PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE EDUCACIÓN
PROGRAMA DE MAGÍSTER EN EDUCACIÓN
MENCIÓN EVALUACIÓN DE APRENDIZAJES

DISEÑO, DESARROLLO Y VALIDACIÓN PARA SU USO EN INVESTIGACIÓN, DE UNA PRUEBA PILOTO DE HABILIDADES CUANTITATIVAS PARA ESTUDIANTES SECUNDARIOS

POR

CARLOS ALEXI HERNÁNDEZ DÍAZ

Proyecto de Magíster presentado a la Facultad de Educación
de la Pontificia Universidad Católica de Chile
para optar al grado académico de Magíster en Educación

Profesora Guía:
María Verónica Santelices Etchegaray

Marzo de 2017
Santiago de Chile

A mi Padre

AGRADECIMIENTOS

Quiero agradecer a los académicos, ayudantes y estudiantes de la universidad que directa o indirectamente contribuyeron a la realización de este proyecto. Particularmente, a los profesores revisores Gabriel Muñoz y Dib Atala por sus asertivas y oportunas sugerencias. Y especialmente, a la profesora Verónica Santelices por su orientación constante y por brindarme la oportunidad de participar en el proyecto de investigación Fondecyt # 1160871 "El Rol de la Información en la Toma de Decisiones de los Alumnos en Tránsito a la Educación Superior".

También, quiero agradecer a la Comisión Nacional de Investigación Científica y Tecnológica por la Beca Magister Nacional para Profesionales de la Educación y a la Facultad de Educación de la Pontificia Universidad Católica de Chile por la Beca Matrícula de Facultades. Ambas ayudas facilitaron en este proyecto una enriquecedora experiencia académica y personal.

Finalmente, agradecer a mis padres por su continuo apoyo, comprensión y paciencia.

RESUMEN

El objetivo de este proyecto fue diseñar y desarrollar una prueba piloto de habilidades cuantitativas para estudiantes de tercero y cuarto medio de la provincia de Santiago. Y validarla para su uso en investigaciones educacionales sobre la transición a la educación superior de estudiantes secundarios.

Los métodos para el desarrollo de la prueba comprendieron: el planteamiento de argumentos de interpretación y uso, la revisión de literatura para definir y modelar el constructo, y el uso de criterios estandarizados para construir los ítems y producir la prueba, que fue aplicada a una muestra de 240 casos. En cuanto a la validación, se integraron múltiples evidencias para fundamentar cada argumento de validez.

Los principales resultados fueron: un diseño articulado desde la descripción de niveles de desempeño hasta las especificaciones para la prueba y sus items, y una prueba coherente con su diseño compuesta de 17 ítems de los cuales 14 cumplieron con los criterios psicométricos preestablecidos. Una confiabilidad estimada por un Alpha de Cronbach de 0,67 y un moderado grado de validez para cada interpretación y uso propuesto para los puntajes de la prueba.

Palabras Claves: Validación de Pruebas, Argumento de Interpretación y Uso, Argumento de Validez, Habilidades Cuantitativas.

ABSTRACT

The goal of this project was to design and develop a quantitative skills pilot test for students in 11th and 12th degree of the province of Santiago. And validate it for using in educational research about the transition to higher education of high school students.

The methods to design and develop the test included: posing of interpretation and use arguments, literature review to define and model the construct, and the use of standardized guidelines to write items and produce the test, wich was applied to a sample of 240 cases. About the validation, multiple evidences was integrated to support each validity argument.

The main results was: an articulated design from performance descriptions levels to the specifications of the test and its items, and a test coherent with its design consisting of 17 items of wich, 14 meet the predefined psychometric criteria. A reliability estimated by a Cronbach Alpha of .67, and a moderate degree of validity for each interpretation and use intended for the test scores.

Keywords: Test Validation, Interpretation and Use Argument, Validity Argument, Quantitative Skills.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

I. INTRODUCCIÓN

La Matemática secundaria ha tenido recientes actualizaciones curriculares como el ajuste implementado el año 2009 y las bases curriculares a incorporarse desde el 2017 en primero medio. Ambas coinciden en la agrupación de los contenidos en cuatro ejes temáticos y en su uso para desarrollar habilidades cognitivas simples y complejas.

Aún así, cabe hacerse la pregunta sobre la consistencia que se ha logrado entre la Matemática que se enseña, aprende y evalúa en la escuela y la que se requiere en el quehacer cotidiano, académico y profesional, particularmente para los estudiantes que no prosiguen estudios y profesiones intensivas en formación matemática.

De este cuestionamiento, han surgido conceptos como el alfabetismo cuantitativo y el razonamiento cuantitativo que se enfocan en la aplicación práctica y contextualizada de la Matemática y que pueden eventualmente complementar iniciativas de actualización y enriquecimiento curricular, y en consecuencia inspirar el desarrollo de mediciones con propósitos iniciales de tipo diagnóstico o formativo. O también, con la intención de usar sus puntuaciones, con niveles conocidos de error, para describir a los estudiantes secundarios y eventualmente predecir su acceso y persistencia en la educación terciaria.

Así, surge el desafío que da el título a este proyecto, que pretende contribuir al desarrollo de una prueba piloto de las habilidades cuantitativas subyacentes al alfabetismo y razonamiento cuantitativo, y entendidas como la capacidad para comprender y razonar con datos cuantitativos contextualizados.

Se pretende reportar resultados, conclusiones y recomendaciones que permitan mejorar la prueba para lograr una versión definitiva aplicable a muestras masivas. Considerando metodologías actualizadas para su desarrollo y orientaciones estandarizadas para su validación. Siempre en el contexto de uso de las puntuaciones como variable descriptiva, predictiva o de control en proyectos de investigación relacionados a la trayectoria educacional y transición a la educación superior de estudiantes secundarios.

II. OBJETIVOS

El objetivo general de este proyecto es diseñar, desarrollar y validar para su uso en investigaciones educacionales, una prueba de habilidades cuantitativas para alumnos de 3º y 4º medio en la modalidad científico-humanista, de la provincia de Santiago.

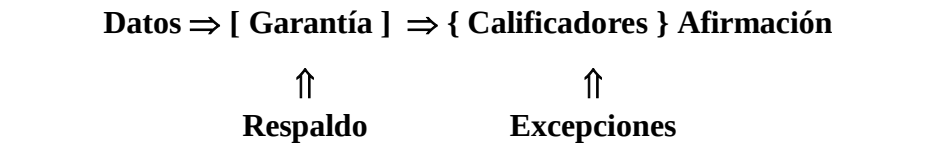
Y los objetivos específicos son:

- Declarar y representar los argumentos de interpretación y uso esperado para las puntuaciones de la prueba.
- Definir y modelar el constructo.
- Especificar y justificar las características de la prueba y de sus ítems.
- Obtener un conjunto de ítems coherentes con sus especificaciones y consistentes con los criterios de construcción.
- Obtener una forma para la prueba piloto coherente con el diseño y consistente con los criterios básicos de producción y documentación.
- Elaborar y fundamentar juicios evaluativos sobre el desempeño psicométrico de los ítems y sobre la estimación de la confiabilidad.
- Elaborar y fundamentar un juicio evaluativo para cada argumento de interpretación y uso.

III. MÉTODOS

El diseño de la prueba comenzó con el planteamiento de los argumentos de interpretación y uso y su representación en diagramas de argumentación del modelo de inferencia de Toulmin presentado en la Figura III.1 (Kane, 2006, 2016; Toulmin, 1958).

Figura III.1. Diagrama de argumentación.



Nota. Este diagrama representa una inferencia en la cual, a partir de un *dato* se hace una *afirmación* con base en una *garantía* (regla para elaborar la inferencia) y su respectivo *respaldo*. Esta garantía puede ser cuantitativa o cualitativamente calificada. Así como la inferencia puede ser restringida bajo ciertas *excepciones* (Kane, 2016).

Luego, con base en los componentes del diseño centrado en evidencia (Mislevy y Haertel, 2006) se definió el constructo a partir de la revisión de: la literatura (Boote y Beile, 2016; Roohr, Graf, y Liu, 2014), el curriculum vigente (MINEDUC, 2009, 2015), y la taxonomía cognitiva actualizada de Bloom (Anderson y Krathwohl, 2001). Posteriormente, se modeló el constructo a través de: las afirmaciones que se harán sobre los examinados, las descripciones de los niveles de desempeño, y el mapa de constructo (Mislevy y Haertel, 2006; Perie, 2008; Perie y Huff, 2016; Wilson, 2005). De esta forma se determinaron y justificaron las características de la prueba y su matriz y tabla de especificaciones.

El desarrollo de la prueba consideró la construcción o adaptación de los ítems orientada por criterios de contenido, formato, estilo, estímulo y opciones (Thomas M. Haladyna y Rogriguez, 2013; Rodriguez, 2016). Y la producción de la prueba fue guiada por los criterios de legibilidad, comprensibilidad y reproductibilidad, además de consideraciones básicas sobre su corrección, reporte, seguridad y documentación (Campion, 2016; S Lane, Raymond, y Haladyna, 2016).

Finalmente, el método de validación se basó en el enfoque de argumentos de validez que proveen una evaluación de las inferencias y supuestos que sustentan cada argumento de interpretación y uso (AERA, APA, y NCME, 2014; Kane, 2016).

Para construir estos argumentos se recogieron los datos aplicando una prueba pre piloto a una muestra de 59 casos, anidados en dos establecimientos, y una prueba piloto a 240 casos anidados en cinco establecimientos, representando aproximadamente un 0,2% de la población objetivo estimada en 96.935 estudiantes[1]. El tipo de muestreo fue aleatorio a nivel de establecimiento, estratificado y no proporcionado por dependencia educacional y por conveniencia a nivel de cursos al interior de cada establecimiento (Thompson, 2012).
En cuanto a la aplicación de las pruebas se siguieron a modo referencial orientaciones estandarizadas para intentar controlar fuentes de varianza irrelevante (McCallin, 2016).

Para el análisis de los datos se utilizaron los métodos que se describen a continuación:

- Para el análisis de los ítems se utilizó el modelo de la teoría clásica de test, bajo el cual se calcularon los estadísticos de: dificultad, discriminación, omisión, distractores y clases de funcionamiento diferencial según procedimiento de Mantel-Haenszel. (T. Haladyna, 2016; Thomas M. Haladyna y Rogriguez, 2013; Meyer, 2014).

- Para el análisis de la confiabilidad, también bajo el modelo de la teoría clásica de test, se calculó el índice Alpha de Cronbach y el error estándar de medición para las puntuaciones y subpuntuaciones de la prueba.

- Para el análisis de la validez de cada argumento de interpretación y uso, se recogieron e integraron múltiples evidencias basadas en:

  - el contenido, evaluando la coherencia de los ítems con sus especificaciones a partir de los juicios de un panel de expertos. Y estimando la representatividad de la prueba sobre el constructo utilizando un cuadro de dos vías de contenidos y habilidades cognitivas (Hogan, 2004).
  - los procesos de respuesta, estimando la capacidad de los ítems para elicitar los procesos cognitivos esperados, a través de una entrevista cognitiva y del análisis de los registros de desarrollo en los cuadernillos de la prueba.

[1] Anexo B.1. Datos de la población y la muestra.

- la estructura interna, estudiando la dimensionalidad del constructo a través de un análisis factorial exploratorio y determinando la cantidad de factores con un análisis paralelo y el gráfico de sedimentación (Cattell, 1966; Horn, 1965). Para luego ejecutar el método de mínimos cuadrados no ponderados y representar los resultados en un diagrama de sendero (IBM, 2011; Jöreskog, 1977; Timmerman y Lorenzo-Seva, 2011).

  Y también estudiando la complejidad cognitiva de los ítems (Schneider, Huff, Egan, Gaines, y Ferrara, 2013) y su relación con la dificultad empírica a través de un modelo de regresión lineal. También se estimó la diferencia entre los porcentajes de logro por categorías de habilidad y contenido, usando tests estadísticos de diferencia de medias para muestras relacionadas.
- las relaciones con otras variables, estimando modelos de regresión y eliminando observaciones influyentes a más de tres desviaciones estándar de la tendencia. Y luego calculando para cada variable criterio el coeficiente de validez desatenuado.
- otros aspectos como, la estandarización de las aplicaciones y la revisión de la calidad constructiva de los ítems con miras a reducir fuentes de varianza irrelevante.

Cabe destacar que la evaluación de los resultados de estos análisis consistió en la emisión de un juicio de valor basado en evidencias sistemáticamente recogidas y contrastadas con criterios predefinidos (Scriven, 1991), fuertemente determinados por el contexto de uso investigativo de las puntuaciones y sus nulas consecuencias para los examinados.

IV. ANTECEDENTES

4.1. Sobre constructos relacionados a las habilidades cuantitativas

Las habilidades cuantitativas requieren de la presentación de tres constructos previos: alfabetismo numérico (numeracy), alfabetismo cuantitativo (quantitative literacy), y razonamiento cuantitativo (quantitative reasoning). Aunque, frecuentemente se usan como sinónimos, una revisión semántica de sus términos devela importantes diferencias conceptuales. Por ejemplo, mientras el alfabetismo hace mención a la capacidad de leer y entender un lenguaje, el razonamiento se refiere a la capacidad de desarrollar lógicamente conclusiones a partir de premisas (VandenBos, 2015).

También se observan diferencias operacionales en la forma en que se han desarrollado estos constructos[2], de las cuales se destacan: la rúbrica de alfabetismo cuantitativo de la Asociación Americana de Universidades[3], las habilidades de alfabetismo cuantitativo enunciadas por la Asociación Americana de Matemática, y la definición de alfabetismo numérico de la Organización para la Cooperación y el Desarrollo Económico (OECD, 2012; Rhodes, 2010; Sons, 1996).

A pesar de su diversidad, estas definiciones tienen en común la aplicación de la Matemática y el pensamiento lógico en contexto, para entender y usar la abundante información cuantitativa presente en nuestro entorno (Steen, 1998; Steen, 2004).

Lo anterior plantea una nueva forma de enseñar y aprender matemática tanto a nivel universitario como escolar, basada en un enfoque más práctico, contextualizado e interdisciplinario (Shavelson, 2008) detallado en la Tabla IV.1.

Incluso, una enseñanza de la Matemática más dirigida por el contexto que por los contenidos, podría lograr mejores resultados en estudiantes que no sigan carreras universitarias relacionadas a la ciencia, tecnología, ingeniería y Matemática (Bennett y Briggs, 2008).

[2] Anexo A.1. Resumen de definiciones de alfabetismo cuantitativo y razonamiento cuantitativo.
[3] Anexo A.2. Rúbrica de alfabetismo cuantitativo de la Asociación Americana de Universidades.

Tabla IV.1. Comparación entre matemática y razonamiento cuantitativo.

Matemática	Razonamiento Cuantitativo
Foco en la abstracción	Foco en la contextualización (auténtica y real)
Énfasis en la generalización	Énfasis en aplicaciones particulares y específicas
Débilmente dependiente del contexto	Fuertemente dependiente del contexto
Métodos algorítmicos	Métodos Ad hoc

Nota. Traducido de "Reflections on Quantitative Reasoning: An assessment Perspective. Calculation vs. Context", Shavelson R. J., 2008, pág. 35.

4.2. Sobre pruebas relacionadas a las habilidades cuantitativas

A nivel internacional se han desarrollado diversos instrumentos para medir el alfabetismo y razonamiento cuantitativo, tal es el caso del Perfil de Desempeño ETS (EPP) aplicado a estudiantes de pregrado en las universidades estadounidenses y la prueba Saber Pro elaborada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), aplicada a los estudiantes al inicio y término de su enseñanza superior (Ver Tabla IV.2).

También se ha desarrollado una propuesta de prueba de alfabetismo cuantitativo basado en la revisión de la literatura y de otros instrumentos relacionados[4] (Roohr et al., 2014).

A nivel nacional la Prueba de Razonamiento Cuantitativo desarrollada por la Pontificia Universidad Católica de Chile es aplicada a los novatos y está relacionada con un curso y un libro homónimo en el que se profundiza su estudio (Mikenberg, 2016).

Otras pruebas miden habilidades similares, pero tienen propósitos formativos o sumativos dirigidos por contenidos curriculares, tal es el caso de la prueba internacional TIMMS y de las pruebas nacionales SIMCE y PSU.

Un caso distinto corresponde a la prueba del Programa Internacional para la Evaluación de los Estudiantes (PISA) que se basa en el concepto de alfabetismo matemático (mathematical numeracy) especificado a partir de procesos y capacidades matemáticas, en combinación con contenidos curriculares y sus contextos de aplicación.

[4] Anexo A.3. Propuesta de prueba de alfabetismo cuantitativo de ETS.

Tabla IV.2. Pruebas de alfabetismo cuantitativo y razonamiento cuantitativo.

	EPP	**SABER PRO**
Nombre de la Prueba	ETS Proficiency Profile (Mathematics)	Saber Pro (Razonamiento Cuantitativo)
Desarrollador	ETS	ICFES
Formato	Opción Múltiple	Opción Múltiple
Aplicación	Papel y Lápiz/Computador	Papel y Lápiz
Tiempo	30 minutos aprox.	60 minutos aprox.
Cantidad de Ítems	27 ítems	35 ítems
Información sobre objetivos de medición	Mide la habilidad para; reconocer e interpretar términos matemáticos; leer e interpretar tablas y gráficos; evaluar fórmulas; ordenar y comparar números grandes y pequeños; interpretar razones, proporciones y porcentajes; leer instrumentos científicos de medición; reconocer y usar fórmulas o expresiones matemáticas equivalentes[a].	Mide las competencias de: comprender y manipular representaciones de datos cuantitativos o de objetos matemáticos, en distintos formatos (textos, tablas, gráficos, diagramas, esquemas); establecer, ejecutar y evaluar estrategias para analizar o resolver problemas; y justificar o dar razón de afirmaciones o juicios a propósito de situaciones que involucren información cuantitativa u objetos matemáticos[b].
Información adicional	Alpha de Cronbach igual a 0,85. Correlación entre 0,14 y 0,27 con promedio de notas en la universidad[c]. Correlación de 0,26 con los créditos aprobados durante el primer semestre en la universidad[d].	La prueba de razonamiento cuantitativo explica el 24% de la varianza del promedio de notas en el primer semestre de universidad[d].

Notas. Información extraida de las páginas web: [a] www.ets.org. [b] www.icfes.gov.co. Y de las publicaciones; [c] Roohr, Liu y Liu 2017, pág. 6; [d] Lakin, Cardenas y Liu, 2012, pág. 746; e [e] ICFES, 2014, pág. 11.

De esta revisión se concluye que estos instrumentos han sido desarrollados principalmente para estudiantes universitarios cuya formación está más directamente relacionada con los conocimientos y habilidades demandas en entornos laborales, pero que lógicamente se pueden traducir en exigencias a la educación secundaria, evidenciándose que "mediciones de calidad deben ser desarrolladas para identificar las fortalezas y debilidades de los estudiantes en alfabetismo cuantitativo cuando entren a la universidad" (Roohr, Graf, y Liu, 2014, p. 2). Además de poder utilizar estas mediciones para analizar su relación con el rendimiento académico en la universidad (ICFES, 2014; Lakin, Elliott, y Liu, 2012; Roohr, Liu, y Liu, 2017).

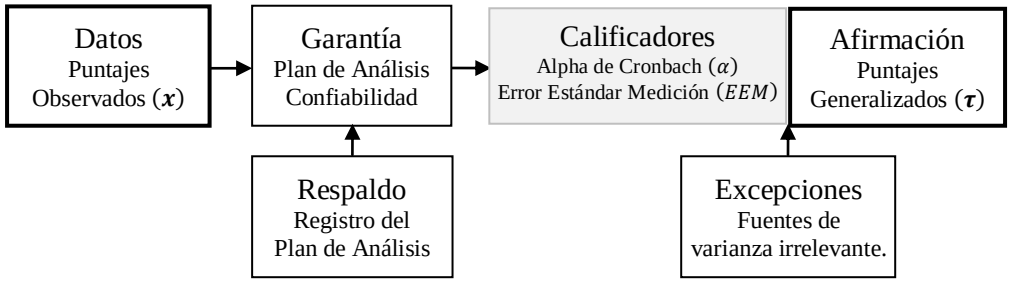
V. DISEÑO Y DESARROLLO DE LA PRUEBA

5.1. Argumentos de interpretación y uso

La prueba pretendió medir el constructo de las habilidades cuantitativas, analizado y modelado en extenso en la Sección 5.2. Y su propósito fue de tipo investigativo, ya que buscó recopilar y analizar sistemáticamente datos para describir estas habilidades y analizar su relación con otras variables de interés (Hernández, Fernández, y Baptista, 2006; Hogan, 2004).

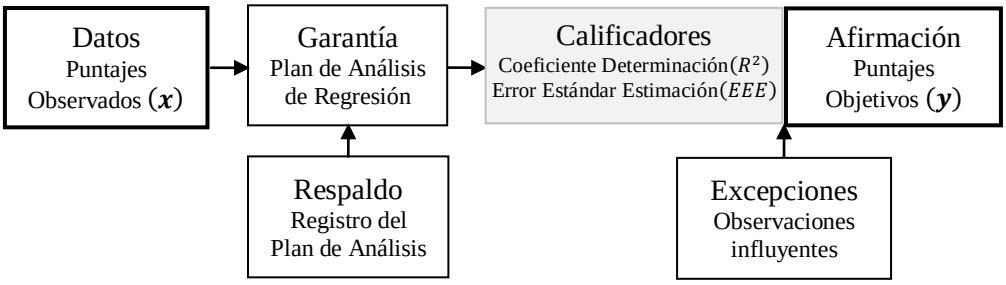
La primera interpretación esperada correspondió a una inferencia de generalización, que relacionó el puntaje del examinado con su puntaje esperado en un universo de condiciones de observación (Kane, 2016). Estableciendo así, un uso descriptivo de caracterización de los estudiantes de la muestra. La garantía de la inferencia estuvo en el plan de análisis de confiabilidad, calificado con el índice Alpha de Cronbach y el error estándar de medición teniendo en cuenta que fuentes no controladas de varianza irrelevante del constructo (Messick, 1989) podían afectar la estimación (Ver Figura V.1).

Figura V.1. Argumento de interpretación y uso descriptivo.



La segunda interpretación correspondió a una inferencia de extrapolación, que relacionó el puntaje del examinado con el puntaje en otras pruebas de interés representativas de un dominio objetivo más amplio (Kane, 2016). Estableciendo así, un uso predictivo de estimación del desempeño en la PSU de Matemática. Para esto, se desarrolló un modelo de regresión lineal, calificado con su coeficiente de determinación y error estándar de estimación, teniendo en cuenta que observaciones influyentes podían afectar el ajuste del modelo (Ver Figura V.2).

Figura V.2. Argumento de interpretación y uso predictivo.



5.2. Definición y modelamiento del constructo

Una revisión detallada de los constructos de alfabetismo numérico, alfabetismo cuantitativo y razonamiento cuantitativo se presenta en la Tabla V.1 (Vacher, 2014).

Tabla V.1. Significados de constructos relacionados con las habilidades cuantitativas.

		Alfabetismo Numérico	**Alfabetismo Cuantitativo**	**Razonamiento Cuantitativo**
S.1	Habilidad con los números y las matemáticas.	X		
S.2	Aptitud para leer, escribir y entender materiales que incluyen información cuantitativa tales como; gráficos, tablas, relaciones matemáticas y estadísticas descriptivas.	X	X	
S.3	Pensamiento lógico y coherente involucrando información cuantitativa tales como relaciones matemáticas y estadísticas descriptivas.		X	X
S.4	Disposición a abordar, en lugar de evitar, información cuantitativa usando las habilidades matemáticas y el conocimiento estadístico propio, en una forma reflexiva y lógica para tomar decisiones apropiadas.	X	X	X

Nota. Traducido de "Looking at the multiple meanings of Numeracy, Quantitative Literacy and Quantitative Reasoning", Vacher, H. L., 2014, pág. 11. S.1, S.2, S.3 y S.4 corresponden a la numeración de los significados asociados a estos constructos.

Descartando la disposición al uso de habilidades matemáticas en la vida diaria (S.4) que es común a los tres constructos y la habilidad numérica (S.1) que solo está presente en el primero de ellos, se pudo asociar alfabetismo cuantitativo y razonamiento cuantitativo con procesos cognitivos simples (S.2) y complejos (S.3), respectivamente. Los que a su vez se relacionan con datos cuantitativos y sus representaciones matemáticas.

Con base en este análisis y en los antecedentes previos, se definió la habilidad cuantitativa como: la capacidad para comprender y razonar con datos cuantitativos contextualizados, representados en forma de; cantidades y modelos matemáticos, medidas geométricas, tablas y gráficos estadísticos.

Las capacidades de comprender y razonar hicieron alusión a procesos cognitivos de baja complejidad (comprender y aplicar) y de alta complejidad (analizar y evaluar). Considerando las habilidades cognitivas de: interpretar, representar, ejecutar, discriminar-integrar y comprobar, según la taxonomía revisada de Bloom[5] (Anderson y Krathwohl, 2001).

Las formas en que se representan los datos cuantitativos se vincularon a los contenidos matemáticos correspondientes a la enseñanza secundaria (MINEDUC, 2009, 2015), seleccionado los tópicos con mayor potencial de contextualización agrupados en cuatro ejes temáticos (Números, Álgebra, Geometría y Datos y Azar)[6]. El detalle de estos contenidos, confirmó que conceptos aprendidos hasta segundo medio pueden servir como base para pruebas de alfabetismo y razonamiento cuantitativo (Roohr et al., 2014). También se observó una selección de contenidos similares en un texto de estudio universitario de razonamiento cuantitativo (Bennett y Briggs, 2008).

La contextualización enfatizó la importancia de evidenciar estas habilidades y contenidos en situaciones cotidianas[7].

La combinación de categorías de habilidades cognitivas (Comprensión y Razonamiento) y de ejes temáticos (Números-Álgebra-Geometría y Datos-Azar) permitieron plantear las afirmaciones generales que se pretendió hacer sobre los examinados a partir de sus puntuaciones, representadas en la descripción de niveles de desempeño (Perie, 2008; Perie y Huff, 2016) de la Tabla V.2.

[5] Anexo B.2. Detalle de las habilidades cognitivas de la prueba.
[6] Anexo B.3. Detalle de los contenidos matemáticos de la prueba.
[7] Anexo B.4. Detalle de los contextos utilizados en la prueba.

Tabla V.2. Descripción de niveles de desempeño.

	Habilidades Cuantitativas
Nivel 3	El estudiante es capaz de comprender y razonar con datos cuantitativos contextualizados, expresados en: cantidades, modelos matemáticos, medidas geométricas, tablas y gráficos estadísticos.
Nivel 2	El estudiante es capaz de comprender datos cuantitativos contextualizados, expresados en al menos tres de las siguientes formas: cantidades, modelos matemáticos, medidas geométricas, tablas y gráficos estadísticos.
Nivel 1	El estudiante es capaz de comprender datos cuantitativos contextualizados, expresados en a lo más dos de las siguientes formas: cantidades, modelos matemáticos, medidas geométricas, tablas y gráficos estadísticos.

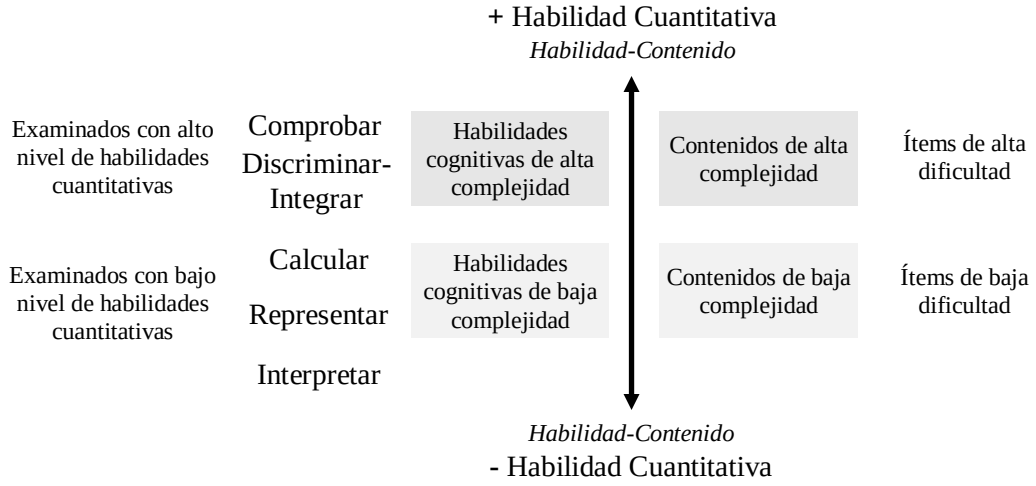
Más detalladamente, la combinación de habilidades cognitivas y contenidos específicos, permitió elaborar un mapa de constructo[8] y redactar objetivos de evaluación más acotados (Wilson, 2004).

