PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE EDUCACIÓN

ANÁLISIS DE EVIDENCIAS DE VALIDEZ DE LOS PUNTAJES DE UN INSTRUMENTO PARA EVALUAR ALFABETIZACIÓN MATEMÁTICA EN ESTUDIANTES CON TALENTO ACADÉMICO

POR

STEFAN CONSTANTINO MOSJOS AGUILAR

Proyecto de Magíster de la Facultad de Educación
de la Pontificia Universidad Católica de Chile Para optar al grado de
Magíster en Educación Mención Evaluación de Aprendizajes

Profesora guía:

VERÓNICA SANTELICES ETCHEGARAY

MARZO, 2022

SANTIAGO DE CHILE

TABLA DE CONTENIDO

AGRADECIMIENTOS

A mi madre, por haber estado siempre.

A la profesora Verónica Santelices, por su apoyo, confianza y paciencia.

A los profesores Andrea Valenzuela y David Torres, por sus valiosísimos aportes a este trabajo.

Al equipo de PENTA UC y LIES, por su generosidad al permitirme trabajar en este proyecto.

A ANID /Magíster en Chile para Profesionales de la Educación/ 2021- 50200097, por el financiamiento de este magíster.

RESUMEN

El objetivo del presente proyecto de magíster es analizar las evidencias de validez de las puntuaciones asociadas a un instrumento diseñado por el Laboratorio Interdisciplinario de Estadística Social de la Pontificia Universidad Católica de Chile (LIES) para PENTA UC con la finalidad de evaluar el nivel base de alfabetización matemática en estudiantes con talento académico. El foco de este estudio está en los análisis psicométricos de la teoría clásica del test y la teoría de respuesta al ítem. Se consideró una muestra de 207 estudiantes a quienes se les aplicó un instrumento de 32 ítems.

Los resultados muestran que el instrumento presenta una dificultad adecuada para evaluar el progreso de estudiantes con talento académico, sus puntuaciones son confiables y no presenta problemas de imparcialidad por género o por edad. De acuerdo con los resultados del Modelo de Crédito Parcial, hay un buen ajuste entre los datos y el modelo teórico de medición (*outfit* e *infit*). Si bien la dependencia de los ítems en el instrumento (por anidamiento debido a tener estímulos en común) puede llevar a cuestionar el uso de modelos como el Modelo de Crédito Parcial, la evidencia muestra que los parámetros de dificultad y habilidad proporcionados por dicho modelo no son muy diferentes de los que se obtienen al usar modelos que no consideran la independencia local como supuesto (*bifactor model*, *Rasch testlet model*).

Palabras clave: Talento académico – Alfabetización matemática – Modelo de Crédito Parcial – *bifactor model* – *Rasch testlet model*

ABSTRACT

This study aims to analyze the validity evidence of the scores associated to an instrument designed by Laboratorio Interdisciplinario de Estadística Social de la Pontificia Universidad Católica de Chile (LIES) for PENTA UC to assess the base level of mathematical literacy in gifted students. The focus of this study is on psychometric analysis of the classical test theory and item response theory. A sample of 207 students was considered to whom a 32-item instrument was applied.

The results show that the instrument presents an adequate difficulty to assess the progress of gifted students, its scores are reliable, and it does not present impartiality problems by gender or age. Although the dependence of the items in the instrument (due to nesting for having common stimuli) may lead to questioning the use of models such as the Partial Credit Model, the evidence shows that the parameters of difficulty and ability provided by that model are not very different from those obtained when using models that do not consider local independence as an assumption (bifactor model, Rasch testlet model).

Keywords: Gifted students – Mathematical literacy – Partial Credit Model – bifactor model – Rasch testlet model

1. INTRODUCCIÓN

Los estudiantes con talento académico han sido considerados como una reserva de riqueza, dado que encierran un gran potencial de capital humano para el desarrollo de la sociedad y sus desafíos avanzados (Arancibia, 2009). Lo anterior explica por qué los sistemas educacionales de los países desarrollados brindan especial atención a los estudiantes con talento académico (Tannenbaum, 2000).

En el contexto de prestar atención a la diversidad y su riqueza, durante las últimas décadas una gran cantidad de países ha creado programas para detectar y potenciar el talento académico. En Chile, múltiples iniciativas ligadas a las universidades han sacado adelante programas de ese tipo en diversas regiones del país (Benavides, Ríos y Marshall, 2004). Una de ellas es el Programa de Estudios y Desarrollo de Talentos de la Pontificia Universidad Católica de Chile (PENTA UC), que desde 2001 desarrolla un programa curricular que busca enriquecer la experiencia de estudiantes con talento académico.

Es importante señalar que, a pesar de la expansión de los programas para atender a estudiantes con talento académico, se ha reportado inequidad en el impacto de tales programas e incluso ausencia de efectos (Redding y Grissom, 2021; Card y Giuliano, 2016, Adelson et al., 2012). Esto último llama la atención a evaluar el efecto que los programas de talento académico tienen para garantizar que efectivamente estén potenciando el desarrollo de todos sus estudiantes.

En línea con lo anterior, PENTA UC se encuentra desarrollando una serie de instrumentos de evaluación diagnóstica que permitan caracterizar las habilidades de los estudiantes con talento académico en relación con el marco curricular del programa. En ese

contexto, surge la necesidad de evaluar las habilidades matemáticas, científicas y de lectura, midiendo la línea de base de los alumnos de sexto básico que ingresan al programa.

El presente estudio tiene por objetivo analizar evidencias de validez asociadas al instrumento diseñado para la evaluación en matemática. Para ello, se considera un análisis desde la teoría clásica del test y desde la teoría de respuesta al ítem. A través de esto se busca contribuir en dar una respuesta más pertinente a los estudiantes con talento académico que asisten a PENTA UC, entendiendo la relevancia social de atender a la diversidad y el enorme potencial que encierran estos estudiantes para aportar en el desarrollo de la sociedad.

En los siguientes dos apartados se explica la problematización que motiva este proyecto de magíster, su relevancia y el objetivo de este estudio. En el apartado 4 se presenta un breve marco teórico del constructo a la base del instrumento diseñado por LIES. En el apartado 5 se describe la metodología abordada desde las etapas propuestas por Wilson (2004) en el modelamiento de constructos. En el apartado 6 se presentan los resultados de los modelos psicométricos usados en el análisis del instrumento, para concluir, en los últimos dos apartados con algunas conclusiones, discusiones y recomendaciones para futuras aplicaciones del instrumento.

2. ANTECEDENTES Y PROBLEMATIZACIÓN

A continuación, se presentan algunos antecedentes relevantes de este estudio y se describe la relevancia del problema. Para ello, se caracteriza a los estudiantes con talento académico y se entregan algunos antecedentes sobre el programa PENTA UC.

2.1 Estudiantes con talento académico

Si bien en la literatura se utilizan los conceptos de superdotación, altas capacidades y talento académico, entre otros, para hacer referencia a estudiantes que presentan altas capacidades cognitivas, no hay consenso en el uso y significado de tales rótulos (Kaufman y Sternberg, 2008).

La idea de talento académico ha estado profundamente vinculada a la concepción de inteligencia, con lo que tanto su conceptualización como su medición han evolucionado estrechamente (López et al., 2002). En el último siglo, la literatura sobre inteligencia ha avanzado desde los trabajos de Galton (1869), Spearman (1904a) y Terman (1925), que concebían la inteligencia como una entidad única y general, hasta las concepciones de talentos o inteligencias múltiples propuestas por Horn & Cattell (1966) y Gardner (1983). En las últimas décadas, las concepciones de inteligencia han reconocido tanto un talento general como específico, y se han incorporado otros factores como la creatividad, la motivación y factores ambientales. Esto último es clave, pues plantea una distinción entre la dotación (determinada genéticamente y catalizada por factores ambientales) y el talento como una expresión del desarrollo sistemático de tales capacidades. Así, la definición del concepto de talento académico ha ido variando desde una conceptualización asociada a un alto nivel de inteligencia general hacia un modelo que considera factores contextuales y socioemocionales (López et al., 2002).

Algunos de los modelos más conocidos en relación con el talento académico son los modelos de Renzulli (1978) y Gagné (1985). El modelo de los tres anillos de Renzulli (1978) describe el talento académico como la combinación de alta capacidad intelectual, motivación y creatividad. Por otra parte, el modelo diferenciador de dotación y talento de Gagné establece una diferencia entre talento y dotación (Gagné, 1985). Mientras la dotación se refiere a la posesión y uso de habilidades naturales no entrenadas, el talento se refiere a un dominio destacado de conocimientos, habilidades o destrezas desarrolladas sistemáticamente. Se ha definido el talento académico como una habilidad significativamente superior que posee una persona en relación con sus pares, en el ámbito académico (López et al., 2002). Dado que es un concepto relativo, es decir, se basa en la comparación con pares, convencionalmente se ha aceptado como referencia estimativa el 10% superior de la población (Gagné, 2004).

Los estudiantes con talento académico conforman un grupo heterogéneo, por lo que cualquier caracterización de ellos debe considerar que no todos poseen las mismas características de desarrollo (Gagné, 2013). En términos generales, destacan en el plano cognitivo por su capacidad e interés de aprender, y por el desarrollo de habilidades de pensamiento superior. Algunas habilidades descritas en la literatura son su memoria y conocimiento de base, útil para conectar el conocimiento nuevo con conocimientos previos, su velocidad en el procesamiento de la información, su flexibilidad en la resolución de problemas, su complejidad o habilidad para usar pensamientos superiores y abstractos, y su autorregulación, entre otras (Salas, 2012).

Los estudiantes con talento académico dan cuenta de necesidades educativas especiales, entre otras razones por su rapidez e interés por aprender (Arancibia, 2009; Blanco et al., 2004). En el ámbito emocional y socioafectivo, se han descrito asincronías del desarrollo y una alta intensidad y sensibilidad emocional (Piechowski, 1997). Esto puede afectarles en el plano social generando dificultades de adaptación con pares de su grupo etario. De esta manera, el talento académico implica necesidades afectivas y socioemocionales que deben ser tomadas en cuenta para promover el desarrollo integral de estos estudiantes.

2.2 Programa PENTA UC

Tanto en Estados Unidos como en la mayoría de los países europeos existen programas especiales para atender a estudiantes con talento académico (Mönks & Katzko, 2005). En la década de los sesenta hubo múltiples presiones sociales contrarias a este tipo de programas, por prejuicios que los asociaban a diferenciación meritocrática, y otras voces partidarias de igualitarismo democrático (Arancibia, 2009). Sin embargo, desde que se publicó el Informe Marland (1971), ha habido una tendencia creciente a reconocer la necesidad de una educación especial para estos niños y niñas.

En general, los principales enfoques para atender a estudiantes con talento académico han sido los programas de aceleración (ingreso temprano a cursos más avanzados) y los programas de enriquecimiento curricular (profundización o ampliación del currículum regular) (Benavides et al., 2004). En Chile se ha optado por estos últimos, en general asociados a programas impartidos por las universidades (Blanco et al., 2004).

El Programa Educacional para Niños con Talentos Académicos (PENTA UC) es un programa educativo pionero que se implementa desde 2001 y que tiene por objetivo

enriquecer la experiencia académica de los estudiantes con talento académico y brindar oportunidades educacionales complementarias que atiendan sus necesidades educativas particulares (López et al., 2002; Flanagan y Arancibia, 2005).

El programa está orientado a estudiantes provenientes de colegios de distintas dependencias que cursen estudios entre sexto básico y cuarto año de enseñanza media. El enfoque está puesto en la implementación de un currículum enriquecido para ciencias, matemáticas y lenguaje. El marco curricular de PENTA UC propone, como parte de su plan formativo, el desarrollo de las siguientes habilidades: metacognición, creatividad, pensamiento crítico, ciudadanía, habilidades para la participación digital, trabajo colaborativo, autonomía y comunicación.

2.3 Relevancia del problema

Uno de los principales desafíos del sistema escolar es la inclusión y la atención a la diversidad (Agencia de la Calidad de la Educación, 2017). La evaluación de aprendizajes, como constitutiva de la enseñanza, también ha sido interpelada para atender la diversidad de estudiantes en el aula. El modelo diferenciador de educación tuvo como foco a los estudiantes que eran considerados de "déficit", es decir, aquellos con dificultades de aprendizaje. Por ello, en general, los estudiantes con talento académico han sido objeto de poca atención en los sistemas educativos (Arancibia, 2009).

Si bien existe una tendencia a creer que invertir en estudiantes con talento académico es "dar a quienes más tienen", lo cierto que, más allá de sus capacidades académicas, estos estudiantes suelen presentar características socioemocionales que pueden hacerlos menos aventajados que sus compañeros (Algaba y Fernández, 2021).

Por otro lado, no brindar una atención adecuada a los estudiantes con talento académico, o la "pérdida" de tales talentos académicos no es problema individual, sino que profundamente social. En un sistema educativo segregado, los estudiantes con talento académico en condición de pobreza reciben menos oportunidades que sus pares de niveles socioeconómicos más altos. Una intervención adecuada puede afectar positivamente la motivación de los estudiantes con talento, disminuir sus posibilidades de deserción escolar, aumentar su autoestima, y disminuir la incidencia de problemas psicológicos y emocionales debido a la estigmatización y discriminación de la que son objeto. Todos estos beneficios reportan externalidades positivas a las comunidades educativas a las que estos estudiantes pertenecen y a la sociedad en su conjunto, particularmente, por su potencial para el desarrollo de capital humano avanzado (Arancibia, 2009).

De acuerdo con Blanco et al. (2004) uno de los aspectos fundamentales para dar una respuesta educativa pertinente a los estudiantes con talento académico es la evaluación de aprendizajes, en tanto "permita identificar sus necesidades educativas específicas y que sirva para la toma de decisiones sobre las adaptaciones del currículum y sobre los recursos y ayudas que hay que proporcionar a cada uno para optimizar el desarrollo de sus capacidades" (p. 50). La evaluación también ha de orientarse a identificar las barreras que limitan el aprendizaje del estudiante.

De esta manera, mediante el diseño de instrumentos de calidad que permitan evaluar el progreso de tales estudiantes, se pueden tomar mejores decisiones instruccionales que permitan una mayor pertinencia de la enseñanza. Por otro lado, es fundamental contar con

herramientas que permitan evaluar el impacto de los programas dirigidos a estudiantes con talento académico para garantizar que estos estén cumpliendo con sus objetivos curriculares.

3. OBJETIVO DEL PROYECTO DE MAGÍSTER

Objetivo general: Analizar evidencias de validez de un instrumento para la evaluación diagnóstica de la alfabetización matemática en estudiantes de sexto básico que ingresan al programa de talento académico de la Universidad Católica (PENTA UC).

Para abordar este objetivo general se implementarán las siguientes actividades:

- Analizar los resultados cuantitativos del pilotaje de un instrumento para la evaluación diagnóstica de la alfabetización matemática en estudiantes de sexto básico del programa PENTA UC, mediante varios modelos psicométricos.
- Proponer ajustes de mejora a la versión de pilotaje de un instrumento para la evaluación diagnóstica de la alfabetización matemática en estudiantes de sexto básico del programa PENTA UC.

4. MARCO TEÓRICO

El constructo evaluado por el instrumento elaborado es la alfabetización matemática. De acuerdo con el marco evaluativo PISA 2021 (OECD, 2018), esta competencia involucra el razonamiento matemático y la resolución de problemas. Si bien el foco del instrumento es la evaluación de la alfabetización matemática, se consideró en su diseño la evaluación de algunas de las habilidades del marco curricular de PENTA UC. En esta sección se describe el constructo, y las habilidades del marco curricular del PENTA UC.

4.1 Alfabetización matemática

El concepto de alfabetización matemática se ha usado, al menos informalmente, desde la década de los 40 en diversos contextos (Stacey & Turner, 2015). Otros conceptos como *numeracy* o *quantitative literacy* han sido usados de forma análoga para hacer referencia al uso de las matemáticas para hacer frente con confianza a las demandas de la vida cotidiana (Cockcroft, 1982; Jablonka, 2003). El uso del concepto de alfabetización científica se ha considerado precursor del concepto de alfabetización matemática. Dicho término denota la familiaridad de una persona común con la ciencia, la que le permita comprender el mundo en que vive y actuar de manera apropiada (Bybee, 1997; DeBoer, 2000).

Cualquier definición de alfabetización matemática debe considerar que no puede conceptualizarse exclusivamente en términos de conocimiento matemático, pues involucra la capacidad de un individuo común para usar y aplicar ese conocimiento, por lo que implica un carácter profundamente funcional (Stacey & Turner, 2015).

