Centro UC
Medición - MIDE

# INFORME TÉCNICO SEPA

## SISTEMA DE EVALUACIÓN DE PROGRESO DEL APRENDIZAJE


**Editores: Jorge Manzi, María Rosa García y María Inés Godoy.**

Santiago, 13 de marzo de 2017

# ÍNDICE

# INTRODUCCIÓN

**Jorge Manzi**

Psicólogo de la Pontificia Universidad Católica de Chile y doctor en Psicología de la Universidad de California, Los Ángeles (E.E.U.U). Actualmente se desempeña como profesor titular de la Escuela de Psicología de la Universidad Católica de Chile y es director del Centro de Medición MIDE UC.

**María Rosa García**

Psicóloga y magíster en Psicología Educacional de la Pontificia Universidad Católica de Chile. Actualmente es profesora asistente adjunta de la Escuela de Psicología de esta misma universidad, y profesional del Centro de Medición MIDE UC.

Desde el año 2007, el Centro de Medición MIDE UC de la Pontificia Universidad Católica de Chile desarrolla un **Sistema de Evaluación de Progreso del Aprendizaje Escolar**, mediante las Pruebas **SEPA.** Su objetivo es entregar a los profesores, equipos directivos y sostenedores de establecimientos escolares, información clara, confiable y oportuna, que sea útil para la toma de decisiones pedagógicas y de gestión. En último término, se espera que esta información permita el mejoramiento de los aprendizajes en el aula.

SEPA evalúa los aprendizajes de los estudiantes a través de pruebas de Lenguaje y Matemática desde 1º Básico hasta 3º Medio. Luego de la primera aplicación de las Pruebas SEPA en un establecimiento educativo, los usuarios reciben información acerca del **Estado** del aprendizaje de los estudiantes por curso, nivel, asignatura y eje temático. A partir de la segunda aplicación, cada año se entregan resultados acerca de cómo estos Estados evolucionan en el tiempo, es decir, del **Progreso** experimentado entre la primera y la segunda medición. Asimismo, se reporta el **Valor Agregado (VA)** del establecimiento, que refleja el aporte que realiza el establecimiento para lograr estos aprendizajes.

Las Pruebas SEPA se diseñan basándose en el Marco Curricular vigente y sus preguntas evalúan los distintos Objetivos de Aprendizaje presentes en los programas y bases curriculares publicados por el Ministerio de Educación. Son aplicadas en noviembre o marzo por los docentes de cada establecimiento a partir de un instructivo elaborado por MIDE UC. La modalidad de evaluación consiste en pruebas escritas de preguntas cerradas (opción múltiple), que son procesadas (corregidas y analizadas) por equipos técnicos del Centro de Medición, el cual aporta capacitación y apoyo antes y durante el proceso, para resguardar que se desarrolle con altos estándares de rigurosidad y calidad.

Para monitorear y medir de manera apropiada el Progreso de los alumnos a través de las pruebas, se requiere hacer comparables sus puntajes de un año con el siguiente, permitiendo ubicar los resultados de cada nivel en una escala común de puntajes (una Escala Vertical). Esto quiere decir que SEPA está construido y diseñado para informar certeramente acerca de cómo evolucionan los aprendizajes de un estudiante o grupo de estudiantes (curso o nivel) a lo largo de su trayectoria escolar (Progreso comparable entre años y entre niveles).

Los resultados se reportan confidencialmente a través de una plataforma web, a la que cada establecimiento educativo accede mediante una clave. Este sistema permite entregar

reportes personalizados con información detallada y comparativa para sostenedores, directivos, profesores, estudiantes y apoderados. Es posible seleccionar y filtrar los datos por distintos criterios y descargar tablas y gráficos de sus resultados en diferentes formatos para su posterior análisis.

En sus casi 10 años de existencia, son más de 700 establecimientos escolares de casi todas las regiones del país, los que han utilizado las pruebas SEPA. Estos incluyen establecimientos de todas las dependencias, incluidas corporaciones municipales, fundaciones, redes de colegios y establecimientos individuales.

Las pruebas SEPA esperan complementar los sistemas de evaluación internos de los establecimientos, brindando información confiable y válida. Esto le permite a los establecimientos usar los resultados con fines formativos, es decir, que a través de un proceso de reflexión y análisis de la información, se retroalimenten las prácticas pedagógicas y se tomen decisiones que potencien los procesos de enseñanza y aprendizaje al interior del establecimiento.

A su vez, se espera que los equipos directivos y sostenedores, puedan tener una visión amplia, detallada y comparable de la realidad de su establecimiento, para orientar de la mejor manera posible el uso de recursos y estrategias que permitan en último término, potenciar el aprendizaje de sus estudiantes. Asimismo, busca que los profesores puedan contar con información detallada que les permita tomar decisiones pedagógicas a nivel de aula, como por ejemplo, modificación de planificaciones y estrategias empleadas en cada curso de acuerdo al perfil de rendimiento de sus alumnos.

En este informe técnico se ha tratado de favorecer que cada lector pueda seleccionar aquellos capítulos que responden mejor a su interés, que permite ser revisado como un documento integrado, o bien escoger las secciones que sean de interés del lector. Si bien se ha buscado dar a conocer la información de este informe de la manera más sencilla posible, es necesario advertir que los Capítulo 3 (Análisis de datos SEPA) y 5 (Evidencia sobre la validez de las pruebas SEPA) tienen conceptos propios de la Teoría de Medición, por ende, es necesario estar familiarizado con esta para comprender la información que se presenta. No obstante, si el lector tiene alguna duda sobre los términos empleados en estas páginas, puede contactarse con

SEPA[1] para recibir la orientación necesaria. A continuación, una breve síntesis de lo que contiene cada capítulo.

En el **Capítulo 1**, **Diseño y construcción de las pruebas**, se informa el riguroso proceso de construcción de las preguntas que son utilizadas en las pruebas. Este proceso se basa en los estándares internacionales para la medición psicológica y educacional (AERA, APA & NCME, 2014). Se detalla cada una de las etapas del proceso de construcción de las pruebas, comenzando por el análisis del marco curricular, la elaboración de especificaciones, la construcción de ítems, revisión de expertos, aplicación de pruebas piloto, análisis psicométrico y ensamblaje de las pruebas definitivas.

En seguida, en el **Capítulo 2**, **Logística y aplicación de las pruebas**, se describe el proceso operativo de aplicación de las pruebas SEPA, el cual es de vital importancia para asegurar la calidad y estandarización de la medición. En particular, se describen las distintas etapas que sigue el proceso operacional, comenzando por la definición del calendario de aplicación, el levantamiento de los datos de los usuarios, la impresión y mecanización de las pruebas, la aplicación de estas, la devolución del material a MIDE UC y, finalmente, el procesamiento de los datos.

El complejo proceso de análisis estadístico se describe en el **Capítulo 3**, **Análisis de datos**, el cual señala los procedimientos de análisis orientados a determinar las propiedades métricas de la prueba SEPA en los dos subsectores y los once niveles medidos. Además, se especifica cómo se obtienen técnicamente los resultados, las diferencias significativas y el modelamiento del Valor Agregado.

El **Capítulo 4**, **Resultados**, describe la escala vertical de puntajes de las pruebas SEPA 2015, junto a las características de la población, y los resultados de Estado, Progreso y Valor Agregado que obtienen los establecimientos evaluados en 2015, en los dos sectores de aprendizaje y en los once niveles medidos, realizando comparaciones pertinentes.

Finalmente, el **Capítulo 5**, **Evidencia sobre la validez de las pruebas SEPA**, presenta un conjunto de análisis de tipo psicométrico desarrollados con el objetivo de aportar evidencia acerca de calidad psicométrica de las pruebas SEPA. Específicamente, se muestra evidencia de la precisión o confiabilidad de las pruebas, de su validez y finalmente, de su justicia.

[1] El correo de contacto es sepa@sepauc.cl

# CAPÍTULO I: DISEÑO Y CONSTRUCCIÓN DE LAS PRUEBAS

**Constanza Iglesias**

Psicopedagoga y Profesora de Educación Básica de la Universidad Andrés Bello, magíster en Innovación Curricular y Evaluación Educativa. Actualmente se desempeña como Coordinadora de Instrumentos del Sistema de Evaluación de Progreso del Aprendizaje (SEPA) de MIDE UC. (ciglesiasl@uc.cl)

**María Inés Godoy**

Ingeniera Estadística de la Universidad de Santiago de Chile, Magíster en Estadística y Doctor en Estadística de la Pontificia Universidad Católica de Chile. Actualmente es investigadora asociada a MIDE UC. (migodoy1@uc.cl)

Este capítulo explica el proceso de diseño y construcción de las pruebas SEPA, que se resume en siete etapas que se muestran a continuación (ver Figura 1).

**Figura 1: Flujo de construcción de las pruebas SEPA**


Las siguientes secciones explican las etapas de construcción de las pruebas SEPA.

## Análisis del marco curricular vigente

Cada año, la elaboración de las Pruebas SEPA comienza mediante la revisión y análisis del marco curricular vigente de cada uno de los niveles y asignaturas evaluadas por SEPA desde 1$^{er}$ año de Educación Básica a 3$^{er}$ año de Educación Media. La revisión curricular involucra el análisis de diferentes documentos elaborados por el Ministerio de Educación, que definen y operacionalizan la propuesta curricular vigente. En particular, se analizan los Objetivos de aprendizaje que se plantean como obligatorios para todos los establecimientos educacionales del país.

Para el 2015, el análisis del marco curricular vigente contempló la revisión de los siguientes documentos:

- Bases curriculares aprobadas de 1° a 6° Básico en el año 2012. (http://www.curriculumenlineamineduc.cl/605/articles-30013_recurso_14.pdf)

- Bases curriculares 2013 y Ajuste curricular 2009 para los niveles de 7° y 8° Básico. (BC2013: http://www.curriculumenlineamineduc.cl/605/w3-article-30013.html y MC2009: http://www.curriculumenlineamineduc.cl/605/articles-34641_bases.pdf)
- Ajuste curricular 2009 para los niveles entre 1° y 3° Medio. (http://www.curriculumenlineamineduc.cl/605/articles-34641_bases.pdf)
- Programas de Estudio 2011. (http://www.curriculumenlineamineduc.cl/605/w3-article-30013.html)
- Niveles de Logro SIMCE (http://www.agenciaeducacion.cl/biblioteca-digital/niveles-de-logro/)
- Textos Escolares MINEDUC 2014 y 2015.
- Marco de evaluación de Lectura y Matemática de prueba internacional PISA 2009 y 2012. (http://www.agenciaeducacion.cl/estudios-e-investigaciones/estudios-internacionales/pisa-programme-for-international-student-assessment/)

El análisis curricular es la primera etapa, que sirve de base para la posterior elaboración de las Tablas de especificaciones, que delimitarán los contenidos, habilidades e indicadores de evaluación específicos que serán evaluados, tal como se describe en la siguiente sección. Este análisis permite especificar los ejes que serán evaluados en cada asignatura (descritos más adelante), como también los mapas de contenidos que serán evaluados en las pruebas de cada nivel.

Ahora bien, es importante advertir que el currículum está sujeto a modificaciones, lo cual incide en que hay años de transición en que hay un currículum que está en retirada mientras otro se está instalando. Esto tiene diversas implicancias. Una de ellas, es que los docentes requieren familiarizarse con los ajustes al currículum, y ajustar sus prácticas de acuerdo a eso, lo cual suele tomar un tiempo desde que se realiza la aprobación desde el nivel central. Asimismo, las evaluaciones estandarizadas como SIMCE, SEPA u otras, requieren ajustar lo que es medido en las pruebas, para que se corresponda con el marco curricular vigente.

Para hacer frente a estos cambios, SEPA ajusta sus pruebas en forma gradual al nuevo currículum, incorporando contenidos compartidos entre ambos currículum, para luego de un par de años dar paso al nuevo. Implementar los cambios de forma paulatina es fundamental para, por una parte, medir aquello que está siendo enseñado en las aulas por los profesores y,

también, favorecer la comparabilidad de las pruebas entre años y niveles. Esto último es de suma importancia en un sistema como SEPA, que mide el progreso en los aprendizajes de los estudiantes entre una medición y otra.

A continuación, se describen los ejes temáticos considerados en las pruebas de Lenguaje y Matemática.

**Descripción de Ejes de Lenguaje**

En Lenguaje, los ítems que componen las pruebas SEPA se clasifican en torno a cinco ejes temáticos: i) Información explícita, ii) Información implícita, iii) Sentido global, iv) Situación comunicativa y elementos estructurales y v) Conocimientos y recursos del lenguaje.

A continuación, se presenta una descripción general de los ejes que componen las Pruebas SEPA de Lenguaje, de 1° Básico a 3° Medio.

a) **Información explícita**: Localización y extracción de información de los textos. Las preguntas asociadas a este eje requieren que el estudiante encuentre una o varias informaciones explícitas en el texto y/o sus relaciones. Esta información puede ser fácilmente distinguible o más compleja de encontrar y, además, puede ubicarse en una o más partes del texto.
b) **Información implícita:** Realización de inferencias e interpretaciones de un texto o de fragmento de textos. Las preguntas asociadas a este eje requieren que el estudiante obtenga información que se deduce a partir de información explícita ubicada en una o más partes del texto. Asimismo, algunos ítems requieren la comprensión del significado de una palabra o expresión, a partir de claves contextuales presentes en los textos.
c) **Sentido global:** Comprensión global de los textos. Las preguntas asociadas a este eje requieren que el estudiante comprenda todo el texto y establezca relaciones entre sus distintas partes, pudiendo así determinar, por ejemplo, el tema, la idea principal, la enseñanza o la conclusión desarrollada en textos literarios o no literarios.
d) **Situación comunicativa y elementos estructurales**: Análisis de los elementos que componen la situación comunicativa y de los elementos propios de los diferentes tipos de textos. Las preguntas asociadas a este eje requieren que el estudiante distinga los componentes de la situación comunicativa inherentes a los distintos tipos de textos literarios

y no literarios, y además elementos estructurales que los caracterizan y permiten diferenciarlos.

e) **Conocimientos y recursos del lenguaje**: Conocimiento de las características de nuestra lengua en función de la comprensión. Las preguntas asociadas a este eje evalúan el dominio de las reglas de ortografía puntual; las nociones elementales de corrección idiomática y el conocimiento de las estructuras gramaticales propias de nuestra lengua y funciones en el lenguaje, así como procedimientos de cohesión textual, manejo de conectores y sustitución de términos.

**Descripción de Ejes de Matemática**

En Matemática, las preguntas de las pruebas SEPA se estructuran en torno a cuatro ejes temáticos, i) Números, ii) Álgebra, iii) Geometría y iv) Estadística y probabilidad. A continuación, se presenta una descripción general de los contenidos que cubre cada uno de los ejes en la asignatura de Matemática, de 1° Básico a 3º Medio.

a) **Números:** Este eje considera el estudio de los distintos conjuntos numéricos. En los primeros niveles, se enfoca en los números naturales y sus operaciones básicas, posteriormente, se estudian los números racionales positivos, los números enteros y, después, los números racionales positivos y negativos. En los últimos niveles, el conjunto numérico se extiende a los números reales y, finalmente, al conjunto de los números complejos. Cada vez que se amplía el conjunto numérico estudiado, se extienden las operaciones básicas y las propiedades de orden. Este eje también incluye las nociones de razón, porcentaje y proporción. Asimismo, considera el trabajo con potencias, raíces y logaritmos. Por su relación con las operaciones básicas de números naturales y racionales positivos, también son parte del eje Números en los primeros niveles, las unidades de medida de tiempo, peso y longitud; las conversiones entre dichas unidades; las secuencias repetitivas, numéricas y no numéricas; y la identificación de reglas de formación en dichas secuencias.

b) **Álgebra:** Este eje está presente a partir de 6° Básico y considera el estudio de ecuaciones: ecuaciones simples de primer grado en los niveles básicos; ecuaciones literales de primer grado iniciando la Enseñanza Media; sistemas de ecuaciones lineales y

ecuaciones de segundo grado en los últimos niveles. También es central en el eje Álgebra el desarrollo del lenguaje algebraico y, en Enseñanza Media, el estudio inicial de funciones reales.

c) **Geometría:** En los niveles iniciales, este eje se enfoca en el estudio de figuras y cuerpos geométricos, en el desarrollo del lenguaje geométrico, de la imaginación espacial y el reconocimiento de representaciones planas de cuerpos geométricos. Luego, paulatinamente, incluye el trabajo con ángulos en polígonos y entre rectas que se intersectan, el estudio de los elementos principales y secundarios de los triángulos y el Teorema de Pitágoras, las definiciones como lugares geométricos de la circunferencia y el círculo y los conceptos de congruencia y semejanza de figuras planas. Este eje también considera el trabajo de mediciones geométricas, comenzando con el conocimiento y uso de unidades no estandarizadas, para luego pasar a las unidades estandarizadas que permiten medir longitudes, superficies y volúmenes. Por último, el eje incluye el estudio de transformaciones isométricas y homotecias de figuras planas y el desarrollo de elementos iniciales de la geometría analítica.

d) **Estadística y probabilidad:** En este eje se considera el estudio de la recolección y organización de datos para obtener información estadística en variados contextos (tablas, pictogramas, diagramas de puntos, entre otros), y de la interpretación de la información representada. Incluye el análisis de datos estadísticos, agrupados y no agrupados en intervalos, mediante medidas de tendencia central, de posición y de dispersión y el estudio de los conceptos de población, muestras representativas y muestreo aleatorio simple. También involucra el desarrollo del lenguaje de las probabilidades, desde la identificación de eventos posibles e imposibles en juegos aleatorios simples; la relación entre la frecuencia relativa y la probabilidad de ocurrencia de eventos; la introducción de la Ley de los Grandes Números; el cálculo de probabilidades mediante el modelo de Laplace, usando distintas técnicas de conteo y operaciones conjuntistas; hasta el trabajo con probabilidades condicionadas. Por último, uniendo ambas líneas de estudio, este eje aborda el estudio de los conceptos de variable aleatoria discreta, función de probabilidad y función de distribución acumulada, el cálculo del valor esperado, la varianza y la desviación estándar de variables aleatorias discretas y

el trabajo con distribuciones de variables aleatorias, en particular, con la distribución binomial.

**Elaboración de especificaciones**

Para proseguir la construcción de las pruebas, el siguiente paso es elaborar las Tablas de Especificaciones para cada uno de los niveles y asignaturas evaluadas. Estas tablas se construyen en base al análisis curricular descrito, atendiendo a dos propiedades de los ítems: i) los dominios o contenidos que serán evaluados, los que se estructuran en ejes temáticos, y ii) las habilidades cognitivas que se requieren para resolver las tareas.

Así, las Tablas de Especificaciones definen la cantidad de preguntas que se asociarán a cada uno de los ejes, procurando representarlos de acuerdo a la importancia relativa que tienen tales contenidos en el documento curricular vigente (bases curriculares o ajuste curricular).

Cada Tabla de Especificaciones define ejes, subejes, objetivos de aprendizaje o contenidos mínimos[2], contenidos temarios, habilidades e indicadores de evaluación. Para su elaboración, se requiere que expertos curriculares analicen el currículo nacional vigente junto a las especificaciones del año anterior, de forma de realizar las actualizaciones que se requieran, para ajustarse a los cambios que pudiera experimentar el currículum. Este experto, vela por la coherencia entre los distintos aspectos definidos en la Tabla de Especificaciones, y también por la cobertura del currículum, la cual debe expresarse en los indicadores de evaluación planteados. Asimismo, evalúa la relevancia y pertinencia de los contenidos que se declaran, y define la ponderación o cantidad de preguntas por eje y subeje que se incluirán en las pruebas.

A continuación, las Tablas 1 y 2 muestran los ejes temáticos considerados en Lenguaje y Matemática para cada nivel, junto con la cantidad de preguntas asociadas a cada uno.

[2] Esto depende del documento curricular vigente en que se base. Para este año (2015), los objetivos de aprendizaje corresponden a los que son definidos en las bases curriculares de 1° a 6° Básico, y los contenidos mínimos obligatorios corresponden a los definidos en el ajuste curricular de 7° Básico a 3° Medio.

**Tabla 1: Distribución de preguntas por eje temático en pruebas SEPA Lenguaje 2015**

| Eje/nivel | 1° B | 2° B | 3° B | 4° B | 5° B | 6° B | 7° B | 8° B | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Información explícita de los textos** | 7 | 8 | 8 | 9 | 9 | 9 | 6 | 6 | 6 | 6 | 8 |
| **Información implícita de los textos** | 9 | 9 | 10 | 12 | 11 | 10 | 16 | 18 | 11 | 11 | 11 |
| **Sentido global de los textos** | 5 | 5 | 6 | 7 | 7 | 7 | 12 | 12 | 13 | 16 | 13 |
| **Situación comunicativa y elementos estructurales de los textos** | 4 | 4 | 6 | 7 | 8 | 8 | 10 | 8 | 15 | 11 | 13 |
| **Conocimiento y recursos del lenguaje** | 0 | 4 | 5 | 5 | 5 | 6 | 6 | 6 | 5 | 6 | 5 |
| **TOTAL de preguntas** | **25** | **30** | **35** | **40** | **40** | **40** | **50** | **50** | **50** | **50** | **50** |

**Tabla 2: Distribución de preguntas por eje temático en pruebas SEPA Matemática 2015**

| Eje/nivel | 1° B | 2° B | 3° B | 4° B | 5° B | 6° B | 7° B | 8° B | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Números** | 15 | 16 | 19 | 22 | 21 | 15 | 19 | 11 | 11 | 13 | 13 |
| **Geometría** | 5 | 8 | 10 | 12 | 11 | 12 | 11 | 16 | 14 | 12 | 12 |
| **Álgebra** | 0 | 0 | 0 | 0 | 0 | 5 | 10 | 10 | 15 | 15 | 15 |
| **Estadística y Probabilidad** | 5 | 6 | 6 | 6 | 8 | 8 | 10 | 13 | 10 | 10 | 10 |
| **TOTAL de preguntas** | **25** | **30** | **35** | **40** | **40** | **40** | **50** | **50** | **50** | **50** | **50** |

**Construcción de ítems**

Las pruebas SEPA están formadas por ítems de selección múltiple, como es usual en pruebas estandarizadas. En ellos el estudiante debe seleccionar la opción que contiene o representa la respuesta correcta. En los niveles iniciales, los ítems presentan tres opciones de respuesta, y desde 3° Básico se incorporan preguntas con cuatro opciones de respuesta (ver Tabla 3).

Para la construcción de los ítems de las pruebas, SEPA conforma comisiones técnicas, las cuales son integradas por el encargado de prueba y el analista de apoyo, quienes son parte del equipo profesional interno de SEPA, junto a tres constructores de preguntas que son contratados para esta tarea, todos ellos docentes con experiencia en aula en el nivel y sector para el cual construyen preguntas. Además, es deseable que los constructores cuenten con experiencia previa en procesos de construcción de ítems para mediciones estandarizadas, y con formación en evaluación.

Para cerciorarse de que todos conozcan los criterios técnicos y los lineamientos de construcción de preguntas que han sido definidos por el equipo SEPA, los constructores asisten

a una capacitación inicial, en la cual se comparten los criterios técnicos de construcción de ítems (asociados al contexto, enunciado, opción de respuesta correcta, plausibilidad de los distractores y justificación de los mismos) y procedimientos involucrados en el proceso, como el uso de la plataforma para la entrega de preguntas, entre otros.

El proceso de construcción de ítems comienza cuando los especialistas construyen sus preguntas de acuerdo a los requerimientos indicados por el equipo SEPA, los criterios considerados en la capacitación, y las Tablas de Especificaciones. Una vez que entregan los primeros ítems, estos son revisados, en primera instancia, por el encargado de prueba y el analista de apoyo, quienes incluyen comentarios y sugerencias de mejora. Posteriormente, los ítems son revisados en reuniones semanales por la comisión técnica, instancia en la cual se analizan conjuntamente y se decide si se aprueban, modifican o rechazan. El ítem se aprueba cuando cumple con todos los criterios acordados, y no se requiere realizar mejora alguna. Se modifica cuando requiere cambios menores que permiten que cumpla con los criterios acordados, sin transformarlo completamente respecto del original. Finalmente, el ítem se rechaza cuando la formulación presenta serias dificultades, que no son factibles de corregir mediante cambios menores a este, sino que requeriría ser construido nuevamente en su totalidad o gran parte de esta.

Los criterios que se contemplan en las revisiones de ítems dicen relación con las dimensiones de:

- Contenido: los ítems deben abordar contenidos propios de la disciplina de forma rigurosa y precisa, junto con aspectos centrales y relevantes de ella, de acuerdo al análisis del marco curricular vigente.
- Contexto: el texto o situación que sirve de estímulo debe ser relevante, claro, ser necesario y suficiente para responder la pregunta, y no presentar sesgos de género, raza, entre otros.
- Enunciado: la pregunta propiamente tal debe estar claramente dirigida, no debe entregar pistas para la respuesta correcta, y tiene que ser precisa.
- Respuesta correcta: debe ser la única o mejor respuesta, resultar coherente con el enunciado, ser clara, precisa, y tener una extensión adecuada.

- Opciones: las posibilidades de respuesta deben ser plausibles, mantener un mismo nivel lógico, tener una extensión adecuada, ser claras, no traslaparse entre sí ni presentar opciones subsumidas, entre otras características.
- Coherencia indicador-pregunta: se revisa que exista coherencia entre el indicador de evaluación y lo que es medido en la pregunta.
- Coherencia habilidad-pregunta: se revisa que exista coherencia entre la habilidad definida en el ítem y lo que es medido a través de la pregunta.

A lo largo del proceso de construcción, se monitorea la construcción de ítems para asegurar que cumplan con los estándares y criterios mencionados, asegurando la entrega de retroalimentación pertinente y precisa cuando se requiera. La Tabla 3 muestra el número de preguntas de cada prueba y la distribución del número de opciones que tiene cada una de las preguntas por nivel escolar.

**Tabla 3: Número de opciones de respuestas y de preguntas por nivel**

| Nivel | 1°B | 2°B | 3°B | 4°B | 5°B | 6°B | 7°B | 8°B | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Número de opciones de respuesta** | 3 | 3 | 3 - 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| **Total de preguntas** | 25 | 30 | 35 | 40 | 40 | 40 | 50 | 50 | 50 | 50 | 50 |

**Revisión de expertos**

Finalizada la etapa de construcción de ítems, y para potenciar la evidencia de validez de contenido de ellos, un experto en la asignatura y con vasta experiencia en evaluación educacional, evalúa los ítems seleccionados.

El experto también podrá aprobar, sugerir modificaciones o rechazar los ítems, de acuerdo a los siguientes criterios:

- Aprobada: la pregunta no requiere ningún cambio importante en el contexto, enunciado ni en las opciones de respuesta. Asimismo, la pregunta corresponde al contenido e indicador declarado, y es coherente con el nivel de habilidad declarado.
- Aprobada con modificaciones: la pregunta requiere algunos cambios menores y/o precisiones en el contexto, enunciado y/u opciones de respuesta, o bien presenta un desajuste menor en relación al indicador declarado o la habilidad medida. En este caso, se solicita explicar claramente cuál o cuáles son las modificaciones que se deben realizar, e idealmente hacer sugerencias concretas sobre las mismas.
- Rechazada: la pregunta presenta deficiencias mayores, ya sea debido a que resulta irrelevante, no mide apropiadamente los indicadores de evaluación señalados en la Tabla de Especificaciones, siendo muy difícil poder reformularla o modificarla para salvaguardar estas falencias.

Para estas revisiones los expertos deben tener en consideración posibles errores disciplinarios o curriculares de cada ítem, y revisar la correspondencia curricular que se le ha asignado a cada uno de ellos. Asimismo, deben revisar las dimensiones de contenido, contexto, enunciado, opción correcta y opciones, descritas anteriormente.

**Ensamblaje y aplicación de pruebas de campo**

Una prueba de campo es un instrumento que se administra a un grupo más reducido de estudiantes y se compone de ítems recién construidos sobre los cuales se desea obtener información acerca de su comportamiento empírico, de forma de poder seleccionar los mejores para incorporarlos en las pruebas definitivas. De esta manera es posible asegurar estándares de calidad que permitan construir pruebas definitivas con buen comportamiento psicométrico.

El ensamblaje de las pruebas de campo, se realiza con ítems que han sido aprobados en las etapas anteriores del proceso de construcción. Se incorporan ítems de cada eje temático de las pruebas de Lenguaje y Matemática de acuerdo a su representación en el currículo vigente específico de cada nivel y asignatura.

Las pruebas de campo son aplicadas a muestras representativas de aproximadamente 250 estudiantes para cada forma de las pruebas, en cada nivel y asignatura evaluados. Para obtener este tipo de muestras, se busca seleccionar establecimientos de distinto tipo de dependencia de acuerdo a la distribución de la totalidad de pruebas aplicadas en los últimos dos

años. De esta manera, se mantiene una proporción de pruebas en cada dependencia similar a la que se espera en la aplicación de las pruebas definitivas. Actualmente, estas proporciones corresponden a 50% de colegios municipales, 20% de particulares subvencionados y 30% de particulares pagados.

Los colegios pertenecientes a la muestra son invitados a participar a cambio de la entrega de algunos resultados asociados a la aplicación de esta prueba. En particular, estos resultados corresponden al porcentaje de logro de los estudiantes en los distintos ejes evaluados, considerando para este cálculo solo los ítems que hayan sido aceptados según los criterios que serán detallados en la siguiente sección.

El proceso de aplicación de las pruebas de campo será abordado en el Capítulo dos de este informe.

**Análisis psicométrico de las pruebas de campo**

A partir de los datos obtenidos del estudio de campo, el equipo procede a analizar las propiedades métricas de los ítems, usando distintos indicadores con el objetivo de seleccionar aquellos ítems con mejor comportamiento y a la vez, descartar aquellos que no muestran un funcionamiento adecuado. Para este análisis psicométrico se consideran los siguientes indicadores:

1. **Grado de dificultad:** este índice mide la dificultad de un ítem, en una escala de 0% a 100% y corresponde a la proporción de individuos que responde correctamente un ítem. De esta manera, mientras más alto es el valor, menor es la dificultad de un ítem. El grado de dificultad de los ítems utilizados en las pruebas SEPA se encuentra entre 15% a 85%, descartando los ítems que se encuentran en los extremos, es decir, muy fáciles o muy difíciles.
2. **Capacidad discriminativa**: este índice da cuenta de la capacidad de un ítem de distinguir entre estudiantes con mayor o menor nivel de habilidad. Corresponde a la correlación entre responder correctamente un ítem y el puntaje obtenido en el total de la prueba. Se espera que quien tuvo una puntación alta en toda la prueba deberá tener una alta probabilidad de contestar bien el ítem, mientras que los estudiantes que

lograron bajo puntaje en toda la prueba deberán tener baja probabilidad de contestar bien el ítem. Se espera que su valor sea a lo menos de 0,2.

3. **Funcionamiento de distractores:** este índice corresponde a la correlación biserial entre la respuesta a un determinado distractor y el puntaje obtenido en la prueba. Se espera que esta correlación sea nula o negativa, para asegurar que quienes los eligen no tengan un alto desempeño en las pruebas.
4. **Plausibilidad de los distractores:** este indicador refiere a la proporción de individuos que selecciona cada una de las opciones incorrectas de respuesta y se utiliza para asegurar que todas las preguntas tienen alternativas de respuesta plausibles, es decir, que son elegidas por una proporción mínima de individuos. Esta plausibilidad debiera ser mayor a 4%.

5. **Grado de omisión**: este índice corresponde a la proporción de individuos que no responde un ítem, es decir, no marca ninguna preferencia de respuesta. Se descartan ítems con muy alta omisión (mayor al 25%) o donde la omisión pareciera reflejar ambigüedad en el ítem (por ejemplo, cuando omiten por igual estudiantes con alto y bajo desempeño en la prueba).

