ISBN: 978-607-02-2345-7

Universidad Nacional Autónoma de México

Instituto de Investigaciones
sobre la Universidad y la Educación

www.iisue.unam.mx/libros

Ángel Díaz-Barriga (2011)
"Teoría del test, nuevos desarrollos en las pruebas a gran
escala y la prueba PISA 2006"
en *La prueba PISA 2006. Un análisis de su visión
sobre la ciencia,*
Ángel Díaz-Barriga (coord.),
IISUE-UNAM, México, pp. 53-79.

TEORÍA DEL TEST, NUEVOS DESARROLLOS EN LAS PRUEBAS A GRAN ESCALA Y LA PRUEBA PISA 2006

*Ángel Díaz-Barriga**

PISA es una prueba a gran escala que tiene un empleo reciente a nivel mundial. No es la primera prueba internacional, puesto que esta clase de instrumentos se emplean desde la década de los cincuenta del siglo xx para determinar el grado en que los estudiantes de un país, región o localidad pueden mostrar ciertos aprendizajes. Sin embargo, la prueba PISA en este momento[1] es la más relevante en el contexto de la globalización, bajo los supuestos de una sociedad del conocimiento.

Las pruebas a gran escala reflejan una cosmovisión de la sociedad contemporánea signada por la competitividad y por la competencia, tanto entre países como entre personas. Al mismo tiempo que reflejan las aspiraciones de lo que se puede considerar el mundo de

* Investigador emérito del IISUE-UNAM.

1 Desde mediados de los años cincuenta del siglo XX la International Association for Evaluation of Educational Achievement empezó a desarrollar a nivel mundial pruebas a gran escala. La más relevante ha sido la prueba TIMSS (Third International Mathematics and Sciencie Studies) que se aplica en México desde la década de los noventa. En esa década el Laboratorio de Medición de la Calidad de la Educación de la OREALC-UNESCO diseñó otro conjunto de pruebas para estudiantes de algunos grados de primaria en América Latina. Sin embargo, la prueba PISA que se aplica desde el año 2000 en los países miembros de la OCDE y en aquellos que aceptan participar en la prueba, se ha distinguido por el grado de su difusión internacional. En otros trabajos la hemos calificado como una prueba a gran escala de tercera generación. A. Díaz Barriga, *Las pruebas masivas. Análisis de sus diferencias técnicas.*

la información, en una aldea global, donde el conocimiento traspasa cualquier frontera; reflejan de igual manera, el desconocimiento de una serie de características singulares, de procesos sociales, económicos y culturales altamente diferenciados entre naciones, regiones y grupos sociales que precisamente desconocen los teóricos de la visión económica global. Las pruebas a gran escala parten del presupuesto de que, con independencia de las condiciones socioeconómicas y sobre todo culturales que signan a cada uno de los países del mundo, las escuelas pueden funcionar de manera homogénea, los alumnos pueden aprender no sólo los mismos conocimientos de diversas disciplinas, sino que pueden entender estos conocimientos con los mismos conceptos, los mismos lenguajes y la misma extensión. La escuela es el lugar de homogeneización de la sociedad del conocimiento.

Las pruebas a gran escala son, en estricto sentido, instrumentos de medición. Esto es, instrumentos que se basan en los desarrollos de la teoría de la medición formulada en el siglo XX, en particular en el campo de las ciencias del comportamiento. Si bien estamos lejos de las formulaciones de Binet-Simon que dieron pie a la construcción de los primeros test de inteligencia, la realidad es que toda prueba está basada en la teoría del test y, por tanto, debe cumplir con una serie de postulados que emanan de ella, para garantizar, sólo desde el punto de vista de la medición, que realmente miden lo que pretenden medir; esto es, que son instrumentos confiables y válidos para determinar una medida, en este caso de aprendizaje, así como una báscula mide el peso de una persona en todo el mundo o un metro es usado para determinar su altura. Aunque este ejemplo es perfectamente imperfecto, porque el sistema de medición de peso a nivel mundial no es homogéneo, ya que al menos en los diversos países se reconocen dos sistemas conformados para ello.

El objetivo de este capítulo es examinar, desde la perspectiva de la teoría de la medición, hasta dónde la prueba PISA cumple con los postulados que emanan de la teoría del test, lo que permitirá estudiar la validez que entraña la prueba en sí misma; es decir, realizar un estudio por medio del cual se determine si esta prueba reúne los requisitos técnicos que se exigen a este tipo de instrumentos, y anali-

ce si los nuevos desarrollos en las pruebas a gran escala se apartan de la teoría del test; en suma, encaminarse hacia una redefinición de los criterios de validez y confiabilidad. Esta redefinición no está exenta de enfrentar nuevos riesgos sobre el uso de ambos criterios, así como sobre la validez de constructo que se ha constituido en los últimos 20 años como el criterio de referencia para valorar tales instrumentos.

El capítulo explora la validez de constructo que subyace en la prueba PISA, tomando como referencia la aplicada en 2006, desde el punto de vista técnico de la teoría del test y de los desarrollos que recientemente se han efectuado en torno a esta teoría. Sin embargo, las preguntas centrales siguen siendo las mismas que orientaron la conformación de la teoría del test. Nos preguntamos, por ejemplo, qué es lo que miden estas pruebas, o sea, cuáles son los aprendizajes que son medidos en estos instrumentos; por otro lado, cuáles son los valores estandarizados de tales mediciones, esto es, hasta dónde se puede afirmar que son instrumentos confiables para medir aprendizajes significativos, pertinentes, derivados de los procesos escolares o, por el contrario, se deben a diversos sesgos culturales, de dificultades de calibración del ítem (problemas técnicos con el contenido, con el vocabulario, con la traducción e incluso, como lo han mostrado diversas investigaciones, con las distintas idiosincrasias que subyacen en las prácticas pedagógicas en distintos países). En este sentido, cabe preguntar también cuál es el valor de un puntaje en una prueba a gran escala, esto es, estudiar si realmente se puede hacer la comparación que los expertos de PISA asumen entre estudiantes de diversos países, o hasta dónde, como resultado de la globalización económica,[2] PISA refleja una aspiración idealista (pero dominante) de lograr la homogeneidad, en un momento en que las ciencias sociales realizan un intenso debate sobre lo diverso, lo singular, lo heterogéneo. Hoy la meta es que los estudiantes de todo el orbe aprendan temas, formas de razonamiento, formas de responder

2 Reconocemos que existen múltiples formas de globalización. En el marco de las ideas, el cristianismo, el socialismo, la democracia son algunas expresiones de ellas; mientras que en el ámbito de la ciencia el positivimo, la teoría de la ciencia constituyen otra de sus manifestaciones (véase A. Mattelard, *Historia de una utopía planetaria).*

ante un interrogante como lo hacen los estudiantes en Finlandia, sin necesariamente analizar las múltiples diferencias que hay entre las culturas de aquél y el resto de los países, sus procesos demográficos, sus proyectos de formación docente, sus sistemas educativos, entre otros factores.[3]

De la teoría del test a las pruebas a gran escala

En los primeros cincuenta años del siglo xx se realizaron desarrollos en el campo de la psicología que tenían como intención fundamental lograr una medición precisa de diversos componentes del comportamiento humano. La psicología científica lograría este estatus en tanto tradujera las cualidades o atributos humanos (inteligencia, interés, actitud, capacidad, aprendizaje) a un atributo capaz de ser medible. Las ciencias del comportamiento en esta orientación se vieron obligadas a acudir al modelo que la física newtoniana había venido construyendo en los dos siglos anteriores. La meta de desarrollar cada vez más y mejores mediciones o comparaciones cuantitativas estaba en su apogeo, y así se fue conformando, a partir de los trabajos de Simon y Binet, la era inicial de la formulación de la teoría de los test, instrumentos de medición de una cualidad específica (inteligencia, aptitud, aprendizaje) que rápidamente se fueron incorporando al sistema educativo estadounidense. Los test de aprendizaje se consideraban una alternativa frente a las valoraciones subjetivas y sin posibilidad de ser replicadas por los docentes. En la calificación de un test no interfieren elementos de apreciación, aspectos subjetivos, ni el efecto de "halo"[4] que se documentaría en las décadas

3 United Nations Educational, Scientific and Cultural Organization, *Docentes como base de un buen sistema educativo. Descripción de la formación y carrera docente en Finlandia 2007*.

