rizo, Felipe
Universidad Autónoma de Aguascalientes

Resumen

El artículo inicia con un breve repaso del desarrollo de la psicometría, que destaca la brecha entre la situación de Estados Unidos y la de otros países, sobre todo de menor desarrollo, en cuanto a la existencia de personal calificado en temas técnicos complejos. En seguida se hacen consideraciones sobre la noción de validez y su importancia, en especial en cuanto a la validez de consecuencias. Luego se describe el impacto de las evaluaciones, con especial atención a las de gran escala y las internacionales, con el caso de PISA, considerando en especial el caso de países con una tradición psicométrica débil. En la conclusión se discute el tema, a partir de la experiencia de México y del Grupo Iberoamericano de PISA.

Fecha de recepción
23 Abril 2016

Fecha de aprobación
22 Junio 2016

Fecha de publicación
23 Junio 2016

Palabras clave:
Pruebas en gran escala; Psicometría; Validez; Validez de consecuencias; PISA.

Abstract

The paper begins with a brief review of the development of psychometrics, which highlights the gap between the situation in the US and in other countries, especially less developed, as to the existence of qualified personnel in complex technical issues. Then considerations are made on the notion of validity and its importance, especially concerning consequential validity. Impact of the tests is then described, with special attention to large scale and international, with the case of PISA, and mainly concerning countries with weak psychometric tradition. In the conclusion the whole issue is discussed, from the experience of Mexico and the Latin American Group of PISA.

Reception Date
2016 April 23

Approval Date
2016 June 22

Publication Date:
2016 June 23

Keywords:
Large-scale tests; Psychometry; Validity; Consequential Validity; PISA

EL DESIGUAL DESARROLLO DE LA PSICOMETRÍA

La situación atípica de los Estados Unidos

La disciplina especializada en medición en ciencias del hombre se desarrolló sobre todo en los Estados Unidos desde fines del siglo XIX y a lo largo del XX. Cattell inventó la palabra *test* con el texto *Mental Tests & Measurements* de 1890. Las pruebas de inteligencia desarrolladas por Binet en Francia fueron adaptadas por Terman en Stanford (1917), y se extendieron al ser utilizadas por el ejército americano. (De Landsheere, 1996: 56-71). En 1904 se había publicado en Nueva York la obra clave de Edward L. Thorndike, *An Introduction to Theory of Mental and Social Measurement*. (Martínez-Arias, 1995).

El *College Board* nació en 1900 (*College Entrance Examination Board*), para racionalizar los procesos de selección para ingresar a universidades del noreste americano, con pruebas de tipo ensayo. La

Autor de contacto / Corresponding author
Martínez-Rizo, Felipe. Universidad Autónoma de Aguascalientes (México): felipemartinez.rizo@gmail.com

dificultad para calificar de manera rápida y confiable cantidades crecientes de ese tipo de exámenes llevó al desarrollo de pruebas llamadas *objetivas*, de respuesta abierta breve o pre-estructurada, especialmente de opción múltiple, las primeras de las cuales, del *Scholastic Aptitude Test* (SAT), fueron administradas por primera vez en 1926. Un avance importante fue la equiparación de versiones y el cuidado de la estabilidad de la prueba a lo largo del tiempo, que comenzaron a hacerse en 1941 (Donlon, 1984).

En 1948 la Universidad de Princeton dio lugar al *Educational Testing Service* (ETS); en la década siguiente surgió el *American College Testing* (ACT) y un centro de la Universidad de Iowa (De Landsheere, 1996:150, nota 4). Para 1950, la teoría de las pruebas en su versión clásica puede considerarse completa con la aparición del libro *Theory of Mental Tests*, de Gulliksen. (Martínez Arias, 1995)

En la segunda mitad del siglo XX el avance siguió, con creciente difusión de las pruebas, asociada a la preocupación por la calidad educativa que provocaron el Informe Coleman (1966) y la baja del promedio de los resultados del SAT. Ya en 1957 el lanzamiento del Sputnik se había interpretado como evidencia de que la URSS estaba superando a los Estados Unidos en la carrera espacial, porque tenía mejores científicos e ingenieros y mejor educación en matemáticas y ciencias (Mathison & Ross, 2008). Varios estados hicieron obligatorio evaluar regularmente a los alumnos de educación obligatoria con pruebas diseñadas con base en *estándares mínimos de desempeño*. En 1982 en 42 de los 50 estados había *pruebas de competencia mínima*, muchas veces deficientes, por lo que no cumplían las expectativas depositadas en ellas y aumentó el número de demandas judiciales que las cuestionaban por discriminatorias, sesgadas y poco fiables. (Baker & Choppin, 1990)

En 1969 se puso en marcha el *National Assessment of Education Progress* (NAEP), para evaluar el nivel de la educación a escala nacional. En 1983 se encomendó su operación al ETS (Walberg, 1990). En ese mismo año se publicó el informe *A Nation at a Risk*, solicitado por el presidente Ronald Reagan, con el que inició el *movimiento de estándares educativos* (Mathison & Ross, 2008). En la *Cumbre Educativa de Charlottesville* (1989), los 50 gobernadores de los estados norteamericanos adoptaron metas para el año 2000. Una meta establecía que para esa fecha los alumnos americanos serían los primeros del mundo, terminando los grados 4°, 8° y 12° con altos niveles de competencia en temas exigentes de inglés, matemáticas, ciencias, historia y geografía (Mathison & Ross, 2008).