Si bien este modelamiento correspondió al de una prueba de aprovechamiento educacional, dado que la habilidad cuantitativa demanda un grado intermedio de entrenamiento específico, se agregaron supuestos teóricos sobre su estructura interna, como en el caso del modelamiento de una prueba de capacidad mental. Esto fue coherente con la ubicación de la prueba en una rango intermedio del continuo de capacidad–aprovechamiento[9] (Hogan, 2004).

Por esto, se planteó un primer supuesto sobre el comportamiento unidimensional del constructo, asumiendo que la construcción de cada ítem se orientó hacia una habilidad y un contenido de similar nivel de complejidad. Es decir, se propuso que la habilidad cuantitativa sería el único factor necesario y suficiente para ordenar a los examinados y a los ítems, y a su vez explicar su variabilidad, esto se representa en la Figura V.3, con base en la estructura de un mapa genérico de constructo (Burke, Mattar, Stopek, y Eve, 2014; Wilson, 2005).

[8] Anexo B.5. Mapa de constructo para las habilidades cuantitativas.
[9] Anexo B.6. La prueba en el continuo de capacidad–aprovechamiento.

Figura V.3. Propuesta de unidimensionalidad para el constructo.



El segundo supuesto planteó que la dificultad empírica de los ítems debería asociarse positivamente con su complejidad cognitiva, asociada al contenido y a la demanda cognitiva y lingüística requerida para responder correctamente (Schneider et al., 2013). Además se propuso una diferencia significativa entre los porcentajes de logro promedio de los ítems de comprensión cuantitativa y razonamiento cuantitativo.

5.3. Especificaciones de la prueba y de los ítems

La prueba de habilidades cuantitativas se clasificó como una prueba aplicada grupalmente y en formato de lápiz y papel (Hogan, 2004) y sus especificaciones se presentan y justifican en la Tabla V.3.

Cabe mencionar que para hacer un uso eficiente del tiempo de aplicación disponible se descartó una mayor cantidad de ítems y el uso de ítems de construcción de respuesta. Además, la eficiencia temporal de los ítems de opción múltiple (Downing y Haladyna, 2006; T M Haladyna y Downing, 1993; Hogan, 2004) y su potencial para recoger evidencias de razonamiento complejo (Huff, Steinberg, y Matts, 2010) fueron argumentos adicionales a favor de su uso.

Tabla V.3. Especificaciones de la prueba.

Característica	Descripción	Criterio	Justificación
Cantidad de Ítems	17 ítems.	Representatividad del dominio de ítems.[a]	Muestreo con al menos un ítem en cada combinación de habilidad y categoría de eje temático.
		Tiempo de aplicación disponible.	Estimación del tiempo máximo de respuesta cercano al tiempo disponible de aplicación.
Tipo de Ítems	De selección de respuesta, con opción múltiple simple.	Eficiencia temporal.[a]	Recogida de información de diversos objetivos de evaluación en menos tiempo que con preguntas de construcción de respuesta.
		Confiabilidad y eficiencia en la calificación.[a]	Rapidez y reducido margen de error en la corrección de las respuestas.
	Cuatro opciones, una clave y tres distractores.	Suficiencia de distractores.[b]	Cantidad de distractores suficientes para tener frecuencias no nulas.
Puntuación de los Ítems	Dicotómica (desde la teoría clásica de tests), 1 punto por respuesta correcta y 0 por respuesta incorrecta u omitida.	Viabilidad del modelo psicométrico.	Puntuación no representa exigencias de tamaño muestral y es fácil de calcular, interpretar y comunicar.
Puntuación de la Prueba	Puntuación natural equivalente a la suma de los puntajes obtenidos en cada ítem.		
Tiempos de Respuesta	35 minutos como máximo.	Tiempo de respuesta por ítem.	Correspondencia con el tiempo promedio de respuesta, de 2 minutos por ítem, según pruebas pre piloteadas.
		Tiempo de aplicación disponible.	Correspondencia con el tiempo de aplicación disponible en los establecimientos.

Nota. La mayoría de los criterios han sido extraídos de las siguientes referencias bibliográficas.

[a] "Pruebas Psicológicas: Una Introducción Prática", Hogan T., 2015, Cap. 6.

[b] "How many options is enough for a multiple choice item?", Haladyna T. y Downing S., 1993.

La cantidad de ítems en cada categoría de habilidad se estableció considerando el nivel de enseñanza de los examinados y en cada categoría de eje temático considerando el potencial de contextualización de sus contenidos específicos. El orden de los ítems fue por eje temático (desde Números hasta Azar) y por habilidad cognitiva (desde representar hasta comprobar). Esto se tradujo en la respectiva matriz y tabla de especificaciones[10]. Las especificaciones de los ítems incluyeron estos mismos componentes más el objetivo de evaluación, mientras que el contexto fue libremente elegido por el constructor-editor según su pertinencia para el objetivo establecido[11].

[10] Anexo B.7. Matriz y tabla de especificaciones.

[11] Anexo B.8. Ejemplo de ficha de construcción de un ítem.

5.4. Construcción de los ítems y producción de la prueba

La construcción y adaptación de los ítems siguió tres principios básicos; tener un contenido relevante, no regalar la respuesta y redactar de manera sencilla y clara (Hogan, 2004). Además de los criterios de construcción específicos categorizados y descritos a continuación:

→ El contenido evitó conocimientos conceptuales y procedimentales complejos.
→ El formato fue vertical para presentar los ítems y sus opciones, excepto en los casos de múltiples gráficos que se presentaron ordenados rectangularmente.
→ El estilo al igual que el contenido evitó lenguaje innecesariamente complejo, por ejemplo sustituyendo la palabra "ingreso per cápita" por "ingreso por persona". También se minimizó la cantidad de lectura en cada ítem restringiéndose la cantidad de palabras en el enunciado y en las preguntas a 45 y 20, respectivamente.
→ Los estímulos frecuentemente usados dada la naturaleza contextualizada de las habilidades cuantitativas fueron principalmente: gráficos y expresiones analíticas de modelos matemáticos, cuerpos geométricos, tablas y gráficos estadísticos. En su mayoría hicieron alusión a datos reales o ficticios de situaciones tales como: las ganancias de una inversión, las probabilidades de fallecer en un accidente de tránsito o el desempeño académico en pruebas de aula o estandarizadas. En ítems específicos, se hicieron modificaciones menores a estos datos para viabilizar algunos distractores.
→ Las distractores fueron justificados con base en los errores esperados de los examinados.

La construcción de los ítems se inició elaborando una idea de ítem (Wilson, 2005) a partir de un contexto, para luego buscar un estímulo que en interacción con el enunciado y las opciones permitiera escribir el ítem, citando la fuente de los datos utilizados y justificando los distractores propuestos. A pesar de esto, la cercanía entre los objetivos de evaluación de esta prueba y los de otras pruebas desarrolladas, se tradujo en una baja originalidad en los ítems construidos[12].

[12] Anexo B.9. Análisis de originalidad de los ítems.

La adaptación de los ítems a partir de un banco de ítems confidenciales (Pontificia Universidad Católica de Chile, 2014), incluyó ajustes menores de estilo y el cambio de formato de respuesta construida a respuesta seleccionada, citando debidamente la fuente del ítem en su ficha de construcción.

La producción de la prueba, se inició con el proceso de ensamblaje definido por la tabla de especificaciones, ordenando los ítems y revisando la presencia equilibrada de la letras correspondientes a las claves. Para reducir la posibilidad de trampas se diseñaron dos cuadernillos con los mismos ítems pero en distinto orden.
La diagramación consideró: la facilidad para reconocer letras y estímulos; la facilidad para leer palabras, frases y estímulos; y la calidad con que se reprodujo el material (Campion, 2016).

Las aplicaciones de la prueba tuvieron en cuenta orientaciones estandarizadas para intentar controlar fuentes de varianza irrelevante (AERA et al., 2014) resumidas en un protocolo de aplicación. En cuanto a la corrección, las respuestas fueron tabuladas manualmente y corregidas computacionalmente, para controlar la calidad de este proceso se observó la coherencia entre los desarrollos escritos y las respuestas seleccionadas en el 88% de los cuadernillos piloteados.

Los reportes de la prueba fueron excepcionales y breves, e informaron el puntaje promedio y los porcentajes de logro por categoría de habilidad cuantitativa a nivel de establecimiento.

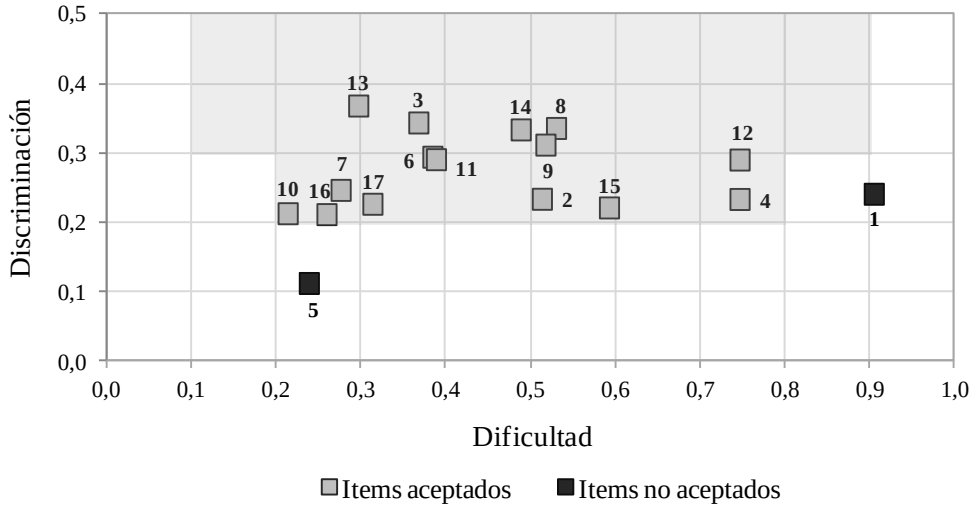
Para contribuir a la seguridad de la prueba se aplicaron, retiraron y registraron cuadernillos debidamente foliados y los archivos digitales con los resultados fueron encriptados. Finalmente, los anexos de este informe constituyen un respaldo de la documentación del proceso de desarrollo de la prueba.

VI. VALIDACIÓN DE LA PRUEBA PARA SU USO EN INVESTIGACIÓN

6.1. Evaluación de los ítems

Después de un proceso iterativo de evaluación[13,14,15,16] se aprobaron los 14 ítems que cumplieron los criterios de desempeño psicométrico preestablecidos[17]. En la Figura VI.1 se presentan los números de los ítems aceptados y no aceptados en términos de dificultad y discriminación.

Figura VI.1. Dificultad y discriminación de los ítems.



El porcentaje promedio de omisión fue de 4,7%, observándose en los ítems Nº 5 y Nº 17 la omisión más alta, del orden del 8% y 9%, respectivamente[18]. Cabe mencionar, que este último ítem es el único correspondiente a contenidos curriculares de cuarto medio.

Además todos los distractores tuvieron una correlación biserial puntual negativa con el puntaje de la prueba y una frecuencia promedio inferior a la frecuencia de la clave (excepto en el ítem Nº 10). Y de los 51 distractores de tenía la prueba se observaron 5 distractores vacíos con una frecuencia inferior al 5%.

[13] Anexo C.1. Proceso de evaluación de los ítems.
[14] Anexo C.2. Revisión de jueces sobre la calidad constructiva de los ítems.
[15] Anexo C.3. Resultados de test y pauta de entrevista cognitiva.
[16] Anexo C.4. Características y resultados de aplicaciones pre piloto.
[17] Anexo C.6. Criterios de evaluación del desempeño psicométrico de los ítems.
[18] Anexo C.5. Características y resultados de aplicaciones piloto.

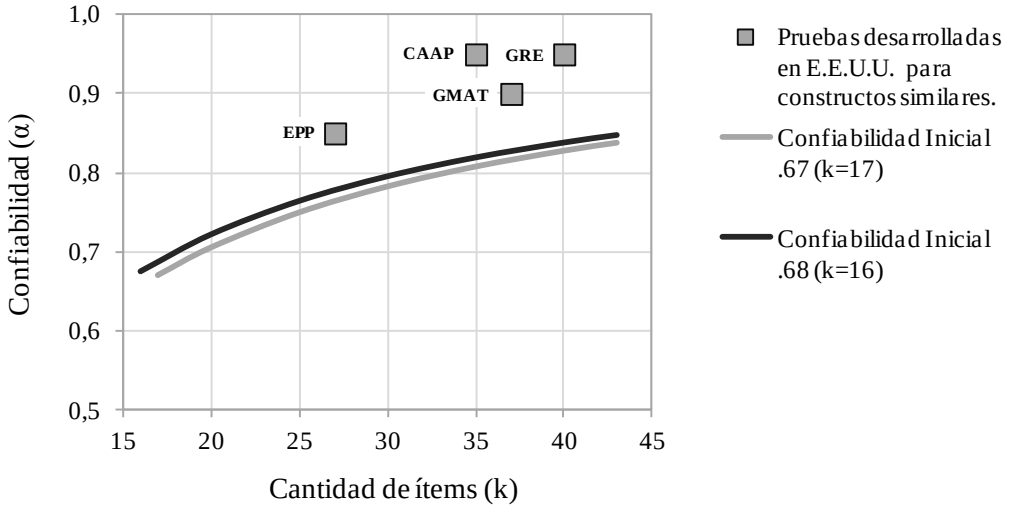
Complementariamente, se observó funcionamiento diferencial clase C a favor de los examinados de género femenino en el ítem Nº 12, lo que se tradujo en su eliminación de la prueba (Zwick y Ercikan, 1989).

6.2. Evaluación de la confiabilidad

El índice Alpha de Cronbach fue de 0,67, con un intervalo de confianza al 95% de [0,61; 0,73]. El error estándar de medición fue de 1,78 puntos y la banda de confianza al 95% para el puntaje promedio de 7,8 puntos, fue [4,8; 10,7]. Cabe recordar que la escala de la prueba va desde 0 a 17 puntos[19].

Este índice implica que duplicando la cantidad de ítems se alcanzaría una estimación de confiabilidad de 0,8. En la Figura VI.2 se grafica la proyección de este índice en función de la cantidad de ítems y se compara con los índices de pruebas diseñadas para medir un constructo similar en estudiantes universitarios estadounidenses (Roohr et al., 2014).

Figura VI.2. Resultado, proyección y comparación de la confiabilidad.



Considerando la amplitud de las habilidades y contenidos modelados, la extensión de la prueba y sobre todo las nulas consecuencias para los examinados, se evaluó este índice como un valor mínimo aceptable para dar inicio al análisis de validez.

[19] Anexo D.1 Proceso de la evaluación de la confiabilidad.

Se reportan adicionalmente, los índices de confiabilidad y error estándar de medición para cada subpuntuación de categoría de habilidades y de ejes temáticos, además del impacto en el Alpha de Cronbach de la eliminación de cada ítem[20].

Cabe destacar que la estimación de confiabilidad para los puntajes de los examinados de tercero y cuarto medio fue de 0,65 y 0,72, respectivamente.

6.3. Evaluación de la validez

Los tipos de evidencias recolectadas dependieron de las inferencias y supuestos considerados en cada interpretación y uso esperado de las puntuaciones (AERA et al., 2014. p. 12; Kane, 2006, 2016).

Así, para la inferencia de generalización se recogieron evidencias sobre: la coherencia de los ítems con sus especificaciones, la representatividad de la prueba sobre el constructo, la capacidad de los ítems para elicitar los procesos cognitivos esperados y el respaldo empírico a los supuestos teóricos de dimensionalidad del constructo y de dificultad de los ítems.

Mientras que, para la inferencia de extrapolación se recogieron evidencias sobre la significancia e intensidad de la asociación entre los puntajes de la prueba y las variables criterio[21] y sobre la representatividad de la prueba en el dominio del criterio principal, la PSU de Matemática.

Evidencias basadas en el contenido

La coherencia de los ítems con sus especificaciones y con el constructo, fue evaluada por dos jueces revisores que aprobaron 22 de los 27 ítems presentados inicialmente. Un segundo grupo de 6 ítems, fue presentado a un juez revisor, que los aprobó sin observaciones sobre este aspecto.

[20] Anexo D.2. Estimación de la confiabilidad y otros resultados.
[21] Anexo E.1. Proceso de evaluación de la validez.

La representatividad de los ítems sobre el dominio de ítems del constructo (Hernández et al., 2006), se tabuló en un cuadro de dos vías, donde cada celda correspondió a una combinación de habilidad cognitiva y contenido específico (Hogan, 2004).

La subrepresentación más importante se produjo en el eje de Números, sin ítems relacionados a contenidos de enseñanza media, y en el eje de Álgebra sin ítems de ecuaciones e inecuaciones lineales, tal como se muestra en la Tabla VI.1.

Tabla VI.1. Representatividad de la prueba sobre el dominio de ítems.

			Habilidades cognitivas						
E	**Contenidos específicos**	**NE**	Interpretar	Representar	Calcular	Dicriminar-Integrar	Comprobar		
Números	Prop., porcentajes y aproximaciones	NB			Item 1	Item 2	Item 3	60%	
	Números racionales	1º							
	Potencias de exponente entero	1º							
	Exponenciales y logaritmos	2º							**15%**
Álgebra	Ec. de primer grado y sistemas	2º							
	Inec. de primer grado y sistemas	4º							
	Ecuaciones de segundo grado	3º							
	Función lineal y afín	1º	Item 4[a]					20%	
	Función exponencial y logartímica	2º		Item 6				20%	
	Función cuadrática	3º				Item 7		20%	
	Función potencia	4º			Item 5[b]			20%	**11%**
Geo.	Semejanza de figuras planas	2º							
	Cuerpos geométricos	4º			Item Nº 8[c]			20%	**10%**
Datos y azar	Tablas y gráficos estadísticos	1º	Item 9	Item 10	Items 11 y 12		Item 15	80%	
	Medidas de tendencia central	1º			Item 13			20%	
	Medidas de posición	1º							
	Medidas de dispersión	2º					Item 16	20%	
	Combinatoria	1º							
	Probabilidad teórica y experimental	1º			Item 14			20%	
	Probabilidad condicional	2º							
	Distribución normal	4º				Item 17		20%	
	Intervalos de confianza	3º							**20%**
			9%	**9%**	**32%**	**14%**	**14%**	**12%**	

Notas. La columna **E** se refiere al eje temático, la columna **NE** especifica el nivel de enseñanza media correspondiente al contenido según el curriculum vigente (NB indica que el contenido corresponde a enseñanza básica). Los porcentajes en negrita corresponden a la cobertura de la prueba sobre cada eje temático y habilidad cognitiva.

[a] Estrictamente se trata de un ítem de una función definida por tramos (contenido no curricular) pero considerado cercano a la función afín y a la función parte entera que era curricular antes del ajuste del año 2009.

[b] Podría clasificarse como un tema de función exponencial si se considera que la variable corresponde el exponente.

[c] Si bien se trata de un poliedro sencillo que se estudia en enseñanza básica, se puede considerar también pertenenciente a la unidad de geometría tridimensional de 4º medio.

También se puede afirmar que la prueba captura el 12 % de los objetivos de evaluación posibles, de tal forma que todas las habilidades cognitivas y ejes temáticos están representados en la prueba con un porcentaje cercano o mayor al 10 % de las posibilidades teóricas.

Estas evidencias constituyeron un sustento moderado a favor de los supuestos de coherencia y representatividad de los ítems de la prueba.

Evidencias basadas en los procesos de respuesta

En la entrevista cognitiva[22], el examinado de mayor desempeño resolvió correctamente un ítem evidenciando parcialmente el proceso cognitivo esperado, mientras que el examinado de menor desempeño lo respondió incorrectamente a través de un procedimiento no matemático. Esto permitió ratificar las características deseables de los distractores y vislumbrar el mismo análisis sobre los desarrollos registrados en los cuadernillos de la prueba[23] (AERA et al., 2014).

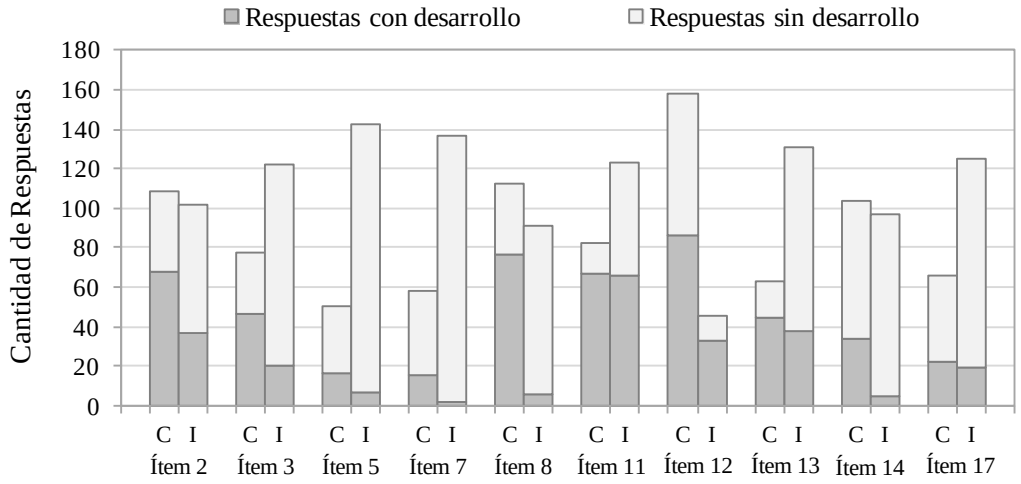
Al analizar 212 cuadernillos de un total de 240, se observó que el 54% de las respuestas correctas y el 21% de las respuestas incorrectas, tenían un desarrollo coherente con la opción seleccionada. En la Figura VI.3 se presenta el detalle a nivel de ítems.

Estos datos constituyeron una fuerte evidencia a favor de la capacidad de la prueba para elicitar los procesos cognitivos esperados en el caso de las respuestas correctas. Para las respuestas incorrectas la evidencia fue más débil ya que para un porcentaje mayoritario de los distractores seleccionados no se observaron registros de los errores esperados.

Estas evidencias junto a las basadas en el contenido constituyeron evidencias de alineación esenciales para el argumento de interpretación y uso descriptivo que pretendió indicar el grado de dominio de los examinados sobre los contenidos y procesos cognitivos modelados (Wise y Plake, 2016).

[22] Anexo C.3. Resultados de test y pauta de entrevista cognitiva.
[23] Anexo E.2. Procesos de respuesta incorrecta y correcta para un ítem.

Figura VI.3. Resultado de análisis de respuestas desarrolladas en los cuadernillos.



Nota. En cada ítem, la primera columna representa la cantidad de respuestas correctas (C) con y sin desarrollo en los cuadernillos, de igual forma la segunda columna representa la cantidad de respuestas incorrectas (I). Los ítems Nº 1, Nº 4, Nº 6, Nº 10, Nº 15 y Nº 16 se excluyeron del análisis porque responderlos correctamente no demanda necesariamente un desarrollo escrito.

Evidencias basadas en la estructura interna

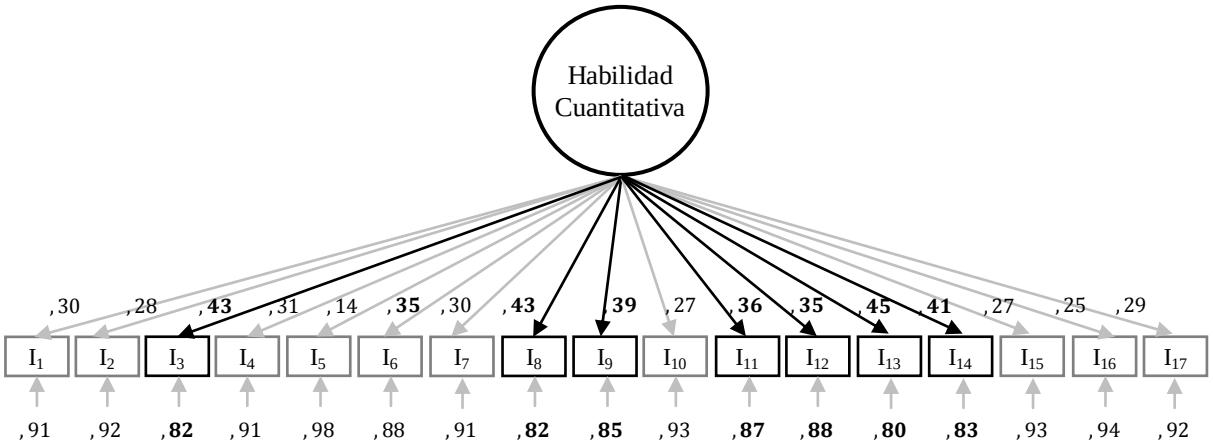
En cuanto a la dimensionalidad, se verificó previamente que varios ítems no cumplieron con el principio constructivo de evaluar un contenido y una habilidad de complejidad similar[24].

Para iniciar el análisis factorial se obtuvo un índice KMO de 0,69, por debajo del mínimo regular sugerido de 0,7 (Hair y Suárez, 1999) y una sugerencia consistente de extraer un factor[25].

Teniendo en cuenta que la varianza explicada según los autovalores fue de 16,4%, por debajo del 50% sugerido (Merenda, 1997), la ejecución del método de Mínimos Cuadrados no Ponderados reportó los resultados[26] resumidos en la Figura VI.4.

[24] Anexo E.3. Índice de complejidad cognitiva y dificultad empírica de los ítems.
[25] Anexo E.4. Gráficos de análisis paralelo y de sedimentación.
[26] Anexo E.5. Resultados de análisis paralelo y factorial con mínimos cuadrados no ponderados.

Figura VI.4. Diagrama para resultados del análisis factorial exploratorio.



Nota. La circunferencia representa el factor extraído denominado Habilidad Cuantitativa, cada recuadro representa un ítem y sobre él se anota la carga factorial y debajo la varianza única. En negrita están las cargas factoriales mayores o iguales que 0,35 y las varianzas únicas menores o iguales que 0,87.

Las cargas factoriales de los ítems promediaron 0,33 no superando el mínimo recomendado de 0,35 (Pérez y Medrano, 2010) y el promedio de comunalidades de 0,11 refleja que el modelo representa más error, que verificación de la unidimensionalidad del constructo (Thissen y Wainer, 2001, p.197). Es decir, no hay evidencia empírica de que las habilidades cuantitativas sean el factor subyacente que explica la variabilidad de los ítems.

En cuanto a la dificultad, para cada ítem se calificó la complejidad del contenido, la habilidad cognitiva y el texto. Con esto se calculó un índice teórico de complejidad cognitiva con el cual se estimó un modelo de regresión estadísticamente significativo, $R^2 = 0{,}32$; $F(1{,}15) = 7{,}3$; $p = 0{,}016$, con una moderada capacidad para predecir la dificultad empírica de los ítems[27].

También, se estimó una diferencia de 11 puntos entre las medias de porcentajes de logro para los ítems de comprensión cuantitativa y razonamiento cuantitativo ($p < 0{,}001$) confirmando el supuesto de que ítems de categorías cognitivas más complejas deben ser en promedio, más difíciles. Y una diferencia de 5 puntos entre las categorías de ejes temáticos ($p < 0{,}001$) cuestionando el planteamiento de que dichas categorías no afectan la dificultad promedio.

[27] Anexo E.6. Análisis de la relación entre la complejidad cognitiva y la dificultad empírica de los ítems.

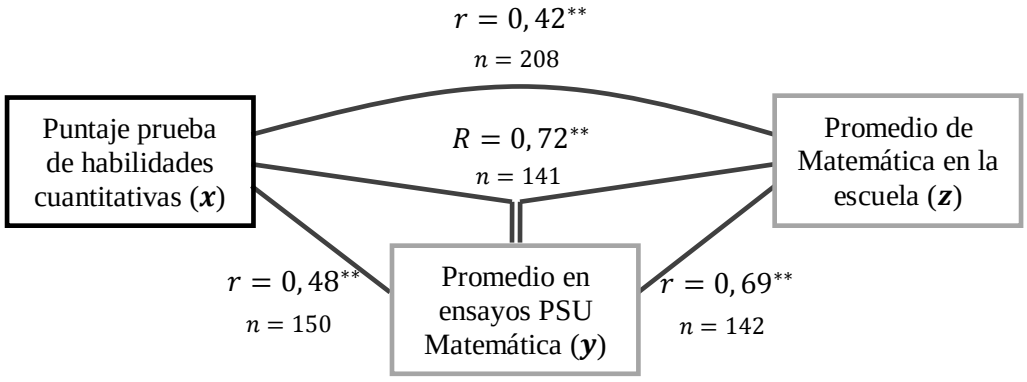
Cabe agregar, que no se observó una diferencia significativa entre las medias de los puntajes para cada cuadernillo aplicado, lo que respalda el supuesto de que serían equivalentes en términos de dificultad[28].

