Informe Final
Evaluación de la PSU Chile
22 de enero de 2013

Contenido:

Resumen Ejecutivo

Introducción

La evaluación internacional de un programa de admisión universitaria provee un marco de referencia para entender desde múltiples perspectivas las propiedades y usos de los resultados de las mediciones. Esta perspectiva multifacética permite a los tomadores de decisiones contar con información adicional para llevar a cabo la planificación, ejecución y control de las mejoras de la Prueba de Selección Universitaria (PSU).

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. Un área era la evaluación de los procesos asociados a la construcción de instrumentos y el análisis de resultados de la PSU. Esta área abarcaba catorce objetivos de evaluación, desde las prácticas utilizadas para el desarrollo de las pruebas de la PSU hasta los procedimientos de puntuación. La otra área cubrió la evaluación de la validez de los puntajes de la PSU; a saber, la estructura interna de la PSU, la recopilación de evidencia acerca de la validez de contenido de las pruebas, las tendencias de los resultados, y la evidencia disponible acerca de la validez predictiva de la batería respecto de los resultados universitarios.

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. Lo que sigue a continuación es un resumen ejecutivo del informe final de esta evaluación.

En la primera sección de este resumen ejecutivo se presenta una visión general de la PSU, explicando su origen, su propósito, las pruebas que la componen y cómo se emplean sus resultados en el proceso de selección universitaria. Esto es seguido por una descripción del propósito y estructura de la evaluación. Luego, se presentan los hallazgos más relevantes respecto de la evaluación de la PSU y, finalmente, se presentan las principales recomendaciones que se desprenden de la evaluación.

Visión General de la PSU

El 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. La estructura de la PSU se basó específicamente en el Marco Curricular de Enseñanza Media bajo la solicitud expresa de la *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior* (DEMRE, 2010a).

El MINEDUC y el CRUCH dieron forma a la PSU como respuesta a la necesidad de revisar los procesos de selección universitaria, a la luz de las reformas curriculares que tuvieron lugar en la década de los 90's. 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 formuló la recomendación de "llevar a cabo un diagnóstico, establecer pautas y proponer enmiendas respecto de las pruebas de selección universitaria, que habrán de permitirle

converger con las nuevas metas y objetivos educacionales planteadas por el currículum de la Reforma Educacional" (DEMRE, 2010a, p. 6).

Luego de un año de reuniones, 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). En cambio, la comisión propuso una nueva configuración que reemplazara la PAA y PCE, mediante un conjunto de cuatro pruebas referidas al currículum de enseñanza media de Chile con relación a las siguientes áreas de contenido:

- Lenguaje y Comunicación
- Matemática
- Historia y Ciencias Sociales
- Ciencias – Biología, Física y Química

El propósito de estas pruebas en el proceso de selección universitaria para las universidades del CRUCH, se refleja en la siguiente declaración:

> El propósito del proceso de selección es el de seleccionar a los candidatos que postulan a ser aceptados en una de las veinticinco instituciones que forman parte del CRUCH. El objetivo del sistema es el de seleccionar a aquellos postulantes que obtengan el mejor desempeño en la batería de pruebas que conforman la PSU, bajo el supuesto que ellos representan las mayores posibilidades de cumplir exitosamente con las tareas exigidas por la educación superior, para que ingresen de acuerdo con sus preferencias, a una de las instituciones que forman parte del CRUCH, a las carreras para las cuales postulan. Tal propósito es logrado mediante la aplicación de instrumentos de medición educacional (PSU), junto con incluir los puntajes promedio de la Enseñanza media (NEM). (MINEDUC, 2011, p. 29)

El desarrollo de la PSU también respondió a otro propósito: la necesidad de respaldar la implementación del currículum de enseñanza media a nivel nacional. Parecía evidente que una prueba de selección universitaria basada en el dominio académico que se nutre de elementos del currículum nacional chileno generaría necesidades educacionales en el sistema, tales como la adopción de dicho currículum.

En el año 2003, el Departamento de Evaluación, Medición y Registros Educacionales (DEMRE) de la Universidad de Chile (grupo responsable del desarrollo y construcción de instrumentos de evaluación y medición de las habilidades de los postulantes a la educación superior) tomó las conclusiones de esta revisión y las implementó durante la creación de la PSU, siendo estas pruebas administradas por primera vez en el proceso de selección universitaria del año 2004.

Por otro lado, ocho universidades privadas se integraron a este sistema de admisión en el año 2012. Algunas de esas universidades ya estaban solicitando como requisito a sus postulantes haber rendido la PSU, pero a partir de ese momento participaron en el proceso completo de selección de la PSU. Las ocho universidades son: U. Diego Portales, U. Mayor, U. Finis Terrae, U. Nacional Andrés Bello, U. Adolfo Ibáñez, U. de los Andes, U. del Desarrollo y la U. Alberto Hurtado.

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. Esta evaluación consideró revisiones de la documentación de la PSU y dos reuniones con personal del DEMRE en Santiago (Educational Testing Service, 2005).

1. La evaluación identificó como principal fortaleza de la PSU la calificación del personal del DEMRE y su dedicación al programa de pruebas de la PSU.

Por otro lado, la evaluación identificó varias áreas de mejoramiento para la PSU:

1. El desarrollo de documentación sobre todos los usos de las pruebas PSU, más allá de su uso respecto de decisiones de selección.
2. La validación de los usos de las pruebas de la PSU.
3. El desarrollo de planes para la equiparación de las pruebas de la PSU.
4. El desarrollo de un plan para el pilotaje de ítems que permitiese una comparación entre administraciones.
5. Adoptar la Teoría de Respuesta al Ítem (TRI) en las actividades de construcción de las pruebas de la PSU.
6. El desarrollo de un plan integral para equiparar los puntajes de las pruebas de la PSU entre años.
7. El desarrollo de investigaciones sobre las pruebas de la PSU, incluyendo la publicación de los informes respectivos.
8. El desarrollo de pautas para la interpretación de los puntajes de las pruebas de la PSU para audiencias relevantes.
9. El desarrollo de planes para la introducción del análisis DIF (Differential Item Functioning) para subgrupos relevantes.

Además, los resultados de la evaluación identificaron áreas específicas para el mejoramiento de la adecuación técnica de las pruebas de la PSU.

1. Ajustar el nivel de dificultad de las pruebas de la PSU al nivel de habilidad de los postulantes. La prueba de Matemática resultó demasiado difícil para la población de postulantes. Por otro lado, la prueba de Lenguaje y Comunicación demostró una dificultad adecuada para los postulantes.
2. Investigar el sesgo de la PSU respecto de todas las subpoblaciones relevantes y usar los resultados para el mejoramiento de las pruebas. La evaluación indicó que algunas subpoblaciones mostraban diferencias en su desempeño.
3. Expandir los análisis de confiabilidad de los puntajes de las pruebas a todas las subpoblaciones relevantes (por ejemplo, tipo de institución educativa y región) y documentar los resultados obtenidos para dichos subgrupos.

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. Los marcos de evaluación de la PSU de 2007 establecen la medición de los contenidos del Plan de Formación General que pueden ser evaluados con una prueba de papel y lápiz que incluye solamente preguntas de selección múltiple.

El DEMRE ha administrado la PSU una vez al año a todos los postulantes a las universidades del CRUCH, y más recientemente a las universidades privadas integradas al sistema. La batería de pruebas se divide en dos pruebas obligatorias ("Matemática" y "Lenguaje y Comunicación") y dos pruebas opcionales ("Ciencias" e "Historia y Ciencias Sociales"). Una de las dos pruebas opcionales debe ser rendida para postular a cada carrera universitaria.

La prueba de Ciencias consiste en una sección común que se aplica con uno de los tres módulos electivos (Biología, Física o Química). Sea cual sea el módulo que el postulante elija, se le asigna un puntaje único en la prueba de Ciencias, mediante un procedimiento que "enlaza" los puntajes del módulo común con los puntajes de un módulo electivo específico.

Todas las preguntas o ítems de la PSU son de *selección múltiple, por* lo que cada postulante debe elegir su respuesta entre cinco alternativas. Los puntajes totales de cada prueba se calculan sumando el número de respuestas correctas, restándole posteriormente un cuarto de punto por cada respuesta incorrecta. A los ítems cuyas respuestas fueron omitidas se les asigna cero punto. Este procedimiento procura ajustar el puntaje por el potencial "efecto de adivinación".

Los puntajes corregidos son transformados a una escala de puntajes con un promedio de 500 puntos y una desviación estándar de 110 puntos. El puntaje máximo de esta escala son 850 puntos y el mínimo 150 puntos. Esto permite que la escala tenga un "techo" y un "piso". El último paso del proceso es "suavizar" la distribución de puntajes para reducir las acumulaciones en el extremo superior de la distribución.

Cómo se utiliza la PSU

El siguiente paso en el proceso de selección universitaria es utilizar los puntajes de la PSU para calcular un *promedio ponderado*. El proceso de ponderación toma en cuenta los pesos ponderados decididos previamente por las universidades respecto de sus carreras, donde cada universidad/carrera le da un peso ponderado al promedio de notas de enseñanza media (NEM) de cada postulante, así como asigna pesos ponderados para cada prueba de la PSU considerada en el proceso de selección universitaria. Los requerimientos para cada carrera son informados en la publicación *Series del CRUCH: Lista Preliminar de Carreras,* a mediados de cada año. El CRUCH proporciona normas acerca de los límites inferiores y superiores de las ponderaciones para el criterio de selección, siendo las universidades, y sus carreras, los únicos responsables de asignar las ponderaciones específicas dentro del rango de ponderaciones permitidas. Los conjuntos específicos de ponderaciones puede que varíen no solamente entre carreras, sino que también dentro de una carrera en el tiempo. El

DEMRE comunica los resultados de las postulaciones a través del portal del DEMRE y a las universidades usando un proceso definido para tal propósito.

El proceso de selección es una responsabilidad compartida entre el CRUCH, las universidades afiliadas al CRUCH, el Comité Técnico Asesor de la PSU (CTA) y el DEMRE. Los roles principales del CTA, DEMRE y del *Consejo Directivo para las Pruebas de Selección y Actividades de Admisión* son los siguientes:

Consejo Directivo para las Pruebas de Selección y Actividades de Admisión (CD).

> Consejo permanente cuya función es la de velar por el desarrollo y gestión del sistema de selección y admisión, en especial de la Prueba de Selección Universitaria PSU. (Consejo Directivo, 2010)

Departamento de Evaluación, Medición y Registro Educacional (DEMRE).

> El DEMRE es el organismo técnico de la Universidad de Chile responsable del desarrollo y construcción de instrumentos de evaluación y medición de las capacidades y habilidades de los egresados de la enseñanza media; la aplicación de dichos instrumentos y la realización de una selección interuniversitaria a nivel nacional en forma objetiva, mecanizada, pública e informada. A su vez, es el organismo encargado de la administración del sistema de selección a la educación superior.
>
> Este departamento, dependiente de la Vicerrectoría de Asuntos Académicos de la Universidad de Chile, administra el proceso de selección para el ingreso a las 25 universidades que conforman el Consejo de Rectores. El DEMRE y sus predecesoras participaron en la creación y administración de la Prueba de Aptitud Académica, y desde el 2003, es el encargado de la Prueba de Selección Universitaria (PSU). (DEMRE, 2013)

Comité Técnico Asesor (CTA).

> El CTA es una agencia del Consejo de Rectores de Universidades Chilenas (CRUCH) cuya misión es la de asistir al Consejo Directivo (CD) respecto de la coordinación y supervisión de las instituciones que rigen todas las dimensiones de la selección y admisión a las universidades, y de actuar como intermediario entre el CD y los equipos técnicos responsables del desarrollo e implementación de la PSU.
>
> Las responsabilidades del CTA incluyen proponer iniciativas respecto de los aspectos de su función técnica relacionados con el CD (procesamiento de información, ponderaciones, conversiones, etc.). Antes del inicio del proceso de selección de cada año, el CTA establece la estrategia para difundir la información de utilidad para todos los postulantes y para el público en general, así como también la versión oficial de la información técnica que se distribuye en los medios masivos y páginas web asociadas (por ejemplo, por parte del CRUCH, DEMRE, Ministerio de Educación, universidades, etc.).
>
> El CTA también monitorea el desarrollo y la administración de las pruebas mediante su colaboración con el DEMRE en la resolución de los problemas técnicos relacionados con la construcción de las pruebas, su aplicación y el procesamiento de los resultados de los mismos. También el CTA conduce los estudios científicos con respecto a las características sociales y académicas de los postulantes. Finalmente, el CTA administra los procesos y las decisiones relacionadas con las técnicas de instrumentación adecuadas, el

> procesamiento y análisis de la información, así como también del proceso de selección de estudiantes a las universidades en general. (Comité Técnico Asesor, 2013)

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. En Chile, el apoyo financiero se encuentra disponible en forma de becas y créditos a los estudiantes que ingresan a las universidades y que califican para recibir tales beneficios. Hay una amplia gama de becas disponibles para los estudiantes que prosiguen hacia la educación postsecundaria. Los criterios para la asignación de becas se basan en los puntajes obtenidos por los postulantes en la PSU, así como también en otros aspectos, tales como su desempeño académico en el pasado y el nivel socioeconómico. El MINEDUC ha definido requerimientos generales para postulación a las becas. Hay más información disponible acerca de los tipos de becas en www.becasycreditos.cl.

El aporte del estado a las instituciones de educación superior se realiza mediante contribuciones fiscales indirectas. El aporte fiscal indirecto está dirigido a las instituciones de educación superior que reciben un incentivo monetario por cada alumno registrado en los 27.500 puntajes superiores de la PSU.

> El Aporte Fiscal Indirecto (AFI) es asignado anualmente por el Estado a todas las Universidades, Institutos Profesionales y Centros de Formación Técnica, reconocidos por el MINEDUC como Instituciones de Educación Superior (IES), que admitan a los 27.500 mejores puntajes de los alumnos matriculados en el primer año de estudios. (MINEDUC, 2012)

La Junta Nacional de Auxilio Escolar y Becas (JUNAEB) proporciona apoyo financiero a estudiantes calificados que se gradúan de la enseñanza media para rendir las pruebas PSU; por ejemplo,

> La Beca JUNAEB para la PSU es un subsidio destinado a financiar el costo total de rendición de la Prueba de Selección Universitaria (PSU), cifrado en $26.000 el 2012, para estudiantes de Establecimientos Educacionales Municipales y Particulares Subvencionados de la promoción del año. De manera especial, pueden postular estudiantes de Establecimientos Educacionales Particulares Pagados, que acrediten una situación socioeconómica que amerite la entrega del beneficio.(JUNAEB, 2012)

La beca JUNAEB ha permitido que un segmento de la población de egresados de la enseñanza media, que tradicionalmente no tenía acceso a rendir la PSU, tome parte en las administraciones de las pruebas de la PSU. Tal esfuerzo de otorgar acceso igualitario a rendir la PSU ha provocado un aumento cercano a un 30% de los estudiantes inscritos para rendir PSU, provenientes en su mayoría de la modalidad Técnico-Profesional.

Estructura de la Evaluación

El empleo de estándares profesionales para guiar la evaluación

El trabajo encomendado por el MINEDUC y el CRUCH fue una evaluación del sistema de pruebas de la PSU. Se utilizaron tres fuentes de estándares profesionales para la conducción de la evaluación de la PSU. La principal referencia a estándares profesionales fueron los *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Una

referencia adicional provino de las *International Guidelines for Test Use* (International Test Commission, 2012) y del *Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011).

Cada uno de estos grupos de estándares profesionales brindó una perspectiva propia de las responsabilidades de los desarrolladores de pruebas y evaluadores de programas. Los *Standards for Educational and Psychological Testing,* informaron sobre pautas en el desarrollo, validación, interpretación y uso de las pruebas. Los *International Guidelines for Test Use,* informaron sobre pautas para la evaluación e interpretación de puntajes de pruebas dentro de un contexto cultural cruzado. Finalmente, los *Program Evaluation Standards,* permitieron establecer 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. La adhesión de Pearson a los principios expuestos en el *Program Evaluation Standards* puede resultar un tanto novedosa en este contexto. No obstante, se considera importante reconocer que la PSU no es una medición aislada, sino que existe dentro de un sistema de recolección de información y de toma de decisiones que es llevado a cabo por el MINEDUC, DEMRE y el CRUCH.

Fuentes de información para la evaluación

En la evaluación se emplearon cuatro fuentes de información:

Documentación formal. El DEMRE, MINEDUC y la Contraparte Técnica suministraron la documentación formal existente acerca de los procedimientos utilizados en el desarrollo de la PSU y en la generación de sus puntajes.

Entrevistas individuales. Se entrevistó a actores clave en el DEMRE, MINEDUC y la Contraparte Técnica a fin de aclarar los procedimientos documentados y para obtener información acerca de procedimientos para los cuales la documentación oficial no fue proporcionada o no estaba disponible.

Información de la PSU. El MINEDUC dio acceso a: información de respuestas de los examinados e información de las administraciones de la PSU, y resultados psicométricos a nivel de ítem y nivel de forma de las mismas administraciones. Resulta importante destacar que aunque la entrega de datos se canalizó a través del MINEDUC, la mayoría de los archivos suministrados correspondió a bases de datos provistas por el DEMRE o las universidades.

Paneles de opinión. Se convocaron paneles que incluyeron a actores claves del sistema, el MINEDUC, el CRUCH y la Contraparte Técnica para obtener información adicional sobre el programa, incluyendo: prácticas de administración, percepciones acerca de ciertas características del programa, la facilidad de uso de los informes de puntajes, calidad de la evaluación, ecuanimidad y consecuencias, entre otros.

Durante febrero y parte de marzo de 2012, Pearson revisó la información proporcionada y las notas tomadas en las entrevistas individuales a fin de lograr una comprensión acabada del proceso de desarrollo de las pruebas llevado a cabo por el DEMRE. El propósito de esta revisión fue generar una comprensión de los métodos, pasos, herramientas, software, roles y responsabilidades de los actores involucrados, y del contexto en el cual se da este proceso. Algunas preguntas se enfocaron en clarificar los procedimientos utilizados, mientras que otras apuntaban a los pasos de procesos esperados que no fueron descritos en la documentación.

Adicionalmente, Pearson presentó una lista de descripciones de roles al MINEDUC, la cual el MINEDUC luego utilizó como base para generar una lista de posibles participantes en entrevistas. Pearson invitó a estos participantes a reuniones que tendrían lugar en Santiago entre el 27 y 29 de marzo de 2012. Los participantes eran grupos de individuos que tomarían parte en entrevistas respecto del proceso del DEMRE para la comunicación de los resultados de selección y validez de contenido de la PSU.

El marco de evaluación

Una revisión exhaustiva de cualquier programa comienza con la identificación de los objetivos 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 se relacionaba con el desarrollo de las pruebas de la PSU, mientras que el segundo y tercer grupo de objetivos (cuatro objetivos cada uno) examinaba los puntajes de las pruebas y su validez, respectivamente. Cada objetivo corresponde a un requerimiento de evaluación establecido en la licitación de la evaluación de la PSU.

Para cada objetivo se definieron facetas basadas en las mejores prácticas en programas de evaluación del logro académico a gran escala. En la medida que las facetas evaluativas se definían, se precisaban los puntos de importancia mediante una serie de reuniones con actores relevantes, tales como la Contraparte Técnica y el DEMRE. Los principales elementos de cada objetivo y faceta fueron codificados para los estándares profesionales a los que se referían usando los *Standards for Educational and Psychological Testing*(AERA, APA, & NCME, 1999).

Los objetivos de evaluación fueron presentados a la Contraparte Técnica y discutidos durante la primera visita del equipo de trabajo a Chile en enero de 2012. El proceso de aclaración de metas duró tres días y proporcionó un medio para que los participantes se familiarizaran con un conjunto inicial de objetivos y facetas, para proponer modificaciones y, en algunos casos, para agregar nuevos elementos definiendo las facetas de evaluación. Se indica un resumen de los objetivos de evaluación en la Tabla 1, para su posterior discusión en la siguiente sección.

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.

Los objetivos que tienen que ver con el desarrollo de las pruebas de la PSU o el uso de la PSU (Objetivos 1.1.a–1.5) requirieron una evaluación faceta por faceta. Respecto de estos objetivos, los procesos de la PSU representados por cada faceta fueron evaluados a la luz de (y codificados a través de) los estándares profesionales del *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Se proporcionó tanto una declaración evaluativa general como una calificación para los elementos de cada faceta, y se hicieron recomendaciones acerca de cómo mejorar procesos específicos de la PSU.

Los objetivos en relación con la validez de la PSU (Objetivos 2.1–2.4) 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.

Visión general de los hallazgos clave y de las recomendaciones por objetivo de la evaluación de las pruebas de la PSU

En las dos secciones siguientes se resumen los hallazgos y las recomendaciones de la evaluación de la PSU. La primera sección consiste en una evaluación de la construcción de las pruebas PSU según los estándares referidos a los Objetivos 1.1.a al 1.5. En la segunda sección se revisan los principales hallazgos de los estudios de validez desarrollados en los Objetivos 2.1 al 2.4.

Dentro de cada sección, se presentan los hallazgos y se entregan recomendaciones objetivo por objetivo. Las recomendaciones en negrita son consideradas particularmente relevantes por el equipo evaluador.

Antes de comenzar con la descripción de los hallazgos y recomendaciones, es importante señalar una recomendación general de mucha importancia: la necesidad de una documentación más extensa y clara asociada a la construcción, aplicación y análisis de las pruebas de la PSU. *Esta necesidad de documentación más detallada es vital, dada su importancia con respecto al mejoramiento del sistema de selección de la PSU.* Esta recomendación se mencionará específicamente en varios de los siguientes objetivos.

Objetivos de evaluación 1.1.a–1.5: La construcción y usos de la PSU

Estos objetivos examinaron el marco y las especificaciones usadas en el proceso de elaboración de ítems (Objetivo 1.1.a), el pilotaje de los ítems de las pruebas (Objetivo 1.1.b), y el criterio para la selección de ítems para las distintas formas de las pruebas operacionales (Objetivo 1.1.c). También incluyeron objetivos relacionados con la administración de las bases de datos que almacenan los ítems PSU y sus estadísticas –el *banco de ítems*– (Objetivo 1.1.d), 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 (Objetivos 1.1.e–1.1.f), y el análisis del comportamiento estadístico de los ítems denominado análisis DIF (Objetivo 1.1.g).

Los dos objetivos siguientes examinan los procedimientos empleados para puntuar las pruebas de la PSU (Objetivo 1.1.h) y evalúan los procedimientos utilizados para estimar la confiabilidad de dichas pruebas (Objetivo 1.1.i). Finalmente se recomienda un método para establecer puntos de corte que permitan asignar beneficios sociales, tales como becas (Objetivo 1.1.j).

En el Objetivo 1.2 se analizó 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. 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), fue el tema del Objetivo 1.3. Un examen del software y de los procesos utilizados para el análisis estadístico y del banco de ítems de las pruebas de la PSU fue el foco del Objetivo 1.4. Finalmente, el Objetivo 1.5 se refirió al proceso de informar puntajes en las pruebas de la PSU.

Objetivo 1.1.a: Estándares de calidad, seguridad y confidencialidad respecto del desarrollo de ítems y pruebas: capacitación de los redactores, redacción de ítems, revisión, ensamblaje, impresión, distribución y aplicación de pruebas

Descripción:

Este objetivo incluyó una revisión de varios aspectos del desarrollo de ítems y procesos de administración usados para la PSU, incluyendo:

- El proceso para desarrollar marcos y especificaciones para guiar la redacción de los ítems (Faceta 1).
- El proceso para seleccionar y capacitar a los redactores de ítems para la PSU (Faceta 2).
- El proceso para la asignación de la redacción de ítems para la PSU (Faceta 3).
- El proceso para la revisión y aceptación de los ítems en borrador de la PSU (Faceta 4).
- Las herramientas de autoría de ítems y banco de ítems (Faceta 5).
- El proceso para la distribución y la administración de las pruebas de la PSU (Faceta 6).
- Los procesos de captura de datos y puntuación de la PSU (Faceta 7).

Para este objetivo, el equipo evaluador verificó los principales participantes y documentó los procedimientos empleados con un foco en las calificaciones de los participantes, la posibilidad de revisión de las especificaciones y de los ítems desarrollados por otro grupo de expertos, la calidad de la seguridad y respaldos del sistema sobre el cual los ítems y la prueba fueron construidos, así como también la calidad y la documentación para los procedimientos de distribución y retorno de materiales.