En 2002 el presidente Georges W. Bush firmó la ley *No Child Left Behind* (NCLB), que obligaba a todos los estados a definir estándares educativos y crear sistemas de evaluación alineados con ellos, con pruebas anuales de inglés, matemáticas y ciencias para todos los alumnos de 4° a 8° grado. La participación en las pruebas del NAEP pasó a ser obligatoria para acceder a fondos federales para programas de mejora educativa. Con la ley NCLB las pruebas se volvieron de alto impacto, porque los resultados de los alumnos en las pruebas eran el criterio para definir si una escuela conseguía o no el avance estipulado para recibir apoyo (*adequate yearly progress*) o incluso para ser cerrada si no lo conseguía. La ley que sustituye a la NCLB (*Every Student Succeeds Act*,ESSA), firmada por el presidente Obama el 10 de diciembre de 2015, reduce el peso que se atribuye a las evaluaciones, pero aún es demasiado pronto para valorar su impacto.

La situación en otros países y las evaluaciones internacionales

Hasta la Segunda Guerra Mundial en otros países industrializados no había avances comparables en el campo de la psicometría. La diferencia era tal que, en 1931, al escuchar que los asistentes a un congreso internacional se referían a los tests como algo típicamente estadounidense, Thorndike reaccionó diciendo que *por el bien de la ciencia y por nuestro bienestar, sería preferible que las pruebas no fueran llamadas exámenes estadounidenses.*

(Joncich, 1968, según De Landsheere, 1986: 68, nota 24)

En la segunda mitad del siglo XX los cambios sociales, los de los sistemas educativos y los avances de la metodología psicométrica trajeron consigo una rápida difusión de las pruebas en gran escala. El NAEP sirvió como referente para sistemas de pruebas para el monitoreo de la calidad educativa en países como Australia (ACER) y Holanda (CITO).

Los hechos que agudizaron la preocupación por la calidad educativa en Estados Unidos, en especial el Sputnik, llevaron también al desarrollo de evaluaciones internacionales. Como la diferencia de currículo y forma de evaluar el aprendizaje impedían comparar resultados de alumnos de distintos países, un grupo encabezado por Torsten Husén planteó en 1958 la idea de evaluaciones que dieran resultados comparables a nivel internacional. En 1959 se organizó un estudio piloto y en 1964 se aplicó el primer estudio sobre matemáticas. En 1966 se creó la *International Association for the Evaluation of Educational Achievement* (IEA), que realizó otros estudios en las décadas de 1960 y 1970. En la de 1980 llevó a cabo un segundo estudio sobre matemáticas, otro sobre ciencias y otro sobre escritura. Hasta mediados de los 90 la IEA condujo otros estudios, en especial el tercero de matemáticas y ciencias (*Third International Mathematics & Science Study*, TIMSS). Luego se han desarrollado otros como el de educación cívica; el TIMSS (ahora *Trends in Mathematics and Science Study*) adoptó un ritmo de aplicaciones de cuatro años, y el estudio de lectura (*Progress in International Reading Literacy Study*, PIRLS) un ritmo de cinco años. (De Landsheere, 1994; Husén & Neville-Postlethwaite, 1996; Postlethwaite, 1985).

A partir del año 2000 comenzaron las aplicaciones de otro proyecto internacional que ha llegado a ser el más conocido, promovido por la Organización para la Cooperación y el Desarrollo Económico (OCDE): *Programme for Institutional Student Assessment* (PISA).

Hoy existen sistemas de evaluación en casi todos los países de la Unión Europea y en otros de desarrollo alto, como Japón, Corea del Sur, Singapur e Israel. Comienzan a implantarse en países árabes, y en África, con apoyo del Instituto Internacional de Planificación de la Educación de la UNESCO, se han desarrollado en países francófonos y en los que han formado el *South African Consortium for the Monitoring of Educational Quality*, SACMEQ (Ross, 1994; SACMEQ, 1995).

Situación en el ámbito Iberoamericano

Como otros países de desarrollo intermedio o bajo, los de América Latina no tienen una fuerte tradición en lo relativo a evaluaciones en gran escala del aprendizaje. En la década de 1960 comenzaron a hacerse pruebas para el acceso a carreras universitarias. En educación básica, con pocas excepciones entre las que destaca el caso de Chile, fue hasta los años 1990 cuando se extendieron en la región las pruebas estandarizadas, en programas tanto nacionales como internacionales (Cfr. Wolff, 2004).