Si bien el concepto de alfabetización matemática (*mathematical literacy*) ha sido usado por diversos autores, ha sido asociado y relevado en su importancia por el marco

evaluativo de PISA. Este programa de la OECD evalúa la formación de estudiantes de 15 años en lectura, matemática y ciencias. Como parte de los antecedentes que dieron origen a su marco curricular en matemáticas, se consideraron los enfoques de matemática realista (Freudenthal, 2012), los trabajos De Lange (1999) y los trabajos sobre competencias de Niss y Jensen (2002).

El constructo evaluado por la prueba PISA es *mathematical literacy*, y en las últimas versiones de este marco curricular se ha traducido como alfabetización matemática (en las primeras versiones se tradujo como competencia matemática). De acuerdo con su último marco evaluativo (OECD, 2018) se define como se indica a continuación:

> La alfabetización matemática es la capacidad de un individuo de razonar matemáticamente y de formular, emplear e interpretar las matemáticas para resolver problemas en una amplia variedad de contextos de la vida real. Esto incluye conceptos, procedimientos, datos y herramientas para describir, explicar y predecir fenómenos. Ayuda a los individuos a conocer el papel que cumplen las matemáticas en el mundo y hacer los juicios y tomar las decisiones bien fundamentadas que necesitan los ciudadanos reflexivos, constructivos y comprometidos del siglo XXI. (p. 7).

La anterior definición apunta a una competencia que se puede tener en un menor o mayor grado, y que es deseable en todo ciudadano aun cuando este no se dedique a carreras científicas o matemáticas. De acuerdo con el marco evaluativo de PISA (OECD, 2018), la alfabetización matemática involucra el razonamiento matemático y la resolución de problemas, tal como se muestra en la Tabla 1.

Tabla 1. Constructo de acuerdo con el marco PISA (OECD, 2018).

Alfabetización matemática	Razonamiento matemático	
	Resolución de problemas matemáticos	Formular situaciones matemáticamente
		Emplear conceptos, hechos, procedimientos y razonamientos matemáticos
		Interpretar, aplicar y evaluar resultados matemáticos

El marco PISA 2021 es enfático en reconocer que la alfabetización matemática no solo se centra en el uso de las matemáticas para resolver problemas del mundo real, sino que también identifica el razonamiento matemático como un aspecto central de ella.

Tanto el razonamiento matemático como la resolución de problemas forman parte de los estándares de *National Council of Teachers of Mathematics* (NCTM). La NCTM considera, dentro de su propuesta curricular, estándares de contenidos y estándares de procesos que los estudiantes deberían aprender a conocer y a ser capaces de usar cuando avancen en su educación. Los cinco estándares de procesos, que representan modos destacados de adquirir y usar el conocimiento, son: resolución de problemas, razonamiento y demostración, comunicación, conexiones, y representación (Marín y Lupiáñez, 2005).

4.1.1 Razonamiento matemático

El concepto de razonamiento matemático se usa en la literatura atendiendo de forma implícita a procesos formales e informales asociados a las demostraciones matemáticas o a la argumentación de ideas matemáticas. En el marco evaluativo de la prueba internacional TIMSS (*Trends in International Mathematics and Science Study*) se considera el razonamiento como uno de los tres dominios cognitivos evaluados (los otros son el

conocimiento y la aplicación). Esta habilidad es definida por el marco de TIMSS (Mullis y Martin, 2017) como se señala a continuación:

> Razonar matemáticamente implica pensamiento lógico y sistemático. Incluye razonamiento intuitivo e inductivo basado en patrones y regularidades que se pueden utilizar para llegar a soluciones a problemas planteados en situaciones nuevas o desconocidas. Estos problemas pueden ser puramente matemáticos o pueden tener escenarios de la vida real. Ambos tipos de elementos implican la transferencia de conocimientos y habilidades a nuevas situaciones; y las interacciones entre las habilidades de razonamiento suelen ser una característica de tales elementos. (p. 24)

Por otro lado, el marco PISA 2021 (OECD, 2018) define el razonamiento matemático como se indica a continuación:

> El razonamiento matemático, tanto inductivo como deductivo, involucra la evaluación de situaciones, la selección de estrategias, extraer conclusiones lógicas, desarrollar y describir soluciones, y reconocer cómo las soluciones pueden ser aplicadas. Un estudiante razona matemáticamente cuando
>
> - Identifica, reconoce, organiza, conecta, representa.
> - Construye, abstrae, evalúa deduce, justifica, explica defiende.
> - Interpreta, realiza juicios, crítica, refuta, cualifica.
>
> (pp. 14, 15)

La habilidad de razonamiento matemático no se encuentra de manera explícita en el currículum nacional, sin embargo, se considera implícita en las habilidades de resolver problemas, modelar y, especialmente, argumentar y comunicar.

4.1.2 Resolución de problemas

Desde los trabajos de Polya (1945), la literatura en educación matemática ha dado cada vez un mayor énfasis a la resolución de problemas, ya sea entendida como contexto, como habilidad o como la capacidad de hacer matemática (Lester, 2003).

El marco PISA 2021 asocia la resolución de problemas al ciclo de modelamiento matemático que involucra la formulación, empleo e interpretación de las matemáticas. Los términos de *formular, emplear* e *interpretar* se utilizan para organizar el proceso matemático que describe lo que las personas hacen para conectar el contexto de un problema con las matemáticas y la resolución de problemas. Estas tres etapas han sido consideradas en PISA como parte de un ciclo de matematización o modelamiento matemático (PISA, 2013). Estas etapas involucran:

- Formular situaciones de manera matemática.
- Emplear conceptos, hechos, procedimientos y razonamientos matemáticos.
- Interpretar, aplicar y evaluar resultados matemáticos.

Es importante señalar que el currículum nacional considera la resolución de problemas como una de las cuatro habilidades centrales en matemática, junto con la de representar, modelar, y argumentar y comunicar. De manera que, en el contexto nacional, tanto SIMCE como la prueba de acceso al sistema universitario evalúan dicho constructo. DEMRE define la resolución de problemas como “la capacidad que tiene el postulante para

solucionar una situación problemática dada, contextualizada o no, rutinaria o no, sin que se le haya indicado necesariamente un procedimiento a seguir" (DEMRE, 2021), en concordancia con las Bases Curriculares vigentes.

4.2 Marco curricular de PENTA UC

El objetivo del programa PENTA UC es enriquecer la experiencia académica de los estudiantes con talento académico a través de un currículum complementario. Este currículum considera las siguientes habilidades: metacognición, creatividad, pensamiento crítico, ciudadanía, habilidades para la participación digital, trabajo colaborativo, autonomía y comunicación. Se presenta una descripción de estas habilidades en la Tabla 2.

Tabla 2. Marco curricular del programa PENTA UC (PENTA UC, 2021a).

Metacognición	Habilidad de reflexionar, comprender y regular los propios procesos cognitivos y de aprendizaje. Se distinguen dos componentes principales, el conocimiento metacognitivo (Declarativo, Procedural, Condicional) y el control metacognitivo (Schraw & Graham, 1997).
Creatividad	Capacidad de buscar y proponer múltiples y novedosas soluciones a un mismo problema, de crear cosas nuevas a partir de lo existente, entendiendo que lo creado tiene una aplicación práctica que funciona, o que responde apropiadamente a un problema (Daskolia, Dimos & Kampylis, 2012). Entre sus componentes se incluyen el pensamiento divergente, pensamiento convergente y, la tolerancia a la ambigüedad y voluntad de tomar riesgos.
Pensamiento crítico	Capacidad de procesar, analizar, evaluar críticamente y reelaborar la información que se recibe, de modo de disponer de una base de sustentación de las propias creencias y hacer juicios razonados sobre el valor o la validez de algo. Incluye un conjunto de habilidades para la conceptualización, aplicación, análisis, interpretación, síntesis y evaluación de la información, y realización de inferencias (Ennis, 1993) como son el pensamiento reflexivo, analítico y evaluativo.
Ciudadanía	Grado en que los jóvenes practican el respeto y la consideración hacia otro/as diversos, independientemente de su condición de género, etnia o NSE, están dispuesto/as a escuchar sus puntos de vista, a negociar las diferencias de manera pacífica, como también a interesarse por asuntos públicos, formarse una opinión informada, hacer oír su voz y participar en la vida de sus comunidades y de la sociedad en general. Entre otros componentes, incluye la responsabilidad social, la reflexión crítica sobre la realidad social y la participación activa (ciudadana o comunitaria).
Habilidades para la participación digital	Habilidades operacionales y cognitivas de orden superior que permiten acceder, comprender y utilizar efectivamente la información y el conocimiento, y participar efectivamente en entornos digitales (Van Laar et al., 2017).

Trabajo colaborativo	Habilidades para trabajar efectiva y responsablemente con otros construyendo conocimiento o resolviendo un problema a través del compromiso de los distintos miembros del grupo con un objetivo común. Incluye la flexibilidad y voluntad para ayudar y hacer los compromisos necesarios para lograr objetivos comunes, asumir responsabilidad compartida por el trabajo colaborativo y valorar las contribuciones individuales realizadas por cada miembro del equipo.
Autonomía	Comprende un conjunto de cogniciones, emociones y comportamientos iniciados y regulados por el sí mismo, que involucran ejercicio de voluntad y elección por parte del individuo. Es la capacidad de actuar agénticamente (Ryan, 1995). Incluye habilidades para establecer metas (académicas/personales), para formular planes y trabajar para lograr metas (a corto, mediano y largo plazo), manejando el tiempo y los recursos; para la toma de decisiones autónoma y expresión de preferencias y, la responsabilidad personal por las propias acciones y decisiones.
Comunicación	La habilidad para articular ideas y pensamientos de manera efectiva, transmitiéndolas de manera oral, escrita, verbal y no-verbal en variadas formas, para distintos propósitos según el contexto. Incluye las capacidades de Escucha efectiva y activa para descifrar significado (conocimiento, valor, actitudes e intenciones), usar la comunicación de manera flexible y estratégica de acuerdo con una variedad de propósitos (informar, enseñar, motivar y persuadir) y, utilizar múltiples medios y tecnologías para comunicar, siendo capaces de juzgar su efectividad y evaluar su impacto. Más allá de los dominios académicos, las dimensiones sociales y emocionales son centrales en la manifestación de los talentos, y en cómo estos se desarrollan en los estudiantes con altas capacidades. Para que niño/as y jóvenes usen sus capacidades intelectuales, motivacionales y creativas de manera que logren una productividad sobresaliente y/o que movilicen sus habilidades interpersonales practicando sus principios éticos y morales, los programas educativos deben promover las habilidades sociales y emocionales (pasión, sentido de propósito, autorregulación) necesarias para gestionar y optimizar mejor los frutos de sus habilidades. Los valores personales proporcionan inspiración para iniciativas creativas y novedosas relevantes orientadas al bien común o dirigidas al mejoramiento de la sociedad.

5. METODOLOGÍA

En esta sección se presentan los aspectos metodológicos asociados a la construcción del instrumento diseñado para la evaluación diagnóstica de alfabetización matemática, y a los procesos de validación de los usos e interpretaciones de sus puntuaciones. Este apartado metodológico se desarrolla considerando las cuatro etapas descritas por Wilson (2004) en relación con el modelamiento de constructos: definición del mapa de constructo, diseño de ítems, generación de espacios de respuesta y selección de un modelo de medición. En relación con la selección de un modelo de medición, se describe la teoría clásica del test (CTT) y la teoría de respuesta al ítem (IRT).

Por último, se presentan algunas ideas sobre la imparcialidad como criterio de validez de una medición, y se describe la población y muestra considerada en el estudio, junto con el plan ético diseñado para el resguardo de los participantes.

5.1 Definición de un mapa de constructo

Esta etapa considera la creación de un mapa teórico que sitúe a los examinados en distintas etapas del desarrollo del constructo. La creación de un mapa de constructo (concepto más preciso que el de *constructo*) considera cómo este se categoriza en desempeños progresivos que reflejan cómo los examinados manifiestan distintos niveles del constructo por medir, y qué conductas o respuestas se espera de ellos según tales niveles (Wilson, 2004).

En una etapa inicial, considerando y adhiriendo al marco PISA (OECD, 2018), se definió la *alfabetización matemática* como el constructo por evaluar. Como ya se mencionó, la alfabetización matemática involucra el razonamiento matemático y la resolución de problemas. Por otro lado, se consideraron cuatro habilidades del currículum de PENTA UC:

metacognición, pensamiento crítico, ciudadanía y comunicación. Estas no se asociaron a ítems específicos, sino que se abordaron de manera transversal en el instrumento.

En relación con los desempeños esperados, no se contó, a priori, con un mapa teórico. Este se elaboró ex post a partir de las evidencias recolectadas en las aplicaciones de pilotaje.

5.2 Diseño de ítems

En esta etapa se construyen los reactivos o ítems considerando la cobertura del constructo elegido. Esto implica el diseño de situaciones reales en las cuales el examinado pueda manifestar el constructo que está siendo evaluado (Wilson, 2004).

Si bien el constructo definido aborda habilidades o competencias, estas se enmarcaron en el área temática de incerteza y datos, en particular, en el uso de estadística en la toma de decisiones. Dada la edad de los participantes, se consideró el uso de frecuencias naturales de probabilidad (Gigerenzer, 1994). El contexto en el que se sitúan los ítems responde a situaciones científicas y cotidianas, un aspecto esencial de la alfabetización matemática. Como una forma de aproximarse a la habilidad de ciudadanía, uno de los contextos involucra análisis de encuestas de opinión.

El instrumento fue elaborado por un docente de matemáticas de PENTA UC, quien es experto en evaluación de estudiantes con talento académico. Este fue presentado a jueces expertos de PENTA UC y LIES (Laboratorio Interdisciplinario de Estadística Social) en una serie de reuniones a lo largo de las cuales los ítems fueron discutidos con la finalidad de adecuarlos tanto desde el vocabulario y la gramática usada, como desde su dificultad y formato de presentación. Se resguardaron, además, aspectos de validez de contenido, la cual Hogan (2004) define como la relación entre el contenido de una prueba y algún dominio bien

definido de conocimiento o conducta. Esto implica una relación o grado de ajuste entre el contenido de la prueba y el dominio pertinente.

El instrumento final quedó conformado por 32 preguntas con un tiempo estimado de respuesta de 90 minutos.

La primera parte de la prueba consiste en 8 preguntas de selección múltiple o respuestas cortas. La segunda parte está formada por 15 preguntas, de las cuales 11 son de respuestas cortas y 4 de respuestas abiertas extensas. La tercera parte consiste en 9 preguntas de respuestas extensas. En estas dos últimas partes se explicita que el estudiante argumente sus respuestas y explique sus razonamientos. El diseño de los ítems sigue una dificultad teórica creciente a lo largo del instrumento.

En todos los casos los ítems consideran contextos cercanos, cotidianos o de las ciencias, que buscan evocar en el estudiante interés por responder.

El instrumento se desarrolla a través de una plataforma virtual, de manera que el estudiante tiene la oportunidad de avanzar interactivamente en la resolución de los problemas planteados. Esto es fundamental en la segunda y tercera parte del instrumento, pues ambas están diseñadas para que a lo largo de las preguntas el estudiante vaya reflexionando sobre sus argumentos. Las preguntas siguen un hilo conductor que orienta al estudiante en el desafío de resolver los problemas asociados, entregando apoyo mediante preguntas orientadoras o ejemplos que apoyan la reflexión.

En términos generales, las características del instrumento se resumen en la tabla 3. En el Anexo 1 se presenta una tabla de especificaciones del instrumento.

Tabla 3. Caracterización del instrumento.

Parte	Número de ítems	Puntaje total	Temática	Tipo de preguntas
Parte 1	8	17	Análisis de gráficos de frecuencia. Aplicación del principio de Laplace. Combinatoria.	4 cerradas de selección múltiple 4 abiertas de respuesta corta
Parte 2	15	22	Análisis de diseños de encuestas. Introducción a tablas de contingencia.	11 abiertas de respuesta corta 4 abiertas de respuesta extensa
Parte 3	9	17	Análisis de tablas de contingencias en un contexto científico.	7 abiertas de respuesta extensa
Total	32	56		

Para indagar en las evidencias de validez asociadas a los procesos de respuesta, se realizaron *think aloud* o entrevistas cognitivas. Esta fuente de validez se relaciona con el análisis teórico y empírico de los procesos de respuesta de los examinados. El objetivo de las entrevistas cognitivas es verificar si los ítems desencadenan los procesos que suponen, y asegurarse de la claridad de las instrucciones y el formato de presentación del instrumento (Willis, 2015).