4 Así fue llamada la reflexión (acompañada seguramente por un componente de emoción) del docente que se caracteriza por intentar ser más benévolo al calificar a un estudiante cuando reconoce que ha sido un buen estudiante durante el curso, o bien, el extrañamiento que puede tener al ver un examen con resultados no tan exitosos como se deseara e identificar el nombre del alumno, cuando el docente espera mejores resultados de éste. A partir de ello se aconsejaba a los profesores no ver el nombre del alumno en el momento de calificar una prueba.

de los cincuenta y sesenta. El test aparece como un instrumento de medición, neutro, objetivo y fundamentalmente científico. Según Cremin,[5] en el sistema educativo estadounidense los test se empezaron a emplear desde la década de los veinte del siglo pasado, mientras que se tiene evidencia de que en México fue la Normal Superior la que los empezó a utilizar desde 1927; específicamente, la teoría del test fue una de las que de manera más temprana se incorporó al sistema educativo mexicano, como se puede identificar en los planes de estudio de las escuelas normales hasta antes de que desapareciera en la reforma de 1997.

Los principios. La teoría clásica

Para que los test se puedan considerar instrumentos de medición deben cumplir con una serie de normas y requisitos en su proceso de formulación. Toda prueba debe reunir una serie de requisitos psicométricos y cumplir con una serie de normas que le den validez y confiabilidad, esto es, que midan lo que deben medir y que sus resultados no estén alterados por cualquier otro elemento que haga dudar del dato que se obtiene de ellos. En su primera etapa,[6] la teoría del test dio origen a lo que posteriormente se denominó teoría clásica del test, teoría que contenía una definición muy clara de las múltiples medidas psicométricas que el instrumento debía poseer. En conjunto, todos estos requisitos llevan a la tipificación de una prueba.

Se trata de un proceso lento y laborioso que exige determinar el atributo (aprendizaje de ciencias) y establecer las operaciones a partir de las cuales se reconoce el atributo (contenidos específicos), lo que significa determinar desde qué lugar se fija cada uno de los contenidos que se van a explorar; contenidos que pueden ir desde un plan o programa de estudios hasta un libro de texto, o la opinión de un grupo de especialistas, y generar las preguntas para efectuar la medición. Estas preguntas requerían ser estandarizadas (hoy se

5 L. Cremin, *La transformación de la escuela. El movimiento escuela progresiva en los Estados Unidos.*

6 Véase R. Tyler, "Examen y valoración del conocimiento, destreza y capacidad adquiridos".

dice calibradas). La estandarización de una prueba se puede hacer de una manera imperfecta o estadística. La imperfecta consiste en someter cada reactivo a un juicio de expertos, quienes discuten si esa pregunta está bien redactada, si el contenido que presenta es relevante, si puede mejorarse y, en su caso, redactarla de una mejor manera o bien rechazarla. El juicio de expertos es imperfecto porque finalmente los juicios que se formulan sobre cada reactivo están condicionados por la percepción de estos especialistas.

Por su parte, la estandarización estadística reclama determinar una serie de medidas sobre cada uno de los reactivos, tales como su índice de dificultad o su poder de discriminación. El índice de dificultad es un valor proporcionalmente inverso al porcentaje de estudiantes que obtuvieron la respuesta correcta en cada reactivo. De tal suerte que un reactivo que responden favorablemente 80 estudiantes de un grupo de 100, significa que es muy fácil porque su índice de dificultad es de 20; mientras que un reactivo que responden satisfactoriamente sólo 30 tiene un nivel de dificultad mayor de 70. Esto significa que se espera que de un grupo de 10 alumnos sólo tres respondan bien el reactivo en la dificultad mayor, o bien que siete tengan el acierto en la dificultad menor. Los expertos en test, dado que en la teoría clásica del test el modelo de referencia está basado en la llamada curva de Gauss,[7] se daban a la tarea de elaborar un examen con un nivel promedio preestablecido de reactivos como una decisión basada en un arbitrio:[8] una prueba dura o difícil tiene un nivel de dificultad de 65 a 80, una mediana de 50 y una baja de alrededor de 35.

7 También la teoría clásica, que se basa en un modelo de estadística descriptiva, tenía formas de mover los intervalos de una medición a partir de la media y de la desviación estándar, con la finalidad de hacer más corta la distancia del conjunto de intervalos, o más extensa, según se quisiera que la curva cumpliese con un efecto modelístico (véase A. Díaz Barriga, "Tesis para una teoría de la evaluación y sus derivaciones en la docencia").

8 Este concepto lo construimos desde 1982. *Arbitrio* es una decisión que toma cualquier especialista en evaluación (aún hoy en la prueba PISA o ENLACE) para determinar cómo selecciona un reactivo, cómo integra el conjunto de reactivos, etc. No es una arbitrariedad; ésta tiene una relación estrecha con el capricho, pero un arbitrio es una decisión racional, fundamentada, que sin lugar a dudas afecta al instrumento, pero sin la cual no se puede formular el instrumento. También la evaluación que hace el docente en el salón de clases está sujeta a este tipo de decisiones (arbitrios) porque de otra forma no se podría trabajar. *Ibid.*

Por otra parte, se analizaba estadísticamente el nivel de discriminación que resultaba de la aplicación de cada pregunta o reactivo, lo que significa determinar la capacidad que tiene cada reactivo para diferenciar a los alumnos que saben, o que han aprendido, de los que no saben. El presupuesto de esta cuestión es que debe ser consistente la tendencia de los alumnos que tienen mayor rendimiento a contestar bien las preguntas, así como, los que tienen menor rendimiento, a responderlas con error. El modelo estadístico empleado permitía confrontar un porcentaje de estudiantes que hubieran resuelto el examen (27%) con mejores puntajes, frente a otro porcentaje (27%) que hubiera tenido los menores aciertos. En cada pregunta se determinaba el grado en que los estudiantes del primer grupo superaba a los del segundo grupo, lo que daba un factor mayor a 1 o menor a 1. Un resultado positivo indicaba un grado de discriminación satisfactorio, mientras que uno negativo determinaba que ese reactivo no discriminaba entre "buenos" y "malos" estudiantes y, por lo tanto, debía ser rechazado en el examen.[9]

