International Division

el 10 de enero de 2005

Prof. Dra. Cecilia Sepúlveda Carvajal
Vicerrectora de Asuntos Académicos
Universidad de Chile
Diagonal Paraguay 265 oficina 2101
Santiago - Chile

Estimada Dra. Sepulveda,

Educational Testing Service (ETS) se complace en enviarle el informe final de la evaluación externa hecha por ETS de la PSU. En adicional, como nos han pedido, le brindamos en la presente nuestras sugerencias con respecto a la capacitación del equipo de DEMRE.

Areas de capacitación

Taller sobre temas sicométricos de tipo aplicación cubriendo los siguientes áreas:
- Interpretación y uso de SPSS y BILOG
- Equiparación
- Construcción de banco de ítems
- Desarrollo de escalas de puntaje
- Assemblaje de pruebas usando deltas equiparadas

Taller en temas relacionados a la elaboración de pruebas:
- Desarrollo y revisión de ítems especialmente los de matemática y ciencia

Consultoria sobre:
- el desarrollo de planes de investigación (research agendas) con respecto a los diferentes tipos de validez y la equidad
- posible inclusión de una prueba de escritura en la PSU
- el desarrollo de planes de comunicación

Estamos dispuestos a entregarle a la fecha que usted diga una propuesta que describe con mayor detalle una serie de oportunidades de capacitación que podía tomar lugar en Santiago o en Princeton durante 2005. También podiamos facilitar que algunos del equipo pudieran explorar la posibilidad de seguir estudios avanzados, de tipo posgrado en una universidad amiga, o sea una que coopera con ETS.

ETS se siente honrada con haber tenido la oportunidad de colaborar con la Universidad de Chile en la evaluación de la PSU. Ojalá que tengamos la oportunidad de continuar esta colaboracion positiva en el futuro. Seguimos a sus órdenes en todo que le ofrece.

Saludos cordiales,

Annabelle Galera Simpson
Directora
ETS Global Institute

Rosedale Road 331
Princeton, NJ 08541

Evaluación Externa de las Pruebas de Selección Universitaria (PSU)

Enero del 2005

Educational Testing Services (ETS)

Evaluación Externa de las Pruebas de Selección Universitaria (PSU)

Enero de 2005

Educational Testing Services (ETS)

Antecedentes

En agosto del año 2003 una delegación representante de la unidad de servicios de evaluación de la Universidad de Chile, DEMRE (Departamento de Evaluación, Medición y Registro Educacional) participó en un seminario del Global Institute (Instituto Global) de Educational Testing Service (ETS), en Princeton, New Jersey, U.S.A., titulado "Métodos Estadísticos para Pruebas Educativas Críticas". Uno de los resultados de este seminario fue la solicitud de información por parte de la Vicerrectora de Asuntos Académicos de la Universidad de Chile, acerca de la posibilidad de que ETS realizara una evaluación externa del nuevo examen de admisión universitaria PSU (Pruebas de Selección Universitaria). Los resultados de las PSU son utilizados por la Universidad de Chile y por otras 24 universidades supervisadas por el H. Consejo de Rectores para la toma de decisiones con respecto a admisiones, al igual que para el otorgamiento de becas y para ayuda financiera. En marzo de 2004, la Universidad de Chile presentó los términos de referencia para la evaluación externa, y en septiembre de 2004, comenzó la evaluación inicial de los documentos de las PSU por parte del equipo de ETS. Se acordó que, en su etapa inicial, la evaluación externa por parte de ETS sólo cubriría las PSU de Lenguaje y Comunicación, y Matemáticas. Durante un período de aproximadamente cuatro meses, el equipo de evaluación externa de ETS revisó una amplia gama de documentos, los cuales reflejaban los procedimientos seguidos por el DEMRE en el desarrollo de los exámenes. El equipo de ETS se reunió dos veces con el equipo del DEMRE en Santiago de Chile. La primera reunión se realizó en octubre de 2004 y estuvo enfocada en la aclaración de temas, tales como el propósito de los exámenes y los procedimientos de escalamiento usados por el DEMRE. Durante la segunda reunión, efectuada en diciembre del 2004 la versión en borrador de este informe fue presentada al equipo del DEMRE.

La Prueba de Selección Universitaria (PSU)

La Universidad de Chile ha sido responsable del desarrollo del proceso de admisión a las universidades representadas en el H. Consejo de Rectores desde los años 1930. En el año 1967, los exámenes escritos fueron cambiados al formato de opción múltiple. La Prueba de Aptitud Académica (PAA) inicialmente medía habilidades verbales y matemáticas, pero luego fueron introducidos exámenes basados en contenidos que medían el conocimiento del postulante en diferentes áreas, tales como: biología, química, matemáticas, física y estudios sociales. En abril de 2001, el Ministerio de Educación financió el desarrollo de una nueva batería de pruebas de admisión: el Sistema de Ingreso a la Educación Superior (SIES), que contribuyó a la elaboración de la actual PSU. La PSU tuvo su primera administración operacional en diciembre de 2003.

La batería de la PSU consiste de dos exámenes obligatorios: Lenguaje y Comunicación, y Matemáticas; y dos exámenes opcionales: Historia y Estudios Sociales, y Ciencias. El examen de Ciencias tiene un módulo común y tres módulos electivos en Biología, Física y Química.

Proceso de Evaluación Externa (Auditoría) de ETS

La base para la evaluación externa de las PSU es el Programa de Auditorías de ETS. En ETS todos los programas, sean de pruebas o no, son sometidos a una auditoría una vez cada tres años, aproximadamente. Cada una de las políticas y las prácticas de los programas son evaluadas de acuerdo con estándares reconocidos internacionalmente que están presentados en *Los Estándares de Calidad y Equidad de ETS (ETS Standards for Quality and Fairness)* y que fueron adoptados como política de la corporación por el ETS Board of Trustees en el año 1981. Estos estándares fueron revisados en el año 2000 (ETS, 2000), siguiendo la revisión de los *Standards for Educational and Psicological Testing (Estándares para las Pruebas Psicológicas y Educativas),* desarrollados por la American Educational Research Association (Asociación Americana de Investigación Educativa), la American Psychological Association (Asociación Americana de Psicología), y el National Council of Measurement in Education (Consejo Nacional de Medición en Educación), (APA, 1999). El propósito de los estándares de ETS es ayudar a ETS a diseñar, desarrollar y proporcionar productos y servicios técnicamente

sólidos, equitativos y útiles, y también ayudar a los auditores en la evaluación de estos productos y servicios. Mediante las auditorías externas o servicios de evaluación, ETS provee de retroalimentación objetiva a las instituciones internacionales en sus procesos de pruebas. Un programa de evaluación como el del DEMRE que acepta someterse al proceso de evaluación externa de ETS, acuerda ser evaluado con respecto a un conjunto de estándares uniforme y riguroso, mediante un proceso debidamente documentado.

En la evaluación de las PSU, ETS acordó centrarse en tres preguntas claves, relacionadas con la validez, la confiabilidad y el uso de las evaluaciones:

1. ¿Reflejan las pruebas de manera precisa el marco teórico y las especificaciones de contenido? El marco teórico y las especificaciones de contenido, ¿sustentan el uso previsto para las pruebas?
2. ¿Son los puntajes de las pruebas una medida precisa de sus contenidos?
3. ¿Sustentan los puntajes de las pruebas, el (los) uso (s) previstos de las mismas?

Es muy importante indicar que el programa de auditorías de ETS no otorga ningún tipo de certificación o acreditación, pero sí ofrece guía y orientación acerca de cómo remediar casos específicos de incumplimiento de los estándares internacionalmente aceptados de pruebas equitativas y válidas.

ETS está comprometido a trabajar con la Universidad de Chile y el DEMRE a medida que ellos procuran asegurarse que las PSU sean pruebas válidas, confiables y equitativas para todos aquellos que las rindan. En el resto de este informe presentamos los resultados de nuestra evaluación de las pruebas de Lenguaje y Comunicación, y Matemáticas. El informe se organiza de acuerdo con los estándares utilizados para la evaluación. Primero, informamos los resultados de nuestra evaluación del programa de acuerdo a los estándares de validez. Posteriormente, discutimos la evaluación del programa de acuerdo a los estándares de confiabilidad y, finalmente, presentamos los resultados de la evaluación del programa de acuerdo a los estándares correspondientes al uso de las evaluaciones. Para cada estándar, describimos primero el estándar y luego consideramos los puntos fuertes del programa, seguido por las áreas por mejorar. Finalmente, proveemos recomendaciones para mejorar el programa en relación al estándar en particular que se está discutiendo. Terminamos nuestro informe con un resumen y un conjunto general de recomendaciones.

Debido a que existe una considerable superposición entre los estándares en cuanto a los requisitos que establecen, el lector notará que también existe cierta superposición o redundancia en los puntos que se presentan en las evaluaciones por separado.

El Apéndice del informe contiene una evaluación detallada de cada una de las pruebas (Lenguaje y Comunicación, y Matemáticas) medidas en términos de cada uno de los estándares de manera individual.

RESULTADOS DE LA EVALUACION EXTERNA

Revisión de la Validez de la Prueba de Lenguaje y Comunicación, y de Matemáticas

Los Estándares de Validez

El propósito de los estándares de validez es ayudar a asegurar que el programa de evaluaciones reúna y documente evidencia que sustente los usos contemplados, las inferencias, y las acciones que puedan estar basadas en los puntajes reportados de las evaluaciones. La validez es un atributo muy importante de una evaluación. Por lo tanto, es importante que un programa de evaluaciones provea evidencia, tanto lógica como empírica, que muestre que la prueba es capaz de cumplir con el propósito que se propone. Aunque se pueden usar muchos tipos de evidencia para demostrar la validez de una evaluación, es muy importante que se suministre la evidencia que se usa para sustentar el propósito de la evaluación. Si los usuarios de los puntajes usan los resultados de las evaluaciones para otros propósitos que no sean los indicados específicamente por el programa de evaluación, esos usuarios serán responsables de evaluar la validez de esos puntajes para esos otros propósitos. El conjunto de estándares de validez está compuesto por los siguientes ocho estándares:

- Describir cómo las evaluaciones cumplen con la misión y los valores del programa.
- Describir claramente el constructo (conocimientos, destrezas, y otras características) a ser medido, el (los) propósito (s) de cada evaluación, la (s) interpretación (es) prevista(s) de los puntajes u otros resultados de las evaluaciones, y la población a quien van dirigidas las pruebas. Poner a disposición del público la información pertinente cuando se la requiera.
- Brindar una justificación para los tipos y cantidades de evidencias recolectadas que respaldan la validez de las inferencias a realizarse en base a las evaluaciones. Para evaluaciones nuevas, brindar un plan de validez indicando los tipos de evidencia que serán recolectados.
- Obtener y documentar la evidencia empírica y/o lógica de que las evaluaciones cumplirán con el (los) propósito (s) que se propone y respaldar la (s) interpretación (es)

propuestas de los resultados para la (s) población (es) a quien (es) van dirigidas la pruebas.

- Advertir a los usuarios potenciales que eviten usos de las evaluaciones para los cuales existe insuficiente evidencia de su validez.
- Si el uso de los resultados de una evaluación tiene consecuencias no intencionadas para un grupo que está siendo estudiado, entonces se debe revisar la evidencia de validez para determinar si esas consecuencias derivan de fuentes inapropiadas de variación. De ser así, revisar las evaluaciones para reducir, en la medida de lo posible, las fuentes inapropiadas de variación.
- Si algunos factores relevantes cambian, entonces se debe reevaluar la evidencia de que la evaluación cumple el (los) propósito (s) previsto (s) y sustenta la (s) interpretación (es) prevista (s) de los resultados de las pruebas para la población a quien van dirigidas las mismas, y reunir nueva evidencia en la medida en que sea necesario.
- Brindar orientación a los usuarios de los puntajes –u otros resultados de las evaluaciones– para ayudarles a reunir e interpretar su propia evidencia de validez.

La validez de las PSU- Lenguaje y Comunicación, y Matemáticas ha sido evaluada usando cada uno de los ocho estándares mencionados anteriormente. Una descripción detallada de la evaluación de la validez de cada una de las pruebas se encuentra en el apéndice de este informe. A continuación se presenta un resumen de las fortalezas, las áreas por mejorar, y recomendaciones para hacer que las prácticas de las PSU se alineen mejor con los ocho estándares de validez.

Fortalezas de las PSU Lenguaje y Comunicación, y Matemática

El personal de las PSU en el DEMRE ha realizado una labor extraordinaria al haber diseñado y creado una evaluación de gran escala y de grandes consecuencias en un período de tiempo extremadamente reducido, y con recursos relativamente limitados. Una de las fortalezas importantes del programa de las PSU es el dedicado personal del DEMRE. El dedicado profesionalismo de este personal es evidente en todo lo que realiza. Todos los documentos que fueron compartidos con el personal de ETS fueron de la más alta calidad. A pesar de las tan difíciles restricciones de tiempo bajo las cuales el personal del DEMRE se encontraba operando para introducir las nuevas PSU, éste logró analizar una gran cantidad de datos e información técnica y reunir y documentar esta información de una manera muy profesional.

