Informe Pearson
Evaluación de la PSU Chile

En Junio de 2011 el Ministerio de Educación de Chile (MINEDUC) y el Consejo de Rectores de las Universidades Chilenas (CRUCH), llamaron a una licitación para evaluar la calidad de la batería completa de pruebas de la PSU. El MINEDUC y el CRUCH estaban interesados en realizar una evaluación que cubriera dos áreas principales:

- Evaluación de los procesos asociados a la construcción de instrumentos y el análisis de resultados de la PSU.
- Evaluación de la validez de los puntajes de la PSU.

En respuesta a esta licitación, el MINEDUC y el CRUCH seleccionaron a Pearson para efectuar una evaluación de la PSU que abarcara diversos aspectos de las pruebas, que van desde su construcción hasta un análisis de validez.

En el resumen ejecutivo se presenta:

- **Visión general de la PSU**: se explica su origen, su propósito, las pruebas que la componen y cómo se emplean sus resultados en el proceso de selección universitaria.
- **Descripción del propósito y estructura de la evaluación.**
- **Presentación de los hallazgos** más relevantes respecto de la evaluación de la PSU y las principales **recomendaciones** que se desprenden de la evaluación.

Visión General de la PSU

- **Origen de la PSU:** La batería de pruebas que conforman la PSU fue creada por mandato del CRUCH en 2001 y se basa en los Objetivos Fundamentales (OF) y en los Contenidos Mínimos Obligatorios (CMO) de la enseñanza media, elaborados por el MINEDUC en 1998. El MINEDUC y el CRUCH invitaron a diversas organizaciones y grupos de profesionales a formar una **comisión** para el análisis de las pruebas de selección universitaria chilenas, considerando las reformas al currículum nacional. La comisión propuso abandonar la idea de una evaluación de aptitudes, sobre la cual se habían basado las pruebas de selección universitaria chilenas en forma previa (Prueba de Aptitud Académica-PAA y la Prueba de Conocimientos Específicos PCE) y propuso una nueva configuración que reemplazara la PAA y PCE, basada en el dominio académico mediante un conjunto

de cuatro pruebas referidas al currículum de enseñanza media de Chile considerando las siguientes áreas de contenido: Lenguaje, Matemáticas, Ciencias Sociales y Ciencias.

- **Evaluación previa de la PSU:** En el año 2004, el Educational Testing Service (ETS) realizó una evaluación externa de PSU. El propósito del estudio fue evaluar la adecuación técnica de las pruebas de Lenguaje y Comunicación y de Matemática en términos de su validez y confiabilidad.
- **Estructura de la PSU:** Los marcos de evaluación de la PSU se ajustaron en la medida que se fue implementando el sistema. Dichos marcos estaban referidos a los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO), del currículum nacional chileno para la enseñanza media. La alineación de los marcos de las pruebas al currículum nacional chileno se realizó durante los primeros tres años de administración de la PSU, aumentando gradualmente su cobertura curricular. Los ajustes de los marcos de las pruebas PSU se completaron para el proceso de selección de 2007 en Matemática, Historia y Ciencias Sociales, y en las Ciencias (Biología, Física y Química); y para el proceso de selección de 2009, en Lenguaje y Comunicación.
- **Uso de la PSU:** los puntajes de la PSU se usan para calcular el puntaje ponderado considerando ponderaciones decididas previamente por las universidades respecto de las pruebas y del promedio de notas de enseñanza media (NEM) de cada postulante para cada carrera. Los requerimientos para cada carrera son informados en la publicación Series del CRUCH: Lista Preliminar de Carreras, a mediados de cada año. Los resultados de las pruebas de selección universitaria también tienen otro uso, el cual consiste en otorgar becas y créditos a los estudiantes que ingresan a la educación superior.