En 1994 se creó el Laboratorio Latinoamericano para la Evaluación de la Calidad Educativa (LLECE), que ha organizado tres *Estudios Regionales Comparativos y Explicativos*, cuyos resultados se difundieron en 1997, 2008 y 2015. Hoy los países de la región han implantado sistemas de pruebas en gran escala, comenzando con Chile en la década de 1980, seguido por casi todos en la década de 1990 (Cfr. Ferrer, 2006; GTEE, 2007-2008; Martínez-Rizo, 2008). Recientemente, y además de Chile, han desarrollado pruebas censales México, Brasil, Colombia, Costa Rica, República Dominicana, Ecuador, El Salvador, Guatemala, Perú y Uruguay. (Cfr. Martínez-Rizo, 2010a)

Por lo que hace a PISA, además de España y Portugal, como miembro de la OCDE México participa desde 2000, cuando lo hizo también Brasil, de manera voluntaria, seguido en

distintos momentos por Argentina, Chile, Perú, Uruguay, Colombia, Costa Rica, Panamá y la República Dominicana.

La débil tradición en el manejo de pruebas en gran escala que es común a los países de Iberoamérica –España y Portugal, Brasil y países hispanohablantes de América Latina— explica que en ellos hay pocos especialistas con formación suficiente para utilizar modelos psicométricos como los de las teorías de Respuesta al Ítem y de la Generalizabilidad, modelos lineales jerárquicos, modelos de ecuaciones estructurales y otros necesarios para el tratamiento adecuado de las complejas bases de datos que general esas pruebas. Tampoco abundan personas con la competencia necesaria para diseñar y llevar a cabo procesos de validación que tengan en cuenta las diversas dimensiones de la noción de validez.

LA CALIDAD DE LAS PRUEBAS Y SU IMPACTO

La validez en general

Hasta mediados del siglo XX la noción de validez se enfocaba a la predicción de un criterio particular: *en un sentido general, una prueba es válida para cualquier cosa con la que se correlaciona* (Guilford, según Messik, 1989: 18). Después las definiciones se centraron en un limitado número de tipos: de contenido, criterio (predictiva y concurrente) y constructo, con preponderancia creciente del último, ya que la validez de constructo subsumía la de contenido y la de criterio. Para Messick la validez es *un juicio evaluativo integrado del grado en que la evidencia empírica y los fundamentos teóricos apoyan la adecuación y conveniencia de las inferencias y acciones basadas en puntajes de pruebas u otras formas de evaluación* (1989: 13). Luego el foco se centró en la interpretación de los puntajes obtenidos con un instrumento de medición. Se mantuvo el énfasis en la validez de constructo, como la esencia de una concepción unitaria de validación y que se resume en la afirmación de Cronbach (1988): *toda validación es una sola*.

Coincidiendo con la definición que se manejaba ya en 1999, la versión más reciente de los *Standards for Educational and Psychological Testing* define validez como *el grado en que evidencia y teoría respaldan las interpretaciones de los puntajes de una prueba para los usos que se pretende hacer de ellos* (AERA-APA-NCME, 2014: 11). Esta definición coincide con las visiones de Messick (1989) y Kane (2006), en el sentido de que el proceso de validación debe enfocarse a la *interpretación* y los *usos* de las puntuaciones obtenidas mediante un instrumento de medición. La validez no es un atributo del instrumento que se utiliza para recabar la información, sino de las interpretaciones y usos que se haga de ella.

Validar una inferencia interpretativa es comprobar el grado en que es sustentada por múltiples tipos de evidencias, mientras que inferencias alternativas están menos soportadas. Validar una inferencia de acción requiere validar no sólo el significado de cierto puntaje en un instrumento de medición, sino el valor de las implicaciones y de los resultados de las acciones, especialmente valorando la relevancia y utilidad de la puntuación en una prueba para un propósito específico, así como las consecuencias sociales de usar una puntuación para tomar decisiones. Aunque hay diferentes fuentes y mezclas de evidencias que soporten las inferencias realizadas a partir de las puntuaciones, *la validez es un concepto unitario que siempre refiere al grado en que la evidencia empírica y el fundamento teórico apoyan lo adecuado de las interpretaciones y acciones realizadas a partir de las puntuaciones de un instrumento* (Messick, 1989: 13).

Hoy se enfrentan posturas contrastantes, como puede verse en Borsboom, Mellenbergh y van Heerden, 2004, o en el número 1 de 2013 del *Journal of Educational Measurement.* Una presentación de Newton (2013) enumera 149 acepciones del término validez, y en algunos casos se llega a proponer que se abandone el concepto (Michell, 2000).

Las versiones más elaboradas de la conceptualización predominante incluyen la idea de la validez como argumento (Kane, 2006 y 2013). Una dimensión reciente es la de *validez cultural*, que destaca la importancia de cuidar desde el diseño y desarrollo de una prueba, la forma en que factores culturales, lingüísticos y sociales distintos de los constructos de interés pueden influir en la forma en que los sujetos interpretan el contenido de los ítems y en que los responden. (cfr. Basterra, Trumbull y Solano-Flores, 2011)

La validez de consecuencias

Al ver cómo evolucionó la noción de validez hasta que la dimensión de constructo fue omnipresente en el proceso de validación, se advertirá que la única fuente de evidencia no explícitamente incorporada a la validez de constructo es la que evalúa las consecuencias sociales, lo que según Messick es irónico, ya que la validez era concebida inicialmente en términos funcionales: qué tan bien hace una prueba la tarea para la que fue diseñada. La validez de consecuencias apareció en los estándares 1999 AERA-APA-NCME, e introdujo una complejidad que para algunos hizo más confuso el panorama en lugar de aclararlo y dio lugar a fuertes discusiones. Otros señalan la necesidad de incluir la nueva dimensión debido al cambio de pasar del ámbito psicométrico al de políticas, que hizo que la valoración de las evaluaciones no pueda limitarse a lo técnico y deba incluir lo relativo a consecuencias.