El documento *Pruebas de Selección Universitaria: Proceso de Admisión 2004, Informe Técnico* ofrece un excelente ejemplo de lo que un informe técnico para un nuevo programa debe incluir. Documentos como *Manual de Procedimiento de Construcción de Preguntas o Itemes* y *Matriz Curricular de Lengua Castellana y Comunicación para la Elaboración de Itemes en la PSU de Lenguaje*, son ejemplos importantes que demuestran el cuidado que se ha tenido en el diseño, desarrollo y ensamblaje de las PSU.

Además, se ha desarrollado un gran número de folletos informativos que fueron distribuidos entre los usuarios de los puntajes, particularmente entre los estudiantes, proveyendo así de valiosa información acerca de la nueva evaluación. El DEMRE también proporcionó un examen de práctica completo para ayudar a los estudiantes a familiarizarse con las nuevas pruebas.

Áreas por Mejorar de las Pruebas de Lenguaje y Comunicación, y Matemáticas

Clarificación del Propósito de la Evaluación. Un número de documentos indica el propósito de las PSU. Varios de los documentos indican claramente que el propósito de las pruebas es para admisión universitaria. Sin embargo, otros documentos también dan a entender que existen usos secundarios de los puntajes de las PSU, tales como el uso de estos puntajes para evaluar la efectividad de programas educativos. Varios documentos producidos por el DEMRE presentan inferencias acerca de las habilidades cognitivas de los examinados que rindieron estas pruebas. Es muy importante que el DEMRE aclare todos los propósitos que

tienen las PSU. Si las PSU tienen propósitos múltiples, más allá de la admisión universitaria, entonces es imprescindible que las pruebas sean apropiadamente validadas para esos propósitos. Si las PSU son únicamente utilizadas para la admisión universitaria, entonces se deben desalentar otros usos de las pruebas, y los usuarios de los puntajes tendrían que ser debidamente informados de que los puntajes de las PSU no han sido validados para otros propósitos.

Desarrollo de Puntajes de Corte. Los puntajes en las PSU son usados, junto con las Notas de Enseñanza Media (NEM), para formar una medida compuesta ponderada. Los pesos o ponderaciones usados en una medida compuesta difieren de acuerdo a cada universidad y a la carrera dentro de la universidad. Los puntajes de corte, que varían de una universidad a otra y de una carrera a otra dentro de cada universidad, se forman usando una medida compuesta ponderada y también se usan para establecer un banco de examinados para ser admitidos a la universidad. Entendemos que el DEMRE no puede influir las decisiones de política que toman las Universidades en relación a los puntajes de corte. Sin embargo, se recomienda que el DEMRE brinde orientación técnica a las Universidades acerca del uso de los puntajes de corte y acerca de cómo validar el uso de los puntajes de corte. Además, es importante que el DEMRE brinde a las Universidades información de los puntajes de corte que sea consistente a través de los años. Para poder realizar esto, el DEMRE necesita implementar un plan de equiparación de puntajes y de equiparación de los delta que asegure que las pruebas que son administradas a través de los años sean de similar dificultad y que los puntajes en esas pruebas estén colocados en la misma escala.

Información Proporcionada a los Usuarios de los Puntajes. Tanto a los estudiantes como a las universidades se les proporciona un gran número de documentos que describen las pruebas, cómo se califican las mismas, y cómo se obtienen los puntajes escalados. Estos documentos no contienen mucha información técnica que pueda ser utilizada por los estudiantes o universidades para interpretar los puntajes obtenidos en las pruebas. Sería muy útil si se les suministrara información tal como: la confiabilidad de los puntajes, la dificultad de las pruebas, la clasificación de los puntajes en subgrupos y los rangos percentiles de los puntajes para una población relevante. Una gran cantidad de información es suministrada a los usuarios de los puntajes en el documento SIRPAES. Sin embargo, alguna de esta información podría ser interpretada equivocadamente, por lo que se insta al DEMRE a que brinde

suficientes advertencias a los usuarios de que la prueba aún no ha sido validada para propósitos de diagnóstico o para propósitos de evaluación de la efectividad de programas educativos.

El DEMRE debe ser elogiado por proporcionar una prueba de práctica completa a los estudiantes que rinden las PSU: la denominada prueba de ensayo. El DEMRE califica las pruebas de ensayo y les entrega a los estudiantes sus puntajes escalados. Un posible problema consiste en que los puntajes en la prueba de ensayo no son equiparados con los puntajes obtenidos en la prueba operacional, por lo cual no tienen el mismo significado. Por lo tanto, los estudiantes pueden de manera equivocada inferir que los puntajes que reciben en la prueba de ensayo reflejan los puntajes que recibirán al rendir la prueba operacional. Se recomienda al DEMRE que equipare los puntajes de la prueba de ensayo con los de la prueba real, o que se asegurare que los puntajes suministrados por la prueba de ensayo no puedan ser confundidos con los de la prueba real.

Planes para Validar las Evaluaciones. Las PSU son un nuevo programa de pruebas y siempre toma tiempo acumular evidencia de validez predictiva. Se recomienda al DEMRE comenzar inmediatamente a recolectar evidencia de la validez de los puntajes de las PSU, incluyendo evidencia de validez predictiva. Se insta al DEMRE a clarificar los propósitos y usos de las pruebas y entonces, basado en ese (esos) propósito(s), desarrollar un plan abarcativo para validar las pruebas para todos los propósitos que se les intente dar e iniciar el plan tan pronto como sea posible. Es importante que el DEMRE involucre a los usuarios de los puntajes en el proceso de validación, particularmente a las universidades. El plan de validación debería también incluir cómo se van a comunicar los resultados de los estudios de validez a las universidades, a los estudiantes y al público en general.

Grado de Adecuación de las Pruebas para el Nivel de Habilidad del Grupo. La prueba de Lenguaje y Comunicación parece estar a un nivel correcto de dificultad para el grupo que rinde esta prueba. Este no es el caso de la prueba de Matemáticas que resulta muy difícil para la población a quien va dirigida. Los puntajes corregidos de las pruebas son extremadamente asimétricos, resultando así en una distribución asimétrica con una gran acumulación de puntajes en el extremo inferior de la escala. La normalización de los puntajes cambió la forma de la distribución de los mismos, pero no remedió el problema de que la evaluación era demasiado difícil para la población a quien iba dirigida. El resultado de la normalización fue una escala de puntajes con algunas disparidades, indicando diferencias bastante grandes en habilidad, cuando la escala de puntajes sin la transformación no muestra esta diferenciación. En

otros casos, los puntajes que deberían haber sido reportados como puntajes separados se redondearon al mismo puntaje escalado. Se insta al DEMRE a examinar y a revisar las especificaciones de la prueba de Matemáticas para que la prueba coincida mejor con el nivel de habilidad de la población a quien va dirigida. Una vez que esto se realice, la prueba puede ser reescalada y la nueva escala debería proporcionar una mejor base para el reporte de los puntajes.

Equiparación de las PSU. El DEMRE había demostrado preocupación acerca de la comparabilidad de los puntajes en las formas múltiples de la misma prueba PSU y llevó a cabo procedimientos de equiparación para asegurarse de que los puntajes de estas formas fueran comparables. El DEMRE también desarrolló una nueva escala para las PSU utilizando diferentes procedimientos, incluyendo normalización, una forma de equiparación equipercentil y una transformación lineal a una escala con una media de 500 y una desviación estándar de 105. Desafortunadamente, el DEMRE reportó los puntajes usando los mismos números utilizados para reportar los puntajes de las PAA, aún cuando los puntajes de las PSU no habían sido equiparados a los puntajes de la PAA. Como resultado, el público esperaba que los puntajes de las PSU tuvieran el mismo significado que los puntajes de la PAA. Esta idea errónea resultó en una confusión considerable cuando se reportaron los puntajes de las nuevas pruebas. Se recomienda al DEMRE que no sólo equipare entre sí los puntajes de formas múltiples de una prueba rendida en la misma administración, sino que también equipare los puntajes de estas formas con respecto a pruebas administradas anteriormente. De esta manera, el DEMRE puede establecer una continuidad en la escala en que se reportan los puntajes y facilitar la interpretación de los puntajes obtenidos en las PSU. Esta continuidad de la escala ayudará al público en la interpretación de los puntajes obtenidos en las pruebas a través del tiempo, y también ayudará en la interpretación de los puntajes de corte que se usan para propósitos de admisión.

Diferencias entre Subgrupos. Las PSU exhiben diferencias de desempeño para varios subgrupos. Este comportamiento en sí mismo no causa necesariamente un problema. Sin embargo, sería muy útil para el DEMRE desarrollar un plan amplio de equidad que incluya la evaluación de Funcionamiento Diferencial del Item (DIF, por sus siglas en inglés: Differential Item Functioning) para todos los subgrupos relevantes, y también planes acerca de cómo los hallazgos de DIF podrían utilizarse para diseñar y desarrollar nuevas formas de las pruebas.

Recomendaciones

La siguiente es una lista de recomendaciones para mejorar la validez de los puntajes para el programa de las PSU. Esperamos que esta lista provea un conjunto provechoso de metas para mejoras al programa en el corto y largo plazo.

- Es importante que el DEMRE trabaje con la Universidad de Chile y otros usuarios de los puntajes para clarificar todos los usos que se intenten dar a los puntajes de las PSU.
- El DEMRE debería brindar asistencia técnica y orientación a las universidades en cuanto al establecimiento y la validación de los puntajes de corte.
- Se debería proveer a los usuarios de materiales adicionales para la interpretación de puntajes, tales como las propiedades psicométricas de las pruebas y los rangos percentiles para los grupos correspondientes.
- Los puntajes de la prueba de ensayo deberían ser equiparados con los puntajes de la prueba operacional, o bien, deberían ser reportados utilizando números que no puedan ser confundidos con los números usados para reportar los puntajes de la prueba operacional.
- El DEMRE debería desarrollar un plan abarcativo para validar las PSU que incluya estudios para validar los puntajes de las pruebas para todos los propósitos que se les intente dar.
- La PSU–Matemáticas parece ser demasiado difícil para el grupo de estudiantes que rinde esta prueba. Se recomienda que el DEMRE revise las especificaciones de la prueba para que así concuerden mejor con el nivel de habilidad de la población a quien va dirigida.
- Actualmente el DEMRE equipara los puntajes de varias formas de una misma prueba rendida en una misma administración. Se recomienda que las pruebas sean equiparadas entre administraciones para proveer continuidad en la interpretación de los puntajes obtenidos en estas pruebas.
- Las PSU exhiben diferencias en los puntajes de los subgrupos principales de la población que rinde las pruebas. Se recomienda que el DEMRE desarrolle un plan de equidad que incluya un análisis DIF para todos los subgrupos principales

donde el tamaño de muestra sea suficiente, y que los resultados de estos análisis sean considerados durante el proceso de desarrollo de las pruebas.

Revisión de la Confiabilidad de las Pruebas de Lenguaje y Comunicación, y Matemáticas

Los Estándares de Confiabilidad

El propósito de los estándares de confiabilidad es ayudar a asegurar que los puntajes u otros resultados reportados de las evaluaciones serán suficientemente confiables para lograr sus propósitos, y que el programa está utilizando procedimientos apropiados para determinar y reportar la confiabilidad. Junto con la validez, la confiabilidad es uno de los dos atributos principales de una evaluación. Más aún, la confiabilidad es una condición necesaria para la validez. Si una prueba no exhibe un grado aceptable de confiabilidad para el propósito que persigue, no tendrá tampoco un grado aceptable de validez. De esta manera, la confiabilidad es una condición necesaria pero no suficiente para la validez. Consecuentemente, es importante que el programa de evaluación suministre tanto evidencia lógica como empírica, mostrando que la prueba es capaz de lograr un grado aceptable de precisión en la medición para el propósito establecido. Aunque se pueden usar muchos tipos diferentes de evidencia para demostrar la confiabilidad de una evaluación, es muy importante que se brinde evidencia que sustente la confiabilidad para el (los) propósito (s) que persigue la evaluación.

El conjunto de estándares de confiabilidad se compone de los siguientes seis estándares:

- Todos los puntajes reportados, incluyendo los sub-puntajes o puntajes desagregados, deben ser suficientemente confiables para apoyar las interpretaciones que se intenta hacer con ellos.
- Los métodos usados para estimar la confiabilidad deben ser apropiados para el uso que se intenta hacer de los puntajes.
- El programa de pruebas debe: suministrar información que permita a los usuarios de los puntajes juzgar si los puntajes son suficientemente confiables para respaldar las interpretaciones que se intentan hacer; incluir estimaciones de los errores estándar de medición; si es apropiado, suministrar información sobre el error estándar de medición y la confiabilidad de diferencias; reportar consistencia de promover/reprobar para los puntajes de corte; suministrar información acerca del error estándar de medición, u otros coeficientes similares, alrededor del puntaje de corte.

