Profesorado
Revista de currículum y formación del profesorado

VOL. 17, Nº 2 (mayo-agosto 2013)
ISSN 1138-414X (edición papel)
ISSN 1989-639X (edición electrónica)
Fecha de recepción 24/04/2013
Fecha de aceptación 06/06/2013

PISA, COMPARACIONES INTERNACIONALES, PARADOJAS EPISTÉMICAS

PISA, international comparisons, epistemic paradoxes

David Scott
Institute of Education, University of London
E-mail: D.Scott@ioe.ac.uk

Resumen:
El artículo analiza críticamente, desde una perspectiva del realismo crítico, las formas de conocimiento que se ponen a prueba en PISA. Después de una diferenciación inicial y fundamental entre dos formas de conocimiento, desenmascara las falsas creencias o suposiciones acerca de los rasgos característicos de estas dos formas de conocimiento y de la problemática relación entre el conocimiento y su evaluación. La relación entre el conocimiento y la evaluación se agrava por diversas "tecnologías de examen", como ya sea un incentivo está unido a la práctica de la prueba, la motivación de los estudiantes para hacer la prueba y el formato de la prueba, que podría favorecer a algunos grupos en comparación con otros. Las evaluaciones de estudiantes comparadas internacionalmente (como PISA) se enfrentan a la dificultad añadida de construir pruebas desligadas del currículum que suponen la idea de una forma universal de conocimiento. La noción presenta varios supuestos reduccionistas y no tiene en cuenta adecuadamente las diferencias culturales que podrían afectar, de varios modos, al rendimiento de la prueba. La última crítica se dirige a la forma como se publican los resultados de PISA se publican en los cuadros nacionales comparativos poniendo énfasis en la posición que se ocupa en lugar de la calificación obtenida
Palabras clave: PISA, Comparaciones internacionales, conocimiento, epistemología, discurso, performatividad

---

Traducción de Mireia Moll Garriga. Revisión técnica de Antonio Bolívar y Antonio Luzón.

Abstract:
This paper critically analyze the forms of knowledge that are tested in PISA from a critical realist perspective. After a initial and fundamental differentiation between two forms of knowledge, unmasks false beliefs or assumptions about the characteristic features of these two forms of knowledge and about the problematic relationship between knowledge and its assessment. The relationship between knowledge and its assessment is further aggravated by various 'examination technologies' such as whether an incentive is attached to the taking of the test, the students' motivation to take the test and the test format, which might favour some groups in comparison with others. International comparative student assessments (like PISA) face the additional difficulty of trying to construct curriculum-free tests underpinning the idea of a universal form of knowledge. This notion makes a number of reductionist assumptions and does not account properly for cultural differences which might affect test performance in several ways. The final criticism is directed at the way PISA results are published in comparative national tables thereby putting emphasis on position rather than score.
Key words: PISA, international comparisons, Knowledge, epistemology, discourse, performativity.

1. Introducción

Se pueden identificar dos formas de conocimiento (a las que llamaremos K_a y K_b). K_a representa aquellos conjuntos de conocimientos, habilidades y actitudes de una persona, conocidos colectivamente como competencias. K_b representa aquellos conjuntos de conocimientos, habilidades y actitudes de una persona que le permiten superar con éxito los exámenes, particularmente, exámenes relevantes. K_a y K_b tienen diferentes características. Si un sistema de educación presenta test relevantes, es decir, pruebas en las que hay beneficios significativos ligados al éxito en la prueba para una persona, una institución o incluso una nación; entonces hay dos consecuencias. La primera es que K_b se convierte en la forma dominante de conocimiento en el currículum y la segunda es que K_a se va transformando con el tiempo para parecerse más a K_b, es decir, cada vez tiene más características similares. Los que elaboran las pruebas o test comúnmente confunden K_a y K_b, y con ello, hacen una serie de suposiciones falsas sobre el conocimiento y su evaluación, con la consecuencia de que estas dos formas de conocimiento se vuelven indistinguibles en las mentes de los que toman decisiones políticas, profesionales de la educación, estudiantes y otros interesados. Además, el conocimiento de las competencias de una persona o de un grupo (por ejemplo una nación, una cohorte de edad o una categoría), a través de sistemas internacionales comparativos de pruebas como el Programa Internacional para la Evaluación de Estudiantes (PISA) (OCDE 2000, 2001, 2006, 2009), es respaldado por una particular y específica noción geo-histórica de comparación.

