

INVESTIGACIONES

Aprender del error es un acierto. Las dificultades que enfrentan los estudiantes chilenos en la Prueba PISA*

Learning from failure is a success. The difficulties faced by Chilean students in the PISA Test

Aprender com o erro é um acerto. Dificuldades enfrentadas por estudantes chilenos na prova do PISA

Verónica Villarroel Henríquez, Catalina García Gómez, Roberto Melipillán Araneda, Eugenio Achondo Mandiola, Alejandro Sánchez Oñate.^a

^a Universidad del Desarrollo, Chile. Fono: 56-41-2268602 / 56-41-2268590.

Correo electrónico: vvillarroel@udd.cl

RESUMEN

La presente investigación buscó caracterizar los errores y aciertos de estudiantes chilenos en su desempeño en la Prueba PISA 2009. Se trabajó con datos secundarios disponibles en la base de datos de la Prueba PISA 2009. La muestra estuvo compuesta por ítems de lectura, matemáticas y ciencias, aplicados a 5.600 estudiantes. La estrategia de selección fue no-probabilística y de tipo intencional, analizándose ítems de alto y bajo acierto. Empleando una estrategia metodológica cuantitativa se concluyó que los estudiantes chilenos presentaron mayores dificultades en ítems con problemas extensos, contextualizados, con más de una variable y que presentan información en distintos formatos. Respecto al tipo de ítem, los más complejos fueron los del tipo respuesta abierta y alternativas complejas múltiples, menos comunes en nuestro sistema escolar. Los resultados de este estudio pueden ser utilizados para perfeccionar los instrumentos de evaluación empleados en el aula.

Palabras clave: prueba PISA, evaluación, aprendizaje.

ABSTRACT

The present study sought to characterize the successes and failures of Chilean students in their performance in 2009 PISA Test. We worked with secondary data available in the database of the 2009 PISA test. The sample consisted of items in reading, mathematics and science, applied to 5,600 students. The selection strategy was non-probabilistic and intentional analyzing items with the higher and lower levels of success. Using a quantitative methodological approach we concluded that the Chilean students have greater difficulties answering items with large problems that have a context, involving more than one variable and presenting information in different formats. Regarding the type of item, the most complex type were open response and complex multiple choice, which are less common in our school system. The results of this study can be used to refine the assessment instruments used in the classroom.

Keywords: PISA test, evaluation, learning.

* Proyecto N°11128-2011 titulado: "Errores y aciertos comunes de estudiantes chilenos en la Prueba PISA 2009", financiado por el MINEDUC a través del Fondo Concursable FONIDE-PISA.

RESUMO

O presente estudo buscou caracterizar erros e acertos de estudantes chilenos em seus desempenhos na prova do PISA 2009. Trabalhou-se com dados secundários disponíveis no banco de dados da referida prova. A amostra foi composta por itens de leitura, matemática e ciências aplicados a 5.600 estudantes. A estratégia de seleção foi não probabilística intencional, analisando-se itens de altos e baixos acertos. Empregando-se uma estratégia metodológica quantitativa, concluiu-se que os estudantes chilenos apresentaram maiores dificuldades em itens com problemas extensos, contextualizados, com mais de uma variável e que apresentam informações em diferentes formatos. Quanto ao tipo de item, os mais complexos foram aqueles com respostas abertas e alternativas complexas de múltipla escolha, menos comum em nosso sistema escolar. Resultados do estudo podem ser utilizados para aperfeiçoar os instrumentos de avaliação utilizados em sala de aula.

Palavras chave: teste PISA, avaliação, aprendizagem.

1. INTRODUCCIÓN

El Programa PISA (*Program for International Students Assessment*) es un proyecto desarrollado por la Organización para la Cooperación y Desarrollo Económico (OCDE), que busca medir el nivel de preparación que muestran estudiantes de 15 años para enfrentar los desafíos futuros. Su objetivo es evaluar los sistemas educacionales del mundo a través de la aplicación de una prueba internacional, cuyo énfasis está puesto en el dominio de procesos, la comprensión de conceptos y la capacidad de desenvolverse en distintas situaciones de la vida diaria. Por ello, la prueba se aplica cuando los estudiantes han cumplido un ciclo de enseñanza obligatoria, ya que en este nivel se preparan para el término de la educación secundaria o el ingreso a la vida laboral (OCDE, 2010a).

El supuesto que subyace a esta evaluación es que aprender no es adquirir conocimientos específicos, como el nombre de plantas y animales, sino la aplicación de ellos para pensar en los grandes problemas en debate dentro de la comunidad adulta. Por esto se plantea que el diseño de la Prueba PISA comparte características de la evaluación auténtica, ya que establece una conexión y una articulación entre los contenidos enseñados en el contexto escolar y las experiencias de la vida real. Los ítems miden tareas valiosas y significativas, favoreciendo un pensamiento de alto nivel y el logro de aprendizajes complejos. De esta forma, reivindica la importancia del contexto, el realismo de la demanda y de la situación instruccional (Monereo, Castelló, Durán y Gómez, 2009; Sanmartí y Sardá, 2007; Rico, 2006; Aguilar y Cepeda, 2005; Bravo y Fernández, 2000).

La prueba PISA evalúa tres dominios: **lectura, matemáticas y ciencias**. Éstos se organizan, a su vez, en tres dimensiones: procesos, contenido y contexto. Los *procesos* se refieren a las tareas o actividades cognitivas que se necesitan realizar para responder al ítem, el *contenido* es el conocimiento que los estudiantes requieren para cada dominio de evaluación, y el *contexto* incluye las circunstancias de aplicación relevantes (OCDE, 2010b). En las Tablas N° 1, N° 2 y N° 3 se describen las competencias, procesos, contenidos y contextos evaluados en cada área.

La Prueba PISA es vista como una de las evaluaciones más importantes a nivel global. Desde el año 2000, más de 70 países han participado en esta medición, los que representan cerca del 90% de la economía mundial. A la luz de sus resultados, se han desarrollado intensos debates que estimulan la revisión de los programas escolares. Por ejemplo, los resultados obtenidos en Alemania generaron un importante debate público, tendiente a proponer una reforma en las políticas educacionales, lo que se denominó "*PISA Shock*". Similares reacciones se han visto en Dinamarca, Japón, Suiza y Gran Bretaña (Breakspear, 2012).