- Suministrar suficiente información acerca de los análisis de confiabilidad para permitir que personas conocedoras evalúen los resultados y repliquen los análisis.
- Ejecutar análisis separados de confiabilidad cuando se hayan hecho modificaciones significativas a la evaluación o a los procedimientos de administración o calificación.
- Estudiar la confiabilidad y los errores estándar de medición de los puntajes reportados para subgrupos de la población.

La confiabilidad de las PSU de Lenguaje y Comunicación y de Matemáticas fue evaluada utilizando cada uno de los seis estándares que se mencionaron anteriormente. Una descripción detallada de la evaluación de la confiabilidad de cada una de las pruebas se encuentra en el apéndice de este informe. El siguiente es un resumen de las fortalezas, las áreas por mejorar y las recomendaciones para hacer que las prácticas de las PSU se alineen mejor con los seis estándares de confiabilidad.

Fortalezas de las PSU de Lenguaje y Comunicación y de Matemáticas

A partir de los documentos suministrados por el DEMRE y sus intercambios con el equipo de evaluación, es evidente que el personal técnico a cargo de las PSU tiene buena capacitación y posee un buen entendimiento de los conceptos y procedimientos sobre confiabilidad, especialmente aquellos derivados de la teoría clásica de pruebas. En la presentación titulada *Proceso de Admisión a Las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados* se suministraron los coeficientes de confiabilidad para todas las pruebas, basados en la teoría clásica de pruebas. Las confiabilidades reportadas en ese documento son apropiadas para propósitos de admisión. Se determinaron coeficientes de confiabilidad por el método de las mitades y se realizaron análisis de confiabilidad con el alfa de Cronbach para las formas únicas de la prueba de Matemáticas y de la prueba de Lenguaje y Comunicación. Si las pruebas no están afectadas por la velocidad, estos métodos son apropiados para el uso que se pretende hacer de los puntajes. Los errores estándar de medición se reportan para todas las pruebas.

El DEMRE también ha demostrado ser conciente de la necesidad de evaluar el efecto de la velocidad de las evaluaciones, aún cuando el procedimiento que se ha utilizado hasta ahora para evaluar este efecto debería ser modificado. En relación con este punto, el DEMRE ha

afirmado que tiene la información y las herramientas para implementar un procedimiento que pueda separar los ítemes "no alcanzados" de los ítemes omitidos para evaluar el grado en que las pruebas están afectadas por la velocidad.

Otra fortaleza de este programa de pruebas es el esfuerzo llevado a cabo para suministrar información acerca de la confiabilidad y el error estándar de medición de los puntajes en documentos que están disponibles a los usuarios de las pruebas.

El documento técnico titulado *Resultados de la Aplicación de Pruebas de Selección Universitaria Admisión 2004* es un excelente ejemplo de una publicación que pone información acerca de las pruebas a disposición de los usuarios, esfuerzo que debería ser continuado y extendido.

El DEMRE tiene la información básica y las herramientas para llevar a cabo análisis separados de confiabilidad para subgrupos significativos de la población. Existe información para ejecutar análisis de confiabilidad separados de acuerdo al género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público y privada sin financiamiento público). Esta es una importante fortaleza técnica que se alienta al DEMRE a ejercer.

Finalmente, en relación con de la implementación de los métodos IRT (Teoría de Respuesta al Item) como un complemento a los análisis de confiabilidad con la teoría clásica, el DEMRE también ha hecho progresos significativos en cuanto a la comprensión de los conceptos, usos y procedimientos asociados a la IRT.

Áreas por Mejorar de las PSU de Lenguaje y Comunicación, y Matemáticas

Añadir Modelos de IRT y de Equiparación de Deltas. Será útil complementar los análisis de confiabilidad de la teoría clásica con análisis IRT. El uso de los modelos IRT brindará información acerca del poder de la medición de la prueba en puntos críticos de la distribución de puntajes. Se sugiere que el DEMRE use IRT para establecer una función de información meta (junto con su función de error de medición) para determinar las especificaciones estadísticas para los ítemes y para las pruebas en su totalidad. La Teoría de Respuesta al Item puede ser utilizada para determinar una distribución de la dificultad de los ítemes que concuerde con la distribución de la habilidad de la población a quien van dirigidos, y que brinde máximo poder de medición en puntos deseados en la distribución de habilidad de la población. Es importante que el DEMRE considere desarrollar especificaciones estadísticas

para las pruebas utilizando procedimientos de IRT para establecer especificaciones para las mismas que presenten una buena concordancia con la distribución de habilidad de los examinados que las rinden.

Además, existe la necesidad de establecer un proceso de pretest que brinde estadísticas estables de los ítemes que sean comparables entre grupos de pretest (deltas equiparados trabajarían bien en esta situación). Los procedimientos de pretest deberían ser suficientemente robustos para construir un banco de ítemes que sea adecuado para sustentar un proceso de desarrollo de calidad para las pruebas.

La situación descrita anteriormente, claramente no sucedió en el caso de la prueba de Matemáticas introducida en el 2003, ya que resultó bastante difícil para la población a quien fue dirigida. Las estimaciones globales de confiabilidad de la evaluación podrían posiblemente ser engañosas, dado que la prueba es obviamente muy difícil para la población. Consecuentemente, a la prueba le falta poder de medición en los rangos medios y bajos de puntaje. La normalización de los puntajes no puede ser un remedio para este problema. El nivel de dificultad de esta prueba debería ser modificado para que se alinee mejor con el nivel de habilidad de la población. Como se explicó anteriormente, este objetivo puede ser logrado con la ayuda de modelos IRT. La equiparación de deltas también se recomienda para controlar diferencias de habilidad entre los grupos de pretest que son utilizados para generar las estadísticas de los ítemes para el ensamblaje operacional de la prueba y el grupo que en realidad rinde la prueba operacional.

Un Procedimiento para Evaluar el Efecto de la Velocidad en las Pruebas. Hasta ahora el DEMRE ha evaluado el efecto de la velocidad en las evaluaciones examinando el número de omitidos para los ítemes cercanos al final de la prueba. Sin embargo, debería emplearse un procedimiento que pueda separar los ítemes "no alcanzados" de los ítemes omitidos para evaluar el efecto de la velocidad. El DEMRE ha afirmado que tiene la información y las herramientas para implementar tal procedimiento.

Un Plan de Equiparación Claro Para Hacer los Puntajes de las PSU Equiparables de Año a Año. Para poder preservar el mismo significado de los puntajes de las pruebas (incluyendo los puntajes de corte) de año a año, se necesita un diseñar plan de equiparación. Los puntajes en las pruebas del año actual necesitarán ser equiparados con los puntajes de las pruebas del año anterior. Los datos recolectados en la administración 2003 pueden no ser apropiados para establecer la escala inicial de la prueba. La razón para esto es que el grupo que

rindió la prueba en la administración inicial puede no haber estado adecuadamente preparado para rendir la nueva prueba y consecuentemente puede no ser representativo de la población a quien va dirigida. Un indicador de que este ha sido el caso es la extremadamente asimétrica distribución de puntajes que fue obtenida por este grupo en la prueba de Matemática.

Producción de un Documento Escrito Dirigido a Estudiantes y Padres. Debería escribirse un documento en un lenguaje accesible y dirigido a los estudiantes, padres y el público en general, explicando la naturaleza de los puntajes y sus errores de medición (incluyendo los puntajes de corte), junto con su correcta interpretación en términos del modelo con referencia a normas sobre el que se basa la construcción y el uso de las pruebas.

Documentación de los Análisis de Confiabilidad. Es importante documentar todos los aspectos de los análisis de confiabilidad, tales como la justificación para la escogencia del método, la muestra utilizada para ejecutar los análisis y cualquier otra información que ayudaría a una persona conocedora a evaluar los análisis.

Análisis de Confiabilidad para Subgrupos. Una manera de asegurar de que la prueba es equitativa para todos los subgrupos significativos de la población, es asegurar que brinda puntajes confiables para estos subgrupos. Esto se puede lograr implementado un procedimiento para llevar a cabo rutinariamente análisis de confiabilidad para subgrupos de la población, cuando sea posible. Al menos dos variables podrían ser consideradas inicialmente para estos análisis de confiabilidad de subgrupos: género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público). También es importante que el DEMRE reflexione sobre la posibilidad de llevar a cabo análisis con otras subpoblaciones posibles tales como las que se definen por región y por zona urbana o rural.

Recomendaciones

La siguiente es una lista de recomendaciones para mejorar la confiabilidad del programa de pruebas PSU de acuerdo con lo especificado en los seis estándares de confiabilidad. Creemos que todas nuestras recomendaciones podrían ser logradas por el DEMRE en el corto y mediano plazo, si se le brinda al DEMRE los recursos necesarios.

- El DEMRE debería diseñar un plan para implementar el uso de modelos IRT y de equiparación de deltas en los procesos de construcción de sus pruebas y en sus análisis de confiabilidad.
- Implementación de un procedimiento para evaluar el grado en que las pruebas están afectadas por la velocidad que pueda separar los ítemes "no alcanzados" de los ítemes omitidos.
- Diseño de un plan de equiparación por parte del DEMRE para hacer los puntajes de las PSU equiparables de año a año.
- Producción de un documento escrito dirigido a estudiantes, padres y al público en general, usando lenguaje accesible, explicando la naturaleza de los puntajes y sus errores de medición (incluyendo los puntajes de corte), junto con su correcta interpretación en términos del modelo con referencia a normas sobre el que se basa la construcción y el uso de las pruebas.
- Producción de un sólo documento escrito que describa todos los aspectos de los análisis de confiabilidad, tales como la justificación para la escogencia del método, la muestra utilizada para lleva a cabo los análisis y cualquier otra información que ayudara a una persona conocedora a evaluar los análisis.
- La adición de análisis de confiabilidad para subgrupos como parte de los análisis psicométricos regulares de las pruebas, comenzando con género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público). También se recomienda al DEMRE que considere llevar a cabo análisis en otras posibles subpoblaciones, tales como las que se definen por región y por zona urbana o rural.

Revisión de los Usos de la Prueba de Lenguaje y Comunicación, y de Matemáticas

Estándares Sobre los Usos de las Evaluaciones

El propósito de los estándares sobre los usos de las evaluaciones es ayudar a asegurar que el programa de pruebas suministre información que describa y promueva el uso apropiado de las mismas y que advierta a los usuarios que eviten usos equivocados comunes de estos resultados. Estos estándares recomiendan a los usuarios de los resultados utilizar las pruebas de una manera justa y apropiada, de acuerdo con evidencia que respalde sus usos. El conjunto de estándares sobre los usos de las evaluaciones consiste de los siguientes cuatro estándares:

- Proveer a los usuarios de las pruebas de la información necesaria para evaluar su adecuación y de la oportunidad de consultar con el personal del programa de pruebas sobre los usos apropiados de las mismas. La información debe incluir los propósitos y la población; el contenido y el formato; la dificultad, la confiabilidad y la validez; la disponibilidad y la aplicabilidad de datos normativos; los requerimientos de administración y puntuación; las políticas de retención y publicación de datos; y las investigaciones representativas y relevantes.
- Hacer un uso apropiado de las pruebas e interpretaciones adecuadas de los resultados de las mismas a través de acciones tales como informar a los usuarios de las pruebas sobre: cómo utilizar los resultados y sobre cómo evaluar diferencias de puntajes entre individuos; posibles explicaciones de desempeños pobres de los examinados;la necesidad de ofrecer a los examinados un número razonable de oportunidades de tener éxito en las pruebas; y la necesidad de familiarizarse con sus responsabilidades, tales como aparecen descritas en *Standards for Educational and Psychological Testing* (APA, 1999). Advertir a los usuarios de evitar usos equivocados comunes que sean razonablemente predecibles, precaviéndoles sobre posibles problemas e informándoles sobre acciones tendientes a evitar dichos problemas.
- Investigar alegatos creíbles de usos equivocados de las evaluaciones, cuando sea posible. Si se encuentra un uso equivocado de las evaluaciones, informar al cliente y al

usuario del uso apropiado de las mismas. Si el uso equivocado continúa, consultar con el cliente acerca de acciones correctivas apropiadas, las cuales pueden incluir no entregar los puntajes u otros resultados de las pruebas a aquellos usuarios que persistan en un uso equivocado perjudicial.

- Proveer información y consejo para ayudar a las partes interesadas a evaluar la adecuación, la utilidad, y las consecuencias de las decisiones tomadas en base a los puntajes y a otros resultados de las pruebas. Los usuarios de los resultados, los individuos e instituciones encargados de políticas educacionales y los clientes se encuentran entre las partes interesadas a considerar. Información relevante y creíble acerca de posibles efectos que devengan de los usos de las evaluaciones, puede representar una gran ayuda a los encargados de políticas educacionales que estén considerando requerir las evaluaciones para algún propósito.

El uso de la PSU-Lenguaje y Comunicación y la PSU-Matemáticas fue evaluado utilizando cada uno de los cuatro estándares mencionados anteriormente. En el apéndice de este informe se encuentra una descripción detallada de la evaluación sobre el uso de estas pruebas. A continuación se resumen las fortalezas, las áreas por mejorar, y las recomendaciones para acrecentar la alineación de las prácticas de las PSU con los cuatro estándares sobre el uso de las evaluaciones.

