



PILOTO PSU

Informe 2017

Departamento de Evaluación,
Medición y Registro Educacional

Enero 2018

CONTENIDO

Presentación	4
Características del Piloto PSU 2017	5
Planificación	6
Aplicación de forma ancla	6
Plan de difusión	7
Diseño Muestral	8
Marco muestral	8
Tipo de muestreo	12
Cierre de muestra	15
Operativa	16
Trabajo de campo	16
Equipo de aplicación	17
Traslado de material	18
Condiciones para rendir las pruebas piloto	18
Personas en situación de discapacidad	19
Resumen de la aplicación	21
Asistencia por prueba	21
Asistencia por sexo y región	22
Asistencia por tipo de establecimiento	23
Muestra efectiva	24

Análisis	26
Informes a establecimientos educacionales y estudiantes	26
Análisis de resultados	27
Teoría Clásica del Test (CTT)	28
Teoría de Respuesta al Ítem (IRT)	31
Procedimiento de análisis	32
Funcionamiento Diferencial del Ítem (DIF)	33
Revisión cualitativa de ítems	40
Finalización del proceso	41
Análisis de los resultados en la forma ancla	41
Bibliografía	43
Anexos	45

Presentación

Cada año el DEMRE realiza procesos de pilotaje de nuevos ítems para la PSU, cuyo fin es validar estadísticamente el comportamiento de las preguntas que se construyen para las distintas pruebas. Esto permite contar con un banco de preguntas disponibles para las pruebas oficiales que se utilizan en cada Proceso de Admisión Universitaria.

La aplicación piloto del año 2017 se realizó entre los meses de septiembre y octubre, con algunas particularidades que la diferencian de procesos anteriores, principalmente relacionadas con cambios en la metodología de muestreo y aplicación. Estos cambios tuvieron por objetivo mejorar la información disponible para los análisis de las preguntas y aumentar la tasa de participación de los estudiantes.

El presente informe describe la metodología y las distintas etapas asociadas a la aplicación piloto 2017, especificando los cambios y ajustes que se realizaron en relación con procesos anteriores. El informe se organiza en 4 capítulos. El primer capítulo muestra las características generales del Piloto. El segundo capítulo muestra la planificación de tareas y actividades para llevar a cabo la aplicación. El tercer capítulo muestra la operativa desplegada para la aplicación. El cuarto capítulo muestra los análisis realizados.

Características del Piloto PSU 2017

La aplicación piloto de la PSU es una tarea crucial en el proceso de construcción de instrumentos de evaluación dentro del Sistema Único de Admisión, ya que en esta se realiza la validación estadística de los ítems que luego formarán parte de las pruebas oficiales. Esta validación se realiza en una muestra de la población que rinde la PSU oficial, bajo estrictos parámetros que aseguran la calidad del producto final.

Para este proceso, el DEMRE difundió el piloto como una oportunidad para que los estudiantes pudiesen poner a prueba su futuro desempeño en la PSU en las mismas condiciones de aplicación que en la prueba oficial. Sumado a lo anterior, y para motivar aún más la participación, se entregó un reporte de rendimiento, tanto a nivel individual como a cada establecimiento educacional.

Para la aplicación, se seleccionaron regiones y comunas que permitiesen caracterizar a la población del país que rinde PSU en términos de tipo de establecimientos, características escolares (como rendimiento y puntajes PSU) y geográficas. Las regiones seleccionadas fueron Antofagasta, Coquimbo, Valparaíso, Metropolitana, O'Higgins, Maule, Bío-Bío y Araucanía.

Tras considerar una serie de variables —como dependencia, rama, tasa de participación en Piloto 2016 y tamaño de establecimiento—, se seleccionaron comunas al interior de cada región¹. Posteriormente se utilizó un método de selección aleatorio asignando pesos, para seleccionar los establecimientos participantes, considerando a la totalidad de sus estudiantes matriculados en IV medio.

La aplicación piloto se realizó en dos etapas, con el objetivo de conseguir la cantidad necesaria de estudiantes para completar la muestra para realizar los análisis. Se realizó una primera etapa de carácter estándar, simulando las condiciones de aplicación de una prueba oficial, los días martes 5 y miércoles 6 de septiembre. Tras analizar las tasas de asistencia, fue necesario generar una segunda etapa, cuya aplicación fue al interior de cada establecimiento educacional seleccionado, en fechas y horarios acordados individualmente con la contraparte de cada establecimiento.

Durante la aplicación piloto se contó con la estrecha colaboración de los Secretarios de Admisión que el DEMRE tiene a lo largo del país para cada una de las sedes que participaron en este proceso. Además, participaron directivos, profesores, profesionales y auxiliares de los establecimientos que facilitaron sus dependencias como locales de aplicación.

Cada uno de los estudiantes convocados para el piloto rindió al menos dos pruebas dependiendo del caso: al menos una de las dos obligatorias —Lenguaje y Comunicación y Matemática— y al menos una prueba electiva (Historia o Ciencias). La asignación de la prueba electiva se realizó primordialmente en base a la información que proporcionaron los establecimientos y a la inscripción a la PSU Oficial Admisión 2018.

1 Las comunas se detallan en la sección de “Muestreo”

Las pruebas piloto de Lenguaje, Matemática e Historia, Geografía y Ciencias Sociales tuvieron 75 preguntas, mientras que las de Ciencias estuvieron compuestas por 80 preguntas. En el piloto se probaron 97 formas, distribuidas en las distintas pruebas que componen la batería PSU. Tal como se muestra en la Tabla 1.

TABLA 1: PRUEBAS PILOTO PSU, CANTIDAD DE FORMAS PROBADAS

	Prueba	Numero de formas
	Lenguaje y Comunicación	19
	Matemática	19
	Historia, Geografía y Ciencias Sociales	19
Ciencias	Biología	10
	Física	10
	Química	10
	Técnico Profesional	10
	TOTAL	97

Planificación

Aplicación de forma ancla

Realizar el pilotaje de preguntas en una muestra que caracteriza a la población que finalmente participará de la aplicación oficial permite suponer que los resultados obtenidos en la aplicación piloto serán similares a los de la aplicación oficial. Este supuesto permite confeccionar las pruebas oficiales con mayor información y anticipar el comportamiento de las preguntas utilizadas.

Uno de los desafíos al realizar el pilotaje de preguntas de forma independiente a la aplicación oficial es asegurar dicha similitud de las condiciones de rendición. Una de las condiciones de rendición que aparecen como inigualables entre la aplicación piloto y la oficial tiene relación con la motivación de los estudiantes para responder correctamente la prueba.

Por otro lado, el fin principal del piloto PSU es nutrir de preguntas de calidad al banco de preguntas. Para esto, es fundamental el uso de preguntas de anclaje entre la aplicación oficial y la piloto con el propósito de calibrar el banco en base a una sola escala.

Para estudiar las diferencias entre los resultados de la aplicación piloto y la oficial junto con la calibración del banco de preguntas, se utilizó para cada disciplina una forma “ancla”. La forma ancla corresponde a una de las formas aplicadas en la prueba oficial en diciembre de 2015 para el proceso de Admisión 2016, que también se aplicó en el piloto PSU del año 2016, con el fin de mantener un link entre los pilotos y el proceso oficial. Así, la cantidad de formas detallada en la Tabla 1 considera una forma ancla para cada disciplina.

En el caso de la prueba de Ciencias, la forma ancla utilizada correspondió a una de las pruebas oficiales Ciencias TP. Esto permite que la forma ancla pueda aplicarse

aleatoriamente a los estudiantes que participaron en la aplicación piloto de las pruebas de Ciencias. Luego, para verificar la estabilidad de los parámetros entre la prueba oficial y la aplicación piloto se comparan los resultados obtenidos en la forma ancla, pues fue utilizada en ambas instancias.

Plan de difusión

Para mejorar la participación de los establecimientos educacionales en el piloto, el DEMRE llevó a cabo una serie de estrategias para involucrarlos intensamente en el proceso. Para ello, se realizaron reuniones informativas comunales, donde se entregó material de difusión como afiches y volantes (ver Anexo 1). La difusión también se realizó mediante el envío de correos electrónicos a establecimientos educacionales y estudiantes, que contenían toda la información relevante de la aplicación piloto.

Las reuniones informativas tuvieron el objetivo de comunicar los aspectos operativos y la relevancia de la participación de los establecimientos en el pilotaje. A estas reuniones se convocó a directivos u orientadores de educación media. Principalmente, se buscó la asistencia de representantes que tuviesen cercanía con los estudiantes para motivar la inscripción temprana en la PSU Oficial y la participación en la aplicación piloto.

En la difusión de la aplicación piloto, el DEMRE enfatizó a los establecimientos la necesidad de contar con información sobre la participación y la prueba electiva que los estudiantes deseaban rendir en la aplicación piloto. Esta última información era crucial para asegurar la asignación de una prueba electiva que respondiera a los intereses de los estudiantes y así evitar el ausentismo. El DEMRE definió dos vías para obtener esta información:

- Inscripción a la PSU
- Información proporcionada por los establecimientos educacionales

Tras confirmar la participación de los establecimientos educacionales convocados para el piloto, el DEMRE habilitó una planilla informática anclada al Portal Colegios², con la nómina de estudiantes de IV medio matriculados en el establecimiento, según los datos reportados al Ministerio de Educación. Esta planilla indicaba, para cada estudiante, su estado en la inscripción PSU Oficial (inscrito / no inscrito) y, en caso de estar inscrito, las pruebas electivas seleccionadas.

El propósito fue que, a través de esta plataforma, cada establecimiento consultara a sus estudiantes y retroalimentara al DEMRE sobre:

- Participación: que los establecimientos notificaran sobre estudiantes que no estuviesen en la nómina, estudiantes retirados del establecimiento, y aquellos que, estando en la nómina, no deseaban participar del piloto.
- Pruebas electivas: que los establecimientos informaran la prueba electiva que deseaban rendir aquellos estudiantes aun no inscritos para la PSU Oficial.

2 Portal habilitado por el DEMRE, destinado a que los establecimientos realicen todos los procedimientos correspondientes a cada Proceso de Admisión.

- **Personas en situación de discapacidad:** que los establecimientos informaran sobre los estudiantes en tal situación para que, desde el DEMRE, se tomaran las medidas pertinentes en la aplicación piloto.

Respecto de la primera etapa de pilotaje, esta información fue recibida hasta el día 8 de agosto. Ese mismo día se inició la asignación de pruebas electivas y de locales de aplicación. A partir de la tercera semana de agosto, el DEMRE envió la información definitiva de la aplicación piloto a los establecimientos educacionales: ubicación del local de rendición asignado, protocolo sugerido por el Ministerio de Educación para informar la participación, nómina de estudiantes seleccionados y pruebas a rendir. A su vez, y en conjunto con el Servicio de Tecnología e Información de la Universidad de Chile, el DEMRE envió un correo informativo a los estudiantes, para que conocieran los detalles del proceso, el local de rendición al que debían acudir y la prueba electiva que iban a rendir.

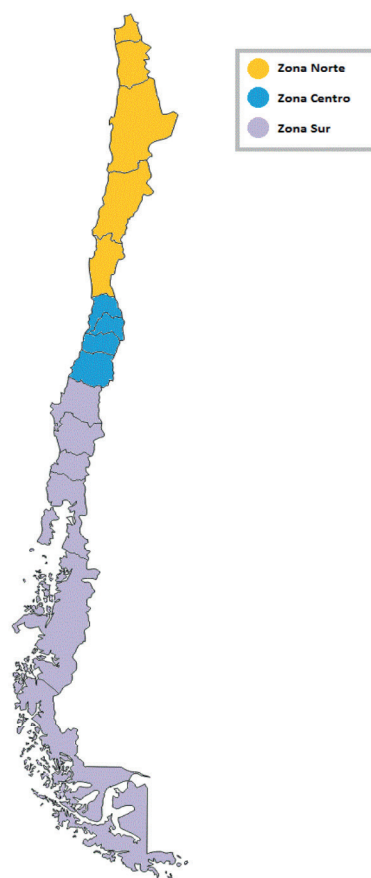
Para la segunda etapa de pilotaje, se contactó individualmente a cada establecimiento seleccionado. En este contacto se determinaron fechas y horarios de aplicación, la cantidad de estudiantes participantes, las pruebas obligatorias y electivas que se rendirían y trabajadores del establecimiento dispuesto como personal de aplicación.

Diseño Muestral

Marco muestral

Previo a determinar los criterios de selección de la muestra y, con el objetivo de tener información de referencia, se analizaron los datos de rendición PSU del año anterior (Admisión 2017) y otros antecedentes educativos. Por razones prácticas y operativas, se tomó la decisión de dividir al país en tres zonas geográficas: norte, centro y sur, de forma tal que cada zona estuviese formada por regiones vigentes a Agosto 2017. La zona norte consideró cinco regiones: de Arica y Parinacota, de Tarapacá, de Antofagasta, de Atacama y de Coquimbo. La zona centro agrupó las regiones: de Valparaíso, del Libertador General Bernardo O'Higgins, del Maule y Metropolitana. Por último, la zona sur correspondió a las regiones del Bío-Bío, de la Araucanía, de los Lagos, Aysén del General Carlos Ibáñez del Campo, de Magallanes y Antártica Chilena y Región de Los Ríos, como se expone en la Figura 1.

FIGURA 1. ZONAS GEOGRÁFICAS, PILOTO PSU 2017



Tras una serie de análisis, se determinó que estas zonas serían representadas por algunas regiones, estas fueron Antofagasta, Coquimbo, Valparaíso, Metropolitana, O'Higgins, Maule, Bío-Bío y Araucanía. Se buscó que las regiones seleccionadas reflejaran lo mejor posible las características de su zona en cuanto a la distribución de las pruebas rendidas, composición de establecimientos educacionales por rama educativa y dependencia administrativa, tal como lo señala la Tabla 2. Las regiones de Antofagasta y Coquimbo caracterizaron a la zona norte, las regiones de Valparaíso, Metropolitana, O'Higgins, Maule a la zona centro, mientras que las regiones de Bío-Bío y Araucanía caracterizaron a la zona sur.

En total, estas regiones alcanzan el 86.7% del total de estudiantes asistentes a la PSU en el Proceso de Admisión 2017 y un 83.7% del total de unidades educativas, tal como se muestra en la Tabla 3. Se entiende como unidad educativa, cada una de las modalidades de enseñanza que pertenezcan a un mismo establecimiento educacional³.

