# EVIDENCIA PRELIMINAR PARA LA REDUCCIÓN DE CONTENIDOS EN LA PSU DE MATEMÁTICAS

Departamento de Evaluación, Medición y Registro Educacional -DEMRE
Universidad de Chile

Santiago, 4 de Octubre de 2017

# ÍNDICE

# LISTA DE ABREVIACIONES

**CA** Contenidos Avanzados. Puntaje de los ítems de la PSU-M que evalúan contenidos curriculares de 3º y 4º Medio

**CB** Contenidos Básicos. Puntaje de los ítems de la la PSU-M que evalúan contenidos curriculares de 1º y 2º Medio.

**CMO** Contenidos Mínimos Obligatorios

**CRUCh** Consejo de Rectores de Universidades Chilenas

**CTA** Comité Técnico Asesor del CRUCh para la PSU

**DIF** Funcionamiento Diferencial de Ítems

**EMHC** Enseñanza Media Humanista Científica

**EMTP** Enseñanza Media Técnica Profesional

**ETS** Educational Testing Service

**IRT** Teoría de Respuesta al Ítem

**NEM** Notas de Enseñanza Media

**OF** Objetivos Fundamentales

**PAA** Prueba de Aptitud Académica

**PCE** Pruebas de Conocimientos Específicos

**PSU-M** PSU de Matemáticas

**SUA** Sistema Único de Admisión

**TCC** Curva Característica de la Prueba

**TCM** Teoría Clásica de Medición

**TIF** Función de Información de la Prueba

# RESUMEN EJECUTIVO

El informe de Pearson Education de la PSU (2013) es la evaluación más exhaustiva a la fecha del funcionamiento de la PSU. El equipo de expertos que llevó a cabo la evaluación, detectó numerosos déficits en materia de capacidad predictiva y equidad. Estos son los dos pilares básicos sobre los cuales se sustentan la legitimidad del uso de una prueba. De ahí la urgencia de abordarlos con la rigurosidad que merece una prueba de tan altas consecuencias como es la PSU.

Dentro de las 124 recomendaciones de cambio emitidas por los expertos de Pearson, hay algunas de central importancia para mejorar su capacidad predictiva y su equidad, en especial para el grupo de la educación técnica profesional. El informe hizo especial hincapié en este grupo, ya que la PSU se enfoca en evaluar contenidos curriculares de los dos últimos años de la enseñanza media humanista científica. Los expertos recomendaron desarrollar estudios para evaluar la posibilidad de reducir los contenidos a evaluar en las PSU, privilegiando contenidos curriculares de primero y segundo medio que fueran relevantes para predecir el éxito universitario, dado que tanto el currículo humanista científica como el técnico-profesional no difieren en estos niveles.

El presente estudio toma como punto de partida el diagnóstico de la PSU realizado por el Educational Testing Service (ETS, 2005) y el de Pearson Education (2013). Con respecto a la PSU-M ambos informes coincidieron en señalar el inadecuado grado de dificultad de esta prueba para el grupo que la rinde e instaron a corregir este problema. El informe de Pearson muestra, además, que hay amplio espacio para mejorar la capacidad predictiva, en particular en algunas áreas de carreras. Asimismo cuestionó el uso de la prueba para el grupo de estudiantes provenientes de la educación técnico-profesional.

En este contexto, el presente trabajo tiene como objetivo recabar evidencia preliminar para estudiar el impacto que podría tener en la calidad de la PSU-M la sugerencia de los expertos de Pearson de focalizar ésta en los contenidos de primero y segundo medio.

Tres fueron las preguntas que se buscó responder a través de este estudio:

1. ¿Es posible reducir contenidos en la PSU-M a fin de corregir el inadecuado grado de dificultad reportado en el informe del ETS (2005) y Pearson (2013), sin comprometer significativamente la confiabilidad de la medición?

2. ¿Es posible circunscribir los contenidos curriculares evaluados en la PSU-M a aquellos prescritos para primero y segundo medio sin perjudicar significativamente la actual capacidad predictiva del instrumento?

3. La focalización de la PSU-M en contenidos curriculares de primero y segundo medio, ¿contribuiría a la reducción de la brecha de rendimiento que se observa entre los postulantes provenientes de establecimientos particulares pagados y municipales, en especial los que asisten a la EMTP?

Los análisis tomaron como base a los alumnos egresados en el año 2012 y que rindieron la PSU-M para la admisión 2013, esto pues se contaba con datos que permitían realizar el estudio de validez predictiva por tipo de ítem, análisis que no se había realizado anteriormente. Por ende, y dado que entre el 2013 y el presente el DEMRE ha incorporado cambios al proceso de construcción de las pruebas y en las

tablas de especificaciones sobre las cuales se construyen éstas, se recomienda replicar este estudio con datos de la admisión 2017.

La evidencia preliminar indica que tanto los contenidos de primero y segundo medio (contenidos básicos) como los de tercero y cuarto medio (contenidos avanzados) presentan un grado de dificultad inadecuado en relación a la población que rinde la PSU-M, a excepción de los ítems de contenidos básicos del eje temático de números. Por tanto, para corregir en forma efectiva el grado de dificultad habría que considerar incluir contenidos de años previos, que fueran predictivos del rendimiento universitario y que no están siendo considerados para la construcción de la prueba en la actualidad.

En cuanto a la capacidad predictiva, para la mayoría de las carreras (aproximadamente 55%) el aporte adicional de los ítems de contenidos avanzados por sobre la NEM y los ítems de contenidos básicos es marginal. Para estas carreras, con una prueba de ítems de contenidos básicos se podría mantener la misma capacidad de predicción que se observa en la actualidad al emplear la PSU-M con todos sus ítems. Vale decir, en materia de capacidad predictiva, los ítems avanzados no aportan información relevante para la mayoría de las carreras. Por tanto, focalizar la PSU-M en contenidos de primero y segundo medio—tal cual fue sugerido en el informe de Pearson—no debería perjudicar la actualmente escasa capacidad predictiva de esta prueba en la mayoría de las carreras ofrecidas por universidades adscritas al Sistema Único de Admisión (SUA). Sin embargo, más que mantener el actual nivel de capacidad predictiva de la PSU-M, habría que estudiar cuáles son las materias relevantes para mejorar la predicción del éxito universitario.

Para aquellas carreras en que los ítems avanzados hacen un aporte significativo a la predicción, se podría considerar mantener una prueba como la que se aplica en la actualidad, privilegiando la evaluación de contenidos avanzados en matemática.

En cuanto al tamaño de las brechas de rendimiento entre los distintos grupos que la rinden, las mayores brechas se presentan entre los alumnos de establecimientos particulares pagados y los de la enseñanza media técnica profesional (EMTP), sea ésta municipal o particular subvencionada. Dentro de los hallazgos, se aprecia que en los ítems de contenidos básicos de los ejes temáticos de números y datos y azar se verifican las menores brechas de rendimiento.

Se realizaron análisis para el subgrupo de alumnos destacados en la enseñanza media (alumnos con al menos 6.0 en notas de enseñanza media). Este subgrupo está compuesto por quienes aprovechan mejor las oportunidades educacionales de su medio escolar y, por ende, tienen mayores posibilidades de ser admitidos a la carrera y universidad de su elección, si es que tienen un buen desempeño en las pruebas de admisión. Idealmente, se esperaría que las brechas por tipos de colegio para alumnos destacados fueran menores que cuando se considera el total de los alumnos. Sin embargo, ello ocurre solo para el caso de los estudiantes EMCH de colegios municipales y particulares subvencionados, y no así para los EMTP.

Finalmente, al diseñar las futuras pruebas hay que prestar atención a la capacidad predictiva de los ítems, privilegiándose aquellos tipos de ítems que contribuyan a la predicción y presenten las menores brechas.

# Prólogo

El presente informe, preparado por Nancy Lacourly, Mónica Silva y Karina Díaz, se realizó por encargo de la dirección del DEMRE[1], con el objetivo de orientar la toma de decisiones en torno a los cambios pendientes en las pruebas de admisión, sugeridos en los dos informes internacionales de la PSU.

El documento se organizará de la siguiente manera: en la primera sección se partirá por describir los antecedentes y recomendaciones que entregan los informes del ETS (2005) y de Pearson (2013) con respecto a la PSU-M. En la segunda sección se describe la metodología de análisis que se empleará y en la tercera se reportan los resultados obtenidos para finalmente entregar conclusiones y algunas sugerencias de cambios para esta prueba.

[1] Se agradece muy sinceramente a los revisores externos: Andrés Sánchez, Director General de Evaluación de Resultados Educativos del Instituto Nacional para la Evaluación de la Educación (INEE) de México y miembro del Comité Técnico Internacional del DEMRE para la PSU, Eugenio González director del Instituto de Investigación (IEA- ETS) y al profesor Jaime San Martín del Centro de Modelamiento Matemático de la U. de Chile por sus sugerencias y comentarios a versiones preliminares de este informe.

# 1. Antecedentes

## 1.1 Pruebas Estandarizadas de Admisión Universitaria

Las pruebas estandarizadas de admisión universitaria surgen en Chile en el año 1967 reemplazando al Bachillerato, sistema de admisión basado en la aprobación de un conjunto de pruebas de desarrollo comunes a todos los postulantes y otras que variaban de acuerdo a la mención del Bachillerato. A ello, se añadía que cada universidad y escuela administraba a sus postulantes otras pruebas adicionales para determinar si eran o no aceptados en la institución. El sistema era gravoso para los estudiantes, en especial para quienes postulaban a más de una carrera o universidad, ya que debían someterse a varias instancias de evaluación. Ello, junto al creciente aumento del interés por proseguir estudios universitarios y las críticas hacia el Bachillerato, particularmente la subjetividad en su corrección, fueron factores facilitadores para transitar hacia el desarrollo de pruebas estandarizadas de admisión, proceso que culminó con la implementación de la Prueba de Aptitud Académica (PAA) y las Pruebas de Conocimientos Específicos (PCE) a fines de la década de los sesenta.

Otros países tienen una historia más larga de uso de pruebas estandarizadas con fines de admisión universitaria, puesto que hay beneficios asociados a su uso para los estudiantes, sus familias, las universidades y la sociedad. Camara (2009), plantea que la estandarización es una ventaja dado que la cobertura del currículo y las prácticas de evaluación varían entre colegios, con lo cual las pruebas estandarizadas ofrecen una fuente confiable de información, más allá de lo que revelan las notas escolares. Las pruebas estandarizadas centralizadas son eficientes para estudiantes y universidades. Para los postulantes, el rendir una prueba única de admisión les ahorra los costos emocionales y financieros de tener que someterse a distintas evaluaciones para postular a diferentes carreras o instituciones. Finalmente, para aquellos estudiantes talentosos que por diversas razones no mostraron un rendimiento acorde a sus capacidades en la educación media, las pruebas de admisión les brindan la posibilidad de poder demostrar que son capaces de rendir bien[2].

Gran parte de las controversias, críticas y rechazo a las pruebas estandarizadas surgen como reacción al mal uso de éstas.

## 1.2 Origen de las PSU

La PSU tiene sus orígenes en la reforma de la educación secundaria de fines de la década del 90 y sustituyó a la antigua batería de pruebas de selección, compuesta por la PAA y las PCE. La PAA era una prueba de razonamiento verbal y matemático basada en conocimientos básicos, mientras que las PCE evaluaban el dominio de conocimientos avanzados requeridos en algunas carreras del sistema.

El cambio se justificó argumentando que las nuevas pruebas, por estar referidas al currículum nacional, contribuirían a mejorar el sistema educacional. Aseguraban sus creadores, que las nuevas pruebas de admisión actuarían como incentivo para que los establecimientos educacionales entregaran una educación de calidad, señalando que ocurriría "una mejora del nivel educacional... en el mediano

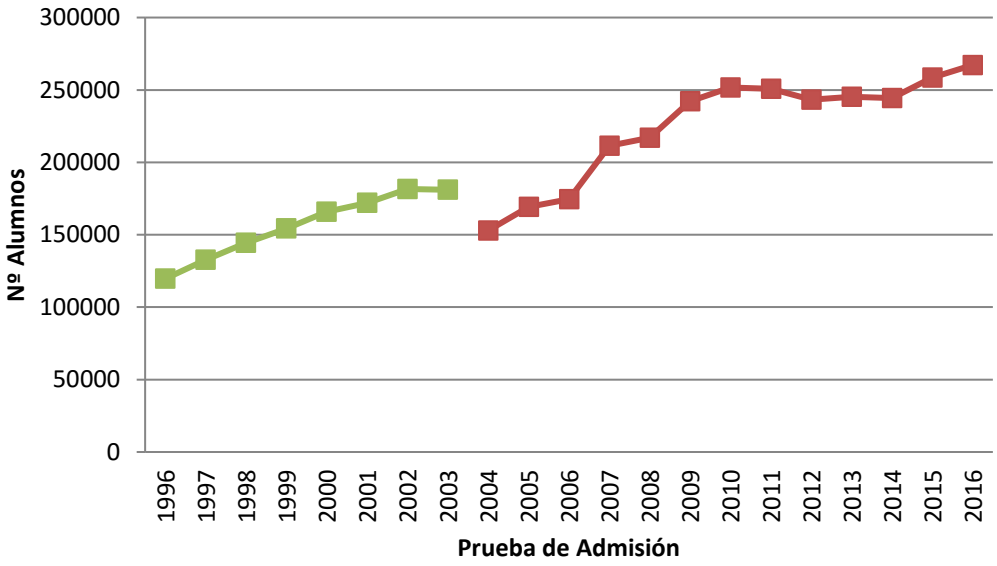
[2] Extraído de W. Camara (2009). "College Admission Testing: Myths and Realities in an Age of Admissions Hype". En R. Phelps (editor). Correcting Fallacies About Educational and Psychological Testing, pg. 147-180. Washington D.C. American Psychological Association.

plazo"[3]. Señalaban además que las nuevas pruebas abrirían espacios de equidad puesto que "mientras más contenidos se incorporen a la evaluación se da la posibilidad de que el factor escuela tenga más importancia y eso es un hecho igualizante [sic]"[4].

Tales expectativas no se cumplieron, tal como quedó reflejado en el informe de Pearson de la PSU(2013).

La primera evidencia de que no se cumplió con la expectativa de abrir mayores espacios de equidad, fue la abrupta caída que se produjo en el número de estudiantes que rendían las pruebas de selección el año en que debutó la PSU. El contingente que rendía las pruebas de selección fue creciente desde la primera hasta la última aplicación de la PAA, aumentando de 30.000 a 180.000 evaluados en su última aplicación. Sin embargo, el número de personas que rindió pruebas de admisión bajó a 153.963 para la primera aplicación de la PSU, una caída de un 15%. Quienes se restaron de participar en el proceso de admisión del año 2004 fueron mayoritariamente estudiantes de colegios municipales y postulantes rezagados que habían estudiado el currículo anterior a la reforma[5]. No fue sino hasta el año 2007, cuando se crearon las becas PSU, que el número de postulantes superó el alcanzado en las dos últimas rendiciones de la PAA como se aprecia en la Figura 1.

Figura 1. Evolución del número de postulantes que rinden las pruebas de dmisión (PAA y PSU)*




Fuente de Datos: Elaboración propia, DEMRE 2016.
*Nota: La primera aplicación de la PSU fue en el proceso de admisión 2004.

