Ruiz-Primo, Maria-Araceli & Li, Min (2016). Items contextualizados de ciencias en PISA: La conexión entre las demandas cognitivas y las características de contexto de los items. *RELIEVE*, *22*(1), art. M11. DOI: http://dx.doi.org/10.7203/relieve.22.1.8280

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Items contextualizados de ciencias en PISA: la conexión entre las demandas cognitivas y las características de contexto de los items

PISA science contextualized items: the link between the cognitive demands and context characteristics of the items

Ruiz-Primo, Maria-Araceli [(1)] & Li, Min [(2)]
(1) Stanford University. (2) University of Colorado Boulder.

Resumen

El uso frecuente de contextos en items de una prueba se basa en la premisa de que agregar contextos a los items es una estrategia eficaz para comprobar si los estudiantes pueden aplicar o transferir sus conocimientos. En este trabajo, seguimos una línea de investigación que se centra en probar esta premisa. Se estudian las características de los contextos en una muestra de items de ciencias de PISA 2006 y 2009 y cómo estas características, así como el desempeño de los estudiantes, pueden estar relacionados con las demandas cognitivas de los items . El estudio se guío por dos preguntas de investigación: (1) ¿Cuáles son las demandas cognitivas de una muestra de items con contexto de la prueba PISA y cuál es el desempeño de los estudiantes en estos items? (2) ¿Están asociadas las demandas cognitivas de los items con ciertas características de los contextos, que previamente han demostrado estar relacionadas con el desempeño de los estudiantes? Se codificaron 52 items liberados y no liberados de PISA en tres dimensiones de los contextos de los items (nivel de abstracción, recursos y naturaleza del contexto) además de las demandas cognitivas de los items. Una regresión logística multinomial, con la demanda cognitiva como variable de resultado, las características contextuales como los predictores, y el porcentaje de respuestas correctas como la covariante, indicó que ciertas características del contexto están vinculados a las demandas cognitivas de los items . Por ejemplo, se encontró que los contextos en los que sólo se presentan ideas concretas están asociadas a items con bajas demandas cognitivas; en estos items es poco probable que se requiera de un conocimiento del contenido para ser respondidos. También se encontró que el tipo de recurso (por ejemplo, tablas, gráficos) se asocia con las demandas cognitivas de los items las representaciones esquemáticas parecen estar vinculadas a items de tipo procedural y no a items de tipo declarativo o esquemático. Se concluye que se necesita más investigación para comprender mejor la influencia que tienen las características de contexto en los procesos cognitivos de los estudiantes y en su desempeño

Palabras clave: PISA; items de ciencias; características contextuales de los items; demandas cognitivas; validez

Fecha de recepción
15 Mayo 2016

Fecha de aprobación
25 Junio 2016

Fecha de publicación
25 Junio 2016

Abstract

The ubiquitous use of contexts in test items is based on the premise that contextualizing items is an effective strategy to test whether students can apply or transfer their knowledge. In this paper, we continue a research agenda focusing on testing this premise. We present a study of the context characteristics in a sample of 2006 and 2009 PISA science items and how these characteristics as well as student performance may be related to the cognitive demands of the items. The study was guided by two research questions: (1) What are the cognitive demands of the sampled PISA contextualized items and what is the students' performance linked to these items? (2) Are the items' cognitive demands associated with certain characteristics of the contexts of the items that proved to be linked

Reception Date
2016 May 15

Approval Date
2016 June 25

Publication Date:
2016 June 25

Autor de contacto / Corresponding author
Ruiz-Primo, Maria-Araceli. Graduate School of Education. 485 Lasuen Mall. Stanford University, Stanford, CA (USA). aruiz@stanford.edu

to students' performance? Using 52 released and secured PISA science items, we captured information about three context dimensions of items (i.e., level of abstraction, resources, and nature of the context) and the cognitive demands of the items. A multinomial logistic regression with cognitive demand as the outcome variable, context characteristics as the predictors, and percent of correct responses as the covariant indicated that certain context characteristics are linked to the cognitive demands of items. For example, we found that items in which contexts involve only concrete ideas were associated with items with low cognitive demands; these items are unlikely to require content knowledge to be responded. We also found that the type of resource (e.g., tables, graphs) was associated with the cognitive demands of the items: schematic representations seem to be linked to items tapping procedural knowledge rather than to items tapping declarative or schematic knowledge. We concluded that further research is needed to better understand the influence that context characteristics have on the cognitive processes in which students are asked to engage and in their performance.

Keywords: PISA; science items; context characteristics of the items; cognitive demands; validity

El argumento de que toda la enseñanza, el aprendizaje y las pruebas deben realizarse en un contexto significativo (véase, por ejemplo Boaler, 1993, 1994; Greeno, 1989; Hembree, 1992; Taber, 2003; Wiggins, 1993), ha llevado a desarrollar ítems que incluyan un contexto que los estudiantes reconozcan como algo realista, interesante, relevante y útil (Haladyna, 1997). La opinión generalizada es que el uso de contextos significativos durante la instrucción (aprendizaje basado en contexto) aumenta la motivación y confianza de los estudiantes (Boaler, 1993, 1994; Kelly, 2007). Si tales contextos hacen mejorar el aprendizaje, se deduce que, al menos hipotéticamente, los contextos pueden hacer lo mismo en los ítems de una prueba. Durante años, la contextualización de ítems ha sido considerada como una estrategia eficaz para medir el pensamiento complejo (Haladyna, 1994) y para probar si los estudiantes pueden aplicar sus conocimientos (Ahmed & Pollitt, 2000, 2007; Boaler, 1994; Haladyna, 1997).

Debido a estas creencias, los ítems con contextos han llegado a utilizarse ampliamente en las pruebas de ciencia. En los Estados Unidos, el 70% de los ítems liberados de 8^{avo} grado de la Prueba Nacional de Progresivo Educativo (National Assessment of Educational Progress, NAEP) y el 71% de la prueba de 4º grado son ítems con contextos (Wang & Li, 2014). De forma similar, el 78% de los ítems liberados de ciencias de TIMSS 2011 (*Trends in International Mathematics and Science Study)* para el 8^{avo} grado también tienen contexto (Wang & Li, 2014). Ciertamente, los ítems con contexto son el principal tipo de ítem en los programas de evaluación a gran escala, tales como el Programa para la Evaluación Internacional de Estudiantes (PISA) (Ruiz-Primo & Li, 2012, 2015; Wang & Li, 2014).

A pesar del uso generalizado de los contextos en los ítems, algunos (Ahmed & Pollitt, 2001; Gerofsky, 1996; Wiliam, 1997) han puesto en duda las suposiciones que les subyacen debido a la falta de conocimiento sobre este tipo de ítems (Haladyna, 1994; pero véase Ahmed & Pollitt, 1999, 2000, 2007; Mevarech & Stern, 1997). Por ejemplo, ¿sabemos qué tipos de contextos se utilizan en ítems de ciencia?, ¿sabemos cómo el grado de contextualización puede facilitar o impedir el desempeño de los estudiantes con capacidades diferentes?, ¿sabemos cómo el grado de contextualización puede afectar el constructo o cómo interactúa con habilidades importantes, tales como el dominio en lectura? En otras palabras, ¿podemos saber qué características o tipos de contextos tienen el potencial de ser una *fuente inválida de dificultad* (Fisher-Hoch & Hughes, 1996; Hembree, 1992; Mevarech & Stern, 1997)? Una revisión de las investigaciones actuales indican que sabemos muy poco acerca de qué tipos de contextos son útiles o perjudiciales para evaluar con precisión el constructo en cuestión. Como Ahmed y Pollitt (2000) indicaron, "Un buen contexto nos permite medir la capacidad del estudiante para aplicar su conocimiento, pero un mal contexto

puede impedir que midamos nada en absoluto" (p. 1). Si los ítems contextualizados contienen algo que los hace funcionar en una dirección diferente de lo esperado, ninguna buena aplicación, buena enseñanza, o juicio racional pueden compensar la inexactitud del ítem (Ahmed & Pollitt, 2007). Si el contexto puede afectar la forma en que se interpretan los ítems y, por lo tanto, la forma de responder a ellos, entonces estamos ante un problema de validez que debe investigarse de forma deliberada. Aprender más sobre el efecto de los contextos y de sus características en el desempeño de los estudiantes es necesario.

Hemos iniciado (Ruiz-Primo & Li, 2012, 2015; Ruiz-Primo, Li, & Minstrell, 2014) un programa de investigación que se centra en aprender más acerca de los contextos en los ítems. Este tipo de investigación tiene el potencial de ofrecer un marco sólido para la caracterización de contextos de ítems y el estudio de sus efectos sobre el desempeño de los estudiantes. Como parte de este programa de investigación, en este estudio nos centramos en cómo las características de los contextos están relacionadas con las demandas cognitivas de ítems de PISA. A diferencia de la mayoría de las pruebas de desempeño (por ejemplo, la NAEP), algunos ítems de PISA tienen hasta *tres niveles de contextos* y por lo tanto constituyen una importante fuente de información para explorar estas ideas. Además, los contextos en los ítems de PISA contienen mucha más información que sus equivalentes en otras pruebas de desempeño.

