Prueba de Selección Universitaria

Informe Técnico

Volumen III

Pilotaje de Instrumentos PSU, aplicaciones hasta 2015

Unidad de Desarrollo y Análisis
Departamento de Evaluación, Medición y Registro Educacional—Universidad de Chile
Septiembre de 2016

CONTENIDO

GLOSARIO

BD Estadístico Breslow-Day

CDR Regla combinada de decisión, por sus siglas en inglés: Combined Decision Rule

CRUCH Consejo de Rectores de las Universidades Chilenas

DIF Funcionamiento diferencial del Item, por sus siglas en inglés: Differential Item Functioning

DEMRE Departamento de Evaluación, Medición y Registro Educacional

ETS Educational Testing Service

MH Estadístico Mantel-Haenszel

PSU Prueba de Selección Universitaria

TCT Teoría Clásica del Test

TRI Teoría de Respuesta al Item

UCP Unidad de Construcción de Pruebas del DEMRE

UDA Unidad de Desarrollo y Análisis del DEMRE

UEI Unidad de Estudios e Investigación del DEMRE, actual UDA

UI Unidad de Informática del DEMRE

PRESENTACIÓN

El pilotaje de una prueba, es el procedimiento experimental mediante el cual se obtiene la validación estadística de los ítems. El objetivo de la aplicación piloto de una prueba es determinar las propiedades psicométricas de los ítems y su adecuación para medir lo que se desea evaluar. De cumplir con los estándares establecidos, los ítems pasarán a formar parte definitiva del Banco de ítems y podrán ser utilizados en futuras pruebas oficiales.

En el caso de la Prueba de Selección Universitaria (PSU), los ítems validados en los procesos de pilotaje nunca se utilizan en la aplicación oficial del mismo año en que se aplicó el pilotaje. Estos ítems quedan disponibles para el ensamblaje de una prueba oficial a partir del Proceso de Admisión siguiente.

A su vez, el Consejo de Rectores (CRUCH) en la Sesión N°544 del 30 de mayo de 2013, acordó incluir ítems de pilotaje dentro de las PSU oficiales a partir del Proceso de Admisión 2014. Estos ítems no son considerados para la asignación de puntaje de los sujetos y pueden ser utilizados en futuras aplicaciones oficiales. Dada la complejidad en términos de la combinación de contenidos de diferentes disciplinas, desde el inicio de esta medida, se optó por no incluir ítems de pilotaje en Ciencias. Para el Proceso de Admisión 2016, las pruebas de Lenguaje y Comunicación, Matemática, e Historia, Geografía y Ciencias Sociales incluyeron 5 ítems de pilotaje en cada una de sus formas.

Consecuentemente, el presente volumen presenta las características principales del Piloto PSU realizado hasta el año 2015[1]. Se presentan los criterios para seleccionar una muestra representativa de la población que rendirá la PSU, los aspectos relevantes de la aplicación de las Pruebas Piloto, los recursos humanos y materiales involucrados, elementos operacionales y los análisis psicométricos realizados.

[1] Desde el año 2016, el DEMRE ha modificado su metodología de pilotaje de ítems PSU.

1. Muestra de la prueba Piloto

El objetivo de la aplicación piloto es validar empíricamente los ítems elaborados durante el año. Con el fin de obtener resultados fiables, esta validación debe hacerse en una muestra lo más similar posible a la población objetivo que rinde la PSU cada año.

La población objetivo está compuesta principalmente por todos aquellos individuos egresados de la educación media que desean rendir la PSU. Según datos históricos, la mayor parte de la población que rinde la PSU está compuesta por estudiantes que egresan de IV° medio el mismo año en que se aplica la prueba. A estos estudiantes se les identifica como "promoción del año" y en el Proceso de Admisión 2015 (aplicado en diciembre de 2014) fueron el 71,35%, mientras que el 28,65% provino de promociones anteriores (DEMRE, 2015).

Por razones operativas y prácticas, para las aplicaciones piloto el DEMRE ha definido la población objetivo solo considerando a los sujetos pertenecientes a la "promoción del año". Así, para los pilotajes realizados hasta el año 2015, se seleccionó una muestra representativa, considerando como **marco muestral** a los estudiantes que cursaban IV° medio durante ese año lectivo. Debido a las diferencias de comportamiento de las distintas pruebas se seleccionaba una muestra para las dos pruebas obligatorias y una muestra para cada prueba electiva. Estas muestras no son excluyentes entre sí.