Hoy las principales conceptualizaciones de validez, comenzando con la versión 2014 de los estándares de AERA, APA y NCME, incluyen las consecuencias –individuales o sociales, deseadas o no, previstas o imprevistas– que trae consigo el uso de una prueba. (Kane, 2013; Moss, 2008; Sireci, 2013)

La validez, por otra parte, es cuestión de grado, no de todo o nada, y al pasar el tiempo las evidencias se fortalecen o debilitan por nuevos hallazgos, incluyendo que las previsiones de las posibles consecuencias sociales de las evaluaciones se transforman a partir de la evidencia sobre las consecuencias reales y las condiciones sociales cambiantes. Entonces la validez es una propiedad en evolución, y la validación un proceso continuo.

IMPACTO DE LAS EVALUACIONES EN GENERAL Y EN GRAN ESCALA

En relación con la evaluación del aprendizaje de los alumnos y sus consecuencias, en especial la que hacen los profesores en el aula, Richard Stiggins señala que hasta hace poco se consideraba normal que sólo una parte de los alumnos alcanzara los aprendizajes esperados, mientras un número importante no lo conseguía, y que el papel de la evaluación consistía en distinguir a unos y otros en forma consistente. Por ello los criterios para valorar la calidad de las evaluaciones eran la validez y la confiabilidad. Hoy, en cambio, se espera que las escuelas hagan que todos los alumnos alcancen altos niveles de competencia para vivir en la sociedad del conocimiento, lo que lleva a reflexionar sobre el papel y las formas apropiadas de evaluar el aprendizaje en este nuevo contexto. Stiggins dice al respecto:

> *Las evaluaciones más válidas y confiables del mundo que tengan como efecto hacer que los alumnos abandonen la tarea desesperanzados no pueden ser consideradas productivas, porque hacen más daño que bien... Los marcos de referencia para el control de la calidad de las evaluaciones no tomaban en cuenta su impacto en el alumno; la nueva visión de la excelencia en lo relativo a evaluación, en cambio, pone en el centro de la escena este criterio de calidad*. (2008: 2 y 3)

El papel que las evaluaciones en gran escala han adquirido en muchos sistemas educativos tiene implicaciones importantes en cuanto a la validez de consecuencias, análogas a las que Stiggins señala para la evaluación en aula. En sus inicios las pruebas en gran escala eran de bajo impacto; sus resultados no influían en las decisiones respecto a cada alumno, ni a

maestros o escuelas individuales. En Estados Unidos esto comenzó a cambiar en la década de 1980, y la tendencia se acentuó en la de 1990, para culminar en las disposiciones de la ley NCLB, con la que las pruebas a gran escala adquirieron un peso sin precedentes en las decisiones relativas a alumnos, maestros o escuelas individualmente.

La aplicación de la ley NCLB evidenció deficiencias y consecuencias contraproducentes, tanto para los maestros al asociarse decisiones fuertes para ellos con base en los resultados de sus alumnos, incluso con Modelos de Valor Agregado, como para las escuelas que no pudieron cumplir las metas de la ley sobre Avance Anual Adecuado (*Adequate Yearly Progress*), y debieron enfrentar consecuencias que podían llegar hasta su cierre. Las metas nada realistas de la ley NCLB promovieron prácticas fraudulentas por parte de maestros en lo individual, pero también de escuelas e incluso distritos y estados completos (Oakes & Lipton, 2007), lo que justificó los cambios de la ley ESSA.

En muchos otros países ocurrió algo similar. La aplicación masiva de pruebas, y el que sus resultados se difundan con ordenamientos (*rankings* o *league tables*) de escuelas basados en los puntajes de los alumnos, sin tener en cuenta el contexto, vuelve de alto impacto los resultados. Ese impacto se presenta por el hecho mismo de la difusión de los *rankings*, aun si no hay disposiciones legales que impliquen consecuencias para las escuelas según su lugar en la lista. Si hay tales disposiciones la situación es más grave, pero incluso sin ello el impacto en escuelas, maestros y, en última instancia, los estudiantes, puede ser fuerte, ya que las pruebas tienden a corromperse al aparecer prácticas negativas, como la preparación de alumnos para la prueba, la subordinación del currículo a las evaluaciones, o la alteración de resultados mediante estrategias más abiertamente deshonestas.

Las personas con débil formación en estos temas ignoran que la precisión de los resultados de toda prueba es limitada, por lo que los ordenamientos basados en ellos son engañosos. Los resultados de un sistema educativo son más precisos y estables, pero los de escuelas individuales pueden variar mucho de una aplicación a otra. Por otra parte, los resultados de las pruebas dependen sólo en parte de escuelas y maestros, en tanto que las condiciones de la familia y el entorno tienen un peso importante. Las metodologías que buscan controlar el efecto de esos factores, como los *Modelos de Valor Agregado*, tampoco tienen la precisión y confiabilidad necesarias para sustentar ordenamientos consistentes en el tiempo.