Estructura de la Evaluación:

- Se emplearon estándares profesionales para guiar la evaluación que fueron:
    - Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). Informaron sobre pautas en el desarrollo, validación, interpretación y uso de las pruebas.
    - International Guidelines for Test Use (International Test Commission, 2012). Informaron sobre pautas para la evaluación e interpretación de puntajes de pruebas dentro de un contexto cultural cruzado.
    - Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011). Establecieron pautas respecto de las responsabilidades de los evaluadores de Pearson y de la importancia de reconocer el propósito del programa y los intereses de las partes involucradas en nuestro trabajo evaluativo.

- Fuentes de información para la evaluación: se emplearon cuatro fuentes de información:
  - Documentación formal.
  - Entrevistas individuales.
  - Información de la PSU.
  - Paneles de opinión.
- Marco de evaluación: En el contexto de evaluación de las pruebas de la PSU, se identificaron 18 objetivos divididos en tres grupos:
  - El primer grupo de objetivos (Ver tabla del 1.1.a al 1.5) se relacionan con el desarrollo de las pruebas de la PSU o el uso de la PSU. Evaluados a la luz de (y codificados a través de) los estándares profesionales del Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999).
  - El segundo y tercer grupo de objetivos (2.1 al 2.4) examina los puntajes de las pruebas y su validez. El foco de estos estudios fueron los análisis de los datos de la PSU que pudieran conducir a nueva información respecto de las pruebas.

Para cada objetivo se definieron facetas basadas en las mejores prácticas en programas de evaluación del logro académico a gran escala.

Tabla 1: Evaluación de objetivos y número de facetas

Objetivos de evaluación	Número de facetas
1.1.a. Desarrollo de ítems	7
1.1.b. Pilotaje de ítems	4
1.1.c. Construcción de pruebas	6
1.1.d. Banco de ítems	4
1.1.e. Muestreo piloto y selección de ítems	3
1.1.f. Desempeño de ítems operacionales vs. pilotos	4
1.1.g. Fuentes de DIF exploratorias	2
1.1.h. Puntajes de pruebas estandarizados	2
1.1.i. Confiabilidad y Error Condicional de Medida (CSEM)	2
1.1.j. Recomendación de un modelo para derivar puntajes de corte	7
1.2. Analizar proceso usado para derivar un puntaje único para Ciencias	6
1.3. Evaluar métodos TRI para calibrar ítems y equiparar puntajes	7
1.4. Evaluar software para el análisis de ítems y bancos de ítems	3
1.5. Evaluar información de puntajes	3
2.1. Estructura interna de constructo	N/A
2.2. Validez de contenidos	N/A
2.3. Cambio en el desempeño del puntaje de la prueba	N/A
2.4. Predicción de resultados universitarios	N/A

Nota: N/A significa no aplicable.

Objetivos de la Evaluación

- **Los Objetivos 1.1.a. al 1.5. : relacionados con el desarrollo de las pruebas.**
  - 1.1.a. evaluó el marco y las especificaciones usadas en el proceso de elaboración de ítems, más exactamente los estándares de calidad, seguridad y confidencialidad respecto del desarrollo de ítems y pruebas; la capacitación de los redactores, redacción de ítems, revisión, ensamblaje, impresión, distribución y aplicación de pruebas.
  - 1.1.b. evaluó el pilotaje de los ítems de las pruebas; cuáles son los estándares de calidad para el pilotaje de las preguntas.
  - 1.1.c. evaluó los criterios hacia la selección de preguntas para el ensamblaje de las pruebas definitivas.
  - 1.1.d. se evaluaron objetivos relacionados con la administración de las bases de datos que almacenan los ítems PSU y sus estadísticas –el banco de ítems– Se vieron los estándares de calidad en la administración del banco de ítems.
  - 1.1.e. y 1.1.f. se evaluaron la calidad y la consistencia de los resultados basados en ítems de prueba de las pruebas piloto y empleados en las versiones operacionales de la PSU. Se observó la calidad de los términos empleados en las aplicaciones operativas, considerando los indicadores usados en su selección y considerando indicadores de funcionamiento de ítems (indicadores de la Teoría Clásica de Pruebas, Teoría de Respuesta al Ítem y análisis DIF) por género, dependencia y modalidad educacional en la muestra experimental y en la población que rinde la prueba. También el grado de consistencia entre los indicadores de funcionamiento de ítem obtenidos en la aplicación sobre la muestra experimental, respecto de aquellos obtenidos en la población que rinde la prueba.
  - 1.1.g. se evaluó el análisis del comportamiento estadístico de los ítems denominado análisis DIF, es decir, se exploraron variables asociadas al DIF, en el caso que se encontraran presentes.
  - 1.1.h se evaluaron los procedimientos empleados para puntuar las pruebas de la PSU, es decir, se hizo análisis de procedimientos para el cálculo de puntajes estandarizados, transformación de puntajes en relación a las distribuciones originales.
  - 1.1.i. se evalúan los procedimientos utilizados para estimar la confiabilidad de dichas pruebas. Se observó la confiabilidad (TCT) y precisión (TRI), incluyendo la función de información, de los diferentes instrumentos que forman parte de la batería de pruebas de la PSU. Se hizo análisis de error condicional de medida para las diferentes secciones de distribución de puntajes, poniendo especial énfasis en los puntajes de corte para la asignación de los beneficios sociales.