Los estudiantes pertenecientes a la promoción del año fueron seleccionados aleatoriamente por el DEMRE para participar en la aplicación de estas pruebas experimentales. La participación en este tipo de instancias es de carácter voluntario. Una vez que el estudiante decide participar, debe cumplir estrictamente las normas del proceso, especialmente las referidas al resguardo de la seguridad de los ítems, pues se aplica la misma normativa que en la PSU oficial.

Con la finalidad de resguardar la representatividad de la muestra elegida se consideraron las siguientes variables:

- **Región y comuna**.
- **Dependencia del Establecimiento Educacional:** Municipal, Particular Subvencionado y Particular Pagado.
- **Modalidad educacional:** Humanístico – Científica (solo diurno), Técnico- Profesional (ramas Comercial, Industrial y Servicios).
- **Sexo:** Masculino, Femenino.
- **Rendimiento PSU:** promedio del puntaje PSU en la aplicación inmediatamente anterior según las variables descritas a continuación.

La muestra piloto se selecciona a partir de un muestreo multi etápico, con un tamaño de muestra definido según requerimientos técnicos, tomando en cuenta las variables mencionadas anteriormente.

1.1 Procedimiento general de muestreo

El procedimiento para determinar la muestra consta de las siguientes etapas

1. Definir el tamaño muestral (n) según las formas que se quieren probar. En el caso de las pruebas de Lenguaje, Matemática e Historia se determina un tamaño muestral de 2500 estudiantes por forma. En el caso de las pruebas de Ciencias, se determina un n de 2000.
2. Seleccionar la fuente de información para generar el marco muestral, ya que esta puede ser matricula, inscripción o una combinación de ambas.

   En el caso del primer proceso piloto del año, realizado en el 1° semestre, se trabaja con el total de matriculados. Para el segundo proceso piloto (2° semestre), ya se dispone de información respecto de los sujetos inscritos en el sistema de admisión.
3. Determinar las características por prueba (promedio y desviación estándar), por región, comuna, dependencia, modalidad, sexo y establecimiento para establecer estratos y calcular sus tamaños.

En una primera etapa se calcula el tamaño muestral por región mediante afijación óptima. Se decide seleccionar todas las regiones del país con el fin mejorar la representatividad de la muestra[2]. Dentro de cada región del país se seleccionan comunas al azar, acorde a criterios de inclusión basados en rendimiento. Una vez escogidas, se calcula su tamaño muestral mediante afijación óptima al interior de ellas.

Luego, se determinan los tamaños muestrales por dependencia y modalidad, con el mismo procedimiento utilizado para las regiones. Posteriormente, se seleccionan establecimientos con distinta probabilidad según su rendimiento, en cada grupo de dependencia y modalidad. Finalmente, dentro de cada establecimiento, se selecciona a los estudiantes mediante un muestreo aleatorio simple.

1.1.1 Tamaño muestral

El cálculo del tamaño muestral en las regiones, comunas, dependencia, modalidad y establecimiento se realiza mediante la afijación óptima de un muestreo estratificado.

Con este método probabilístico se estratifica por variable. De ese modo, el cálculo del tamaño de la muestra se realiza primero por región, después por comuna, dependencia y después por establecimiento. Para definir el número de sujetos por estrato (*n*) se aplica la fórmula de afijación óptima que se muestra en la Ecuación 1.

[2] No obstante, una región puede quedar fuera del proceso por razones administrativas, como la falta de locales, problemas presupuestarios, o difícil acceso.

Ecuación 1.

$$n_i = n_m \frac{S_i N_i}{\sum_{i-l}^{k} S_i N_i}$$

- n_i = Total de seleccionados en la muestra en el estrato *i.*
- S_i = Desviación estándar del puntaje en la prueba del año anterior, en el estrato *i.*
- N_i = Número de individuos que pertenecen al estrato *i* (promoción de IV° medio).
- n_m = Tamaño muestral definido en cada etapa.
- Para iniciar el proceso (primer muestreo; por región), se determina un tamaño muestral de 2500 por forma en el caso de las pruebas de Lenguaje, Matemática e Historia. En el caso de las pruebas de Ciencias, se determina un *n* de 2000.
- Cuando se obtiene el tamaño por región (n_m), se calcula nuevamente un tamaño en las comunas seleccionadas, obteniéndose un nuevo valor n_m. Este proceso se repite hasta llegar a los establecimientos y obtener el número de estudiantes a seleccionar.