3 Cada "Unidad Educativa" es un elemento único dentro de la población, y está compuesta por un Rol Base de Datos ("RBD") y un código de enseñanza. Por ejemplo, si un establecimiento educacional polivalente (con un RBD "9999").

TABLA 2: COMPOSICIÓN DE ESTABLECIMIENTOS POR RAMA DE ENSEÑANZA Y DEPENDENCIA ADMINISTRATIVA, SEGÚN ZONA DEL PAÍS Y REGIONES SELECCIONADAS

	Modalidad de enseñanza						
	HC Diurna	HC Vespertina	TP Comercial	TP Industrial	TP Servicios	TP Agrícola	TP Marítimo
Zona Norte	46,71%	12,82%	10,62%	15,18%	10,62%	3,20%	0,84%
Antofagasta	55,65%	12,17%	8,70%	13,91%	8,70%	0,87%	0,00%
Coquimbo	52,44%	16,44%	8,44%	11,11%	7,11%	4,00%	0,44%
Zona Centro	51,53%	15,13%	10,84%	10,47%	9,28%	2,53%	0,22%
Valparaíso	55,48%	16,08%	7,60%	10,25%	7,42%	2,12%	1,06%
O'Higgins	46,27%	21,96%	10,98%	7,84%	8,63%	4,31%	0,00%
Maule	38,46%	13,46%	8,97%	14,10%	14,10%	10,90%	0,00%
Metropolitana	53,58%	13,99%	12,38%	10,25%	9,09%	0,71%	0,00%
Zona Sur	41,56%	18,98%	7,97%	9,49%	14,04%	4,55%	3,42%
Bío-Bío	44,26%	16,76%	9,98%	12,99%	9,98%	5,08%	0,94%
Araucanía	33,03%	21,02%	10,51%	12,01%	14,71%	8,11%	0,60%

	Dependencia administrativa		
	Particular Pagado	Particular Subvencionado	Municipal
Zona Norte	6,91%	51,26%	41,82%
Antofagasta	17,39%	32,17%	50,43%
Coquimbo	4,89%	59,56%	35,56%
Zona Centro	11,74%	53,09%	35,17%
Valparaíso	10,78%	51,94%	37,28%
O'Higgins	5,10%	41,96%	52,94%
Maule	3,53%	43,27%	53,21%
Metropolitana	14,83%	57,32%	27,85%
Zona Sur	4,17%	47,82%	48,01%
Bío-Bío	5,46%	40,87%	53,67%
Araucanía	3,30%	55,26%	41,44%

TABLA 3: PORCENTAJE DE UNIDADES EDUCATIVAS Y ESTUDIANTES POR REGIÓN, PROCESO DE ADMISIÓN 2017

Región	Unidades educativas	Estudiantes IV Medio	Estudiantes III Medio	Inscritos PSU Admisión 2017 (promoción del año)
Antofagasta	115	8.813	7.613	6.667
Coquimbo	225	11.702	8.945	9.167
Valparaíso	566	26.962	20.919	22.288
Metropolitana	1.551	96.348	80.084	81.379
O'Higgins	255	13.616	10.962	10.804
Maule	312	14.169	12.688	11.400
Bío-Bío	531	29.243	24.758	24.105
Araucanía	333	14.934	12.378	11.211
Total Nacional	4.668	252.024	207.298	204.040

Posteriormente, se escogieron las comunas detalladas en la Tabla 4. Estas fueron seleccionadas por cumplir con los siguientes criterios:

- Distribución porcentual de establecimientos educacionales según rama y dependencia, de forma similar a su zona y región.
- Alta asistencia al piloto 2016 (en las comunas que participaron).
- Poseer alta densidad poblacional en la región.

TABLA 4: COMUNAS SELECCIONADAS PILOTO PSU 2017

Región de Antofagasta	Región de Coquimbo	Región de Valparaíso	Región de O'Higgins	Región del Maule
» Antofagasta » Calama	» La Serena » Coquimbo » Ovalle	» Valparaíso » Quillota » Quilpué	» Rancagua » Machalí » San Fernando	» Talca » Curicó
Región del Bío-Bío	Región de la Araucanía	Región Metropolitana		
» Concepción » Los Ángeles » Chillán	» Temuco » Villarrica » Angol	» Santiago » El Bosque » La Pintana » Las Condes	» Maipú » Ñuñoa » Providencia » Quilicura	» San Miguel » Puente Alto » Colina » San Bernardo » Talagante

Tipo de muestreo

Una vez determinadas las regiones y comunas que representarían a cada zona del país, se seleccionaron los establecimientos que participarían en la aplicación piloto. Al respecto, se consideró un solo criterio de exclusión dentro de selección de las unidades educativas. Las ramas Técnico Agrícola y Marítimo, y las modalidades nocturnas⁴ (tanto HC como TP) no fueron incluidas en el marco muestral, ya que representan una proporción muy menor de la población que rinde la prueba Oficial.

Dentro de cada comuna, la población muestral consideró establecimientos educacionales de todas las dependencias administrativas, de las ramas HC Diurno, y Técnico Comercial, Industrial y Servicios. Para seleccionar la muestra, se utilizó un método de selección asignando pesos (Weighted Random Sampling). La razón de asignar pesos a los establecimientos se debe a que la distribución de las pruebas electivas no es homogénea, es decir, aproximadamente la mitad de los estudiantes rinden la PSU de Historia y en la prueba de Ciencias, alrededor del 50 % rinde el módulo de Biología, por lo que se produce la necesidad de asegurar la cantidad de estudiantes que rinden las pruebas con menos frecuencia, con el fin de obtener el tamaño de muestra suficiente para hacer los análisis psicométricos. Cabe destacar que, si bien, el módulo Técnico Profesional también muestra menor porcentaje de rendición respecto de las pruebas señaladas, es menos complejo llegar a la cantidad necesaria, ya que el diseño muestral selecciona primero las unidades educativas TP.

Diseño muestral

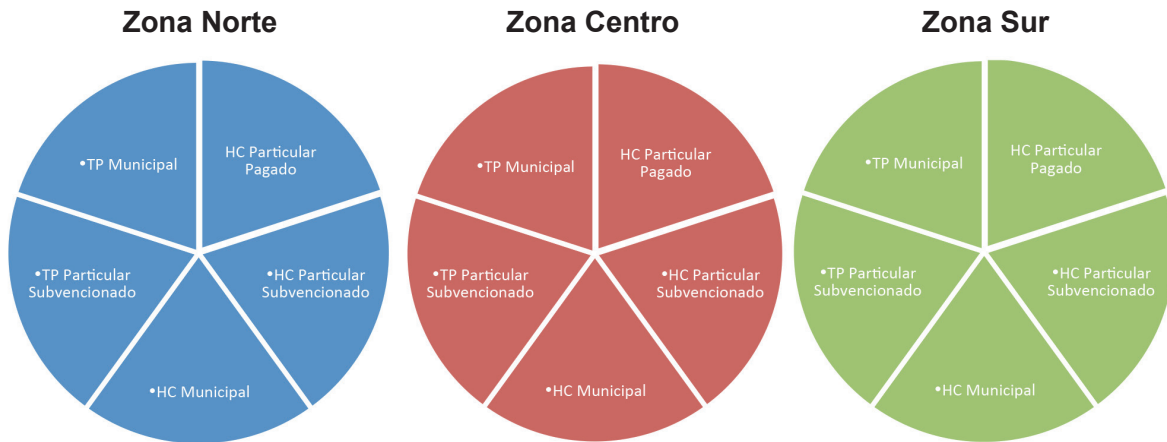
En este piloto se consideró como unidad muestral al establecimiento educacional. Para el diseño, se dividió al país en las 3 zonas antes mencionadas, las que se consideran como universos para seleccionar la muestra, a su vez, estas 3 zonas se segmentan en los siguientes 5 grupos conformados por unidades educativas pertenecientes a las comunas seleccionadas.

- HC PP: Humanista-Científico Particular Pagado
- HC PS: Humanista-Científico Particular Subvencionado
- HC MU: Humanista-Científico Municipal
- TP PS: Técnico Profesional Particular Subvencionado
- TP MU: Técnico Profesional Municipal

Finalmente se aplicó un diseño de muestreo estratificado, donde los estratos son de la forma Zona-Dependencia-Rama, teniéndose un total de 15 estratos, como se expresa en la Figura 2.

4 Respecto de las modalidades nocturnas, estas no fueron seleccionadas además por presuntos problemas de disponibilidad y compatibilidad horaria, lo que podría impactar en la asistencia al pilotaje.

FIGURA 2. ESQUEMA DEL DISEÑO MUESTRAL



Algoritmo de selección

Para la selección de la muestra, se utilizó el algoritmo WRS (Weighted Random Sampling) (Efraimidis & Spirakis, 2008). Este algoritmo parte de la base de que existe una población en la que cada unidad muestral i perteneciente a la población tiene asignado un peso según criterios definidos, denominado W_i . Para cada uno de las unidades muestrales se calcula un número aleatorio basado en una distribución uniforme entre 0 y 1, para luego calcular una “clave” (k_i) en base al peso asignado. Luego, los objetos se ordenan de tal manera que sus claves estén de mayor a menor y se selecciona a los m primeros objetos necesarios. El algoritmo se describe a continuación:

Algoritmo

Input: Una población V de m unidades muestrales con pesos asignados.

Output: Una muestra WRS de tamaño n

1. Para cada $v_i \in V$, hacer $u_i = \text{random}(0,1)$ y $k_i = u_i \frac{1}{w_i}$
2. Seleccionar los m unidades muestrales con mayor k_i como un WRS
Donde w_i son los pesos de cada objeto i y $u_i \sim \text{Uniforme}(0,1)^2$

Para el piloto PSU las unidades muestrales corresponden a las unidades educativas⁵, y los pesos asignados a cada una se calculan considerando 3 factores, el primero, corresponde a la pertenencia de la UE a una solución óptima calculada mediante un método de optimización conocido como “Programa de Programación Lineal” (PPL), el segundo, es la matrícula de 4° medio en la UE y el tercero, un ponderador β calculado en base a la disponibilidad de estudiantes para cada prueba electiva, respecto a la disponibilidad de la prueba de Química. Finalmente, el peso es:

- 1) β * **Matricula 4° medio** si la UE pertenece a la solución óptima.
- 2) **Matricula 4° medio** si la UE no pertenece a la solución óptima.

En el caso de los establecimientos polivalentes, para asegurar que sean seleccionados en su totalidad, es necesario verificar si hay unidades educativas seleccionadas en un estrato, cuyos RBD estén en un estrato diferente. De ser así, se debe seleccionar la unidad en el otro estrato y se debe descontar de la cantidad necesaria en ese estrato.

Sean:

- Z_j = Zona; con $j = 1, 2, 3$, Talque
 - » Z_1 = Zona Norte,
 - » Z_2 = Zona Centro y
 - » Z_3 = Zona Sur
- W_k = Grupo; $k = 1, 2, 3, 4, 5$
 - » W_1 = Humanista-Científico Particular Pagado:
 - » W_2 = Humanista-Científico Particular Subvencionado
 - » W_3 = Humanista-Científico Municipal
 - » W_4 = Técnico Profesional Particular Subvencionado
 - » W_5 = Técnico Profesional Municipal
- Y_{jk} = Estrato, correspondiente a Z_j del Grupo W_k
- X_i = Unidad educativa
- $n_{Y_{jk}}$ = Tamaño muestral del estrato Y_{jk}

Algoritmo

1. Para cada Y_{jk}
2. Para cada X_i disponible en Y_{jk}
 - a) Generar $u_i = \text{random}(0,1)$
 - b) Hacer $k_i = u_i^{\frac{1}{w_i}}$

5 Cada “Unidad Educativa” es un elemento único dentro de la población, y está compuesta por un Rol Base de Datos (“RBD”) y un código de enseñanza. Por ejemplo, si un establecimiento educacional polivalente (con un RBD “9999”), tiene dos modalidades de enseñanza (Científico Humanista y Técnico Industrial), entonces éste se consideraba como dos unidades distintas dentro de la población: “9999_310” y “9999_510”.

3. Ordenar los valores k_i de forma decreciente y seleccionar la cantidad $n_{Y_{jk}}$ de establecimientos necesarios en el estrato Y_{jk} con los valores de k_i más altos.
4. Si $W_i \in Y_{jk}$ y Y_{jk} , $jk \neq jk'$ entonces seleccionar $W_i \in Y_{jk'}$ y hacer $n_{Y_{jk'}} = n_{Y_{jk}} - 1$
5. Repetir para cada estrato

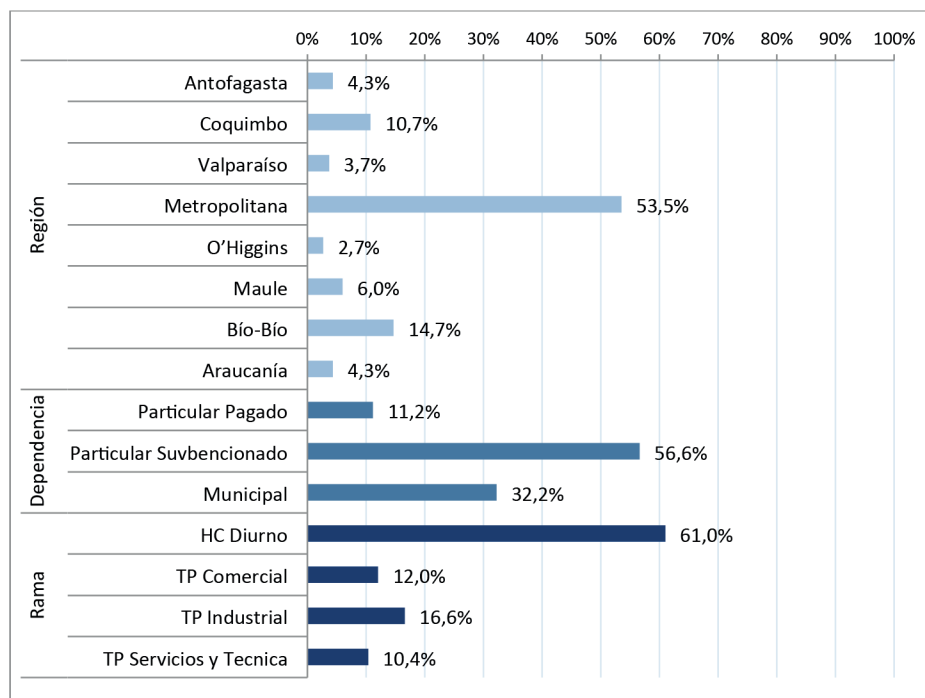
Como se mencionó, se requiere la participación de los establecimientos completos, incluyendo todas sus posibles unidades educativas. Esto es, si se selecciona una unidad de cierto establecimiento en un estrato y este establecimiento posee otra unidad en un estrato diferente, es necesario agregar el paso 4.