El análisis de las propiedades métricas de los ítems se realiza en base al comportamiento que muestran los índices mencionados anteriormente, en cada uno de los ítems. A partir de ello, se procede a tipificarlos para distinguir aquellos que poseen características métricas adecuadas, y se seleccionan aquellos que cumplen con las exigencias técnicas para incorporarlos a las pruebas definitivas.

Aquellos que no muestran un funcionamiento adecuado pueden ser nuevamente experimentados en futuras pruebas de campo, si se evalúa que resultan posibles de modificarse. Esto ocurre cuando presentan dificultades en una de sus opciones de respuesta, la que muestra baja elegibilidad (menor al 4%) o bien, una correlación biserial positiva. En estos casos, se modifica el distractor y la pregunta puede ser nuevamente incorporada en una prueba de campo y evaluada en su comportamiento métrico.

**Ensamblaje de pruebas definitivas SEPA**

Para ensamblar las formas de las pruebas definitivas, se utilizan los ítems que han sido aprobados a partir del análisis de sus propiedades psicométricas, junto con otros criterios que se detallan a continuación.

Primeramente, para ensamblar las pruebas definitivas, se seleccionan las preguntas a incluir considerando las cantidades definidas en las Tablas de especificaciones para los distintos ejes y subejes de contenido, y habilidades. Esto permite representar el marco curricular vigente en proporciones adecuadas de acuerdo a la importancia relativa de los distintos dominios de la asignatura.

Asimismo, la prueba se estructura en dos bloques, en los que se incluyen preguntas asociadas a cada eje y subeje, de forma de que cada eje es consultado en dos oportunidades, una en la primera mitad de la prueba, y otra en la segunda mitad de esta.

Sumado a lo anterior, se consideran criterios psicométricos de la prueba completa, que aportan evidencia asociada a la comparabilidad de estas. Entre estos criterios se encuentran: i) asegurar que el grado de dificultad de la prueba sea mediano; ii) asegurar que exista una proporción de preguntas de grado de dificultad bajo, mediano y alto, similar entre las distintas pruebas.

Por otra parte, las pruebas definitivas requieren asegurar comparabilidad entre años, y también, entre pruebas de niveles sucesivos. Esto es especialmente importante en un sistema de medición como SEPA, puesto que los resultados se reportan comparando datos del año anterior y siguiente, y en niveles sucesivos entre un nivel y otro, cuando se reportan datos de Progreso y Valor Agregado. Para que estos datos sean fiables y adecuados, se requiere que las pruebas se encuentren equiparadas.

Una solución convencional para resolver este desafío es utilizar preguntas ancla, es decir, preguntas comunes que se comparten entre las pruebas de un nivel y otro, de forma de realizar análisis de comparabilidad (conocidos como *equating*) entre años y niveles. Las preguntas ancla que se incluyen en cada nivel, corresponden a preguntas propias del año inmediatamente inferior, que se comparten en ambas pruebas.

Para seleccionar las preguntas ancla, se consideran criterios técnicos y psicométricos exigentes. Entre estos, se contempla que muestren propiedades psicométricas adecuadas en todos los indicadores considerados (grado de dificultad, capacidad discriminativa, correlación

biserial distractores, plausibilidad de los distractores, y omisión). Además, se requieren anclas que representen a los distintos ejes, en la proporción que dicho eje tiene en la prueba del nivel. La posición en la prueba debe corresponderse de forma muy cercana con la posición que utiliza en la prueba del año y nivel anterior, evitando estar al final de la prueba. A lo largo del tiempo, estas preguntas anclas se renuevan, de forma que cada cuatro años se reemplazan en forma completa.

El número de preguntas ancla varía de acuerdo al nivel escolar, como se detalla en Tabla 4 para el año 2015.

**Tabla 4: Número de preguntas ancla según nivel escolar**

| Nivel | 1°B | 2°B | 3°B | 4°B | 5°B | 6°B | 7°B | 8°B | 1º M | 2º M | 3º M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Nº de preguntas ancla | - | 10 | 10 | 12 | 12 | 12 | 14 | 14 | 14 | 14 | 14 |
| Nº de preguntas totales | 25 | 30 | 35 | 40 | 40 | 40 | 50 | 50 | 50 | 50 | 50 |

# CAPÍTULO II: LOGÍSTICA Y APLICACIÓN DE LAS PRUEBAS

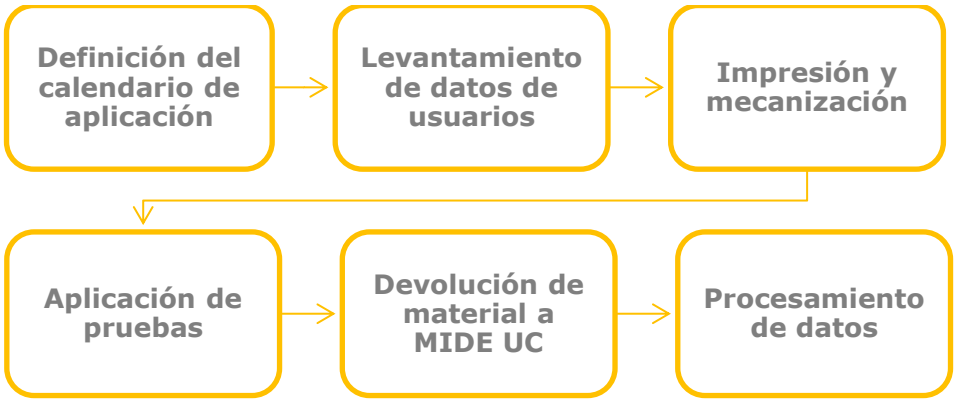
**Francisca Brito**

Psicóloga Organizacional de la Universidad Diego Portales. Actualmente se desempeña como Encargada de Operaciones del Sistema de Evaluación de Progreso del Aprendizaje (SEPA) de MIDE UC. (mbritol@uc.cl)

Los procesos referidos a la aplicación de las Pruebas SEPA son de gran importancia para asegurar la calidad, pues de estos depende que se minimicen posibles fuentes de error en la puntuación de los datos, lo que es de suma relevancia para la validez de los resultados que se informan a los establecimientos educativos.

Este capítulo detalla el proceso operativo de aplicación de las Pruebas SEPA, el cual es resumido en las siguientes seis etapas mostradas en la Figura 1.

**Figura 1: Flujo de logística y aplicación Pruebas SEPA**



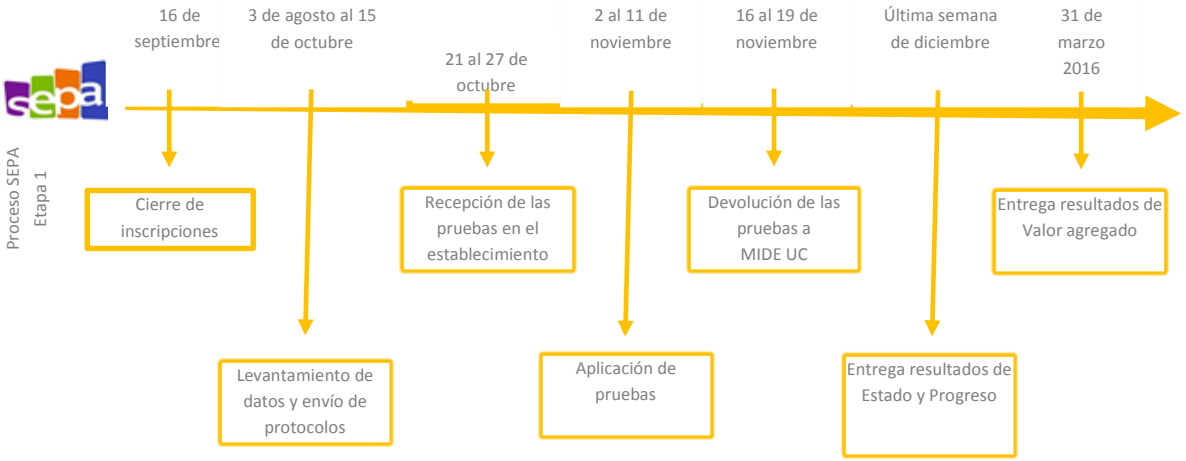

- Definición del calendario de aplicación: es la etapa en la que se establece el calendario asociado a los procesos referidos a la aplicación de las Pruebas SEPA 2015.
- Levantamiento de datos de usuarios: señala el registro de datos de cada establecimiento, curso y cantidad de estudiantes que participarán en SEPA.
- Impresión y mecanizado: describe el proceso de impresión y mecanizado de las pruebas y posterior despacho a los establecimientos.
- Aplicación de pruebas: detalla el proceso de aplicación de pruebas que es realizado localmente en cada establecimiento siguiendo los protocolos e instructivos para garantizar una aplicación estandarizada.
- Devolución de material a MIDE UC: detalla el proceso de despacho de las pruebas desde los establecimientos a MIDE UC para su posterior digitalización.
- Procesamiento de datos: detalla el proceso mediante el cual los datos son capturados desde las hojas de respuesta y registrados en bases de datos, junto a los procedimientos de seguridad y control que se implementan para asegurar la confiabilidad de estos.

## Definición del calendario de aplicación

Las Pruebas SEPA son aplicadas al final de cada año escolar, durante noviembre, y excepcionalmente en algunos establecimientos se aplican al comienzo del siguiente año, en el mes de marzo. Así, cada establecimiento educacional se inscribe para aplicar en uno de estos dos períodos. En ambas instancias el objetivo es realizar la aplicación en un momento del año escolar que permita recabar información sobre los aprendizajes que los alumnos han alcanzado durante todo un año escolar. Así, tanto en la aplicación a fin de año como en la de comienzo del período escolar, se aplican las mismas pruebas de Lenguaje y Matemática para evaluar los aprendizajes del nivel en curso o del nivel anterior respectivamente. Por ejemplo, si se evaluarán los aprendizajes de 6° Básico, se aplican las pruebas de este nivel a los estudiantes al término de 6° año, o bien, al comienzo de 7° grado.

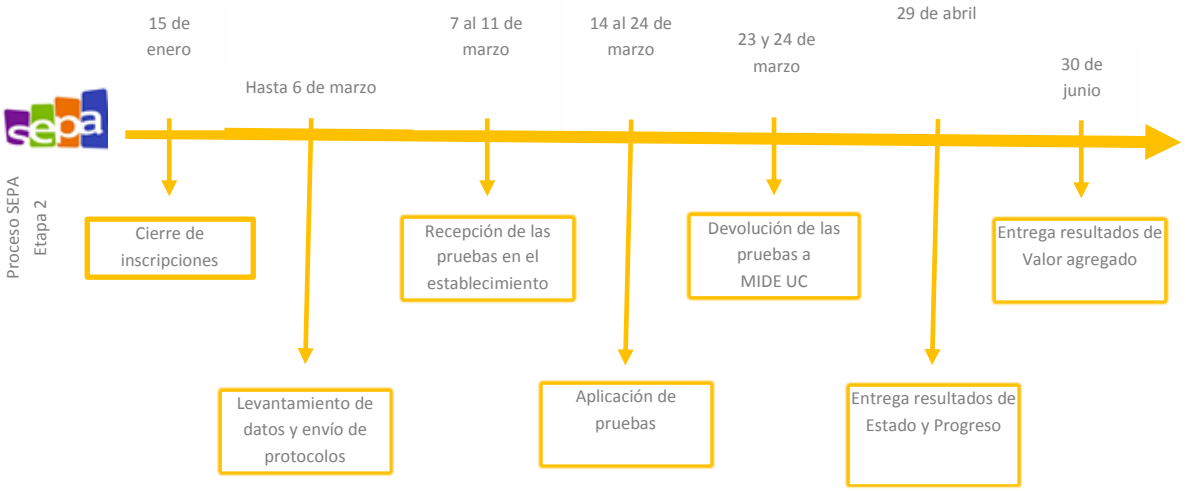
El calendario de aplicación, para ambas etapas, comienza con el cierre de inscripciones y finaliza con la entrega de resultados de las pruebas. Las Figuras 2 y 3 detallan los hitos del proceso de aplicación de las pruebas con sus correspondientes fechas para el año 2015, primera etapa noviembre 2015 (ver Figura 2) y segunda etapa marzo 2016 (ver Figura 3) respectivamente.

**Figura 2: Calendario de aplicación etapa 1 (noviembre 2015)** [3]




[3] Progreso y Valor Agregado es entregado sólo para establecimientos que tengan dos evaluaciones SEPA consecutivas.

**Figura 3: Calendario de aplicación etapa 2 (marzo 2016)**




### Levantamiento de datos de usuarios

Al momento que un establecimiento confirma su participación en SEPA, se le contacta para darle la bienvenida, recordar las fechas que comprende el período de aplicación, y solicitar completar un formulario en línea que recoge información sobre la ubicación del establecimiento, y los datos de contacto del director, del Jefe de la Unidad Técnico Pedagógica (UTP) y de quién asumirá el rol de Colaborador SEPA. Este último es el interlocutor permanente entre el establecimiento y el equipo SEPA, para coordinar el proceso de aplicación dentro del colegio y manejar el material evaluativo.

Durante la aplicación de las pruebas, se solicita a cada Colaborador enviar las listas de estudiantes de los cursos evaluados, para la posterior revisión de los datos de identificación.

Durante todo este período, SEPA se comunica de forma constante con los Colaboradores a través de correos electrónicos y se organiza un Centro de llamados que presta atención constante durante todo el proceso de aplicación.

### Impresión y mecanizado

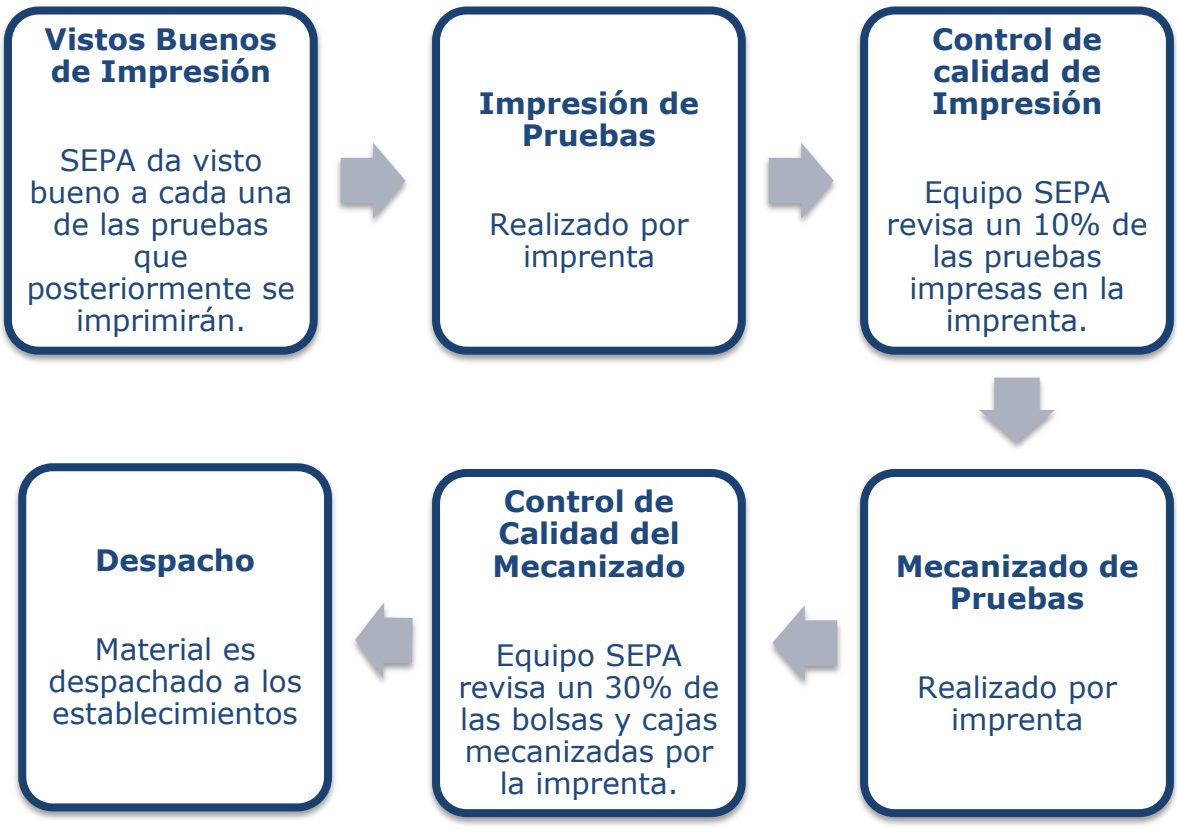
La impresión y mecanizado de las Pruebas SEPA, deben cumplir con ciertos estándares que garanticen que todos los estudiantes que participan en la aplicación reciban pruebas con las mismas características y condiciones, evitando errores que puedan afectar sus resultados, como por ejemplo: errores de compaginación, imágenes muy oscuras que dificulten la compresión de

una pregunta, manchas en las alternativas que puedan incidir en que el estudiante marque una alternativa, etc.

Para comenzar, se selecciona un proveedor de impresión y mecanizado a través de una licitación abierta. Este proveedor es el mismo para las fechas de aplicación. Las propuestas recibidas son evaluadas de acuerdo a una serie de criterios relacionados con la experiencia técnica de las instituciones imprimiendo y mecanizando instrumentos de evaluación de similares características a SEPA. Estos criterios consideran, por ejemplo, la existencia de protocolos de seguridad que garanticen la confidencialidad del instrumento, existencia de controles de calidad que aseguren la impresión y mecanizado según los estándares acordados, instalaciones adecuadas para la realización del proceso, experiencia en grandes volúmenes de impresión, entre otros. De acuerdo a estos criterios se selecciona al proveedor con mejor perfil.

Posteriormente, el proceso de impresión y mecanizado consta de seis etapas cronológicas, las cuales se nombran en el siguiente flujo (Figura 4) y detallan a continuación.

**Figura 4: Proceso de impresión y mecanizado**



a) **Vistos buenos de impresión:** Antes de comenzar la impresión, se envían a la imprenta las pruebas definitivas para visualizar un ejemplar impreso de cada una de ellas. En esta instancia el equipo SEPA verifica que las pruebas correspondan con lo esperado, es decir, que todas las preguntas y respuestas correspondan a lo enviado, que las imágenes sean claras, los signos matemáticos correctos, etc., y se entrega el visto bueno para comenzar la impresión.

b) **Impresión de pruebas:** La imprenta realiza la impresión de las 22 pruebas, cuyas cuartillas varían según el nivel y asignatura. Además, se imprimen las hojas de respuesta que contienen las respuestas de Lenguaje y de Matemática de un mismo estudiante. Tanto las pruebas como las hojas de respuesta se imprimen en color y grosor necesarios para asegurar legibilidad y nitidez de las pruebas.

c) **Control de calidad de impresión**: En esta etapa se revisa un porcentaje de las pruebas impresas, para verificar que la calidad de la impresión sea adecuada, considerando que las imágenes se visualicen claramente, no existan manchas de tinta, que el guillotinado de la prueba sea parejo y que no haya errores de compaginación. Esta revisión se realiza en las distintas cajas de impresión, previo al mecanizado, y en caso de detectarse errores mayores se rechaza la partida de impresión y la imprenta debe volver a imprimir ese nivel-prueba.

d) **Mecanizado de pruebas:** El material de aplicación es mecanizado por la imprenta a partir de la información entregada por SEPA, de acuerdo a los cursos y la cantidad de material correspondiente para cada asignatura, junto a folios únicos de identificación para cada una de las bolsas y sobres. Las pruebas se ingresan en bolsas para cada curso y asignatura evaluada, y las hojas de respuesta se ingresan en sobres para cada curso evaluado.

Tanto las bolsas de pruebas como los sobres de hojas de respuesta son etiquetados para permitir un fácil reconocimiento del material, como puede verse en la Figura 5:

**Figura 5: Ejemplos de etiquetado de las bolsas de pruebas y hojas de respuesta**


Una vez que se preparan las bolsas y sobres para cada curso de un establecimiento, estos materiales son ingresados en cajas. El orden para el embalaje está dado por el nivel y asignatura de la prueba, ingresando las bolsas y sobres de aplicación según este criterio, las bolsas de los cursos ordenados alfabéticamente primero Lenguaje y luego Matemática. Por ejemplo, caja 1 contiene Bolsa 1° A Lenguaje, 1° B Lenguaje, 1° C Lenguaje, 1° A Matemática, y caja 2 tiene bolsas 1° B Matemática, 1° C Matemática, sobres 1° A, 1° B, 1° C.

Cada caja tiene etiquetas en todas sus caras laterales, en que se identifica al establecimiento (ver Figura 6), y en la cara frontal se coloca la etiqueta que detalla el material que contiene en su interior (bolsas, sobres).

**Figura 6: Ejemplo de etiquetado de caja curso**

**a) Etiqueta lateral**


**b) Etiqueta Frontal**


Es importante mencionar que cada bolsa de pruebas mecanizada incluye pruebas extra con respecto a la cantidad de estudiantes que tiene el curso y cada sobre de hojas de respuesta lleva también hojas extra, por posibles inconvenientes.

**e) Control de calidad del mecanizado:** Para garantizar la calidad del material enviado se revisa un porcentaje de las cajas que contiene el pallet, verificando que se haya ingresado en las cajas, las bolsas y sobres especificados en su etiqueta (Figura 6). Luego se revisan todas las bolsas al interior de la caja y se comprueba que la cantidad de pruebas sean las necesarias, y que estas correspondan al nivel y a la asignatura señalada en la etiqueta (Figura 2).

**f) Despacho:** Para distribuir las Pruebas SEPA en las distintas regiones del país en que se encuentran los establecimientos educacionales que participan, se usan dos modalidades de despacho: vía *courier*, y a través de transportista. En ambos casos, se avisa con anticipación mediante correo electrónico a cada establecimiento la fecha en que llegará

el material evaluativo, la modalidad en que este será despachado y la cantidad de cajas que deben recibir. Además, los operadores del Centro de llamados se comunican con cada establecimiento para recordar la llegada del material, de forma de asegurarse de que estén atentos a su recepción.

Posteriormente, los operadores del Centro de llamados vuelven a contactar a cada establecimiento para verificar que hayan recibido correctamente todo el material que corresponde. Es decir, se comprueba que la cantidad de pruebas por curso corresponda a la matrícula que los cursos poseen en ese momento. En caso de que falten pruebas, se evalúa la posibilidad de utilizar el material extra que se dispone para cada curso, y de ser necesario, se prepara nuevo material para los cursos en esa situación y se despacha.

Para dejar registro del despacho del material, se le solicita a cada establecimiento que envíe un correo electrónico confirmando la recepción conforme del material evaluativo solicitado. Adicionalmente a esto, en ambas modalidades de despacho se cuenta con archivos que permiten corroborar que el material fue recepcionado por cada establecimiento (como son las guías de despacho en el caso del transportista, y foto del nombre y firma de la persona que recibe, en el caso del despacho a través de *courier*).

**Aplicación de pruebas**

SEPA envía a los establecimientos instructivos y protocolos que buscan entregar pautas que orienten a los usuarios a realizar una aplicación estandarizada de las pruebas. Se envían los siguientes protocolos:

a. **Introducción "¿Qué es SEPA?"**: este instructivo detalla el programa SEPA, en qué consisten las pruebas y los resultados que los establecimientos recibirán al finalizar el proceso.
b. **Protocolo del Colaborador SEPA**: indica las acciones y orientaciones que debe seguir el Colaborador SEPA[4] para la coordinación y organización del proceso de aplicación dentro de su establecimiento.

[4] Colaborador SEPA: Es el interlocutor permanente entre el establecimiento y el equipo SEPA, para coordinar el proceso de aplicación dentro del establecimiento y manejar el material evaluativo.

c. **Protocolo del Aplicador**: dirigido al profesor que realiza la aplicación, los cuales son docentes que no realizan clases al curso, ni a la asignatura que le corresponde aplicar.

En este protocolo se entrega toda la información necesaria para realizar la aplicación de las pruebas de manera estandarizada. Como el tiempo de aplicación máximo de las pruebas: 60 minutos para 1° y 2° Básico y 70 minutos desde 3° Básico a III medio. A la vez, se les informa el uso de las hojas de respuesta, cómo borrar una respuesta incorrecta, cómo anular una hoja de respuesta en caso de ser necesario y revisar la correcta identificación de los estudiantes en ella**.** Entre las indicaciones para los profesores aplicadores se incluyen:

- No pueden responder ninguna pregunta asociada al contenido de la prueba, solo preguntas relacionadas con el uso de la hoja de respuestas.
- No pueden leer las preguntas en voz alta bajo ninguna circunstancia. En casos de estudiantes con necesidades educativas especiales permanentes (en adelante, NEEP) se les señala que la aplicación se debe realizar de la misma forma que habitualmente son evaluados.
- Se solicita que todo el material de apoyo pedagógico de las salas de clase sea retirado antes de realizar la evaluación.

d. **Principales errores en el uso de la hoja de respuesta**: muestra errores en el uso de la hoja de respuesta que se han encontrado en aplicaciones pasadas, se especifican las consecuencias que este error tiene para el estudiante afectado, y se entrega además información sobre lo que deben hacer para evitarlos y/o solucionarlos.

Además de enviar los protocolos de aplicación, SEPA organiza distintas instancias de capacitación, cuyo objetivo es reforzar el contenido de los protocolos enviados, así como aclarar todas las posibles dudas que los profesionales de los establecimientos y profesores aplicadores puedan tener. A estas instancias, se invita a participar de forma presencial a los establecimientos de la Región Metropolitana y de regiones cercanas y, a través de videoconferencia a los establecimientos más distantes.

Durante los procesos de aplicación de las Pruebas SEPA, se dispone de un Centro de llamados para centralizar las comunicaciones con los establecimientos. Este apoyo tiene como funciones contactar a los establecimientos para verificar la llegada de los protocolos de aplicación, confirmar la recepción correcta del material evaluativo y, además, recibir todo

tipo de consultas desde los establecimientos asociadas a la aplicación de las pruebas, al manejo del material evaluativo y a la devolución de este a MIDE UC.

**Devolución de material a MIDE UC**

Luego de finalizar el proceso de aplicación de las pruebas, se envía a cada establecimiento un comunicado informando la fecha y modalidad (transportista o *courier*) en que el material será retirado para ser despachado a MIDE UC.

Una vez que el material llega a MIDE UC, la recepción y revisión de este, en ambos periodos de aplicación, se realiza en dos grandes etapas:

- En una primera instancia, se revisa que todas las bolsas y sobres enviados a los establecimientos se recepcionen en MIDE UC. En caso de detectar que falta el material de algún curso se contacta inmediatamente al establecimiento para solicitar que envíen este material faltante.
- Posterior a la primera revisión, cada una de las pruebas y hojas de respuestas recibidas son ingresadas al sistema de inventario de SEPA el que permite, mediante el código de barras, identificar la llegada en detalle del material evaluativo.

Adicionalmente a esto, el sistema permite ingresar el estado en que cada una de las hojas de respuesta es recepcionada, dejando un registro de todas aquellas que fueron usadas, es decir, todas las que tienen respuestas completadas.

**Procesamiento de datos**

Para la captura de los datos desde las hojas de respuesta, se trabaja con un proveedor de digitalización con experiencia en el rubro y que cuenta con los controles de seguridad y de manejo de datos necesarios para resguardar tanto la confiabilidad de estos, como la integridad del material entregado para su procesamiento.

Los datos que se capturan desde las hojas de respuesta son los siguientes: Folio, RUT, Nombres, Apellidos, Sexo, Respuestas de Lenguaje, Respuestas de Matemática.

Para garantizar la confiabilidad de los datos, el proveedor llevó a cabo las siguientes acciones:

- Digitación manual de los datos de identificación de los estudiantes: RUT, nombres y apellidos.

- Revisión y digitación manual de todos los casos que tienen doble marca como respuesta. Esto ocurre en ocasiones cuando el sistema considera que hay dos respuestas marcadas cuando en realidad una de ellas está borrada y dejó algo de gris en la hoja. En esos casos, al pasar por un revisor manual, se puede identificar cuál es la respuesta válida. Ahora bien, cuando hay dos marcas igualmente señaladas, se registra como respuesta inválida.

Una vez procesadas las hojas de respuestas por el proveedor, SEPA recibe una base de datos con toda la información capturada, la que es validada para verificar que todas las hojas de respuesta hayan sido digitalizadas, y se aplican criterios de control de calidad.

Posteriormente a la primera validación de la captura, comienza el proceso de validación general de la base de datos, cuyo objetivo es lograr una base de datos depurada para llevar a cabo los análisis psicométricos y producir los resultados que serán comunicados a los establecimientos educacionales. Al contar con estas bases, finaliza el proceso de aplicación de las Pruebas SEPA, y comienza el proceso de análisis de datos que se detalla en el Capítulo 3.

# CAPÍTULO III: ANÁLISIS DE DATOS DE LAS PRUEBAS SEPA

**María Inés Godoy**

Ingeniera Estadística de la Universidad de Santiago de Chile, magíster en Estadística y doctora en Estadística de la Pontificia Universidad Católica de Chile. Actualmente es investigadora asociada a MIDE UC. (migodoy1@uc.cl)

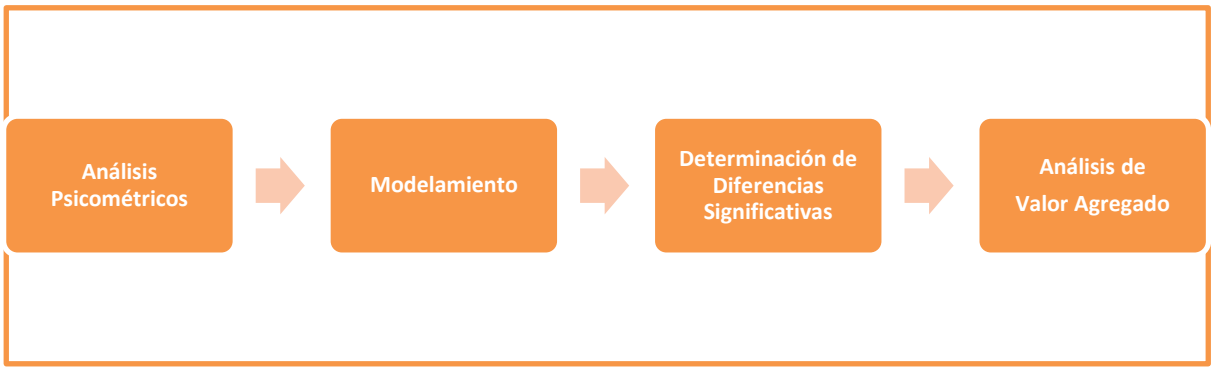
**Andrea Abarzúa**

Psicóloga y magíster en Psicología Educacional de la Pontificia Universidad Católica de Chile. Actualmente se desempeña como Coordinadora de la Unidad de Análisis de MIDE UC, y como docente en el área de formación. (raabarzu@uc.cl)

Este capítulo señala los procedimientos de análisis estadístico orientados a determinar las propiedades métricas de la prueba SEPA en los dos subsectores y los once niveles medidos. Además, se especifica cómo se obtienen técnicamente los resultados, las diferencias significativas y el modelamiento del Valor Agregado. Advertimos al lector que la adecuada comprensión de este capítulo supone familiaridad con los modelos y herramientas psicométricas a pesar de que se ha hecho un esfuerzo por presentar la información de la forma más comprensiva posible.

Los análisis de los datos pueden ser resumidos en el siguiente flujo, donde cada actividad será detallada dentro de las secciones siguientes (ver Figura 1).