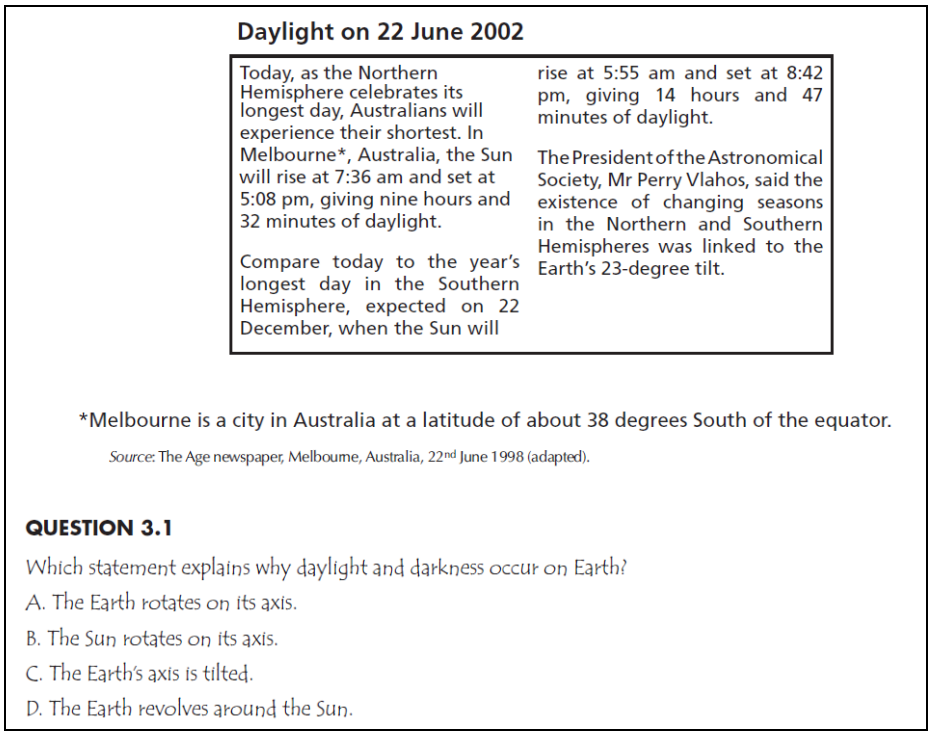
El estudio presentado en este artículo se centra en las siguientes preguntas de investigación (1) *¿Cuáles son las demandas cognitivas de una muestra de ítems con contexto de la prueba PISA y cuál es el desempeño de los estudiantes en estos ítems?* (2) *¿Están asociadas las demandas cognitivas de los ítems con ciertas características de los contextos, que previamente han demostrado estar relacionadas con el desempeño de los estudiantes?* Para obtener información sobre las características de los contextos, se utilizó la aproximación de *perfiles de contexto*, para obtener información acerca de los ítems, se utilizó un *análisis lógico*. El estudio se enfoca en la codificación de una muestra de los ítems de ciencias de PISA 2006 y 2009 que incluye ítems no-liberados y liberados. El estudio se centra en ciertas características de contexto que anteriormente se encontró que estaban relacionados con el desempeño de los estudiantes. Se explora cómo estas características se relacionan con las demandas cognitivas de los ítems. El estudio ofrece evidencia exploratoria sobre la posible relación entre las características del contexto, las demandas cognitivas de los ítems y el desempeño de los evaluados. El artículo contribuye al campo en, por lo menos, dos formas: (1) permite mejorar nuestra comprensión de los ítems con contexto y de su impacto en el desempeño de los estudiantes; y (2) contribuye a un mejor entendimiento del desarrollo de ítems específicamente en ciencia, pero potencialmente en otras disciplinas de STEM. En los siguientes párrafos, primero se presenta información acerca de los ítems con contexto. Después se describe la aproximación de codificación utilizada para analizar los ítems de contexto, seguido de la descripción del estudio.

Items con contexto

En artículos anteriores, (Ruiz-Primo & Li, 2012, 2015) se define el *contexto de un ítem* como un componente de información adicional que precede o sigue a la pregunta del ítem, por ejemplo, una descripción de un experimento de laboratorio, la descripción de un fenómeno natural, o un problema práctico[i]. Términos comúnmente usados para referirse a ítems con contexto incluyen escenario, viñeta o historia. Los ítems con contextos han sido llamados *ítems contextualizados* (por ejemplo, Ahmed & Pollitt, 2007), *ítems con escenarios* (por ejemplo, Fulkerson, Nichols, Haynie, & Mislevy, 2009; McMartin, McKenna, & Youssefi, 2000), *ítems con narrativa* (por ejemplo, Terry, 1980), *ítems realistas* (por ejemplo, Cooper & Dunne, 2000; Hembree, 1992), o *ítems dependientes de un contenido*

(por ejemplo, Haladyna, Downing, & Rodríguez, 2002)

Se ha argumentado que agregar contextos a los ítems los haces más concretos y menos difíciles (Ahmed & Pollitt, 2000, 2007). La figura 1 muestra un ejemplo de un ítem contextualizado[ii].



Daylight on 22 June 2002

Today, as the Northern Hemisphere celebrates its longest day, Australians will experience their shortest. In Melbourne*, Australia, the Sun will rise at 7:36 am and set at 5:08 pm, giving nine hours and 32 minutes of daylight.

Compare today to the year's longest day in the Southern Hemisphere, expected on 22 December, when the Sun will rise at 5:55 am and set at 8:42 pm, giving 14 hours and 47 minutes of daylight.

The President of the Astronomical Society, Mr Perry Vlahos, said the existence of changing seasons in the Northern and Southern Hemispheres was linked to the Earth's 23-degree tilt.

*Melbourne is a city in Australia at a latitude of about 38 degrees South of the equator.

Source: The Age newspaper, Melbourne, Australia, 22nd June 1998 (adapted).

QUESTION 3.1

Which statement explains why daylight and darkness occur on Earth?

A. The Earth rotates on its axis.

B. The Sun rotates on its axis.

C. The Earth's axis is tilted.

D. The Earth revolves around the Sun.

Figura 1. Ejemplo de un ítem contextualizado

Desafortunadamente, los ítems contextualizados se construyen principalmente basados ya sea en la sabiduría convencional o en reglas que se han establecido para desarrollar ítems que son no-contextualizados. Por lo tanto, los contextos usados pueden interferir con el constructo que es el blanco del ítem. Los contextos pueden potencialmente evocar en los estudiantes los conocimientos pertinentes, mejorando la comprensión del contenido y/o procesos que se requieren para responder a lo que los ítems están pidiendo. Sin embargo, los contextos también pueden evocar información irrelevante y confundir a los estudiantes de manera que pueden dar lugar a respuestas incorrectas debido a la mala interpretación de la tarea requerida, lo cual puede conducir a conclusiones inexactas sobre el aprendizaje del estudiante (Ahmed & Pollitt, 2001; Leighton & Gokiert, 2005).

En la literatura del desarrollo de ítems, el uso del contexto se ha asociado con ítems que se presentan en grupo, también llamados paquetes de ítems, clusters de ítems, o testlets Usamos el término testlet (Wainer & Kiely, 1987) para referirnos al grupo de ítems (conjunto o cluster) relacionados con un solo contexto. Un testlet tiene uno o más párrafos introductorios que establecen el problema o describen un escenario, seguido por un conjunto de ítems que se pueden presentar en cualquier formato. Un testlet puede contener de cinco a diez ítems (Haladyna, 1994). Sin embargo, también son comunes los ítems individuales con su propio contexto. En PISA, los testlets pueden tener hasta tres niveles de contextos: (a) *contexto general* que introduce toda el testlet; (b) el *contexto subtestlet* que introduce un subconjunto de, por ejemplo, dos ítems del testlet; y (c) *contexto del ítem* que se presenta a nivel de ítem individual.

Aproximación para analizar los contextos de los ítems

La aproximación que propusimos anteriormente (Ruiz-Primo & Li, 2012) para describir el *perfil de un contexto* utiliza el análisis lógico basado en elementos de juicio como evidencia. La aproximación propone dimensiones y aspectos para guiar el juicio de las características del ítem, del contexto del ítem, y de la relación del contexto con la pregunta del ítem[iii].

La codificación de las *características del ítem* se centra en cinco aspectos del ítem: formato, diseño, nivel del contexto, demandas cognitivas, y la ubicación del ítem en el testlet. El *perfil del contexto (*caracterización del contexto*)* se centra en cuatro dimensiones principales: la complejidad, la abstracción, los recursos y la naturaleza. Cada una de estas dimensiones incluye diferentes aspectos que describen de manera más detalladas las características de los contextos de los ítems. Por ejemplo, la dimensión de complejidad incluye características de extensión del contexto, el tipo de vocabulario, el número de las ideas principales, y la carga de lectura. Estas dimensiones permiten la codificación de características del contexto sin tener en cuenta la pregunta del ítem. A diferencia de los dos primeros conjuntos de códigos (de ítem y del contexto), el tercer conjunto de códigos requiere la codificación de contexto en relación con la pregunta del ítem. Estos códigos se centran en juzgar la función y las propiedades del contexto en relación con cada una de las preguntas de los ítems en un testlet. La Tabla 1 proporciona una breve descripción de las dimensiones utilizadas en este estudio. Una explicación detallada de los códigos se proporciona en Ruiz-Primo y Li (2012, 2015).

Cuando se utilizan testlets, es importante juzgar el papel del contexto general (o el subcontexto y/o contexto del ítem) con respecto a cada uno de los ítems que forman un testlet. La aproximación para describir el perfil de un contexto, a cualquier nivel, usa el ítem como unidad de análisis porque el contexto debe ser analizado en relación con el ítem al que está vinculado. Esto es porque finalmente la información proporcionada en los contextos necesita ser procesada o filtrada por los examinados para generar las respuestas a cada ítem individual. Por lo tanto, la aproximación propuesta capta las características de los contextos en cada nivel identificado en un testlet dado.

En trabajos anteriores que utilizan esta aproximación de codificación, se ha reconocido la dificultad y la subjetividad involucrada en describir los perfiles de los contextos. La aproximación debe ser pensada como una estrategia exploratoria para identificar aquellas características del contexto que merecen más atención como posibles factores que pueden estar relacionados con el desempeño de los estudiantes. Creemos que esto es un paso necesario que no se ha hecho anteriormente. La aproximación de codificación debe ser vista “como el primer paso para operacionalizar estas dimensiones” (Ruiz-Primo & Li, 2015, p. 14). Se necesitan estudios que utilicen protocolos de verbalización (talk-aloud) con estudiantes, entrevistas y diseño experimental de ítems para mejorar y refinar el enfoque de codificación.