- 1.1.j se propone un método para establecer puntos de corte que permitan asignar beneficios sociales desde la perspectiva de la Teoría Clásica de los Tests (TCT) así como también de la Teoría de Respuesta al Ítem (TRI).
- 1.2. evaluó el uso de un puntaje único para la prueba de Ciencias, considerando que esta prueba incluye módulos electivos de ítems de Biología, Física o Química.
- 1.3. se analizó la evaluación con base en un enfoque estadístico diferente para el desarrollo de las pruebas PSU y sus puntajes, la llamada Teoría de Respuesta al Ítem (TRI).
- 1.4. se realizó un examen del software y de los procesos utilizados para el análisis estadístico y del banco de ítems de las pruebas de la PSU.
- 1.5. se refirió al proceso de informar puntajes en las pruebas de la PSU. Evaluación del proceso de entrega y de la claridad de la información en cuanto a los examinados y los diferentes usuarios del sistema de selección.

- **Los objetivos 2.1–2.4 tienen relación con la validez de la PSU.** Estos objetivos adoptaron la forma de estudios independientes. El foco de estos estudios fueron los análisis de los datos de la PSU que pudieran conducir a nueva información respecto de las pruebas. Estos estudios tomaron la forma estándar de publicaciones profesionales: una introducción, una sección de metodología, descripción de datos, resultados y discusión.

Hallazgos

- **Principales hallazgos favorables:**
  - La administración y el manejo del banco de ítems observarían criterios claros y protocolos de seguridad que dan confiabilidad a la confidencialidad de los ítems antes de su aplicación.
  - Planes de contingencia para los cuadernillos perdidos o extraviados, están dentro del nivel de seguridad requerido para este tipo de examen de gran relevancia.
  - Procedimiento utilizado para seleccionar la muestra se ajusta a criterios aceptados, teniendo en consideración los estratos que son importantes para la PSU (dependencia, modalidad curricular, etc.).
  - Los criterios establecidos para la revisión de los parámetros estadísticos de los ítems son razonables. Están alineados con lo que es visto internacionalmente; específicamente, la literatura y los manuales de software usados comúnmente en la psicométrica y evaluación de instrumentos.
  - Hay un nivel básico de conocimiento psicométrico entre los miembros del equipo y, con respecto a la selección de ítems, hay una orientación hacia los objetivos estadísticos de la Teoría Clásica de los Tests (TCT) y los estándares internacionales (niveles de aceptación de los indicadores estadísticos).