Tabla N° 1. Competencias, procesos, contenidos y contextos evaluados en Lectura

| Lectura | | | |
|---|---|---|--|
| Competencia | Procesos | Contenidos | Contextos |
| Comprender, utilizar y analizar textos escritos para desarrollar el conocimiento y participar en la sociedad. | <ol style="list-style-type: none"> 1. Recuperar información a través de la revisión, búsqueda, localización y selección de datos. 2. Interpretar textos para procesar la información a través de la comparación y el contraste de ella. 3. Reflexionar y evaluar textos, relacionando la nueva información con conocimientos procedentes de otras fuentes. | <ol style="list-style-type: none"> 1. Textos continuos, organizados en oraciones y párrafos, y prosas como descripción, narración, exposición, argumentación y prescripción. 2. Textos discontinuos que presentan información en formato diferente en función de su estructura (avisos, anuncios, listas, gráficos, formularios, diagramas o tablas). | <ol style="list-style-type: none"> 1. Privado. 2. Público. 3. Laboral. 4. Educativo. |

Fuente: elaboración propia.

Tabla N° 2. Competencias, procesos, contenidos y contextos evaluados en Matemáticas

| Matemáticas | | | |
|--|--|--|--|
| Competencia | Procesos | Contenidos | Contextos |
| Identificar y comprender el papel que juegan las matemáticas en el mundo, emitir juicios bien fundados y utilizarlas eficazmente para dar respuesta a las necesidades de una persona como ciudadano. | <ol style="list-style-type: none"> 1. Reproducción, que corresponde a la ejecución de cálculos simples que requieren del conocimiento y la definición de hechos, la representación y la solución de problemas rutinarios. 2. Conexión, que requiere la reunión de ideas y procedimientos matemáticos no rutinarios y que involucran la construcción de modelos, la interpretación y la solución de problemas. 3. Reflexión, que implica enlazar conocimientos de distintas procedencias, la formulación y la solución de problemas complejos, la comprensión y la reflexión profunda, el desarrollo de una aproximación matemática original y la capacidad de generalización. | <ol style="list-style-type: none"> 1. Cantidad. 2. Espacio y forma. 3. Cambio y relaciones. 4. Probabilidad. | <ol style="list-style-type: none"> 1. Personal. 2. Educativa o laboral. 3. Pública. 4. Científica. |

Fuente: elaboración propia.

Tabla N° 3. Competencias, procesos, contenidos y contextos evaluados en Ciencias

| Ciencias | | | |
|---|---|--|---|
| Competencia | Procesos | Contenidos | Contextos |
| Utilizar el conocimiento para identificar problemas, adquirir nuevo conocimiento, explicar fenómenos científicos, sacar conclusiones fundadas, comprender la ciencia como generación de conocimiento. | 1. Describir, explicar y predecir fenómenos científicos. 2. Comprender la investigación científica, que involucra la capacidad de reconocer y comunicar preguntas que pueden ser investigadas científicamente. 3. Utilizar evidencias científicas, dando sentido a los hallazgos científicos de modo que se puedan probar afirmaciones. | 1. Estructura y propiedades de la materia. 2. Cambio atmosférico, cambios físicos y químicos. 3. Transformaciones de la energía, fuerzas y movimiento. 4. Cambio fisiológico. 5. Biodiversidad. 6. Ecosistemas. 7. Cambio geológico. | 1. Ciencias de la vida y la salud. 2. Ciencias de la tierra y del ambiente. 3. Ciencias de la tecnología. |

Fuente: elaboración propia.

A partir de sus resultados, PISA establece tres posibles áreas de influencia en los procesos educativos de un país. La primera es la influencia en las mismas políticas y prácticas evaluativas en educación, la segunda se refiere a los estándares del currículo y, por último, la influencia en los indicadores de desempeño. De acuerdo a un estudio realizado en 37 países adscritos a la evaluación, el 54% de ellos reconoció que el sólo hecho de saber la posición de su país en el ranking global influyó en el cambio de políticas o prácticas educativas. Más del 80% de los países reconoce haber realizado cambios mayores o parciales en sus políticas o prácticas educativas (Breakspear, 2012).

En Chile, los resultados de PISA 2009 fueron muy relevantes para realizar cambios, por ejemplo, en el currículo de la instrucción del lenguaje. A partir esto, hubo un giro en el programa, cambiando el énfasis tradicional en la literatura y la gramática, hacia la comprensión lectora y el desarrollo de habilidades de comunicación, competencias evaluadas en la Prueba PISA (OCDE, 2010b).

El resultado obtenido por los estudiantes de un país permite contar con información de gran valor para la orientación y el mejoramiento de sus prácticas educativas. Para ello, es relevante avanzar desde una mirada centrada en los rankings que comparan países o escuelas según el puntaje o la calificación obtenida, hacia un análisis que permita la comprensión profunda del tipo de aprendizaje logrado y no logrado por los estudiantes (Hargreaves, 2005; Dann, 2002). Dado que cada día es mayor el esfuerzo que se pide a los centros escolares para que participen en evaluaciones educativas, podría resultar motivador y útil proporcionarles información específica acerca de los puntos fuertes y débiles de los alumnos, lo que permitiría mantener o corregir el rumbo del proceso educativo (Navas y Urdaneta, 2011).

En este sentido, la presente investigación consideró de interés para los colegas conocer, por ejemplo, el tipo de preguntas y habilidades medidas en la prueba PISA y compararlas con las que realizan en aula, o analizar las dificultades que presentan los estudiantes en los distintos ítems y áreas evaluadas, de modo de poder incorporar en sus evaluaciones la estructura de este tipo de ítems, intencionando la medición de competencias.

2. LA CALIDAD DE LA EDUCACIÓN Y DE LAS PRÁCTICAS DE EVALUACIÓN EN AULA

Distintas investigaciones han mostrado que las prácticas del docente son un factor crítico en la mejora del aprendizaje de los estudiantes (por ejemplo, Bellei, 2001; Cardemil, 2001). Esta relación entre prácticas docentes y aprendizaje de los alumnos ha sido probada a nivel básico (Goldhaber & Anthony, 2007) y en enseñanza media (Muñoz & Chang, 2007; McCaffrey, Lockwood, Koretz, Louis & Hamilton, 2004). Para Díaz y Poblete (2009), resulta evidente que los profesores requieren manejar estrategias de resolución de problemas para enseñar las matemáticas. Lo paradójico es que las mayores resistencias de los docentes están asociadas a aprender esta metodología. Esto concuerda con lo encontrado por Espinoza, Barbé y Gálvez (2011), quienes observaron que la forma de enseñar matemáticas impide que los estudiantes exploren auténticamente posibles caminos de abordaje y solución de los problemas.

Se puede suponer que los docentes competentes en el diseño de evaluaciones de aula son capaces de distinguir los tipos de preguntas que deben conformar una prueba y las distintas habilidades de pensamiento que el alumno requiere desplegar para resolverlos (William, Lee, Harrison & Black, 2004). Por lo tanto, un profesor hábil en construir una prueba, también debería ser competente para responderla, como lo plantean las investigaciones de Bayrakdar, Deniz, Akgün & Isleyen (2011).