## 1.3 Evaluaciones de la PSU

Desde su implementación, la PSU fue objeto de siete evaluaciones nacionales encargadas por el CRUCh a su Comité Técnico Asesor (CTA)[6] y de dos evaluaciones internacionales para examinar su

[3] Ver (SIES: Los Peligros de lo que No Es, 2002) Artes y Letras, El Mercurio.

[4] (El SIES Comenzará a ser aplicado a contar del 2003)Diario Austral 21/7/2002.

[5]Ver (Koljatic, 2006) Equity issues associated with the college admission tests in Chile. Equal Opportunities International, Vol. 25 Iss: 7, pp.544 - 561

[6]Las siete evaluaciones oficiales del CTA del CRUCh se incluyen en las referecias (CTA2004; CTA 2005; CTA 2006;CTA 2008(a); CTA2008(b); CTA2010(a) y CTA2010(b)). Más antecedentes en el Anexo A.

funcionamiento, llevadas a cabo por expertos del Educational Testing Service (ETS, 2005) y de Pearson Education (2013). Ambos informes internacionales presentan un diagnóstico semejante entre sí y diametralmente distinto de aquel que ofrecen las evaluaciones oficiales del CTA del CRUCh. Tanto el informe del ETS como el de Pearson incluyeron numerosas recomendaciones para mejorar las pruebas que componen el sistema de selección chileno, que en la actualidad presentan falencias en los dos pilares básicos de toda prueba de admisión: su capacidad predictiva y su equidad.

Algunas de las modificaciones sugeridas en la última evaluación internacional de la PSU (Pearson, 2013) han sido implementadas, pero otras están aún pendientes. Específicamente, este documento se focaliza en las sugerencias de cambios formuladas por los expertos internacionales con el fin de mejorar la calidad predictiva y equidad de la PSU-M y que no han sido abordadas.

La primera evaluación de la PSU fue llevada a cabo por expertos norteamericanos del ETS en el año 2005, luego de la primera aplicación de las pruebas. El informe del ETS corroboró la apreciación del personal técnico del DEMRE de la época en cuanto a que la PSU-M tenía un grado de dificultad inadecuado para la población que la rendía. Sin embargo el DEMRE no contó en su momento con la autonomía necesaria para corregir el problema[7].

Casi diez años después de la evaluación del ETS, el nuevo informe de la PSU realizado por expertos de Pearson Education: (a) coincidió con el juicio de los expertos de ETS y del DEMRE en cuanto al inadecuado grado de dificultad de dicha prueba, (b) detectó déficits en su capacidad predictiva y (c) cuestionó su equidad como instrumento de selección para los grupos de la educación TP (constatando asimismo en otras PSU algunos de los problemas enunciados).

En relación al inadecuado grado de dificultad de la PSU-M, el informe de Pearson señaló explícitamente que dicho problema fue exacerbado con el aumento de contenidos a evaluar en dicha prueba y señaló que este problema de larga data debía ser abordado y corregido. Cabe recordar aquí que los contenidos a evaluar en la PSU-M se incrementaron entre su primera aplicación y el año 2007, y posteriormente el CTA del CRUCh tomó la decisión de incrementarlos nuevamente en la admisión 2011, con lo cual se agudizó el problema[8].

La decisión de aumentar contenidos no mejoró la calidad predictiva de la PSU, como se puede apreciar a partir de datos consignados en el informe de Pearson.

En las dos dimensiones centrales para emitir un juicio acerca de la legitimidad del uso de una prueba—calidad predictiva y equidad—los expertos de Pearson ofrecieron recomendaciones claras, las que serán abordadas a continuación.

[7]Una situación semejante se verificó en el caso de la PSU de Ciencias, acerca de la cual las opiniones de los expertos del DEMRE y ETS coincidían en que debían calcularse puntajes separados para Biología, Física y Química. No obstante, el cambio no pudo realizarse por la oposición del CTA del CRUCh, como se puede constatar en Caruman, S. (2004). "Informe Técnico Etapa de Aplicación de Pruebas PSU 2004 para el Proceso de Admisión 2005". DEMRE.

[8]Señala textualmente el informe de (Pearson, 2013):"The difference in difficulty between the Language and Mathematics tests is an empirical issue that **has been present since ETS's review of the PSU in 2005**, where that evaluation team noted that while PSU **Mathematics test was too difficult** for the population of applicants, the PSU Language and Communication test showed adequate difficulty for the population of applicants. **That difference has been recently exacerbated, in part, by the fact that the 2011 Mathematics test included five additional items of high difficulty in order to provide for a higher ceiling on the test to distinguish among applicants at the upper tail of the score distribution.** Given the persistence of this difference over time, the PSU program should address it …". ( p. 143; sin resaltar en el original).

## 1.4 Capacidad predictiva de la PSU-M

Cuando se cambió el sistema de admisión basado en la PAA y las PCE por la actual PSU, no se realizaron estudios para evaluar qué contenidos y habilidades eran relevantes para tener éxito en las distintas carreras unversitarias. Se optó en vez por convertir las pruebas de admisión en instrumentos para evaluar los conocimientos adquiridos en la enseñanza media humanista científica (EMHC). Esto se tradujo en un aumento de los contenidos a evaluar en las pruebas, fueran o no éstos de utilidad para predecir el rendimiento universitario.

No es por tanto de extrañar que en materia de capacidad predictiva los expertos de Pearson concluyeran que las PSU están bajo los límites inferiores de predictibilidad observados internacionalmente en instrumentos de selección universitaria. En este punto, aun cuando las pruebas más débiles del sistema chileno son las de Lenguaje e Historia, Geografía y Ciencias Sociales, en las PSU-M y de Ciencias hay amplio espacio para mejorar, puesto que muestran una capacidad predictiva menor que la de otras pruebas de admisión internacionales[9].

En particular, con respecto a la PSU-M, el aumento de contenidos que se verificó entre los años 2004 y 2007 no se tradujo en un aumento de su capacidad predictiva, como se aprecia en datos reportados en el informe de Pearson (Tabla 1).

Tabla 1. Capacidad Predictiva: PSU Matemática[10]

| Año | Correlación con rendimiento universitario (corregida por restricción de rango) |
|---|---|
| 2004 (menos contenidos) | .38 |
| 2005 | .35 |
| 2006 | .35 |
| 2007 (más contenidos) | .36 |

Fuente de Datos: Pearson Report, Tabla 157, pg.453.
Nota: La leyenda que indica la cantidad de contenidos incluida en las pruebas no se encuentra en la tabla original.

En consecuencia, la evidencia contenida en el informe de Pearson apunta a que la reducción de contenidos en la PSU-M no debería perjudicar su actual capacidad predictiva. No obstante, es deseable realizar estudios más finos, como se sugiere en el informe, para estudiar qué contenidos deben privilegiarse con miras a mejorar la capacidad predictiva de esta prueba.

## 1.5 PSU y "*oportunidad de aprender*"

Los estándares internacionales de medición recientes de la American Education Research Asssociation, Americal Psychological Association y el National Council for Measurement in Education (1999 y 2014), destacan la *oportunidad de aprender* como un elemento fundamental a considerar en la construcción

[9] Al respecto señala textualmente el informe: "When analyzing individual PSU test prediction validity over years, it becomes evident the low predictive power of PSU Language and Communication and History and Social Sciences tests. Systematically prediction validity index showed magnitudes indicative of small relationship with university FYGPA. On the other hand, PSU Mathematics and Science tests showed medium values of predictive validity. In none instance PSU tests achieve prediction validity indexes closer to the lower bound observed internationally" (Pearson, 2013 pg. 452; sin resaltar en el original).

[10] Cabe recordar que en la primera aplicación de la PSU, como concesión a la primera generación que la rendía, se incluyó una menor cantidad de contenidos a evaluar, los cuales fueron aumentando en los años posteriores hasta culminar el año 2007.

y uso de pruebas[11]. En este aspecto, el informe de Pearson cuestiona el que la PSU se enfoque en el currículo HC.

Dado que la PSU fue concebida y diseñada para evaluar el currículum HC[12], dejó en desventaja a quienes egresan de la Enseñanza Media Técnica Profesional (EMTP), a pesar de que ellos tienen una alta participación en el sistema de admisión y la PSU. En el año 2016, aproximadamente un 30% de quienes rindieron la PSU provenían de la EMTP.

La evidencia contenida en el informe de Pearson indica que el crecimiento de la brecha en las pruebas de selección responde a un aumento sostenido de los puntajes promedios obtenidos por estudiantes de la educación media particular pagada (humanista científica) a lo largo del tiempo, mientras que los promedios correspondientes a alumnos de la educación municipal y técnico-profesional se han mantenido estables en el tiempo[13].

Atendiendo a las necesidades de mejorar la capacidad predictiva y equidad de las pruebas, el informe de Pearson señala la necesidad de realizar estudios para reducir contenidos de las actuales PSU, incluida la PSU-M. Explícitamente, los expertos señalan la necesidad de alejarse de la pretensión de medir contenidos curriculares de la EMHC que no sean relevantes para alcanzar el éxito en la universidad. Con miras a este fin, los expertos de Pearson recomendaron realizar estudios que aportaran información para acotar los contenidos a evaluar, priorizando aquellos cubiertos en los dos primeros años de la enseñanza media, donde los contenidos especificados para la EMHC y la EMTP son los mismos[14]. Asimismo señalaron que se debe generar un marco que describa las aptitudes requeridas por los estudiantes para tener éxito en la universidad y desarrollar las pruebas conforme a éste[15].

En la actualidad, transcurridos cuatro años desde la emisión del informe de Pearson, la PSU se sigue diseñando sobre los "objetivos fundamentales" (OF) y los extensos "contenidos mínimos obligatorios" (CMO), con un énfasis en aquellos cubiertos durante el tercer y cuarto año de la EMHC, con la consiguiente ventaja para este grupo, según advierte el informe de Pearson[16]. Por tanto, la decisión de

---

[11] Ver al respecto el capítulo tres de AERA, APA & NCME (2014) ,"Standards for Educational and Psychological Testing", pp.49-72.

[12] Ello quedó registrado en el documento que da origen al cambio de pruebas, donde se señala que las nuevas pruebas tienen "aplicabilidad plena solo a la matricula de la modalidad humanista científica y no a la técnico-profesional". Ver Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior,"Informe Sometido en Consulta Previa a la Ministra de Educación" de fecha 22 de noviembre del 2000, página 41.

[13] "Trend lines disaggregated by school type and curricular branch showed that scores steadily increased over time for private schools and schools with a Scientific-Humanistic curricular branch, while scores stayed flat for public and technical schools" (Pearson, 2013, pág. 417).

[14] El texto señala lo siguiente: "…the recommendation is to advance studies on the effect that the decision to approach the test with a priority on the **CMOs of grades 1 and 2 in high school** education may have, as well as determining the effects of the fact that the test may have greater weight for the Scientific–Humanistic curricular branch than for the Technical-Professional one…" (Pearson, 2013, pág. 12) el (Pearson, 2013, pg. 12; sin resaltar en el original).

[15] El informe señala: **"We recommend a review of the policy of using the Curricular Framework as the basis for the development of the PSU test frameworks.** As a part of this review, we recommend the development of a framework that describes the aptitudes (e.g., abilities) and **relevant** non-cognitive variables (e.g., study skills and motivation) needed by students in order to be successful at the university. Such a framework would focus the PSU on the aptitudes necessary to succeed at the university and complement the measure of high school achievement found in NEM and combined together in the postulation score."(Pearson, 2013, pg. 54; sin resaltar en el original).

[16] Señala al respect el informe de Pearson:"[T]he national curriculum has two curricular branches: the Scientific-Humanistic and the Technical-Professional. Even though both curricular branches have a considerable amount of content in common, they differ especially during the third and fourth years of high school. (These differences pose a dilemma for the PSU, which is based on curriculum. While the national curriculum was designed to address the needs of two different groups of students, the PSU assessment frameworks were designed to address the needs of one of these two groups. Thus, **the specifications matrices for the PSU came to target the third and fourth years of the Scientific-Humanistic curricular branch**, while **neglecting those of the Technical-Professional** curricular branch.)"(Pearson, 2013, pg. 153; sin resaltar en el original).

focalizar la prueba en los dos últimos años del currículo humanista científico constituye una desventaja para los estudiantes que asisten a la EMTP que debe ser corregida.

# 2. Metodología

## 2.1 Objetivo general del estudio

El presente estudio aborda los temas pendientes en las evaluaciones internacionales en relación a la PSU-M, en particular la sugerencia de los expertos de Pearson Education de focalizar la prueba en los contenidos cubiertos en los dos primeros años de la enseñanza media y explora:

(a) si con ello se lograría corregir o mejorar su inadecuado grado de dificultad con respecto a la población que la rinde

(b) si se afectaría la capacidad predictiva de la PSU-M al no incluir ítems de contenidos avanzados

(c) si varía la magnitud de las brechas de rendimiento entre alumnos provenientes de la educación particular pagada y aquellos que provienen de otros tipos de colegios.

Finalmente, sobre la base de los resultados, se formularán recomendaciones acerca de si es factible acotar los contenidos de la PSU-M a aquellos cubiertos hasta segundo medio.

## 2.2 Métodos

El estudio consistió en un análisis de datos secundarios de la PSU-M como instrumento de selección, usando métodos descriptivos y correlacionales. Asimismo, se emplearon técnicas psicométricas propias de la teoría clásica de medición (TCM) y teoría de respuesta al ítem (IRT, por sus siglas en inglés)[17].

La mayoría de los estudios de pruebas de altas consecuencias examinan la relación entre los puntajes totales de las pruebas sobre uno o más indicadores de rendimiento. Aun cuando no es frecuente encontrar estudios de confiabilidad o validez que contemplen grupos de ítems[18], hay quienes plantean que dado que los ítems son la unidad básica de observación de una prueba, es importante recolectar evidencia de validez tanto para los ítems individuales, grupos de ítems y los puntajes totales[19].

Por lo anterior y para fines del presente estudio se calculó, además de los puntajes totales en la PSU-M, subpuntajes basados en grupos de ítems por nivel y eje temático, de manera de poder responder a la sugerencia de los expertos de Pearson de estudiar la factibilidad de transitar hacia una prueba

[17] La principal ventaja de los métodos derivados de la TCM es que son simples y fueron la base del desarrollo de pruebas estandarizadas por décadas, de ahí que son más asequibles y fáciles de comprender para quienen no son expertos. No es el caso de los modelos IRT que corresponden a un desarrollo más reciente y son más poderosos, pero más complejos de entender para quienes no han sido entrenados en ellos. Los resultados de ambos análisis se reportan en este documento.

[18] Una excepción en este sentido es el trabajo de Kobrin (2012). Modeling the predictive validity of SAT Mathematics ítems using ítem characteristics. Educational and Psychological Measurement, 72(1), 99-119.

[19] Ver por ejemplo (Haladyna, 2004). Developing and Validating Multiple Choice Items. Routledge.

construida sobre la base de contenidos seleccionados hasta el segundo medio, donde los currículos humanista científico y técnico-profesional no difieren.

El nivel de un ítem se refiere al año de la enseñanza media en que se debiera enseñar el contenido evaluado, esto es si examinaba un contenido curricular correspondiente al nivel de 1°, 2°, 3° o 4° medio, según el currículo oficial del Ministerio de Educación. Los ejes temáticos definidos son aquellos contemplados en la construcción de la PSU-M: números, geometría, álgebra y datos y azar. La información relativa al nivel y eje temático fue provista por profesionales del DEMRE.