En este punto, se debe cuidar que las agrupaciones realizadas se hagan bajo el criterio de rendimiento escolar equivalente, para así mantener la representatividad de la muestra. Por ejemplo, si no se eligiera la Primera Región, la cantidad de estudiantes que le corresponde se deben agregar a otra región elegida que tenga un rendimiento similar.

2. Aplicación de las pruebas piloto

La aplicación de las pruebas experimentales requiere de la coordinación de los distintos departamentos del DEMRE, en conjunto con el Equipo externo de aplicación de pruebas.

En la Región Metropolitana, se realizan reuniones en el DEMRE con los encargados (Delegados, Jefes de Local y Coordinadores Técnicos), donde se detallan las normas del proceso, las características y funciones del equipo, los documentos asociados, y las consideraciones técnicas y protocolares. Es responsabilidad de los encargados transmitir esta información a su equipo de trabajo. Los encargados o delegados, según sea el caso, recogen el material de aplicación del DEMRE, el que es trasladado a su respectivo local de rendición.

En el caso de regiones, las reuniones informativas son dirigidas por las Secretarías de Admisión[3], o el Delegado designado para el lugar. El material se traslada desde el DEMRE hacia las Secretarías, quienes contactan a los distintos delegados y Jefes de Local para coordinar la entrega y el traslado de los materiales, hacia los locales de rendición.

Al finalizar la jornada, todo el material es reintegrado en su totalidad al DEMRE, asegurando así, la seguridad y organización del proceso. En el caso de la aplicación en regiones, los delegados deberán tomar contacto con las Secretarías de Admisión para la entrega del material de aplicación y su posterior retorno al DEMRE.

2.1 Control de calidad

Para la ejecución de las pruebas piloto, se utilizan un conjuntos de documentos de aplicación, que sirven para tener en orden y controlar los detalles del proceso. Estos documentos son físicos y electrónicos y se gestionan por el DEMRE para el control de calidad del pilotaje.

Dentro de estos documentos físicos destacan: el **Acta de Aplicación**, el **Listado de Aplicación**¸ y el **Formulario de Asistencia de Postulantes por Sala.** El primero, registra los datos de la aplicación de las pruebas en cada sala, documentando todos los detalles del proceso, además de cualquier anomalía que ocurra durante la aplicación. El segundo documento es el registro oficial de los sujetos que rinden el piloto, individualizando a todos quienes participan del proceso y sus respectivas pruebas a rendir. Por último, el Formulario de Asistencia, complementa la información consignada en el Listado de Aplicación, consignando la información de rendición a nivel de Local.

[3] Las Secretarías de Admisión son representantes del DEMRE en todo el país. La nómina completa de secretarías, encargados y datos de contacto, ordenadas por región pueden encontrarse en: DEMRE. "*Secretarías de Admisión*". En línea, disponible en: http://psu.demre.cl/proceso-admision/secretarias-admision/ [Consultado el 4 de septiembre de 2016].

A su vez, se hace uso de un **Acta Electrónica**, la cual debe ser llenada por el Jefe de Local, señalando toda la información anterior, y otras de interés. Esto con el fin de agilizar la recepción de información, optimizar los controles y resultados de la aplicación.

2.2 Equipo de Aplicación de Pruebas Piloto[4]

El equipo que aplica las pruebas está compuesto preferentemente por profesores titulados, sin perjuicio del nombramiento de otros funcionarios. Para formar los equipos, se invita a los trabajadores de los establecimientos educacionales que participarán del piloto, como también a otros invitados desde el DEMRE, las Secretarías de Admisión o Universidades. Aun así, hay criterios excluyentes para formar parte del equipo, en ningún caso podrán ser parte, menores de 21 años o personas objetadas por su desempeño en procesos anteriores.

La composición del equipo en orden jerárquico, dependiendo del número de locales y salas que tenga la sede, será la siguiente (se detallan además sus principales funciones en el Anexo 1.

1. **Delegado de las Universidades Chilenas:** funcionario universitario, designado por el DEMRE, que actúa como representante directo de las Universidades Chilenas en las sedes de aplicación.
2. **Jefe de Local:** responsable del funcionamiento del Local de aplicación de pruebas piloto.
3. **Coordinador Técnico:** colaborador directo del Jefe de Local en todas las actividades del proceso de aplicación de pruebas piloto.
4. **Coordinador de Local:** apoya en sus labores al Jefe de Local y al Coordinador Técnico.
5. **Examinadores (Jefe de Sala, Coexaminador, suplente):** directamente responsables de la aplicación de las pruebas piloto en sala.
6. **Auxiliares:** encargados del aseo, orden en los Locales de aplicación. Además prestan apoyo en el traslado de materiales en el local.