Estos resultados no permitieron sustentar la estructura unidimensional propuesta, ni confirmar la predictibilidad de la dificultad de los ítems a partir de su complejidad cognitiva. Estableciendo una débil evidencia a favor de estos supuestos.

Evidencias basadas en las relaciones con otras variables

Los coeficientes de correlación se calcularon entre las siguientes variables: puntaje en la prueba de habilidades cuantitativas (***x***), promedio estimado y auto-reportado de puntajes en ensayos PSU de Matemática (***y***), y promedio auto-reportado de notas de Matemática en la escuela durante el primer semestre (***z***). Después de eliminar las observaciones influyentes[29], se obtuvieron los resultados resumidos en la Figura VI.5, que indican una "moderada asociación positiva" (Hernández et al., 2006, p. 453) entre la prueba y los criterios.

Figura VI.5. Diagrama para la asociación entre la prueba y los criterios.



Nota. r corresponde al coeficiente de correlación simple, R al coeficiente de correlación múltiple y n al número de casos válidos para el cálculo de estos coeficientes. ** $p < 0{,}001$ (unilateral).

[28] Anexo E.7. Análisis de diferencias en porcentajes de logro y puntajes totales.
[29] Anexo E.8. Matriz de correlaciones y error estándar de medición antes y después de eliminar observaciones influyentes.

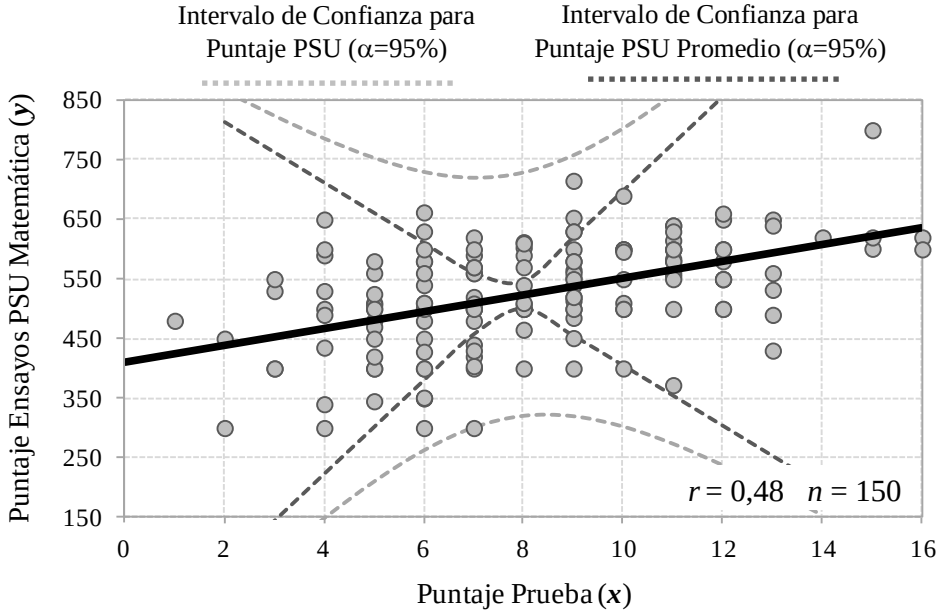
El primer modelo de regresión entre las variables $\boldsymbol{x}$ e $\boldsymbol{y}$ fue estadísticamente significativo[30], $R^2 = 0{,}23\ ;\ F(1{,}148) = 44{,}7\ ;\ p < 0{,}001$, y quedó determinado por la ecuación

$$\boldsymbol{y}(\boldsymbol{x}) = 411 + 14{,}1 \cdot \boldsymbol{x} \tag{1}$$

Que establece la equivalencia de un punto en la prueba de habilidades cuantitativas con 14 puntos en el puntaje estimado en la PSU de Matemática, con una base de 411 puntos.

El error estándar de estimación de 82 puntos, cuantifica la imprecisión de predicciones individuales. No obstante, la predicción a nivel promedio quedó estimada por un intervalo de confianza obviamente más acotado, graficado junto al diagrama de dispersión en la Figura VI.6.

Figura VI.6. Diagrama de dispersión para el primer modelo de regresión.



Por ejemplo, para un examinado de cuarto medio que obtuvo 8 puntos en la prueba, su estimación de puntaje PSU de Matemática es de 524 puntos y el intervalo de confianza al 95% es [318; 730]. En tanto, el promedio de su establecimiento educacional de 7,6 puntos se asocia con una estimación de 518 puntos y un intervalo al 95% de [499; 538] que contiene al puntaje promedio de 515 puntos obtenido por el establecimiento en el proceso de admisión 2016.

[30] Anexo E.9. Resultados de la regresión entre $\boldsymbol{x}$ e $\boldsymbol{y}$.

El coeficiente de validez desatenuado por la confiabilidad de la prueba y de los ensayos PSU de Matemática, este último estimado por el Alpha de Cronbach de 0,96 de la prueba oficial (Antivilo, Contreras, y Hernández, 2014), resultó equivalente a $r_{x'y'} = 0{,}6$.

El segundo modelo de regresión entre las variables $\boldsymbol{x}$ y $\boldsymbol{z}$ fue estadísticamente significativo[31], $R^2 = 0{,}17$; $F(1{,}206) = 37{,}4$; $p < 0{,}001$, y quedó determinado por la ecuación

$$\boldsymbol{z}(\boldsymbol{x}) = 4{,}3 + 0{,}14 \cdot \boldsymbol{x} \tag{2}$$

El error estándar de estimación de 1 punto, refleja la imprecisa relación entre las notas y el puntaje de la prueba. Mientras que, el coeficiente de validez desatenuado por la confiabilidad de la prueba, resultó equivalente a $r_{x'z} = 0{,}6$.

Se debe mencionar la mayor intensidad de la asociación entre el promedio de notas y el puntaje en ensayos PSU de Matemática, que se refleja en un error estándar de estimación de 65 puntos y un modelo de regresión estadísticamente significativo[32] ($p < 0{,}001$).

Adicionalmente, el modelo que combina las tres variables en una regresión lineal múltiple, descontadas las observaciones influyentes, resultó estadísticamente significativo[33] ($p < 0{,}001$) con una considerable correlación positiva de 0,72.

Complementariamente, para indagar en la pertinencia de usar el puntaje PSU como medida auto-reportada se comparó el promedio de esta variable a nivel de establecimiento, con el puntaje obtenido en el proceso de admisión 2016, hallándose una diferencia de 75, 7, 63 y 147 puntos en los cuatro establecimientos con examinados de tercero medio y de 28 puntos en el establecimiento con examinados de cuarto medio.

Cabe destacar que, si bien las medias de los puntajes de los examinados según nivel de enseñanza no difieren significativamente[34] ($p < 0{,}001$), si lo hacen sus coeficientes de correlación con los puntajes en ensayos PSU de Matemática.

[31] Anexo E.10. Resultados y diagrama de dispersión para la regresión entre ***x*** y ***z***.
[32] Anexo E.11. Resultados y diagrama de dispersión para la regresión entre ***z*** e ***y***.
[33] Anexo E.12. Resultados de la regresión múltiple entre ***x*** y ***z*** con ***y***.
[34] Anexo E.13. Test de diferencia entre medias de puntajes, según nivel de enseñanza.

Mientras para tercero medio esta correlación fue 0,37 para cuarto medio fue 0,63 traduciéndose en coeficientes de validez desatenuados equivalentes a 0,47 y 0,76, respectivamente[35].

Al comparar las especificaciones de la prueba con la PSU de Matemática, se observaron entre las similitudes: el uso intensivo del mismo tipo de ítems, el tiempo asignado por ítem, la vinculación con el mismo marco curricular, y la especificación con base en las mismas categorías de contenido. Entre las diferencias se observaron: las ponderaciones en la matriz de especificaciones[36], la cantidad de ítems contextualizados, y el nivel de entrenamiento específico requerido.

En síntesis, se considera que estos datos constituyen una moderada evidencia a favor de la inferencia de extrapolación a nivel individual. Teniendo presente que esta inferencia está mejor calificada para examinados de cuarto medio y que tiene un error más acotado a nivel de establecimiento educacional.

[35] Anexo E.14. Diagramas de dispersión y ajustes de regresión, según nivel de enseñanza.
[36] Anexo E.15. Comparación entre las especificaciones de la prueba y la PSU de Matemática.

VII. DISCUSIÓN

El planteamiento inicial de los argumentos de interpretación y uso anticipó que las evidencias para evaluarlos se recogerían durante todo el proceso de diseño, desarrollo y aplicación de la prueba. Y que estas serían de diverso tipo y relevancia según su rol en cada inferencia.

Se ha omitido la inferencia de puntuación debido a la simplicidad del modelo utilizado y se han descartado inferencias basadas en teoría por encontrarse fuera del alcance de este proyecto y demandar un mayor desarrollo teórico del constructo. Esta carencia de relaciones teóricas fundadas con otras mediciones cognitivas y educacionales impulsaron la naturaleza semántica de la definición del constructo que conjugó las habilidades cognitivas y los contenidos matemáticos con sus contextos de aplicación. El modelamiento obtenido estableció un dominio muestreable con ítems de selección de respuesta y vinculado con el dominio objetivo de la PSU de Matemática, contribuyendo así al uso descriptivo y predictivo de los puntajes, respectivamente.

La evaluación de los ítems contrastó datos mixtos provenientes de las opiniones de los jueces revisores y de los estadísticos de desempeño psicométrico, de los cuales el indicador crítico resultó ser la capacidad discriminativa por su incidencia en el Alpha de Cronbach y en las cargas factoriales. De hecho se pudo estimar que un aumento del 25% en la discriminación promedio aumentaría el Alpha de Cronbach a 0,75.

El argumento de validez para el uso descriptivo debe sopesar el sustento a su favor provisto por las evidencias de alineación, con la baja calificación que representó el error estándar de medición. Y el débil apoyo entregado por las evidencias de estructura interna sugiere revisar el cumplimiento del supuesto inicial sobre la construcción de los ítems. En tanto, para el argumento de validez del uso predictivo, el procedimiento de eliminación de observaciones influyentes permitió depurar el ajuste de los modelos de regresión y suprimir casos atípicos que reportaron respuestas de baja coherencia entre sí. En tanto, la mejor calificación de la inferencia para examinados de cuarto medio guarda sentido con el mayor grado de preparación y conocimiento de la PSU de Matemática de este grupo.

VIII. CONCLUSIONES

En cuanto al diseño y desarrollo, se confirma el rol estructurante del argumento de interpretación y uso, además de la importancia de un modelamiento detallado y fundado del constructo que ilumine coherentemente la especificación y construcción de los ítems. También se constató la dificultad para alcanzar un modelamiento e items originales, en consideración del gran número instrumentos desarrollados para objetivos de evaluación similares.

En cuanto a la validación, la verificación de las condiciones necesarias para el uso descriptivo, en términos de confiabilidad, calidad constructiva y desempeño psicométrico de los ítems y estandarización del sistema de aplicación, permitió iniciar la recolección de evidencias que proporcionaron un sustento moderado a la coherencia y representatividad de los ítems, un fuerte apoyo a su capacidad para elicitar los procesos cognitivos esperados en las respuestas correctas y un bajo respaldo a las hipótesis de unidimensionalidad y dificultad. Con lo cual, **se concluye un grado moderado de validez** para sustentar la interpretación del puntaje de la prueba como estimador de un puntaje generalizado en el dominio de ítems.

Del mismo modo, los coeficientes de determinación y correlación, descontadas las observaciones influyentes, indicaron una moderada asociación positiva entre los puntajes de la prueba y los puntajes PSU auto-reportados, que permitió calcular el coeficiente de validez desatenuado de 0,6 y concluir como moderado el grado de validez en que se sustenta la interpretación del puntaje de la prueba como predictor este criterio[37]. Distinguiendo un grado de validez moderadamente alto para este inferencia en los examinados de cuarto medio.

[37] Anexo E.16. Diagramas para argumentos de validez.

IX. RECOMENDACIONES

La recomendación principal es utilizar los resultados de este proyecto para perfeccionar la prueba, en términos de aumentar su grado de validez para cada interpretación y uso esperado. Más específicamente:

- Ajustar el plan de muestreo para aumentar la representatividad de la muestra, resguardando que el porcentaje de casos según género y dependencia educacional, se acerque a los parámetros poblacionales.
- Detallar la definición conceptual y ejemplificar la definición operacional de las habilidades cognitivas evaluadas en los diferentes contenidos del constructo, para elaborar un documento de referencia para los jueces revisores y el constructor de los ítems.
- Construir o adaptar una mayor cantidad de ítems de tal forma de ensamblar más de una forma y abarcar en su conjunto un mayor porcentaje del dominio de ítems, además de eventualmente producir formas diferenciadas y equiparables para tercero y cuarto medio.
- Aumentar las exigencias al desempeño psicométrico de los ítems. Acotando su rango de dificultad entre 0,4 y 0,7 para generar más información en torno a puntuaciones intermedias. Y elevando el rango de discriminación promedio por sobre 0,3 para aumentar la estimación de confiabilidad.
- Utilizar en caso de estar disponibles, datos reales para el cálculo de correlaciones entre las puntuaciones de la prueba y las variables criterio. O en su defecto, estudiar la distorsión de los valores de estas variables al ser auto-reportadas.
- Declarar y documentar el cumplimiento o incumplimiento de estándares internacionales para la aplicación de pruebas educacionales que sea pertinente considerar como referencia.

X. REFERENCIAS

AERA, APA, y NCME. (2014). *Standards for educational and psychological testing*.

Anderson, L. W., y Krathwohl, D. R. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, 1–336.

Antivilo, A., Contreras, P., y Hernández, J. (2014). *Estudio de confiabilidad de las pruebas de selección universitaria, admisión 2013.* Santiago de Chile.

Bennett, J. O., y Briggs, W. L. (2008). *Using and Understanding Mathematics : A Quantitative Reasoning Approach* (4th Ed.). Boston: Pearson Education.

Boote, D. N., y Beile, P. (2016). Research Preparation Features Scholars Before Researchers : On the Centrality of the Dissertation Literature Review in Research Preparation. *Educational Researcher*, *34*(6), 3–15.

Burke, M., Mattar, J., Stopek, J., y Eve, H. (2014). Modelling complex performance tasks. Paper presented at the Annual Metting of the National Council of Measurement in Education. Philadelphia, PA.

Campion, D. (2016). Test production. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development* (2nd Ed.). New York: Routledge.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Downing, S. M., y Haladyna, T. M. (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates.

Hair, J., y Suárez, M. (1999). *Análisis multivariante*. Madrid: Prentice Hall.

Haladyna, T. (2016). Item Analysis for Selected-Responde Test Items. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development*. New York: Routledge.

Haladyna, T. M., y Downing, S. M. (1993). How many options is enough for a multiple choice item? *Educational and Psychological Measurement*, *53*, 999–1010.

Haladyna, T. M., y Rogriguez, M. C. (2013). *Developing and validating test items.* Routledge.

Hernández, R., Fernández, C., y Baptista, P. (2006). *Metodología de la investigación* (4a Ed.). México: McGraw-Hill Interamericana.

Hogan, T. (2004). *Pruebas Psicológicas. Una introducción práctica* (2a Ed.). Mexico: Manual Moderno.

Horn, J. L. (1965). A rationale and test for the numbers of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Huff, K., Steinberg, L., y Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*, 310–324.

IBM. (2011). *IBM SPSS Statistics 20 core system user's guide*.

ICFES. (2014). *Medición de Preparación Universitaria con SABER PRO: Un examen de la validez predictiva*. Bogotá, Colombia.

Jöreskog, K. G. (1977). Factor analysis by least squares and maximum likelihood methods. In K. Enslein (Ed.), *Statistical Methods for Digital Computers*. New York: Wiley.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th Ed.). Westport, CT: Praeger.

Kane, M. (2016). Validation Strategies: Delineating and Validating Proposed Interpretations an d Uses of Test Scores. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development* (2nd Ed.). New York: Routledge.

Lakin, J. M., Elliott, D. C., y Liu, O. L. (2012). Investigating ESL Students' Performance on Outcomes Assessments in Higher Education. *Educational and Psychological Measurement*, *72*(5), 734–753.

Lane, S., Raymond, M., y Haladyna, T. (2016). *Handbook of test development* (2nd. Ed.). New York: Routledge.

McCallin, R. (2016). Test Administration. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development* (2nd Ed.). New York: Routledge.

Merenda, P. (1997). Methods, plainly speaking: A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Evaluation*, *30*, 156–163.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed.). New York: American Council of Education.

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. New York: Routledge.

Mikenberg, I. (2016). *Razonamiento Cuantitativo*. Santiago de Chile: Ediciones UC.

MINEDUC. (2009). *Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Básica y Media. Actualización 2009*. Santiago de Chile.

MINEDUC. (2015). *Bases Curriculares 7º básico a 2º medio*. Santiago de Chile.

Mislevy, R., y Haertel, G. (2006). Implications of evidence centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

OECD. (2012). *The survey of adult skills: Reader's companion.* Paris.

Pérez, E. R., y Medrano, L. (2010). Análisis factorial exploratorio : Bases conceptuales y metodológicas. *Revista Argentina de Ciencias Del Comportamiento*, *2*(1889), 58–66.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29.

Perie, M., y Huff, K. (2016). Determining Content and Cognitive Demans for Achievement Tests. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development* (2nd Ed.). New York: Routledge.

Pontificia Universidad Católica de Chile. (2014). Prueba de Razonamiento Cuantitativo (Formas 2011-1, 2011-2, 2012-B, 2013-A, 2014-A). Santiago de Chile.

Rhodes, T. L. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics.* Association of American Colleges and Universities.

Rodriguez, M. (2016). Selected-Response Item Development. In S. Lane, M. Raymond, y T. Haladyna (Eds.), *Handbook of test development* (2nd Ed). New York: Routledge.

Roohr, K. C., Graf, E. A., y Liu, O. L. (2014). Assessing Quantitative Literacy in Higher Education: An Overview of Existing Research and Assessments With Recommendations for Next-Generation Assessment. *ETS Research Report Series*.

Roohr, K. C., Liu, O. L., y Liu, H. (2017). Investigating Validity Evidence for the ETS ® Proficiency Profile. *ETS Research Report Series*.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., y Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, *18*(2), 99–121.

Scriven, M. (1991). *Evaluation thesaurus*. Sage.

Shavelson, R. J. (2008). Reflections on Quantitative Reasoning: An Assessment Perspective. *Calculation vs. Context*, 27–47.

Sons, L. (1996). *Quantitative reasoning for college graduates: A supplement to the standards.* Mathematical Association of America.

Steen, L. A. (1998). Why numbers count: Quantitative literacy for tomorrow's America. *NASSP Bulletin*, *82*(600), 120.

Steen, L. A. (2004). *Achieving Quantitative Literacy: An Urgent Challenge for Higher Education*.

Thissen, D., y Wainer, H. (2001). *Test Scoring*. New York: Routledge.

Thompson, S. (2012). *Sampling* (3rd Ed.). New York: Wiley.

Timmerman, M. E., y Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

Vacher, H. L. (2014). Looking at the Multiple Meanings of Numeracy, Quantitative Literacy, and Quantitative Reasoning. *Numeracy*, *7*(2), 1–14.

VandenBos, G. (2015). *APA dictionary of psychology* (2nd Ed.). Washington: American Psychological Association.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New Jersey: Lawrence Erlbaum Associates.

Wise, L., y Plake, B. (2016). Test Design and Development Following the Standards for Educational and Psychological Testing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd Ed.). New York: Routledge.

Zwick, R., y Ercikan, K. (1989). Analysis of differential item functioning in NAEP history assesment. *Journal of Education Measurement*, *26*(1), 55–66.

XI. ANEXOS

Anexo A. Antecedentes complementarios

A.1. Resumen de definiciones de alfabetismo cuantitativo y razonamiento cuantitativo.

Organización	Definición de Alfabetismo o Razonamiento Cuantitativo[a]
Asociación Americana de Universidades	"Alfabetismo Cuantitativo - también conocido como Alfabetismo Numérico o Razonamiento Cuantitativo - es un "hábito mental", competencia, y comodidad trabajando con datos numéricos. Individuos con fuertes habilidades de Alfabetismo Cuantitativo poseen la abilidad para razonar y resolver problemas cuantitativos desde un amplio rango de contextos auténticos y situaciones de la vida diaria. Ellos entienden y pueden crear sofisticados argumentos apoyados por evidencia cuantitativa y pueden claramente comunicar estos argumentos en una variedad de formatos (usando apropiadamente palabras, tablas, gráficos, ecuaciones matemáticas, etc.)"[a].
Asociación Americanan de Matemática	"Un estudiante universitario considerado cuantitativamente alfabetizado debería ser capaz de: 1. Interpretar modelos matemáticos como fórmulas, gráficos, tablas y esquemas y esbozar inferencias a partir de ellos. 2. Representar información matemática simbólicamente, visualmente, numericamente y verbalmente. 3. Usar métodos aritméticos, algebraicos, geométricos y estadísticos para resolver problemas. 4. Estimar y verificar respuestas a problemas matemáticos en orden a determinar su pertinencia, identificar alternativas y seleccionar resultados óptimos. 5. Reconocer que los métodos matemáticos y estadísticos tienen límites"[b].
Organización para la Cooperación y el Desarrollo Económico	"La habilidad para acceder, usar interpretar y comunicar ideas e información matemática en orden a abordar y gestionar las exigencias matemáticas de un rango de situaciones en la vida adulta. Para este fin, el Alfabetismo Numérico involucra gestionar una situación o resolver un problema en un contexto real, respondiendo a contenido/informacion/ideas matemáticas representadas en múltiples formas"[c].

Nota. Traducido de "Assessing Quantitative Literacy in Higher Education: An Overview of Existing Research and Assessment With Recommendations for Next-Generation Assessment", Roohr et al., 2014, pág. 4.

[a] Rhodes, T. L. (Ed.) (2010). Assessing outcomes and improving achievement: Tips and tools for using rubrics.Washington,DC: Association of American Colleges and Universities. [b] Sons, L. (Ed.) (1996). Quantitative reasoning for college graduates: A complement to the standards. Washington, DC: Mathematical Association of America. [c] Organisation for Economic Co-Operation and Development. (2012b). The survey of adult skills: Reader's companion. Paris, France: Author.

A.2. Rúbrica de alfabetismo cuantitativo de la Asociación Americana de Universidades.

Habilidad	Descripción	Máximo Nivel de Logro
Interpretación	Habilidad para explicar información presentada en formas matemáticas (por ejemplo: ecuaciones, gráficos, diagramas, tablas, palabras).	Provee explicaciones precisas de información presentada en forma matemática. Hace inferencias apropiadas basadas en esta información.
Representación	Habilidad para convertir información relevante en variadas formas amtemáticas (por ejemplo: ecuaciones, gráficos, diagramas, tablas, palabras).	Habilmente convierte información relevante en una descripción matemática significativa, en una forma que contribuye a un entendimiento mejor o más profundo.
Cálculo		Los calculos realizados son todos esencialmente correctos y suficientemente comprensivos para resolver el problema. Estos cálulos son también presentados elegantemente (claramente, concisamente, etc.)
Aplicación/Análisis	Habilidad para hacer juicios y esbozar conclusiones apropiadas basadas en el análisis cuantitativo de datos, reconociendo los límites de este análisis.	Usa el análisis cuantitativo de datos como la base para juicios profundos y fundados, esbozando conclusiones signiticativas y cuidadosamente calificadas de su trababjo.
Supuestos	Habilidad para hacer y evaluar suposiciones importantes en la estimación, modelamiento y análisis de datos.	Explicitamente describe supuestos y provee justificaciones convincentes de porque cada supuesto es apropiado. Muestra con claridad que la confianza en las conclusiones finales está limitada por la precisión de los supuestos.
Comunicación	Expresar evidencia cuantitativa en sustento del argumento o propósito del trabajo (en términos de que evidencia es usada y como esta es formateda, presentada y contextualizada).	Usa información cuantitativa en conexión con el argumento o propósito del trabajo, presentándola en un formato efectivo, y explicándolo consistentemente con alta calidad.

Nota. Traducido de la Rúbrica de Alfabetismo Cuantitativo del Programa "Evaluación Válida del Aprendizaje en la Educación de Pregrado" de la Asociación Americana de Universidades (E.E.U.U.).

A.3. Marco de referencia para prueba de razonamiento cuantitativo de ETS (fragmento).

Área de Res. de problemas	*Descripción breve*
(a) Interpretación	El entedimiento y explicación de información matemática, como la habilidad para entender datos, leer gráficos, esbozar conclusiones y reconocer fuentes de error.
(b) Conocimiento y razonamiento estratégico	La formulación y evaluación de problemas matemáticos usando heurísticas, y la habilidad para reconocer relaciones acerca de conceptos y situaciones matemáticas.
(c) Modelamiento	El proceso de capturar relaciones presentes en el entorno o en formas matemáticas, y expresión de modelos en una o mas representaciones matemáticas.
(d) Cálculo	El proceso de identificación y manejo apropiado de manipulaciones algebraicas y aritméticas necesarias para resolver un problema.
(e) Comunicación	La presentación de conceptos e ideas de nivel superior (por ejemplo; argumentos y modelos matemáticos) tan bien como soluciones a problemas y otros procedimientos estandarizados; la comunicación puede tomar varias formas matemáticas y personalizarse para hacerla apropiada a la audiencia objetio.
Área de contenido	
(a) Números y operaciones	Números reales, propiedades de orden y cantidades físicas. Operaciones aritméticas con números reales. Estimación y razonamiento proporcional.
(b) Algebra	Variables, expresiones algebraicas y su uso en representar cantidades. Funciones, sus tipos y propiedades y su uso en la resolución de problemas. Ecuaciones, inecuaciones y su uso en la resolución de problemas.
(c) Geometría y Medición	Figuras geométricas en una, dos y tres dimensiones. Mediciones de figuras geométricas (por ejemplo, área, distancia, longitud, volumen, ángulos) para resolver un problema.
(d) Estadística y Probabilidad	Interpretación y representación de datos. Estadística descriptiva. Probabilidad básica.
Contexto	
(a) Vida cotidiana	Manejar dinero y presupuestos; compras; gestión del tiempo; viajes personales; juegos de azar; estadísticas deportivas; lecutra de mapas; medidas en la cocina; reparaciones en el hogar; o hobbies personales; calcular una propina; completar un formulario de compra; entender y evaluar la salud personal; balancear un talonario de cheques.
(b) Lugar de trabajo	Manejar horarios, presupuestos y recursos de proyectos; usar hojas de cálculo; hacer y registrar mediciones; llevar contabilidad y remuneraciones; seguimiento de gastos; predicción de costos; toma de decisiones relacionadas al trabajo.
(c) Sociedad	Cambios poblacionales; tasas de desempleo; sistemas de votación; transporte público; gobierno; políticas públicas; demografía; publicidad; estadísticas nacionales; economía; información cuantitativa en los medios; recaudación de fondos para una organización; interpretación de estudios de investigación; tendencias o asuntos ambientales.

Nota. Traducido de "Assessing Quantitative Literacy in Higher Education: An Overview of Existing Research and Assessment With Recommendations for Next-Generation Assessment", Roohr et al., 2014, págs. 15, 16 y 17.

Anexo B. Sobre el diseño y desarrollo de la prueba

B.1. Datos de la población y la muestra.