Las entrevistas fueron efectuadas a través de videoconferencias por un equipo de profesionales, incluyendo psicólogas y expertos disciplinarios. Estas fueron grabadas con la autorización de los participantes. Durante la entrevista se aplicó a cada estudiante sólo una o dos de las tres partes que componen la prueba.

La primera parte de la prueba (8 ítems de respuesta corta) fue aplicada a tres estudiantes mujeres de primero medio. Los tiempos de respuesta registrados fueron 14, 16 y 18 minutos. La segunda y tercera parte de la prueba (15 y 9 ítems, respectivamente) se aplicaron a tres estudiantes mujeres: una de sexto básico, una de séptimo básico y una de primero medio. Los tiempos de respuestas registrados fueron 90, 60 y 84 minutos, respectivamente.

Debido a las posibilidades de acceso, todas las entrevistas cognitivas se aplicaron a mujeres, por lo que los resultados deben ser leídos con esa consideración.

A partir de las entrevistas, en la primera parte de la prueba solo se evidenciaron problemas en dos preguntas, en una por un histograma cuya visualización en 3D dificultaba su interpretación, y en otra por un aspecto de redacción. Ninguna estudiante demoró más de 20 minutos en responder los 8 ítems de esta parte.

En la segunda y tercera parte de la prueba, se evidenciaron importantes problemas asociados a redacción y falta de claridad en las instrucciones. Las estudiantes señalaron no comprender qué tareas debían hacer en algunos ítems, y manifestaron obstáculos asociados a ciertas palabras cuyos significados no les parecen claros a la luz del contexto. Por ejemplo, en la prueba se presenta una encuesta que los estudiantes deben analizar, sin embargo, por confusión, entendían que debían contestarla. Otro problema importante, se presentó en una serie de ítems cuyo contexto era el análisis de resultados de PCR (test diagnóstico de covid-19). En este caso, se observó que las participantes respondieron a partir de sus experiencias personales o conocimientos previos en relación con el tema. Dada la varianza irrelevante implicada en esto último, en una segunda versión se cambió el contexto por un test de positividad de vitaminas.

5.3 Generación de espacios de respuesta

En esta etapa se recolectan evidencias en relación con los procesos de respuesta. Se consideran tanto las respuestas cerradas como las respuestas abiertas. De acuerdo con Wilson (2004), el objetivo de esta etapa es categorizar y asignar puntuación a las observaciones como indicadores del constructo.

El instrumento mejorado en base a los resultados de la entrevista cognitiva fue aplicado a un grupo de 258 estudiantes. Se seleccionaron 56 de ellos con el objetivo de utilizar sus respuestas para pilotear el funcionamiento de la rúbrica diseñada para asignar puntuación a las respuestas abiertas. Esta aplicación permitió ajustar los descriptores y criterios de la rúbrica en base a los desempeños observados. El objetivo de esta aplicación preliminar de la rúbrica fue analizar si los descriptores de las rúbricas eran consistentes con las respuestas de los examinados y si esta permite ser aplicada a distintos niveles de desempeño.

Una vez obtenida una versión final de la rúbrica, se inició la corrección de todas las pruebas (N=258). Estas fueron corregidas por tres correctores con conocimientos especializados en matemática. Durante su capacitación (dos jornadas), los correctores resolvieron individualmente la prueba y luego fueron capacitados por el equipo elaborador de la rúbrica para efectuar la corrección. Cada uno utilizó la rúbrica por separado para asignar puntuaciones a los desempeños del estudiante. La corrección implicaba asignar puntajes a las preguntas abiertas (ubicarla en un nivel desempeño), pero también asignar la respuesta en una variable nominal que la asocia a una categoría dentro de un mismo nivel. Así, cada respuesta está asignada a un nivel (de 0 a 2 o 4, dependiendo de la pregunta) y a un código (A, B, C, D, …) que permite agrupar respuestas similares.

Para medir la concordancia entre las puntuaciones de los correctores, se determinó el coeficiente kappa de Cohen (Cohen, 1960). Este mide la proporción de concordancias observadas sobre el total de observaciones, ajustando por azar. Este coeficiente fue mayor que 0,85 en cada ítem para los niveles de desempeño. Al calcularse el coeficiente por código,

se obtuvo un mínimo de 0,76 (ítem 4). Los coeficientes obtenidos pueden consultarse en el Anexo 2.

5.4 Selección de un modelo de medición

En esta etapa se relacionan las respuestas observadas (categorizadas y asociadas a un puntaje) con el constructo, haciendo uso de un modelo estadístico. El objetivo de esta etapa es contrastar los comportamientos de los examinados con el modelo teórico subyacente (Wilson, 2004).

Si bien, como parte de los procesos de validación se consideran evidencias basadas en el contenido de la prueba y las asociadas a los procesos de respuesta, el foco de los análisis presentados en este informe se centra en las evidencias de estructura interna. Esto considera un análisis del comportamiento de las puntuaciones y de cada ítem. Para estos análisis se consideró la teoría clásica del test y la teoría de respuesta al ítem.

En esta etapa, como parte del análisis psicométrico, se consideró el Modelo de Crédito Parcial (Masters, 1982; Masters y Wright, 1997) y el Modelo de Crédito Parcial Generalizado (Muraki, 1992). También se usaron el *bifactor model* (Reise, 2012) y el *Rasch testlet model* (Wang y Wilson, 2005), ambos multidimensionales.

5.4.1 Teoría clásica del test (CTT)

La teoría clásica del test es el modelo más antiguo y común en los análisis de las puntuaciones de una prueba. Fue iniciada a principios del siglo XX por los trabajos de Spearman (1904b), y se consolidó con los aportes de Gulliksen (1950), y Lord y Novik (1966).

La familiaridad y sencillez de esta teoría en psicometría se fundamenta en el uso de un modelo lineal (Muñiz, 2010). Es decir, asume que la puntuación empírica de una persona (X) es su puntuación verdadera (V) más un error de medición (e).

$$X = V + e \tag{1}$$

A partir de la ecuación anterior, y asumiendo que los errores son aleatorios e independientes de ambas puntuaciones, se establece que:

$$\mathrm{Var}(X) = \mathrm{Var}(V) + \mathrm{Var}(e) \tag{2}$$

A partir de esta última ecuación, considerando $\frac{\mathrm{Var}(X)}{\mathrm{Var}(V)}$ como la fiabilidad del test, se puede estimar el error como:

$$e = \sigma_X\sqrt{1 - \text{fiabilidad}} \tag{3}$$

Como parte de este enfoque clásico, se consideró al alfa de Cronbach (α) como estimador de la fiabilidad del test. Entre más cercano a 1 es esta medida, mayor consistencia presentan entre sí los ítems del test (Hogan, 2004). A partir de este se estimó el error estándar de la medida como: $e = \sigma\sqrt{1 - \alpha}$, considerando la desviación estándar σ. Esta medida permite estimar la puntuación verdadera a partir de la puntuación observada (Gempp, 2006).

Para los análisis de los resultados del instrumento desde la CTT, se consideró el puntaje total de cada estudiante como la suma simple del puntaje obtenido en cada ítem. Los datos perdidos se imputaron como puntaje cero, siempre que el examinado hubiera contestado más de la mitad de la prueba. Se analizó la media de las puntuaciones por personas, y se analizaron las diferencias entre grupos con ANOVA de un factor. Todos los análisis se realizaron con el *software IBM SPSS Statics 21*.

5.4.2 Teoría de respuesta al ítem (modelos unidimensionales)

La teoría de respuesta al ítem (IRT) agrupa varias líneas de investigación psicométricas iniciadas por Rasch (1960) y Birnbaum (1968). A diferencia de la CTT, en este enfoque se considera el ítem como unidad de análisis. Los modelos de IRT relacionan el comportamiento de un sujeto frente a un ítem con el rasgo latente responsable de ese comportamiento. Para ello se modeliza, con funciones matemáticas, la probabilidad de dar una determinada respuesta al ítem para cada nivel del rasgo medido.

A continuación, se describe el Modelo de Rasch para ítems dicotómicos (Rasch, 1960), el Modelo de Crédito Parcial (Masters, 1982) y el Modelo de Crédito Parcial Generalizado (Muraki, 1992), estos últimos para ítems politómicos. En este estudio se consideraron los dos últimos modelos, sin embargo, cuando el ítem es dicotómico, el Modelo de Crédito Parcial se comporta para dicho ítem como un modelo de Rasch, por lo que este también se describe a continuación.

A la base de los tres modelos mencionados en este apartado están los supuestos de unidimensionalidad e independencia local (Demars, 2010). De acuerdo con Demars (2010), la unidimensionalidad se refiere a que, a la base del modelo, una única variable latente es

suficiente para explicar la mayoría de las variaciones en las respuestas al ítem; la independencia local se refiere a que, controlando por variable latente, las respuestas a un ítem son independientes de las respuestas a otro ítem.

Modelo de Rasch

El modelo de Rasch es el más simple de una familia de modelos, y es el utilizado para el análisis de ítems dicotómicos, es decir, ítems cuyas respuestas solo admiten dos categorías (Wright & Mok, 2004).

Este modelo utiliza la suma de puntuaciones para calcular la localización de una persona (lo que es interpretado como su habilidad o logro) y la localización del ítem (interpretada como su dificultad) en una escala lineal que representa la variable latente (escala logit). La diferencia entre el logro de una persona y la dificultad del ítem se usa para calcular la probabilidad de haber respondido correctamente $(X = 1)$ o incorrectamente ese ítem $(X = 0)$, de acuerdo con el siguiente modelo:

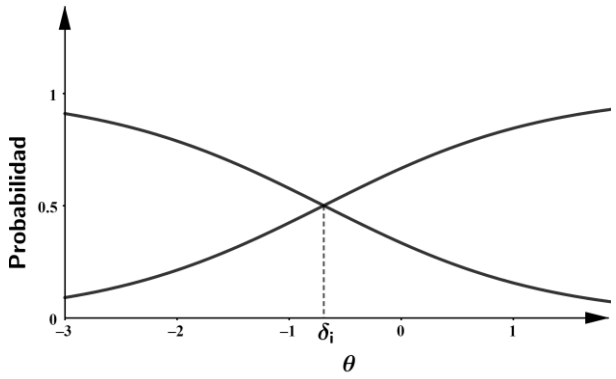
$$P_{ni}(X = 1) = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}} \quad (4)$$

$$P_{ni}(X = 0) = \frac{1}{1 + e^{\theta_n - \delta_i}} \quad (5)$$

En las ecuaciones (4) y (5) $P_{ni}(X = 1)$ es la probabilidad de que la persona n responda correctamente el ítem i, θ_n es la localización de la persona n y δ_i es la localización del ítem.

La Figura 1 muestra la curva característica para un ítem dicotómico. El punto de donde se intersecan ambas curvas corresponde al nivel de rasgo latente para el cual la probabilidad de contestar el ítem correctamente es igual a la de contestarlo incorrectamente, esto es: $P(X = 1) = P(X = 0) = 0{,}5$. Este valor se denomina localización del ítem y es considerado una estimación de la dificultad de este. Si un ítem tiene un δ_i elevado, se requerirá un nivel alto de rasgo latente para contestarlo correctamente.

Figura 1. Curvas características para un ítem dicotómico.



Modelo de Crédito Parcial

El Modelo de Crédito Parcial es una extensión del modelo de Rasch para ítems politómicos. Masters (1982) lo desarrolló aplicando el modelo dicotómico de Rasch a pares adyacentes de categorías de puntuación.

De acuerdo con Wu, Tam y Yen (2016), la función de probabilidad detrás de este modelo es:

$$P(X_{ni} = x) = \frac{e^{\sum_{k=0}^{x}(\theta_n - \delta_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h}(\theta_n - \delta_{ik})}} \quad (6)$$

En donde $P(X_{ni} = x)$ es la probabilidad de que la persona n obtenga una puntuación x en el ítem i, con m_i la puntuación máxima en ese ítem. En el caso de tres categorías, por ejemplo, se tiene que:

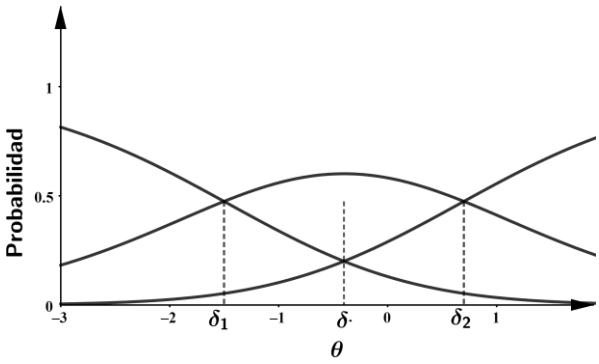
$$P(X = 0) = \frac{1}{1 + e^{(\theta_n - \delta_1)} + e^{(2\theta_n - (\delta_1 + \delta_2)}} \quad (7)$$

$$P(X = 1) = \frac{e^{(\theta_n - \delta_1)}}{1 + e^{(\theta_n - \delta_1)} + e^{(2\theta_n - (\delta_1 + \delta_2)}} \quad (8)$$

$$P(X = 2) = \frac{e^{(2\theta_n - (\delta_1 + \delta_2)}}{1 + e^{(\theta_n - \delta_1)} + e^{(2\theta_n - (\delta_1 + \delta_2)}} \quad (9)$$

En este caso, los valores δ_k (*threshold* o umbrales) representan el valor del rasgo latente para el cual la probabilidad de estar en la categoría i o $i - 1$ es igual, como se puede observar en los gráficos de las curvas características del ítem (Figura 2). Estos valores corresponden a lo que Masters (1982) denomina *step parameters*.

Figura 2. Curvas características para un ítem con tres categorías (niveles de desempeño).



El modelo de crédito parcial puede ser parametrizado de manera equivalente considerando los siguientes valores:

$$\delta_\cdot = \sum_{i=1}^{m} \frac{\delta_k}{m} \tag{10}$$

$$\tau_k = \delta_\cdot - \delta_k \tag{11}$$

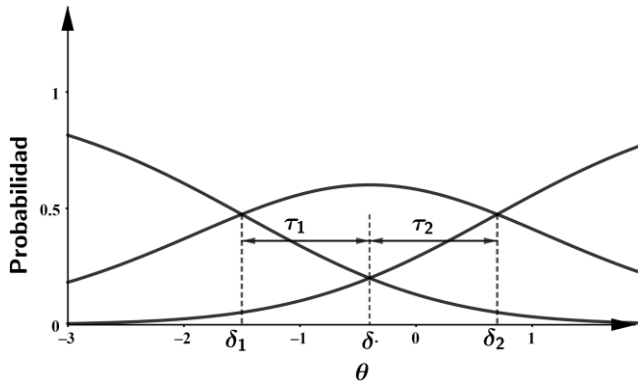
De acuerdo con Wu, Tam y Yen (2016), $\delta_\cdot$ (*delta dot*) corresponde a una especie de "dificultad promedio", por lo que es útil como un único indicador que permite una aproximación a la dificultad del ítem.

Por otro lado, τ representa la distancia entre una categoría de puntaje de crédito parcial y la dificultad del ítem "promedio". Este último permite parametrizar la ecuación 6 como:

$$P(X_{ni} = x) = \frac{e^{\sum_{k=0}^{x}(\theta_n - (\delta_\cdot - \tau_{ik}))}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h}(\theta_n - (\delta_\cdot - \tau_{ik}))}} \tag{12}$$

La representación gráfica de los parámetros δ. y τ puede observarse en la Figura 3.

Figura 3. Curvas características para un ítem con tres categorías (niveles de desempeño).



Dado que tanto en el modelo de Rasch dicotómico como en el Modelo de Crédito Parcial subyacen ciertos supuestos, se han desarrollado diversos estadísticos para verificar la bondad de ajuste del modelo, es decir, la relación entre los datos obtenidos y los datos esperados de acuerdo con el modelo. El estadístico basado en residuos, que se describe a continuación, considera las diferencias entre el puntaje del ítem observado y el puntaje esperado para formar residuos para cada ítem y persona.

Los estadísticos que se presentan a continuación, elaborados por Wright (1977) se basan en los siguientes residuos:

$$z_{ni} = \frac{x_{ni} - E(X_{ni})}{\sqrt{Var(X_{ni})}} \tag{13}$$

En donde x_{ni} es el valor observado para la persona n en el ítem i, y X_{ni} la variable aleatoria.