Una vez seleccionados los establecimientos, el DEMRE se comunicó con ellos para informar los detalles de la aplicación piloto (ver sección “Plan de difusión”). En el caso de que algún establecimiento declinara la invitación, se estableció una estrategia de reemplazo que buscó una nueva unidad educativa de similares características dentro de la misma comuna.

Cierre de muestra

En base a todos los procedimientos anteriormente descritos, para la aplicación piloto se seleccionó un total de 38.307 estudiantes, distribuidos en 484 unidades educativas. En el Gráfico 1, se presentan las características principales de estos establecimientos. La desagregación por regiones puede verse en el Anexo 2.

GRÁFICO 1. CARACTERÍSTICAS DE LOS ESTABLECIMIENTOS SELECCIONADOS PARA PILOTAJE



En el gráfico se observa que del total de establecimientos convocados, la mayor cantidad corresponde a establecimientos Humanista Científico de modalidad diurna (61.0%). En tanto, establecimientos de rama Técnico Profesional corresponden a un 39.0% del total, en las tres modalidades señaladas.

La mayoría de establecimientos seleccionados son de dependencia particular subvencionada (56.6%), lo cual es consistente con la distribución de estudiantes en el sistema escolar. A su vez, se seleccionaron 156 establecimientos municipales y 54 particulares pagados, que representan el 32.2% y 11.2% de la muestra, respectivamente. Por razones relacionadas con la distribución de la población en el país, el 53.5% de los establecimientos que se seleccionaron fueron de la Región Metropolitana.

Con estos establecimientos y cursos seleccionados, la muestra final convocada a rendir el piloto estuvo compuesta por 38.307 estudiantes. Estos se distribuyeron por región, sexo y prueba electiva de acuerdo con lo que se presenta en la Tabla 5.

TABLA 5. MUESTRA TEÓRICA PILOTO PSU 2017, SEGÚN REGIÓN, SEXO Y PRUEBA ELECTIVA

Indicador	Categoría	Frecuencia	Porcentaje
Región	Antofagasta	1.974	6,2%
	Coquimbo	3.239	7,1%
	Valparaíso	978	3,1%
	Metropolitana	21.931	54,9%
	O'Higgins	791	2,5%
	Maule	1.885	6,0%
	Bío-Bío	5.787	14,7%
	Araucanía	1.722	5,5%
Sexo	Mujer	19.611	47,0%
	Hombre	18.014	51,2%
Prueba electiva	Historia	17.591	39,2%
	Ciencias-Biología	8.898	19,9%
	Ciencias-Física	5.620	12,5%
	Ciencias-Química	6.149	13,7%
	Ciencias-Técnico Profesional	6.575	14,7%

Operativa

Trabajo de campo

Para llevar a cabo la aplicación, tanto del piloto como de la PSU oficial, se divide el país en unidades geográficas regionales y comunales denominadas Sedes de Rendición. A su vez, las sedes de rendición se constituyen por uno o más locales de aplicación de pruebas. La aplicación del piloto PSU se llevó a en 70 locales de aplicación, los días martes 5 y miércoles 6 de septiembre.

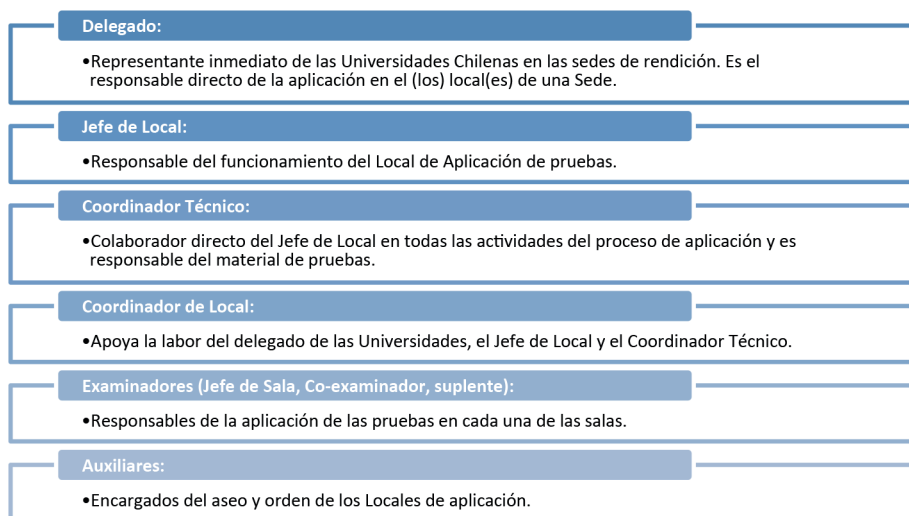
Para la realización del piloto se contó con la autorización oficial del Ministerio de Educación, considerándose esta instancia como un “Cambio de Actividad Lectiva”. Para justificar la asistencia de los estudiantes a la aplicación piloto y responder al control de subvenciones, el DEMRE hizo entrega de la asistencia oficial a los establecimientos educacionales. Un ejemplo de listado de asistencia puede verse en el Anexo 4.

Para evaluar el posible efecto en el orden de rendición de las pruebas obligatorias (Lenguaje y Matemática), algunas comunas iniciaron la aplicación con la prueba de Matemática, mientras otras comenzaron con la de Lenguaje y Comunicación (ver Anexo 3 para mayor detalle). Respecto de las pruebas electivas, estas se mantuvieron fijas: todos los establecimientos rindieron Ciencias el día 5 y la prueba de Historia fue rendida el día 6.

Equipo de aplicación

El equipo de aplicación de pruebas estuvo compuesto, preferentemente, por los docentes de los establecimientos que funcionaron como locales de aplicación. Solo quienes trabajan en preuniversitarios, o hayan sido objetados por su desempeño en procesos anteriores, están inhabilitados para participar. La composición del equipo en orden jerárquico se muestra en la Figura 3.

FIGURA 3: COMPOSICIÓN DEL EQUIPO DE APLICACIÓN



Una de las responsabilidades principales durante la aplicación es resguardar del material y asegurar un correcto desarrollo del proceso. Por lo tanto, el equipo aplicador no está autorizado para abrir, leer o revisar los folletos, usar celulares, cámaras u otro dispositivo electrónico, ni desarrollar otra tarea durante el proceso de aplicación.

Para asegurar el control de calidad de la aplicación, existe una serie de documentos que ayuda a supervisar y monitorear la adecuada realización del piloto. Previo a la aplicación piloto, se realizó una reunión en cada local de aplicación para informar al equipo las reglas del proceso y para establecer los protocolos y responsabilidades de cada uno. Luego de esa reunión se identificaron los lugares relevantes dentro del local (e.g. para procedimientos de

evacuación en caso de emergencia y salas de acopio de material), se prepararon las salas para la aplicación, y se informó sobre los procedimientos de emergencia –por ejemplo, en caso de sismo o incendio.

Traslado de material

En la Región Metropolitana, el traslado del material fue responsabilidad de los Coordinadores Técnicos. Los coordinadores se presentaron en el DEMRE para retirar las cajas con el material correspondiente a cada día de aplicación y trasladarlo al local. Una vez finalizado el día de aplicación, el coordinador retornó todo el material aplicado a las dependencias del DEMRE.

En regiones, el material fue despachado con algunos días de anticipación y quedó bajo la custodia de Carabineros de Chile, en una comisaría cercana al(los) local(es) de aplicación. En estos casos, fue responsabilidad del Delegado reunirse con el representante de carabineros a cargo del proceso y coordinar el resguardo y traslado del material a cada local de aplicación. El Delegado también fue el responsable de gestionar el traslado de material con los encargados de cada local. Al interior del local de aplicación, el Coordinador Técnico fue responsable de resguardar correctamente el material. Los Coordinadores Técnicos recibieron los folletos de las pruebas correspondientes a su local ordenados en lotes agrupados por sala y en cajas selladas. Cada folleto y su respectiva hoja de respuesta fueron entregados en un envase sellado.

Condiciones para rendir las pruebas piloto

El horario de citación de los estudiantes fue a las 8:00 am para ambos días de aplicación. Además, el DEMRE informó a colegios y a estudiantes -el local de aplicación al que debía concurrir cada estudiante, y la sala en la que rendirían las pruebas. Esto facilitó el ingreso e identificación de los estudiantes. Para ingresar a rendir cada prueba, los estudiantes debieron identificarse con su cédula de identidad u otro documento con fotografía que permitiera establecer su identidad (e.g. comprobante de solicitud de cédula de identidad, pase escolar). Se aceptó el ingreso de estudiantes a las salas hasta media hora después de comenzada la aplicación.

Los estudiantes no citados al piloto que se presentaron a rendir las pruebas y que no pertenecían a un establecimiento en la muestra, debían ser autorizados por el Jefe de Local, quien debía comunicarse con el DEMRE para autorizar el ingreso del estudiante. Estos estudiantes fueron autorizados a rendir solo si cursaban IV medio, presentaban un documento de identificación apropiado y el local contaba con material de aplicación.

Los estudiantes firmaron un documento donde expresaron su autorización o rechazo a que el DEMRE entregara sus resultados de la prueba piloto a su establecimiento educacional.

Cambios de prueba electiva

En algunos casos, los estudiantes solicitaron cambiar la prueba electiva que se les había asignado. Esta solicitud fue atendible siempre y cuando existiese material disponible en el

local (por estudiantes ausentes o material de reserva). Los estudiantes pudieron solicitar cambio de prueba electiva entre Ciencias e Historia, Geografía y Ciencias Sociales —o viceversa—, o dentro de los módulos electivos de Ciencias⁶.

Personas en situación de discapacidad

El DEMRE implementó ajustes que permitieran a estos estudiantes participar de la aplicación piloto, bajo una lógica de inclusión. Esto se realizó considerando los recursos disponibles y la colaboración de los establecimientos para proporcionar condiciones especiales a estos estudiantes.

A diferencia de los trámites y protocolos que deben seguir los casos especiales para la rendición de la PSU oficial⁷, para el piloto solo se solicitó a los establecimientos informar de la situación al DEMRE. Para ello, cada establecimiento informó en la Plataforma Piloto (ver Figura 4) los estudiantes en esta situación y las sugerencias para resguardar la igualdad de oportunidades en la rendición de las pruebas. Dentro de las facilidades otorgadas se cuentan las siguientes:

- Realización de la prueba en sala con adecuaciones
- Modificación del tiempo de rendición (se agregó un 25% por sobre el tiempo de cada prueba) y permitir pausas.
- Ayuda de un tutor o especialista que colaboró en el registro de datos personales, lectura de la prueba, o completación de la hoja de respuestas. La ayuda recibida consistió en la asistencia directa de dos examinadores externos o de personal especializado proveniente del establecimiento del estudiante.
- Modificación de las hojas de respuestas o folletos. A los estudiantes en situación de discapacidad que informaron un déficit visual, se les imprimieron folletos y hojas de respuesta con letra ampliada.
- Uso de lámpara, atril, lupa o lente especial y audífonos.

6 En Ciencias, los estudiantes podían realizar cambios entre los módulos de Biología, Física o Química. Únicamente los estudiantes egresados de establecimientos Técnico-Profesionales podían cambiar entre uno de los módulos antes mencionados y el módulo Técnico-Profesional.

7 Ver: DEMRE (s.f). "Solicitud de Ajustes, Adecuaciones o Apoyos en la Rendición para Postulantes en Situación de Discapacidad". En línea, disponible en: <http://psu.demre.cl/inscripcion/postulantes-con-discapacidad> [Consultado el 26 de enero de 2018]

FIGURA 4. MÓDULO DE SELECCIÓN DE AJUSTES PARA ESTUDIANTES EN SITUACIÓN DE DISCAPACIDAD. PLATAFORMA PILOTO PSU

Todos los estudiantes que recibieron algún tipo de ajuste para la rendición del piloto firmaron una declaración jurada donde declaraban estar en conocimiento de que el ajuste proporcionado para la rendición de las pruebas piloto no era vinculante para la rendición de la PSU en el Proceso Oficial de Admisión 2018.

En total, 37 estudiantes recibieron ajustes durante la aplicación piloto, la Tabla 6 muestra el tipo de discapacidad de estos estudiantes y la Tabla 7 los ajustes entregados.

TABLA 6. ESTUDIANTES EN SITUACIÓN DE DISCAPACIDAD, PILOTO PSU 2017

Tipo de discapacidad	Cantidad de estudiantes
Auditiva	2
Física (Motora)	10
Intelectual (Cognitiva)	11
Psíquica	3
Visual	11

TABLA 7. AJUSTES ENTREGADOS A ESTUDIANTES EN SITUACIÓN DE DISCAPACIDAD, PILOTO PSU 2017

	Cantidad de estudiantes solicitando el ajuste
Ayuda de tutor	13
Impresión ampliada de prueba y hoja de respuesta (macrotipo)	9
Lectura de prueba en voz alta	5
Sala con adecuaciones (1° piso, para una o pocas personas, etc.)	15
Tiempo extra	20
Uso de accesorios (lámpara, atril, lupa, audífono, etc.)	5
Otras medidas	6

Resumen de la aplicación

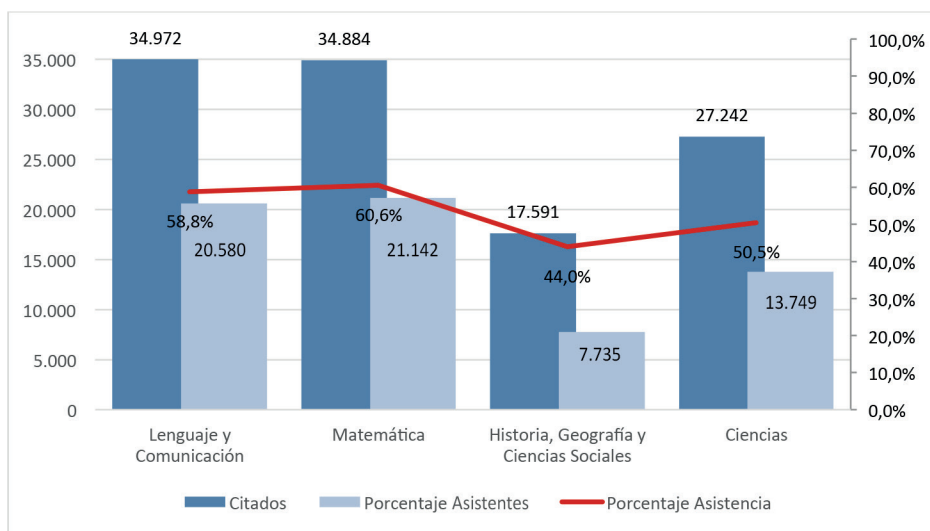
A continuación, se presentan datos sobre la asistencia al pilotaje de forma resumida, según variables de contexto y de tipo de establecimiento. En primer lugar, se presenta la asistencia y su porcentaje por prueba. En segundo lugar, se muestra la asistencia en variables relativas a las características de los estudiantes: sexo y región de origen. En tercer lugar, se muestra la asistencia por variables relativas a los establecimientos educacionales de los estudiantes: dependencia administrativa y rama de enseñanza. Por último, se describe la muestra efectiva, es decir, la distribución empírica de la muestra teórica definida en la sección “Cierre de Muestra”.