Para todos los efectos penales, civiles y administrativos, el Jefe de Local, el Coordinador Técnico, el Coordinador de Local y los Examinadores son los responsables de la custodia y buen uso de los materiales de la aplicación. Los procedimientos de seguridad y resguardo de la prueba son muy importantes en este proceso, por lo que hay ciertos protocolos que deben cumplirse por parte del equipo de aplicación de pruebas, como por ejemplo: no romper los sellos de las pruebas, no reproducir por ningún medio el contenido de las pruebas, y no ingresar a la sala con bolsos, carteras, celulares o dispositivos electrónicos.

[4] Esta información fue extraída y sintetizada del *"Manual de Aplicación de Pruebas Experimentales. Proceso de Admisión 2016".*

3. Análisis de los resultados de la Prueba Piloto[5]

La Unidad de Informática (UI) proporciona los datos y los resultados de las pruebas la Unidad de Estudios e Investigación (UEI) del DEMRE (Actual Unidad de Desarrollo y Análisis –UDA-). Con esta información, se realizan los análisis psicométricos para determinar la calidad de los ítems y por tanto, la factibilidad de ser ensamblados en pruebas oficiales.

A continuación, se describen los distintos análisis psicométricos realizados a las Pruebas Piloto. Se exponen los análisis, las formulas 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 mínimos de calidad para una prueba estandarizada de altas consecuencias.

En la evaluación de calidad de los ítems se utiliza la **Teoría Clásica del Test** (TCT) que establece *modelos* capaces de evaluar las *propiedades psicométricas* de los instrumentos de medición. Específicamente, estudia 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 (Kramp, 2006; Muñiz J. , Psicometría, 1996):

1. Respecto de su comportamiento en tanto escala: refiere al estudio de la confiabilidad y validez del instrumento[6].
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 (DIF, por sus siglas en inglés, *Differential Item Functioning*).
3. Una combinación de ambas.

En el análisis de la prueba piloto del año correspondiente, así como de la prueba oficial, los criterios estadísticos considerados corresponden a la TCT. Desde el Proceso de Admisión 2017, el

[5] Toda la información de este apartado fue extraída y sintetizada de: Contreras, Hernández, & Kramp. (2011). Directrices Psicométricas para el Análisis de Ítemes PSU. Documento de Trabajo. DEMRE, Universidad de Chile.

[6] Los análisis de confiabilidad y validez se realizan a los resultados de la Prueba Oficial.

análisis de los ítems y su comportamiento en la prueba oficial, será realizado con TCT, complementado con el modelo de Teoría de Respuesta al Ítem (TRI)[7].

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.

3.1 Teoría Clásica del Test (TCT)

En cuanto a su formulación general, la TCT 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 2.

Ecuación 2.

$$X_i = V_i + e_i$$

Idealmente, en la TCT deben existir como mínimo dos formas paralelas[8] (*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 misma[9] 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.

[7] Para mayor información respecto de las diferencias y complementariedades de los modelos TCT y TRI, véase: Navas, M. J. (1994). Teoría Clásica de los Test versus Teoría de Respuesta al Ítem. *Psicológica 15*. 175-208

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

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

3.1.1 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 3.

Ecuación 3.

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

- *pi*= dificultad del ítem.
- *Ai*= 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.

Es posible transformar el valor *p* a una escala basada en puntuaciones Z[10]. La escala de transformación más difundida es la escala delta (Δ), propuesta por el Educational Testing Service (ETS) (Martínez-Arias, 1996), señalada en la Ecuación 4:

Ecuación 4.

$$\Delta = 13 + 4z$$

- *z* = el valor en la distribución normal de un porcentaje de respuestas correctas (ETS, 2010).

A diferencia de *p*, un Δ alto indica la presencia de un ítem difícil. La Tabla 1, resume los puntos de corte utilizados por el DEMRE para valorar un ítem como fácil, mediano o difícil, según los índices *p* y Δ.

[10] La puntuación Z estandariza la distribución de un conjunto de valores numéricos que describen alguna característica de una población. A su vez, permite determinar a cuantas unidades de desviaciones estándar esta un puntaje de la media poblacional.

