Anexo N° 13/2013

Evaluación de la PSU: Principales Observaciones y Recomendaciones de Informe Pearson

Comité Técnico Asesor

Marzo 28, 2013

Esta presentación:

- Principales comentarios y hallazgos de la evaluación:
  - apreciaciones positivas;
  - apreciaciones críticas.
- Recomendaciones:
  - Para perfeccionamientos (la mayoría);
  - Para innovaciones relevantes;
  - Complejas y que requieren mayor análisis.
- Desafíos institucionales

PRINCIPALES COMENTARIOS Y HALLAZGOS DE LA EVALUACIÓN

Apreciaciones positivas sobre proceso de construcción de pruebas

- Calificación de constructores (experticia en áreas disciplinarias y existencia de personal con preparación en aspectos psicométricos)
- Proceso de elaboración, incluyendo filtros expertos
- Criterios para seleccionar ítems luego de su evaluación en el pre-test
- Software para el manejo del banco de ítems

Apreciaciones positivas sobre calidad de los instrumentos

- Ítems cumplen con exigencias psicométricas y en su gran mayoría no presentan evidencia de funcionamiento diferencial entre grupos (sesgo)
- Pruebas muestran adecuada confiabilidad y error de medición
- Análisis factorial muestra que en todas las pruebas domina una dimensión latente (lo que facilita su escalamiento y comparabilidad en el tiempo)
- Poca evidencia de funcionamiento diferencial. Es decir, los puntajes muestran razonable invarianza entre grupos.
- Validez diferencial entre grupos es razonable y equivalente a lo reportado en EEUU (leve presencia de validez diferencial por género, casi nula evidencia de validez diferencial por NSE)

Apreciaciones críticas sobre proceso de construcción de pruebas

- Bajo alineamiento de las preguntas con el marco curricular.
- Escasa diversidad institucional y geográfica de constructores de preguntas y expertos.
- Necesidad de documentar procesos centrales en la construcción de pruebas.
- Falta de un plan integral para el pre-test, que garantice que se obtenga evidencia sobre distintos aspectos de las pruebas y que asegure que las nuevas preguntas se calibren con las ya existentes en el banco de ítems.
- Se cuestiona la existencia de diferencias significativas en los parámetros de los ítems entre el pre-test y la aplicación operacional (lo que puede derivar de las distintas reglas para la puntuación y diferente motivación de los examinados)
- Ensamblaje de las pruebas no garantiza que se logren metas adecuadas de precisión (no se consideran errores condicionales de medición). También se cuestiona que no sea totalmente automatizada, lo que conlleva riesgos.

Apreciaciones críticas sobre los puntajes de las pruebas

- Se cuestiona repetidamente el empleo del puntaje corregido por la probabilidad de responder al azar (que no tiene respaldo en la experiencia internacional y que plantea diversas dificultades técnicas).
- Se cuestiona que no se haya establecido un procedimiento para asegurar la comparabilidad de puntajes entre aplicaciones.
- Se critica que no se reporte el grado de precisión de las pruebas en diversas regiones de puntajes (empleando, por ejemplo, errores condicionales de medición).

Apreciaciones críticas sobre los puntajes de las pruebas

- Se comenta críticamente la ausencia de documentación para sustentar la forma en que se transforman las NEM a puntaje estándar. También se cuestiona que su distribución no sea comparable a la de las pruebas.
- Se cuestiona el cálculo de un puntaje común en la prueba de ciencias, indicando que ello no se basa en una equivalencia de las 3 versiones, por lo que en rigor no se trataría de “equating”, sino que de “linking”. Además, se sugiere modificar el método para calcular un puntaje común.

Apreciaciones críticas sobre validez de las pruebas

- Bajo alineamiento con el marco curricular afecta la validez de contenido de las pruebas.
- Se observa una trayectoria levemente ascendente de los puntajes en el tiempo, con una mayor diferencia a partir de 2007 entre los estudiantes de establecimientos Particulares Pagados y Científico-Humanistas (en comparación con los otros grupos de estudiantes). Sin embargo, hay escasa evidencia de sesgo de medición.
- Se sostiene que la validez predictiva de las PSU es inferior a la observada en el contexto internacional (aunque solo se cita evidencia de EEUU). El juicio crítico de Pearson no considera las diferencias en la enseñanza universitaria de ambos países como posible explicación. Al igual que lo reportado en estudios del CTA, se constata que las PSUM, PSUC, NEM y Ranking presentan niveles equivalentes de validez predictiva (superiores a los de las PSUL y PSUH).