Fortalezas de la PSU-Lenguaje y Comunicación y la PSU-Matemáticas

- El DEMRE ha hecho un trabajo extraordinario para proveer a los usuarios de las pruebas de información clara sobre la población, el contenido, y el formato de estas pruebas; sobre los requerimientos de admisión para cada programa de pregrado de las Universidades del H. Consejo de Rectores, incluyendo las ponderaciones asignadas a cada prueba y al NEM para calcular el puntaje compuesto final de los candidatos para su admisión; y sobre el proceso de admisión a las Universidades del H. Consejo de Rectores en general, a través de varias publicaciones, tales como *Documentos Oficiales* (publicados antes de las administraciones del 2003 y 2004), y *Análisis de los Resultados* (publicado después de la administración del 2003).

Además el DEMRE ha llevado a cabo varios talleres informativos a través del país. En estos talleres se provee de información y consejo a los usuarios de las pruebas para ayudar a evaluar la adecuación, la utilidad y las consecuencias de las decisiones tomadas en base a los resultados de las mismas. Para el proceso de admisión del 2005, el DEMRE ha implementado un programa extensivo de publicaciones informativas, a través de un periódico de alta circulación nacional. El DEMRE ha hecho esfuerzos muy significativos para mejorar la comunicación de la información antes mencionada. El uso de la tecnología para proveer servicios y para facilitar el acceso a la información a los usuarios de las pruebas ha sido un complemento muy importante a los esfuerzos de comunicación del programa. Los usuarios de los resultados de las pruebas, las instituciones e individuos encargados de políticas educativas y los clientes en general han sido provistos de acceso al personal del DEMRE a través de la "Mesa de ayuda" y los "Portales", los cuales se encuentran disponibles en la página de Internet del DEMRE.

- Se han desarrollado, y se encuentran disponibles para los usuarios de las pruebas, un número de documentos valiosos que contienen información sobre los estándares curriculares, los contenidos y las habilidades medidas por las PSU *(Documentos Oficiales Proceso de Admisión 2004 y 2005; Matriz Curricular de Lengua Castellana y Comunicación para la Elaboración de Itemes de la PSU de Lenguaje; Matriz Curricular para la Elaboración de Itemes de la PSU de Matemática).*
- El DEMRE ha desarrollado un informe técnico para las Universidades del H. Consejo de Rectores, el cual contiene información sobre los resultados de las PSU. Este informe será distribuido anualmente a las 25 Universidades del H. Consejo de Rectores.
- Toda la información mencionada anteriormente es provista a un nivel que puede ser entendida por los usuarios e incluye la información necesaria para contactar al personal del DEMRE para consultas sobre las pruebas.
- Las pruebas de ensayo representan una oportunidad valiosa para que los estudiantes se familiaricen con el formato y los contenidos de las pruebas.
- El DEMRE ha identificado claramente a todos los usuarios y todos los usos de las PSU.

- El documento "SIRPAES" también provee de información valiosa a instituciones educativas. En su presentación preliminar, el documento fomenta un uso apropiado de los resultados de las PSU.

Áreas por Mejorar Sobre el Uso de la PSU-Lenguaje y Comunicación y la PSU-Matemática

Clarificación del (los) Propósito (s) de las Pruebas. Como se mencionó anteriormente en este informe, los propósitos previstos de las PSU son claramente establecidos en varios documentos, esto es, que las PSU deben ser utilizadas únicamente para propósitos de admisión a la universidad *(Antecedentes Generales, Documentos Oficiales, Modelo de Documento SIRPAES-Presentación Preliminar).* Sin embargo, un número importante de documentos también dan a entender que existen usos secundarios de las PSU, tales como evaluación profesoral y evaluación de la efectividad de programas educativos. Varios documentos producidos por el DEMRE proveen inferencias, basadas en los resultados de las pruebas, sobre habilidades cognitivas de los examinados. Es fundamental que el DEMRE clarifique todos los propósitos de las pruebas. Como también se mencionó anteriormente, si las pruebas tienen propósitos múltiples, más allá de la admisión a la universidad, entonces es imperativo que las pruebas sean validadas para tales propósitos. Si las pruebas se utilizan únicamente para propósitos de admisión a la universidad, entonces los otros usos de las pruebas deben ser desalentados y los usuarios de los resultados de las pruebas deben ser informados de que los resultados de las pruebas no han sido validados para estos propósitos adicionales.

Planes para Validar la Prueba y Comunicación de los Resultados de las Validaciones. Como también se mencionó anteriormente en este informe, las PSU son un programa de evaluación nuevo y siempre toma tiempo acumular evidencia de validez predictiva. Sin embargo, parte de la evidencia de validez predictiva así como otros tipos de validez ya se pueden comenzar a recolectar. Se insta al DEMRE a desarrollar un plan amplio para validar la prueba para todos los usos que se le intente dar. Es importante que el DEMRE involucre a los usuarios de los resultados de las pruebas en el proceso de validación, particularmente a las universidades. El plan de validación debería incluir cómo serán comunicados los resultados de los estudios a las universidades, a los estudiantes y al público en general.

La Información que se Provee a los Usuarios de los Resultados. A los estudiantes y a las universidades se les provee de un gran número de documentos que describen las pruebas,

cómo se califican las mismas, y cómo son obtenidos los puntajes a escala. Como fue mencionado anteriormente, estos documentos no contienen una gran cantidad de información técnica que pueda ser utilizada por los estudiantes o las universidades para interpretar los resultados de las pruebas. Sería de mucha utilidad si se proveyera información tal como la confiabilidad de los puntajes, la dificultad de la prueba, un análisis de los resultados por subgrupo y los rangos percentiles para una población relevante.

Políticas para la Retención de Datos. El DEMRE ha formulado su política para una diligente publicación de los resultados de las pruebas y ha aplicado esta política consistentemente a través del tiempo. El DEMRE, sin embargo, debería comunicar a todos los usuarios de los resultados de las pruebas cuál es su política para la retención de datos. Debería desarrollarse un conjunto claro de pautas acerca de la duración de la retención de los registros de los individuos, y de la disponibilidad y utilización de esos registros a través del tiempo.

Información de Investigación Relevante. Se recomienda al DEMRE que comunique toda información de investigación relevante tan pronto como ésta esté disponible. Referencias a cualquier tipo de material que provea más detalles acerca de investigaciones realizadas por el DEMRE o por investigadores independientes, deberían ser citadas en los documentos que se publiquen y deberían estar fácilmente disponibles para los usuarios y revisores de las pruebas.

Interpretación de los Resultados. Se insta al DEMRE a que desarrolle una guía para la interpretación de los resultados. La información que se provea en la guía debería estar presentada a tal nivel que sea factible de ser comprendida por todos los usuarios. También debe incluirse la información necesaria para que los usuarios de las pruebas puedan comunicarse con el personal del DEMRE. La guía debería informar a los usuarios acerca de la adecuación, utilidad y consecuencias de las decisiones tomadas en base a los resultados de las PSU. Se recomienda al DEMRE establecer una política para evitar el uso de resultados que hayan quedado obsoletos e incorporar información acerca de esta política en la guía.

Diferencias en los Resultados y Posibles Explicaciones para Desempeños Pobres. Las PSU exhiben diferencias en el desempeño para varios subgrupos. Los usuarios de los resultados se podrían beneficiar de tener acceso a información que provea posibles explicaciones para las diferencias entre subgrupos y para desempeños pobres, y que haga mención acerca de los múltiples factores que pueden afectar los resultados de las pruebas.

Las Oportunidades de los Estudiantes de Tener Éxito. Como se ha mencionado anteriormente, al no ser los puntajes de la prueba de ensayo equiparados con los resultados de

la prueba real, éstos no tienen el mismo significado. Consecuentemente, los estudiantes pueden erróneamente inferir que los resultados que obtengan en la prueba de ensayo sean similares a los resultados que obtengan posteriormente en la prueba real. Se recomienda al DEMRE que equipare los puntajes de la prueba de ensayo con los de la prueba real, o bien que reporte puntajes de la prueba de ensayo que no puedan ser confundidos con los de la prueba real. El uso diferenciado de estrategias para la toma de pruebas que no estén relacionadas con el dominio que se está midiendo, y que se encuentren que mejoran o afectan el desempeño en las pruebas, podría afectar la validez y confiabilidad de la interpretación de los resultados. Antes de la administración de las pruebas, debería proporcionarse a todos aquellos que las rendirán, información acerca de estrategias para la toma de pruebas e información acerca de la interpretación apropiada de los puntajes de la prueba de ensayo.

Responsabilidades de los Usuarios de las Pruebas. Se recomienda que el DEMRE informe a los usuarios de las pruebas acerca de la necesidad de que se familiaricen con sus responsabilidades como usuarios de evaluaciones tal como están descriptas en *Standards for Educational and Psychological Testing* (APA, 1999).

Comunicación con los Usuarios y con el Público en General. Como los resultados de las PSU son dados a conocer al público y a aquellos que se encargan de desarrollar políticas, el DEMRE debería proveer y explicar cualquier información adicional que minimice posibles interpretaciones inadecuadas de los resultados. Brindar información preliminar con anterioridad a la publicación de los resultados de las pruebas, daría a los medios de comunicación la oportunidad de asimilar datos relevantes y evitar posibles interpretaciones incorrectas de los resultados de las PSU y posibles consecuencias negativas no intencionadas.

Recomendaciones.

La siguiente es una lista de recomendaciones para el perfeccionamiento del programa de evaluación PSU. Entendemos que el DEMRE tiene recursos limitados y tal vez no pueda tratar la lista completa que estamos proporcionando aquí, pero esperamos que ésta brinde un conjunto útil de metas para mejoras del programa a corto y largo plazo.

- El DEMRE debería desarrollar una guía para la interpretación de los resultados. Esta guía debería recordar a los usuarios acerca del (los) propósito (s) de las PSU y del uso adecuado de las pruebas, y proveer suficiente información y asesoramiento para ayudar a todos los usuarios de las PSU a evaluar la adecuación, utilidad y consecuencias de las decisiones que se tomen en base a los resultados de la prueba. El DEMRE puede recomendar a los usuarios de las PSU que verifiquen periódicamente que sus interpretaciones de los datos de las pruebas continúen siendo adecuadas, dado que pueden surgir cambios significativos en la población que toma la prueba, en la administración de la misma, y en sus propósitos. Las guías desarrolladas para las pruebas SAT y ACT podrían ser utilizadas como un modelo para el desarrollo de la guía que aquí se propone.
- Se recomienda al DEMRE que desarrolle un plan de comunicación que ayude a educar a todos los usuarios, usuarios potenciales y al público en general acerca de las PSU como pruebas de gran escala. El DEMRE debería advertir a los usuarios de los resultados de las PSU que como tales tienen la responsabilidad de informarse adecuadamente acerca de los usos apropiados de las pruebas.
- El DEMRE debería desalentar activamente usos de los resultados de las PSU que carezcan de evidencia que los respalde. Si una determinada apreciación profesional conduce a un uso de las PSU para el cual existe poca documentación de su validez, el DEMRE debería advertir al usuario que interprete los resultados cautelosamente y que tenga cuidado de no dar a entender que las decisiones o inferencias que se hagan en base a los resultados de las pruebas están bien documentadas con respecto a confiabilidad y validez. En esos casos, el DEMRE debería requerir al usuario que reúna la evidencia de validez necesaria que respalde el uso procurado.
- El DEMRE debería promover estudios de investigación relacionados con las PSU. Los resultados de estos estudios deben ser publicados tan pronto como estén disponibles.
- El DEMRE debería proveer a todos los que tomen la prueba de iguales oportunidades para que se desempeñen óptimamente. Antes de la administración de las pruebas, debería proporcionarse a todos aquellos que las rendirán, información acerca de estrategias para la toma de pruebas e información acerca de la interpretación apropiada de los puntajes de la prueba de ensayo.

- El DEMRE debería comunicar a todos los usuarios de las pruebas sobre su política de retención de datos, así como también sobre su política de caducación de los resultados.

RESUMEN Y RECOMENDACIONES

Para resumir, el programa de pruebas PSU tiene un número importante de fortalezas. Creemos que una de las mayores fortalezas del programa es el personal altamente calificado y dedicado del DEMRE. El personal del DEMRE ha trabajado muy arduamente para implementar un nuevo programa de pruebas en un período de tiempo muy corto y con recursos muy limitados, haciendo su trabajo de una manera sumamente profesional.