Tabla 1. *Ejemplos de las Dimensiones y los Aspectos Usados en Este Estudio para Capturar las Característica de los Contextos*

Dimensiones y Aspectos	**Descripción**	**Ejemplos de Códigos y Criterios de Codificación**
	Características del ítem. Juzgadas con base en…	
Demanda cognitiva	Juicio del tipo de conocimiento tratado en el ítem	• Razonamiento general • Declarativo • Procedimental • Esquemático
	Características del Contexto. Juzgadas con base en…	
Nivel de Abstracción		
Nivel de abstracción de las ideas principales	Presencia de ideas abstractas en el contexto.	• Todas las ideas concretas • Algunas ideas abstractas otras concretas • Todas las ideas abstractas
Uso de referentes concretos	Si las ideas abstractas identificadas en el contexto tienen o no un referente concreto tales como objetos o eventos.	• Hay vinculación entre todas las ideas abstractas y referentes concretos • Hay vinculación entre algunas ideas abstractas y referentes concretos • No hay ninguna vinculación en las ideas abstractas y referentes concretos
Recursos		
Tipo de recurso	Cualquier otro material en el contexto que sea no verbal	• Diagrama • Gráfico • Mapa
Naturaleza del Contexto		
Escenario	El tipo de escenario que se describe en el contexto.	• Actividad de la vida diaria en el hogar • Trabajo profesional o lugar de trabajo • Información científica o Informativo

Contexto del estudio

El estudio presentado en este documento utiliza 52 ítems de ciencias no-liberados y liberados de PISA que fueron utilizados en un estudio exploratorio previo (Ruiz-Primo y Li, 2015)[iv]. En este estudio, se analizan las demandas cognitivas de estos ítems y cómo estas demandas están vinculadas a las características del contexto y al desempeño de los estudiantes. El análisis de las dimensiones cognitivas de los ítems se realizó con un marco conceptual que se ha usado en el pasado y cuyas categorías se han probado empíricamente en el contexto de las pruebas de ciencia (Li, 2002; Li & Shavelson, 2001; Li, Ruiz-Primo, y Shavelson, 2006; Yin, 2005). Se optó por centrarse únicamente en las características del contexto general (nivel más alto) por dos razones: (1) cualquier análisis de ítems con múltiples niveles de contextos reducirían el tamaño de la muestra de los ítems analizados de 52 a 40, lo cual reduce aún más el poder estadístico para evaluar adecuadamente la asociación entre la demanda cognitiva de los ítems y el contexto, y (2) porque creemos que los resultados tienen un mayor potencial para influir en las decisiones de la construcción de ítems y de sus contextos.

El Marco Conceptual de PISA. El marco conceptual actual de ciencias de PISA (OECD, 2015, por sus siglas en inglés - Organisation for Economic Cooperation and Development), difiere del marco conceptual utilizado en las administraciones del 2006 y 2009. Para los propósitos de este estudio, se describen el marco utilizado para los ítems aplicados en 2006 y 2009 que se estudian en este estudio[v]. En general, el marco conceptual de ciencias de PISA se centra en un conocimiento general de la ciencia (scientific literacy): "¿Qué es importante que los ciudadanos sepan, valoren y sean capaces de hacer en situaciones que involucren conocimiento científico y tecnología?" (OECD, 2009, p 126). El marco conceptual se divide en tres componentes: conocimientos, competencias y actitudes. Los ítems utilizados en este estudio tocan las siguientes dimensiones del marco: (1)

áreas disciplinarias (sistemas físicos, sistemas vivos, sistemas terrestres y espaciales, y sistemas tecnológicos), (2) las prácticas de la ciencia (por ejemplo, la investigación científica y las explicaciones científicas), y (3) competencias (es decir, identificar cuestiones científicas, explicar fenómenos científicamente y utilizar la evidencia científica). En este estudio, nos centramos en una de las dimensiones de los ítems de PISA, las competencias debido su importancia en el constructo de interés, el conocimiento general de la ciencia.

La importancia del conocimiento general de la ciencia en el marco conceptual de PISA hace que el uso del contexto en el ítem sea un aspecto crítico en la prueba. Estos contextos se organizan en dos dimensiones cruzadas en el marco conceptual: tipos de contexto – personal (uno mismo, familia y compañeros), social (la comunidad), y global (la vida en todo el mundo), y (2) áreas de aplicación - salud, recursos naturales, medio ambiente, riesgos, y las fronteras de la ciencia y la tecnología.

Características del contexto y desempeño de los estudiantes. En un estudio exploratorio previo (Ruiz-Primo & Li, 2012, 2015), se comprobó la importancia de las características de los contexto de los ítems de PISA. Se encontró que en el caso de los ítems de PISA, los contextos deben ser analizados por nivel (general, sub-testlet, ítem). Este resultado fue bastante consistente a través de muchas de las dimensiones y aspectos codificados. Ya sea que un contexto se encuentre en el nivel general o en el específico, sus características están conectadas con diferencias notables en el desempeño de los estudiantes. Por ejemplo, un contexto general puede no ser necesario para responder a un ítem de un testlet, pero el contexto a nivel de ítem puede estar directamente relacionado con la pregunta del ítem.

También aprendimos que el rol que los contextos juegan varían según el nivel de contexto y que ítems con contextos con diferentes roles tienden a estar correlacionados de manera diferente con el desempeño de los estudiantes. Para los contextos generales, el desempeño del estudiante fue estadísticamente menor en los ítems con los contextos descriptivos (43.61 por ciento de respuestas correctas [PRC]) que en los contextos con escenarios (59.28 PRC) con una D de Cohen = 1,01; esta diferencia fue insignificante para contextos específicos, aunque también con la misma tendencia.

Los *recursos* y el *nivel de abstracción* de los contextos son dos dimensiones que vale la pena considerar como fuentes fundamentales de dificultad de los ítems. En contextos generales (el nivel más alto), el número de recursos (ya sean dibujos, gráficos, mapas o tablas) y la relevancia del recurso para la pregunta que se hace en el ítem están significativamente correlacionados con el nivel de dificultad de los ítems pero en direcciones opuestas; es decir, un ítem es más difícil cuando se presentan más recursos en su contexto, pero es más fácil si los recursos son pertinentes para la pregunta planteada en el ítem. Este resultado fue el mismo en el análisis por género. El nivel de abstracción también afecta a la dificultad del ítem, al menos para los contextos específicos; los contextos de ítems con ideas abstractas, sobre todo que carecen de referentes a objetos o eventos concretos, se asociaron con un índice de dificultad medio bajo. Las otras dos dimensiones, complejidad del contexto y conectividad del contexto, requieren de un mayor desarrollo, elaboración, reexaminación, y perfeccionamiento de los aspectos (códigos) para poder entender su relación con el desempeño de los estudiantes.

La siguiente sección describe el estudio exploratorio utilizando los ítems de PISA y explica cómo se aproximó el estudio de las características del contexto y de sus relaciones con el desempeño de los estudiantes.

Metodología

Ítems. La muestra original incluyó 52 ítems organizados en 17 testlets de ciencias de PISA 2006 y PISA 2009, incluyendo 39 ítems no-liberados y 13 ítems liberados. La muestra fue de 4 testlets administrados en PISA 2006, 11 administrados tanto en PISA 2006 como en

PISA 2009, y 2 administrados en PISA 2009 y en administraciones anteriores de PISA. El 65% de los ítems requerían que los estudiantes seleccionaran una opción, y el 23% requerían que los estudiantes construyeran su respuesta. Cuarenta y ocho por ciento de los ítems fueron clasificados por PISA como "explicar fenómenos científicamente", el 36% como "usar evidencia científica", y el resto como "identificar cuestiones científicas."

Para el desempeño de los estudiantes, se realizó el análisis con siete países de habla inglesa: Australia, Canadá, Inglaterra, Hong Kong-China, Singapur, Nueva Zelanda y Estados Unidos. El examen de los patrones de datos a través de grupos de estudiantes nos permitió examinar cómo las características de los estudiantes interactúan con las características de contexto. Por simplicidad, sólo se presentan y analizan los resultados de la muestra de los Estados Unidos, pero se describen las diferencias observadas con otros países al final de la sección de resultados.

Codificación de las demandas cognitivas. El análisis de las demandas cognitivas de los ítems se realizó durante la codificación de los ítems. La aproximación utilizada para codificar las demandas cognitivas se basa en tres tipos interdependientes de conocimiento. En esta aproximación se supone que el tipo de conocimiento se encuentra en un continuo desde lo concreto a lo abstracto, de bits de información desconectados a altos niveles de organización del conocimiento. Es de suponer que el desempeño en los niveles más altos de conocimiento general de la ciencia debería estar vinculado a ciertos tipos de conocimiento. Las demandas cognitivas fueron codificados utilizando tres tipos de códigos para tres tipos de conocimiento:

1. **Conocimiento declarativo - Saber qué.** Este tipo de conocimiento se extiende desde elementos de contenido discreto y aislado, tales como terminología o hechos, a formas de conocimiento más organizado, tales como declaraciones, definiciones y conocimiento de clasificaciones, categorías y principios (por ejemplo, la masa es una propiedad de un objeto).
2. **Conocimiento procedimental - Saber cómo.** Este tipo de conocimiento implica el conocimiento de habilidades, algoritmos, técnicas y métodos. Por lo general, este conocimiento toma la forma de reglas de producción "si esto-entonces lo otro" o de una secuencia de pasos que son el resultado del consenso, acuerdo, o normas disciplinarias (Anderson et al., 2001). Este tipo de conocimiento se ve como un conocimiento de habilidades (“skills”, Royer, Ciscero, & Carlo, 1993) que oscila entre procedimientos motores (por ejemplo, medir la masa de un objeto con una balanza) y procedimientos más complejos (por ejemplo, la aplicación de una secuencia de varios pasos para encontrar la densidad de un objeto cuando no se conoce ni la masa ni el volumen). El conocimiento procedimental puede automatizarse mediante la práctica, lo que permite que pueda recuperarse y utilizarse sin una atención de la memoria deliberada.
3. **Conocimiento esquemático - Saber por qué.** Este tipo de conocimiento incluye información que está interconectada con cuerpos de conocimiento más organizados, tales como esquemas, modelos mentales, o "teorías" (implícitas o explícitas). Este tipo de conocimiento permite a los estudiantes definir problemas nuevos y resolverlos de manera novedosa, explicar fenómenos, o predecir resultados. El conocimiento esquemático y el conocimiento procedimental se combinan en el proceso de razonamiento para diseñar experimentos o para resolver problemas matemáticos.