Finalmente, se ponderaba el papel de los distractores en cada pregunta de opción múltiple (se llama distractor en una pregunta de opción múltiple a las respuestas que no se pueden considerar correctas). La lógica de estos instrumentos es que para cada cuestión hay una respuesta que es correcta o incluso que es más correcta que las otras. A los distractores se les exige que tengan una lógica que los haga aceptablemente correctos, y ello incluye aspectos de forma (cuidar singulares y plurales, en castellano atender al género de los artículos), así como de fondo (que sean posibles respuestas correctas). Cuando todas las respuestas de una preguntan se concentran en una o dos opciones, los otros distractores no están cumpliendo su papel, razón por la cual técnicamente se recomienda eliminar la

9 Varios autores hacen excelentes explicaciones sobre esta situación; en México, destaca la que elaboró F. García, *Paquete de autoenseñanza de evaluación del aprendizaje*. Otro estudio que tiene referencias muy importantes al respecto es el que formula Ch. Stage, "Teoría clásica de medición o teoría de respuesta al ítem. La experiencia sueca", p. 188. Una cuestión importante en el planteamiento de esta autora es que sostiene que en el fondo no existe una diferencia significativa en la construcción de pruebas apoyadas en la teoría clásica o en la teoría de respuesta al ítem.

pregunta o mejorar la calidad de los distractores (entiéndase por ello, hacer redacciones que tengan algún grado de factibilidad para ser correctas).

Esta última situación hace que el proceso lógico de pensamiento que se requiere para resolver con mejores probabilidades una prueba de este tipo esté basado en sistemas de recuerdo, razonamiento lógico y discriminación. Esto, aunado a otra serie de consejos que se derivan de los mismos criterios de construcción de una prueba, permitió formular recomendaciones muy generales, pero efectivas, para resolver esta clase de pruebas; tales recomendaciones son la base de muchos cursos de preparación de este tipo de exámenes. Ahora bien, quizá el primero en plantear los excesos que todo ello producía en el ámbito educativo fue Vanhecke, en un pequeño ensayo de mediados de los años setenta que lleva por título "Brasil. La computadora atonta la enseñanza",[10] en el cual recreaba una serie de estrategias, entre las que se pueden mencionar: primero resolver todas las preguntas de las que se está seguro, en una segunda revisión a la prueba buscar preguntas que se encuentran entre dos cuya respuesta fue dada como segura y eliminar los distractores que corresponden a dicha respuesta, por ejemplo "c", si en la respuesta anterior esa opción es la correcta y en la siguiente lo es la "a", entonces limitarse a analizar las opciones que quedan libres para la pregunta intermedia, esto es b y d, en cuatro opciones, o bien b, d y e, en cinco opciones; luego, empezar a descartar la menos lógica hasta concentrarse en una o dos, y en caso de no tener claridad sobre cuál es la respuesta apostar sencillamente a una de ellas, pues la probabilidad de acierto aumentó a un 33 o 50 por ciento. Es importante recordar estos señalamientos, porque ello apunta a los errores de medición que son muy difíciles de determinar en una prueba, precisamente porque hay pocos estudios sobre los procesos de pensamiento que exigen en su resolución. Algo sí podemos afirmar: los alumnos que obtienen más altos puntajes han desarrollado estrategias adecuadas para la resolución de este tipo de pruebas, lo que no necesariamente significa que tengan mejores aprendizajes.

10 En A. Díaz Barriga, *El examen: textos para su historia y debate*.

Hemos dado un breve bosquejo de la teoría clásica del test, señalando los criterios de medición que la caracterizan, así como las limitaciones que se fueron generando en su evolución. Sólo debemos señalar que en México la teoría clásica del test, en los pocos planes de estudio donde se enseña, sigue teniendo un valor predominante.[11] Del conjunto de críticas que se formularon señalando las limitaciones de este tipo de medición, y a partir del desarrollo de nuevos modelos matemáticos, así como del empleo de diversos programas de cómputo *(software)*, se avanzó en el establecimiento de nuevos desarrollos en la teoría del test.

La búsqueda de alternativas: la teoría de respuesta al ítem

Algunos autores[12] sostienen que el desarrollo de la teoría de la respuesta al ítem (TRI) se realiza en una etapa que tiene cierto paralelismo temporal con el desarrollo de la teoría clásica; en ciertos casos se argumenta que se trata de emplear distintos criterios estadísticos para la ponderación de los criterios de validez que postula la teoría clásica. Lo que es claramente cierto es que esta última se basa en un modelo de distribución de frecuencias que tiene como referencia la campana de Gauss, en la que se cuenta con algunos elementos para moderar la curva a partir del manejo de intervalos, utilizando como referencia la medidas de la estadística descriptiva (media, mediana, moda y desviación estándar). Por su parte, la teoría de la respuesta al ítem se apoya en un concepto denominado *criterial*, esto es, que busca dar cuenta del logro o no logro de una ejecución, y su modelo estadístico es más cercano a uno inferencial que se apoya en coefi-

11 En otro trabajo mostramos cómo la tradición mexicana de pruebas a gran escala, hasta el año 2005, se fundamentaba en la teoría de los test de los años cincuenta, teoría clásica, así como lo lejos que se encontraba de apoyarse en nuevos desarrollos en este campo. Realizamos en ese momento una comparación entre el EXANI I que aplica el CENEVAL (Centro Nacional de Evaluación) a aspirantes para realizar estudios de bachillerato (jóvenes de 15-16 años en promedio) y la prueba PISA 2000, que en ese momento había publicado 300 reactivos (véase A. Díaz Barriga, "Las pruebas masivas. Análisis de sus diferencias técnicas").

12 J. Muñiz, *Introducción a la teoría de respuesta a los ítems;* A. Tristán, *Análisis de Rasch para todos. Una guía simplificada para evaluadores educativos.*

cientes de regresión. Sin lugar a dudas, el desarrollo de diversos programas de cómputo, como instrumentos que realizan operaciones complejas, acortando significativamente el periodo de tiempo que demanda esta tarea, ha jugado un papel fundamental, tanto para desarrollar pruebas a gran escala para grandes cantidades de poblaciones, como para nacionalizar, regionalizar e internacionalizar dicho tipo de pruebas.

De alguna forma podemos afirmar que los elementos básicos para analizar el test se mantienen, puesto que en su sistema de calibración se conservan los indicadores centrales que permiten identificar el poder de discriminación, así como el índice de dificultad, como valores estadísticos centrales; aunque con algunas modificaciones en su concepción, puesto que en la TRI la aceptación de un reactivo para una prueba es resultado solamente de su índice de dificultad (sin tomar en consideración el nivel de discriminación).[13] Mientras que con respecto al índice de discriminación la TRI sostiene "que no hay bases para determinar a qué corresponde un grupo de máximo y mínimo rendimiento",[14] la teoría clásica resuelve tal aspecto a partir del puntaje máximo obtenido y la construcción de la campana de Gauss.

Sin embargo, la tarea es mucho más compleja de lo que parece inicialmente, dado que varias lógicas subyacen en el momento de calibrar los reactivos en la teoría de la respuesta al ítem. Una de ellas es apoyarse en los análisis de Rasch para construir el instrumento a partir de los diversos desempeños que tiene un grupo piloto en el que se aplica el instrumento. Si la base de la teoría clásica era la campana de Gauss, la base en la teoría de la respuesta al ítem es la denominada curva característica del reactivo (curva en forma de "s"), que "relaciona la medida de las personas y la probabilidad que tienen para responder un reactivo"[15] a partir de una pendiente, y sobre ambas unidades se establecen los valores esperados de respuesta a un examen.