Como en cualquier programa de pruebas nuevo, hay varias áreas en las que pueden hacerse mejoras. Hemos desarrollado una lista comprensiva de estas áreas como parte de las recomendaciones incluidas en cada sección del informe. Quisiéramos reiterar aquí algunas de las recomendaciones más esenciales. Creemos que es muy importante que el DEMRE cuente con los recursos y el apoyo suficientes para lograr todas las mejoras recomendadas en este informe. Sin embargo, creemos que las siguientes mejoras deberían ser una prioridad para el DEMRE:

- Es esencial que todos los propósitos y usos de los resultados de las PSU sean clarificados y que se desarrolle un plan para reunir evidencia para validar los puntajes de la evaluación para todos los propósitos que se les intente dar a las pruebas.
- El DEMRE debería desarrollar un plan de equiparación que incluya la equiparación de todas las formas de la evaluación pasadas y presentes, incluso cualquier prueba de práctica (prueba de ensayo).
- Se recomienda al DEMRE desarrollar un plan para el pretest que incluya un método (tal como la equiparación de deltas) para colocar todas las estadísticas de los ítemes en una misma escala.
- El DEMRE debería considerar el uso de métodos IRT para desarrollar especificaciones estadísticas para todas las pruebas con niveles de dificultad que sean apropiados para los niveles de habilidad de los estudiantes que las rinden. Esto es particularmente importante para la prueba de Matemáticas.
- El DEMRE debería desarrollar un plan comprensivo de equidad que incluya la evaluación de la confiabilidad de la prueba para distintos subgrupos, así como también un análisis de DIF.

- Se recomienda al DEMRE realizar investigaciones sobre las pruebas y que ponga los resultados de estas investigaciones a disponibilidad del público.

Nuestra recomendación final es que se provea al DEMRE con el tiempo, el personal y los recursos para lograr llevar adelante las recomendaciones que se brindan en este informe. Es claro para nosotros que el personal del DEMRE tiene gran cantidad de responsabilidades en estos momentos, y si va a haber un progreso real en lo relacionado a las recomendaciones de este informe, su personal con experiencia debe ser aliviado de algunas de sus responsabilidades presentes para que pueda trabajar en estas recomendaciones. Por consiguiente, creemos que al personal actual del DEMRE debe brindársele capacitación en servicio y que el personal necesitará ser incrementado para que se puedan llevar a cabo las recomendaciones brindadas en este informe. Esperamos que esto sea posible en un futuro cercano.

REFERENCIAS

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Tests.* Washington, DC: American Psychological Association.

ETS Standards for Quality and Fairness. Princeton, NJ: Educational Testing Service.

APENDICE

Revisión Técnica de la PSU-Lenguaje y Comunicación

Validez

Asegurar que los programas recolecten y documenten apropiadamente la evidencia que apoye las inferencias que intentan hacer y las acciones basadas en los resultados reportados de las evaluaciones.

Estándares		Resultados de la Revisión
1. ¿Existe una descripción de cómo la evaluación se ajusta a la misión y los valores del programa?		• Existe una serie de documentos en donde se establece el propósito de las PSU. (Ver, por ejemplo, *Información General, Validez para las dos pruebas: Lenguaje y Comunicación, Matemáticas; Compendio Estadístico Proceso de Admisión Año Académico 2004, y Proceso de Admisión a las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados*). También se suministra información acerca del propósito de las pruebas en folletos que se distribuyen a los estudiantes y a las universidades. Esta información claramente establece que los usuarios de las pruebas son las universidades que pertenecen al H. Consejo de Rectores y los estudiantes. Varios de los documentos también establecen claramente que el propósito primario de las pruebas es seleccionar candidatos para admisión a las 25 universidades del H. Consejo de Rectores. Sin embargo algunos de los documentos también dan a entender que hay usos secundarios para los puntajes de las PSU. • Algunos informes (Ver *Proceso de Admisión a las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados y Análisis de Resultados, Junio de 2004*) incluyen un análisis de cómo se desempeñaron los estudiantes en diferentes áreas de destrezas y contenido de las PSU. En varios documentos suministrados por el programa de pruebas, se brinda información y se hacen inferencias acerca de las destrezas cognitivas de los estudiantes que rinden las pruebas. • Debería tenerse cuidado de no interpretar los resultados de la prueba para ningún otro propósito (como evaluación de programas o diagnósticos) que no sea el propósito que se ha establecido públicamente: admisión a las universidades. A los usuarios de los puntajes, universidades y estudiantes, se les debería advertir que no interpreten los puntajes de las PSU como útiles para ningún otro propósito que no sea el de un indicador de la habilidad del estudiante para ser exitoso en una universidad, a menos que la prueba sea validada para estos propósitos adicionales. Si se decide que los puntajes de

		las PSU serán usados para otros propósitos además de las admisiones a la universidad, por ejemplo, evaluación de la efectividad de las escuelas o profesores, entonces debe recolectarse evidencia de la validez de la prueba para este propósito. Hasta que tal evidencia sea reunida, se debería advertir a los usuarios de los puntajes para que no usen los puntajes para ningún otro propósito que no sea la admisión a la universidad. • Las PSU se usan también para conceder becas a la universidad. • Los puntajes de las pruebas PSU se usan, junto con la NEM, para formar una medida compuesta ponderada. Las ponderaciones usadas en la medida compuesta difieren de acuerdo con la universidad y la carrera dentro de la universidad. Los puntajes de corte, que varían entre universidades y entre carreras dentro de universidades, se forman usando la medida compuesta ponderada y se usan para establecer un banco de examinados para admisión a la universidad. Se recomienda al DEMRE brindar orientación técnica a las Universidades para asistirlas en el establecimiento de los puntajes de corte y en la validación de los puntajes de corte, una vez que se han establecido. Además, es esencial que el DEMRE diseñe e implemente planes para la equiparación de los delta y equiparación de puntajes, de manera que puedan suministrar a las Universidades puntajes de corte consistentes a través de los años.
2. ¿Se describen claramente las siguientes piezas de información: -el constructo, es decir, el conocimiento, destrezas o características a ser medidas? -el propósito de la evaluación? -las interpretaciones que se intentan hacer a partir de los puntajes u otros resultados de la evaluación? -las poblaciones a quienes van dirigidas las pruebas? ¿Está esta información disponible al público en caso de ser solicitada?		• Existe un documento, *La PSU de Lenguaje y Comunicación y su relación con el Marco Curricular de Lengua Castellana y Comunicación* que describe las tres secciones de la prueba y brinda una justificación para el contenido de la prueba. Las diferencias entre la PAA-V y L y C también se describen en este documento. Además, existe una matriz curricular para el desarrollo de los ítemes de la prueba de Lenguaje y Comunicación que suministra un marco de referencia de contenido para referenciar cruzadamente los ítemes con el currículo. • Los documentos referidos anteriormente brindan información acerca del propósito de la prueba. Otros documentos que se suministran a los estudiantes describen cómo se califica la prueba y brindan un cuadro que muestra cómo los puntajes corregidos se convierten a puntajes estándar. A los estudiantes también se les brindan pruebas de ensayo. • Un problema con las pruebas de ensayo es que los puntajes derivados de estas pruebas no están en la misma escala que los puntajes obtenidos en las PSU operacionales. Consecuentemente, los estudiantes podrían interpretar que los puntajes que ellos reciben en la prueba de práctica son indicadores de los puntajes que recibirán en la prueba operacional. Ya que los puntajes de la prueba de ensayo no están en la misma escala que los puntajes de la prueba

		operacional, esta interpretación puede ser engañosa. Si los puntajes en la prueba de ensayo son suministrados usando los mismos números de la escala de los puntajes de la prueba operacional, es imperativo que los puntajes en la prueba de ensayo sean equiparados a los puntajes de la prueba operacional. • Aunque una gran cantidad de información se suministra a los estudiantes y a las universidades acerca de las pruebas, se brinda muy poca información técnica que pueda ser utilizada para interpretar los puntajes. Es importante suministrar información interpretativa a los estudiantes y a las universidades, por ejemplo, lo que mide la prueba, estadísticas de la prueba tales como la confiabilidad, la dificultad, el grado en que la prueba puede estar afectada por la velocidad e información de validez, cuando ésta sea obtenida. Deberían suministrarse los rangos percentiles de los puntajes para un grupo de referencia significativo y bien descrito. • En la presentación en PowerPoint, *Proceso de Admisión a Las Universidades del H. Consejo de Rectores 2004*, se afirma que los examinados se desempeñan mejor en preguntas que involucran procesos mentales menos complejos. Ha sido establecido en varias publicaciones que el único propósito de la prueba es para admisión a la universidad. Las afirmaciones o interpretaciones diagnósticas acerca de destrezas cognitivas deberían evitarse, a menos que la prueba haya sido validada para este propósito.
3. ¿Se suministra una justificación para las cantidades y tipos de evidencia recolectadas para apoyar las inferencias que se van a hacer? Si se trata de una nueva evaluación, ¿existe un plan que indique los tipos de evidencia que se recolectarán?		• La PSU es un nuevo programa de pruebas y siempre toma tiempo acumular evidencia de validez predictiva. Sin embargo, la evidencia para otros tipos de validez podría ser recolectada ahora. Algunas de las publicaciones del programa implican usos adicionales de los puntajes de la prueba tales como la evaluación de las destrezas cognitivas del estudiante o la efectividad de un ambiente académico. Se insta al DEMRE, primero, a clarificar los varios propósitos de las PSU y segundo, a desarrollar un plan abarcativo para recolectar y analizar datos y proveer evidencia empírica de que la evaluación es válida para cada propósito. Es importante que el DEMRE involucre a los usuarios de los puntajes en el proceso de validación. Esto es particularmente cierto en el caso de las universidades. Se debería alentar a las universidades para que realicen estudios de validez predictiva de los puntajes de las PSU de manera rutinaria. El plan de validación de las PSU desarrollado por el DEMRE debería incluir cómo serán comunicados los resultados de los estudios a las universidades, a los estudiantes y al público en general. • En una comunicación del DEMRE, se mencionó que se estaban considerando estudios de análisis de factores. Se insta al DEMRE a que lleve adelante estos estudios lo antes posible, de manera que éstos puedan informar de manera

		oportuna los esfuerzos de construcción futura de pruebas.
4. La evidencia y la documentación existente (tanto lógica como empírica), ¿indica que la prueba logra el propósito que busca y apoya las interpretaciones de los resultados para la población a quien va dirigida?		• Un número de documentos muy bien preparados se han suministrado describiendo análisis de ítemes y análisis de los resultados de la prueba, al igual que las especificaciones de la prueba. • La prueba de Lenguaje y Comunicación parece haber sido construida con un nivel apropiado de dificultad de acuerdo con la población a quien va dirigida. • Parece que los puntajes de la prueba de Lenguaje y Comunicación son reportados usando los mismos números que la PAA-V, aún cuando estos puntajes no fueron equiparados. Además, parece que los puntajes de corte usados con la PAA permanecieron iguales para las PSU, aún cuando los puntajes no fueron equiparados. Esta práctica resultó en una confusión considerable por parte de los usuarios de los puntajes que se observó en la primera administración de las PSU. Es importante que se equiparen los puntajes en formas diferentes de las PSU que sean rendidas en la misma administración y también que se equiparen los puntajes de formas que se rinden a través de los años en diferentes administraciones, si estos puntajes van a ser comparados entre formas y entre administraciones.
5. ¿Se advierte a los usuarios potenciales para que eviten posibles usos probables de la evaluación para los cuales no hay suficiente evidencia de validez?		• Los puntajes de las PSU se reportan a un número de instituciones, pero no son enviados directamente a los estudiantes individuales. Los puntajes de los estudiantes son publicados en periódicos protegiendo su confidencialidad. Existe el riesgo de que los puntajes de las PSU sean usados para propósitos diferentes a los de admisión, tales como evaluación escolar. Se debe prevenir a los usuarios contra este tipo de uso, hasta que se recolecte evidencia de la validez de los puntajes de las PSU para propósitos de evaluación y diagnóstico. Es importante comunicar directamente a las universidades, estudiantes y otros usuarios de los puntajes acerca de los usos válidos e inválidos de los puntajes de las PSU.
6. ¿Ha tenido el uso de la evaluación consecuencias no intencionadas para un grupo? Si es así, ¿ha sido revisada la evidencia de validez para determinar si las consecuencias no intencionadas surgen de fuentes irrelevantes de variación?		• Las PSU muestran diferencias en desempeño para subgrupos importantes, como se esperaría en cualquier evaluación de gran escala. • Se sugiere que el DEMRE desarrolle planes para estudiar estas diferencias entre subgrupos para asegurar que no son el resultado de cualquier tipo de inequidad o sesgo en la evaluación. Actualmente se usan los procedimientos de Mantel-Haenszel para estudiar el funcionamiento diferencial de los ítemes entre hombres y mujeres en la evaluación. Se alienta al DEMRE a expandir estos estudios e incluir otros subgrupos de la población y utilizar los resultados de los estudios en el diseño y desarrollo de pruebas futuras.

7. ¿Se han hecho cambios sustantivos en la evaluación desde que la evidencia de validez fue documentada? Si es así, ¿ha sido la evidencia reevaluada para ver si la prueba todavía logra el propósito que busca y apoya las interpretaciones que se intentan hacer para la población a quien va dirigida?		• Las PSU son nuevas pruebas y la evidencia de validez aún no ha sido recolectada. Es importante que el DEMRE desarrolle un plan abarcativo para recolectar evidencia de la validez de las pruebas para todos los usos que se intenta dar a los puntajes.
8. ¿Se suministra orientación a los usuarios para ayudarles a recolectar e interpretar su propia evidencia de validez?		• Los usuarios principales de los puntajes son las universidades y los estudiantes. Como fue mencionado previamente, se recomienda al DEMRE involucrar a los usuarios de los puntajes, especialmente a las universidades, a medida que desarrollen planes para validar la evaluación.