La filosofía del realismo crítico proporciona una alternativa a esta posición. Los realistas críticos sostienen las siguientes tres afirmaciones: existen diferencias significativas entre el campo transitivo del saber y el campo intransitivo del ser; el mundo social está sistemáticamente abierto; y los investigadores y observadores deben comprender la profundidad ontológica de la realidad. La primera de estas afirmaciones se refiere a la distinción entre el mundo intransitivo del ser (el ámbito ontológico) y el mundo transitivo del saber (el ámbito epistemológico), por lo que mezclarlos resulta ilegítimo, ya sea al alza, cayendo en la falacia epistémica, o a la baja, dando lugar a la falacia óntica (Bhaskar, 1989). De esto se derivan dos implicaciones. Los objetos sociales, aunque reales, cambian constantemente, y por lo tanto, es el objeto cambiante el que relativamente perduran, incluso hasta el punto de que el objeto se ha transformado tan completamente que apenas es reconocible en relación a lo que fue. La segunda implicación es más importante, y es que, en determinadas circunstancias y bajo ciertas condiciones, los objetos sociales del ámbito transitivo pueden penetrar al ámbito intransitivo y ser "cosificados". De ello se desprende

que, en principio, la medida de las capacidades de un individuo o de un grupo de individuos puede activar propiedades emergentes del constructo a medir y cambiar ese constructo.

Esto también sugiere que puede darse una disyunción entre ambos campos que provoque una desincronización entre ellos. Bhaskar (1989) identifica cuatro razones para que esto ocurra: en el mundo existen objetos sociales tanto si éstos son conocidos como si no; el conocimiento es falible ya que cualquier afirmación epistémica puede ser refutada; hay verdades trans-fenoménicas referidas al mundo empírico que no consideran los niveles más profundos de la realidad social, es decir, la labor de los mecanismos sociales; y lo más importante, hay verdades contra-fenoménicas en las que aquellas estructuras profundas pueden de hecho estar en conflicto con su apariencia. La mezcla entre ambas nos lleva a la confusión y a la apropiación indebida.

La segunda afirmación es que el mundo social está sistemáticamente abierto. Los sistemas cerrados se caracterizan por dos condiciones: los objetos funcionan de manera coherente y no cambian su naturaleza esencial. Ninguna de estas condiciones se da en sistemas abiertos. En sistemas cerrados, las regularidades medidas son sinónimo de mecanismos causales. Por lo tanto, la experimentación es innecesaria ya que las condiciones experimentales están presentes de manera natural. Hay dos alternativas: el cierre artificial y el uso de métodos y estrategias que se ajusten a la apertura sistémica, incluyendo, aunque no exclusivamente, juicios inferenciales a partir del análisis de evidencias indirectas. La primera de estas alternativas, el cierre artificial, hace una serie de suposiciones sin fundamento: las transferencias más allá del contexto se pueden realizar aún cuando el conocimiento original se construya bajo condiciones artificiales; y este conocimiento original está correctamente relacionado con la constitución del objeto, es decir, el resultado de la evaluación es isomorfo con la capacidad del individuo, tanto si ésta se expresa como un conjunto de conocimientos, una habilidad o una actitud. Por lo tanto, nos quedamos con los métodos y estrategias que se ajustan al principio de apertura sistémica.

La tercera afirmación es que la realidad social tiene una profundidad ontológica. Los objetos sociales son la manifestación real de las tipos idealizados utilizados en el discurso y son el centro de cualquier investigación. Se estructuran de varias formas y, debido a ello, poseen poderes. Los poderes que estas estructuras (o mecanismos) ejercen pueden ser uno de los tres tipos siguientes (Brown et al., 2002): pueden ser poderes que se posean, que se ejerzan o que se actualicen. Los poderes *que se poseen* son poderes que los objetos tienen, tanto si éstos son activados por las circunstancias como si no. Sus efectos pueden no ser evidentes en cualquier fenómeno observable. Los poderes *que se ejercen* han sido provocados y tienen efectos en un sistema abierto, y como resultado están interactuando con otros poderes de otros mecanismos dentro de su esfera de influencia. Estos poderes ejercidos pueden no dar lugar aún a ningún fenómeno observable ya que estos otros poderes pueden estar actuando contra ellos. Los poderes que han sido *actualizados* generan efectos; dentro del sistema abierto, estos poderes trabajan junto con otros, pero en este caso no han sido suprimidos o contrarrestados. Las estructuras personificadas, institucionales o discursivas pueden ser poseídas y no ejercidas o actualizadas, poseídas y ejercidas, o poseídas y actualizadas. Como consecuencia, un modelo causal basado en constantes conjunciones se rechaza y se sustituye por uno generativo-productivo, y los objetos y las relaciones entre los mismos (como en los sistemas educativos o regímenes de pruebas) tienen propiedades emergentes.

A partir de esta perspectiva crítica realista se deducen tres proposiciones. La primera es que cualquier descripción que hagamos de la acción humana y de sus capacidades depende

de "la causalidad intencional o la causalidad de la razón" (Bhaskar et al., 2010, p. 14). En segundo lugar, estas descripciones deben tener en cuenta "el materialismo sincrónico de los poderes emergentes" (ibid.), es decir, el cambio secuencial en el tiempo del poder de los objetos, ya sean discursivos o personificados; y en tercer lugar, hay una necesidad de reconocer "la(s) implicación(es) *valorativa y crítica* del discurso factual" (ibid., en cursiva). Estos tres principios tienen implicaciones importantes para el desarrollo de una amplia explicación acerca de los regímenes de pruebas transnacionales y transculturales como PISA.