RECOMENDACIONES

Sobre las recomendaciones

- El texto incluye 124 recomendaciones (en realidad 121, pues 3 de ellas son comentarios)
- En su gran mayoría implican mejorías incrementales (que pueden ser adoptadas sin mayor dificultad):
  - Documentar procesos (28 recomendaciones)
  - Mejorar procesos (27 recomendaciones)
  - Incorporar nuevos indicadores o técnicas de análisis (16 recomendaciones)
  - Llevar a cabo estudios (19 recomendaciones)

Recomendaciones reiteradas en el informe y que conllevan ajustes e innovaciones relevantes

- Eliminar el uso de puntuación que penaliza una proporción de respuestas correctas
- Adoptar el modelo IRT en lugar de la Teoría Clásica de la Medición. Este cambio facilita procesos claves contenidos en otras recomendaciones:
  - Permite mantener una escala comparable en el banco de ítems (controlando discrepancias entre pre-test y aplicaciones operativas)
  - Facilita la selección de ítems y el ensamblaje de las pruebas (garantizando metas de precisión) (43).
  - Permite calcular, usar y comunicar errores de medición (estándar y condicionales)
  - Facilita la mantención de escalas de puntajes comparable entre aplicaciones.

Otros cambios relevantes

- Documentar y difundir propósitos y usos de las PSU (1)
- Validar las especificaciones que se produzcan para orientar la construcción de ítems (3; 38)
- Establecer un propósito explícito para el pre-test (23; 70)
- Documentar mejor reglas de selección de ítems, incluyendo procedimientos cuando se observa cumplimiento solo parcial de las reglas (62)
- Abordar el tema del escalamiento de las NEM para que su escala sea equivalente a la de las pruebas (91)
- Revisar la metodología para el cálculo de un puntaje común en la prueba de ciencias (95)

Recomendaciones complejas, que requieren amplio análisis

- Estudiar la posibilidad de reemplazar las NEM por una medición estandarizada del desempeño en la EM (SIMCE) (82)
- Revisar en forma completa el sistema de puntajes de postulación (ponderados por carrera), adoptando la perspectiva de puntajes compuestos (83)
- Reemplazar la prueba de Ciencias por pruebas separadas para Biología, Física y Química con puntajes independientes (94)
- Fundamentar las PSU en un marco que identifique aspectos cognitivos y no cognitivos relevantes para el éxito universitario, en lugar de basarla en el marco curricular nacional (117)

En suma

- Se trata de un informe amplio y ambicioso tanto en los aspectos que aborda, como en las recomendaciones que plantea.
- No está exento de inconsistencias o de juicios parciales, pero incluye mucha ideas valiosas para mejorar el sistema de admisión chileno.
- Es importante advertir que, probablemente por provenir de una institución extranjera, el informe ignora aspectos relevantes de la educación chilena (escolar y universitaria), lo que se constata especialmente en algunas de sus recomendaciones más complejas, que se emiten sin contemplar las condicionantes educativas e institucionales que las explican, o las consecuencias educacionales que tendrían.

En suma

- Mayoritariamente el informe contiene recomendaciones para producir mejoras incrementales al sistema de admisión.
- El informe no debiera entenderse directamente como una agenda de cambios, sino como un conjunto de antecedentes que deben ser analizados y ponderados con criterios técnicos, educacionales e institucionales para definir una agenda de mejoramientos para el corto y mediano plazo.

DESAFÍOS INSTITUCIONALES

Desafíos Institucionales

- Se requiere un análisis pormenorizado de los aspectos evaluados, la evidencia que fundamenta los juiclos de Pearson y las recomendaciones asociadas, para poder establecer una agenda clara, priorizada, pública y realista de los cambios, con su respectiva calendarización.
- Dadas las interrelaciones existentes entre distintas dimensiones del proceso de selección, no es conveniente iniciar ajustes y cambios sin un análisis global de sus implicancias, interdependencias y costos.
- El CRUCH debe establecer una instancia político-técnica para hacer este análisis, tomar las decisiones y acompañar su implementación.