Revisión Técnica de la PSU-Lenguaje y Comunicación

Confiabilidad

Asegurar que los puntajes y otros resultados reportados de la evaluación serán suficientemente confiables para lograr sus propósitos, y que el programa utilizará procedimientos apropiados para determinar y reportar la confiabilidad.

Estándares	Resultados de la Revisión
1. ¿Son todos los puntajes reportados suficientemente confiables para apoyar las interpretaciones que se intentan hacer, incluyendo los subpuntajes o puntajes desagregados?	• Análisis con el método de las mitades y con el coeficiente alfa de Cronbach se han suministrado para las formas 101 y 102 de Lenguaje y Comunicación. • Se suministran coeficientes de confiabilidad para todas las pruebas, basados en la teoría clásica de las pruebas, en la presentación en PowerPoint: *Proceso de Admisión a Las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados*. Las confiabilidades reportadas en este documento parecen ser apropiadas para propósitos de admisión. Sin embargo, sería útil complementar estos análisis con la teoría clásica de las pruebas con análisis IRT (Item Response Theory, Teoría de Respuesta al Item). El uso de modelos IRT brindará información acerca del poder de la medición en puntos críticos de la distribución de puntajes. Se sugiere que el DEMRE use la IRT para establecer una función de información meta (junto con su función de error de medición) para determinar las especificaciones estadísticas de los ítemes y para la prueba como un todo. La Teoría de Respuesta al Item puede usarse para determinar una distribución de la dificultad de los ítemes que concuerde con la distribución de la habilidad de la población a quienes van dirigidos, y que brinde máximo poder de medición en puntos deseados de la distribución de habilidad de la población. • Por otra parte, se recomienda al DEMRE que establezca un proceso de prestest que brinde estadísticas estables de los ítemes que sean comparables entre grupos de pretest (los deltas equiparados trabajarían bien en esta situación). Los procedimientos de pretest deberían ser lo suficientemente robustos para construir un banco de ítemes que sea adecuado para sostener un desarrollo de pruebas de calidad.

2. Los métodos utilizados para estimar la confiabilidad, ¿son apropiados para el uso que se intenta hacer de los puntajes?		• Se suministran los análisis de confiabilidad par las formas 101 y 102 de L y C. Se determinaron estimaciones de confiabilidad por el método de las mitades y por KR 20. Si las pruebas no están afectadas por la velocidad, estos métodos son apropiados para el uso que se intenta hacer de los puntajes. Los errores estándar de medición se reportan para todas las pruebas. • El DEMRE ha evaluado el efecto de la velocidad en las evaluaciones examinando el número de respuestas omitidas para ítems cercanos al final de la prueba. Se recomienda emplear un procedimiento que pueda separar los ítems "no alcanzados" de los ítems omitidos para evaluar el efecto de la velocidad en las pruebas. El DEMRE ha afirmado que ya tiene la información y las herramientas para implementar tal procedimiento.
3. ¿Se suministra información que permita a los usuarios de los puntajes juzgar si los puntajes son suficientemente confiables para apoyar las interpretaciones que se intentan hacer? ¿Se incluyen los errores de medición? Si es apropiado, ¿se suministra información sobre el error estándar y la confiabilidad de las diferencias? ¿Se reporta la consistencia de "promover/reprobar" para los puntajes de corte? ¿Se suministra información acerca del error estándar de medición u otros coeficientes similares alrededor del puntaje de corte?		• Se suministra información a los estudiantes acerca de la confiabilidad y el error estándar de medición de las pruebas. • Los puntajes de corte se publican en documentos que están disponibles a los estudiantes y están basados en los puntajes de corte de años anteriores. Para preservar el mismo significado de año a año de los puntajes de corte (no necesariamente el mismo valor numérico), los puntajes de las pruebas del año actual necesitarán ser equiparados a los puntajes de las pruebas del año anterior. Se recomienda al DEMRE que desarrolle un claro plan de equiparación para hacer los puntajes de las PSU equiparables de año a año. Los datos recolectados en la administración 2003 pueden no ser apropiados para establecer la escala inicial de la prueba. La razón para esto es que el grupo que rindió la prueba en la administración inicial puede no haber sido adecuadamente preparado para rendir la nueva prueba y consecuentemente puede no ser representativo de la población a quien va dirigida la prueba. • En lo concerniente a la comunicación con el público, se recomienda al DEMRE ofrecer un documento escrito dirigido a estudiantes, padres y el público en general, usando lenguaje accesible, explicando la naturaleza de los puntajes y sus errores de medición (incluyendo los puntajes de corte), junto con su correcta interpretación en términos del modelo con referencia a normas sobre el que se basa la construcción y el uso de las pruebas.
4. ¿Se suministra suficiente información acerca de los análisis de confiabilidad para permitir a personas conocedoras evaluar los resultados y replicar los análisis?		• Análisis con el método de las mitades y con el coeficiente Alfa de Cronbach se han suministrado para las formas 101 y 102 de Lenguaje y Comunicación. • Es importante documentar todos los aspectos de los análisis de confiabilidad tales como la justificación para la escogencia del método, la muestra utilizada para llevar a cabo los análisis y cualquier otra información que ayudara a una persona conocedora a evaluar los análisis. Además deberían presentarse los análisis del efecto de la velocidad. • Se recomienda al DEMRE que lleve adelante un análisis de confiabilidad para todos los subgrupos significativos para los

		cuales haya datos disponibles, comenzando con género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público). También se recomienda al DEMRE que considere llevar a cabo análisis en otras posibles subpoblaciones, tales como las que se definen por región y por zona urbana o rural.
5. ¿Se han llevado a cabo análisis de confiabilidad separados cuando se realizaron modificaciones significativas a la prueba, a la administración o a los procedimientos de calificación?		• La PSU puede ser considerada una modificación mayor. Se ha realizado un trabajo para examinar la confiabilidad de la evaluación modificada. Sin embargo, se recomienda la adición de modelos IRT y análisis para subgrupos.
6. ¿Se ha estudiado la confiabilidad y el error estándar de medición de los puntajes reportados en subgrupos de la población?		• El DEMRE ha dicho que los análisis de confiabilidad no son ejecutados para subgrupos porque las universidades no están interesadas en los antecedentes de los estudiantes. Aún así, una manera de asegurar que la prueba es equitativa para todos los subgrupos mayores de población es asegurar que provea puntajes confiables para estos subgrupos. Se recomienda al DEMRE implementar un procedimiento para llevar a cabo análisis de confiabilidad rutinarios para subgrupos de la población, cuando sea posible. Al menos dos variables podrían ser consideradas inicialmente para estos análisis de confiabilidad de subgrupos: género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público).

Revisión Técnica de la PSU-Matemáticas

Validez

Asegurar que los programas recolecten y documenten apropiadamente la evidencia que apoye las inferencias que intentan hacer y las acciones basadas en los resultados reportados de las evaluaciones.

Estándares		Resultados de la Revisión
1. ¿Existe una descripción de cómo la evaluación se ajusta a la misión y los valores del programa?		• Existe una serie de documentos en donde se establece el propósito de las PSU. (Ver, por ejemplo, *Información General, Validez para las dos pruebas: Lenguaje y Comunicación, Matemáticas; Compendio Estadístico Proceso de Admisión Año Académico 2004, y Proceso de Admisión a las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados*) También se suministra información acerca del propósito de la prueba en folletos que se distribuyen a los estudiantes y a las universidades. Esta información claramente establece que los usuarios de la prueba son las universidades que pertenecen al H. Consejo de Rectores y los estudiantes. Varios de los documentos también establecen claramente que el propósito primario de la prueba es seleccionar candidatos para admisión a las 25 universidades del H. Consejo de Rectores. Sin embargo algunos de los documentos también dan a entender que hay usos secundarios para los puntajes de las PSU. • Algunos informes (Ver *Proceso de Admisión a las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados y Análisis de Resultados, Junio de 2004*) incluyen un análisis de cómo se desempeñaron los estudiantes en diferentes áreas de destrezas y contenido de las PSU. En varios documentos suministrados por el programa de pruebas, se brinda información y se hacen inferencias acerca de las destrezas cognitivas de los estudiantes que rinden las pruebas. • Debería tenerse el cuidado de no interpretar los resultados de la prueba para ningún otro propósito (como evaluación de programas o diagnósticos) que no sea el propósito que se ha establecido públicamente, es decir admisión a las universidades. A los usuarios de los puntajes, universidades y estudiantes, se les debería advertir que no interpreten los puntajes de las PSU como útiles para ningún otro propósito que no sea el de un indicador de la habilidad del estudiante para ser exitoso en una universidad, a menos que la prueba sea validada para estos propósitos adicionales. Si se decide que los puntajes de las PSU serán usados para otros propósitos además de las admisiones a la universidad, por

		ejemplo, evaluación de la efectividad de las escuelas o profesores, entonces debe recolectarse evidencia de la validez de la prueba para este propósito. Hasta que tal evidencia sea recolectada, se debería advertir a los usuarios de los puntajes para que no usen los puntajes para ningún otro propósito que no sea el de admisiones a la universidad. • Las PSU se usan también para conceder becas a la universidad. • Los puntajes de las pruebas PSU se usan, junto con la NEM, para formar una medida compuesta ponderada. Las ponderaciones usadas en la medida compuesta difieren de acuerdo con la universidad y la carrera dentro de la universidad. Los puntajes de corte que varían entre universidades y entre carreras dentro de universidades, se forman usando la medida compuesta ponderada y se usan para establecer un banco de examinados para admisión a la universidad. Se recomienda al DEMRE brindar orientación técnica a las Universidades para asistirlas en el establecimiento de los puntajes de corte y en la validación de los puntajes de corte, una vez que se han establecido. Además, es esencial que el DEMRE diseñe e implemente planes para la equiparación de deltas y equiparación de puntajes, de manera que pueda suministrar a las Universidades puntajes de corte consistentes a través de los años.
2. ¿Se describen claramente las siguientes piezas de información: -el constructo, es decir, el conocimiento, destrezas o características a ser medidas? -el propósito de la evaluación? -las interpretaciones que se intentan hacer a partir de los puntajes u otros resultados de la evaluación? -las poblaciones a quienes van dirigidas las pruebas? ¿Está esta información disponible al público en caso de ser solicitada?		• Existe un documento, *Matriz Curricular de Matemáticas para la Elaboración de Ítemes en la PSU de Matemáticas*, que da una descripción del contenido de la prueba. El documento *La PSU de Matemáticas y su Relación con el Marco Curricular del Sector de Matemática* describe una justificación para el contenido de la prueba. • Los documentos mencionados anteriormente brindan información acerca del propósito de la prueba. Otros documentos que se suministran a los estudiantes describen cómo se califica la prueba y brindan un cuadro que muestra cómo los puntajes corregidos se convierten a puntajes estándar. A los estudiantes también se les suministran pruebas de ensayo. • Un problema con las pruebas de ensayo es que los puntajes derivados de estas pruebas no están en la misma escala que los puntajes obtenidos en las PSU operacionales. Consecuentemente, los estudiantes podrían interpretar que los puntajes que ellos reciben en la prueba de ensayo son indicadores de los puntajes que recibirán en la prueba operacional. Ya que los puntajes de la prueba de práctica no están en la misma escala que los puntajes de la prueba operacional, esta interpretación puede ser engañosa. Si los puntajes en la prueba de ensayo son suministrados utilizando los mismos números de la escala de los puntajes de la prueba operacional, es imperativo que los puntajes en la prueba de ensayo sean equiparados a los puntajes de la prueba

		operacional. • Aunque una gran cantidad de información se suministra a los estudiantes y a las universidades acerca de las pruebas, se brinda muy poca información técnica que pueda ser utilizada para interpretar los puntajes. Es importante suministrar información interpretativa a los estudiantes y a las universidades, por ejemplo, lo que mide la prueba, estadísticas de la prueba tales como la confiabilidad, la dificultad, el grado en que la prueba puede estar afectada por la velocidad e información de validez, cuando ésta sea obtenida. Deberían suministrarse los rangos percentiles de los puntajes para un grupo de referencia significativo y bien descrito. • En la presentación en PowerPoint, *Proceso de Admisión a Las Universidades del H. Consejo de Rectores 2004*, se afirma que los examinados se desempeñan mejor en preguntas que involucran procesos mentales menos complejos. Ha sido establecido en varias publicaciones que el único propósito de la prueba es para admisión a la universidad. Las afirmaciones o interpretaciones diagnósticas acerca de destrezas cognitivas deberían evitarse, a menos que la prueba haya sido validada para este propósito.
3. ¿Se suministra una justificación para las cantidades y tipos de evidencia recolectadas para apoyar las inferencias que se van a hacer? Si se trata de una nueva evaluación, ¿existe un plan que indique los tipos de evidencia que se recolectarán?		• La PSU es un nuevo programa de pruebas y siempre toma tiempo acumular evidencia de validez predictiva. Sin embargo, la evidencia para otros tipos de validez podría ser recolectada ahora. Algunas de las publicaciones del programa dan a entender que existen usos adicionales de los puntajes de la prueba tales como la evaluación de las destrezas cognitivas del estudiante o la efectividad de un ambiente académico. Se insta al DEMRE, primero, a clarificar los varios propósitos de las PSU y segundo, a desarrollar un plan abarcativo para recolectar y analizar datos y proveer evidencia empírica de que la prueba es válida para cada propósito. Es importante que el DEMRE involucre a los usuarios de los puntajes en el proceso de validación. Esto es particularmente cierto en el caso de las universidades. Se debería alentar a las universidades para que realicen estudios de validez predictiva de los puntajes de las PSU de manera rutinaria. El plan de validación de las PSU desarrollado por el DEMRE debería incluir cómo serán comunicados los resultados de los estudios a las universidades, a los estudiantes y al público en general. • En una comunicación del DEMRE, se mencionó que se estaban considerando estudios de análisis de factores. Se insta al DEMRE a que lleve adelante estos estudios lo antes posible, de manera que éstos puedan informar, de manera oportuna, futuros esfuerzos de construcción de las pruebas.