Se utilizaron otros dos códigos para capturar características de los ítems relacionados con las demandas cognitivas:

4. **Razonamiento general** - Este código se utilizó para dar cuenta de aquellos ítems que podían responderse, incluso correctamente, sin o con un mínimo de conocimiento. Estos ítems requieren únicamente una cuidadosa lectura del ítem

o un razonamiento lógico para encontrar la respuesta correcta.

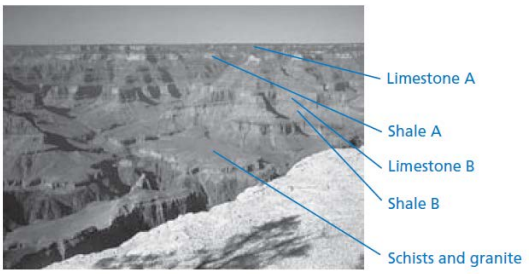
5. **Difícil de saber**. Este código se utilizó para dar cuenta de aquellos ítems en los que era difícil saber exactamente qué tipo de conocimiento se requería para responderlo, pero seguro no podría ser respondido con un razonamiento general.

Los códigos son mutuamente excluyentes; sólo un código por ítem puede ser seleccionado. La Figura 2 presenta ejemplos de cuatro ítems y los códigos de demanda cognitiva.

(a) Ejemplo de ítem PISA codificado como Razonamiento General

The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.



QUESTION 7.3

About five million people visit the Grand Canyon national park every year. There is concern about the damage that is being caused to the park by so many visitors.

Can the following questions be answered by scientific investigation? Circle "Yes" or "No" for each question.

Can this question be answered by scientific investigation?	Yes or No?
How much erosion is caused by use of the walking tracks?	Yes / No
Is the park area as beautiful as it was 100 years ago?	Yes / No

(b) Ejemplo de ítem PISA codificado como Conocimiento Declarativo

Read the following section of an article about the ozone layer.

The atmosphere is an ocean of air and a precious natural resource for sustaining life on the Earth. Unfortunately, human activities based on national/personal interests are causing harm to this common resource, notably by depleting the fragile ozone layer, which acts as a protective shield for life on the Earth.

Ozone molecules consist of three oxygen atoms, as opposed to oxygen molecules which consist of two oxygen atoms. Ozone molecules are exceedingly rare: fewer than ten in every million molecules of air. However, for nearly a billion years, their presence in the atmosphere has played a vital role in safeguarding life on Earth. Depending on where it is located, ozone can either protect or harm life on Earth. The ozone in the troposphere (up to 10 kilometres above the Earth's surface) is "bad" ozone which can damage lung tissues and plants. But about 90 percent of ozone found in the stratosphere (between 10 and 40 kilometres above the Earth's surface) is "good" ozone which plays a beneficial role by absorbing dangerous ultraviolet (UV-B) radiation from the Sun.

Without this beneficial ozone layer, humans would be more susceptible to certain diseases due to the increased incidence of ultra-violet rays from the Sun. In the last decades the amount of ozone has decreased. In 1974 it was hypothesised that chlorofluorocarbons (CFCs) could be a cause for this. Until 1987, scientific assessment of the cause-effect relationship was not convincing enough to implicate CFCs. However, in September 1987, diplomats from around the world met in Montreal (Canada) and agreed to set sharp limits to the use of CFCs.

Source: Connect, UNESCO International Science, Technology & Environmental Education Newsletter, Section from an article entitled 'The Chemistry of Atmospheric policy', Vol. XXII, No. 2, 1997 (spelling adapted)

QUESTION 2.3

Lines 14 and 15 state: "Without this beneficial ozone layer, humans would be more susceptible diseases due to the increased incidence of ultra-violet rays from the Sun."

Name one of these specific diseases.

.....

(c) Ejemplo de ítem PISA codificado como Conocimiento Procedimental

A team of British scientists is developing "intelligent" clothes that will give disabled children the power of "speech". Children wearing waistcoats made of a unique electrotextile, linked to a speech synthesiser, will be able to make themselves understood simply by tapping on the touch-sensitive material.

The material is made up of normal cloth and an ingenious mesh of carbon-impregnated fibres that can conduct electricity. When pressure is applied to the fabric, the pattern of signals that passes through the conducting fibres is altered and a computer chip can work out where the cloth has been touched. It then can trigger whatever electronic device is attached to it, which could be no bigger than two boxes of matches.

"The smart bit is in how we weave the fabric and how we send signals through it – and we can weave it into existing fabric designs so you cannot see it's in there," says one of the scientists.

Without being damaged, the material can be washed, wrapped around objects or scrunched up. The scientist also claims it can be mass-produced cheaply.

Source: Steve Farrer, 'Interactive fabric promises a material gift of the garb', The Australian, 10 August 1998.

QUESTION 6.2

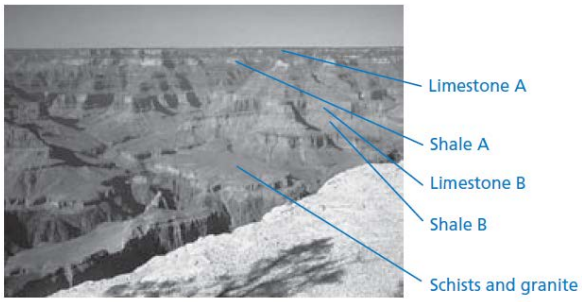
Which piece of laboratory equipment would be among the equipment you would need to check that the fabric is conducting electricity?

A. Voltmeter
B. Light box
C. Micrometer
D. Sound meter

(d) Ejemplo de ítem PISA codificado como Conocimiento Esquemático

The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.



QUESTION 7.1

The temperature in the Grand Canyon ranges from below 0 °C to over 40 °C. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks?

A. Freezing water dissolves warm rocks.
B. Water cements rocks together.
C. Ice smoothes the surface of rocks.
D. Freezing water expands in the rock cracks.

Figura 2. Ejemplos de items de PISA con diferentes demandas cognitivas

Los códigos de demanda cognitiva fueron juzgados con base en un análisis lógico de lo que los estudiantes necesitan usar de su conocimiento después de examinar tanto el contexto como la tarea planteada en la pregunta del ítem. Como muestra, el ejemplo en la Figura 2a se codificó como razonamiento general porque la redacción de las dos preguntas ofrece a los estudiantes múltiples pistas para llegar a las respuestas correctas sin necesidad de un conocimiento científico (por ejemplo, la primera pregunta que indica una relación causal y utiliza un término científico, mientras que la segunda pregunta sobre todo implica un juicio subjetivo). El ejemplo en la Figura 2b requiere que los estudiantes recuerden piezas de conocimiento sobre hechos relacionados con las enfermedades asociadas posiblemente a la exposición a los rayos ultra-violeta, por lo que se codificó como conocimiento declarativo. El ejemplo de la Figura 2c se codificó como un ítem de conocimiento procedimental porque pide a los estudiantes identificar el equipo de laboratorio necesario para determinar la conductividad eléctrica de los materiales. El ejemplo en la figura 2d se codificó como un conocimiento esquemático ya que los estudiantes necesitan explicar el fenómeno descrito mediante la aplicación de su modelo mental de cambio de estado, temperatura y volumen de agua.

Codificación para identificar los perfiles de los contextos. Cuatro testlets de ítems liberados de 12 ítems fueron utilizados para entrenar a los codificadores y refinar la aproximación de codificación. Debido al refinamiento del sistema de codificación, no fue posible probar formalmente la consistencia entre codificadores antes de que los ítems fueran codificados en el National Center for Education Statistics (NCES) en los EE.UU. La codificación se llevó a cabo durante 12 horas en 1,5 días, el tiempo máximo otorgado por NCES para codificar los ítems. Los ítems que se codificaron fueron presentados por NCES en carpetas por año. Para asegurar un acuerdo de codificación, los dos codificadores llevaron a cabo la codificación por consenso con el 50% de los ítems, ocho testlets de 26 ítems. Un único codificador codificó los testlets restantes.

Códigos de contexto seleccionados. Se eligió estudiar cuatro aspectos de los contextos, tres de los cuales resultaron estar significativamente relacionados con el desempeño del estudiante, dos de la dimensión *nivel de abstracción*, y uno de la dimensión de *recursos*. También se incluyó una cuarta dimensión, la *naturaleza del escenario* del contexto porque se consideró que esta característica contextual era factor importante a tomarse en cuenta durante la fase de redacción de los ítems. Además, se estudiaron las *competencias* de PISA, que fueron tomadas directamente de la descripción del ítem tal como se definió por los diseñadores de los ítems de acuerdo con los documentos de especificaciones de ítems de PISA.