- El proceso de construcción de pruebas se basa en la capacitación de los participantes, de los profesionales del equipo del DEMRE así como de los demás miembros de la comisión; el buen nivel de capacitación de los profesionales del DEMRE apoya la seguridad de la calidad del proceso.
- Proceso de ensamblaje de la prueba contempla las cantidades de ítems establecidos para cada área de la matriz de especificaciones e incorpora ítems que cumplen con los criterios estadísticos establecidos como aceptables. Debido a que los ensambladores de las pruebas son miembros del comité que ha participado en el diseño de la prueba, su criterio informado asegura seleccionar preguntas que den respuesta a aquello que se pretende que sea evaluado.
- Existe información general y aparentemente completa sobre el banco de ítems. Cómo se organiza el banco y las interacciones entre aquellos que lo operan y el software es claramente entendido.
- En general, el DEMRE usa criterios claros en la selección de ítems, es decir, indicadores de la Teoría Clásica de Pruebas y TRI (2 parámetros). En casi todos ellos, los criterios de aceptación establecidos corresponden a rangos aceptados internacionalmente.
- La documentación del DEMRE presenta: cómo realiza estudios de DIF, cómo procesa los datos para los análisis y cómo los resultados son resumidos. La documentación también presenta criterios utilizados en la clasificación de ítems con DIF (Funcionamiento Diferencial de Ítems).

- **<u>Principales hallazgos desfavorables:</u>**
  - Proceso de ingreso de ítems en el banco podría ser mucho más detallada. La documentación no incluye una descripción detallada de los criterios para la actualización del banco de ítems.
  - No hay documentación respecto de los procedimientos de auditoría a ser implementados periódicamente con el fin de detectar posibles puntos de filtración de información. Aunque las herramientas para administrar el banco de ítems han sido consideradas adecuadas, el proceso de selección de ítems, es decir, los procedimientos para la toma de decisiones en la selección de ítems revisados en el Objetivo 1.1.b. no son adecuados.
  - Graves problemas en el uso de la "corrección por adivinación" adoptada por la PSU. Debido al rol esencial que tales puntajes cumplen en el informe de los puntajes de la PSU, el equipo internacional de evaluación considera inadecuada su utilización porque atenta contra la validez de los puntajes de la PSU y de los resultados de la administración de la prueba de la PSU en terreno.

- No está claro el propósito general del pilotaje de ítems y las expectativas psicométricas de los resultados del piloto.
    - Si el propósito de la prueba piloto es el de coleccionar datos de ítems para su posterior análisis por grupos de revisores en sesiones de revisión de datos de ítems, los procedimientos deberían establecer claramente los límites del desempeño psicométrico esperado para los ítems y la naturaleza y representación de los paneles revisores.
    - Si la expectativa es la de estimar el desempeño piloto de los ítems para informar la construcción de pruebas sin involucrar reuniones de revisión de datos, algo que es necesario para una prueba de elevada importancia tal como la PSU, hay evidencia de que este propósito no se cumple, porque los datos indican cambios drásticos en las propiedades de los ítems entre las administraciones piloto y operacionales.
- A pesar que la PSU es una prueba relativamente nueva, no existen estudios disponibles que detallen la percepción de los diferentes usuarios (directos e indirectos). Tales estudios ya podrían estar suministrando información para decidir cómo ajustar el contenido de la PSU, así como también sus aspectos formales (edición), sus condiciones de administración y la difusión y uso de los resultados de las pruebas. Más cuando existen usos adicionales de los resultados de las pruebas de la PSU que no son intencionados, tales como aquellos en que se emplea el informe SIRPAES para formular juicios acerca de la calidad de las instituciones educativas.
- Falta un manual de construcción de pruebas, como un documento técnico y orientador de instrucción para aquellos que participan de la construcción de pruebas. De esa forma, se aseguraría la estandarización en la comunicación de la aceptación y rechazo de ítems y de las pautas y sus recomendaciones de construcción.
- La prueba pone mayor énfasis sobre la modalidad Científico-Humanista del currículum de enseñanza media que sobre la modalidad Técnico-Profesional. Debería destacarse que el nivel de alineamiento de las matrices con respecto al currículum implementado que tiene lugar en las salas de clases actuales no es conocido.
- Falta un revisor final en representación de la educación de la enseñanza media que conozca de cerca la población objetivo (un profesor de este nivel educacional que no haya participado en el proceso de construcción de preguntas para asegurar su independencia y objetividad) para validar aspectos tales como claridad de las preguntas para los estudiantes y cuestionar su pertinencia con respecto al currículum cubierto en la sala de clases.