Método:

La información para este objetivo se obtuvo a partir de entrevistas realizadas a funcionarios del DEMRE, entre el 19 y 20 de marzo de 2012, incluyendo a su director, al jefe del departamento de construcción de pruebas, a los coordinadores de los comités de construcción de pruebas de las cuatro áreas temáticas, al jefe de la unidad de investigación y su equipo, así como también al jefe de las unidades logísticas, de los registros educacionales, la unidad computacional, del proceso de selección y al coordinador general. También se consultó la documentación formal disponible del DEMRE en la preparación de la evaluación de este objetivo.

Hallazgos:

Los equipos técnicos a cargo de redactar las especificaciones tienen una capacitación académica y experiencia profesional acorde a su participación en el desarrollo de las especificaciones de la prueba. Los marcos y las especificaciones de las pruebas son el producto de un análisis curricular que también ha contemplado las demandas de la educación superior, a fin de plantear una propuesta de evaluación objetiva y seria. Sin embargo, ni en la construcción del marco ni de las especificaciones se encuentra evidencia de la participación de expertos externos al DEMRE en la revisión y análisis de la pertinencia y relevancia de estas especificaciones, para los fines de la prueba y para el interés de la población evaluada.

Aun cuando los documentos de los marcos teóricos y las especificaciones para cada prueba existen, la profundidad y el detalle de la cantidad de información no se encuentran estandarizadas entre las distintas pruebas y sus comités.

El equipo técnico, a cargo de coordinar la selección y evaluación de los redactores de ítems, cuenta con la información técnica y académica adecuada para llevar a cabo estas labores. El proceso de selección se caracteriza por ser abierto, lo cual contribuye a la transparencia del proceso y a la diversidad de los equipos de redactores.

Al parecer se le ha otorgado un período de tiempo adecuado al proceso de redacción de ítems, lo cual es positivo para el proceso porque asegura que se pueda dedicar una cantidad de tiempo adecuada al análisis y ajuste de cada ítem. Los desarrolladores de ítems los redactan desde las instituciones donde se desempeñan, lo cual permite su disponibilidad al desarrollarlos y asistir a reuniones de revisión de ítems. El número de ítems asignado a cada redactor es razonable. Sin embargo, el proceso de revisión podría mejorarse si el coordinador de cada comisión llevase a cabo una revisión previa de ítems, antes de que sean llevados a las reuniones de revisión conjuntas, haciendo uso de listas de cotejo estandarizadas para garantizar el cumplimiento de los criterios básicos de calidad de los ítems. Esta revisión facilitaría la retroalimentación a los redactores en la identificación de aspectos de los ítems que no cumplen con el estándar establecido y optimizaría el aprovechamiento del tiempo en las reuniones de revisión.

La plantilla utilizada para la redacción de ítems tiene información clave para la caracterización del ítem elaborado, aunque también podría incluir información más precisa sobre, por ejemplo, la persona de la comisión que está a cargo de llevar a cabo la primera revisión, etc. Más tarde esta información podría ser incorporada al banco de ítems y facilitaría los subsiguientes procesos de selección para ensamblaje de pruebas y auditorías de calidad.

Se reconoce la existencia de un procedimiento sistemático para la revisión y aprobación de los ítems. Sin embargo, aunque las observaciones de los revisores están documentadas, no hay evidencia de documentación que incluya los estándares de calidad de los ítems para cada prueba. De hecho, las escalas de calificación para la aprobación o el rechazo de ítems que se utilizan en las diferentes pruebas no son las mismas, lo cual es evidencia que se están empleando diferentes criterios para la valoración de la calidad de los ítems.

En general, la administración y el manejo del banco de ítems parecen obedecer criterios claros y protocolos de seguridad que dan confiabilidad a la confidencialidad de los ítems antes de su aplicación. Sin embargo, la documentación que describe el proceso de ingreso de ítems en el banco podría ser mucho más detallada. También es evidente que la documentación no incluye una descripción detallada de los criterios para la actualización del banco de ítems.

Respecto de los protocolos de seguridad descritos en la documentación, está claro que el acceso restringido al espacio físico del banco, junto con las restricciones de perfil en cuanto al acceso al sistema, ha jugado un rol importante para mantener la confidencialidad de los ítems. Sin embargo, una vez más, este positivo hallazgo no significa que no haya cabida para mejoras; por ejemplo, no hay documentación respecto de los procedimientos de auditoría a ser implementados periódicamente con el fin de detectar posible puntos de filtración de información. Más aún, 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.

Se reconoce que se han establecido protocolos de seguridad adecuados para el manejo de los materiales de los exámenes durante el proceso de impresión, y que los mismos pueden ser garantes de la no divulgación, total o parcial, del material de prueba antes de la administración operacional. Sin embargo, la documentación no brinda evidencia de la existencia de protocolos para el control del material, en su distribución a los lugares de administración de pruebas.

Con respecto a los planes de contingencia para los cuadernillos perdidos o extraviados, creemos que este proceso se encuentra dentro del nivel de seguridad requerido para este tipo de examen de gran relevancia. Este proceso es comparable con aquellos exámenes de gran relevancia a nivel internacional; por ejemplo, en los Estados Unidos, el sistema de entrega de los cuadernillos de las pruebas utiliza un número único de identificación para cada cuadernillo para propósitos de seguimiento. Este proceso de numeración permite el rastreo de los cuadernillos de pruebas despachados y devueltos.

Los evaluadores reconocen 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. Para una evaluación más detallada de este procedimiento, ver la discusión en relación al Objetivo 1.1.h.

Recomendaciones:

1. **A fin de proveer un diseño de prueba más riguroso, recomendamos que los documentos declaren 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.** Debido a que a menudo se refiere a la PSU como una batería, es pertinente esperar un documento de especificaciones consolidado para todas las pruebas que incluye. El comité internacional de evaluación desearía enfatizar la importancia del desarrollo de tal documentación con un usuario-destinatario en mente.
2. En relación al rigor que la definición y explicación del dominio de la prueba debe tener, la recomendación es la de adelantar estudios sobre el efecto que pueda tener la decisión de alinear las pruebas a los CMO de los primeros 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. Sería deseable fomentar la materialización de las políticas referidas durante las entrevistas con el equipo técnico en el sentido de 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.
3. **De acuerdo con la necesidad de incluir una revisión internacional experta de las especificaciones, se recomienda 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.** Esta independencia permitiría la retroalimentación en cuanto a aspectos tales como la adecuación de la cobertura de contenidos (ejes), así

como también la decisión en cuanto a la inclusión o no inclusión de los CMO en el mismo. Parte del proceso de validación del marco debería llevarse a cabo con la participación de expertos y con la documentación pertinente. La descripción de la experiencia de los revisores es importante de recolectar e informar dentro de la documentación de marcos de prueba de la PSU. Esta documentación debería cubrir descripciones y ejemplos de materiales evaluados, con particular énfasis sobre el dominio de la prueba, uso de la prueba, y poblaciones destinatarias de examinados. Finalmente, recomendamos encarecidamente desarrollar, administrar y resumir las respuestas de los participantes a una encuesta luego del ejercicio de evaluación y emplear sus recomendaciones para el mejoramiento del proceso de evaluación en los años siguientes.

4. Recomendamos 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. El uso de tecnologías de la información podría ser de utilidad para este fin, permitiendo la participación a distancia de redactores de ítems para quienes desplazarse a la ciudad de Santiago es difícil.

5. **Deberían tener lugar procesos de capacitación y certificación de redactores de ítems para asegurar la calidad de las pruebas y la validez del proceso de evaluación.** Los participantes deberían responder a criterios desafiantes que formalizan el proceso, para que todos los redactores de ítems se involucren en el trabajo bajo condiciones similares. Deberían tener un dominio básico de la disciplina en la cual habrán de construir preguntas; un conocimiento básico de los propósitos, del marco teórico y de las especificaciones de las pruebas; y finalmente, el conocimiento técnico básico del proceso de construcción de ítems para pruebas. Cada comité asigna una cantidad diferente de tiempo al proceso de capacitación y a los tiempos de trabajo y el énfasis en los temas tratados en las reuniones de inducción para los redactores de ítems no se encuentran estandarizados. A fin de asegurar un nivel de competencia adecuado para un redactor de ítems, es necesario estandarizar el proceso de capacitación con respecto a los objetivos, tópicos, tiempos, materiales y el resto de los recursos, así como también mecanismos para el control de los procesos y evaluación. Es necesario darle a cada redactor la oportunidad de desarrollar sus habilidades de redacción y recibir retroalimentación oportuna sobre los ítems antes de que ese redactor empiece a redactar ítems para el piloto. El comité de Lenguaje y Comunicación sigue el proceso de desarrollo de ítems más formalizado. Valdría la pena generalizar este proceso a las demás áreas de contenidos a fin de asegurar una capacitación uniforme de los miembros de la comisión.

6. También recomendamos la implementación de un sistema de certificación para redactores de ítems que complete el proceso de capacitación formal con el propósito de desarrollar una base de redactores de ítems certificados. Esto podría llevarse a cabo directamente por el DEMRE, o por una institución que pueda ofrecer capacitación a los redactores. En cualquier caso, el proceso de capacitación debería estar diseñado para incluir:

    - Verificación de que los redactores de ítems tengan dominio de la disciplina, enfatizando el dominio de los contenidos evaluados.
    - Capacitación teórica en los aspectos técnicos del diseño de la prueba, tales como matrices de especificaciones, lineamiento para la construcción de ítems, escalas, e información de resultados y conceptos de medición tales como validez y confiabilidad.

- Capacitación en la redacción de ítems mediante talleres con retroalimentación detallada respecto de los errores cometidos individualmente en la construcción.
- Capacitación en análisis de ítems, incluyendo aspectos conceptuales de indicadores psicométricos y ejercicios de interpretación práctica de los mismos, enfatizando la relación entre las características de la construcción de ítems y los indicadores obtenidos.
- Capacitación con respecto a los propósitos particulares de la PSU, su trasfondo y usos.

La capacitación deberá incluir evaluaciones de los participantes para verificar un nivel mínimo de comprensión de los temas así como también de la calidad de los ítems producidos.

El propósito final será el de otorgar a cada participante, status como redactor certificado de ítems. También, dependiendo de los cambios introducidos en la prueba o en sus procedimientos de desarrollo, debería contemplarse procesos de actualización periódica para los redactores de ítems. Considerando la cantidad y las características de los temas recomendados para la certificación de redactores de ítems, podríamos estimar que un curso completo de capacitación de redactores de ítems, tal como el descrito anteriormente, tomaría por lo menos entre 30 y 40 horas.

7. Recomendamos llevar a cabo chequeos sistemáticos sobre los supuestos de que los redactores de ítems están cumpliendo con los principios de confidencialidad y derechos de autor. Estos chequeos se llevarían a cabo con personal autorizado del DEMRE, quienes verificarían al azar los contenidos y el estado del arte de los ítems de las pruebas respecto de las fuentes de materiales protegidos por los derechos de autor.
8. Según lo establecido en el Estándar 3.7, "los procedimientos empleados para desarrollar, revisar y ensayar ítems, y para la selección de ítems, a partir del banco de ítems, deberían estar documentados". Recomendamos documentación clara y transparente del proceso para consultar el banco de ítems respecto de la identificación de la cantidad de ítems a ser comisionados. En esta misma línea, nos gustaría recomendar que se especifique con mayor precisión cómo son seleccionados los redactores de ítems.
9. **Recomendamos 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.** Aunque el procedimiento establecido por la UCP para aumentar el tamaño del tipo de letra y la gráfica para las personas con visión limitada es relevante y representa un elemento importante en asegurar la equidad en el proceso de evaluación, aún hay una necesidad importante de explorar mecanismos alternativos, en la construcción de pruebas y en el proceso de implementación para asegurar mayores condiciones de equidad para todos los postulantes; por ejemplo, ¿cuáles son las condiciones físicas locales para la administración de la prueba o la distancia que debe recorrer la gente con discapacidad hasta los lugares donde se realiza la prueba? **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.**

10. En general, las herramientas utilizadas para la autoría de ítems son las apropiadas para la tarea requerida. Ellas proporcionan los medios y el ambiente seguro requerido para este tipo de desarrollo de pruebas de alto impacto. Recomendamos continuar con esta práctica de utilización.
11. Recomendamos aumentar la eficiencia del proceso de revisión de los ítems borradores iniciales mediante el involucramiento del personal de contenidos de mayor antigüedad del DEMRE con anterioridad a la revisión completa del comité. El propósito de esta revisión por parte de los expertos es el de verificar el cumplimiento con las especificaciones de los ítems, es decir, con la relevancia de contenidos de los ítems, lo apropiado de los ítems para las diferentes poblaciones y la aplicación de especificaciones editoriales.
12. Recomendamos que el DEMRE formalice el proceso de retroalimentación a los redactores de ítems de forma clara y objetiva. Dicha documentación podría proporcionar información para analizar problemas comunes durante la redacción de ítems y, por lo tanto, orientar la capacitación futura de los redactores y revisores de ítems.
13. Recomendamos que los procesos para el desarrollo, revisión, pilotaje y selección de ítems del banco de ítems sea documentado.
14. Recomendamos la utilización de un panel de revisores de ítems independiente del panel de desarrolladores de ítems. Este panel de revisores debería estar compuesto por un grupo de redactores de ítems calificados que no hayan participado en el desarrollo de los ítems en revisión.
15. El banco de ítems resume las características esenciales de los ítems dentro de un ambiente seguro. Sin embargo, dado el permanente progreso tecnológico en materias de administración de información, sería de utilidad implementar periódicamente sistemas de auditoría internos o externos para los procesos de control de la seguridad del banco, a fin de identificar posibles puntos vulnerables con respecto a la seguridad del ítem, así como también aumentar la eficiencia en los procesos de almacenamiento, consultas y atomización de los procesos de ensamblaje.
16. Recomendamos documentación para los planes de empaquetado y distribución, tanto para pilotajes como para formatos operacionales.
17. Recomendamos protocolos de control de calidad más claros: verificación de identidad, control de copiado, y el manejo de eventos fortuitos (crisis, enfermedad, mal tiempo, etc.).
18. Recomendamos catalogar las desviaciones del proceso de administración respecto del estándar, a fin de que el DEMRE cuente con dicha información para sus decisiones. Recomendamos emplear tal catálogo para evaluar el proceso de administración de pruebas y proporcionar capacitación al personal que participa de tales procesos. El desarrollo profesional puede permitir que los participantes del proceso de administración (por ejemplo, jefes de locales o delegados) usen su experiencia como administradores y coordinadores de la PSU para brindar sugerencias al DEMRE acerca de cómo el proceso de administración de la prueba puede ser mejorado en el futuro.
19. Realizar estudios para descartar el efecto del tiempo y de otras variables de la aplicación (tipos de letras, diseño del cuadernillo de prueba, instrucciones dadas a los estudiantes, condiciones físicas de los locales, etc.) que puedan afectar el desempeño en la prueba por parte de los estudiantes.

20. Aunque el proceso de lectura electrónica de las respuestas (escaneo) es meticuloso, los procedimientos utilizados no han sido documentados ni tampoco hay informes con respecto a temas que surgen durante cada administración. Recomendamos que estos procesos técnicos sean documentados por escrito y que los informes anuales, que registran los temas del escaneo más recientes y su resolución, sean producidos.
21. A la fecha, el proceso de captura de datos realizado por el DEMRE involucra inspecciones mecánicas y manuales de las respuestas múltiples para cada pregunta. El hecho que el proceso involucra dos niveles de resolución, más una verificación manual, es encomiable ya que reduce las fuentes de variancia no relacionadas. Sin embargo, aunque esta información es utilizada principalmente para la resolución de respuestas de ítem de puntajes individuales, recomendamos el uso de estos análisis de tachado a un nivel agregado para un mayor respaldo de la integridad del proceso de administración de la prueba. Debido al elevado perfil de la PSU, también recomendamos un mayor análisis de las amenazas potenciales a la integridad de los puntajes de las pruebas que surjan de comportamientos anti éticos (por ejemplo, copiado de las respuestas).
22. **En términos generales, no hay estudios que proporcionen evidencia que apoyen las decisiones para ajustar los puntajes brutos mediante la corrección por adivinación. Recomendamos implementar estudios prospectivos de investigación para documentar decisiones acerca del empleo de varios estudios en apoyo de las decisiones tomadas. El equipo internacional de evaluación también recomienda una serie de estudios retrospectivos para evaluar cualquier efecto negativo sobre los puntajes de la PSU y sobre las estadísticas de terreno del banco de ítems; por ejemplo, las decisiones tomadas en el pasado debido al uso de fórmulas de puntajes.**

Objetivo 1.1.b: Estándares de calidad para el pilotaje de las preguntas

Descripción:

Este objetivo incluyó una revisión de diversos aspectos del pilotaje de ítems para la PSU, incluyendo:

- El diseño de estudios piloto, es decir, especificaciones, pautas y criterios (Faceta 1).
- El proceso de toma de decisiones y los criterios para la selección de ítems que son piloteados (Faceta 2).
- El proceso de toma de decisiones para determinar el estado del banco de ítems en preparación para el pilotaje de ítems (Faceta 3).
- El proceso de revisión de desempeño de ítems pilotados (Faceta 4).

Para este objetivo, el equipo evaluador verificó la calidad de la muestra del pilotaje, la relevancia del procedimiento establecido para la inclusión de ítems en el piloto, la relativa eficacia del proceso de revisión del banco para obtener ítems para pilotaje, y el criterio para la revisión de ítems una vez que ya han sido piloteados.

Método:

La información para este objetivo se obtuvo originalmente por medio de entrevistas con personas relevantes del DEMRE el 21 de marzo de 2012, incluyendo al director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también

el jefe del proceso de selección. La información documentada del DEMRE que estuvo disponible también se consultó en la preparación de la evaluación de este objetivo.

Hallazgos:

El 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.). Sin embargo, 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. Cuando los estudiantes rinden voluntariamente las pruebas piloto sin ninguna consecuencia para ellos, es posible que su motivación sea muy diferente de la motivación de la población objetivo de las pruebas. Esta disparidad en la motivación podría explicar algunas de las diferencias en el comportamiento estadístico de los ítems presentadas en el Objetivo 1.1.f.

En general, 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.

Según se señaló con anterioridad, aunque las herramientas para administrar los bancos de ítems que fueron revisadas en el Objetivo 1.1.a. se consideraron adecuadas, no ocurre necesariamente lo mismo con los procesos de selección, esto es, porque los procedimientos para la toma de decisiones de selección de ítems no parecen ser adecuados. La documentación al respecto es insuficiente y durante las entrevistas no se logró información adicional. Las decisiones con respecto a la selección de ítems para el pilotaje parecieron dirigirse a un único objetivo, el cual es cerrar la brecha entre conteos de ítems actuales y esperados en el banco. Sin embargo, este proceso también necesita reflejar las características psicométricas de los ítems esperados del pilotaje. La descripción de este proceso no permite la identificación de criterios adicionales que orienten la selección de ítems para el piloto. Estas consideraciones adicionales incluyen explorar el efecto psicométrico de diferentes formatos de ítems o el comportamiento estadístico de ítems basado en su ubicación dentro del cuadernillo de pruebas, entre otros.

De acuerdo con la documentación técnica, el proceso de banco de ítems requiere de un análisis con respecto a la preparación de la construcción de ítems y del pilotaje que tiene lugar periódicamente al inicio de cada año. Esta revisión existente parece enfocarse fundamentalmente sobre la falta de cobertura de matriz de especificaciones de cada prueba, lo cual constituye un criterio válido e importante. Sin embargo, deja de lado la posibilidad de diseñar el pilotaje sobre una base científica para estudiar los efectos psicométricos sobre los ítems, largo de ítems, edición del cuadernillo, etc. Este estudio enriquecería la toma de decisiones desde el diseño hasta la administración. La revisión del banco de ítems es llevada a cabo independientemente por personal responsable de la prueba, y la documentación no indica que se sigan criterios estandarizados al realizar tales revisiones.

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. Sin embargo, el límite superior de la tasa de omisión aceptada es mayor que la que se ve en otros programas. Por ejemplo, la tasa de omisión aceptada en evaluaciones internacionales (tales como PISA) no es tan alta como la aceptada para la PSU. De acuerdo al reporte técnico PISA 2006 (Organización para la Cooperación y Desarrollo Económico, 2009, p. 219), el promedio ponderado de los ítems omitidos fue 5.41%. En el reporte técnico PISA 2009 (Organización para la Cooperación y Desarrollo Económico, 2012, p. 200), el promedio de ítems omitidos fue 4.64%, ligeramente menor que en 2006. Esto sugiere que un límite superior para omisiones podría ser del orden de 10%.

Recomendaciones:

23. **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.** Desarrollar, por ejemplo, cuotas del tamaño de muestras que tengan en cuenta las tasas esperadas de no participación –y que esas tasas puedan ser diferentes para diferentes asignaturas– a fin de que se cumpla uniformemente la meta de 1500 participantes. Aún más, analizar el impacto de las tasas de no participación sobre la representación de variables socio demográficas mayores. **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.** De los resultados de los Objetivos 1.1.f. y 1.1.g., la administración piloto tiene poco valor más allá del análisis marginal de la calidad de los ítems; de ahí, la baja calificación en algunos aspectos.
24. Recomendamos que se suministre documentación adicional para las siguientes áreas: la racionalidad detrás del estudio piloto; recolección y análisis de datos para el estudio piloto; y el proceso para predecir el número de ítems y el número de pasajes de lectura a ser administrados.
25. **Aunque los criterios estadísticos del plan de muestreo para la pre prueba han sido documentados, es decir, modalidad curricular y tipo de institución educativa, recomendamos una mejor articulación de las variables de estratificación.**
26. Recomendamos que la documentación de los criterios para la selección de ítems del piloto sea complementada con una articulación sistemática de las razones para la selección de ítems, que son determinadas por la experticia de los participantes. De acuerdo con el Estándar 3.7, la documentación de estos procesos aseguraría la replicabilidad de los mismos aún cuando los grupos expertos involucrados en el desarrollo varíen, aumentando así la confiabilidad del proceso implementado.
27. En este sentido, recomendamos proporcionar mayores detalles, incluyendo una fundamentación estadística, con respecto a la ubicación de ítems piloto comunes incrustados en más de una forma piloto.
28. La recomendación es para que en la planificación de las aplicaciones piloto se incluyan objetivos claros e intencionados para la verificación de aspectos tales como el efecto psicométrico de usar los mismos ítems con bloques de grupos de ítems diferentes, o del efecto del cambio de posición de un ítem en diferentes cuadernillos, etc. Estos estudios deben estar documentados y deberían proporcionar retroalimentación al diseño de pruebas.

29. Recomendamos documentar el propósito del piloto y la racionalidad para determinar qué ítems son necesarios de acuerdo con especificaciones estandarizadas. También la documentación del piloto debe incluir los requerimientos para el muestreo y análisis de los ítems luego de la administración y los elementos de TCT o de TRI que se consideren relevantes hacia el diseño del piloto.
30. Realizar análisis de la documentación para cada ciclo de pilotaje y medir su cumplimiento con los procedimientos descritos en la sección anterior. Este análisis debería realizarse luego de la administración piloto y debería estar bien documentada. Se pueden elaborar listas de cotejo para documentar el cumplimiento de las especificaciones del pilotaje, los procesos y las etapas. En general, debería ubicarse un sistema de verificación de calidad para monitorear la calidad y la utilidad de los componentes y resultados de los pilotos.
31. Recomendamos documentar el proceso de diagnóstico del banco de ítems y el establecimiento de criterios estandarizados que conduzcan tales procesos para todas las pruebas, o en caso de ser necesario, la justificación a fin de que ese proceso tenga lugar en forma diferente para cada prueba. El contar con manuales para proporcionar la racionalidad hacia los análisis piloto asegura que las decisiones tomadas acerca de aspectos del pilotaje no desatiendan los criterios estadísticos para la selección de ítems.
32. Recomendamos documentar el plan para la comunicación de los resultados de la inspección del banco a los grupos funcionales. Los resultados de los estudios piloto deberían ser emitidos de una manera sistemática entre los equipos responsables de las pruebas. Para que el pilotaje sea efectivo, debería contribuir al mejoramiento del diseño, construcción, revisión y ensamblaje de la prueba.