**Figura 7: Flujo de análisis de datos**




Para el análisis de mediciones educacionales se requiere utilizar modelos psicométricos, los que permiten cuantificar y obtener estimaciones de los niveles de habilidad de los estudiantes evaluados según el marco de referencia de la medición. Dentro de las teorías utilizadas se encuentran la Teoría Clásica de Test (TCT) y la Teoría de Respuesta al Ítem (TRI). En el caso de las pruebas SEPA se realiza un modelamiento en el marco de la Teoría de Respuesta al Ítem, específicamente un modelo Rasch de un parámetro con regresión latente. Esta clase de modelo da la posibilidad de describir en forma separada a los ítems y a los estudiantes, pues considera que un estudiante responde correctamente un ítem según el grado de dificultad de este y la habilidad que tenga en el área evaluada en un nivel educacional determinado. Además, una de las ventajas de estos modelos es que permiten mediante un proceso de equiparación (equating) la comparabilidad de puntajes a través del tiempo y entre

niveles educacionales, aunque se utilicen pruebas distintas. Esta metodología se encuentra en línea con la mayoría de las pruebas estandarizadas a gran escala, como es SIMCE en el caso nacional y en casos internacionales como PISA, TIMMS, TERCE y PIRLS entre otras.

**Análisis psicométricos**

Este análisis tiene dos grandes objetivos, el primero es asegurar que el puntaje de cada uno de los examinados es suficientemente confiable como para orientar la toma de decisiones, para lo cual se realizan análisis de consistencia interna, error condicional y función informativa del test, y análisis del comportamiento psicométrico de los ítems. Un segundo objetivo de estos análisis es asegurar un comportamiento adecuado de los ítems ancla, que permita generar resultados comparables entre niveles sucesivos y entre años. Más adelante se describen los análisis realizados para este segundo objetivo.

En esta primera etapa los insumos para los análisis son los siguientes:

- Tablas de especificaciones de cada una de las pruebas, en la cual se señala la información de contenido de los ítems, su posición en la prueba y su comportamiento previo en las pruebas de campo.
- Bases de datos con las respuestas de los estudiantes examinados, donde cada pregunta tiene una puntuación de 0 (incorrecta) o 1 (correcta).
- Marco muestral orientado a asegurar que los casos que se usan para los análisis psicométricos son similares a la composición del universo de estudiantes, en términos de dependencia del establecimiento y sus resultados en SIMCE (se detalla en la siguiente sección). Esto permite que el comportamiento de la prueba no dependa de las características específicas del grupo de establecimientos escolares que participa en SEPA cada año, grupo que presenta variaciones entre años.

**Definición de la muestra representativa**

Una muestra representativa es un subconjunto de estudiantes que participaron en SEPA 2015, la cual es obtenida con la intención de inferir propiedades en la totalidad de la población SEPA y realizar comparaciones entre años.

La muestra representativa es utilizada en el establecimiento de valores de referencia para las distintas comparaciones entre grupos que se realizan con los resultados obtenidos por los estudiantes.

Como un primer paso, se realiza un procedimiento de selección muestral entre los casos válidos y se determina un subconjunto de examinados que forma parte de la muestra representativa. El procedimiento consiste en los siguientes pasos:

1. **Validación de casos**: Se procede a la validación de los casos examinados que incluye eliminación de casos pertenecientes a educación de necesidades educativas especiales permanentes (NEP), casos con alta omisión (>20). Luego se procede a verificar que todos los registros ingresados sean válidos y únicos.

2. **Mínimo de estudiantes por establecimientos:** Los casos que cuentan con menos de 20 examinados válidos por colegio-nivel son descartados del muestreo.

3. **Estratificación y vinculación con SIMCE:** A cada caso se vincula con la información de la Prueba SIMCE del año en curso, en específico: los datos de desempeño en la prueba SIMCE para el sector (Lenguaje/Matemática), grupos socioeconómico (GSE) y dependencia del establecimiento educativo. Lo anterior se realiza con las siguientes consideraciones: Para los cursos de 1° a 4° Básico de la prueba SEPA se utiliza el valor SIMCE de 4° Básico, para los niveles de 5° a 8° se usa el valor SIMCE de 8° Básico, y para 1° a 4° Medio se utiliza el valor de la prueba SIMCE de 2° Medio. Si bien ya está disponible la medición de 2º y 6º básico, no se utilizaron dichos valores para los niveles respectivos, ya que existía una tasa importante de colegios que no poseían los datos. Adicionalmente, la variable GSE se reduce de 5 a 3 niveles socioeconómicos: alto, medio (medio más medio alto) y bajo (medio bajo más bajo), y la dependencia se considera según los 3 tipos más generales: particular pagado, particular subvencionado y municipal. Una vez caracterizado cada colegio-nivel con su GSE y dependencia, se realiza un muestreo estratificado proporcional según nivel, dependencia y GSE, de acuerdo a la distribución de estos. Esta operación se realiza 3 veces para obtener 3 escenarios aleatorios distintos, cuya validez se juzga a partir de su distancia con los puntajes SIMCE

nacionales en los respectivos estratos y con ello se selecciona la versión que resulte más válida.

Los respectivos estratos de colegio-nivel de curso que fuesen seleccionados son identificados en una nueva variable (con MR=1), la que se lleva a la base de datos a nivel alumno que se utilizará para los análisis. Si una determinada combinación colegio-nivel presenta más de 100 alumnos, estos son ponderados para mantener su peso en ese valor, y que de esta forma no se sobre represente ese estrato dentro del conjunto general.

**Confiabilidad**

Una medida relevante de la calidad métrica de los instrumentos de medición es su confiabilidad. Este concepto alude al grado en que un instrumento de medición genera puntuaciones consistentes, es decir, con bajo nivel de error de medición. Como aproximación para estimar la confiabilidad de las pruebas SEPA, se emplea el coeficiente alfa de Cronbach (1951). Este analiza el grado en que las respuestas a todas las preguntas muestran consistencia (Kuder & Richardson, 1937). Para juzgar la calidad del coeficiente Alfa de Cronbach, usualmente se considera aceptable a partir de 0,7, aunque es recomendable que supere a 0,8.

La confiabilidad de las pruebas SEPA, a lo largo de los años, ha sido adecuada en todos los niveles evaluados, superando con holgura el valor 0,8 en casi todos los casos. La Tabla 1 muestra los índices de confiabilidad asociados a la aplicación de las pruebas SEPA en 2015.

**Tabla 1: Confiabilidad de las pruebas SEPA 2015**

| | 1°B | 2°B | 3°B | 4°B | 5°B | 6°B | 7°B | 8°B | 1°M | 2°M | 3°M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Matemática** | 0,79 | 0,84 | 0,88 | 0,91 | 0,88 | 0,90 | 0,89 | 0,90 | 0,91 | 0,91 | 0,87 |
| **Lenguaje** | 0,82 | 0,84 | 0,88 | 0,89 | 0,90 | 0,85 | 0,88 | 0,89 | 0,87 | 0,89 | 0,87 |

**Función Informativa**

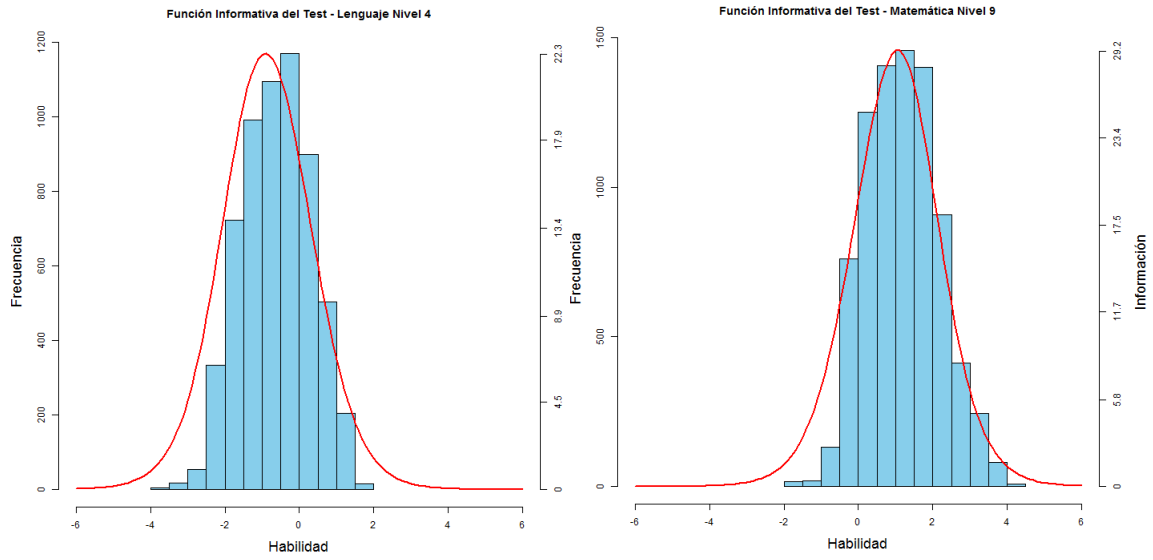
En Teoría de Respuesta al Ítem, la *información* es una generalización del concepto de confiabilidad de la Teoría Clásica. De esta forma, la información está directamente relacionada con la precisión en los resultados e inversamente relacionada con el error de medición.

Una de las características de la información es que no es constante a lo largo de la escala de puntajes, por lo que permite cuantificar el error de medición en cada nivel de la

habilidad latente que se está estudiando y así, sirve para cuantificar de mejor manera la precisión de los resultados entregados sobre los estudiantes.

Debido que las pruebas SEPA buscan informar de manera general acerca del desempeño de los estudiantes examinados, se espera que la información se encuentre optimizada en aquella región de la escala de puntaje donde haya mayor concentración de estudiantes. La Figura 2 muestra las funciones informativas de dos pruebas junto al histograma con los resultados obtenidos por los estudiantes que rindieron las respectivas pruebas. En ellas se aprecia que la información entregada por cada test es mayor donde la cantidad de evaluados es mayor y decrece en los extremos, que es donde la cantidad de evaluados disminuye. Lo observado en estos dos casos es equivalente a lo que se constata en cada uno de los niveles evaluados, según se puede observar en el anexo A.

**Figura 8: Funciones informativas de dos pruebas SEPA: Lenguaje 4° Básico (izquierda) y Matemática I° Medio (derecha) 2015. La línea roja indica la información del test.**



**Análisis clásico del ítem**

El análisis clásico que se realiza a las pruebas SEPA cumple con un rol informativo respecto a la etapa de calibración de las pruebas. Se analizan los cinco indicadores siguientes, los cuales son resumidos en el anexo B.

1. **Parámetros que describen la dificultad de cada ítem:** corresponde a la proporción de examinados que respondieron correctamente ese ítem, cuando la pregunta posee solo una opción correcta (Crocker & Algina, 1986; Haladyna, 2004; Kaplan & Saccuzzo, 2001; Livinston, 2006). Conocer esta propiedad de los ítems permite evitar el incluir en una prueba preguntas que, por ser en extremo fáciles o difíciles, no contribuyen a obtener información sobre los examinados (Livinston, 2006). Ahora bien, el nivel óptimo de dificultad de los ítems depende del uso o propósito de la prueba y el tipo de ítems, y por lo general se recomienda que en una misma prueba haya ítems de distintos grados de dificultad (Kaplan & Saccuzzo, 2001).

2. **Parámetros que describen la capacidad discriminativa de cada ítem**: dado que el propósito de cualquier prueba es proveer información acerca de las diferencias individuales de los examinados en un ámbito específico (en este caso Matemática y Lenguaje), se examina la capacidad discriminativa de los ítems, que corresponde al grado en que cada ítem puede diferenciar entre examinados que tienen un mayor nivel de habilidad con respecto a quienes tienen un menor nivel de habilidad. Este análisis típicamente se lleva a cabo correlacionando la forma en que los examinados responden una pregunta y el puntaje que ellos obtienen en el conjunto de preguntas de la prueba (Puntaje Total) (Crocker & Algina, 1986; Haladyna, 2004; Kaplan & Saccuzzo, 2001; Livinston, 2006).

3. **Parámetros que permiten examinar la tasa de omisión de cada ítem**: es usual que los examinados dejen algunas preguntas sin responder. El grado en que esto ocurre depende de diversos factores, como las instrucciones que se entregan (que pueden incentivar o desincentivar la omisión), la longitud de la prueba y la dificultad de la pregunta. En el caso de SEPA las instrucciones no promueven la omisión, por lo que el

foco principal del análisis de la omisión está en determinar si la falta de respuesta responde a la longitud de la prueba (lo que se constata cuando la omisión se incrementa en las preguntas localizadas hacia el final de la prueba), así como estableciendo el grado en que la omisión refleja la dificultad de la pregunta (lo que se manifiesta cuando examinados con mejor desempeño presentan menor omisión que examinados con peor desempeño) (Haladyna, 2004).

4. **Parámetros que permiten examinar la calidad de los distractores:** Thissen, Steinberg y Fitzpatrick (1989 citado en Haladyna, 2004) establecieron que los distractores de una pregunta de respuesta múltiple deben ser considerados como parte importante del ítem. Estadísticamente, Haladyna (2004) recomienda analizar esto observando su capacidad discriminativa y el puntaje promedio de quienes seleccionan cada distractor. Otro aspecto que conviene analizar de los distractores es la proporción de examinados que los elige. Bajo el supuesto que hay examinados de distintos niveles de habilidad o conocimiento, un distractor que rara vez es seleccionado, o bien que es demasiado popular, debería ser eliminado o, idealmente reemplazado en la formulación del ítem (Nunnally & Bernstein, 1994).

**Análisis de funcionamiento de ítems ancla**

Para asegurar la comparabilidad de las puntuaciones entre años y niveles consecutivos, se realiza un procedimiento de equiparación de puntajes. La equiparación es un procedimiento que permite comparar las puntuaciones de los examinados en distintas pruebas. La equiparación se basa en el empleo de ítems comunes (ancla) entre formas de las pruebas (por ejemplo, entre formas que corresponden a niveles escolares consecutivos, o entre formas referidas a un mismo nivel, pero que se aplican en diferentes años). A partir de lo anterior, el procedimiento de equiparación ("equating" como se le denomina en inglés) permite poner en una misma métrica estas pruebas, permitiendo entre otras cosas estimar el progreso de puntajes en el tiempo o comparar el desempeño de estudiantes que rinden la prueba de un mismo nivel o grado en distintos años.

Para realizar una correcta equiparación de puntajes es importante que los ítems comunes tengan un comportamiento psicométrico adecuado. Para verificar esto se realiza un

análisis particular centrado en este conjunto de ítems. A continuación, se describen los dos análisis principales que se realizan para verificar el funcionamiento de los ítems de anclaje.

Cuando los ítems de anclaje se usan entre niveles consecutivos de enseñanza, se espera que ellos muestren un grado de dificultad menor en el nivel escolar superior. Para ilustrar lo dicho, en la Figura 3 se muestra un ejemplo del comportamiento de los ítems ancla, que aparecen en las pruebas de 5° y 6° Básico en la prueba de Matemática. Tal como se esperaba, los estudiantes del nivel superior (color rojo) contestan con mayor frecuencia correctamente tales ítems que los estudiantes de nivel inferior (color azul).

Adicionalmente, se examina la correlación entre las dificultades de los ítems ancla en cada par de niveles. Se espera que el ordenamiento de los ítems ancla sea equivalente al aplicarlos en distintos grados o años, por lo que la correlación debe ser alta. Siguiendo con el mismo ejemplo recién mencionado, en la Figura 4 se muestra un dispersiograma con los grados de dificultad en ambos niveles. Tal como puede observarse, la correlación del grado de dificultad para los ítems ancla usados en las pruebas de 5° y 6° Básico es muy alta (0,95 en este caso).

Este análisis se realiza para cada par de niveles y en los casos en que algún ítem ancla presenta un comportamiento anómalo, se decide descartarlo en su rol de anclaje. Una vez tomadas estas decisiones, se implementa la equiparación de puntajes entre niveles de enseñanza.

**Figura 9: Porcentaje de respuestas correcta a los ítems anclas, Matemática 5° y 6° Básico 2015**

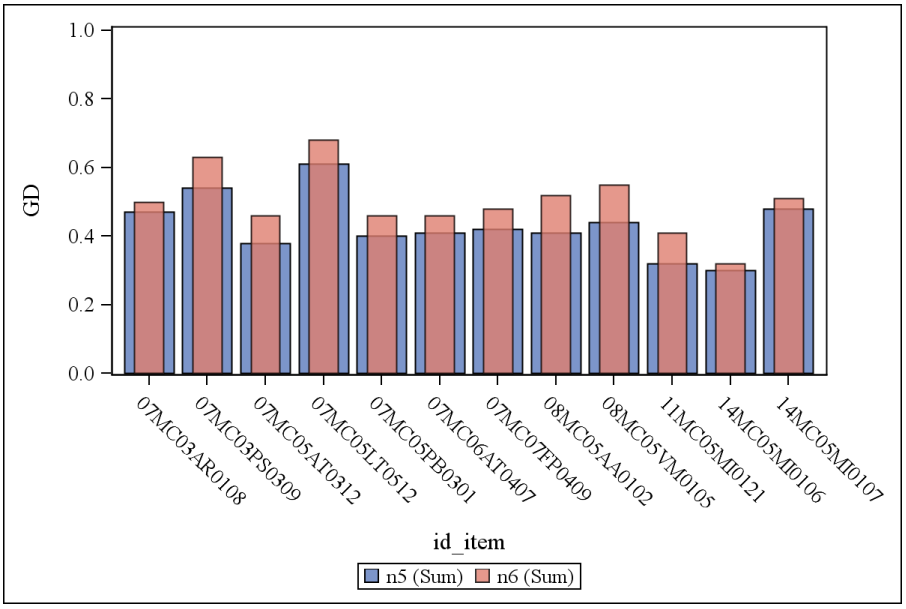
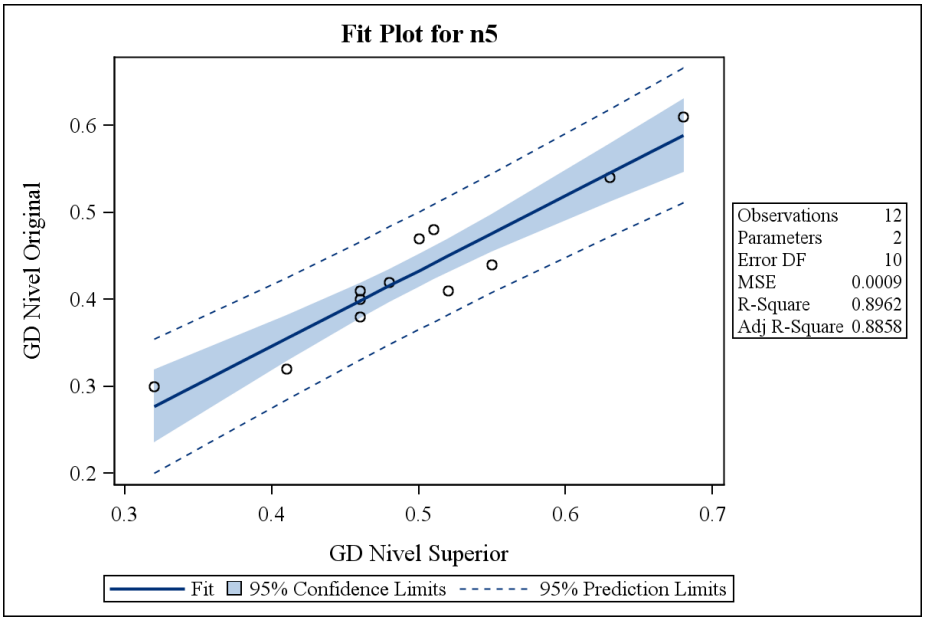



**Figura 10: Gráfico de dispersión entre los grados de dificultad de los ítems ancla, Matemática 5° y 6° Básico 2015**




## Modelamiento

La prueba SEPA es modelada a partir de la Teoría de Respuesta al Ítem (TRI), específicamente un modelo Rasch de un parámetro con regresión latente. El modelo Rasch plantea que la probabilidad de contestar correctamente un ítem depende de su dificultad ($\beta_i$) y

del nivel de habilidad individual del examinado ($\theta_p$) en una prueba determinada, por ejemplo, la habilidad del estudiante en Lenguaje o Matemática.

Para definir rigurosamente la habilidad individual de los estudiantes (examinados) y la dificultad de un ítem, es necesario considerar dos hechos:

1. Al observar las respuestas a un ítem en una determinada población de individuos, por ejemplo, la población de estudiantes SEPA, se constata que dichas respuestas son heterogéneas.
2. Al observar las respuestas que un estudiante da a un conjunto de ítems, también se observa un comportamiento heterogéneo.

La heterogeneidad puede ser representada formalmente por medio de probabilidades. Específicamente, definamos $\pi_{pi} = P(Y_{pi} = 1|\theta_p, \beta_i)$ como la probabilidad que tiene el estudiante $p$ de contestar correctamente el ítem $i$, tal que su habilidad latente es $\theta_p$ y el grado de dificultad del ítem es $\beta_i$. Ahora, tanto la habilidad de una persona, como la dificultad de un ítem, deben ser interpretadas con respecto a un proceso generador de datos, el cual se formaliza en términos probabilísticos. Este proceso no hace sino representar la heterogeneidad que se observa tanto en las respuestas que una persona da a un conjunto de ítems, así como la heterogeneidad que se observa cuando un grupo de estudiantes responde un determinado ítem. Asumido esto, el modelo Rasch asume que la probabilidad de responder correctamente un ítem está dada por,

$$\pi_{pi} = P\left(Y_{pi} = 1\middle|\theta_p, \beta_i\right) = \frac{e^{\theta_p - \beta_i}}{1+e^{\theta_p - \beta_i}} \qquad (1)$$

La ecuación (1) es una función de $\theta_p$ y $\beta_i$ llamada curva característica del ítem (CCI[5])

Donde,

- $Y_{pi}$ es una variable aleatoria que toma el valor 1 cuando la persona $p$ responde correctamente el ítem $i$, y toma el valor 0 en caso contrario.

- Consecuentemente la probabilidad de responder incorrectamente un ítem es

[5] CCI es conocida como ICC que es la sigla en inglés de Ítem characteristic curve.

$$P(Y_{pi} = 0|\theta_p, \beta_i) = 1 - \pi_{pi} = \frac{1}{1 + e^{\theta_p - \beta_i}}$$

**Modelo Estadístico**

Para el modelamiento inicial se contempla que la probabilidad de responder correctamente está dada por la ecuación (1), en este contexto como cada respuesta al ítem $i$ puede ser "correcta" o "incorrecta", se dice que

$$Y_{pi}|\pi_{pi} \sim Bern\,(\pi_{pi})$$

El modelo Rasch usado en las pruebas SEPA incluye una regresión latente, de tal forma que la habilidad de los estudiantes depende de su nivel educacional, $Z_p$, es decir

$\theta_p = \lambda Z_p + \epsilon_p$, $\epsilon_p \sim N\,(0, \sigma^2)$

De esta manera, el modelo estadístico tiene la forma:

$$P(Y_{p1} = y_1, \dots, Y_{pn} = y_n|\beta_1, \dots, \beta_I, \theta) = \prod_{i=1}^{I} \left(\frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}}\right)^{y_{pi}} \left(\frac{1}{1 + e^{\theta_p - \beta_i}}\right)^{1 - y_{pi}}$$

Puesto que los parámetros que interesa estimar involucran tanto a las dificultades de los ítems como las habilidades, parece relevante mencionar dos propiedades implicadas por la representación simultánea de dificultades y habilidades: comparación de dificultades de ítems invariante con respecto a los individuos y comparación de habilidades de individuos invariante con respecto a los ítems. Más precisamente, en la escala de modelo se tiene que,

$$\boldsymbol{ln\left(\frac{\pi_{pi}}{1-\pi_{pi}}\right) = \theta_p - \beta_i} \qquad (2)$$

La ecuación (2) es llamada logaritmo de odds, así, la diferencia en logaritmo de los odds para cualquier ítem es simplemente la diferencia entre las dos habilidades,

$$ln\left(\frac{\pi_{1i}}{1 - \pi_{1i}}\right) - ln\left(\frac{\pi_{2i}}{1 - \pi_{2i}}\right) = \theta_1 - \theta_2$$

Por lo que, las habilidades de individuos son invariante con respecto a los ítems. Similarmente, para un mismo individuo *i* y dos ítems diferentes, se tiene que

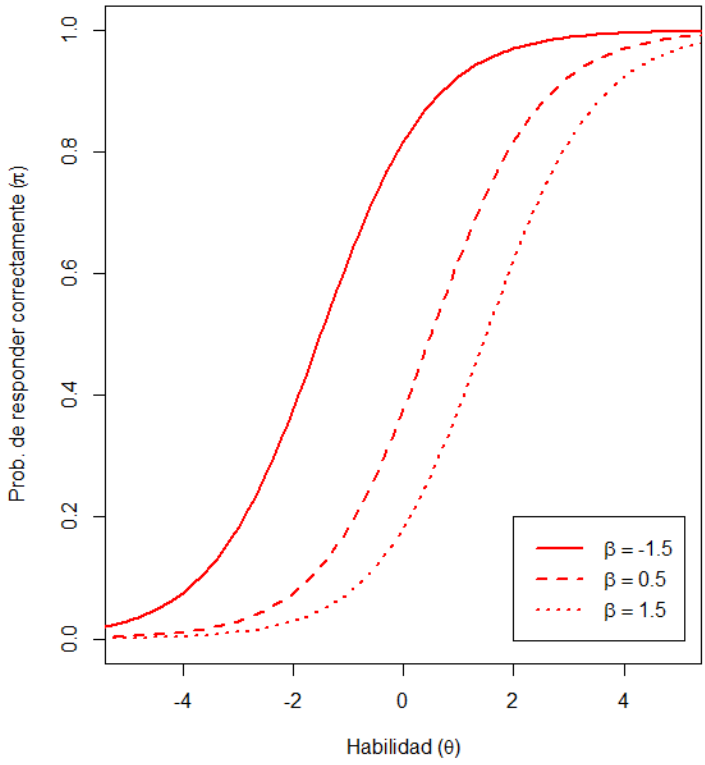
$$ln\left(\frac{\pi_{P1}}{1-\pi_{P1}}\right) - ln\left(\frac{\pi_{P2}}{1-\pi_{P2}}\right) = \beta_2 - \beta_1$$

Es decir, la diferencia entre los logaritmos de los odds para los ítems 1 y 2 es simplemente la diferencia de sus respectivas dificultades, es decir, las dificultades de los ítems son invariante con respecto a los individuos.

**Interpretación de los parámetros**

Notemos que hay una curva característica para cada ítem (CCI), esta CCI es una función $\theta_p$ y tiene un parámetro fijo $\beta_i$ como se mostró en la ecuación (1). Una forma de ilustrar la interpretación de estos parámetros es observando diferentes CCI con distintos valores de $\beta$. En la figura 4 se aprecia los CCI a tres valores de beta, la interpretación es bastante inmediata: mientras más grande sea el valor de $\beta$, como $\beta = 1.5$, la CCI está hacia la derecha, por lo que para que la probabilidad de que un estudiante conteste correctamente un ítem sea cercana a uno, se requiere un valor de $\theta$ mayor, es decir, es un ítem que requiere mayor habilidad del estudiante para responder correctamente. Por otra parte, cuando el valor de $\beta$ es menor, como $\beta = -1.5$, la CCI está hacia la izquierda, por lo que para que la probabilidad de que un estudiante conteste correctamente un ítem sea cercana a uno requiere poca habilidad del estudiante.

**Figura 11: Curvas características de ítems para distintas dificultades de estos**



**Estimación de los parámetros**

La estimación de los parámetros es el paso que nos permite llegar de las respuestas conocidas de las personas a los ítems, a los valores desconocidos de los parámetros de los ítems ($\beta$) y de los niveles de habilidad ($\theta$). Para obtener las estimaciones se aplica fundamentalmente el método de máxima verosimilitud. La lógica general de la estimación consiste en encontrar los valores de los parámetros que hagan más probable las respuestas obtenidas de los estudiantes. Si se consideran las respuestas de $N$ examinados a los $I$ ítems, la función de verosimilitud es la siguiente

$$L(\beta_1, \dots, \beta_I, \theta) = \prod_{p=1}^{N} \prod_{i=1}^{I} {\pi_{pi}}^{y_{pi}} (1 - \pi_{pi})^{1-y_{pi}}$$

Así, para obtener las estimaciones de los parámetros es necesario resolver la siguiente ecuación para cada parámetro de dificultad,

$$\frac{\partial \ln(L(\beta_1, \dots, \beta_I, \theta))}{\partial \beta_i} = 0$$

Para la puntuación de los estudiantes ($\theta_p$) se emplea la estimación EAP (Expected A Posteriori). De acuerdo a lo planteado por Mislevy y Bock (1982) existen diversas ventajas de este método por sobre otros como Máxima Verosimilitud (ML) o Máxima Verosimilitud a Posteriori (MAP), entre las que destacan: la facilidad de cálculo, dado que no es un proceso iterativo y no es necesario obtener la derivada de la función de respuesta, por otro lado, la estabilidad para test de distinta extensión. Este tipo de estimación no posee contradicciones entre el método utilizado para equating y el método utilizado para estimar la habilidad.

**Equiparación de puntajes**

Para monitorear y medir de manera apropiada el progreso de los alumnos a través de las pruebas, se requiere hacer comparables los puntajes de las pruebas entre niveles y de un año con el siguiente, permitiendo ubicar los resultados de cada nivel en una escala común. Para lograr esta comparabilidad de puntajes entre niveles y a través del tiempo, las pruebas SEPA contienen preguntas comunes entre un nivel y el anterior. Así, por ejemplo, la prueba SEPA de Matemática de 7° Básico incluye algunas preguntas que evalúan contenidos de 6° Básico que son medidos en ambas pruebas. De esta manera, cada prueba se compone de ítems de su

propio nivel y del anterior, siendo la prueba de 1° Básico la única que contiene sólo ítems de su propio currículo (Figura 5).

La equiparación de puntajes usada en las pruebas SEPA corresponde a un proceso estadístico bajo los modelos de teoría de respuesta al ítem. Específicamente, se utilizan dos métodos:

- Para establecer la escala vertical se utiliza el método de calibración concurrente, que se caracteriza por estimar los parámetros de los ítems de todos los niveles de forma simultánea (Lord & Wingersky, 1984). De esta manera, los parámetros de los ítems ancla se estiman utilizando la información aportada por las pruebas de los dos niveles en los que se presenta cada uno de ellos y se obtiene un solo valor para su parámetro que es utilizado para la puntuación de todos los estudiantes en los dos niveles. Debido a que el desempeño esperado en los ítems ancla es, en general, mejor en los estudiantes del nivel superior que en los estudiantes del nivel inferior, los ítems ancla permiten evidenciar la evolución en el proceso de aprendizaje de los estudiantes, lo que se traduce en posicionar de forma natural a los estudiantes del nivel superior por encima de los estudiantes de nivel inferior en la escala de puntajes SEPA generando el escalamiento vertical de las pruebas.
- Para establecer la comparabilidad de los puntajes a través de los años se realiza un procedimiento de calibración del test con fijación de parámetros. Esto consiste en utilizar los valores de los parámetros de los ítems ancla obtenidos en el proceso del año anterior e imputarlos en la calibración de los parámetros de la prueba, previo al proceso de estimación de parámetros. De esta forma, la estimación de los parámetros de los ítems y la posterior puntuación de los estudiantes se realiza habiendo fijado ya un subconjunto de ítems en la escala del año anterior, con lo que se logra la comparabilidad de puntajes entre dos años consecutivos.

**Figura 12: Calibración Concurrente entre niveles**

| Prueba / Items | 1° | 2° | 3° | 4° | 5° | 6° | 7° | 8° | I° | II° | III° |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grado 1° | X | X | | | | | | | | | |
| Grado 2° | | X | X | | | | | | | | |
| Grado 3° | | | X | X | | | | | | | |
| Grado 4° | | | | X | X | | | | | | |
| Grado 5° | | | | | X | X | | | | | |
| Grado 6° | | | | | | X | X | | | | |
| Grado 7° | | | | | | | X | X | | | |
| Grado 8° | | | | | | | | X | X | | |
| Grado I° | | | | | | | | | X | X | |
| Grado II° | | | | | | | | | | X | X |
| Grado III° | | | | | | | | | | | X |

**Diferencias significativas**

Junto con los resultados SEPA, se reporta si existen o no diferencias significativas en los puntajes y en los porcentajes de logros obtenidos en las pruebas entre distintos grupos de estudiantes y/o establecimientos, las comparaciones reportadas corresponden a:

a. **Comparación de nivel escolar de un colegio en particular con la referencia en el mismo nivel.** La referencia se compone de todos los alumnos que fueron seleccionados en la muestra representativa (MR), explicada previamente.

b. **Comparación entre los cursos de un mismo establecimiento.** Se toma el promedio de un curso, para un mismo nivel y se compara con el promedio del puntaje de todos los alumnos del nivel.