Algunos sectores de la opinión pública son partidarios de medidas fuertes basadas en los resultados de pruebas censales, partiendo de la idea de que su aplicación traerá mejoras educativas rápidas gracias a los *rankings* que, al difundirse, producen una competencia que presiona a escuelas y maestros, haciendo que se esfuercen más, lo que hará mejorar los resultados. El supuesto implícito del argumento es que mejorar el rendimiento de cualquier alumno, incluyendo el de los más pobres, es algo sencillo, que no se consigue simplemente por la negligencia de los maestros o la falta de esfuerzo de los alumnos, lo que se corregiría fácilmente con las medidas mencionadas:

> Los sistemas de rendición de cuentas basados en pruebas se basan en la creencia de que la educación pública puede mejorar gracias a una estrategia sencilla: haga que todos los alumnos presenten pruebas estandarizadas de rendimiento, y asocie consecuencias fuertes a las pruebas, en la forma de premios cuando los resultados suben y sanciones cuando no ocurra así. (Hamilton, Stecher & Klein, 2002, p. iii)

Una consecuencia especialmente desafortunada de la tendencia a sobrevalorar las pruebas en gran escala, es que se tiende a perder de vista la importancia de las evaluaciones que lleva a cabo cotidianamente cada maestro en el aula, con sus alumnos. Por la escala misma en que deben aplicarse, y al menos en el estado actual de la tecnología, las

pruebas estandarizadas deben estar formadas por preguntas cuyas respuestas se pueden calificar de manera automática, computarizada. Por esto difícilmente sirven para medir los aspectos más complejos del currículo, e incluso áreas completas del mismo, como la expresión oral y la escrita, o los aspectos actitudinales y valorales. Los avances de los estudiantes en estos aspectos sólo pueden ser valorados con la precisión y la frecuencia necesarias para orientar las prácticas de enseñanza y las de aprendizaje gracias al trabajo del maestro en el aula, que por ello no pueden sustituir las pruebas en gran escala.

Posibilidades y limitaciones de PISA

PISA puede aportar elementos valiosos para mejorar los sistemas educativos, al permitir la comparación entre países (y regiones, en países con muestras subnacionales, como México y Brasil, además de España), posibilitando análisis de niveles de aprendizaje que tengan en cuenta factores de la escuela y el entorno, explorando los niveles de equidad social.

PISA permite distinguir dos tipos de problemas: los que se presentan cuando hay una alta proporción de alumnos de bajo rendimiento, los más pobres, y los que surgen si hay muy pocos alumnos en los niveles más altos. En el primer caso se trata del desafío de que todos los jóvenes tengan el nivel de competencias suficiente para una vida satisfactoria como trabajadores, pero también como ciudadanos. En el segundo caso se trata del reto de la formación de las futuras élites de ingenieros y científicos, y de quienes ocuparán los puestos directivos en los sectores empresarial y político. A partir de la aplicación 2009 de PISA, es posible saber qué son capaces o no de hacer los alumnos que se ubican abajo del Nivel 1 en las escalas definidas en PISA, gracias a la opción de aplicar el cuadernillo con ítems de baja dificultad, innovación promovida por los países del GIP. Los resultados cognitivos, en relación con los de los cuestionarios de contexto y con las preguntas que exploran actitudes y otros aspectos no cognitivos, pueden aportar elementos útiles para estudiar los factores que influyen en los niveles de las competencias estudiadas.

Por otra parte, por su diseño matricial y su carácter muestral PISA no puede dar resultados por escuela y menos por alumnos, sino sólo sobre el sistema educativo en conjunto, por lo que sus resultados son relevantes de manera directa para los responsables de las políticas públicas, y sólo de manera indirecta para directores de escuela y maestros, ya que no puede ofrecer información suficiente para orientar las prácticas de enseñanza y las de aprendizaje, pero además PISA es insuficiente por si sola como base para el diseño de políticas, ya que es necesario integrar muchos otros elementos del contexto, y muchos otros indicadores de aspectos educativos no derivados de pruebas, para tener una visión suficiente para sustentar la toma de decisiones. La obra de Axel Rivas (2015) es un excelente ejemplo al respecto.

CONCLUSIÓN

Una consecuencia positiva de PISA y las pruebas en gran escala en general, han sido que han llamado la atención de la sociedad sobre la importancia de hacer esfuerzos para mejorar los niveles de aprendizaje que se alcanzan en el sistema educativo. Otro impacto positivo ha sido la consolidación de instancias nacionales de evaluación, cuyo panorama es hoy muy distinto al que había hace 15 años. La capacidad técnica ha avanzado mucho en Chile, en el área especializada del estado y en el Centro de Medición de la Universidad Católica; en México con el Instituto Nacional para la Evaluación de la Educación (INEE), fundado en 2002 y que adquirió plena autonomía en 2013; en Uruguay, con la Unidad de Medición de Resultados Educativos y desde 2013 con su propio Instituto Nacional para la Evaluación de la Educación (INEEd); en Brasil y Colombia con la consolidación de instancias creadas años antes, el Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira (INEP) y el Instituto Colombiano de Fomento de la Educación

Superior (ICFES), que desde 2010 se transformó en el Instituto de Evaluación. La participación en PISA fue uno de los factores que contribuyó a esa consolidación.