4. La evidencia y la documentación existente (tanto lógica como empírica), ¿indica que la prueba logra el propósito que busca y apoya las interpretaciones de los resultados para la población a quien va dirigida?		• Un número de documentos muy bien preparados se han suministrado describiendo análisis de ítemes y análisis de los resultados de la prueba, al igual que las especificaciones de la prueba. • La PSU-Matemáticas parece ser bastante difícil para la población a quien va dirigida. La normalización de los puntajes "estira" y "comprime" la escala de puntajes en una forma que podría provocar que puntajes que representan el mismo nivel de habilidad sean interpretados como que están representando niveles diferentes de habilidad y viceversa. Parece que se llevó a cabo la normalización como respuesta a la distribución asimétrica de los puntajes brutos obtenidos en la prueba. Se insta al DEMRE a que examine las especificaciones de la prueba y modifique estas especificaciones de tal forma que la prueba tenga una mejor concordancia con el nivel de habilidad de la población a quien va dirigida. Si la prueba tiene una mejor concordancia con el nivel de habilidad de la población, los puntajes brutos serán menos asimétricos y la normalización, u otros procedimientos de escalamiento, no resultarán en distorsiones de la escala de puntajes. • Al parecer, los puntajes de la PSU-Matemáticas son reportados usando los mismos números que la PAA-M, aún cuando estos puntajes no fueron equiparados. Además, parece que los puntajes de corte usados con la PAA permanecieron iguales para las PSU, aún cuando los puntajes no fueron equiparados. Esta práctica resultó en una confusión considerable por parte de los usuarios de los puntajes que se observó en la primera administración de las PSU. Es importante que se equiparen los puntajes en formas diferentes de las PSU que sean rendidas en la misma administración y también que se equiparen los puntajes de formas que se rinden a través de los años en diferentes administraciones, si estos puntajes van a ser comparados entre formas y entre administraciones.
5. ¿Se advierte a los usuarios potenciales para que eviten posibles usos probables de la prueba para los cuales no hay suficiente evidencia de validez?		• Los puntajes de las PSU se reportan a un número de instituciones, pero no son enviados directamente a los estudiantes individuales. Los puntajes de los estudiantes son publicados en periódicos protegiendo su confidencialidad. Existe el riesgo de que los puntajes de las PSU sean usados para propósitos diferentes a los de admisión, tales como evaluación escolar. Se debe prevenir a los usuarios contra este tipo de uso, hasta que se recolecte evidencia de la validez de los puntajes de las PSU para propósitos de evaluación y diagnóstico. Es importante comunicar directamente a las universidades, estudiantes y otros usuarios de los puntajes acerca de los usos válidos e inválidos de los puntajes de las PSU.

6. ¿Ha tenido el uso de la evaluación consecuencias no intencionadas para un grupo? Si es así, ¿ha sido revisada la evidencia de validez para determinar si las consecuencias no intencionadas surgen de fuentes irrelevantes de varición?		• Las PSU muestran diferencias en desempeño para subgrupos importantes, como se esperaría en cualquier evaluación de gran escala. • Se sugiere que el DEMRE desarrolle planes para estudiar estas diferencias entre subgrupos para asegurar que no son el resultado de cualquier tipo de inequidad o sesgo en la evaluación. Actualmente se utilizan los procedimientos de Mantel-Haenszel para estudiar el funcionamiento diferencial de los ítemes entre hombres y mujeres en la evaluación. Se recomienda al DEMRE expandir estos estudios e incluir otros subgrupos de la población y utilizar los resultados de los estudios en el diseño y desarrollo de pruebas futuras.
7. ¿Se han hecho cambios sustantivos en la prueba desde que la evidencia de validez fue documentada? Si es así, ¿ha sido la evidencia reevaluada para ver si la prueba todavía logra el propósito que busca y apoya las interpretaciones que se intentan hacer para la población a quien va dirigida?		• Las PSU son nuevas pruebas y la evidencia de validez aún no ha sido recolectada. Es importante que el DEMRE desarrolle un plan abarcativo para recolectar evidencia de la validez de la prueba para todos los usos que se intenta dar a los puntajes.
8. ¿Se suministra orientación a los usuarios para ayudarles a recolectar e interpretar su propia evidencia de validez?		• Los usuarios principales de los puntajes son las universidades y los estudiantes. Como fue mencionado previamente, se recomienda al DEMRE involucrar a los usuarios de los puntajes, especialmente a las universidades, a medida que desarrollen planes para validar la evaluación.

Revisión Técnica de la PSU-Matemática

Confiabilidad

Asegurar que los puntajes y otros resultados reportados de la evaluación serán suficientemente confiables para lograr sus propósitos, y que el programa usará procedimientos apropiados para determinar y reportar la confiabilidad.

Estándares		Resultados de la Revisión
1. ¿Son todos los puntajes reportados lo suficientemente confiables para apoyar las interpretaciones que se intentan hacer, incluyendo los subpuntajes o puntajes desagregados?		• Se suministran coeficientes de confiabilidad para todas las pruebas, basados en la teoría clásica de las pruebas, en la presentación en PowerPoint: *Proceso de Admisión a Las Universidades del H. Consejo de Rectores 2004 Análisis de Resultados*. Las confiabilidades reportadas en este documento son apropiadas para propósitos de admisión. Sin embargo, sería útil complementar estos análisis de la teoría clásica de las pruebas con análisis IRT (Item Response Theory, Teoría de Respuesta al Item). El uso de modelos IRT brindará información acerca del poder de la medición en puntos críticos de la distribución de puntajes. Se sugiere que el DEMRE utilice la IRT para establecer una función de información meta (junto con su función de error de medición) para determinar las especificaciones estadísticas de los ítemes y para la prueba como un todo. La Teoría de Respuesta al Item puede utilizarse para determinar una distribución de la dificultad de los ítemes que concuerde con la distribución de la habilidad de la población a quienes van dirigidos, y que brinde máximo poder de medición en puntos deseados de la distribución de habilidad de la población. La prueba de Matemáticas del 2003 es claramente bastante difícil para la población a quien fue dirigida. Es importante que el DEMRE considere el desarrollo de especificaciones estadísticas para la prueba, usando los procedimientos de IRT descritos anteriormente, para establecer especificaciones para la prueba que tengan mejor concordancia con la distribución de la habilidad de los examinados que rinden la prueba. • Por otra parte, se recomienda al DEMRE que establezca un proceso de prestest que brinde estadísticas estables de los ítemes que sean comparables entre grupos de pretest (los deltas equiparados trabajarían bien en esta situación). Los procedimientos de pretest deberían ser lo suficientemente robustos para construir un banco de ítemes que sea adecuado para sostener un desarrollo de pruebas de calidad, en el caso de la prueba de Matemáticas.

2. Los métodos utilizados para estimar la confiabilidad, ¿son apropiados para el uso que se intenta hacer de los puntajes?		• Se determinaron estimaciones de confiabilidad por el método de las mitades y por KR 20 para ambas formas únicas de la prueba de Matemática. Si las pruebas no están afectadas por la velocidad, estos métodos son apropiados para el uso que se intenta hacer de los puntajes. Los errores estándar de medición se reportan para todas las pruebas. • El DEMRE ha evaluado el efecto de la velocidad en las evaluaciones examinando el número de respuestas omitidas para ítems cercanos al final de la prueba. Se recomienda emplear un procedimiento que pueda separar los ítems "no alcanzados" de los ítems omitidos para evaluar el efecto de la velocidad en las pruebas. El DEMRE ha afirmado que ya tiene la información y las herramientas para implementar tal procedimiento.
3. ¿Se suministra información que permitirá a los usuarios de los puntajes juzgar si los puntajes son suficientemente confiables para apoyar las interpretaciones que se intentan hacer? ¿Se incluyen los errores de medición? Si es apropiado, ¿se suministra información sobre el error estándar y la confiabilidad de las diferencias? ¿Se reporta la consistencia de "promover/reprobar" para los puntajes de corte? ¿Se suministra información acerca del error estándar de medición u otros coeficientes similares alrededor del puntaje de corte?		• Se suministra información a los estudiantes acerca de la confiabilidad y el error estándar de medición de las pruebas. • Las estimaciones globales de la confiabilidad de la evaluación podrían posiblemente ser engañosas en el sentido de que la prueba es claramente muy difícil para la población. Consecuentemente, a la prueba le está faltando potencia de medición en el rango de puntajes que va desde los medios hasta el extremo inferior. Normalizar los puntajes no puede ser un remedio para este problema. Se recomienda al DEMRE que considere modificar el nivel de dificultad de esta prueba para que esté más de acuerdo con el nivel de habilidad de la población. Como se explicó en el Estándar 1, esto puede lograrse con la ayuda de modelos IRT. Se recomienda también la equiparación de deltas para controlar diferencias de habilidad entre el grupo de pretest que se usa para generar las estadísticas de los ítemes para el ensamblaje de la prueba operacional y el grupo que en realidad rinde la prueba operacional. • Los puntajes de corte se publican en documentos que están disponibles a los estudiantes y están basados en los puntajes de corte de años anteriores. Para preservar el mismo significado de año a año de los puntajes de corte (no necesariamente el mismo valor numérico) los puntajes de las pruebas del año actual necesitarán ser equiparados a los puntajes de las pruebas del año anterior. Se recomienda al DEMRE que desarrolle un claro plan de equiparación para hacer los puntajes de las PSU equiparables de año a año. Los datos recolectados en la administración 2003 pueden no ser apropiados para establecer la escala inicial de la prueba. La razón para esto es que el grupo que rindió la prueba en la administración inicial puede no haber sido adecuadamente preparado para rendir la nueva prueba y consecuentemente puede no ser representativo de la población que rinde la prueba. Una indicación de que este es el caso es la

		marcadamente asimétrica distribución de puntajes que fue obtenida por este grupo en la prueba de Matemáticas. • En lo que concierne a la comunicación con el público, se recomienda al DEMRE ofrecer un documento escrito dirigido a estudiantes, padres y el público en general, utilizando un lenguaje accesible, explicando la naturaleza de los puntajes y sus errores de medición (incluyendo los puntajes de corte), junto con su correcta interpretación en términos del modelo con referencia a normas sobre el que se basa la construcción y el uso de las pruebas.
4. ¿Se suministra suficiente información acerca de los análisis de confiabilidad para permitir a personas conocedoras evaluar los resultados y replicar los análisis?		• Análisis con el método de las mitades y con el coeficiente alfa de Cronbach han sido llevados a cabo para la prueba de Matemática. • Es importante documentar todos los aspectos de los análisis de confiabilidad tales como la justificación para la escogencia del método, la muestra utilizada para llevar a cabo los análisis y cualquier otra información que ayudaría a una persona conocedora a evaluar los análisis. Además deberían presentarse los análisis del efecto de la velocidad. • Se recomienda al DEMRE que ejecute análisis de confiabilidad para todos los subgrupos significativos para los cuales haya datos disponibles, comenzando con género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público). También se recomienda al DEMRE que considere llevar a cabo análisis en otras posibles subpoblaciones, tales como las que se definen por región y por zona urbana o rural.
5. ¿Se han llevado a cabo análisis de confiabilidad separados cuando se realizaron modificaciones significativas a la prueba, a la administración o a los procedimientos de calificación?		• La PSU puede ser considerada una modificación mayor. Se ha realizado un trabajo para examinar la confiabilidad de la prueba modificada. Sin embargo, se recomienda la adición de modelos IRT y análisis para subgrupos.
6. ¿Se ha estudiado la confiabilidad y el error estándar de medición de los puntajes reportados en subgrupos de la población?		• El DEMRE ha dicho que los análisis de confiabilidad no son realizados para subgrupos porque las universidades no están interesadas en los antecedentes de los estudiantes. Aún así, una manera de asegurar que la prueba es equitativa para todos los subgrupos mayores de población es asegurar que provea puntajes confiables para estos subgrupos. Se recomienda al DEMRE implementar un procedimiento para llevar a cabo análisis de confiabilidad para subgrupos de la población rutinariamente, cuando sea posible. Al menos dos variables podrían ser consideradas inicialmente para estos análisis de confiabilidad de subgrupos: género (masculino y femenino) y tipo de escuela (pública, privada con financiamiento público, y privada sin financiamiento público).