2. Falsas creencias

La postura que, por defecto, adoptan aquéllos que trabajan dentro de la tradición psicométrica de conocer otras mentes, es la de que una persona tiene una serie de competencias (esto es, un conjunto de conocimientos, habilidades y actitudes) que podemos describir como el contenido de la mente de esa persona y que posteriormente podemos caracterizar usando métodos de experimentación y pruebas. Por lo tanto, potencialmente hay una puntuación verdadera para cada persona, y esta puntuación verdadera representa en términos simbólicos su capacidad en el dominio particular que está siendo evaluado. Por varias razones, durante el proceso de construcción de dicha puntuación verdadera pueden cometerse errores, pero éstos son corregibles, es decir, pueden corregirse mediante el uso de diferentes (y por lo tanto, implícitamente mejores) métodos y enfoques. Los errores pueden deberse a una mala elección en el tipo de instrumento usado para medir la puntuación verdadera de una persona o porque el estado emocional y afectivo de ésta es tal que crea una falsa impresión de sus capacidades. Por otra parte, me gustaría sugerir que hay una serie de falsas suposiciones hechas aquí, quizás mejor expresadas como falsas creencias.

La primera de ellas es la de que una persona tiene un conjunto de conocimientos, habilidades o disposiciones que se configura de una determinada manera (es decir, tiene una gramática), y es este conjunto, o al menos algunos de sus elementos, el que se evalúa directamente cuando esa persona es puesta a prueba. Por el contrario, cualquier prueba que se lleva a cabo con el propósito de determinar si un individuo posee, no posee o incluso parcialmente posee estos atributos, siempre supone un proceso *indirecto* de evaluación, donde el elemento adicional es una conjetura, una deducción lógica o una suposición. Además, la actuación requerida que se obtiene durante la prueba está específicamente relacionada con la tecnología usada para la misma; así si, por ejemplo, se elige un test de respuesta múltiple, la respuesta correcta y, por lo tanto, la construcción correcta del problema, se estructura para adaptarse a esta tecnología. Con el fin de obtener una medida real de la capacidad de esa persona (es decir, K_a), y no –conviene destacarlo–, una medida comparativa del constructo a evaluar a nivel individual o grupal (es decir, K_b), se tendría que usar entonces un modo retroductivo de inferencia para identificar cuál debe haber sido el caso para dar lugar al evento observado (es decir, el sujeto evaluado contestando un test de respuesta múltiple en una prueba estandarizada).

La segunda creencia falsa es que esta gramática se organiza en elementos, que existen relaciones entre los mismos y que cada elemento se puede escalar, por lo que puede ser directamente investigado. Esto se puede contrastar con una posición que sugiere que, en la aplicación del conjunto de conocimientos, habilidades o disposiciones, ya sea para los fines de una prueba o para su uso en la vida diaria, se invocan otra serie de elementos de conocimiento, habilidades y actitudes. Este hecho no debería confundirse con la idea de que el contenido del currículum no puede desconectarse del propósito de evaluar, lo que lleva a

una creencia propiamente holista (vid. Curren, 2006, para una refutación). Lo que aquí, en contraposición, se discute es que durante la aplicación de un conjunto de conocimientos, habilidades o actitudes, ya sea para los fines de una prueba o para otros diferentes, se necesitan otro tipo de conocimientos y habilidades, y puede que el sujeto evaluado no posea suficiente conocimiento en estas materias o no sea suficientemente habilidoso en relación a las mismas. Por ejemplo, la aplicación de habilidades matemáticas de un nivel superior, como puede ser la resolución de ecuaciones algebraicas, lleva aparejados un conocimiento y una capacidad en habilidades matemáticas de un nivel inferior, como la suma y la resta.