Es importante mencionar que para aquellos códigos en los cuales solamente se encontraron dos o menos ítems de PISA, se recodificaron los valores uniendo categorías. Por ejemplo, la dimensión naturaleza del escenario que originalmente tenía cinco códigos (por ejemplo, clase o escuela, casa-vida diaria), se recodificó en dos códigos: *escenario personal* que combina tanto la vida diaria, roles, y el entorno escolar (12 y 4 ítems, respectivamente) y *escenario no-personal* que combina escenario profesional/lugar de trabajo y textos científicos/de información (34 y 2 ítems, respectivamente). Al final, se recodificaron todos los aspectos de los contextos en las tres dimensiones: abstracción, los recursos y la naturaleza del contexto. Esta recodificación permitió evitar celdas con menos de dos observaciones y así cumplir con los supuestos necesarios para llevar a cabo el análisis estadístico seleccionado. No se recodificaron las demandas cognitivas o las competencias.

Todos los códigos utilizados en el análisis son variables nominales. También incluimos el PRC, una variable continua. Un PRC más alto indica que los ítems son relativamente más fáciles o los estudiantes se desempeñaron mejor en comparación con los valores más

bajos de PRC. La Tabla 2 proporciona un resumen de las variables utilizadas en este estudio organizada por variables de resultado, predictores y covariables.

Tabla 2. *Lista de Variables Incluidas en el Análisis de Datos y sus Valores*

	Variables Usadas	Descripción	Valores
Resultado	Demanda cognitiva	Juicio del tipo de conocimiento tratado en el ítem.	• Razonamiento general • Declarativo • Procedimental • Esquemático
Predictores	**Nivel de Abstracción del Contexto**		
	Abstracción de las ideas principales del contexto[a]	Presencia de ideas abstractas	• Todas las ideas son concretas • Al menos algunas ideas abstractas con algunas ideas concretas o con ninguna idea concreta
	Facilidad de conexión entre ideas[a]	Si las diferentes ideas presentadas en el contexto pueden vincularse facialmente	• Fáciles de conectar • Algunas de las ideas abstractas son difíciles de conectar para el estudiante
	Recursos en el Contexto		
	Tipo de recurso involucrado[a]	Cualquier otro material en el contexto que sea no verbal	• Toda la información es textual • Hay representación visual • Hay representación esquemática (por ejemplo, tablas con datos o gráficos)
	Naturaleza del Contexto		
	Tipo de escenario[a]	El tipo de escenario que se describe en el contexto.	• Contexto personal • Contexto no-personal (es decir, contexto social o global clasificado por PISA)
	Estructura PISA		
	Competencias de PISA	Categorías de la ejecución de los estudiantes de acuerdo con el marco conceptual de PISA	• Explicar fenómenos científicamente • Identificar cuestiones científicas • Utilizar la evidencia científica
Covariable	Porcentaje de respuestas correctas (PRC)	Porcentaje de examinados que contestaron correctamente a un ítem	Rango entre 0 y 100

[a] Los valores fueron recodificados para colapsar niveles, de manera tal que todos los niveles de la variable tuvieran un número razonable de ítems.

Con el fin de evaluar cómo la demanda cognitiva de los ítems está vinculada a las características del contexto y desempeño de los estudiantes, primero se obtuvo una tabla cruzada con información descriptiva de los ítems por demandas cognitivas. Después se llevó a cabo una regresión logística multinomial con la demanda cognitiva como variable de resultado, las características contextuales como los predictores y el PRC como covariable. Como una extensión de la regresión logística, la regresión logística multinomial es adecuada para variables nominales. Este método permitió determinar la relación entre las variables categóricas mediante la predicción de las probabilidades de las diferentes demandas cognitivas de los ítems, dadas las características de sus contexto y el PRC como covariable. En otras palabras, el análisis de regresión estima la probabilidad de la demanda cognitiva de un ítem de acuerdo con un perfil dado de características del contexto y la dificultad del ítem.

Resultados

Los análisis se llevaron a cabo tanto con el estadístico delta (el parámetro de dificultad estimado de acuerdo con el enfoque de la Teoría de Respuesta al Ítem; véase PISA 2006 informe técnico) como con el porcentaje de respuestas correctas (PRC), que equivale a la tasa de estudiantes que respondieron correctamente al ítem (ver PISA 2006 informe técnico). La correlación entre los dos

indicadores, el estadístico delta y el porcentaje de respuestas correctas, fue muy alta (r = -0,994), lo que sugiere que los dos indicadores ordenan de manera muy similar los ítems en términos de dificultad. Para responder a las preguntas de investigación, se reporta el porcentaje de respuestas correctas porque la interpretación de este indicador es más sencilla.

La primera pregunta de investigación, *¿cuáles son las demandas cognitivas de una muestra de ítems con contexto de la prueba PISA y cuál es el desempeño de los estudiantes en estos ítems?*, se responde presentando el porcentaje de ítems por tipo de conocimiento. Esta información se presenta en la Tabla 3. De los 52 ítems codificados, sólo uno resultó ser demasiado ambiguo para determinar su demanda cognitiva en términos del conocimiento subyacente requerido. El porcentaje más alto de ítems se encontró para el conocimiento procedimental y el más bajo para el conocimiento declarativo. Es interesante observar que se encontraron porcentajes similares para los ítems que requieren un conocimiento esquemático y un razonamiento general para responderlos. De acuerdo con los resultados del análisis unidireccional de varianza, no se observa ninguna diferencia significativa en el porcentaje de respuestas correctas por tipo de demanda cognitiva de los ítems (F (3, 47) = 0,418, p = 0,741) .

Tabla 3. *Ítem de PISA No Liberados y Liberados y PRC por Tipo de Conocimiento (N = 52)*

Tipo de Conocimiento	Frecuencia	Porcentaje	Promedio PCR
Declarativo	8	15.4	50.97
Procedimental	16	30.8	55.96
Esquemático	14	26.9	50.67
Razonamiento General	13	25.0	48.98
Difícil de Saber[a]	1	1.9	-

a. El ítem codificado como “difícil de saber” no se incluyó en ningún análisis.

Para abordar la segunda pregunta, *¿están asociadas las demandas cognitivas de los ítems con ciertas características de los contextos, que previamente han demostrado estar relacionadas con el desempeño de los estudiantes?*, nos centramos en las cuatro características del contexto y las competencias de PISA. Primero se presentan los estadísticos descriptivos sobre los predictores desglosados por demanda cognitiva. Después se describen los resultados de la regresión logística multinomial en el que se evaluó el PRC como covariable y las características de los contextos de los ítems se asocian con las demandas cognitivas de los ítems.

Perfiles de los contextos de los ítems de PISA de acuerdo con las demandas cognitivas

Nivel de abstracción de los contextos. Para esta dimensión, nos centramos en la abstracción de las ideas principales y el uso de referencia(s) concreta(s) en el(los) contexto(s) de un ítem. En nuestro estudio anterior se encontró que ambas características se asociaban significativamente con los patrones de desempeño de los estudiantes. Los ítems contextuales con ideas abstractas, sobre todo aquellos que carecían de referentes concretos como objetos o eventos, se asociaron con un índice de dificultad media baja. La hipótesis que se planteo fue que si las ideas presentadas en el contexto son demasiado abstractas y su vinculación con objetos concretos o eventos no se aprecia fácilmente, podrían interferir potencialmente en la comprensión de los estudiantes de la información que se presenta en el contexto del ítem y, por lo tanto, con la comprensión de la pregunta planteada en el ítem y de su respuesta. La Tabla 4 proporciona información sobre la distribución conjunta desglosada por tipo de conocimiento y abstracción del contexto. Se encuentra que la

mayoría de los contextos que presentan ideas concretas que implican referencias a objetos concretos o eventos se encuentran en ítems que evalúan razonamiento general o conocimiento procedimental. Es interesante observar que no hay mucha diferencia en los perfiles entre los ítems de conocimiento procedimental y esquemático.

Tabla 4. *Ítems Clasificados de Acuerdo con el Tipo de Conocimiento y el Nivel de Abstracción de las Ideas en el Contexto de los Ítems*

Tipo de Conocimiento	Nivel de Abstracción de las Ideas Principales			Uso de Referentes Concretos		
	Todas las ideas son concretas	Al menos algunas ideas son ideas abstractas	Total	Todas las ideas están vinculadas a un objeto o evento concreto	Alguna o ninguna vinculación con a un objeto o evento concreto	Total
Razonamiento general	9	4	13	3	10	13
Declarativo	2	6	8	4	4	8
Procedimental	6	10	16	4	12	16
Esquemático	5	9	14	6	8	14
Total	22	29	51	17	34	51

Recursos de los contextos. Denominamos como recursos contextuales a todos los diagramas, tablas, imágenes, gráficos o fotografías que formaban parte del contexto general de los testlets. En nuestro estudio anterior se planteó la hipótesis de que los recursos podrían potencialmente afectar el desempeño de los estudiantes en cualquier dirección, dirigiendo o desviando la atención de los estudiantes a aspectos apropiados del ítem o a recursos ambiguos innecesarios para responder a los ítems. La hipótesis fue confirmada. Se encontró una relación entre el desempeño de los estudiantes y el número de recursos en el contexto, así como con la relevancia del recurso (codificado de tres maneras: el recurso es innecesario para entender el contexto, es meramente un complemento para la comprensión del contexto; o es esencial para la comprensión del contexto y proporciona información no transmitida en el texto del contexto).