A partir del parámetro anterior, se pueden definir estadísticos para determinar un índice por persona o por ítem. Wright y Masters (1982) propusieron un estadístico ponderado y no ponderado que se definen a continuación:

$$\text{Unweighted mean square (outfit)} = \frac{\sum_n z_{ni}^2}{N} = \frac{1}{N}\sum_n \frac{\left(x_{ni} - E(X_{ni})\right)^2}{Var(X_{ni})} \quad (14)$$

$$\text{Weighted mean square (infit)} = \frac{\sum_n z_{ni}^2 \, Var(X_{ni})}{\sum_n Var(X_{ni})} = \frac{\sum_n \left(x_{ni} - E(X_{ni})\right)^2}{\sum_n Var(X_{ni})} \quad (15)$$

En ambos casos, se espera en general un valor cercano a 1 como indicador de un buen ajuste del ítem.

Modelo de Crédito Parcial Generalizado

El Modelo de Crédito Parcial Generalizado fue formulado por Muraki (1992) basado en los trabajos de Masters (1982). A diferencia del Modelo de Crédito Parcial, este permite un parámetro de discriminación variable entre los ítems, pero constante para todos los *threshold* de un mismo ítem (Muraki, 1992, Muraki y Muraki, 2016). La ecuación (16) representa este modelo, en esta los parámetros son los mismos que en la ecuación (6), con la diferencia de que se añade un *slope parameter* (a_i) o índice de discriminación.

$$P(X_{ni} = x) = \frac{e^{\sum_{k=0}^{x} a_i(\theta_n - \delta_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h} a_i(\theta_n - \delta_{ik})}} \quad (16)$$

Como índice de confiabilidad en el uso de estos modelos, Adams (2005) propone los siguientes:

$$\text{EAP reliability} = 1 - \frac{\sigma}{\sigma + v} \quad (17)$$

$$\text{WLE reliability} = 1 - \frac{\sigma}{v} \quad (18)$$

En ambas fórmulas σ denota el promedio de los errores cuadráticos y v varianza de las dificultades estimadas. El método de estimación usado en cada una es EAP (*estimated a posteriori*) y WLE (*weighted maximum likelihood estimation*), respectivamente.

Para efectuar los análisis se utilizaron las librerías TAM (Robitzsch et al., 2021) y eRm (Mair et al., 2021). Ambas difieren en sus parametrizaciones, mientras TAM considera una estimación de máxima verosimilitud marginal, eRm estima considerando una máxima verosimilitud condicional.

5.4.3 Teoría de respuesta al ítem (modelos multidimensionales)

Algunos modelos IRT consideran supuestos que no siempre se cumplen en el desarrollo de las pruebas. Uno de estos es el supuesto de independencia local, es decir, el hecho de que la probabilidad de responder correctamente un ítem es independiente de la probabilidad de responder otros, para un mismo nivel de habilidad (Geramipour, 2021).

Hay varias razones que implican transgresiones del supuesto de independencia local, una de ellas es cuando la prueba consta de ítems que se responden a partir de un mismo estímulo. Este anidamiento de ítems se denomina *testlet* (Wainer y Kiely, 1987; Wainer et al., 2007). Algunos ejemplos comunes de esto son las pruebas de idiomas, en donde se reconoce la influencia de efecto de método; o pruebas de comprensión lectora en las cuales varios ítems se responden a partir de un estímulo común (Rijmen, 2010). A continuación, se describen dos de estos modelos.

Bifactor model (BM)

El *bifactor model* es una estructura latente en la que los ítems se cargan en un factor general, por un lado, y en una estructura de factores específicos por el otro. El nombre de *bifactor* refiere a esa doble carga de cada ítem. El factor general representa la variable latente de interés central, mientras que los otros factores se incorporan para tener en cuenta las dependencias entre los ítems que pertenecen a un mismo grupo (Rijmen, 2010). Este modelo se interpreta como los otros métodos analíticos de factores, considerando que cada factor corresponde a un *testlet* (Reise, 2012).

Para el caso de un modelo logístico, la probabilidad de éxito en un ítem se puede expresar para ítems politómicos como:

$$P(x_i = x|\theta) = \frac{e^{\sum_{k=0}^{x} z_i}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h} z_i}} \tag{19}$$

Para N *testlets*, z_i se define como:

$$z_i = \alpha_{iGen}(\theta_{Gen}) + \alpha_{iDim1}(\theta_{Dim1}) + \alpha_{iDim2}(\theta_{Dim2}) + \cdots + \alpha_{iDimN}(\theta_{DimN}) + \gamma_i$$

En esta ecuación α_{iGen} es el parámetro de discriminación del ítem sobre el factor general, $\alpha_{iDim\,n}$ es el parámetro de discriminación en el factor específico, y γ_i es un parámetro de intercepto que se relaciona negativamente con la dificultad del ítem (a mayor γ_i, más fácil es el ítem). Los valores θ representan el rasgo latente del individuo.

Rasch testlet model (RTM)

Al igual que el *bifactor model*, el *Rasch testlet model* no considera como requisito el supuesto de independencia local. Este modelo fue introducido y desarrollado por Wang y Wilson (2005) tanto para ítems dicotómicos como politómicos. Este corresponde a un tipo de *bifactor model* en el cual los parámetros de discriminación (carga) están restringidos a ser iguales en cada *testlet* (Wang y Wilson, 2005, Jiao et al., 2013).

En este caso, las curvas se modelan mediante la siguiente ecuación:

$$P(x_i = x|\theta) = \frac{e^{\sum_{k=0}^{x} z_i}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h} z_i}} \quad (20)$$

Para N *testlets*, z_i se define como:

$$z_i = \alpha_{iGen}\left(\theta_{Gen} + c_1(\theta_{Dim1}) + c_2(\theta_{Dim2}) + \cdots + c_N(\theta_{DimN})\right) + \gamma_i$$

Los parámetros representan lo mismo que en el *bifactor model*, pero en este caso se considera una constante de proporcionalidad c_i. Es decir, la diferencia fundamental entre el *bifactor model* y el *Rasch testlet model* es que, en este último, se asume que las cargas a los

factores específicos son proporcionales a la carga al factor general. Con esto las cargas son restringidas a ser iguales en cada factor o *testlet* (Wang y Wilson, 2005).

Para ambos modelos se utilizó la librería TAM (Robitzsch et al., 2021).

5.5 Imparcialidad

La imparcialidad es un aspecto fundamental en el contexto de medición educativa, pues los resultados de tales mediciones suelen reflejar brechas asociadas a inequidades o diferencias en el acceso a oportunidades. De acuerdo con los Estándares para pruebas educativas y psicológicas (AERA, APA, NCME, 2014), la imparcialidad es una cuestión fundamental de validez que requiere atención en todas las etapas del desarrollo y uso de pruebas. Cuando examinados con iguales capacidades difieren en sus probabilidades de responder correctamente el ítem de una prueba, se habla de funcionamiento diferencial del ítem (DIF). Sin embargo, el DIF solo se considera sesgo cuando hay una causa que explica tales diferencias. Para hablar de sesgo debe haber un desempeño diferencial que no se explique por el nivel de habilidad del examinado, sino por su pertenencia a un grupo específico (Camilli, 2006).

Para los análisis de DIF solo se consideraron los ítems dicotómicos, debido a las restricciones de número de personas por categoría. Se consideraron como subgrupos el género (femenino y masculino), la condición de ser o no estudiante de PENTA UC, y el nivel educativo. Debido a restricciones del tamaño muestral no se consideró la nacionalidad como variable. Analizar las brechas de género es de interés dado que en Chile estas brechas se han evidenciado sistemáticamente en evaluaciones estandarizadas como SIMCE, TIMS y PISA (Agencia de la Calidad de la Educación 2019, 2020a, 2020b). Por otro lado, analizar la presencia de DIF según nivel educativo permite identificar si hay varianza irrelevante de

constructo asociada el avance de los estudiantes en el currículum nacional. Por último, dado que el grupo objetivo de la prueba son los estudiantes con talento académico, es fundamental estudiar si hay diferencias entre el comportamiento de este grupo y la población general.

Para estudiar la presencia de DIF se utilizó prueba de Wald, la cual es un test de hipótesis que examina si una variable independiente presenta diferencias estadísticamente significativas en relación con la variable dependiente (Glas y Verhelst, 1995). Así, para determinar si las diferencias entre subgrupos son significativas o no, se utilizó el siguiente test de hipótesis (Wright y Masters, 1982):

$$z = \frac{d_1 - d_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{21}$$

En donde d_1 y d_2 son las dificultades específicas de los ítems para los subgrupos 1 y 2, respectivamente; y σ_1^2 y σ_2^2 son las varianzas de las dificultades específicas de los ítems para cada grupo. Si $z < -2$ o $z > 2$, se considera que las diferencias son significativas. Es necesario una revisión del ítem para concluir si tales diferencias pueden atribuirse a sesgo o imparcialidad.

Para efectuar este análisis se utilizó la librería eRm (Mair et al., 2021).

5.6 Población y muestra

Dado que el uso del instrumento se enmarca en una investigación sobre el efecto del programa PENTA UC, los criterios de selección de la muestra se definen por el universo de estudiantes que postulan al programa.

La población de estudio corresponde a la totalidad de estudiantes que ingresa, en sexto básico, a PENTA UC en el área de matemática, pues el instrumento se aplicará como diagnóstico censal para caracterizar su perfil de ingreso. En este caso, las corporaciones municipales que se encuentran afiliadas al programa contribuyen con más del 80% de la matrícula de cada año. En este contexto, los colegios fueron invitados a participar de manera voluntaria, informando a sus estudiantes la posibilidad de participar en el proceso de análisis de los instrumentos del programa.

La muestra se seleccionó a partir de los estudiantes de estos establecimientos que se encuentran entre quinto y octavo básico, realizándose una invitación abierta y un muestreo por bola de nieve. La prueba fue rendida por 258 estudiantes, sin embargo, fueron excluidos del análisis 8 estudiantes que rindieron una aplicación especial, mediada por un profesor de equipo PENTA; y 5 estudiantes que informaron presentar alguna discapacidad cognitiva. Además, se excluyeron los casos en los que los estudiantes desertaron de rendir la prueba, siempre que hayan contestado menos del 50% (38 casos). De esta manera, el análisis consideró finalmente un $N = 208$, los cuales pertenecen a 4 colegios municipales de la Región Metropolitana. En la tabla 4 se muestra la distribución de los estudiantes por género y nivel.

Tabla 4. Caracterización de la muestra de estudiantes a quienes se aplicó el instrumento por género y nivel educativo.

	5°	**6°**	**7°**	**8°**	**Total**
Femenino	19	9	35	21	84
Masculino	27	24	41	30	122
No contesta	0	1	1	0	2
Total	46	34	77	51	208

5.7 Plan ético

El equipo de PENTA UC, en colaboración con el Comité Ético Científico de Ciencias Sociales, Artes y Humanidades de la Pontificia Universidad Católica de Chile, desarrolló los protocolos pertinentes para resguardar los aspectos éticos en el diseño, desarrollo y difusión del proyecto.

Los estudiantes que participaron como examinados fueron invitados por los colegios a participar de manera voluntaria. En dicho proceso se señalaron explícitamente los objetivos del estudio y el carácter voluntario de la instancia.

Una vez aceptada la invitación por los colegios, se solicitó a las instituciones enviar la invitación por correo electrónico a los apoderados y estudiantes. Además de dicha invitación, y con la finalidad de resguardar la voluntariedad de participación de los estudiantes y sus apoderados, se envió una invitación desde la coordinación del LIES.

Es importante señalar que, tanto a los colegios como a los apoderados y estudiantes, se les explicitó la posibilidad de retirarse de este en cualquier fase del proceso sin ningún tipo de consecuencias.

La recopilación de los consentimientos y asentimientos se realizó de manera digital. En el caso de los apoderados se envió un formulario con las respectivas cartas. A los

estudiantes se les solicitó completar el asentimiento en la misma plataforma en donde se alojan las pruebas.

Los colegios son retribuidos por su participación mediante la devolución de un informe de resultados por curso evaluado. Para resguardar la confidencialidad de los resultados de los participantes, dicho informe presenta la información de manera agregada

En la aplicación de las pruebas se recogió información sociodemográfica de los estudiantes (género, fecha de nacimiento, nacionalidad, autorreporte de promedio de notas, nacionalidad de la madre y nivel educacional). Ninguno de los elementos presupone un riesgo para los participantes. Los resultados son analizados a nivel agregado y su participación no posee ningún tipo de consecuencia a futuro. La voluntariedad de la participación y la posibilidad de retirarse en cualquier momento de la evaluación resguarda que los examinados no estén obligados a realizar una tarea que no deseen.

6. RESULTADOS

En esta sección se presentan los principales hallazgos de la investigación, de acuerdo con el modelo considerado y los resultados del análisis DIF.

6.1 Teoría clásica del test (CTT)

En la tabla 5 se presentan algunos estadísticos descriptivos generales del instrumento.

Tabla 5. Caracterización de la muestra de estudio.

Tamaño de la muestra	208
Número de ítems	32
Puntuación máxima teórica	56
Puntuación máxima obtenida	40
Alfa de Cronbach	0,89
Media de puntuaciones	13,39
Desviación estándar de las puntuaciones	8,67
Error estándar de la medida	2,86

De los 56 puntos (máximo teórico) se obtuvo una media de 13,4 puntos y una desviación estándar de 8,67 puntos. Esto representa un porcentaje de logro de 24,2%. El error estándar de la medida es de 2,86 puntos, el cual es un valor alto considerando la media obtenida. El alfa de Cronbach indica una alta fiabilidad del instrumento (0,89).

Se observan diferencias significativas ($p - value < 0{,}05$) al agrupar por colegio, en donde el desempeño de los estudiantes de PENTA sobresale en relación con todos los otros grupos. Sin embargo, no se observan diferencias significativas al agrupar por nivel cursado, o por género. En las tablas 6 y 7 se resumen los estadísticos de análisis.

Tabla 6. Puntajes por colegio.

Colegio	Media	N	Desv. típ.
C1	17,45	20	7,19
C2	11,51	95	7,37
C3	12,45	22	7,16
C4	9,57	49	5,91
PENTA	27,27	22	7,20
Total	13,39	208	8,67

Tabla 7. Puntajes por nivel educativo.

Curso	Media	N	Desv. típ.
5°	9,02	46	5,80
6°	13,21	34	8,23
7°	14,94	77	8,82
8°	15,12	51	9,69
Total	13,39	208	8,67

6.2 Modelo de Crédito Parcial (PCM)

Idealmente el Modelo de Crédito Parcial se aplicaría en una situación en la cual para cada ítem se observan estudiantes con puntuación en todos los niveles. Dado que en 5 ítems no se alcanzó el nivel máximo, los resultados deben leerse con esa precaución. Por otro lado, es importante señalar que solo un estudiante obtuvo una puntuación en la máxima categoría del ítem 31, sin embargo, ninguno obtuvo puntuación en la categoría anterior a ella. Esto último impide que el programa compile exitosamente los datos de ese ítem, por lo cual el

estudiante con máxima puntuación en el ítem 31 fue excluido de los análisis con este modelo[1].

En las tablas 8 y 9 se presenta el parámetro θ (localización de la persona), el cual es interpretado como la habilidad del examinado. Esta fue estimada por el método de estimación EAP. Se observan diferencias significativas al agrupar por colegio y nivel educativo, y estas son consistentes con las reportadas en el apartado anterior. No se observan diferencias significativas por género.

Tabla 8. Habilidad (logit) por colegio.

Colegio	Media	N	Desv. típ.
C1	-1,22	20	1,29
C2	-2,07	95	1,34
C3	-1,83	22	1,13
C4	-2,35	49	1,25
PENTA	0,06	22	0,82
Total		208	

Tabla 9. Habilidad (logit) por nivel educativo.

Curso	Media	N	Desv. típ.
5°	-2,47	46	1,27
6°	-1,75	34	1,22
7°	-1,58	77	1,45
8°	-1,56	51	1,49
Total		208	

[1] Otra posibilidad es recodificar la respuesta del estudiante, asignándole una puntuación correspondiente al nivel anterior al que efectivamente alcanzó según la rúbrica. Esto se realizó y se comparó con el efecto de eliminar los datos de ese estudiante. Los resultados del modelo no variaron significativamente.

En relación con la confiabilidad, en este modelo se obtuvieron los siguientes índices: EAP reliability $= 0{,}869$ y WLE reliability $= 0{,}871$. Para verificar el ajuste de los ítems, se verificó el *unweighted mean square* y el *weighted men square*. En el caso del *infit*, todos los valores fluctuaron entre 0,64 y 1,43, lo que evidencia un buen ajuste del modelo a los datos. Los tres ítems de menor *infit* son los ítems 20, 22 y 23 (*infit* menor que 0,7). Solo 5 ítems presentan un *outfit* fuera del rango esperado, como se muestra en la tabla 10.