Por otro lado, hay un conjunto de factores que pueden dar lugar a varianza irrelevante del constructo (Messick, 1989), esto es, una varianza entre una población de sujetos que están siendo evaluados como resultado de factores que no tienen nada que ver con el constructo que se evalúa. Incluso si el conocimiento de ese constructo o la competencia en el mismo está igualmente distribuida en esta población, algunos sujetos lo harán mejor que otros (esto es, en sus puntuaciones reales) y ello no se debe a que tengan un mayor conocimiento o a que sean más competentes en el constructo que está siendo evaluado. Esto podría implicar una representación insuficiente del constructo o una sobrerrepresentación del mismo (William, 2010) y –dentro de los límites de la prueba en sí misma– es imposible determinar cuál de ellas ha ocurrido. El reto para los examinadores es –entonces– eliminar esta varianza irrelevante del constructo. Sin embargo, esto no está exento de problemas. En primer lugar, no podemos decir a ciencia cierta cuál podría ser la varianza porque no sabemos qué es una puntuación verdadera para un individuo o una puntuación verdadera global para un grupo, y por lo tanto no tenemos nada con qué compararla. Pueden hacerse comparaciones analíticas, y en PISA se hacen, i) en el tiempo (entre T_1 y T_2, donde T representa un momento concreto), ii) entre diferentes capacidades (si un individuo es experto en C_a, entonces también lo será en C_b, donde C representa una capacidad), iii) entre diferentes constructor (Co_1 tiene el mismo nivel de dificultad que Co_2, donde Co se refiere al concepto/constructo), iv) entre diferentes contextos de realización (R_1 se considera isomorfa a R_2, donde R se refiere a la realización), v) durante la misma prueba, en dos momentos diferentes (ésta es una medida externa de fiabilidad, R_a), vi) con diferentes ítems de la misma prueba, en un momento concreto (ésta es una medida interna de fiabilidad, R_b) y vii) en pruebas comparables en dos momentos diferentes (ésta es otra medida externa de fiabilidad, R_c). Con i) se supone que no se activan propiedades emergentes del constructo que se evalúa, y además, que el aprendizaje no tiene lugar, como resultado de la prueba o de otra cosa, entre T_1 y T_2. Con respecto a ii) se hace el supuesto de que la experiencia en capacidades específicas automáticamente se transfiere a una experiencia en general. Con la tercera comparación analítica (iii), la hipótesis que se hace es la de que todos los constructos medibles tienen el mismo nivel de dificultad en su adquisición y en su aplicación. La cuarta de nuestras medidas (iv) tiene por objeto confirmar la validez de una calificación en una prueba examinando si esa aptitud puede aplicarse a otros contextos espacio-temporales fuera del entorno de la prueba. Se asume que el constructo que se está evaluando tiene características transferibles y no está conectado específicamente a una realización representativa particular. Finalmente, con v), vi) y vii) se hace la suposición de que si una puntuación en una prueba es fiable entonces también es válida. Cada una de estas comparaciones analíticas se basa en suposiciones o creencias que a su vez necesitan ser verificadas, o por lo menos pueden dar confianza en su uso. Y de este modo, debe proporcionarse una justificación adicional para cada uno de estos supuestos.

Un segundo problema que aparece al eliminar la varianza irrelevante del constructo es que ésta no puede conseguirse mediante la sustitución de una competencia por un

constructo de conocimiento, a pesar de que ésta es la clara intención de los creadores de la prueba de PISA. Por ejemplo, PISA 2006 trató de evaluar tres competencias del campo de las ciencias: "i) Identificación de temas científicos; ii) Explicación científica de fenómenos; y iii) Uso de pruebas científicas" (OCDE 2006: 12). Esto se debe a que los problemas relacionados con la varianza irrelevante del constructo se aplican por igual al constructo del conocimiento y de la competencia y, además, con relación a la evaluación de constructos de competencias, está el problema de las múltiples interpretaciones que se hacen. Tradicionalmente, esto se describe como un problema relacionado con la fiabilidad del examinador.

Los diseñadores de las pruebas enfrentados por el problema de la varianza irrelevante del constructo pueden tratar de reformular el constructo, para que aquellas materias que pueden considerarse separadas del mismo, como el elemento tiempo para resolver un problema en un test, ahora se convierta en parte del constructo, es decir, ahora la evaluación se relaciona con la capacidad para resolver el problema dentro de un período de tiempo definido y no sólo con la capacidad para resolver dicho problema. Esto introduce un elemento de representación dentro del constructo en sí mismo. Una vez más, este movimiento está plagado de problemas, ya que debilita la idea de que la experiencia personal en ese constructo puede trasladarse a otros ámbitos porque ahora es más dependiente del contexto como una evaluación. Lo que se ha debilitado es la validez predictiva de la evaluación. En entornos de prueba transnacionales como PISA algunos de esos elementos de la realización pueden estandarizarse, es decir, las pruebas se realizan en condiciones más o menos similares. Sin embargo, lo que no puede estandarizarse es la relación entre lo que se enseña y lo que se evalúa, cómo este conocimiento evaluado se relaciona con su uso en otros entornos y la capacidad de enfrentarse a una prueba del individuo o del grupo.

Una tercera falsa creencia es que en el uso de un conjunto de conocimientos o en el desempeño de una habilidad, o en la aplicación de una actitud o disposición, no tiene lugar ninguna transformación interna. (De hecho, tanto las transformaciones internas como las externas no se tienen en cuenta dentro de las consideraciones psicométricas tradicionales). Por el contrario, dentro de la mente de una persona se activan dos conjuntos de conocimientos. El primero es el conjunto de conocimientos original (K_a); y el segundo es el conjunto transformado (K_b). Además de esto, K_b no es sólo el resultado de un mecanismo causal en el trabajo sino que también puede, en diferentes momentos, influir y transformar K_a, es decir, tiene la capacidad de doblarse sobre sí mismo y actuar recursivamente para cambiar su forma original.

También existe un proceso de transformación externa en el trabajo y, así, una cuarta falsa creencia es la de que evaluar los conocimientos, habilidades y aptitudes de una persona no tiene efectos colaterales ("washback") en K_a, el constructo original de conocimiento, o en K_b, el conjunto de conocimientos transformado internamente preparado para la prueba. Por contra, el proceso bien documentado de efectos colaterales ("washback") funciona precisamente en este sentido (Stobart, 2009), por lo que en lugar de que la evaluación actúe simplemente como un dispositivo descriptivo, también lo hace de una variedad de maneras para transformar el constructo que se pretende medir. Los efectos "washback" funcionan sobre una serie de objetos y de diferentes maneras. Así, por ejemplo, hay efectos "washback" en el currículum, en la enseñanza y el aprendizaje, en la capacidad del individuo y más fundamentalmente en las estructuras del conocimiento, aunque estos cuatro mecanismos se confunden con frecuencia en las mentes de los agentes educativos.

