# Evaluación de la Prueba de Selección Universitaria (PSU)
## *Resumen de los principales hallazgos de la evaluación*

Minuta elaborada por la Contraparte Técnica CRUCH-MINEDUC de la Evaluación de la PSU

29 de Enero de 2013

La Evaluación de la PSU, adjudicada vía licitación pública a Pearson, contó con una Contraparte Técnica (CT) conformada, de acuerdo a lo establecido en las bases de licitación, paritariamente por representantes del Consejo de Rectores de las Universidades Chilenas (CRUCH) y del Ministerio de Educación (MINEDUC).

Participaron de la CT como representantes del CRUCH: María Elena González Plitt, Doctor en Ciencias de la Educación y académico de la Universidad de Tarapacá; Héctor Manuel Allende Olivares, Doctor en Estadísticas y académico de la Universidad Técnica Federico Santa María; y, Victor Hugo Salinas Torres, Doctor en Ciencias Estadísticas y académico de la Universidad de Santiago de Chile. Por parte del MINEDUC participaron: María Francisca Dussaillant Lehmann, Doctor en Economía y consultora del Programa de las Naciones Unidas para el Desarrollo; Juan Rafael Bravo Miranda, Magíster en Ciencias de la Educación y Jefe de la División de Evaluación de Logros de Aprendizaje de la Agencia de Calidad de la Educación; y, Francisco Lagos Marín, Psicólogo Educacional y Jefe del Centro de Estudios del MINEDUC.

Las principales funciones de la CT fueron revisar los informes de avance y el informe final de la evaluación, realizando observaciones y/o recomendaciones a cada uno de estos; colaborar y coordinar la entrega de documentos y bases de datos requeridos por equipo evaluador de Pearson; y supervisar el normal desarrollo de la evaluación, velando por el cumplimiento de los objetivos del estudio.

La CT fue asesorada, a su vez, por dos expertos internacionales en el ámbito de la medición educacional y psicometría, a saber, Christina Stage y Ronald Hambleton. La Dra. Stage es Ph.D. en Educación y Master en Psicología, y actualmente se desempeña como académica en Umeå University. Además, ella lideró el proyecto “Swedish Scholastic Assessment Test” (SweSAT). El Dr. Hambleton es Ph.D en Métodos Psicométricos v se desempeña actualmente como Director Ejecutivo del Centro de Medición Educacional de la Universidad de Massachusetts. Ambos especialistas han realizado numerosas publicaciones académicas sobre medición educacional y pruebas de admisión para la educación superior.

La siguiente minuta resume los principales hallazgos contenidos en el informe final de evaluación entregado por Pearson el 22 de enero del 2013. Con miras a la mejora de la PSU, la minuta se focaliza en los principales hallazgos y recomendaciones del informe.

El proceso de evaluación de la PSU se llevó a cabo entre enero de 2012 y enero de 2013, y se enfocó en tres áreas: evaluación de los procesos de construcción de las pruebas; análisis de la puntuación de las pruebas, comunicación y uso de los resultados; y, estudio de la validez.

Respecto a la evaluación de los procesos de construcción de las pruebas, el estudio consideró aspectos como la redacción de ítems, calidad del pilotaje de preguntas y administración de los bancos de preguntas. El análisis de la puntuación de las pruebas, comunicación y uso de los resultados, revisó el cálculo de los puntajes estandarizados, propuso un modelo para el tratamiento de los puntos de corte para beneficios sociales, y examinó la adecuación de puntaje único de las pruebas de ciencias, entre otros. La evaluación de la validez de la PSU integró diversas fuentes de evidencia basadas en la estructura interna de las pruebas, la definición de los contenidos de éstas, su capacidad predictiva y las consecuencias asociadas a su uso (validez predictiva y diferencial, considerando género, modalidad y dependencia). El equipo evaluador tuvo el encargo de pronunciarse acerca de la idoneidad de las pruebas como mecanismo de selección para todos los subgrupos de la población que las rinde.

Los estándares utilizados para la evaluación de la PSU están referidos a: *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999); *International Guidelines for Test Use* (International Test Commission, 2012) y *Program Evaluation Standards* (Yarbrough, Shulha, Hopson & Caruthers, 2011).

A continuación se presentan los principales hallazgos y recomendaciones del equipo evaluador.