Bayrakdar et al. (2011) estudiaron la relación entre la capacidad de resolver problemas y el éxito en ítems de la prueba PISA de estudiantes de pedagogía en matemáticas. Los estudiantes mostraron un mejor rendimiento en ítems en que se pedía encontrar un patrón o fórmula, y un desempeño más bajo cuando debían utilizar pensamiento divergente. Un estudio similar realizó Sáenz (2009), encontrando que los profesores obtenían un buen resultado cuando las tareas evaluaban reproducción del saber. No obstante, las dificultades aparecían cuando implicaba habilidades de relación o reflexión porque carecían de redes cognitivas que caracterizan un aprendizaje profundo.

En Chile, el área más deficitaria de la práctica pedagógica es la aplicación de pruebas en aula. La calidad de los instrumentos de evaluación construidos por los docentes, las habilidades evaluadas y el uso de la información entregada por ellos, han resultado ser las dimensiones peor evaluadas (Manzi y Flotts, 2007). Pellegrino, Chudowsky & Glaser (2001) observaron grandes discrepancias entre los aprendizajes valorados por las mediciones de gran escala y los aprendizajes que son evaluados por los docentes a través de pruebas en sala de clases, siendo las primeras significativamente más complejas que las últimas.

Un ejemplo de ello se observa en el estudio que analizó las habilidades matemáticas medidas en octavo básico por la prueba internacional TIMSS, la prueba nacional SIMCE, y ejercicios de pruebas aplicadas en la sala de clases, encontrándose que el 15% de las

preguntas de TIMMS y el 10% de las preguntas de la Prueba SIMCE medían conocimientos y procedimientos. En cambio, un 67% de las pruebas de aula evaluaban estas habilidades. Por otro lado, la capacidad de resolver problemas tenía una ponderación del 40% en TIMSS, 55% en SIMCE y 19% en evaluaciones de aula. La habilidad más compleja, que involucraba razonar matemáticamente, concentró el 25% de los ítems en TIMMS, el 17% en SIMCE y sólo el 6% de las evaluaciones de aula (Merckes, 2007).

Una tendencia similar a lo observado en Chile se ha visto en otros países. En España, Colomina, Onrubia y Naranjo (2000) analizaron la evaluación del aprendizaje en matemáticas de profesores de enseñanza media. Los resultados mostraron que las pruebas estaban formadas por tareas independientes entre sí y de naturaleza intra-matemática, evaluando prioritariamente contenidos que remiten a niveles bajos de exigencia cognitiva. Otro ejemplo es la investigación de Anagnostopoulou, Hatzinika & Christidou (2010) en Grecia, quienes compararon pruebas construidas por los propios colegios y la prueba PISA en el área de ciencias. Encontraron que la habilidad más frecuentemente evaluada por los colegios es “recordar conocimiento declarativo”, en contraste con “usar evidencia científica”, que es una competencia altamente medida en la prueba PISA y donde los estudiantes muestran baja familiaridad, lo que explica un bajo desempeño en este tipo de ítems. En otro estudio similar, Anagnostopoulou, Hatzinika, Christidou & Domopoulos (2013) encontraron que el desempeño de los estudiantes era más bajo en la prueba internacional que en las pruebas locales, debido a que en la prueba PISA se tendía a expresar el mensaje científico principalmente en formato visual más que lingüístico, mientras que en las evaluaciones de aula las pruebas se basan en textos.

Uno de los aspectos diferenciadores entre las pruebas de aula y las evaluaciones estandarizadas radica en la capacidad de la prueba para medir profundidad y aplicación del conocimiento, habilidades que son ampliamente evaluadas en pruebas estandarizadas como TIMSS y PISA. Las fortalezas de estas pruebas radican en ítems que cuentan con un contexto de aplicación, las preguntas son representadas en distintos formatos y su intención es medir la generalización o la transferencia del conocimiento a nuevas situaciones, lo que implica un alto procesamiento cognitivo para ser resueltas.

Estas características responden a lo que Webb (2002) denomina profundidad del conocimiento, distinguiendo cuatro niveles: recuerdo, desarrollo de conceptos y habilidades, pensamiento estratégico y pensamiento amplio. Bajo este modelo, las evaluaciones que se realizan en el aula corresponderían a un nivel inicial de la profundidad de conocimiento, referido a la medición de recuerdo y el reconocimiento de información, dos de los ámbitos menos evaluados en SIMCE y TIMSS. Esto podría explicar el rendimiento alcanzado por los estudiantes chilenos en las pruebas estandarizadas, donde el 50% se desempeña en el nivel mínimo y sólo el 1% muestra desempeños destacados.

En la evaluación tradicional de aula, el profesor diseña evaluaciones que buscan constatar la existencia de ciertos conocimientos en los estudiantes. Frente a este tipo de evaluaciones, los alumnos responden correctamente cuando repiten adecuada y específicamente esos repertorios de información. El estudiante recordará datos e información, pero a través de ellos no construirá conocimiento ni habilidades de pensamiento transferibles a otros contextos y, probablemente, no podrá aplicarlos para explicar fenómenos de la vida cotidiana, tal como los ítems de la prueba PISA lo solicitan. Para Wiggins (1990), el peligro de estas evaluaciones convencionales es que a través de ellas se pierdan o empobrezcan las habilidades intelectuales de los estudiantes.

Mejorar la calidad de los aprendizajes de los estudiantes implica reconocer que la forma de evaluar lo aprendido no está logrando que aprendan de manera significativa y profunda. Es necesario promover el desarrollo de evaluaciones que vayan más allá de la reproducción textual de contenidos fragmentados y carentes de sentido, y que se avance hacia la evaluación de habilidades de pensamiento de orden superior en formatos contextualizados, relacionados con problemas de la vida diaria. Esto implica que la evaluación debe ser concebida como parte de la enseñanza y articulada en forma íntima con ella, dando cuenta de la complejidad de los aprendizajes (Wyse & Viger, 2011; Gulikers, Bastiaens & Kirschner, 2004; Monereo, 2009).

Considerando estos supuestos, la presente investigación buscó caracterizar los errores y los aciertos de estudiantes chilenos en su desempeño en la Prueba PISA 2009. El objetivo fue describir las dificultades que presentan los alumnos para resolver correctamente los ítems asociados a dichos aciertos y errores, con la finalidad de sugerir cambios en los sistemas de evaluación de aula.

3. MÉTODO

El enfoque teórico-metodológico escogido para esta investigación fue cuantitativo, ya que se necesitaba describir y explicar los tipos de errores y aciertos bajo la lógica hipotético-deductiva, a través de técnicas cuantificables y replicables. Además, se aspiraba a la generalización de los resultados.

El alcance de los objetivos de esta investigación fue descriptivo, permitiendo caracterizar el fenómeno de interés y conocer su variabilidad. Se trató de un estudio de tipo transversal y no experimental, ya que se analizaron datos recogidos en un tiempo particular y no se manipularon ni controlaron variables. Se utilizaron datos secundarios, el muestreo fue no probabilístico y de tipo intencional (casos extremos), seleccionando los ítems de mayor acierto y error para los estudiantes chilenos.