13 M. A. Rosado, "Análisis de ítems. Teoría Clásica y Teoría de Respuesta al Ítem", p. 331.

14 *Ibid.*, p. 332.

15 A. Tristán, *op. cit.*, p. 25.

En la integración de los reactivos finales de la prueba se busca que ésta contenga reactivos que se comporten de diversa manera en la pendiente, buscando que vayan incrementando gradualmente su grado de dificultad. Esto no significa que se encuentren en este orden en el examen. El presupuesto central, en esta perspectiva, es que los sujetos que responden un examen tienen aprendizajes diferentes, unos más escasos o reducidos (la parte inicial de la pendiente y de la curvas que corresponde a valores negativos en el eje de las abscisas), y otros, los menos, mucho mayores (los puntos altos en el eje de las ordenadas y positivos en el de las abscisas). Cualquier reactivo o estudiante que tenga un desempeño distinto significa que hay una dificultad en la discriminación que presenta dicho examen.

El acceso a los sistemas de cómputo permite realizar una multitud de operaciones que de otra forma no sería factible; por ejemplo, integrar en una plantilla los resultados de los estudiantes en cada uno de los reactivos a partir de los alumnos que obtuvieron el más alto puntaje; en esa planilla electrónica se señalarán las respuestas correctas (1) e incorrectas (0) que tuvieron los alumnos a cada pregunta. Eso permitirá comparar el desempeño del estudiante contra sí mismo, pues empezarán a saltar aquellos casos donde un estudiante de bajo desempeño responda preguntas de mayor grado de dificultad, cuando su patrón de respuestas es que sólo responde las más fáciles, lo que los especialistas denominan el Escalograma de Guttman.[16]

Este instrumento es empleado en una lógica estadística para determinar los errores de acierto y los errores de falla a partir de clasificar las respuestas que presenta un estudiante como respuesta correcta inesperada o correcta esperada, así como respuestas incorrectas esperadas e incorrectas no esperadas. En el fondo, podemos afirmar que entre la teoría clásica y la teoría de la respuesta al ítem se conserva la misma lógica de construcción: una prueba no está elaborada para buscar que los estudiantes obtengan la máxima calificación, sino para mostrar las diferencias de aprendizaje que hay entre un grupo de estudiantes. Esto significa que una prueba a gran

16 A. Tristán, *op. cit.*, p. 85.

escala existe siempre y sólo puede existir, siempre y cuando logre mostrar que hay sujetos que destacan y hay otros que tienen deficiencias. La diferencia entre ambas teorías es que el TRI se apoya en modelos estadísticos de probabilidad que reclaman una mayor elaboración, lo que permite, a partir de instrumentos de cómputo, construir varias comparaciones de desempeño de un grupo frente a otros grupos, de todos los que integran el grupo o subgrupo, y de cada sujeto respecto de su patrón de respuestas. Para cada sujeto se fija el grado de respuestas esperadas y se prenden focos rojos cuando aparecen respuestas inesperadas. Máxime cuando éstas corresponden a preguntas que tienen mayor grado de dificultad.[17]

Estos procedimientos permiten calibrar un reactivo, esto es, medir el grado de desempeño que tiene cada pregunta de un test, estableciendo diversas valoraciones estadísticas. No es un procedimiento sencillo, ni mucho menos, que se pueda realizar en un corto plazo. Se trata de una cuestión que podemos suponer se realiza en la prueba PISA, pero sobre la cual prácticamente no se ofrece información. Todo lo cual nos puede llevar a suponer que la prueba está bien calibrada, o bien, que pese a la importancia que la política educativa le concede a los resultados PISA, este organismo no se considera obligado a dar cuenta de la parte técnica de su propio instrumento.

Otros elementos aparecen con mayor énfasis, o bien, con aspectos novedosos. El primero, el que ocupa un lugar central en el debate actual de la teoría del test, es la determinación de la validez de constructo; mientras que entre los segundos se puede considerar la formulación de la tabla de especificaciones, la conformación de un esquema de reactivos de "multi-ítems", en los que a partir de una información central se formula un esquema de 6 o 7 preguntas (lo que convierte la prueba en un instrumento que reclama una gran inversión de tiempo para su resolución por la cantidad de texto que

17 El tema ya lo habíamos examinado desde el ángulo de la teoría clásica del test, cuando afirmamos que estos instrumentos sólo pueden construirse pensando en que los estudiantes tienen que caer (equivocarse) forzosamente en alguno de los distractores (de otra forma, esos distractores no cumplen su función), y cuando enunciamos que son pruebas construidas para lograr que no todos los estudiantes muestren un dominio (véase A. Díaz Barriga, "Tesis para una teoría de la evaluación y sus derivaciones en la docencia").

hay que leer como base de cada grupo de preguntas), así como los estudios sobre las dificultades y tiempos que reclama la traducción de los reactivos de una prueba. Dichas dificultades en la traducción cobran relevancia en el caso de las llamadas pruebas internacionales, esto es, en aquel tipo de instrumentos que se aplica a estudiantes de diversos países.

Con estos aspectos se puede finalmente reconocer que el grado de validez y confiabilidad que tiene una prueba constituye un factor que determina con claridad hasta dónde se puede generalizar sus resultados en poblaciones que tienen diferencias culturales (según estudios de los psicólogos piagetianos, la palabra *tío* no tiene los mismos significados para un alumno suizo que para un asiático de origen musulmán), así como reconocer las mismas dificultades para traducir un vocablo que puede ser muy claro en la cultura de origen de una pregunta y un tanto oscuro en un país que tiene otras construcciones lingüísticas.

Todos estos aspectos constituyen una parte técnica fundamental de una prueba a gran escala; se trata del núcleo técnico de una prueba, que permite legitimar las inferencias que se realicen sobre ella. Reclaman un riguroso dominio técnico, por lo cual resulta lamentable que la mayoría de los informes internacionales (y claramente los nacionales) no hagan ninguna referencia a estos aspectos o se limiten a proponer a la sociedad que acepte ciegamente los trabajos técnicos que se hicieron al respecto. Lo que hemos afirmado reclama una exposición analítica un poco más detallada, con la finalidad de ponderar el valor de estos temas.

Antes de concluir este rubro conviene tener presente que, junto con la aplicación de un número mayor de exámenes a gran escala en el medio nacional y tomando como referencia los desarrollos actuales en la teoría del test, esas nuevas formulaciones están lejos de convertirse en contenidos curriculares para la formación de expertos en educación. Lamentablemente, el estudio de la teoría del test ha desaparecido de los planes de estudio de las escuelas normales,[18] no

18 Históricamente los planes de estudio de las escuelas normales tuvieron en el siglo XX una asignatura, "Psicotécnica Pedagógica", que es donde se analizaba tanto la teoría de la medición

existe tampoco como un contenido en los planes de estudio de las licenciaturas en pedagogía y/o ciencias de la educación (pese al papel no necesariamente positivo que ha jugado el Examen de Egreso de Licenciatura, EGEL), y sólo se aborda en algunos cursos de teoría de la medición en las licenciaturas en psicología.