Objetivo 1.1.c: Criterios hacia la selección de preguntas para el ensamblaje de las pruebas definitivas

Descripción:

Este objetivo incluyó una revisión de los procedimientos para el ensamblaje de la PSU, incluyendo:

- Los usos intencionados y no intencionados de los puntajes de la PSU y la población destinataria de examinados (Faceta 1).
- El diseño y las especificaciones de la PSU que continúan con el ensamblaje de una forma operativa (Faceta 2).
- Las especificaciones para la construcción de la PSU (Faceta 3).
- Matriz de especificaciones para orientar el ensamblaje de una forma operativa (Faceta 4).
- El proceso de ensamblaje de una forma operativa de la PSU y los criterios utilizados para ello (Faceta 5).
- Proceso y criterios para la revisión y aprobación de una forma operativa de la PSU (Faceta 6).

Respecto de este objetivo, el equipo evaluador determinó el contexto histórico para la decisión inicial del desarrollo de la PSU, su aplicación como prueba referida a norma para el proceso de selección, su supuesta base en el currículum de la enseñanza media, su correspondencia con estándares internacionales respecto de la explicación de la adecuada interpretación de los puntajes de la PSU, las especificaciones y prácticas adoptadas en el

desarrollo de las formas de la PSU, y si es que las pruebas construidas finalmente pueden ser consideradas confiables.

Método:

La información para este objetivo se obtuvo mediante entrevistas con personas relevantes del DEMRE el 21 de marzo de 2012, incluyendo el jefe del departamento de construcción de pruebas, los coordinadores de los comités de construcción de pruebas de las cuatro áreas temáticas, el director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección. La documentación formal del DEMRE que estuvo disponible también fue consultada en la preparación de la evaluación de este objetivo.

Hallazgos:

Se reconoce que la PSU es una prueba relativamente nueva, teniendo en cuenta su primera administración, y el hecho que su desarrollo ha sido completado progresivamente en el tiempo (DEMRE, 2010a). Por esta razón no es posible hablar acerca de cambios sustanciales en el significado y los usos de los puntajes de las pruebas a través de su historia. Aún así, se podría esperar que para esta fecha ya hubiera estudios disponibles que detallarían 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.

Hay 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. Según lo indicado en nuestra descripción general, el DEMRE ha realizado algunos intentos de proporcionar advertencias acerca del uso del informe; por ejemplo, no extraer conclusiones acerca de la calidad de la educación entregada en instituciones en particular a partir de los resultados de la PSU. Sin embargo, no es suficiente publicar una advertencia en un sitio web; tales mensajes deben ir integrados en el informe mismo. Las advertencias incluidas en el informe son insuficientes para comunicar claramente el uso pretendido de los resultados de la PSU. Como resultado, la posibilidad del mal uso de los resultados de la PSU, por ejemplo, la desagregación de los puntajes por institución educativa y las comparaciones con otras instituciones, es elevada.

Desde una perspectiva internacional, la experiencia del equipo evaluador con la diseminación de los resultados de exámenes universitarios (tales como SIRPAES) es que tal distribución es limitada (más allá de aquella apuntada a las universidades mismas) a los postulantes individuales y a sus consejeros de la enseñanza secundaria. El principal propósito de esta información es el de mirar hacia adelante para la admisión universitaria de cada estudiante, más que mirar hacia atrás a la calidad de la enseñanza de la institución secundaria. En los Estados Unidos, la calidad de la enseñanza secundaria es adjudicada por evaluaciones a nivel de los estados y específicamente diseñadas para dicho propósito.

Con respecto a la toma de muestras, el equipo de diseño de la PSU evidencia una adecuada consideración de las características de la población objetivo. 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). Esto es, el DEMRE documenta debidamente las estadísticas TCT recolectadas que se emplean para la construcción de la PSU. Sin embargo, durante las entrevistas el

DEMRE informó acerca de la existencia de casos donde algunos indicadores (por ejemplo, nivel de dificultad promedio o índice de discriminación) se alejan del criterio deseable. La documentación del DEMRE no da cuenta del análisis de estos desajustes.

Es evidente que aunque hay documentación que describe el criterio considerado en el diseño de las pruebas, sería deseable contar con el respaldo de referencias técnicas bibliográficas, así como también de estudios llevados a cabo con los datos de las pruebas, para cada una de las decisiones referidas a tal criterio. Una prueba con consecuencias, como la PSU, también debería incluir medidas de precisión como parte de los criterios para su construcción. Tales criterios, como por ejemplo la estimación de errores de medición estándar en el marco de TCT o de TRI, permitirían a los desarrolladores de las pruebas enfocar y minimizar los errores en tramos específicos de la escala de puntajes.

El equipo evaluador ha examinado las características de los ítems operacionales empleados durante pilotajes previos que han sido etiquetados como "anclas". El equipo evaluador los considera por debajo de los estándares internacionales. Estos ítems de ninguna manera son ítems "ancla". Aunque en la documentación de la PSU, el DEMRE se refiere a ellos como "conjuntos ancla", en la práctica estos conjuntos no se utilizan para la calibración de ítems y la equiparación de los puntajes. Aún si fueran utilizados para calibración y equiparación, los números absolutos de ítems en los conjuntos ancla son tan bajos que no serían suficientes para lograr la tarea de una manera válida y confiable. (Ver Objetivo 1.3 para una discusión más amplia de este punto).

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. Si bien es cierto que un buen nivel de capacitación de los profesionales del DEMRE apoya la seguridad de la calidad del proceso, es claro que un manual de construcción de pruebas sería útil, como 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. Adicionalmente, un proceso de capacitación, incluyendo más tiempo de entrenamiento para futuros desarrolladores (tal como el que de acuerdo con la documentación, se lleva a cabo con respecto a los desarrolladores del área de Lenguaje y Comunicación) aseguraría la adecuada apropiación del marco teórico para los otros dominios de las pruebas, de sus especificaciones de pruebas y de las pautas de construcción de pruebas entre los constructores y ampliarían las oportunidades para ver y analizar ejemplos y modelos de ítems de los diferentes formatos utilizados para esos dominios de pruebas.

De acuerdo con lo que se informó durante las entrevistas del equipo técnico, 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. Pero una vez más, esto debería verificarse.

De acuerdo con la documentación revisada, el 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. Adicionalmente, la revisión por parte de un experto del ambiente universitario como revisor final de la prueba ensamblada agrega seguridad respecto de la pertinencia del instrumento dentro de

un proceso de selección universitario. El proceso de revisión podría ser enriquecido si incluyera 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.

En cuanto al sistema de ensamblaje, aunque está parcialmente automatizado, con un sistema que preselecciona ítems en función de criterios dados, de acuerdo con la descripción que se encuentra en la documentación técnica, una gran parte del proceso de selección para el ensamblaje es manual en gran medida, lo cual reduce la eficiencia del proceso (requiere de más tiempo y recursos humanos contar con una prueba ensamblada). Un sistema de ensamblaje de pruebas más automatizado reduciría el riesgo de duplicación de ítems dentro de una prueba y la interferencia de criterios subjetivos cuando se debe tomar una elección respecto de uno entre muchos ítems con posibilidades similares para completar un ensamblaje. Tal sistema se beneficiaría de gran manera con el involucramiento de un marco TRI en lugar de un marco TCT que se usa actualmente para la PSU. Un marco TRI permitiría focalizar las pruebas a los niveles de habilidad de los postulantes de una manera sistemática. De acuerdo a la evaluación previa de la PSU, el ETS reportó una diferencia desproporcionada en la dificultad de la prueba de la PSU debido a la falta del estándar de construcción que tome en consideración los niveles de habilidad de los postulantes. (Educational Testing Service, 2005).

Recomendaciones:

33. **Recomendamos una 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.**
34. Recomendamos documentación de los criterios para la construcción de pruebas. Esta documentación debería listar:
    a. características y calificaciones de los participantes;
    b. definiciones claras de los usos primarios y secundarios de los puntajes de las pruebas; y
    c. análisis de las consecuencias sobre los puntajes de las pruebas al apartarse de los criterios de construcción de pruebas.
35. **Recomendamos que se empleen 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. Recomendamos actualizar las especificaciones de los conjuntos ancla del DEMRE para cumplir con los estándares internacionales.**
36. **Recomendamos proporcionar un manual para la construcción de pruebas.**
37. **Recomendamos que el proceso de capacitación para el ensamblaje de pruebas sea unificado mediante la generación de pautas estandarizadas que se les enseñen a todos.** Estos ejemplos deben ilustrar errores que deberían

evitarse y aspectos que deberían ser considerados para lograr cumplir con los criterios de aceptación establecidos. **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.**

38. Respecto de la implementación de nuevas tablas de especificaciones, dado el cambio curricular del año 2009, recomendamos la introducción de un proceso de validación de las respectivas tablas de especificaciones con equipos de expertos en la enseñanza media y superior (primer semestre) para agregar validez externa al proceso, enfatizando aspectos tales como pertinencia y relevancia de los aspectos incluidos en dichas tablas.

39. Recomendamos someter la decisión de colocar más énfasis en la modalidad Científico-Humanista que sobre la modalidad Técnico-Profesional a una validación externa e incluir en los marcos teóricos de las pruebas la justificación de esta decisión.

40. El proceso de construcción de pruebas ha sido bien descrito dentro del marco TCT. Recomendamos que la capacitación sea formalizada y documentada, por ejemplo, capacitación en herramientas de construcción y procesos para desarrollar objetivos estadísticos.

41. Recomendamos considerar la automatización del ensamblaje de la prueba para evitar los riesgos de seguridad que puedan surgir en el futuro, a partir de la repetición de las preguntas o de las inconsistencias que se encuentren dentro de la tabla de especificaciones.

42. Se sugiere incluir un revisor de la prueba ensamblada proveniente del nivel de enseñanza media, en contraste con el revisor proveniente del nivel universitario.

43. **Recomendamos una transición hacia el marco TRI 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.**

44. Recomendamos documentar en mayor detalle el tratamiento de los borradores o del material de pruebas cambiado durante los sucesivos procesos de revisión.

45. Recomendamos que las revisiones externas del proceso de la revisión de la prueba operacional representen una mayor diversidad institucional (esto es, que no todos los revisores sean exclusivamente de la Universidad de Chile).

46. Recomendamos documentar más extensamente el procedimiento a seguir cuando uno de los revisores externos sugiera eliminar o reemplazar un ítem de una prueba preensamblada.

47. La recomendación es de documentar de una manera precisa las instrucciones sobre ensamblaje de pruebas, indicando la distribución ideal de las preguntas en función de los indicadores estadísticos que se toman en cuenta, esto es, el número de ítems máximo y mínimo aceptable con un cierto nivel de discriminación, etc., a fin de asegurar la comparabilidad de la prueba entre diferentes aplicaciones.

Objetivo 1.1.d: Estándares de calidad en la administración del banco de ítems

Descripción:

Este objetivo incluyó una revisión de los estándares de la administración del banco de ítems de la PSU, incluyendo:

- La estructura del banco de ítems (es decir, el diseño lógico, las plataformas, campos y registros) (Faceta 1).
- Las herramientas del banco de ítems (Faceta 2).
- Los protocolos de acceso y procesos de seguridad (Faceta 3).
- El flujo de procesos para la actualización y agregar registros al banco de ítems (Faceta 4).

Para este objetivo, el equipo evaluador investigó el banco de ítems, el software Pregunta Segura utilizado para la construcción de pruebas, la asignación de valores estadísticos en el banco luego del análisis de datos, así como también la seguridad y flujo de procesos de los grupos que tienen acceso al banco de ítems (la Unidad de Tecnologías de la Información, la Unidad de Estudios e Investigación y la Unidad de Construcción de Pruebas).

Método:

La información para este objetivo se obtuvo mediante entrevistas con personas relevantes del DEMRE el 21 de marzo de 2012, incluyendo el jefe del departamento de construcción de pruebas, los coordinadores de los comités de construcción de pruebas de las cuatro áreas temáticas, el director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección. La documentación formal disponible del DEMRE también fue consultada en la preparación de la evaluación de este objetivo.

Hallazgos:

Los documentos revisados aportan 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. Sin embargo, la información es presentada desde la perspectiva de la arquitectura del software del banco de ítems más que de la perspectiva de la psicométrica, lo cual tiende a ser de importancia para este proceso.

A pesar de tener una estructura clara y una base de datos poderosa, con mucha información, no se menciona si la información estadística relacionada con el uso del banco está siendo producida. El uso psicométrico de este tipo de información podría conducir los desarrollos de la PSU en el futuro inmediato y en el mediano plazo. Vale la pena desarrollar reportes que informen a los diseñadores acerca de los comportamientos de las pruebas y no solamente acerca del uso de ítems y sus estadísticas.

En los documentos no hay mención acerca de los criterios empleados en la actualización del banco de ítems más allá de la cantidad de ítems por región de la tabla de especificaciones, ni hay mención acerca de la capacidad del banco.

En la información suministrada hay alguna mención acerca de protección del sistema, sistemas de respaldo para la información y auditoría del sistema. Sin embargo, la información de los bancos de datos no tiene respaldos en otros archivos externos,

aumentando el riesgo de la pérdida total de información debido a un accidente. No hay procedimientos ni recursos asignados para la protección del sistema contra posibles ataques de virus computacionales. No hay nada programado para la actualización y mantención del banco de ítems en el sentido de incluir nuevas tecnologías que podrían habilitar su desarrollo en relación a la PSU.

Recomendaciones:

48. **Se encontró insuficiente información técnica que describa el banco de ítems más allá de aquella relacionada con la base arquitectónica. Recomendamos suministrar la información faltante respecto de sus módulos, su funcionalidad y características.**
49. Aunque los aspectos mencionados anteriormente sobre el banco de ítems están claramente expuestos, recomendamos la producción de características técnicas más precisas del proceso de elaboración de la prueba; específicamente, se necesita generar manuales técnicos y de uso para facilitar una comprensión de lo que tiene lugar y del verdadero alcance y limitaciones del banco.
50. Aunque no hay un estándar específico acerca de cuáles indicadores deberían ser incluidos en un banco de ítems, recomendamos agregar indicadores de uso de ítems adicionales al banco de ítems; por ejemplo, el saber la historia de administración de un ítem nos permitiría calcular su tasa de exposición.
51. Recomendamos un documento técnico que describa las características técnicas completas del software.
52. Recomendamos una descripción más clara de las reglas específicas acerca de cómo se realizan las asignaciones a nivel de usuario.
53. Es necesario llevar a cabo una inspección detallada del sistema del banco de ítems para determinar las necesidades de actualización y las modificaciones dados los desarrollos tecnológicos.
54. Con respecto a respaldos para los sistemas de información, recomendamos, si bajo mayor investigación se descubre que el banco de ítems no está respaldado con sistemas redundantes,
    a. proteger contra interrupciones del servicio y efectuar la mantención de las ediciones más recientes, estableciendo un servicio redundante (por ejemplo, servidores), y
    b. proteger contra fallas catastróficas, programando respaldos de medios incrementales diarios y respaldos completos de medios semanalmente de la base de datos.
55. Actualmente la revisión de ítems se lleva a cabo basada en la experiencia de los participantes en sus respectivas sesiones (se busca consenso). Sin embargo, no hay mención a manuales o estándares a seguir. **Por lo tanto, recomendamos documentar los criterios para la revisión de ítems.**
56. Más allá del hecho que los miembros del comité parecen pensar que un ítem es bueno y que posee una cierta dificultad, no se aplican otros elementos estadísticos o psicométricos. No se mencionan manuales de análisis de información estadística. **Por lo tanto, recomendamos que los comités en forma adicional analicen 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.**

Objetivo 1.1.e: 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

Descripción:

Este objetivo incluyó una revisión del proceso para el análisis del desempeño de los ítems seleccionados para y eventualmente usados en las formas operacionales de pruebas de la PSU, incluyendo:

- Los criterios de calidad para juzgar los ítems administrados operacionalmente (muestras piloto de estudiantes y población de estudiantes) (Faceta 1).
- El proceso para la selección de los ítems operacionales para rendir puntajes de pruebas (Faceta 2).
- El proceso de revisión y aprobación de ítems operacionales seleccionados (Faceta 3).

Respecto de este objetivo, el equipo evaluador revisó los criterios de la Teoría Clásica de Pruebas empleados y los criterios de dificultad y discriminación de la Teoría de Respuesta al Ítem consultados durante la determinación del conjunto de ítems a colocar en las formas operacionales de la PSU. Se le prestó especial atención a los criterios aplicados y a las prácticas seguidas por cada uno de los comités de áreas temáticas y si es que los comités y los revisores externos habían documentado sistemáticamente cómo habían priorizado los diferentes indicadores de ítems.

Método:

La información para este objetivo se obtuvo mediante entrevistas con personas relevantes del DEMRE el 21 de marzo de 2012, incluyendo el jefe del departamento de construcción de pruebas, los coordinadores de los comités de construcción de pruebas de las cuatro áreas temáticas, el director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección. La documentación formal disponible del DEMRE también fue consultada en la preparación de la evaluación de este objetivo.

Hallazgos:

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. Las excepciones son las 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. La diferenciación en los criterios aplicados para las diferentes pruebas se explica. Sin embargo, no es respaldada por estudios que evidencien que tales diferencias no tienen ningún efecto sobre el proceso de evaluación, tomando en cuenta el propósito de la prueba, la población evaluada y el objeto de evaluación. Además, no hay documentación que describa, ya sea la base para el criterio empleado en la selección de ítem o un procedimiento a seguir para seleccionar un ítem, cuando se cumplen ciertos criterios y otros criterios no son cumplidos.

Las decisiones sobre la selección de ítems para pruebas operacionales son compartidas entre los miembros del equipo técnico quienes, según ya ha sido descrito, cuentan con las calificaciones académicas y psicométricas que les permiten realizar esta labor. Las decisiones se basan en etapas de revisión y discusiones en equipo que le confieren confiabilidad al proceso. Los procedimientos para tratar con ítems que demuestran comportamientos extremos son aceptables porque tales ítems son, de hecho, revisados adecuadamente por el comité de contenidos del DEMRE, utilizando criterios relevantes.
En general, el proceso parece ser el adecuado y cumple con las expectativas mínimas.

Los equipos responsables de la selección de ítems para las pruebas operacionales se basan en la documentación respecto de los indicadores y características de los ítems que deben ser tomados en cuenta. Sin embargo, el DEMRE no ha establecido un procedimiento para seleccionar ítems cuando un ítem cumple con algunos criterios, pero no con otros. Esto deja espacio para una mayor subjetividad en la selección de los ítems y puede afectar la comparabilidad de las formas de prueba entre administraciones.

En la documentación revisada no hay evidencia de que los procedimientos de selección de preguntas incluyen comparaciones entre el comportamiento de los ítems en aplicaciones piloto y operacionales.

Las pautas del DEMRE explícitamente permiten que los ítems pilotados sean editados o de otra forma, 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.

Recomendaciones:

57. Recomendamos documentar la razón de los cambios de los criterios de selección de ítems entre los años 2005 y 2011; específicamente, documentar la historia de la prueba en sus procedimientos técnicos y las razones para realizar cambios en los mismos.
58. Recomendamos especificar cómo los indicadores TRI son analizados: CCI y Función de Información. ¿Son ellos los criterios para aceptación o rechazo?
59. Recomendamos documentar la justificación para establecer un rango mucho más amplio (desde -5 a +5) de aceptación del valor de dificultad que aquel indicado como aceptable por la literatura TRI (desde -3 a +3), así como también describir y respaldar con mayor precisión la decisión de aplicar criterios de dificultad diferentes (p) para pruebas diferentes, incluyendo controles en el caso que sean necesarios para impedir que tales decisiones sean contraproducentes respecto de los propósitos de la prueba.
60. Recomendamos revisar y reconciliar los criterios que difieren para valores de discriminación TRI aceptables (esto es, $a \geq 0.6$ versus $a \geq 0.65$), dado que los ítems seleccionados con ese criterio podrían ser catalogados a un nivel bajo de discriminación, de acuerdo con la tabla de clasificación para este indicador.
61. Recomendamos revisar el actual criterio utilizado para marcar ítems con altas tasas de omisión, considerando un estándar extraído de las observaciones y experiencia del equipo evaluador respecto de evaluaciones internacionales (por ejemplo, 10% omisiones).
62. **Recomendamos 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, nosotros recomendamos documentar la racionalidad que determina cuál indicador deberá tener prioridad frente al resto.

63. **Recomendamos 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.**
64. Recomendamos documentar las razones que fueron consideradas para establecer algunas diferencias en la revisión de ítems por parte del comité de Lenguaje y del resto de comités. También recomendamos analizar la posibilidad de estandarizar estos procedimientos para todas las pruebas, lo más posible. Cuando esto no sea posible, recomendamos que los argumentos sean documentados y los controles anticipados para que las diferencias no afecten los resultados o de otra forma, sean contraproducentes con respecto al propósito de la prueba.
65. También recomendamos que la elección de participantes para los procesos de revisión sea efectuado deliberadamente para aumentar la diversidad institucional y geográfica. Finalmente, recomendamos documentar las políticas con respecto a las contingencias, tales como el número de ítems no aprobados por los criterios establecidos, debido a que son más elevados que lo que se encuentra regularmente.
66. **Recomendamos 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.**

Objetivo 1.1.f: 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

Descripción:

Este objetivo incluyó un análisis de los factores asociados al desempeño diferencial de ítems entre sus administraciones piloto y operacionales de la PSU.

Para el objetivo, el equipo evaluador evaluó variables de la subpoblación asociadas a la mayor variabilidad en estadísticas de ítem desde el piloto hasta el uso operacional.

Método:

La información para este objetivo se desarrolló mediante un análisis del equipo evaluador de datos a nivel de ítem proporcionados por el DEMRE.

Hallazgos:

Con base en los resultados de análisis llevados a cabo tanto con el desempeño piloto y operacional de ítems a lo largo de años, hay diferencias significativas en los indicadores de desempeño de ítems entre la administración piloto y la administración operacional.

Considerando los resultados de los valores de la teoría clásica de pruebas en su conjunto (dificultad, correlación biserial y omisión) las diferencias son grandes y significativas cuando se analizan los valores de ítem para el conjunto de todos los años y en todas las pruebas. La categoría de omisiones, por ejemplo, aumenta sustancialmente en la administración final; el

supuesto es que aquellos evaluados, luego de enterarse de que la calificación final usa una fórmula de corrección, omiten aquellas respuestas respecto de las cuales tienen una duda razonable. Este hecho afecta, por sí solo, en una gran medida los valores de todas las estimaciones estadísticas, ocasionando que los valores de ítem (no solamente aquellos de TCT, sino que también aquellos de TRI) sean subestimados o sobrestimados.

El mismo efecto ocurre con los valores TRI, los cuales son diferentes entre ambas administraciones de la prueba. Esto contradice el fundamento teórico del TRI de que estos valores son independientes de la muestra de la cual son obtenidos, si son representativos de la misma población.

Para todos los años, los valores de asociación más bajos entre el piloto y la administración final corresponden a la correlación biserial; específicamente, para género, dependencia y modalidad. Es necesario notar que estos valores de discriminación son afectados por el puntaje total obtenido de todos los ítems de pruebas y las estrategias empleadas para responder. Esto es, los valores de discriminación son afectados por los ítems con problemas técnicos o por cómo es manejado el piloto, o aún por el hecho de que los estudiantes saben acerca de la corrección por adivinación usada para los puntajes de la prueba. En este respecto, se espera que los valores de la correlación biserial entre el piloto y la administración final cambien sustancialmente, lo cual se ve en las tablas y cifras. Podría también ser el caso que los cómputos biseriales sean afectados por los niveles de dificultad de los ítems y así tornarse bajos como resultado de aquel artefacto metodológico.

Un elevado nivel de tasas de omisión y el uso de la corrección por adivinación también puede haber contribuido a las discrepancias en el desempeño de ítems entre las administraciones piloto y operacionales. Debido a que la corrección por adivinación es conocida para los postulantes, los índices del TCT piloto pueden no ser aproximaciones confiables del desempeño de ítems en el contexto operacional.

Resumiendo, el piloto sí proporciona información importante acerca de la calidad de los ítems que pueden usarse para tomar decisiones acerca de ellos en términos de su inclusión o no en el banco para su utilización en cualquier administración final. Sin embargo, no está claro que los datos puedan usarse como valores precisos de los diferentes índices estudiados. Específicamente, los cambios en los valores obtenidos en el piloto y en la administración final son más o menos grandes y significativos. Estos cambios ocurren principalmente debido a 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.

En segundo lugar, 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.

En tercer lugar, el hecho que el piloto es una administración voluntaria modifica la muestra seleccionada.