3.1. MUESTRA

La muestra del estudio estuvo compuesta por 66 ítems de Lectura, Matemáticas y Ciencias, aplicados a 5.600 estudiantes de 84 colegios municipales, 97 particulares subvencionados y 20 particulares pagados, lo que corresponde al 35% del total de ítems aplicados en Chile. Estos datos están disponibles en la base de datos PISA 2009.

El criterio de selección de los ítems fue el porcentaje de acierto y error que mostraron los estudiantes en ellos. Es así como el alto acierto involucraba ítems contestados correctamente por un porcentaje de alumnos igual o mayor al 70%, mientras que los ítems de bajo acierto eran contestados correctamente por un porcentaje de estudiantes igual o inferior al 30%. De esta forma, la muestra estuvo compuesta por 20 ítems que mostraron alto acierto y 46 ítems en que los estudiantes presentaron bajo acierto.

3.2. PROCEDIMIENTO

Como se indicó anteriormente, primero se realizó la selección de ítems de respuesta abierta y cerrada de mayor y menor acierto dentro de la población chilena. Todos los ítems

seleccionados fueron analizados considerando la respuesta del alumno y las posibles explicaciones del error o el acierto de esa pregunta. Para indagar en sus posibles causas, se analizaron las siguientes características de cada ítem: *tipo de ítems, habilidad cognitiva evaluada, intención de la pregunta, claridad de la instrucción y redacción de la pregunta*. Para ello, se contaba con la pauta de corrección de PISA, donde aparecían todos los ítems y sus respuestas correctas.

El análisis del error en ítems de bajo acierto requirió de procedimientos adicionales en preguntas de respuesta abierta y cerrada. En el caso de ítems de respuesta cerrada se analizó el funcionamiento de los distractores, es decir, las alternativas de respuesta que no corresponden a la opción correcta o clave.

En los ítems de respuesta abierta, la tabulación de PISA mostraba un puntaje que indicaba si la respuesta estuvo completamente mala (cero puntos), medianamente equivocada (1 punto) o correcta (2 puntos). Por estos motivos, se seleccionaron aleatoriamente 40 ítems en que los estudiantes contestaban con puntuación cero, fotografiándose 812 respuestas (289 de Matemáticas, 276 de Lectura y 247 de Ciencias), cuyo análisis enriqueció la caracterización de ítems de bajo acierto en respuestas abiertas.

3.3. INSTRUMENTO

El instrumento analizado corresponde a 66 ítems de la Prueba PISA 2009. La prueba constaba de cinco tipos de ítems, entre ellos: *Multiple Choice, Complex Multiple Choice, Closed Constructed Response, Short Response y Open Response*. Los dos primeros corresponden a ítem de respuesta cerrada y los tres últimos a respuesta abierta. Los cinco ítems fueron aplicados en Lectura y Matemáticas. En Ciencias no se utilizaron ítems de *Short Response*. A continuación se describen los distintos tipos de ítems utilizados, incorporando un ejemplo de cada uno de ellos, lo que permite comprender mejor el sentido de las preguntas.

Tabla N° 4. Descripción de ítem *Multiple Choice*

| Descripción del Ítem | Ejemplo: Ítem S256Q1 |
|---|---|
| Estas preguntas corresponden a ítems de selección múltiple con 4 opciones de respuesta. Los estudiantes deben seleccionar la mejor respuesta. | Una cuchara de metal, una cuchara de madera y una cuchara de plástico se ponen en agua caliente. ¿Cuál de las tres cucharas se sentirá más caliente después de 15 segundos? A. La cuchara de metal. B. La cuchara de madera. C. La cuchara de plástico. D. Las tres cucharas se sentirán igualmente calientes. |

Fuente: elaboración propia.

Tabla N° 5. Descripción de ítem Complex Multiple Choice

| Descripción del Ítem | Ejemplo: Ítem R455Q05 |
|--|---|
| <p>En estas preguntas se presentan varias afirmaciones, y para cada una de ellas, el estudiante debe determinar su veracidad respecto a un contexto.</p> | <p>Los beneficios para la salud de la epicatecina, un compuesto que se encuentra en el cacao natural, son tan espectaculares que pueden competir con los de la penicilina, según un informe publicado en Química e Industria. El profesor Norman Hollenberg, de la Escuela de Medicina de Harvard, pasó años estudiando a los Kuna, pueblo consumidor de cacao líquido de Panamá, y descubrió que su riesgo de desarrollar derrame cerebral, fallas cardíacas, cáncer y diabetes era menor al 10%. “La epicatecina es tan importante para la salud que debería ser considerada una vitamina”, señala. La ciencia actual todavía no reconoce un rol tan esencial de la epicatecina. Sin embargo, la implicancia de que detrás de estas enfermedades comunes podría haber algún tipo de deficiencia, podría llevar a su desarrollo como suplemento nutricional. Usa la información del texto para decidir si las siguientes afirmaciones son verdaderas o falsas. Encierra en un círculo “Verdadero” o “Falso” para cada una de las afirmaciones.</p> <p style="text-align: center;">Afirmación ¿Verdadero o falso?</p> <p style="text-align: center;">El pueblo Kuna de Panamá come mucho chocolate. Verdadero / Falso</p> <p style="text-align: center;">La epicatecina se encuentra en la cocoa. Verdadero / Falso</p> <p style="text-align: center;">Es una creencia extendida que la epicatecina es esencial para la salud. Verdadero / Falso</p> <p style="text-align: center;">La epicatecina es reconocida ampliamente como una vitamina. Verdadero / Falso</p> |

Fuente: elaboración propia.

Tabla N° 6. Descripción del ítem Closed Constructed Response

| Descripción del Ítem | Ejemplo: Ítem M474Q01 |
|--|---|
| <p>En estas preguntas de respuesta abierta se le pide al estudiante construir una respuesta numérica o contestar con una palabra o frase corta, con límites restringidos de espacio. En este marco, se debe elaborar una respuesta con una única opción de respuesta; opción que se encuentra literalmente en el contexto entregado. Es un ítem similar a una completación donde el alumno construye la respuesta a partir de la información entregada explícitamente en el contexto de la pregunta.</p> | <p>Ocho estudiantes corren 100 metros. Los siguientes tiempos, en segundos, fueron registrados por los estudiantes.</p> <p style="text-align: center;">15,2 16,02 16 17,4 15,65 17,05 16,1 15,03</p> <p>¿Cuál es el tercer tiempo más rápido? Encierra en un círculo la respuesta correcta.</p> |

Fuente: elaboración propia.