	Población			
		Estudiantes		
	Establecimientos	Enseñanza Media		3º y 4º Medio[b]
Provincia de Santiago[a]	**2.236**	**193.869**		**96.935**
Municipal	497	51.331	26%	25.666
Particular Subvencionado	1.460	102.071	53%	51.036
Particular Pagado	279	40.467	21%	20.234

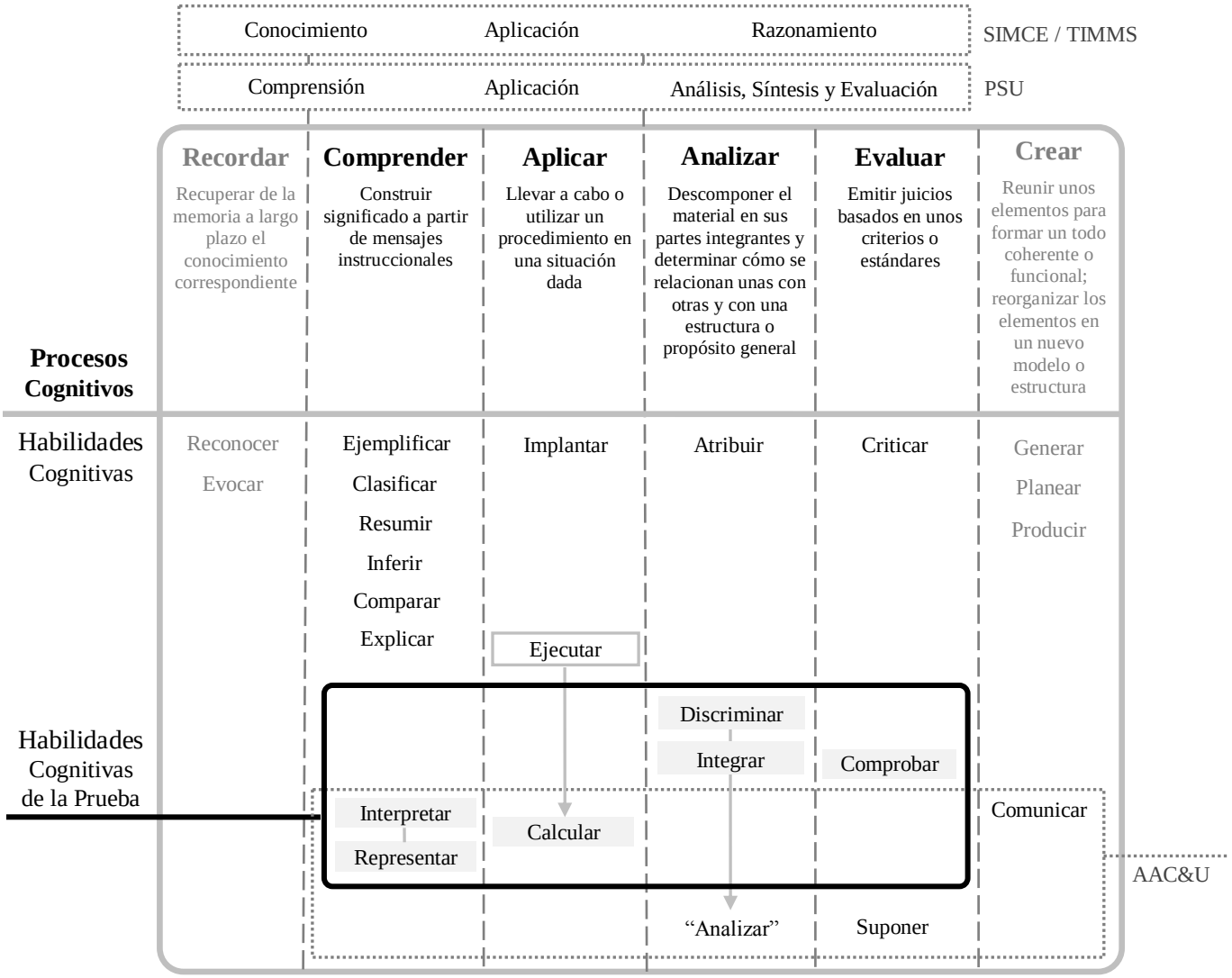
Nota. La cantidad de establecimientos corresponde a los catalogados como "funcionando" y la cantidad de estudiantes corresponde a la matrícula de "enseñanza media científico-humanista de niños y jóvenes", según base de datos "Resumen de matrícula oficial por establecimiento, 2015 " de la información estadística disponible en el Centro de Estudios Mineduc (centroestudios.mineduc.cl/).

[a] Los datos de la provincia de Santiago corresponde a la suma de las 32 comunas que conforman su división administrativa actual. [b] La estimación de la cantidad de estudiantes en tercero y cuarto medio equivale al 50% de la matrícula de enseñanza media.

	Muestra				
		Estudiantes[a]			
	Establecimientos	Mujeres	Hombres	Total	%
Prepilotaje	2	42	17	59	100%
Particular Subvencionado	1	31	-	31	53%
Particular Pagado	1	11	17	28	47%
Pilotaje	**5**	82	158	**240**	100%
Municipal	2	30	65	95	40%
Particular Subvencionado	1	31	41	72	30%
Particular Pagado	2	21	52	73	30%

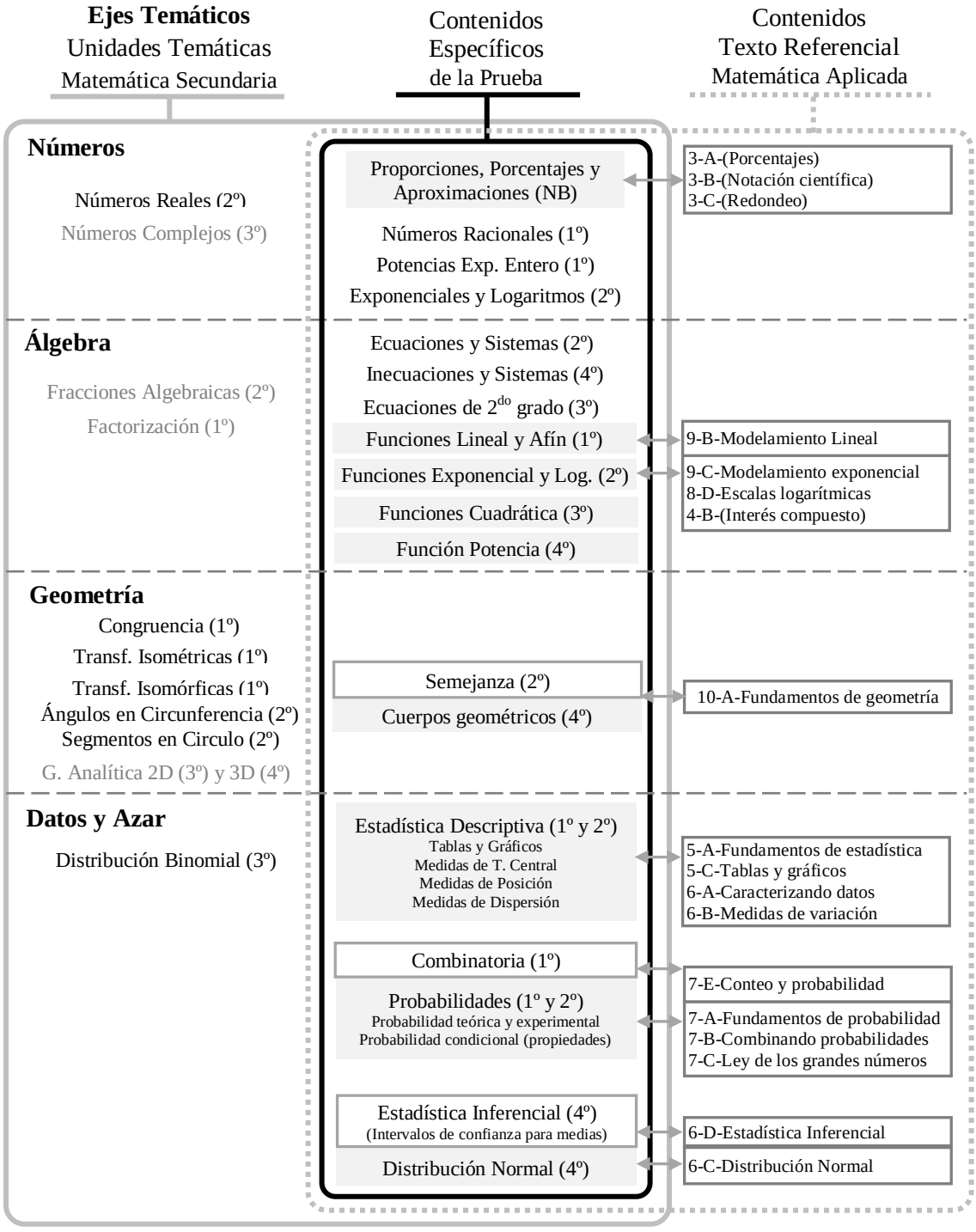
Nota. La cantidad de estudiantes que participaron del pilotaje incluye siete casos de diferentes establecimientos, que particiaparon en una aplicación adicional.

B.2. Detalle de las habilidades cognitivas de la prueba.



Nota. A través de este diagrama se observa que la taxonomía revisada de Bloom (Anderson et al., 2001. p. 41-42) es lo suficientemente amplia y específica para contener las habilidades cognitivas evaluadas en pruebas estandarizadas como PSU, SIMCE y TIMMS. También subyace al modelamiento del razonamiento cuantitativo de la Asociación Americana de Universidades. Y se relaciona con el modelamiento de la prueba de habilidades cuantitativas diferenciando los mismos procesos cognitivos y vinculándose a nivel de habilidades específicas (por ejemplo; calcular en relación con ejecutar y "analizar" en relación con discriminar e integrar).

B.3. Detalle de los contenidos matemáticos de la prueba.



Nota. A través de este diagrama se observa que los contenidos de la prueba son un subconjunto de los contenidos del curriculum vigente de matemática secundaria (MINEDUC, 2009). Y su mayor potencial de contextualización, se verifica por su inclusión mayoritaria en un texto referencial de Matemática Aplicada, que siguiendo un enfoque de razonamiento cuantitativo, desarrolla y ejercita estos contenidos en contexto (Bennett y Briggs, 2008).

B.4. Detalle de los contextos utilizados en la prueba.

Datos Cuantitativos en…

Cifras

Cotidianas; Precios, cantidades y porcentajes de descuento en productos.
Científicas; Magnitudes grandes (distancias astronómicas) y pequeñas (pesos atómicos).

Modelos

De situaciones de cambio; lineal (tarifas de servicios), exponencial (crecimiento poblacional, interés compuesto), logarítmico (escalas de intensidad sonora dB, intensidad sísmica, acidez pH) y cuadráticas (caída libre, movimiento balístico)

Medidas

De objetos cotidianos tales como; cajas, frascos, conos y pelotas.
De planos o maquetas que representan lugares u objetos a escala.

Estadísticas

Económicas; Ingreso, crecimiento, ahorro, inversión, inflación, empleo.
Demográficas; Natalidad, mortalidad, esperanza de vida.
Académicas; Puntajes en pruebas estandarizadas, notas en pruebas de aula.
Deportivas; Resultados en competencias, rendimiento de deportistas.

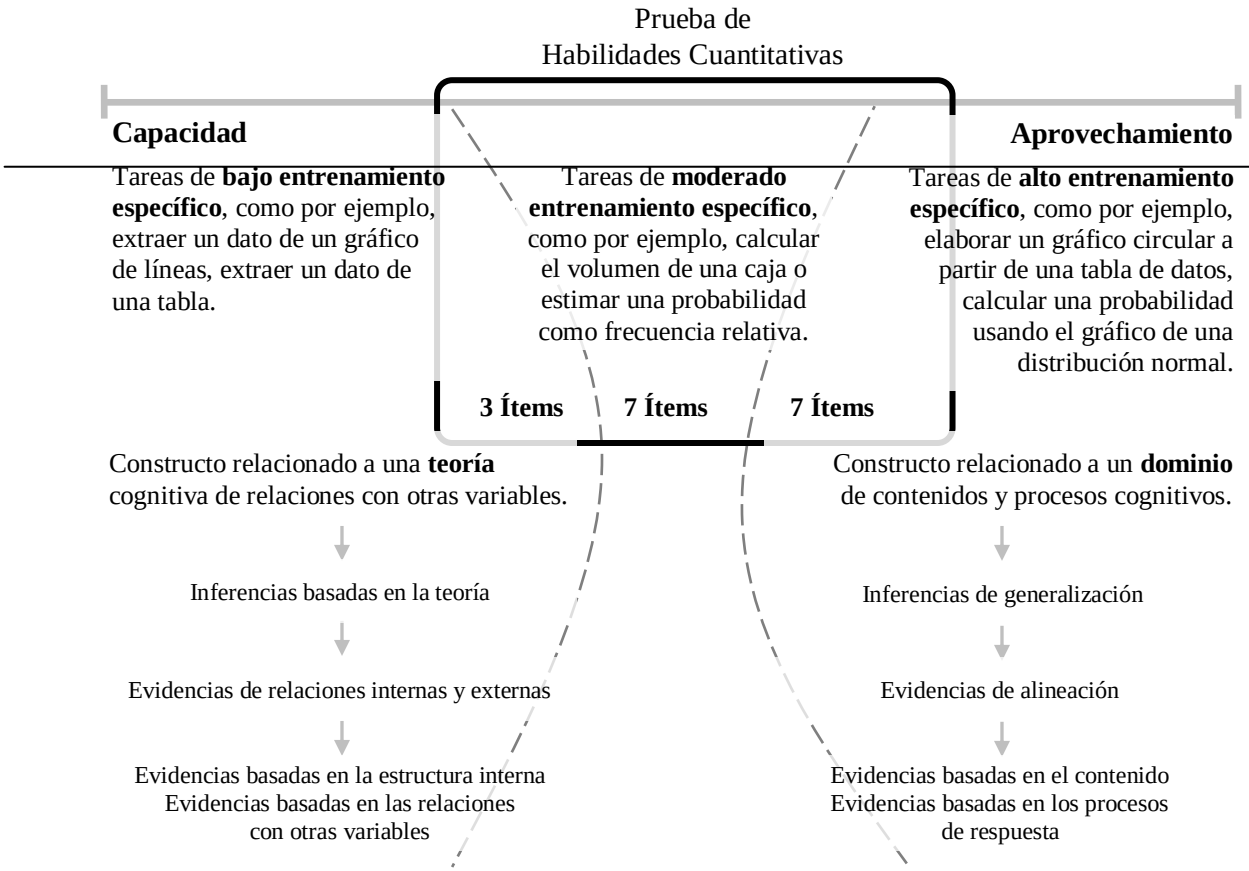
Azar

En situaciones cotidianas; juegos y apuestas.
En situaciones de riesgo; accidentes u otro tipo de siniestros.

B.5. Mapa de constructo para las habilidades cuantitativas.

		Números, Funciones y Geometría	**Datos y Azar**
Razonamiento	Comprobar (Evaluar)	Comprueba la pertinencia de afirmaciones basadas en aproximaciones, porcentajes, modelos matemáticos y cuerpos geométricos.	Comprueba la pertinencia de afirmaciones basadas en tablas y gráficos estadísticos.
	Discriminar-Integrar. (Analizar)	Elabora afirmaciones a partir de datos representados en porcentajes, modelos matemáticos y cuerpos geométricos.	Elabora afirmaciones a partir de datos en tablas y gráficos estadísticos.
Comprensión	Calcular (Aplicar)	Obtiene aproximaciones, porcentajes, valores de un modelo y medidas geométricas a partir de cantidades, funciones y cuerpos geométricos.	Obtiene datos, frecuencias, promedios y probabilidades a partir de tablas y gráficos estadísticos.
	Representar. (Comprender)	Expresa datos en forma de porcentajes, proporciones, modelos matemáticos y cuerpos geométricos.	Expresa datos tabulados en forma gráfica o viceversa.
	Interpretar. (Comprender)	Extrae información de porcentajes, proporciones, modelos matemáticos y cuerpos geométricos.	Extrae información de tablas y gráficos estadísticos.

B.6. La prueba en el continuo de capacidad–aprovechamiento.



Nota. Este esquema es una propuesta integradora basada en: el continuo de capacidad y aprovechamiento para clasificar pruebas (Hogan, 2014), la clasificación de las inferencias de los argumentos de interpretación y uso (Kane, 2016) y la clasificación de las evidencias para sustentar argumentos de validez (Wise y Plake, 2016). Permite visualizar la prueba de habilidades cuantitativas en un rango intermedio, considerando que se podrían clasificar 3, 7 y 7 de sus ítems como tareas de bajo, moderado y alto entrenamiento específico, respectivamente. Implicando una mayor relevancia para las evidencias de alineación.

B.7. Matriz y tabla de especificaciones.

Matriz de especificaciones

Categorías / Habilidades Cognitivas (Proceso Cognitivo)	**Categorías** / Ejes Temáticos: Números	Álgebra	Geometría	**Números, Álgebra y Geometría**	**Datos y Azar**		
Comprensión	**1**	**3**	**1**	**5**	**6**	**11**	**65%**
Interpretar (Comprender)		1		1	1	2	11,8%
Representar (Comprender)		1		1	1	2	11,8%
Calcular (Aplicar)	1	1	1	3	4	7	41,2%
Razonamiento	**2**	**1**		**3**	**3**	**6**	**35%**
Discriminar-Integrar (Analizar)	1	1		2	1	3	17,6%
Comprobar (Evaluar)	1			1	2	3	17,6%
	3	4	1	**8**	**9**	**17**	
	18%	23%	6%	**47%**	**53%**		

Tabla de especificaciones

Nº Ítem	**Eje Temático**	**Categoría de Habilidad**	**Proceso Cognitivo**	**Habilidad Cognitiva**	**Nivel**
1	Números	Comprensión Cuantitativa	Aplicar	Calcular	NB
2	Números	Razonamiento Cuantitativo	Analizar	Discriminar-Integrar	NB
3	Números	Razonamiento Cuantitativo	Evaluar	Comprobar	1º Medio
4	Álgebra	Comprensión Cuantitativa	Comprender	Interpretar	1º Medio[a]
5	Álgebra	Comprensión Cuantitativa	Aplicar	Calcular	2º Medio[b]
6	Álgebra	Comprensión Cuantitativa	Comprender	Representar	2º Medio
7	Álgebra	Razonamiento Cuantitativo	Evaluar	Comprobar	3º Medio
8	Geometría	Comprensión Cuantitativa	Aplicar	Calcular	NB[d]
9	Datos	Comprensión Cuantitativa	Comprender	Interpretar	NB
10	Datos	Comprensión Cuantitativa	Comprender	Representar	NB
11	Datos	Comprensión Cuantitativa	Aplicar	Calcular	NB
12	Datos	Comprensión Cuantitativa	Aplicar	Calcular	1º Medio
13	Datos	Comprensión Cuantitativa	Aplicar	Calcular	1º Medio
14	Datos	Comprensión Cuantitativa	Aplicar	Calcular	1º Medio
15	Datos	Razonamiento Cuantitativo	Evaluar	Comprobar	1º Medio
16	Azar	Razonamiento Cuantitativo	Evaluar	Comprobar	2º Medio
17	Azar	Razonamiento Cuantitativo	Analizar	Discriminar-Integrar	4º Medio

Nota. La columan titulada Nivel especifica el nivel de enseñanza asociado al ítem. NB indica enseñanza básica.

[a] Contenido de función definida por tramos, cercano a la función lineal y afín.

[b] Contenido relacionado al interés compuesto, como aplicación de la función exponencial.

[c] Contenido que no se cataloga en 4º Medio por tratarse del volumen de un poliedro sencillo.

B.8. Ejemplo de ficha de construcción de un ítem.

Ítem Nº 5.1b
"Vida en nuestra galaxia"

Especificaciones	
Categoría de habilidad	Razonamiento
Habilidad cognitiva	Comprobar (Evaluar)
Eje temático	Números
Contenido	Proporciones con potencias de base 10
Contexto	Científico: grandes cantidades en cifras astronómicas
Objetivo de Evaluación	*Comprueba la pertinencia de afirmaciones basadas en aproximaciones y proporciones.*

Ítem

Una persona estima que habría vida en cuatro millones de planetas de nuestra galaxia. Asumiendo que por cada medio millón de estrellas habría un planeta con vida, ¿cuántas estrellas consideró en su estimación?

A) $4 \cdot 10^6$

B) $2 \cdot 10^{11}$

C) $2 \cdot 10^{12}$

D) $8 \cdot 10^{12}$

Justificaciones	
Opción A	Asocia la cantidad de estrellas con la cantidad de planetas con vida, omitiendo la tasa de estrellas por planeta con vida. $\{X = 4.000.000 = 4 \cdot 10^6\}$
Opción B	Expresa en notación científica sin ajustar el exponente $\{20 \cdot 10^{11} = 2 \cdot 10^{11}\}$
Opción C	CLAVE.
Opción D	Expresa "medio millón" como $2 \cdot 10^{11}$ $\{X = 4 \cdot 10^6 \cdot 2 \cdot 10^6\}$

Observaciones del Constructor
Este planteamiento es una simplificación de la ecuación de Drake. La cantidad de estrellas corresponde al mínimo valor entre los 100 mil y 400 mil millones de estrellas, que se estima actualmente.

Observaciones del Experto						
¿El ítem cumple con las especificaciones?		Sí		No		
¿Cuáles son sus correcciones y sugerencias?						
Decisión del Experto	Aprueba		Modifica	¿Por qué?		Rechaza

B.9. Análisis de originalidad de los ítems.

Ítems similares en ensayos oficiales PSU

Ítem Construído			Ítem similar(es) encontrado(s) en publicaciones DEMRE											
Nº	**Título**	**Idea**	**Cantidad**	p2017	p2016	p2015	p2014	p2013	p2012	p2011	p2010	p2009	p2008	p2007
1	Las monedas de cinco pesos[a]	Eliminación de las monedas de $1 y $ 5 anunciada por las autoridades.	**0**	-	-	-	-	-	-	-	-	-	-	-
3	La vida en nuestra galaxia	Estimación de la probabilidad de encontrar vida en el universo a través de la ecuación de Drake.	**2**	Ítem 4	-	Ítem 6	-	-	-	-	-	-	-	-
4	La tarifa del estacionamiento	Tarificación frecuente de los estacionamientos en zonas urbanas del país.	**4**	-	-	-	Ítem 26	Ítem 28	-	-	-	Ítem 29	-	Ítem 32
6	Los datos informáticos	Estimación del crecimiento explosivo de los datos informáticos, en particular según el modelo de la Corporación EMC.	**2**	-	-	-	Ítem 36	-	-	-	-	Ítem 34	-	-
7	La pelota de tenis	Situación comúnmente modelada a través de una función cuadrática.	**2**	Ítem 33	-	-	Ítem 35	-	-	-	-	-	-	-
8	La cajita de jugo[a]	Situación de cálculo sencillo del volumen de un objeto.	**0**	-	-	-	-	-	-	-	-	-	-	-
9	El ingreso per-capita[a]	Representación gráfica de trayectoria de crecimiento económico del país durante los últimos 15 años.	**0**	-	-	-	-	-	-	-	-	-	-	-
10	El medallero olímpico	Representación gráfica de datos estadísticos para los juegos olímpicos de Rio 2016.	**2**	-	-	-	Ítem 66	-	Ítem 65	-	-	-	-	-
11	El climograma de Santiago	Interpretación de información geográfica propia de la zona de Santiago.	**1**	-	-	-	-	-	-	-	-	-	Ítem 62	
12	El ahorro para la jubilación	Interpretación del nivel de ahorro previsional que tienen los chilenos en el contexto del debate nacional por el sistema de ahorro previsional.	**2**	Ítem 62	-	Ítem 61	-	-	-	-	-	-	-	-
13	El puntaje ponderado	Situación común de cálculo sencillo de un promedio ponderado, en el contexto de la postulación a una universidad.	**0**	-	-	-	-	-	-	-	-	-	-	-
14	Las causas de accidentes	Divulgación de la imprudencia de peatones y conductores como una causa relevante de los accidentes de tránsito.	**2**	-	-	-	-	Ítem 61		Ítem 61				
17	El puntaje en la PSU de Lenguaje[b]	Uso de propiedades básicas de la distribución normal para calcular probabilidades de puntajes en pruebas estandarizadas.	**0**	-	-	-	-	-	-	-	-	-	-	-

Nota. Se han excluído los ítems Nº 2, Nº 5, Nº 15 y Nº 16 por su categoría de reservados ya que pertenecen a preguntas de formas confidenciales de la prueba de razonamiento cuantitativo de la Pontificia Universidad Católica de Chile, aplicada entre el 2012 y el 2014.

[a] No se encontraron ítems similares por corresponder a contenidos de enseñanza básica, no por que tengan mayor nivel de originalidad.

[b] No se encontraron ítems similares debido a que corresponde a un contenido recientemente incorporado a la enseñanza media.

Ideas de ítem similares en texto de referencia de Matemática Aplicada

Ítem Construído			Temas y ejemplos de ideas similares en texto de Matemática Aplicada		
Nº	**Título**	**Idea**	**Capítulo**	**Sección**	Ejemplos relacionados
1	Las monedas de cinco pesos[a]	Eliminación de las monedas de $1 y $ 5 anunciada por las autoridades.	-	-	-
3	La vida en nuestra galaxia	Estimación de la probabilidad de encontrar vida en el universo a través de la ecuación de Drake.	3. Números en el mundo real.	B. Poniendo números en perspectiva.	El ejemplo Nº 8, trata de la conversión de distancias astronómicas usando potencias de 10 y notación científica.
4	La tarifa del estacionamiento[a]	Tarificación frecuente de los estacionamientos en zonas urbanas del país.	-	-	-
6	Los datos informáticos	Estimación del crecimiento explosivo de los datos informáticos, en particular según el modelo de la Corporación EMC.	9. Modelando nuestro mundo	C. Modelamiento Exponencial	El ejemplo Nº 1, aborda el fenómeno del crecimiento poblacional a través de funciones exponenciales.
7	La pelota de tenis[a]	Situación comúnmente modelada a través de una función cuadrática.	-	-	-
8	La cajita de jugo	Situación de cálculo sencillo del volumen de un objeto.	10. Modelando con geometría	A. Fundamentos de Geometría	Los ejemplos Nº 5 y Nº 6, se refieren al cálculo de volúmenes para objetos reales como recipientes cúbicos o cilíndricos.
9	El ingreso per-capita	Representación gráfica de trayectoria de crecimiento económico del país durante los últimos 15 años.	5. Razonamiento Estadístico	C. Tablas y gráficos estadísticos (Gráficos de líneas)	El ejemplo Nº 8, presenta la utilidad de los gráficos de líneas para representar series de tiempo de variables económicas como el precio del oro.
10	El medallero olímpico	Representación gráfica de datos estadísticos para los juegos olímpicos de Rio 2016.	5. Razonamiento Estadístico	C. Tablas y gráficos estadísticos (Gráficos de torta)	-
11	El climograma de Santiago	Interpretación de información geográfica propia de la zona de Santiago.	5. Razonamiento Estadístico	D. Gráficos en los medios (Gráficos de datos geográficos)	El ejemplo Nº 3, muestra un gráfico con información geográfica que describe las líneas de temperaturas promedio en las diferentes zonas de un país.
12	El ahorro para la jubilación	Interpretación del nivel de ahorro previsional que tienen los chilenos en el contexto del debate nacional por el sistema de ahorro previsional.	5. Razonamiento Estadístico	C. Tablas y gráficos estadísticos (Tablas de Frecuencias)	-
13	El puntaje ponderado[a]	Situación común de cálculo sencillo de un promedio ponderado, en el contexto de la postulación a una universidad.	-	-	-
14	Las causas de accidentes	Divulgación de la imprudencia de peatones y conductores como una causa relevante de los accidentes de tránsito.	7. Probabilidad: Viviendo con posibilidades	A. Fundamentos de Probabilidad	El ejemplo Nº 4, explica el uso de datos empíricos para estimar la probabilidad de cada resultado, al lanzar dos monedas.
17	El puntaje en la PSU de Lenguaje[a]	Uso de propiedades básicas de la distribución normal para calcular probabilidades de puntajes en pruebas estandarizadas.	6. Poniendo las estadísticas a trabajar	C. La distribución normal	El ejemplo Nº 3, explica el uso de la distribucion normal y sus propiedades para el cálculo de probabilidades de puntajes en la prueba estandarizada SAT.

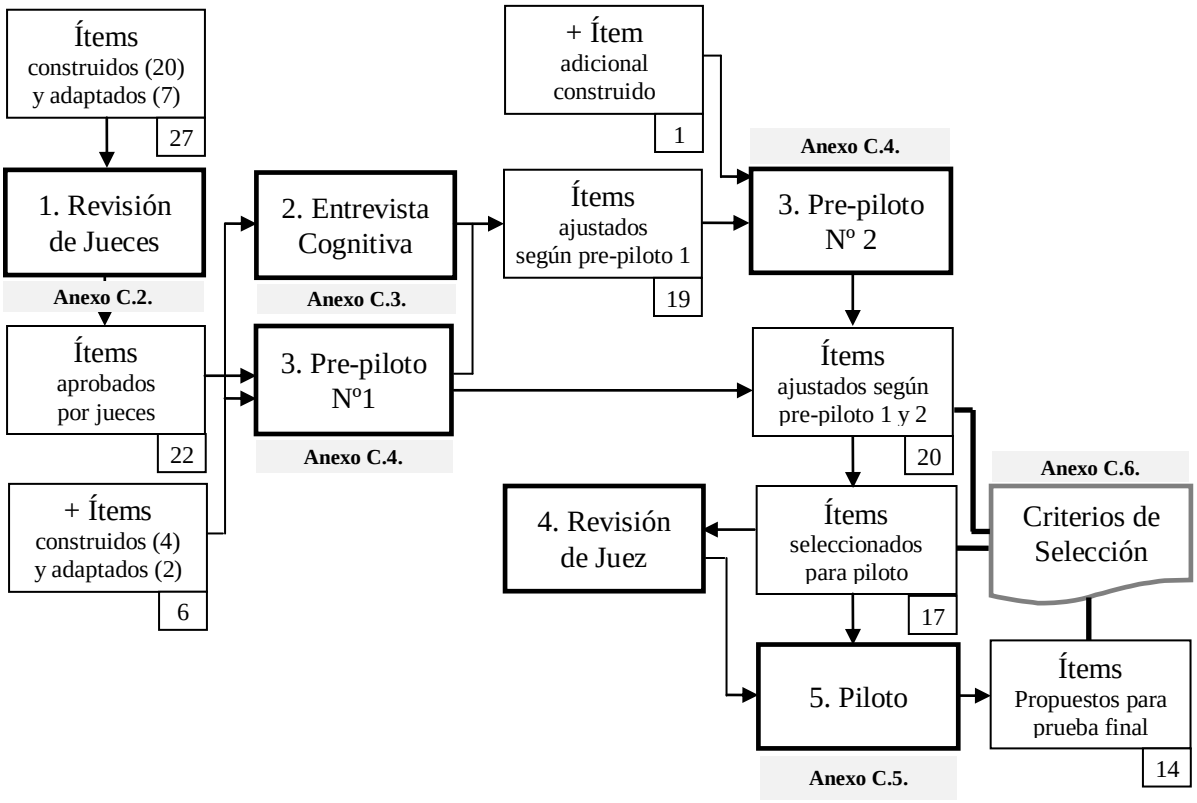
Nota. El texto de referencia utilizado en este análisis corresponde a "Usando y entendiendo la Matemática, un enfoque de razonamiento cuantitativo" (Bennett & Briggs, 2008).