Asistencia por prueba

De un total de 38.307 estudiantes citados, 25.837 se presentaron a rendir al menos una prueba. Es decir hubo una tasa de asistencia al Piloto PSU del 67,4%.

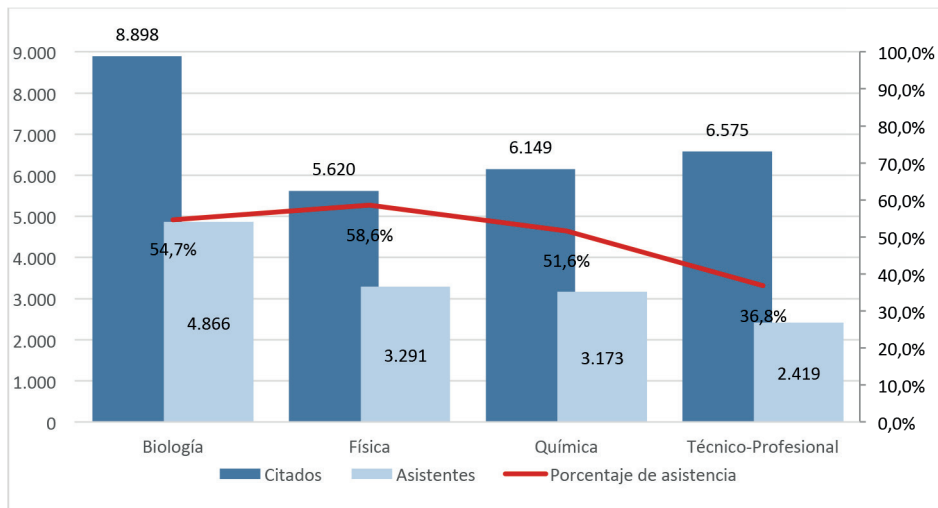
El Gráfico 2 muestra la cantidad de estudiantes asistentes al piloto respecto de los citados y el porcentaje de asistencia por disciplina. Como se observa, la tasa de asistencia fue de un 58,8% en Lenguaje y Comunicación, 60,6% en Matemática, en Historia, Geografía y Ciencias Sociales 44,0% y 50,5% en Ciencias.

GRÁFICO 2. ASISTENCIA A PILOTO PSU, POR PRUEBA



El Gráfico 3 presenta la asistencia a cada módulo electivo que compone la prueba de Ciencias. Se observa una tasa de asistencia sobre 54,7% en Biología, 58,6% en Física, 51,6% en Química y 36,8% en Técnico Profesional.

GRÁFICO 3. ASISTENCIA A PILOTO PSU, PRUEBAS DE CIENCIAS



Asistencia por sexo y región

A continuación, el Gráfico 4 presenta las características de los asistentes según sexo de los estudiantes, donde el porcentaje de asistencia es similar para ambos sexos, 69.3% para hombres y 68.1% en mujeres.

GRÁFICO 4. ASISTENCIA A PILOTO PSU, SEGÚN SEXO

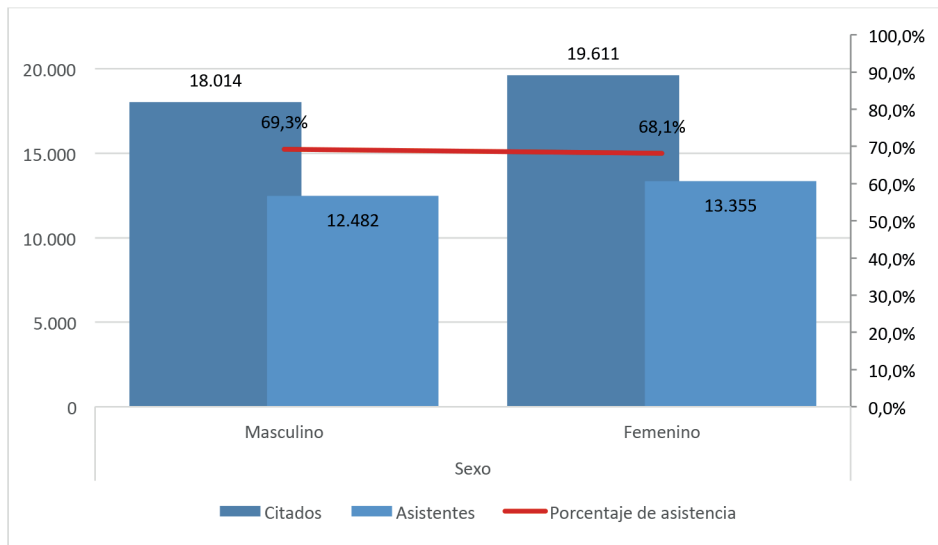
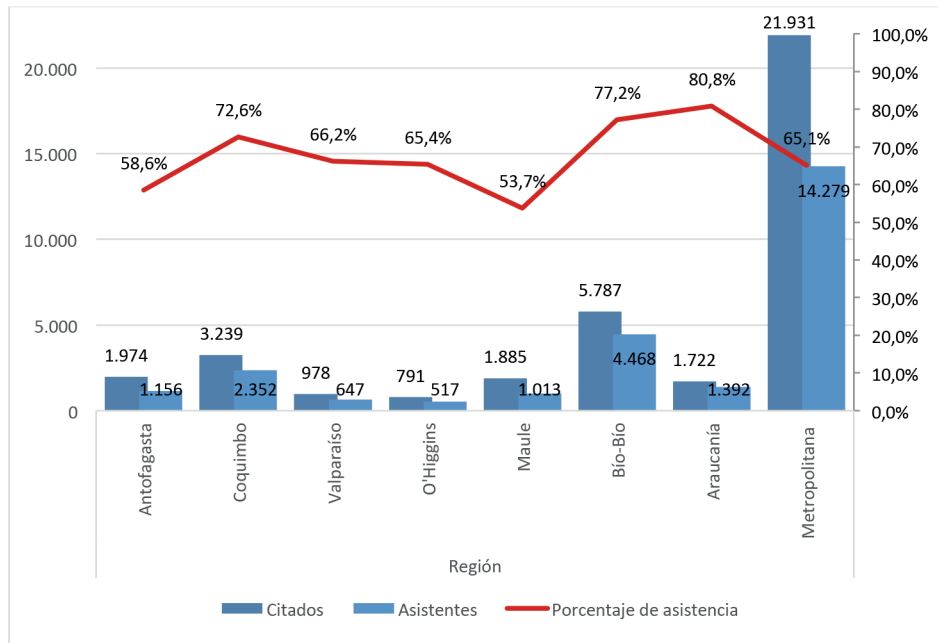


Gráfico 5 presenta las características de los asistentes según la región de origen de los estudiantes, donde el porcentaje de asistencia varía entre un 53.7% a un 80.8%, donde la región del Maule fue la que tuvo el menor porcentaje de asistencia, mientras que la región de la Araucanía el mayor porcentaje.

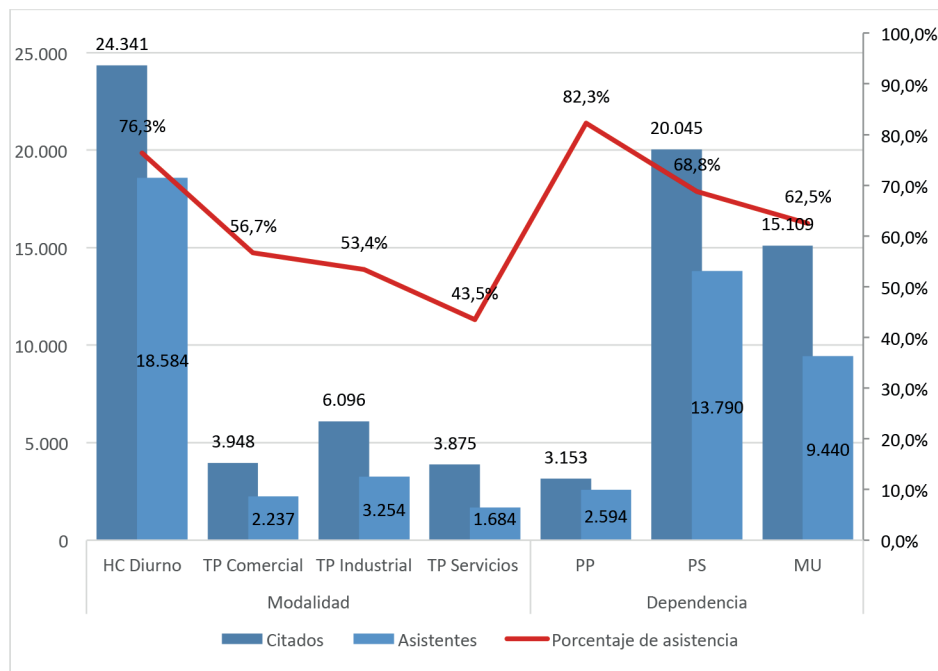
GRÁFICO 5. ASISTENCIA A PILOTO PSU, SEGÚN REGIÓN



Asistencia por tipo de establecimiento

A continuación, se presentan las características de los establecimientos educacionales de los estudiantes que participaron en el piloto. Tal como se muestra en el Gráfico 6, la asistencia se distribuyó de distinta forma, dependiendo del establecimiento educacional de origen.

GRÁFICO 6. ASISTENCIA A PILOTO PSU, SEGÚN MODALIDAD DE ENSEÑANZA Y DEPENDENCIA ADMINISTRATIVA



Al realizar el análisis por modalidad educativa, se observa que la mayor participación ocurre entre los estudiantes provenientes de establecimientos científico humanista, seguido por aquellos provenientes de establecimientos técnicos de rama comercial. Este resultado podría vincularse a que son los estudiantes HC quienes mayoritariamente se preparan para la PSU, con la expectativa de acceder a la educación superior por esta vía.

Respecto de la dependencia administrativa de los establecimientos, se observa que son los estudiantes de establecimientos particulares quienes tuvieron una mayor asistencia en el piloto PSU, con sobre el 82.3%. La Tabla 8 ilustra cómo se comportan las variables de modalidad de enseñanza y dependencia del establecimiento respecto del ausentismo.

TABLA 8. PORCENTAJE DE AUSENTES, SEGÚN COMBINACIONES DE MODALIDAD DE ENSEÑANZA CON DEPENDENCIA ADMINISTRATIVA

			Dependencia		
			Particular Pagado	Particular Subvencionado	Municipal
Modalidad de enseñanza	Humanista Científico Diurno	Citados	3.153	12.656	8.532
		Ausentes	2.594	9.688	6.302
		Tasa de inasistencia	18%	23%	26%
	Técnico Comercial	Citados	0	2.333	1.615
		Ausentes	0	1.378	859
		Tasa de inasistencia	0%	41%	47%
	Técnico Industrial	Citados	0	2.920	3.176
		Ausentes	0	1.699	1.555
		Tasa de inasistencia	0%	42%	51%
	Técnico Servicios	Citados	0	2.136	1.739
		Ausentes	0	1.025	659
		Tasa de inasistencia	0%	52%	62%

En la tabla se observa que la mayor cantidad de ausentes se concentra en la dependencia municipal, de modalidad técnica en servicios. No obstante, en la misma modalidad, pero de dependencia particular subvencionada, la inasistencia de los estudiantes bordea el promedio. Se observa, también, que la menor proporción de estudiantes ausentes se encuentra en la modalidad humanista científico de dependencia privada, tanto particular como subvencionada.

Muestra efectiva

A continuación, la Tabla 9 presenta la composición de la muestra efectiva de estudiantes asistentes al piloto. En la tabla se muestra el total de citados, por cada variable y categoría, y el porcentaje de cada categoría respecto del total de los citados. Lo mismo para los

asistentes, se muestra el total y la proporción de cada categoría respecto del total de asistentes. La columna variación muestra la diferencia porcentual, por cada categoría, entre los asistentes y los citados.

TABLA 9. MUESTRA EFECTIVA SEGÚN VARIABLES DE CARACTERIZACIÓN

Variable	Categoría	Citados		Asistentes		Variación
		Cantidad	% respecto del total de citados	Cantidad	% respecto del total de asistentes	
Rama*	HC Diurno	24.341	63,6%	18.584	72,1%	8,5%
	TP Comercial	3.948	10,3%	2.237	8,7%	-1,6%
	TP Industrial	6.096	15,9%	3.254	12,6%	-3,3%
	TP Servicios	3.875	10,1%	1.684	6,5%	-3,6%
Total		38.260	100,0%	25.759	100,0%	
Dependencia	Particular Pagado	3.153	8,2%	2.594	10,0%	1,8%
	Particular Subvencionado	20.045	52,3%	13.790	53,4%	1,1%
	Municipal	15.109	39,4%	9.440	36,6%	-2,8%
Total		38.307	100,0%	25.824	100,0%	
Sexo**	Femenino	18.014	47,9%	12.482	48,3%	0,4%
	Masculino	19.611	52,1%	13.355	51,7%	-0,4%
Total		37.625	100%	25.837	100,0%	
Región	Antofagasta	1.974	5,2%	1.156	4,5%	-0,7%
	Coquimbo	3.239	8,5%	2.352	9,1%	0,7%
	Valparaíso	978	2,6%	647	2,5%	0,0%
	O'Higgins	791	2,1%	517	2,0%	-0,1%
	Maule	1.885	4,9%	1.013	3,9%	-1,0%
	Bío-Bío	5.787	15,1%	4.468	17,3%	2,2%
	Araucanía	1.722	4,5%	1.392	5,4%	0,9%
	Metropolitana	21.931	57,3%	14.279	55,3%	-2,0%
Total		38.307	100%	25.824	100%	

* No se consideran 65 estudiantes asistentes de otras ramas de enseñanza

** Existen 682 registros de estudiantes sin especificar su sexo

Como se observa en la tabla precedente, la mayor variación se encuentra en los establecimientos HC diurnos, los cuales se encuentran sobre representados, en detrimento de los establecimientos técnicos, los que se encuentran sub representados. Otra diferencia se observa en la dependencia municipal con una leve sub representación de estudiantes en la muestra efectiva. La variación en el resto de las categorías es baja, por lo tanto, se concluye que la distribución de la muestra efectiva es cercana a la muestra teórica.