Así, podemos definir cuatro criterios de diferencias reportadas por las pruebas SEPA en todos sus niveles y en ambas áreas de aprendizaje:

i) Comparar los puntajes promedio de los establecimientos con su promedio de referencia. Se usa un test de comparación de medias para poblaciones independientes.

ii) Para un establecimiento y nivel dados, comparar los puntajes promedio de los cursos respecto al promedio del mismo nivel (ej. 4° A con el nivel 4º).

iii) Para un nivel dado, comparar los porcentajes de logro promedio de los ejes temáticos de los establecimientos (ej. 4° Números de un establecimiento con 4° Números de la muestra de referencia de la misma dependencia).

iv) Para un establecimiento y nivel dados, comparar los porcentajes promedio de los ejes temáticos de los cursos respecto del nivel.

Para obtener las diferencias significativas se usa un test t para muestras independientes, que se detallará en el Anexo B.

**Valor Agregado**

En términos generales, el análisis de Valor Agregado intenta cuantificar la contribución de los establecimientos educacionales al aprendizaje de los estudiantes. De esta manera, los índices que obtiene cada establecimiento muestran, para cada nivel y sector, si el establecimiento ha contribuido en promedio, lo mismo que otros establecimientos de similares características, o si, por el contrario, ha hecho un aporte significativamente mayor o menor a lo que se esperaría. El índice de Valor Agregado no indica por sí mismo qué es aquello que ha realizado el establecimiento escolar para agregar valor, sino que invita a que puedan realizar una reflexión interna que considere las características propias de la institución, para comprender así sus resultados. Esto puede conectarse así, con las prácticas pedagógicas y la gestión que resulta más o menos efectiva para contribuir a los aprendizajes de los estudiantes.

El Valor Agregado reportado por SEPA, es estimado en el contexto de efectividad escolar. Un enfoque estándar para el cálculo del VA de la escuela son los modelos lineales mixtos jerárquicos (en adelante HLM[6]), o modelos multinivel, (Raudenbush y Bryk, 2002; Snijders y Bosker, 1999). Esta clase de modelos se ajusta a una función específica de los datos educativos normalmente utilizados para realizar análisis de Valor Agregado, es decir, la estructura jerárquica de los datos, en que los estudiantes se anidan en las escuelas. En este contexto, el valor añadido de una escuela es, por un lado, modelado como un efecto aleatorio y, por otra

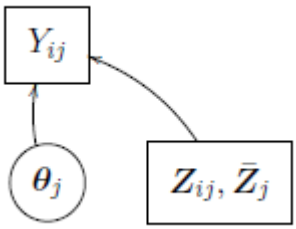
[6] HLM: Hierarchical Lineal Mixed

parte, calculado como la predicción estimado de ella (Aitkin & Longford, 1986; Longford, 2012; Raudenbush y Willms, 1995; Tekwe et al., 2004).

Desde la perspectiva de los modelos HLM, definiremos las siguientes variables

- $Y_{ij}$: Puntaje en la prueba SEPA 2015 del estudiante $i$ que pertenece a la escuela $j$ en un área determinada y un nivel escolar específico, por ejemplo, Matemática o Lenguaje de 5° Básico. De tal manera que $i \in \{1, \dots, n_j\}$ con $n_j$ el número de alumnos en la escuela $j$
- $Z_{ij}$: Puntaje en la prueba SEPA 2014 del estudiante $i$ que pertenece a la escuela $j$ en un solo nivel.
- $\bar{Z}_{.j}$: El efecto composicional de la escuela, entiéndase por este el puntaje promedio del establecimiento en el año 2014. Este efecto captura la heterogeneidad que existe en los establecimientos educacionales.
- $\theta_j$: El efecto escuela, esta es una variable aleatoria no observable.

**Figura 13: Relación entre el puntaje ($Y_{ij}$), Covariables ($Z_{ij}, \bar{Z}_j$) y efecto escuela ($\theta_j$) en un modelo HLM**




La estructura del modelo es lineal, entonces se asume que la puntuación individual final esperada depende linealmente de variables y el efecto de la escuela (ver Figura 6). Es decir, para cada estudiante $i$ perteneciente a la escuela $j$, se supone que existe un vector de parámetro, denotado por $\beta = (\beta_0, \beta_1, \beta_2)'$, de manera que,

$$E(Y_{ij} \mid \theta_j) = \beta_0 + \beta_1 \boldsymbol{Z_{ij}} + \beta_2 \bar{Z}_j + \theta_j$$

lo cual es equivalente a escribir

$$Y_{ij} = \beta_0 + \beta_1 \boldsymbol{Z_{ij}} + \beta_2 \bar{Z}_j + \theta_j + \epsilon_{ij}$$

Siguiendo con la especificación de los modelos HLM,

$$\boldsymbol{Y_{ij}|Z_{ij},\overline{Z}_j,\theta_j \sim N(\beta_0 + \beta_1 Z_{ij} + \beta_2 \overline{Z}_j + \theta_j ; \sigma^2)},$$

$$\boldsymbol{\theta_j|Z_{ij},\overline{Z}_j \sim N(0 ; \tau^2)},$$

Tal que, $\boldsymbol{i = 1,\dots,n_j}$, $n\_j$ es el número de estudiantes de la escuela $j$ y $\sigma^2 > 0,\ \tau^2 > 0,$

Ahora, heurísticamente hablando, la estrategia para definir el efecto escuela consiste en la descomposición del puntaje 2015 del estudiante $i$ que pertenece a la escuela $j$, $\boldsymbol{Y_{ij}}$ , y posteriormente identificar cuál de estas componentes están bajo el control de la escuela. Desde el punto de vista de la modelización, la puntuación de cada $\boldsymbol{Y_{ij}}$ se ve afectada tanto por las covariables como el efecto escuela. Por lo tanto, el puntaje se puede descomponer en tres componentes, cada uno de ellos está representado en términos de esperanza condicional, como sigue:

$$\boldsymbol{Y_{ij} = E(Y_{ij}|Z_{ij},\overline{Z}_j) + \left(E(Y_{ij}|Z_{ij},\overline{Z}_j,\theta_j) - E(Y_{ij}|Z_{ij},\overline{Z}_j)\right) + \left(Y_{ij} - E(Y_{ij}|Z_{ij},\overline{Z}_j,\theta_j)\right)} \quad \textbf{(3)}$$

Las tres componentes de la descomposición (3) son por construcción incorrelacionadas entre ellas. De esta manera, cada una de ellas representa una específica contribución al puntaje del estudiante, $\boldsymbol{Y_{ij}}$. Cada una de estas contribuciones tiene un significado específico:

i) El primer componente mide la contribución del vector de covariables $\boldsymbol{Z_{ij},\overline{Z}_j}$ en el puntaje $\boldsymbol{Y_{ij}}$.

ii) El segundo componente corresponde a la contribución de la escuela, $\boldsymbol{\theta_j}$, sobre $\boldsymbol{Y_{ij}}$, luego de tener en cuenta la contribución del vector de covariables $\boldsymbol{Z_{ij},\overline{Z}_j}$ en $\boldsymbol{Y_{ij}}$.

iii) El tercer componente es conocido como el error idiosincrático, esto es la "parte" de $\boldsymbol{Y_{ij}}$ que no es explicado estadísticamente ya sea por el efecto de la escuela $\$\boldsymbol{\theta_j}$, o por el vector de covariables $\boldsymbol{Z_{ij},\overline{Z}_j}$.

De la descomposición (3), es aceptable que el segundo componente es el que depende de la escuela, por consiguiente, está bajo su control. Esto lleva a definir el efecto de la escuela

de $j$ como el promedio de dicho componente del Valor Agregado, ver (Manzi, San Martín & Van Bellegem, 2014), a saber,

$$VA_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(E\left(Y_{ij}|\boldsymbol{Z_{ij}}, \overline{\boldsymbol{Z}}_{\boldsymbol{j}}, \boldsymbol{\theta_j}\right) - E\left(Y_{ij}|\boldsymbol{Z_{ij}}, \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\right)\right)$$

El primer término representa un promedio de la puntuación esperada en una escuela específica, después de controlar por las covariables. El segundo término corresponde a una media de la puntuación esperada en la escuela de referencia, después de controlar por las covariables; esta última interpretación se basa en la siguiente propiedad general: $E\left(Y_{ij}|\boldsymbol{Z_{ij}}, \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\right) = E\left(\, E\left(Y_{ij}|Z_{ij}, \bar{Z}_j, \theta_j\right)|Z_{ij}, \overline{\boldsymbol{Z}}_{\boldsymbol{j}}\,\right)$, es decir, el efecto es la integración de la escuela con respecto a su distribución.

En la literatura el Valor Agregado de la escuela es definido en términos de una escuela promedio o de referencia (Timmermans, Doolaard & de Wolf, 2011; Raudenbush, 2004; Raudenbush & Willms, 1995).

Bajo la definición de Valor Agregado en un modelo HLM, este corresponde al efecto escuela, $VA_j = \theta_j$. La predicción de efectos escuelas es determinada por el método de Bayes empírico en el que los parámetros desconocidos son reemplazados por sus estimadores, que corresponde a,

$$\widehat{\theta_j} = \left(\frac{\widehat{\tau^2}}{\widehat{\sigma^2} + n_j\widehat{\tau^2}}\right)\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{\beta}_0 + \hat{\beta}_1\boldsymbol{Z_{ij}} + \hat{\beta}_2\bar{Z}_j\right)$$

Cabe señalar que el VA calculado por SEPA es por área de conocimiento y por nivel escolar. Así, el modelo definido recientemente es aplicado a todos los niveles escolares desde 2° Básico y en Lenguaje y Matemática, esto quiere decir que, si un establecimiento participa en más de un nivel escolar en el año 2015, donde sus estudiantes tienen puntajes previos, puntajes 2014, entonces se les informa un VA por nivel escolar y área de conocimiento.

**Referencias**

Aitkin, M. & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Haladyna, T. (2004). *Developing and validating multiple-choice test items.* New York: Routledge.

Kaplan, R. & Saccuzzo, D. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA, US: Wadsworth.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151-160.

Livinston, S. (2006). Item analysis. En S. Downing & T. Haladyna. (Eds.), *Handbook of test development* (pp. 421-444). New Jersey: Lawrence Erlbaum Associates.

Longford, N. (2012). 'Which model?' is the wrong question. *Statistica neerlandica, 66*(3), 237-252.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement, 8,* 453-461.

Mislevy, R. J. & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*(3), 725-737.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3ra ed.). New York: Mc-Graw Hill.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121-129. doi:10.3102/10769986029001121

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage Publications, Inc.

Raudenbush, S. W. & Willms, D. (1995). The estimation of school effects*. Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

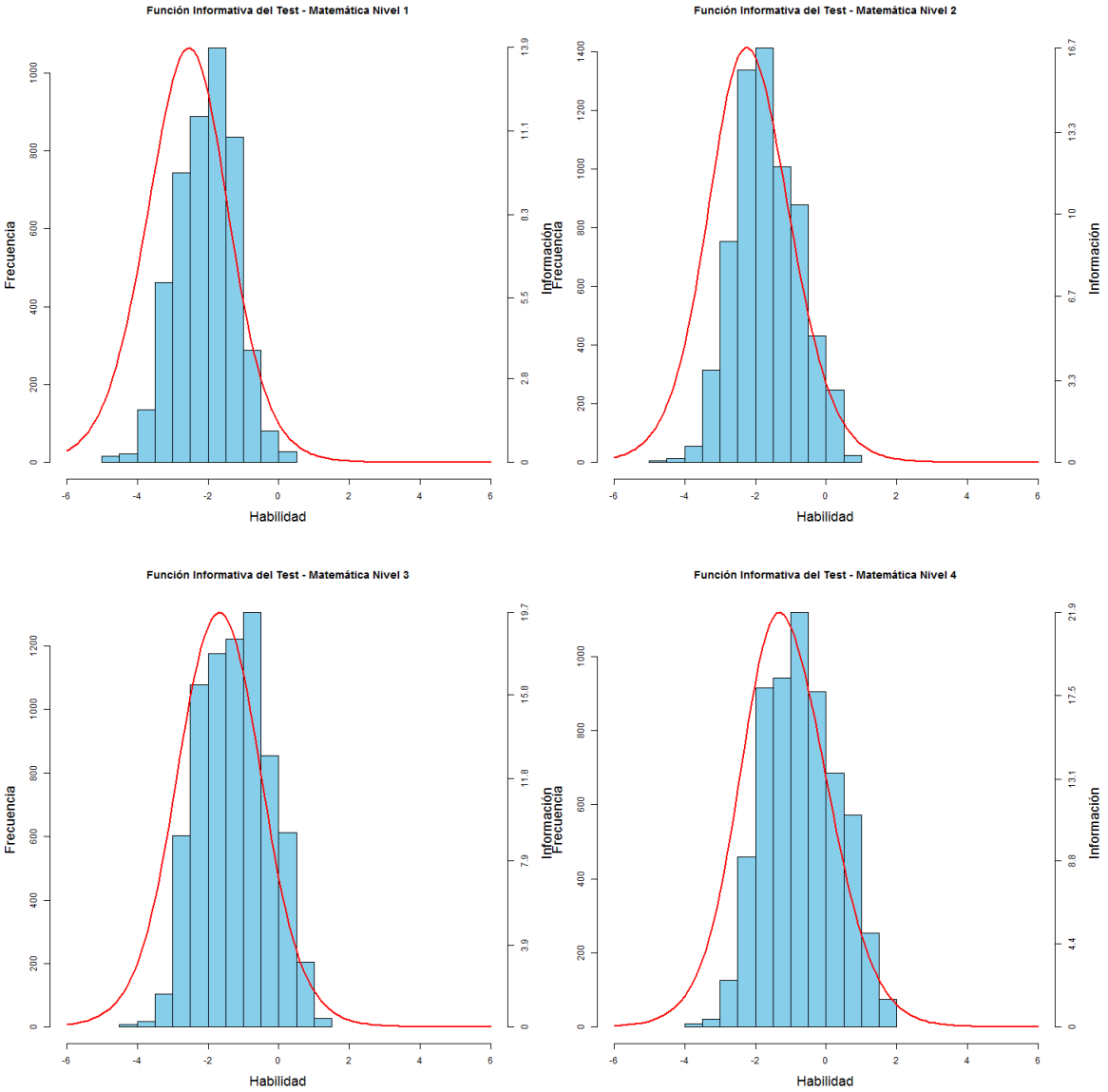
Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). London: Sage Publications, Inc.

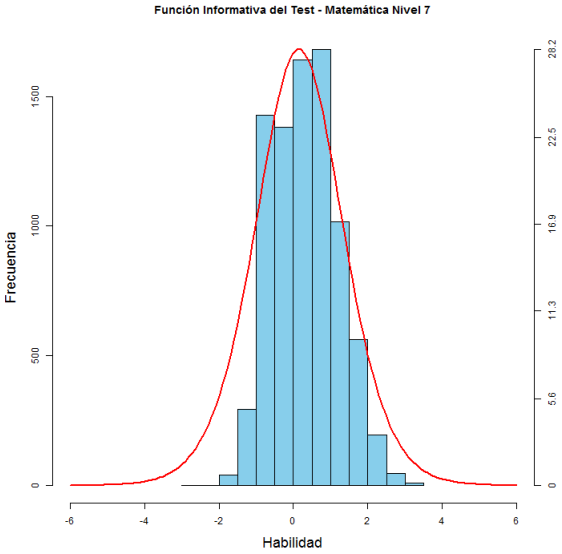
Tekwe, C.D., Carter, R.L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., and Resnick, M.B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. doi:10.3102/10769986029001011

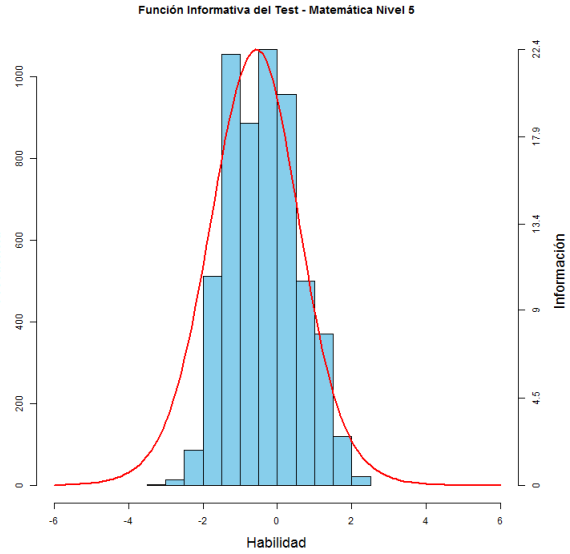
Timmermans, A. C., Doolaard, S. & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement, 22*(4), 393-413.

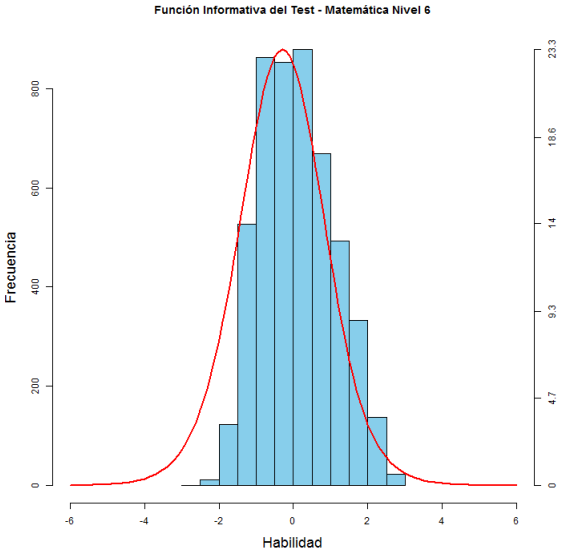
## Anexo A: Función informativa del test

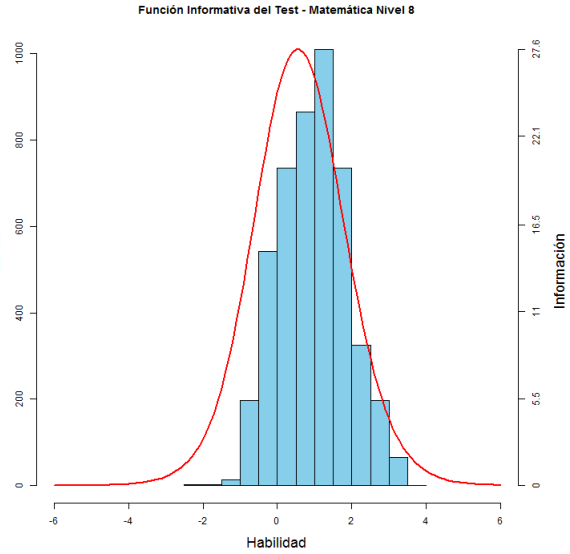
## Matemática 2015

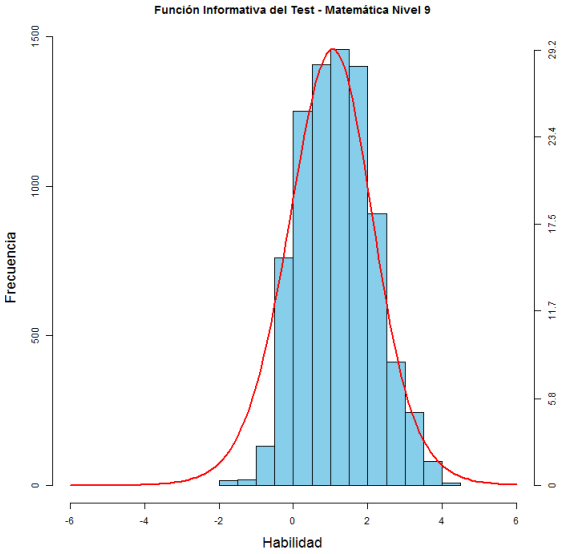


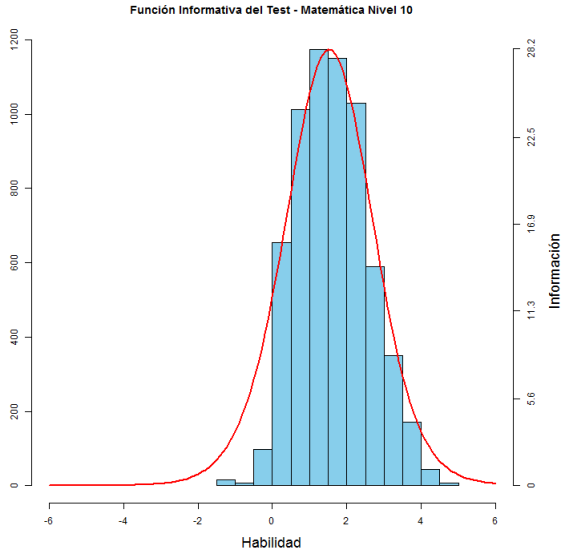

Función Informativa del Test - Matemática Nivel 7
Frecuencia
Información
Habilidad

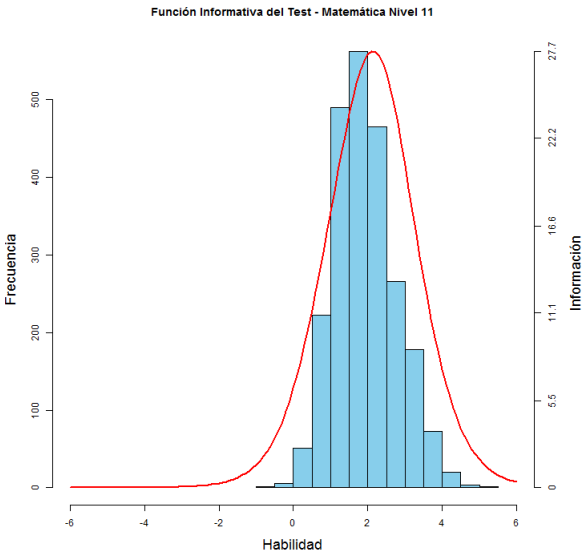

Función Informativa del Test - Matemática Nivel 5
Frecuencia
Información
Habilidad


Función Informativa del Test - Matemática Nivel 6
Frecuencia
Información
Habilidad

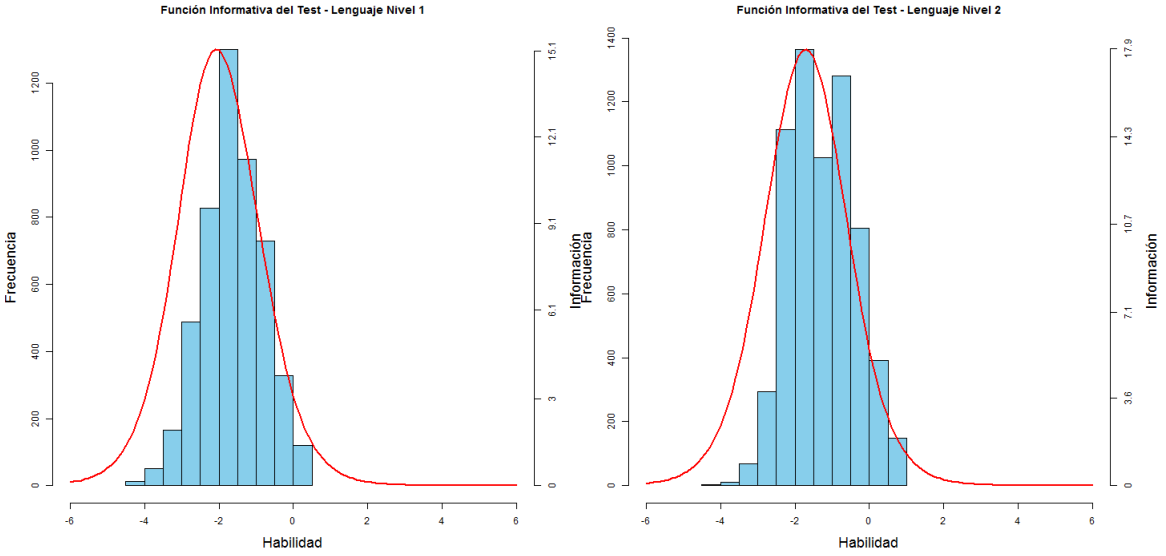

Función Informativa del Test - Matemática Nivel 8
Frecuencia
Información
Habilidad


Función Informativa del Test - Matemática Nivel 9
Frecuencia
Información
Habilidad


Función Informativa del Test - Matemática Nivel 10
Frecuencia
Información
Habilidad

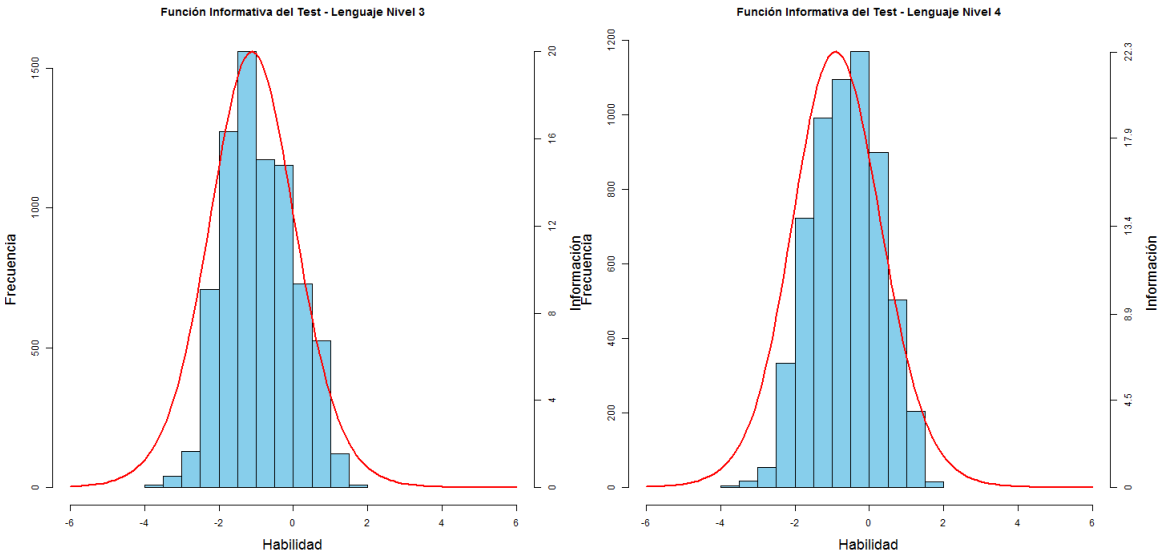

Función Informativa del Test - Matemática Nivel 11
Frecuencia
Información
Habilidad

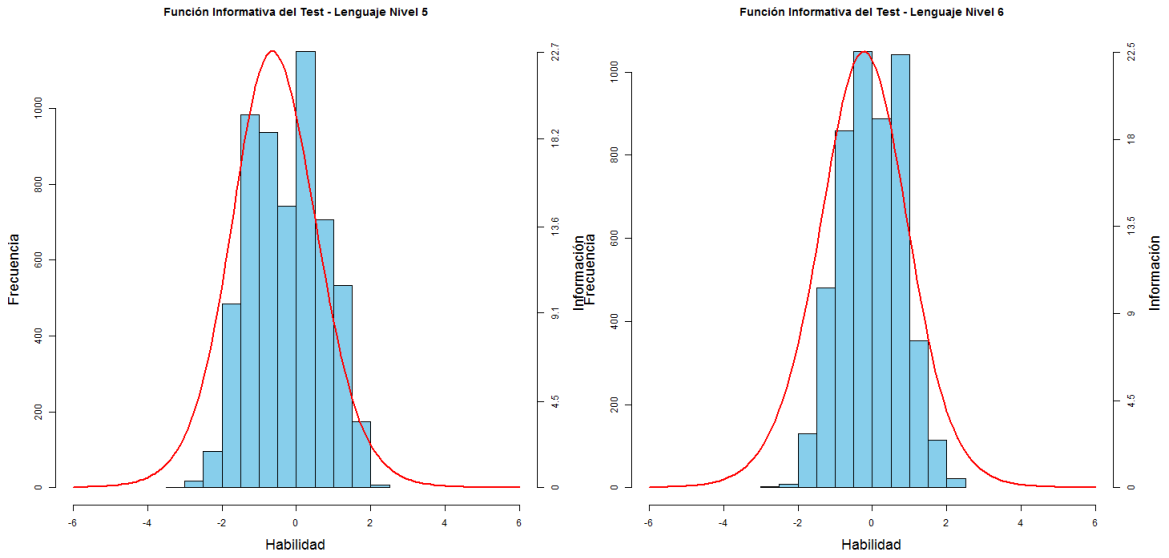
## Lenguaje 2015



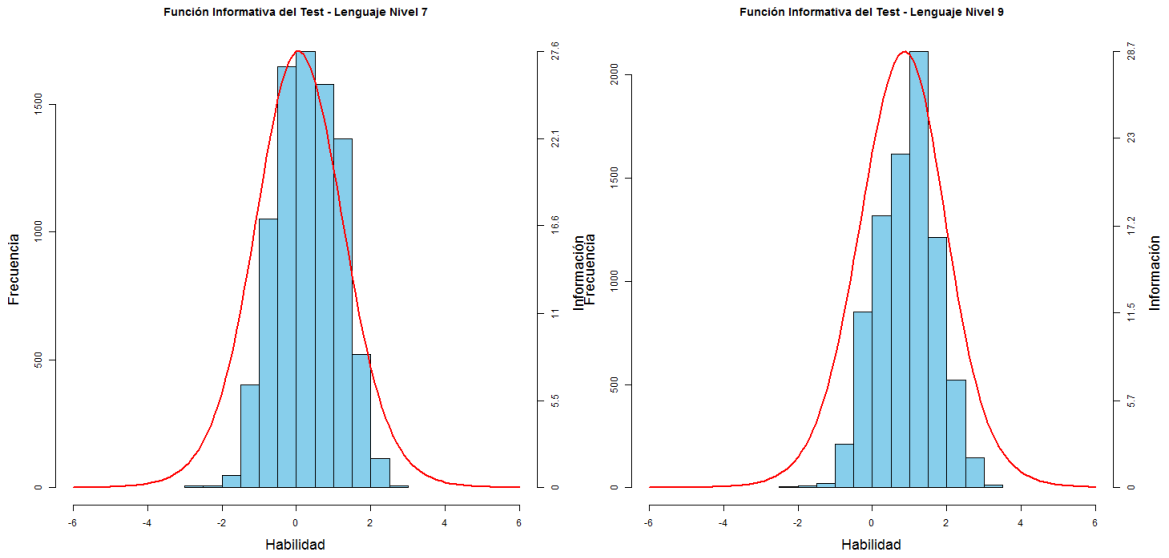




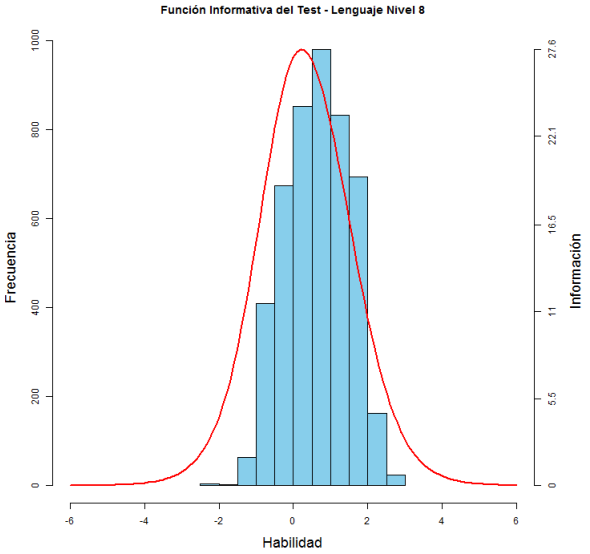


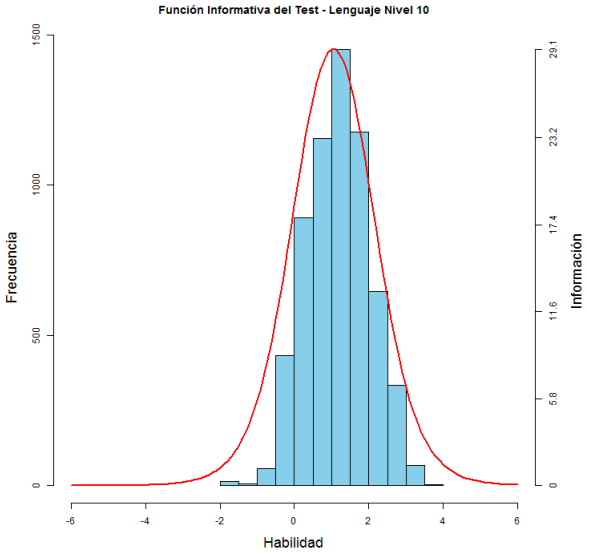

Función Informativa del Test - Lenguaje Nivel 5
Frecuencia
Información
Habilidad

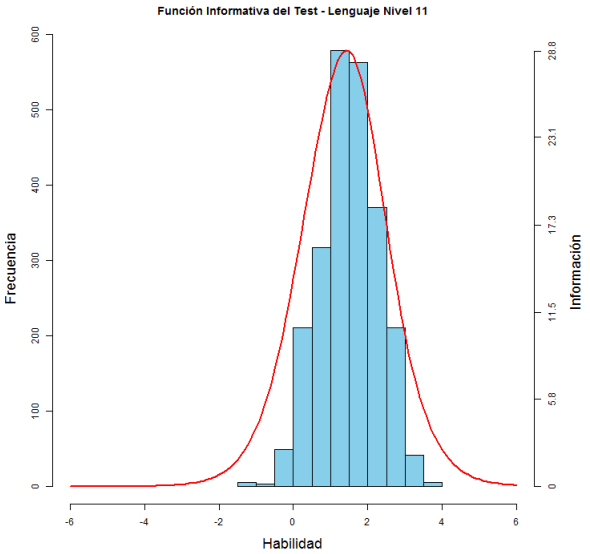

Función Informativa del Test - Lenguaje Nivel 6
Información
Frecuencia
Información
Habilidad


Función Informativa del Test - Lenguaje Nivel 7
Frecuencia
Información
Habilidad


Función Informativa del Test - Lenguaje Nivel 9
Información
Frecuencia
Información
Habilidad


Función Informativa del Test - Lenguaje Nivel 8
Frecuencia
Información
Habilidad


Función Informativa del Test - Lenguaje Nivel 10
Frecuencia
Información
Habilidad


Función Informativa del Test - Lenguaje Nivel 11
Frecuencia
Información
Habilidad

**Anexo B: Metodología**

**Test de comparación de medias:** Este test busca encontrar evidencia estadística para concluir si la diferencia entre el promedio de una agrupación de interés, $\bar{x}$, y el promedio con el cual se desea comparar, $\bar{y}$ es igual a 0 o no. Si, $\bar{x}$ e $\bar{y}$ provienen de la misma población, la diferencia tendría que ser pequeña, en caso contrario, si provienen de poblaciones distintas, la diferencia será grande.