[a] Ítem referido a contenidos no abordados por el texto de referencia.

Anexo C. Sobre la evaluación de los ítems

C.1. Proceso de evaluación de los ítems

El proceso de evaluación de los ítems se desarrollo en un plazo de aproximadamente 4 meses (Agosto-Diciembre) y consideró las etapas descritas en el siguiente flujo.



C.2. Revisión de jueces sobre la calidad constructiva de los ítems.

Las principales observaciones en relación a la calidad constructiva de los ítems fueron:

- Corregir la redacción de los enunciados y opciones, de tal forma de reducir su redundancia y/o complejidad lingüística.
- Cambiar la presentación de gráficos a ordenamientos cuadrangulares y verificar la reproductibilidad de los mismos.

		Decisión del Juez Nº 1			
		Aceptar	Modificar	Eliminar	Total
Juez Nº 2	Aceptar	13	4	2	19
	Modificar	3	2	0	5
	Eliminar	2	0	1	3
	Total	18	6	3	

Cada juez revisor hizo énfasis en corregir aspectos distintos de los ítems, lo que sumado al limitado tiempo de revisión disponible, podría explicar la baja consistencia entre sus decisiones, con un coeficiente de correlación de Pearson de 0,06, sin significancia estadística.

La formación y experiencia de cada juez se describe en la siguiente tabla.

Juez Nº 1	
Formación	Licenciado en Matemática de la Pontificia Universidad Católica de Chile Magíster en Matemática de la Pontificia Universidad Católica de Chile
Experiencia relacionada	Coordinador de la unidad de diagnóstico de la Facultad de Matemática de la misma universidad, con experiencia en la aplicación y análisis de sus pruebas de razonamiento cuantitativo.

Juez Nº 2	
Formación	Licenciado en Matemática de la Universidad de Santiago Profesor de Matemática de la Universidad de Santiago
Experiencia relacionada	Profesor de aula con experiencia en preparación para la prueba de selección universitaria y creación y administración de bancos de ítems.

C.3. Resultados de test y pauta de entrevista cognitiva.

La entrevista consistió en una primera etapa, de la aplicación de 8 ítems a un grupo de 11 examinados. Los resultados se detallan en la siguiente tabla.

Examinado	# Correctas	% de Logro	Tiempo Utilizado (Minutos)
1	6	86%	15
2	4	57%	18
3	3	43%	25
4	4	57%	23
5	3	43%	18
6	3	43%	20
7	3	43%	22
8	3	43%	17
9	2	29%	16
10	1	14%	16
11	3	43%	21
Promedio	**3,2**	**45%**	**19,2**

En una segunda etapa se condujo una discusión sobre aspectos generales de la prueba (tiempo y apreciación de dificultad) y aspectos específicos de los ítems (contenidos, enunciados, estímulos y distractores), sobre lo que surgieron las siguientes observaciones:

- El tiempo asignado fue suficiente y la prueba tuvo una dificultad general baja "no estaba difícil", además de develar en cierto grado que la habilidad matemática de los examinados, se relaciona positivamente con su autoconcepto de capacidad para la asignatura.

  Ejemplo de esto son los casos Nº 1 y Nº 10 que obtuvieron 86% y 14% de logro y que se auto-describen como "*a mí me va bien en matemática"* y *"yo no soy buena para las matemáticas"*, respectivamente.

- La dificultad de los enunciados se ve aumentada por la mayor presencia de expresiones algebraicas (Ítems Nº 4: Función Cuadrática y Nº 5: Función Logarítmica).
- La legibilidad y comprensibilidad fue insuficiente en algunos estímulos y opciones de ítems específicos (Ítems Nº 2: Objetos a escala y Nº 6: Gráficos Circulares).

La pauta de la entrevista se resumen en la siguiente tabla.

Preparación

Consigna
(Estructura)

1º Agradecimiento por la participación en la entrevista.
2º Explicación sobre la confidencialidad de la entrevista.
Entrega de asentimiento informado
3º Explicación del objetivo principal de la entrevista.

(Texto)

"Buenos días, junto con agradecerles su participación quiero comentarles que la entrevista es confidencial y se enmarca en el proceso de desarrollo de una prueba de habilidades matemáticas, en el cual todas sus respuestas y comentarios son muy bienvenidos. ¿Tienen alguna consulta o comentario antes de comenzar?. Bien comencemos…

Pauta

Objetivos
Recolectar los juicios de los alumnos sobre los ítems presentados, en términos de; dificultad, tiempo requerido para responder, contenido, claridad del enunciado y plausibilidad de los distractores, además de otros aspectos relevantes que surjan en la discusión.
Recoger evidencias sobre las habilidades cognitivas y los conocimientos utilizados por los estudiantes que están presentes (o ausentes) durante el desarrollo correcto (o incorrecto) de los ítems.

Preguntas
(Temas y Palabras Clave)

Preguntas	**Tipo de Pregunta**	**Temas y palabras claves de la Pregunta**	**Tiempo Estimado**
¿Qué les parecieron las preguntas?, ¿Por qué?	Introducción		2'
¿Qué tan fáciles o difíciles encontraron las preguntas?, ¿Tuvieron suficiente tiempo para responderlas?, ¿Qué dirían ustedes que se necesitaba saber para responderlas?, ¿Los enunciados eran claros y comprensibles? ¿Las imágenes y gráficos eran legibles?	Directa	Juicios generales sobre el conjunto de ítems. Dificultad, Tiempo, Contenido/Habilidad, Enunciados y Estímulos.	5'
¿Cuál consideran que fue el ítem más fácil? ¿Porqué? ¿Y el más difícil? ¿Porqué?	Directa	Juicios específicos sobre los ítems. Dificultad.	1'
Se ofrece la palabra a un estudiante para que explique su respuesta sobre un ítem que haya respondido correctamente (sobre un conjunto de ítems pre-configurado y con una mini-pizarra dispuesta por el moderador).			
¿Podrías explicarnos como llegaste a la respuesta correcta?	Profundización	Proceso de Respuesta sobre un ítem.	5'
¿Qué es lo que hay que saber para llegar a la respuesta correcta? ¿Dónde/Cuándo lo aprendieron?	Profundización	Proceso de Respuesta sobre un ítem.	2'
Se retoma la discusión grupal y se comienza a cerrar la entrevista.			
¿Hay algún otro ítem que les haya llamado particularmente la atención? ¿Porqué?	Indirecta	Juicios específicos sobre los ítems.	3'
¿Hay algo más que quieras compartir o comentar sobre esta experiencia?	Cierre		2'

C.4. Características y resultados de aplicaciones pre piloto.

La prueba pre piloto se aplicó en dos establecimientos educacionales, caracterizados en la siguiente tabla.

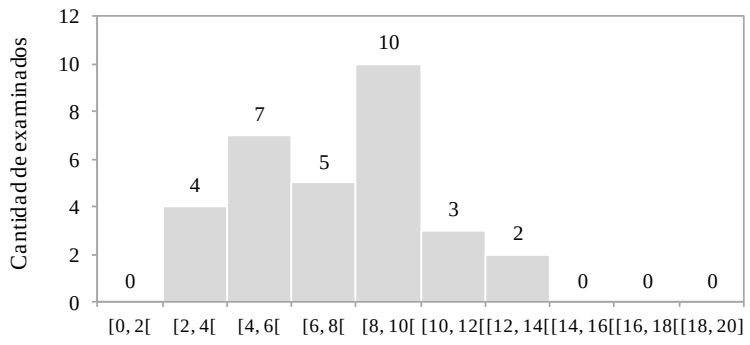
Características de los establecimientos participantes del pre pilotaje.

Características	Establecimiento **Nº 1**	Establecimiento **Nº 2**
Dependencia	P. Subvencionado	P. Pagado
Estudiantes examinados	**31**	**28**
Datos PSU[a]		
Estudiantes que la rinden	58	113
Puntaje Matemática	554	582
Datos SIMCE[b]		
Puntaje Matemática	309	387
Grupo Socioeconómico	Medio	Alto

[a] Datos correspondientes al proceso de admisión 2016. Fuente: www.demre.cl

[b] Datos correspondientes a la prueba aplicada durante el 2015. Fuente: www.simce.cl

Distribuciones de puntajes y estadísticos descriptivos para la prueba pre piloto Nº 1.

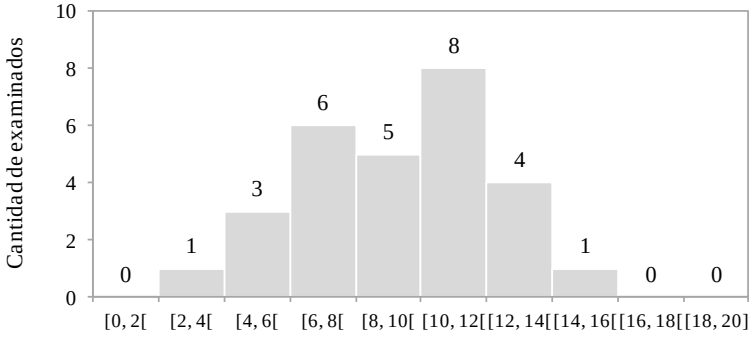


DESCRIPTIVE STATISTICS
PREPILOTO #1

Puntaje Total

Statistic	Value
N	31,0000
Valid N	31,0000
Min	2,0000
Max	12,0000
Mean	7,0323
Std. Dev.	2,8575
Skewness	-0,0897
Kurtosis	-0,8705
First Quartile	5,0000
Median	7,0000
Third Quartile	9,0000
IQR	4,0000

Distribuciones de puntajes y estadísticos descriptivos para la prueba pre piloto Nº 2.



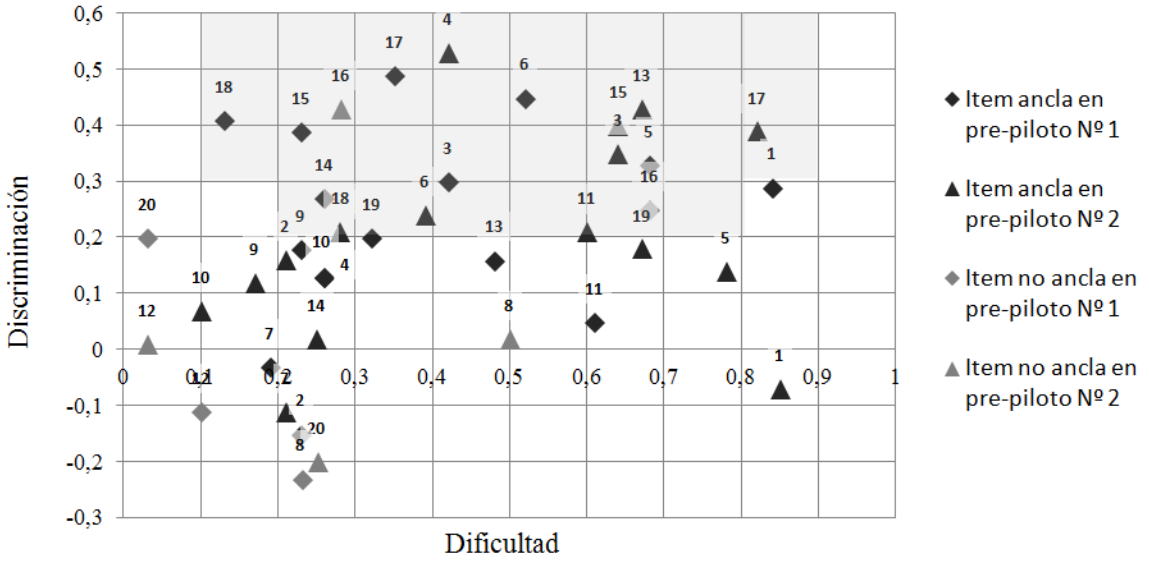
DESCRIPTIVE STATISTICS
Prepiloto #2

Puntaje Total

Statistic	Value
N	28,0000
Valid N	28,0000
Min	3,0000
Max	14,0000
Mean	8,8571
Std. Dev.	2,8767
Skewness	-0,1578
Kurtosis	-0,6006
First Quartile	7,0000
Median	9,0000
Third Quartile	10,7500
IQR	3,7500

En términos de dificultad y discriminación los ítems se representan en el siguiente diagrama, donde la cantidad de ítems fuera de la zona sombreada de aceptación, enfatizó la importancia de corregirlos para aumentar su discriminación.

Dispersión de los ítems pre-piloteados, según dificultad y discriminación.



La comparación detallada del desempeño de los ítems con los criterios de selección permitió fundamentar la siguientes decisiones: aceptar 6 ítems (1, 3, 5, 9, 13 y 18), aceptar 8 ítems después de ajustar sus distractores (4, 6, 11, 14, 15, 16, 17 y 19) , revisar 2 ítems (2 y 10) y eliminar 4 ítems (7, 8, 12 y 20).

Al ejecutar el procedimiento de Maentel–Hanzel para estimar el funcionamiento diferencial de los 13 ítems ancla, se obtuvo para el ítem Nº 17 un funcionamiento diferencial clase C ($\chi^2_{dif} = 7{,}2$) a favor de los examinados de género masculino. Esto justificó modificar algunas características del contexto del ítem.

Resultados de funcionamiento diferencial según género.

DIF ANALYSIS

Item	Chi-square	p-value	Valid N	E.S. (95% C.I.)	Class
1	0,50	0,48	21	3,22 (0,17, 62,55)	A
2	1,43	0,23	30	0,39 (0,08, 1,90)	A
3	0,00	0,96	38	0,97 (0,26, 3,55)	A
5	0,95	0,33	25	2,36 (0,37, 15,05)	A
6	0,55	0,46	38	1,80 (0,40, 8,16)	A
7	1,61	0,20	34	3,47 (0,54, 22,29)	A
9	1,53	0,22	34	3,19 (0,51, 19,79)	A
11	0,18	0,67	33	1,29 (0,36, 4,69)	A
13	0,01	0,94	34	0,94 (0,20, 4,44)	A
15	2,18	0,14	26	0,25 (0,04, 1,63)	A
17	7,22	0,01	24	0,03 (0,00, 0,59)	C+
18	0,53	0,47	22	0,49 (0,08, 3,23)	A
19	0,40	0,53	38	1,62 (0,39, 6,67)	A

Options

Matching Variable: pbruto
DIF Group Variable: género
Focal Group Code: Masculino
Reference Group Code: Femenino

C.5. Características y resultados de aplicaciones piloto.

Características y resultados de aplicación piloto, por establecimiento.

	Establecimientos					
Características	**Nº 1**	**Nº 2**	**Nº 3**	**Nº 4**	**Nº 5**	
Dependencia	P. Subv.	P. Pagado	P.Pagado	Municipal	Municipal	Global
Cant. de examinados (*n*)	68	38	32	65	30	240
Estadística PHC						
Media	7,6	7,5	8,1	7,7	7,4	7,81
Desv. Estándar	3,3	3,6	2,5	2,8	2,9	3,15
Mínimo	1,0	0,0	4,0	2,0	2	0,00
Cuartil 1	5,0	5,0	6,0	6,0	5	6,00
Mediana	7,0	7,0	8,0	8,0	7	7,00
Cuartil 3	9,8	10,0	10,0	10,0	10	10,00
Máximo	16,0	16,0	13,0	13,0	13	16,00
Curtosis	0,6	0,4	-0,9	-0,7	-8,9	-0,30
Asimetría	-0,1	-0,2	0,4	-0,2	0,0	0,31
α de Cronbach	0,72	0,76	0,49	0,60	0,61	0,67
Media PSU	487	507	588	540	517	
Media MAT	5,2	5,1	5,7	5,9	4,6	
Correlaciones						
PHC-PSU	0,55** (61)	0,70* (7)	0,62** (32)	0,20 (36)	-0,24 (13)	0,38** (154)
PHC-MAT	0,40** (62)	0,50** (34)	0,45** (31)	0,37** (49)	0,32* (29)	0,39** (210)
MAT-PSU	0,55** (60)	0,64 (6)	0,68** (31)	0,56** (35)	-0,08 (13)	0,50** (150)
Datos PSU[a]						
Puntaje Matemática	515	582	581	603	664	
Datos SIMCE[b]						
Puntaje en Matemática	308	387	333	334	322	
Grupo Socioeconómico	Medio	Alto	Alto	Medio Alto	Medio Alto	

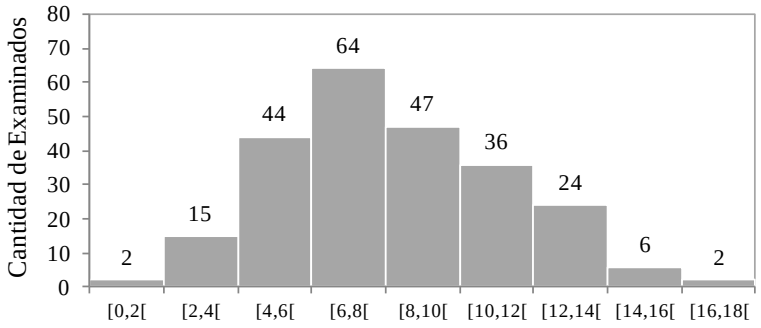
Nota. Las abreviaturas usadas son PHC: Puntaje en Prueba de Habilidades Cuantitativas, PSU: Puntaje promedio en ensayos PSU de Matemática (auto-reportado), MAT: Promedio de Notas de Matemática durante el primer semestre (auto-reportado). Entre paréntesis se muestra el número de casos utilizados para calcular cada coeficiente de correlación.

[a] Datos correspondientes al proceso de admisión 2016. Fuente: www.demre.cl

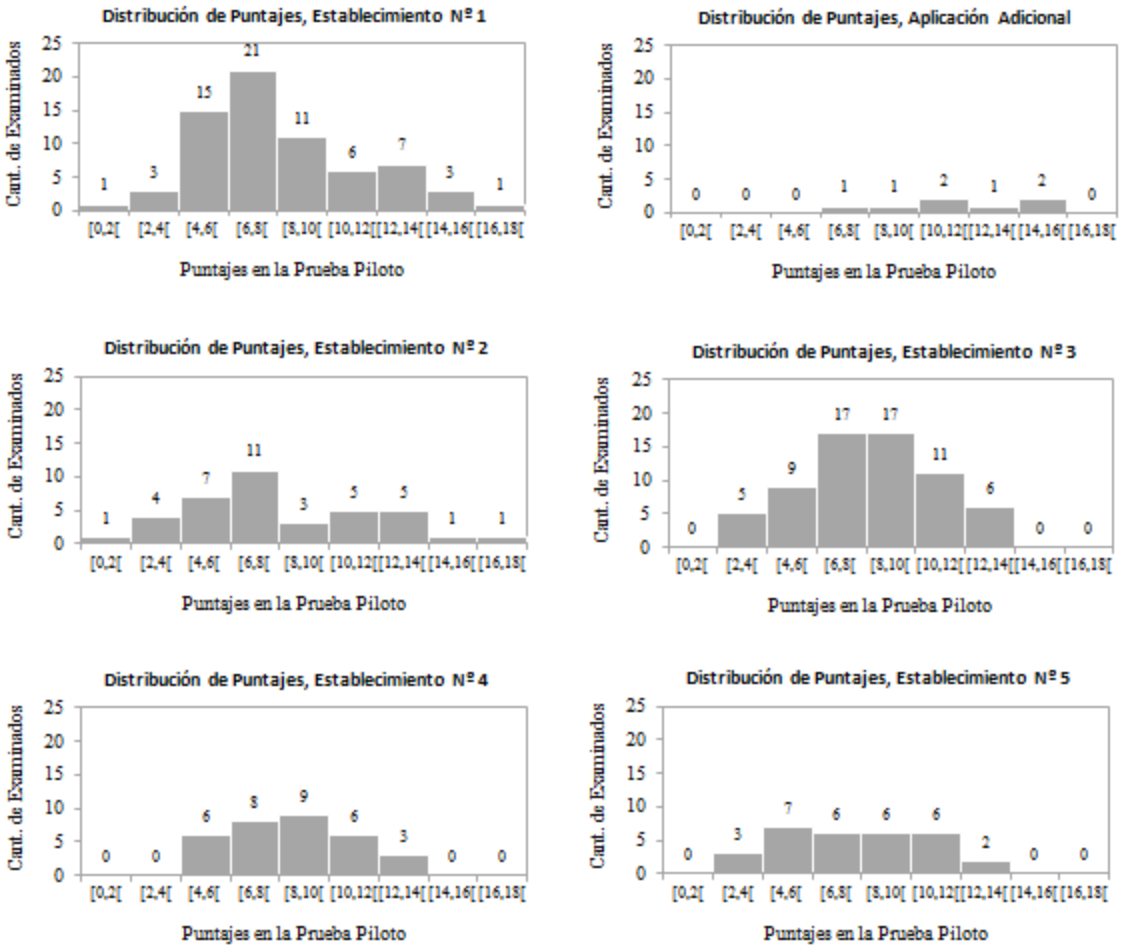
[b] Datos correspondientes a la prueba aplicada durante el 2015. Fuente: www.simce.cl

** La correlación es significativa al nivel 0,01 unilateral. * La correlación es significativa al nivel 0,05 unilateral.

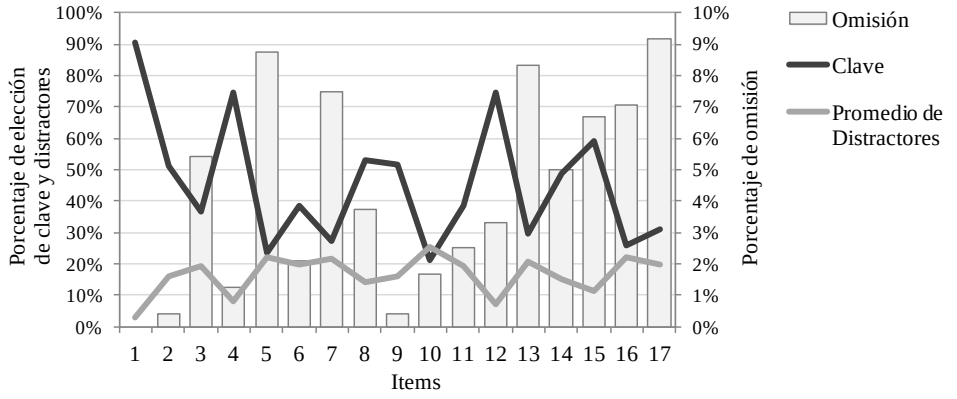
Distribuciones de puntajes y estadísticos descriptivos para la prueba piloto.



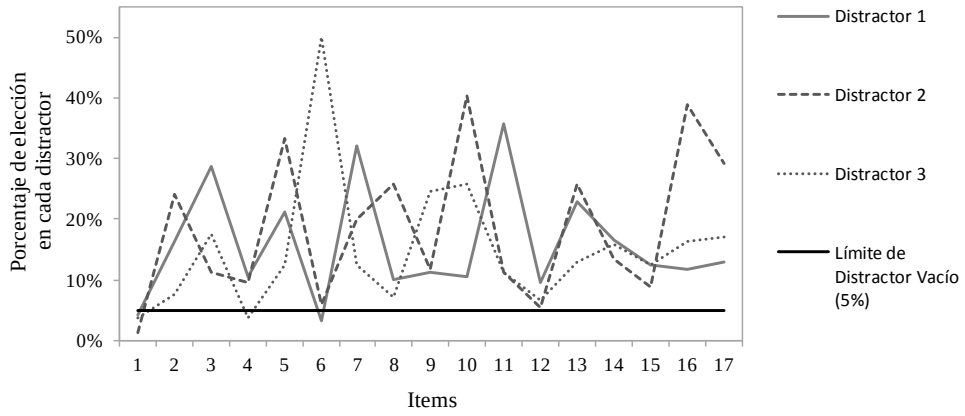
Statistic	Value
N	240,0000
Valid N	240,0000
Min	0,0000
Max	16,0000
Mean	7,7625
Std. Dev.	3,1233
Skewness	0,2866
Kurtosis	-0,3359
First Quartile	5,0000
Median	7,0000
Third Quartile	10,0000
IQR	5,0000



Omisión, frecuencia de la clave y frecuencia promedio de los distractores.



Frecuencia relativa de los distractores.



Funcionamiento diferencial de los ítems y de la prueba.