En base a estas dos dimensiones —eje temático y nivel— se construyeron subpuntajes e indicadores empleados en los análisis:

1. Los ítems correspondientes a materia que se cubre en 1° y 2° medio, denominado subpuntaje de *contenidos básicos.*
2. Los ítems de 3° y 4° medio, incluidos los cinco ítems correspondientes a contenidos agregados a partir del año 2011 para aumentar el grado de dificultad de la PSU-M, denominado subpuntaje de *contenidos avanzados*[20].
3. Otros puntajes basados en combinaciones de ítems por nivel y eje temático.

## 2.3 Datos y Muestra

Se emplearon las bases de datos de la PSU-M proporcionadas por el DEMRE y por el SUA para los estudiantes que egresaron de la enseñanza media en el 2012 y que rindieron la PSU-M para el proceso de admisión 2013 (N= 170,391), por ser la cohorte más reciente en la que se contaba con información acerca del rendimiento universitario de los postulantes admitidos a las universidades del SUA.

Las bases de datos del DEMRE contenían información de rendimiento de los postulantes en la enseñanza media (NEM y ranking) e información de rendimiento en la PSU-M desagregada a nivel de ítem para las dos formas empleadas en el proceso de admisión 2013, las cuales contenían ítems idénticos en distinta posición. El total de ítems válidos en ese año fue de 74.

La base de datos del SUA había sido compilada previamente para un estudio de validez predictiva encargado por el CRUCh para evaluar la validez del ranking como predictor del rendimiento. Contenía información de los estudiantes ingresados a las 33 universidades que estaban adscritas al SUA en aquella fecha. Cada universidad reportó diversas medidas de rendimiento de sus estudiantes[21]. No obstante, la versión entregada por el SUA —probablemente una versión previa a la empleada en el estudio del ranking— debió ser extensamente revisada y acondicionada por contener numerosos registros duplicados y/o con información incompleta.

[20] En estricto rigor, estos ítems de contenidos avanzados no evalúan el currículo prescrito en la enseñanza media. Esta situación fue advertida en el estudio de alineamiento del informe de Pearson. Sus autores señalan déficits en el alineamiento entre la PSU y el currículo nacional, siendo esta falencia más evidente en el caso de la EMTP.

[21] (Grau, 2016). Estudio acerca de la validez predictiva del ranking de notas. SUA, Consejo de Rectores de Universidades Chilenas (CRUCH).

Dentro de los indicadores de rendimiento universitario que contenía la base del SUA, se empleó como indicador de rendimiento el promedio ponderado (de cursos aprobados y reprobados) obtenido por los alumnos en su primer año de universidad.

Para los análisis de validez predictiva se fusionaron ambas bases de datos resultando un total de casos válidos de 45.530 para alumnos de la promoción 2012 que ingresaron a las universidades del Consejo de Rectores y a universidades privadas adscritas al sistema único de admisión (SUA) en el año 2013. De ellos, 34.552 ingresaron a carreras con 30 o más alumnos, constituyendo la muestra ésta para el análisis de regresión múltiple.

## 2.4 Estudios

A continuación se describen los estudios que se llevaron a cabo para estimar el impacto de la reducción de contenidos en la PSU-M en (a) su grado de dificultad y confiabilidad, (b) capacidad predictiva y (c) magnitud de las brechas entre grupos, por tipo de colegio.

### 2.4.1 Focalización en contenidos de 1° y 2° medio: grado de dificultad y confiabilidad

***Objetivo***

Este estudio aporta evidencia para responder a la pregunta de si es posible reducir contenidos en la PSU-M a fin de corregir el inadecuado grado de dificultad reportados en el informe del ETS (2005) y Pearson (2013), sin comprometer significativamente la confiabilidad de la medición[22].

***Método***

Para evaluar el impacto de la reducción de contenidos en la PSU-M con miras a la corrección de su inadecuado grado de dificultad, se utilizaron las bases de datos desagregadas a nivel de ítems para todos los estudiantes que egresaron de la enseñanza media en el 2012 y que rindieron la PSU-M (N=170.391). Para este estudio se emplearon los subpuntajes por nivel y eje temático. La cantidad de ítems según nivel y eje temático se detalla en la Tabla 2.

Tabla 2. Número de ítems según eje temático y nivel. Admisión 2013.

| Nivel | Números | Álgebra | Geometría | Datos y Azar | Total |
|---|---|---|---|---|---|
| Contenidos básicos (1° y 2° medio) | 5 | 12 | 8 | 5 | 30 |
| Contenidos avanzados ( 3° y 4° medio) | 6 | 18 | 14 | 6 | 44 |
| **Total** | 11 | 30 | 22 | 11 | 74 |

[22] Es conveniente tener presente que la confiabilidad es un requisito necesario, aunque no suficiente para evaluar la calidad de un instrumento. Si bien la confiabilidad es un eslabón clave en el proceso de validación--ya que si no existe suficiente evidencia de que la prueba es confiable, reportar acerca de su validez pierde sentido--una prueba puede mostrar índices altos de confiabilidad, sin ser necesariamente un instrumento válido para medir aquello que se pretende medir. Ver al respecto Crooks, Kane y Cohen. (1996). "Threats to the valid use of assessments", en Assessment in Education: Principle Policy and Practice, 3: 265-286 para una discusión al respecto. Asimismo, (Nunnally, 1972) en su libro " Psychometric Theory" ofrece ejemplos de cómo altos índices de confiabilidad de una prueba no garantizan la validez de su uso.

Fuente de Datos: Elaboración propia, DEMRE 2016.

Se reportan resultados desagregados para el total de ítems de la prueba y para los subconjuntos de ítems según nivel y eje temático, empleando métodos de la teoría clásica de la medición y modelos IRT[23].

### 2.4.2 Focalización en contenidos de 1° y 2° medio: capacidad predictiva

***Objetivo***

Este estudio busca aportar evidencia para responder si es posible reducir contenidos en la PSU-M sin perjudicar significativamente su actual capacidad predictiva.

***Método***

Para estimar el impacto de la reducción de contenidos en la validez predictiva de los subpuntajes de la PSU-M se empleó la base fusionada con datos del DEMRE y SUA.

Para estimar la capacidad predictiva de la PSU-M, se ajustaron distintos modelos de regresión múltiple para predecir el rendimiento universitario (definido operacionalmente como el promedio ponderado en función de sus créditos de ramos aprobados y reprobados de primer año de universidad[24]), siendo el foco principal el examen del impacto en la capacidad predictiva al no incluir contenidos avanzados en la PSU-M.

Para estos análisis se consideró solo carreras con al menos 30 estudiantes inscritos en el primer año[25]. Se ajustaron tres modelos de regresión múltiple con los predictores de interés: notas de enseñanza media, contenidos básicos y contenidos avanzados. Analizados en conjunto, los resultados de los tres modelos permiten comparar el aporte relativo de los diferentes predictores a la predicción del rendimiento universitario de primer año[26].

Se reportan los resultados de tres modelos de regresión que incluyen los siguientes predictores:

1. Solo NEM
2. NEM + Contenidos Básicos (CB)
3. NEM + Contenidos Básicos (CB) + Contenidos Avanzados (CA)

Los modelos de regresión fueron estimados a nivel de carrera y agregados por áreas de carreras y universidad. Se reportan los resultados agregados dado que otros informes previos (del CTA del CRUCh y SUA) así lo hacen, sin embargo, la agregación en áreas y universidad debe ser analizada con

[23] En relación al análisis IRT, se ajustaron modelos de uno (Rasch), de dos y de tres parámetros respectivamente. Se determinó la mejora relativa en la proporción de varianza explicada al usar uno u otro modelo utilizando la heurística de Ayala, 2009, optándose por un modelo de tres parámetros[23] (Ver Anexo C).

[24] Ver Grau M. (2016), pg. 14.

[25] En este caso , se empleó el criterio de (Harrell, 2001) que sugiere un mínimo de 10 observaciones por variable predictora para llevar a cabo un análisis de regresión múltiple.

[26] Un procedimiento semejante fue empleado (Geiser & Studley, 2001)para estimar el aporte relativo de usar el SATI y/o el SAT II para fines de selección en la Universidad de California. (Artículo accesible en :http://www.cshe.berkeley.edu/sites/default/files/shared/publications/docs/UC%20and%20the%20SAT_Geiser.pdf).

cautela, dado que existe una gran heterogeneidad en los resultados que se obtienen a partir de las carreras que las conforman.

Se codificaron áreas de carreras tomando como referente la categorización empleada por los expertos de Pearson, fusionando algunas de sus categorías[27]. Las áreas definidas fueron las siguientes:

1. Agronomía y Forestal
2. Arquitectura
3. Artes (musicales, plástica, visual y teatral)
4. Ciencias
5. Humanidades
6. Letras
7. Química y Farmacia
8. Construcción
9. Ciencias Sociales (sociología, psicología)
10. Derecho
11. Diseño y Publicidad
12. Pedagogía en Educación Parvularia
13. Pedagogía General Básica
14. Pedagogía en Educación Diferencial y Psicopedagógica
15. Otras Pedagogías
16. Pedagogía Media Científica
17. Pedagogía Media Humanista
18. Idiomas
19. Enfermería y Carreras de la Salud (kinesiterapia, nutrición, biotecnología)
20. Ingeniería Civil
21. Ingeniería Comercial y Economía
22. Otras Ingenierías
23. Medicina
24. Odontología
25. Periodismo
26. Técnico Universitario en Administración (contabilidad y auditoría)
27. Técnico Universitario otros
28. Ingeniería en Ejecución
29. Veterinaria

No se estimó la validez diferencial y los sesgos de predicción de los puntajes de contenidos básicos y avanzados según tipo de escuela (HC y TP) por escasa o nula representación de estudiantes de la rama TP en la mayoría de ellas.[28]

[27] Grau (2016) definió 10 categorías basándose en la agrupación Cine Unesco, pero para fines de este estudio se optó por una mayor desagregación de áreas. Informaciones detalladas al respecto, contactar autora principal.

[28] A modo de ejemplo, en el área de Medicina (6 estudiantes de un total de 912), Odontología (14 estudiantes de un total de 1005) y Periodismo (16 estudiantes de un total de 648), etc..

### 2.4.3 Focalización en contenidos de 1° y 2° medio: brechas de rendimiento

***Objetivo***

El estudio busca responder si la eventual focalización de la PSU-M en ítems de contenidos básicos contribuiría o no a la reducción de la brecha de rendimiento que se observa entre los postulantes provenientes de establecimientos particulares pagados y municipales, en especial los que asisten a la EMTP.

***Método***

Este estudio compara el rendimiento de los postulantes según el tipo de colegio del cual egresan. Se estiman las brechas o distancia existente en el rendimiento entre los grupos de egresados de establecimientos particulares pagados (grupo de referencia) y aquellos provenientes de la educación particular subvencionada y municipal (HC y TP respectivamente), en todas las categorías de ítems definidas por nivel y eje temático.

Se calcularon brechas para el grupo total de egresados que rindió la PSU-M y para el subgrupo constituido por alumnos destacados, definidos como aquellos que tienen un promedio igual o mayor a 6,0 en NEM, y que corresponde aproximadamente al 25% superior de la distribución de rendimiento escolar. Este subgrupo de alumnos destacados está compuesto por quienes aprovechan mejor las oportunidades educacionales que el medio escolar les ofrece y, por ende, tienen mayores posibilidades de ser admitidos a la carrera y universidad de su elección, si es que tienen un buen desempeño en las pruebas de admisión. Idealmente, se esperaría que las brechas por tipos de colegio para este grupo selecto de alumnos fueran menores que para la población total.

Para efectos del cálculo, se estimaron las brechas como la diferencia entre el promedio del grupo de referencia (particular pagado) y el promedio de los grupos provenientes de la educación municipal y particular subvencionada (HC y TP, respectivamente), dividida por la desviación estándar combinada de los grupos comparados. Así, por ejemplo, una brecha de 1.0 implica que el rendimiento del grupo particular pagado en ese tipo de ítems se sitúa a una desviación estándar por sobre el promedio obtenido por los integrantes del grupo con el cual se compara[29].

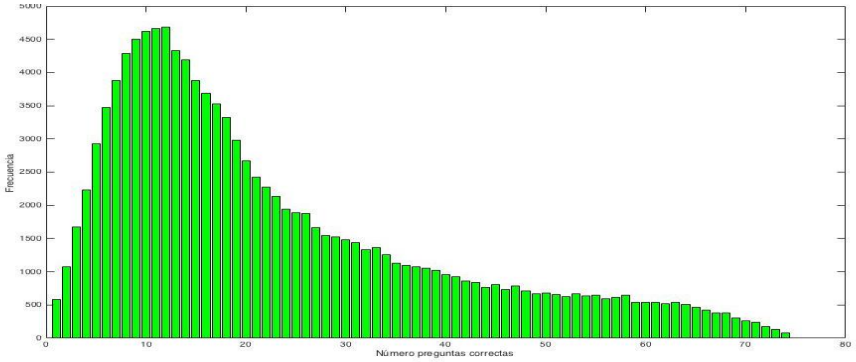
[29] En el anexo F se incluye la fórmula para estimar las desviaciones estándar en el cálculo de las brechas. En el mismo anexo se adjuntan además algunos datos relativos al funcionamiento diferencial de ítems (DIF, por sus siglas en inglés) que muestra la medida en que los ítemes pueden estar midiendo diferentes habilidades para miembros de distintos subgrupos. Se espera que personas de diferentes grupos con una misma habilidad tengan la misma probabilidad de dar una respuesta correcta ante el mismo ítem. Se empleó el método de Mantel-Haenszel y regresión logística. Cabe señalar al respecto, que la presencia de DIF no necesariamente indica la existencia de sesgos en favor de un grupo, sino más bien es una señal de alerta que indica que los ítems en cuestión deben ser revisados.

# 3. Resultados

## 3.1 Grado de dificultad y confiabilidad del instrumento[30]

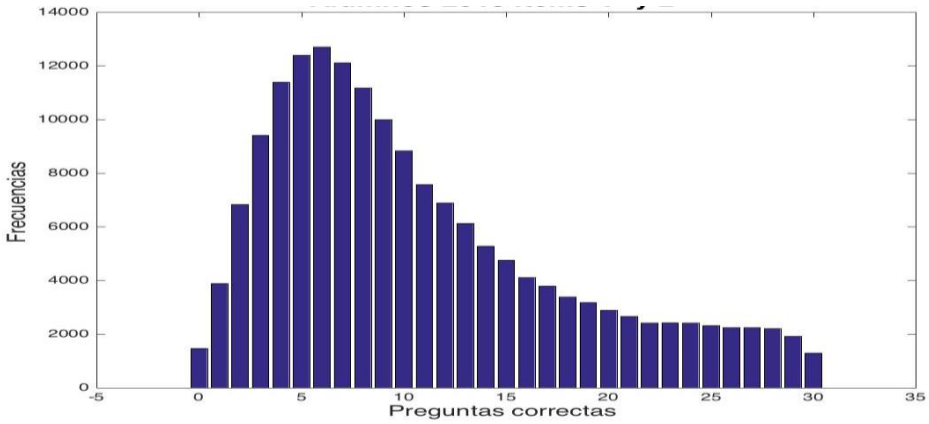
Los análisis estadísticos empleando la metodología clásica (TCM) indican que el inadecuado grado de dificultad de la PSU-M detectado en las evaluaciones del ETS y Pearson persiste al estudiar los datos correspondientes a la prueba rendida para la admisión del año 2013.