- La información del banco de ítems es presentada desde la perspectiva de la arquitectura del software del banco de ítems más que de la perspectiva psicométrica.
- Dificultades con el criterio TRI, el cual abarca un rango superior que lo aceptado comúnmente, y el criterio de nivel de omisión, que suele estar bastante elevado, aunque no es el mismo para todas las pruebas.
- Las pautas del DEMRE explícitamente permiten que los ítems pilotados sean editados o cambiados previo al uso operacional. Esta práctica contradice las mejores prácticas en el desarrollo de formas de pruebas operacionales respecto de que los ítems no deberían ser editados o cambiados, a no ser que los ítems sean piloteados nuevamente.
- Hay diferencias significativas en los indicadores de desempeño de ítems entre la administración piloto y la administración operacional por la falta de consideración de todas las variables (por ejemplo, género) en la muestra de la población para la administración piloto, o en algunos casos, la falta de participantes fuerza al plan de muestras (la modalidad, por ejemplo) a ser reconstruido.
- El efecto del sistema de asignación de puntajes por la corrección por adivinación puede inducir a diferentes estrategias entre los estudiantes al participar de las administraciones piloto y final. Los análisis llevados a cabo sobre las tasas de omisión indicaron mayores tasas de omisión en la administración operacional que en las administraciones piloto.
- El hecho que el piloto es una administración voluntaria modifica la muestra seleccionada.
- Elevadas tasas de DIF (Funcionamiento Diferencial de Ítems) manifiestan problemas significativos con las condiciones piloto; por ejemplo, autoselección, tasas de motivación diferenciales, y la representatividad de la muestra piloto. Estas condiciones pueden introducir una condición de sesgo al puntaje total de la prueba, afectando así su uso como una variable de cotejo para la comparación de grupos de referencia y de foco en un análisis DIF. Existe la decisión de enfatizar los resultados DIF operacionales sobre los resultados DIF piloto. Falta una pauta de políticas que dirija la selección de grupos de referencia y focales para los análisis DIF.
- La precisión de la escala PSU con la cual se informan los resultados no está estimada. La confiabilidad de la prueba y la medición del error estándar se estiman a partir de puntajes brutos usando la Teoría Clásica de Pruebas. La precisión de la escala de puntajes, error típico y error condicionado, no es parte de los procesos de la PSU.
- Los puntajes NEM se basan en prácticas de asignación de puntaje que pueden o no ser comparables entre instituciones educativas (Privadas,

Subvencionadas y Municipales) y las modalidades curriculares (Científico-Humanista y Técnico-Profesional), por ejemplo.

- Falta de medición de error estándar condicional respecto de los puntajes de postulación.
- Ciertos usos de puntajes requieren una mayor confianza en la exactitud de la prueba que otros usos de los puntajes de las pruebas PSU; por ejemplo, en Chile, se toman decisiones importantes con los puntajes de la prueba PSU tales como aceptar postulaciones universitarias y otorgar becas.
- El equipo evaluador considera que no es sostenible informar un puntaje único de la PSU Ciencias porque depende de un supuesto cuestionable de equivalencia (es decir, significado) de puntajes de partes de pruebas (Biología, Física y Química).
- El equipo evaluador rechaza la documentación TRI y procesos actualmente utilizados en el programa de pruebas de la PSU. La documentación revisada es equivocada. Los procesos que tienen lugar necesitan ser etiquetados adecuadamente, y los procesos que no tienen lugar necesitan ser identificados.
- Los software y hardware requieren ser optimizados tanto para el procesamiento de datos como para el trabajo con los ítems por diferentes usuarios especializados que no siempre pueden acceder a ellos por el tema de la seguridad.