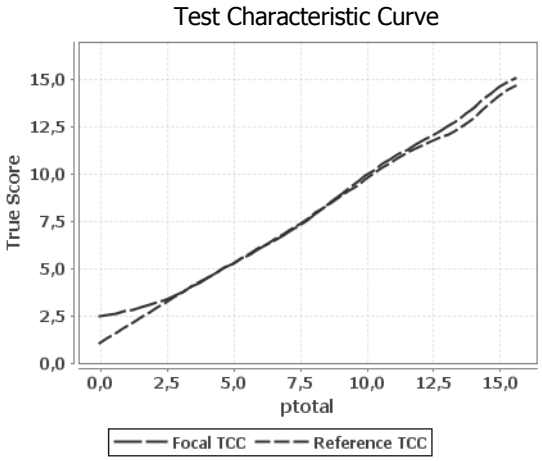
Funcionamiento diferencial, según género.
(Grupo Focal: Masculino, Grupo Referencial: Femenino)

DIF ANALYSIS

Item	Chi-square	p-value	Valid N	E.S. (95% C.I.)	Class
1	0,29	0,59	170	0,74 (0,26, 2,13)	A
2	0,95	0,33	219	0,75 (0,42, 1,34)	A
3	2,58	0,11	208	0,59 (0,31, 1,12)	A
4	0,57	0,45	222	0,78 (0,40, 1,51)	A
5	1,38	0,24	230	0,68 (0,36, 1,28)	A
6	1,33	0,25	225	0,70 (0,39, 1,29)	A
7	0,77	0,38	230	1,35 (0,70, 2,62)	A
8	0,59	0,44	230	1,26 (0,69, 2,30)	A
9	0,75	0,39	230	0,77 (0,42, 1,41)	A
10	0,19	0,66	230	1,17 (0,58, 2,37)	A
11	0,38	0,54	225	1,21 (0,66, 2,24)	A
12	12,15	0,00	222	3,51 (1,73, 7,13)	C-
13	1,81	0,18	207	0,63 (0,32, 1,24)	A
14	0,11	0,74	225	1,11 (0,60, 2,05)	A
15	7,63	0,01	225	2,38 (1,29, 4,41)	B-
16	2,95	0,09	220	0,57 (0,30, 1,08)	A
17	0,42	0,52	235	1,25 (0,65, 2,40)	A

Options

Matching Variable: ptotal
DIF Group Variable: genero
Focal Group Code: Masculino
Reference Group Code: Femenino



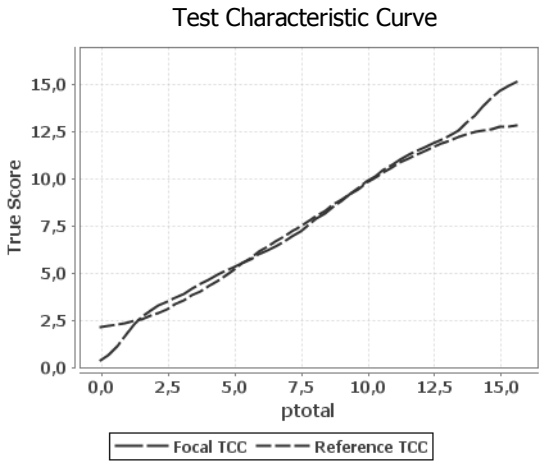
Funcionamiento diferencial, según dependencia educacional del establecimiento.
(Grupo Focal: Particular Pagado, Grupo Referencial: Municipal)

DIF ANALYSIS

Item	Chi-square	p-value	Valid N	E.S. (95% C.I.)	Class
1	0,87	0,35	98	1,76 (0,53, 5,77)	A
2	1,77	0,18	149	1,63 (0,80, 3,34)	A
3	0,21	0,65	149	0,83 (0,38, 1,82)	A
4	2,91	0,09	155	2,08 (0,91, 4,75)	A
5	0,13	0,71	160	1,16 (0,52, 2,60)	A
6	2,53	0,11	160	0,57 (0,29, 1,14)	A
7	0,37	0,54	152	1,27 (0,59, 2,76)	A
8	6,17	0,01	160	0,39 (0,18, 0,83)	B+
9	0,55	0,46	144	0,75 (0,36, 1,58)	A
10	0,45	0,50	149	0,75 (0,34, 1,68)	A
11	0,36	0,55	160	0,80 (0,39, 1,65)	A
12	1,30	0,25	144	0,61 (0,26, 1,41)	A
13	0,01	0,91	136	0,95 (0,42, 2,19)	A
14	0,06	0,80	149	0,91 (0,45, 1,86)	A
15	2,64	0,10	147	1,78 (0,88, 3,61)	A
16	7,32	0,01	152	3,63 (1,42, 9,24)	B-
17	0,31	0,58	152	0,81 (0,39, 1,69)	A

Options

Matching Variable: ptotal
DIF Group Variable: dep1
Focal Group Code: 3.0
Reference Group Code: 1.0



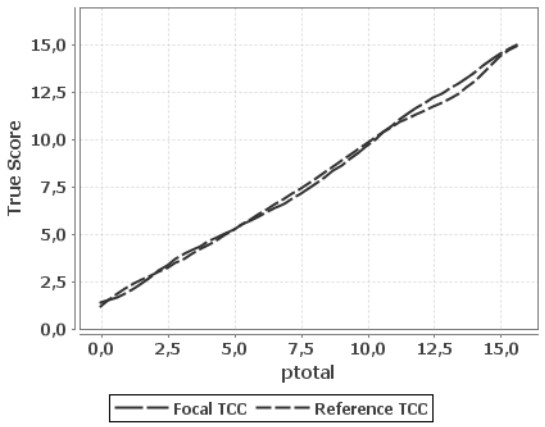
Funcionamiento diferencial, según nivel de enseñanza.
(Grupo Focal: Cuarto Medio, Grupo Referencial: Tercero Medio)

DIF ANALYSIS

Item	Chi-square	p-value	Valid N	E.S. (95% C.I.)	Class
1	0,21	0,65	170	1,27 (0,46, 3,53)	A
2	1,39	0,24	219	0,69 (0,37, 1,28)	A
3	1,96	0,16	208	0,62 (0,31, 1,24)	A
4	1,22	0,27	222	1,46 (0,74, 2,88)	A
5	0,83	0,36	230	0,72 (0,37, 1,43)	A
6	2,31	0,13	225	1,78 (0,86, 3,66)	A
7	1,84	0,17	230	1,64 (0,79, 3,42)	A
8	3,13	0,08	230	1,80 (0,93, 3,47)	A
9	0,39	0,53	230	1,22 (0,64, 2,32)	A
10	2,38	0,12	232	1,92 (0,83, 4,46)	A
11	0,40	0,53	225	0,82 (0,43, 1,55)	A
12	0,63	0,43	222	0,74 (0,36, 1,55)	A
13	7,22	0,01	209	0,35 (0,17, 0,75)	B+
14	1,22	0,27	225	1,43 (0,75, 2,74)	A
15	3,80	0,05	225	0,54 (0,29, 1,03)	A
16	0,14	0,70	220	0,88 (0,44, 1,74)	A
17	0,36	0,55	235	1,25 (0,61, 2,54)	A

Options

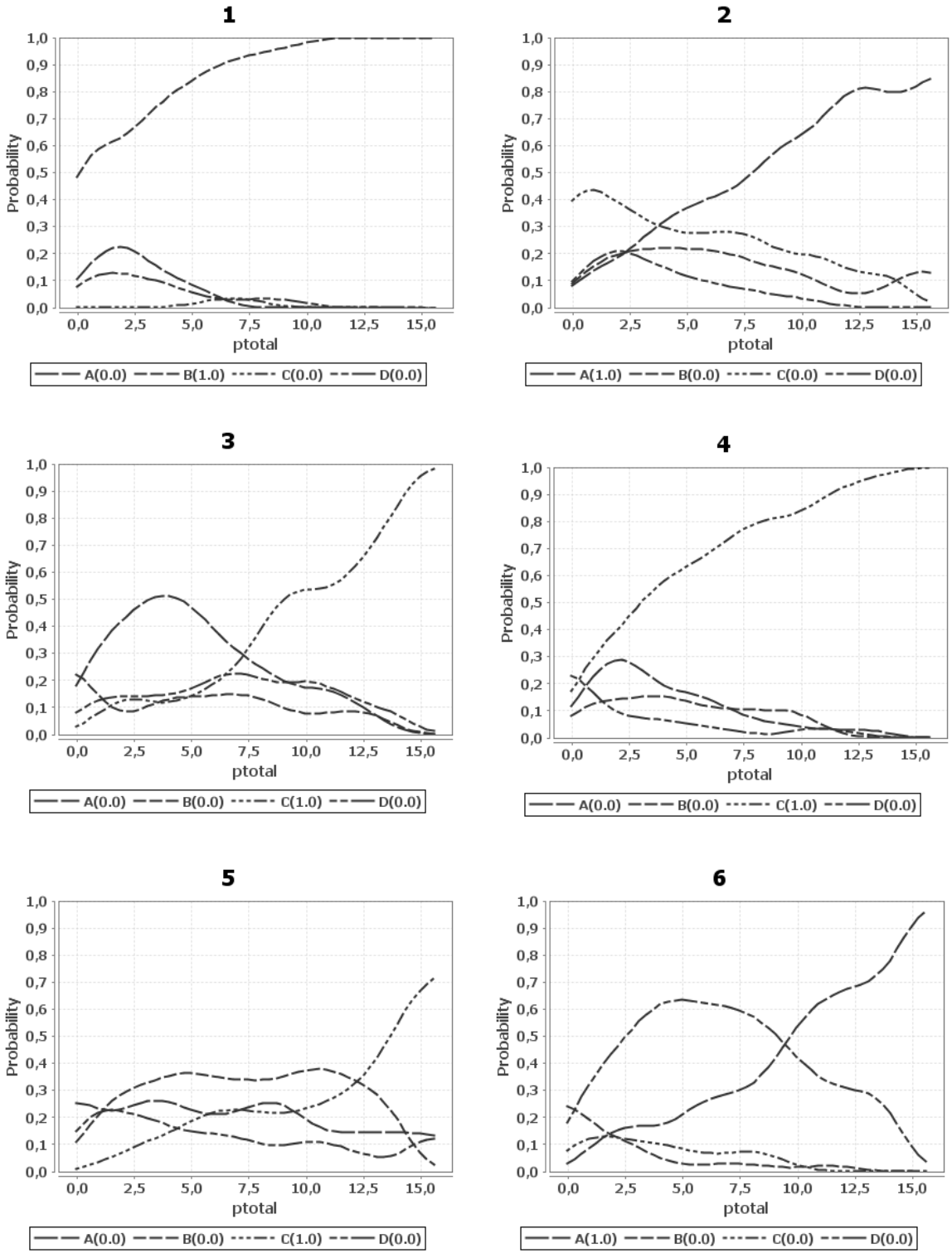
Matching Variable: ptotal
DIF Group Variable: nivel
Focal Group Code: 4.0
Reference Group Code: 3.0



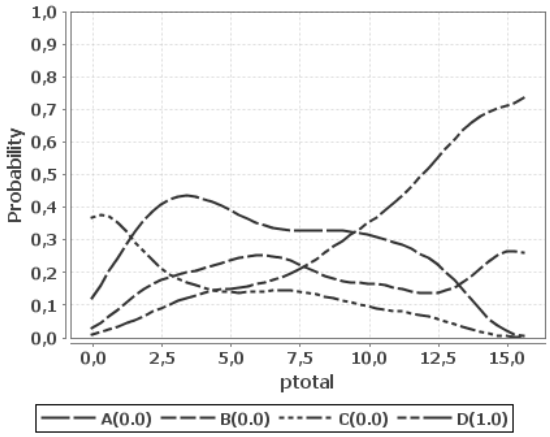
Indicadores psicométricos de los ítems.

Item	Option (Score)	Difficulty	Std. Dev.	Discrimin.
1	Overall	0,9042	0,2950	0,2415
	A(0.0)	0,0417	0,2002	-0,3167
	B(1.0)	0,9042	0,2950	0,2415
	C(0.0)	0,0125	0,1113	-0,0750
	D(0.0)	0,0375	0,1904	-0,2193
2	Overall	0,5125	0,5009	0,2343
	A(1.0)	0,5125	0,5009	0,2343
	B(0.0)	0,1625	0,3697	-0,2425
	C(0.0)	0,2417	0,4290	-0,3029
	D(0.0)	0,0750	0,2639	-0,2853
3	Overall	0,3667	0,4829	0,3442
	A(0.0)	0,2875	0,4535	-0,4310
	B(0.0)	0,1125	0,3166	-0,1859
	C(1.0)	0,3667	0,4829	0,3442
	D(0.0)	0,1750	0,3808	-0,1588
4	Overall	0,7458	0,4363	0,2340
	A(0.0)	0,1042	0,3061	-0,3215
	B(0.0)	0,0958	0,2950	-0,2379
	C(1.0)	0,7458	0,4363	0,2340
	D(0.0)	0,0375	0,1904	-0,1849
5	Overall	0,2375	0,4264	0,1130
	A(0.0)	0,2125	0,4099	-0,2086
	B(0.0)	0,3333	0,4724	-0,1877
	C(1.0)	0,2375	0,4264	0,1130
	D(0.0)	0,1250	0,3314	-0,2073
6	Overall	0,3833	0,4872	0,2949
	A(1.0)	0,3833	0,4872	0,2949
	B(0.0)	0,0333	0,1799	-0,1979
	C(0.0)	0,0583	0,2349	-0,2297
	D(0.0)	0,5000	0,5010	-0,3638
7	Overall	0,2750	0,4474	0,2471
	A(0.0)	0,3208	0,4678	-0,2910
	B(0.0)	0,2000	0,4008	-0,1585
	C(0.0)	0,1250	0,3314	-0,2576
	D(1.0)	0,2750	0,4474	0,2471
8	Overall	0,5292	0,5002	0,3363
	A(0.0)	0,1000	0,3006	-0,2572
	B(1.0)	0,5292	0,5002	0,3363
	C(0.0)	0,2583	0,4386	-0,4303
	D(0.0)	0,0708	0,2571	-0,2145
9	Overall	0,5167	0,5008	0,3123
	A(1.0)	0,5167	0,5008	0,3123
	B(0.0)	0,1125	0,3166	-0,1448
	C(0.0)	0,1167	0,3217	-0,1892
	D(0.0)	0,2458	0,4315	-0,4986
10	Overall	0,2125	0,4099	0,2135
	A(0.0)	0,1042	0,3061	-0,1918
	B(0.0)	0,4042	0,4918	-0,3267
	C(0.0)	0,2583	0,4386	-0,1340
	D(1.0)	0,2125	0,4099	0,2135
11	Overall	0,3875	0,4882	0,2913
	A(0.0)	0,3583	0,4805	-0,3265
	B(0.0)	0,1125	0,3166	-0,2872
	C(1.0)	0,3875	0,4882	0,2913
	D(0.0)	0,1125	0,3166	-0,1982
12	Overall	0,7458	0,4363	0,2908
	A(0.0)	0,0958	0,2950	-0,2596
	B(0.0)	0,0542	0,2268	-0,2512
	C(0.0)	0,0667	0,2500	-0,2381
	D(1.0)	0,7458	0,4363	0,2908
13	Overall	0,2958	0,4574	0,3686
	A(0.0)	0,2292	0,4212	-0,3470
	B(1.0)	0,2958	0,4574	0,3686
	C(0.0)	0,2583	0,4386	-0,2623
	D(0.0)	0,1292	0,3361	-0,2673
14	Overall	0,4875	0,5009	0,3342
	A(1.0)	0,4875	0,5009	0,3342
	B(0.0)	0,1667	0,3735	-0,2721
	C(0.0)	0,1333	0,3406	-0,3102
	D(0.0)	0,1583	0,3658	-0,2713
15	Overall	0,5917	0,4926	0,2218
	A(0.0)	0,1250	0,3314	-0,2844
	B(0.0)	0,0875	0,2832	-0,2877
	C(0.0)	0,1250	0,3314	-0,1367
	D(1.0)	0,5917	0,4926	0,2218
16	Overall	0,2583	0,4386	0,2122
	A(0.0)	0,1167	0,3217	-0,1932
	B(1.0)	0,2583	0,4386	0,2122
	C(0.0)	0,3875	0,4882	-0,1963
	D(0.0)	0,1625	0,3697	-0,2494
17	Overall	0,3125	0,4645	0,2271
	A(0.0)	0,1292	0,3361	-0,2369
	B(0.0)	0,2917	0,4555	-0,1766
	C(1.0)	0,3125	0,4645	0,2271
	D(0.0)	0,1708	0,3771	-0,2373

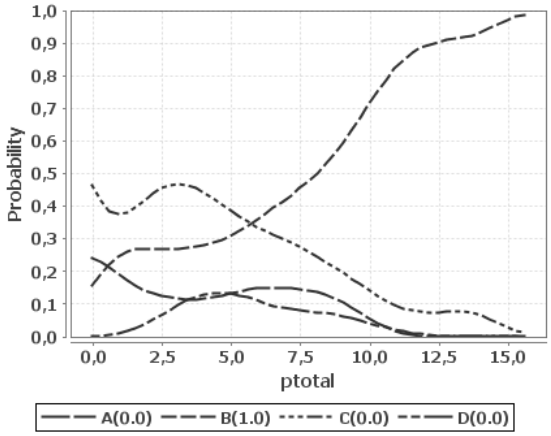
Curvas características de los ítems.



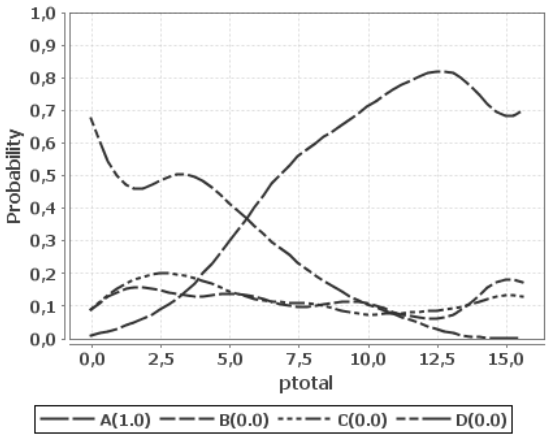
7


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(0.0) B(0.0) C(0.0) D(1.0)

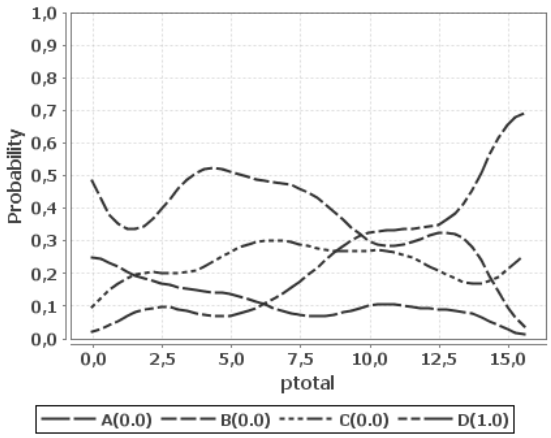
8


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(0.0) B(1.0) C(0.0) D(0.0)

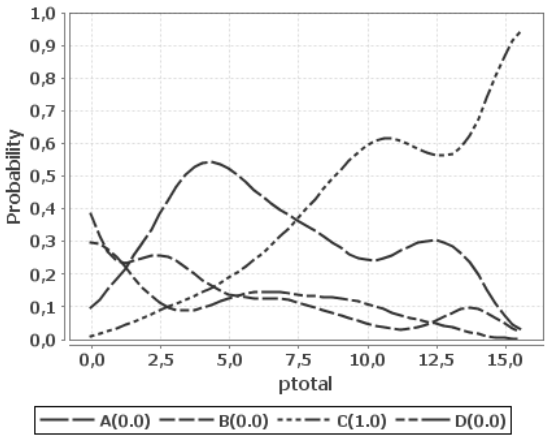
9


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(1.0) B(0.0) C(0.0) D(0.0)

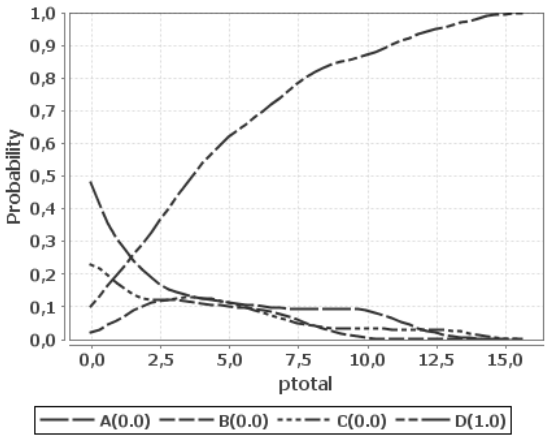
10


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(0.0) B(0.0) C(0.0) D(1.0)

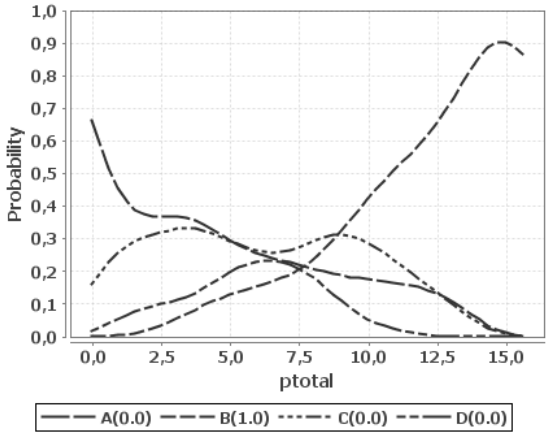
11


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(0.0) B(0.0) C(1.0) D(0.0)

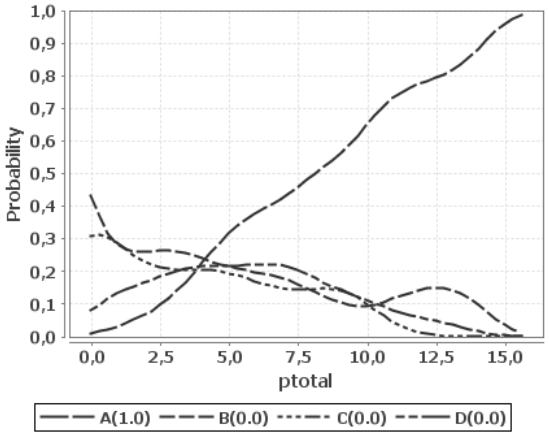
12


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0 2,5 5,0 7,5 10,0 12,5 15,0
ptotal
A(0.0) B(0.0) C(0.0) D(1.0)

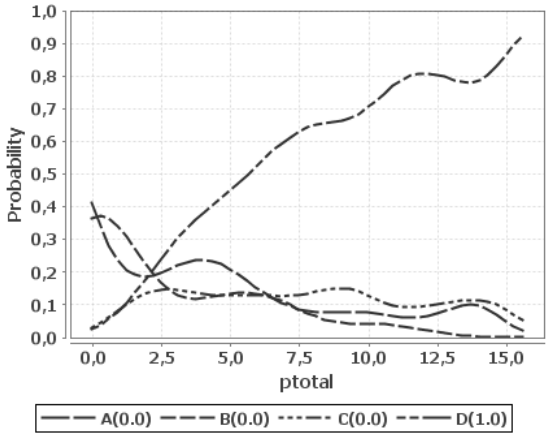
13


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
ptotal
A(0.0)
B(1.0)
C(0.0)
D(0.0)

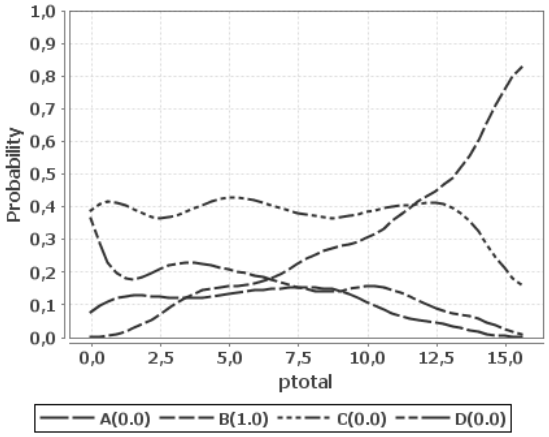
14


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
ptotal
A(1.0)
B(0.0)
C(0.0)
D(0.0)

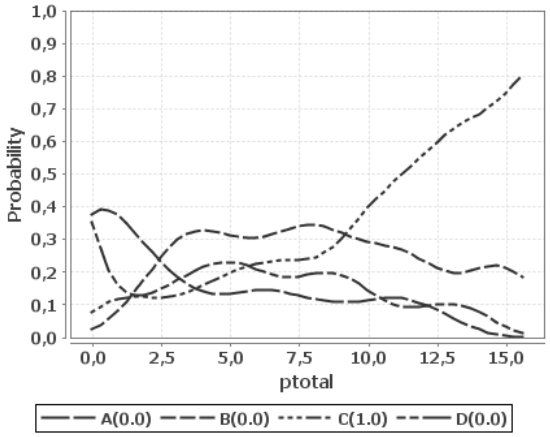
15


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
ptotal
A(0.0)
B(0.0)
C(0.0)
D(1.0)

16


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
ptotal
A(0.0)
B(1.0)
C(0.0)
D(0.0)

17


Probability
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
ptotal
A(0.0)
B(0.0)
C(1.0)
D(0.0)

C.6. Criterios de evaluación del desempeño psicométrico de los ítems.

Indicador	Recorrido	Criterio	Justificación
Dificultad ($\boldsymbol{p}$)	$0 < \boldsymbol{p} < 1$	$0{,}2 < \boldsymbol{p} < 0{,}8$	Rango de dificultad en el cual los ítems alcanzan una mayor discriminación[a]. Se pueden incluir ítems con dificultades entre 0,1 y 0,2 ó entre 0,8 y 0,9, siempre que tengan una capacidad discriminativa superior a 0,3.
Capacidad Discriminativa ($\mathbf{r}_{pbis}$)	$-1 < r_{pbis} < 1$	$r_{pbis} > 0{,}2$	Valor esperado para obtener nivel estimado mínimo de confiabilidad (α>0,6)[a].
Distractores ($\boldsymbol{d}$)	$0 < \boldsymbol{d} < 1$	$\boldsymbol{d} > 0{,}05$	Valor mínimo de frecuencia para considerar al distractor no-vacío.[b]
	$0 < \boldsymbol{d_p} < 0{,}33$	$\boldsymbol{d_p} < \boldsymbol{p}$	Valor que el promedio de las frecuencias de los distractores ($\boldsymbol{d_p}$) no debería superar.[c]
	$-1 < r_{pbis} < 1$	$r_{pbis} < 0$	Valor coherente con una respuesta incorrecta, que se relaciona negativamente con el puntaje total de la prueba.[c]
Omisión ($\boldsymbol{o}$)	$0 < \boldsymbol{o} < 1$	$\boldsymbol{o} < 0{,}05$	Valor esperado considerando que en la prueba no se aplica descuento por respuestas incorrectas.
Funcionamiento Diferencial	A, B ó C	A ó B	Clases que indican que el ítem puede considerarse libre de funcionamiento diferencial (A) ó con un funcionamiento diferencial moderado que permite usarlo discrecionalmente, si no hay un item equivalente para reemplazarlo (B).[d]

[a] Según datos obtenidos en el pre-pilotaje.

[b] Según Haladyna y Downing, 1993.

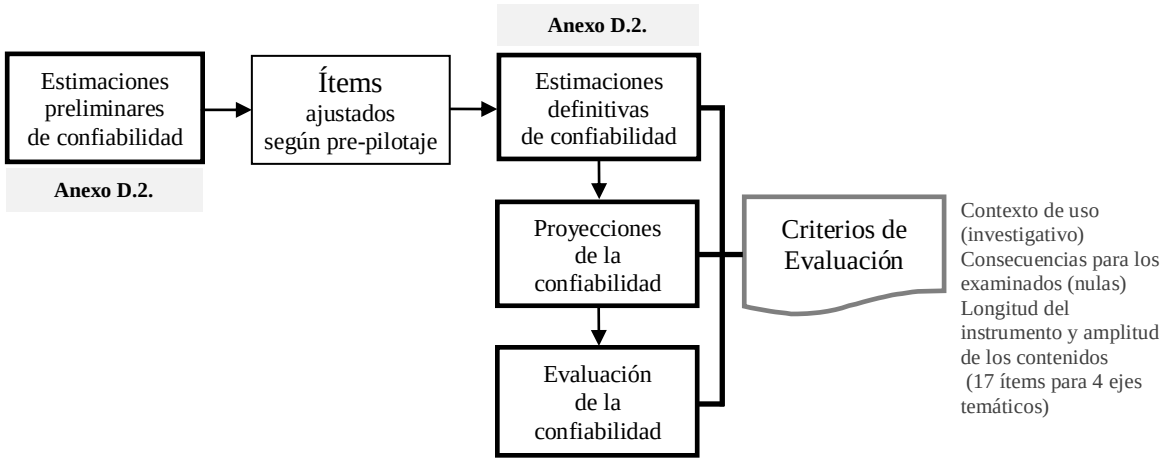
[c] Según Haladyna en Lane 2014.

[d] Según Zwick y Erickan, 1989.

Anexo D. Sobre la evaluación de la confiabilidad

D.1. Proceso de evaluación de la confiabilidad

El proceso de evaluación de la confiabilidad implicó estimaciones preliminares en la etapa de pre-pilotaje y estimaciones finales en la etapa de pilotaje a través del índice Alpha de Cronbach.



D.2. Detalles de la estimación de la confiabilidad.

Estimación preliminar de la confiabilidad (pre-pilotaje).

	Cantidad de Ítems	Alpha de Cronbach	Discriminación Promedio[a]
Prepiloto Nº 1 (*n* =31)	20	0,56	0,19
Prepiloto Nº 2 (*n* =28)	20	0,57	0,18
Prepiloto Nº 1 Reducido[b]	15	0,70	0,30
Prepiloto Nº 2 Reducido[b]	15	0,67	0,27

[a] Prueba después de eliminar los cinco ítems con menor capacidad discriminativa.

[b] Promedio de la capacidad discriminativa de los ítems. Calculada a través de su correlación biserial puntual con el puntaje de la prueba, corregida su espureidad.

Estimación de la confiabilidad (pilotaje).

	Cantidad de Ítems	Alpha de Cronbach	Desviación Estándar	Error Estándar de Medición	Discriminación Promedio[b]
Piloto	17	0,67	3,12	1,79	0,27
Piloto Reducido[a]	16	0,68	3,05	1,72	0,28
Sub-puntuación de categoría de habilidad					
Comprensión cuantitativa	11	0,59	2,21	1,41	
Razonamiento cuantitativo	6	0,38	1,40	1,10	
Sub-puntuación de categorías de ejes de contenidos					
Números, Funciones y Geometría	8	0,49	1,69	1,21	
Datos y Azar	9	0,55	1,96	1,32	

[a] Prueba después de eliminar el ítem con menor capacidad discriminativa.

[b] Promedio de la capacidad discriminativa de los ítems. Calculada a través de su correlación biserial puntual con el puntaje de la prueba, corregida su espureidad.

Impacto en la confiabilidad de la eliminación de cada ítem.