Análisis

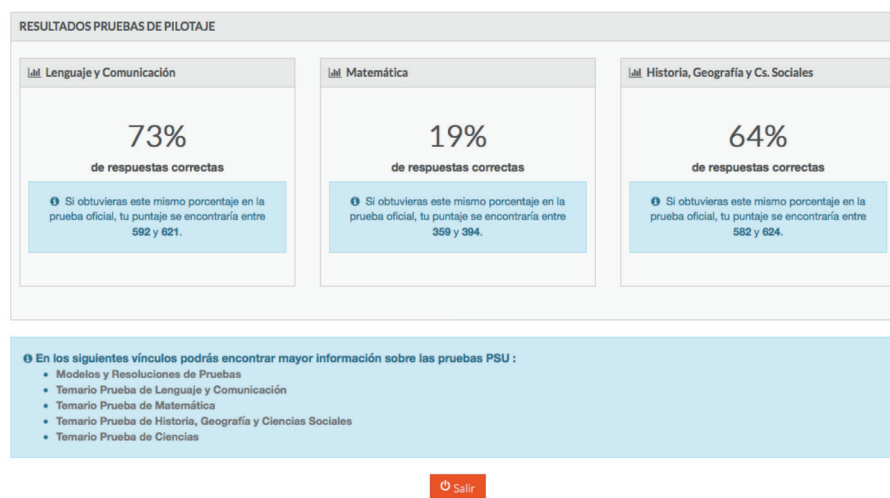
Informes a establecimientos educacionales y estudiantes

Tras obtener los resultados del piloto para cada estudiante (porcentaje de respuestas correctas), el DEMRE generó reportes individuales y agregados de desempeño. Los primeros fueron enviados a los estudiantes y los reportes agregados fueron enviados a los establecimientos educacionales.

El reporte individual presentó el porcentaje de respuestas correctas obtenido en cada prueba rendida por el estudiante. Dado que en la aplicación piloto se testean preguntas para utilizar en aplicaciones futuras, no es posible reportar puntajes con la precisión de una aplicación oficial. Sin embargo, el DEMRE entiende la necesidad de los estudiantes por contar con este tipo de información.

Para entregar información técnicamente correcta y de utilidad a los estudiantes, se entregó un rango de puntaje PSU de referencia asociado al porcentaje de respuestas correctas obtenido. Estos rangos de puntaje fueron calculados en referencia a la escala de puntajes de los años 2015, 2016 y 2017 para ese porcentaje de respuestas correctas. Para visualizar los resultados, se envió un correo electrónico a cada estudiante que contenía un link a la plataforma de resultados. Un ejemplo de la pantalla visualizada por los estudiantes se encuentra en la Figura 5.

FIGURA 5. PLATAFORMA ENTREGA DE RESULTADOS APLICACIÓN PILOTO A ESTUDIANTES



El reporte a nivel agregado presentó los resultados por prueba, a nivel nacional, regional, comunal y de cada establecimiento. Además, se entregó como referencia, el resultado global obtenido por establecimientos educacionales con características similares. Por último, se entregaron los resultados del establecimiento en cada prueba, según áreas/ejes temáticos y habilidades cognitivas. Cada establecimiento recibió, además, una nómina con

los resultados individuales de aquellos estudiantes que accedieron a entregar sus datos mediante la firma de un consentimiento informado.

Como las preguntas del piloto estaban probándose, los resultados debían tratarse e interpretarse con sumo cuidado. Con este objetivo, el DEMRE especificó que los resultados del piloto responden a la motivación y esfuerzo que el/la estudiante dedicó al responder cada prueba. Además, se enfatizó que dado que la prueba piloto no es igual a la PSU oficial, los puntajes debían usarse de modo referencial.

Sobre los resultados entregados a los establecimientos —y por los mismos argumentos ya mencionados—, estos no debían usarse como indicador de desempeño de los estudiantes. Tal como con los resultados de la PSU oficial, este tipo de evaluaciones no pueden dar cuenta de todo lo que se hace en la escuela, ni menos deberían utilizarse para clasificar establecimientos o generar rankings (Sánchez, 2016). Además, debido a los objetivos de la PSU y el piloto, estos resultados no hablan de calidad, puesto que no se evalúan todas las asignaturas ni contenidos cubiertos en la educación media.

Análisis de resultados

Teniendo los datos y los resultados de las pruebas, la Unidad de Desarrollo, Análisis e Investigación del DEMRE realizó los análisis psicométricos para determinar la calidad de los ítems y, por tanto, la factibilidad de estos para ser utilizados en pruebas oficiales. A continuación, se describen los distintos análisis psicométricos realizados a las Pruebas Piloto. Se exponen los análisis, las fórmulas y procedimientos para llevarlos a cabo. A su vez, se muestran los criterios definidos por el DEMRE que aseguran que los ítems de la PSU cumplen con los estándares de calidad para una prueba estandarizada de altas consecuencias.

En la evaluación de calidad de los ítems se utiliza la Teoría Clásica del Test (CTT, *Classic Theory of Tests*) y se complementa este análisis con el modelo de Teoría de Respuesta al Ítem (IRT, *Item Response Theory*). De esta forma, es posible establecer *modelos* capaces de evaluar las *propiedades psicométricas* de los instrumentos de medición. Específicamente, se estudian aquellos factores que influyen sobre las puntuaciones obtenidas en los test y sus ítems, proponiendo modelos que permitan controlar y minimizar los factores de error. Estos factores de error inciden en las estimaciones realizadas a partir del instrumento de medición.

El concepto *propiedades psicométricas* refiere al análisis de las características métricas del ítem, que dan cuenta de la idoneidad del instrumento para medir lo que se desea medir, minimizando el error. En un sentido amplio, lo anterior puede ser definido como un proceso que se centra en el análisis del instrumento, en los siguientes tres niveles (Muñiz, Teoría Clásica de los Test, 2003):

1. Respecto de su comportamiento en tanto escala: refiere al estudio de la confiabilidad y validez del instrumento⁸.

8 Los análisis de confiabilidad y validez se realizan a los resultados de la Prueba Oficial.

2. Respecto de las características de sus ítems: se orienta a los análisis de las características propias de cada ítem, tales como su dificultad, discriminación, omisión, comportamiento de los distractores y funcionamiento diferencial del ítem.
3. Una combinación de ambas.

La evaluación de los ítems, tiene por objetivo apoyar el proceso de ensamblaje de las pruebas PSU, que se realiza con ítems previamente validados en una muestra representativa de la población. Este proceso de pilotaje asegura que los ítems utilizados cumplen los estándares mínimos y suficientes para asegurar la calidad técnica del instrumento.

Teoría Clásica del Test (CTT)

En cuanto a su formulación general, la CTT propone un modelo lineal en el que se asume que la puntuación obtenida por el sujeto i en un test (X o puntuación empírica) se compone de dos elementos aditivos: la puntuación verdadera obtenida por el sujeto (V) y el error de medida presente en las puntuaciones observadas (e). Formalmente, lo anterior queda definido en la Ecuación 1.

Ecuación 1.

$$X_i = V_i + e_i$$

Idealmente, en la TCT deben existir como mínimo dos formas paralelas⁹ (j y k) de un mismo test para comprobar el modelo. Dos formas de un test son consideradas paralelas si la varianza (σ^2) de los errores (e) es la misma¹⁰ para las dos formas (j y k), y si las puntuaciones verdaderas (V) obtenidas tras la aplicación de las dos formas es igual [$V_j = V_k$].

La TCT ha formulado distintos criterios de valoración de la calidad de los ítems, entre los que destacan por su utilidad los siguientes: índice de dificultad, índice de discriminación y análisis de los distractores.

9 El paralelismo en este caso se entiende como los ítems comunes o de anclaje entre las distintas formas que componen un test.

10 La expresión: [$\sigma^2(e_j) = \sigma^2(e_k)$] implica que la distribución de los errores es homogénea.

Índice de Dificultad

El *índice de dificultad* de un ítem (p) se define como la proporción de sujetos que responde correctamente al mismo, en función del total de individuos que abordaron el ítem. Definido en la Ecuación 2.

Ecuación 2.

$$p_i = \frac{\sum A_i}{N}$$

- p_i : dificultad del ítem.
- A_i : personas que acertaron el ítem.
- N : número de individuos que intentaron responder el ítem.

El índice de dificultad de un ítem (p) admite valores dentro de un intervalo que va de 0 a 1. Cuando p se acerca al valor 1, indica que muchos individuos han contestado correctamente el ítem, por lo que este resulta fácil. Por el contrario, a medida que p se aproxima o alcanza el valor 0, indica que el ítem en cuestión resulta difícil para los sujetos de la muestra o población en que fue aplicado.

TABLA 10. PARÁMETROS DE DIFICULTAD

p	Clasificación
0,00 – 0,39	Difícil
0,40 – 0,59	Mediano
0,60 – 1,00	Fácil

Para efectos de los ítems que componen la batería de pruebas PSU, el índice de dificultad utilizado por el DEMRE se encuentra entre 0,10 y 0,80.

Índice de Discriminación

De manera amplia, el *índice de discriminación* puede ser definido como la correlación que se establece entre las puntuaciones que obtienen los sujetos en un ítem particular y la puntuación total en el test.

Según Muñiz *et. al.* (2005), una pregunta tiene poder de discriminación si es capaz de distinguir entre los sujetos que puntúan alto en una prueba de aquellos que puntúan bajo. Por lo tanto, “es condición de calidad de un ítem el que sea contestado correctamente por los estudiantes con mayor puntuación” (pág. 61).

El DEMRE establece los índices de discriminación de los ítems que componen los instrumentos de la batería de pruebas PSU, por medio de correlaciones. Específicamente, dadas las características de los ítems de selección múltiple y los requerimientos de la CTT, se utiliza el índice de correlación biserial (r_b). Este permite relacionar respuestas de tipo dicotómicas y discretas (acierto versus no acierto), con una escala de tipo continua

(puntuación total sobre la escala o prueba), evaluando así el grado de asociación y, por extensión, de discriminación de los ítems (Ecuación 3).

Ecuación 3.

$$r_b = \frac{\bar{x}_c - \bar{x}_t}{s_t} * \frac{p}{y}$$

- \bar{x}_c : promedio en la prueba del grupo que contesta correctamente el ítem.
- \bar{x}_t : promedio del grupo total en la prueba.
- s_t : desviación estándar del grupo total.
- p : proporción de sujetos que contesta correctamente la pregunta.
- y : ordenada correspondiente al valor de la puntuación típica (z) que deja por debajo un área igual a p .

Los criterios internacionales para clasificar un índice de correlación biserial, son expuestos por Muñiz, *et al.* (2005). En la Tabla 11 se muestran los puntos de corte utilizados para clasificar el índice de correlación biserial (r_b).

TABLA 11: CLASIFICACIÓN DEL ÍNDICE DE CORRELACIÓN BISERIAL

r_b	Clasificación del ítem
Igual o mayor que 0,40	Discrimina muy bien
Entre 0,30 y 0,39	Discrimina bien
Entre 0,20 y 0,29	Discrimina poco
Entre 0,10 y 0,19	Limite. Se debe mejorar
Menor de 0,10	Carece de utilidad para discriminar

Análisis de las opciones incorrectas o “distractores”

En un ítem, se denomina opción incorrecta o distractor a sus opciones incorrectas de respuesta. Como señalan Muñiz *et al.* (2005), es fundamental que todas las opciones incorrectas incluidas, en tanto opciones de respuesta al ítem, resulten “(...) igualmente atractivas para las personas evaluadas que desconocieren la respuesta correcta” (pág. 70). Analizar la distribución de las respuestas de los examinados, explica el funcionamiento de los distractores.

Por ejemplo, en un ítem, un índice de discriminación bajo puede explicarse porque alguno de los distractores fue elegido tanto por los individuos con bajo desempeño como por los de alto desempeño. En este caso, es conveniente cambiar dicho distractor por uno más adecuado y volver a pilotear el ítem. Además, es posible que algún distractor no sea elegido por los examinados (lo que se denomina *distractor vacío*), lo que también afecta el poder discriminativo del ítem.

Teoría de Respuesta al Ítem (IRT)

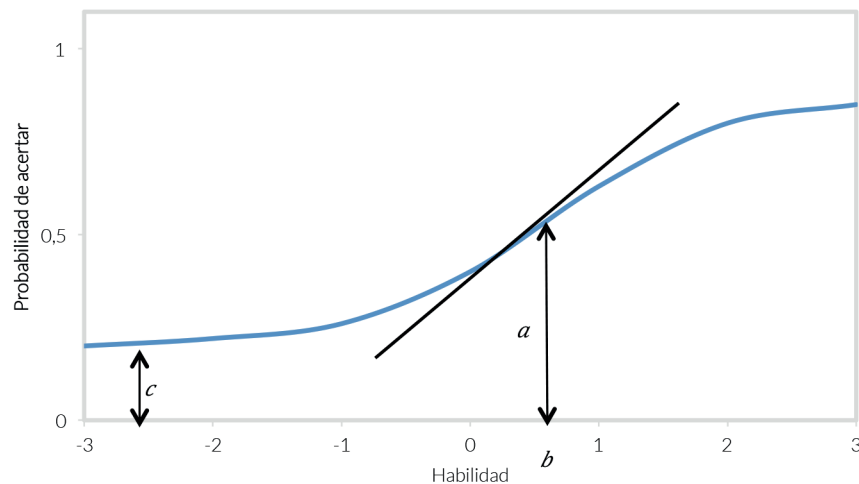
La IRT, constituye un nuevo enfoque dentro de la psicometría, cuyo objetivo es resolver problemáticas que escapan del alcance de la CTT. Tal como su nombre lo indica, el enfoque IRT se basa en las propiedades de los ítems más que en el test en términos globales.