La validez de constructo en la prueba PISA

Hemos afirmado que las pruebas a gran escala son instrumentos de medición y que como tales deben cumplir con determinados requisitos a fin de que los datos que se obtienen tengan un grado de certidumbre y confiabilidad. Para ello se requiere poder emitir un juicio sobre las cuestiones técnicas de una prueba, en particular sobre su validez de *constructo*, término aceptado ampliamente en la literatura especializada sobre el tema.

No se trata solamente de la aplicación de modelos matemáticos al análisis de funcionamiento de un reactivo, sino de un conjunto de decisiones técnicas que permiten realizar la tarea de calibrar cada una de las preguntas y calibrar el instrumento en su conjunto. Sobre esta cuestión se indaga muy poco en las pruebas a gran escala; en la mayoría de ellas las organizaciones que las elaboran no se consideran obligadas a dar cuenta de esta cuestión. Reconocemos que con muchas dificultades se encuentra una información sobre el caso de la prueba PISA.

---

como los criterios para la elaboración de pruebas llamadas objetivas. Por muchos años el libro del profesor Herrera y Montes, junto con el *Manual de psicotécnica*, de Manuel Villalpando, eran los libros de texto que se consideraban referencia obligada. En la reforma de 1984, la que concedió el título de licenciatura en educación primaria, esto se reemplazó por una asignatura denominada "Evaluación del aprendizaje", con un enfoque mucho más cercano al debate didáctico y más alejado de la teoría del test. Posteriormente, en la reforma de 1997, la que tuvo como bandera "abajo el teoricismo", sencillamente se eliminaron todos los contenidos de evaluación, así como los de didáctica, psicología educativa, evolutiva y del aprendizaje, teoría pedagógica e historia de la educación. La aberración consiste en que formemos un maestro empíricamente, en la práctica. Nadie asocia los resultados de los exámenes a gran escala al sistema de formación de maestros en México.

Los reportes que se editan para el público suelen sólo presentar generalidades referidas a estos problemas. Esto es, no mencionan con detalle cuáles son las especificaciones que exige el consorcio que maneja la prueba, no plantean cómo se construyen los reactivos ni mucho menos cómo se traducen y calibran para el caso mexicano.

Un ejemplo que se puede retomar de la industria automotriz se refiere a que las especificaciones técnicas de consumo de combustible y la consecuente emisión de contaminantes cambia de país a país, acorde no sólo con legislaciones particulares, sino también con las condiciones climáticas generales en las que se desempeñará dicho vehículo. Los laboratorios de ciencia saben que una clave fundamental en la realización de una investigación es la calibración de sus instrumentos, esto es, que los materiales de medición como pipetas, buretas y matraces estén aforados para lo que se desea medir, de tal manera que la medición sea confiable y reproducible, o bien, que los equipos automatizados de medición estén calibrados de acuerdo con las condiciones climatológicas y de laboratorio, con la finalidad de que la condición de confiabilidad refleje lo que se está midiendo; un ejemplo de esto último es el equipo para medir pH —potenciómetro— cuya calibración requiere una temperatura estándar y una solución estándar de acidez para que posteriormente se realicen las mediciones de diferentes soluciones que reflejen las propiedades o composición de las soluciones problema; en cada caso, en cada momento, este tema salta en la investigación científica.

Al parecer, la prueba PISA es un examen mundial en el que estos "pequeños detalles" no importan mucho: los alumnos de todo el mundo pueden responder al mismo instrumento, sin considerar si hay diferencia entre los planes de estudio de todos los países, en los libros de texto que se utilizan, en las formas de enseñanza o en las concepciones explícitas o implícitas de aprendizaje que reflejan planes y programas o que emplea cada docente. Pareciera que en la era mundial se puede medir de la misma forma a todos los estudiantes.

En ocasiones se llega a conclusiones apresuradas por parte de los responsables nacionales cuando afirman que los alumnos mexicanos tuvieron en su escolaridad acceso tanto a los contenidos como a las formas de aprendizaje que demanda la resolución del examen.

Así, se expresa:

> Desde el punto de vista curricular, podría afirmarse que los estudiantes cuentan con los conocimientos y habilidades necesarias para demostrar un buen desempeño en la evaluación de PISA [...] Aunque los objetivos curriculares están planteados en términos de conocimiento, se podría decir que el logro de éstos permitiría a los estudiantes desarrollar las competencias evaluadas por PISA.[19]

Si bien esta afirmación se basa en un cuadro de página y media en el que dichos responsables comparan competencias PISA y Objetivos Curriculares del Plan de Estudios de 1993 de la escuela secundaria en México, ello es insuficiente para dar cuenta con rigor académico de la afirmación que realizan. En el fondo, contribuyen a impulsar la cuestionable perspectiva en el debate de las competencias de que un comportamiento conductual es equiparable a aquéllas.[20]

Esta afirmación es contradictoria a las recomendaciones 2 y 3 que los expertos de la OCDE formulan al gobierno mexicano cuando analizan los resultados de la prueba PISA, y sugieren:

> Establecer con total claridad los estándares de rendimiento esperados para los estudiantes de los diferentes niveles del sistema, en áreas clave (como el alfabetismo, nociones elementales de cálculo aritmético y tecnología de la información), así como alinear el plan de estudios a estas áreas clave y elaborar materiales educativos de alta calidad para apoyar el trabajo de profesores.[21]

19 M. A. Díaz Gutiérrez *et al.*, *PISA 2006 en México*, pp. 53 y 55. En la página 54 hacen un comparativo entre competencias PISA y Objetivos Curriculares del Plan de Estudios de Secundaria mexicano.

20 Véase P. Perrenoud, *Construir competencias desde la escuela*.

21 "Establish absolute clarity about the standards expected in key areas (such as literacy, numeracy and information techonology) requiered for students at various levels in the system and align the curriculum these key areas and produce high quality and practical materials to support the work of teachers". Éstas son parte de las 12 recomendaciones que los expertos de Londres, a nombre del consorcio, formulan para México, a partir de los resultados obtenidos en la prueba PISA 2006. D. Hopkins *et al.*, *An analysis of the mexican school system in light of PISA 2006*, pp. 4-5.

En todo caso, encontramos un primer problema en la validez de constructo de la prueba PISA 2006. Si bien el consorcio PISA no tiene la intención de que sea una prueba alineada al currículo, de todas formas es arbitrario, por principio, establecer que tanto los contenidos en el campo de las ciencias, como las formas de enseñanza que se utilizan en la escuela responden a un concepto de formación de aprendizajes y destrezas para la vida. Aunque PISA utiliza en sus documentos recientes con mayor frecuencia el término *competencias*, éste tiene varias acepciones en los diferentes reportes que construye.

En algunos documentos su énfasis se encuentra en la capacidad de transferir la información a situaciones nuevas,[22] también se refiere a ello como "la capacidad de los alumnos para identificar cuestiones científicas, explicar fenómenos de manera científica y utilizar pruebas científicas al encontrarse, interpretar y resolver problemas y tomar decisiones en situaciones de la vida real que tienen que ver con la ciencia y la tecnología".[23] Asimismo, considera que su concepto abarca "las competencias que se valoran en las sociedades modernas, y que implican muchos aspectos de la vida, desde el éxito en el trabajo hasta la ciudadanía activa".[24]

A simple vista, no queda claro que todas estas intenciones reflejen de la mejor manera lo que el sistema educativo mexicano establece en sus planes de estudio, ni en las estrategias que se aplican en el aula.