Estadísticos de fiabilidad

Alfa de Cronbach	N de elementos
,672	17

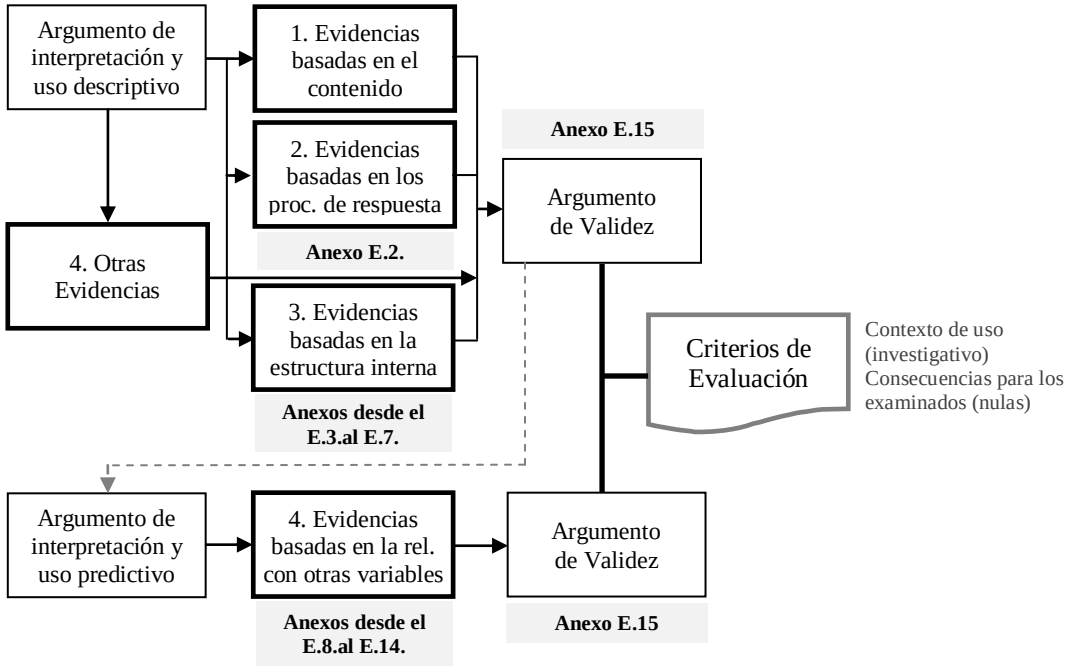
Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
it1	6,86	9,235	,242	,662
it2	7,25	8,808	,234	,662
it3	7,40	8,550	,344	,648
it4	7,02	8,954	,234	,662
it5	7,53	9,280	,113	,675
it6	7,38	8,672	,295	,654
it7	7,49	8,895	,247	,660
it8	7,23	8,523	,336	,649
it9	7,25	8,588	,312	,652
it10	7,55	9,060	,213	,664
it11	7,38	8,679	,291	,655
it12	7,02	8,811	,291	,655
it13	7,47	8,560	,369	,645
it14	7,28	8,527	,334	,649
it15	7,17	8,862	,222	,664
it16	7,50	9,004	,212	,664
it17	7,45	8,910	,227	,663

Anexo E. Sobre la evaluación de la validez

E.1. Proceso de evaluación de la validez

El proceso de evaluación de la validez implicó la recogida de diversas evidencias durante el desarrollo y aplicación de la prueba, descritas en el siguiente flujo.



E.2. Procesos de respuesta incorrecta y correcta para un ítem.

Ítem

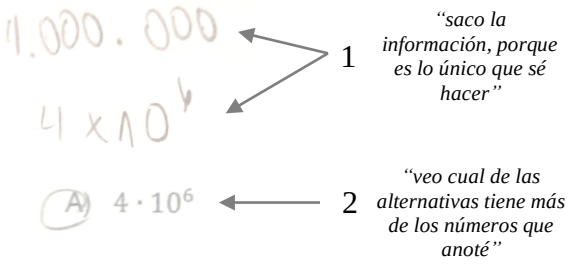
Una persona estima que habría vida en cuatro millones de planetas de nuestra galaxia. Asumiendo que por cada medio millón de estrellas habría un planeta con vida, ¿cuántas estrellas consideró en su estimación?

Procesos de respuesta incorrecta

A) $4 \cdot 10^6$

Un entrevistado de bajo desempeño en el test declaró: *"saco la información, porque es lo único que sé hacer…y veo cual de las alternativas tiene más de los números que anoté"*. Esto reivindica la importancia de considerar distractores que capturen respuestas de examinados, que intentan llegar a la clave por procedimientos no matemáticos.

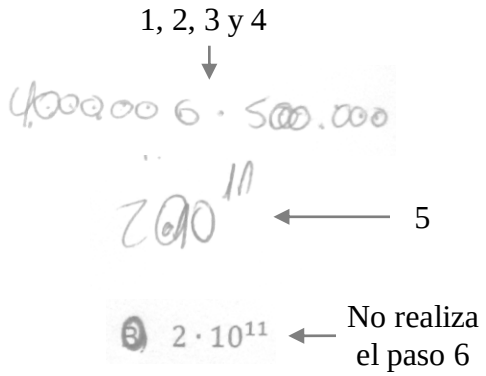
PC	Paso	Proceso de respuesta incorrecta
Comprensión	1	Identificar las datos presentes en el problema.
	2	Relacionar a través de un procedimiento no matemático, estos datos con una de las opciones del ítem.



B) $2 \cdot 10^{11}$

Se espera que examinados con un nivel más alto de habilidad cuantitativa cometan un error menor en el desarrollo de su respuesta, como por ejemplo no representar el resultado con la potencia correcta.

PC	Paso	Proceso de respuesta incorrecta
Comprensión	1	Identificar las dos variables presentes en el enunciado; estrellas y planetas con vida.
	2	Expresar las cantidades mencionadas para esta variables, con potencias de base 10.
Análisis	3	Identificar la incógnita correspondiente a la pregunta.
	4	Plantear la proporción correspondiente.
	5	Resolver la proporción planteada.
Evaluación	6	No verificar que la solución es pertinente con las características del problema.

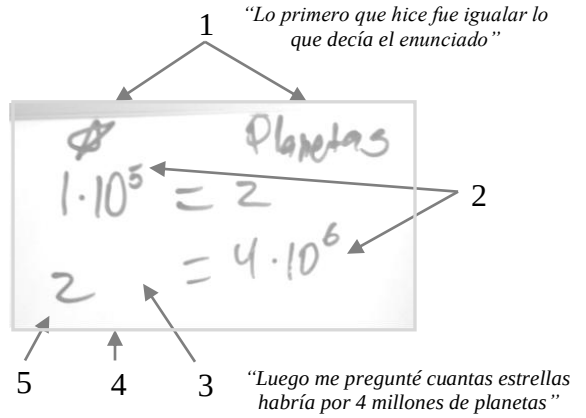


Proceso de respuesta correcta

C) $2 \cdot 10^{12}$

PC	Paso	Proceso de respuesta correcta
Comprensión	1	Identificar las dos variables presentes en el enunciado; estrellas y planetas con vida.
	2	Expresar las cantidades mencionadas para esta variables, con potencias de base 10.
Análisis	3	Identificar la incógnita correspondiente a la pregunta.
	4	Plantear la proporción correspondiente.
	5	Resolver la proporción planteada.
Evaluación	6	Verificar que la solución es pertinente con las características del problema.

Nota. PC indica el proceso cognitivo involucrado en la resolcuón correcta del ítem, según la taxonomía de Bloom revisada (Anderson, 2001).



Aunque el entrevistado no escribe la letra *x* para señalar la incógnita, ni expresa la proporción como igualdad de razones, ni calcula el resultado final completo, su desarrollo evidencia que ejecuta comprensivamente estos procesos, sin darse el tiempo para escribirlos. De hecho, omite anotar la operación realizada para llegar a la potencia ($10^5 \cdot 10^6 = 10^{11}$) que es parte de la clave ($2 \cdot 10^{11}$).

E.3. Índice de complejidad cognitiva y dificultad empírica de los ítems.

Ítem[a]	Complejidad del Contenido[b]	Complejidad de la Habilidad[c]	Complejidad del Texto[d]	Complejidad Cognitiva[e]	Dificultad Empírica[f]
1*	1	3	1	5	0,90
2	1	4	2	7	0,51
3	2	5	2	9	0,37
4	2	1	2	5	0,75
5	3	3	3	9	0,24
6	3	2	2	7	0,38
7	3	5	3	11	0,28
8	2	3	2	7	0,53
9	2	1	1	4	0,52
10*	2	2	3	7	0,21
11	3	3	2	8	0,39
12*	2	3	2	7	0,75
13	3	3	2	8	0,30
14	2	3	2	7	0,49
15*	3	5	3	11	0,59
16	3	5	3	11	0,26
17	3	4	3	10	0,31

[a] Los ítems marcados con asterisco, se alejan más de 1 desviación estándar de la tendencia.

[b] Valor asignado por un juez revisor que puede ser; 1: Baja, 2: Mediana y 3:Alta complejidad. Según descripición en tabla adjunta.

[c] Valor definido en la especificación del ítem que puede ser; 1: Interpretar (Comprender), 2: Representar (Comprender), 3: Calcular (Aplicar), 4: Discriminar (Analizar) y 5: Comprobar (Evaluar).

[b] Valor asignado por constructor del ítem que puede ser; 1: Baja, 2: Mediana y 3:Alta complejidad. Según descripición en tabla adjunta.

[e] Índice de complejidad cognitiva equivalente a la suma de los tres valores anteriores.

[f] Dificultad empírica de cada ítem como porcentaje de respuestas correctas en la muestra.

Complejidad del Contenido

Nivel	Descripción
Alta	El contenido necesario para resolver el ítem corresponde a conceptos matemáticos que requieren compresión previa de fórmulas y propiedades y/o procedimientos algorítmicos específicos de múltiples pasos, que requieren ejercitación previa.
Media	El contenido necesario para resolver el ítem corresponde a conceptos matemáticos que pueden ser comprendidos a partir de conceptos básicos previos y/o procedimientos algorítmicos breves.
Baja	El contenido necesario para resolver el ítem corresponde a conceptos matemáticos básicos que no requieren estudio específico previo y/o procedimientos algorítmicos directos, u operaciones aritméticas fundamentales.

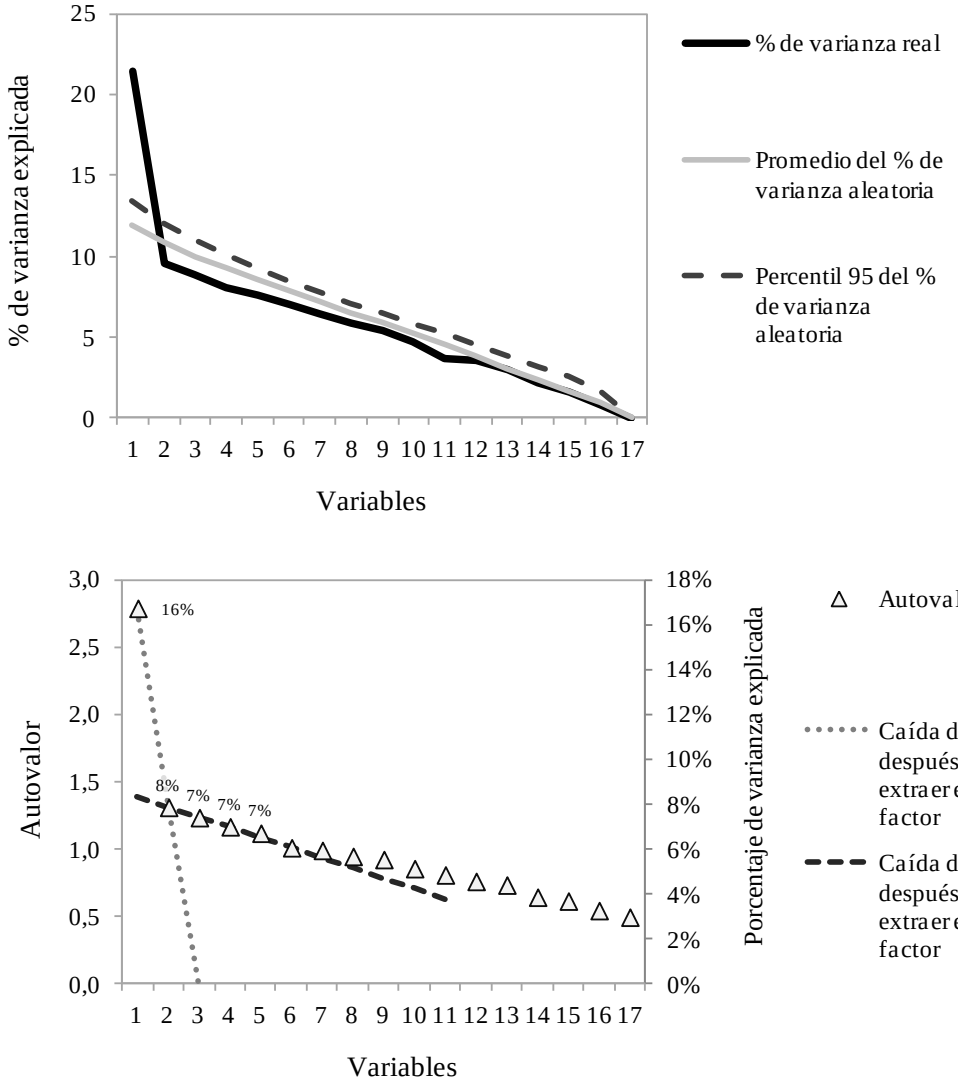
Adaptado de los tipos de conocimientos en la taxonomía revisada de Bloom (Anderson, 2001) y de la complejidad cogntiva de un ítem (Schneider, 2013).

Complejidad del Texto

Nivel	Descripción
Alta	El ítem incluye una gran cantidad de texto (por ejemplo: muchas frases), del cual la mayoría es linguísticamente complejo o matemáticamente complejo debido a la terminología y conceptos utilizados. Elementos en el texto podrían estar interrelacionados, y requieren de los examinados entender estas interrelaciones para responder el ítem. El formato del texto del ítem y de las imágenes podrían ser complejas. Por ejemplo, el ítem podría requerir a los examinados leer, procesar y entender varias secciones del texto y sus imágenes para lograr seleccionar o construir la respuesta correcta.
Media	El ítem incluye una baja cantidad de texto e imágenes. Sin embargo, requieren de los examinados usar múltiples pasos para procesar y responder. Y podría incluir frases cortas y andamiaje para conducir a los examinados a través del texto para seleccionar la respuesta correcta.
Baja	El ítem incluye una pequeña cantidad de texto (por ejemplo: una frase o una frase introductoria) relacionada únicamente con la Matemática requerida para responder. La forma en que los examinados procesan el item es obvia o está explícitamente declarada.

Traducido y adaptado de Ferrara el al. (2011, p.7). La categoría texto incluye enunciado, estímulo y pregunta del ítem.

E.4. Gráfico de análisis paralelo y gráfico de sedimentación.



E.5. Resultados de análisis paralelo y análisis factorial con mínimos cuadrados no ponderados.

```
                    F A C T O R

                Unrestricted Factor Analysis
              Release Version 10.3.01  x32bits
                        July, 2015
                Rovira i Virgili University
                    Tarragona, SPAIN

                       Programming:
                   Urbano Lorenzo-Seva

                Mathematical Specification:
                   Urbano Lorenzo-Seva
                     Pere J. Ferrando

--------------------------------------------------------------------------------

DETAILS OF ANALYSIS

Participants' scores data file                      : D:\III-MAGISTER\4-Proyecto\Items-Testeo\Pilotaje\ALL\PHC-TABULACION-ALL_FACTOR.dat
Variable labels file                                : D:\III-MAGISTER\4-Proyecto\Items-Testeo\Pilotaje\ALL\PHC-ETIQUETAS-Reducidas.txt
Method to handle missing values                     : Hot-Deck Multiple Imputation in Exploratory Factor Analysis (Lorenzo-Seva & Van Ginkel, 2015)
Missing code value                                  : 999
Number of participants                              : 240
Number of variables                                 : 17
Variables included in the analysis                  : ALL
Variables excluded in the analysis                  : NONE
Number of factors                                   : 1
Number of second order factors                      : 0
Procedure for determining the number of dimensions  : Optimal implementation of Parallel Analysis (PA) (Timmerman, & Lorenzo-Seva, 2011)
Dispersion matrix                                   : Pearson Correlations
Method for factor extraction                        : Unweighted Least Squares (ULS)
Rotation to achieve factor simplicity               : Promin (Lorenzo-Seva, 1999)
Clever rotation start                               : Weighted Varimax
Number of random starts                             : 10
Maximum mumber of iterations                        : 100
Convergence value                                   : 0.00001000

--------------------------------------------------------------------------------

UNIVARIATE DESCRIPTIVES
Variable                                            Mean       Confidence Interval     Variance   Skewness     Kurtosis
                                                               (95%)                               (Zero centered)

Item 1                                              0.904     (   0.86     0.95)       0.087     -2.758        5.577
Item 2                                              0.512     (   0.43     0.60)       0.250     -0.050       -1.993
Item 3                                              0.367     (   0.29     0.45)       0.232      0.556       -1.688
Item 4                                              0.746     (   0.67     0.82)       0.190     -1.134       -0.715
Item 5                                              0.237     (   0.17     0.31)       0.181      1.239       -0.467
Item 6                                              0.383     (   0.30     0.46)       0.236      0.482       -1.765
Item 7                                              0.275     (   0.20     0.35)       0.199      1.012       -0.976
Item 8                                              0.529     (   0.45     0.61)       0.249     -0.117       -1.982
Item 9                                              0.517     (   0.43     0.60)       0.250     -0.067       -1.991
Item 10                                             0.213     (   0.14     0.28)       0.167      1.411       -0.012
Item 11                                             0.387     (   0.31     0.47)       0.237      0.464       -1.782
Item 12                                             0.746     (   0.67     0.82)       0.190     -1.134       -0.715
Item 13                                             0.296     (   0.22     0.37)       0.208      0.898       -1.192
Item 14                                             0.488     (   0.40     0.57)       0.250      0.050       -1.993
Item 15                                             0.592     (   0.51     0.67)       0.242     -0.375       -1.856
Item 16                                             0.258     (   0.19     0.33)       0.192      1.109       -0.771
Item 17                                             0.313     (   0.24     0.39)       0.215      0.812       -1.339

Polychoric correlation is advised when the univariate distributions of ordinal items are
asymmetric or with excess of kurtosis. If both indices are lower than one in absolute value,
then Pearson correlation is advised. You can read more about this subject in:

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables.
British Journal of Mathematical and Statistical Psychology, 38, 171-189.
Muthén, B., & Kaplan D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables:
A note on the size of the model. British Journal of Mathematical and Statistical Psychology, 45, 19-30.

--------------------------------------------------------------------------------

MULTIVARIATE DESCRIPTIVES
Analysis of the Mardia's (1970) multivariate asymmetry skewness and kurtosis.

                                        Coefficient        Statistic     df       P

Skewness                                     34.519         1380.759    969     1.0000
SKewness corrected for small sample          34.519         1399.955    969     1.0000
Kurtosis                                    304.285           -5.704            0.0000**

** Significant at 0.05

--------------------------------------------------------------------------------

STANDARIZED VARIANCE / COVARIANCE MATRIX (PEARSON CORRELATION)

Variable     1        2        3        4        5        6        7        8        9       10       11       12       13       14       15       16       17
V   1        1.000
V   2        0.136    1.000
V   3        0.130    0.137    1.000
V   4        0.168    0.062    0.107    1.000
V   5        0.015    0.055    0.104   -0.101    1.000
V   6        0.053    0.220    0.165    0.027    0.084    1.000
V   7        0.042    0.097    0.151    0.102    0.051    0.205    1.000
V   8        0.062    0.065    0.181    0.197    0.036    0.177    0.207    1.000
V   9        0.167    0.058    0.217    0.240    0.089    0.197   -0.002    0.190    1.000
V  10        0.100    0.119    0.133    0.163    0.021    0.156    0.022    0.123    0.095    1.000
V  11        0.143    0.160    0.140    0.052   -0.022    0.164    0.066    0.133    0.085    0.151    1.000
V  12        0.103    0.062    0.087    0.099    0.078    0.027    0.124    0.140    0.221    0.046    0.189    1.000
V  13        0.180    0.048    0.208    0.085    0.110    0.090    0.132    0.246    0.115    0.110    0.122    0.127    1.000
V  14        0.119    0.101    0.209    0.129    0.063    0.088    0.127    0.185    0.126    0.145    0.131    0.148    0.171    1.000
V  15        0.162    0.038    0.087    0.002    0.105   -0.025    0.132   -0.036    0.096   -0.024    0.139    0.196    0.167    0.217    1.000
V  16       -0.002    0.080    0.025    0.170    0.029    0.102    0.042    0.080    0.037    0.042    0.136    0.104    0.160    0.053    0.180    1.000
V  17       -0.025    0.100    0.103    0.043    0.004    0.134    0.068    0.150    0.094   -0.021    0.091    0.105    0.292    0.116    0.048    0.116    1.000

--------------------------------------------------------------------------------

ADEQUACY OF THE CORRELATION MATRIX

Determinant of the matrix      = 0.226900069130288
Bartlett's statistic           =   344.9 (df =   136; P = 0.000010)
Kaiser-Meyer-Olkin (KMO) test = 0.69121 (mediocre)

--------------------------------------------------------------------------------
```

EXPLAINED VARIANCE BASED ON EIGENVALUES

Variable	Eigenvalue	Proportion of Variance	Cumulative Proportion of Variance
1	2.78899	0.16406	0.16406
2	1.91315	0.11254	
3	1.23716	0.07277	
4	1.16921	0.06878	
5	1.12149	0.06597	
6	1.01373	0.05963	
7	0.99409	0.05848	
8	0.94972	0.05587	
9	0.92648	0.05450	
10	0.85897	0.05053	
11	0.81229	0.04778	
12	0.76424	0.04496	
13	0.73739	0.04338	
14	0.64724	0.03807	
15	0.61866	0.03639	
16	0.54812	0.03224	
17	0.49908	0.02936	

PARALLEL ANALYSIS (PA) BASED ON MINIMUM RANK FACTOR ANALYSIS
(Timmerman & Lorenzo-Seva, 2011)

Implementation details:

Correlation matrices analized: Pearson correlation matrices
Number of random correlation matrices: 500
Method to obtain random correlation matrices: Permutation of the raw data (Buja & Eyuboglu, 1992)

Variable	Real-data % of variance	Mean of random % of variance	95 percentile of random % of variance
1	21.4*	11.9	13.4
2	9.8	10.9	12.0
3	8.9	10.0	11.0
4	8.1	9.3	10.1
5	7.6	8.6	9.3
6	7.0	7.9	8.5
7	6.5	7.2	7.8
8	5.9	6.5	7.1
9	5.4	5.9	6.5
10	4.7	5.2	5.8
11	3.7	4.5	5.2
12	3.6	3.8	4.5
13	3.0	3.1	3.9
14	2.2	2.4	3.2
15	1.6	1.7	2.6
16	0.8	1.0	1.7
17	0.0	0.0	0.0

* Advised number of dimensions: 1

GOODNESS OF FIT STATISTICS

Chi-Square with 119 degrees of freedom = 140.620 (P = 0.085724)
Chi-Square for independence model with 136 degrees of freedom = 344.855
Non-Normed Fit Index (NNFI; Tucker & Lewis) = 0.88
Comparative Fit Index (CFI) = 0.90
Goodness of Fit Index (GFI) = 0.96
Adjusted Goodness of Fit Index (AGFI) = 0.95
Goodness of Fit Index without diagonal values (GFI) = 0.80
Adjusted Goodness of Fit Index without diagonal values(AGFI) = 0.77

EIGENVALUES OF THE REDUCED CORRELATION MATRIX

Variable	Eigenvalue
1	1.922465671
2	0.411958679
3	0.339762283
4	0.275135770
5	0.219849161
6	0.131809107
7	0.120196257
8	0.038354479
9	0.019661957
10	-0.006997416
11	-0.058585682
12	-0.111916152
13	-0.138542184
14	-0.259678220
15	-0.273454643
16	-0.312707792
17	-0.394846118

UNROTATED LOADING MATRIX

Variable	F 1	Communality
Item 1	0.300	0.090
Item 2	0.279	0.078
Item 3	0.428	0.183
Item 4	0.307	0.094
Item 5	0.139	0.019
Item 6	0.348	0.121
Item 7	0.301	0.091
Item 8	0.430	0.185
Item 9	0.391	0.153
Item 10	0.271	0.073
Item 11	0.357	0.127
Item 12	0.347	0.121
Item 13	0.451	0.203
Item 14	0.410	0.168
Item 15	0.269	0.072
Item 16	0.246	0.060
Item 17	0.288	0.083

```
DISTRIBUTION OF RESIDUALS

Number of Residuals = 136
Summary Statistics for Fitted Residuals

Smallest Fitted Residual = -0.1519
  Median Fitted Residual = -0.0016
 Largest Fitted Residual =  0.1619
    Mean Fitted Residual = -0.0001
Variance Fitted Residual =  0.0032

              Root Mean Square of Residuals (RMSR) =  0.0567
Expected mean value of RMSR for an acceptable model =  0.0647 (Kelley's criterion)
(Kelley, 1935,page 146; see also Harman, 1962, page 21 of the 2nd edition)

Histogram for fitted residuals

 Value           Freq
                   |
   -0.1519         2  |  **
   -0.1206         3  |  ***
   -0.0892         8  |  **********
   -0.0578        18  |  ************************
   -0.0264        28  |  ************************************
    0.0050        31  |  ****************************************
    0.0363        26  |  **********************************
    0.0677        11  |  ***************
    0.0991         6  |  *******
    0.1305         2  |  **
    0.1619         1  |  *
                      +-----------+---------+---------+-----------+
                      0        7.8      15.5      23.3       31.0

Summary Statistics for Standardized Residuals

Smallest Standardized Residual = -2.35
  Median Standardized Residual = -0.02
 Largest Standardized Residual =  2.50
    Mean Standardized Residual = -0.00

Stemleaf Plot for Standardized Residuals

  -2 | 32
  -1 | 987554222221000
  -0 | 99998888887777666655555444444433333222222222111111
   0 | 000111111112222223333333334455555566777777788888888899
   1 | 0000222335666899
   2 | 5

Largest Positive Standardized Residuals
Residual for Var  17 and Var  13     2.50

------------------------------------------------------------------------------
 DESCRIPTIVES RELATED TO MISSING DATA
 Missing value code          :       999
 No missing data was observed in your data
------------------------------------------------------------------------------

Largest Positive Standardized Residuals
Residual for Var  17 and Var  13     2.50

------------------------------------------------------------------------------
 DESCRIPTIVES RELATED TO MISSING DATA
 Missing value code          :       999
 No missing data was observed in your data
------------------------------------------------------------------------------
```

References

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. Multivariate Behavioral Research, 27(4), 509-540.

Harman, H. H. (1962). Modern Factor Analysis, 2nd Edition. University of Chicago Press, Chicago.

Kelley, T. L. (1935). Essential Traits of Mental Life, Harvard Studies in Education, vol. 26. Harvard University Press, Cambridge.

Lorenzo-Seva, U., & Van Ginkel, J. R. (2015). Multiple Imputation of missing values in exploratory factor analysis of multidimensional scales : estimating latent trait scores.Anales, in press.

McDonald, R.P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum.

Mardia, K. V. (1970), Measures of multivariate skewnees and kurtosis with applications. Biometrika, 57, 519-530.

Mislevy, R.J., & Bock, R.D. (1990). BILOG 3 Item analysis and test scoring with binary logistic models. Mooresville: Scientific Software.

Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. Psychometrika, 46, 201-213.

Ten Berge, J.M.F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69, 613-625.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. Psychological Methods, 16, 209-220.

Woodhouse, B. & Jackson, P.H. (1977). Lower bounds to the reliability of the total score on a test composed of nonhomogeneous items: II. A search procedure to locate the greatest lower bound. Psychometrika, 42, 579-591.


```
 FACTOR is based on CLAPACK.
 Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling,
 S., McKenney, A., & Sorensen, D. (1999). LAPACK Users' Guide. Society for Industrial and Applied Mathematics. Philadelphia, PA

 FACTOR can be refered as:
 Lorenzo-Seva, U., & Ferrando, P.J. (2013). FACTOR 9.2 A Comprehensive Program for Fitting Exploratory and Semiconfirmatory
 Factor Analysis and IRT Models. Applied Psychological Measurement, 37(6), 497-498.
 Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model.
 Behavioral Research Methods, Instruments and Computers, 38(1), 88-91.

For furhter information and new releases go to:
psico.fcep.urv.cat/utilitats/factor

------------------------------------------------------------------------------
FACTOR completed
Computing time     : 0.06666667 minutes.
Matrices generated : 442597
```

E.6. Análisis de la relación entre la complejidad cognitiva y la dificultad empírica de los ítems.

Matriz de correlaciones entre componentes de complejidad cognitiva y dificultad empírica de los ítems

		Complejidad Cognitiva[b]		
	Dificultad[a]	Contenido	Habilidad	Texto
Dificultad	-			
Contenido	-0,63**	-		
Habilidad	-0,27	0,25	-	
Texto	-0,64**	0,62**	0,51*	-

Nota. Las correlaciones fueron calculadas usando el coeficiente de Pearson.

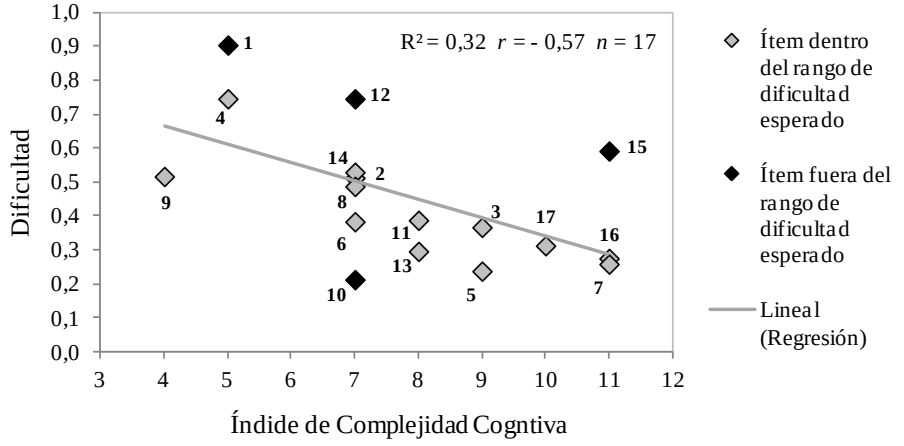
[a] La dificultad corresponde al porcentaje de respuestas correctas de cada ítem.