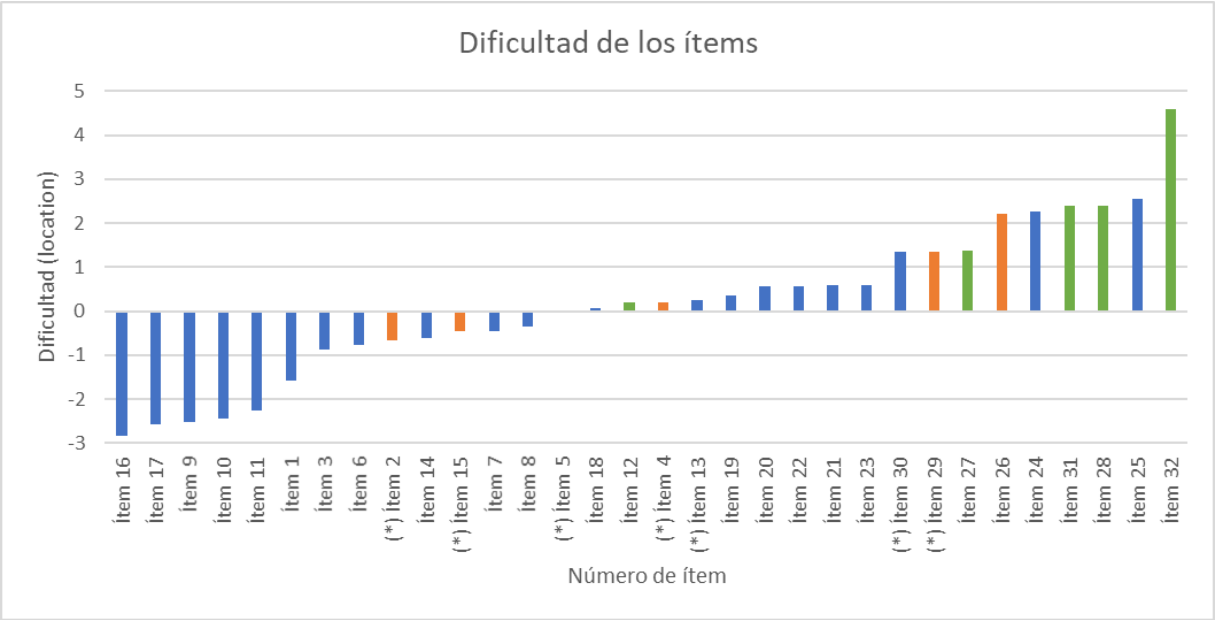
Tabla 10. Ítems con infit y outfit fuera del rango 0,7 – 1,3.

Ítem	***Outfit***	**Índice de discriminación**	**Ítem**	***Infit***	**Índice de discriminación**
Ítem 2	1,59	0,58	**Ítem 20**	0,64	0,76
Ítem 4	3,57	0,52	**Ítem 22**	0,64	0,72
Ítem 15	3,76	0,52	**Ítem 23**	0,68	0,75
Ítem 26	1,71	0,29			
Ítem 29	8,17	0,36			

Los valores altos de *outfit* implican que en esos ítems existen respuestas con puntaje inesperadamente alto por parte de personas con bajo rasgo latente.

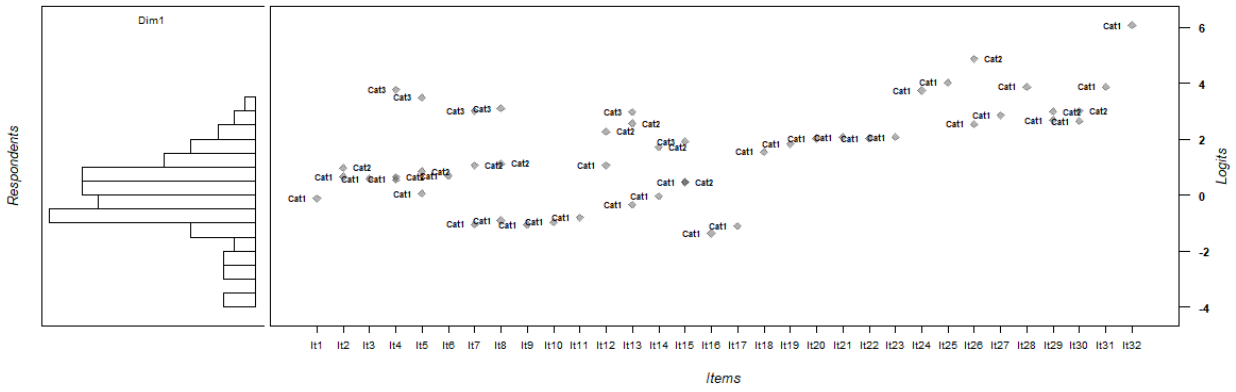
En la Figura 4 se muestra la dificultad de los ítems de acuerdo con el PCM.

Figura 4. Dificultad de los ítems (location). En anaranjado se marcan los ítems con outfit mayor que 1,5 y en verde los ítems en los que no se alcanzaron las categorías superiores de respuesta. Los ítems marcados con asteriscos presentan thresholds desordenados.



En la figura 5 se presenta el *Wright Map*. Esta representación muestra la distribución de rasgo latente por ítem y persona (Wilson, 2011). La distribución muestra que la prueba tiene una alta dificultad para los examinados y que la dificultad empírica del instrumento está distribuida de forma creciente. Esto último es parte de la expectativa teórica, pues el instrumento se diseñó dejando las preguntas de mayor dificultad teórica hacia el final.

Figura 5. Wright Map (librería TAM). Cat 1, Cat 2, Cat 3 corresponden a los thresholds por ítem.



A continuación, se presentan algunos comentarios generales de cada parte de la prueba.

6.2.1 Parte 1 (ítems 1 a 8)

Al ordenar los ítems según dificultad, los 8 ítems de la primera parte se encuentran en la mitad más fácil de la prueba.

Los tres primeros ítems evalúan lectura de gráficos. Todos muestran una dificultad similar. El único ítem no dicotómico (el ítem 2) presenta un nivel intermedio que no funciona adecuadamente (*threshold* desordenados).

El ítem 4 es una pregunta de combinatoria. El nivel 4 de la rúbrica describe que el estudiante responde correctamente y justifica adecuadamente, el nivel 3 describe a los estudiantes que responden correctamente, pero no justifican. El nivel 2, que presenta un comportamiento inadecuado, describe a estudiantes que en su respuesta reflejan una comprensión parcial del problema. Este ítem, a pesar de solicitar una aplicación directa del principio multiplicativo, es el más difícil de esta parte.

Los ítems 5, 6, 7 y 8 evalúan probabilidad clásica. El ítem 6 es dicotómico y es el más fácil de este grupo de ítems. Los ítems 5, 7 y 8 presentan rúbricas análogas de 4 niveles. Los ítems 7 y 8 presentan una muy buena adecuación de cada nivel de la rúbrica. En el ítem 5, a diferencia de los ítems 7 y 8, uno de los niveles no se comporta adecuadamente (presentan *thresholds* desordenados) a pesar de que su rúbrica recoge niveles de desempeños análogos.

Los ítems dicotómicos de esta parte presentan una dificultad baja (ítems 1, 3 y 6).

6.2.2 Parte 2 (ítems 9 a 23)

Respuestas cerradas

Los tres primeros ítems de esta parte (9, 10 y 11) están dentro de los 5 ítems más fáciles del instrumento y responden a una tarea cerrada de conteo simple.

Los ítems 16 y 17 son los de menor dificultad de todo el instrumento (74% y 71% de porcentaje de logro, respectivamente); en cambio los ítems 18 y 19 presentan una dificultad alta (23% y 19% de porcentaje de logro, respectivamente). Los 4 ítems forman parte de la lectura de un mismo estímulo (Figura 6) y responden a la tarea matemática de conteo simple. Los dos primeros ítems preguntan por el conteo de cierta condición de la tabla (por ej. *¿Cuántos están a favor de participar en las marchas?*), mientras que los otros dos preguntan por dos condiciones (por ej. *¿Cuántos están a favor de la participación en las marchas y seleccionan la primera razón (demanda por una educación de calidad)?*).

Figura 6. Estímulo de los ítems 16 a 19. Tomado de PENTA UC (2021b).

Encuesta	A favor participación	Demanda educación de calidad	Vivencia nueva experiecia	Seguir otros grupos
A	Sí	X		
B	Sí		X	
C	Sí			X
D	No	X		
E	No	X		
F	No	X		
G	No			X
H	No			X

Tal como se comentó, los ítems 9 al 11 presentan una baja dificultad, y solicitan del estudiante una tarea idéntica a la de los ítems 16 y 17, por lo que se comportan de manera muy similar.

Por otro lado, los ítems 20 a 23 forman solicitan la tarea de completar una tabla de contingencia a partir del mismo estímulo (Figura 7). Son ítems de alta dificultad y se comportan parecidos a los ítems 18 y 19.

Resumiendo lo anterior, los ítems 9 al 11, 16 y 17 se comportan de manera similar y tienen baja dificultad; y los ítems 18 al 23, tienen comportamiento similar y una dificultad mucho mayor.

Respuestas abiertas

Los ítems 12, 13, 14 y 15 son preguntas abiertas. Los ítems 12, 13 y 15 tienen cuatro niveles, sin embargo, en uno de ellos no se alcanza el nivel superior (ítem 12) y en los otros dos hay niveles que no funcionan adecuadamente de acuerdo con sus *thresholds*.

En el caso del ítem 12, los dos niveles más altos de desempeño sólo difieren en el formato de la respuesta: en el nivel 3 se consideran respuestas en las que el estudiante usa un lenguaje de razón matemática (por ej. *3 de 4*), mientras que en el nivel superior se consideran respuestas en donde el resultado se expresa como porcentaje (por ej. *75%*). Dado que no se solicita en la tarea el cálculo de porcentaje como parte de la argumentación, se sugiere revisar si es adecuado considerarlo como un nivel de desempeño, pues este no surge espontáneamente en las respuestas de los estudiantes.

6.2.3 Parte 3 (ítems 24 a 32)

Estos 9 ítems son, sin excepción, los más difíciles de toda la prueba. Todos consideran el uso del estímulo de la figura 7. A pesar de su dificultad, solo en 3 de estos ítems ningún estudiante alcanza el nivel más alto de la rúbrica.

Figura 7. Estímulo de los ítems en la parte 3. Tomado de PENTA UC (2021b).

	Test positivo	Test negativo
Persona con vitamina	30	12
Persona sin vitamina	5	26

Los ítems 24 y 25 solicitan que el postulante indique el nivel de positividad (razón o porcentaje) de las personas con o sin vitaminas. Los ítems 26 y 27 solicitan interpretar el significado de un bajo o alto porcentaje de positividad en personas con vitaminas, respectivamente; de estas, solo en el ítem 26 se alcanza el nivel de desempeño más alto.

Los ítems 29 y 30 solicitan interpretar el significado de un bajo o alto porcentaje de positividad en personas sin vitaminas, si bien en ambos se alcanzan los tres niveles, el ordenamiento de los *thresholds* muestra un mal funcionamiento del segundo nivel.

6.3 Modelo de Crédito Parcial Generalizado (GPCM)

En este modelo, se consideran parámetros de discriminación variables entre ítems, pero fijos para los *thresholds* de un ítem. Estos se presentan en la tabla 11. Se obtuvo una confiabilidad de 0,912 (EAP – reliability). Este modelo presenta parámetros de dificultad y de persona similares a los que ofrece el PCM. Es importante señalar que los parámetros de discriminación de los ítems 20, 22 y 23 son los más altos del instrumento.

Tabla 11. Parámetro de discriminación de acuerdo con el Modelo de Crédito Parcial Generalizado (GPCM).

Ítem	Parámetro de discriminación	Ítem	Parámetro de discriminación
Ítem 1	1,03	Ítem 17	1,78
Ítem 2	1,14	Ítem 18	1,40
Ítem 3	0,62	Ítem 19	2,17
Ítem 4	0,65	Ítem 20	11,68
Ítem 5	0,62	Ítem 21	3,42
Ítem 6	0,80	Ítem 22	11,07
Ítem 7	0,81	Ítem 23	8,17
Ítem 8	0,82	Ítem 24	3,73
Ítem 9	2,04	Ítem 25	3,36
Ítem 10	2,32	Ítem 26	0,59
Ítem 11	2,08	Ítem 27	0,69
Ítem 12	1,55	Ítem 28	1,26
Ítem 13	1,03	Ítem 29	0,71
Ítem 14	1,20	Ítem 30	0,91
Ítem 15	0,60	Ítem 31	1,64
Ítem 16	2,67	Ítem 32	1,40

En la tabla 12 se comparan los índices de ajuste del PCM y GPCM, considerando log-likelihood (función logaritmo de verosimilitud evaluada en la máxima verosimilitud estimada), NP (número de parámetros estimados), AIC (Criterio de información de Akaike:

$-2\log - \text{likelihood} + 2\text{NP}$; Akaike, 1974) y BIC (Criterio de información Bayesiano: $-2\log - \text{likelihood} + \text{NP} \cdot \log(\text{tamaño muestral})$; Schwarz, 1978).

De acuerdo con DeMars (2012), cuando se tienen modelos anidados, es decir, si uno de los modelos se puede obtener eliminando o fijando los parámetros del otro modelo, es posible comparar con una prueba de χ^2 los índices de bondad de ajuste para determinar si éstas son significativas. En este caso la diferencia en log-likelihood se distribuye como χ^2 con grados de libertad iguales al número de parámetros libres adicionales. Esta prueba también se denomina log-likelihood ratio, porque la diferencia en los log-likelihood es el logaritmo de la razón de las verosimilitudes.

Como se puede verificar en la tabla 12, el GPCM muestra mejores índices de ajuste que el PCM. Estas diferencias son significativas ($\chi^2 = 332$, $\text{p} - \text{value} < 0{,}05$).

Tabla 12. Comparación de Modelo de Crédito Parcial (PCM) y Modelo de Crédito Parcial Generalizado (GPCM).

	EAP reliability	**NP**	**log-likelihood**	**AIC**	**BIC**	χ^2 **−difference test**
PCM	0,90	51	−3304,88	6711,77	6881,74	$\chi^2(31) = 332{,}90^*$
GPCM	0,91	82	−3138,43	6440,87	6714,15	

NP = Número de parámetros; $^*p < 0{,}05$

6.4 Modelos multidimensionales

Como ya se ha señalado, el instrumento utilizado para la evaluación de la alfabetización matemática consiste en tres partes. La primera parte de la prueba (ítem 1 a 8) es una serie de ítems de respuesta abierta corta o cerrada que cumplen con el supuesto de independencia local. La segunda parte de la prueba (ítem 9 a 23) es una serie de ítems anidados con estímulos y contexto común. La tercera parte (ítem 24 a 32) también consiste

en ítems anidados con contexto común que no cumplen con el supuesto de independencia local.

Debido a la dependencia entre ítems entre partes, se consideraron análisis multidimensionales y se compararon con el modelo unidimensional de crédito parcial. Para el anidamiento de ítems, se consideraron las siguientes opciones:

- **Modelo de dos factores específicos**: se consideró la segunda y tercera parte cada una como un factor, y los ítems de la primera parte se cargaron solo al factor general. Esto es:
    - Factor general: todos los ítems
    - Factor 1: Parte 2 de la prueba
    - Factor 2: Parte 3 de la prueba

- **Modelo de tres factores específicos**: se consideró cada parte de la prueba como un factor. Esto es:
    - Factor general: todos los ítems
    - Factor 1: Parte 1 de la prueba
    - Factor 2: Parte 2 de la prueba
    - Factor 3: Parte 3 de la prueba

En las tablas 13 y 14 se presentan los resultados de bondad de ajuste y otros parámetros asociados a cada modelo. Hay mejores índices de ajuste en *bifactor model* y *Rasch testlet model* que en el modelo unidimensional, es decir, al considerar que los ítems cargan en un factor general y dos o tres factores específicos se obtiene mejor bondad de ajuste que al considerar un modelo unidimensional. En el caso del modelo unidimensional (PCM) se obtuvo un AIC (6711,77), siendo este el peor ajuste al ser comparado con los otros modelos. Todas estas diferencias son estadísticamente significativas de acuerdo con el test de hipótesis. Estos resultados sugieren evidencia de efecto *testlet* asociado a las distintas partes de la prueba.

Tabla 13. Comparación de Modelo de Crédito Parcial (PCM) y modelos de dos factores: bifactor model (BM) y Rasch testlet model (RTM).