Figura 2. Frecuencia de respuestas correctas (74 items): Rinden 2012, promoción del año (N=170.391)




Fuente de Datos: Elaboración propia, DEMRE 2016.

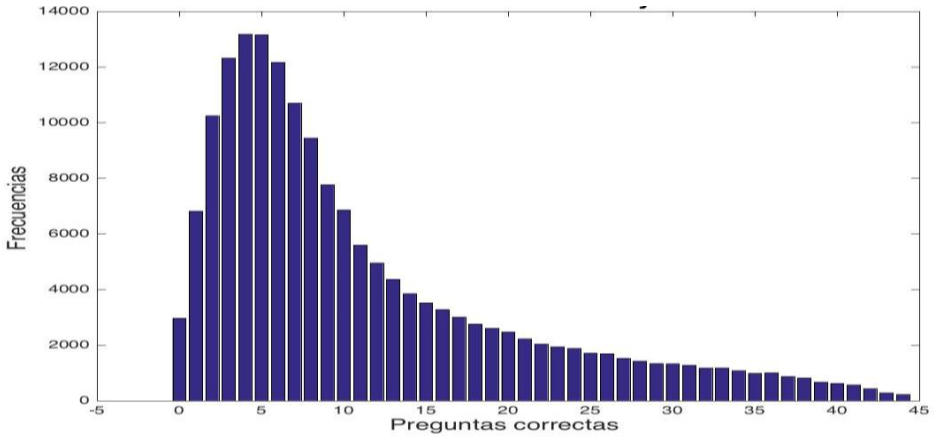
Figura 3. Frecuencia de respuestas correctas ítems básicos (1° y 2° Medio; 30 items): Rinden 2012, promoción del año (N=170.391)




Fuente de Datos: Elaboración propia, DEMRE 2016.

[30] En esta sección se reportan los coeficientes de confiabilidad puesto que se solicitó este análisis como parte del informe, a pesar de que estos coeficientes pueden ser engañosamente altos, por el inadecuado grado de dificultad de la PSU. Al respecto señala textualmente el informe del ETS (2005) : "Las estimaciones globales de la confiabilidad [de la PSU-M] podrían posiblemente ser engañosas en el sentido de que la prueba es claramente muy difícil para la población que la rinde. Consecuentemente, a la prueba le está faltando potencia en el rango de puntajes que va desde los medios hasta el extremo inferior" (pg. 45).

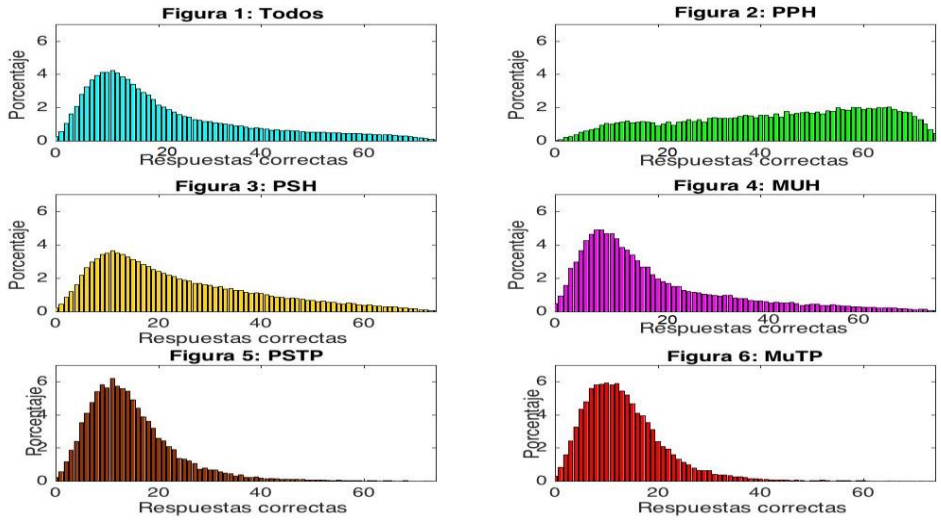
Figura 4. Frecuencia de respuestas correctas ítems avanzados (3° y 4° Medio; 44 items): Rinden 2012, promoción del año (N=170.391)




Fuente de Datos: Elaboración propia, DEMRE 2016.

Como se aprecia en las figuras 2, 3 y 4, no existe una distribución simétrica de los puntajes brutos, ni aun para los ítems básicos. Tampoco ello se observa en las distribuciones por tipo de colegio y rama educacional, como se puede apreciar en el conjunto de gráficos de la Figura 5.

Figura 5. Frecuencia de respuestas correctas. Rinden 2012 (promoción del año), por rama y dependencia.




PPH: particular pagado HC
PSH: particular subvencionado HC
PSTP: particular subvencionado TP
MUH: municipal HC
MUTP: municipal TP

Fuente de Datos: Elaboración propia, DEMRE 2016.

En términos de la TCM el grado de dificultad se define como el porcentaje de respuestas correctas. El promedio para una prueba de 5 alternativas debería estar en torno a un grado de dificultad de 0.60[31]. Ello no se cumple, como se puede apreciar en las tablas 3 y 4 que reportan el grado de dificultad de la PSU-M, tanto para el puntaje total como para los subpuntajes de contenidos básicos y avanzados para la muestra total y por tipo de colegio respectivamente.

Tabla 3. Grado de dificultad de ítems y confiabilidad por nivel y eje temático.

| Items | TCM | |
|---|---|---|
| **Todos los ítems** | Dificultad Promedio | 0.30 |
| **74 ítems** | Desviación estándar de la media | 0.18 |
| | Alpha Cronbach | 0.96 |
| **Contenidos básicos(1° y 2° medio)** | Dificultad Promedio | 0.36 |
| **30 ítems** | Desviación estándar de la media | 0.17 |
| | Alpha Cronbach | 0.91 |
| **Contenidos avanzados** | Dificultad Promedio | 0.25 |
| **( 3° y 4° medio)** | Desviación estándar de la media | 0.16 |
| **44 ítems** | Alpha Cronbach | 0.94 |
| **Contenidos básicos Número** | Dificultad Promedio | 0.55 |
| **5 ítems** | Desviación estándar de la media | 0.05 |
| | Alpha Cronbach | 0.67 |
| **Contenidos básicos Álgebra** | Dificultad Promedio | 0.32 |
| **12 ítems** | Desviación estándar de la media | 0.15 |
| | Alpha Cronbach | 0.85 |
| **Contenidos básicos Geometría** | Dificultad Promedio | 0.26 |
| **8 ítems** | Desviación estándar de la media | 0.08 |
| | Alpha Cronbach | 0.76 |
| **Contenidos básicos Datos y azar** | Dificultad Promedio | 0.42 |
| **5 ítems** | Desviación estándar de la media | 0.21 |
| | Alpha Cronbach | 0.52 |
| **Contenidos avanzados Número** | Dificultad Promedio | 0.31 |
| **6 ítems** | Desviación estándar de la media | 0.22 |
| | Alpha Cronbach | 0.62 |
| **Contenidos avanzados Álgebra** | Dificultad Promedio | 0.27 |
| **18 ítems** | Desviación estándar de la media | 0.15 |
| | Alpha Cronbach | 0.89 |
| **Contenidos avanzados Geometría** | Dificultad Promedio | 0.20 |
| **14 ítems** | Desviación estándar de la media | 0.12 |
| | Alpha Cronbach | 0.86 |
| **Contenidos avanzados Datos y azar** | Dificultad Promedio | 0.28 |
| **6 ítems** | Desviación estándar de la media | 0.24 |
| | Alpha Cronbach | 0.62 |

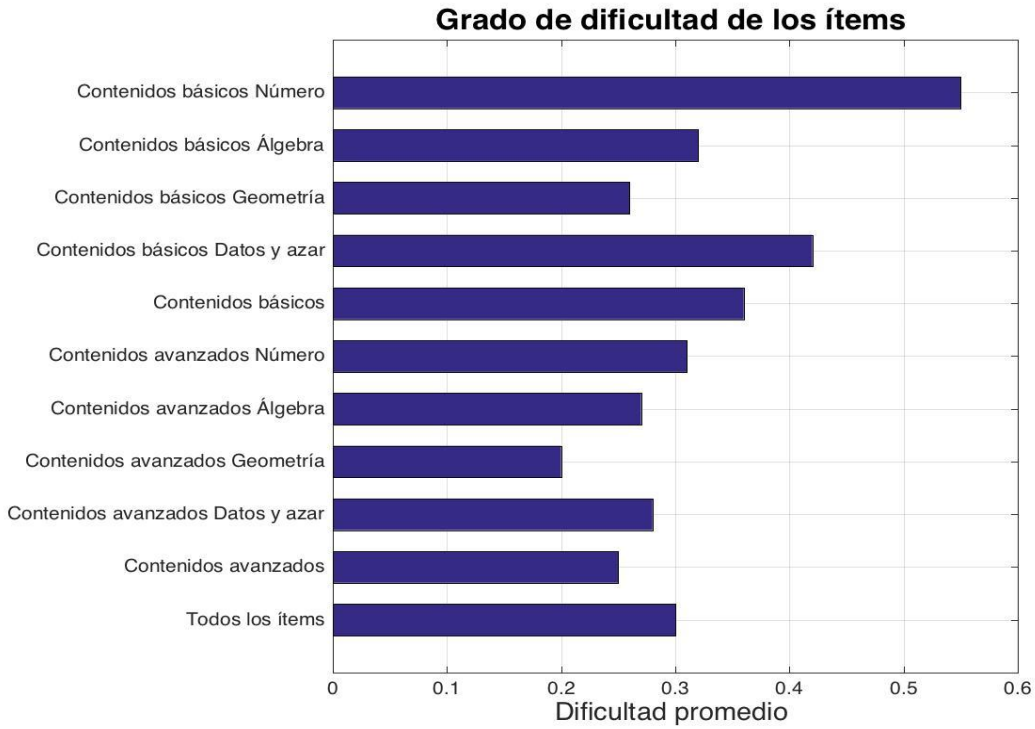
Fuente de Datos: Elaboración propia, DEMRE 2016.

En la Tabla 3, al analizar el grado de dificultad para subconjuntos de ítems por nivel y eje temático se observa que existe variabilidad entre ellos. El grado de dificultad del subconjunto de ítems de "contenidos básicos de números" es el más cercano al recomendado en la literatura para ítems con cinco alternativas de respuesta. Todas las restantes combinaciones de eje temático por nivel presentan

[31]Ver al respecto (Lord, 1952). The relation of the reliability of multiple choice tests to the distribution of item difficulties. Psychometrika,17: 181-192; (Henrysson)(1971). Gathering, analyzing , and using data on test ítems. En R. L. Thorndike (editor): Educational Measurement, segunda edición. American Council on Education, Washington D.C.

un grado de dificultad inadecuado para la población recién egresada de la enseñanza media que la rinde. La Figura 6, a continuación, ilustra los grados de dificultad de los ítems en forma visual para la muestra total.

Figura 6. Grado de dificultad de los ítems por eje y nivel




Fuente de Datos: Elaboración propia, DEMRE 2016.

Para la muestra total (tabla 3) la confiabilidad del conjunto de los ítems de nivel básico es de .91 con sólo treinta ítems. El subconjunto de ítems de nivel avanzado, que consta de 44 ítems, presenta un coeficiente de confiabilidad de .94. La confiabilidad de la escala total (74 items) es de .96. Sin embargo, si se duplicara el número de ítems que evalúa contenidos básicos empleando la fórmula de Spearman Brown la confiabilidad de esta escala sería de .96, exactamente la misma confiabilidad que tiene la escala actual con 74 ítems. Esto indicaría que, *ceteris paribus* y con sólo sesenta ítems de contenidos básicos, se podría tener un instrumento igualmente confiable, sin necesidad de incorporar ítems de tercero y cuarto medio[32]**.**

Con respecto a la confiabilidad, cabe advertir que su magnitud está directamente relacionada con la longitud de la escala, por lo que las estimaciones de confiabilidad que se reportan en las tablas 3 y 4 no son directamente comparables entre sí por el dispar número de ítems que componen las distintas categorías definidas en función del eje temático y el nivel. El número de ítems varía entre 5 ítems de "contenidos básicos de datos y azar" y 12 ítems para "contenidos básicos de álgebra".

[32]Spearman-Brown, citado en (Crocker, 1986) Introduction to classical and Modern Test Theory. HBJ (Orlando, Florida). Ver fórmula en el Anexo B.

Al analizar los datos desagregados por tipo de colegio (tabla 4), se advierten diferencias en el grado de dificultad de la escala dependiendo del grupo, siendo los indicadores del grupo particular pagado los más cercanos a los esperados para una prueba con cinco alternativas de respuesta.

Tabla 4. Grado de dificultad de ítems y confiabilidad por nivel y eje temático por tipo de colegio[33]

| ITEMS/PRUEBA | ESTADÍSTICA | PPHC | PSHC | MuHC | PSTP | MuTP |
|---|---|---|---|---|---|---|
| Todos los ítems | Dificultad media | 0.58 | 0.32 | 0.26 | 0.20 | 0.18 |
| 74 ítems | Desviación estándar de la media | 0.21 | 0.19 | 0.16 | 0.15 | 0.14 |
| | Alpha de Cronbach | 0.97 | 0.96 | 0.96 | 0.88 | 0.86 |
| Contenidos básicos (1° y 2° medio) | Dificultad media | 0.65 | 0.39 | 0.31 | 0.26 | 0.24 |
| 30 ítems | Desviación estándar de la media | 0.17 | 0.19 | 0.16 | 0.17 | 0.16 |
| | Alpha de Cronbach | 0.93 | 0.90 | 0.90 | 0.77 | 0.74 |
| Contenidos avanzados | Dificultad media | 0.53 | 0.27 | 0.22 | 0.16 | 0.14 |
| ( 3° y 4° medio) | Desviación estándar de la media | 0.23 | 0.19 | 0.15 | 0.13 | 0.12 |
| 44 ítems | Alpha de Cronbach | 0.96 | 0.93 | 0.93 | 0.81 | 0.77 |
| Contenidos básicos Número | Dificultad media | 0.83 | 0.60 | 0.50 | 0.44 | 0.41 |
| 5 ítems | Desviación estándar de la media | 0.05 | 0.06 | 0.06 | 0.07 | 0.08 |
| | Alpha de Cronbach | 0.67 | 0.65 | 0.66 | 0.51 | 0.49 |
| Contenidos básicos Álgebra | Dificultad media | 0.64 | 0.35 | 0.27 | 0.20 | 0.18 |
| 12 ítems | Desviación estándar de la media | 0.17 | 0.18 | 0.14 | 0.14 | 0.12 |
| | Alpha de Cronbach | 0.88 | 0.83 | 0.83 | 0.57 | 0.53 |
| Contenidos básicos Geometría | Dificultad media | 0.56 | 0.28 | 0.22 | 0.16 | 0.14 |
| 8 ítems | Desviación estándar de la media | 0.11 | 0.09 | 0.08 | 0.09 | 0.09 |
| | Alpha de Cronbach | 0.80 | 0.74 | 0.75 | 0.54 | 0.49 |
| Contenidos básicos Datos y Azar | Dificultad media | 0.64 | 0.44 | 0.37 | 0.36 | 0.34 |
| 5 ítems | Desviación estándar de la media | 0.23 | 0.23 | 0.20 | 0.21 | 0.19 |
| | Alpha de Cronbach | 0.60 | 0.49 | 0.50 | 0.34 | 0.33 |
| Contenidos avanzados Número | Dificultad media | 0.55 | 0.33 | 0.27 | 0.21 | 0.19 |
| 6 ítems | Desviación estándar de la media | 0.33 | 0.26 | 0.20 | 0.16 | 0.15 |
| | Alpha de Cronbach | 0.67 | 0.60 | 0.61 | 0.39 | 0.36 |
| Contenidos avanzados Álgebra | Dificultad media | 0.55 | 0.28 | 0.22 | 0.14 | 0.13 |
| 18 ítems | Desviación estándar de la media | 0.23 | 0.18 | 0.14 | 0.11 | 0.10 |
| | Alpha de Cronbach | 0.91 | 0.88 | 0.87 | 0.67 | 0.61 |
| Contenidos avanzados Geometría | Dificultad media | 0.49 | 0.22 | 0.17 | 0.12 | 0.11 |
| 14 ítems | Desviación estándar de la media | 0.19 | 0.13 | 0.10 | 0.09 | 0.09 |
| | Alpha de Cronbach | 0.88 | 0.83 | 0.83 | 0.57 | 0.52 |
| Contenidos avanzados Datos y Azar | Dificultad media | 0.56 | 0.32 | 0.26 | 0.23 | 0.21 |
| 6 ítems | Desviación estándar de la media | 0.22 | 0.24 | 0.21 | 0.21 | 0.19 |
| | Alpha de Cronbach | 0.74 | 0.61 | 0.60 | 0.40 | 0.36 |
| Número de alumnos | | 18846 | 67812 | 34105 | 23585 | 25275 |

[33] Las diferencias entre el número de alumnos en ambas tablas responde a que no todos tenían información con respecto al tipo de colegio (rama y/o dependencia) del cual egresaron.