La mayor contribución de este enfoque es que permite obtener mediciones invariantes respecto de los instrumentos utilizados y los sujetos implicados. Es decir, con el uso de IRT se obtendrán mediciones que no cambian en función del test usado e instrumentos de medida con propiedades independientes de los sujetos evaluados (Muñiz, 1997).

IRT busca estimar las características de los ítems, y además, estimar la probabilidad de que un estudiante responda correctamente. Para esto, los modelos IRT asumen la existencia de una relación funcional entre los valores de la variable que miden los ítems (dicho de otra forma, las puntuaciones de los sujetos en la variable medida) y la probabilidad de acertarlos, denominando a dicha función “Curva Característica del Ítem” (CCI).

En el Gráfico 7, se observa la CCI. En el eje de abscisas se representan los valores de la variable que mide el ítem o “habilidad”, la que se denota como θ . En el eje de ordenadas se representa la probabilidad de responder correctamente el ítem, la que se denota como $P(\theta)$. Como se observa, sujetos con mayor habilidad, poseen mayores probabilidades de acertar el ítem.

GRÁFICO 7. CURVA CARACTERÍSTICA DEL ÍTEM



Fuente: Muñiz (1997)

La CCI se compone de tres parámetros, los cuales pueden analizarse en conjunto o por separado. Si bien son los mismos criterios usados en CTT, su significado no es exactamente el mismo. El primer parámetro, a , es el índice de discriminación y su valor es proporcional a la pendiente de la recta tangente a la CCI. Es decir, cuanto mayor sea la pendiente, mayor será el índice de discriminación.

Por su parte, el parámetro b , o índice de dificultad, es el valor de θ correspondiente al punto de máxima pendiente de la CCI, por lo tanto, la dificultad se mide en la misma escala de θ . Por último, el parámetro c , es la probabilidad de acertar el ítem al azar o “adivinando”.

Procedimiento de análisis

Tras explicitar los modelos principales que se utilizan en los análisis psicométricos, esta sección describe el procedimiento realizado por el DEMRE para analizar, aceptar y rechazar preguntas en el piloto PSU 2017.

Previo al análisis, se eliminan los casos de estudiantes que omitieron toda la prueba. Además, se identifican ítems cuyos distractores tienen un biserial superior a 0.2, estos se envían a los comités constructores para que verifiquen las claves y/o que no exista un error de impresión. Posteriormente, la primera fase, que llamaremos de “purificación del instrumento”, busca obtener resultados más limpios y robustos. Para cada una de las pruebas, se analizan todas las preguntas a través de una correlación puntual con análisis de CTT. La correlación se realiza con el porcentaje de respuesta correcta de cada individuo que contesta el ítem. Tras el cálculo, se procede a eliminar todos los ítems cuyos biserales puntuales son menores a 0,15.

La segunda fase se realiza con las preguntas aceptadas tras la purificación. En esta fase, se implementa un análisis IRT de un parámetro, específicamente el modelo de Rasch. La principal característica de este modelo es que los ítems solo difieren en el parámetro de dificultad, asumiendo que la discriminación para todos los ítems es igual a 1. La forma tradicional de estimar el modelo de Rasch, se presenta en Ecuación 4

Ecuación 4.

$$P(\theta) = \frac{1}{1 + e^{-(\theta-b)}}$$

En el análisis IRT, a través de este modelo, lo que se hace es calcular la *habilidad individual* de cada sujeto, en vez de calcular un puntaje propiamente tal. Para cada sujeto, esta habilidad se calcula sumando todas las probabilidades de responder correctamente (cada ítem). Es decir, si una prueba tiene 75 preguntas, la habilidad individual de una persona se calcula sumando 75 probabilidades de responder correctamente.

El cálculo de esta habilidad individual es necesario para la tercera fase, que consiste en analizar los resultados psicométricos de los ítems piloteados. Para la aceptación o rechazo de ítems de pilotaje, el principal criterio utilizado es el índice de discriminación calculado mediante CTT. Esta vez, la correlación no se realiza con el porcentaje de respuesta correcta de cada individuo, sino con el puntaje IRT (o habilidad individual), calculado previamente.

Los criterios para aceptar o rechazar ítems se exponen en la Tabla 12. Cabe destacar que, la decisión de aceptar un ítem, también cuenta con la condición previa de obtener una dificultad en Rasch-IRT entre -2.75 y 2.75.

TABLA 12. CRITERIOS PARA LA CLASIFICACIÓN DE ÍTEMS

Clasificación		Dificultad	Discriminación	Omisión	Porcentaje respuesta distractores	Biserial distractores
Aprobada		Entre 10-80% (Inclusive)	$\geq 0,25$	$\leq 10\%$	Todos $\geq 5\%$	$\leq 0,05$
Aprobada con comentarios	Distractor vacío	Entre 10-80% (Inclusive)	$\geq 0,25$	$\leq 10\%$	Uno $< 5\%$	$\leq 0,05$
	Distractor compite	Entre 10-80% (Inclusive)	$\geq 0,3$	$\leq 10\%$	Todos $\geq 5\%$	Uno entre 0,05-0,15
	Distractor vacío - compite	Entre 10-80%	$\geq 0,3$	$\leq 10\%$	Uno $< 5\%$	Entre 0,05-0,15
Rechazada	Dificultad	Fuera de 10%-80%				
	Omisión			$> 10\%$		
	Distractor compite					$> 0,2$
	Distractores vacíos				Al menos 2 $< 5\%$	
	Biserial		$< 0,25$			

Cualquier otra situación, lleva al ítem al estado de “revisar” en el que se decide su estado final, lo que implica una mirada extensiva no solo de los estadísticos, sino también de sus componentes (como formato, contenido, clave, etc.).

Funcionamiento Diferencial del Ítem (DIF)

Además del análisis CTT e IRT realizado a los ítems piloto, se realiza un estudio sobre el funcionamiento diferencial de los ítems (DIF, *Differential Item Functioning*). Este análisis DIF busca detectar posibles sesgos analizando la equivalencia entre grupos comparables de individuos que rinden la prueba.

Desde la perspectiva CTT, se dice que un ítem funciona diferencialmente cuando examinados de igual nivel en la variable medida por el test, pertenecientes a diferentes grupos, tienen distinta probabilidad de resolverlo correctamente. Si un ítem no presenta DIF, implica que no hay sesgo, pero si el ítem presenta DIF existen dos posibles causas. Esto puede ocurrir por las diferencias reales que existen entre los grupos en el rasgo subyacente, llamado impacto, o porque el ítem está sesgado. Una de las investigaciones ante la presencia de DIF debe ser un análisis de contenido por parte de expertos en la materia, ya que es imprescindible estudiar las causas y encontrar una explicación teórica de la ocurrencia del mismo.

El proceso inicial para analizar el funcionamiento diferencial de los ítems toma como punto de referencia las variables que se consideran susceptibles de diferencias. Cada variable se categoriza en dos grupos diferentes: *grupo focal* y *grupo referencial*. Es arbitrario establecer la categorización de cada grupo, pero suele reservarse el término focal para el grupo minoritario o que, a priori, se considera posiblemente perjudicado por alguno de los ítems (Muñiz, 2003). Para el caso del análisis del pilotaje de la PSU, en la Tabla 13 se muestran las variables y grupos analizados que –considerando la realidad nacional– podrían presentar DIF.

TABLA 13. VARIABLES Y GRUPOS CONSIDERADOS PARA EL ANÁLISIS DIF

Variable	Grupo Focal	Grupo Referencial
Sexo	Femenino	Masculino
Dependencia	Municipal	Particular Subvencionado
	Particular Subvencionado	Particular Pagado
	Municipal	Particular Pagado
Modalidad	Técnico-Profesional	Humanista-Científico
Zona	Norte (regiones: XV, I a VI)	Metropolitana
	Sur (regiones VII a XIV)	Norte (regiones: XV, I a VI)
	Sur (regiones VII a XIV)	Metropolitana

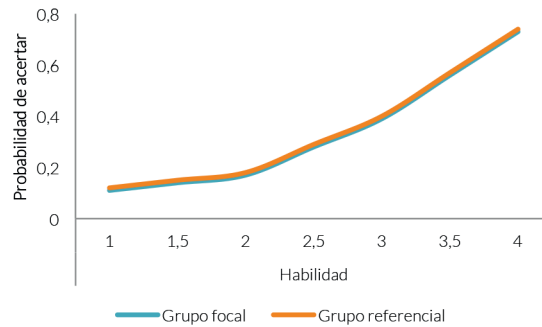
Para el análisis, se utilizan métodos para detectar comportamiento diferencial uniforme y no uniforme. Es decir, para evidenciar si existen diferencias de probabilidad de respuesta correcta y si esta probabilidad es constante o no entre los grupos estudiados.

Funcionamiento Diferencial del Ítem uniforme y no uniforme

Como se mencionó, un ítem presenta un funcionamiento diferencial cuando la probabilidad de ser resuelto correctamente por los estudiantes que poseen el mismo nivel de habilidad varía en función de su grupo de pertenencia. Esto es observable cuando la curva de acierto de un ítem es diferente para distintas poblaciones (Moreira–Mora, 2008).

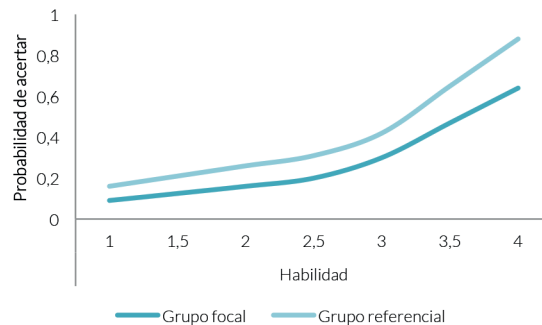
Ahora, la forma de esta curva varía dependiendo de la naturaleza del DIF. A modo de ejemplo, el Gráfico 8 muestra la curva de un ítem sin DIF. Como se observa, no hay brechas entre los grupos y a medida que aumenta la habilidad, aumenta también la probabilidad de acertar el ítem.

GRÁFICO 8. EJEMPLO DE ÍTEM SIN DIF



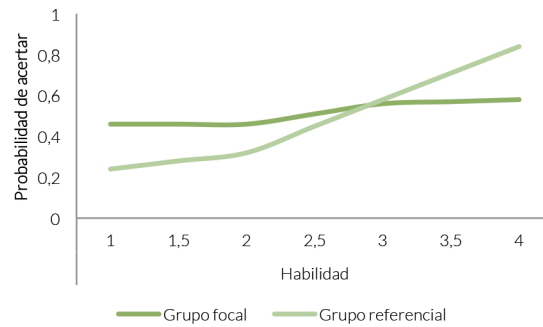
En contraposición, cuando no existe interacción entre el nivel del atributo medido y la pertenencia a un determinado grupo, se presenta funcionamiento diferencial *uniforme*. Es decir, cuando el ítem proporciona una ventaja constante para el mismo grupo, de un extremo a otro en el rango de desempeño (Penfield & Camilli, 2006). En este caso, las curvas de acierto del ítem son paralelas tal como lo muestra el Gráfico 9. En el gráfico se observa que la curva del grupo referencial se encuentra por sobre la curva del grupo focal, lo que indica que el ítem es más fácil para el primer grupo. La brecha entre ambas curvas supone que el ítem presenta DIF, ya que para una misma habilidad, la probabilidad de responder correctamente es siempre superior para el grupo de referencia.

GRÁFICO 9. EJEMPLO DE ÍTEM CON DIF UNIFORME



Por otra parte, el funcionamiento diferencial *no uniforme* se observa cuando la diferencia en las probabilidades de responder correctamente al ítem no es la misma a lo largo del *continuum* del atributo medido (Moreira–Mora, 2008) y, por consiguiente, las curvas no son totalmente paralelas. Dentro del DIF no uniforme existe una modalidad conocida como *crossing*, expuesta en el Gráfico 10, donde el grupo focal tiene ventaja en el extremo de menor habilidad, mientras que el referencial posee ventaja en el extremo de mayor habilidad.

GRÁFICO 10. EJEMPLO DE ÍTEM CON DIF NO UNIFORME



Con el fin de detectar DIF uniforme, se empleó el método de Mantel-Haenszel. Se calculan, además, estimadores que determinan si el ítem favorece al grupo focal o referencial, e indican la magnitud de las diferencias entre ellos. En lo que respecta a magnitud, se utiliza la clasificación DIF promovida por el ETS. Finalmente, se calcula el estadístico de Breslow-Day, que es efectivo cuando existen diferencias no uniformes en los niveles de habilidad de los grupos. Todos estos cálculos, que se describen a continuación, fueron complementados con análisis de Regresión Logística.

Mantel-Haenszel

La lógica que subyace al procedimiento Mantel-Haenszel (en adelante “MH”) es la siguiente: si el ítem no presenta DIF, la razón entre el número de personas que aciertan el ítem y las que lo fallan debe ser la misma en los dos grupos comparados a lo largo de todos los niveles de puntuación (Pérez Gil, 2003). De tal modo, el método MH (1959), distribuye los datos de los grupos en tantas tablas de contingencia como niveles de habilidad de los sujetos, con el propósito de comparar las probabilidades de acierto de un ítem (ver Tabla 14).

TABLA 14. FRECUENCIAS ABSOLUTAS Y MARGINALES DE GRUPOS EN EL NIVEL J

Grupos	Tipo de respuesta		
	Aciertos (1)	Errores (0)	Marginales
Grupo de Referencia (R)	A_j	B_j	n_{Rj}
Grupo Focal (F)	C_j	D_j	n_{Fj}
Marginales	n_{1j}	n_{0j}	N_j

- A_j : Es la frecuencia absoluta del grupo referencial que acierta el ítem para el nivel j .
- B_j : Es la frecuencia absoluta del grupo referencial que no acierta el ítem para el nivel j .
- C_j : Es la frecuencia absoluta del grupo focal que acierta el ítem para el nivel j .
- D_j : Es la frecuencia absoluta del grupo focal que no acierta el ítem para el nivel j .
- n_{Rj} : Cantidad de individuos del grupo referencial para el nivel j .

- n_{Fj} : Cantidad de individuos del grupo focal para el nivel j .
- n_{1j} : Cantidad de individuos que acierta el ítem para el nivel j .
- n_{0j} : Cantidad de individuos que no acierta el ítem para el nivel j .
- N_j : Número total de la muestra.

Con el fin de probar la hipótesis nula, correspondiente a la ausencia de DIF, se postula que la proporción de respuesta correcta del grupo referencial y focal es el mismo para cada nivel de habilidad j . Mientras, la hipótesis alternativa indica que al menos en un nivel es distinto y por tanto, hay presencia de DIF. Esta hipótesis nula se somete a prueba mediante el estadístico MH, asociado a una prueba de significación, que distribuye según una X^2 (Chi-cuadrado) con un grado de libertad descrito en la Ecuación 5.