Tabla 1. Parámetros de dificultad

p	Δ	**Clasificación**
0,00 – 0,39	25,0 – 14, 1	Difícil
0,40 - 0,59	14,0 – 12, 1	Mediano
0,60 – 1,00	12,0 – 1,0	Fácil

Para efectos de los ítems que componen la batería de pruebas PSU, el índice de dificultad utilizado por el DEMRE se encuentra en los rangos:

- Expresado en *p*: 0,10 ≤ p ≤ 0,80
- Expresado en escala delta: 9,6 ≤ Δ ≤ 18,1

3.1.2 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 (Muñiz, Fidalgo, Cueto, Martinez, & Moreno, 2005).

Según los autores, 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" (2005, 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 TCT, 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 5.

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

- $\overline{x}_c$ = promedio en la prueba del grupo que contesta correctamente el ítem.
- $\overline{x}_t$= 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 2 se muestran los puntos de corte utilizados para clasificar el índice de correlación biserial (r_b). Para efectos de los ítems utilizados en los instrumentos que componen la batería PSU, el índice de correlación biserial mínimo aceptado por el DEMRE es de r_b ≥ 0,250.

Tabla 2. Clasificación del índice de correlación biserial

rb	**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

3.1.3 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*). Eso también afecta el poder discriminativo del ítem.

Para efectos de los ítems utilizados en las pruebas que componen la batería de pruebas PSU, el DEMRE utiliza los siguientes criterios para valorar el comportamiento de los distractores:

1. Deben ser elegidos por al menos 2% o más de los sujetos que abordaron la pregunta.
2. Deben presentar un coeficiente de correlación biserial (r_b) negativo.
3. El promedio del grupo que elige el distractor debe ser menor que el promedio del grupo que contesta la clave (respuesta correcta) y menor que el promedio (de respuestas correctas) del grupo total.

3.2 Análisis Diferencial de los Ítems (DIF)

En adición al análisis del comportamiento de los ítems en los instrumentos que componen la batería PSU descrito anteriormente, en DEMRE se realiza un análisis de funcionamiento diferencial de los ítems (DIF).

El análisis DIF busca detectar posibles *sesgos,* analizando la equivalencia entre grupos comparables de individuos que rinden la prueba. Dado que un instrumento de medición no debe estar afectado por las características del objeto medido.

Desde la perspectiva TCT, se dice que un ítem funciona diferencialmente cuando los estudiantes que tienen 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

Para el caso de la PSU, en la Tabla 3 se muestran las variables y grupos analizados por el DEMRE que, basados en la realidad nacional, podrían presentar DIF.

Tabla 3. 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.

Con el fin de detectar DIF uniforme, se emplea el método de Mantel-Haenszel. Este método utiliza tablas de contingencias para estudiar diferencias entre grupos comparables a través de un estadístico. 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 entrega el estadístico de Breslow-Day, que es efectivo cuando existen diferencias no uniformes en los niveles de habilidad de los grupos.

Hasta el año 2015, los estimadores fueron calculados a través de los programas estadísticos SPSS 23 y DIFAS 5.0. A continuación se indican los criterios de análisis para determinar el funcionamiento diferencial de un ítem.

3.2.1 Mantel-Haenszel

El método de Mantel-Haenszel (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 4).

Tabla 4. 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 la 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.

Por lo tanto, la hipótesis nula, correspondiente a la ausencia de DIF, 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 son distintos y por tanto, hay presencia de DIF.

La hipótesis nula se somete a prueba mediante el estadístico Mantel-Haenszel (MH), asociado a una prueba de significación, que distribuye según una χ^2 (Chi-cuadrado) con un grado de libertad descrito en la Ecuación 6.

Ecuación 6.

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

- $\sum_j A_j =$ 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 $n_{Rj}n_{1j}/N_j$.
- $\sum_j Var(A_j)$ = es la suma de las varianzas de A e igual a $n_{Rj}n_{Fj}n_{1j}n_{0j}/N_j^2(N_j-1)$

Para rechazar o no la hipótesis nula, el DEMRE utiliza un nivel de significación (α) al 2,5%. Específicamente, si el estadístico de MH (χ^2_{MH}) es mayor que una $\chi^2_{0.975,1}$ equivalente a 5,02389, se rechaza la hipótesis nula. Por consiguiente, existe evidencia estadística significativa para afirmar que el ítem analizado posee DIF.

3.2.2 Características de un ítem DIF

El método Mantel-Haenszel, 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. Utilizando la información de la Tabla 4, el estimador es el señalado en la Ecuación 7.

Ecuación 7.