	EAP reliability	**NP**	**log-likelihood**	**AIC**	**BIC**	χ^2 **−difference test**
PCM	0,90	51	−3304,88	6711,77	6881,74	$\chi^2(2) = 160{,}59^*$
RTM (2 factores)	FG: 0,81 F1: 0,65 F2: 0,35	53	−3224,59	6555,18	6731,82	$\chi^2(53) = 397{,}78^*$
BM (2 factores)	FG: 0,86 F1: 0,64 F2: 0,36	106	−3025,70	6263,40	6616,67	

NP = Número de parámetros; $^*p < 0{,}05$

Tabla 14. Comparación de Modelo de Crédito Parcial (PCM) y modelos de tres factores: bifactor model (BM) y Rasch testlet model (RTM).

	EAP reliability	**NP**	**log-likelihood**	**AIC**	**BIC**	χ^2 **−difference test**
PCM	0,90	51	−3304,88	6711,77	6881,74	$\chi^2(3) = 157{,}52^*$
RTM (3 factores)	FG: 0,80 F1: 0,04 F2: 0,64 F3: 0,35	54	−3226,12	6560,25	6740,21	$\chi^2(60) = 621{,}60^*$
BM (3 factores)	FG: 0,86 F1: 0,80 F2: 0,62 F3: 0,38	114	−2915,32	6058,65	6438,57	

NP = Número de parámetros; $^*p < 0{,}05$

Para el caso de un *bifactor model* de tres factores se obtiene el menor AIC (6058,65). Estos resultados pueden responder a que las restricciones asociadas a la proporcionalidad de cargas del *Rasch testlet model* son demasiado estrictas (Rijmen, 2010), es decir, los *bifactor model* ajustan mejor al tener mayor libertad en la estimación de las cargas. Dado que los modelos con más parámetros ajustan mejor, es importante comparar los valores de log-likelihood considerando una prueba de χ^2 con grados de libertad iguales al número de

parámetros libres adicionales. Considerando lo anterior, se tiene que el *bifactor model* ajusta significativamente mejor que los otros modelos.

Para comparar la forma en que ambos modelos estiman los parámetros de ítem (dificultad) y persona (habilidad) se graficaron las correlaciones entre los modelos. En la Figura 8 se muestran las correlaciones asociadas a los parámetros de dificultad para cada modelo. *Bifactor model* (de dos o tres factores específicos) muestra algunos parámetros de dificultad distorsionados en relación con los del modelo de crédito parcial unidimensional (PCM) o *Rasch testlet model*. Como puede observarse, *Rasch testlet model* muestran correlaciones casi perfectas con el modelo unidimensional. Las dificultades estimadas por el *bifactor model* muestran correlaciones menores a 0,43 con los otros modelos. Sin embargo, dicho coeficiente se duplica si no se consideran los ítems 8, 20 y 22, en los cuales la dificultad del *bifactor model* se desajusta fuertemente. Debe considerarse que, teóricamente y desde su comportamiento psicométrico, los ítems 7 y 8 son similares (aluden a una misma tarea). En el caso de los ítems 20 a 23 se tiene la misma situación (los cuatro corresponden a la completación de una misma tabla de contingencia). Por ello, los valores de dificultad entregados por el *bifactor model* para los ítems 8, 20 y 22 (superiores a 14 logits) deben considerarse desajustados.

Para indagar en las bajas correlaciones entre el *bifactor model* y los otros modelos, se grafican las dificultades de estos modelos con las proporcionadas por el modelo unidimensional (Figura 9). Es importante notar cómo en ambos *bifactor models,* los ítems 20 y 22 presentan problemas en la estimación de su dificultad.

Figura 8. Correlaciones entre las dificultades de los 5 modelos considerados (PCM: Partial credit model; RTM: Rasch testlet model; BM: Bifactor model).

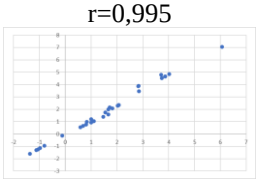
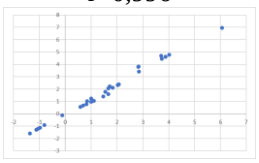
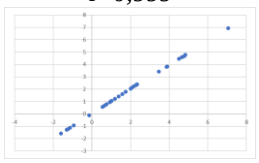
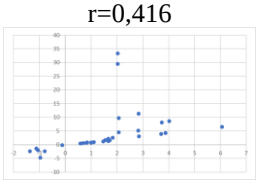
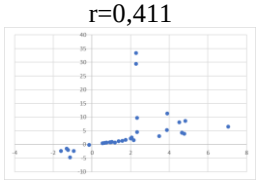
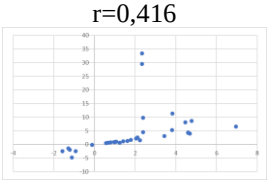
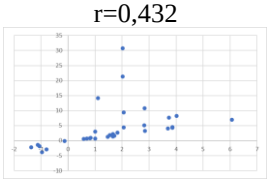
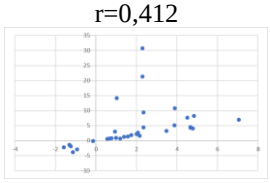
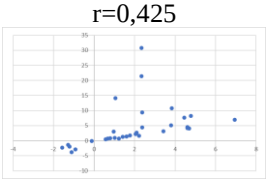
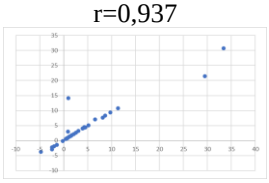
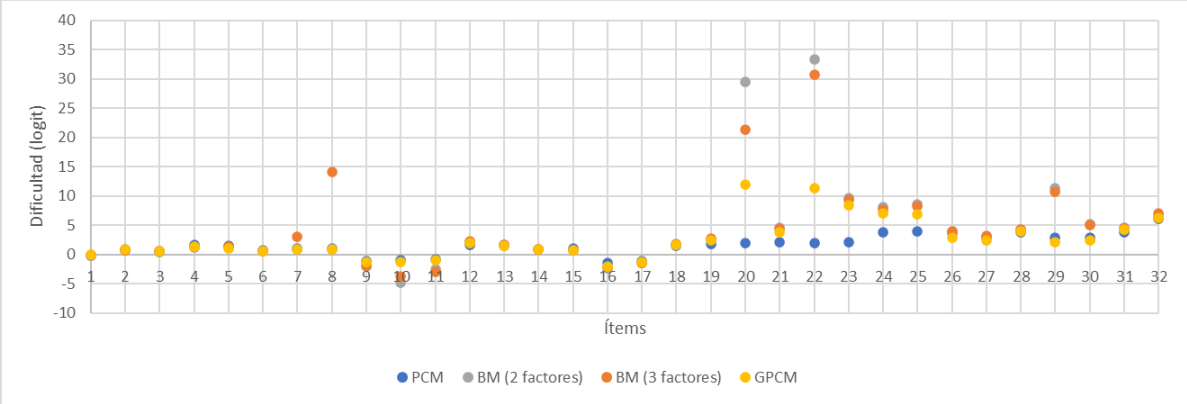
	PCM	**RTM (2 factores)**	**RTM (3 factores)**	**BM (2 factores)**
RTM (2 factores)	r=0,995 			
RTM (3 factores)	r=0,996 	r=0,999 		
BM (2 factores)	r=0,416 	r=0,411 	r=0,416 	
BM (3 factores)	r=0,432 	r=0,412 	r=0,425 	r=0,937 

Figura 9. Dificultades de los modelos considerados (PCM: Partial credit model, unidimensional; BM: Bifactor model; GPCM: Generalized Partial Credit Model).



Por otro lado, se compararon los parámetros por persona (habilidad) entre el PCM (unidimensional) y los otros modelos, como se muestra en Figura 10. Como se puede observar, se obtienen correlaciones mayores que 0,95, es decir, independiente del modelo usado las estimaciones para el rasgo latente (habilidad) por persona son similares.

Algunos autores sugieren evaluar los efectos *testlet* comparando las cargas de los factores específicos con las cargas del factor general (Min y He, 2014; Byung y Lee, 2016). Si las cargas a los factores específicos son mayores, se sugiere mayor fuerza del efecto *testlet*. En las Figuras 11 y 12 se presentan las cargas (específicas y generales) para los modelos de tres factores. Estas se comparan con el índice de discriminación estimado por *eRm package* en un modelo unidimensional (PCM). Para los modelos multidimensionales se tiene, en general, cargas específicas menores que las cargas al factor general en la parte 1 y 2 de la prueba (Figuras 11 y 12). Solo en la parte 3 se observan varias cargas específicas mayores a la carga general, lo que sugiere un efecto *testlet* más pronunciado en esa parte de la prueba.

Figura 10. Correlaciones entre el parámetro por persona (habilidad). (PCM: Partial credit model, unidimensional; RTM: Rasch testlet model; BM: Bifactor model; 2F es el modelo con dos factores específicos y un factor general; 3F es el modelo con tres factores específicos y un factor general).

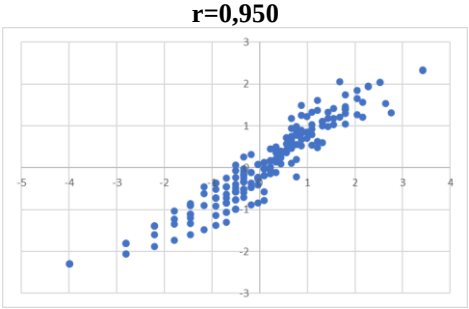
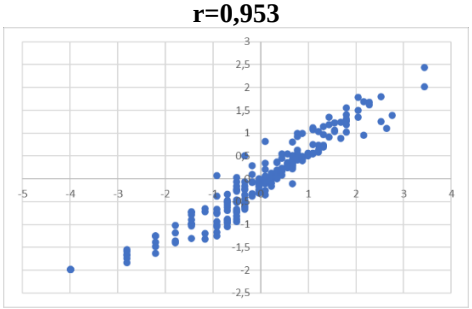
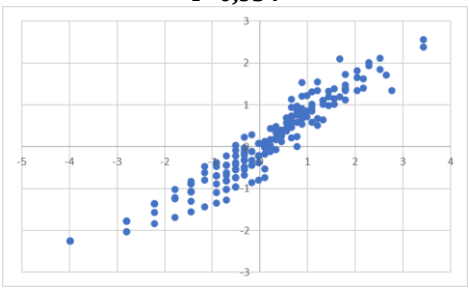
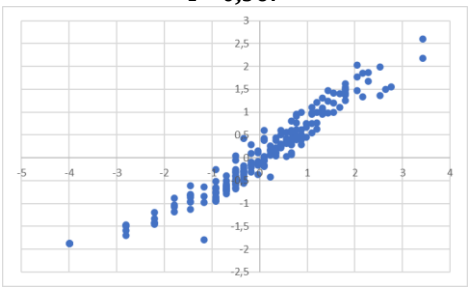
	PCM		PCM
RTM (2 factores)	**r=0,950** 	**BM (2 factores)**	**r=0,953** 
RTM (3 factores)	**r=0,954** 	**BM (3 factores)**	**r= 0,967** 

Figura 11. Cargas factoriales para bifactor model de 3 factores. Las líneas verticales separan las tres partes de la prueba.

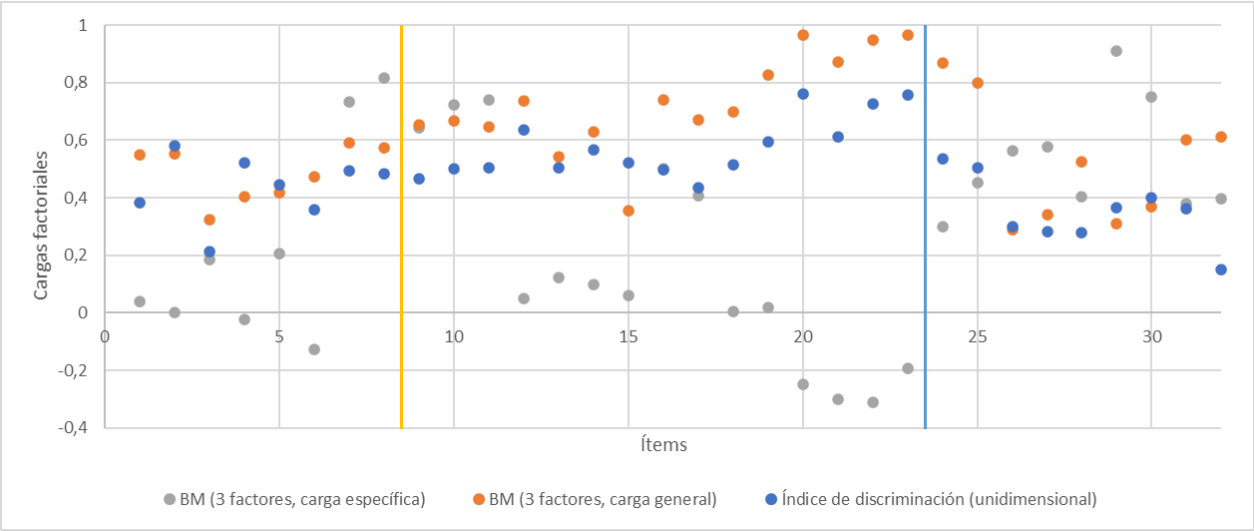
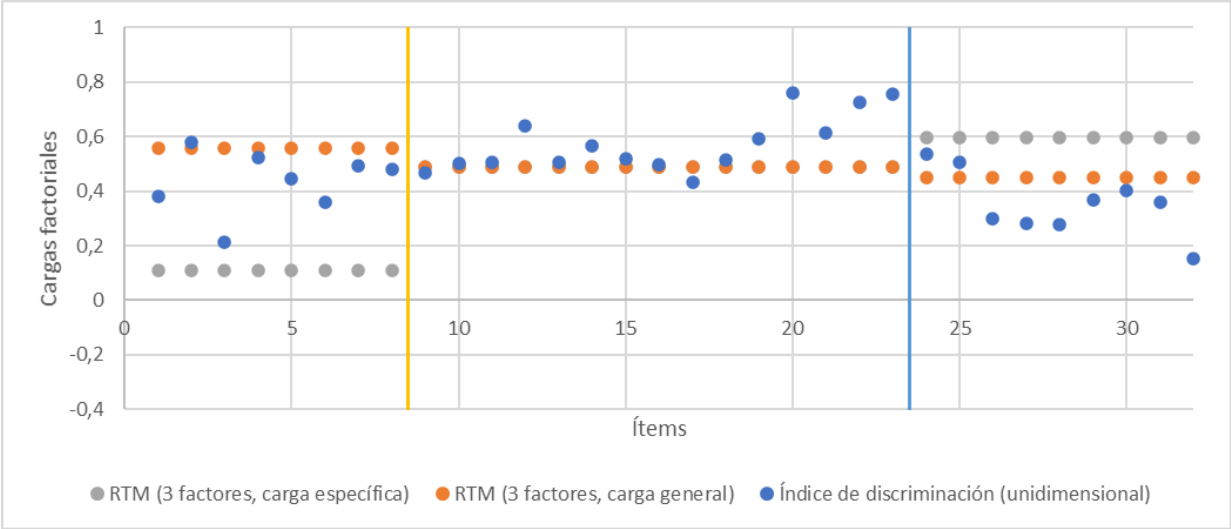


Figura 12. Cargas factoriales para Rasch testlet model (RTM) de 3 factores. Las líneas verticales separan las tres partes de la prueba.



6.5 Funcionamiento diferencial del ítem

Es importante considerar que los análisis de DIF solo se efectuaron considerando los ítems dicotómicos debido a las restricciones asociadas al tamaño de la muestra. A pesar de no ser dicotómico, el ítem 32 no fue analizado debido a su baja tasa de respuesta.

En términos de género, se observa DIF a favor del grupo masculino en la pregunta 9 (1,29 logit de diferencia; $\mathrm{p-value} < 0{,}05$). Para evaluar si el DIF implica sesgo, se debe evaluar el ítem para analizar si hay elementos del reactivo que puedan favorecer a un grupo por sobre otro. No hay evidencia de que el ítem 9 presente elementos que puedan generar sesgos de género, en particular porque otros ítems similares (10 y 11) no presentan un DIF.

En relación con el nivel, se compararon los estudiantes de quinto y sexto básico con los de séptimo y octavo. Se observó DIF a favor de estos últimos en el ítem 27 (1,39 logit de diferencia, $\mathrm{p-value} < 0{,}05$).