Los efectos "micro-washback" trabajan directamente sobre la persona, mientras que los "macro-washback" lo hacen sobre las instituciones y sistemas, que posteriormente tienen

impacto en los individuos que se encuentran dentro de esas instituciones y sistemas. Por ejemplo, a un nivel global, los decretos y leyes pueden conllevar cambios en los planes de estudio nacionales y en los sistemas nacionales de pruebas, lo que a su vez dará lugar a cambios en el currículo y en la evaluación a nivel de las escuelas y por consiguiente a cambios en lo que se aprende y en lo que un individuo considera como conocimiento realizado. Por lo tanto, lo que se considera un conocimiento realizado apropiado ha cambiado como resultado de cambios a nivel global, nacional y escolar. Los efectos colaterales o "washback" no funcionan de una manera determinista, ya que hay un gran número de actividades que deben ser coordinadas durante la secuencia de eventos para lograr el resultado deseado, y mecanismos como éstos tienen propiedades emergentes porque operan en sistemas abiertos (Bhaskar, 1989).

El argumento que, por lo tanto, dan los psicólogos cognitivos y los constructores de las pruebas es el de que no se producen procesos internos o externos de transformación cuando los conocimientos, habilidades o actitudes de una persona se ponen a prueba; es decir, esa persona sabe A, o tiene la habilidad B, o la disposición C, y en el momento en que muestra ese conocimiento, usa esa habilidad o permite que se dé esa disposición no se produce ningún cambio en el constructo original de conocimiento, conjunto de habilidades o actitudes, a fin de que esa persona responda de manera adecuada a la situación que se enfrenta. Por el contrario, me gustaría sugerir que hay un proceso de transformación y que puede tomar diversas formas, a saber, un aumento y por lo tanto retención del dominio original del conocimiento, habilidad o actitud; o una subsunción, donde el dominio original del conocimiento es incluido dentro de un nuevo dominio más amplio y, por lo tanto, pierde su identidad; o una sustracción de manera que las partes son descartadas para dar cabida a las contingencias de la nueva configuración.

Lo que esto también destaca es que en el proceso de determinar si una persona sabe esto, puede hacer aquello, o tiene la disposición necesaria, se necesita un proceso inferencial para que el observador pueda pasar de las evidencias, es decir, el resultado del test, a una descripción del estado real. Se supone que si una persona puede hacer X durante la prueba, entonces también puede hacerlo en diferentes situaciones, o si esa persona sabe algo durante la prueba, entonces también lo sabrá en otras situaciones. Es, en resumen, el problema de la transferencia (de T_1 a T_2 o de C_1 a C_2, donde T se refiere a un momento en el tiempo y C al contexto de aplicación) y es problemático porque es prospectivo y morfogenético. Se puede desarrollar una medida del éxito predictivo para determinar si una persona o un grupo de personas pueden hacer X en otros escenarios fuera del entorno de la prueba, sin embargo, es una medida poco fiable por dos razones. Eventos, acontecimientos y sucesos imprevistos durante el intervalo entre los dos momentos concretos (T_1 – condiciones de la prueba y T_2 – el contexto de aplicación) no pueden controlarse; y las dos actividades diferentes no son comparables.

Una quinta falsa creencia es que el proceso de evaluar funciona en una sola dirección de manera lineal. Por ejemplo, una persona sabe X, esa persona se somete a un examen que está diseñado para evaluar los indicios de X en una población de conocedores con características similares, y se registra una puntuación en relación con ese constructo indicando que esa persona lo conoce, no lo conoce o lo conoce hasta cierto punto. No se tiene en cuenta la bidireccionalidad, que incorpora flujos hacia adelante y hacia atrás, de manera que la toma de la prueba y el registro de la calificación impactan e influyen en el constructo original de conocimiento. Esto cambia la estructura del constructo (tanto cuantitativa como cualitativamente) y sus posibilidades, tomando la determinación original de la misma y poco fiable.

Una sexta falsa creencia es que diferentes tipos de conocimiento, incluyendo los que están en niveles de abstracción diferentes, pueden ser evaluados usando los mismos procesos algorítmicos. Por ejemplo, evaluar un conocimiento de hechos y evaluar la capacidad para sintetizar hechos básicos son procesos diferentes. Y esto es así porque en el primer caso, los ítems de la prueba se refieren directamente al constructo que se evalúa, mientras que en el último se refieren a un ejemplo de dicho constructo y el dominio exitoso de dicho constructo tiene que ser inferido del dominio exitoso del ejemplo. Además, este último proceso tiene que satisfacer, por lo tanto, criterios tales como la relevancia, la calidad y la fuerza probatoria para esa relación inferencial entre el ejemplo y el constructo que se considera válido.