Tabla N° 7. Descripción del ítem *Short Response*

| Descripción del Ítem | Ejemplo: Ítem M464Q01 |
|---|---|
| <p>En estas preguntas de respuesta abierta se requiere que el estudiante genere una respuesta breve con un rango limitado de respuestas posibles para tener éxito. A diferencia del ítem <i>Closed Constructed Response</i>, la respuesta no se encuentra literalmente en el contexto de la pregunta, sino que debe ser resuelta a partir del conocimiento previo del alumno.</p> | <p>María tiene 120 metros de cerca. Ella quiere hacer tres recintos cuadrados para encerrar a sus chanchos, como se muestra en el diagrama.</p> <p>¿Cuál es el área de cada uno de los recintos?</p>  <p>Área:.....m².</p> |

Fuente: elaboración propia.

Tabla N° 8. Descripción del ítem *Open Response*

| Descripción del Ítem | Ejemplo: Ítem R403Q03-019 |
|---|---|
| <p>En estas preguntas de respuesta abierta, contestar acertadamente permite la posibilidad de elaborar una respuesta personal que, cumpliendo ciertos criterios, puede estar correcta. Así, existe más de una respuesta posible, dependiendo de la manera en que se argumenta o aborda el problema. Se espera una escritura más extensa, que frecuentemente requiere alguna explicación o justificación. Una respuesta estará correcta cuando exista coherencia entre la solución, la argumentación entregada y la información del problema presentado.</p> | <p>¿Se vuelven los dientes más y más limpios mientras más tiempo y con más fuerza los lavamos? Investigadores británicos dicen que no es cierto. Ellos probaron muchas alternativas diferentes hasta encontrar la manera perfecta de lavarse los dientes. Un cepillado de dos minutos, sin demasiada fuerza, es lo que da mejores resultados. Si uno cepilla demasiado fuerte, se daña el esmalte de los dientes y las encías sin remover los residuos de alimentos ni el sarro.</p> <p>Bente Hansen, experta en cepillado de dientes, señala que es una buena idea sostener el cepillo de la misma forma en que se sostiene un lápiz. “Empiece en un extremo y cepille progresivamente a lo largo de toda la dentadura”, nos dice. “¡Y no se olvide de la lengua! De hecho, ésta puede contener grandes cantidades de bacterias que podrían causar mal aliento”.</p> <p>Usa este artículo para responder a la siguiente pregunta. Según Bente Hansen, ¿por qué hay que cepillarse la lengua?</p> <p>.....</p> <p>.....</p> <p>.....</p> <p>.....</p> <p>.....</p> <p>.....</p> |

Fuente: elaboración propia.

Los tipos de ítems que se presentan más frecuentemente, en las tres áreas evaluadas, son los de *Multiple Choice* y *Open Response*. Sin embargo, en la prueba de Ciencias toman fuerza los ítems de *Complex Multiple Choice*, y en Matemáticas se observan menos preguntas del tipo *Closed Constructed Response*.

4. PLAN DE ANÁLISIS

Los análisis hicieron uso de la variable de ponderación (peso final de cada estudiante), así como de los 80 pesos replicados (BRR con ajuste de Fay). Se utilizó el Programa STATA 11.2. Complementariamente, en los ítems de respuesta abierta se incorporó el análisis de contenido, donde tres investigadores analizaron, de forma independiente, las respuestas considerando los siguientes criterios: errores en la respuesta y elementos faltantes en la respuesta. Luego las tres evaluaciones se sistematizaron y sintetizaron, buscando los aspectos comunes que explicaban el error de los alumnos en esas preguntas.

5. RESULTADOS

5.1. DATOS DESCRIPTIVOS DE LOS ÍTEMS DE ALTO Y BAJO ACIERTO SELECCIONADOS EN CADA ÁREA

De las 102 preguntas aplicadas en Lectura, 15 ítems (14.7%) fueron contestados erróneamente por más del 70% de los estudiantes. De estas 15 preguntas, 4 corresponden a ítems de respuesta cerrada del tipo *Complex Multiple Choice*, lo que representa el 27% del error. El resto del error (73%) se distribuye en los 11 ítems de respuesta abierta (8 *Open Response*, 1 *Short Response* y 2 *Closed Constructed Response*).

De las 36 preguntas de Matemáticas, el 47% fueron contestadas erróneamente por más del 70% de los estudiantes chilenos, lo que equivale a 17 preguntas. De estas 17 preguntas, el 29% de ellas corresponde a preguntas de respuesta cerrada (2 son de *Multiple Choice* y 3 de *Complex Multiple Choice*), y el 71% a preguntas de respuestas abiertas (2 son de *Closed Constructed Response*, 5 *Short Response* y 5 *Open Response*).

Finalmente, de las 53 preguntas de Ciencias, el 26% fueron contestadas erróneamente por más del 70% de los estudiantes chilenos, lo que equivale a 14 ítems. De las 14 preguntas, el 43% de ellas son preguntas cerradas (2 de *Multiple Choice* y 4 de *Complex Multiple Choice*) y el 57% son de respuesta abierta (1 de *Closed Constructed Response* y 7 de *Open Response*). La Tabla N° 9 resume los resultados encontrados en ítems de bajo acierto en las tres pruebas antes mencionadas.

Tabla N° 9. Descripción de ítems de bajo acierto

| | Lectura | | Matemáticas | | Ciencias | |
|------------------------|---------|---|-------------|----|----------|----|
| | N | % | N | % | N | % |
| <i>Multiple Choice</i> | 0 | 0 | 2 | 12 | 2 | 14 |

| | | | | | | |
|------------------------------------|---|----|---|----|---|----|
| <i>Complex Multiple Choice</i> | 4 | 27 | 3 | 18 | 4 | 29 |
| <i>Closed Constructed Response</i> | 2 | 13 | 2 | 12 | 1 | 7 |
| <i>Short Response</i> | 1 | 7 | 5 | 29 | - | - |
| <i>Open Response</i> | 8 | 53 | 5 | 29 | 7 | 50 |

Fuente: elaboración propia.

Respecto al acierto, en las tres áreas, el tipo de ítem donde los estudiantes muestran mayor acierto es el *Multiple Choice*. Esto ocurre en el 61% de los ítems de mayor acierto de Lectura, el 100% de los ítems de mayor acierto en Matemáticas y el 60% de las preguntas de mayor acierto en Ciencias. La Tabla N° 10 resume los resultados encontrados en ítems de alto acierto en las tres pruebas antes mencionadas.

Tabla N° 10. Descripción de ítems de alto acierto

| | Lectura | | Matemáticas | | Ciencias | |
|------------------------------------|---------|----|-------------|-----|----------|----|
| | N | % | N | % | N | % |
| <i>Multiple Choice</i> | 8 | 61 | 2 | 100 | 3 | 60 |
| <i>Complex Multiple Choice</i> | 0 | 0 | 0 | 0 | 1 | 20 |
| <i>Closed Constructed Response</i> | 4 | 31 | 0 | 0 | 0 | 0 |
| <i>Short Response</i> | 1 | 8 | 0 | 0 | - | - |
| <i>Open Response</i> | 0 | 0 | 0 | 0 | 1 | 20 |

Fuente: elaboración propia.