En la ronda del año 2000 la participación de México y los otros países de Iberoamérica se limitó a lo básico: traducir los reactivos enviadas por el consorcio a cargo de las pruebas; pilotearlos y hacer la aplicación con la muestra mínima; calificar las respuestas abiertas, enviar los resultados y esperar el análisis internacional, sin participar en la planeación de las pruebas ni en el diseño de reactivos, ni hacer análisis propios de los resultados. En la ronda 2003 México utilizó una muestra ampliada para tener resultados por entidad federativa, y elaboró un informe nacional propio, que se difundió al mismo tiempo que el informe internacional (Vidal & Díaz, 2004), lo que supuso la participación del equipo del INEE a cargo de PISA en México en talleres sobre Teoría de Respuesta al Ítem y Modelos Jerárquicos Lineales, entre otros. Chile y Uruguay hicieron informes nacionales, y el segundo de estos países difundió boletines que analizan puntos particulares.

Desde el inicio de la preparación de la ronda 2006, México se propuso incrementar el nivel de su participación en PISA, por dos razones: por una parte, la consideración de que en evaluaciones en que participan numerosos países, de lenguas, culturas y niveles de desarrollo diversos, el riesgo de sesgo cultural es alto, y que para reducirlo importa que los países no se limiten a aplicar instrumentos hechos por otros, sino que se involucren activamente en su diseño y en el análisis de los resultados; y porque participar en PISA, al lado de países con mayor tradición en psicometría, era una oportunidad de aprendizaje para equipos técnicos de países con menos experiencia. Poco después inició una etapa de intensa colaboración entre equipos técnicos a cargo de PISA en países de Iberoamérica, con la formación del Grupo Iberoamericano de PISA (GIP), liderado por México y España. (Martínez Rizo & Roca, 2009)

Inicialmente la colaboración consistió en compartir las traducciones de las versiones originales de los instrumentos de PISA y los manuales para su aplicación, del inglés y el francés, al español y al portugués. Tras la aplicación de 2006, los responsables nacionales se apoyaron para la codificación de respuestas a las preguntas abiertas y la depuración de la información. El intercambio de experiencias entre países del GIP fue útil también para la preparación de informes nacionales de PISA 2006. Se llevaron a cabo encuentros y talleres formativos con apoyo del secretariado de la OCDE.

La colaboración incluyó la preparación de unidades de ítems de lectura para PISA 2009, comenzando con un taller de capacitación impartido por expertos del consorcio a cargo del desarrollo de PISA, y con un intenso intercambio durante meses, en los que los responsables de cada país del GIP intercambiaron las unidades desarrolladas en cada uno, antes de enviarlas al consorcio internacional, lo que llevó a que las pruebas de 2009 tuvieron unidades de ítems desarrolladas en Iberoamérica. El GIP influyó también para que se ofreciera la opción de aplicar unidades de baja dificultad que, sin disminuir el nivel de las pruebas ni imposibilitar la comparación con los resultados anteriores, permitan describir con mayor precisión que en el pasado las competencias de los jóvenes que no alcanzan el nivel más bajo medido por los instrumentos desarrollados hasta ahora. Un aspecto sobresaliente de la colaboración fue la preparación de un informe sobre los resultados de PISA 2006 en los ocho países iberoamericanos que participaron en esa ronda, así como en las 10 comunidades autónomas de España y los estados federales de Brasil y México. (Martínez Rizo & Roca, 2009)

Por otra parte PISA y, en general, las pruebas en gran escala, han tenido también consecuencias negativas. En México, destacan las de dos pruebas en gran escala. El caso de las pruebas nacionales denominadas ENLACE es el más claro: su aplicación masiva anual a todos los alumnos de los cuatro últimos grados

de educación primaria y tres de secundaria (4° a 9° de la clasificación internacional CINE), y su asociación con estímulos económicos importantes para los maestros, y ventajas para las escuelas, hizo que proliferaran prácticas corruptas, como preparación de los alumnos para las pruebas, estrechamiento curricular, y falsificación de los resultados. Esto llevó a la cancelación de las pruebas a partir de 2014.

En el caso de PISA, y pese a que su aplicación muestral y su diseño matricial impedían dar resultados por escuela y alumno, la visibilidad internacional de las pruebas y el peso de la OCDE hicieron que obtener resultados más altos en las siguientes aplicaciones se volviera la meta principal del Programa del Sector Educativo del sexenio gubernamental 2007-2012, con lo que esto significa como distorsión y estrechamiento de las políticas educativas.