La dificultad media de los distintos puntajes indica que la PSU-M es sustancialmente más difícil para los alumnos que egresan de los otros tipos de colegios, en especial los de la EMTP. También se observan diferencias en la confiabilidad estimada según el tipo de colegio, siendo más alta cuando se estima sobre la base de los alumnos provenientes de la educación particular pagada.

El análisis IRT muestra resultados consistentes a los de la TCM en cuanto al grado de dificultad de la prueba, sea que se empleen modelos de 1, 2 o 3 parámetros. A continuación, en la tabla 5, se presentan los resultados promedio del modelo de tres parámetros. (En el anexo C se entregan antecedentes acerca del ajuste de los distintos modelos).

Tabla 5. Indicadores IRT.

| N= 170.391 | IRT (3PL) | | | |
|---|---|---|---|---|
| **ITEMS** | **Número de ítems** | **Pseudo-Adivinación** | **Dificultad** | **Discriminación** |
| Todos los ítems | 74 | 0.06 | 1.08 | 2.33 |
| Contenidos básicos (1° y 2° medio) | 30 | 0.07 | 0.79 | 2.04 |
| Contenidos avanzados ( 3° y 4° medio) | 44 | 0.05 | 1.27 | 2.53 |
| | | | | |
| Contenidos básicos Número | 5 | 0.14 | 0.02 | 1.73 |
| Contenidos básicos Álgebra | 12 | 0.07 | 0.90 | 2.58 |
| Contenidos básicos Geometría | 8 | 0.04 | 1.19 | 1.86 |
| Contenidos básicos Datos y azar | 5 | 0.09 | 0.64 | 1.34 |
| | | | | |
| Contenidos avanzados Número | 6 | 0.05 | 1.14 | 2.66 |
| Contenidos avanzados Álgebra | 18 | 0.06 | 1.19 | 2.62 |
| Contenidos avanzados Geometría | 14 | 0.04 | 1.49 | 2.61 |
| Contenidos avanzados Datos y azar | 6 | 0.02 | 1.16 | 2.00 |

Fuente de Datos: Elaboración propia, DEMRE 2016.

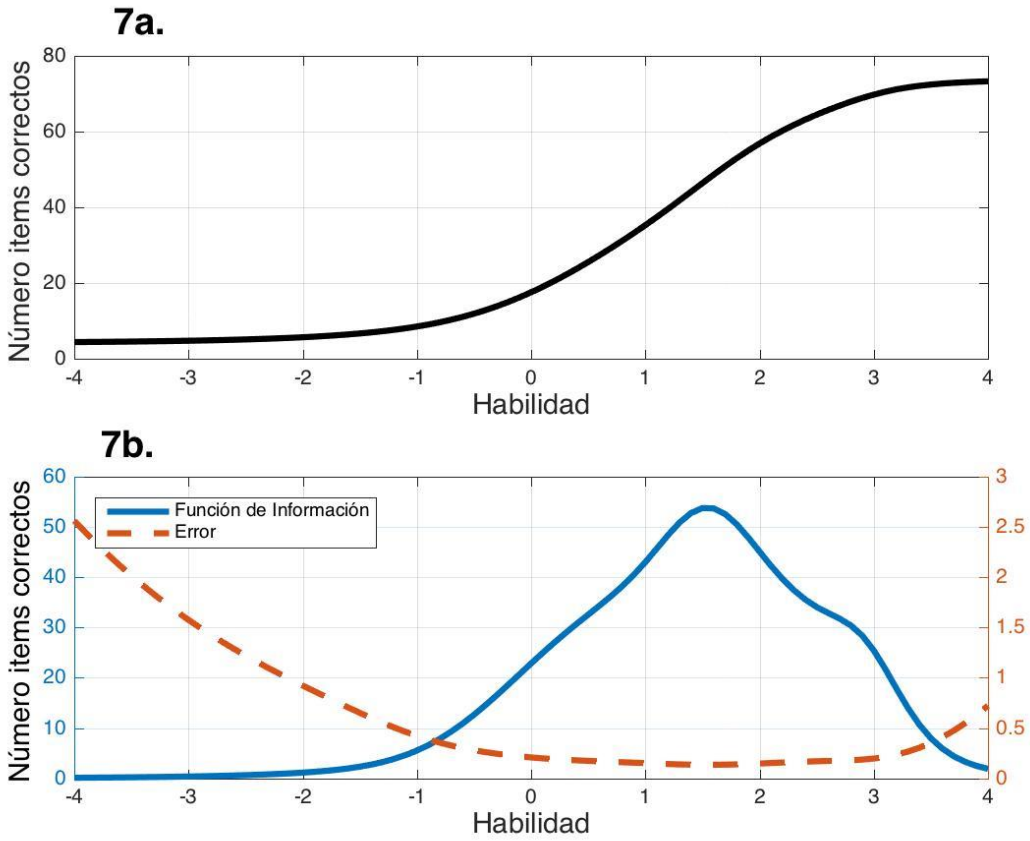
La tabla 5 muestra que los subconjuntos de ítems de contenidos básicos de números y de datos y azar son los de menor dificultad. Los de geometría, tanto los de nivel básico, son los de mayor grado de dificultad, en concordancia con los resultados obtenidos en el análisis TCM.

En las figuras 7, 8 y 9 se consignan las curvas características y las funciones de información (TCC y TIF, por sus siglas en inglés) para el conjunto de todos los ítems y para las subpruebas constituidas por los subconjuntos de ítems básicos y avanzados respectivamente. La TCC muestra el número de respuestas correctas esperadas (eje Y) en función de cada nivel de habilidad (eje X).

La TIF, por su parte, muestra la contribución sumativa de cada ítem de la prueba a la información total que entrega ésta.

En la TCC (Figura 7a) se observa que quienes se sitúan en la escala de habilidad en el valor 2 tienen un puntaje esperado en la PSU-M de aproximadamente 58 ítems acertados (78% de respuestas correctas). La curva de información (Figura 7b), por su parte, muestra que la prueba total alcanza su máximo de información entre los puntos de habilidad 1 y 2, pero aporta muy poca información en otros puntos de la escala, donde existen puntos de corte relevantes que se emplean para la toma de decisiones en materia de admisión y acceso a financiamiento y becas (por ejemplo en el nivel de habilidad 0, que corresponde en términos aproximados a puntajes escalados en torno a los 500 puntos).

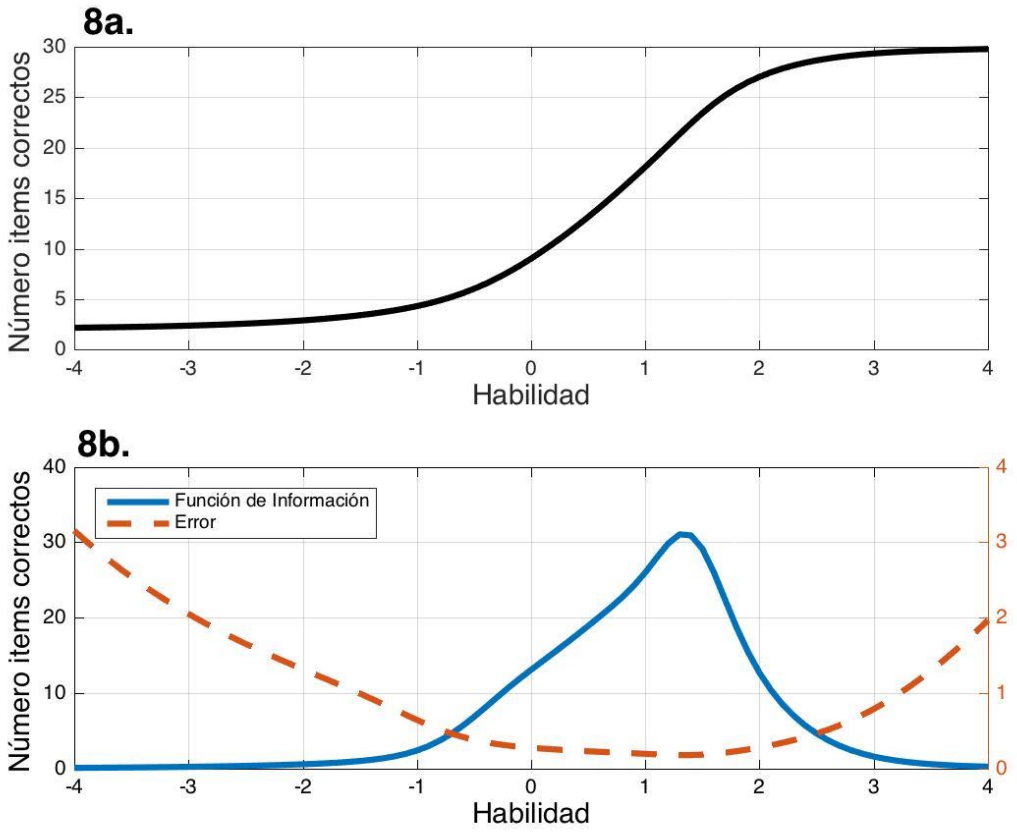
Figura 7. Curva Característica de la Prueba (7a) y Función de Información (7b). Admisión 2013: Todos los items ($N_{ítems}$=74; N=170.391).




Fuente de Datos: Elaboración propia, DEMRE 2016.

En relación a los ítems de contenidos básicos (Figura 8a), se advierte en la TCC que quienes se sitúan en la escala de habilidad en el punto 2 tendrían un desempeño esperado de aproximadamente 26 ítems acertados de un total de treinta (87% de respuestas correctas). La curva de información, por su parte, muestra que la subprueba de los ítems de contenidos básicos alcanza su máximo de información (Figura 8b) entre los puntos de habilidad 1 y 2. Al igual que sucede cuando se considera el conjunto total de 74 items, los ítems de contenidos básicos arrojan poca información en torno a puntos más bajos en la escala que se emplean para la toma de decisiones de altas consecuencias en materia de admisión y acceso a financiamiento y becas. Por tanto habría que estudiar la opción de incluir ítems que arrojen información en torno a los niveles de habilidad comprendida entre -1 y 0. En tal sentido, no bastaría con aumentar el número de ítems básicos si son de caraterísticas similares a los aquí analizados.

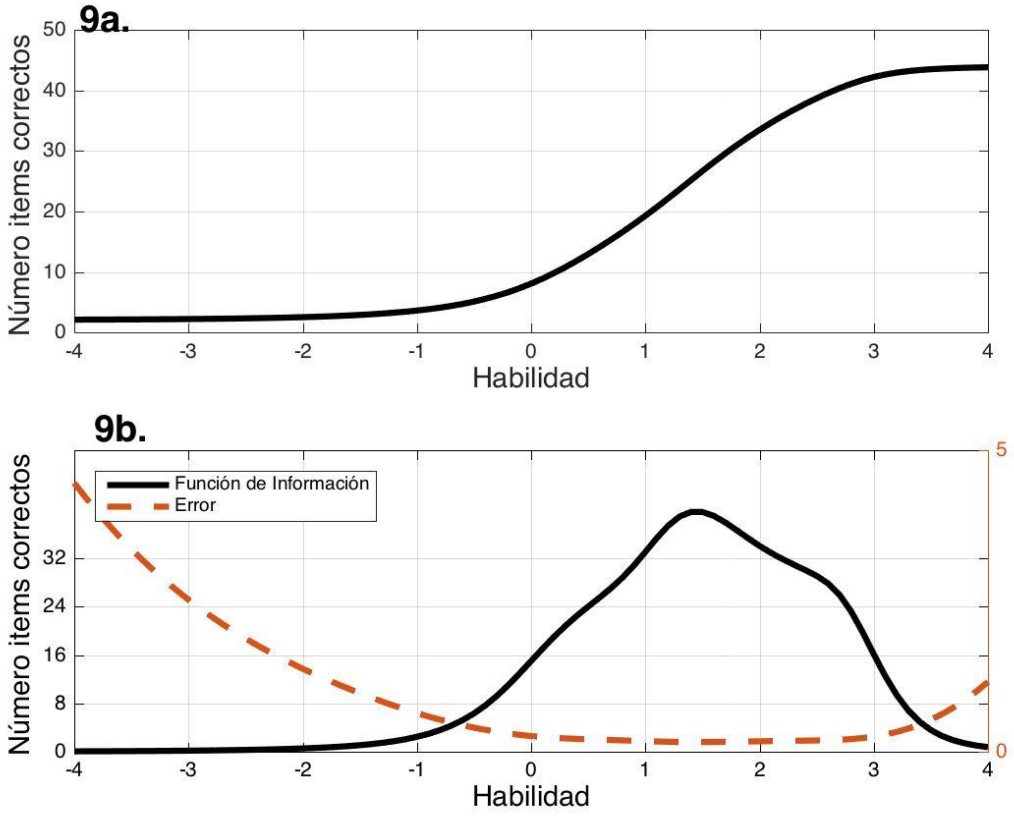
Figura 8.Curva Característica de la Prueba (8a) y Función de Información(8b).Admisión 2013: Items Básicos ($N_{ítems}$=30; N=170.391)




Fuente de Datos: Elaboración propia, DEMRE 2016.

Cuando se considera el subconjunto constituido por los ítems de contenidos avanzados (Figura 9a), en el nivel de habilidad 2 se observa que el número de respuestas correctas esperadas sería de 32 sobre un total de 44 (73%). Al igual que en los casos anteriores, el máximo de información (Figura 9b) se ubica entre 1 y 2 puntos en la escala de habilidad, aunque también se observa un buen nivel de información en el rango de habilidad entre 2 y 2,5 aproximadamente, lo cual es esperable, por tratarse del subconjunto de ítems avanzados.