***Construcción de las pruebas***

- Redacción de ítems. Si bien en el proceso de elaboración de ítems se reconoce la experiencia profesional de los equipos a cargo de este proceso, y que los marcos y especificaciones de pruebas resultan de un análisis curricular que contempla de algún modo las demandas de la educación superior, es importante señalar que no existe participación de especialistas externos al DEMRE en los procesos de revisión y análisis de pertinencia de los marcos y especificaciones y de los ítems de las pruebas. Asimismo, otro aspecto relevante es la falta de homogeneidad de los procedimientos llevados a cabo para asegurar la calidad de los ítems y de las pruebas, por parte de los comités responsables de cada prueba. En cuanto a la documentación que guía y respalda estos procesos, si bien ésta existe parcialmente, puede ampliarse y mejorarse. Lo mismo se señala con relación a los criterios y protocolos de seguridad.
  Las principales recomendaciones se centran en la necesidad de establecer *a priori* los estándares de calidad de los procesos y productos asociados a la construcción de la PSU, y en la participación de expertos externos al DEMRE en las etapas críticas del proceso (definición de marcos y especificaciones, revisión y aprobación de ítems, revisión de las formas operacionales, entre otras). En este contexto, se recomienda una mejora sustantiva de la documentación que guía y respalda los procesos, la realización de estudios específicos, el desarrollo de procesos de capacitación y certificación de los redactores de ítems, la inclusión de especialistas en currículum y profesores de enseñanza media que representen a más de una institución de educación superior, y la realización de auditorías externas e internas en forma periódica.

- Pilotaje de las pruebas. La evaluación realizada por Pearson al pilotaje indica que éste se aleja de los estándares y prácticas internacionales, principalmente en términos de su propósito, la selección de los ítems, la representatividad de la muestra, y las expectativas psicométricas de los resultados del pilotaje. El proceso requiere estar mejor documentado con respecto a la planificación de la administración del piloto y los criterios utilizados para definir los tamaños de la muestra y las variables utilizadas para la estratificación, ya que no es claro que sea representativa de la población de estudiantes que rinde la PSU. El piloto es voluntario, y los resultados obtenidos por los estudiantes que lo rinden no tienen consecuencias para ellos, lo que podría afectar los parámetros estadísticos de los ítems y por ende la construcción de las pruebas operacionales. Asimismo, los ítems piloteados no son presentados en formas múltiples. Todo lo anterior tendría como consecuencia un análisis de los ítems no adecuado. Finalmente, el pilotaje no considera los cambios en las características psicométricas de los ítems cuando éstos son utilizados en diferentes formas de prueba, ni los efectos en los cambios de posición de un ítem en diferentes cuadernillos.
  Las recomendaciones se enfocan a definir claramente el propósito del pilotaje y, sobre esta base, rediseñar este proceso. En forma complementaria, se recomienda mejorar la documentación de los procesos asociados a esta etapa de la construcción de las pruebas.

- Armado de las pruebas. Los criterios establecidos para la selección de los ítems que serán incluidos en las formas operacionales de la PSU están referidos al uso de la Teoría Clásica de los Tests (TCT) y en este contexto son razonables, excepto en lo relativo a las tasas de omisión aceptadas. Sin embargo, la prueba pone mayor énfasis sobre la modalidad Científico-Humanista (HC) del currículum de educación media que sobre la modalidad Técnico-Profesional (TP). Además, el sistema de ensamblaje de las pruebas es manual, lo cual reduciría la eficiencia del proceso y aumentaría el riesgo de error.
  Las recomendaciones consideran, además de documentar los criterios utilizados para armar las pruebas, utilizar un marco de Teoría de Respuesta al Ítem (TRI) que considere el uso de medidas de precisión. Esto a fin de que las pruebas sean acordes a los niveles de habilidad de los postulantes en los tramos de puntaje asociados a la toma de decisiones. Por otro lado, también recomiendan un sistema de ensamblaje de pruebas automatizado, el cual reduciría el riesgo de que se apliquen criterios subjetivos y/o se produzcan errores.

- Banco de ítems. Respecto de la calidad del banco de ítems, en la evaluación se señala que, a pesar de que existe una estructura clara y una base de datos completa, la información es presentada

exclusivamente desde la perspectiva de la arquitectura del software del banco de ítems. En esta línea, no hay claridad respecto de los criterios psicométricos utilizados para la mantención y actualización del banco. Además, la gestión del banco de ítems no se condice con las mejores prácticas a nivel internacional; puesto que, por ejemplo, se permite que los ítems sean modificados después del pilotaje, una vez que ya fueron ingresados al banco.
El equipo evaluador sugiere contar con mejor información estadística sobre los ítems e incorporar, entre otros, un seguimiento a las distintas versiones de los ítems, incluyendo la identificación de los responsables de sus revisiones.