Por último, solo se observa DIF a favor de los estudiantes que con talento académico en el ítem 3 (2,73 logit de diferencia, $\mathrm{p-value} < 0{,}05$). Este requiere responder una pregunta en la que se solicita al participante elegir una opción correcta entre una serie de afirmaciones sobre frecuencias relativas porcentuales, las cuales no están explícitas en el gráfico que sirve de estímulo (este considera frecuencias absolutas). Es importante señalar que la representación de datos en gráficos de frecuencias relativas y un trabajo más aplicado de porcentajes corresponden a objetivos de aprendizaje de séptimo básico, nivel en el que cursan 12 de los 22 estudiantes con talento académico. De los otros 10 estudiantes, 4 cursan sexto básico; y 8, octavo básico, por lo que es posible el avance curricular explique el DIF.

7. DISCUSIÓN DE LOS RESULTADOS Y CONCLUSIONES

La evaluación de aprendizajes es un campo cada vez más relevante en educación. En el caso de los estudiantes con talento académico, permite el seguimiento de su progreso y crecimiento, y la evaluación de las intervenciones educativas que se les brindan (Cao et al., 2017). La batería de instrumentos diseñada por LIES apunta en esa dirección. En particular, la prueba de matemática tiene por objetivo evaluar el nivel base de los estudiantes en alfabetización matemática, en relación con el marco curricular del programa. Este último considera, entre otras habilidades, la metacognición, el pensamiento crítico, la ciudadanía y la comunicación.

Considerando que el desarrollo de pruebas debe guiarse por estándares de calidad que garanticen la confiabilidad, validez e imparcialidad de los resultados obtenidos o de sus usos, en este apartado se presentan algunas discusiones sobre los resultados obtenidos.

7.1 Confiabilidad de las puntuaciones del instrumento diseñado

AERA, APA, NCME (2014) definen la confiabilidad como la coherencia de puntajes entre replicaciones de un procedimiento de evaluación, independientemente de cómo se estime o reporte. En el contexto de la teoría clásica del test se consideró el uso del alfa de Cronbach (α) como medida de consistencia interna. El valor obtenido da cuenta de una alta confiabilidad del instrumento ($\alpha = 0{,}89$), los valores menores que 0,90 son indicativos de no redundancia de elementos en el instrumento (Streiner, 2003). Además, se consideraron formas de estimaciones alternativas (EAP y WLE). En ambos casos, para el modelo de crédito parcial, el instrumento diseñado presenta alta confiabilidad ($\text{EAP reliability} = 0{,}896$; $\text{WLE reliability} = 0{,}871$). En el caso del GPCM se obtuvo un EAP reliability de $0{,}912$. En *bifactor model* y *Rasch testlet model* se obtienen siempre confiabilidades sobre

0,8 para el factor general, sin embargo, la confiabilidad baja si se consideran las partes de la prueba por sí solas. En relación con esto, se sugiere seguir aplicando el instrumento sin eliminar ninguna de sus partes.

7.2 Validez asociada a los usos instrumento diseñado

De acuerdo con AERA, APA, NCME (2014), la validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para sus usos propuestos. De acuerdo con estándares, las fuentes de evidencia de validez abarcan el contenido de la prueba, los procesos de respuesta, la estructura interna, la relación del constructo medido con otras variables y la interpretación de las puntuaciones, en particular en relación con las consecuencias asociadas a sus usos.

Es importante señalar que el instrumento diseñado no tiene por objetivo diagnosticar el talento académico, sino que un nivel de base de la alfabetización matemática para evaluar el progreso de los estudiantes. Es decir, el instrumento no fue diseñado para ser aplicado en la población general, sino que, para estudiantes con talento académico. Dado que las pruebas estandarizadas de uso común tienen un techo muy bajo, esto dificulta que los estudiantes con talento puedan demostrar crecimiento en las mediciones con ese tipo de instrumentos (VanTassel-Baska et al., 2002). En esa línea, la evidencia muestra que la prueba diseñada tiene una dificultad alta para la población general. Sin embargo, a pesar de su dificultad, solo 5 de los 32 ítems no registran respuestas en todos los niveles, es decir, la prueba es abordable para evaluar el crecimiento del constructo involucrado. Por último, en relación con la dificultad, la prueba presenta una dificultad empírica creciente, lo que confirma la expectativa teórica, pues esa graduación de dificultad era parte del diseño del instrumento.

Como índices de ajuste del modelo se suelen examinar el comportamiento del *outfit* (*unweighted mean square*) e *infit (unweighted mean square)* a nivel de ítem. El comportamiento de los ítems de acuerdo con un modelo estadístico esperado aportará evidencia a favor o en contra de la válida interpretación de los puntajes de un instrumento. Generalmente, se aceptan valores de *infit* y *outfit* entre 0,7 y 1,3 para pruebas de selección múltiple, o entre 0,8 y 1,2 si esta es de altas consecuencias (Bond y Fox, 2013), a pesar de que otros autores consideran rangos más amplios (Zubairi y Kassim, 2016). Sin embargo, esto es una regla general no exenta de críticas, pues se trata de rangos que no consideran las propiedades de distribución de las estadísticas de ajuste como el tamaño de la muestra, el número de ítems, la diferencia entre la distribución de las habilidades de la persona y la distribución de las dificultades de los ítems, y la varianza y la curtosis de las respuestas de los ítems (Smith, 1991; Wang & Chen, 2005).

En el caso del *infit*, los valores de todos los ítems fluctuaron entre 0,66 y 1,43, lo que representa un buen ajuste del modelo a los datos. Solo 5 de los 32 ítems presentaron un *outfit* mayor que 1,5. En el caso del *infit*, este le da relativamente más peso al desempeño de las personas ubicadas más cerca del valor de dificultad del ítem. El *outfit* es una medida no ponderada y, por lo tanto, es relativamente más sensible a la influencia de puntajes de personas distantes de la ubicación del ítem. Es por esta razón se suele prestar más atención a los valores de *infit* que a los de *outfit*, pues el primero da una idea más sensible del desempeño de ese ítem (Bond y Fox, 2013).

En general, el uso de preguntas de selección múltiple en pruebas estandarizadas no ha estado exento de discusión en torno a su capacidad para medir habilidades de alto nivel e impedir que los examinados demuestren un mayor repertorio en la resolución de problemas matemáticos (Scully, 2017). Dado que la prueba diseñada está principalmente compuesta por preguntas abiertas, permite que las respuestas sean evaluadas en un rango de categorías, sin embargo, estas deben validarse en relación con la forma en que las puntuaciones de la rúbrica permiten diferenciar y calificar las respuestas. De las 16 preguntas politómicas en el instrumento, 3 ítems terminaron funcionando como dicotómicos por no tener puntuaciones en la categoría más alta (ítems 27, 28 y 32). Algo similar ocurrió con los ítems 12 y 31 (ambos de 4 niveles), los cuales funcionaron con menos niveles de los propuestos teóricamente. Así, de los 32 ítems que componen la prueba, en cinco de ellos no se alcanza uno de los niveles descritos en la rúbrica. Los ítems restantes presentaron un funcionamiento adecuado de la rúbrica. Si bien 8 ítems presentan problemas con sus *thresholds* (ítems 2, 4, 5, 13, 15, 29, 30, 31), el análisis de la rúbrica en todos esos casos no permite sugerir el colapso de categorías. Debe verificarse que el comportamiento de las categorías sea similar en el caso de aplicarse el instrumento a un alto número de estudiantes de la población objetivo (solo 22 estudiantes con talento académico rindieron la evaluación). Todos los problemas descritos deben ser considerados para ajustar la rúbrica, pero el tamaño muestral no permite proponer colapso de categorías.

Como parte de los requerimientos en la elaboración de instrumentos para evaluación en matemática, Tout y Spithill (2015) señalan que las respuestas de los estudiantes debieran ser calificadas de manera consistente y que el rango de dificultad de los ítems debiera abarcar

una gama que permita a todos los examinados poder encontrar ítems que les permitan demostrar sus destrezas. Ambos requerimientos se evidencian en el instrumento diseñado tanto por su buena consistencia entre correctores (coeficiente kappa de Cohen mayor que 0,85), como por su rango de dificultad.

En el diseño de instrumentos para evaluar la alfabetización matemática se suelen especificar ítems para cada una de las etapas del ciclo de modelamiento: *formular*, *emplear* e *interpretar* (Tout y Spithill, 2015; Munadi y Febriyanti, 2020). En el instrumento diseñado, la tabla de especificaciones no asocia los ítems con cada uno de los procesos del ciclo de modelamiento, pues estos no fueron diseñados en correspondencia unívoca con cada uno de ellos. Una de las razones de especificar a qué proceso del ciclo se asocia cada ítem es equilibrar el peso de los procesos. En el caso de PISA (OECD, 2018), por ejemplo, se intenta equilibrar el peso de los procesos de conexión (formular e interpretar), con el peso del proceso *emplear* (junto con el razonamiento matemático). Por otro lado, si los ítems del instrumento se asociaran a cada uno de los procesos matemáticos involucrados, se podría contar con información más pormenorizada para el diseño de acciones pedagógicas pertinentes. Esto último permitiría, por ejemplo, evaluar si hay algún proceso del ciclo de modelamiento en el que se deba enfatizar el trabajo académico con los estudiantes.

A pesar de que PISA es una de las evaluaciones prestigiosas del mundo en alfabetización matemática, y cuenta con un diseño psicométrico sólido, algunos autores argumentan que los métodos psicométricos que se utilizan con mayor frecuencia (como los modelos de Rasch) pueden ser demasiado limitantes para tal constructo (Ekmekci y Carmona,

2014). En esta línea, los enfoques que permiten la multidimensionalidad pueden proporcionar evaluaciones más válidas.

En el diseño de instrumentos para evaluar alfabetización matemática, Tout y Spithill (2015) sugieren que los ítems deben ser los más independientes los unos de los otros, en particular, que la respuesta de un ítem no influya en la respuesta de otro, siendo esto considerado un requerimiento del modelo psicométrico. Si bien es cierto que IRT requiere independencia local de los ítems, los resultados obtenidos muestran que al aplicar modelos multidimensionales que no consideran ese supuesto, sí se obtienen puntuaciones coherentes y consistentes. Esto se puede considerar como una oportunidad para el diseño de ítems anidados en los que se sigue un enfoque en donde el examinado tenga la posibilidad de ir reflexionando sobre sus respuestas anteriores para ir resolviendo un problema.

Si bien tradicionalmente se ha asumido la alfabetización matemática como unidimensional, algunas investigaciones han demostrado que aplicar modelos multidimensionales puede ser una herramienta poderosa para construir perfiles para cada estudiante de acuerdo con sus puntuaciones en cada dimensión (Wu & Adams, 2006; Frey et al., 2013). Dos modelos que permiten abordar la multidimensionalidad cuando hay ítems que se responden a partir de un estímulo común son el *bifactor model* y *Rasch testlet model*. Wainer y Wang (2000) demostraron que cuando se ignoraba la dependencia local que se deriva de la estructura de tipo *testlet*, las dificultades de los ítems no se alteran significativamente, aunque sí se afectan importantemente los parámetros de discriminación: ignorar el efecto *testlet* lleva a que las discriminaciones de ítems se subestimen para los ítems de *testlet* y se sobrestimen para los ítems independientes. Los resultados obtenidos en el

instrumento aplicado son coherentes con tales hallazgos en tanto no observan alteraciones generales en los parámetros de dificultad o habilidad, aunque sí en los parámetros de discriminación.

Si bien el uso de GPCM, *bifactor model* y *Rasch testlet model* añaden complejidad al análisis, no están proporcionando parámetros de dificultad y habilidad muy distintos que los del PCM. La mejor bondad de ajuste de los *bifactor models* en relación con los *Rasch testlet model* debe leerse con precaución, considerando que los ajustes de este último modelo pueden responder a las restricciones impuestas sobre las cargas factoriales y, en el caso del PCM, a las restricciones impuestas sobre los parámetros de discriminación. A pesar de que los mejores índices de ajuste del *bifactor model* (AIC, BIC) sugieren efecto de método, tal modelo presenta problemas en la estimación de dificultad de algunos ítems (ítems 8, 20 y 22). Además, el GPCM muestra problemas en la estimación del parámetro de discriminación de los ítems 20, 22 y 23. Considerando que estos tres ítems son los de menor *infit* de todo el instrumento, se sugiere la posibilidad de redundancia de ítems (Bond y Fox, 2013). Los ítems 20 a 23 corresponden a la completación de una misma tabla en el contexto de una tarea de conteo muy similar.

A modo de resumen, de acuerdo con los valores de *infit* y *outfit* del PCM hay un buen ajuste de los datos al modelo estadístico. Si bien la validez se define sobre la interpretación de las puntuaciones del instrumento, un buen ajuste del modelo evidencia consistencia en la estructura interna del instrumento, con lo que entrega información respecto a evidencias de validez. Por otro lado, si bien la dependencia entre los ítems puede llevar a cuestionar el uso de un modelo unidimensional, la evidencia muestra que los parámetros de dificultad y

habilidad proporcionados por dicho modelo no son muy diferentes de los que se obtienen al usar modelos de más parámetros. Aunque 8 de los 11 ítems que funcionaron de manera politómica presentan *thresholds* desordenados, no hay evidencia suficiente para proponer colapso de categorías.

Por último, es importante señalar que si bien la prueba se construyó considerando la incorporación transversal de las habilidades del marco curricular de PENTA UC (metacognición, pensamiento crítico, ciudadanía y comunicación), la evidencia no permite dar cuenta de la medición de tales habilidades.

7.3 Imparcialidad de los ítems del instrumento diseñado

La imparcialidad es fundamental en el contexto de medición educativa, y su resguardo busca que las brechas asociadas a inequidades o diferencias en el acceso a oportunidades no se reflejen en las puntuaciones del instrumento (AERA, APA, NCME, 2014).

En los análisis realizados se compararon los subgrupos de acuerdo con las variables de género (femenino versus masculino), nivel (quinto y sexto versus séptimo y octavo), y talento académico (estudiantes de PENTA versus estudiantes regulares). Es importante señalar que los análisis de DIF se realizaron considerando sólo los ítems dicotómicos, por lo que los resultados deben ser leídos con esa precaución.

En relación con los resultados obtenidos, no se observa sesgo de género en las puntuaciones.

En el caso de la variable de talento académico, solo se observa DIF a favor de los estudiantes de PENTA en un único ítem de la prueba (ítem 3, selección múltiple). Sin embargo, en ambos casos, es importante considerar que en los análisis el tamaño de la muestra de estudiantes con talento es bajo ($n = 22$). En futuros estudios de funcionamiento

diferencial debería resguardarse una muestra que considere un mayor número de estudiantes con talento académico.

Un solo ítem presenta DIF a favor de los estudiantes de cursos mayores. Dado que el instrumento se desea aplicar en sexto básico, se sugiere modificar los ítems que involucran el uso del concepto de porcentaje, pues dicho concepto se aborda por primera vez en ese nivel, o reformularlos de tal manera que no requieran el uso de ese contexto. De confirmarse progreso en tales ítems, este se podría explicar a partir del avance curricular regular. De esta manera se evita que el contenido propio del currículum nacional sea fuente de varianza irrelevante en la evaluación del constructo.

8. RECOMENDACIONES PARA EL DISEÑO DE LA PRUEBA Y RÚBRICAS DE CORRECCIÓN

En relación con el instrumento general y sus aplicaciones, se sugiere lo siguiente:

- En lo posible, aumentar el tiempo de aplicación del instrumento. La evidencia muestra que el instrumento tiene una dificultad creciente, debe resguardarse que esta responde a la naturaleza de los ítems y no a la prisa de los examinados por responder debido a problemas de tiempo. Se sugiere realizar entrevistas cognitivas en las que se aplique el instrumento completo y no solo partes de él, para pesquisar información sobre la longitud de la prueba.
- Las entrevistas cognitivas solo se realizaron a estudiantes mujeres. Se sugiere realizarlas con un enfoque paritario para resguardar que no se reproduzcan sesgos de género.
- Por razones de tamaño muestral, para el análisis de DIF por niveles educativos se consideró a los estudiantes de quinto y sexto como un solo grupo, y a los de séptimo y octavo como otro. Se sugiere que en nuevos pilotajes se procure un tamaño muestral que permita efectuar el análisis de DIF considerando cada uno de los cuatro niveles como un solo grupo. Esto último es de relevancia para poder aislar los efectos del avance de los estudiantes en el currículum nacional, de los efectos del currículum de PENTA.