El CTA y los desafíos institucionales

- Cabe mencionar que en sus 8 años de existencia el CTA ha trabajado en varios temas recomendados en el informe.
- Entre esos temas se puede mencionar:
  - La incorporación del modelo IRT
  - El aseguramiento de puntuaciones comparables entre aplicaciones de las pruebas
  - El ajuste de la escala para transformar las NEM a puntaje estándar
- La dificultad que ha existido para resolver estas materias en el pasado refuerza la necesidad de usar la oportunidad que brinda este informe para establecer un nuevo marco institucional que asegure la creación de una agenda estable de mejoramientos al sistema de admisión.

ANEXO
PRINCIPALES OBSERVACIONES EN CADA CAPÍTULO DEL INFORME

Sobre desarrollo de ítems

- El personal a cargo de construir las especificaciones y supervisar la producción de ítems tiene calificaciones adecuadas.
- Falta presencia de expertos externos al DEMRE
- La documentación no está suficientemente estandarizada entre las pruebas
- La producción de ítems se realiza en condiciones adecuadas (constructores, especificaciones, template, tiempo, revisión de expertos, seguridad, etc.)

Sobre pre-test de preguntas

- Aunque el muestreo se basa en criterios aceptados, falta un plan claro para el pre-test (incluyendo el tipo de expectativas psicométricas y tipo de revisión posterior de los ítems)
- Dado que los participantes del pre-test son voluntarios, su motivación puede diferir de la de los examinados en la aplicación operacional (lo que puede explicar las diferencias en el funcionamiento de los ítems en los dos contextos).
- Se sostiene que el principal criterio para seleccionar ítems parece ser cerrar la brecha entre el stock actual y esperado en el banco de ítems, con insuficiente consideración de aspectos psicométricos. Además, no se usa el pre-test para evaluar aspectos como el formato o la localización de los itemes.
- Los criterios estadísticos para la selección de ítems son consistentes con prácticas internacionales (aunque se cuestiona la tolerancia alta para la omisión)

Sobre los criterios para seleccionar ítems para las pruebas definitivas

- Se valora que el equipo a cargo de la selección tiene un manejo psicométrico adecuado, quienes se basan en criterios razonables para la selección (en el marco de la teoría clásica de la medición)
- Sin embargo, se recomienda contar con criterios adicionales, como los errores condicionales de medición para apoyar la selección.
- El entrenamiento de los constructores es serio, pero se recomienda documentar el proceso con un manual de construcción.
- Se afirma que durante las entrevistas se mencionó que las pruebas ponen mayor énfasis en la formación C-H que T-P. Se recomienda verificar esto.
- El proceso de ensamblaje es parcialmente automatizado, pero se sostiene que sería muy deseable que se automatice, para evitar errores de ensamblaje. Esta automatización sería facilitada por la adopción del enfoque IRT.

Sobre el manejo del banco de ítems

- Hay buena documentación del banco de ítems desde la perspectiva de la arquitectura del software, pero falta documentar la dimensión psicométrica del banco.
- Falta información acerca de los criterios empleados para actualizar el banco.
- Se mencionan potenciales riesgos asociados a la pérdida de información, por falta de respaldo a la información en los bancos de datos.

Sobre criterios para analizar funcionamiento de ítems en aplicaciones operacionales

- En general el DEMRE emplea criterios claros para la selección de ítems, basados en CTT e IRT.
- Las decisiones son tomadas por equipos con calificaciones académicas y psicométricas empleando procedimientos adecuados.
- Existe un conjunto de criterios para la selección, pero no existen procedimientos para los casos en que se produce un cumplimiento parcial de los criterios.

Sobre consistencia del funcionamiento de ítems en pre-test y aplicación operacional

- Se constatan diferencias significativas entre ambas aplicaciones, probablemente causadas por el uso de puntuación que penaliza respuestas erradas en aplicación operacional (lo que induce diferentes estrategias en los examinados). También puede deberse a las diferencias en la composición de la muestra en ambas ocasiones.
- La mayor diferencia se produce en la capacidad discriminativa de los ítems (correlación biserial), lo que es esperable.