Recomendaciones:

67. **Recomendamos tomar medidas a fin de que los valores obtenidos en el piloto y la administración final sean más cercanos entre sí;** por ejemplo, mayor consideración a variables como género deben ser tomadas en cuenta para el muestreo de la población de la administración piloto.
68. **De acuerdo con la recomendación anterior, y de otra evidencia en la evaluación, se recomienda que el DEMRE reconsidere 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.**
69. **Recomendamos 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.**
70. **Recomendamos 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.** Esta redefinición incluye considerar otras formas de pilotaje de ítems tales como la inclusión de grupos de ítems en la administración operacional. Estos grupos de ítems no serían asignados puntajes o utilizados para obtener resultados de aquellos respondiéndolos, pero eso proporcionaría estadísticas muy cercanas a los datos de las administraciones operacionales, ya que los evaluados no sabrían cuáles ítems están siendo pilotados.
71. Aunque la muestra que participa en el piloto es voluntaria y hay un compromiso de parte de las instituciones seleccionadas para que sus estudiantes participen, es importante analizar el impacto de la no participación en la representación pretendida de grupos socioeconómicos mayores de manera tal de dar cuenta de un posible sesgo en los resultados. Una vez realizado este análisis, se podría dar cuenta de las tasas históricas de no participación mediante la sobremuestra de dichos grupos.

Objetivo 1.1.g: Exploración de variables asociadas al DIF, en el caso que se encuentre presente

Descripción:

Este objetivo incluyó una exploración de las variables asociadas al DIF en la PSU.

Para este objetivo, el equipo evaluador suministró (1) análisis experto de procesos DIF documentados en los informes del DEMRE y aclarados durante entrevistas al personal del DEMRE, (2) una inspección analítica de variables relacionadas con el DIF, cubriendo el modelaje estadístico de los resultados de DIF con información relevante de nivel de ítem y (3) una demostración empírica de computación de DIF con datos del proceso de selección del año 2012.

Método:

La información para este objetivo se desarrolló mediante análisis de datos proporcionados por el DEMRE por parte de los equipos evaluadores y aclarados a través de preguntas en entrevistas.

Hallazgos:

Funcionamiento Diferencial de Ítems (DIF) se refiere a establecer la dificultad relativa de una pregunta de una prueba para un grupo de candidatos versus otro grupo cuando se equiparan esos candidatos dentro de grupos con un mismo nivel de habilidades. Este proceso estadístico permite a los desarrolladores identificar aquellos ítems con un potencial de sesgo. Tales ítems son luego revisados para determinar si esos efectos diferenciales son irrelevantes al constructo objetivo de la prueba. 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.

Hay algunos aspectos del enfoque del DEMRE en relación con el análisis DIF que han llamado la atención del equipo evaluador. Un aspecto es la decisión del DEMRE de descartar información con respecto a la elevada prevalencia de tasas de DIF en las administraciones piloto. Las elevadas tasas de DIF 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.

El segundo aspecto de los procedimientos DIF del DEMRE que plantea una preocupación es su decisión de enfatizar los resultados DIF operacionales sobre los resultados DIF piloto. El equipo evaluador considera que estas decisiones son problemáticas porque los análisis y resultados DIF son de importancia para la evaluación proactiva de la equidad durante el pilotaje de ítems. El abordar el DIF a través de resultados operacionales es riesgoso porque las mejores prácticas exigen examinar el ítem en cuanto a sesgo antes de ser presentados a los estudiantes durante las administraciones operacionales. En un mundo ideal, los procesos de control de calidad son establecidos para detectar ítems anómalos antes de que se conviertan en operacionales. En escenarios aplicados, retirar ítems que indican DIF requiere de interpretación humana experta de las fuentes de varianza de constructo irrelevantes detrás de las banderas estadísticas.

Un tercer aspecto es la falta de una pauta de políticas que dirija la selección de grupos de referencia y focales para los análisis DIF. El hecho que el DIF es calculado solamente para género y dependencia también es problemático. Internacionalmente, el concepto de clases de examinados protegidos ha influido la práctica de definir los grupos para análisis DIF. Recientemente, el concepto fue ampliado para incluir variables para entender de mejor manera los factores detrás del DIF (es decir, exposición al currículum). El análisis del DEMRE debería incluir también factores tales como nivel socioeconómico, modalidad curricular de enseñanza y región. También, se recomienda tener precaución sobre la dependencia de comparaciones múltiples (es decir, Privado vs. Municipal, Privado vs. Subvencionado y Municipal vs. Subvencionado) sin hipótesis particulares a verificar. Es obvio esperar que las comparaciones múltiples pudieran tener un efecto sobre la tasa de error Tipo I y afectar la eficiencia del proceso. La documentación del DEMRE no da una racionalidad hacia el uso de comparaciones múltiples. Las banderas DIF no son

necesariamente indicativas de sesgo y los análisis DIF no deberían llevarse a cabo mecánicamente.

Un aspecto final que plantea preocupación respecto de los procesos DIF de DEMRE es su decisión de usar más de un enfoque para explorar el DIF. Este enfoque hace recordar la historia del hombre que tenía dos relojes que no sabía la hora exacta. Al usarlo, ya sea individualmente o en forma conjunta, los enfoques apuntan a la detección del DIF relativamente al conjunto de ítems de prueba, con diferentes niveles de error estadístico (Tipo I y Tipo II) y de poder estadístico. La documentación del DEMRE no parece incluir ninguna racionalidad para el empleo de más de un método de análisis DIF o ninguna declaración acerca de la relativa superioridad de los enfoques.

El análisis DIF es una tarea que va más allá del procesamiento de datos y el uso de criterios estándar acerca de tales datos. El desarrollo de procedimientos que permitan la detección de explicaciones plausibles acerca de la existencia de DIF debería agregarse al repertorio de DEMRE. El análisis del equipo evaluador de datos DIF archivados de la PSU indicó una manera directa de explorar hacia fuentes potenciales asociadas al DIF empleando la regresión logística. Este tipo de análisis podría ampliarse para acomodar otros atributos de ítem, tales como el uso de palabras, presencia o ausencia de arte y la naturaleza de distractores.

La demostración proporcionada por el equipo evaluador ejemplifica la computación de DIF para un conjunto más amplio de variables demográficas u otras agrupaciones. El seguimiento a esta demostración involucraría a grupos de especialistas en contenidos calificados y de docentes en reuniones en las cuales los ítems con marcadores de alerta DIF sean analizados en mayor profundidad. Los resultados de esas reuniones son importantes porque amplían la comprensión de los factores que afectan la calidad de ítems, lo cual puede mejorar la capacitación en el desarrollo de ítems e informar las especificaciones de desarrollo de ítems en el futuro.

Para el piloto, siempre que el DIF estadístico de un ítem haya sido analizado por un comité de revisión de sesgo y estos expertos hayan encontrado que el marcador de alerta no estaba asociado a ningún sesgo significativo, entonces debería permitirse el uso del ítem en caso de ser necesario para llenar brechas de contenido en la prueba.

Para la prueba operacional, si hay marcador de alerta de DIF para un ítem, no necesariamente significa que el ítem tenía sesgo. Se necesita de una revisión para determinar por qué se marcó el ítem; por ejemplo, algo que no se encontró con anterioridad (por ejemplo, llaves dobles) y fundamentalmente error de formateo o de impresión podría originar el marcador de alerta de DIF. Las tasas de omisión también pueden confundir la información.

Entre otros resultados, la demostración encontró que:

- La mayoría de los ítems indicaban DIF despreciable (A), con muy pocos ítems evidenciando DIF débil o fuerte (B o C).
- La PSU de Ciencias (porciones comunes y electivas) indicaron un mayor número de banderas C de DIF que la PSU de Matemática, Lenguaje y Ciencias Sociales.
- La variable de género mostró el mayor número de banderas C de DIF para una porción común de la prueba de Ciencias (seis favoreciendo a los hombres y tres favoreciendo a las mujeres).

- Las variables SES, región, modalidad curricular, y modalidad, indicaron menos banderas C de DIF, con el mayor número de banderas C de DIF en la prueba de Química para la modalidad Científico-Humanista versus la modalidad Técnico-Profesional (tres favoreciendo la Científico-Humanista y cuatro favoreciendo la Técnico-Profesional).

Recomendaciones:

72. **Recomendamos 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.** Una vez que los datos hayan sido analizados, debería agregarse un registro de las decisiones logradas en la revisión de datos a la documentación de ítems asociada, indicando la decisión de usarlos o de no usarlos en la administración operacional.
73. **Recomendamos ampliar los análisis DIF a subgrupos relevantes que históricamente no han sido parte de los análisis DIF del DEMRE.** Como mínimo, los análisis DIF deberían ampliarse a los siguientes subgrupos: región, status socioeconómico y modalidad curricular.
74. **Recomendamos establecer una política para definir los grupos de referencia.** Actualmente el proceso que se sigue involucra comparaciones múltiples entre categorías de la variable de subgrupo, que no solamente es ineficiente, sino que aumenta la tasa de error Tipo I para los resultados DIF.
75. **Recomendamos elegir el método de DIF de Mantel-Haenszel en vez de usar múltiples métodos de DIF.** El uso del método Mantel-Haenszel Chi cuadrado está bien documentado y permite el empleo de las reglas de clasificación para el DIF de ETS. Si por cualquiera razón se necesita un método de respaldo, el equipo evaluador recomienda el método de regresión logística. La dependencia de un proceso único debería estar claramente declarada en la documentación. El uso de métodos múltiples se torna problemático porque los diferentes métodos tienen diferentes tasas de error Tipo I.
76. **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. Recomendamos que el programa de pruebas de PSU cuente con criterios para invalidar ítems piloto con resultados DIF, tales como marcadores C. El proceso debería diferenciar entre la identificación estadística de DIF y la identificación de las fuentes de contenido de DIF. El equipo evaluador recomienda evitar el uso de múltiples métodos estadísticos DIF.**
77. El equipo evaluador recomienda tomar en cuenta las delimitaciones establecidas en las políticas educativas y las prácticas de medición al elegir grupos focales y de referencia para los análisis DIF.
78. El programa de la PSU debería investigar las fuentes de DIF y utilizar los resultados para afinar sus prácticas de elaboración de ítems, modelos de construcción de pruebas y el proceso de puntuación. Enfoques analíticos pueden ser usados para ganar mayor comprensión de variables relacionadas con marcadores DIF.
79. **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.**

Objetivo 1.1.h: Análisis de procedimientos para el cálculo de puntajes estandarizados, transformación de puntajes en relación a las distribuciones originales

Descripción:

Este objetivo incluyó un análisis de los puntajes estandarizados de la PSU. Como parte de este objetivo, el equipo evaluador calculó puntajes brutos corregidos de la PSU, escalas de puntaje de la PSU y puntajes escalados y suavizados de la PSU a fin de comparar los puntajes estandarizados de la PSU y la transformación de los puntajes en relación a la distribución original de puntajes brutos corregidos.

Método:

La información respecto de este objetivo fue obtenida por medio de análisis llevados a cabo por parte del equipo evaluador y mediante entrevistas con personal relevante del DEMRE el 23 de marzo de 2012, incluyendo al director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección.

Hallazgos:

El equipo evaluador internacional reconoce los esfuerzos realizados para proporcionar puntajes estandarizados de la PSU y sus procedimientos computacionales correspondientes. Los pasos seguidos para calcular las escalas de puntajes de la PSU se parecen, en conjunto, a la estructura canónica de procesos para calcular puntajes de escala en contextos de referencia de norma. Los pasos responden a la política de selección ordenada por el CRUCH y el papel del DEMRE como entidad procesadora de datos para los puntajes de postulación. Conforme a lo ordenado por el CRUCH, los puntajes de la prueba PSU son usados para ordenar a los postulantes que buscan ser admitidos a estudios de nivel universitario de acuerdo a rangos.

La documentación de los procesos de la PSU para desarrollar escalas de puntajes proporciona un buen punto de partida, pero esta documentación necesita mayor elaboración e inclusión de evidencia de respaldo hacia: (1) la corrección por adivinación; (2) la racionalidad para elegir la desviación promedio y estándar de las escalas de puntaje de la PSU; (3) las decisiones de truncar las escalas de puntajes de la PSU; y (4) la mantención de la escala de la PSU.

Los evaluadores reconocieron problemas graves con el uso de la fórmula para corregir por adivinación en la PSU. Debido al papel esencial que tales puntajes juegan en la información de los puntajes de la prueba de la PSU, el equipo evaluador internacional considera su empleo como problemático porque desafía la validez de los puntajes de la PSU y los resultados de la administración de pruebas piloto de la PSU. Esto puede ser remediado en los próximos años mediante la adopción de modelos de Teoría de Respuesta al Ítem que dan cuenta de la adivinación. La contribución de los puntajes corregidos al mejoramiento de la confiabilidad de los puntajes de la prueba PSU y la validez predictiva de los puntajes, y la opinión pública no están documentados.

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. Esto es una limitación

grave considerando que las decisiones se llevan a cabo con los puntajes escalados. Las magnitudes de los errores típicos condicionales de las escalas de puntaje involucradas en las decisiones de selección universitaria deben ser estimadas y comunicadas a las audiencias de la PSU.

La mayor preocupación en el desarrollo de la escala de la PSU es la falta de un mecanismo para mantener la escala de la PSU a lo largo de los procesos de admisión. Tal como se mencionó anteriormente, en Chile, el desempeño en la prueba de la PSU se informa con la escala de la PSU, pero no se realiza una equiparación para mantener la escala a lo largo de los años de selección. Esto constituye un problema grave que debe ser atendido para asegurar que la prueba de selección universitaria de Chile sea justa para los examinados y asegurar comparaciones válidas de los puntajes de las pruebas entre las administraciones de las pruebas. Para mayor información sobre la evaluación de la equiparación de la PSU, favor de ver el Objetivo 1.3 de este informe.

Finalmente, a partir de las entrevistas, el equipo evaluador obtuvo información más precisa de las características principales del proceso de estandarización del promedio de notas de la enseñanza media (NEM). El NEM es calculado promediando las notas finales logradas en los estudios de la enseñanza media. El sistema de calificaciones va desde una nota mínima(1.0) a una nota máxima(7.0). Sin embargo, el DEMRE no fue capaz de profundizar nuestra comprensión proporcionando mayor información acerca de la estructura del NEM y del proceso para el desarrollo de normas para el NEM debido a la falta de documentación acerca de esos temas. Debido al papel que el NEM juega en el cálculo de los puntajes de postulación, es problemático no saber las propiedades psicométricas de los puntajes NEM. De especial significancia es la comparabilidad de sentidos de los puntajes NEM comparando múltiples subpoblaciones. 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.

El equipo evaluador también reconoce graves problemas en la documentación y adecuación técnica de las escalas, lo cual ha ocasionado que el equipo desapruebe los procesos asociados para el desarrollo de escalas. La mayoría de los estándares profesionales listados para evaluar este objetivo no se cumplieron (estándares profesionales 2.2, 2.14, 4.2, 4.5, 4.6, 4.8). Estas deficiencias pueden ser mejoradas en el futuro cercano.

Ni las propiedades psicométricas ni un método de interpretación propuesto del puntaje de postulación han sido documentados. Mientras los criterios de selección estipulan el uso de puntajes de la PSU y puntajes NEM (dentro de un conjunto de ponderaciones y consideraciones de políticas) como elementos para el puntaje de postulación, poco se sabe acerca de las características psicométricas del puntaje de postulación tales como, por ejemplo, la unidad de origen y de dispersión. El programa PSU imputa un sentido normativo a los puntajes individuales en las pruebas, sin embargo, esta práctica no presenta evidencia para el puntaje de postulación.

Otro problema que detectamos es la falta de información acerca de la precisión en la medición de los puntajes de postulación. Al clasificar los puntajes de postulación, las diferencias numéricas entre los puntajes de postulación no son de consideración a pesar de su potencial falta de significado práctico. Un puntaje de 672.15 puntos, por ejemplo, se clasifica sobre un puntaje de 672.13 puntos; sin embargo, ambos puntajes indican diferencias en el segundo decimal. Sin mediciones de precisión de puntajes disponibles para entender las diferencias entre los puntajes, es plausible que las diferencias del tamaño de unos pocos puntos decimales pueda interpretarse y usarse para tomar decisiones acerca de

quién es seleccionado y quién no. Resumiendo, la falta de medición de error estándar condicional respecto de los puntajes de postulación es un inconveniente mayor que necesita ser abordado en el futuro cercano. Se incluye más discusión acerca de las necesidades de calcular mediciones de exactitud y precisión en la sección que cubre al Objetivo 1.1.i (confiabilidad y medición del error estándar condicional).

Los evaluadores encontraron algo de documentación de las descripciones respecto de los procesos para derivar escalas de puntajes de la PSU para pruebas de la PSU individuales. La documentación revisada y la información a partir de las entrevistas ayudaron a entender el proceso para desarrollar la escala y el proceso para asignar significado a los puntos de la escala; por ejemplo, mientras que la documentación resumía las características de la escala de la PSU, las entrevistas contribuían a que las evaluaciones obtuvieran información afinada sobre la escala de la PSU. La documentación del DEMRE necesita ser ampliada para cubrir la descripción de la precisión de medición (por ejemplo, la medición del error estándar condicional)de pruebas individuales de la PSU, lo cual hasta la fecha actual no ha sido ni calculado ni informado a los usuarios. También se recomienda que se proporcionen resúmenes de las limitaciones de las escalas derivadas de puntajes donde se aplique; por ejemplo, encontramos el proceso de suavizamiento manual de las colas superiores de las distribuciones (1%) problemáticas por las siguientes razones: Los procesos seguidos por el DEMRE dependen del juicio humano y de la falta de chequeos de control de calidad. Además, el DEMRE no ha producido evidencia acerca de los efectos del suavizamiento sobre la disminución del sesgo.

Los evaluadores encontraron poca o ninguna documentación del proceso para derivar normas para las NEM. Durante las entrevistas con el personal del DEMRE, los evaluadores internacionales descubrieron la existencia y el uso de normas NEM. Las entrevistas con personas relevantes pusieron en evidencia que las normas NEM han sido usadas desde el año 2003. A partir de la entrevista también se supo que hay tres conjuntos de normas NEM: (1) Científico-Humanista (diurna), (2) Científico-Humanista (vespertina), y (3) Técnico-Profesional, respectivamente. A lo largo de las entrevistas, el personal del DEMRE ha expresado su falta de conocimiento de las condiciones bajo las cuales las normas fueron realizadas y comentaron acerca de la falta de disponibilidad de informes técnicos sobre los estudios del establecimiento de las normas. En el momento de las entrevistas, el DEMRE no fue capaz de profundizar nuestro conocimiento por medio de suministrar más información sobre la estructura del NEM y del proceso para desarrollar normas del NEM y las propiedades psicométricas de los puntajes NEM son problemáticos. La falta de evidencia que respalde la comparabilidad de los criterios para calificar a los estudiantes en las salas de clases es otra preocupación. Los puntajes NEM se basan en prácticas de calificación que pueden y no pueden ser comparables entre las instituciones educativas (Privadas, Subvencionadas y Municipales) y las modalidades curriculares (Científico-Humanista y Técnico-Profesional), por ejemplo.

Recomendaciones:

80. En Chile, la contribución de la corrección por adivinación para el mejoramiento de la confiabilidad de los puntajes en las pruebas de la PSU, la validez predictiva de los puntajes de la PSU, y la opinión pública sobre la PSU no se encuentra documentada. Internacionalmente, el uso de la corrección por adivinación enfrenta desafíos en estas áreas según se presenta en la evaluación sumativa de esta faceta. Adicionalmente, cuando se consideran las omisiones como no alcanzadas, los puntajes corregidos varían entre los estudiantes que poseen un mismo puntaje neto, pero difieren en sus tasas de omisión. La corrección por adivinación puede conducir a los estudiantes a usar una estrategia para abordar la prueba que no tiene

que ver con sus conocimientos, así reduciendo la exactitud de la predicción respecto del desempeño universitario. **Finalmente, a la luz de la presencia internacional de la corrección por adivinación en los programas de selección universitarios, el equipo de evaluación internacional recomienda abandonar la práctica de la corrección por adivinación para las administraciones futuras.**

81. **Recomendamos considerar la teoría de respuesta al ítem como un enfoque alternativo para tratar el comportamiento de adivinación del postulante.** Los enfoques actuales de puntaje de la PSU usan los procesos de corrección por adivinación que se han encontrado con graves limitaciones en la literatura. Como se ha mencionado anteriormente, el proceso le agrega capas de complejidad a múltiples aspectos de los procesos tal como la calibración de las respuestas a las pruebas piloto, el cálculo de estadísticas de ítems y estadísticas de puntajes en las pruebas. El marco de la teoría de respuesta al ítem hace que las características incorporadas den cuenta de la cantidad de la adivinación (esto es, pseudo adivinación) presente en las respuestas del examinado. **También recomendamos que, en preparación de una transición desde el contexto de la corrección por adivinación, si así se decidiera, el DEMRE prepare y someta un plan de transición a un grupo externo de revisores expertos.** El plan, entre otros aspectos, debería involucrar un análisis de riesgo y de factibilidad y un marco temporal hacia la introducción de los cambios necesarios de procesos críticos del programa de pruebas de la selección universitaria, tales como la construcción de la prueba de la PSU, el banco de ítems de la PSU, los pilotos de la PSU, mantención de la escala de la PSU, validez y confiabilidad de la PSU, informes de puntajes de la PSU. **El equipo evaluador internacional también recomienda una serie de estudios retrospectivos para evaluar cualquier efecto potencial sobre las tendencias históricas de la PSU y las estadísticas en terreno del banco de ítems; por ejemplo, las decisiones tomadas en el pasado debido al uso de asignación de puntajes por fórmulas.**

82. Debido a la inclusión del promedio de notas estandarizado de la enseñanza media (NEM) en el puntaje de postulación, el equipo evaluador recomienda la evaluación de sus datos normativos. Las tablas de conversión del NEM se utilizaron por primera vez en el proceso de selección de 2003. Debido a que el NEM es uno de los dos elementos que definen el puntaje de postulación para la mayoría de las carreras del CRUCH y las ocho universidades privadas afiliadas, recomendamos profundizar la información disponible sobre la estructura del NEM y el proceso para calcularlo. En esta misma línea, recomendamos estudiar la validez de las inferencias extraídas a partir del conjunto de datos normativos que ya tiene diez años y que está basado en un currículum nacional que casi ha sido reemplazado por el currículum nacional actual. Según estas líneas de investigación y documentación, el equipo evaluador internacional recomienda estudiar la validez de los puntajes NEM estandarizados. De especial significado es la comparabilidad de sentido de los puntajes NEM en las pruebas. Los puntajes NEM se basan en la generalización de prácticas de calificación desconocidas en los distintos tipos de instituciones educativas (Privadas, Subvencionadas y Municipales) y modalidades curriculares (Científico-Humanista y Técnico-Profesional). El equipo evaluador internacional propone estudiar la posibilidad de reemplazar el NEM con mediciones estandarizadas de desempeño académico en la enseñanza media tales como puntajes a partir de pruebas administradas nacionalmente. Es de crucial importancia equilibrar apropiadamente los marcos de la prueba de la PSU y su referencia al currículum nacional de Chile para evitar sobre enfatizar medidas del desempeño académico de la enseñanza media mientras sacrifica medidas de aptitudes escolares generales.

83. **Recomendamos revisar completamente el sistema de puntajes de postulación usando la perspectiva de puntuaciones compuestas.** Los procedimientos para calcular las puntuaciones compuestas se encuentran bien documentados en la literatura, y hay diversos métodos disponibles para los practicantes (Feldt & Brennan, 1989; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). Estos esfuerzos deberían ser instituidos para suministrar información precisa a esas regiones de la escala de puntajes de postulación donde son tomadas decisiones importantes (por ejemplo, decisiones de selección), mientras se reconocen las ponderaciones compuestas diferenciales usadas por las universidades y sus carreras.
84. **Recomendamos introducir la precisión de la medición de puntajes estandarizados informados para las pruebas individuales dentro del sistema de puntaje y postulación de la PSU.** El Programa de la PSU debería desarrollar pautas sobre los usos pretendidos e interpretación de las escalas de puntaje de la PSU y puntajes estandarizados con un énfasis en la delineación de las limitaciones del uso e interpretación de puntajes derivados. Los esfuerzos deberían mantener en perspectiva las ponderaciones compuestas diferenciales usadas en las universidades y sus carreras.
85. Recomendamos proporcionar documentación técnica sobre la fijación de normas sobre puntajes NEM con un énfasis en las descripciones de las poblaciones destinatarias de la prueba de la PSU, procedimientos para muestras, tasas de participación, enfoques de ponderaciones (de ser usados), fechas de las pruebas e información descriptiva de variables de trasfondo.
86. **Recomendamos mantener una agenda de investigación para estudiar la estabilidad año tras año de las escalas primarias y secundarias de la PSU.**

Objetivo 1.1.i: 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 — 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

Descripción:

Este objetivo incluyó una revisión del proceso para la estimación de la confiabilidad del puntaje de la PSU a partir de marcos TCT y TRI para calcular los errores condicionales de las escalas de puntajes de la PSU. Los diversos aspectos investigados incluyeron el efecto de la política de corrección por adivinación así como también el significado de la confiabilidad de los puntajes en la PSU al ser usados para la selección y las becas.