Figura 9. Curva Característica de la Prueba (9a) y Función de Información (9b). Admisión 2013:Ítems Avanzados. ($N_{ítems}$=44; N=170.391)




Fuente de Datos: Elaboración propia, DEMRE 2016.

En resumen, en base a los resultados de los análisis TCM e IRT, tanto los ítems avanzados como los de contenidos básicos incluidos en la PSU-M del 2013 resultan muy difíciles para la población que la rinde, en especial para el grupo proveniente de la educación técnica profesional. Es importante que el conjunto de pruebas que compongan el sistema de admisión provea de suficiente información para todos los grupos que la rinden, tanto en los tramos altos de la distribución de habilidad —donde seleccionan las universidades más prestigiosas a los estudiantes que aspiran a ser admitidos a sus programas de elite— como en tramos más bajos de la distribución donde hacen su selección las universidades y programas menos selectivos. La PSU-M del 2013 aporta poca información en torno a áreas claves de selección —en el rango entre 450 a 500 puntos— que constituyen puntos de corte para admisión y asignación de becas y beneficios estudiantiles. (Ver Anexo C para curvas características y funciones de información por eje temático y nivel).

## 3.2 Capacidad predictiva de ítems de contenidos básicos y avanzados

Para explorar si, para fines de predicción, es necesario incorporar los contenidos más avanzados en las pruebas, se realizó un estudio empleando modelos de regresión por área de carreras, por universidad y por carreras.

Las Tablas 6 y 6a muestran estadísticas descriptivas para las variables de interés por área de carrera y universidad, respectivamente.

Tabla 6. Estadísticas descriptivas para estudiantes de la promoción 2012 matriculados en universidades del SUA según área (con la desviacón estándar en paréntesis).

| ÁREA DE CARRERA | Promedio Primer Año Universidad | Notas Enseñanza Media (NEM) | Puntaje NEM | Ranking del Alumno en Enseñanza Media | Proporción de items correctos | | |
|---|---|---|---|---|---|---|---|
| | | | | | Nivel Básico | Nivel Avanzado | Todos los ítems |
| AGRONOMIA y FORESTAL (n=683) | 4.15 (std=0.95) | 5.85 (std=0.42) | 589.78 (std=86.15) | 613.35 (std=108.96) | 0.59 (std=0.2) | 0.46 (std=0.19) | 0.53 (std=0.19) |
| ARQUITECTURA (n=1105) | 4.27 (std=1.17) | 5.89 (std=0.43) | 597.82 (std=89.15) | 619.47 (std=109.33) | 0.62 (std=0.2) | 0.48 (std=0.21) | 0.55 (std=0.2) |
| ARTE (n=586) | 4.81 (std=1.11) | 5.82 (std=0.41) | 582.53 (std=83.92) | 601.69 (std=101.56) | 0.51 (std=0.21) | 0.38 (std=0.2) | 0.44 (std=0.2) |
| CIENCIAS (n=1990) | 4.51 (std=1.11) | 6 (std=0.44) | 619.77 (std=89.66) | 647.83 (std=111.62) | 0.65 (std=0.24) | 0.52 (std=0.24) | 0.59 (std=0.24) |
| CIENCIAS QUI FARM (n=648) | 4.37 (std=1) | 6.01 (std=0.4) | 621.71 (std=82.5) | 653 (std=107.33) | 0.62 (std=0.19) | 0.49 (std=0.19) | 0.56 (std=0.18) |
| CIENCIAS SOCIALES (n=2739) | 4.77 (std=0.99) | 5.86 (std=0.42) | 591.27 (std=86.1) | 612.89 (std=107.04) | 0.49 (std=0.2) | 0.35 (std=0.19) | 0.42 (std=0.19) |
| CONSTRUCCION (n=827) | 4.03 (std=0.98) | 5.79 (std=0.39) | 576.96 (std=79.89) | 601.2 (std=103.44) | 0.58 (std=0.19) | 0.45 (std=0.18) | 0.52 (std=0.18) |
| DERECHO (n=2798) | 4.44 (std=1.01) | 5.96 (std=0.44) | 611.42 (std=90.73) | 633.84 (std=110.27) | 0.55 (std=0.22) | 0.41 (std=0.21) | 0.48 (std=0.21) |
| DISENO Y PUBLICIDAD (n=879) | 4.54 (std=1.12) | 5.76 (std=0.42) | 570.93 (std=85.95) | 586.09 (std=103.04) | 0.54 (std=0.21) | 0.40 (std=0.21) | 0.47 (std=0.2) |
| EDUC BASICA (n=535) | 5.1 (std=0.95) | 5.8 (std=0.42) | 578.94 (std=86.79) | 597.05 (std=105.89) | 0.52 (std=0.2) | 0.38 (std=0.19) | 0.45 (std=0.19) |
| EDUC DIFERENCIAL y PSICO (n=518) | 5 (std=0.89) | 5.76 (std=0.38) | 570.66 (std=78.1) | 590.97 (std=100.1) | 0.42 (std=0.15) | 0.29 (std=0.14) | 0.35 (std=0.14) |
| EDUC MEDIA CS (n=551) | 4.31 (std=1.15) | 5.98 (std=0.43) | 615 (std=89.02) | 646.87 (std=115.3) | 0.62 (std=0.2) | 0.5 (std=0.2) | 0.56 (std=0.19) |
| EDUC MEDIA HU (n=1280) | 4.63 (std=1.18) | 5.81 (std=0.41) | 581.99 (std=84.02) | 607.14 (std=108.19) | 0.44 (std=0.17) | 0.29 (std=0.15) | 0.36 (std=0.15) |
| EDUC PARVUL (n=329) | 5.15 (std=1.02) | 5.71 (std=0.43) | 559.8 (std=88.05) | 576.99 (std=109.99) | 0.46 (std=0.18) | 0.32 (std=0.17) | 0.39 (std=0.17) |
| EDUC OTRAS (n=809) | 4.65 (std=1.1) | 5.65 (std=0.4) | 548.51 (std=81.69) | 565.44 (std=102.01) | 0.47 (std=0.18) | 0.33 (std=0.16) | 0.4 (std=0.16) |

| ÁREA DE CARRERA | Promedio Primer Año Universidad | Notas Enseñanza Media (NEM) | Puntaje NEM | Ranking del Alumno en Enseñanza Media | Proporción de items correctos | | |
|---|---|---|---|---|---|---|---|
| | | | | | Nivel Básico | Nivel Avanzado | Todos los ítems |
| ENFERMERIA Y OTROS (n=5227) | 4.89 (std=0.82) | 6.07 (std=0.39) | 634.94 (std=79.5) | 668.3 (std=103.81) | 0.57 (std=0.19) | 0.43 (std=0.18) | 0.50 (std=0.18) |
| HUMANIDADES (n=500) | 4.80 (std=0.88) | 5.91 (std=0.45) | 601.59 (std=92.45) | 626.05 (std=117.15) | 0.62 (std=0.23) | 0.49 (std=0.23) | 0.55 (std=0.23) |
| IDIOMAS (n=156) | 4.93 (std=0.96) | 5.96 (std=0.35) | 612.69 (std=70.97) | 641.37 (std=97.24) | 0.42 (std=0.18) | 0.29 (std=0.15) | 0.36 (std=0.16) |
| INGENIERIA CIVIL (n=10296) | 4.29 (std=1.02) | 6.05 (std=0.45) | 630.92 (std=91.66) | 662.75 (std=114.4) | 0.73 (std=0.2) | 0.62 (std=0.22) | 0.68 (std=0.21) |
| INGENIERIA COM (n=3579) | 4.5 (std=0.85) | 5.97 (std=0.44) | 614.87 (std=90.89) | 639.07 (std=111.37) | 0.68 (std=0.22) | 0.55 (std=0.23) | 0.62 (std=0.22) |
| INGENIERIA EN EJECUCION (n=878) | 4.05 (std=0.98) | 5.82 (std=0.4) | 583.5 (std=82.5) | 611.43 (std=108.66) | 0.57 (std=0.19) | 0.43 (std=0.19) | 0.50 (std=0.18) |
| INGENIERIA OTROS (n=2328) | 4.10 (std=1.03) | 5.73 (std=0.42) | 565.58 (std=86.88) | 588.22 (std=110.6) | 0.52 (std=0.19) | 0.39 (std=0.18) | 0.46 (std=0.18) |
| LETRAS (n=199) | 4.86 (std=1.17) | 6.05 (std=0.42) | 629.93 (std=85.33) | 663.27 (std=109.9) | 0.59 (std=0.2) | 0.43 (std=0.19) | 0.51 (std=0.19) |
| MEDICINA (n=912) | 5.55 (std=0.61) | 6.68 (std=0.17) | 759.27 (std=35.88) | 809.83 (std=42.31) | 0.92 (std=0.08) | 0.84 (std=0.11) | 0.88 (std=0.09) |
| ODONTOLOGIA (n=1005) | 4.82 (std=0.89) | 6.3 (std=0.34) | 681.75 (std=69.55) | 716.58 (std=89.48) | 0.75 (std=0.15) | 0.62 (std=0.16) | 0.68 (std=0.15) |
| PERIODISMO (n=648) | 4.75 (std=0.93) | 5.78 (std=0.43) | 574.07 (std=89.1) | 589.79 (std=106.63) | 0.49 (std=0.21) | 0.36 (std=0.2) | 0.43 (std=0.20) |
| TECNO ADM (n=1321) | 4.50 (std=0.92) | 5.83 (std=0.42) | 585.82 (std=86.06) | 615.49 (std=112.6) | 0.49 (std=0.18) | 0.35 (std=0.17) | 0.42 (std=0.17) |
| TECNO NO ADM (n=1719) | 4.46 (std=1.06) | 5.9 (std=0.43) | 598.99 (std=89.17) | 631.38 (std=114.76) | 0.52 (std=0.2) | 0.39 (std=0.19) | 0.45 (std=0.19) |
| VETERINARIA (n=485) | 4.29 (std=0.78) | 5.83 (std=0.41) | 585.16 (std=85.06) | 608.51 (std=107.86) | 0.56 (std=0.19) | 0.42 (std=0.19) | 0.49 (std=0.18) |
| TOTAL (n=45530) | 4.50 (std=1.05) | 5.96 (std=0.45) | 612.13 (std=92.59) | 639.78 (std=115.29) | 0.61 (std=0.23) | 0.48 (std=0.23) | 0.54 (std=0.22) |

Fuente de Datos: Elaboración propia, DEMRE 2016.

Se observa que por área de carrera, hay diferencias sustantivas en los antecedentes de rendimiento en la enseñanza media y rendimiento en los dos subpuntajes de ítems de la PSU-M. Los más altos promedios y la menor dispersión corresponden al área de medicina, lo que revela que los alumnos admitidos en esta área presentan un rendimiento alto y homogéneo tanto en la enseñanza media como en su desempeño en la PSU-M.

Al igual que sucede al agrupar por áreas de carrera, se observan importantes diferencias en materia de rendimiento en la enseñanza media y en los puntajes de la PSU-M entre los postulantes que

ingresan a las distintas universidades del sistema. Las universidades Católica y de Chile son las que atraen a los estudiantes de más alto rendimiento.

Tabla 6a. Estadísticas descriptivas estudiantes promoción 2012 matriculados en Instituciones del SUA según universidad (desviación estándar en paréntesis)

| Institución | Promedio Primer Año Universidad | Nota Enseñanza Media (NEM) | Puntaje NEM | Ranking del Alumno en Enseñanza Media | Proporción de items correctos | | |
|---|---|---|---|---|---|---|---|
| | | | | | Nivel Básico | Nivel Avanzado | Todos los ítems |
| UCH (n=3003) | 4.66 (std=1.19) | 6.33 (std=0.31) | 687.18 (std=62.85) | 729.57 (std=82.06) | 0.82 (std=0.16) | 0.71 (std=0.19) | 0.77 (std=0.17) |
| PUC (n=3278) | 4.98 (std=0.71) | 6.4 (std=0.29) | 702.64 (std=59.56) | 739.64 (std=77.77) | 0.84 (std=0.15) | 0.74 (std=0.19) | 0.79 (std=0.16) |
| UDEC (n=3025) | 4.37 (std=1.08) | 6.08 (std=0.43) | 636.85 (std=89.08) | 665.09 (std=109.78) | 0.64 (std=0.21) | 0.51 (std=0.22) | 0.58 (std=0.21) |
| PUCV (n=1812) | 4.2 (std=0.99) | 5.99 (std=0.4) | 617.68 (std=82.1) | 645.45 (std=104.94) | 0.64 (std=0.18) | 0.5 (std=0.18) | 0.57 (std=0.18) |
| UFSM (n=2174) | 3.96 (std=1.19) | 6.12 (std=0.42) | 644.42 (std=86.8) | 681.73 (std=109) | 0.72 (std=0.23) | 0.61 (std=0.24) | 0.66 (std=0.23) |
| USACH (n=1923) | 4.81 (std=0.69) | 6.16 (std=0.31) | 652.83 (std=64.71) | 703.9 (std=91.39) | 0.67 (std=0.2) | 0.54 (std=0.2) | 0.61 (std=0.2) |
| UACH (n=1643) | 4.3 (std=1.35) | 5.92 (std=0.42) | 604.66 (std=85.81) | 634.69 (std=111.44) | 0.55 (std=0.2) | 0.42 (std=0.2) | 0.49 (std=0.19) |
| PUCN (n=1302) | 4.49 (std=0.87) | 5.93 (std=0.42) | 606.03 (std=85.73) | 631.79 (std=109) | 0.58 (std=0.2) | 0.45 (std=0.19) | 0.51 (std=0.19) |
| UV (n=1523) | 4.53 (std=1.04) | 5.97 (std=0.45) | 613.49 (std=93.57) | 642.46 (std=117.04) | 0.57 (std=0.19) | 0.43 (std=0.18) | 0.5 (std=0.18) |
| UMCE (n=413) | 4.78 (std=1.1) | 5.89 (std=0.34) | 596.86 (std=70.01) | 632.22 (std=97.94) | 0.55 (std=0.17) | 0.4 (std=0.17) | 0.47 (std=0.16) |
| UTEM (n=852) | 4.17 (std=0.97) | 5.62 (std=0.37) | 542.8 (std=75.9) | 568.58 (std=104.57) | 0.48 (std=0.16) | 0.34 (std=0.15) | 0.41 (std=0.15) |
| UTA (n=813) | 4.53 (std=1.1) | 5.94 (std=0.41) | 608.41 (std=83.71) | 627.8 (std=100.79) | 0.44 (std=0.17) | 0.32 (std=0.16) | 0.38 (std=0.16) |
| UAP (n=337) | 4.73 (std=0.77) | 5.87 (std=0.38) | 594.38 (std=77.82) | 622.08 (std=101.95) | 0.43 (std=0.19) | 0.31 (std=0.17) | 0.37 (std=0.17) |
| UANT (n=775) | 3.95 (std=1.5) | 5.78 (std=0.48) | 575.5 (std=99.33) | 598.79 (std=123.01) | 0.46 (std=0.2) | 0.33 (std=0.19) | 0.4 (std=0.19) |
| ULS (n=985) | 4.07 (std=1.2) | 5.89 (std=0.42) | 598.34 (std=87.12) | 628.18 (std=114.09) | 0.55 (std=0.2) | 0.42 (std=0.19) | 0.48 (std=0.19) |
| UPLA (n=664) | 3.87 (std=1.59) | 5.66 (std=0.42) | 549.59 (std=85.98) | 569.02 (std=109.32) | 0.39 (std=0.15) | 0.26 (std=0.13) | 0.33 (std=0.13) |
| UDA (n=477) | 4.14 (std=1.1) | 5.73 (std=0.44) | 565.65 (std=90.19) | 600.78 (std=120.06) | 0.49 (std=0.19) | 0.38 (std=0.17) | 0.43 (std=0.18) |
| UBB (n=1410) | 4.66 (std=0.84) | 5.97 (std=0.4) | 613.83 (std=82.71) | 639.41 (std=104.37) | 0.52 (std=0.19) | 0.38 (std=0.18) | 0.45 (std=0.18) |