En este trabajo nos centramos en el tipo de recurso recodificando los ítems cuyos contextos *no tienen ningún recurso* o *solamente contienen información textual*. Esta recodificación permitió incluir todos los ítems en el análisis en lugar de incluir sólo aquellos ítems con al menos un recurso. La Tabla 5 proporciona información sobre la distribución conjunta entre las dos variables. Los ítems en los que se incluyen representaciones esquemáticas en contextos aparecen principalmente en ítems para evaluar el conocimiento procedimental en contraste con ítems de conocimiento declarativo y esquemático. En las pruebas de ciencias de PISA es común que los ítems de conocimiento procedimental le pidan a los estudiantes que interpreten los datos que se presentan en una tabla o un gráfico.

Tabla 5. *Ítems Clasificados por Tipo de Conocimiento y por Tipo de Recurso en el Contexto*

Tipo de Conocimiento	Información Textual Únicamente	Representación(nes) Gráfica(s) Incluida(s)	Representación(nes) Esquemáticas(s) Incluida(s)	Total
Razonamiento general	4	7	2	13
Declarativo	4	1	3	8
Procedimental	5	4	7	16
Esquemático	3	8	3	13
Total	16	20	15	51

Escenarios de los Contextos. Los contextos se agruparon en dos tipos de escenarios: (1) *personal* - escenarios familiares a los estudiantes como los relacionados con el hogar o el trabajo escolar de los estudiantes; y (2) *no personal* – escenarios que describen acontecimientos, fenómenos, o actividades en un lugar de trabajo profesional o escenarios que utilizan textos científicos o informativos. Los escenarios no personales se pueden alinear fácilmente a los contextos sociales y globales del marco conceptual de PISA. Nuestra categoría de "no personal" incluye a ambas. En un estudio previo encontramos que los ítems con contexto con información científica tienden a ser más difíciles que los que tienen escenarios relacionados con el hogar, un hallazgo que podría estar relacionado con las demandas de lectura de los textos o con la familiaridad de los estudiantes con escenarios domésticos de la vida diaria. En el presente estudio, nos enfocamos en la asociación del tipo de escenario en los contextos de los ítems y el tipo de conocimientos que el ítem mide. La Tabla 6 proporciona información sobre la distribución de ítems por tipo de conocimiento y por tipo de escenario en el contexto. Los escenarios personales son aparentemente menos adecuados para evaluar conocimiento esquemático en comparación con los escenarios de contextos que se ocupan de asuntos sociales o globales; lo contrario se observó en los escenarios no personales. La frecuencia mayor de los escenarios no personales se observa para los ítems de tipo de conocimiento esquemático.

Tabla 6. *Ítems Clasificados por Tipo de Conocimiento y por Escenarios del Contexto*

Tipo de Conocimiento	**Personal**	**No-personal**	**Total**
Razonamiento general	4	9	13
Declarativo	5	3	8
Procedimental	6	10	16
Esquemático	0	14	14
Total	15	36	51

La vinculación de las demandas cognitivas de los ítems, las característica de los contextos de los ítems y el desempeño de los estudiantes

En esta sección, llevamos a cabo un examen más detallado de la posible asociación entre la demanda cognitiva de los ítems y las características del contexto de los ítems y el desempeño de los estudiantes. El modelo completo con las cuatro características del contexto y las competencias de PISA como predictores y el PRC como la covariable fue estadísticamente significativa, lo que indica que las características del contexto, en general, se asociaron con las demandas cognitivas de

los ítems (x^2 [24, N = 51] = 78.55, p = .000). La R^2 de Nagelkerke fue de 0,841, lo que indica una fuerte relación entre la predicción y la agrupación de acuerdo con las cuatro demandas cognitivas.

La Tabla 7 presenta los efectos principales de los predictores y la covariable en la predicción del nivel de demanda cognitiva de los ítems. Las cinco características del contexto o de ítems excepto el uso de la referencia concreta contribuyen de manera significativa a la predicción de la demanda cognitiva de ítems, entre los cuales el ajuste del contexto y las competencias de desempeño PISA muestran la asociación más fuerte con los niveles de exigencia cognitiva.

Tabla 7. *Pruebas de Cociente de Probabilidad* (*Likelihood Ratio Tests) para los Efectos Principales en la Predicción de Tipo de Conocimiento*

Efecto Principal Evaluado	Pruebas de Cociente de la Prueba de Chi-Cuadrada	gl	Significancia Estadística
Predictores			
Nivel de abstracción de las ideas principales	10.366	3	.016
Uso de referentes concretos	1.695	3	.638
Tipo de recurso incluido	15.222	6	.019
Escenario del contexto	21.974	3	.000
Competencias de PISA	40.523	6	.000
Covariable			
PRC (porcentaje de respuestas correctas)	9.416	3	.024

El resumen detallado de los resultados de la regresión se presentan en la Tabla 8, donde cada nivel (código) de cada aspecto de los contextos (por ejemplo, para el aspecto "*nivel de abstracción de las ideas principales del contexto*," el código incluido es "todas las ideas en el contexto son concretas") se evaluó frente a cada nivel cognitivo de los ítems (por ejemplo, el conocimiento declarativo) para determinar cómo cada código se asocia a las demandas cognitivas de los ítems. Es decir, para cada código utilizado para cada aspecto de los contextos evaluados, se llevaron a cabo comparaciones por pares entre los cuatro niveles de demanda cognitiva. Para evitar la presentación de todas las posibles comparaciones, el cuadro incluye únicamente los resultados encontrados con significancia estadística (p <0,05) y con tendencia estadística (p <0,10)[vi]. Por ejemplo, para el código "todas las ideas en el contexto son concretas", en vez de presentar las seis comparaciones (todas las comparaciones posibles con todos los tipos de demanda cognitiva), sólo se proporcionan aquellas que muestran un valor de p <0,10. Por lo tanto, cuando el contexto tiene la característica de "todas las ideas son concretas", la probabilidad de ser clasificado como un ítem de razonamiento general es aún mayor de 30 veces más que ser clasificado como un ítem de conocimiento declarativo.

Tabla 8. *Comparaciones por Pares de los Efectos Principales Selectas: Nivel de Abstracción, Recursos, Escenarios de los Contextos, Competencias, y Porcentaje de Respuestas Correctas**

Tipo de Demanda Cognitiva Focal	Tipo de Demanda Cognitiva Comparada	B[a]	Error Estándar de B	Prueba Wald	Significancia Estadística (gl = 1)	Exp(B) cuando cambia la Probabilidad[b] (Exp(B) as the Change In Odds)[b]
Nivel de abstracción de las ideas principales: Todas las ideas concretas						
Razonamiento	Declarativo	8.939	4.392	4.143	.042	>30
Razonamiento	Procedimental	3.389	1.649	4.223	.040	29.623
Razonamiento	Esquemático	5.968	3.513	2.886	.089	>30
Recursos: Solamente representaciones visuales						
Razonamiento	Declarativo	10.765	5.165	4.344	.037	>30
Razonamiento	Procedimental	2.678	1.448	3.421	.064	14.561
Declarativo	Esquemático	-9.047	4.615	3.844	.050	.000
Escenario de los Contextos: Personal						
Razonamiento	Esquemático	27.051	1.302	31.663	.000	>30
Declarativo	Esquemático	30.106	2.321	68.220	.000	>30
Procedimental	Declarativo	-4.118	2.321	3.147	.076	.016
Competencias de PISA: Explicar fenómenos científicamente						
Declarativo	Procedimental	29.290	3.551	68.037	.000	>30
Porcentaje de respuestas correctas						
Razonamiento	Esquemático	.109	.064	2.910	.088	1.115
Procedimental	Esquemático	.160	.066	5.874	.015	1.174
Procedimental	Declarativo	.137	.066	4.334	.037	1.146

* La tabla solamente reporta resultados con una $p < .10$.

a. B, como el coeficiente de inclinación del logit, indica el efecto de una unidad de cambio en la variable predictora en el logit pronosticado manteniendo las otras variables del modelo constantes.

b. Los cocientes de probabilidad son los efectos de una unidad de cambio en la variable predictora en el cociente de probabilidad de la variable pronosticada cuando se mantienen las otras variables constantes.

Encontramos que los ítems con contextos en los que las ideas principales son todas concretas son más propensos a ser clasificados como ítems de razonamiento general y no de otro tipo de conocimiento. Esto sugiere que los diseñadores de ítems deben incorporar algunas ideas abstractas con el fin de mantener la demanda cognitiva de los ítems. Cuando los contextos tienen un escenario personal, los resultados indican una probabilidad más alta de que los ítems se asocien con un razonamiento general o un conocimiento declarativo más que con un conocimiento esquemático.

Con respecto al tipo de recurso (por ejemplo, tablas), los resultados indican que los contextos con representaciones gráficas tienen más probabilidad de ser asociados con ítems de razonamiento general, que con ítems de conocimiento declarativo o procedimental. Por último, los ítems con representaciones visuales (por ejemplo, dibujos) tienen una probabilidad estadísticamente mayor de ser clasificados como ítems que suscitan conocimiento esquemático más que ítems que suscitan conocimiento declarativo.

Las *competencias* propuestas por PISA son un fuerte predictor para las demanda cognitivas de los ítems. Los ítems diseñados para evaluar en los estudiantes la competencia de *explicar fenómenos científicamente* se asociaron significativamente con el conocimiento declarativo más que con el conocimiento procedimental. Por último, es interesante observar que la ejecución de los estudiantes es relativamente mejor en los ítems

de procedimientos; en otras palabras, los ítems con demandas de conocimiento procedimental fueron más fáciles para los estudiantes que los que ítems que requirieren conocimiento declarativo y esquemático. Una tendencia similar se identificó para los ítems que requieren un razonamiento general cuando son comparados con ítems de conocimiento esquemático[vii].