Sean $X_1, X_2, \dots, X_{n_1}$ una muestra aleatoria de una distribución $N(\mu_x\,;\sigma_x^2\,)$ e $Y_1, Y_2, \dots, Y_{n_2}$ una muestra aleatoria de una distribución $N(\mu_y\,;\sigma_y^2\,)$ e independientes entre si. Un estimador para $\boldsymbol{\mu_x - \mu_y}$ es $\bar{X} - \bar{Y}$ y sabemos que

$$\bar{X} - \bar{Y} \sim N\left(\boldsymbol{\mu_x - \mu_y}\,; \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}\right)$$

Entonces si deseamos contrastar hipótesis sobre $\boldsymbol{\mu_x - \mu_y}$ , donde

$$H_0: \mu_x - \mu_y = 0$$

el estadístico de prueba cuando $H_0$ es verdadera será

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} \sim N(0;1)$$

La región de rechazo para conocer si las medias de dos poblaciones son diferentes corresponde a $RC: \left\{ z > z_{\frac{\alpha}{2}}\, o\, z < -z_{\frac{\alpha}{2}} \right\}$ [7]. Por otra parte, también podemos modificar la hipótesis alternativa, $H_A$, para conocer en cuál de las poblaciones la media es mayor, así, podemos resumir las siguientes hipótesis

[7] $z_{\frac{\alpha}{2}}$ Corresponde al valor de la normal estándar de un nivel de $\frac{\alpha}{2}$

| Hipótesis alternativa | Región de rechazo para un nivel $\alpha$ |
|---|---|
| $H_A: \mu_x - \mu_y \neq 0$ | $RC: \left\{ z > z_{\frac{\alpha}{2}} \; o \; z < -z_{\frac{\alpha}{2}} \right\}$ |
| $H_A: \mu_x - \mu_y > 0$ | $RC: \left\{ z > z_{\frac{\alpha}{2}} \right\}$ |
| $H_A: \mu_x - \mu_y < 0$ | $RC: \left\{ z < -z_{\frac{\alpha}{2}} \right\}$ |

# CAPÍTULO IV: RESULTADOS DE PRUEBAS APLICADAS EN 2015

**María Inés Godoy**

Ingeniera Estadística de la Universidad de Santiago de Chile, magíster en Estadística y doctor en Estadística de la Pontificia Universidad Católica de Chile. Actualmente es investigadora asociada a MIDE UC. (migodoy1@uc.cl)

Este capítulo resume los principales resultados observados en la prueba SEPA durante su aplicación en 2015. En la primera parte se muestra la evolución que presenta la escala de puntajes SEPA (la que se emplea para reportar puntajes de Estado y Progreso). Luego se presentan en secuencia los resultados de Estado, Progreso y Valor Agregado. En cada apartado se resumen los resultados generales observados con todos los estudiantes examinados en 2015, lo que se complementa con comparaciones de los puntajes según grupo socioeconómico y sexo. La información que presenta este capítulo sirve para conocer cómo se distribuyen los distintos puntajes que reporta SEPA, lo que a su vez permite contar con un marco de referencia para interpretar puntajes obtenidos por establecimientos escolares, cursos y examinados individuales.

La información técnica acerca de los puntajes que reporta este capítulo aparece en el capítulo 3, donde se informa acerca de los análisis y chequeos psicométricos que fundamentan los puntajes. Complementariamente, el Capítulo 5 entrega evidencia acerca de la validez de las puntuaciones SEPA.

**Escala SEPA**

La escala de puntajes SEPA se calcula empleando la Teoría de Respuesta al Ítem (IRT), (ver capítulo 3). Con esta escala se busca medir y monitorear de manera apropiada el nivel de logro que demuestran los estudiantes en cada nivel escolar evaluado, así como estimar su progreso a medida que avanzan en su escolaridad. Para lograr lo anterior, se requiere hacer comparables los puntajes de las pruebas entre niveles y años de aplicación, permitiendo ubicar los resultados de cada nivel en una escala común. Esta comparabilidad de puntajes se hace posible a través del empleo de preguntas comunes entre niveles consecutivos y entre años (por ejemplo, la prueba SEPA de Matemática de 7° Básico incluye algunas preguntas que evalúan contenidos de 6° Básico que son medidos en ambas pruebas y, además, la prueba SEPA de Matemática de 7° Básico 2015 contiene preguntas de contenidos de 7° Básico que fueron utilizadas en este nivel el año 2014).

La escala SEPA es una escala vertical de puntajes que va aumentando de nivel en nivel, donde cada porcentaje de logro está asociado a un puntaje estándar determinado en cada nivel. Entiéndase por porcentaje de logro como el porcentaje de respuestas correctas que obtienen los estudiantes en la prueba. Las Figuras 1 y 2 muestran los puntajes estandarizados (en la escala

vertical) que obtendrían estudiantes que tuvieran un 50% de logro en cada prueba. Las figuras ilustran que a medida que aumenta el nivel escolar, aumenta también el puntaje estándar de las pruebas de Matemática y Lenguaje. Es importante advertir que la diferencia de puntajes estándar entre niveles no es constante, pues depende del desempeño demostrado por el conjunto de estudiantes que rinde cada nivel de las pruebas.

**Figura 1: Puntaje SEPA de Lenguaje asociado a un 50% de logro en cada nivel educacional**

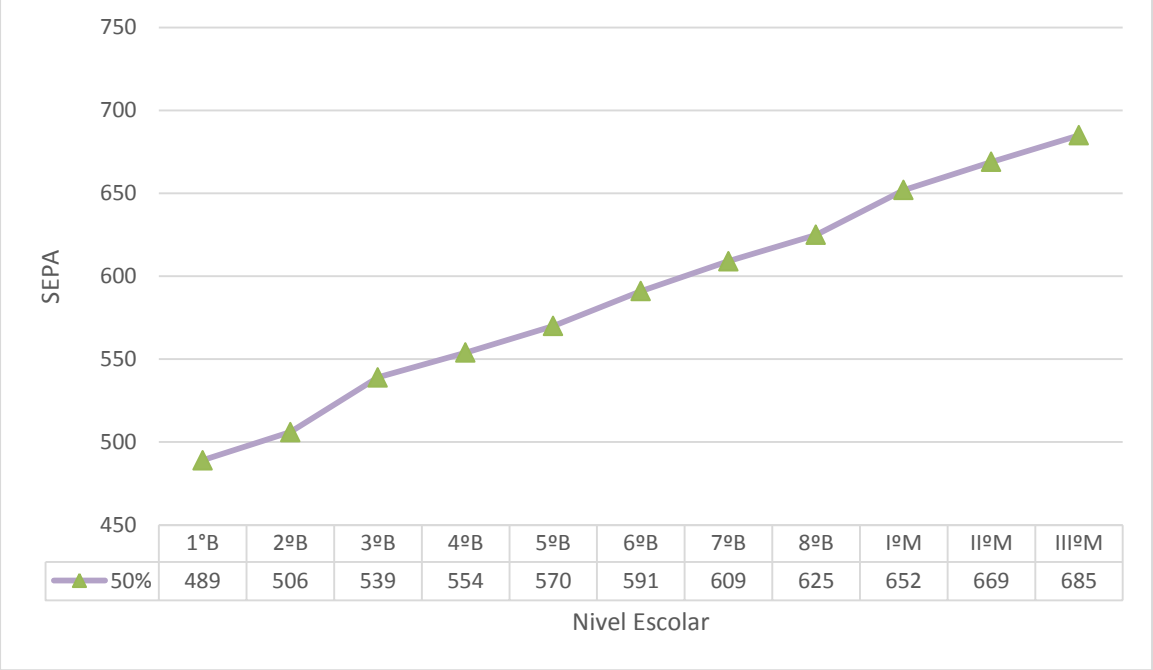
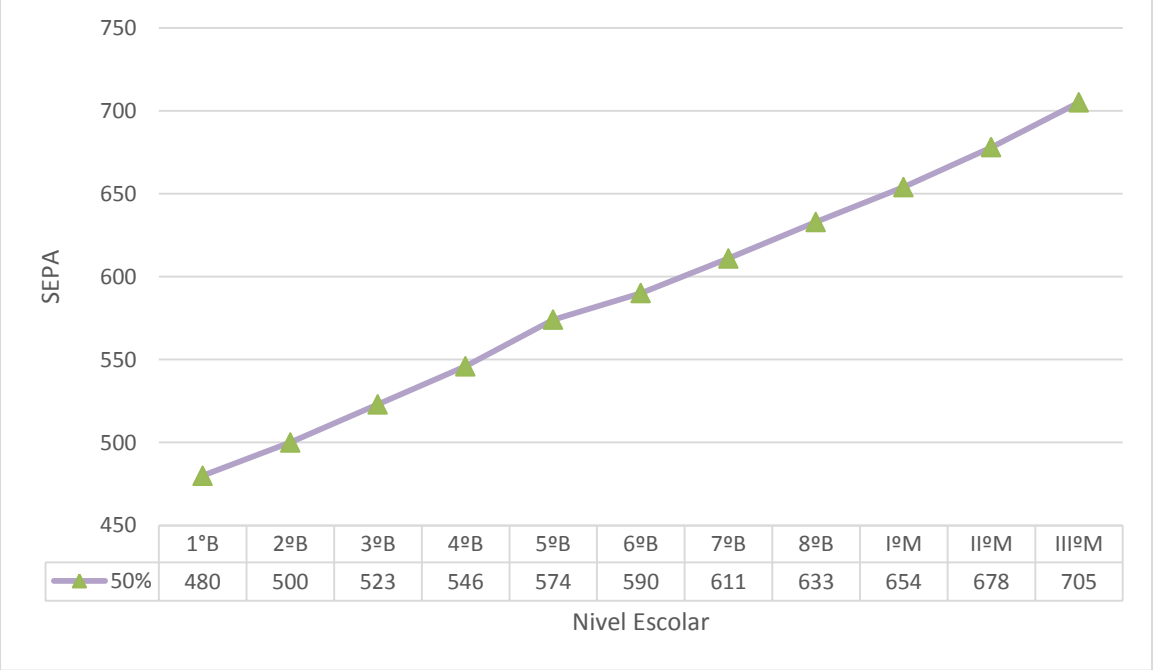


**Figura 2: Puntaje SEPA de Matemática asociado a un 50% de logro en cada nivel educacional**




**Población SEPA 2015**

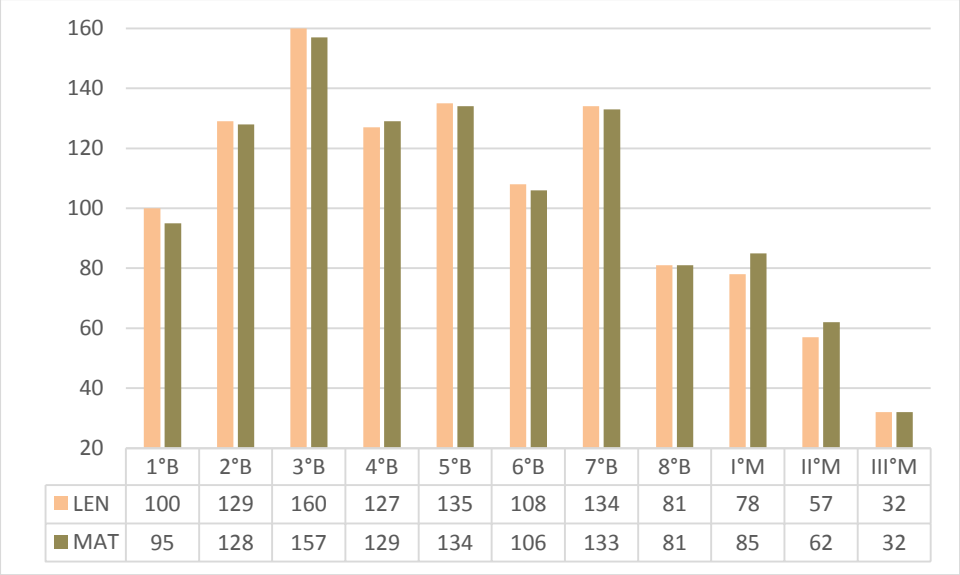
En SEPA 2015 se evaluó un total 227 establecimientos educacionales y 68,028 estudiantes desde 1° Básico hasta 3° Medio en al menos una de las dos pruebas SEPA, Lenguaje o Matemática. La Tabla 1 muestra la cantidad y proporción de establecimientos y estudiantes evaluados en ambas pruebas, observándose que, de los 227 establecimientos participantes, el 90,7% fue evaluado en Lenguaje, y el 96% en Matemática. Con respecto a los estudiantes, el 96,3% rindió la prueba de Lenguaje y el 94,9% la de Matemática, y el 91,2% rindió ambas pruebas.

**Tabla 1: cantidad de establecimientos educacionales y estudiantes evaluados en Pruebas SEPA**

| | Establecimientos evaluados | | Estudiantes evaluados | |
|---|---|---|---|---|
| **Asignatura** | **N** | **%** | **N** | **%** |
| Lenguaje | 206 | 90,7% | 65.481 | 96,3% |
| Matemática | 218 | 96% | 64.574 | 94,9% |

La Figura 3, muestra la cantidad de establecimientos que fueron evaluados por SEPA en Matemática y Lenguaje, en los distintos niveles, en el año 2015. En general, se observa una mayor cantidad de establecimientos evaluados en la enseñanza Básica, en comparación a los evaluados en la enseñanza Media, siendo 3° Básico el nivel en que se concentra la mayor cantidad de colegios evaluados, y III° Medio el nivel de menor participación.

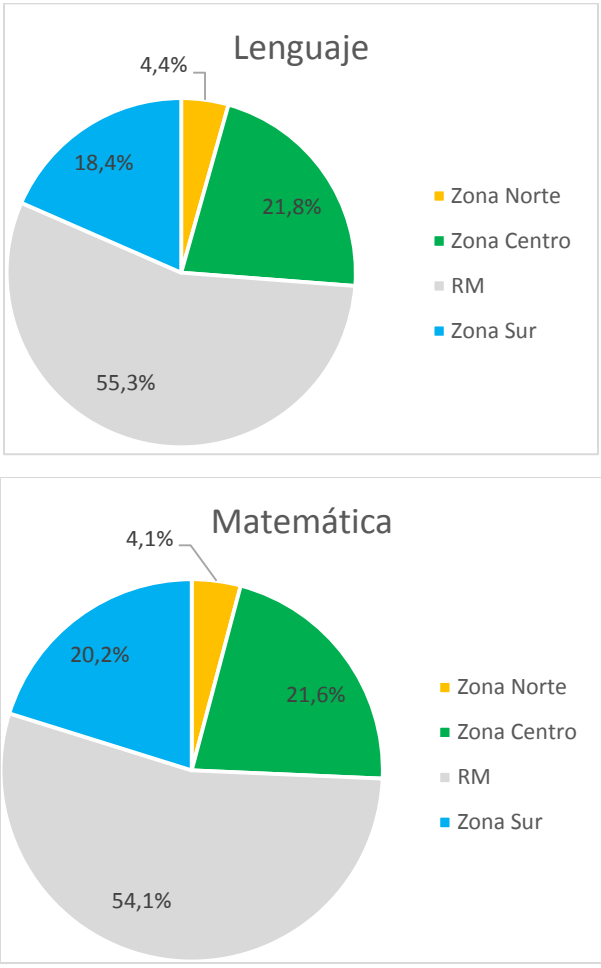
**Figura 3: Cantidad de establecimientos evaluados en Pruebas SEPA según nivel escolar**




Con respecto a la distribución geográfica de los establecimientos evaluados por SEPA en el año 2015, puede observarse que hay establecimientos educativos evaluados en trece de las quince regiones del país (las únicas regiones ausentes son las de Coquimbo y Arica-Parinacota). Como se aprecia en la Figura 4, la región Metropolitana de Santiago (RM), abarca el mayor

número de establecimientos SEPA con alrededor de un 55%. La sigue la zona central de Chile (Regiones de Valparaíso, del Libertador General Bernardo O'Higgins, del Maule y del Bío-Bío) con un 22%, junto a la zona sur (Regiones de la Araucanía, de Los Ríos, de Los Lagos, de Aysén del General Carlos Ibáñez del Campo y de Magallanes y de la Antártica Chilena) que representa un 19% de la población SEPA 2015. Finalmente, la zona norte (Regiones de Tarapacá, Antofagasta y Atacama) es la que representa la menor población evaluada, con alrededor del 4%.

**Figura 4: Distribución de establecimientos SEPA según región**




Por otra parte, con respecto a la proporción de establecimientos según su dependencia (ver Tabla 2), la población SEPA 2015 en su mayoría corresponde a establecimientos

municipales (alrededor de un 59%), seguida por colegios particulares pagados (23% aproximadamente) y finalmente los particulares subvencionados (cercanos al 17%).

**Tabla 2: Proporción de establecimientos evaluados según dependencia**

| Dependencia | Lenguaje | Matemática |
|---|---|---|
| MUN | 59,7% | 58,3% |
| PS | 16% | 18,8% |
| PP | 24,3% | 22,9% |

En cuanto a la proporción de estudiantes evaluados en cada dependencia, se observa que alrededor de un 43% pertenece a establecimientos municipales, un 35% a particulares pagados, y un 21% a particulares subvencionados (Tabla 3).

**Tabla 3: Proporción de estudiantes evaluados según dependencia del establecimiento**

| Dependencia | Lenguaje | Matemática |
|---|---|---|
| MUN | 43,2% | 43,6% |
| PS | 20,9% | 21,5% |
| PP | 35,9% | 34,8% |

Finalmente, con respecto a la distribución de establecimientos evaluados según grupo socioeconómico[8] (GSE), se observa que alrededor de un 48% son de un GSE bajo o medio bajo, cerca del 26% son de GSE medio, y el 26% restante corresponden a establecimientos de GSE medio alto o alto (Tabla 4).

[8] El grupo socio económico de los establecimientos educacionales fue imputado a partir de la información publicada por SIMCE 2013.

**Tabla 4: Proporción de establecimientos evaluados según grupo socioeconómico**

| GSE | Lenguaje | Matemática |
|---|---|---|
| Bajo | 19,3% | 20,3% |
| Medio Bajo | 28% | 28,2% |
| Medio | 26,4% | 25,9% |
| Medio Alto | 6,7% | 6,8% |
| Alto | 19,7% | 18,8% |

La proporción de estudiantes evaluados según el grupo socioeconómico de los establecimientos, muestra que alrededor del 46% son estudiantes de establecimientos de GSE Medio Alto o Alto, cercano a un 30% de los estudiantes son de establecimientos con GSE Medio, y el 24% restante son de establecimientos de GSE Bajo o Medio Bajo (Tabla 5).

**Tabla 5: Proporción de estudiantes evaluados según grupo socioeconómico**

| GSE | Lenguaje | Matemática |
|---|---|---|
| Bajo | 4,6% | 4,7% |
| Medio Bajo | 19,1% | 19,4% |
| Medio | 29,9% | 30,4% |
| Medio Alto | 10,3% | 10,5% |
| Alto | 36,1% | 35,1% |

**Estado del Aprendizaje**

Los resultados promedio obtenidos los estudiantes de la población SEPA 2015 (Figuras 5 y 6) muestran un incremento sostenido a medida que se aumenta en el nivel escolar. El comportamiento es parecido en ambos sectores, con un crecimiento similar entre pares de niveles y un rango de valores levemente superior en las pruebas de Matemática.

**Figura 5: Escala SEPA Lenguaje**

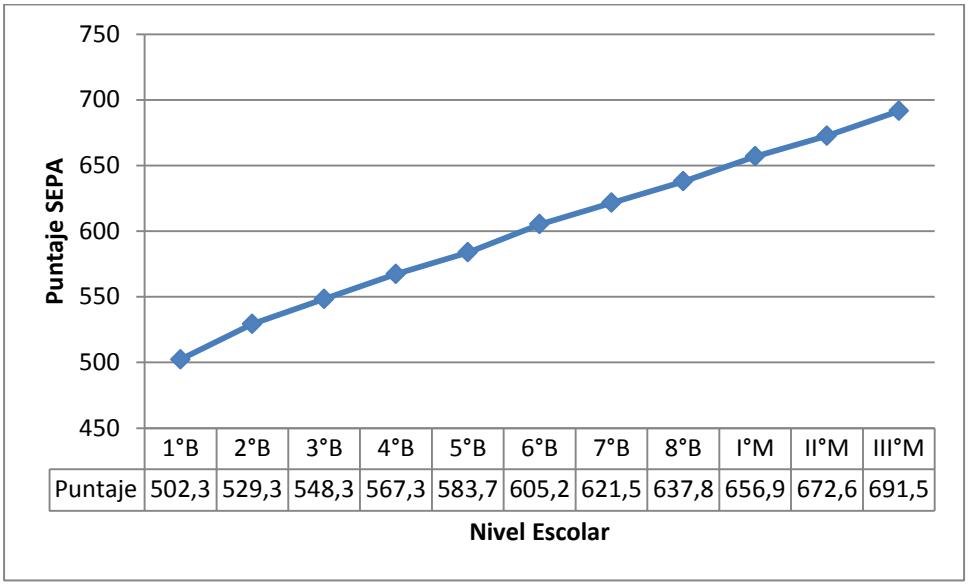
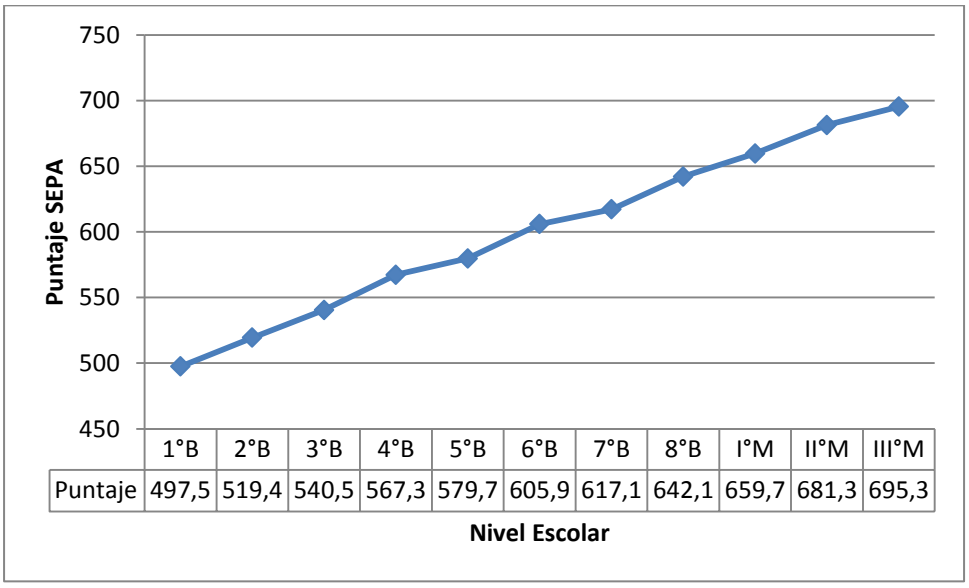



**Figura 6: Escala SEPA Matemática**

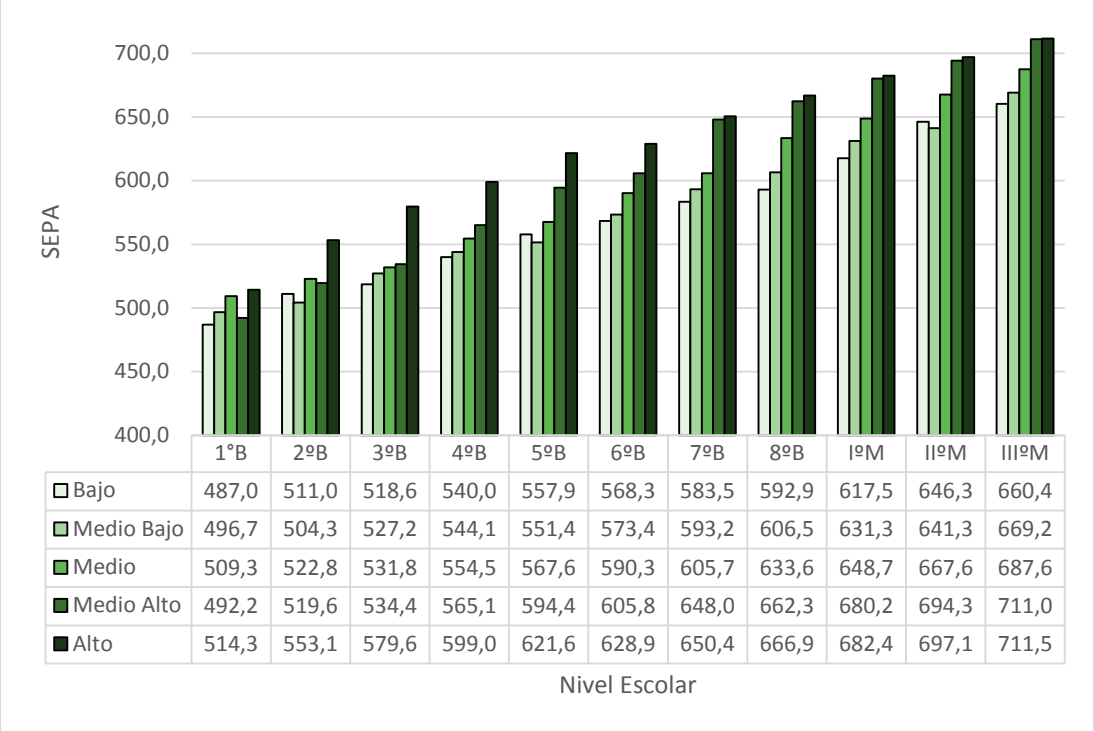



**Resultados Estado SEPA en los distintos niveles escolares por grupo socioeconómico**

Tanto en Lenguaje como en Matemática, los puntajes SEPA muestran diferencias relacionadas con el grupo socioeconómico del establecimiento: tal como se observa en todas las mediciones educacionales en Chile, mientras mayor es el grupo socioeconómico, mayor es el puntaje SEPA. Esta tendencia se observa más claramente desde 3° Básico en adelante.

Cabe agregar que el grupo socioeconómico alto muestra diferencias mayores con el siguiente grupo (medio-alto) entre segundo y sexto. Desde 7° en adelante los dos grupos superiores (alto y medio-alto) muestran un desempeño muy similar.