Una séptima falsa creencia es que el rendimiento en la prueba representa en mayor o menor medida (dado que la persona puede haber estado distraída o limitada de algún u otro modo) lo que la persona evaluada puede hacer o mostrar, en lugar de haber una diferencia cualitativa entre el rendimiento en la prueba y el constructo, habilidad o actitud de la persona evaluada. Un individuo puede tener que redefinir su conjunto de conocimientos para adecuarlos a la prueba, y por lo tanto la evaluación de su dominio del constructo no determina su capacidad en relación con el constructo original pero sí que determina si la persona evaluada ha entendido satisfactoriamente cómo revisar su capacidad para adaptarla a las demandas de la tecnología usada para la prueba.

Una octava falsa creencia es que una prueba puede construirse al margen de la cultura o libre de esas cuestiones que perjudican a algunos tipos de alumnos a costa de otros. Este mecanismo funciona de varias maneras: los constructores de las pruebas pueden usar material de base que sea desconocido para algunas de las personas evaluadas pero no lo sea para otras; los ítems de la prueba pueden haber sido enseñados de diferentes maneras a diferentes grupos de personas, es decir, se les han dado diferentes valores, o se han enseñado en un orden diferente o incluso pueden no haberse enseñado en absoluto; y la tecnología usada para la prueba puede ser desconocida para ellos a causa de factores que son periféricos a la articulación o al uso del constructo particular, pero centrales a la tecnología de la prueba usada para evaluarlo.

3. Tecnologías usadas en las pruebas

Si no existe un incentivo ligado a la realización de la prueba, es decir, un beneficio personal como conseguir el ingreso en una institución de educación superior, o una recompensa monetaria, o el fomento de la trayectoria de aprendizaje de un estudiante, o una ventaja a nivel nacional, entonces no es probable que el estudiante se la tome muy en serio. El valor que se le atribuya es siempre cuestión de percepción, más que designación, y esto significa que diferentes tipos de estudiantes estarán motivados para hacerlo bien en diferentes grados. Psicólogos cognitivos y constructores de pruebas argumentan que estas características individuales de los examinandos se tienen en cuenta a nivel de grupo, y el argumento que se hace entonces es que estas características, como la propensión a perder la concentración durante una prueba o no dar cuenta real de sus capacidades porque la tecnología usada para la prueba no les ofrece ningún incentivo para hacerlo bien, o tener un estilo de presentación que sea incompatible con las posibilidades de la tecnología usada para la prueba, se distribuyen al azar entre los miembros de cualquier grupo, y por lo tanto no afectan a las puntuaciones a nivel de grupo. Como resultado, los grupos pueden compararse con otros sin problema. Sin embargo, la suposición de que estas características de los

miembros del grupo se distribuyen uniformemente es falsa (cf. Mac Ruairc, en este monográfico) y, además, ésta es una medida de fiabilidad más que una validez del constructo. Es más, estas características pueden ser las características definitorias del grupo.

Como ejemplo, tomemos un test de respuesta múltiple. La tecnología sólo permite un rango limitado de respuestas; por lo tanto hay una alta probabilidad de errores falsos negativos y falsos positivos (Wood & Power, 1987), a pesar de que se insertan preguntas engañosas como preguntas para permitir que se realicen controles de fiabilidad. Solamente un número limitado de elementos de conocimiento y procesos pueden ser potencialmente evaluados porque se piden respuestas correctas, y esas respuestas se enmarcan de manera que no permiten digresiones o respuestas equívocas. Como resultado, esta tecnología tiene el efecto de ensanchar la brecha entre la capacidad del individuo y su desempeño (tanto interna como externamente), porque la prueba se construye para que tenga algunas de las características del constructo original de conocimiento y potencialmente su aplicación. En definitiva, a las personas evaluadas se les da una escasa capacidad de decisión y, por lo tanto, al menos en principio, los test de respuesta múltiple tienen una mayor propensión al efecto colateral ("washback") en el currículum. Además, las características de la tecnología usada para los test de respuesta múltiple favorecen a algunos grupos en comparación con otros, es decir, los chicos pueden tener ventaja sobre las chicas.

Por contraste, un ejemplo es el uso de un formato de redacción libre para determinar la capacidad comparativa de un grupo. A cada candidato se le da una amplia capacidad de decisión, aunque los efectos de la falta de fiabilidad del marcador pueden ser elevados. La evaluación no se centra en hechos discretos sino en competencias generales, es decir, en la habilidad para sostener un argumento. Por lo tanto, en principio, puede ser más capaz de medir habilidades de alto nivel. La validez puede ser fuerte si es entendida como un alineamiento entre los conocimientos, las habilidades y las disposiciones de la persona y la descripción que se hace de ellos. Debido a que el marcador de discreción es alto y a que se permite al candidato más libertad en la formulación de sus respuestas, la posibilidad de un efecto "washback" significativo es reducida.