Sobre funcionamiento diferencial de los ítems

- DIF alude a potenciales diferencias en el desempeño de examinados de diferentes grupos que poseen habilidad equivalente. La detección de DIF puede indicar sesgo en los ítems.
- Se considera problemático que el DEMRE no considere la evidencia de DIF en el pre-test y que se concentre en el DIF durante la aplicación operacional (en el pre-test la proporción de ítems con DIF es mayor).
- Se manifiesta preocupación por la ausencia de una política explícita para seleccionar variables para este análisis. Se sugiere ampliar tales variables para incluir NSE, modalidad de EM y región.
- También se considera preocupante que se emplee más de un método para detectar DIF
- Los análisis realizados por Pearson indicaron que una proporción muy baja de ítems presentaron DIF débil o fuerte

Sobre los procedimientos para calcular puntajes estandarizados

- Se reconoce que el cálculo de los puntajes se basa en procedimientos convencionales para pruebas de este tipo.
- Se cuestiona la falta de evidencia para (1) corrección por adivinanza, (2) selección de promedio y DS de las escalas, (3) la decisión de truncar puntajes en los extremos de las escalas, y (4) la mantención de las escalas en el tiempo.
- También se cuestiona que no se reporte información acerca de la precisión de los puntajes (error estándar y condicional)
- Lo que más preocupa es la ausencia de un mecanismo para asegurar la comparabilidad de puntajes entre aplicaciones.
- También se comenta críticamente la ausencia de documentación para sustentar la forma en que se transforman las NEM a puntaje estándar (que no poseen la misma distribución que las pruebas)
- Se comenta críticamente que no se entrega información acerca del promedio y dispersión del puntaje de postulación

Sobre la precisión y confiabilidad de los puntajes

- Se plantea que las pruebas deben demostrar adecuada precisión especialmente en la región donde se toman decisiones (puntajes superiores a 500). El cálculo de un índice general de confiabilidad (coeficiente alfa) provee una base parcial para lo anterior. Se sugiere que se agregue información acerca de la decisión de aceptar o rechazar postulantes.
- Se cuestiona la falta de información acerca de la precisión de los puntajes de postulación.
- Se propone corregir el escalamiento de las NEM para que se base en los mismos criterios de las pruebas (actualmente ello no ocurre)

Sobre el cálculo de un puntaje común en ciencias

- Se plantea que el puntaje único en ciencias es seriamente cuestionable, pues se basa en un supuesto de equivalencia entre las 3 disciplinas (no se discute, sin embargo, el contexto curricular y educacional chileno que explica la creación de una prueba general de ciencias)
- La producción de un puntaje único no corresponde a los que la teoría de la medición define como “equating”, por lo que debe ser referido como “linking” (esto ha sido reconocido así en el informe técnico del CTA).
- Se plantean cuestionamientos al método empleado para estimar el puntaje único (que conecta el puntaje en los módulos optativos con el módulo común), aunque se reconoce que la alta correlación observada de los tres módulos optativos con la parte común es consistente con el enfoque adoptado. Advierte, sin embargo, que esta correlación pudiera cambiar en el futuro, por lo que recomienda el uso de un método convencional (como el equipercentil encadenado).
- Se sugiere estudiar la validez predictiva de las tres opciones de esta prueba (correlacionándola con notas universitarias), para respaldar el supuesto de equivalencia de los puntajes únicos.

Sobre el uso de métodos IRT (Teoría de Respuesta al Ítem) para el desarrollo y análisis de las pruebas

- Aunque el informe indica que el fundamento actual para el desarrollo y análisis de las pruebas es la Teoría Clásica de la Medición, tuvo acceso a análisis efectuados con el enfoque IRT. Al respecto, emite un juicio crítico, sosteniendo que hay un uso impropio de ciertos conceptos clave (como ítems ancla o ajuste del modelo), o una implementación también alejada de estándares convencionales (como el tamaño del set de ítems comunes)
- Se cuestiona también que no se use este enfoque para asegurar que los ítems que se pre-testean queden en el banco de ítems calibrados con los ya existentes (lo que supone emplear herramientas de equating)
- Se plantea un cuestionamiento a la ausencia de un esfuerzo para asegurar la equivalencia de los puntajes entre años, especialmente en el contexto en que los puntajes son válidos por dos años.
- A partir delo anterior, se emiten recomendaciones para adoptar el enfoque IRT para el desarrollo, calibración y equating de las PSU. De hecho, el uso de IRT está detrás de 2 de las 3 recomendaciones finales de Pearson.