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

Los valores de $\hat{\alpha}_{MH}$ oscilan entre cero e infinito. Valores mayores que 1 indican que el ítem favorece al grupo referencial, mientras que valores menores favorecen al focal.

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 8.

Ecuación 8.

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

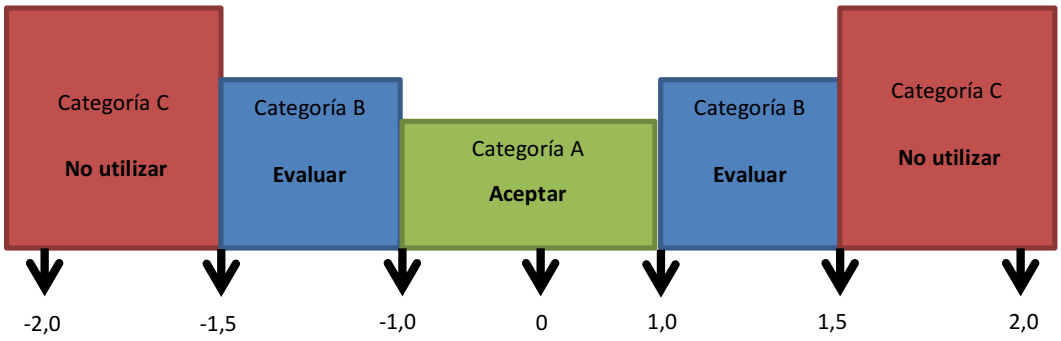
En esta escala, un valor negativo indica que el ítem favorece al grupo de referencia, mientras que un valor positivo, al grupo focal. De igual manera, un valor igual a cero indica ausencia de DIF.

Complementariamente a lo señalado, el ETS 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 estadísticos, lo que se grafica en la Figura 1.

Figura 1. Rangos de la magnitud ∆MH de un ítem DIF



3.2.3 Breslow-Day

La prueba de Breslow-Day (BD) (Penfield, 2003), 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, el estadístico es el señalado en la Ecuación 9.

Ecuación 9.

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

Donde,

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

Del mismo modo que para el estadístico Mantel-Haenszel, el DEMRE utiliza un nivel de significación (α) al 2,5%. Específicamente, si el estadístico BD es mayor que una $\chi^2_{0.975,1}$ equivalente a 5,02389, se rechaza la hipótesis nula. Es por eso que existe evidencia estadística significativa para afirmar que el ítem analizado posee DIF no uniforme.

3.2.4 Valores entregados por el programa DIFAS

El programa utilizado para el cálculo del DIF es DIFAS 5.0 y entrega los valores de los estadísticos MH y BD (ver ecuaciones 6 y 9). Además, entrega los siguientes estadísticos, presentados a continuación.

Una transformación del estadístico de Mantel-Haenszel es el MH LOR, que es la transformación logarítmica (ln) del coeficiente $\hat{\alpha}_{MH}$, el cual permite obtener una escala simétrica para el análisis.

Los valores positivos de éste indican DIF a favor del grupo de referencia, y los valores negativos al grupo focal. En cuanto a magnitud, la clasificación para el DIF del ETS es transformada a

- *Irrelevante:* si |MH LOR| ≤ 0,425
- *Moderado:* si 0,425 ≤ |MH LOR| ≤ 0,638
- *Severo:* si |MH LOR| ≥ 0,638

Otro indicador es el LOR Z, que se calcula mediante el cociente entre la estimación logarítmica de Mantel-Haenszel (MH LOR) y la estimación del error estándar. Este estimador indica que un valor superior a 2,0 o menor que -2,0 puede ser considerado evidencia de la presencia de DIF.

El MH es conocido por ser el test más potente para el DIF uniforme (Cox, 1988), pero ha sido demostrado ser relativamente ineficiente al momento de detectar DIF no uniforme, especialmente en ítems de dificultad media (Narayanan & Swaminathan, 1996). Además, simulaciones piloto indican que el poder de BD para detectar DIF no uniforme tiende a ser

relativamente alto cuando el ítem estudiado es de dificultad media (Penfield, 2003). Así, una regla combinada de decisión (CDR) basada en MH y BD podría mantener la potencia y un adecuado error tipo I a través de los niveles de dificultad de los ítems. Bajo ésta hipótesis, la CDR del programa muestra:

- *OK*: si acepta la hipótesis nula de no presencia de DIF en MH y DB. Es decir, ninguno de estos estadísticos es significativo al 2,5%.
- *Flag*: si rechaza una de las hipótesis de no DIF. Es decir, el ítem posee DIF en al menos uno de los dos estadísticos.