Una prueba es siempre una representación. Los examinandos adecúan sus respuestas a la prueba en términos de lo que ellos perciben como la respuesta correcta. Esto opera a nivel inconsciente y es poco notable. Cuando mantenemos una conversación con otra persona, adecuamos nuestras respuestas y nuestro modo de responder a cómo pensamos que nuestros mensajes van a ser recibidos. Con relación a la evaluación, hay un elemento adicional, y es que las personas evaluadas adecúan sus respuestas desde el punto de vista de su percepción de lo que consideran que es la respuesta correcta. Si, por ejemplo, hay alguna ambigüedad en la pregunta, el examinando se hace la pregunta: ¿qué tipo de respuesta debería dar que sea más probable que tenga como resultado la adjudicación de una calificación máxima? Los constructores de pruebas abogan por escribir preguntas o construir problemas que se respondan con la menor ambigüedad posible. Esto se consigue (aunque rara vez satisfactoriamente) reduciendo el ámbito de la pregunta/problema a resolver o bien reduciendo la respuesta que el examinando tiene que hacer, y esto supone una reformulación del constructo de conocimiento, aunque puede todavía contener residuos de su forma original.

4. Pruebas imparciales

Los constructores de la prueba PISA han elegido medir las competencias en lugar de los conjuntos de conocimientos sobre la base de que estos últimos son específicos de cada país en particular, mientras que las competencias tienen características universales. Hay dos problemas con esto. En primer lugar, esas características nacionales y locales de los ámbitos del conocimiento se aplican por igual a las habilidades, competencias (habilidades expresadas como capacidades individuales) y disposiciones o actitudes (configuraciones de las capacidades individuales que se pueden expresar como posibilidades). En segundo lugar, existe una cadena inferencial más larga y compleja involucrada en la medición de las competencias de la que hay en la medición de la adquisición de conocimiento, y por lo tanto hay una mayor probabilidad de que se produzca una varianza del constructo irrelevante.

PISA ha intentado llevar a cabo la difícil tarea de construir unas pruebas desligadas del currículum; el ejemplo más notorio es el examen +11 en el Reino Unido (Torrance, 1981, para una evaluación crítica). La razón para ello es que hacer comparaciones entre los resultados de las pruebas de estudiantes de diferentes países, con diferentes currículos y con diferentes enfoques y métodos de enseñanza, requiere la selección de preguntas que no reflejen el currículum o los métodos pedagógicos nacionales. Así que estas pruebas comparativas internacionales, y esto incluye elementos que se refieren a las condiciones socio-económicas del estudiante y a datos actitudinales (como en el último conjunto de pruebas PISA centradas en las Ciencias), no son una medida de su currículum, ni de lo que se les enseña, ni tampoco una medida de lo que han aprendido en sentido formal. Esto significa que el contenido de las preguntas y su presentación son propensos a favorecer a algunos países a expensas de otros.

Las diferencias culturales pueden adoptar diferentes formas, como por ejemplo, atribuir diferentes valores y diferentes grados a dichos valores, a los elementos culturales, o la determinación de la naturaleza, calidad, fuerza probativa, relevancia del valor y alcance de las pruebas, o centrarse en las prácticas que pueden ser más familiares para la gente de algunos países y menos para la de otros. Sin embargo, más importante aún, las diferencias culturales con respecto a la selección de las cuestiones de las pruebas se refieren a la expresión del problema a resolver. Si, por ejemplo, los diferentes modismos nacionales, las diferentes formas de pensar incrustadas en las formas del lenguaje, y los diferentes valores entretejidos en la trama de discursos nacionales se ignoran, entonces la presentación de los elementos reales de las pruebas así como el rango de posibles respuestas que pueden darse pueden favorecer a los estudiantes de una nación a expensas de los estudiantes de otra.

Éste es el problema de la comparación equitativa. Y realizar una comparación justa, no puede ser sólo una cuestión de traducir las palabras que se usan, es decir, sustituir un conjunto (palabras, frases y estructuras lingüísticas) por otro, sino de hacer una transposición del ejemplo y del problema, de modo que refleje mejor su nueva base epistémica. Respaldar el concepto de una prueba internacional es la idea de una forma de conocimiento universal, es decir, desligado de la cultura, que puede adaptarse para que las diferencias superficiales entre las naciones se eliminen. No obstante, nunca es suficiente decir que una prueba simplemente evalúa las capacidades y las construcciones de conocimiento de un grupo (en este caso un grupo transnacional) de estudiantes. Lo que hace una prueba transnacional es una serie de suposiciones reduccionistas sobre las bases del conocimiento que se evalúan, lo que da lugar a una caricatura imperfecta de las bases del conocimiento nacional en cuestión.