El análisis de regresión se replicó con otros seis países con la muestra completa, así como con cada uno de los dos grupos de género. Curiosamente, los resultados de los cuatro países de la OECD (es decir, Australia, Canadá, Inglaterra y Nueva Zelanda) son comparables a los resultados de la muestra de Estados Unidos. La ejecución de los estudiantes de 15 años en Singapur y Hong Kong-China fue diferente en ítems con diferente demanda cognitiva; los patrones de desempeño en ítems de diferentes demandas cognitivas fueron más aparentes que en los estudiantes de Estados Unidos y los otros cuatro países. Por esta razón, en el modelo de regresión la dificultad de los ítems fue un predictor fuerte del nivel de demanda cognitiva de los ítems. Por ejemplo, para estudiantes de Singapur no se encontró una asociación significativa entre los escenarios de los contextos de los ítems y las demandas cognitivas de ítems. Del mismo modo, para los estudiantes de Hong Kong la presencia de representaciones visuales en contextos no fue un predictor de las demandas cognitivas de los ítems.

Discusión

A pesar de los años de experiencia en el desarrollo de pruebas y de las normas y principios desarrollados para la construcción de ítems de evaluación, lo que se sabe acerca de cómo diseñar buenos ítems y buenos contextos sigue siendo un misterio. De hecho, hasta el 40% de nuevos ítems no funcionan como estaba previsto o como se intentó originalmente (Haladyna, 1994). El diseño de pruebas con ítems eficaces sigue siendo más un arte que una ciencia. Todavía no se sabe cómo escribir ítems que de forma inequívoca pasen todas las pruebas de calidad. La revisión de los ítems de las pruebas supone para las compañías de pruebas (testing companies) un gasto mayor de tiempo y dinero. Las diferentes pautas, modelos o marcos propuestos a lo largo del tiempo para desarrollar ítems y pruebas (Anderson, 1972; Downing, 2006; Haladyna, 1994; Haladyna & Downing, 1989; Haladyna, Downing, & Rodríguez, 2002; Mislevy & Riconscente, 2006; Shoemaker , 1975) no dan una orientación clara acerca de cómo elaborar ítems de evaluación de manera efectiva. Más aún, ninguno de los modelos o marcos propuestos abordan cómo desarrollar ítems con contextos adecuados. Incluso usar a redactores experimentados no garantiza la producción de ítems de alta calidad (Haladyna, 1997). Como indicó Welch, "Teniendo en cuenta que los artículos son la columna vertebral de la industria de la evaluación, hay relativamente poca investigación sobre la redacción de los items" (2006, p. 306). De hecho, ni siquiera en las distintas ediciones recientes de los manuales de *Medición para la Educación* se ha tratado cuidadosamente el proceso de escribir ítems. No hay duda de que la elaboración de ítems no ha recibido la misma atención que tienen los modelos estadísticos de medición (Haladyna, 1994; Osterlind, 1998), incluso la atención recibida es todavía menor para la elaboración de los contextos de los ítems de manera efectiva. De las reglas para escribir ítems usualmente citadas en la literatura, sólo unas pocas están empíricamente fundamentadas (Haladyna, 1994), ya que la mayoría son un conjunto de reglas basadas en las experiencias y el sentido común.

Con estos antecedentes en mente, no debería ser una sorpresa, entonces, lo poco que sabemos sobre el diseño y uso del *contexto* en ítems. Se ha argumentado que los contextos en los ítems hace que las preguntas sean más concretas y menos demandantes (Ahmed y Pollitt, 2000, 2007). Por otra parte, la contextualización de los ítems ha sido considerada como una estrategia eficaz para comprobar si los estudiantes pueden aplicar (transferir) sus conocimientos (Ahmed y Pollit,

2000, 2007; Boaler, 1994; Haladyna, 1997) y para medir el pensamiento complejo (Haladyna, 1994). Sin embargo, se sabe poco acerca de si alguna de estas afirmaciones o más de una son ciertas. No hay prácticamente ninguna evidencia empírica acerca de qué tanto contexto es necesario en los ítems para poder obtener una información más rica y precisa acerca de la aplicación o de la transferencia del conocimiento de los estudiantes. En el caso de PISA, este tipo de evidencia empírica es crítica con el fin de justificar el uso de contextos en tantos niveles diferentes y con tanta abundancia.

Con este estudio, contribuimos a la acumulación de evidencias sobre las características de los contextos y su impacto en el desempeño de los estudiantes, evidencia que esperamos ayude al campo del desarrollo de pruebas. Reconocemos que el tamaño de la muestra de los ítems codificados es muy pequeño y que los resultados presentados deben ser tratados de manera exploratoria. Por ejemplo, por lo limitado del número de ítems, no hemos podido examinar los efectos de las interacciones entre las características del contexto, a pesar de que en la práctica muchas características en conjunto puedan afectar de una manera compleja el desempeño de los estudiantes. Aun así, nuestro estudio demuestra que los ítems propician diferentes tipos de procesos cognitivos en los examinados (que van desde la recuperación de la información memorizada a conjeturar, interpretar, justificar, etc.). ¿Qué características del contexto están relacionadas con ciertos procesos cognitivos?, ¿Están asociadas estas características y demandas cognitivas con el desempeño de los estudiantes? Estas son las principales preguntas que tratamos de responder en este estudio.

Primero, vinculamos diferentes características del contexto con las demandas cognitivas de los ítems. Los resultados indicaron que los ítems con contextos que involucran sólo ideas concretas es poco probable que requieran el uso del conocimiento del contenido en la respuesta. Los ítems de tipo conocimiento procedimental y esquemático son los que tienen contextos en los que se presentan algunas ideas abstractas. Si este resultado se puede replicar, entonces podemos empezar a desarrollar algunas pautas para ayudar a los que desarrollan ítems. ¿Por qué es importante este tipo de información? Debido a que el contexto utilizado en un ítem puede reducir las demandas cognitivas previstas.

Los resultados también indicaron la importancia del tipo de recurso que se utiliza en definir las demandas cognitivas de los ítems: las representaciones esquemáticas parecen estar vinculadas a ítems de conocimiento procedimental y no a ítems de conocimiento declarativo o esquemático. Aprender acerca de las características específicas de los recursos utilizados en los contextos de los ítems ayuda a adaptar mejor el tipo de conocimiento que es necesario para que los estudiantes utilicen los procesos cognitivos apropiados. También es interesante saber cómo el tipo de escenario, “personal o no personal”, está vinculado a ciertas demandas cognitivas. Este estudio proporciona evidencias de que los escenarios personales son menos adecuados que los escenarios no personales para evaluar conocimiento esquemático. Por último, es interesante saber que la competencia de “uso de evidencia” se asoció no solamente, como era de esperarse, con ítems de conocimiento procedimental, sino también se asoció con ítems de razonamiento general. ¿Qué necesitan los contextos de los ítems que intentan medir conocimientos procedimentales para evitar que se respondan correctamente sólo con un razonamiento general?

Independiente del marco conceptual que se utilice para codificar las demandas cognitivas de los ítems, este estudio exploratorio ha demostrado la importancia de aprender más acerca de las características de los contextos y la forma en que se relacionan con los procesos cognitivos en los que tienen que ocuparse los estudiantes. Aprender más acerca de las

características del contexto de ítems es esencialmente una cuestión de validez que debe ser considerada por el equipo de desarrollo de pruebas de la OECD. Además, este equipo debe compartir e informar de los resultados de pruebas piloto relacionados con los contextos para estimular a la comunidad de medición a preguntar y explorar por qué algunos contextos son poco inútiles e incluso engañosos. Permitir que los investigadores tengan idealmente acceso incondicional a los ítems no liberados ayudará al área de desarrollo de pruebas a moverse hacia adelante. Limitar el tiempo de acceso al estudio de los ítems no apoya un aprendizaje más profundo acerca de las características de los ítems de PISA y la forma en que se pueden mejorar.

Referencias

Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment. (2015). Promoting and evaluating cultural validity in the activities performed by the National Institute for Educational Evaluation (INEE). Submitted to the National Institute for Educational Evaluation. Mexico City, Mexico, January 16.

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

Bialystok, E. (2002). Cognitive processes of L2 users. En V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 145-165). Buffalo, NY: Multilingual Matters.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387-399.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Paper Number 71*. Consultado en OECD website: http://www.oecd-ilibrary.org/education/the-policy-impact-of-pisa_5k9fdfqffr28-en

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

Camilli, G. (2006). Test fairness. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger.

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

Capacity Development Group (2007, May). *Capacity assessment methodology: User's guide*. Bureau for Development Policy, United Nations Development Program. New York, September 2005. Consultado en the United Nations Development Programme website: https://www.unpei.org/sites/default/files/PDF/institutioncapacity /UNDP-Capacity-Assessment-User-Guide.pdf

Carnoy, M. (2015). International test score comparisons and educational policy. Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. Boulder, CO: National Education Policy Center. Consultado en http://nepc.colorado.edu/publication/international-test-scores

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. Consultado en the World Bank website: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Darling-Hammond, Linda (2014). What can PISA tell us about U.S. education policy? *New England Journal of Public Policy*: *26*(1), Art. 4. Consultado en http://scholarworks.umb.edu/nejpp/vol26/iss1/4

Dogan, E., & Circi, R. (2010). A blind item-review process as a method to investigate invalid moderators of item difficulty in translated assessment. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 3) (pp. 157-172). Hamburg: IERI.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, *117*(1), 1-28.