| Institución | Promedio Primer Año Universidad | Nota Enseñanza Media (NEM) | Puntaje NEM | Ranking del Alumno en Enseñanza Media | Proporción de items correctos | | |
|---|---|---|---|---|---|---|---|
| | | | | | Nivel Básico | Nivel Avanzado | Todos los ítems |
| UFRO (n=1179) | 4.56 (std=0.78) | 5.96 (std=0.42) | 612.69 (std=86.4) | 644.09 (std=110.23) | 0.58 (std=0.2) | 0.45 (std=0.2) | 0.51 (std=0.19) |
| ULL (n=385) | 4.74 (std=0.95) | 5.72 (std=0.38) | 562.86 (std=78.94) | 595.17 (std=113.03) | 0.39 (std=0.15) | 0.26 (std=0.13) | 0.32 (std=0.13) |
| UMAG (n=246) | 4.66 (std=1.47) | 5.83 (std=0.39) | 585.26 (std=80.2) | 603.28 (std=100.29) | 0.44 (std=0.18) | 0.31 (std=0.16) | 0.38 (std=0.16) |
| UTAL (n=1378) | 4.33 (std=0.97) | 6.16 (std=0.37) | 652.32 (std=75.09) | 687.69 (std=96.67) | 0.57 (std=0.19) | 0.43 (std=0.18) | 0.5 (std=0.18) |
| UCM (n=553) | 4.69 (std=0.74) | 5.99 (std=0.39) | 618.65 (std=79.74) | 646.58 (std=102.39) | 0.52 (std=0.2) | 0.39 (std=0.19) | 0.45 (std=0.19) |
| UCSC (n=817) | 4.51 (std=1.02) | 5.88 (std=0.45) | 595.62 (std=91.67) | 617.18 (std=111.51) | 0.49 (std=0.18) | 0.36 (std=0.18) | 0.43 (std=0.17) |
| UCT (n=1106) | 4.65 (std=0.88) | 5.7 (std=0.41) | 557.56 (std=84.08) | 584.38 (std=113.58) | 0.4 (std=0.17) | 0.27 (std=0.15) | 0.33 (std=0.15) |
| UDP (n=1411) | 4.55 (std=0.78) | 5.8 (std=0.35) | 578.78 (std=72.29) | 596.65 (std=92.42) | 0.65 (std=0.17) | 0.51 (std=0.17) | 0.58 (std=0.17) |
| UMAY (n=1893) | 4.6 (std=1.21) | 5.74 (std=0.42) | 566.46 (std=86.32) | 585.26 (std=107.74) | 0.54 (std=0.2) | 0.4 (std=0.19) | 0.47 (std=0.19) |
| UFT (n=519) | 4.61 (std=0.74) | 5.75 (std=0.48) | 568.45 (std=98.05) | 589.46 (std=121.1) | 0.54 (std=0.21) | 0.41 (std=0.21) | 0.47 (std=0.2) |
| UAB (n=4898) | 4.54 (std=0.9) | 5.67 (std=0.41) | 553.1 (std=83.74) | 572.04 (std=105.93) | 0.49 (std=0.18) | 0.35 (std=0.17) | 0.42 (std=0.17) |
| UAI (n=1241) | 4.64 (std=0.68) | 5.98 (std=0.35) | 616.23 (std=72.48) | 630.26 (std=87.23) | 0.79 (std=0.17) | 0.67 (std=0.19) | 0.73 (std=0.18) |
| UAND (n=1130) | 4.86 (std=0.77) | 6.07 (std=0.38) | 635.29 (std=77.86) | 654.24 (std=96.05) | 0.77 (std=0.17) | 0.64 (std=0.19) | 0.71 (std=0.17) |
| UDD (n=1833) | 4.5 (std=1.14) | 5.78 (std=0.38) | 574.45 (std=78.33) | 583.7 (std=90.54) | 0.65 (std=0.18) | 0.51 (std=0.18) | 0.58 (std=0.18) |
| UAH (n=532) | 4.09 (std=1.47) | 5.63 (std=0.32) | 543.26 (std=66.16) | 557.68 (std=83.9) | 0.49 (std=0.16) | 0.35 (std=0.15) | 0.42 (std=0.15) |

Fuente de Datos: Elaboración propia, DEMRE 2016.

En las tablas 7a y 7b se reportan estadísticas descriptivas y correlaciones para los predictores empleados en el análisis de regresión y la variable criterio: promedio ponderado de notas obtenidas en ramos aprobados y reprobados en el primer año de universidad[34] (Se adjunta tabla de correlaciones por área en el Anexo D).

[34] Grau (2016) define y emplea dicho indicador.

Tabla 7a. Estadísticos descriptivos y correlaciones entre la variable criterio y predictores para el grupo de estudiantes de la promoción 2012 matriculados en universidades adscritas al SUA en 2013 (N=45530).

| Variables | Promedio Primer año universidad | Notas Enseñanza Media (NEM) | Ranking del Alumno en Enseñanza Media | Puntaje Items | | |
|---|---|---|---|---|---|---|
| | | | | Nivel Básico (1° y 2° Medio) | Nivel Avanzado (3° y 4° Medio) | Todos los Ítems |
| Promedio 1er Año Universidad<br>Media: 4.50<br>Desviación Estándar: 1.05 | 1 | | | | | |
| Notas Enseñanza Media (NEM)<br>Media: 5.96<br>Desviación Estándar: 0.45 | 0.31 | 1 | | | | |
| Ranking del Alumno en E. Media<br>Media:: 639.78<br>Desviación Estándar: 115.30 | 0.29 | 0.97 | 1 | | | |
| Puntaje Items Nivel Básico<br>Media: 0.61<br>Desviación Estándar: 0.22 | 0.18 | 0.45 | 0.39 | 1 | | |
| Puntaje Items Nivel Avanzado<br>Media:0.48<br>Desviación Estándar: 0.23 | 0.19 | 0.47 | 0.40 | 0.92 | 1 | |
| Puntaje todos ítems<br>Media:0.54<br>Desviación Estándar: 0.22 | 0.19 | 0.47 | 0.40 | 0.98 | 0.98 | 1 |

Fuente de Datos: Elaboración propia, DEMRE 2016.
Nota: Todas las correlaciones son significativas (p-valor menor o igual a 0.05).

Las variables que muestran más alta correlación con el promedio de notas de primer año de universidad son NEM y Ranking (0.31 y 0.29 respectivamente). Los puntajes de ítems básicos y avanzados muestran una correlación menor con la variable criterio (0.18 y 0.19 respectivamente).

Las más altas correlaciones se observan entre NEM y Ranking (.97) y entre los puntajes de los ítems avanzados y todos los ítems (0.98). Ello indica un alto grado de redundancia de información aportada por estas variables [35].

La Tabla 7b muestra las correlaciones entre el rendimiento universitario y los predictores para los alumnos según tipo de colegio.

[35] Las notas de enseñanza media y ranking están muy altamente correlacionadas en la muestra total (r=0.97), por lo cual no existen diferencias significativas si se emplea uno u otro como indicador de rendimiento en la educación media. En el estudio de capacidad predictiva se ajustaron modelos de regresión separados usando las notas de enseñanza media y ranking. Dado que las diferencias observadas eran mínimas, sólo se reportan los resultados de los modelos con NEM.

Tabla 7b. Correlaciones entre la variable criterio y predictores para el grupo de estudiantes de la promoción 2012 admitidos en universidades adscritas al SUA en 2013

| Variable Predictora | EMHC | | | EMTP | |
|---|---|---|---|---|---|
| | Particular Pagado (n=12.250) | Particular Subvencionado (n=21.402) | Municipal (n=7.905) | Particular Subvencionado (n=1991) | Municipal (n=1905) |
| Notas Enseñanza Media (NEM) | 0.39 | 0.29 | 0.26 | 0.26 | 0.25 |
| Ranking del Alumno en EM | 0.37 | 0.27 | 0.24 | 0.26 | 0.24 |
| Puntaje Items Nivel Básico | 0.23 | 0.13 | 0.19 | 0.09 | 0.1 |
| Puntaje Items Nivel Avanzado | 0.24 | 0.14 | 0.20 | 0.09 | 0.09 |
| Puntaje todos ítems | 0.24 | 0.14 | 0.20 | 0.09 | 0.1 |

Fuente de Datos: Elaboración propia, DEMRE 2016.
Nota: Todas las correlaciones son significativas (p-valor menor o igual a 0.05).

Como se puede observar en la tabla 7b, hay variabilidad en la magnitud de las correlaciones al desagregar por tipo de colegio. El grupo particular pagado es el grupo en el cual se observa la correlación más alta entre la variable criterio y las variables predictoras, sin excepción.

Al igual que cuando se considera el grupo total, al desagregar por tipo de colegio el mayor poder predictivo corresponde a las NEM y ranking.

A continuación se presentan los resultados de los modelos de regresión por área de carrera (Tabla 8). El modelo I incluye sólo NEM; el modelo II incluye NEM y puntaje de ítems básicos, y el modelo III añade a los anteriores el puntaje de ítems avanzados.[36]

[36] No se presentan los resultados de losmodelos de regresión por área desagregados según tipo de colegio debido a la escasa presencia de alumnos de colegios técnico profesionales en muchas de ellas

Tabla 8. Contribución incremental a la varianza explicada ($R^2$) del rendimiento universitario por las NEM y los subpuntajes de ítems de nivel básico y avanzado, por área de carrera.

| Área de carrera | NEM ($R^2$) | NEM + contenidos básicos($R^2$) | Incremento en $R^2$ al incorporar contenidos básicos (II-I) | NEM + Contenidos básicos+ contenidos avanzados($R^2$) | Incremento en $R^2$ al incorporar contenidos avanzados (III-II) |
|---|---|---|---|---|---|
| | (I) | (II) | | (III) | |
| Agronomía y Forestal (n=471) | 9.95 | 18.39 | 8.44* | 19.67 | 1.28* |
| Arquitectura (n=1029) | 6.79 | 6.93 | 0.14 | 7.69 | 0.76* |
| Arte (n=194) | 12.03 | 13.02 | 0.99 | 13.02 | 0 |
| Ciencias (n=1529) | 8.74 | 10.23 | 1.49* | 10.27 | 0.04 |
| Humanidades (n=422) [a] | 30.34 | -- | -- | 45.23 | -- |
| Letras (n=126) | 9.57 | 11.33 | 1.76 | 13.09 | 1.76 |
| Ciencias Quim y Farmacia (n=388) | 14.84 | 25.59 | 10.75* | 28.37 | 2.78* |
| Construccion (n=711) | 4.79 | 9.29 | 4.5* | 9.45 | 0.16 |
| Ciencias sociales (n=1923) | 13.3 | 14.45 | 1.15* | 14.74 | 0.29* |
| Derecho (n=2695) | 12.48 | 13.91 | 1.43 | 14.19 | 0.28 |
| Diseño y publicidad (n=494) | 13.49 | 13.75 | 0.26 | 13.75 | 0 |
| Pedagogia en Educ. Parvularia (n=48) | 15.1 | 26.38 | 11.28* | 26.42 | 0.04 |
| Pedagogia general básica (n=272) | 15.31 | 17.27 | 1.96* | 17.27 | 0 |
| Pedagogía en Educ. Diferencial (n=215) | 2.17 | 2.68 | 0.51 | 3.25 | 0.57 |
| Otras Pedagogías (n=497) | 7.55 | 7.91 | 0.36 | 8.04 | 0.13 |
| Pedagogia media cs (n=77) | 9.58 | 19.39 | 9.81* | 21.37 | 1.98 |
| Pedagogia media hu (n=446) | 11.4 | 11.79 | 0.39 | 12.05 | 0.26 |
| Idiomas (n=61) | 10.05 | 33.75 | 23.7* | 33.75 | 0 |
| Enfermeria y otros (n=4183) | 8.65 | 13.06 | 4.41* | 13.37 | 0.31* |
| Ingenieria civil (n=9175) | 11.36 | 15.04 | 3.68* | 16.24 | 1.2* |
| Ingenieria com (n=3407) | 10.1 | 11.86 | 1.76* | 11.91 | 0.05 |
| Ingenieria otros (n=1086) | 9.32 | 12.08 | 2.76* | 12.9 | 0.82* |
| Medicina (n=794) | 6.25 | 11.96 | 5.71* | 13.01 | 1.05* |
| Odontologia (n=954) | 5.13 | 8.54 | 3.41* | 9 | 0.46* |
| Periodismo (n=411) [a] | 8.43 | -- | -- | 8.63 | -- |
| Tecno Univ en adm (n=810) | 7.57 | 11.41 | 3.84* | 11.98 | 0.57 |

| Área de carrera | NEM ($R^2$) (I) | NEM + contenidos básicos($R^2$) (II) | Incremento en $R^2$ al incorporar contenidos básicos (II-I) | NEM + Contenidos básicos+ contenidos avanzados($R^2$) (III) | Incremento en $R^2$ al incorporar contenidos avanzados (III-II) |
|---|---|---|---|---|---|
| Tecno Univ otros (n=1113) | 13.98 | 18.54 | 4.56* | 19.89 | 1.35* |
| Ingenieria en ejecución (n=553) | 2.45 | 7.14 | 4.69* | 7.19 | 0.05 |
| Veterinaria (n=468) | 9.99 | 13.8 | 3.81* | 15.83 | 2.03* |

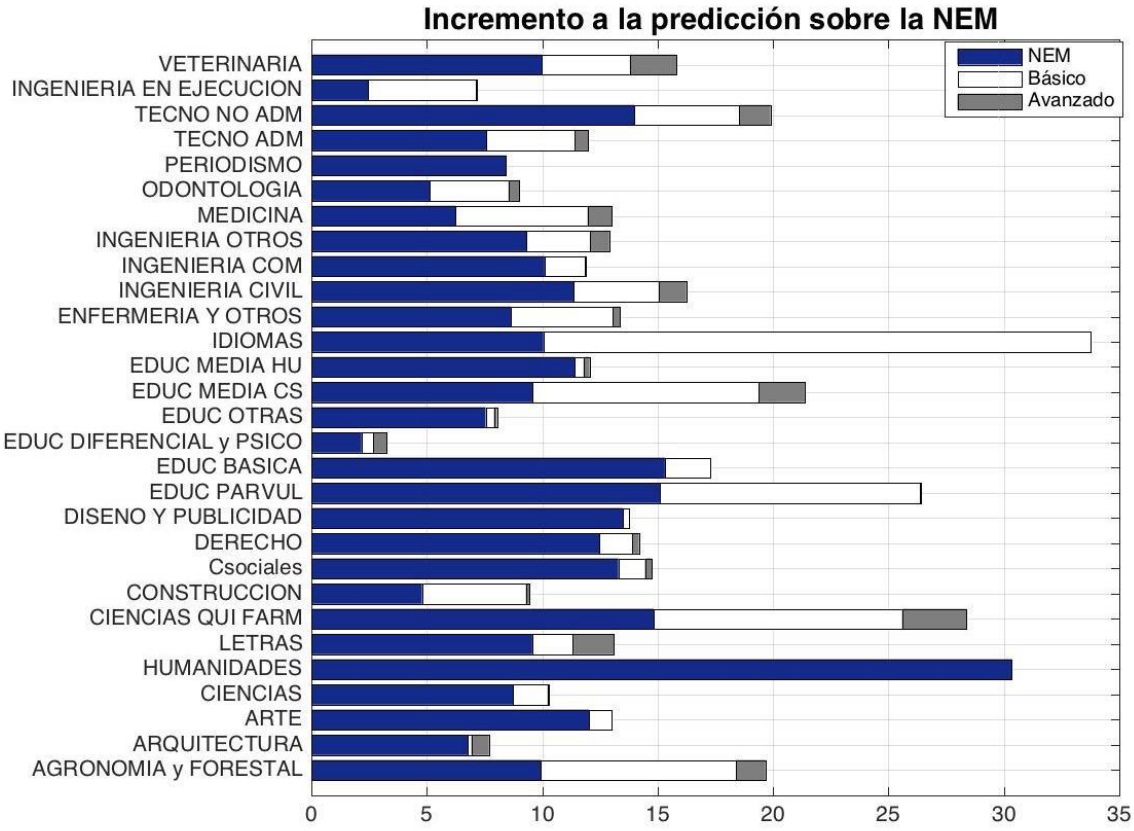
*Incremento estadísticamente significativo (p<.05)

[a] Uno o ambos coeficientes beta asociados a puntajes en la PSU son de signo negativo.