Para este objetivo, el equipo evaluador también proporcionó una demostración del cálculo del Error Condicional de Medición (CSEM) de los puntajes de la PSU bajo el marco de la Teoría de Respuesta de Ítem (TRI).La estimación del Error Condicional de Medida permite estimar la magnitud del error de medición a lo largo de la escala de la prueba. Los desarrolladores de pruebas usan al CSEM para lograr un nivel de precisión más elevado en la medición para regiones específicas de la escala donde tienen lugar las decisiones educacionales más importantes. El no cumplimiento de esta meta indica que las pruebas tienen niveles de precisión más bajos en las áreas de la escala hacia donde se apuntó. Hasta esta fecha, los análisis de CSEM no han sido parte de los procesos psicométricos del DEMRE. En el futuro, este tipo de análisis debería agregarse al programa de pruebas de la PSU.

Método:

La información para este objetivo se obtuvo mediante análisis llevados a cabo por el equipo evaluador y mediante entrevistas con personas relevantes del DEMRE el 23 de marzo de 2012, incluyendo al director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección. El equipo evaluador también leyó el reporte de confiabilidad de la PSU (DEMRE, 2010b).

Este objetivo también provee una demostración del cómputo del Error Condicional de Medición (CSEM) de los puntajes de la prueba de la PSU bajo el marco de la Teoría de Respuesta al Ítem (TRI). El DEMRE brinda confiabilidad y medición del error estándar a partir de la Teoría Clásica de Pruebas. La demostración utilizó información del proceso de selección de la PSU del año 2012.

Hallazgos:

Es un principio sólido de desarrollo y uso de pruebas documentar la confiabilidad, el error de medición estándar, la medición de error estándar condicional en cada puntaje y su combinación dentro de un puntaje único compuesto. Al evaluar estimaciones de confiabilidad de los puntajes de la prueba PSU, es importante considerar el uso y la interpretación de los puntajes de esta prueba. 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. Si esas decisiones habrán de ser ejecutadas adecuadamente, deben tener en cuenta las características del proceso de selección universitaria y del proceso para otorgar becas. Aunque hay un conjunto de puntajes de corte que conducen las decisiones a lo largo de dos avenidas, la mayor parte de las decisiones tienden a realizarse en la región de la escala de puntajes de la prueba sobre el centro de la escala de la PSU (500 puntos). Cuando se usan puntajes de corte, la cantidad de información que la prueba produce debería ser maximizada un tanto en esos puntajes, particularmente cuando se están tomando decisiones importantes.

La dependencia de las medidas de consistencia internas (por ejemplo, coeficiente alfa) y el error de medición de la Teoría Clásica de Pruebas proporciona una cobertura parcial de lo que se espera de una prueba de alta relevancia en términos de estándares internacionales.

El proceso para estimar la confiabilidad se enfoca en informar estimaciones para las pruebas PSU individuales; sin embargo, las decisiones de selección se realizan con las ponderaciones de puntajes compuestos de puntajes individuales en la prueba PSU y del promedio de notas de la enseñanza media. El informe de confiabilidad de la PSU de DEMRE (DEMRE, 2010b) no proporciona una racionalidad para saltarse la información de confiabilidad del puntaje de selección universitario compuesto, que por último es el trozo de información sobre el cual las decisiones de selección universitaria son hechas. Además de la falta de una estimación de confiabilidad del compuesto de la selección universitaria, no hay trozos de información acerca de la medición de error para el puntaje compuesto o en las bandas alrededor de los percentiles asociados informados.

El proceso para estimar la confiabilidad involucra la formulación típica del coeficiente alfa para pruebas de selección múltiple calificadas sin la corrección por adivinación. En Chile, los puntajes de la prueba PSU se producen con una fórmula de asignación de puntaje que, en primer lugar, extrae de una respuesta de opción múltiple correcta un cuarto de punto de cada respuesta equivocada de opción múltiple y, en segundo lugar, deja sin afectar el

puntaje de número correcto cuando los postulantes omiten el ítem. Aún cuando el DEMRE se basa en el puntaje observado por la "corrección por adivinación" anterior al calcular los puntajes de la prueba PSU, los coeficientes de confiabilidad calculados de la PSU se basan en puntajes brutos de la PSU. Es interesante que los puntajes brutos no dan cuenta de la corrección por adivinación.

El alcance de las estimaciones de confiabilidad de la PSU que encontramos es limitado en cuanto a proporcionar una racionalidad para depender solamente del coeficiente alfa sin la corrección por adivinación e ignorar las medidas de precisión de las clasificaciones (aceptados vs. rechazados). Cuando los puntajes continuos son interpretados con respecto a uno o más puntajes de corte, el coeficiente alfa y la medición de error estándar puede producir información que puede no estar relacionada con la siguiente pregunta: "¿Cuán consistente es la clasificación aprobar/reprobar?" Debido a que el uso principal de los puntajes de la prueba PSU es el de tamizar entre postulantes a carreras universitarias que logran los puntajes de selección y los postulantes que no llegan a los puntajes de selección, la exactitud de las clasificaciones es un importante trozo de información que debería ser incluido como parte del informe, si la audiencia para tal informe ha de entender el grado de consistencia de decisión logrado. Tales decisiones de aprobar/reprobar se encuentran mejor informadas con enfoques de consistencia de clasificación y de exactitud de clasificación, los cuales son prácticas psicométricas estándares involucrando a una sola prueba.

Los estándares nacionales e internacionales recomiendan usar la medición del error estándar como un indicador para comparar grupos en vez de simples comparaciones entre estimaciones de confiabilidad. Es bien sabido que las estimaciones de confiabilidad son dependientes del grupo mientras que la medición del error no lo es. El informe de confiabilidad proporciona información acerca del error de medición para los postulantes típicos (medición de error estándar). Sin embargo, en lo que se refiere a las mediciones mismas, sería mejor considerar la precisión de la medición. Técnicamente esto es el inverso de la varianza de error de las mediciones individuales. En la Teoría Clásica de Pruebas la medición del error estándar supone que el error es el mismo sobre toda la escala de evaluación. Es más realista suponer que los errores estándares son menores a niveles de pericia donde una gran cantidad de ítems están concentrados. Esto implica que la precisión se concentra en el medio de una distribución de pericia y más bajan las colas de distribución donde se encuentran relativamente pocos ítems. Cuando los errores estándar son trazados en relación a la pericia, esto produce una curva en forma de U.

Respecto de los exámenes de selección universitarios, es importante que el centro de esta curva en U esté posicionado sobre el rango en la distribución de pericia donde las decisiones de selección son probables de ser tomadas. Uno puede pensar de esta posición como un cierto rango de percentiles dentro de la población de examinados. Agregar información histórica sobre los porcentajes de examinados seleccionados a carreras a la medida de la precisión de los puntajes de la prueba PSU, las universidades y sus carreras podrían incluir información adicional para orientar sus decisiones de selección.

Debido a la importancia del puntaje de postulación sobre el proceso de selección, el hecho de que la escala de puntajes NEM y la escala de puntaje de la PSU difieren en su dispersión, genera preocupación. Con el NEM reportado en una escala con menor dispersión (por ejemplo, una desviación estándar de 100), la varianza de NEM en el puntaje de postulación sería pequeña y esta decisión de escala entraría dentro de la varianza del puntaje de postulación y su confiabilidad. Los puntajes de postulación son compuestos ponderados que involucran puntajes de la prueba PSU y del NEM y conjuntos de ponderaciones adoptadas. La varianza de los puntajes es una función de (1) las varianzas de los puntajes de la PSU y de los puntajes del NEM multiplicadas por el cuadrado de la ponderación para cada uno de

los puntajes, respectivamente y (2) la covarianza ponderada de los puntajes de la PSU y NEM. El equipo evaluador internacional recomienda abordar la escala del NEM para administraciones futuras y al computar normas actuales para tal variable. Una decisión que consideramos razonable es la de establecer la escala NEM con los mismos parámetros de la escala PSU (por ejemplo, promedio y desviación estándar).

Recomendaciones:

87. El informe de confiabilidad de la PSU de DEMRE (DEMRE, 2010b) es limitado en cuanto a proporcionar justificación para los enfoques de estimación de la confiabilidad que se han utilizado. En los análisis informados, el coeficiente alfa se implementó para estimar la confiabilidad. El coeficiente alfa muestra fuentes específicas de error de medición relevantes a algún tipo de decisiones. Específicamente, el coeficiente alfa indica error de medición debido a error de medición asociado a ítems.

88. La discusión de fuentes sistemáticas plausibles de errores en los puntajes de la PSU también está ausente en el informe de confiabilidad de la PSU. Al informe le falta discusión acerca de los efectos y del tratamiento para acomodar la corrección por adivinación y la omisión. Los efectos de estas dos condiciones merecen mayor estudio. Se notaron desafíos similares para la estimación de la medición de error estándar, que dependió de la desviación estándar de puntajes de pruebas brutas no corregidas.

89. **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).** Según se aprecia en nuestra demostración, la Teoría de Respuesta al Ítem provee un marco para examinar de manera valiosa la magnitud del Error Estándar de Medida a lo largo de la escala PSU. **Recomendamos que estos análisis sean agregados, en el futuro, al programa de la PSU.**

90. **El informe de confiabilidad de la PSU es limitado en cuanto a que no proporciona la descripción de la cantidad del error típico de medición para las regiones críticas de la escala de la PSU en las que se llevan a cabo las decisiones de admisión y otorgamiento de becas (por ejemplo, seleccionado/rechazado y becado/no becado).** Las medidas de consistencia/exactitud son importantes trozos de información actualmente ausentes de la estimación de confiabilidad y precisión del puntaje de la PSU. **Adicionalmente, los procesos existentes no explican la precisión de los puntajes de la PSU para las decisiones primarias y otras (por ejemplo, la entrada a la universidad y becas, respectivamente). *The Standards for Psychological and Educational Testing* (AERA, APA, NCME, 1999) sugieren informar precisión en los puntajes de la escala de la cual se toman las decisiones.**

91. Recomendamos abordar el tema de una escala para el NEM para futuras administraciones y computar las actuales normas para dicha variable. Una decisión podría ser la de establecer la escala de NEM de tal manera que no solamente reduzca la posibilidad de llevar a cabo interpretaciones erróneas de los puntajes, sino que también mantenga la contribución del NEM en el proceso de ponderación establecido por el CRUCH.

Objetivo 1.1.j: Propone un modelo para el tratamiento de los puntos de corte para los beneficios sociales, desde la perspectiva de la Teoría Clásica de los Tests (TCT) así como también de la Teoría de Respuesta al Ítem (TRI)

Descripción:

Este objetivo incluyó proponer un modelo para derivar puntajes de corte de la PSU para los beneficios sociales.

Método:

Basado en su amplia experiencia con el conjunto de puntajes de corte para evaluaciones educacionales de elevada importancia, el equipo evaluador revisó las circunstancias actuales en las cuales los beneficios sociales son asignados en Chile con respecto a la PSU y de ahí recomendó un método (el método Hofstee) que podría ser seguido por las partes responsables de tales distribuciones de beneficios sociales en Chile.

Hallazgos:

No hay hallazgos específicos para el Objetivo 1.1.j, solamente las recomendaciones a continuación.

Recomendaciones:

92. **Con respecto a proponer un enfoque para definir los puntajes de corte en la escala de la PSU para otorgar beneficios sociales en la forma de becas, recomendamos un enfoque que considere el manejo del dominio de la PSU y las consecuencias sociales.** Lo anterior apunta a identificar el nivel de conocimientos, habilidades y destrezas definidas por un panel de profesores universitarios y formuladores de políticas del MINEDUC y la de traducir tal definición en un puntaje de selección de la PSU. El último toma en cuenta consideraciones de políticas, consecuencias sociales y datos históricos para afinar el puntaje de corte. Un foco sobre las consecuencias sociales implicaría formuladores de políticas del MINEDUC convocando un panel para el uso de dicha información como la información histórica acerca del número de becas disponibles cada año, el número de estudiantes que recibe becas, su desempeño, sus tasas de deserción y sus tasas de graduación.
93. **El método específico que recomendamos para fijar el puntaje de corte para fines de becas es el método Hofstee.** El método Hofstee es un ejemplo de un método fijador de acuerdos de estándares. Hofstee (1983) acuñó este término cuando desarrolló su método para capturar la naturaleza dual de fijaciones de estándares. Esto es, aún los juicios fijadores de estándares referenciados a criterios son moderados por expectativas referenciadas a normas. Su método hace uso explícito de la información referenciada a criterios y a normas para derivar puntajes de corte. El término "acuerdo" no connota la disminución o dilución de cualquiera de los dos criterios de juicio; más bien, significa la integración de ambos de una manera que es más multifacética que la aplicación individual de ambos criterios. Esto es importante en el otorgamiento de becas porque hay recursos finitos para distribuir y la selectividad referenciada a normas apunta esos recursos a estudiantes de una manera que debería ser considerada junto al dominio del área temática referenciada por criterio. Aunque el enfoque no está exento de críticas, así como ocurre con cualquier otro proceso fijador de estándares, creemos que fijar un puntaje de corte para otorgar becas requiere de una mezcla de consideraciones: algunas escolásticas, algunas monetarias y algunas sociales. La virtud del enfoque de Hofstee depende de

su capacidad de tomar todas estas perspectivas en cuenta para facilitar las discusiones y decisiones de acuerdos dentro de un marco temporal razonable. El enfoque es lo suficientemente flexible para ser utilizado con el enfoque actual al escalamiento de la PSU (esto es, usando la Teoría Clásica de Pruebas) así como también cualquier cambio predecible al sistema (esto es, usando la Teoría de Respuesta al Ítem).

Objetivo 1.2: Análisis de la suficiencia de un puntaje único para la prueba de Ciencias y de los procedimientos para calcular dicho puntaje, considerando que esta prueba incluye bloques electivos de Biología, Física y Química

Descripción:

Este objetivo incluyó un análisis del proceso usado para derivar un puntaje único para Ciencias de la PSU. Para este objetivo, el equipo evaluador proporcionó un análisis de la pertinencia del puntaje único para Ciencias de la PSU. El equipo evaluador llevó a cabo análisis demostrativos para abordar dos preguntas fundamentales:¿Cuán razonable es el informar un puntaje único? y ¿Cuáles son las alternativas? El equipo evaluador también discutió el tema de la dimensionalidad de los puntajes de las pruebas.

Método:

El equipo evaluador leyó la documentación del Comité Técnico Asesor del proceso de "equiparación" de Ciencias de la PSU y se reunió con personal del DEMRE para obtener información adicional acerca del proceso para calcular el puntaje único de Ciencias. Las entrevistas con las personas relevantes interesadas del DEMRE tuvieron lugar el 26 de marzo de 2012, e incluyeron al director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, así como también el jefe del proceso de selección.

Hallazgos:

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 proceso que se está siguiendo para lograr un puntaje único de Ciencias no es equiparación en el sentido estricto definido por Kolen y Brennan (2004) porque los examinados toman pruebas con contenidos diferentes basado en las secciones opcionales (módulos alternativos). Debido a que estos módulos alternativos traen diferencias de contenido, los puntajes de los estudiantes que toman los diferentes módulos opcionales no pueden ser considerados como equiparados. Sin embargo, tales puntajes pueden ser referidos como "enlazados", y el proceso seguido se lo puede referir como "enlace". La terminología asociada con el puntaje único en Ciencias necesita cambiarse desde "equiparación" a "enlace".

El proceso empleado actualmente para desarrollar un puntaje para cada examinado involucra enlazar el puntaje de cada sección opcional con el puntaje en la porción común usando un método de regresión no lineal. El puntaje único es la suma del puntaje en la porción común y el puntaje enlazado en la sección opcional. Se usan regresiones no lineales separadas para las tres secciones opcionales para derivar la porción de puntaje enlazado del puntaje único. Este proceso está descrito en detalle en la documentación técnica.

El procedimiento de regresión no lineal usa un conjunto de nodos fijos (puntajes de la porción común) de las pruebas y encuentra una regresión no lineal de puntajes opcionales para los puntajes de la sección común. El procedimiento de regresión no lineal parece calzar con una relación lineal entre nodos y parece resultar en una función lineal por tramos.

No se da ninguna racionalidad para la elección particular de nodos en el procedimiento de regresión. Además, no se entrega ninguna racionalidad para el uso de lo que aparentemente es una función de regresión lineal por tramos. Los conjuntos alternados de nodos probablemente conduzcan hacia diferentes resultados de enlaces. Además, un procedimiento de regresión de *splines* cúbicos (es decir, *splines* suavizados) es probable que sería una mejora con respecto al procedimiento usado aquí porque las *splines* cúbicos (ver Kolen y Brennan, 2004, respecto de una discusión acerca del uso de *splines* cúbicos en la equiparación) producen una función de regresión curvilineal continua y hay criterios en la literatura para la selección de los nodos y de los parámetros del suavizado.

No se proporciona una racionalidad estadística respecto del cálculo de un puntaje único mediante la suma de los puntajes en la sección común y los puntajes de la sección opcional enlazada. La correlación entre la porción común y la porción opcional tendrá un efecto sustancial sobre la variabilidad de los puntajes totales. Para estas pruebas, las correlaciones entre la porción común y las porciones opcionales eran casi iguales para las tres secciones opcionales, por lo que la variabilidad de los puntajes totales ha sido probablemente similar. Sin embargo, si en algún punto en el tiempo estas correlaciones hubieren de diferir sustancialmente, la variabilidad de los puntajes sumados podría ser bastante diferente respecto de los examinados que tomen diferentes secciones opcionales.

Los criterios estadísticos usados no están declarados respecto del procedimiento, local hace surgir las siguientes preguntas: ¿Qué se pretende que el método logre desde una perspectiva estadística o psicométrica? ¿Qué supuestos estadísticos o psicométricos se están haciendo?

De los análisis presentados, es difícil verificar la extensión de la comparabilidad de los puntajes totales respecto de los estudiantes que toman módulos diferentes. Basados en la forma en que el procedimiento es implementado, parece ser que el puntaje único para un estudiante que tomó un módulo opcional de Biología es considerado comparable con un estudiante que tomó un módulo opcional de Química o Física. La racionalidad para esta comparación no está clara.

Sería informativo regresar variables de resultados tales como promedio de notas universitarias (en cursos de ciencias universitarios comparables) en el puntaje único de Ciencias para los grupos que toman los módulos opcionales de Biología, Química y Física. Si los puntajes totales son comparables, estas tres regresiones deberían tener como resultado coeficientes de regresión aproximadamente iguales.

Recomendaciones:

94. El equipo evaluador internacional de la PSU considera que la racionalidad del enlazado del puntaje en la prueba PSU de Ciencias cae por debajo de los estándares internacionales. Este proceso no solamente está etiquetado de forma equivocada, sino que también la documentación está incompleta y la evidencia de la adecuación técnica es insuficiente para decisiones de impacto elevado. **El equipo evaluador recomienda 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.** Cada una de estas

pruebas sería informada en diferentes escalas de la PSU, siguiendo procesos estándares ya disponibles para las pruebas de la PSU. Más aún, una vez que el engorroso proceso de enlace actual haya sido reemplazado, la mantención año tras año de las nuevas escalas de Ciencias de la PSU mediante la equiparación sería más rigurosa y, por lo tanto, más defendibles.

95. **Hasta que nuestra recomendación pueda ser implementada, hay una necesidad de más documentación para las actuales pruebas de Ciencias que informa al público y a los revisores técnicos, acerca de la decisión de políticas actuales de reportar un puntaje único para Ciencias, su racionalidad y la investigación que informó dicha decisión.** La práctica de enlazar los puntajes de las pruebas de diferentes áreas de contenidos se ha llevado a cabo para lograr comparabilidad mediante el escalamiento. Esta forma de enlazado es débil en comparación a la equiparación. Por ese motivo, debería entregarse evidencia de la generalización de las tablas de conversión para los subgrupos, ocasiones y pruebas. **Otras recomendaciones respecto de las pruebas actuales de Ciencias de la PSU son las siguientes:**

    a. **Referirse al proceso como "enlazado" más que "equiparación."**

    b. **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.**

    c. **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.** Comparar meticulosamente los resultados para todos los métodos considerados. Proporcionar criterios estadísticos y psicométricos que indican qué es lo que el procedimiento pretende lograr.

    d. Con los actuales métodos de enlace hay un supuesto implícito que los puntajes totales son, de alguna manera, equivalentes sin tener en cuenta el módulo opcional tomado. Este supuesto, que parece ser irrealista, necesita ser investigado meticulosamente. **Para cualquier procedimiento considerado, incluyendo estos estándares (por ejemplo, equipercentil encadenado y equipercentil de estimación de frecuencia), es importante verificar acerca de la comparabilidad de puntajes.** Variables de resultados de regresiones tales como promedios de notas universitarias sobre puntajes totales para los grupos que toman los módulos opcionales de Biología, Química y Física. Si los puntajes totales son comparables, estas tres regresiones deberían ser aproximadamente iguales.

    e. **Estimar la confiabilidad, las mediciones de error estándar y las mediciones de errores estándar condicionales usando procedimientos que han sido desarrollados para puntajes compuestos. Calcular errores estándar de los puntajes enlazados usando procedimientos de muestreo probabilístico con reemplazo (*bootstrap*).**

    f. **Documentar procesos que describen el aseguramiento de la calidad y verificaciones de calidad para enlaces de puntajes de las pruebas de Ciencias de la PSU.**

Objetivo 1.3: Evaluación de modelos TRI para la calibración de ítems, el desarrollo de pruebas y los propósitos de equiparación

Descripción:

Este objetivo incluye una evaluación de los métodos TRI para calibrar ítems y del prospecto de que las sucesivas formas de la PSU puedan ser equiparados algún día.

Método:

El equipo evaluador leyó la documentación del DEMRE e información relacionada acerca del proceso de calibración de la PSU. Se obtuvo información adicional para este objetivo mediante entrevistas con personas relevantes en el DEMRE el 26 de marzo de 2012, e incluyó al director del DEMRE, al coordinador general, el jefe de la unidad de investigación y su equipo, también como al jefe del proceso de selección. En la preparación de la reunión, el equipo evaluador creó una matriz de datos sintética con 3,000 simulaciones y 50 ítems con propiedades de parámetros conocidas usando una metodología de simulación estándar.

Hallazgos:

Luego de considerar todas las facetas y elementos inspeccionados para el Objetivo 1.3, 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; por ejemplo, el equipo evaluador se dio cuenta del uso equivocado de la terminología de equiparación (es decir, preguntas de anclaje, bondad de ajuste del modelo) y corrigió su comprensión original. Específicamente, el programa de pruebas de la PSU no mantiene una escala de información sobre una base anual ni calibra los ítems piloto en una escala común.

Una descripción clara de un proceso para calibrar ítems pilotados es una práctica importante que se encuentra a menudo en programas de pruebas maduros que involucran evaluaciones de alto impacto. Al desarrollar y evaluar ítems de prueba usando la Teoría de Respuesta al Ítem (TRI), los practicantes diseñan procesos de manera tal que los parámetros de ítems piloto puedan compartir una escala común con los parámetros de ítems piloto de los bancos de ítems. Producir estimaciones de parámetros de ítems pilotados dentro de una escala común permite comparaciones de "manzanas con manzanas" de los parámetros de ítems y la información TRI relacionada, ambos entre formas piloto múltiples en cualquier año dado y a lo largo de años de administraciones operacionales.

En el contexto del proceso de selección universitario de Chile, el equipo evaluador recomienda corregir las imprecisiones presentes en la documentación de la PSU para equiparación piloto. El proceso existente para desarrollar un conjunto ancla es descrito en documentos oficiales como no convencional y no refleja los estándares internacionales; por ejemplo, el largo del conjunto ancla está bajo la relación estándar entre conjuntos ancla y largo de la prueba. La relación mencionada en la documentación para la PSU no llega al mínimo del 25% del largo total de la prueba (Kolen & Brennan, 2004). Además, los conjuntos ancla están construidos sin pautas sobre cómo lograr la representación de contenido. Como resultado, estos conjuntos ancla quedan cortos como para representar completamente y en forma precisa las características totales de la prueba.