La excesiva atención prestada a estas dos pruebas trajo consigo consecuencias que se han sintetizado en los siguientes puntos (Martínez Rizo, 2010b):

- *Banalización del debate público sobre la calidad educativa, reducido a discusiones superficiales de los rankings, perdiendo de vista la complejidad del tema.*
- *Mercadotecnia engañosa de las escuelas, sobre todo de sostenimiento privado, para atraer alumnos basadas en esos ordenamientos.*
- *Empobrecimiento del currículo, por la tendencia a enseñar para la pruebas, descuidando aspectos que no serán evaluados, aunque sean importantes.*
- *Cansancio y desaliento en escuelas que, pese a su esfuerzo, no consiguen resultados comparables con planteles de condiciones más favorables, y actitud negativa de los alumnos frente a una educación centrada en prepararlos para las pruebas.*
- *Empobrecimiento de las políticas públicas, que tienden a buscar soluciones fáciles a problemas complejos, descuidando aspectos fundamentales, como la equidad.*

La situación de los países iberoamericanos en cuanto al uso de pruebas en gran escala, y en particular en cuanto a PISA, ha sido similar a la de México. Hace 15 años los países de la región comenzaban a incursionar en el campo pero, con excepción del SIMCE de Chile, el impacto de las pruebas era bajo e incluso nulo, por la escasa difusión de los resultados (Martínez Rizo & Roca, 2009). En la segunda década del siglo XXI, además de Chile, se aplicaban pruebas censales en Uruguay, México, Brasil, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Perú y República Dominicana, y su impacto es cada vez mayor, de manera similar a la descrita antes para el caso mexicano. Hace 15 años los resultados de las pocas evaluaciones que se hacían tenían poca importancia. Hoy las evaluaciones proliferan, atraen fuertemente la atención y se han vuelto referente importante de la política educativa, pero las acecha el riesgo de prácticas derivadas de una comprensión inadecuada de sus alcances y límites, que lleva a esperar resultados casi milagrosos para las escuelas por la sola aplicación de pruebas, sin reconocer sus verdaderos alcances.

La combinación de consecuencias positivas y negativas de las pruebas en gran escala hace que la noción de validez de consecuencias cobre especial relevancia. En un reciente estudio sobre las principales pruebas en gran escala aplicadas en México (Martínez Rizo, 2015) la noción se concretó precisando puntos que las instancias a cargo del desarrollo y/o la utilización de resultados deberán atender satisfactoriamente:

- Informar a los usuarios sobre propósitos y características de la prueba, precisando lo que puede o no medir y los usos y consecuencias previstas, con argumentos teóricos y evidencia empírica que respalden unos y otras, advirtiendo sobre usos para los que no haya suficiente evidencia de validez, tratando de identificar los más probables
- Reportar resultados en plazos razonables a las partes interesadas utilizando lenguaje

claro y preciso, sin jerga técnica innecesaria, con información para minimizar la posibilidad de interpretaciones incorrectas y usando categorías que no estigmaticen.

- Ofrecer el marco normativo para evaluar el desempeño de los examinados; describir el perfil y las características de la población de referencia.
- Apoyar a instituciones y usuarios para desarrollar la capacidad necesaria para la adecuada interpretación y utilización de los resultados.
- Documentar y valorar el grado en que se producen las consecuencias previstas y/o deseables de la prueba, así como la existencia de usos o consecuencias imprevistas, sean adecuadas o inadecuadas; si hay evidencia de usos inapropiados se investigan, y si persisten se informa a los usuarios y se intenta aplicar medidas correctivas

¨ Aunque las pruebas PISA tienen alta calidad psicométrica, y su documentación técnica señala, en general, los alcances y límites de los resultados, parece posible sostener que no se ha cuidado suficientemente lo relativo a la validez de consecuencias, que precisan los cinco puntos que se acaban de citar.

Puede argumentarse que, más que a las instancias técnicas a cargo del desarrollo de las pruebas y del análisis de los resultados a nivel internacional, dicho cuidado correspondería más bien a las instancias técnicas nacionales, así como a los organismos del estado que tienen la responsabilidad del sistema educativo. Aceptando este argumento, la todavía no plena consolidación de las áreas especializadas en muchos países de desarrollo medio o bajo, junto a la gran carga de trabajo que muchas de esas instancias tienen, explica las deficiencias en el cuidado de la validez de consecuencias a la que se alude.

Por el peso ya mencionado de PISA en las políticas educativas de muchos países, la OCDE debe atender con mayor cuidado estos aspectos. En los países con mayor capacidad técnica instalada puede ser menos importante, pero lo es mucho más en países que, precisamente por su menor desarrollo, tienen menos elementos para prevenir o corregir usos inadecuados, y en los que el riesgo de consecuencias negativas es mayor.

Para maximizar el potencial positivo y minimizar el negativo de las poderosas herramientas que son las pruebas estandarizadas es indispensable que se consoliden las competencias técnicas de las instancias especializadas, y muy especialmente que se atienda lo relativo a la validez de consecuencias, involucrando tanto a las instancias a cargo de los sistemas educativos, como a la sociedad civil. Los organismos internacionales que manejan pruebas en gran escala podrían contribuir más a tal consolidación.

Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, Authors.

Baker, E. & Choppin, B. (1990). Minimum competency testing. En Walberg, H y Haertel, H. (Eds.). *The International Encyclopedia of Educational Evaluation*. Nueva York: Pergamon Press, pp. 499-502.

Basterra, M. Rosario, E. Trumbull & G. Solano, eds. (2011). *Cultural validity in assessment: Addressing linguistic & cultural diversity*. New York: Routledge.

Borsboom, D., G. J. Mellenbergh & J. van Heerden (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071. DOI: http://dx.doi.org/10.1037/0033-295X.111.4.1061

Cronbach, L. J. (1988). Five perspectives on validity argument. En Wainer, H & Braun, H (Eds.), *Test validity* (pp. 3–17). Princeton: IEA.