5.2. RELACIÓN ENTRE TIPO DE ÍTEM Y DESEMPEÑO EN ESTUDIANTES CHILENOS EN CADA UNA DE LAS ÁREAS EVALUADAS

Para determinar si las diferencias en los porcentajes de acierto de los ítems en cada una de las tres pruebas se asociaban al tipo de ítem, se llevó a cabo un ANOVA de un factor. Los supuestos de ANOVA fueron testeados y se verificó el cumplimiento de cada uno de ellos. Se analizó el supuesto de la independencia de los residuos a través de la Prueba X² para independencia, el supuesto de la normalidad de los residuos a través del Test de Kolmogorov-Smirnov y el de la homocedasticidad a través del Test de Levene.

En Lectura el análisis muestra que existen diferencias estadísticamente significativas entre el tipo de ítem y el desempeño de los alumnos ($F_{(4, 89)}=8,20, p < 0,001$). La prueba *post hoc* HSD de Tukey indicó tres diferencias estadísticamente significativas, las cuales se presentaron entre los tipos *Multiple Choice* y *Complex Multiple Choice*, *Complex Multiple Choice* y *Short Response*, así como entre *Complex Multiple Choice* y *Closed Constructed Response*. Esto significa que en Lectura, los estudiantes chilenos muestran un desempeño

significativamente menor en el tipo de ítem de *Complex Multiple Choice* en comparación a los ítems de *Multiple Choice*, *Short Response* y *Closed Constructed Response*.

En Matemáticas no se incluyeron los ítems de *Closed Constructed Response*, debido a su bajo número. Los resultados mostraron que existen diferencias estadísticamente significativas entre los tipos de ítems ($F_{(3, 28)}=6,39, p< 0,01$). Para determinar cuáles promedios resultaban diferentes se llevó a cabo un análisis *post hoc* empleando la prueba HSD de Tukey. Se observaron dos diferencias estadísticamente significativas, correspondientes a la comparación entre los tipos *Multiple Choice* y *Short Response*, y entre *Multiple Choice* y *Open Response*. Se concluye que el porcentaje de acierto de los alumnos chilenos en las preguntas de *Multiple Choice* de la prueba de Matemáticas es significativamente mayor al logrado en las preguntas de *Short Response* y *Open Response*. El rendimiento más bajo lo obtienen en los ítems *Open Response*.

En Ciencias tampoco se incluyeron los ítems de *Closed Constructed Response*. Los resultados mostraron que existen diferencias estadísticamente significativas entre los tipos de ítem ($F_{(2, 47)}=4,62, p< 0,05$). Un análisis en base a la prueba *post hoc* HSD de Tukey permitió identificar una diferencia estadísticamente significativa, correspondiente a la comparación entre los tipos *Complex Multiple Choice* y *Open Response*. Esto quiere decir que, en Ciencias, los estudiantes chilenos mostraron un desempeño significativamente inferior en los ítems de *Open Response* que en los de *Complex Multiple Choice*. La Tabla N° 11 muestra los resultados obtenidos en las tres pruebas.

Tabla N° 11. Comparación de porcentajes de acierto de los ítems de Lectura, Matemáticas y Ciencias según tipo de ítem

| | Lectura | | | Matemáticas | | | Ciencias | | |
|------------------------------------|---------|----|------|-------------|----|------|----------|----|------|
| | N | M | DE | N | M | DE | N | M | DE |
| <i>Multiple Choice</i> | 35 | 69 | 0,17 | 9 | 49 | 0,19 | 18 | 51 | 0,19 |
| <i>Complex Multiple Choice</i> | 9 | 31 | 0,22 | 7 | 31 | 0,07 | 17 | 42 | 0,16 |
| <i>Short Response</i> | 9 | 62 | 0,27 | 8 | 24 | 0,18 | - | - | - |
| <i>Closed Constructed Response</i> | 11 | 69 | 0,27 | - | - | - | - | - | - |
| <i>Open Response</i> | 38 | 50 | 0,19 | 8 | 16 | 0,17 | 17 | 34 | 0,13 |

Fuente: elaboración propia.

5.3. RELACIÓN ENTRE TIPO DE HABILIDAD Y DESEMPEÑO EN ESTUDIANTES CHILENOS

Se llevó a cabo un ANOVA de un factor para determinar si las diferencias en los porcentajes de acierto se asociaban al tipo de proceso cognitivo evaluado, no encontrándose diferencias estadísticamente significativas. Es así como en Lectura, los resultados de la prueba ANOVA correspondieron a $F_{(2, 90)}=2,15, p=0,12$. En Matemáticas, al analizar las diferencias en los promedios de los tres tipos de habilidad, se obtuvo como resultado $F_{(2, 32)}=0,58, p=0,57$.

Finalmente, en Ciencias, al analizar las diferencias en los promedios de los tres tipos de ítems, se obtuvo como resultado $F_{(2, 50)}=0,45, p=0,64$.

5.4. CARACTERIZACIÓN DE ÍTEMS EN QUE LOS ALUMNOS PRESENTAN BAJO ACIERTO

Se observa una relación entre el tipo de preguntas y el área de aplicación (Lectura, Matemáticas y Ciencias). En Lectura, los ítems que resultaron más complejos para los estudiantes fueron los de respuesta cerrada, particularmente, el ítem de *Complex Multiple Choice*. En cambio, en Matemáticas, los más difíciles fueron los de respuesta abierta, en particular, el *Open Response*. En Ciencias, el error fue más equilibrado entre preguntas de respuesta abierta y cerrada, siendo el error levemente superior en las preguntas de *Open Response*.

No sólo influye el tipo de ítem en la dificultad que enfrentan los alumnos para responder correctamente a ellos, sino el área en que se aplica ese ítem. En el caso de Matemáticas, la alta magnitud de error que presentan los estudiantes en ítems *Open Response*, hace pensar que no están acostumbrados a responder preguntas matemáticas en un formato abierto en el que tienen que justificar, argumentar o reflexionar sobre un procedimiento o problema de carácter numérico. Lo contrario sucede en el área de Lectura, donde el mayor error ocurre en preguntas *Complex Multiple Choice*, lo que hace pensar que los estudiantes no están familiarizados en utilizar el pensamiento lógico y pragmático para resolver contenidos de lenguaje.