Sobre el uso de software para el análisis estadístico y banco de ítems

- Se reconoce la existencia de un buen programa para manejar el banco de ítems, que incluye niveles adecuados de seguridad, buen manejo de permisos a distintos tipos de usuarios y buena información. Solo se solicita que el software permita incorporar la historia de uso de un ítem (y sus respectivos estadísticos).
- Se comenta que los programas de análisis empleados (SAS, BILOG, DIFAS), son adecuados, aunque para algunos propósitos es necesario considerar alternativas (a BILOG) y en general se solicita que se elija programas puedan ser corridos en modo batch para facilitar la automatización de los análisis.

Sobre la claridad de la información a usuarios

- A partir de entrevistas con muestras (sin carácter de representatividad estadística) de estudiantes, profesores y encargados de admisión, se analizó su entendimiento de diversos dispositivos informativos de las PSU. En general se sostiene que hay una apreciación más positiva de los encargados de admisión, mientras que los otros actores no valoran o comprenden adecuadamente al menos parte de la información contenida en esos dispositivos.

Sobre la estructura interna de las PSU

- Empleando técnicas estadísticas como el análisis factorial, se encuentra evidencia de un claro factor único en todas las pruebas, lo que otorga respaldo para el futuro uso de modelos IRT unidimensionales para mantener la comparabilidad de puntajes en el tiempo.
- También se analiza la invarianza factorial de las PSU, comparando el funcionamiento de las pruebas para distintos subgrupos. La concusión general es que las “PSU muestran poca evidencia de funcionamiento diferencial de los tests (DTF)” (p. 385). Hay casos en que se constata DTF, pero siempre cerca del valor límite mínimo. Sin embargo, es recomendable que este tipo de análisis se efectúe regularmente, para evitar potenciales sesgos.

Sobre validez de contenido de las pruebas

- Se llevó a cabo un estudio del alineamiento curricular de la prueba en que expertos en las disciplinas entrenados en una metodología para el análisis de alineamiento (método de Webb).
- Se concluye que el alineamiento de las pruebas es uniformemente bajo, aunque se reconoce que hay varios aspectos del currículo que no pueden ser evaluados con preguntas de selección múltiple. Pese a lo último, se plantea que existe mucho espacio para mejorar el alineamiento de las pruebas con el marco curricular.
- El informe también menciona comentarios hechos por profesores universitarios en entrevistas, donde aluden a que las PSU no capturan todas las habilidades necesarias para el éxito en la universidad (pero mencionan como ejemplo que no miden la motivación –un aspecto no cognitivo- que sería importante para el éxito académico)

Sobre las trayectorias en el tiempo de subgrupos

- Se analizó la trayectoria de diversos subgrupos entre 2004 y 2011, constatándose que los puntajes son bastante consistentes, con una ligera elevación a partir de 2007 (año en que se otorgó gratuidad para las PSU). La diferencia la explica principalmente la elevación de los puntajes de los particulares pagados y rama C-H (en comparación con puntajes relativamente estables de los otros grupos).
- Adicionalmente se constató que la relación entre NEM y PSU variaba según la dependencia de los establecimientos educacionales. Dicha relación fue mayor en establecimientos particulares pagados y C-H.

Sobre validez predictiva

- Se analiza la validez predictiva y diferencial (según subgrupos) de las PSU con respecto al desempeño de los estudiantes en la universidad. Para estos análisis se emplean bases de datos diferentes de las que ha usado el CTA para sus análisis (aunque las tendencias son equivalentes).
- Se concluye que las PSU poseen cierta validez predictiva. Esta sería mayor para las PSUM, PSUC, NEM (y ranking) que para las PSUL y PSUH.
- Se sostiene que los resultados observados son inferiores a los constatados internacionalmente (aunque solo se citan estudios del norteamericanos el SAT y ACT). No se analizan las diferencias en la formación universitaria norteamericana y chilena que pudieran explicar las diferencias constatadas.
- Con respecto a la validez diferencial, se constatan leves diferencias, semejantes a las observadas en estudios norteamericanos (leve subpredicción del éxito de las mujeres, casi nulas diferencias en la predicción según grupo socioeconómico)