- **Principales recomendaciones del equipo evaluador:**
  - Los documentos debieran declarar claramente los propósitos y usos de los puntajes de las pruebas y la naturaleza de las decisiones que se tomarán a partir de los puntajes (por ejemplo, referencia de norma/referencia de criterios), la población objetivo de examinados, la definición del dominio de interés, y los procesos destacando el desarrollo de marcos y especificaciones de las pruebas.
  - Necesidad de incluir una revisión internacional experta de las especificaciones y que los marcos teóricos y las especificaciones de las pruebas sean sometidas a validación por parte de actores ajenos a los miembros de comités, que no tengan relación con la construcción de los ítems.
  - Adelantar estudios sobre el efecto que pueda tener la decisión de alinear las pruebas a los CMO de los dos primeros grados de la enseñanza media, así como también determinar los efectos del hecho que la prueba pueda tener una ponderación mayor para la modalidad Científico-Humanista que para la modalidad Técnico-Profesional. Incluir aspectos de la modalidad Técnico-Profesional dentro de la PSU, o estudiar alternativas para una evaluación equitativa de las poblaciones formadas bajo ambas ramas curriculares.

- Incluir más documentación explícita en cuanto a la participación de redactores de ítems, capturando niveles de especialidad educacional, tiempo de experiencia docente y región de origen dentro del país.
- Considerar procesos de capacitación y certificación de redactores de ítems para asegurar la calidad de las pruebas y la validez del proceso de evaluación.
- Realizar estudios que identifiquen las características de los ítems que pueden adaptarse durante el desarrollo y confección de éstos (tales como materiales gráficos incluidos, tipos de letras, aspectos de diagramación y edición en general, entre otros), de modo que puedan facilitar el acceso a los mismos, por parte de poblaciones con discapacidades especiales. También es importante investigar si las instalaciones ofrecidas se asemejan a las de las salas de clases regulares. Adaptaciones de instalaciones, cuando no se asemejan a las condiciones del aula, pueden introducir variancia irrelevante de constructo a la prueba, en lugar de eliminarla.
- Es necesario establecer un propósito explícito y claro para el piloto.
    - Primero, repensar el proceso de pilotaje completo mediante la definición de metas y el uso y procedimientos a ser llevados a cabo de acuerdo con esta definición.
    - Luego, encontrar maneras socialmente aceptables de aumentar la motivación de los estudiantes para mostrar su mejor desempeño en las administraciones piloto.
    - Finalmente, identificar claramente la calidad de ítems esperada y obtener valores preliminares de los parámetros que sean consistentes con la administración final.
- Aunque los criterios estadísticos del plan de muestreo para la pre prueba han sido documentados, es decir, modalidad curricular y tipo de institución educativa, se recomienda una mejor articulación de las variables de estratificación.
- El proceso de pilotaje de ítems necesita estar mejor documentado con respecto a la planificación de la administración piloto y el criterio para definir los tamaños de población para cada prueba piloto.
- Mejor definición de las metas de construcción de pruebas del DEMRE; por ejemplo, un nivel de tolerancia para el error estándar condicionado de medición. El programa de la PSU debería identificar los tramos de la escala de puntajes donde se requiere de mayor precisión y construir la prueba en concordancia con este objetivo.
- Emplear ítems ancla para sus verdaderos propósitos, esto es, enlazar las formas para facilitar su calibración y equiparación. El programa de la PSU también debería revisar los criterios para la selección de ítems ancla — incluyendo el nivel de cobertura de las celdas de las matrices de especificaciones o, al menos la distribución de estos ítems a través de los ejes

temáticos de cada prueba— para alcanzar los estándares internacionales. Se recomienda actualizar las especificaciones de los conjuntos ancla del DEMRE para cumplir con los estándares internacionales.