Hay varias maneras en que el programa de pruebas de la PSU no logra un nivel de análisis generalmente esperado al equiparar una evaluación de impacto elevado. La irregularidad

más notable fue la información errada que se encontró en la documentación del programa de evaluación: específicamente, la información en cuanto a la equiparación año tras año y las calibraciones piloto. El equipo evaluador internacional reitera que estas actividades no están teniendo lugar, aunque la documentación provista pareciera indicar que sí lo están llevando a cabo. El equipo evaluador enfatiza la necesidad de que las actividades de equiparación sean llevadas a cabo con respecto a las pruebas de la PSU.

Durante la revisión de la documentación y las entrevistas, el equipo evaluador se dio cuenta del uso suelto de la terminología de equiparación y de los procesos. En el contexto de la equiparación del programa de pruebas de la PSU hay una discrepancia entre lo que se encuentra documentado y lo que es practicado. A partir de las entrevistas, es evidente cuál es el propósito del así llamado conjunto ancla en el programa; por ejemplo, el término "conjunto ancla" está siendo usado para referirse a un grupo de ítems que se agregan a una forma piloto con un propósito diferente a la calibración de las administraciones piloto a una escala común.

Hay un gran inconveniente en el proceso de calibración de los ítems de la PSU actual: las calibraciones año a año no están referenciadas a una escala común. La ausencia de dicho esfuerzo de calibración crea preocupación acerca de la comparabilidad de parámetros de ítems, de las actividades de construcción de pruebas asociadas y de la calidad de la información de los bancos de ítems. También están ausentes del proceso de calibración de ítems de la PSU los análisis de dimensionalidad. Sería encomiable agregar este tipo de análisis para verificar los supuestos de modelo.

Los ítems ancla son esencialmente inútiles, ya que no cumplen con su propósito. Aún en la etapa de diseño, el conjunto ancla muestra múltiples defectos. El contenido del conjunto ancla y las características estadísticas son deficientes a la luz de los estándares internacionales. Las anclas son demasiado breves y subrepresentan el contexto total de la prueba. La implementación de tales especificaciones anclas, de ser obedecidas, rendirían resultados de sesgo equiparado y un gran error de equiparación. También el diseño del conjunto ancla muestra falta de información respecto de las acciones tomadas, por lo menos desde una perspectiva de diseño, para disminuir los efectos de contexto (es decir, reteniendo la posición de ítem de los conjuntos ancla) sobre el desempeño de los ítems ancla y de los procesos para llevar a cabo verificaciones en la deriva de parámetros de los ítems ancla. Para minimizar los efectos de contexto, los ítems ancla deberían retener su posición en las pruebas. La falta de planes para incorporar la selección de derivación de parámetros de ítems presenta amenazas potenciales a la precisión de la equiparación.

Recomendaciones:

El equipo evaluador internacional considera que la equiparación de la PSU se encuentra bajo los estándares internacionales. La documentación respecto de la equiparación de la PSU no es solamente poco clara, sino que también es incompleta e imprecisa en algunas áreas. Respecto de un examen nacional de selección de impacto elevado para miles de postulantes, la adecuación técnica es insuficiente, lo cual significa que pueden ocurrir resultados equivocados (por ejemplo, decisiones). El equipo evaluador propone un reenfoque de los esfuerzos que abordan las siguientes mejoras recomendadas.

96. A fin de reponer el banco de ítems en la medida que se crean nuevas pruebas cada año, 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. La calibración de los parámetros de ítems de prueba en terreno puede llevarse a cabo con enfoques revisados por Kolen y Brennan (2004).

97. **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.** Una equiparación estadística simplemente establece la relación entre dos formas de prueba. Típicamente, esto se logra mediante el uso de un elemento en común a lo ancho de las administraciones de la prueba —ya sea personas comunes o ítems comunes. En algunos casos, donde sea apropiado, se puede hacer un supuesto de que dos grupos separados, tomando dos formas de prueba separados, son equivalentes al azar. En la mayoría de los contextos de pruebas de selección universitaria —donde la meta es equiparar las formas de pruebas año tras año— un diseño de personas en común no es factible típicamente. Kolen y Brennan (2004) tienen un amplio tratamiento de los enfoques alternativos para conducir equiparación de pruebas que pueden ser consultados. Es importante enfatizar que la equiparación de pruebas no es la solución a los problemas de construcción de pruebas. El proceso de construcción de pruebas apunta a desarrollar una forma de pruebas que sea equivalente en contenido y dificultad a otras formas administradas en años anteriores. La equiparación es una herramienta que compensa las diferencias en dificultades de la prueba que no podrían haber sido controladas durante la construcción de pruebas.

98. **Recomendamos que la PSU equipare formatos de pruebas a través de administraciones.** La falta de puntajes equiparados socava la habilidad de desarrollar evaluaciones que son justas para los examinados. La ecuanimidad podría estar en riesgo cuando los estudiantes que rinden la prueba de la PSU en el año 1 tienen ventajas con respecto a aquellos que rindieron otra prueba de la PSU en un año a continuación. Para que una evaluación sea considerada justa, los puntajes de las pruebas no deberían depender de la forma de prueba en particular rendida. En Chile, los puntajes de la prueba PSU se pueden utilizar hasta dos años consecutivos como parte del proceso de selección. La equivalencia entre los puntajes de la PSU entre formas es una condición necesaria para respaldar dicho uso emergente.

99. 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.

100. El modelo 2PL es actualmente utilizado para el análisis de ítems en el programa de la PSU sin mayor racionalidad. **Si este modelo continuará siendo usado en el futuro para análisis de ítems o para otros propósitos, el equipo evaluador recomienda encarecidamente seguir la práctica internacional validando la pertinencia de su uso por sobre las alternativas típicamente usadas del modelo de Rasch o 3PL.**Similarmente, el equipo evaluador recomienda desarrollar documentación de calibración de ítems que se encuentren disponibles para el personal que participe de la calibración de ítems. Tal documentación debería ser empleada para capacitar al personal en los procesos de calibración de ítems de la PSU que hayan sido aprobados por el DEMRE y por el comité técnico asesor (CTA).

Objetivo 1.4: Evaluación de software y de procesos usados en el análisis estadístico y banco de ítems

Descripción:

Este objetivo incluyó una evaluación del software usado para análisis de ítems y para el banco de ítems de la PSU. El equipo evaluador verificó si el programa proporciona el nivel de seguridad adecuado, así como un nivel adecuado de control de versiones de los ítems, del estado de los ítems y de los datos psicométricos asociados.

Método:

El equipo evaluador analizó los paquetes de software usados por el DEMRE para el banco de ítems y el análisis estadístico. Se obtuvo información adicional para este objetivo mediante entrevistas con personal relevante interesado en el DEMRE el 26 de marzo de 2012, e incluyó al director del DEMRE, el coordinador general, el jefe de la unidad de investigación y su equipo, también como al jefe del proceso de selección.

Hallazgos:

La base de datos del banco de ítems parece estar bien diseñada con respecto a la seguridad de los ítems. Las limitaciones impuestas a los usuarios minimizan la posibilidad de brechas de seguridad para los ítems operacionales y las formas; por ejemplo, solamente los autores de pruebas pueden ver los ítems y solamente pueden ver los ítems asociados a sus áreas temáticas, se requiere de llaves basadas en hardware para tener acceso a imágenes de ítems, los psicométricos no pueden ver ítems ni llaves, y los técnicos de IT, incluyendo los administradores de bases de datos, tampoco pueden ver ítems ni llaves. Hay opciones para que los usuarios autorizados exporten las imágenes de ítems a archivos*.DOC que se guardan en sus equipos locales, por lo que la seguridad de las pruebas debería incluir la seguridad de los computadores de esos usuarios autorizados. Estas medidas podrían incluir, por ejemplo, encriptación de discos duros de los equipos, habilitación de contraseñas para los atenuadores de pantalla y delimitación de las redes de los equipos a una red de área local interno solamente.

Está presente el control de versiones de imágenes de ítems, pero parece estar con base débil según la documentación disponible. Solamente los usuarios autorizados pueden hacer cambios, pero no parece que estos cambios sean registrados de alguna manera o que las versiones anteriores de los ítems sean retenidas. Los ítems están bloqueados en ciertas etapas del ciclo de desarrollo y empleo de ítems.

No hay documentación del control de versión de las estadísticas de ítems más allá de un indicador del estado que muestra cuántas veces ha sido usado un ítem operacionalmente. Parece ser que solamente las estadísticas de la administración más reciente están retenidas en la base de datos.

Los trozos de software BILOG 3.11 y DIFAS ambos usan procedimientos estadísticos bien investigados y reconocidos para estimar parámetros TRI y estadísticas de DIF de Mantel-Haenszel respectivamente. El BILOG 3.11 está limitado a ítems dicotómicos el cual es el formato actualmente en uso para las pruebas de la PSU. Tal como hemos declarado anteriormente, el DEMRE depende de la versión 3.11 versión del software BILOG 3.11. El equipo evaluador recomienda desarrollar un plan de actualización para las versiones de software. Puede que sea necesario reemplazar el software BILOG 3.11 por otro, dependiendo de decisiones futuras. Si en el futuro se decide incluir preguntas de ensayo en

la prueba de la PSU, el BILOG 3.11 quedaría corto al manejar este tipo de formato de ítem. El software disponible comercialmente que permite este tipo de preguntas (por ejemplo, MULTILOG) puede ser considerado y evaluado. Asimismo, si en el futuro se decide incluir la equiparación TRI, permitiendo la estimación de parámetros de ítems para más de un grupo, el software BILOG 3.11 podría ser reemplazado por software que permitiera tal tipo de análisis (por ejemplo, BILOG MG).

El DIFAS acomoda a ítems multipunto, pero solamente provee medidas estadísticas de DIF sin banderas heurísticas para tales ítems. El banco de ítems necesitaría ser ampliado para proveer dichas banderas en el caso de que ítems multipunto sean agregados a las pruebas en algún punto del futuro. Además, el equipo evaluador recomienda desarrollar un plan para la actualización de la versión del software. (Nota: se puede encontrar más información sobre los desafíos al involucrar metodología DIF múltiples como parte del Objetivo 1.1.g. de esta evaluación).

El SAS es un sistema robusto y está bien adecuado para su uso para los análisis de equiparación de Ciencias. El código mismo del SAS tiene suficiente y adecuada documentación, y están siendo usados SAS PROCs adecuados.

Recomendaciones:

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. La evaluación indica que aunque puede que haya justo la automatización suficiente *dentro* de un grupo funcional, no hay suficiente *entre* grupos funcionales. Respecto de un examen de selección nacional con una administración operacional única de seis evaluaciones, el ambiente de procesamiento puede ser tolerable. Sin embargo, el programa de pruebas de la PSU tendría un desafío si se le exigiera permitir administraciones múltiples de pruebas, algo que ocurre regularmente en muchos programas de selección universitaria internacionalmente. El equipo evaluador propone reenfocar los esfuerzos que abordarían las siguientes mejoras recomendadas.

101. Los sistemas actuales parecen estar un tanto dislocados, requiriendo de mucha manipulación manual de ítems y de datos de pruebas. Uno de los pasos de verificación durante la construcción de pruebas es chequear los códigos de los ítems para verificar que existan dentro de la base de datos, implicando que el autor de la prueba debe digitar los códigos de ítem manualmente. Este es un lugar donde un error puede ocurrir si el autor de la prueba digita mal un código de ítem y el código equivocado coincide con otro ítem del banco (no deseado). El código SAS, usado para implementar la equiparación de la prueba de Ciencias, parece requerir edición manual para cada año sucesivo. Hay muchas referencias de la "exportación" e "importación" de datos hacia y desde la base de datos; sin embargo, en la medida que estas funciones requieren manipulación manual, estos son pasos en el proceso donde pueden ocurrir errores. **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.**

102. **Las versiones de ambas imágenes de ítems y estadísticas de ítems pueden mejorarse.** Las ID de usuarios que efectúan modificaciones a los ítems deberían ser rastreables. Además, las versiones anteriores deberían ser conservadas, ambos para proporcionar un registro histórico de los cambios efectuados a un ítem y también como un resguardo para permitir revertir a una versión anterior si se lo necesita.

103. **Deberían conservarse las estadísticas de todas las administraciones de un ítem, y los usuarios deberían ser capaces de verlos juntos en un orden cronológico.** Los grandes cambios en las estadísticas de ítems desde una administración a otra pueden indicar cosas tales como exposición de ítems, errores de impresión u otros problemas. La documentación indica que se efectúan análisis de verificación de claves para ítems piloteados; se recomiendan llevar a cabo los análisis de las claves para todos los ítems. Esta medida permitirá controlar cambios en los ítems que no hayan sido registrados, detectar errores de impresión o de producción, errores en la importación y exportación de datos hacia y desde la base de datos, así como otros errores no predichos.

104. El software BILOG 3.11 puede correr en el modo por lotes, y se pueden producir archivos de comando automáticamente mediante software escritos a la medida o por medio de programas SAS; ambos son recomendados. El BILOG es un programa relativamente antiguo, y softwares más recientes podrían traer beneficios. Si el uso de TRI para la equiparación es posible en el futuro, podría valer la pena la actualización de softwares más recientes.

105. El DIFAS no permite el uso de archivos de comando y, por lo tanto, no puede correr en modo *batch*. Los análisis Mantel-Haenszel pueden codificarse fácilmente en software a la medida o pueden correr en SAS. El reemplazo de DIFAS con una solución que pueda automatizarse más fácilmente reduciría la necesidad de la mano de obra involucrada en correr DIFAS, guardando los resultados y luego importándolos a la base de datos.

106. El proceso general para producir los puntajes normalizados fue descrito en los documentos disponibles para esta revisión, pero los procedimientos y softwares específicos usados para lograr estos análisis no lo fueron. En general, las recomendaciones precedentes también se aplican a los procesos y softwares usados para la derivación de los puntajes normalizados. Hasta donde sea posible, estos procesos deberían estar automatizados a fin de que pueda correr sin la intervención de usuario manual, y debería usarse un software susceptible a correr en modo *batch* (tal como SAS) para estas derivaciones.

Objetivo 1.5: 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

Descripción:

Este objetivo incluyó una evaluación de la información del puntaje de la PSU que implicó el resumen y análisis de las respuestas de entrevistas obtenidas de parte de personas clave interesadas en la PSU (estudiantes, docentes de enseñanza media y oficiales de selección universitaria) con respecto a características particulares de los informes de puntaje de la PSU.

Método:

El equipo evaluador entrevistó a grupos de audiencias destinatarias de los informes para recolectar información acerca de si estas personas interesadas entendían la información incluida en los informes actuales y si ellos pensaban que podría haber alguna necesidad de mejoramiento de los informes y procesos. Las preguntas de la entrevista apuntó a tres grupos distintos basado en la información que reciben actualmente del DEMRE: estudiantes, profesores de enseñanza media y oficiales de selección universitaria.

Hallazgos:

El tema general a partir de las entrevistas con los estudiantes respecto de su reacción a la *Entrega del Informe de Resultados de la PSU* es que mientras ellos entienden la intención y el propósito básico del informe, ellos no pudieron recolectar mucha información útil a partir de él. Los estudiantes no creían que el informe contenía información que pudiera ayudarlos a contestar preguntas cuantitativas acerca de los puntajes en el informe. Muchos de los estudiantes deseaban una retroalimentación y un diagnóstico más específico del informe.

Con respecto al *Reporte de Admisión a las Universidades Chilenas*, los estudiantes no sintieron que el informe diera información relevante para ser capaces de responder a cualquier pregunta cuantitativa. Básicamente, ellos sintieron que este informe era de poca ayuda y que no suministraba suficiente información.

Las respuestas de los docentes a las preguntas que inquirían acerca de los usos y aplicaciones de los *Informes Estadísticos de la PSU* entregaron una variedad de respuestas. Algunos profesores sintieron que los informes podían utilizarse para proyectar cómo les iría a los estudiantes en el futuro, mientras que otros creían que podían usarse para medir cuán bien los estudiantes conocían el contenido al momento de la prueba. Las siguientes preguntas aplicadas a los profesores respecto de usar la información de los *Informes,* indicaron una dificultad general para encontrar la información apuntada. Dos solicitudes del panel fueron la distribución más veloz y más eficiente de los informes.

En general, los oficiales de selección universitaria indicaron que estaban "satisfechos" con el proceso actual usado para cada etapa específica del proceso. Estos sujetos dieron diversas calificaciones a la importancia de los distintos segmentos de información de los informes, pero en general expresaron un deseo de obtener esta información antes, debido a lo restringido de los plazos. Uno también mencionó que la información que se les daba era mucho mejor que la que recibían con anterioridad.

Recomendaciones:

107. **Debido a que los estudiantes no entendían las escalas de puntajes de la PSU que se les presentaron, recomendamos que el DEMRE provea información interpretativa adicional explicando los Informes de Entrega de Resultados de la PSU.**

108. **Aunque el DEMRE publica información sobre la ponderación y los puntajes de selección, recomendamos que esta información también debería estar cortada a la medida para su inclusión en cada informe para estudiantes.**

109. Recomendamos que los informes sean rediseñados para facilitar encontrar la información por parte de los estudiantes, tal como el número de vacantes disponibles en las diversas sedes universitarias.

110. **Recomendamos que 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 sea suspendido hasta que los resultados hayan sido escrutados cuidadosamente para asegurar la confiabilidad y validez de dicha información.**

111. **Debido a que los docentes indican que ellos usan los resultados de la prueba de la PSU para otros propósitos que la selección universitaria, recomendamos que los Informes Estadísticos de la PSU expliquen cuáles son los usos que se pretenden con las pruebas de la PSU y adviertan contra los usos no pretendidos.**

112. Aunque el informe contiene una gran cantidad de información, la gran mayoría muy valiosa, es difícil encontrar datos específicos para responder a preguntas específicas; por ejemplo, el apéndice del informe provee el número de estudiantes seleccionados a la universidad de su institución educativa en particular. Sin embargo, los docentes no pudieron ubicar esta información. Recomendamos incluir una tabla de contenidos detallada que mejoraría el valor de este informe, haciendo que esta información esté disponible más fácilmente para los docentes.

Objetivos de evaluación 2.1–2.4: Los estudios de validez de la PSU

Todos los objetivos de esta sección están relacionados con la validez de las pruebas de la PSU. La forma en que los ítems de pruebas de la PSU se relacionan unos con otros—la estructura interna de la prueba— es examinada en el Objetivo 2.1. Los análisis para determinar si el contenido de la prueba de la PSU refleja plenamente el dominio que está intentando evaluar —la validez de contenido— se entrega en el Objetivo 2.2. El Objetivo 2.3 considera la trayectoria de los puntajes de la prueba de la PSU en el tiempo. Finalmente, la habilidad de los puntajes de la PSU para predecir otras mediciones importantes tales como las calificaciones universitarias y las tasas de graduación es el tema de estudio del Objetivo 2.4.

Objetivo 2.1.Estructura interna de los exámenes de la PSU: bondad de ajuste de los puntajes de las pruebas de la PSU analizada con modelos de análisis de factor de ítems y teoría de respuesta de ítems

Descripción:

Al desarrollar pruebas y usos previstos para las pruebas, a menudo es útil desarrollar primero las definiciones conceptuales y luego encontrar maneras de definirlas operacionalmente. Cronbach y Meehl (1955) introdujeron el concepto de validez de constructo para estudiar si los puntajes de las pruebas son suficientes para recuperar componentes destacados en las definiciones conceptuales de las pruebas.

La investigación que se ha llevado a cabo para el programa PSU en el área de validez y dimensionalidad y su invariancia a lo ancho de las subpoblaciones es escasa. Hasta la fecha no se cuenta con estudios en los cuales la validez se haya establecido. Asimismo, modelos modernos de medición (análisis de factorial a nivel de la pregunta y la teoría de respuesta al ítem) no han sido incorporados para investigar las propiedades de los puntajes de las pruebas de la PSU. Finalmente, no se han llevado a cabo estudios de la invariancia de una estructura unidimensional entre las poblaciones que rinden la PSU.

El propósito de este estudio fue examinar la estructura interna de ítem de la batería de pruebas de la PSU con marcos analíticos de factor de ítems y de TRI, usando datos de la PSU del proceso de administración del año 2012.

Las siguientes preguntas de investigación orientaron a la investigación:

- ¿Cuál es la dimensionalidad de las pruebas de la PSU?
- ¿En qué medida la estructura del factor de la prueba de la PSU generaliza sobre las subpoblaciones que rinden la PSU tales como:
    - Género: masculino o femenino.
    - Regiones: Norte (códigos 1, 2, 3, 4, 15), Central (5, 13 [Metro]), o Sur (6, 7, 8, 9, 10, 11, 12, 14).
    - Status socioeconómico: cinco quintiles de la variable SES— el quintil A define el grupo más bajo; el quintil B define grupo promedio bajo; el quintil C define el grupo promedio; el quintil D define al grupo sobre el promedio; y el quintil E define el grupo superior. El SES se calculó usando información de los ingresos familiares de los postulantes y de la educación parental.
    - Modalidad Curricular: Científico-Humanista o Técnica-Profesional.
    - Tipo de financiamiento: ¿Privado, Subvencionado o Municipal?

Método:

Análisis factorial a nivel de la pregunta se condujo separadamente en todas las pruebas de la PSU. El propósito de los análisis fue determinar si cada una de las pruebas de la PSU abrumadoramente representa un factor subyacente único o si cualquiera de estas pruebas indica una evidencia consistente de una estructura multifactorial. La presencia de una fuerte dimensión para una prueba de la PSU contribuye evidencia para respaldar el enlace de los ítems de pruebas que subyacen una variable latente testeada (Lord & Novick, 1968).De otra forma, indica que los ítems no pueden escalarse a lo largo de una dimensión única, debilitando así el significado y el uso de los puntajes de las pruebas. Este análisis se llevó a cabo comparando los tamaños de los tres primeros valores propios (valores propios) estimados a partir de matrices de correlaciones tetracóricas.

El Funcionamiento Diferencial de Pruebas (DTF) se llevó a cabo para evaluar la invariancia de estructura de factor a lo ancho de subpoblaciones relevantes. El DTF es un proceso psicométrico para investigar la diferencia relativa sobre los puntajes totales de las pruebas entre dos grupos de examinados luego de dar cuenta de la diferencia en el desempeño de la prueba en los niveles de los grupos. El DTF le permite a los desarrolladores de pruebas identificar si es que la prueba favorece a un grupo sobre otro y brinda líneas de evidencia en apoyo (o rechazo) de la equidad de la prueba. Una meta al desarrollar pruebas es desarrollar mediciones que funcionen igual de bien para todos los grupos. El no cumplimiento de esta meta constituye evidencia de validez en contra de la prueba. Hasta la fecha de este escrito, los análisis de DTF no han formado parte de los procesos psicométricos del DEMRE. En el futuro, este tipo de análisis debería agregarse a los análisis de nivel del ítem de la PSU.

El equipo evaluador internacional dependió de un marco TRI univariado, luego de verificar las soluciones unidimensionales de la PSU, para el análisis de los datos. La ausencia de DTF es indicativa de que los puntajes de la PSU son directamente comparables en las subpoblaciones, mientras que la presencia de DIF mostrará que los puntajes de ítem son inconsistentes entre las subpoblaciones. Se puede argumentar que el DTF es más importante que el Funcionamiento Diferencial de Ítem porque el primero habla del impacto, mientras que el último puede ser significativo para un ítem, pero puede que no tenga mucho impacto práctico sobre los resultados de la prueba (Templin, 2009).

Se llevaron a cabo análisis de factores y análisis de funcionamiento diferencial de pruebas con datos del proceso de admisión del año 2012.

Hallazgos:

Los hallazgos de este estudio respaldaron la presencia de una fuerte dimensión latente para cada prueba de la PSU. Los análisis revelaron una única dimensión subyacente para cada una de las pruebas de la PSU (Lenguaje y Comunicación, Matemática, Historia y Ciencias Sociales, Ciencias-Común, Ciencias-Biología, Ciencias-Física y Ciencias-Química). Tal hallazgo es alentador y apoya el uso futuro de modelos de teoría de respuesta al ítem unidimensionales para fijar y mantener la escala de puntajes de las pruebas de la PSU sobre años de administraciones de las pruebas y para la equiparación de los puntajes en las pruebas de la PSU. En Chile, los puntajes de la prueba PSU de una administración dada son válidos durante dos años; así, la comparabilidad entre puntajes de las pruebas es necesaria hacia la construcción de un programa de pruebas de selección nacional justo. La existencia de una "dimensión subyacente única" para cada una de las pruebas es una condición necesaria pero insuficiente para validez de la prueba. En nuestro análisis el equipo evaluador encontró una dimensión subyacente única para cada una de las pruebas de la

PSU. El equipo evaluador nota que las dimensiones subyacentes son interpretables dentro de una prueba de la PSU y, así, el conjunto de dimensiones subyacentes únicas puede que no sea el mismo, incluso para las diversas pruebas de Ciencias.