Ecuación 5

$$X_{MH}^2 = \frac{(|\sum_j A_j - \sum_j E(A_j)| - 0,5)^2}{\sum_j Var(A_j)}$$

- $\sum_j A_j$: es la suma de los aciertos del grupo referencial de cada una de los niveles j .
- $\sum_j E(A_j)$: es la suma de las esperanzas matemáticas de A e igual a $\frac{n_{0j}n_{1j}}{N_j}$
- $\sum_j Var(A_j)$: es la suma de las varianzas de A e igual a $\frac{n_{0j}n_{Fj}n_{1j}n_{0j}}{N_j^2(N_j - 1)}$

Para rechazar o no la hipótesis nula, el DEMRE utiliza un nivel de significación (α) al 2,5%. Específicamente, si el estadístico de MH (X_{MH}^2) es mayor que una $X_{0,975,1}^2$ equivalente a 5,02389, se rechaza la hipótesis nula. Por consiguiente, existe evidencia estadística significativa para afirmar que el ítem analizado posee DIF.

El método MH, además, proporciona un estimador numérico que indica la dirección de las diferencias encontradas, es decir, cuál es el grupo favorecido cuando existe un funcionamiento diferencial. El estimador es el coeficiente $\hat{\alpha}_{MH}$, cuyos valores oscilan entre cero e infinito (Ecuación 6).

Ecuación 6

$$\hat{\alpha}_{MH} = \frac{\sum_j \frac{A_j D_j}{N_j}}{\sum_j \frac{B_j C_j}{N_j}}$$

Con el fin de obtener una forma más práctica de interpretación, se propone una transformación logarítmica del coeficiente $\hat{\alpha}_{MH}$ (Holland & Thayer, 1985) a una escala simétrica con origen en cero, señalada en la Ecuación 7. En esta escala, un valor negativo indica que el ítem favorece al grupo de referencia, mientras que un valor positivo, al grupo focal. De igual manera, un valor igual a cero indica ausencia de DIF.

Ecuación 7

$$\Delta_{MH} = -2,35 \ln(\hat{\alpha}_{MH})$$

Donde el estimador $\hat{\alpha}_{MH}$ es el señalado en la Ecuación 6. Complementariamente a lo señalado, el ETS (Zwick, 2012) propuso una escala jerárquica para los distintos valores del coeficiente Δ_{MH} el cual indica la magnitud de las diferencias.

- $|\Delta_{MH}| < 1$: es un ítem Categoría A, considerado con DIF *Despreciable o Irrelevante*.
- $1 \leq |\Delta_{MH}| < 1,5$: es un ítem Categoría B, considerado con DIF *Moderado*.
- $|\Delta_{MH}| \geq 1,5$: es un ítem Categoría C, considerado con DIF *Severo*.

Según estas categorizaciones, se sugieren ciertas decisiones asociadas al valor del estadístico, lo que se grafica en la Figura 6.

FIGURA 6. RANGOS DE LA MAGNITUD Δ_{MH} DE UN ÍTEM DIF



Breslow Day

La prueba de Breslow-Day (BD) se utiliza para detectar DIF no uniforme (Camilli & Shepard, 1994). Esta prueba determina si la asociación entre la respuesta del ítem y los grupos es homogénea sobre el rango de valores del total de los puntajes. Basado en una distribución Chi-cuadrado con un grado de libertad y con la misma notación usada en el método Mantel-Haenszel (Penfield R., 2003), el estadístico es el señalado en la Ecuación 8.

Ecuación 8.

$$BD = \frac{[\sum_j X_j (A_j - \alpha_j)]^2}{\sum_j X_j Var(A_j) - \frac{[\sum_j X_j Var(A_j)]^2}{Var(A_j)}}$$

- $$a_j = \frac{n_{1j} - n_{Fj} - \psi(n_{1j} + n_{Rj}) \pm \sqrt{(n_{Fj} - n_{1j} + \psi n_{1j} + \psi n_{1j})^2 + 4(1 - \psi)(n_{Rj} n_{1j} \psi)}}{2(1 - \hat{\psi})}$$
- $$\psi = \frac{a_j(n_{Fj} - n_{1j} + a_j)}{(n_{Rj} - a_j)(n_{1j} - a_j)}$$
- $$Var(A_j) = \left(\frac{1}{a_j} + \frac{1}{n_{Rj} - a_j} + \frac{1}{n_{1j} - a_j} + \frac{1}{n_{Fj} - n_{1j} + a_j} \right)^{-1}$$

Del mismo modo que para el estadístico Mantel-Haenszel, el DEMRE prueba la hipótesis nula con un nivel de significación (α) al 2,5%. Específicamente, si el estadístico BD es mayor que una Chi-cuadrado equivalente a 5,02389, se rechaza la hipótesis nula. En efecto, existe evidencia estadística significativa para afirmar que el ítem analizado posee DIF no uniforme.

Regresión logística

Además de la detección de DIF mediante MH y BD, el análisis de funcionamiento diferencial se complementa con regresión logística. La regresión logística (RL) es un modelo lineal generalizado que, a través de una función, pretende explicar y predecir los valores de una variable dependiente dicotómica, a partir de los valores conocidos de una o varias variables independientes.

Mediante esta técnica, se busca determinar si en la función matemática necesaria para predecir las respuestas dicotómicas a un ítem (correcta o errada), se debe incluir un término referido al nivel de habilidad de los sujetos independiente de su grupo de pertenencia, o al grupo de pertenencia, o a la interacción entre el grupo y la habilidad (Pérez Gil, 2003). Las fórmulas para calcular el DIF con RL se ilustran en las ecuaciones 9 y 10.

Ecuación 9.

$$p_j = \frac{f(x)}{1 + f(x)}$$

- p_j : probabilidad de responder correctamente el ítem j (entre 0 y 1, donde 1 = "responde correctamente", 0 = "responde incorrectamente").

Ecuación 10.

$$f(x) = e^{\tau_0 + \tau_1\theta + \tau_2G + \tau_3\theta G}$$

- e : exponencial
- θ : habilidad observada (puntaje en la prueba)
- G : grupo variable (0 = "referencial" o 1 = "focal")
- τ_0 : representa el peso asociado a la intersección
- $\tau_1\theta$: indica las diferencias de habilidades entre grupos de examinados, respecto de la probabilidad de acertar el ítem
- τ_2G : probabilidad de acertar respecto de la pertenencia a un grupo
- $\tau_3\theta G$: interacción entre el grupo y la habilidad. Es decir, la probabilidad de no responder correctamente no se ve influenciada únicamente por pertenecer a un grupo, sino también por diferencias de habilidad (*crossing*).

Bao *et. al.* (2009), señalan las características de los coeficientes en términos de su significancia. Si $\tau_1\theta$ es estadísticamente significativo, eso implica que los examinados de habilidad superior tienen mayor probabilidad de responder el ítem correctamente. Si τ_2G es significativo, quiere decir que la probabilidad de responder correctamente es diferente en cada uno de los dos grupos, observándose DIF uniforme. Por último, si $\tau_3\theta G$ es estadísticamente significativo, el ítem muestra amplias diferencias en el desempeño del grupo en ciertos niveles de habilidad, pudiendo observarse DIF no uniforme.

Con el fin de clasificar el DIF, se generan tres modelos sucesivos de regresión, para evaluar cada uno de los coeficientes de forma separada. Los valores de los coeficientes son los mismos de la Ecuación 10.

Ecuación 11. Modelo 1

$$f(x) = e^{\tau_0 + \tau_1 \vartheta + \tau_2 G}$$

Ecuación 12. Modelo 2

$$f(x) = e^{\tau_0 + \tau_1 \vartheta + \tau_2 G + \tau_3 \vartheta G}$$

Ecuación 13. Modelo 3

$$f(x) = e^{\tau_0 + \tau_1 \vartheta + \tau_2 G + \tau_3 \vartheta G}$$

Considerando estos modelos, Zumbo (1999) proporciona una forma para medir el tamaño del efecto DIF, denominado ΔR^2 , que es la diferencia en los valores de R-cuadrado en cada uno de los modelos. Para el análisis DIF del piloto de la PSU, interesa únicamente saber si el ítem posee DIF, y en qué magnitud, por lo que esta diferencia se realiza entre el Modelo 3 y el 1, como se expone en la Ecuación 14.

Ecuación 14.

$$\Delta R^2 = R^2(M3) - R^2(M1)$$

Tras calcular este ΔR^2 , el DEMRE utiliza el criterio propuesto por Jodoin y Gierl (2001), para categorizar el DIF, con una lógica similar a la mostrada en la Figura 8.

- $\Delta R^2 < 0,035$: es un ítem Categoría A, considerado con DIF Despreciable o Irrelevante.
- $0,035 \leq \Delta R^2 \leq 0,070$: es un ítem Categoría B, considerado con DIF Moderado. Se rechaza la hipótesis nula (no hay diferencias entre grupos).
- $\Delta R^2 > 0,070$: es un ítem Categoría C, considerado con DIF Severo e hipótesis nula rechazada.

Regla combinada de decisión

Como se mencionó, MH es conocido por su eficacia al detectar DIF uniforme, mientras que BD es útil para la identificación de DIF no uniforme. De tal forma, se combinan ambos estadísticos a un nivel de significancia de 2,5 con el resultado obtenido en RL para evaluar la presencia y el tipo de comportamiento diferencial. Es decir, si ninguno de estos estadísticos es significativo, el ítem no posee DIF.

Revisión cualitativa de ítems

Tras el análisis estadístico de los ítems piloto, estos fueron revisados cualitativamente para indagar en las causas que pudieron alterar su comportamiento en la población. El análisis cualitativo se utiliza como una estrategia para reconocer la posible influencia del contenido curricular u otros aspectos de la prueba, como el formato o el lenguaje, en las respuestas de

los examinados. De tal modo, y sobre todo tras el análisis DIF, es importante contemplar la posibilidad de realizar un estudio que incorpore la interacción cultura evaluación (Moreira–Mora, 2008).

En el DEMRE, se realiza un análisis posterior, donde se confirman las claves y se evalúan los folletos, para confirmar que no existan problemas de impresión de imágenes, figuras o textos. Durante la revisión, también se confrontan los estadísticos con el acta de certificación y ensamblaje¹¹. En estas actas, los constructores y revisores realizan comentarios respecto de las preguntas y sus características, previo a la aplicación piloto.

Por último, los ítems considerados “límitrofes”, es decir, cuyos estadísticos se encuentren al límite de lo aceptado o rechazado, se revisan en conjunto entre la UDAI y la UCP, combinando elementos disciplinares y técnicos.

Finalización del proceso

Tras la identificación de los ítems con características estadísticas óptimas para ser ensambladas, la UDAI informa a la UCP de sus resultados. Los ejes de análisis y las recomendaciones incluyen una mirada extensiva, considerando todos los parámetros mencionados anteriormente. Se incluye necesariamente una visión conjunta sobre su dificultad, discriminación, distractores y posibles sesgos.

De tal forma, la UDAI entrega a la UCP, la nómina de ítems aceptados y sus características. Esto quiere decir que estos son óptimos según los criterios estadísticos y pueden ensamblarse en una prueba definitiva.

Análisis de los resultados en la forma ancla

Como se indicó en la sección 1, en el piloto 2017, al igual que en el piloto del proceso anterior, se aplicó en cada disciplina una forma de las utilizadas en la prueba PSU Oficial de la Admisión 2016 (diciembre 2015). Esta forma, denominada “ancla”, se distribuyó aleatoriamente entre los estudiantes. El objetivo fue contrastar los resultados psicométricos de esta forma ancla con los de la aplicación oficial y evaluar así el supuesto de estabilidad de los parámetros de los ítems. Este supuesto es esencial para asumir que las preguntas piloteadas se comportarán de manera similar en la aplicación oficial.

Las preguntas comparadas fueron clasificadas de acuerdo con la diferencia en dificultad y discriminación entre ambas aplicaciones. Por una parte, si existían 10 o menos puntos porcentuales de diferencia en dificultad, la pregunta se clasificó como diferencia aceptable. Por otra parte, para comparar el parámetro de discriminación entre ambas aplicaciones se utilizó la transformación Z de Fisher (Kullback, 1959), detallada en la Ecuación 15. Dado que la discriminación es un parámetro basado en correlaciones, se utiliza esta transformación

11 Para mayor información respecto de los procesos de ensamblaje de pruebas piloto, ver página 14 del *Informe Técnico PSU. Proceso de Construcción y Ensamblaje de Pruebas*. En línea, disponible en: <http://www.portaldemre.demre.cl/estadisticas/documentos/informes/2016-vol-2-proceso-de-construccion-y-ensamblaje-de-pruebas.pdf> [Consultado el 26 de diciembre de 2016].

para realizar el test de comparación de muestras. Esto permite evaluar si la diferencia entre la correlación observada en la muestra piloto y la correlación poblacional observada en la aplicación oficial es o no estadísticamente significativa al 95% (utilizando el p-valor).

Ecuación 15

$$r' = \frac{1}{2} \log \left| \frac{1 - r_b}{1 + r_b} \right|$$

- r_b : correlación biserial

Ecuación 16

$$Z_k = \frac{(r'_{1k} - r'_{2k})}{\sqrt{\frac{1}{N_{1k} - 3} + \frac{1}{N_{2k} - 3}}}$$

- r'_{1k} : transformación de la correlación biserial del ítem k obtenida en forma ancla
- r'_{2k} : transformación de la correlación biserial del ítem k obtenida en forma oficial
- N_{1k} : número de sujetos que respondieron el ítem k en forma ancla
- N_{2k} : número de sujetos que respondieron el ítem k en forma oficial

La Tabla 15 muestra la cantidad de preguntas que cumplen los criterios descritos en el párrafo anterior. Así, por ejemplo, se observa que en el caso de matemática 73 preguntas se encuentran dentro del rango aceptable cuando se considera el criterio de dificultad, pues no tienen más de 10 puntos porcentuales de diferencia en este parámetro. En relación al parámetro de discriminación, en esta misma disciplina hay 70 preguntas sin diferencia significativa entre la aplicación piloto y la oficial.