5. Propiedades comparativas emergentes

Los resultados de PISA se expresan en tablas comparativas entre países en lugar de hacerlo en forma de calificaciones obtenidas por los participantes. Se centra en la posición más que en la puntuación, a pesar de que las mejoras significativas que hace un país entre dos momentos en el tiempo pueden ser enmascaradas por las mejoras hechas por otros países. Si a esta idea se añade la de que hay algo de incertidumbre o poca fiabilidad en las puntuaciones (es decir, marcador de error, resultados pobres por parte de los examinandos, efectos culturales sesgados, diferencias epistémicas, incapacidad para transformar el conocimiento interno en conocimiento performativo, etc.) es difícil creer que estas clasificaciones puedan proporcionar información muy útil a un país. Sin embargo, lo que tenemos aquí es un mecanismo de visualización (localizado inicialmente en el nivel transitivo, pero también penetrando y, por lo tanto asumiendo capacidad de operar en el nivel intransitivo). Este mecanismo de visualización tiene claramente aspiraciones científicas (Habermas, 1971), añadiendo además la necesidad de introducir elementos críticos y de evaluación en todas las cuentas hechas, tanto si se refieren a individuos, grupos dentro de naciones o naciones en sí mismas.

Michel Foucault (1978, p. 196) sugirió que el examen transformaba al individuo en un objeto para una rama del conocimiento:

> El caso no es ya, como en la casuística o la jurisprudencia, un conjunto de circunstancias que califican un acto y que pueden modificar la aplicación de una regla; es el individuo tal y como se le puede describir, juzgar, medir, comparar a otros y esto en su individualidad misma; y es también el individuo cuya conducta hay que encauzar o corregir, a quien hay que clasificar, normalizar, excluir, etc.

Por primera vez un individuo podía ser científicamente y objetivamente categorizado y caracterizado a través de una modalidad de poder, donde la diferencia se convierte en el factor más relevante.

Además, el instrumento (PISA) es un dispositivo performativo, en tanto que su intención no es sólo describir las habilidades/disposiciones de los niños sino promover y, por lo tanto, contribuir al diseño de las políticas nacionales. Ciertas formas de conocimiento performativo llegar a ser la norma. El instrumento para medir los niveles de conocimiento y habilidades de los niños se convierte en un instrumento para determinar lo que esos niveles de conocimiento y habilidades deberían ser y cómo debieran ser aprendidos. El mecanismo que respalda esta serie de acciones es un ejemplo de materialismo de poderes sincrónicos emergentes (Bhaskar, 2010) y, como resultado, opera como un dispositivo estandarizado/de normalización con respecto a estas materias (es decir, crea una norma) y no debe ser entendido como un dispositivo para hacer juicios justos, razonables y precisos acerca de las capacidades de los grupos de estudiantes de diferentes países. Hay una conclusión final que se hace, y es que el lugar que ocupa una nación en estos rankings se convierte en parte del lugar folclórico que la nación da a sí misma y de sí misma. Debido a que este resultado es una parte importante de la identidad de una nación, entonces tener éxito en una prueba internacional como PISA se convierte en algo aún más importante.

Referencias bibliográficas

Bhaskar, R. (1989). *Reclaiming reality*. London: Verso.

Bhaskar, R., Frank, C., Hoyer, K.-G., Naess, P. & Parker, J. (2010). *Interdisciplinarity and climate change.* London: Routledge.

Brown, A., Fleetwood, S. & Roberts, J. (2002). *Critical realism and Marxism.* London and New York: Routledge.

Curren, R. (2006). Connected learning and the foundations of psychometrics: A rejoinder. *Journal of Philosophy of Education,* 40(1), 17-29.

Foucault, M. (1979). *Discipline and punish: the birth of the prison*. New York: Vintage. [Edic. cast.

Habermas, J. (1971). *Knowledge and human interests.* Boston: Beacon Press [Edic. cast.: *Conocimiento e interés*. Madrid: Taurus, 1982].

Messick, S. (1989). Validity. In R. Linn (Ed.) *Educational measurement* (3rd Edition). American Council on Education, Washington, D.C.

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2000). *Manual for the PISA 2000 Database.* Paris: OECD.

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2001). *Knowledge and Skills for Life: First Results from PISA.* Paris: OECD [Edic. cast.: *Conocimientos y aptitudes para la vida. Primeros resultados del PISA 2000 de la OCDE*. Madrid: Santillana-MEC, 2002].

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2006). *Knowledge and Skills for Life: PISA.* Paris: OECD [Edic. cast.: *Informe PISA 2006. Competencias científicas para el mundo del mañana*. París: OCDE, 2006].

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2009). *Knowledge and Skills for Life: PISA.* Paris: OECD.

Stobart, G. (2008). *Testing times: the uses and abuses of assessment.* London: Routledge [Edic. cast.: *Tiempos de pruebas: los usos y abusos de la evaluación*. Madrid. Morata: 2010].

Torrance, H. (1981). The origins and development of mental testing in England and the United States. *British Journal of the Sociology of Education,* 2 (1), 45-59

Vigilar y castigar. El nacimiento de la prisión. Madrid: Siglo XXI. 1978]

Wiliam, D. (2007). Balancing dilemmas: traditional theories and new applications. In A. Haynes, & L. McDowell (Eds.), *Balancing Dilemmas in assessment and learning in contemporary education.* London: Taylor and Francis.

Wood, R. y Power, C. (1987). Aspects of the competence-performance distinction: Educational,psychological and measurement issues. *Journal of Curriculum Studies,* 19 (5), 409-24.