En general, las pruebas de la PSU indican alguna evidencia de Funcionamiento Diferencial de Pruebas (DTF). Bajo estas circunstancias, es razonable concluir que la invariancia de factor de estructura de las pruebas por grupos de subpoblaciones ha sido parcialmente lograda. En particular, la evidencia más fuerte del DTF es referida a los estudiantes Técnico-Profesionales de desempeño más bajo en relación a los estudiantes que realizan estudios Científico-Humanistas de desempeño más elevado en las pruebas de Ciencias-Biología, Lenguaje y Comunicación y Matemática, y para estudiantes de establecimientos privados de aún mucho mejor desempeño, y de estudiantes de establecimientos subvencionados de desempeño algo superior en relación a los estudiantes de establecimientos municipales de menor desempeño en las pruebas de Lenguaje y Comunicación y Matemática. Hay casos específicos que pueden ameritar mayor consideración y revisión.

Con respecto a los resultados por tema en la prueba de la PSU, Matemática indicó tener el mayor número de banderas DTF, con ocho banderas de las diez comparaciones de subgrupos. Esta fue seguida por Lenguaje y Comunicación con cinco banderas DTF.

Recomendaciones:

113. **El equipo evaluador recomienda adoptar el marco TRI para las actividades de la construcción de pruebas, los análisis de nivel de ítem, el escalamiento y la mantención de escalas.**

114. **El equipo evaluador recomienda seleccionar ítems operacionales durante las actividades de la construcción de pruebas para lograr altos niveles de precisión en los puntos de decisión críticos de la escala de puntajes.**

115. **El equipo evaluador recomienda usar los resultados del análisis factorial (que ha señalado la presencia de una dimensión) para incorporar el uso de la TRI tanto en los procesos de construcción de pruebas como en la equiparación de los puntajes de PSU.**

116. **Como resultado de los análisis realizados por el equipo evaluador en el Funcionamiento Diferencial de Pruebas (DTF), el equipo evaluador recomienda que el programa de la PSU realice análisis adicionales para entender mejor el DTF entre las instituciones educativas privadas y subvencionadas versus las instituciones educativas municipales, en particular para las pruebas de Lenguaje y Comunicación y Matemática. Este es un estándar reconocido (Estándar 7.3, AERA, APA, & NCME, 1999) para pruebas de alto nivel en todo el mundo, donde la equidad de la prueba a lo ancho de las diferentes subpoblaciones es un tema.**

Objetivo 2.2.Validez de contenidos: Análisis lógicos y empíricos para determinar si el contenido de las pruebas refleja plenamente el dominio y si la prueba tiene la misma relevancia en cuanto a la interpretación de los puntajes en las poblaciones de los subgrupos

Descripción:

El desarrollo de una adecuada validez de contenido siempre ha sido una preocupación para los desarrolladores y los usuarios de pruebas (AERA, APA, NCME, 1999). El método más común para asegurar la validez de contenido ha sido el uso de expertos en contenido en el proceso de desarrollo de pruebas. Sin embargo, el alineamiento de una prueba con un conjunto de estándares de logro es una estrategia relativamente nueva para determinar la validez de contenido de una evaluación.

Para que la PSU esté alineada con el currículum nacional chileno, es esencial que las pruebas de la PSU midan la profundidad y el ancho del currículum nacional en Lenguaje y Comunicación, Matemática, Historia y Ciencias Sociales, y Ciencia. Las evaluaciones que dedican un número proporcional de ítems a lo ancho de subconjuntos de contenido y habilidades especificadas en el currículum están mejor alineadas en relación a aquellas que se enfocan sobre contenidos periféricos y/o equilibrios desproporcionados.

Los propósitos del estudio fueron (1) documentar el grado de alineamiento de las pruebas de la PSU con el dominio pretendido de la PSU, (2) lograr una comprensión más profunda del alineamiento de la PSU desde la perspectiva del profesor de enseñanza media y del profesor universitario y (3) resumir la posición de la división de Currículum del Ministerio de Educación de Chile.

Respecto del primer propósito, la siguiente pregunta condujo la parte del alineamiento del estudio:

- ¿Cuál es el grado de alineamiento de las pruebas de la PSU con su dominio pretendido?

En cuanto al segundo propósito, los siguientes temas proporcionaron una estructura para la reunión con las personas interesadas.

- Grado de alineamiento percibido del dominio de la PSU con la instrucción en la sala de clases; y
- Preparación percibida de los estudiantes seleccionados a nivel de ingreso universitario al comienzo de la instrucción universitaria.

Método:

Se llevaron a cabo análisis lógicos y empíricos para la investigación del grado de cobertura del dominio pretendido de las pruebas de la PSU administradas en el proceso de selección del año 2012. Este esfuerzo tomó en consideración el currículum nacional chileno, los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) para las ramas curriculares Científico-Humanista y Técnico-Profesional. Los análisis se llevaron a cabo considerando la Capacitación General y la Capacitación Diferenciada (Científico-Humanista y Técnico-Profesional), usando la metodología de alineamiento de Webb (1997). La metodología entrega una visión integral del potencial de una prueba para evaluar a los

estudiantes sobre el material requerido. El método Webb no solamente mide el grado al cual los estándares (CMO y OF) son abordados por las preguntas de la prueba, sino que adicionalmente examina el nivel de complejidad cognitiva requerido por las preguntas de la prueba, la amplitud del conocimiento requerido por las preguntas de la prueba y la uniformidad de la cobertura de estándares (CMO y OF) por parte de una prueba.

Pearson utilizó las cinco dimensiones del proceso de alineamiento de Webb para juzgar el alineamiento entre los estándares de contenidos curriculares de la enseñanza media y las pruebas de la PSU para responder a las siguientes preguntas:

- **Concurrencia Categórica:** ¿Mide la PSU lo que los estándares curriculares declaran que los estudiantes deberían saber y ser capaces de hacer?
- **Profundidad del Conocimiento:** ¿Refleja la PSU la exigencia y profundidad cognitiva de los estándares curriculares?
- **Rango del Conocimiento:** ¿Refleja la PSU el ancho de los estándares curriculares?
- **Equilibrio de Representación:** ¿Refleja la PSU el rango completo de los estándares curriculares?
- **Fuente de Desafío:** ¿Refleja la PSU exigencias cognitivas ajenas a aquellas de los estándares curriculares?

Se llevó a cabo un estudio de alineamiento con un equipo de especialistas en áreas de contenido de Pearson, todos hispanoparlantes. De los seis panelistas, dos han viajado extensamente por el territorio nacional o han vivido en Chile. Cuatro de los panelistas son profesores con un promedio de ocho años trabajando en las salas de clases o en el desarrollo de mallas curriculares. Cuatro de los panelistas han obtenido grados avanzados (maestrías o doctorados) en sus áreas temáticas.

El alineamiento de Webb con la PSU se llevó a cabo según se presenta a continuación. En primer lugar, el jefe de proyecto del equipo asignó las diversas porciones de la PSU (Matemática, Ciencia, etc.) a los especialistas del área de contenido que habían sido extensamente capacitados en el método Webb de alineamiento. Luego, la primera tarea de los especialistas del área de contenido fue asignar calificaciones en la profundidad de conocimiento (DOK) a los objetivos del Ministerio de Educación de Chile respecto de los estándares de contenido de la enseñanza secundaria. De igual forma, ellos asignaron calificaciones DOK a los ítems de evaluación de la PSU. Luego, los panelistas determinaron la concurrencia categórica, la consistencia DOK, la correspondencia del rango de conocimientos, el equilibrio de representación y la fuente de criterios de desafío según definido anteriormente. Una vez completado, los especialistas de área de contenidos enviaron su trabajo al líder de equipo del proyecto, quien a su vez lo envió fuera de Pearson para ser revisado en forma independiente.

En un esfuerzo separado, profesores universitarios y docentes de enseñanza media participaron de una serie de entrevistas. El equipo evaluador propuso usar grupos accesibles de personas interesadas identificadas a partir de una lista inicial del MINEDUC para recolectar su información anecdótica sobre el dominio de la PSU y la relación con niveles de conocimiento y habilidades relevantes para que los estudiantes que ingresan sean exitosos. Un total de 27 personas interesadas participaron de las entrevistas. Se entrevistó a once profesores universitarios. Todos eran de universidades estatales metropolitanas del CRUCH. Había un director docente, cuatro profesores de Matemática, un profesor de Física, dos profesores de Química, dos profesores de Biología y un profesor de Lenguaje y Comunicación. Se entrevistó a dieciséis profesores de enseñanza media. Los profesores provenían de dos establecimientos secundarios privados metropolitanos con modalidad curricular Científico-Humanista. Entre los profesores había cuatro que se especializaban en

Lenguaje y Comunicación, cuatro en Matemática, tres en Historia y Ciencias Sociales y cinco en Ciencias. Los profesores de Ciencias consistían de dos profesores de Química, dos profesores de Física y un profesor de Biología.

Las reuniones de entrevistas se iniciaron con una introducción de alto nivel del propósito y de las reglas del juego, seguido de una visión panorámica general de la evaluación de la PSU. Luego de esta presentación se les solicitaba a los entrevistados familiarizarse con los OF y CMO de las ramas curriculares de la enseñanza media de Chile (Científico-Humanista y Técnico-Profesional) y obedecer las indicaciones declaradas en los protocolos de las entrevistas. El facilitador del equipo evaluador fomentó la discusión y los puntos de vista alternativos por parte de los miembros del panel. Las respuestas a las entrevistas fueron analizadas por separado por un grupo de personas interesadas y los resultados fueron categorizados en hallazgos mayores y menores en tablas resumen.

Hallazgos:

Los resultados del estudio de alineamiento indican que para casi todas las pruebas de la PSU, el nivel de alineamiento de la PSU con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del currículum chileno era uniformemente bajo. Un aspecto a tener en cuenta al interpretar estos resultados es que hay ciertas ramas dentro de los conjuntos estándar que son imposibles de evaluar en un formato de examen de opción múltiple de la PSU. Esto tendería a disminuir el alineamiento. Sin embargo, aunque la existencia de estas ramas dentro de los estándares habrá de resultar en puntajes de alineamiento para la prueba de la PSU artificialmente más bajos, los resultados del estudio demuestran que hay muchas mejoras que se podrían hacer.

Adicionalmente, los análisis de información anecdótica de los docentes de enseñanza media y de los profesores universitarios indicaron que hay una desconexión fundamental entre el propósito y uso de la PSU para seleccionar estudiantes para las universidades y el contenido de la PSU que se encuentra basado en el Marco Curricular de la enseñanza media. Esta desconexión entre el propósito y el uso se encontró que era más fuerte para los estudiantes de la modalidad curricular Técnico-Profesional que para aquellos de la modalidad curricular Científico-Humanista.

En cuanto al uso de la PSU como un predictor del éxito en la educación superior, un tema mayor de las entrevistas fue que la PSU no captura todas las aptitudes (o habilidades) necesarias para que le vaya bien en la educación superior; por ejemplo, los profesores universitarios declararon que las pruebas de la PSU no capturan la motivación del estudiante u otras cualidades importantes y que los estudiantes a quienes no les va bien en las pruebas de la PSU pueden continuar y ser exitosos en la universidad. Otros temas que fueron capturados incluyeron la percepción de que los estudiantes de status socioeconómico inferior (SES) estaban en desventaja en las pruebas de la PSU y que los profesores de los grados 11 y 12 se enfocan más en enseñar para las pruebas de la PSU que en enseñar el currículum.

Finalmente, la Unidad de Currículum del MINEDUC entregó sus propios análisis por fuera del trabajo de evaluación sobre el Currículum Nacional de Chile y su relación con la PSU.

- *Alineamiento consistente*: la principal preocupación, que deriva del informe solicitado por el mismo DEMRE, es que el marco de evaluación de la PSU utiliza como referencia una sección del currículum, los CMO, que no necesariamente dan cuenta de la totalidad de los aprendizajes del currículum. En otras palabras, la PSU utiliza una metodología que la alinea solo superficialmente del currículum, dejando de lado lo central y los aprendizajes fundamentales

- *Coordinación con los cambios curriculares*: el currículum chileno, como cualquier currículum moderno, ha tenido últimamente un proceso de actualización y revisión. Específicamente, el currículum de Educación Media fue modificado en forma importante en 2009; este cambio ha sido implementado gradualmente año a año. Es fundamental entonces que exista certidumbre y transparencia, para todo el sistema educativo, en relación a qué currículum se está evaluando y cómo se construyen las intersecciones entre dos currículos.

- *Formación TP*: A la fecha, 45% de la matrícula de la enseñanza media corresponde a la formación Técnico-Profesional, un número creciente de egresados de esta modalidad rinde la PSU con miras a entrar a la educación superior. Desde la Unidad de Currículum hay preocupación por la distancia entre lo declarado (la PSU como evaluación de la formación general) y lo real (la PSU como evaluación de la formación general y diferenciada, que enfatiza más bien a la rama curricular Científico-Humanista). [MINEDUC, comunicación personal, enero 2013]

Recomendaciones:

117. **Recomendamos 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.** Como parte de esta revisión, recomendamos el desarrollo de un marco que describa las aptitudes (por ejemplo, habilidades) y variables no cognitivas relevantes (por ejemplo, habilidades de estudio y motivación) que los estudiantes necesitan a fin de ser exitosos en la universidad. Tal marco enfocaría la PSU sobre las aptitudes necesarias para tener éxito en la universidad y complementaría la medida de logro de la enseñanza media que se encuentra en el NEM y que se combinan en el puntaje de postulación.

118. Aunque el equipo evaluador ha recomendado alinear las pruebas de la PSU con estándares para el éxito en la universidad, reconocemos la urgencia de desarrollar las formas de pruebas de la PSU para el 2013–2014 basados en la plena implementación de la reforma curricular del año 2009. Hacia tal fin, el equipo evaluador recomienda llevar a cabo un estudio de alineamiento de estas formas de pruebas de la PSU. Los resultados de este estudio deberían informar la recomendación más amplia de redirigir el énfasis de la PSU al éxito universitario.

119. Recomendamos revisar los tipos de ítems usados en las pruebas de la PSU para abordar el nivel bajo de complejidad cognitiva percibido que se encuentra en las pruebas debido al foco sobre el uso exclusivo de habilidades de pensamiento de orden inferior y los Contenidos Mínimos Obligatorios.

Objetivo 2.3. Análisis de trayectorias de puntajes de la PSU para las subpoblaciones en el tiempo, considerando dependencia, modo y género

Descripción:

Debido a que la admisión a las universidades involucra el uso de información de pruebas de selección, el análisis de la trayectoria de los puntajes de admisión a las universidades a lo largo del tiempo se convierte en un elemento importante del portafolio de estudios de validez institucional para el respaldo de las generalizaciones a lo largo del tiempo. El análisis de las trayectorias de los puntajes de admisión contribuye a desarrollar una mayor comprensión del desempeño en las pruebas por parte de los subgrupos a lo largo de años de administraciones de las pruebas y a localizar tendencias descendentes y ascendentes sospechosas. Cuando los análisis de puntajes en las pruebas son desagregados por

subpoblaciones específicas, los análisis de las tendencias de los puntajes en las pruebas se convierten en una poderosa herramienta para monitorear el desempeño en las pruebas de selección universitaria y para informar las decisiones sobre políticas para cerrar las brechas observadas en el desempeño en las pruebas por parte de diversas subpoblaciones.

Los principales propósitos de la investigación fueron analizar las trayectorias de los puntajes de la PSU a través del tiempo y determinar con precisión las variables que moderaron dichas trayectorias. Un análisis de la trayectoria de los puntajes de las pruebas de selección universitaria en el tiempo es un elemento importante de los estudios de validez institucional, porque dichos análisis longitudinales ayudan hacia la identificación de tendencias estables sobre el desempeño en las pruebas y determinar con precisión brechas en el desempeño de las pruebas respecto de subpoblaciones relevantes.

El equipo evaluador investigó las siguientes preguntas:

- ¿Cuál es la tendencia en los puntajes de las pruebas de la PSU para las siguientes subpoblaciones?
    - Género: masculino o femenino.
    - Región: Norte (códigos 1, 2, 3, 4, 15), Central (5, 13 [RM]), o Sur (6, 7, 8, 9, 10, 11, 12, 14).
    - Status socioeconómico: cinco quintiles de la variable SES— el quintil A define el grupo más bajo; el quintil B define grupo promedio bajo; el quintil C define el grupo promedio; el quintil D define al grupo sobre el promedio; y el quintil E define el grupo superior. El SES se calculó usando información de los ingresos familiares de los postulantes y de la educación parental.
    - Modalidad Curricular: Científico-Humanista o Técnico-Profesional.
    - ¿Qué variables a nivel de institución educativa moderan la relación entre los puntajes de la PSU y el NEM?

Método:

El estudio dependió de los conjuntos de datos longitudinales, abarcando los procesos de postulación desde el año 2004 hasta el año 2011.El DEMRE suministró las bases de datos con la información demográfica de los postulantes, los puntajes en las pruebas de la PSU y el promedio de notas de la enseñanza media (NEM). Se calculó y reportó información descriptiva tal como "conteos", promedios, y desviación estándar para escalas de puntajes de la PSU (y puntajes brutos) por prueba y año de administración para el total de la población de postulantes y para cada una de las subpoblaciones. Como parte de las estadísticas descriptivas, se generaron e interpretaron gráficos de escalas de puntajes promedio (con bandas de confianza del 95% alrededor de las escalas de puntajes promedio). Adicionalmente se llevaron a cabo análisis verificadores de hipótesis con análisis de varianza con una variable dependiente para testear por diferencias nulas en las escalas de puntajes promedio de la PSU para el año y por subpoblación. Los análisis de resultados de varianza fueron resumidos en tablas e interpretados.

Para la segunda pregunta de investigación, los efectos de las características del nivel de enseñanza media sobre la covariación entre los puntajes de subpruebas de la PSU y del NEM se estudiaron con el modelo lineal jerárquico (HLM) [Kreft & De Leeuw, 1998; Raundenbush & Bryk, 2002]. El modelo permitió una investigación de la extensión a la cual las características del nivel de institución educativa (es decir, tipo, modalidad curricular)

afectaban la relación ente los puntajes del NEM y de la PSU. Las siguientes variables se encontraban en el modelo.

- Criterio: Puntajes PSU
- Unidad de análisis nivel 1: Estudiante
    - Predictor de nivel de estudiante: NEM
- Unidad de análisis nivel 2: Enseñanza Media
    - Predictores de nivel institución educativa: Nivel SES de institución educativa, Modalidad Curricular, Tipo de institución educativa, Región y % de la población estudiantil que es femenino.

Hallazgos:

Los resultados del análisis de tendencia indicaron que, en promedio, los puntajes de la PSU permanecieron bastante constantes en el tiempo con una leve tendencia al alza empezando en el año 2007. Un examen a las subpoblaciones indicó que esta tendencia al alza se debe en gran medida al desempeño de los establecimientos Privados y a los establecimientos con modalidad curricular Científico-Humanista. Las líneas de tendencias desagregadas por tipo de institución educativa y modalidad curricular indicaron que los puntajes aumentaban en forma sostenida en el tiempo respecto de establecimientos Privados y establecimientos con ramas curriculares Científico-Humanista, mientras que los puntajes se mantenían planos para los establecimientos de tipo Municipal y para establecimientos con una modalidad curricular Técnico-Profesional. Además de las diferencias de puntajes debido a los tipos de institución educativa y modalidad curricular, el género, el status socioeconómico (SES) y la región en la cual el estudiante residía moderaba significativamente la tendencia de los puntajes de la PSU. Los patrones de puntajes de las pruebas de la PSU a lo ancho de variables demográficas relevantes son parecidos a los patrones observados internacionalmente con personas mayores que postulan a la universidad. Por otro lado, el tamaño de la brecha es grande para la población chilena de estudiantes orientados a la universidad, particularmente para las subpoblaciones basadas en tipo de institución educativa y status socioeconómico.

Un foco secundario de esta investigación fue investigar la covariancia entre el promedio de notas del desempeño en la enseñanza media y los puntajes en la PSU. Se descubrió que mientras el NEM predecía el desempeño en todas las subpruebas de la PSU, se desempeñaba particularmente bien para Matemática y Ciencias. Las variables a nivel de institución educativa que moderan esta relación entre los puntajes NEM y los puntajes de la prueba de la PSU también fueron examinadas. El tipo de institución educativa y la modalidad curricular fueron moderadores particularmente fuertes de esta relación, siendo la pendiente del NEM sustancialmente más aguda para las instituciones educativas Privadas y para las instituciones que proporcionan la modalidad curricular Científico-Humanista. El SES, la región, y el porcentaje de estudiantes femenino en el nivel medio también moderaron la relación entre el NEM y los puntajes de la PSU.

Recomendaciones:

120. **El equipo evaluador recomienda llevar a cabo una equiparación del puntaje de la prueba anualmente.** En esta línea, recomendamos una cuidadosa inspección de la comparabilidad de los puntajes de la prueba de la PSU entre los años. En Chile los puntajes de la prueba de la PSU pueden ser usados durante los procesos de postulación en dos años consecutivos. Luego de dar cuenta de la falta de equiparación de los puntajes en las pruebas y de los cambios en las especificaciones

de las pruebas de la PSU (es decir, la prueba de Matemática aumentó su largo en el proceso de postulación del año 2012), la equidad del puntaje en la prueba de la PSU entre años adyacentes está en riesgo. Debería ser un tema indiferente para los postulantes si toman la prueba de la PSU en el 2011 o en el 2012.

121. **El equipo evaluador recomienda inspeccionar la invariancia de las funciones de equiparación entre subpoblaciones relevantes de postulantes.** Estos tipos de verificaciones son pertinentes para el desarrollo de evidencia de validez sobre el significado de los puntajes de las pruebas. Los resultados fuertes de equiparación deberían ser invariables entre las subpoblaciones; de otra forma, deberían realizarse estudios de enlazado alineando las escalas de puntaje para permitir comparaciones entre puntajes de la prueba de la PSU.

Objetivo 2.4. Validez predictiva de la PSU: Para complementar la validez predictiva sobre grupos de la población a través de los años de administración, considerando las diferencias experimentadas en aquellos que rindieron la PSU y las variaciones de la prueba desde su implementación (2004), que habrá de contemplar un análisis de validez diferencial y posible predicción diferencial de la PSU por año y tipo de carrera, considerando subgrupos definidos por género, dependencia y modo educacional

Descripción:

El propósito de este estudio fue triple: (1) documentar la habilidad de los puntajes de la prueba de la PSU y del desempeño académico durante la enseñanza media (NEM y la clasificación de la enseñanza media) para predecir los resultados académicos de los estudiantes universitarios; (2) documentar el valor de predicción incremental del ranking variable; y (3) examinar hasta dónde los puntajes en la prueba de la PSU y el desempeño académico durante la enseñanza media exhiben predicción diferencial respecto de variables demográficas relevantes. El estudio documentó el desempeño de la validez predictiva para todo (todas las carreras) y niveles de tipo de carrera.

Para la validez predictiva, el estudio buscó dar una respuesta a la siguiente pregunta de investigación:

- ¿Cuál es la validez predictiva de los puntajes de la prueba de la PSU y del promedio de notas de la enseñanza media (NEM) sobre el promedio de notas del primer año universitario, el promedio de notas del segundo año y para la graduación universitaria?

La validez predictiva incremental es el grado en el cual una variable predice de mejor manera el resultado que una alternativa variable. Se busca que el análisis de validez predictiva incremental responda a la siguiente pregunta:

- ¿Cuál es la validez predictiva incremental de la variable de ranking (medida como una variable *proxy* del NEM) sobre y más allá de los puntajes de la prueba de la PSU y NEM sobre el promedio de notas del primer año universitario, el promedio de notas del segundo año y para la graduación universitaria?