Para realizar un balance de otras cuestiones técnicas nos encontramos con una dificultad adicional. No podemos recurrir al informe técnico de 2006, dado que en la página donde la OCDE da a conocer la información sobre PISA afirma que este informe está en elaboración. Esta razón nos orilló a tomar el informe 2003, que nos permite establecer algunos elementos que pueden ser indicativos del proceso que se siguió en el ejercicio 2006. Para éste recurrimos al documento

22 "La característica principal que ha guiado la prueba PISA es su concepto innovador de la capacidad de los alumnos para extrapolar todo lo aprendido y aplicar sus conocimientos y destrezas en materias clave, su relevancia para la formación a lo largo de la vida y su regularidad", OCDE, *Competencias científicas para un mundo del mañana: informe pisa 2006*, p. 3.

23 *Ibid.*, p. 37.

24 *Loc. cit.*

que establece las orientaciones generales para los Administradores Nacionales del Examen, en nuestro caso el INEE. Estas orientaciones determinan las reglas de operación: para traducción y calibración de reactivos, para determinación del tamaño de la muestra de acuerdo con un mínimo exigido, de las condiciones para clasificar las escuelas con la finalidad de establecer rangos que permitan una aplicación de la prueba en diversos sectores. Por esta razón construimos esta sección con base en dos documentos *PISA 2003 technical report* y *Main study national. Project manager's manual*, 2005; de igual manera tomamos otros reportes de la información PISA, tales como el *Informe Pisa 2006. Competencias científicas para el mundo del mañana*, todos ellos publicados por la OCDE.

Una primera cuestión salta a la vista en la elaboración de la prueba PISA. Los reactivos y preguntas que la componen son elaborados por cuatro instituciones que se encuentran en los países desarrollados.[25] Esto significa que la construcción de los reactivos responde a la cosmovisión del conocimiento, aprendizaje, uso social y contexto de esos países. En un mundo globalizado eso no es de extrañar, pero en una sociedad con códigos socioculturales y lingüísticos tan diversos como la de nuestro país, es un tema que al menos debería invitar a una reflexión y análisis referidos a la manera como esta condición afecta el desempeño de los jóvenes mexicanos de 15 años a los que se les aplicó el examen. Si esta prueba se aplicó a estudiantes de 21 países, de los cuales siete son iberoamericanos, resulta llamativo que el papel que se les asigna a los miembros de estos países sea traducir reactivos y manuales, validarlos en un trabajo de campo, entrenar a los que aplicarán y codificarán la prueba y, en algunos casos, elaborar el reporte nacional.

Esto puede explicar por qué la información de algunas preguntas refleja más una visión eurocentrista o estadounidense, que del

25 OCDE, *PISA 2003 technical report*. La coordinación de esta tarea la realiza el Australian Council for Educational Research (ACER), con apoyo de Netherlands National Institute for Educational Measurement (CITO), Educational Testing Service (ETS, Estados Unidos), National Institute for Educational Policy Research (NEAR, Japón) y Westat (Estados Unidos) (Véase OCDE, *Main study national. Project manager's manual;* también, M. A. Díaz Gutiérrez *et al.*, *Pisa 2006 en México*, p. 17.

contexto latinoamericano. Véase en el caso de las preguntas liberadas donde una parte central son la Acrópolis de Atenas,[26] el Gran Cañón, la historia de Mary Montagu, la oveja Dolly y el diario de Semmelweiss. La cuestión de fondo consiste en reconocer que se presenta una situación desigual en estos exámenes al formular preguntas de aspectos culturales que pueden resultar más familiares para un grupo social que para otro. Por lo menos no sabemos hasta qué grado esta situación influye en los resultados que obtienen los que resuelven el examen en distintos lugares del mundo y del país.

Una tercera cuestión la podemos referir a los tiempos que PISA dedica a la confección de la prueba, su calibración, su traducción, la elaboración de la versión final y la aplicación en cada uno de los países. Varias cosas podemos aprender de esta actividad. La etapa de diseño de la versión PISA 2003 requirió tres años, más otro para su interpretación y elaboración de reporte final. Podemos afirmar que indudablemente no es una prueba improvisada, ya que uno de sus méritos es dedicar 10 meses al desarrollo de sus fundamentos, o sea, lo que en nuestra investigación denominamos *teoría del contenido y teoría del aprendizaje*, lo que constituye, sin lugar a dudas, una de las aportaciones significativas de este examen a la teoría del test.[27] Máxime cuando vemos en nuestro medio nacional que en dos o seis meses se elabora una prueba a gran escala y que sus resultados se ofrecen en el mejor de los casos tres meses después.[28]

26 Pensamos si no es más significativo para los jóvenes mexicanos que el ejemplo sea la degradación que sufren las pirámides de Monte Albán, o la cantera de la catedral de Morelia o de Guadalajara.

27 A. Díaz Barriga, "Las pruebas masivas. Análisis de sus diferencias técnicas".

28 Cada año se elabora una versión de la prueba ENLACE (2006-2008) que se aplica en cuatro grados de la educación básica, y se reportan resultados tres meses después de ello. La prueba ENLACE 2008 para bachillerato fue elaborada en menos de seis meses, mientras que el Examen de Habilidades Docentes 2008 fue elaborado en menos de 40 días y sus resultados se entregaron en menos de un mes.

Cuadro 1

Etapas de desarrollo del test para 2003

Tareas centrales	Tiempo de dedicación
Desarrollos de los fundamentos de la prueba	Septiembre 2000-julio 2001
Desarrollo de los reactivos	Septiembre 2000-octubre 2001
Presentación de los reactivos a los países participantes	Febrero-julio 2001
Revisión nacional de los ítems	Febrero-octubre 2001
Distribución del material a cada país	Noviembre-diciembre 2001
Traducción a lenguas nacionales	Diciembre 2001-febrero 2002
Codificación de la prueba en cada caso	Febrero 2002
Validación de la prueba en campo en los países participantes	Febrero 2002-julio 2002
Selección de los ítems que formarán parte del estudio	Julio-octubre 2002
Elaboración de la versión final y de sus cuadernillos (presentación en inglés y francés)	Octubre-noviembre 2002
Distribución de las versiones para aplicación del examen	Diciembre 2002
Entrenamiento en los códigos	Febrero 2003
Estudio en los países participantes	Febrero-octubre 2003

Tomado de: OECD, *PISA 2003 technical report.*

En las siguientes etapas, sin embargo, observamos varios problemas serios. Si utilizáramos términos económicos ubicaríamos en el marco de la teoría de la dependencia el papel que se asigna a todos los países que participan en la prueba: a los equipos nacionales les corresponde revisar los ítems, hacer ejercicios de validación, traducir cada reactivo al lenguaje nacional, y esperar a que el consorcio presente, en los términos que informa en 2003, la versión definitiva del examen en inglés y francés; si bien es necesario precisar que en el informe 2003 se especifica que sólo fueron enviados a un país de América Latina,[29] y que las pruebas piloto se realizaron en Australia, Japón y Noruega,[30] lo cual vuelve mucho más crítica esta cuestión.