- Selección de ítems del piloto para la forma operacional. En general, el DEMRE usa criterios claros y aceptados internacionalmente para la selección de ítems, considerando indicadores TCT e TRI en forma independiente. Sin embargo, por un lado, existen excepciones a las normas internacionales, tales como el rango de dificultad TRI, el cual abarca un rango más amplio que lo aceptado comúnmente, y la tasa de omisión, que excede el rango aceptado normalmente en pruebas estandarizadas de este tipo. Por otro lado, no hay un procedimiento establecido para resolver aquellos casos de ítems que cumplen algunos criterios y no otros. Además, tal como se indica previamente, los ítems pilotados pueden ser editados previo al uso operacional, lo que contradice las mejores prácticas en construcción de pruebas.
  Con respecto a lo anterior, se recomienda, en primer lugar, revisar y reconciliar los criterios que difieren y se alejan de las normas internacionales. Adicionalmente, documentar en mayor detalle los procesos de selección de ítems considerando conjuntamente los distintos criterios. Finalmente, se recomienda que los ítems pilotados no sean modificados previo al uso operacional, a no ser que vayan a ser pilotados nuevamente.

- Funcionamiento de los ítems entre el pilotaje y la administración operacional. Los resultados de los análisis del piloto y de las formas operacionales evidencian diferencias significativas en el funcionamiento de los ítems, lo cual se aleja de las prácticas utilizadas a nivel internacional. Las altas tasas de omisión aceptadas y el uso de la corrección por adivinación también pueden estar contribuyendo a generar dichas diferencias.
  Se recomienda rediseñar la administración del piloto para posibilitar una mayor consistencia en el funcionamiento de los ítems y reconsiderar el uso de la corrección por adivinación.

- Exploración de variables asociadas al DIF.[1] La evaluación señala que el análisis y tratamiento de la información asociada al comportamiento diferencial de los ítems del piloto es problemático, puesto que no se utiliza para detectar aquellos ítems con potenciales sesgos para determinados grupos de estudiantes (género y tipo de establecimiento), lo cual podría estar afectando la validez de la prueba.
  Se recomienda considerar los resultados DIF del piloto como parte de los criterios de selección de ítems para las formas operacionales de la prueba. Por otro lado, se recomienda utilizar en los análisis DIF otras variables relevantes, tales como modalidad de enseñanza (TP-HC) y nivel socioeconómico.

***Puntuación de las pruebas, comunicación y uso de los resultados***

- Cálculo de los puntajes estandarizados. Si bien se reconocen los esfuerzos realizados para proporcionar puntajes estandarizados de la PSU y sus procedimientos computacionales correspondientes, se observan problemas en este ámbito. La principal preocupación es la falta de un método (equating) que permita que la escala de puntajes de la PSU sea equivalente entre distintas aplicaciones y de un año a otro. Además, no se calcula la precisión de los puntajes, lo cual podría estar afectando las decisiones asociadas al proceso de admisión. Otros problemas relacionados con el cálculo de los puntajes estandarizados son el uso de la corrección por adivinación y la estandarización de las notas de enseñanza media (NEM).
  Se recomienda la introducción de metodologías TRI que permitan que los puntajes sean comparables de un año a otro. Asimismo, se recomienda calcular la precisión de los puntajes,

[1] El indicador DIF identifica ítems en los que examinados del mismo nivel de habilidad, pero que pertenecen a diferentes grupos, muestran diferente probabilidad de responder correctamente tales itemes.

utilizando indicadores TRI, tales como el error condicional de medida (CSEM). Por otra parte, el equipo evaluador también recomienda abandonar la corrección por adivinación. Finalmente, es importante estudiar las propiedades estadísticas de las NEM y del puntaje de postulación, dado su uso en el proceso de admisión.

- Confiabilidad y precisión de los puntajes de las pruebas. El equipo evaluador considera que las medidas de consistencia interna calculadas por el DEMRE (coeficiente alfa de Cronbach) son insuficientes para dar cuenta de la precisión de los puntajes de las pruebas, considerando que la PSU es una prueba de altas consecuencias.
  Se recomienda el uso de TRI puesto que esto permitirá estimar indicadores de precisión, tales como el CSEM y los errores de clasificación.

- PSU y asignación de becas y créditos. Con respecto a proponer un enfoque para definir los puntajes de corte en la escala de la PSU para otorgar beneficios sociales en la forma de becas, El equipo evaluador de Pearson recomienda un enfoque que considere tanto el dominio de la PSU como las consecuencias sociales. El método específico que se recomienda para fijar el puntaje de corte para fines de becas es el método Hofstee.