- Proporcionar un manual para la construcción de pruebas.
- Realizar proceso de capacitación para el ensamblaje de pruebas y que sea unificado mediante la generación de pautas estandarizadas que se les enseñen a todos.
- El proceso de capacitación también debería otorgar tiempo suficiente para verificar que los nuevos desarrolladores comprendan los marcos y las especificaciones de las pruebas antes de comenzar la tarea de ensamblaje propiamente tal.
- Favorecer una transición hacia el marco TRI (Teoría de Respuesta al Item) para la construcción de pruebas. Esta transición posicionaría a las actividades de construcción de pruebas de mejor manera para alinear las pruebas de la PSU a los niveles de habilidad de los postulantes de una forma sistemática. El marco TRI también proporcionaría mayor precisión y, por lo tanto, confiabilidad en tramos de la escala de la PSU donde se toman las decisiones importantes.
- Suministrar la información faltante del banco de ítems respecto de sus módulos, su funcionalidad y características, criterios para su revisión, análisis de ítems respecto de posibles criterios de discriminación, de la opción correcta o de las opciones inválidas (distractores), o en el entendimiento que los ítems deberían funcionar de una forma en particular.
- Documentar en mayor detalle los procesos de selección de ítems en cuanto a qué pasos están involucrados en su planeamiento y cómo se aplican criterios específicos a cada prueba. Adicionalmente, en casos donde hay un cumplimiento parcial de los criterios psicométricos por parte de un ítem, se recomienda documentar la racionalidad que determina cuál indicador deberá tener prioridad frente al resto.
- Modificar el software empleado en la construcción de ítems a fin de que los ítems desarrollados puedan ser cargados en el banco con la historia de sus modificaciones y usos en la administración.
- Se recomienda enfáticamente que los ítems pilotados no sean editados o modificados en forma previa a la aplicación operacional, a no ser que los ítems sean pilotados nuevamente.
- Tomar medidas a fin de que los valores obtenidos en el piloto y la administración final sean más cercanos entre sí.
- El DEMRE debiera reconsiderar el uso de la “corrección por adivinación” en el contexto de la PSU. Tal corrección se basa en supuestos teóricos con débil respaldo y los programas de selección universitaria internacionales han abandonado su uso o están seriamente considerando retirarlo de sus procesos.

- Analizar el impacto de las tasas de no participación sobre la representación pretendida de las variables socio demográficas mayores durante el proceso de muestreo del piloto.
- Redefinir los elementos del diseño de muestra de la administración piloto, tomando en cuenta el propósito de dicha administración, el propósito de la PSU, la teoría psicométrica a ser empleada en el ítem y análisis de pruebas y en la escala de puntaje a ser utilizada.
- Evaluar el significado de los resultados DIF piloto como parte de los procesos de revisión de datos y previo a poner los ítems en el banco. Los ítems con marcadores de riesgo C deberían ser escrutados respecto de su sesgo potencial por parte de paneles revisores de datos.
- Elegir el método de DIF de Mantel-Haenszel en vez de usar múltiples métodos de DIF.
- Una vez que el DEMRE seleccione un método único para calcular el DIF, debería involucrar expertos de contenido para examinar aquellos ítems que han sido marcados por DIF.
- El programa de la PSU debería complementar la información obtenida por los análisis de detección de DIF, con la participación de expertos en contenido y educadores.
- Considerar la teoría de respuesta al ítem como un enfoque alternativo para tratar el comportamiento de adivinación del postulante.
- Revisar completamente el sistema de puntajes de postulación usando la perspectiva de puntuaciones compuestas.
- Introducir la precisión de la medición de puntajes estandarizados informados para las pruebas individuales dentro del sistema de puntaje y postulación de la PSU.
- Mantener una agenda de investigación para estudiar la estabilidad año tras año de las escalas primarias y secundarias de la PSU.
- El informe de confiabilidad de la PSU es limitado en cuanto a que no proporciona información del Error Estándar Condicionado de Medida o de la racionalidad para establecer el tamaño aceptable de tales errores respecto de los usos primarios (admisión) y otros usos de los puntajes de la PSU (becas). Se recomienda que estos análisis sean agregados, en el futuro, al programa de la PSU.
- Proponer un enfoque o modelo para definir los puntajes de corte en la escala de la PSU para otorgar beneficios sociales en la forma de becas y un enfoque que considere el manejo del dominio de la PSU. El método específico que se recomienda para fijar el puntaje de corte para fines de becas es el método Hofstee.