En el documento *Main study national*, que publica el consorcio en 2005, además de señalar lo complejo que se ha hecho la formulación de la prueba PISA cuando se aplica simultáneamente en 21 países, se anotan las fechas de algunas reuniones, tales como: una

29 OECD, *PISA 2003 technical report*, p. 20.

30 *Ibid.*, p. 21.

reunión en Melbourne entre el 28 de septiembre y el 6 de octubre, en donde los expertos se reunieron con quienes diseñaron la prueba para después seleccionar los reactivos que se integran a ella.[31] En ese momento también acuerdan el calendario previsto para establecer el idioma que utilizarán y la forma como se realizará la verificación nacional. Para el 19 de diciembre ya se tenía que tener todo el material dispuesto para poder ser revisado nuevamente por el consorcio. Cada administrador nacional haría su envío. Llama la atención que no exista ninguna mención sobre los tiempos que se destinan a traducir el material, o que estos tiempos se encuentren implícitos en una etapa referida a la traducción de la prueba y realizar la verificación del material. Existen varios señalamientos sobre los criterios a emplear en la selección de las escuelas y los alumnos que resolverán la prueba; sobre la codificación de las respuestas; sobre la capacitación de quienes calificarán las respuestas, pero la información sobre traducción en este documento, como en el anterior, es muy pobre.

En torno a este tema se afirma que la prueba es formulada en inglés y francés y traducida a 36 idiomas, desde las llamadas lenguas modernas: alemán, italiano, ruso, portugués, español, hasta lenguas europeas menos conocidas: holandés, finlandés, sueco, búlgaro, eslovaco, esloveno, danés, estonio, croata, checo e islandés, entre otras, o lenguas de España como catalán y gallego. De los países asiáticos: coreano, japonés, chino, tailandés, y del oriente: árabe, hebreo, turco y griego. Sin duda, se trata de un esfuerzo mundial muy notable, pero precisamente esta tarea de traducción reclama ser mirada con una lupa especial. En este rubro el problema que reconoce el documento es la dificultad para revisar la equivalencia de las traducciones, dado que cada idioma tiene una escritura larga o corta.[32] Los investigadores son mucho más precisos en esta cuestión, ya que enuncian de manera clara que es inevitable que existan errores en la traducción de una pregunta a otro idioma, y que ello se debe en primer término

31 OECD, *Main study national*, p. 25.

32 *Ibid.*, p. 29.

a que cada idioma tiene una lógica de construcción propia (algunos dicen que tiene su propia epistemología).[33]

El español o castellano que se utiliza en América Latina está lleno de modismos y de formas peculiares de utilizar el lenguaje: la frase "cancelar un viaje o un hotel", significa cosas totalmente distintas si se emplea en Ecuador o Colombia o si se utiliza en México; afirmar que un "motor se rompió" refiere a temas diferentes en Argentina o México. Utilizar el "maíz como pienso", como se encuentra en un reactivo español, puede ser ininteligible para ciertos sectores mexicanos. Otras sutilezas conviene que sean exploradas, tales como los modismos y regionalismos que se dan en un país. Aquí valen las preguntas: ¿cuánto tiempo tarda el consorcio PISA en traducir una prueba?, ¿cuánto tiempo dedica a validar dicha traducción? El tema tiene distintas respuestas. Si uno se ciñe al reporte técnico 2003, parece que dos meses (diciembre-enero) son suficientes para realizar una tarea de esta envergadura, aunque en el reporte también se puede interpretar que esta actividad reclama la atención de seis meses, hasta que concluye la fase de validación de los reactivos en cada país. Por ahora no hay forma de determinar qué impacto tiene esta situación en los resultados que obtienen los alumnos mexicanos al resolver esta prueba, pero sí es conveniente hacer otra pregunta: ¿cuál es la razón por la que el reporte del INEE no considera relevante dar a conocer estos procesos técnicos de la prueba?

Sin lugar a dudas podemos considerar que esta situación ha abierto una discusión sobre los problemas de traducción de pruebas a gran escala (lo cual se puede considerar como una aportación colateral) pero, al mismo tiempo, es necesario reconocer el papel subordinado que el consorcio PISA establece para los países de habla hispana y, muy específicamente, para los siete latinoamericanos donde se aplica dicho examen.

La traducción es un tema fundamental en la tarea que la OCDE le deja a nuestros países. Lamentablemente, la escasa información

33 G. Solano *et al.*, "Traducción y adaptación de pruebas: lecciones aprendidas y recomendaciones para países participantes en *Trends in International Mathematics and Science Study*, TIMSS, Programa Internacional para la Evaluación de Estudiantes o Informe PISA y otras comparaciones internacionales".

que sobre este asunto se ofrece en el ámbito internacional también se reproduce en el nacional. En concreto, no podemos conocer cómo realizó esta tarea el INEE, como administrador nacional de la prueba a mayor detalle; cuánto tiempo, cuántos especialistas dedicó a su realización. Lo que expresa en su informe[34] se refiere a actividades como: traducción y adaptación de los materiales, diseño de la muestra de escuelas, selección al azar de los alumnos, aplicación de la prueba, codificación y captura de resultados y aplicación de reportes de validez. No podemos inferir si el Instituto realizó la traducción y calibración de los reactivos o cuánto tiempo dedicó a cada fase. Podemos suponer, eso sí, que utilizó la estrategia que han desarrollado en el propio instituto para analizar las dimensiones de error de una traducción, que funciona bajo la lógica de "juicio de expertos".

No cabe duda, pues, que se trata de la generación de un instrumento de registro y control de una serie de dimensiones, tales como: problemas de estilo, formato, semántica, información, currículum, etcétera, en un análisis de especialistas, pero que adolece de una información de base estadística que permita realizar inferencias más generalizables, aunque en este trabajo los especialistas reconocen que es inevitable el error y mencionan dos tipos de errores: un error común y un error fatal. Asimismo, reconocen que su modelo es "deficitario; es decir, está diseñado para detectar fallas, más que simplemente decidir si los ítems son o no aceptables. Está basado en el supuesto de que la traducción perfecta de una prueba es virtualmente imposible".[35]

En este rubro merece una mención el papel subordinado que tiene América Latina. Es interesante observar los objetivos que se asignó el Grupo Iberoamericano Pisa formado en 2005: a) realizar una ayuda mutua de los países iberoamericanos que participan en PISA, b) el Instituto Nacional de Evaluación Educativa se compromete a crear una página web de acceso restringido, c) España buscará

34 M. A. Díaz Gutiérrez *et al.*, *PISA 2006 en México*, p 63. En el anexo 3 el texto menciona a 17 profesores de secundaria o bachillerato como "especialistas en la elaboración de reactivos de ciencias".

35 L. Solano *et al.*, *op. cit.*, p. 17.

fondos para realizar cursos especializados en evaluación educativa, d) los miembros del grupo se comprometen a enviar a Argentina y Colombia las versiones nacionales de los *links* ítems y, e) Argentina, Chile y México compartirán con los miembros del grupo los ítems que han enviado a PISA.[36]

Esto indicaría que para las nuevas versiones de las pruebas se estarían incorporando reactivos elaborados en América Latina,[37] algo que muestra que este tema es preocupación del consorcio y confirma que existe un problema de validez de constructo en los reactivos que deviene de problemas de traducción, así como de los procesos socioculturales de cada nación o grupo social en específico (temas que han sido ampliamente estudiados y documentados por la literatura especializada).