De Landsheere, G. (1994). *Le pilotage des systèmes d'éducation*. Bruselas: De Boeck.

De Landsheere, G. (1996). *La investigación educativa en el mundo*. México: Fondo de Cultura Económica [Edición original en francés de 1986].

Donlon, T. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. Nueva York: College Entrance Examination Board.

Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges*. Washington: PREAL.

Grupo de Trabajo Sobre Estándares y Evaluación. (2007-2008). Evaluaciones nacionales. En *Observatorio regional de políticas de evaluación educativa*. Santiago, PREAL.

Hamilton, L., Stecher, B. & Klein, S. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Mónica: RAND.

Husén, T. & Postlethwaite, T. S. N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy & Practice, 3*(2), 129-141. DOI: http://dx.doi.org/10.1080/0969594960030202

Joncich-Clifford, G. (1968). *The Sane Positivist: A Biography of Edward L. Thorndike*. Middletown: Wesleyan University Press.

Kane, M. (2006). Validation. En R. Brennan (ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport: American Council on Education & Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1): 1–73. DOI: http://dx.doi.org/10.1111/jedm.12000

Martínez-Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Martínez-Rizo, F. (2008). Las evaluaciones educativas en América Latina. En Instituto Nacional para la Evaluación de la Educación (Coord.), *Cuadernos de investigación*. México: INEE.

Martínez-Rizo, F. (2010a). Assessment practice in policy context: Latin American countries. En P.Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 479-485). Nueva York: Elsevier-Academic Press.

Martínez-Rizo, F. (2010b). Usos y abusos de la evaluación. *Este País*, agosto (232), 24-27.

Martínez Rizo, F. (Coord.). (2015). *Las pruebas ENLACE y Excale. Un estudio de validación*. México, INEE.

Martínez Rizo, F. & Roca, E., (Coords.) (2009). *Iberoamérica en PISA 2006*. Madrid. Santillana.

Mathison, S. & Ross, E. (2008). *The Nature and Limits of Standards-Based Reform and Assessment*. Nueva York: Teachers College Press.

Messick, S. (1989). Validity. En R. L. Linn, ed. *Educational Measurement* (3rd ed., pp. 13-103). New York, American Council on Education & Macmillan.

Michell, Joel (2000). Normal Science, Pathological Science and Psychometrics. *Theory & Psychology, 10*(5): 639-667. DOI: http://dx.doi.org/10.1177/0959354300105004

Moss, P. (2008). A critical review of the validity research agenda of the NBPTS at the end of its first decade. En L. Ingvarson y J. Hattie (Eds.), *Assessing teachers for professional certification: the first decade of the National Board for Professional Teaching Standards* (pp. 257–312). Oxford, Elsevier.

Newton, Paul E. (2013). *Does it matter what 'validity' means?* Presentación en el Departamento de Educación de la Universidad de Oxford, febrero 4.

Postlethwaite, T. S. N. (1985). International Association for the Evaluation of Educational Achievement. En T. Husén & T. S. N.

Postlethwaite (Eds.). *The International Encyclopedia of Education*. Nueva York: Elsevier, pp. 2645-2646.

Rivas, Axel (2015). *América Latina después de PISA: Lecciones aprendidas de la educación en siete países (2000-2015)*. Buenos Aires: Fundación CIPPEC.

Ross, K. (1994). *The Establishment of a Southern Africa Consortium for the Monitoring of the Quality of Education*. Paris: IIEP.

Sireci, S. G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement*, *50*(1): 99–104. DOI: http://dx.doi.org/10.1111/jedm.12005

Southern Africa Consortium for Monitoring Educational Quality (1995). *Southern Africa Consortium for Monitoring Educational Quality*. París: IIPE.

Stiggins, R. (2008). *Assessment Manifesto: A Call for the Development of Balanced Assessment Systems*. Portland: ETS-ATI.

Vidal, R. & Díaz, M. A. (2004). *Resultados de las pruebas PISA 2000 y 2003 en México. Habilidades para la vida en estudiantes de 15 años*. México: INEE.

Walberg, H. (1990). National assessment of educational progress: retrospect and prospect. En H. Walberg & G. Haertel (Eds.), *The International Encyclopedia of Educational Evaluation*. (Pp. 435-440). Oxford, New York: Pergamon Press.

Wolff, L. (2004). Educational Assessments in Latin America: The State of the Art. *Applied Psychology: An International Review*, *53*(2), 192-214. DOI: http://dx.doi.org/10.1111/j.1464-0597.2004.00168.x

Autor | **To know more / Saber más**

Martínez-Rizo, Felipe (felipemartinez.rizo@gmail.com).

Profesor de la Universidad Autónoma de Aguascalientes (1974-2016), de la que fue Rector. Ha escrito 57 libros y más de 190 artículos o capítulos. Miembro del Sistema Nacional de Investigadores y la Academia Mexicana de Ciencias de México. Fundador y primer Director General del Instituto Nacional para la Evaluación de la Educación de México de 2002 a 2008. Es Doctor Honoris Causa por la Universidad de Valencia (España).

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]