- Desarrollar pruebas separadas para Biología, Física y Química con propósitos específicos y con las poblaciones destinatarias en mente a fin de que los puntajes tengan sentidos no ambiguos.
- Respecto de las pruebas actuales de Ciencias de la PSU se recomienda:
    - Referirse al proceso como "enlazado" más que "equiparación."
    - Enlazar los puntajes totales, más bien que usar el proceso actual de enlazar puntajes en secciones opcionales y luego sumar los puntajes enlazados con los puntajes de la porción común.
    - Considerar el uso de métodos típicos de enlace, tales como equipercentil encadenado y equipercentil de estimación de frecuencia. Los métodos de suavizado podrían utilizarse con estos procedimientos.
- Los ítems recién desarrollados deben ser verificados en terreno y equiparados sobre la escala de la forma original. Una vez que se administren los ítems para pruebas en terreno, es necesario colocar sus parámetros de ítems sobre la misma escala de la forma original de la prueba a fin de permitir la pre equiparación durante el proceso de ensamblaje de la prueba.
- Con el fin de retener las propiedades de escala y permitir la comparabilidad de los puntajes de las pruebas entre los años de las administraciones de las pruebas, las pruebas de la PSU recién administradas deben ser equiparadas para compensar las diferencias de dificultad.
- El diseño del conjunto ancla debería cumplir con estándares internacionales. El diseño debería describir metas de cobertura de contenidos y representación psicométrica del conjunto ancla, de tal manera que el conjunto ancla pueda verse como una mini versión de la prueba total. El diseño debería describir medidas de control para los efectos de contenido y la derivación potencial de los ítems ancla.
- El equipo evaluador internacional de la PSU considera que el conjunto de herramientas de software disponible para analizar el programa de pruebas de selección universitaria de la PSU está bajo los estándares internacionales, por eso se recomienda que los procesos comunes se automaticen lo más posible y que los análisis sean estandarizados para eliminar o reducir la cantidad de intervención manual requerida.
- Deberían conservarse las estadísticas de todas las administraciones de un ítem, y los usuarios deberían ser capaces de verlos juntos en un orden cronológico.
- El DEMRE debiera proveer a los estudiantes de información interpretativa adicional explicando los Informes de Entrega de Resultados de la PSU como también, información sobre la ponderación y los puntajes de selección, que debería estar cortada a la medida para su inclusión en cada informe para estudiantes.

- La información suministrada respecto de las áreas de fortalezas y debilidades de los examinados en cada prueba de la PSU en los Informes Estadísticos de la PSU para los docentes debiera ser suspendido hasta que los resultados hayan sido escrutados cuidadosamente para asegurar la confiabilidad y validez de dicha información.
- Los Informes Estadísticos de la PSU debieran explicar cuáles son los usos que se pretenden con las pruebas de la PSU y advertir contra los usos no pretendidos.
- Se recomienda una revisión de la política de usar el Marco Curricular como la base para el desarrollo de los marcos para las pruebas de la PSU y centrarse en una habilidad o rasgo teórico subyacente (TRI, teoría de Respuesta al Item).
- El equipo evaluador recomienda inspeccionar la invariancia de las funciones de equiparación entre subpoblaciones relevantes de postulantes.

Fuente:

- "Informe Final Evaluación de la PSU Chile" 22 de enero de 2013, Pearson.

Elaboración:
Departamento de Orientación PDV
Agosto, 2016