- Puntaje único para la prueba de Ciencias. La evaluación indica que el uso de un puntaje único para la prueba de Ciencias es inadecuado porque se basa en un supuesto de equivalencia entre los contenidos de aprendizaje de las disciplinas evaluadas. Este supuesto es cuestionable y, de acuerdo con el equipo de evaluación, no responde a estándares internacionales.
  El equipo evaluador recomienda desarrollar pruebas separadas para Biología, Física y Química con propósitos específicos, y reportar sus resultados en forma independiente.

- Uso de los modelos TRI para calibrar, equiparar y puntuar las pruebas. Desde el año 2011 los puntajes de la prueba se consideran válidos por dos años, por lo tanto es fundamental garantizar la comparabilidad de las pruebas entre distintos años. El equipo evaluador señala que existen diversas áreas en que el diseño de comparabilidad de la prueba no cumple con lo esperado para una prueba de altas consecuencias. En este sentido, afirma que no serían válidas las comparaciones de puntajes entre evaluaciones de distintos años.
  Se recomienda la introducción de métodos que permitan equiparar los resultados de las pruebas entre años y formas operacionales aplicadas en un mismo año.

- Entrega y claridad de la información a usuarios del proceso. En general, se considera que los reportes de resultados de la PSU no son interpretados adecuadamente por las distintas audiencias a los que están dirigidos (principalmente, estudiantes y docentes).
  El equipo evaluador recomienda que en cada uno de los reportes se entregue información adicional que permita interpretar los puntajes y otros resultados de manera adecuada. Además, se recomienda descontinuar la entrega de información de resultados por ejes de contenidos en los informes para docentes.

***Validez de las pruebas y de sus resultados***

- Estructura interna. Los análisis realizados permitieron identificar la presencia de un factor principal o dimensión latente para cada prueba de la PSU. Sin embargo, las pruebas presentan alguna evidencia de Funcionamiento Diferencial de Pruebas (DTF[2]), afectando especialmente a los estudiantes de la modalidad TP. Además, el equipo evaluador advierte que la prueba de Matemáticas es la que presenta más DTF.
  El equipo evaluador recomienda que se realicen análisis adicionales para entender mejor el DTF entre distintas poblaciones de estudiantes, especialmente en el caso de las pruebas de Lenguaje y Comunicación y Matemática.

[2] El DTF es un análisis estadístico que permite identificar diferencias de puntajes entre dos grupos de estudiantes, permitiendo a los desarrolladores de pruebas identificar si es que éstas favorecen a un grupo por sobre otro.

- Validez de contenido. Los resultados del estudio de alineamiento indican que la alineación de la mayoría de las pruebas con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del currículum chileno es baja.
  Se recomienda una revisión de la política de usar el Marco Curricular como la base para el desarrollo de los marcos para las pruebas de la PSU.

- Análisis de trayectorias. El análisis de trayectoria presentado por Pearson, no obstante sus limitaciones, muestra que, al desagregar los puntajes por dependencia y modalidad educacional, los puntajes han aumentado para el caso de los establecimientos particulares pagados y la modalidad HC, en desmedro de la modalidad TP. Por otro lado, de acuerdo a estos análisis, la brecha de acuerdo a tipo de establecimiento y nivel socioeconómico es más alta que lo observado internacionalmente.
  Para superar las limitaciones propias de este tipo de análisis el equipo evaluador recomienda implementar métodos que permitan equiparar los resultados de las pruebas entre años y monitorear las tendencias de los resultados en poblaciones relevantes de estudiantes.

- Capacidad predictiva. La evaluación indica que la PSU tiene cierta capacidad para predecir desempeño académico, en particular respecto de los promedios de notas del primer y segundo año. Sin embargo, los valores de predicción encontrados fueron menores que aquellos informados internacionalmente.
  El equipo evaluador recomienda continuar desarrollando estudios que respalden el uso de la PSU en los procesos de admisión, incorporando eventualmente nuevas variables a este proceso.

Finalmente, Pearson entrega como recomendación general documentar de mejor manera -tanto en términos de extensión como de claridad- los procesos asociados con la construcción, aplicación y análisis de las Pruebas de Selección Universitaria. La necesidad de documentación más detallada es vital para una prueba de altas consecuencias como es el caso de la PSU. Por otra parte, proponen utilizar TRI como modelo base para el análisis y toma de decisiones respecto a la prueba, posibilitando simultáneamente las comparaciones interanuales.