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

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

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: PREAL. Consultado en http://www.uis.unesco.org/Education/Documents/Ferrer.pdf

Figazzolo, L. (2009). *Impact of PISA 2006 on the education policy debate*. Consultado en http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009-00036-01-E.pdf

Gebril, A. (2016). Educational assessment in Muslim countries: Values, policies, and practices. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Back support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). Consultado en http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/Gilmore_Impact_PIRLS_TIMSS.pdf

Hamano, T. (2011).The globalization of student assessments and its impact on education policy [English version]. *Proceedings*, *13*, 1-11. (Originalmente apareció en japonés en 2008 en el *Annual Bulletin of JASEP (Japan Academic Society for Educational Policy)*, *15*, 21-37). Consultado en http://teapot.lib.ocha.ac.jp/ocha/bitstream/10083/51418/1/Proceedings13_01Hamano.pdf

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

Husén, T. (1983). *An incurable academic: Memoirs of a professor*. Oxford, UK: Pergamon Press.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25. doi: http://dx.doi.org/10.1086/648471

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement*, *6*, 125-160. doi: http://dx.doi.org/10.1177/014662168200600201

Kane, M. T. (2006). *Validation*. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian-heritage societies. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Lingard, B., & Lewis, S. (2016). Globalization of the American approach to accountability: The high price of testing. En G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Martínez-Rizo, F. (2015). Las pruebas ENLACE y EXCALE: Un estudio de validación. Consultado en http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

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

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: http://dx.doi.org/10.1037/0003-066X.50.9.741

Mullis, I. V. S., & Martin, M. O. (2011). *TIMSS 2011 item writing guidelines.* Consultado en http://timssandpirls.bc.edu/methods/pdf/T11_Item_writing_guidelines.pdf

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011: Assessment frameworks.* Consultado en http://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

National Project Managers' Meeting (2010, October). Translation and adaptation guidelines for PISA 2012. Doc: NPM10104e. PISA Consortium. Budapest, Hungary. Consultado en https://www.oecd.org/pisa/pisaproducts/49273486.pdf

Organisation for Economic Co-operation and Development (OECD). (n.d.). *Programme for international student assessment (PISA): Results from PISA 2012, Country note: United States.* Consultado en http://www.oecd.org/pisa/keyfindings/PISA-2012-results-US.pdf

Organisation for Economic Co-operation and Development (2006). *PISA released items: Mathematics*. Consultado en http://www.oecd.org/pisa/38709418.pdf

Organisation for Economic Co-operation and Development (2010). *Translation and adaptation guidelines for PISA 2012*. Consultado en http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). *The experience of middle-income countries participating in PISA 2000-2015* (PISA series). Washington, D.C. & Paris: The World Bank & OECD Publishing. doi: http://dx.doi.org/10.1787 /9789264246195-en

Ravela, P. (Ed.). (2001). Los próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina? Document No. 20. Working Group on Assessment and Standards. Santiago: PREAL. Consultado en http://campus-oei.org/calidad/grade.PDF

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-8, 13. doi: http://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible. In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *According to PISA—Does PISA keep what it promises?* Berlin: LIT Verlag.

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

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Basterra, E. Trumbull,

and G. Solano-Flores, *Cultural validity in assessment* (pp. 3-21). New York: Routledge.

Solano-Flores, G. (2016). Generalizability. En L. E. Suter, D. Wyse, E. Smith, & N. Selwyn (Eds.), *The BERA/SSAGE Handbook of Educational Research* (chap. 47). London: Sage.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa (REDIE), 8(*2). Consultado en http://redie.uabc.mx /redie/article/download/143/246

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L.A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78-91.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. En M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research in the context of the programme for international student assessment* (pp. 71-85). Springer Verlag.

Solano-Flores, G., & Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. En M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scaleassessment in education: Theory, issues, and practice* (pp. 87-109). New York: Routledge.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

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

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 533-573. doi: http://dx.doi.org/10.1002/tea.1018

Stachelek, A. J. (2010). Exploring motivational factors for educational reform: Do international comparisons dictate educational policy? *Journal of Mathematics Education at Teachers College, 1*, 52-55.

Suter, Larry E. (2000). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research, 70*(4), 529-545. doi: http://dx.doi.org/10.3102/00346543070004529

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312. doi: http://dx.doi.org/10.1016/0959-4752(94)90003-5

Tatto, M. T. (2006). Education reform and the global regulation of teachers' education, development and work: A cross-cultural analysis. *International Journal of Educational Research, 45*, 231-241. doi: http://dx.doi.org/10.1016/j.ijer.2007.02.003

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment* (Chap. 21). New York: Routledge.

van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. En G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*, (Chap. 25).New York: Routledge.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wuttke, J. (2007). Uncertainties and bias in PISA. En S. T. Hopmann, G. Brinek, and M. Retzl (Eds.), *According to PISA – Does PISA keep what it promises?* Berlin: LIT Verlag.

Authors / Autores

To know more / Saber más

Ruiz-Primo, Maria-Araceli (aruiz@stanford.edu).

Profesora asociada en la Escuela de Graduados en Educación de la Universidad de Stanford (USA). Su trabajo, financiado principalmente por la Fundación Nacional de la Ciencia y por el Instituto de Ciencias de la Educación, examina la evaluación del aprendizaje de estudiantes tanto en el salón de clases como en programas de evaluación a gran escala y las prácticas evaluativas de los maestros. Sus publicaciones se enfocan en el desarrollo y la evaluación técnica de estrategias diversas para evaluar el aprendizaje y el estudio de las prácticas de evaluación formativa formales e informales de los maestros en el salón de clases. Su trabajo reciente se centra en el desarrollo y la evaluación técnica de pruebas sensbles a la instrucción y de instrumentos para medir las prácticas de evaluación formativa de los maestros. Su dirección postal es: Graduate School of Education. 485 Lasuen Mall. Stanford University, Stanford, CA

Li, Min (minli@uw.edu)

Profesora asociada en el Colegio de Educación en la Universidad de Washington, Seattle (USA). Su interés es comprender como el aprendizaje de los estudiantes puede ser evaluado de manera más adecuada y con mayor exactitud tanto en el ámbito de las evaluaciones a gran escala como en el aula. Plantea una combinación de aproximaciones cognitivas y psicométricas, incluyendo el examen de las demandas cognitivas de items de pruebas a gran escala, el uso de cuadernos de ciencias como instrumentos de evaluación, la parametrización del diseño de tareas contextualizadas, el análisis de prácticas evaluativas de los maestros en el salón de clase y la validación de tareas complejas para maestros y para líderes educativos. Su dirección postal es: University of Washington - College of Education. Seattle, WA 98195-3600. Página personal: https://education.uw.edu/people/faculty/minli

RELIEVE

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

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]

NOTAS

[i] Las opciones de los ítems también pueden contener contextos. En este artículo solamente nos enfocamos en ítems contextualizados cuyo contexto se encuentra en la base del ítem.

[ii] Los ejemplos de los ítems que se presentan en este artículo fueron tomados de la liga "PISA take the test: sample questions from OECD's PISA assessment" (http://www.oecd.org/pisa/pisaproducts/pisatakethetestsamplequestionsfromoecdspisaassessments.htm).

[iii] En otros artículos (Li, 2001; Ruiz-Primo, 2003, 2007; Shavelson, Ruiz-Primo, Li, & Ayala, 2002) hemos usado el término análisis lógico para describir nuestra aproximación para analizar las características de los ítems. Contrastamos el análisis lógico con el análisis empírico; éste último involucra colectar y resumir las respuestas de los estudiantes a los ítems de evaluación. Este análisis examina empíricamente tanto las actividades cognitiva evocadas por los ítems de la prueba así como la ejecución observada de los estudiantes.

[iv] Tener acceso a los ítems no liberados para el primer estudio fue extremadamente difícil. Dando por hecho que el procedimiento no ha cambiado, obtener nuevos o más ítems no liberados requeriría varios meses antes de que el permiso se obtuviera por la Organisation for Economic Cooperation and Development (OECD).

[v] El marco conceptual de PISA de ciencias de 2015 se refiere a las tres competencias como "explicar un fenómeno científicamente", "evaluar y diseñar investigación científica", e "interpretar datos y evidencia científicamente. Además, incluye tres tipos de conocimiento científico, de contenido, procedural, y epistémico.

[vi] En la interpretación de los resultados, tomamos en consideración que:

a. Dada la muestra extremadamente pequeña de ítems (n=51) para la regresión multinomial, la significancia estadística es poco probable que esté inflada por el tamaño de la muestra.
b. Los resultados para comparar los niveles de demandas cognitivas en la regresión son de naturaleza exploratoria. Por lo tanto, la aplicación de cualquier procedimiento clásico de comparaciones múltiples, como la corrección de Bonferroni (que con frecuencia llega a ser muy conservadora cuando el número de comparaciones es grande y cuando las pruebas no son independientes), podría llevar a un poder estadístico reducido para detectar cualquier resultado significativo.
c. Las estimaciones de las probabilidades esperadas proveen información como un tamaño del efecto no estandarizado, lo cual ayuda a asegurar la relevancia práctica de los resultados. Más aún, se observaron patrones similares de resultados en la regresión multinomial en todos los países al igual que cuando el análisis se llevó a cabo separadamente por género, lo cual provee una seguridad adicional de la interpretación de los resultados reportados.

[viivii] El resultado acerca de la vinculación entre la demanda cognoscitiva del ítem y la dificultad del ítem puede ayudar a aclarar las comparaciones de los patrones de desempeño de los estudiantes en los diferentes países. Por ejemplo, entre los siete países que se analizaron, la ejecución de los estudiantes de Hong-Kong fue abrumadoramente mejor en aquellos ítems que requerían razonamiento general, conocimiento procedimental, y conocimiento esquemático, pero la ejecución es mediocre en ítems de conocimiento declarativo. En cambio, la ejecución de los estudiantes de Singapur sobresale en ítems de razonamiento general, de conocimiento declarativo, y de conocimiento procedimental, pero es menor en los ítems de conocimiento esquemático, lo cual es comparable con la ejecución de los estudiantes de E.E.U.U. y menor que la ejecución de los estudiantes de los otros cinco países. Estas observaciones se basaron en los estadísticos descriptivos sin considerar los pesos de las muestras para los análisis estadísticos inferenciales