La validez predictiva diferencial es otro tipo de estudio de validez institucional dirigido hacia la investigación del grado de similitud y diferencia en los resultados predichos entre subpoblaciones relevantes. Con respecto a la validez predictiva diferencial, este estudio buscó contestar la siguiente pregunta:

- ¿Cuál es la validez predictiva diferencial de los puntajes de las pruebas de la PSU y del promedio de notas de la enseñanza media sobre el promedio de notas del primer año universitario, el promedio de notas del segundo año y para la graduación universitaria para las siguientes variables?:

    - Género,
    - Nivel socioeconómico,
    - Región,
    - Modalidad Curricular de la enseñanza media, y
    - Tipo de financiamiento del establecimiento de enseñanza media

Método:

La investigación hizo uso de conjuntos de datos longitudinales para postulaciones universitarias que abarcó desde el año 2004 al 2012. El DEMRE suministró las bases de datos con los puntajes de la prueba de la PSU y el promedio de notas de la enseñanza media (NEM), y el MINEDUC proveyó las bases de datos con los resultados académicos universitarios de los estudiantes y su puntaje de ranking de enseñanza media. El MINEDUC también proporcionó una lista con la clasificación de las carreras universitarias por tipo de carrera usado en la investigación.

Se corrieron análisis de regresión lineal y logística separadamente para cada carrera dentro de una universidad y fueron resumidos entre las carreras y universidades. Se aplicaron correcciones para restricción de rango, involucrando varianzas y desviaciones estándar de puntajes de la prueba de la PSU de la población de personas de mayor edad en rumbo a la universidad (esto es, población de postulantes a la universidad), a los coeficientes de validez de Pearson de la población de estudiantes universitarios. Los resultados de validez predictiva de la PSU fueron ponderados a fin de asignar mayor peso a tamaños de muestras más grandes. Se calculó la validez de predicción incremental de la variable ranking mediante el ajuste de dos modelos: base y revisado. El modelo revisado usó el ranking como un predictor adicional. El efecto de la variable ranking se documentó mediante el cálculo de la diferencia en la reducción de la varianza de los resultados universitarios (es decir, el modelo revisado menos el modelo basado).Los análisis de validez diferencial se llevaron a cabo mediante variables demográficas con estimaciones de residuales estandarizados calculados dentro de las carreras y desagregados por variables demográficas. Para propósitos sumativos, los residuales de los estudiantes individuales se promediaron entre las carreras y años de selección antes de su desagregación por variables demográficas.

Hallazgos:

Los hallazgos para el estudio de predicción indican que las pruebas de la PSU hasta un cierto punto tienen la habilidad de predecir los resultados universitarios, en particular respecto de los promedios de notas del primer y segundo año. Sin embargo, los valores de predicción encontrados fueron menores que aquellos informados internacionalmente. La variable "ranking en la enseñanza secundaria" contribuyó a la reducción de la incertidumbre de predecir los resultados universitarios luego de controlar los puntajes de la prueba de la PSU y el NEM. La mayor cantidad de reducción de varianza tuvo lugar para la finalización de la universidad. En su conjunto, los puntajes de las pruebas de la PSU y las mediciones del desempeño durante la enseñanza media parecen tener como resultado montos comparables de validez predictiva diferencial para variables demográficas mayores.

Los resultados a nivel de tipo de carrera indicaron tendencias similares a las observadas en los análisis generales. Al examinar los resultados de validez predictiva e incremental por tipo de carrera, vimos patrones de predicción similares a aquellos de los análisis generales; por ejemplo, los puntajes en la PSU de Matemática y Ciencias y el desempeño académico en la enseñanza media (NEM y ranking) mostraron una mayor capacidad predictiva que los puntajes en la PSU para Lenguaje y Comunicación e Historia y Ciencias Sociales. Además, el ranking en la enseñanza media de los estudiantes mostró una validez predictiva incremental respecto de los resultados universitarios (más allá de los puntajes de la prueba de la PSU y el NEM), aunque su contribución fue menor que la que se encontró a partir de los análisis generales.

Recomendaciones:

122. **El equipo evaluador recomienda continuar desarrollando líneas de evidencia de respaldo para el uso y sentido de las medidas de la PSU.** Deberían agregarse nuevas variables al criterio de selección en las revisiones futuras luego de su cuidadosa evaluación para reducir la cantidad de incertidumbre al predecir los resultados universitarios. **En este contexto, recomendamos llevar a cabo estudios de validez para establecer líneas de evidencia para respaldar el proceso de toma de decisiones en anticipación a la adopción de dichas mejoras.**

123. Recomendamos investigar criterios alternativos para la investigación de la validez predictiva más allá del promedio de notas del primer año universitario o tasas de graduación mediante la inclusión de medidas de continuación de estudios en instituciones educativas de posgrado, de ser contratado en ocupaciones relacionadas con la carrera y del salario a nivel de novato.

124. Recomendamos investigar si es que las prácticas de asignación de notas en la universidad son uniformes dentro de una carrera en diferentes universidades y a lo ancho de la misma carrera en diferentes universidades, ya que esta información contribuiría a solidificar aún más los hallazgos de las medidas de validez predictiva que emplean el promedio de notas universitarias como una variable dependiente.

Prueba de la PSU. Evaluación por pregunta de validez y recomendaciones generales

En esta sección, el equipo evaluador resume, en base a una serie de preguntas, la validez de la batería de pruebas de la PSU. La discusión no es un resumen de los resultados de la evaluación integral. Más bien, las preguntas planteadas en esta sección se enfocan en aspectos de la PSU relacionadas con el contenido de la prueba su confiabilidad, los constructos medidos por esta y los usos e interpretaciones de la prueba. Estas preguntas proporcionan un marco de referencia internacional y destacan las áreas donde la prueba debería ser mejorada.

Pregunta #1: ¿Están las pruebas de la PSU fuertemente alineadas al dominio previsto?

Esta pregunta fue en parte investigada mediante un estudio de alineamiento de las pruebas de la PSU con sus dominios previstos. Adicionalmente se realizaron entrevistas con actores relevantes (profesores universitarios y docentes de enseñanza media), para determinar su percepción sobre el grado de alineamiento de cada dominio de la PSU con la instrucción en la sala de clases. Los hallazgos de los estudios y de las entrevistas, anotados en el Objetivo 2.2, indican que ninguna de las pruebas de la PSU estaba alineada fuertemente con su dominio previsto.

Pregunta #2: La dificultad de las pruebas PSU, ¿apunta adecuadamente al nivel de habilidad del postulante?

Esta pregunta fue investigada mediante una demostración llevada a cabo por el equipo evaluador en el Objetivo 1.1.i., el cual examinó las distribuciones de habilidad latente de las pruebas de la PSU. El equipo evaluador encontró que la única prueba que apuntaba adecuadamente al nivel de habilidad del postulante era la prueba de Lenguaje y Comunicación.

Pregunta #3: Las pruebas de la PSU, ¿miden con precisión las habilidades del postulante en las áreas de habilidad de más importancia?

Para esta pregunta el equipo evaluador analizó en el Objetivo 1.1.i el Error Condicional de Medición para cada prueba de la PSU. Los resultados de esta demostración indicaron que el máximo de información (es decir, CSEM bajo) se obtuvo en el nivel alto (es decir, selectivo) de la escala para cada prueba, en el cual las decisiones de admisión eran probables de tener lugar. Se encontró que Ciencias y Matemática cumplían con este criterio de CSEM bajo dentro del rango que es relativamente selectivo. Sin embargo, el CSEM aumentó dramáticamente fuera de este rango (por ejemplo, en el rango más selectivo de la escala, aún para Ciencias y Matemática).

Pregunta #4: Cada una de las pruebas de la PSU, ¿refleja un rasgo unidimensional subyacente?

Los hallazgos del análisis factorial al nivel de la pregunta (llevados a cabo separadamente para todas las pruebas de la PSU) respaldaron la afirmación de que hay una dimensión latente fuerte para cada prueba de la PSU. Los análisis en Objetivo 2.1 revelaron una dimensión subyacente única para cada una de las pruebas de la PSU (Lenguaje y Comunicación, Matemática, Historia y Ciencias Sociales, Ciencias–Común, Ciencias-Biología, Ciencias-Física, y Ciencias-Química).

Pregunta #5: Los ítems y pruebas de la PSU, ¿se comportan de la misma manera a lo ancho de las mayores subpoblaciones?

Los análisis de Funcionamiento Diferencial de Ítems (DIF) encontrados en el Objetivo 1.1.g y de Funcionamiento Diferencial de Pruebas (DTF) encontrados en el Objetivo 2.1, fueron utilizados para investigar esta pregunta para las siguientes subpoblaciones: Género, Status socioeconómico (SES), Región (Metropolitana, Norte y Sur), modalidad (Público, Privado y Subvencionado), y modalidad curricular (Científico-Humanista y Técnico-Profesional). En general, aún cuando la mayoría de los ítems de la PSU evidencian un DIF insignificante, las pruebas de la PSU indican evidencia parcial de DTF por modalidad y modalidad curricular para todas las asignaturas exceptuando Ciencias Sociales.

Pregunta #6: Las pruebas de la PSU, ¿predicen con efectividad los resultados universitarios de los postulantes?

Esta pregunta fue desarrollada en el Objetivo 2.4. Aunque las pruebas de la PSU de Matemática y de Ciencias mostraron valores medios de validez predictiva, en ninguna instancia lograron un índice de validez predictiva que se acercase al límite inferior de índices de validez predictiva observados internacionalmente.

Pregunta #7: La escala y los puntos de corte de la PSU, ¿permanecen constantes a través de los años?

Los estándares internacionales exigen que las distintas formas de las pruebas sean equiparadas a fin de permitir comparaciones de las escalas de puntajes a través de los años. En el Objetivo 1.3, el equipo evaluador confirmó que la equiparación de puntajes de pruebas de las administraciones año tras año, no se ha llevado a cabo en toda la existencia del programa de pruebas de la PSU. Este hallazgo implica que la escala y los puntajes de corte de la PSU *no son* constantes a lo largo de los años.

Pregunta #8: ¿Son los informes de puntajes de la PSU útiles y claros para las audiencias previstas?

Se llevaron a cabo entrevistas con personas interesadas de la PSU (profesores universitarios, profesores de enseñanza media y estudiantes) en el Objetivo 1.5 para explorar sus opiniones respecto de la utilidad de los informes de puntajes de la PSU que ellos reciben. Estas personas interesadas notaron una falta de claridad y de utilidad en los informes de puntaje de la PSU que ellos examinaron.

Pregunta #9: ¿Deberían usarse los puntajes de las pruebas de la PSU para asignar becas?

Respecto de esta pregunta, el equipo evaluador consideró los hallazgos descritos en el Objetivo 1.1.h y en el Objetivo 1.3. Estos objetivos indican que hay problemas significativos con las escalas de puntajes de la PSU y los cortes para la asignación de becas. En este momento, los puntajes de las pruebas de la PSU se utilizan para otorgar beneficios sociales basados en puntajes de cortes que necesitan mayor validación.

Recomendaciones Generales

Estos juicios retratan un programa de pruebas que se está desarrollando aún y que necesita mejorar en varias áreas. Una imagen más matizada de este programa puede ser recogida de mejor manera estudiando el Informe de Evaluación completo o revisando las principales recomendaciones presentadas por objetivo en este Resumen Ejecutivo. Sin embargo, las dos recomendaciones generales que siguen pueden ser utilizadas para proporcionar un plano hacia el mejoramiento de la PSU.

Recomendación General 1: La base para el desarrollo de pruebas de la PSU debiese desplazarse desde la Teoría Clásica de Pruebas (TCT) hacia la Teoría de Respuesta al Ítem (TRI).

La Teoría de Respuesta al Ítem (TRI) es una teoría psicométrica que define y modela el desempeño de los examinados en términos de una habilidad o rasgo teórico subyacente; por ejemplo, en un examen de Matemática esto significaría que el desempeño de los examinados en ítems de prueba específicos y en la prueba en su totalidad sería definido por un rasgo subyacente denominado "habilidad matemática". Lo que hace que la teoría TRI sea tan poderosa es que el nivel del rasgo o habilidad no observable puede ser inferido y estimado en base al desempeño en un conjunto de ítems de pruebas observables. Usar la teoría de respuesta al ítem como base en el desarrollo de la PSU permitiría abordar una serie de deficiencias del programa de pruebas tal como se describe en el cuerpo del informe de evaluación.

En primer lugar, TRI modela específicamente la relación entre la habilidad del examinado y la dificultad de un ítem dado. Esto es una gran ventaja para la construcción de pruebas, ya que permite que el desarrollador de pruebas apunte explícitamente la dificultad de una prueba a la habilidad de un grupo de examinados. Esto alivia el problema de crear una prueba que pueda ser demasiado fácil o demasiado difícil para la población destinataria, preocupación planteada en la pregunta N°2 con respecto a las pruebas de la PSU de Matemática, Historia y Ciencias Sociales y Ciencias. Además, cuando existe un gran conjunto de ítems, el uso de las Curvas Características del Test (CCT) les permite a los desarrolladores de pruebas hacer calzar las propiedades estadísticas de una prueba nueva con las anteriores en un alto grado.

En segundo lugar, los métodos TRI permiten examinar las curvas de la Función de Información del Test (FIT) y las curvas de Error Condicional de Medición (CSEM). Estas curvas indican el grado de precisión de la medición como función de la habilidad del examinado. Esto significa que los desarrolladores de pruebas podrían determinar qué rango de la escala de puntaje de la PSU está midiendo con mayor exactitud. Mediante un proceso iterativo de selección de ítems y el reemplazo y examen de estas curvas, el desarrollador de pruebas podría enfocar el mayor nivel de precisión de medición en el lugar de la escala de pruebas donde se debería tomar las decisiones más importantes (Pregunta N°3). En el caso de las pruebas de la PSU, este lugar se ubicaría alrededor de los puntos de corte en la escala de puntaje de la PSU que son utilizados para las decisiones de selección universitaria.

Un tercer punto importante es que TRI proporcionaría un marco sólido para la calibración de ítems piloto ubicándolos en la escala del programa de pruebas. La dificultad de ítem obtenida mediante TRI y los parámetros de discriminación proporcionan una alternativa al empleo de valores-*p y* coeficientes biseriales de la

Teoría Clásica de medición (TCT). Además, ambas medidas son independientes del nivel de habilidad de las muestras. Esto significa que cuando los ítems están calibrados adecuadamente, los valores de las estadísticas de ítem TRI no dependen de la muestra específica que se utilizó para calibrarlos. Este no es el caso de las estadísticas de ítem cuando se usa TCT, el enfoque que la PSU utiliza hoy en día.

Un cuarto uso de TRI permitiría a la PSU eliminar la corrección por adivinación que se emplea actualmente. El modelo logístico de 3-parámetros (3PL) es un modelo TRI muy flexible que explícitamente toma en cuenta la dificultad de ítem, la discriminación de ítem y una corrección por adivinación. Mediante el modelamiento de la adivinación dentro del marco TRI, los desarrolladores de pruebas podrían entender de mejor manera cómo esto afecta la probabilidad de un postulante de responder correctamente un ítem.

El programa de pruebas de la PSU está usando análisis TRI ahora, aunque sea de una manera somera. El desplazamiento hacia el uso de TRI como marco del programa completo puede ser gradual. Esta transición de las pruebas de la PSU desde un programa basado en CTT a uno basado en TRI se puede acelerar con los recursos técnicos, políticas y recursos administrativos adecuados.

Recomendación General 2: Las pruebas de la PSU deberían ser equiparadas utilizando TRI.

Cada año, las pruebas de la PSU son administradas en formas de prueba nuevas. Aun cuando desde un punto de vista de seguridad esto es esencial, puede dar lugar a otro problema. De acuerdo con Kolen y Brennan (2004):

> El uso de diferentes formas de pruebas en ocasiones conlleva a otra preocupación: las formas pueden ser algo distintas respecto de su dificultad. La *equiparación* es el proceso estadístico que se emplea para ajustar los puntajes en las formas de pruebas a fin de que los puntajes en las formas puedan ser utilizados de modo intercambiable. La equiparación ajusta respecto de diferencias de dificultad entre formas que tienen similar dificultad y contenido. (Kolen & Brennan, 2004, p. 2)

Un enfoque para realizar la equiparación es el de basar el ajuste estadístico necesario en el desempeño de un conjunto de ítems en común a lo largo de las formas de pruebas de las dos administraciones. Estos conjuntos de ítems son referidos como conjuntos ancla.

Actualmente, el programa de la PSU no equipara sus pruebas, lo cual significa que los puntajes de esas pruebas no son comparables año a año. Aunque el programa de la PSU emplea lo que ellos llaman "conjuntos ancla", que aparecen a lo largo de las formas de pruebas, dichos grupos de preguntas no se emplean para llevar a cabo la equiparación.

Otro problema es que el programa de la PSU intenta equiparar las pruebas de Ciencias que surgen de diferentes combinaciones de la sección común con diferentes secciones opcionales (Biología, Física y Química). Debido a que estas secciones opcionales acarrean diferencias de contenido, los puntajes para los estudiantes que toman las diferentes secciones opcionales no pueden ser considerados como equiparados.

Una solución para estos problemas es el uso de TRI, el cual provee un marco bien desarrollado para la equiparación de las formas de pruebas usando conjuntos ancla. Esto ayudaría al programa de la PSU de varias maneras.

En primer lugar, el usar conjuntos ancla TRI, que de verdad funcionen como se pretende, permitiría que diferentes formas de pruebas piloto u operacionales sean equiparadas. Esto permitiría comparaciones directas de estadísticas de ítems TRI en las formas piloto y operacionales y ayudaría de gran manera a la construcción de nuevas pruebas.

En segundo lugar, el uso de conjuntos de ítems ancla a lo largo de las administraciones de pruebas operacionales y TRI permitiría que las pruebas de la PSU sean de verdad equiparadas año a año. Esto permitiría comparar las habilidades de los postulantes a lo ancho de las administraciones y asegurar que la escala de puntajes de corte de la PSU para decisiones de selección en una administración corresponda a los puntajes de corte de administraciones anteriores.

Finalmente, crear pruebas de la PSU separadas para Biología, Física y Química reconocería el hecho que el contenido es bastante diferente y que las secciones opcionales no pueden ser equiparadas unas con otras. El hecho de proceder con pruebas separadas permitiría una mejor medición de los constructos subyacentes y la verdadera equiparación de las pruebas entre administraciones.

El cambio hacia la equiparación de las formas de las pruebas de la PSU empleando el TRI podría ser introducido junto con los otros usos del TRI descritos anteriormente.

Recomendación General 3: La PSU debería desarrollar un programa de investigación continuo para validar los usos e interpretaciones de las pruebas.

El concepto de validez es considerado por el equipo evaluador como la piedra fundamental de la evaluación de la PSU. De acuerdo con la edición actual del *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), la validez "se refiere al grado al cual la evidencia y la teoría respaldan la interpretación de puntajes de pruebas requerida para los usos propuestos de los puntajes" (1999, p. 9). Por lo tanto, la validez se refiere a un proceso de desarrollo de evidencia para respaldar interpretaciones y usos pretendidos de puntajes de pruebas (Cronbach, 1984; Kane, 2006; Vernon, 1963). El equipo evaluador llevó a cabo numerosos estudios relacionados con la validez de las pruebas de la PSU, y estos estudios pueden ser utilizados como modelos para desarrollar un programa continuo de investigación que busque examinar temas claves de validación tal como se indica arriba.

La pregunta #1 cuestiona si las pruebas de la PSU se alinean con el dominio que pretenden medir. Para evaluar la competencia académica de asistir a la universidad con alguna medida de éxito, podemos extraer niveles razonables de inferencias a partir de los desempeños de los postulantes en las pruebas de selección desarrolladas para medir los estándares de contenidos que definen dicha competencia académica. Si una muestra adecuada de los ítems de una prueba ha sido evaluada de alguna manera y el error de medición está dentro de límites tolerables, entonces los puntajes de las pruebas pueden aceptarse razonablemente como medidas de un nivel de competencia logrado en el dominio de la prueba pretendida; de otra manera el grado de respaldo en evidencias no sería tan fuerte. La metodología en el estudio del equipo evaluador del Objetivo 2.2 podría ser utilizado para el propósito de recolectar esta evidencia de

validez relacionada con contenidos. Esta evidencia será especialmente importante si la base de contenido para las pruebas de la PSU cambia desplazándose hacia adelante.

La pregunta #2 se enfoca sobre la unidimensionalidad de las pruebas de la PSU y fue examinada por el equipo evaluador en el Objetivo 2.1. Tal como se describe más arriba, la evidencia respaldó la afirmación de la presencia de una dimensión fuerte para cada una de las pruebas de la PSU. Sin embargo, el programa de la PSU va a necesitar continuar estudiando este tema, especialmente si se busca implementar un marco basado en TRI para el desarrollo de pruebas y la equiparación de sus puntajes, ya que la unidimensionalidad es una condición necesaria para usar modelos TRI con una sola dimensión.

La evidencia de la validez predictiva está en el corazón de la pregunta #6. De hecho, la respuesta a esta pregunta en muchos aspectos podría ser la más importante fuente de evidencia para la validación de las pruebas de la PSU: Si las pruebas de la PSU no predicen los resultados universitarios, entonces ¿por qué las estamos utilizando? Actualmente, los coeficientes de validez predictiva de las pruebas de la PSU están bajos con respecto a aquellos que se ven internacionalmente. Esto indica la necesidad de continuar explorando en profundidad la relación entre los puntajes de las pruebas de la PSU y las variables asociadas al éxito universitario tales como el promedio de notas de primer año y las tasas de graduación.

La pregunta #8 se enfoca sobre la calidad de los informes de puntaje, lo cual está directamente relacionado con la interpretación del puntaje de la prueba, mientras que la pregunta #9 mira directamente al uso de la prueba PSU. En cada caso, la evidencia respecto de las consecuencias asociadas con las pruebas de la PSU necesita ser recolectada. Esto es especialmente crucial si la PSU habrá de ser utilizada para fines alejados de su propósito original.

En cualquier evento, el proceso de validación de pruebas es continuo. Los *Estándares* indican, "En tanto la validación procede, y nueva evidencia acerca del sentido de los puntajes de las pruebas se hace disponible, se pueden necesitar revisiones de la prueba, en el marco conceptual que le da forma, y aún en el constructo subyacente de la prueba" (AERA, APA, & NCME, 1999, p. 9). El programa debería poner en su lugar formalmente un proceso mediante el cual las pruebas de la PSU puedan validarse sobre una base continuada.

BIBLIOGRAFÍA

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Comité Técnico Asesor. (2013) Objetivo. Extraído de: http://www.cta-psu.cl/objetivo.asp

Consejo Directivo. (2010). *Consejo directivo para las pruebas de selección y actividades de admisión.* Extraído de: http://www.consejoderectores.cl/site/GobTrans/activa/documentos/ConsejoDirectivoPruebas.pdf

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York, Harper and Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

DEMRE. (2010a). *Prueba de Selección Universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE (2010b). Es*tudio de Confiabilidad de las pruebas de selección universitaria*. Admisión del 2010. Santiago Chile: Autor.

DEMRE (2012). *DEMRE. Departamento de evaluación, medición y registro educacional.* Extraído de: http://www.demre.cl/demre.htm

Educational Testing Service. (2005). *Evaluación externa de las pruebas de selección universitaria (PSU)*. Princeton, NJ: ETS Global Institute.

Feldt, L., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp. 105-146). New York: American Council on Education and Macmillan.

Hambleton, R., & Swaminathan, H. (1985).*Item response theory: Principles and applications.* Norwell MA: Kluwer Academic Press.

Hofstee, W. K .B. (1983). The case for compromise in educational selection and grading. In S. B. Andersen & J. S. Helmick (Eds.) *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

International Test Commission. (2012). *ITC guidelines for quality control in scoring, test analysis, and reporting test scores*. ITC: Author.

JUNAEB. (2012). *Beca JUNAEB para la PSU*. Extraído de: http://www.junaeb.cl/prontus_junaeb/site/artic/20100114/pags/20100114174738.html)

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Kolen, M., & Brennan, R. (2004).*Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

MINEDUC. (2011). *Aprueba bases administrativas, bases técnicas y anexos de licitación pública, sobre servicio de evaluación de la Prueba de Selección Universitaria (PSU)* (ID Nº 592-44-LP11). Santiago, Chile: Autor.

MINEDUC. (2012). *Educación superior.Aporte Fiscal Indirecto*. Extraído de: http://www.superior.mineduc.cl/index2.php?id_portal=38&id_seccion=3063&id_contenido=12223

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006 technical report.* OECD Publishing. Extraído de: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Organisation for Economic Co-Operation and Development. (2012). *PISA 2009 technical report.* OECD Publishing. Extraído de: http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Templin, J. (2007). *Introduction to differential item functioning*. [PowerPoint]. A presentation to the American Board of Internal Medicine for an Item Response Theory Course. Extraído de: http://jtemplin.coe.uga.edu/files/irt/irt07abim/irt07abim_lecture10.pdf

Vernon, P. (1963). *Personality assessment*. London, Methuen.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.