3.3 Finalización del Proceso

Tras la identificación de los ítems con características estadísticas óptimas para ser ensambladas, la UEI informaba a la Unidad de Construcción de Pruebas del DEMRE (UCP) de sus resultados. Los ejes de análisis y las recomendaciones incluyen una mirada extensiva, considerando los parámetros mencionados anteriormente. Es decir, el análisis de un ítem incluye necesariamente una visión conjunta sobre su dificultad, discriminación, distractores y sesgos.

En última instancia, es la UCP con sus respectivos Comités quienes deciden la inclusión o exclusión de ítems, tras las recomendaciones entregadas por la UEI. La aceptación de los ítems bajo estos parámetros, implica que éstos son óptimos según los criterios y pueden ensamblarse en una prueba definitiva.

4. Bibliografía

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks: Sage.

Contreras, P., Hernández, J., & Kramp, U. (2011). *Directrices Psicométricas para el Análisis de Ítemes PSU.* DEMRE.

Cox, D. (1988). *Analysis of binary data* . London: Methuen.

DEMRE. (2015). *Compendio Estadístico Proceso de Admisión Año Académico 2015.* Recuperado el 10 de diciembre de 2015, de Departamento de Evaluación, Medición y Registro Educacional. Universidad de Chile: http://psu.demre.cl/estadisticas/documentos/p2015/2015-compendio-estadistico.pdf

ETS. (2010). *Educational Testing Service.* Recuperado el 4 de septiembre de 2016, de Comparison of Content, Item Statistics and Test-Taker Performance for the Redesigned and Classic TOEIC Listening and Reading Tests: https://www.ets.org/Media/Research/pdf/TC-10-04.pdf

Holland, P., & Thayer, D. (1985). *An alternative definition of the ETS delta scale of item difficulty. Educational Testing Service, Research Report* . NJ: Princeton.

Kramp, U. (2006). *Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad [Effects of the number of response on the psychometric properties of personality rating scales]. Ph Doctor.* Barcelona: University of Barcelona.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*.

Martínez-Arias, R. (1996). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Moreno, R., Martínez, R., & Muñiz, J. (2005). *Análisis de los ítems.* Madrid : La Muralla.

Muñiz, J. (1996). *Psicometría.* Madrid: Editorial Universitaria.

Muñiz, J. (2003). *Teoría Clásica de los Test.* Madrid: Ediciones Pirámide.

Muñiz, J., Fidalgo, A. M., Cueto, E., Martinez, R., & Moreno, R. (2005). *Análisis de los ítems.* Madrid: La Muralla .

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 257-274.

Navas, M. J. (1994). Teoría Clásica de los Test versus Teoría de Respuesta al Ítem. *Psicológica*, 175-208.

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*.

5. Anexos

Anexo 1. Composición y principales funciones del equipo de aplicación de pruebas

Delegado de las Universidades Chilenas (solo en aplicaciones de prueba oficial)

Funcionario universitario con jornada completa o un docente de Enseñanza Media, designado por el DEMRE, que actúa como representante directo de las Universidades Chilenas en las sedes de aplicación. Es el responsable directo de la aplicación en el (los) local(es) de una Sede. Depende de la Dirección del DEMRE, durante la aplicación.

Debe coordinarse activamente con los Jefes de Local y Coordinadores Técnicos. En ese sentido, debe responder por el funcionamiento del equipo que dirige y coordina, cautelando la correcta aplicación de las pruebas y la uniformidad del procedimiento. Debe visitar periódicamente todos los locales de aplicación durante el desarrollo de las pruebas

Dentro de sus funciones, las más destacadas son:

- Conocer íntegramente las funciones, responsabilidades y atribuciones de todo el personal de aplicación.
- Retirar diariamente, desde la custodia de Carabineros, el material de aplicación correspondiente a cada día y, posteriormente, devolverlo al mismo lugar.
- Desempeñar, paralelamente, las funciones de Jefe de Local y Coordinador Técnico, en aquellas sedes con menos de 300 inscritos.
- Responder consultas de los medios de comunicación, si fuese necesario, con la información oficial que el DEMRE proporciona sobre el proceso.
- Elaborar un informe escrito sobre el desarrollo del proceso en la Sede de la que estuvo a cargo.