En el caso de las preguntas cerradas, en el tipo de ítem *Complex Multiple Choice* que consiste en el análisis de sentencias o afirmaciones sobre un contexto, la principal dificultad es que, para tener el puntaje correcto, el estudiante debe evaluar correctamente cada una de las afirmaciones. Son ítems “todo o nada”. En cambio, las preguntas *Open Response* requieren que el estudiante construya respuestas propias, las que deben apuntar a conclusiones derivadas, tanto de los datos presentados en el contexto de la pregunta, como de la información previa del alumno. El contexto es de mayor extensión y se requiere contar con capacidad de abstracción. Además, se necesita comprensión lectora, habilidad de redacción y utilización de un esquema argumentativo (ya sea explícito o implícito) para dar respuesta a la pregunta. En este tipo de preguntas existe más de una respuesta posible, por lo que es relevante ser capaz de pensar en alternativas nuevas a las ya señaladas en el contexto, ya sea a nivel de hipótesis o ideas novedosas.

A continuación, en la Tabla N° 12 se presentan las características comunes de los ítems de mayor complejidad para las tres áreas evaluadas.

6. DISCUSIÓN Y CONCLUSIONES

Analizar el error que cometen los estudiantes chilenos en la Prueba PISA permite poner el foco en las habilidades cognitivas que los alumnos no logran poner en juego para resolver problemas, como también en el desafío que implica para los docentes construir pruebas de aula que evalúen tales competencias. Conocer el tipo de ítem que genera mayor dificultad para los estudiantes, permite que los docentes se familiaricen con ellos y los comiencen a utilizar con mayor frecuencia, modificando y complejizando las pruebas con las cuales se evalúa a los estudiantes en las salas de clases.

Tabla N° 12. Características comunes de los ítems de menor acierto

| Extensión | Variables | Formato | Contexto | Habilidad requerida |
|--|--|---|---|---|
| Contextos largos (entre ½ y 1 página). | Más de un cálculo o fórmula para resolver el problema. Presentar más de una variable en una tabla o gráfico. Presentación de más de una hipótesis, experimentos, argumentos. | Datos en distinto formato (noticias, narraciones, opiniones, entrevistas). Datos en distintos formatos, por ejemplo: gráficos, tablas, problema, fórmulas. Problemas con datos cuantitativos y cualitativos en un mismo contexto. | Problemas relacionados con el diario vivir. Contexto con datos imaginarios o hipotéticos. Problema con información que puede parecer absurda o no lógica. | Pregunta que requiere interpretar e inferir datos no explicitados en el contexto entregado. Se requiere abstraer la lógica o el procedimiento de una situación y transferirla a otra. Solicita integrar información desde distintos formatos. Integración de conocimiento y experiencias. Requiere reflexión sobre información entregada en el problema. Requiere evaluar datos del contexto para decidir. |

Fuente: elaboración propia.

Este estudio mostró que los estudiantes chilenos presentan dificultades para resolver problemas complejos que requieren de un pensamiento analítico, integrativo y abstracto. Ellos se equivocan más frente a preguntas con problemas de larga extensión, con más de una variable, con información en distintos formatos no continuos, y con contextos de la vida diaria.

Al parecer, los estudiantes no están acostumbrados a resolver *problemas largos* donde deben analizar información extensa. El error aumenta cuando, además, se trata de *problemas con más de una variable*. Se observa que el razonamiento es adecuado para una variable, pero no logran integrar datos de las otras variables en su respuesta. Los alumnos consideran sólo algunos de los aspectos, pero no todos los que son necesarios.

De esa manera el resultado es incorrecto porque es incompleto. En estas preguntas, se les dificulta analizar un problema desde diferentes ángulos y ser creativos en sus soluciones. Para revertir esta debilidad, las prácticas pedagógicas deberían enfrentar a los alumnos a problemas que puedan resolver a través de varios caminos, como también donde deban integrar sus variables y elementos constituyentes.

Algo similar ocurre cuando el ítem *entrega datos en distintos formatos*. Los ítems de la prueba PISA utilizan contextos de la vida real, como lo es un extracto de un diario de difusión masiva, donde aparecen distintas noticias breves en una misma página y el lector debe seleccionar lo que lee, diferenciar noticias e integrarlas con otras entrevistas en recuadros distintos sobre un mismo tema, cuando corresponda. Frente a este tipo de textos discontinuos (con gráficos, tablas, recortes de narraciones), el alumno no es hábil en seleccionar, discriminar e integrar información para entregar una respuesta adecuada. Una posible explicación de esta dificultad radica en que la forma tradicional de enseñar es más simple, atomizada y focalizada, sin desafiar a los alumnos a enfrentar formatos que integran distinto tipo de información.

Con respecto a los *problemas contextualizados*, se observa que los alumnos no saben cómo proceder para resolver un problema real y modifican la manera de enfrentarlo basándose en su intuición o conocimiento popular, y no en los datos del caso o en el conocimiento curricular. Podría decirse que hay una especie de disociación entre los problemas escolares y los problemas “reales”, que lleva a los alumnos a enfrentar los primeros con contenido académico y los segundos desde el sentido común. Es posible pensar, entonces, que la educación chilena debe apuntar a la integración no sólo de variables sino también de contextos. Dicho de otro modo, debe mostrar el valor de los conocimientos, las fórmulas, la evidencia científica o empírica para dar respuesta a fenómenos y situaciones del diario vivir.

La prueba PISA está construida para evaluar competencias para la vida y, precisamente, la principal dificultad que muestran los alumnos chilenos es en este tipo de preguntas. Los resultados obtenidos hacen pensar que en Chile el proceso de enseñanza-aprendizaje sigue desarrollándose en paralelo a la vida real, es decir, la transmisión de conocimiento sigue siendo descontextualizada: lo que se aprende en el colegio no sirve o no se relaciona con lo requerido para resolver problemas de la vida diaria. Cuando la pregunta de la Prueba PISA pide resolver o explicar una situación cotidiana a través de contenidos de aprendizaje, los alumnos muestran un bajo manejo de contenidos, ya sea porque no se conocen, o bien, porque no se sabe “qué hacer con ellos”. De esta forma, al solicitar un uso “real” de cierto contenido, muchos alumnos fracasan. Es posible hipotetizar que en la medición del mismo contenido, pero en un formato memorístico y no aplicado, los alumnos presentarían menor dificultad.

En esta misma línea, a los estudiantes se les hace más difícil analizar con una mirada científica o basada en evidencia la información disponible. Muestran debilidades cuando se les pide distinguir lo que es generalizable y lo que no, lo que se basa en datos empíricos y lo que es creencia popular. Esto afecta fuertemente el desempeño en Ciencias, pero también en las otras áreas cuando la contextualización del problema tiene este fin.