En base a la evidencia disponible, se sugieren las siguientes recomendaciones en relación con algunos ítems:

- Se sugiere en el ítem 12 eliminar el cuarto nivel que describe respuestas que involucran el uso de porcentaje, en particular porque no es solicitado en la pregunta y no surge de manera espontánea en los examinados.
- En los ítems 27, 28, 31 y 32 no se sugiere eliminar categorías, pues otros ítems que solicitan tareas similares si registran estudiantes en cada nivel. El hecho de que no se alcanzara el nivel más alto puede estar relacionado con la alta dificultad de los ítems y el bajo número de examinados que los contestó.
- Por otro lado, siete de los 11 ítems politómicos del instrumento presentan *thresholds* desordenados (ítems 2, 4, 5, 13, 15, 29 y 30). Esto no necesariamente es indicativo de un mal funcionamiento de las categorías, en particular porque rúbricas con categorías análogas no presentan los mismos problemas por las categorías. Se sugiere considerar muestras de mayor tamaño para estudiar el colapso de categorías.
- Se sugiere reformular los ítems que utilizan el concepto de porcentaje para evitar la varianza irrelevante que aporta al constructo (ítem 3, y 27 a 31). Es posible que el DIF a favor algunos estudiantes respondan a la varianza añadida por el avance curricular.
- Los ítems 20 a 23 ($infit < 0{,}7$) parecen mostrar redundancia, lo que explicaría los problemas en su estimación de parámetros de dificultad y discriminación. Considerando que responden a una misma tarea matemática, se sugiere estudiar la posibilidad de considerarlos como un solo ítem, asignándole un puntaje de 0 a 4.

9. REFERENCIAS BIBLIOGRÁFICAS

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162-172.

Adelson, J. L., McCoach, D. B., & Gavin, M. K. (2012). Examining the effects of gifted programming in mathematics and reading using the ECLS-K. Gifted Child Quarterly, 56(1), 25–39.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational ResearchAssociation.

Agencia de la calidad de la educación (2017). Informe Nacional de la Calidad de la Educación.

Agencia de la calidad de la educación (2019). PISA 2018 Entrega de Resultados Competencia Lectora, Matemática y Científica en estudiantes de 15 años en Chile.

Agencia de la calidad de la educación (2020a). Resultados Educativos 2019.

Agencia de la calidad de la educación (2020b). TIMSS 2019 Estudio Internacional de Tendencias en Matemática y Ciencias.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Algaba, A., y Fernández, T. (2021). Características socioemocionales en población infanto-juvenil con altas capacidades: Una revisión sistemática. *Revista de psicología y educación.*

Arancibia, V. (2009). La educación de alumnos con talentos: una deuda y una oportunidad para Chile.

Benavides, M., Ríos C. y Marshall C. (2004). La educación de niños con talento en Chile. In La educación de Niños con Talento en Iberoamérica (pp. 105-114). Ediciones UNESCO.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. Lord & M. Novick (Eds.). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Blanco, R., Ríos, C. G., y Benavides, M. (2004). Respuesta educativa para los niños con talento. *La educación de niños con talento en Iberoamérica*, 49-60.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Brody, L. E., & Stanley, J. C. (2005). Youths who reason exceptionally well mathematically and/or verbally: Using the MVT:D4 model to develop their talents. In R. J.

Bybee, R. (1997). Achieving scientific literacy: From purposes to practices. Portsmouth: Heinemann.

Byun, J. H., & Lee, Y. W. (2016). Investigating topic familiarity as source of testlet effect in reading tests: bifactor analysis. *외국어교육*, *23*(3), 79-109.

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221-256.

Cao, T. H., Jung, J. Y., & Lee, J. (2017). Assessment in gifted education: A review of the literature from 2005 to 2016. *Journal of Advanced Academics*, 28(3), 163-203.

Card, D., & Giuliano, L. (2016). Universal screening increases the representation of low-income and minority students in gifted education. Proceedings of the National Academy of Sciences, 113(48), 13678–13683.

Cockcroft, W. (1982). Mathematics counts. Report of the committee of inquiry into the teaching of mathematics in schools under the chairmanship of Dr W. H. Cockcroft. London: Her Majesty's Stationery Office.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Daskolia, M., Dimos, A., & Kampylis, P. G. (2012). Secondary Teachers' Conceptions of Creative Thinking within the Context of Environmental Education. *International Journal of Environmental and Science Education*, 7(2), 269-290.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. Journal of Research in Science Teaching, 37 (6), 582–601.

De Lange, J. (1999). Framework for classroom assessment in mathematics. *Utrecht: Freudenthal Institute and National Center for Improving Student Learning and Achievement in Mathematics and Science*.

De Lange, J. (1999), Framework for classroom assessment in mathematics, National Center for Improving Student Learning and Achievement in Mathematics and Science.

DeMars, C. (2010). Item response theory. Oxford University Press.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104-121.

DEMRE (2021).Temario. Contenidos de las Pruebas de Transición a la Educación Superior. Prueba obligatoria de matemática. Proceso de admisión 2022. Recuperado de: https://demre.cl/publicaciones/2022/2022-21-04-26-demre-temario-matematica

Ekmekci, A., & Carmona, G. (2014). Studying Mathematical Literacy through the Lens of PISA's Assessment Framework. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Ennis, R. H. (1993). Critical thinking assessment. Theory into practice, 32(3), 179-186.

Flanagan, A., y Arancibia, V. (2005). Talento académico: Un análisis de la identificación de alumnos talentosos efectuada por profesores. *Psykhe (Santiago)*, *14*(1), 121-135.

Freudenthal, H. (2012). Mathematics as an educational task. Springer Science & Business Media.

Frey, A., Seitz, N. N., & Kröhne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In *Research on PISA* (pp. 103-120). Springer, Dordrecht.

Gagné, F. (1985). Giftedness and Talent: Reexamining a Reexamination of the Definitions. Gifted Child Quarterly, 29(3), 103–112.

Gagné, F. (2004). Transforming gifts into talents: the DMGT as a developmental theory. *High Ability Studies*, 15(2), 119-147.

Gagné, F. (2013). The DMGT: Changes within, beneath, and beyond. *Talent Development & Excellence*, 5(1), 5-19.

Galton, F. (1869). Hereditary genius. London: Macmillan.

Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. New York: Basic Books.

Gempp, R. G. (2006). El error estándar de medida y la puntuación verdadera de los tests psicológicos: Algunas recomendaciones prácticas. *Terapia psicológica*, *24*(2), 117-129.

Geramipour, M. (2021). Rasch testlet model and bifactor analysis: how do they assess the dimensionality of large-scale Iranian EFL reading comprehension tests? *Language Testing in Asia*, 11(1), 1-23.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability* (pp. 129-161). Wiley.

Glas, C. A., & Verhelst, N. D. (1995). Testing the Rasch model. In *Rasch models* (pp. 69-95). Springer, New York, NY.

Gulliksen, H. (1950). Theory of mental tests. Nueva York: Wiley.

Hogan, T. (2004). *Pruebas Psicológicas*. El Manual Moderno.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57(5), 253–270.

Jablonka, E. (2003). Mathematical literacy. *In Second international handbook of mathematics education* (pp. 75-102). Springer, Dordrecht.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50(2),* 186–203.

Kaufman, S. B. y Sternberg, R. J. (2008). Conceptions of giftedness. En S. Pfeiffer (Ed.), Handbook of giftedness in children (pp. 71-91). Tallahassee, FL: Springer.

Lester Jr, F. K. (2003). From problem solving to modeling: The evolution of thinking about research on complex mathematical activity. *Beyond constructivism: Models and modeling perspectives on mathematical problem solving, learning, and teaching*.

López, V., Bralic, S., y Arancibia, V. (2002). Representaciones sociales en torno al talento académico: Estudio cualitativo. *Psykhe* 11(1), 183-201.

Lord, F. M., y Novick, M. R. (1968). Statistical theories of mental test scores. New York: Addison-Wesley.

Mair P, Hatzinger R, Maier MJ (2021). *eRm: Extended Rasch Modeling*. 1.0-2, https://cran.r-project.org/package=eRm.

Marín, A., y Lupiáñez, J. L. (2005). Principios y estándares para la educación matemática: una visión de las matemáticas escolares. *Suma: Revista sobre Enseñanza y Aprendizaje de las Matemáticas*, 48, 105-110.

Marland Jr, S. P. (1971). Education of the Gifted and Talented-Volume 1: Report to the Congress of the United States by the US Commissioner of Education.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). Springer, New York, NY.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*(4), 453-477.

Mönks, F., (1992). Education of the gifted in Europe: Theoretical and research Issues. Report of the educational research workshop held in Nijmegen, the Netherlands, july 23-26, 1991. Taylor & Francis Inc.

Mönks, F. & Katzko M., (2005). Giftedness and gifted education. En: Sternberg, R.J. & Davidson, J.E. (Eds.). Conceptions of giftedness. Cambridge, MA: Cambridge University Press.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). TIMSS 2019 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center. Recuperado de timssandpirls.bc.edu/timss2019/frameworks.com.

Munadi, S., & Febriyanti, W. D. R. (2020). Design and validation of mathematical literacy instruments for assessment for learning in Indonesia. *Design and Validation of Mathematical Literacy Instruments for Assessment for Learning in Indonesia*, *9*(2), 865-875.

Muñiz Fernández, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo: Revista del Colegio Oficial de Psicólogos*.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30.

Muraki, E., & Muraki, M. (2016). Generalized partial credit model. In *Handbook of Item Response Theory, Volume One* (pp. 155-166). Chapman and Hall/CRC.

Niss, M., & Jensen, T. H. (2002). Kompetencer og matematiklæring, Uddannelsesstyrelsen Temahæfteserie, nr. 18.

OECD (2013). PISA 2012 Assessment and Analytical Framework, OECD Publishing, http://dx.doi.org/10.1787/9789264190511-en.

OECD (2018). PISA 2021 Mathematics Framework (Draft).

PENTA UC (2021a). Habilidades del marco curricular de PENTA UC (documento de uso interno).

PENTA UC (2021b). Instrumento diseñado para evaluar alfabetización matemática (documento de uso interno).

Piechowski, M., 1997. Emotional giftedness: The measure of intrapersonal intelligence. En Colangelo, N. & Davis, G. Handbook of gifted education. Boston: Allyn and Bacon.

Polya, G. (1945), How to solve it, Princeton, Princeton University Press.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhague: The Danish Institute for Educational Research.

Redding, C., & Grissom, J. A. (2021). Do Students in Gifted Programs Perform Better? Linking Gifted Program Participation to Achievement and Nonachievement Outcomes. *Educational Evaluation and Policy Analysis*.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5).

Renzulli, J. S. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan*, 60 (3), 180-184.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

Robitzsch A, Kiefer T, Wu M (2021). *TAM: Test Analysis Modules*. R package version 3.7-16, https://CRAN.R-project.org/package=TAM.

Ryan, R. M. (1995). Psychological needs and the facilitation of integrative processes. *Journal of Personality*, 63, 397-427

Salas, M. (2012). Crecer con Talento: Cómo los padres pueden apoyar el desarrollo de sus hijos/as con talento académico. Santiago: Pontificia Universidad Católica de Chile.

Schraw, G., & Graham, T. (1997). Helping gifted students develop metacognitive awareness. *Roeper Review*, *20*(1), 4-8.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4.

Spearman, C. (1904a). "General intelligence," objectively determined and measured. American Journal of Psychology, 15(2), 201–293.

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.

Stacey, K., & Turner, R. (2015). The evolution and key concepts of the PISA mathematics frameworks. In: *Assessing mathematical literacy* (pp. 5-33). Springer, Cham.

Sternberg & J. E. Davidson (Eds.), Conceptions of giftedness (2nd ed., pp. 20–38). Cambridge, UK: Cambridge University Press.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, *80*(1), 99-103.

Tannenbaum, A.J., 2000. A history of giftedness in school and society. En: Heller, K.A., Mönks, F.J., Sternberg, R.J. & Subotnik, R.F. (Eds.). *International handbook of giftedness and talent* (2da. Edición). Oxford: Pergamon, 23-53.

Terman, L. M. (1925). Genetic studies of genius: Vol. 1. Mental and physical traits of a thousand gifted children. Stanford, CA: Stanford University Press.

Tout, D., & Spithill, J. (2015). The challenges and complexities of writing items to test mathematical literacy. In *Assessing Mathematical Literacy* (pp. 145-171). Springer, Cham.

Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. Computers in human behavior, 72, 577-58.

VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). A curriculum study of gifted-student learning in the language arts. *Gifted Child Quarterly*, *46*(1), 30-44.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). What's a Testlet and Why Do We Need Them? *In Testlet Response Theory and Its Applications*, 44-59. Cambridge University Press.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149.

Willis, G. B. (2015). Analysis of the cognitive interview in questionnaire design. Oxford University Press.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis: Rasch Measurement. MESA Press.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. *Introduction to Rasch measurement*, 1-24.

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics education research journal*, 18(2), 93-113.

Wu, M., Tam, H. P., & Jen, T. H. (2016). Educational measurement for applied researchers. *Theory into practice*.

Wu, M. L., Tam, H. P., & Jen, T. H. (2016). Partial credit model. En: Wu, M. L., Tam, H. P., Jen, T. H. (Eds.), *Educational measurement for applied researchers* (pp. 159–185). Springer.

Zubairi, A. M., & Kassim, N. A. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, *2*(1), 1-20.

ANEXO 1: Tabla de especificaciones del instrumento

	Código		Tipo de respuesta		Puntaje	Niveles	
Parte 1	P1_P1L	Ítem 1	Cerrada	Respuesta extensa	1	2	15
	P1_P2.1L	Ítem 2	Abierta	Respuesta corta	2	3	
	P1_P2.2L	Ítem 3	Cerrada	Respuesta extensa	1	2	
	P1_P3L	Ítem 4	Abierta	Respuesta corta	3	4	
	P1_P4L	Ítem 5	Abierta	Respuesta corta	3	4	
	P1_P5L	Ítem 6	Cerrada	Respuesta extensa	1	2	
	P1_P6.1L	Ítem 7	Cerrada	Respuesta extensa	3	4	
	P1_P6.2L	Ítem 8	Abierta	Respuesta corta	3	4	
Parte 2	P2_P1.1L	Ítem 9	Abierta	Respuesta corta	1	2	22
	P2_P1.2L	Ítem 10	Abierta	Respuesta corta	1	2	
	P2_P1.3L	Ítem 11	Abierta	Respuesta corta	1	2	
	P2_P2L	Ítem 12	Abierta	Respuesta extensa	3	4	
	P2_P3L	Ítem 13	Abierta	Respuesta extensa	3	4	
	P2_P4L	Ítem 14	Abierta	Respuesta extensa	2	3	
	P2_P5L	Ítem 15	Abierta	Respuesta extensa	3	4	
	P2_P6.1L	Ítem 16	Abierta	Respuesta corta	1	2	
	P2_P6.2L	Ítem 17	Abierta	Respuesta corta	1	2	
	P2_P6.3L	Ítem 18	Abierta	Respuesta corta	1	2	
	P2_P6.4L	Ítem 19	Abierta	Respuesta corta	1	2	
	P2_P7.1L	Ítem 20	Abierta	Respuesta corta	1	2	
	P2_P7.2L	Ítem 21	Abierta	Respuesta corta	1	2	
	P2_P7.3L	Ítem 22	Abierta	Respuesta corta	1	2	
	P2_P7.4L	Ítem 23	Abierta	Respuesta corta	1	2	
Parte 3	P3_P1L	Ítem 24	Abierta	Respuesta extensa	1	2	17
	P3_P2L	Ítem 25	Abierta	Respuesta extensa	1	2	
	P3_P3L	Ítem 26	Abierta	Respuesta extensa	2	3	
	P3_P4L	Ítem 27	Abierta	Respuesta extensa	2	3	
	P3_P5L	Ítem 28	Abierta	Respuesta extensa	2	3	
	P3_P6L	Ítem 29	Abierta	Respuesta extensa	2	3	
	P3_P7L	Ítem 30	Abierta	Respuesta extensa	2	3	
	P3_P8L	Ítem 31	Abierta	Respuesta extensa	3	2	
	P3_P9L	Ítem 32	Abierta	Respuesta extensa	2	3	
							56

ANEXO 2: Índices de Kappa para el instrumento

Parte	Código	Kappa nivel	Kappa código
Parte 1	Ítem 2	0,98	0,9
	Ítem 4	0,85	0,76
	Ítem 5	0,92	0,91
	Ítem 7	0,87	0,88
	Ítem 8	0,89	0,9
Parte 2	Ítem 12	0,9	0,87
	Ítem 13	0,97	0,93
	Ítem 14	0,9	0,87
	Ítem 15	0,91	0,88
Parte 3	Ítem 24	1	1
	Ítem 25	1	1
	Ítem 26	0,97	0,93
	Ítem 27	1	0,95
	Ítem 28	0,97	0,91
	Ítem 29	0,98	0,91
	Ítem 30	1	0,94
	Ítem 31	0,93	0,9
	Ítem 32	1	0,91