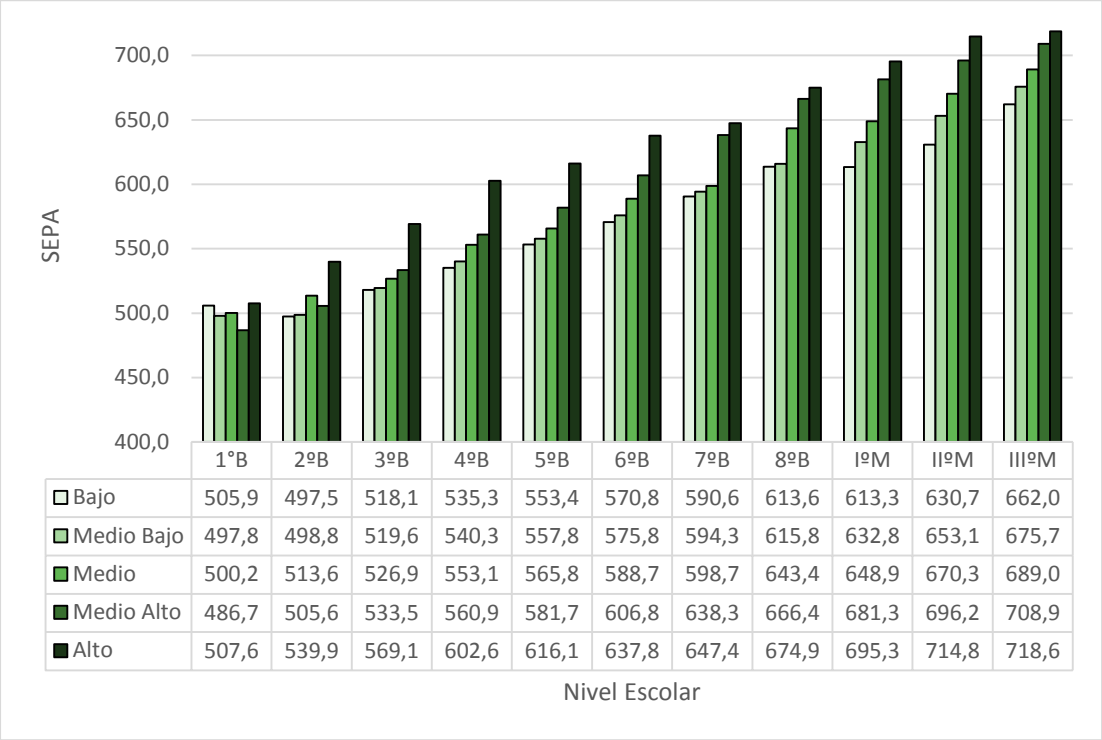
**Figura 7: Promedio en SEPA Lenguaje según grupo socioeconómico y nivel escolar**




| | 1°B | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bajo | 487,0 | 511,0 | 518,6 | 540,0 | 557,9 | 568,3 | 583,5 | 592,9 | 617,5 | 646,3 | 660,4 |
| Medio Bajo | 496,7 | 504,3 | 527,2 | 544,1 | 551,4 | 573,4 | 593,2 | 606,5 | 631,3 | 641,3 | 669,2 |
| Medio | 509,3 | 522,8 | 531,8 | 554,5 | 567,6 | 590,3 | 605,7 | 633,6 | 648,7 | 667,6 | 687,6 |
| Medio Alto | 492,2 | 519,6 | 534,4 | 565,1 | 594,4 | 605,8 | 648,0 | 662,3 | 680,2 | 694,3 | 711,0 |
| Alto | 514,3 | 553,1 | 579,6 | 599,0 | 621,6 | 628,9 | 650,4 | 666,9 | 682,4 | 697,1 | 711,5 |

Nivel Escolar

**Figura 8: Promedio en SEPA Matemática según grupo socioeconómico y nivel escolar**



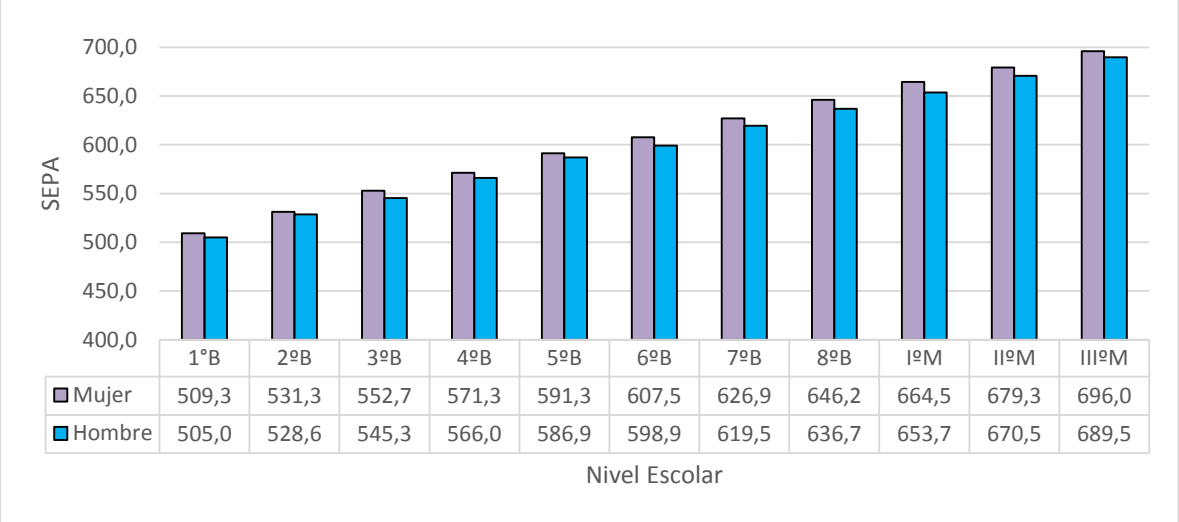

| | 1°B | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | IºM | IIºM | IIIºM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bajo | 505,9 | 497,5 | 518,1 | 535,3 | 553,4 | 570,8 | 590,6 | 613,6 | 613,3 | 630,7 | 662,0 |
| Medio Bajo | 497,8 | 498,8 | 519,6 | 540,3 | 557,8 | 575,8 | 594,3 | 615,8 | 632,8 | 653,1 | 675,7 |
| Medio | 500,2 | 513,6 | 526,9 | 553,1 | 565,8 | 588,7 | 598,7 | 643,4 | 648,9 | 670,3 | 689,0 |
| Medio Alto | 486,7 | 505,6 | 533,5 | 560,9 | 581,7 | 606,8 | 638,3 | 666,4 | 681,3 | 696,2 | 708,9 |
| Alto | 507,6 | 539,9 | 569,1 | 602,6 | 616,1 | 637,8 | 647,4 | 674,9 | 695,3 | 714,8 | 718,6 |

Nivel Escolar

**Resultados Estado SEPA en los distintos niveles educacionales por género**

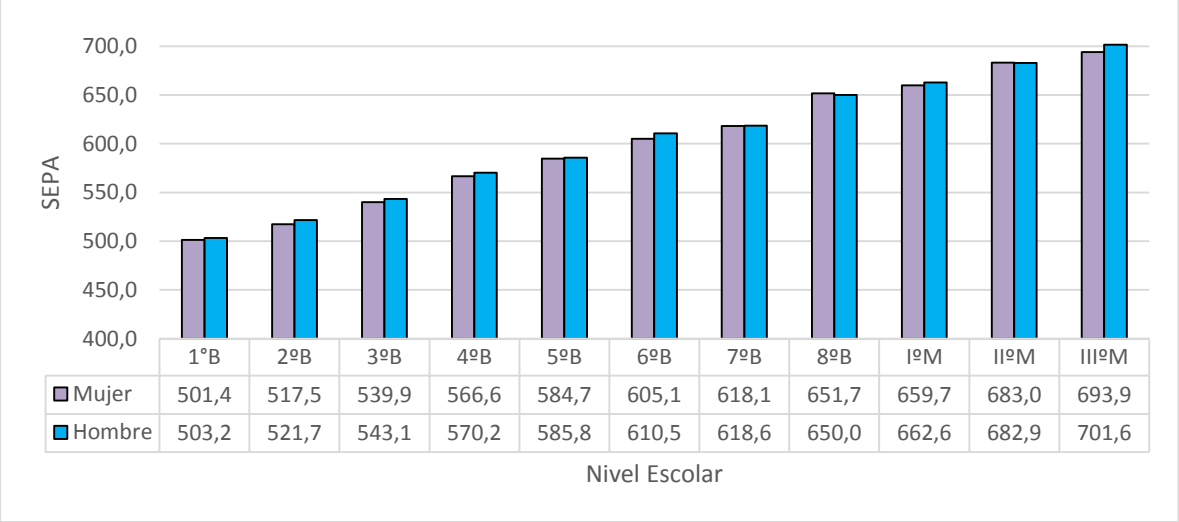
La Figura 9 presenta los resultados SEPA 2015 en Lenguaje por género de los estudiantes, donde los puntajes promedio de las mujeres son levemente mayor a los de los hombres a lo largo de todos los niveles escolares, la brecha fluctúa entre 10,8 puntos en I° Medio, y 2,7 en 1° Básico.

**Figura 9: Promedio en SEPA Lenguaje según género y nivel escolar**




En el caso de Matemática (Figura 10) ocurre el fenómeno contrario pues los resultados en general favorecen levemente a los hombres, aunque con una diferencia menor a la observada en lenguaje. Las diferencias entre los puntajes promedio varían entre 1,1 puntos en 5° Básico a 7,7 puntos en III° Medio.

**Figura 10: Promedio en SEPA Matemática según género y nivel escolar**



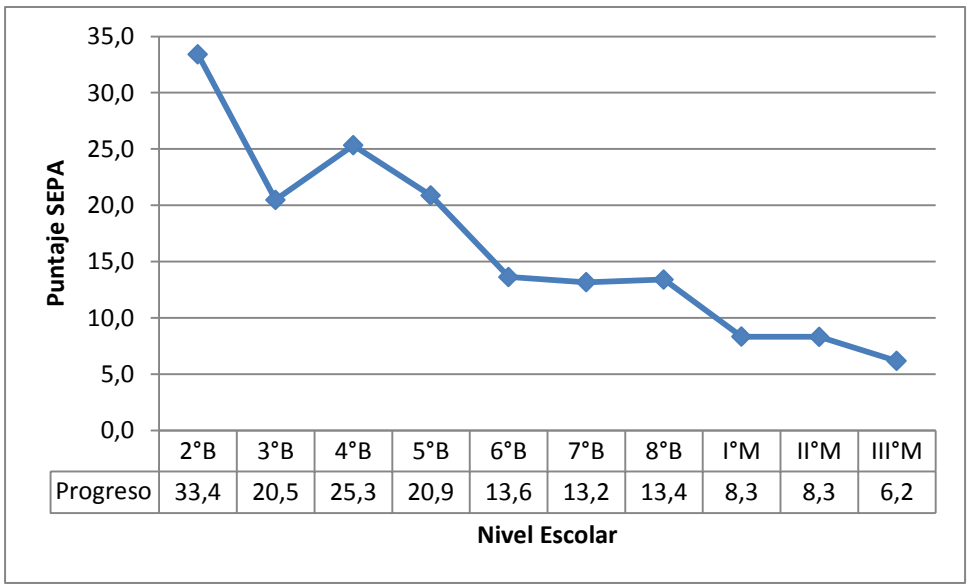
## Progreso del Aprendizaje[9]

Esta sección presenta el Progreso del aprendizaje de los estudiantes SEPA 2015 en relación a los mismos estudiantes que participaron en SEPA 2014.

## Resultados Progreso SEPA

Los progresos promedio obtenidos por los estudiantes de la población SEPA 2015 (Figuras 11 y 12) muestran una tendencia a la disminución conforme aumenta el nivel escolar, lo que se aprecia con mayor claridad en las pruebas de Lenguaje.

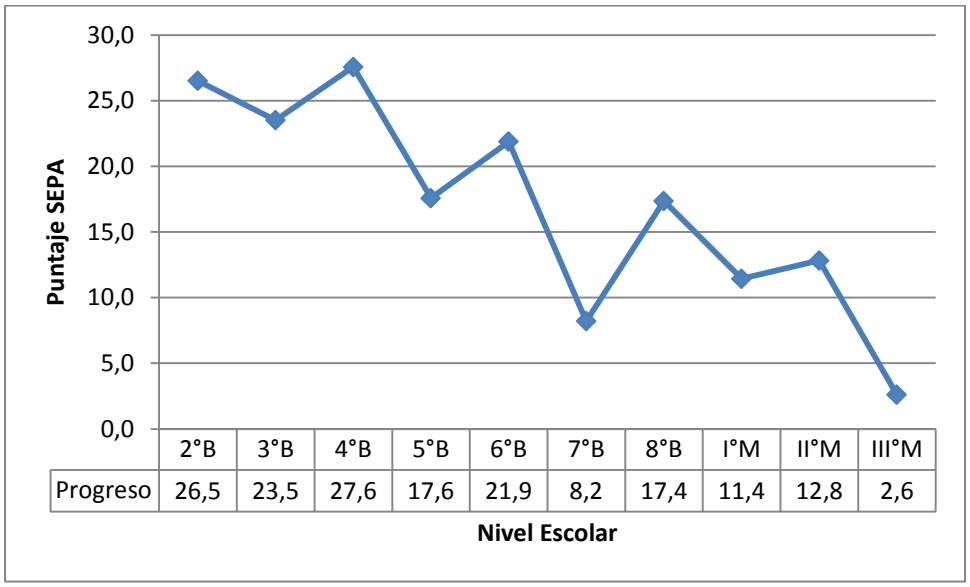
**Figura 11: Progreso SEPA Lenguaje**




| | 2°B | 3°B | 4°B | 5°B | 6°B | 7°B | 8°B | I°M | II°M | III°M |
|---|---|---|---|---|---|---|---|---|---|---|
| Progreso | 33,4 | 20,5 | 25,3 | 20,9 | 13,6 | 13,2 | 13,4 | 8,3 | 8,3 | 6,2 |

[9] Para efectos de esta sección, Progreso del Aprendizaje, cuando se hable de un nivel escolar, corresponde a estudiantes que participaron en SEPA 2015 en un nivel escolar y SEPA 2014 en el nivel anterior. Por ejemplo, si se habla de 2° Básico, se entiende que son estudiantes que están en 2° B el año 2015 y en 1° B el 2014.

**Figura 12: Progreso SEPA Matemática**




**Resultados Progreso SEPA en los distintos niveles educacionales por grupo socioeconómico**

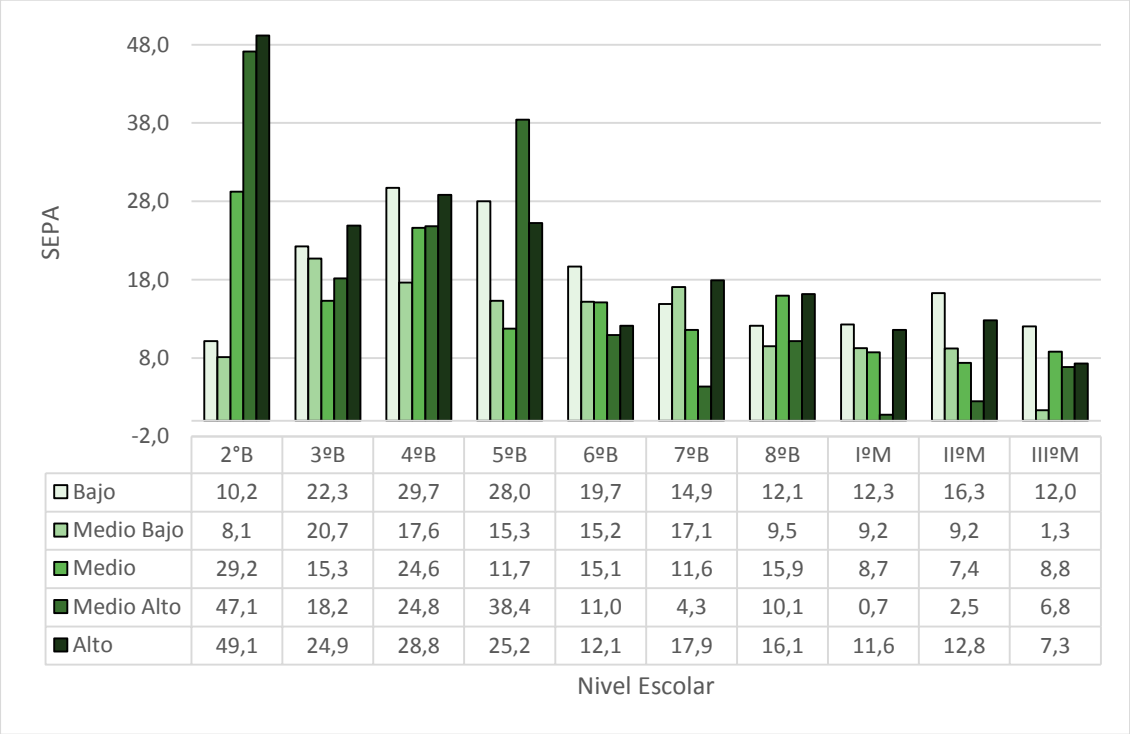
En general, no se observa que el Progreso a través de los niveles escolares difiera según el grupo socioeconómico. Esto quiere decir que en todos los grupos socioeconómicos el Progreso promedio disminuye a medida que se avanza en el nivel de escolaridades menor, independiente del grupo socioeconómico.

En la Figura 13 se observa que en 2° Básico es donde ocurre el mayor Progreso promedio en la prueba de Lenguaje, destacándose los establecimientos de GSE medio alto y alto que logran en promedio un importante Progreso de 47,1 y 49,1 puntos respectivamente, mientras que los establecimientos de nivel Medio Bajo obtienen un promedio de 8,1 puntos. Esta mayor brecha en los dos grupos socioeconómicos superiores probablemente refleja que el dominio de la lectura es más avanzado en establecimientos que atienden a estudiantes que provienen de hogares más favorecidos.

En cuanto al Progreso de Matemática (Figura 14), también se observan diferencias mayores entre los dos grupos socioeconómicos superiores y el bajo en los primeros años de escolaridad. A diferencia de lenguaje, donde la diferencia se concentra en el Progreso entre 1° y 2° Básico en Matemática, esto ocurre hasta 4° Básico. En contraste, en los dos cursos

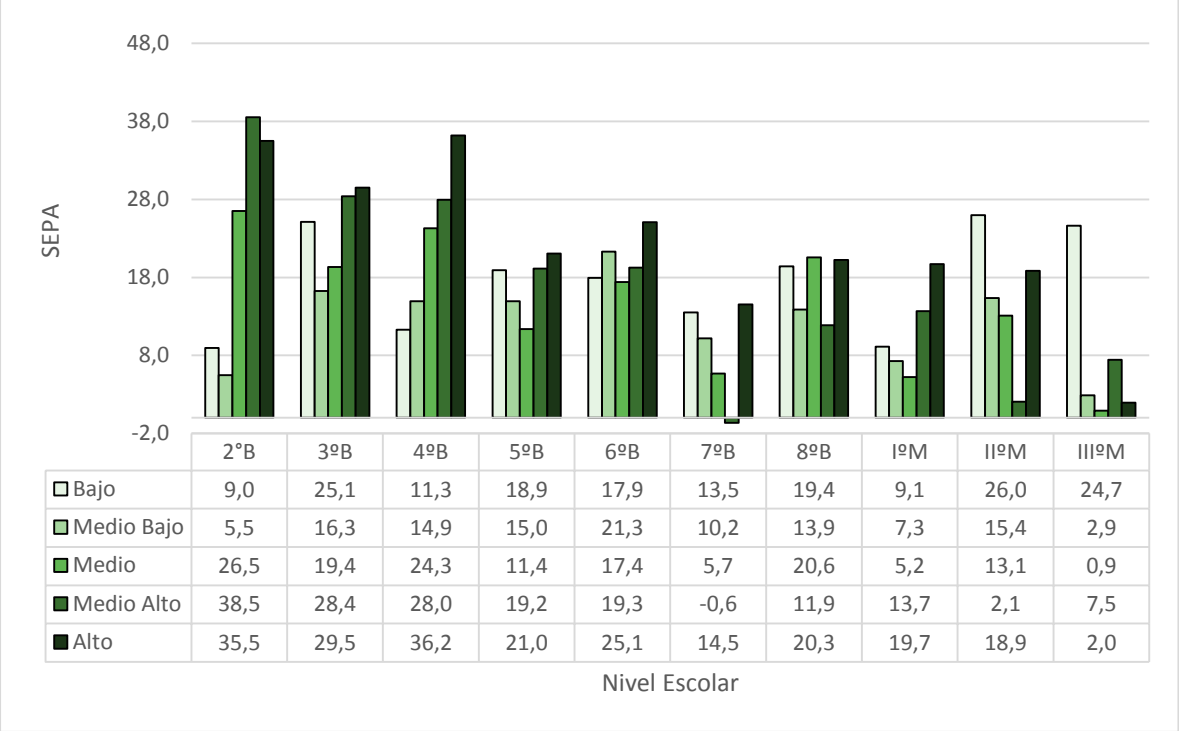
superiores de la enseñanza Media se constatan Progresos superiores en los grupos socioeconómicos bajos.

**Figura 13: Progreso promedio en SEPA Lenguaje según grupo socioeconómico y nivel escolar**




| | 2°B | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | IºM | IIºM | IIIºM |
|---|---|---|---|---|---|---|---|---|---|---|
| Bajo | 10,2 | 22,3 | 29,7 | 28,0 | 19,7 | 14,9 | 12,1 | 12,3 | 16,3 | 12,0 |
| Medio Bajo | 8,1 | 20,7 | 17,6 | 15,3 | 15,2 | 17,1 | 9,5 | 9,2 | 9,2 | 1,3 |
| Medio | 29,2 | 15,3 | 24,6 | 11,7 | 15,1 | 11,6 | 15,9 | 8,7 | 7,4 | 8,8 |
| Medio Alto | 47,1 | 18,2 | 24,8 | 38,4 | 11,0 | 4,3 | 10,1 | 0,7 | 2,5 | 6,8 |
| Alto | 49,1 | 24,9 | 28,8 | 25,2 | 12,1 | 17,9 | 16,1 | 11,6 | 12,8 | 7,3 |

**Figura 14: Progreso promedio en SEPA Matemática según grupo socioeconómico y nivel escolar**



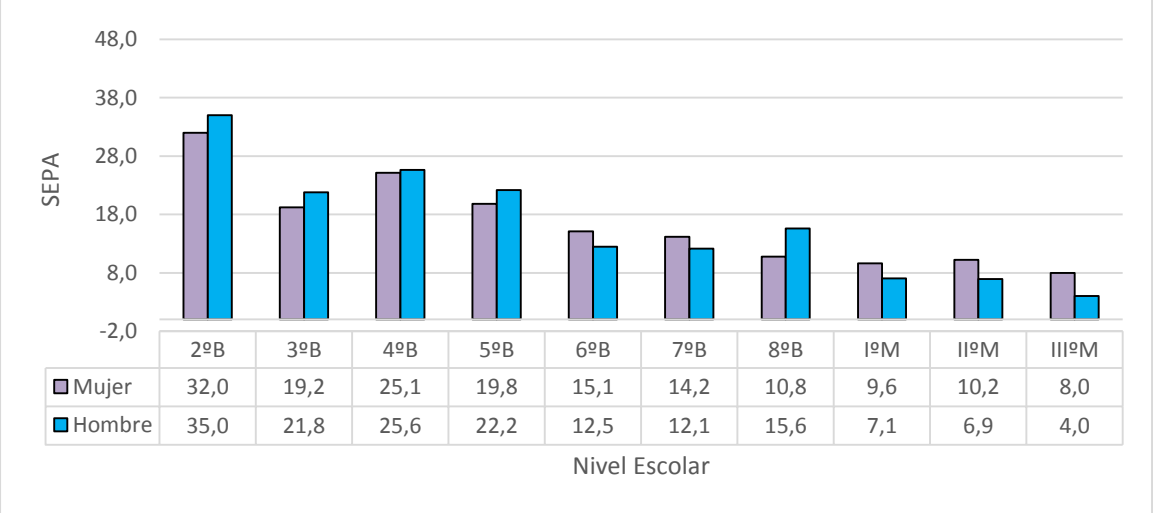

| | 2°B | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | IºM | IIºM | IIIºM |
|---|---|---|---|---|---|---|---|---|---|---|
| Bajo | 9,0 | 25,1 | 11,3 | 18,9 | 17,9 | 13,5 | 19,4 | 9,1 | 26,0 | 24,7 |
| Medio Bajo | 5,5 | 16,3 | 14,9 | 15,0 | 21,3 | 10,2 | 13,9 | 7,3 | 15,4 | 2,9 |
| Medio | 26,5 | 19,4 | 24,3 | 11,4 | 17,4 | 5,7 | 20,6 | 5,2 | 13,1 | 0,9 |
| Medio Alto | 38,5 | 28,4 | 28,0 | 19,2 | 19,3 | -0,6 | 11,9 | 13,7 | 2,1 | 7,5 |
| Alto | 35,5 | 29,5 | 36,2 | 21,0 | 25,1 | 14,5 | 20,3 | 19,7 | 18,9 | 2,0 |

Nivel Escolar

**Resultados Progreso SEPA en los distintos niveles educacionales por género**

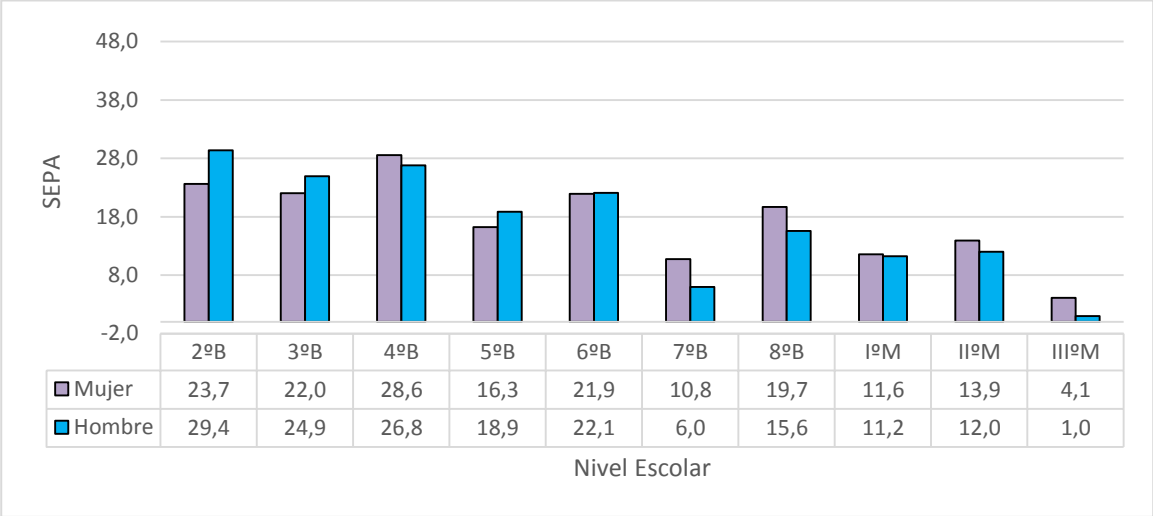
En ambas pruebas se observa que, en general, los hombres tienen mejor desempeño que las mujeres en los primeros niveles, tendencia que se invierte a partir de 6° Básico en Lenguaje y 7° Básico en Matemática, en que las mujeres obtienen en promedio mejores resultados en los niveles superiores (Figura 15 y Figura 16). La diferencia entre hombres y mujeres varía entre 0,5 y 3,93 puntos en Lenguaje y entre 0,2 y 5,7 en Matemática.

**Figura 15: Progreso promedio en SEPA Lenguaje según género y nivel escolar**




| | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|
| Mujer | 32,0 | 19,2 | 25,1 | 19,8 | 15,1 | 14,2 | 10,8 | 9,6 | 10,2 | 8,0 |
| Hombre | 35,0 | 21,8 | 25,6 | 22,2 | 12,5 | 12,1 | 15,6 | 7,1 | 6,9 | 4,0 |

**Figura 16: Progreso promedio en SEPA Matemática según género y nivel escolar**




| | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | Iº M | IIº M | IIIº M |
|---|---|---|---|---|---|---|---|---|---|---|
| Mujer | 23,7 | 22,0 | 28,6 | 16,3 | 21,9 | 10,8 | 19,7 | 11,6 | 13,9 | 4,1 |
| Hombre | 29,4 | 24,9 | 26,8 | 18,9 | 22,1 | 6,0 | 15,6 | 11,2 | 12,0 | 1,0 |

## Valor Agregado

Valor Agregado (en adelante VA) es una estimación del aporte que realiza el establecimiento al aprendizaje de sus estudiantes, permitiendo identificar en qué medida los aprendizajes que alcanza cada colegio se deben a la labor de la escuela, en contraposición a logros que no dependen de su labor pedagógica (por ejemplo, asociados al capital cultural o nivel socioeconómico de los estudiantes que atiende). Esta medida permite que los establecimientos cuenten con un índice que les permita reflexionar acerca de las prácticas que resultan más y menos efectivas para aportar sustantivamente a los aprendizajes de sus estudiantes.

El índice de VA se obtiene a partir de un análisis estadístico de los resultados logrados por un mismo grupo de estudiantes en dos mediciones sucesivas (ver Capítulo 3). En los resultados que aquí se exponen, los estudiantes rindieron SEPA en el año 2014 y 2015.

VA se reporta para cada nivel en una escala en la cual el promedio está fijado en 10. Los índices que estén 2 puntos o más sobre 10, indican que el establecimiento hace un aporte mayor que un "establecimiento promedio" a los aprendizajes de sus estudiantes. Asimismo, índices de VA que estén 2 puntos o más por debajo de 10, indican que el establecimiento hace un aporte menor que un "establecimiento promedio" a los aprendizajes de sus estudiantes.

## Relación entre Valor Agregado e información de Estado en SEPA

Por su descrito, cabe esperar que la estimación del Valor Agregado aporte información diferente de la que se refleja en el puntaje de Estado de las pruebas. Esta última representa una estimación del nivel de logro de estudiantes y escuelas, sin que sea posible establecer el grado en que tales puntajes responden a la acción educativa de la escuela. En consecuencia, se espera que la correlación entre la información de Estado y la de Valor Agregado sea baja. Esto es lo que se constata en las Figuras 17 y 18, que representan dicha relación para las dos pruebas. En estas Figuras cada punto corresponde a la estimación de Estado y Valor Agregado para cada nivel en que participa un establecimiento escolar que haya aplicado las pruebas SEPA en dos años sucesivos. La relación que se verifica en ambos casos es tenue (correspondiente a una correlación de 0,28 en Lenguaje y 0,26 en Matemática), lo que reafirma que un establecimiento puede presentar altos o bajos niveles de Valor Agregado tanto cuando sus puntajes de Estado son bajos como altos.

**Figura 17: Gráfico de dispersión correspondiente a la relación entre puntajes de Estado y Valor Agregado en Lenguaje**

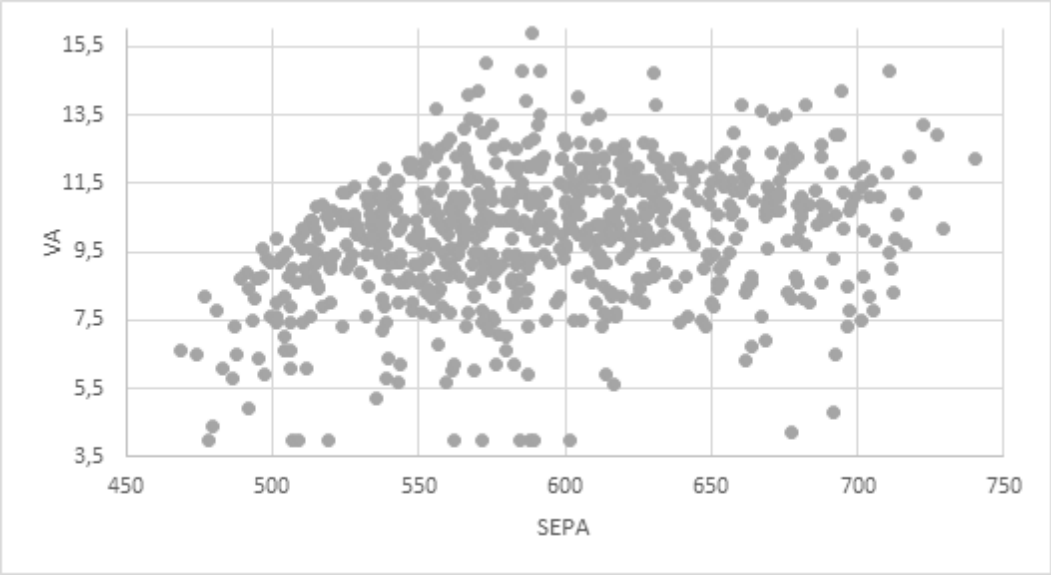
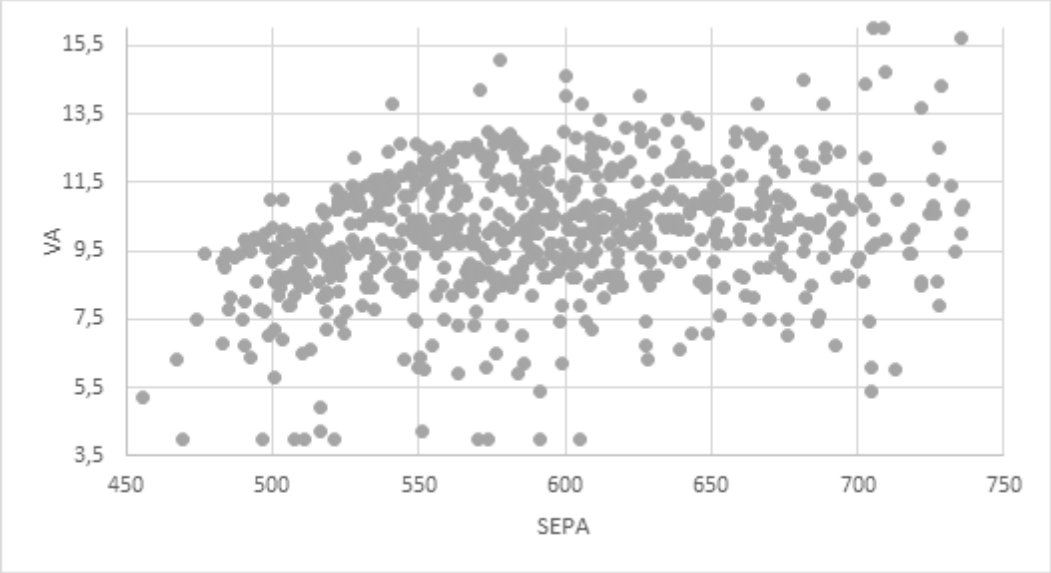



**Figura 1814: Gráfico de dispersión correspondiente a la relación entre puntajes de Estado y Valor Agregado en Matemática**



**Relación entre Valor Agregado e información de Progreso en SEPA**

Por otra parte, las Figuras 19 y 20, presentan los resultados del VA de los establecimientos según los resultados promedio de su Progreso en Lenguaje y Matemática. En este caso se observa que existe una clara relación lineal entre ambos tipos de puntaje. Esto quiere decir que establecimientos que obtienen Progresos promedios altos tienden a tener mayor VA. La correlación es en este caso de 0,81 en Lenguaje y 0,84 en Matemática. Esta mayor correlación, en comparación con la que involucra a la información de Estado, es esperable, precisamente porque el Progreso es la base de la información que se analiza cuando se lleva a cabo la estimación del Valor Agregado.