Jefe de Local

El Jefe de Local es la persona responsable del funcionamiento del Local de Aplicación de pruebas. Esta función puede ser ejercida por un académico o profesional de Educación Superior o, en su defecto, un docente directivo de Educación Media, con experiencia previa en los Procesos de Selección y Admisión. Los Jefes de Local son designados por los Secretarios de Admisión y/(o) por los Rectores de las Universidades participantes.

Dentro de sus funciones están:

- Conocer íntegramente las funciones, responsabilidades y atribuciones de todo el personal de aplicación.

- Nominar, con la autorización del Secretario de Admisión, a un Coordinador de Local y al 50% de los examinadores.
- Reemplazar al personal de aplicación que se ausenta en los días de aplicación de pruebas.
- Controlar la aplicación de las pruebas en cada una de las salas del local.
- Resolver los problemas de identificación o situaciones especiales que se presenten en el local.
- Preocuparse de que la señalización de servicios higiénicos y zonas de seguridad al interior del local, sea clara y esté bien destacada.

Al término de la aplicación de las pruebas, el Jefe de Local debe elaborar un informe completo y detallado, evaluando las condiciones del local de aplicación así como del personal participante. Este informe es utilizado para considerar el uso de los locales y participación del personal de aplicación en futuros procesos.

Coordinador Técnico

El Coordinador Técnico es el colaborador directo del Jefe de Local en todas las actividades técnicas del examen. Este cargo puede ser ejercido por un profesional de la educación o un funcionario universitario. La selección es realizada por el Secretario de Admisión, de acuerdo a los antecedentes en procesos anteriores.

Dentro de sus funciones están:

- Conocer íntegramente las funciones, responsabilidades y atribuciones de todo el personal de aplicación.
- Instruir a los examinadores respecto al trabajo que deben realizar en la sala.
- Recibir, custodiar, distribuir y controlar el material de aplicación de pruebas del local.
- Recibir, contabilizar y entregar al Examinador Jefe de Sala, el material de aplicación (folletos, actas, listados, hojas de respuesta y material de escritorio).
- Mantener, bajo su control, los folletos de reserva y los no utilizados.
- Reemplazar en sus funciones al Jefe de Local si fuera necesario.
- Controlar, junto al Jefe de Local, la correcta aplicación de las pruebas en cada una de las salas del local.
- Al término de la aplicación de cada prueba, debe recibir y contabilizar el material utilizado en cada sala.

Coordinador de Local

El Coordinador de Local es la persona que apoya en sus labores al delegado de las Universidades, al Jefe de Local y al Coordinador Técnico. Para ejercer esta función, la persona contratada puede

ser un profesional de la educación, un funcionario universitario o un estudiante universitario de cursos superiores.

Dentro de sus tareas están:

- Apoyar en sus funciones al Delegado de las Universidades, al Jefe de Local y al Coordinador Técnico.
- Colaborar en la señalización del Local.
- Asumir las funciones y labores que le asigne el Jefe de Local.

Examinadores (Jefe de Sala, Coexaminador, suplente)

Los Examinadores son las personas responsables de la aplicación de las pruebas a los estudiantes en cada una de las salas. Para ejercer este cargo, las personas contratadas pueden ser profesionales de la educación o estudiantes universitarios de cursos superiores.

En cada sala debe haber un mínimo de dos examinadores, uno de los cuales será Examinador Jefe de Sala y el otro, Coexaminador, de acuerdo a la designación realizada por el Jefe de Local.

Las principales obligaciones de esta función son:

- Custodiar el material de aplicación recibido.
- Verificar la identidad de los inscritos asignados a la sala.
- Entregar al estudiante el material de pruebas.
- Retirar el material a los examinados una vez concluida la prueba.
- Completar los documentos asociados a la aplicación: acta de aplicación y listado de aplicación.
- Entregar el material recibido al Coordinador Técnico.

Auxiliares

Son las personas encargadas de preparar las salas de aplicación, mantener el aseo y orden del local, transportar el material desde los vehículos hasta la sala donde se guardará, vigilar la puerta, transportar el material hasta el vehículo que lo retirará del local y dejar el local en condiciones de devolverlo una vez terminada la etapa de aplicación de pruebas.

Serán los mismos funcionarios de servicio del establecimiento, salvo que exista algún impedimento.

Deben ponerse a disposición del Jefe de Local y Coordinador Técnico en todo momento para apoyar el proceso, finalizando su labor una vez que el material esté en el vehículo de transporte y el local de aplicación quede limpio y ordenado.