Revisión Técnica de la PSU-LyC y la PSU-Matemática Uso de las Evaluaciones	
Asegurarse que el programa de pruebas proporcione información que describa y promueva el uso apropiado de las mismas y que advierta a los usuarios de los resultados evitar usos erróneos comunes.	
Estándares	Resultados de la Revisión
1. ¿Se les ha proporcionado a los usuarios de las pruebas la información que necesitan para evaluar la adecuación de las mismas?¿Se les ha proporcionado a los usuarios de las pruebas la oportunidad de consultar con el programa de pruebas sobre usos apropiados de las mismas?¿Se les ha proporcionado a los usuarios de las evaluaciones la siguiente información -el propósito (s) y población de las evaluaciones? -el contenido y formato de las evaluaciones? -la dificultad, la confiabilidad y validez de las evaluaciones? -la disponibilidad y pertinencia de los datos normativos? -administración y requisitos de calificación? -políticas para la	• El DEMRE ha presentado una lista completa y exhaustiva de los usuarios y los usos de las PSU, así como también la información compartida con todos los usuarios de los resultados de las mismas. • Varios documentos les proporcionan información clara a los usuarios de las pruebas sobre la población, el contenido, el formato de las pruebas, requisitos para la admisión a cada programa de académico, incluyendo la ponderación dada a cada prueba y el al NEM para calcular la nota compuesta final de admisión del candidato, y sobre el proceso de admisión en general. Además, el DEMRE ha realizado varios talleres a lo largo del país. En estos talleres, se ha provisto información y consejo para ayudar a los usuarios a evaluar la adecuación, la utilidad y las consecuencias de las decisiones tomadas en base a los resultados. Para el proceso de admisión del 2005, el DEMRE implementó un extenso programa de publicaciones a través de un periódico de alta circulación nacional.El DEMRE ha hecho esfuerzos significativos para realzar la comunicación de la información citada anteriormente. El uso de tecnología para proporcionar servicios y hacer la información más prontamente disponible y accesible para los usuarios de los resultados ha sido una adición muy importante a los esfuerzos de comunicación del programa. Los usuarios de los resultados, los encargados de políticas educacionales y los clientes han sido provistos de acceso al personal del DEMRE para consultas por medio de la Mesa de Ayuda y de los Portales de la página de Internet del DEMRE. • Varios documentos indican que existen usos secundarios de las PSU tales como evaluación profesoral y evaluación de la calidad de la educación. Es fundamental que el DEMRE clarifique todos usos que se esperan de las PSU. Si las PSU tienen propósitos múltiples mas allá de la admisión a la universidad, entonces es indispensable que las pruebas se validen para estos propósitos. Si las pruebas tienen el único propósito de admisión a la universidad, entonces otros usos

retención de los datos y entrega de los mismos? investigación representativa pertinente?	de las pruebas deben ser desalentados y los usuarios de los resultados deben ser notificados que los resultados de las PSU no han sido validados para estos propósitos adicionales. • Aunque se proporciona mucha información a los estudiantes y a las universidades sobre las pruebas, a los usuarios de los resultados se les proporciona poca información técnica que pueda ser utilizada para interpretar los resultados. Es importante proporcionar información técnica a los estudiantes y a las universidades que pueda ser utilizada para interpretar los resultados de las pruebas; por ejemplo, lo que miden las pruebas, estadísticas de las pruebas tales como su confiabilidad, dificultad, rapidez e información de su validez, cuando se obtenga. Se deben proveer rangos percentiles de los resultados para un grupo de referencia significativo claramente descrito. • Algunos informes (Ver *Proceso de Admisión a las Universidades del H. Concejo de Rectores 2004, Análisis de Resultados y Análisis de Resultados, Junio de 2004*) incluyen un análisis del desempeño de los estudiantes en diferentes habilidades y áreas de contenido de las PSU. En varios documentos proporcionados por el DEMRE se proporciona información y se hace referencia acerca de las habilidades cognitivas de los estudiantes que toman dichas pruebas.En la presentación de PowerPoint, *Proceso de Admisión a las Universidades del H. Consejo de Rectores 2004*, se hace la observación de que los candidatos contestan mejor las preguntas que involucran un proceso mental menos complejo. Se ha establecido en varias publicaciones que el único propósito de las pruebas es el de admisión a la universidad. Deben evitarse declaraciones de diagnóstico sobre las habilidades cognitivas de los estudiantes a menos que la prueba se haya validado para este propósito. • El DEMRE formuló una política sobre la pronta entrega de los resultados de las PSU y ha aplicado esta política consistentemente a través del tiempo. Sin embargo, el DEMRE debe comunicar a todos los usuarios de los resultados cuál es su política en lo referente a la retención de los datos. Debe desarrollarse un esquema claro de pautas de políticas sobre la duración de la retención de los resultados de un individuo, sobre su disponibilidad y sobre el uso en el tiempo de estos datos. • Se sugiere que el DEMRE formalice una política sobre la cantidad de tiempo que los resultados de las pruebas siguen siendo una medida válida de las habilidades o destrezas de los candidatos. • Se sugiere que el DEMRE comunique a todos los usuarios de las PSU toda información sobre investigaciones

	representativas relevantes sobre las pruebas realizadas por el personal del DEMRE o por investigadores independientes. Esta información debe ser citada en documentos publicados y debe ser de fácil acceso para los usuarios.
2. ¿Ha recomendado el DEMRE uno uso apropiado de las pruebas tomando acciones tales como el informar a los usuarios: - cómo usar los puntajes de las pruebas u otros resultados junto con alguna otra información pertinente (si tal información es pertinente y útil)? -cómo evaluar las diferencias de los resultados entre individuos (o entre sub-puntajes para el mismo individuo)? -de explicaciones alternativas creíbles para desempeño pobres? -de la necesidad de brindar a los estudiantes de un número razonable de oportunidades para que tengan éxito? - de la necesidad que se familiaricen con las responsabilidades de los usuarios de evaluaciones como se describen en *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999)?	• Los resultados de PSU se usan, junto con el NEM, para formar una nota final de admisión compuesta. Las ponderaciones usadas en el compuesto difieren de acuerdo a la universidad y los programas académicos dentro de cada universidad. Los puntajes de corte, que varían en cada universidad y de acuerdo a programas académicos dentro de las universidades, se forman utilizando el resultado compuesto y se usan para establecer un grupo de candidatos para admisión a la universidad. Se recomienda al DEMRE trabajar con las universidades para clarificar sus metas al establecer los puntajes de corte y diseñar estudios que les permitan utilizar los puntajes de corte para lograr estas metas y validar su uso. • El DEMRE debería desarrollar una guía para la interpretación de los resultados de las pruebas.Esta guía debe recordar a los usuarios del (los) propósito (s) de las PSU, informar a los usuarios del uso apropiado de la evaluación, y proporcionar suficiente información y asesoría para ayudar a todos los usuarios de las PSU a evaluar la adecuación, utilidad y consecuencias de las decisiones hechas en base a los resultados de las pruebas.El DEMRE puede recomendar a los usuarios de las PSU que verifiquen periódicamente que sus interpretaciones de los resultados de las pruebas continúen siendo adecuadas, dado que pueden surgir cambios significativos en la población que toma la prueba, en la administración de la misma, y en sus propósitos. • Las PSU exhiben diferencias en el desempeño para varios subgrupos. Los usuarios de los resultados se podrían beneficiar de tener acceso a información que provea posibles explicaciones para las diferencias entre subgrupos y para desempeños pobres, y que haga mención acerca de los múltiples factores que pueden afectar los resultados de las pruebas. Los puntajes de las pruebas de ensayo no están equiparados con los de la prueba real, no tienen el mismo significado. Por consiguiente, los estudiantes pueden erróneamente inferir que los resultados que obtengan en la prueba de ensayo sean similares a los resultados que obtengan posteriormente en la prueba real. Se recomienda al DEMRE que equipare los puntajes de la prueba de ensayo con los de la prueba real, o bien que reporte puntajes de la prueba de ensayo que no puedan ser confundidos con los de la prueba real. El DEMRE debe también explicar estrategias

¿Ha tratado el DEMRE de desalentar usos erróneos de los puntajes de las pruebas u otros resultados de evaluación advirtiendo a los usuarios de posibles problemas y se les ha dicho a los usuarios qué hacer para evitar estos problemas?	para la toma de pruebas y sus implicaciones a todos los que toman la prueba antes que la prueba se administre. • Se recomienda al DEMRE que informe a los usuarios de las PSU de la necesidad de que se familiaricen con las responsabilidades de los usuarios de evaluaciones como se describen en *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). • Debe tenerse cuidado de no interpretar los resultados de las pruebas para ningún otro propósito, por ejemplo: evaluación de programas o profesores, o diagnósticos, que no sea el propósito públicamente indicado, es decir, admisiones a las universidades. Se debe alertar a los usuarios de las evaluaciones, universidades y estudiantes, de no interpretar los resultados de las PSU para ningún otro propósito que no sea el de indicador de la habilidad del estudiante de tener éxito en una universidad, a menos que la prueba sea validada para esos propósitos adicionales. Si se toma la decisión de usar los resultados de las evaluaciones de las PSU para algún otro propósito además de las admisiones universitarias, entonces debe reunirse evidencia de la validez de las pruebas para ese otro propósito. Hasta que tal evidencia se haya reunido, los usuarios de las evaluaciones deberán ser alertados de no usar los resultados de las pruebas para ningún otro propósito que no sea el de admisión a la universidad. El DEMRE debería activamente desalentar usos de los resultados de las PSU que carezcan de evidencia que los respalden. • Se recomienda al DEMRE que desarrolle un plan de comunicación que ayude a educar a todos los usuarios y al público en general acerca de las PSU como pruebas de gran escala. El DEMRE debería advertir a los usuarios de los resultados de las PSU que como tales tienen la responsabilidad de informarse adecuadamente acerca de los usos apropiados de las pruebas.
3. ¿Se han investigado alegaciones creíbles sobre usos equivocados de las PSU?	• Cualquier consecuencia que devenga del uso de las pruebas, ya sea con o sin intención, debe ser examinada por el DEMRE. • El DEMRE debe alertar a los usuarios de las pruebas acerca de posibles interpretaciones erróneas de los resultados de las mismas y de posibles consecuencias no intencionales de su uso; DEMRE debe tomar medidas para minimizar o evitar interpretaciones erróneas previsibles y consecuencias negativas no intencionales. • Los resultados de las PSU se reportan a varias instituciones pero no se envían directamente a los estudiantes.Los resultados de los estudiantes se publican en periódicos con la confidencialidad de estos resultados protegida. Existe cierto

	peligro que los resultados de las PSU se usen para otros propósitos que no sea el de admisión a las universidades. Debe advertirse a los usuarios contra este tipo de uso, hasta que se haya reunido evidencia de la validez de los resultados de las PSU para otros propósitos de evaluación y/o diagnóstico. Es importante comunicar directamente a las universidades, estudiantes y cualquier otro usuario de los resultados sobre los usos válidos e inválidos de los resultados de las PSU. • El DEMRE debe verificar periódicamente que su interpretación de los datos de las pruebas continúa siendo apropiada, dado cualquier cambio significativo de la población de las pruebas, de su forma de administración y de sus propósitos.
4. ¿Ha proporcionado el DEMRE información y asesoría para ayudar a las partes interesadas a evaluar la adecuación, la utilidad, y las consecuencias de las decisiones hechas en base a los puntajes de las pruebas u otros resultados?	• Como los resultados de las PSU son dados a conocer al público y a aquellos que se encargan de desarrollar políticas, el DEMRE debería proveer y explicar cualquier información adicional que minimice posibles interpretaciones inadecuadas de los datos. Brindar información preliminar con anterioridad a la publicación de los resultados de las pruebas, daría a los medios de comunicación la oportunidad de asimilar datos relevantes y evitar posibles interpretaciones incorrectas de los resultados de las PSU y posibles consecuencias negativas no intencionadas. • El DEMRE debe ayudar a todas las partes interesadas a entender que las decisiones legítimas con respecto al uso de las pruebas e interpretación de los resultados involucran un elemento de juicio profesional. Los procedimientos de recopilación de información para validar el uso apropiado de una prueba a menudo involucran experiencia que no es fácil de cuantificar ni verbalizar.El DEMRE debe advertir que la responsabilidad del uso de la prueba debe ser asumida o delegada sólo a aquellos individuos que tienen el entrenamiento, credencial profesional y experiencia necesaria para hacer uso de esta responsabilidad. • El DEMRE debería promover estudios de investigación relacionados con las PSU. Los resultados de estos estudios deben ser publicados tan pronto como estén disponibles.