**Figura 19: Gráfico de dispersión correspondiente a la relación entre puntajes de Progreso y Valor Agregado en Lenguaje**

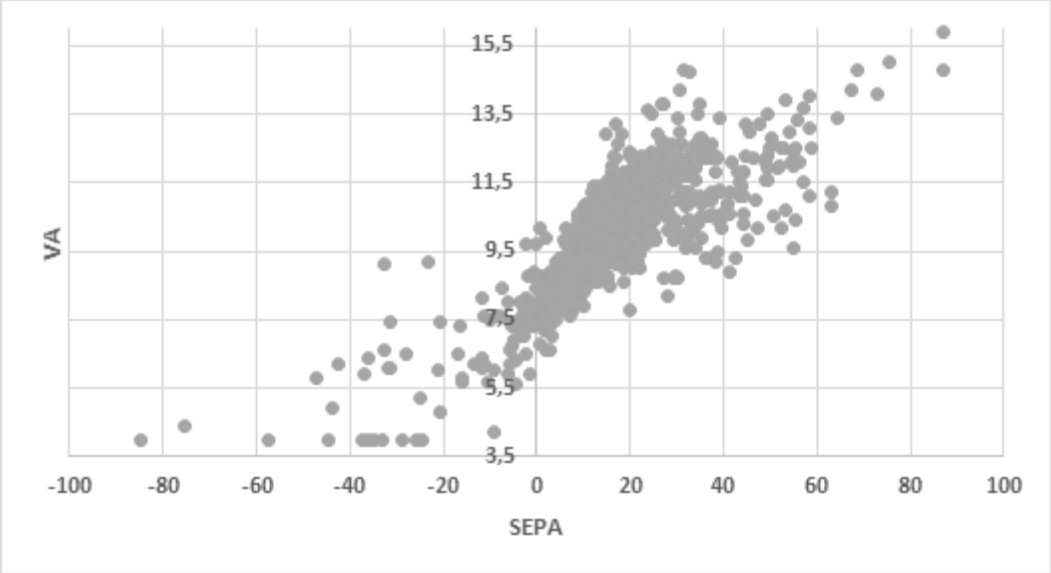
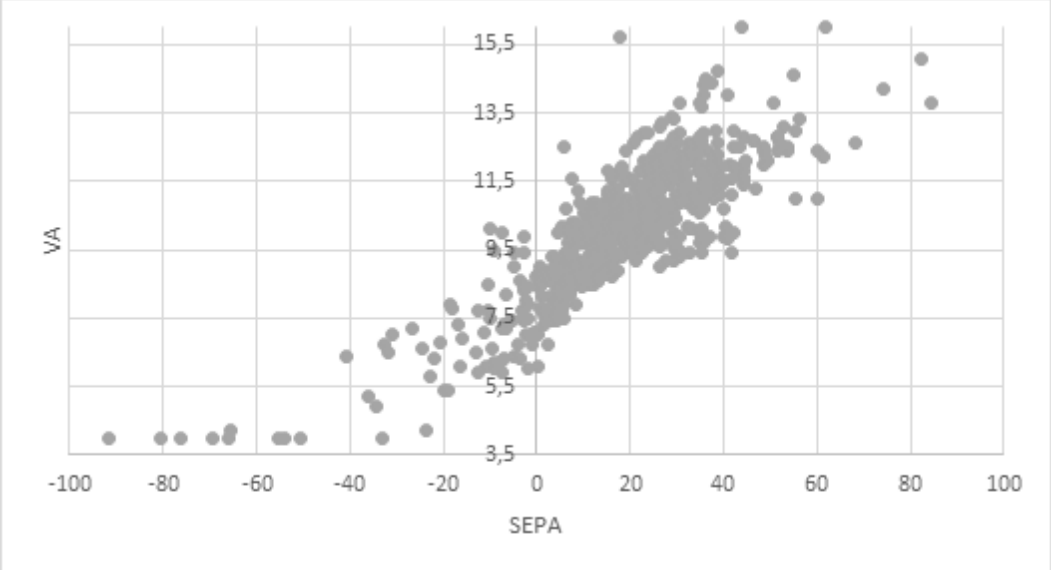


**Figura 20: Gráfico de dispersión correspondiente a la relación entre puntajes de Progreso y Valor Agregado en Matemática**




**Relación entre Valor Agregado y condición socioeconómica de los establecimientos educacionales**

Por último, se presenta información acerca de la distribución de los puntajes de Valor Agregado según la condición socioeconómica de los establecimientos. Es sabido que las puntuaciones de Estado (como las de SEPA o SIMCE) muestran una alta correlación con la condición socioeconómica de los establecimientos. Sin embargo, dado que el Valor Agregado intenta establecer la efectividad de cada establecimiento a partir de las condiciones de entrada que presentan los alumnos, ello debiera conducir a una relación muy atenuada con la condición socioeconómica. Ello es lo que se constata en las Figuras 21 y 22, donde la mayor parte de los grupos socioeconómicos muestran niveles equivalentes de Valor Agregado. La principal excepción es el grupo socioeconómico alto, cuyo Valor Agregado medio es significativamente superior al de los otros grupos, aunque la diferencia es claramente inferior a la que se observa en la información de Estado.

**Figura 21: Comparación de las estimaciones de Valor Agregado según la condición socioeconómica de los establecimientos escolares en la prueba de Lenguaje**

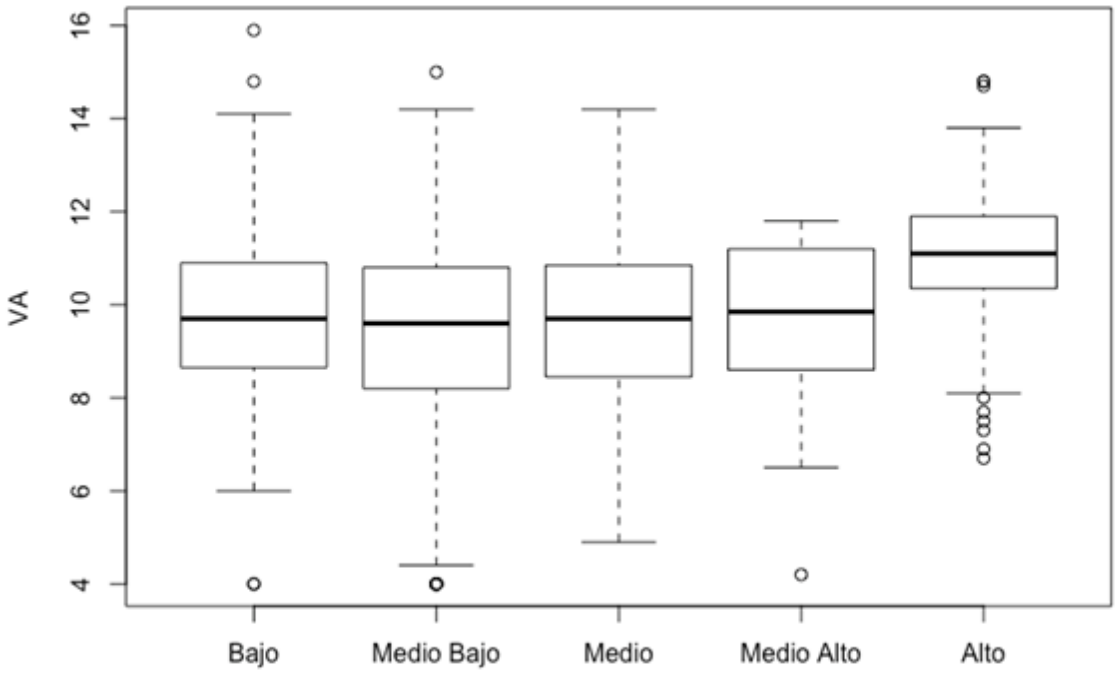
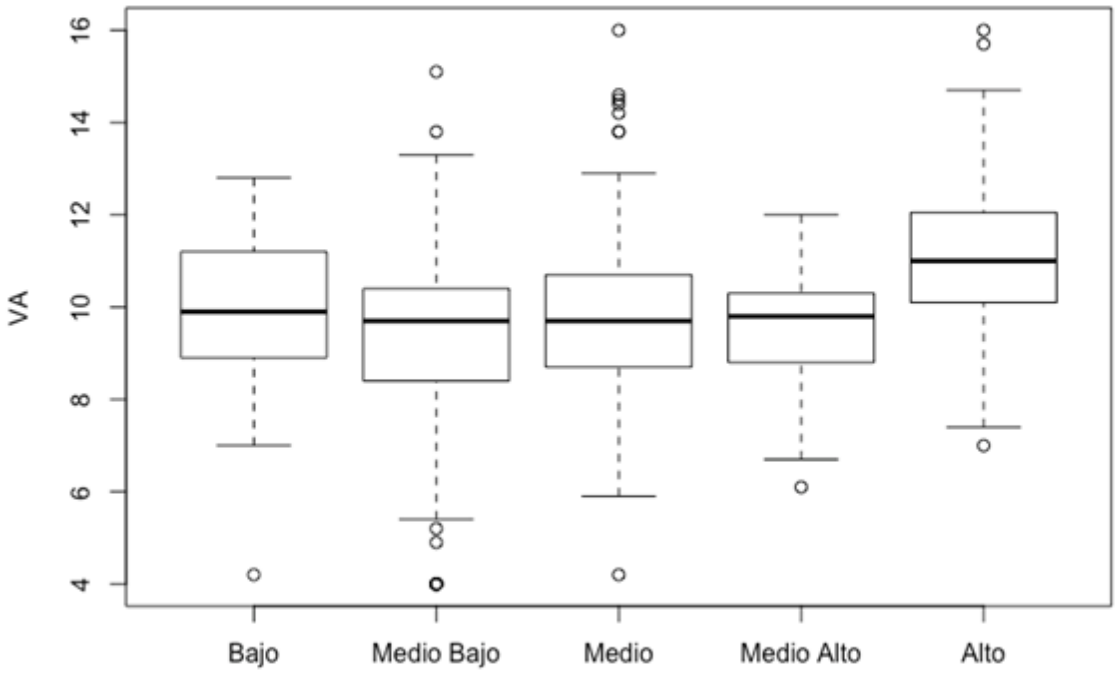


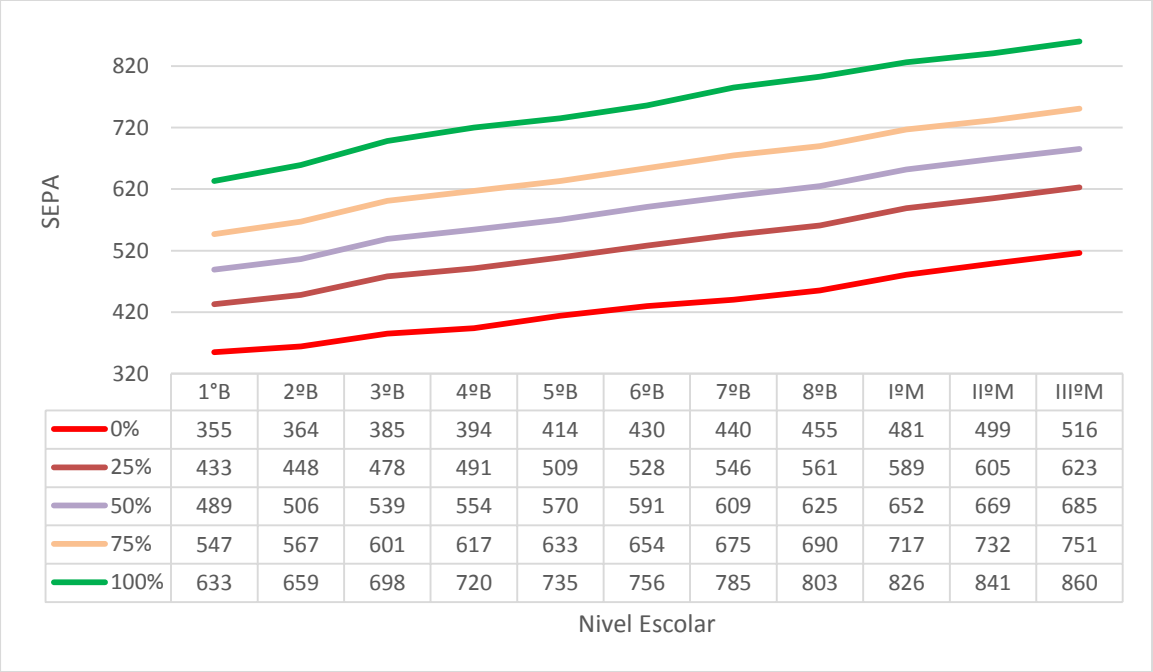
**Figura 22: Comparación de las estimaciones de Valor Agregado según la condición socioeconómica de los establecimientos escolares en la prueba de Matemática**



## Anexos

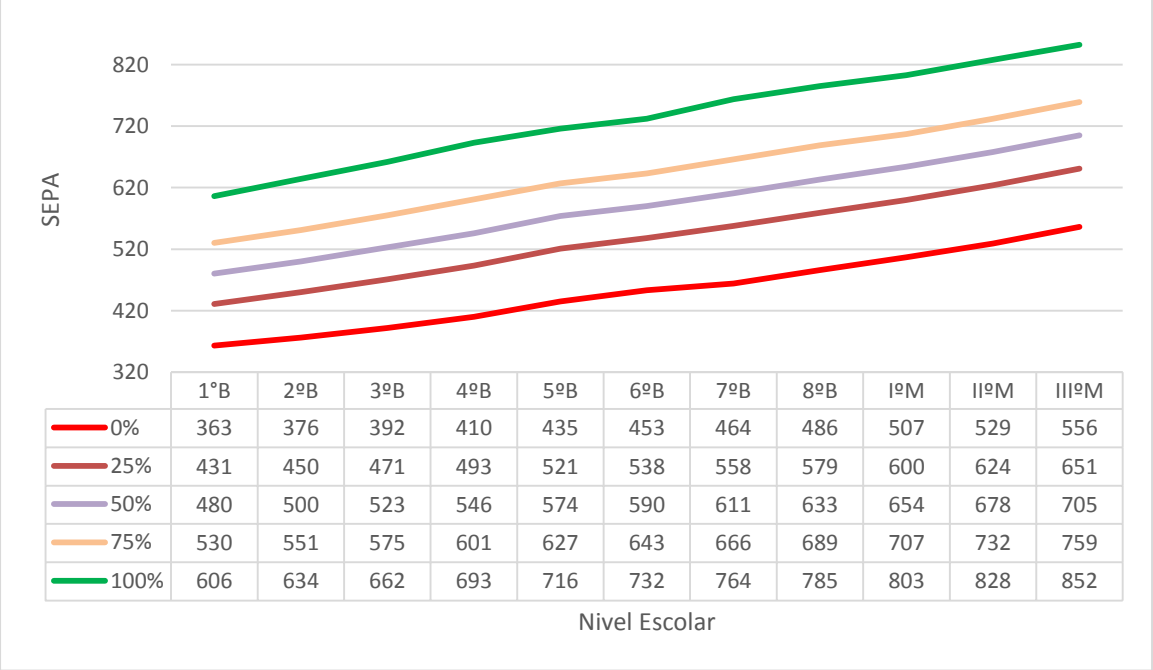
### Anexo 1: Puntajes SEPA para distintos porcentajes de logro

**Figura 23: Comportamiento de los puntajes SEPA 2015 de Lenguaje por nivel educacional**




| | 1°B | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | IºM | IIºM | IIIºM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0% | 355 | 364 | 385 | 394 | 414 | 430 | 440 | 455 | 481 | 499 | 516 |
| 25% | 433 | 448 | 478 | 491 | 509 | 528 | 546 | 561 | 589 | 605 | 623 |
| 50% | 489 | 506 | 539 | 554 | 570 | 591 | 609 | 625 | 652 | 669 | 685 |
| 75% | 547 | 567 | 601 | 617 | 633 | 654 | 675 | 690 | 717 | 732 | 751 |
| 100% | 633 | 659 | 698 | 720 | 735 | 756 | 785 | 803 | 826 | 841 | 860 |

**Figura 24: Comportamiento de los puntajes SEPA 2015 de Matemática por nivel educacional**




| | 1°B | 2ºB | 3ºB | 4ºB | 5ºB | 6ºB | 7ºB | 8ºB | IºM | IIºM | IIIºM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0% | 363 | 376 | 392 | 410 | 435 | 453 | 464 | 486 | 507 | 529 | 556 |
| 25% | 431 | 450 | 471 | 493 | 521 | 538 | 558 | 579 | 600 | 624 | 651 |
| 50% | 480 | 500 | 523 | 546 | 574 | 590 | 611 | 633 | 654 | 678 | 705 |
| 75% | 530 | 551 | 575 | 601 | 627 | 643 | 666 | 689 | 707 | 732 | 759 |
| 100% | 606 | 634 | 662 | 693 | 716 | 732 | 764 | 785 | 803 | 828 | 852 |

# CAPÍTULO V: EVIDENCIA SOBRE LA VALIDEZ DE LAS PRUEBAS SEPA

**Andrea Abarzúa**

Psicóloga y magíster en Psicología Educacional de la Pontificia Universidad Católica de Chile. Actualmente se desempeña como Coordinadora de la Unidad de Análisis de MIDE UC, y como docente en el área de formación. (raabarzu@uc.cl)

En este capítulo se presenta un conjunto de análisis acerca de la validez de las Pruebas SEPA con el objetivo de aportar evidencia acerca de calidad psicométrica de estas. Estos análisis se basan en los estándares internacionales para la medición psicológica y educacional (AERA, APA & NCME, 2014). Este conjunto de estándares establece los siguientes criterios orientadores para juzgar la calidad de un instrumento de medición: confiabilidad, validez y justicia. El cumplimiento de estos estándares supone la realización de estudios regulares que permitan acumular evidencia acerca de estos tres criterios.

La confiabilidad, originalmente referida a la replicabilidad de una medición, es una propiedad de los puntajes obtenidos de un test y puede ser entendida como el grado de precisión de los resultados de determinado instrumento (Cook & Beckman, 2006). Una de las formas más usuales de evaluar la confiabilidad de un instrumento es estimar su consistencia interna (empleando el coeficiente alfa de Cronbach). Dicha evidencia aparece en el Capítulo 3; en esta sección se abordará la confiabilidad desde la perspectiva de la precisión de la información que entregan los puntajes de la escala vertical de las pruebas SEPA, como perspectiva complementaria a estas medidas de consistencia interna.

Por su parte, la validez refiere al grado en que la evidencia y la teoría apoyan las interpretaciones de los puntajes entregados por un test para un determinado propósito o uso (AERA, APA & NCME, 2014). La validez no debe ser entendida como un concepto dicotómico (en términos de si se cumple o no), sino que debe ser argumentada y juzgada a partir de diversos tipos de evidencia, recabadas tanto antes como después de la medición, con la perspectiva de conformar un argumento convincente acerca de un uso particular (Kane, 2013; AERA, APA & NCME, 2014). En esta sección se revisará la estructura interna del instrumento, a través de un análisis de sus componentes principales, en los que específicamente se revisará el grado en que los ítems de una prueba guardan relación con un constructo sobre el cual se deberían basar los puntajes de los examinados (AERA, APA & NCME, 2014). Al mismo tiempo, se desarrollarán otros argumentos comparando el instrumento con otros que aborden el mismo constructo, realizando un análisis correlacional.

La justicia de una medición (*fairness*) refiere al grado en que un instrumento es capaz de entregar resultados que no se encuentran influenciados por características de los examinados que sean diferentes de lo que se busca medir. Estas pueden incluir el género, la condición socioeconómica, la pertenencia a una minoría étnica, pertenecer a una comunidad lingüística diferente de la mayoritaria, etc. Dado que la experiencia en otros programas de medición nacionales (como SIMCE y PSU) han mostrado que el género es potencialmente una de las fuentes más relevantes de sesgo de medición (incluso más que la condición socioeconómica), en el caso de las pruebas SEPA, se emplearon técnicas para detectar sesgo de género a nivel de ítems y pruebas según el género de los estudiantes.

**Precisión de la información en las pruebas SEPA**

En el capítulo de análisis de datos ya se entregó información acerca de la confiabilidad de los puntajes obtenidos por los examinados en las pruebas SEPA, estimada mediante el índice alfa de Cronbach, la que muestra para todas las pruebas índices satisfactorios. En el caso de las pruebas SEPA, donde la estimación de los puntajes de los examinados se hace mediante un modelamiento propio de la Teoría de Respuesta al Ítem (TRI), la estimación de confiabilidad puede ser complementada con índices propios del enfoque TRI. En particular, analizamos la función de información resultante, que nos indica el grado de precisión de las puntuaciones en distintas regiones de puntaje. Típicamente un instrumento posee diferente capacidad informativa para puntuaciones bajas, medias o altas, debido a que como es usual la mayor parte de los ítems son de dificultad intermedia (lo que se traduce en mayor información en la región media). Dado que la precisión es diferencial, el error de medición no es de magnitud constante, como se asumía en la teoría clásica de la medición (Lord, 1984; Manzi & San Martín, 2003).

Con este tipo de análisis, es además posible estimar el grado de correspondencia entre la precisión de un instrumento y la distribución de habilidades del grupo que es evaluado con dicho instrumento (lo deseable es que el instrumento posea mayor precisión o información en la zona donde se concentra la mayor parte de los examinados). Lo anterior es posible debido a que con el modelamiento TRI se puede obtener un mapa del constructo latente en el que ubica a personas e ítems en una misma escala (de Ayala, 2009). Este mapeo, que en la práctica se

traduce en un ordenamiento de personas e ítems en torno al constructo medido, permite establecer de forma más precisa la dificultad de los ítems y habilidad de las personas de lo que lograríamos haciendo uso de técnicas exclusivamente basadas en el marco de la teoría clásica[10]. En el caso de las pruebas SEPA, el constructo latente es, para cada prueba, los conocimientos en Lenguaje y Matemática entendidos como cualitativamente distintos para cada nivel de enseñanza, y a la vez articulados de tal forma que permiten mostrar una progresión a lo largo de estos niveles.

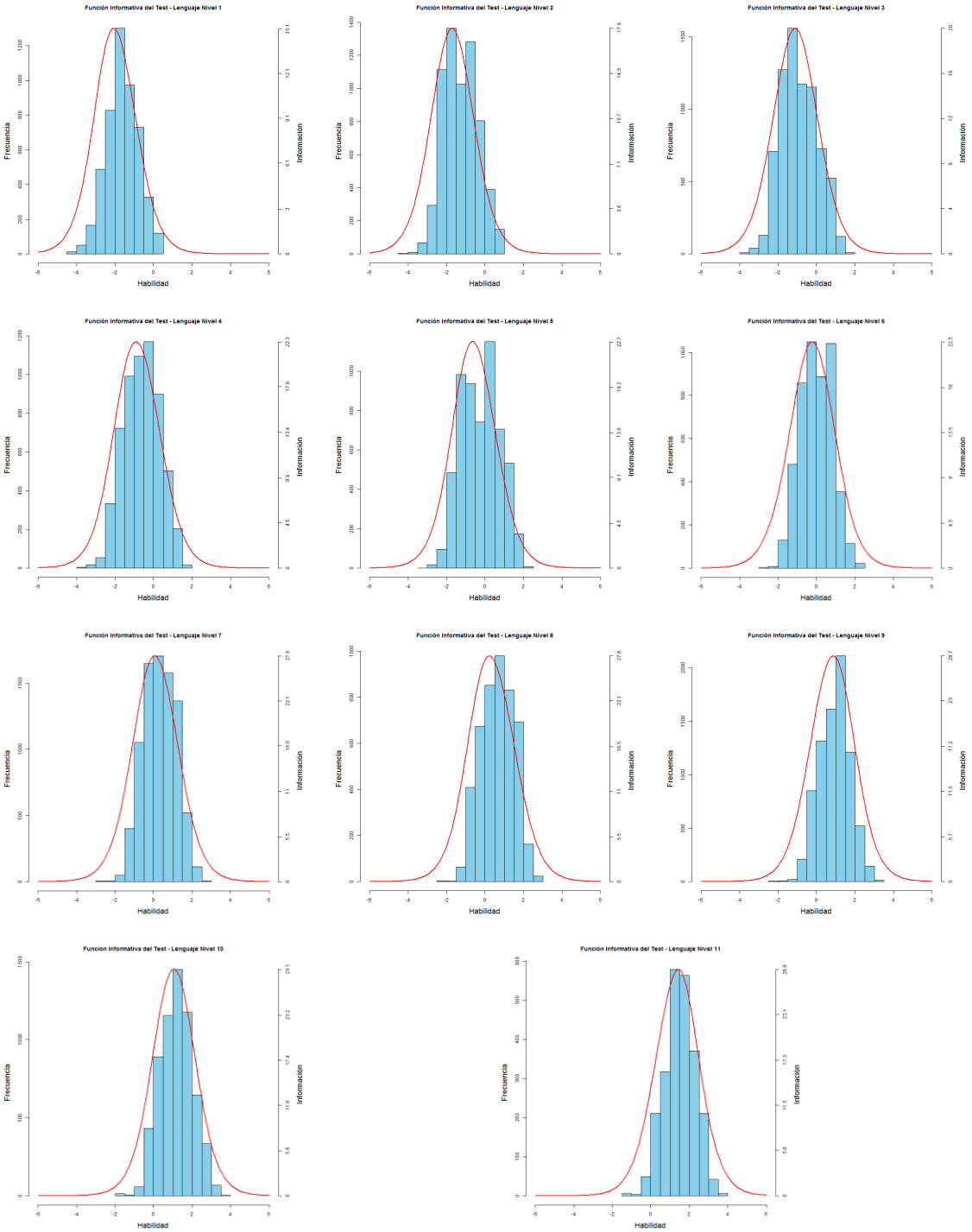
Dado el uso de estas técnicas de análisis, las pruebas SEPA tienen una escala de puntajes que se sustenta sobre la información que entrega cada uno de los ítems en cada uno de los once niveles medidos. La información que aportan los ítems está relacionada con su grado de dificultad (o probabilidad de acierto en el modelo TRI). En términos simples, los ítems fáciles, o con alta probabilidad de acierto, aportan información en la zona de bajo nivel de habilidad del constructo, permitiendo que la prueba entregue información más precisa sobre los examinados con bajo nivel de logro. En contraposición, ítems muy difíciles entregan información acerca de los examinados de un alto nivel de habilidad, permitiendo que la prueba resulte más precisa entre los examinados de alto nivel de logro (de Ayala, 2009). Hipotéticamente, si la prueba estuviese formada únicamente por ítems fáciles y difíciles, no aportaría información sobre la generalidad de los examinados, sino solamente de estos grupos extremos. Es por esto que se espera que la prueba aporte información en los distintos niveles de habilidad medido (recordemos, aprendizaje del currículo de Lenguaje y Matemática) y que aporte mayor información en las zonas del constructo donde se concentra mayor cantidad de personas, ya que, si bien la precisión siempre es deseable, es ahí justamente donde se la requiere en mayor

[10] El modelo TRI a la base de los resultados de los estudiantes en las pruebas SEPA incluye la información de los once niveles que participan en la medición y permite, mediante un proceso de calibración conjunta con ítems ancla entre pares de niveles (ver Capítulo 3), obtener resultados en una misma escala para todos los examinados. De esta forma, es posible afirmar que los resultados de las distintas cohortes que participan en la medición son comparables entre sí. En el modelo TRI de elección para SEPA, específicamente un modelo Rasch (Rasch, 1960), las estimaciones de habilidad de los examinados son una función directa de su puntaje bruto, por ende, no se pierde la posibilidad de establecer una relación única y directa entre los puntajes estandarizados y los porcentajes de logro que se derivan de la cantidad de respuestas correctas. La ventaja de este modelamiento radica en que los puntajes que son resultado de este modelamiento (puntajes estándar) son más precisos que los que se alcanzarían si solamente se hiciera uso de los procedimientos propios de la teoría clásica (puntajes brutos o porcentajes de logro) ya que se sustentan sobre la información que entrega cada ítem por separado y se fundamentan en procedimientos iterativos ente ítems y personas hasta que se logra la convergencia del mejor modelo (de Ayala, 2009).

medida. Tal como establece de Ayala (2009), las funciones informativas de los ítems pueden ilustrar características que son específicas de la prueba, mostrando su capacidad para discriminar a personas en los distintos niveles de habilidad con independencia de cuántas de ellas se ubican efectivamente en cada nivel de habilidad. Por ende, el éxito en el desarrollo de un instrumento radica en el grado de consistencia que exista entre su función informativa y la distribución de habilidades de los examinados. En el caso de SEPA, cuyo propósito es describir la diversidad de niveles de desarrollo del aprendizaje de los estudiantes en función de un currículum común, se espera que el análisis de las gráficas de información y personas muestre para cada prueba y nivel una superposición tal que la capacidad informativa de la prueba se encuentre alineada con la distribución de habilidades de los examinados que responden dichas pruebas.