[b] La complejidad cognitiva del ítem se analiza desde los tres componentes ya mencionados: complejidad del contenido, de la habilidad cognitva y del texto.

** La correlación es significativa al nivel 0,01 (unilateral).

*La correlación es significativa al 0,05 (unilateral).

Regresión entre la complejidad cognitiva y la dificultad empírica de los ítems



Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,573[a]	,328	,284	,16944

a. Variables predictoras: (Constante), Complejidad

b. Variable dependiente: Dificultad

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,211	1	,211	7,336	,016[b]
	Residual	,431	15	,029		
	Total	,641	16			

a. Variable dependiente: Dificultad

b. Variables predictoras: (Constante), Complejidad

Diagnósticos por caso[a]

Número de casos	Residuo típ.	Dificultad	Valor pronosticado	Residual
1	1,712	,90	,6099	,29013
10	-1,724	,21	,5020	-,29205
12	1,463	,75	,5020	,24795
15	1,792	,59	,2864	,30360

a. Variable dependiente: Dificultad

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	,879	,161		5,461	,000	,536	1,223
	Complejidad	-,054	,020	-,573	-2,709	,016	-,096	-,011

a. Variable dependiente: Dificultad

E.7. Análisis de diferencias en porcentajes de logro y puntajes totales.

Test de diferencia de medias de porcentaje de logro, por categorías de habilidad cognitiva

pqlpl : Porcentaje de logro en los ítems de comprensión cuantitativa.
pqrpl : Porcentaje de logro en los ítems de razonamiento cuantitativo.

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	pqlpl	,4951	240	,20074	,01296
	pqrpl	,3861	240	,23321	,01505

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	pqlpl y pqrpl	240	,473	,000

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	pqlpl - pqrpl	,10896	,22453	,01449	,08041	,13752	7,518	239	,000

Test de diferencia de medias de porcentaje de logro, por categoría de eje temático.

p123pl : Porcentaje de logro en los ítems de la categoría Números, Álgebra y Geometría.
p4pl : Porcentaje de logro en los ítems de la categoría Datos y Azar.

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	p123pl	,4943	240	,21123	,01364
	p4pl	,4231	240	,21814	,01408

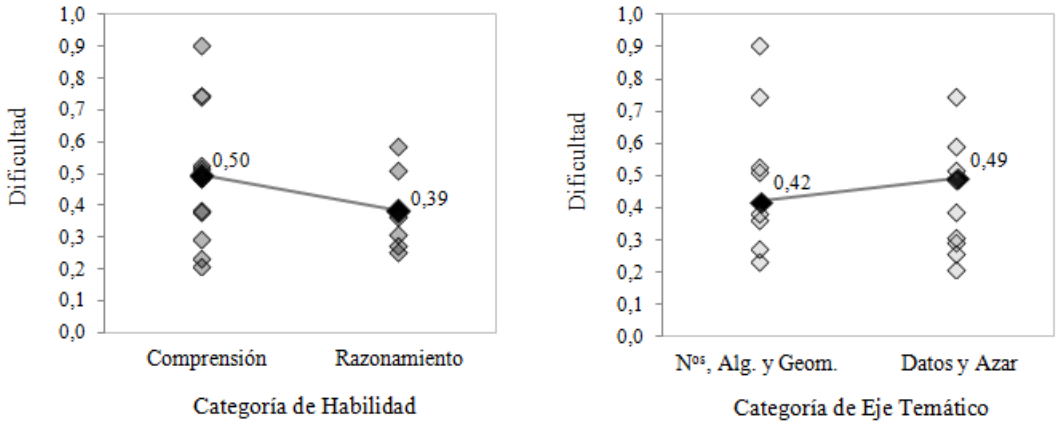
Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	p123pl y p4pl	240	,459	,000

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	p123pl - p4pl	,07112	,22341	,01442	,04271	,09953	4,932	239	,000

Dificultad promedio de los ítems, según categorías de habilidad y eje temático.



Test de diferencia de medias de puntaje total, por cuadernillo de prueba aplicado (Tipos D y E).

D : Puntaje total obtenido en el cuadernillo forma D.
E : Puntaje total obtenido en el cuadernillo forma E.

Estadísticos de grupo

	forma	N	Media	Desviación típ.	Error típ. de la media
ptotal	D	128	7,95	3,053	,270
	E	112	7,54	3,202	,303

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
ptotal	Se han asumido varianzas iguales	1,222	,270	1,011	238	,313	,408	,404	-,388	1,205
	No se han asumido varianzas iguales			1,008	230,414	,315	,408	,405	-,390	1,207

Estimación de confiabilidad, según cuadernillo aplicado (Tipos D y E).

Estadísticos de fiabilidad[a]

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,663	,664	17

a. forma = D

Estadísticos de fiabilidad[a]

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,684	,681	17

a. forma = E

Secuenciación de los ítems en cada cuadernillo.

El cuadernillo E corresponde al cuadernillo D alternando los ítems 2 y 3, 5 y 7, 12 y 13, y 15 y 16.

	Número del ítem en el banco de ítems															
Cuadernillo D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Cuadernillo E	1	3	2	4	7	6	5	8	9	10	11	13	12	14	16	15	17

E.8. Matriz de correlaciones y error estándar de medición, antes y después de eliminar obs. influyentes.

	x			*y*			*z*
	r	*EEM*	*n*	*r*	*EEM*	*n*	
1. Puntaje en prueba de Habilidades Cuantitativas (***x***)		-					
2. Promedio autoreportado de ensayos PSU de Matemática (***y***)	0,38 0,48	95,3 79,1	154 150[a]		-		
3. Promedio autoreportado de notas en Matemática (***z***)	0,39 0,42	0,98 0,97	210 208[b]	0,50 0,69	89,2 62,0	150 142[c]	-

Nota. *r* corresponde a la correlación de Pearson, *n* al número de casos y *EEM* al error estándar de estimación. En cada celda, la primera fila corresponde a los resultados iniciales y la segunda fila a los resultados sin casos atípicos (a más de tres desviaciones estándar de la recta de regresión). Todas las correlaciones son significativas al nivel 0,01 (unilateral).

Se excluyeron los casos: [a] 9, 215, 238 y 75; [b] 80 y 211 y [c] 9, 38, 101, 162, 205, 215 y 238.

E.9. Resultados de la regresión entre *x* e *y*.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,379[a]	,144	,138	95,580

a. Variables predictoras: (Constante), ptotal

b. Variable dependiente: puntajes

Diagnósticos por caso[a]

Número de casos	Residuo típ.	puntajes	Valor pronosticado	Residual
9	3,373	850	527,64	322,359
215	3,460	820	489,33	330,669
238	-4,218	150	553,18	-403,181

a. Variable dependiente: puntajes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,466[a]	,217	,212	82,331

a. Variables predictoras: (Constante), ptotal

b. Variable dependiente: puntajes

Diagnósticos por caso[a]

Número de casos	Residuo típ.	puntajes	Valor pronosticado	Residual
75	3,355	802	525,76	276,237

a. Variable dependiente: puntajes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,482[a]	,232	,227	79,405

a. Variables predictoras: (Constante), ptotal

b. Variable dependiente: puntajes

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	281924,210	1	281924,210	44,714	,000[b]
	Residual	933156,350	148	6305,110		
	Total	1215080,560	149			

a. Variable dependiente: puntajes

b. Variables predictoras: (Constante), ptotal

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados			Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	411,093	18,417		22,322	,000	374,699	447,487
	ptotal	14,103	2,109	,482	6,687	,000	9,935	18,271

a. Variable dependiente: puntajes

E.10. Resultados y diagrama de dispersión para la regresión entre *x* y *z*.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,409[a]	,167	,163	,9778

a. Variables predictoras: (Constante), ptotal

b. Variable dependiente: notas

Diagnósticos por caso[a]

Número de casos	Residuo típ.	notas	Valor pronosticado	Residual
80	-3,478	1,7	5,101	-3,4012
211	-3,008	2,3	5,241	-2,9414

a. Variable dependiente: notas

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,416[a]	,173	,169	,9307

a. Variables predictoras: (Constante), ptotal

b. Variable dependiente: notas

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	37,414	1	37,414	43,195	,000[b]
	Residual	178,430	206	,866		
	Total	215,844	207			

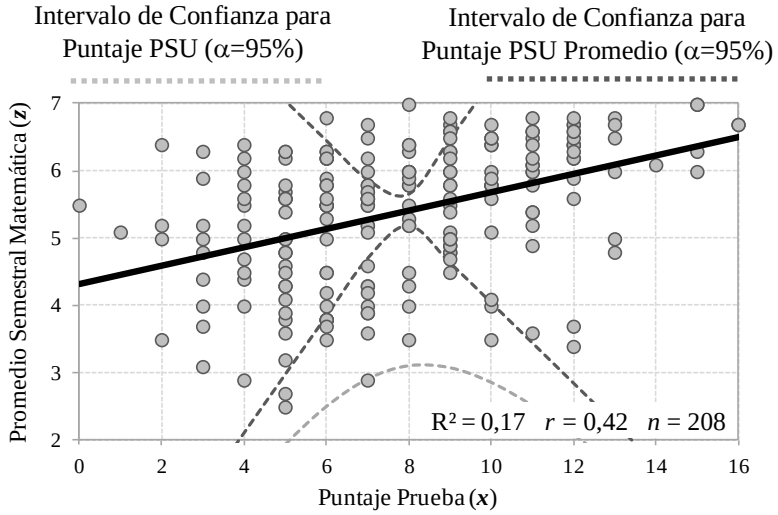
a. Variable dependiente: notas

b. Variables predictoras: (Constante), ptotal

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	4,324	,174		24,860	,000	3,981	4,667
	ptotal	,136	,021	,416	6,572	,000	,095	,177

a. Variable dependiente: notas



E.11. Resultados y diagrama de dispersión para la regresión entre *z* e *y*.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,515[a]	,265	,260	89,024

a. Variables predictoras: (Constante), notas

b. Variable dependiente: puntajes

Diagnósticos por caso[a]

Número de casos	Residuo típ.	puntajes	Valor pronosticado	Residual
9	3,393	850	547,97	302,032
38	-3,164	300	581,67	-281,669
215	3,687	820	491,80	328,200
238	-4,344	150	536,73	-386,734

a. Variable dependiente: puntajes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,622[a]	,386	,382	71,864

a. Variables predictoras: (Constante), notas

b. Variable dependiente: puntajes

Diagnósticos por caso[a]

Número de casos	Residuo típ.	puntajes	Valor pronosticado	Residual
162	-3,303	300	537,40	-237,395
231	3,104	600	376,95	223,054

a. Variable dependiente: puntajes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,671[a]	,450	,446	66,855

a. Variables predictoras: (Constante), notas

b. Variable dependiente: puntajes

Diagnósticos por caso[a]

Número de casos	Residuo típ.	puntajes	Valor pronosticado	Residual
101	-3,176	300	512,31	-212,313
205	-3,176	300	512,31	-212,313

a. Variable dependiente: puntajes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,693[a]	,481	,477	62,286

a. Variables predictoras: (Constante), notas

b. Variable dependiente: puntajes

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	502573,580	1	502573,580	129,546	,000[b]
	Residual	543129,857	140	3879,499		
	Total	1045703,437	141			

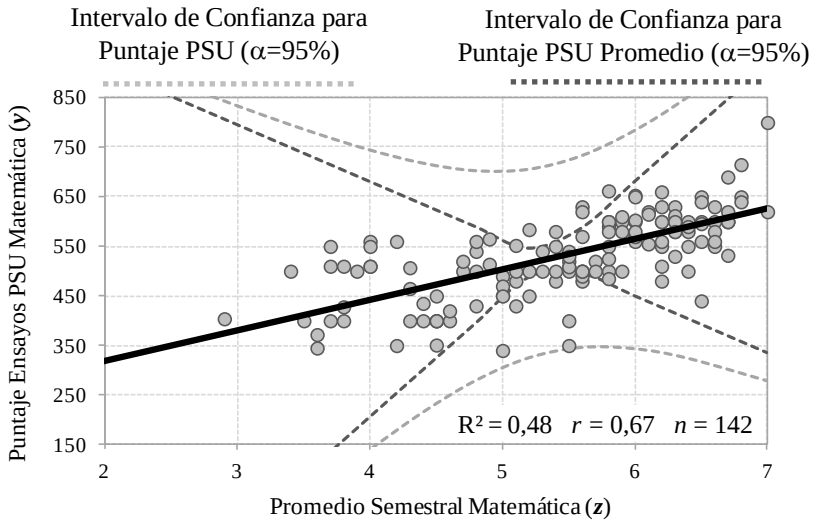
a. Variable dependiente: puntajes

b. Variables predictoras: (Constante), notas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	186,906	30,982		6,033	,000	125,654	248,158
	notas	63,211	5,554	,693	11,382	,000	52,231	74,191

a. Variable dependiente: puntajes



E.12. Resultados de la regresión múltiple entre *x* y *z* con *y*.

Con observaciones influyentes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,549[a]	,301	,292	87,100

a. Variables predictoras: (Constante), ptotal, notas

b. Variable dependiente: puntajes

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	480754,290	2	240377,145	31,685	,000[b]
	Residual	1115196,003	147	7586,367		
	Total	1595950,293	149			

a. Variable dependiente: puntajes

b. Variables predictoras: (Constante), ptotal, notas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	211,907	42,010		5,044	,000	128,886	294,929
	notas	47,700	8,124	,437	5,872	,000	31,645	63,755
	ptotal	6,966	2,525	,205	2,759	,007	1,976	11,956

a. Variable dependiente: puntajes

Sin observaciones influyentes

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,719[a]	,517	,510	58,374

a. Variables predictoras: (Constante), ptotal, notas

b. Variable dependiente: puntajes

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	503395,001	2	251697,501	73,866	,000[b]
	Residual	470233,311	138	3407,488		
	Total	973628,312	140			

a. Variable dependiente: puntajes

b. Variables predictoras: (Constante), ptotal, notas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	188,517	29,347		6,424	,000	130,489	246,545
	notas	53,343	5,695	,601	9,367	,000	42,082	64,603
	ptotal	6,201	1,759	,226	3,525	,001	2,723	9,680

a. Variable dependiente: puntajes

E.13. Test de diferencia entre medias de puntaje, según nivel de enseñanza.

Estadísticos de grupo

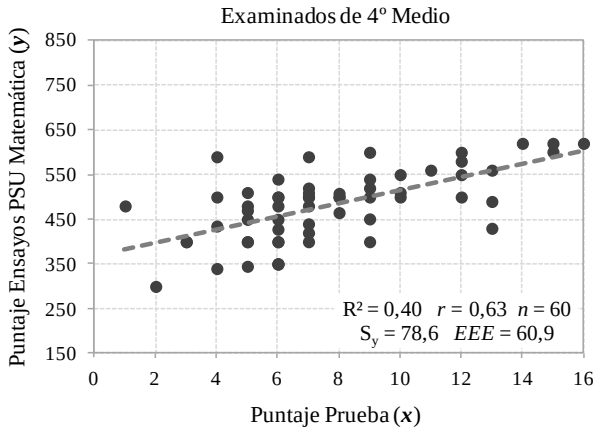
	nivel	N	Media	Desviación típ.	Error típ. de la media
ptotal	3	172	7,82	3,051	,233
	4	68	7,62	3,319	,402

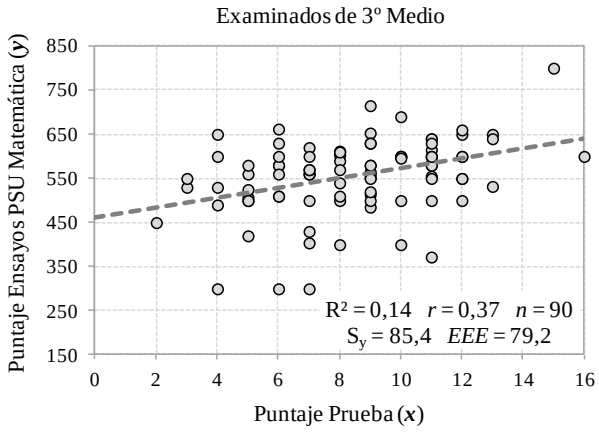
Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
ptotal	Se han asumido varianzas iguales	,388	,534	,451	238	,652	,202	,448	-,681	1,085
	No se han asumido varianzas iguales			,435	114,238	,665	,202	,465	-,719	1,123

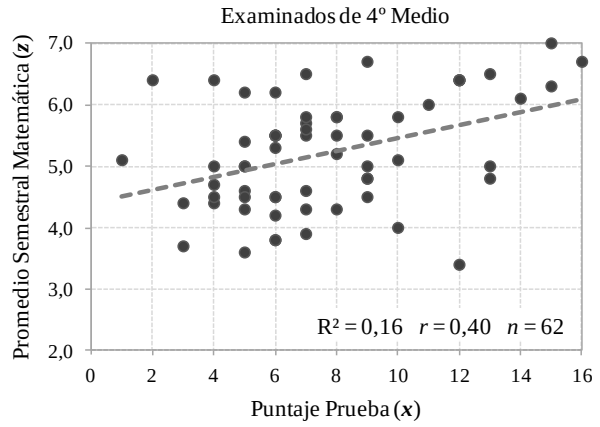
E.14. Diagramas de dispersión y ajuste de las regresiones, según nivel de enseñanza.

*Diagrama de dispersión y ajuste de la regresión entre **x** e **y***





*Diagramas de dispersión y ajuste de la regresión entre **x** y **z***



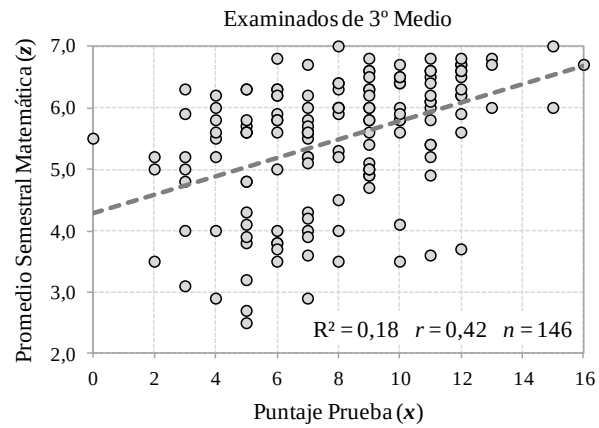
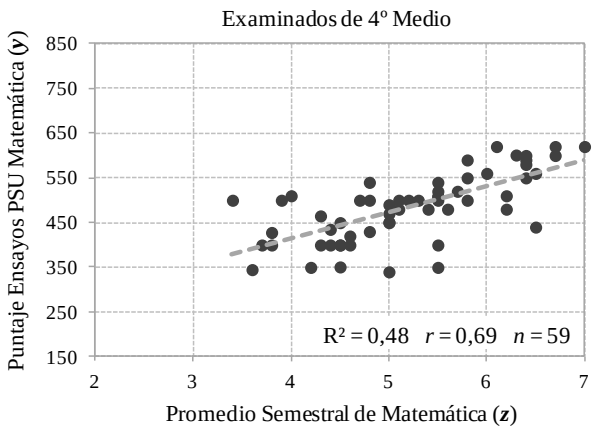
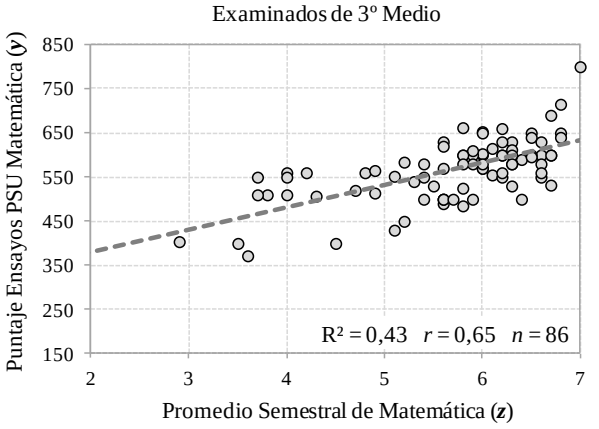


*Diagrama de dispersión y ajuste de la regresión entre **z** e **y***





E.15. Comparación entre las especificaciones de la prueba y la PSU de Matemática.

	Prueba de Habilidades Cuantitativas	PSU de Matemática
Cantidad de Ítems	17 ítems.	80 Ítems (Considerando los 5 ítems de experimentación).
Tipo de Ítems	De selección de respuesta, con opción múltiple simple.	De selección de respuesta, con opción múltiple simple y compuesta (68 ítems). De selección de respuesta, con suficiencia de datos (7 Ítems).
	Cuatro opciones, una clave y tres distractores.	Cinco opciones, una clave y cuatro distractores.
Puntuación de los Ítems	Dicotómica (desde la teoría clásica de tests), 1 punto por respuesta correcta y 0 por respuesta incorrecta u omitida.	Dicotómica, 1 punto por respuesta correcta y 0 por respuesta incorrecta u omitida.
Puntuación de la Prueba	Puntuación natural equivalente a la suma de los puntajes obtenidos en cada ítem.	Puntuación natural equivalente a la suma de los puntajes obtenidos en cada ítem.
Tiempos de Respuesta	35 minutos como máximo. (2 minutos promedio por ítem).	160 minutos como máximo. (2 minutos promedio por ítem).

Nota. Las especificaciones de la PSU de Matemática se han extraído de las publicaciones "Modelo de prueba de Matemática" y "Significado de los puntajes y claves del modelo de prueba de Matemática" correspondientes al proceso de admisión 2017,

Prueba de Habilidades Cuantitativas

	Ejes Temáticos				
Categorías de Procesos Cognitivos	Números	Álgebra	Geometría	Datos y Azar	
Comprensión					**65%**
Razonamiento					**35%**
	18%	**23%**	**6%**	**53%**	

PSU de Matemática

	Ejes Temáticos				
Categorías de Procesos Cognitivos	Números	Álgebra	Geometría	Datos y Azar	
Comprender-Aplicar					**60%-70%**
Analizar-Sintetizar-Evaluar					**30%-40%**
	21%	**24%**	**27%**	**28%**	

Nota. Las tabla de especificaciones de la PSU de Matemática se ha extraído de la publicación "Temario de la prueba de Matemática" correspondiente al proceso de admisión 2017, publicado en www.demre.cl.

E.16. Diagramas para los argumentos de validez.

Diagrama de argumento de validez para interpretación y uso descriptivo

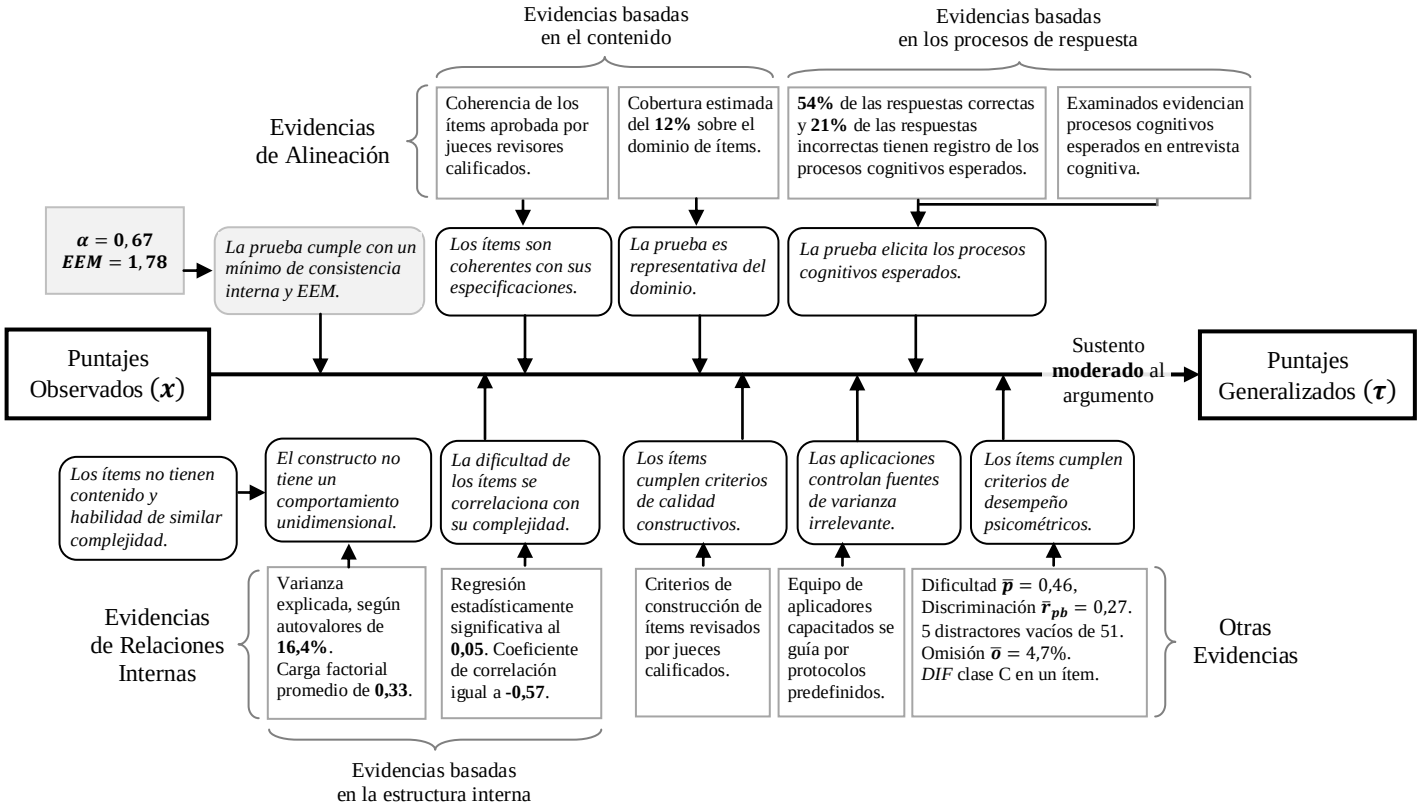
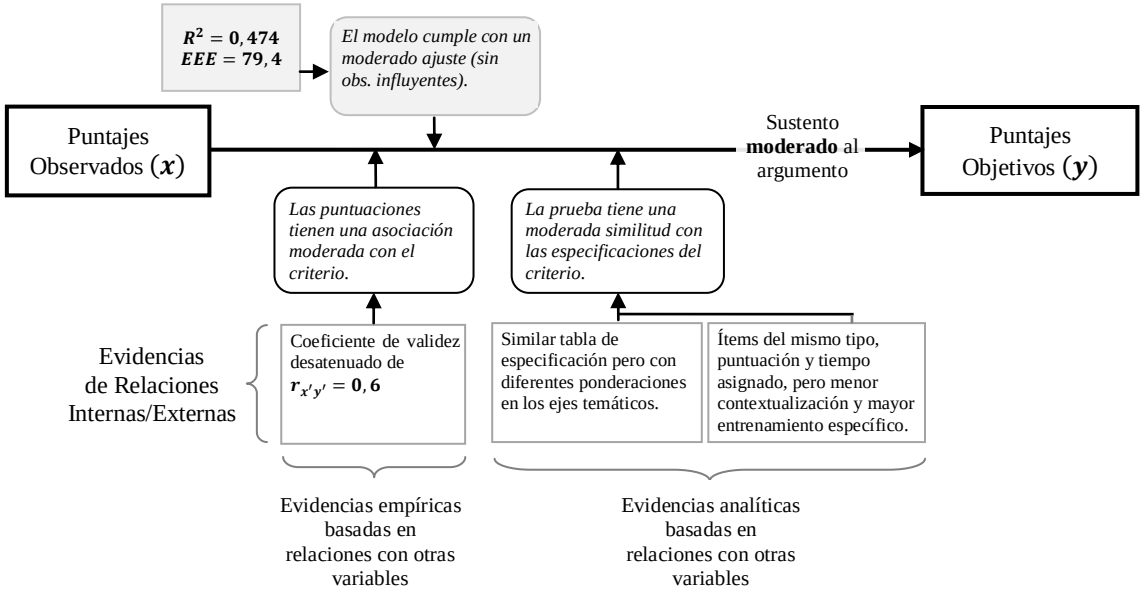


Diagrama de argumento de validez para interpretación y uso predictivo



Nota. Estos diagramas combinan elementos de un diagrama de argumentación (Toulmin, 1958) con componentes de un argumento interpretativo en un proceso de medición (Kane, 2006, p.33).