TABLA 15. CANTIDAD DE PREGUNTAS EN RANGOS ACEPTABLES POR DISCIPLINA

	Total preguntas	Dificultad	Discriminación
Lenguaje	75	74	72
Matemática	75	38	53
Historia, Geografía y Cs. Sociales	75	63	74
Ciencias módulo común	54	54	54

En base a los resultados de este estudio comparativo se puede concluir que la prueba ancla ha variado en sus resultados. El principal problema se aprecia en Matemática, sin embargo, en todas las pruebas se aprecia que la prueba se ha tornado más fácil para los estudiantes.

Este resultado ha implicado un cambio en la estructura de anclaje de las pruebas para próximos procesos.

Bibliografía

- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.
- Adriola, W. (2003). Descripción de los principales métodos para detectar el funcionamiento diferencial del ítem (DIF) en el área de la evaluación educativa. *Revista de Pedagogía Bordón*, 177 – 189.
- Bao, H., Dayton, M., & Hendrickson, A. (2009). Differential Item Functioning Amplification and Cancellation in a Reading Test. *Practical Assessment, Research & Evaluation*, 14(19).
- DEMRE. (s.f). *Casos especiales*. Recuperado el 26 de septiembre de 2016, de Prueba de Selección Universitaria: <http://psu.demre.cl/inscripcion/casos-especiales>
- Holland, P., & Thayer, D. (1985). An alternative definition of the ETS delta scale of item difficulty. Educational Testing Service, Research Report . NJ: Princeton.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluation Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 329-349.
- Kullback, S. (1959). *Information Theory and Statistics*. New York: Dover Publications.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*.
- MINEDUC. (2013). *Bases Curriculares de la Formación Diferenciada Técnico-Profesional aprobadas por Consejo Nacional de Educación*. Recuperado el 28 de octubre de 2016, de Unidad de Currículum y Evaluación. Ministerio de Educación: http://ww2.educarchile.cl/UserFiles/P0001/File/CR_Articulos/Especial_Curriculum_2013/nuevas_bases_curriculares_tp_julio2013.pdf
- Moreira–Mora, T. (2008). El funcionamiento diferencial del ítem: un asunto de validez y equidad. *Avances en Medición*(6), 5-16.
- Moreno, R., Martínez, R., & Muñiz, J. (2005). *Análisis de los ítems*. Madrid : La Muralla.
- Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.
- Muñiz, J. (2003). *Teoría Clásica de los Test*. Madrid: Ediciones Pirámide.
- Penfield, D., & Camilli, G. (2006). Differential Item Functioning and Item Bias. En C. R. (Eds.), *Handbook of Statistics Psychometrics* (págs. 125-167). Amsterdam: Elsevier.
- Penfield, R. (2003). Applying the Breslow-Day test of trend in Odds Ratio heterogeneity to the analysis of nonuniform DIF. *The Alberta Journal of Educational Research*.

- Pérez Gil, J. A. (2003). Funcionamiento Diferencial de los Ítems (DIF). Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica. Sevilla: Dpto. de Psicología Experimental. Universidad de Sevilla.
- Sánchez, A. (2016). Comunicación Ética de Resultados de Pruebas: El Plan Nacional para la Evaluación de los Aprendizajes (Planea) de México. [Documento interno]. INEE: Instituto Nacional para la Evaluación de la Educación, México.
- Vivanco, M. (2005). *Muestreo Estadístico. Diseño y aplicaciones*. Santiago de Chile: Universitaria.
- Zumbo, B. (1999). A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa: Directorate of Human Resources Research and Evaluation Department of National Defense.
- Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. Research Report ETS RR-12-08. Princeton: Educational Testing Service.

Anexos

ANEXO 1. AFICHE Y VOLANTES INFORMATIVOS

DEMRE
UNIVERSIDAD DE CHILE

LA PREVIA CON EL DEMRE

Tu preparación oficial

5 y 6 de septiembre · 8:00 hrs

Invitamos a los 4tos Medios de tu colegio a participar en el Pilotaje de la PSU[®] que desarrollará el DEMRE en tu comuna.

La aplicación permitirá que se preparen en condiciones similares a la Prueba Oficial, con preguntas elaboradas por el DEMRE, la misma institución que construye y aplica la PSU[®].

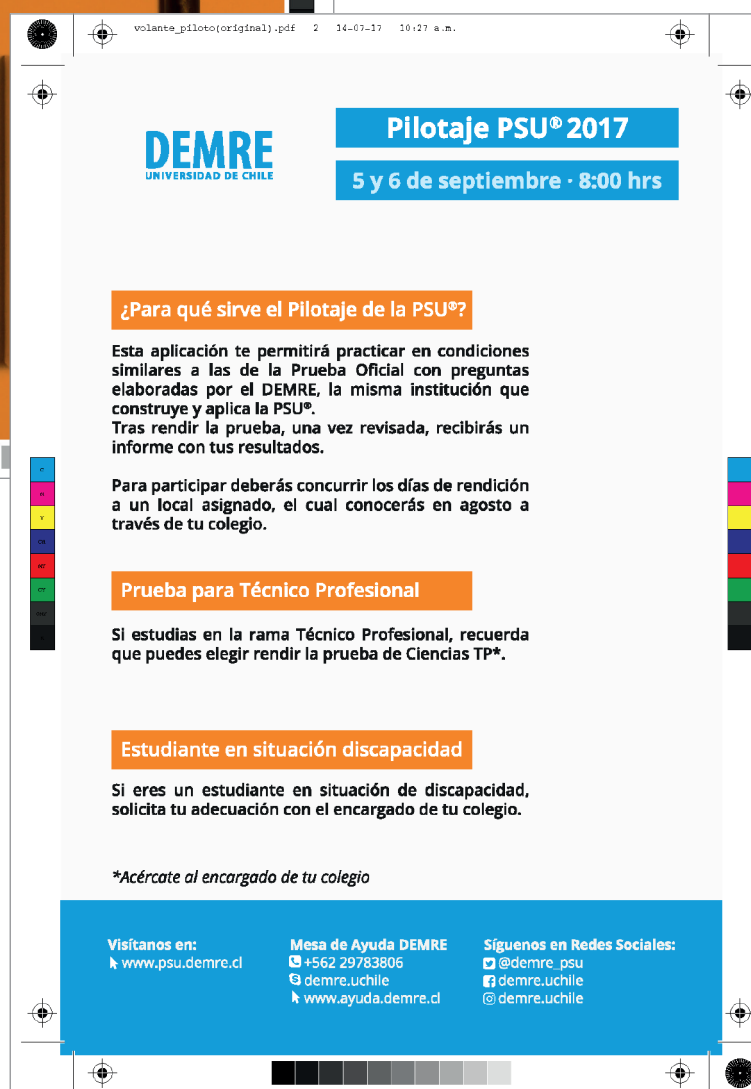
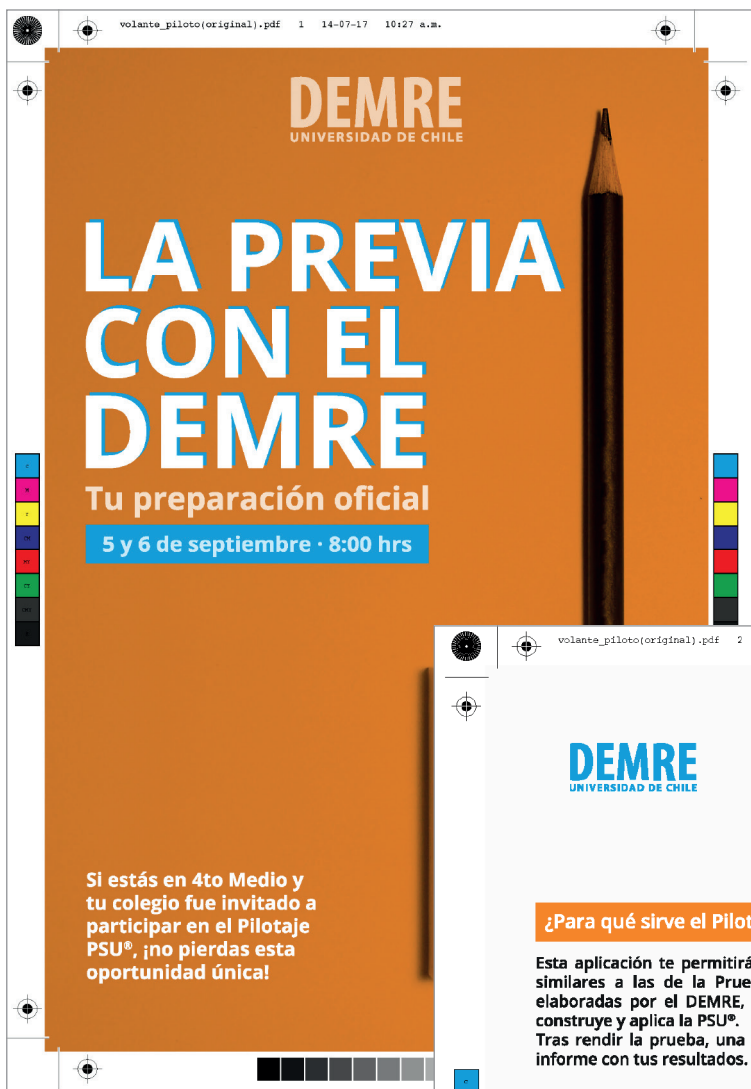
Deberán concurrir los días de aplicación al local asignado, el cual se informará en el mes de agosto.

Infórmate de todos los detalles de la aplicación consultando con el encargado de tu colegio.

Visítanos en:
www.psu.demre.cl

Mesa de Ayuda DEMRE
+562 29783806
demre.uchile
www.ayuda.demre.cl

Síguenos en Redes Sociales:
[@demre_psu](https://www.instagram.com/demre_psu)
[demre.uchile](https://www.facebook.com/demre.uchile)
[@demre.uchile](https://www.facebook.com/demre.uchile)



ANEXO 2. CARACTERÍSTICAS DE ESTABLECIMIENTOS EDUCACIONALES ELEGIDOS PARA PILOTAJE, EN REGIONES SELECCIONADAS

Región	Dependencia						Modalidad de enseñanza							
	Particular pagado		Particular subvencionado		Municipal		Humanista Científico diurno		Técnico comercial		Técnico industrial		Técnico servicios	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Antofagasta	2	3,7%	7	2,6%	12	7,7%	15	5,1%	1	1,7%	3	3,8%	2	4,0%
Coquimbo	6	11,1%	30	10,9%	16	10,3%	37	12,6%	4	6,9%	10	12,5%	1	2,0%
Valparaíso	0	0,0%	6	2,2%	12	7,7%	10	3,4%	1	1,7%	4	5,0%	3	6,0%
Metropolitana	30	55,6%	157	57,3%	72	46,2%	145	49,3%	38	65,5%	44	55,0%	32	64,0%
O'Higgins	5	9,3%	5	1,8%	3	1,9%	9	3,1%	1	1,7%	1	1,3%	2	4,0%
Maule	5	9,3%	15	5,5%	9	5,8%	13	4,4%	7	12,1%	6	7,5%	3	6,0%
Bío-Bío	4	7,4%	42	15,3%	25	16,0%	50	17,0%	5	8,6%	9	11,3%	5	10,0%
Araucanía	2	3,7%	12	4,4%	7	4,5%	15	5,1%	1	1,7%	3	3,8%	2	4,0%

ANEXO 3. DISTRIBUCIÓN DE LA PRIMERA PRUEBA EN LA APLICACIÓN PILOTO, POR COMUNA

Comunas que iniciaban actividades con la prueba de Lenguaje y Comunicación	Comunas que iniciaban actividades con la prueba de Matemática
Antofagasta	Angol
Calama	Colina
Chillan	El Bosque
Concepción	La Cisterna
Coquimbo	La Pintana
Curicó	Las Condes
La Serena	Maipú
Los Ángeles	Melipilla
Machalí	Nuñoa
Ovalle	Providencia
Quillota	Puente Alto
Quilpué	Quilicura
Rancagua	San Bernardo
San Fernando	San Miguel
Talca	Santiago
Valparaíso	Talagante
	Temuco
	Villarrica

ANEXO 4. REGISTRO DE ASISTENCIA A PILOTO PSU 2017



UNIVERSIDAD DE CHILE
VICERRECTORIA DE ASUNTOS ACADÉMICOS
Departamentos de Evaluación, Medición y Registro Educativo
DEMRE

Fecha: 07/09/2017
Página: 1 de 1

PRUEBAS DE PILOTAJE - PROCESO DE ADMISION 2018 NÓMINA DE ASISTENCIA POR UNIDAD EDUCATIVA

UNIDAD EDUCATIVA	LICEO LUIS CRUZ MARTINEZ
PRUEBAS APLICADAS	Lenguaje y Comunicación - Ciencias / Matemática - Historia, geografía y cs. sociales
FECHA / HORA	Martes 5 de Sept - 08:30 Hrs - 11:30 Hrs / Miércoles 6 de Sept - 08:30 Hrs - 11:45 Hrs

RUT	NOMBRES	LENGUAJE Y C.	MATEMÁTICA	HISTORIA	CIENCIAS
	BARRIENTOS LUCIANO CARLA SOFÍA	P	P	A	P
	CASTILLO RODRÍGUEZ YAMARA ESTEFANA	P	A	A	P
	CASTRO ROLAN RAFAEL EDUARDO	A	A	A	A
	CASTRO LUANA JORGE EDUARDO	A	A	A	A
	CHAVEZ WALDE CONSTANZA SOLEDAD	A	A	A	
	COPIETI CONTRERAS NATALIA DANIEL	A	A	A	A
	FARIAS ROLAN JUAN GUILLERMO	A	A	A	A
	HERNANDEZ SEPULVEDA ANGELA NATALIA	P	A	A	P
	HERRERA HERRERA ANIBAL DANIEL FABIAN	A	A	A	A
	INOSTROZA VAQUET LUIS IVAN	A	A		A
	LARRE WALDE CATALINA ESTER	P	P		P
	LARA ARRIAGUE RENEE MICHEL	A	A		A
	MAURER CASTRO JAVIERA ISABEL ESTEFANA	A	A	A	
	MONTENEGRO GONZALEZ NICOLAS ESTEBAN	A	A		A
	ORTIZ OLGA FERNANDO ANTONIO	A	A	A	
	PONLETE REYES JULIAN ANDRÉS	P	P		P
	VALDERRAMA PENA SILVANA ALICIA	A	A		A

NOTA: P = PRESENTE ; A = AUSENTE



Av. José Pedro Alessandri 685, Ñuñoa - Santiago, Chile

 /demre.uchile  /demre_uchile  /DEMREuchile  /demre.uchile  www.demre.cl