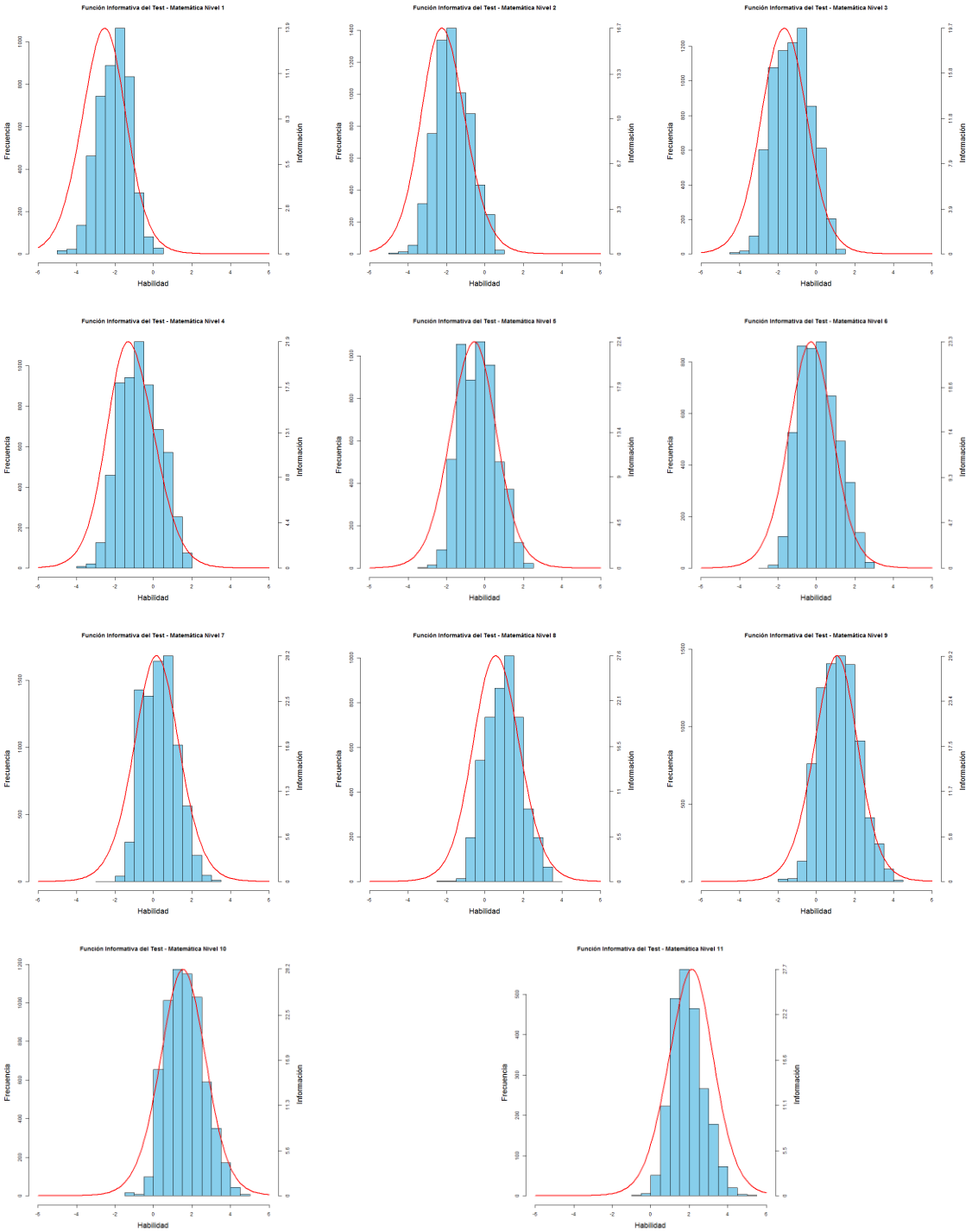
A continuación, las Figuras 1 y 2 presentan de manera gráfica la relación entre la información de las pruebas y las distribuciones de los examinados en la prueba SEPA 2015. En este informe se grafica en una misma escala tanto a las personas (histograma) como la información de las pruebas, a partir de la distribución de sus ítems (la función de información es representada mediante la curva en rojo).

**Figura 1: Función informativa y distribución de puntajes de estudiantes examinados en prueba SEPA Lenguaje 2015, para cada nivel.**




Nota: La línea roja es la función informativa y el histograma azul representa a las personas.

**Figura 2: Función informativa y distribución de puntajes de estudiantes examinados en prueba SEPA Matemática 2015, para cada nivel.**




Nota: La línea roja es la función informativa y el histograma azul representa a las personas.

Hecho este análisis, es posible afirmar que el nivel de precisión en la información que entregan las pruebas SEPA es satisfactorio ya que en todos los niveles de cada sector medido se puede apreciar un adecuado alineamiento entre la información que provee el conjunto de ítems de cada prueba rendida por los estudiantes, con los niveles de habilidad de ellos (en este caso, grado de conocimiento del currículum en cada grado de enseñanza evaluado).

Es importante destacar que esto es una buena noticia en tanto el propósito de esta prueba es la descripción de los distintos niveles de desarrollo en el aprendizaje de los contenidos del currículum, encontrándose una mayor magnitud de información (y menor error) donde se ubica la mayor parte de los examinados. Este análisis muestra que los resultados de las pruebas SEPA son confiables, especialmente cuando los resultados son tratados a nivel de grupos de estudiantes (sea curso, o nivel, o red de colegios de un nivel), contando con un nivel de precisión adecuada para considerarlos insumo para la toma de decisiones a nivel institucional[11].

## Evidencia acerca de la validez de las pruebas SEPA

A continuación, se presentará evidencia asociada a la validez de las pruebas SEPA, en específico, se reportará primeramente evidencia sobre la estructura interna de las pruebas y, posteriormente, evidencia de validez convergente de las pruebas.

### 1) Validez de estructura interna: análisis de componentes principales

Un supuesto básico acerca de la validez de las inferencias acerca de un instrumento es la correspondencia o ajuste entre los atributos que se desea medir y los puntajes o respuestas reales que obtienen los participantes de la medición. Para evaluar esto una aproximación más común es la utilización de la familia de técnicas de análisis factorial, técnicas de tipo multivariado, cuyo propósito es identificar la estructura subyacente de un conjunto de datos observados, empleando todas las variables (en este caso ítems) de forma simultánea (AERA, APA & NCME, 2014; Brown, 2006).

[11] Es importante advertir, como ocurre con cualquier medición estandarizada, que el nivel de precisión de una prueba es siempre menor cuando se la emplea para medir a un estudiante individualmente, que cuando se la emplea para estimar el grado de logro de un grupo de estudiantes.

Para el caso de las pruebas SEPA, que son resumidas en un puntaje único por cada examinado, se hace relevante verificar que se trate de pruebas unidimensionales. En términos sencillos, esto quiere decir que se esté midiendo primariamente un único constructo en estas pruebas, el cual en este caso refiere al conocimiento de los contenidos y habilidades del currículum en cada sector y grado evaluado.

Esta verificación se realizó mediante un Análisis de Componentes Principales en cada sector y nivel. Esta técnica de análisis permite identificar componentes que se basen en las relaciones que muestren los ítems entre sí[12] (Dunteman, 1989).

Para establecer el grado de unidimensionalidad de las pruebas, se analizó la importancia relativa del primer componente, considerando tanto el porcentaje de la variabilidad total que dicho componente explica, así como su relación con el segundo componente. La unidimensionalidad supone que el primer componente explique por sí mismo un porcentaje relevante de la variabilidad total (por ejemplo, más del 20%) y que el primer componente sea claramente superior al segundo (por ejemplo, cuando el primero es al menos 4 veces la magnitud del segundo[13]).

Las Tablas 1 y 2 presentan de modo sintético los resultados de los análisis de componentes principales para cada una de las pruebas analizadas:

[12] Concretamente, esta técnica combina linealmente las variables originales (en este caso, las preguntas de la prueba) y a partir de estas combinaciones extrae una serie de nuevas variables, llamadas componentes principales, con la restricción de que el primer componente debe explicar la máxima proporción de varianza observada posible, y el que le sigue debe resultar ortogonal, o no correlacionado, con el anterior y debe explicar la mayor cantidad de varianza posible de la varianza restante (Abdi & Williams, 2010).

[13] Para comparar la magnitud de cada componente se emplea el valor del *eigenvalue* asociado a cada uno.

**Tabla 2: Resultados de Verificación Unidimensionalidad Prueba SEPA Lenguaje 2015, para cada Nivel**

| Nivel | Eigenvalue primer componente (1) | Variaza explicada primer componente | Eigenvalue segundo componente (2) | Razón entre Eigenvalue (1) y (2) |
|---|---|---|---|---|
| 1 | 6,98 | 0,28 | 1,69 | 4,12 |
| 2 | 8,49 | 0,28 | 1,59 | 5,35 |
| 3 | 11,17 | 0,32 | 1,52 | 7,33 |
| 4 | 12,36 | 0,31 | 1,73 | 7,13 |
| 5 | 12,89 | 0,33 | 1,74 | 7,43 |
| 6 | 10,02 | 0,25 | 1,55 | 6,46 |
| 7 | 11,84 | 0,24 | 1,71 | 6,91 |
| 8 | 12,79 | 0,26 | 1,72 | 7,41 |
| 9 | 11,18 | 0,22 | 1,48 | 7,58 |
| 10 | 12,80 | 0,26 | 1,84 | 6,97 |
| 11 | 11,18 | 0,22 | 2,11 | 5,31 |

**Tabla 3: Resultados de Verificación Unidimensionalidad Prueba SEPA Matemática 2015, para cada Nivel**

| Nivel | Eigenvalue primer componente (1) | Variaza explicada primer componente | Eigenvalue segundo componente (2) | Razón entre Eigenvalue (1) y (2) |
|---|---|---|---|---|
| 1 | 6,37 | 0,25 | 1,20 | 5,32 |
| 2 | 8,32 | 0,28 | 1,45 | 5,74 |
| 3 | 10,76 | 0,31 | 1,48 | 7,28 |
| 4 | 13,83 | 0,35 | 1,77 | 7,81 |
| 5 | 11,60 | 0,29 | 1,50 | 7,71 |
| 6 | 12,93 | 0,32 | 1,38 | 9,36 |
| 7 | 12,77 | 0,26 | 1,58 | 8,08 |
| 8 | 14,17 | 0,28 | 1,73 | 8,18 |
| 9 | 15,22 | 0,30 | 1,82 | 8,36 |
| 10 | 14,90 | 0,30 | 2,21 | 6,76 |
| 11 | 11,85 | 0,24 | 2,05 | 5,78 |

Tal como se muestra en estos resultados, es posible afirmar que existe preponderancia del primer componente extraído por esta técnica en cada una de las pruebas analizadas, puesto que en todos los casos el primer componente explica un porcentaje de la varianza total que supera

al 20% y se constata que el primer componente es marcadamente superior a la importancia del segundo.

**2) Validez convergente: Análisis de correlación entre las pruebas SEPA**

Una segunda evidencia de validez de una medición puede juzgarse a partir de la correlación de los resultados de un instrumento con los de otro instrumento aplicado a al mismo grupo. Idealmente se debiera considerar tanto evidencia acerca de relaciones convergentes (cuando se emplean mediciones que evalúan un mismo constructo, por ejemplo, Matemática), así como discriminante (cuando se emplean mediciones que evalúan constructos diferentes, por ejemplo, Matemática y Lenguaje). La idea de este tipo de análisis, siguiendo la lógica propuesta por Campbell y Fiske (1959), es que las relaciones de tipo convergente sean superiores a las de tipo discriminante. Aplicado a nuestro contexto esto quiere decir que los puntajes de dos pruebas de Matemática debieran mostrar una correlación superior a los puntajes de una prueba de Lenguaje y otra de Matemática que hayan sido aplicadas a los mismos estudiantes.

En el caso de las pruebas SEPA, se realizó un análisis basado en el criterio recién expuesto, comparando diversos indicadores de tipo convergente y discriminante. A continuación, se identifican los indicadores que fueron comparados, según la magnitud hipotetizada de las correlaciones que se debiera observar. El primero de ellos corresponde a evidencia convergente, mientras que los dos siguientes son de tipo discriminante.

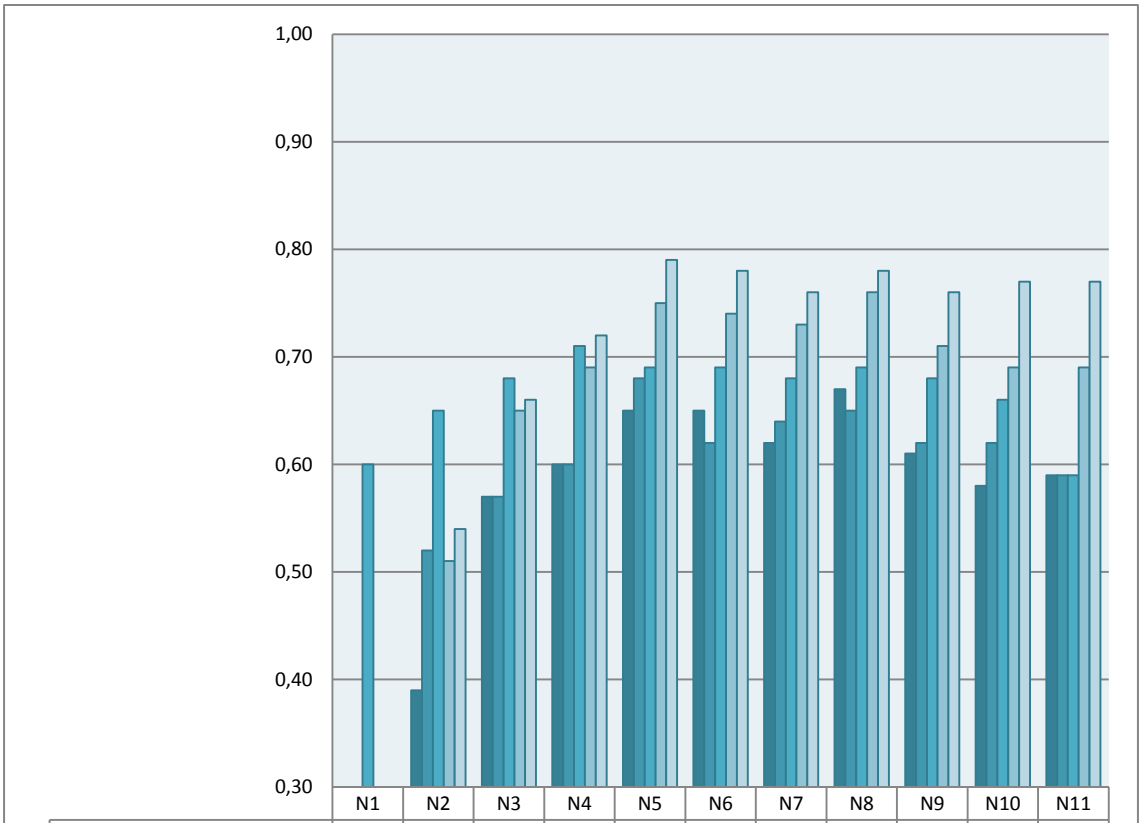
i) Las correlaciones entre los resultados obtenidos entre pruebas de Lenguaje aplicadas en dos años consecutivos (por ejemplo, Len14-Len15) (lo mismo para pruebas de Matemática en dos años consecutivos) debieran ser las más altas. En la Figura 3 corresponde a Len14-Len15 y Mat14-Mat15.

ii) La correlación, en un mismo año, entre los resultados de Lenguaje y Matemática debería ser positiva (por ejemplo, Len14-Mat14), pero comparativamente más baja que la correlación de una misma prueba entre años consecutivos (lo que se observe en (i)). En la Figura 3 se incluye una fila que corresponde al promedio de las dos correlaciones que se pueden calcular para esta situación (el promedio de Len14-Mat14 y el de Len15-Mat15).

iii) Las correlaciones entre Lenguaje y Matemática en años distintos (por ejemplo, Len14-Mat15) deberían ser las más bajas de estas comparaciones. En la Figura 3 corresponde a Len14-Mat15 y Mat14-Len15.

Para facilitar la lectura de estas comparaciones, la Figura 3 resume las correlaciones que corresponden a cada hipótesis:

**Figura 3: Correlaciones entre sectores y años de las Pruebas SEPA**




| | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 | N11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Len14 - Mat15 | | 0,39 | 0,57 | 0,60 | 0,65 | 0,65 | 0,62 | 0,67 | 0,61 | 0,58 | 0,59 |
| Mat14 - Len15 | | 0,52 | 0,57 | 0,60 | 0,68 | 0,62 | 0,64 | 0,65 | 0,62 | 0,62 | 0,59 |
| Len14-Mat14 y Len15-Mat15 | 0,60 | 0,65 | 0,68 | 0,71 | 0,69 | 0,69 | 0,68 | 0,69 | 0,68 | 0,66 | 0,59 |
| Len14 - Len15 | | 0,51 | 0,65 | 0,69 | 0,75 | 0,74 | 0,73 | 0,76 | 0,71 | 0,69 | 0,69 |
| Mat14 - Mat15 | | 0,54 | 0,66 | 0,72 | 0,79 | 0,78 | 0,76 | 0,78 | 0,76 | 0,77 | 0,77 |

Como se puede observar en la Figura 3, las hipótesis planteadas se cumplen al menos parcialmente. Desde 5º Básico en adelante se cumple el criterio de que las correlaciones para una prueba del mismo sector de aprendizaje que haya sido aplicada en años consecutivos, sea la más alta. Las correlaciones son algo superiores entre pruebas de Matemática que entre las de

Lenguaje, pero en todos los casos ambas superan a las restantes correlaciones (desde 5º Básico en adelante). También se observa que las correlaciones entre Matemática y Lenguaje en un mismo año se ubican en un valor intermedio, y a la vez mayor con respecto a las correlaciones que se constatan entre ambos sectores cuando se las aplica en años consecutivos. Puede llamar la atención que la correlación entre Matemática y Lenguaje pueda parecer relativamente alta, pero es importante advertir que ello se observa en la mayor parte de los programas de medición a nivel nacional e internacional (en el contexto nacional ello se observa tanto con SIMCE como con la PSU). Esta relación en parte refleja habilidades generales de los estudiantes que se manifiestan en su aprendizaje de dominios verbales y numéricos, pero también se basa en habilidades cognitivas más específicas (como las de tipo fonológico y semántico), que se requieren para decodificar la información que aparece en preguntas de Lenguaje y Matemática, tal como ha sido recientemente observado por Singer (2016). Es posible que lo observado por Singer sea relevante para explicar la mayor correlación entre ambos dominios que se observa en los primeros años de escolaridad, asumiendo que en esa etapa estas habilidades cognitivas fueran más preponderantes.

**Evidencia acerca de la justicia del test**

Finalmente, el tercer atributo a considerar para juzgar la calidad de las pruebas SEPA, es el grado en que pueden entregar información del nivel de habilidad de los estudiantes sin distorsiones asociadas a características de los examinados que sean diferentes de lo que se evalúa. La mayor parte de las mediciones educacionales muestra diferencias importantes entre grupos (por ejemplo, entre estudiantes pertenecientes a distintos grupos socioeconómicos o entre hombres y mujeres). En principio, estas diferencias no deben ser consideradas como evidencia de sesgo de medición (es decir, producidas por distorsiones en los instrumentos de evaluación), puesto que pueden reflejar diferencias efectivas entre tales grupos[14]. Para que el sesgo de medición pueda ser invocado como posible explicación para tales diferencias, se debe demostrar que, a nivel de los ítems de una prueba o en los puntajes globales de las mismas, se

[14] Una diferencia efectiva no conlleva la connotación de que sea natural. Por ejemplo, aunque en la mayor parte de los países del mundo se observa que los hombres tienen mejor desempeño que las mujeres en pruebas de Matemática, tal diferencia es ampliamente interpretada como resultado de una socialización diferencial de hombres y mujeres con respecto al dominio numérico (que entre otras cosas se refleja en diferentes oportunidades de aprendizaje para niños y niñas).

favorece el desempeño de un grupo determinado. Este tipo de análisis, denominados de sesgo de ítems o pruebas, se basan en principios y técnicas claramente establecidos en la teoría de la medición y que son necesarios en la determinación de la calidad de las mediciones a gran escala (Camilli, 2006).

En este apartado se revisará la evidencia acerca de la justicia de las pruebas SEPA según género, mediante un análisis a nivel de ítems (comportamiento diferencial) y mediante un análisis a nivel de prueba (específicamente estudiando el posible sesgo predictivo de las pruebas). Atendiendo a evidencia previa, que ha mostrado que en el caso de Chile hay mayor evidencia acerca de sesgo de género que socioeconómico en pruebas educacionales (ver Manzi et al, 2010, para el caso de las PSU), en este caso se ha concentrado el análisis en el potencial sesgo de género.

**1) Comportamiento diferencial de ítems según género**

Preliminarmente, se estudió el comportamiento diferencial de ítems según género, como forma de pesquisar distorsiones a nivel de las preguntas de un test que pudieran estar beneficiando o perjudicando a hombres o mujeres (Brennan & National Council on Measurement in Education, 2006). Cuando estas diferencias aparecen, ello conlleva una distorsión que afecta la validez de los puntajes, puesto que implicaría que en determinadas preguntas, hombres y mujeres muestran un desempeño diferencial que no se explicaría por su nivel de desempeño en el dominio que se estuviera evaluando (Lenguaje o Matemática). Los estándares de medición educacional vigentes recomiendan que este tipo de análisis se lleve a cabo para detectar preguntas que pudieran estar afectando la validez de las mediciones, en la medida que este tipo de sesgos indicaría que un atributo diferente de la habilidad que está siendo medida podría afectar el desempeño de los examinados (AERA, APA & NCME, 2014; Fraillon, Schulz, Friedman, Ainley & Gebhardt, 2015; Schulz, Ainley & Fraillon, 2011).

En concreto, se habla de presencia de comportamiento diferencial de un ítem (DIF) cuando dos (o más) grupos de examinados, que poseen un mismo nivel de habilidad, muestran una probabilidad diferente de responder correctamente una determinada pregunta.

Entre los diversos procedimientos para detectar comportamiento diferencial de ítems, para el caso de las pruebas SEPA se ha decidido utilizar un análisis de regresión logística, que permite

revisar si la pertenencia a un grupo particular (en este caso ser hombre o mujer) es un predictor significativo de la probabilidad de acierto en los ítems de la prueba. Complementando el uso de esta técnica, se han empleado los criterios establecidos por el Educational Testing Service (ETS) para clasificar ítems según el grado en que muestren evidencia de funcionamiento diferencial. La Tabla 3 muestra los tres niveles establecidos por ETS, donde el C sería el más serio (Zwick, 2012).

**Tabla 3: Descripción de las Categorías de Comportamiento Diferencial de Items establecidas por el Educational Testing Service utilizadas en Análisis de las Pruebas SEPA**

| Categoría | Comportamiento diferencial | Criterio |
|---|---|---|
| A | Despreciable | $P(\chi^2_{MH}) > 0.05$ o $\lvert\alpha_{MH}\rvert \leq 1$ |
| B | Ligero a Moderado | $P(\chi^2_{MH}) \leq 0.05$ y $1 < \lvert\alpha_{MH}\rvert < 1.5$ |
| C | Moderado a Grande | $P(\chi^2_{MH}) \leq 0.05$ y $\lvert\alpha_{MH}\rvert \geq 1.5$ |

Una vez obtenidos los resultados de este análisis, se constató que entre todos los itemes correspondientes a las 22 pruebas analizadas (11 de Matemática y 11 de Lenguaje), solamente un ítem de todos ellos mostró un funcionamiento diferencial por género que correspondiera a la categoría C (moderado a grande), en este caso favorable a los niños. El contenido de este ítem fue revisado por expertos en su contenido, y fue descartado en futuros ensamblajes de las pruebas SEPA. Por otra parte, se levantaron alertas de funcionamiento diferencial ligero a moderado en un número muy acotado de ítems (9 de 332 ítems en Lenguaje y 13 de 332 ítems en Matemática). Todos estos ítems fueron también revisados por expertos en contenido, buscando determinar causas posibles de este comportamiento que pudiesen orientar futuros procesos de construcción. En todos los casos no fue posible atribuir una causa clara basada en el contenido, por lo que se decidió seguir monitoreando el comportamiento de estos ítems en eventuales futuras aplicaciones de los mismos.

### 2) Sesgo predictivo de las pruebas SEPA según género

El análisis de comportamiento diferencial de ítems nos entrega una información relevante, pero parcial acerca del sesgo de medición, por lo que es conveniente complementarla con

estudios acerca del potencial sesgo asociado a los puntajes globales de las pruebas. Para un análisis de funcionamiento diferencial a nivel de una prueba completa se requiere incorporar alguna medida adicional acerca del desempeño de los estudiantes que permitan establecer si las puntuaciones de la prueba predicen de manera diferencial (en este caso para hombres y mujeres), el desempeño de los examinados en dicha medida adicional (Kobrin & Barbuti, 2008).

Para implementar este tipo de análisis se usó la información disponible a partir de las mismas pruebas SEPA en dos años consecutivos. En cada caso el análisis se enfocó en el posible sesgo de la primera medición, empleando la segunda como criterio externo. En concreto, se buscaba establecer si la primera prueba permitía predecir de manera equivalente para hombres y mujeres el desempeño en la segunda. El procedimiento consistió en ajustar un modelo de regresión lineal común para todos los estudiantes de un nivel, estableciendo como predictor el puntaje en el año 1 y como variable resultado el puntaje promedio en Lenguaje y Matemáticas en el año 2. Una vez calculada la ecuación de regresión empleando simultáneamente los datos de hombres y mujeres, se calcularon los residuos estandarizados de hombres y mujeres por separado[15]. Por definición, en todo modelo de regresión, el promedio de los residuos a partir de la muestra completa es cero. Sin embargo, si el promedio de los residuos de alguno de los grupos es distinto de cero se puede afirmar que existe predicción diferencial. En este caso, valores en promedio negativos indican que existe una sobrepredicción del rendimiento, es decir, la medición está sobreestimando la habilidad medida en el segundo año. En contraposición, si hay valores positivos, se considera que la primera medición está subestimando la habilidad de un grupo determinado respecto de la segunda medición (Bravo et al., 2010; Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008).

Los resultados de estos análisis se muestran en las Tablas 4 y 5. Se presentan para las pruebas de 1º Básico a 2º Medio (nivel 10), puesto que en el caso de las pruebas que se rinden en 3º Medio no existe un criterio posterior con el que se las pueda analizar. En este caso, la

[15] El residuo es la diferencia entre el puntaje que obtiene un examinado en el criterio (en este caso el promedio en las pruebas SEPA de Lenguaje y Matemática en el año 2), con respecto al puntaje que predice el modelo de regresión. El residuo estandarizado permite hacer comparaciones entre residuos al ponerlos en la misma escala para todos los análisis. Cuando el residuo es positivo, quiere decir que el desempeño del examinado es superior a lo que predice el modelo de regresión, indicando que dicho modelo subpredice su desempeño. En el análisis de sesgo de pruebas se compara el residuo promedio que obtienen hombres y mujeres. Cuando uno de los grupos, por ejemplo, las mujeres, tienen residuos promedio positivos, se dice que la prueba subpredice su rendimiento, lo que podría constituir un sesgo en este caso desfavorable para las mujeres.

estimación se basa en la predicción de las pruebas rendidas en 2014 con respecto a los desempeños observados un año después (en SEPA 2015).

**Tabla 4: Promedio de residuos estandarizados prueba SEPA Lenguaje, por género y nivel**

| Nivel | Género | N | Media |
|---|---|---|---|
| 2 | Femenino | 1988 | -0,05 |
| | Masculino | 2075 | 0,04 |
| 3 | Femenino | 1985 | -0,05 |
| | Masculino | 2087 | 0,05 |
| 4 | Femenino | 2005 | -0,03 |
| | Masculino | 2085 | 0,03 |
| 5 | Femenino | 1547 | -0,04 |
| | Masculino | 1621 | 0,04 |
| 6 | Femenino | 1717 | -0,03 |
| | Masculino | 1748 | 0,03 |
| 7 | Femenino | 2351 | 0,00 |
| | Masculino | 2802 | 0,00 |
| 8 | Femenino | 1660 | -0,03 |
| | Masculino | 1711 | 0,03 |
| 9 | Femenino | 2249 | -0,01 |
| | Masculino | 3041 | 0,00 |
| 10 | Femenino | 1949 | 0,02 |
| | Masculino | 2686 | -0,01 |
| 11 | Femenino | 744 | -0,03 |
| | Masculino | 609 | 0,03 |

*Nota.* Valores residuales promedio positivos indican una subestimación del grupo y valores negativos indican una sobreestimación del grupo.

**Tabla 5: Promedio de residuales estandarizados prueba SEPA Matemática, por género y nivel.**

| Nivel | Género | N | Media |
|---|---|---|---|
| 2 | Femenino | 1969 | 0,00 |
| | Masculino | 2088 | 0,00 |
| 3 | Femenino | 1958 | 0,03 |
| | Masculino | 2093 | -0,03 |
| 4 | Femenino | 2009 | 0,06 |
| | Masculino | 2080 | -0,06 |
| 5 | Femenino | 1547 | 0,01 |
| | Masculino | 1612 | -0,01 |
| 6 | Femenino | 1718 | 0,06 |
| | Masculino | 1735 | -0,06 |
| 7 | Femenino | 2358 | 0,11 |
| | Masculino | 2781 | -0,09 |
| 8 | Femenino | 1656 | 0,10 |
| | Masculino | 1714 | -0,10 |
| 9 | Femenino | 2224 | 0,09 |
| | Masculino | 2976 | -0,07 |
| 10 | Femenino | 1922 | 0,09 |
| | Masculino | 2654 | -0,06 |
| 11 | Femenino | 756 | 0,08 |
| | Masculino | 619 | -0,09 |

*Nota.* Valores residuales promedio positivos indican un subestimación del grupo y valores negativos indican una sobreestimación del grupo.

Globalmente las Tablas 4 y 5 muestran que la magnitud de las diferencias observadas, cuando aparecen, es moderada. En el caso de Lenguaje, en ninguno de los niveles se observan diferencias superiores a +- 0,05. En el caso de Matemática las diferencias son algo superiores a partir de 7º básico, acercándose a +- 0,10. En cuanto a la dirección de las diferencias, se observa que en Lenguaje las leves diferencias tienden a mostrar subestimación de los hombres, mientras que en Matemática todas las diferencias observadas a partir de 7º básico corresponden a subestimación de las mujeres. En términos generales estos resultados indican una muy leve presencia de sesgo de género en las pruebas SEPA. En Lenguaje se favorece levemente a las mujeres, mientras que en Matemática se favorece en una magnitud moderada a los hombres. Es interesante hacer presente que la magnitud de las diferencias aquí consignadas son

equivalentes, aunque algo inferiores, a las constatadas en pruebas de admisión a las universidades, tanto en el caso norteamericano con el SAT (Mattern et al, 2008), como en el caso chileno con la PSU (Manzi, et al. 2010).

**Referencias**

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

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

Bravo, D., Bosch, M. A., Del Pino, G., Donoso, G., Manzi, J., Martinez, M., & Pizarro, R. (2010). *Validez diferencial y sesgo de predictividad de las Pruebas de Admisión a las Universidades Chilenas.* Santiago: CTA-PSU.

Brennan, R. L., & National Council on Measurement in Education. (2006). *Educational Measurement*. Westport, CT: Praeger Publishers.

Brown T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

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

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, *119*(2), 166-175.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Publications.

Dunteman, G. H. (1989). *Principal components analysis* (No. 69). Newbury Park: Sage.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). *ICILS 2013 Technical Report.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average.* Research Report No. 2008-5. College Board.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *ETS Research Report Series*, *1984*(1), 1-11.

Manzi, J., Bosch, A., Bravo, D., del Pino, G., Donoso, G. & Pizarro, R. (2010). Validez diferencial y sesgo en la predictividad de las pruebas de Admisión a la Universidades Chilenas (PSU). *Revista Iberoamericana de Evaluación Educativa, 3*(2), 30-48.

Manzi, J., & San Martín, E. (2003). La necesaria complementariedad entre teoría clásica de la medición (TCM) y teoría de respuesta al ítem (TRI): Aspectos conceptuales y aplicaciones. *Estudios Públicos*, *90*, 145-183.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential Validity and Prediction of the SAT®. Research Report No. 4.* New York: College Board.

Moses, T., Miao, J., & Dorans, N. J. (2010). A comparison of strategies for estimating conditional DIF. *Journal of Educational and Behavioral Statistics, 35*(6), 726-743.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

Schulz, W., Ainley, J., & Fraillon, J. (2011). *ICCS 2009 Technical Report.* Amsterdam: International Association for the Evaluation of Educational Achievement.

Singer, V. (2016). *Más convergente que divergente: El desempeño en aritmética y lectura en la etapa escolar.* Tesis de Doctorado para la obtención del título de Doctor en Psicología de la Pontificia Universidad Católica de Chile, Santiago, Chile.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1).