El análisis permitió identificar el tipo de ítems en que la mayoría de los estudiantes presentaban más errores, como son las preguntas *Open Response* y los ítems *Complex Multiple Choice*. Esta información es relevante de considerar porque posiblemente los ítems de mayor error son los menos familiares. Un dato de interés es que el alumno tiene un bajo desempeño en preguntas *Open Response* en Matemáticas y Ciencias, pero éste no

es un ítem de bajo desempeño en Lectura, es decir, es un ítem familiar para Lectura pero no para las otras dos áreas. Asimismo, en las tres áreas los alumnos obtienen mal rendimiento en el ítem *Complex Multiple Choice*, el que efectivamente no es utilizado en nuestro sistema escolar. Esto podría perjudicar el desempeño de los alumnos chilenos, por lo que es importante comenzar a ocupar este tipo de ítems en las pruebas de aula. Por otro lado, los alumnos presentan mayor éxito en ítems *Multiple Choice*, donde la habilidad cognitiva solicitada implica reconocer información y no promueve la construcción de respuestas, lo que, al parecer, es más común en nuestro sistema escolar.

El análisis de las respuestas correctas e incorrectas de los estudiantes informa sobre las habilidades cognitivas menos utilizadas y nos lleva a reflexionar sobre el tipo de preguntas que se deben incorporar en las pruebas de aula para poder desarrollarlas. Estos resultados permitirán orientar las prácticas docentes y avanzar hacia una educación de calidad.

REFERENCIAS BIBLIOGRÁFICAS

- Aguilar, M. y Cepeda, B. (2005). *PISA para docentes: La evaluación como oportunidad de aprendizaje*. México D.F.: Ediciones INE.
- Anagnostopoulou, K., Hatzinika, V., Christidou, V. & Domopoulos, K. (2013). PISA Test and School-based Examinations in Greece: Exploring the Relationship between Global and Local Assessment Discourses. *International Journal of Science Education*, vol.35 (4), 636-662.
- _____. (2010). Assessed Students' Competencies in the Greek School Framework and the PISA Survey. *Review of Science, Mathematics and ICT Education*, vol.4 (2), 43-61.
- Bayrakdar, Z., Deniz, D., Akgün, L. & Isleyen, T. (2011). Problem Solving Approaches of Mathematics Teacher Candidates in PISA 2003. *Procedia Social and Behavioral Sciences*, vol.15, 3334-3337.
- Bellei, C. (2001). *¿Ha tenido impacto la Reforma Educativa chilena?* Santiago: BID.
- Bravo, A. y Fernández, J. (2000). La evaluación convencional frente a los nuevos modelos de evaluación auténtica. *Psicothema*, vol.12 (2), 95-99.
- Breakspear, S. (2012). The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. *OECD Education Working Papers*, (71), 1-120.
- Cardemil, C. (2001). *Prácticas de enseñanza y aprendizaje en las aulas de enseñanza básica: Una cuenta pendiente de la reforma educativa en Chile*. Santiago: CIDE.
- Colomina, R., Onrubia, J. y Naranjo, M. (2000). Las pruebas escritas y la evaluación del aprendizaje matemático en la educación obligatoria. *Revista Electrónica Interuniversitaria de Formación del Profesorado*, vol.3 (2), 54-69.
- Dann, R. (2002). *Promoting Assessment as Learning: Improving the Learning Process*. London: Routledge Falmer.
- Díaz, V. y Poblete, A. (2009). Perfeccionamiento en matemáticas basado en competencias para docentes de escuelas básicas municipalizadas de la región de Los Lagos y de Los Ríos. *Estudios Pedagógicos*, vol.35 (2), 13-34.
- Espinoza, L., Barbé, J. y Gálvez, G. (2011). Limitaciones en el desarrollo de la actividad matemática en la escuela básica: El uso de la aritmética escolar. *Estudios Pedagógicos*, vol.38 (1), 105-125.
- Goldhaber, D. & Anthony, E. (2007). Can Teacher Quality be Effectively Assessed? National Board Certification as a Signal of Effective Teaching. *Review of Economics and Statistics*, vol. 89 (1), 134-150.
- Gulikers, J., Bastiaens, T. & Kirschner, P. (2004). A Five Dimensional Framework for Authentic

Assessment. *Educational Technology, Research and Development*, vol.52 (3), 67-86.

Hargreaves, E. (2005). Assessment for Learning? Thinking Outside the (Black) Box. *Cambridge Journal of Education*, vol.35 (2), 213-224.

Manzi, J. y Flotts, P. (2007). *Sistema de evaluación del desempeño profesional docente de Chile*. Santiago: MIDE UC.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, vol.29 (1), 23-38.

Merckes, L. (2007). Evaluación y estándares: Logros y desafíos para incrementar el impacto en la calidad educativa. *Revista Pensamiento Educativo*, vol.40 (1), 351-371.

Monereo, C. (2009). *La evaluación auténtica en enseñanza secundaria y universitaria*. Barcelona: Edebé.

_____, Castelló, M., Durán, D. y Gómez, I. (2009). Las bases psicoeducativas del proyecto PISA como guía para el cambio en las concepciones y prácticas del profesorado de secundaria. *Infancia y Aprendizaje*, vol.32 (3), 421-447.

Muñoz, M. & Chang, F. (2007). The Elusive Relationship between Teacher Characteristics and Student Academic Growth: A Longitudinal Multilevel Model for Change. *Journal of Personnel Evaluation in Education*, vol.20 (3-4), 147-164.

Navas, M. J. y Urdaneta, E. (2011). PISA y el triángulo de la evaluación. *Psicothema*, vol.23 (4), 701-706.

OCDE (2010a). PISA 2009 Results. Learning Trends: Changes in Student Performance since 2000. *OECD Publishing*, (5), 1-272.

____ (2010b). PISA 2009 Results. What Students Know and Can Do – Student Performance in Reading, Mathematics and Science. *OECD Publishing*, (1), 1-272.

Pellegrino, J., Chudowsky, N. & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Rico, L. (2006). Marco teórico de evaluación en PISA sobre matemáticas y resolución de problemas. *Revista de Educación*, Número Extraordinario, 275-294.

Sáenz, C. (2009). The Role of Contextual, Conceptual and Procedural Knowledge in Activating Mathematical Competencies (PISA). *Educational Studies in Mathematics*, vol.71 (2), 123-143.

Sanmartí, N. y Sardá, A. (2007). Luces y sombras en la evaluación de competencias: El caso PISA. *Cuadernos de Pedagogía*, (370), 60-63.

Webb, N. (2002). *Alignment Analysis of State F Language Arts Standards and Assessments Grades 5, 8, and 11*. Washington, DC: Council of Chief State School Officers.

Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research & Evaluation*, vol.2 (2), 28-37.

William, D., Lee, C., Harrison, C. & Black, P. (2004). Teachers Developing Assessment for Learning: Impact on Student Achievement. *Assessment in Education: Principles, Policy and Practice*, vol.11 (1), 56-69.

Wyse, A. & Viger, S. (2011). How Item Writers Understand Depth of Knowledge. *Educational Assessment*, vol.16 (4), 185-206.