A MANERA DE CONCLUSIÓN

La prueba PISA es una prueba a gran escala muy ambiciosa en sus objetivos: "determinar el grado en que los estudiantes de 15 años pueden aplicar los conocimientos y habilidades adquiridos en su escolarización para resolver problemas cotidianos". Refleja la cosmovisión que la globalización económica ha venido construyendo sobre el nuevo ciudadano del mundo: un sujeto cuya formación escolar le permita, de manera homogénea, resolver una cantidad de problemas que tiene la sociedad contemporánea. En cierto sentido, se puede afirmar que la prueba PISA es la versión globalizada de las intenciones de la pedagogía comparada, que surge a fines del siglo XIX, cuando se están estructurando los sistemas educativos nacionales con la finalidad de comparar los contenidos que se enseñan en cada sistema educativo. La prueba PISA busca valorar lo que se aprende, pero no bajo la forma de contenido, sino en la perspectiva de capacidad para resolver situaciones concretas del mundo actual.

36 Información sobre el Grupo Iberoamericano PISA, <http://www.gip.inee.edu.mx/Informacion_general.html>.

37 *Loc. cit.*.

Lejos de lo que el mismo consorcio considera, y también distante de la opinión de la mayoría de los administradores nacionales de la prueba, los objetivos de PISA y la forma de estructurar las preguntas demandan modificaciones centrales en el modo de organizar los contenidos en los planes de estudio, así como en las formas de enseñanza. Podemos afirmar que constituyen una nueva lucha para vencer el enciclopedismo, así como el tratamiento escolar de los contenidos.

La construcción de una prueba a gran escala como la prueba PISA constituye un reto importante que es necesario reconocer. Reclama el tratamiento de los problemas técnicos del test con una nueva óptica; asistimos, sin lugar a dudas, a una reconfiguración de la teoría del test para posibilitar nuevos desarrollos en sus planteamientos. La elaboración de marcos teóricos sobre los contenidos constituye un paso más allá de la tradicional tabla de especificaciones utilizadas en la construcción de estos instrumentos; por su parte, la determinación de niveles de dificultad del contenido de igual manera representa un avance. En este sentido, se puede afirmar que, desde el punto de vista técnico, la prueba PISA hace una importante aportación al desarrollo de la teoría del test.

Pero como investigadores también estamos obligados a mostrar sus límites. El primero deriva de sus propios objetivos: comparar las habilidades y destrezas para la vida que en el campo de las ciencias han adquirido los estudiantes de los más de cincuenta países donde se aplica el examen. Comparar concibiendo a un ciudadano del mundo global, a un ciudadano que desarrolla determinadas habilidades con independencia de su inserción social, esto es, con independencia de la historia y cultura inmediata desde donde tiene su experiencia de "mundo y vida", puede llevar a importantes errores de apreciación. En la construcción de la prueba no parece importar lo que singulariza al sujeto; de hecho no hay mediaciones, ni las que provienen de la historia, o de la cultura, ni las que emanan de las condiciones socioeconómicas. Según la prueba, los ciudadanos de la nueva era global generan las mismas habilidades y estrategias cognitivas. Y es claro que los países del tercer mundo se consideran como espacios neocolonizables por los expertos PISA; en nuestros países hay "administradores nacionales de la prueba", las preguntas son

elaboradas por especialistas de cuatro países (Noruega, Japón, Estados Unidos y Australia), y nos corresponde validar las traducciones de la prueba en periodos cortos de tiempo.

Esto origina que algunos problemas de validez de constructo salten a la vista en la elaboración y aplicación de este examen, si bien se debe reconocer que hay un trabajo muy minucioso al elaborar los marcos teóricos (la teoría del contenido) de cada examen, lo que se manifiesta en una dedicación a esta etapa de más de un año, frente a dos meses para traducción de los reactivos y cuatro para su validación.

Las preguntas son elaboradas por especialistas de cuatro países; indudablemente, éstas reflejan la perspectiva que tienen estos especialistas, tanto de la ciencia como de los problemas cotidianos en los que un estudiante puede requerir determinada información, así como de la generación de alguna estrategia (habilidad y destreza) de resolución de un problema. Pareciera que en América Latina no existen códigos científicos, no existen problemas del entorno que puedan ser objeto de análisis, pues los reactivos de la prueba suponen que los estudiantes de todos los sectores sociales del mundo utilizan "bloqueadores solares", conocen las aportaciones de Mary Montagu, o saben qué es un "Herr-doktor" en un hospital alemán, etcétera.

Los especialistas que elaboran los reportes nacionales de los resultados del examen no se consideran obligados a dar cuenta de estos temas. Es más fácil informar el puntaje promedio que tuvieron los estudiantes de secundarias y/o bachilleratos públicos o privados, los resultados por entidad federativa, el comportamiento en el examen de estudiantes de acuerdo con diversos grupos sociales, o llenar la cabeza del público en general con números, con cifras, lo que en realidad refleja de manera muy limitada lo que actualmente se denomina evaluación educativa.

No hay en el reporte nacional de PISA, ni en los reportes internacionales, una debida atención a problemas de validez de constructo que permitiría establecer si los resultados que reporta un examen se pueden sostener, al mismo tiempo que determinar de manera cuantitativa el grado de error de tales mediciones. Los problemas de traducción son muchos, ciertamente la tarea es harto más compleja de

lo que parece en primer término, pero son problemas clave que pueden explicar en cierto grado los resultados deficientes que obtienen los alumnos. Igualmente graves son los problemas que emanan del ámbito sociocultural. Éste es un tema que se estudió con bastante detalle en el siglo pasado en los Estados Unidos, y constituye un punto que sigue siendo de interés fundamental para los investigadores del campo de la evaluación, pero sencillamente no es considerado en los reportes nacionales.

De alguna manera se envía el mensaje de que los profesores mexicanos sean formados en instituciones de los países que elaboran estas pruebas con la finalidad de que puedan enseñar como se espera que trabajen los alumnos en este tipo de exámenes. Pero, para señalar sólo un ejemplo de lo discordante que un supuesto tal puede llegar a ser, pensemos en que en nuestro medio importan las culturas precolombinas: baste recordar el intenso debate nacional que se generó cuando se pretendió que no formaran parte de los contenidos de la educación secundaria en la reforma propuesta en 2004.

Finalmente, es de relevancia reconocer que gran parte de la información de la prueba PISA, sobre todo la que guarda relación con su estructura técnica, es considerada por los expertos del consorcio y por los administradores nacionales como confidencial, información a la que no se tiene acceso. A ello se añade una visión particular de los responsables de esta tarea en nuestro país. Sólo recordemos que mientras los responsables PISA señalaban que no había autorización del consorcio para informar en qué escuelas se aplicaría la prueba, las autoridades educativas daban la lista de estas instituciones. Lo mismo acontece con otras informaciones. Negar el acceso a la información técnica de la prueba no sólo dificulta el trabajo de investigación, sino que demuestra una visión cortoplacista de los administradores nacionales.

Más allá de sus méritos, la prueba PISA reclama convertirse en objeto de investigación; al mismo tiempo, es necesario revisar el papel que el consorcio asigna a la mayor parte de los países del mundo en su elaboración, aplicación e interpretación.