Fuente: Elaboración propia, DEMRE 2016.

La Figura 10 muestra en forma gráfica los resultados de la Tabla 8.

Figura 10. Contribución a la Predicción del Rendimiento Universitario de las NEM, Contenidos Básicos y Avanzados por Área de Carrera




Fuente de Datos: Elaboración propia, DEMRE 2016.

En líneas generales, la primera conclusión que se puede extraer a partir del análisis es que hay una amplia variabilidad en cuanto al aporte relativo de los predictores a la explicación de la varianza. El aporte de las NEM, como predictor único, fluctua entre 2.17% y 30.3%. La contribución adicional de

los ítems de contenidos básicos de la PSU-M por sobre las NEM varía entre 0.14% hasta un 23,7%. El aporte de los contenidos avanzados por sobre las NEM y contenidos básicos fluctúa entre 0% y 2.78%.

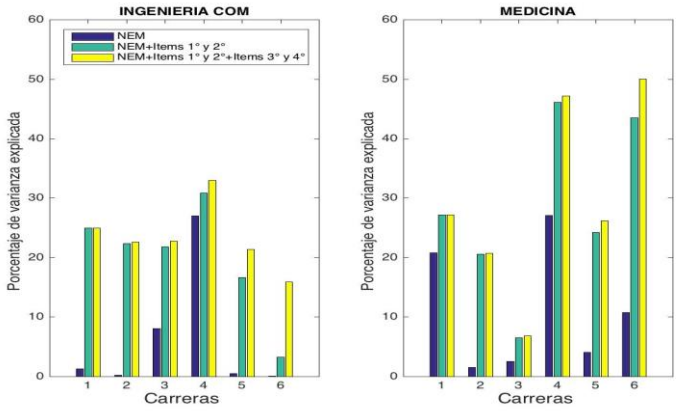
La segunda observación es que hay áreas donde la PSU-M está entregando muy poca o nula información adicional a lo que aportan las NEM, como por ejemplo Periodismo, Arte, Ciencias Sociales, Arquitectura y Diseño, entre otras.

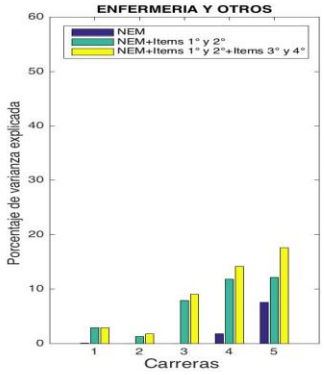
Finalmente, los puntajes de ítems de contenidos avanzados contribuyen muy poco a la explicación de la varianza, una vez que las notas y el puntaje de contenidos básicos están en la ecuación, variando su aporte entre 0% a 2.78%, observándose el mayor aporte a la predicción en el área de Química y Farmacia. Ello se explica por la alta redundancia de información que entregan los puntajes de ítems de nivel básico y avanzado, donde la correlación entre ambos es de 0.92. Una vez incorporado uno de estos predictores en la regresión, el segundo aportará poco a la explicación de la varianza por sobre el primero.

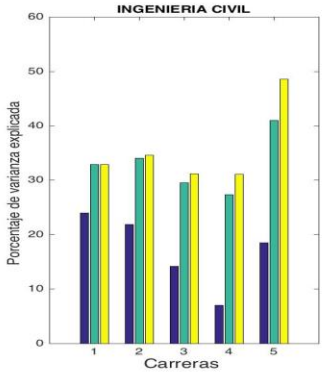
No obstante, al realizar el mismo análisis desagregado por carreras se advierte que hay una importante variabilidad entre carreras ofrecidas por distintas universidades, que no son evidentes al inspeccionar los datos agregados por área. Por ejemplo, el aporte de los contenidos avanzados a la explicación de la varianza va desde 0% a 28.7%. No obstante para un 55% de las carreras , el aporte de los contenidos avanzados no alcanza al 2% por sobre la NEM y los contenidos básicos.

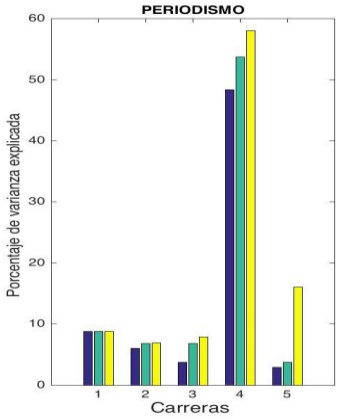
A continuación se presenta  en forma gráfica el incremento en el porcentaje de varianza explicada de algunas carreras seleccionadas dentro de sus respectivas áreas, que muestran la variabilidad existente entre carreras ofrecidas por distintas universidades.

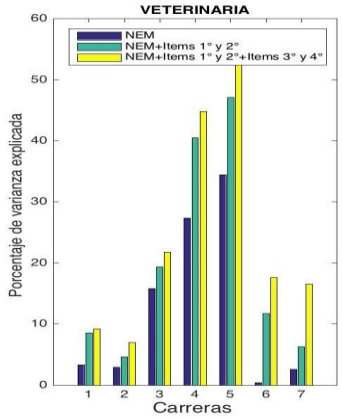
Figura 11. Incremento en el Porcentaje de Varianza Explicada entre Modelos 1, 2 y 3 para algunas carreras y áreas seleccionadas.

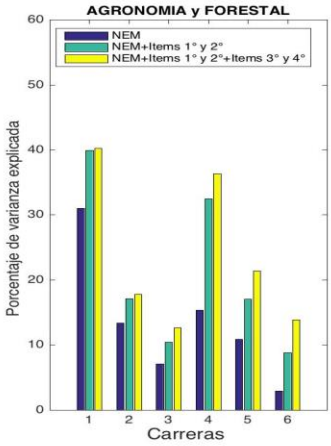


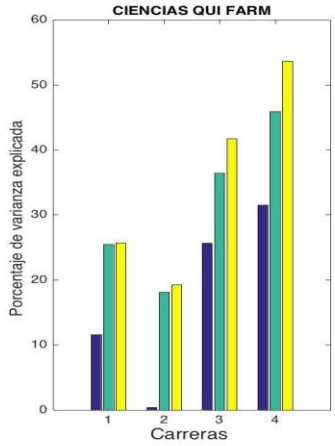

ENFERMERIA Y OTROS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

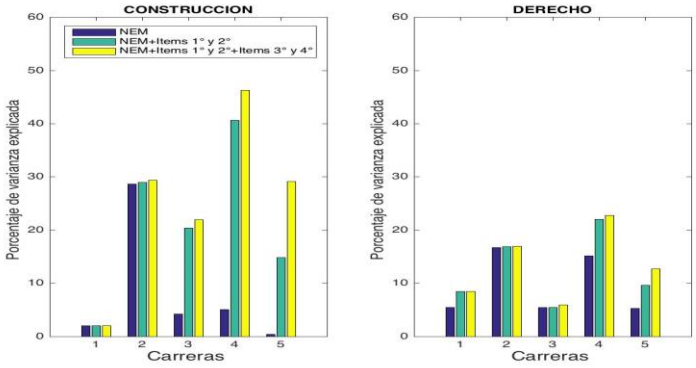

INGENIERIA CIVIL
Porcentaje de varianza explicada
Carreras


PERIODISMO
Porcentaje de varianza explicada
Carreras


VETERINARIA
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras


AGRONOMIA y FORESTAL
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras


CIENCIAS QUI FARM
Porcentaje de varianza explicada
Carreras




A modo de ejemplo, en el área de construcción, se da que en la carrera 1 ni las NEM ni la PSU-M aportan a la predicción. En la carrera dos, las NEM presentan una moderada capacidad predictiva y los ítems básicos y avanzados aportan poco o nada a la predicción, más allá de lo que las NEM logran predecir. En la carrera 4, sin embargo, las NEM tienen una muy baja capacidad de predicción, mientras que los ítems de conocimientos básicos hacen un aporte adicional importante y los de conocimientos avanzados hacen asimismo un aporte adicional significativo, aunque menor, a la predicción.

En ingeniería civil, también se advierte un patrón variable. En las carreras 1 y 2 los ítems de contenidos avanzados aportan cero información adicional más allá de la que aportan las NEM y el puntaje de contenidos básicos. Sin embargo, en la carrera 5, los ítems de contenidos avanzados hacen un aporte significativo y sustantivo a la predicción. Dentro de las carreras de ingeniería algunas pueden requerir de ítems que evalúen en mayor profundidad los contenidos avanzados de la actual PSU-M. (En el AE se consignan gráficos adicionales de carreras).

La conclusión general que puede extraerse es que, dentro de la misma área, hay una alta variabilidad entre carreras en cuanto al aporte de los ítems de la PSU-M en la predicción del rendimiento del primer año. Hay carreras en las cuales los ítems de matemática de ambos niveles hacen un aporte a la predicción del rendimiento. En particular, con respecto a los ítems avanzados hay carreras en las cuales éstos contribuyen a mejorar la predicción mientras que en otras su contribución a la predicción del rendimiento es nula. Por tanto, sería conveniente considerar el desarrollo de instrumentos que evalúen contenidos más y menos avanzados en matemáticas para responder a la diversidad de carreras ofrecidas por las universidades del SUA. La evidencia indica que contar con dos instrumentos — que evalúen contenidos básicos y avanzados respectivamente— debiera ser un aporte a la predicción del rendimiento, y la decisión de cuál o cuáles pruebas emplear debe ser analizada tomando en cuenta el contexto de la carrera y de la universidad en la que se imparte.

## 3.4 Contenidos y Brechas de Rendimiento

A fin de estudiar si existen diferencias entre los rendimientos de los alumnos que egresan de distintos tipo de colegio en los ítems de la PSU-M agrupados según eje y nivel, se estimaron brechas estandarizadas. Éstas constituyen una métrica para cuantificar las diferencias en el desempeño entre dos grupos (por ejemplo, particular pagado y particular subvencionado CH). En este estudio se estimaron las brechas como la diferencia estandarizada entre el grupo egresado de la educación particular pagada (grupo de referencia) y los egresados de los otros tipos de colegio:

$$(P_r - P_c)/s$$

donde:

$P_r$ es la proporción media de respuesta correcta del grupo de referencia,
$P_c$ es la proporción media de respuesta correcta del grupo con el cual se compara y
s es la desviación estándar

Así expresada, la magnitud de la brecha representa la diferencia de rendimiento que existe entre el grupo particular pagado y algún otro tipo de colegio, en función de su desviación estándar. Así, una brecha de magnitud 1.0 indica que el rendimiento del grupo particular pagado está a una desviación estándar por sobre la media del grupo con el cual se compara (ver en anexo F la fórmula de la desviación estándar).

Se estimaron las brechas para diversos puntajes basados en los ítems de la PSU-M categorizados según eje temático y nivel. Se reportan las brechas de rendimiento para todos los estudiantes como también para el subconjunto de alumos destacados (definido como aquellos que tienen un promedio de NEM mayor o igual a 6.0).

La Tabla 9 reporta las brechas de rendimiento estandarizadas que se observan en la PSU-M entre el grupo de estudiantes de la educación particular pagada (grupo de referencia) y los grupos de alumnos que asisten a la educación subvencionada y municipal, desagregando por rama educacional (HC y TP). Se incluye además información para el subgrupo de alumnos destacados (que obtienen un promedio de NEM de 6,0 o más). La figura 12 muestra los datos de la tabla en forma gráfica.

Tabla 9. Brechas PSU Matemáticas 2013: diferencias estandarizadas de la proporción media de respuestas correctas obtenida por el grupo de referencia (PPHC) y los estudiantes egresados de la EMCH y EMTP municipal y subvencionada.

| Grupos | | Núm. Básico | Núm. Avanzado | Alg. Básico | Alg. Avanzado | Geo Básico | Geo Avanzado | Datos y Azar Básico | Datos y Azar Avanzado | Ítems Básicos | Ítems Avanzado |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **PPHC - MuHC** | Todos | 1.0 | 1.1 | 1.3 | 1.4 | 1.3 | 1.4 | 1.0 | 1.2 | 1.4 | 1.4 |
| | *Destacados* | *0.8* | *1.0* | *1.2* | *1.2* | *1.2* | *1.3* | *1.0* | *1.2* | *1.2* | *1.3* |
| **PPHC - MuTP** | Todos | 1.3 | 1.4 | 1.7 | 1.7 | 1.6 | 1.6 | 1.1 | 1.4 | 1.7 | 1.8 |
| | *Destacados* | *1.3* | *1.6* | *1.8* | *1.9* | *1.7* | *1.8* | *1.4* | *1.5* | *1.9* | *1.9* |
| **PPHC - PSHC** | Todos | 0.7 | 0.9 | 1.1 | 1.1 | 1.1 | 1.2 | 0.8 | 0.9 | 1.1 | 1.2 |
| | *Destacados* | *0.4* | *0.7* | *0.8* | *0.8* | *0.8* | *0.9* | *0.7* | *0.9* | *0.8* | *0.9* |
| **PPHC - PSTP** | Todos | 1.2 | 1.4 | 1.6 | 1.7 | 1.5 | 1.6 | 1.0 | 1.3 | 1.6 | 1.7 |
| | *Destacados* | *1.1* | *1.4* | *1.7* | *1.7* | *1.6* | *1.7* | *1.2* | *1.4* | *1.7* | *1.8* |

Las brechas reportadas en la tabla 9 son de gran magnitud y consistentemente favorecen a los alumnos de colegios particular pagado.

En todos los puntajes por eje y nivel, los alumnos de la educación subvencionada HC son los que presentan las menores brechas de rendimiento con respecto a los alumnos de establecimientos particulares pagados. La menor distancia en el rendimiento se observa para los ítems de números de nivel básico (0.7 D.S. para todos los alumnos y 0.4 D.S. para los destacados).

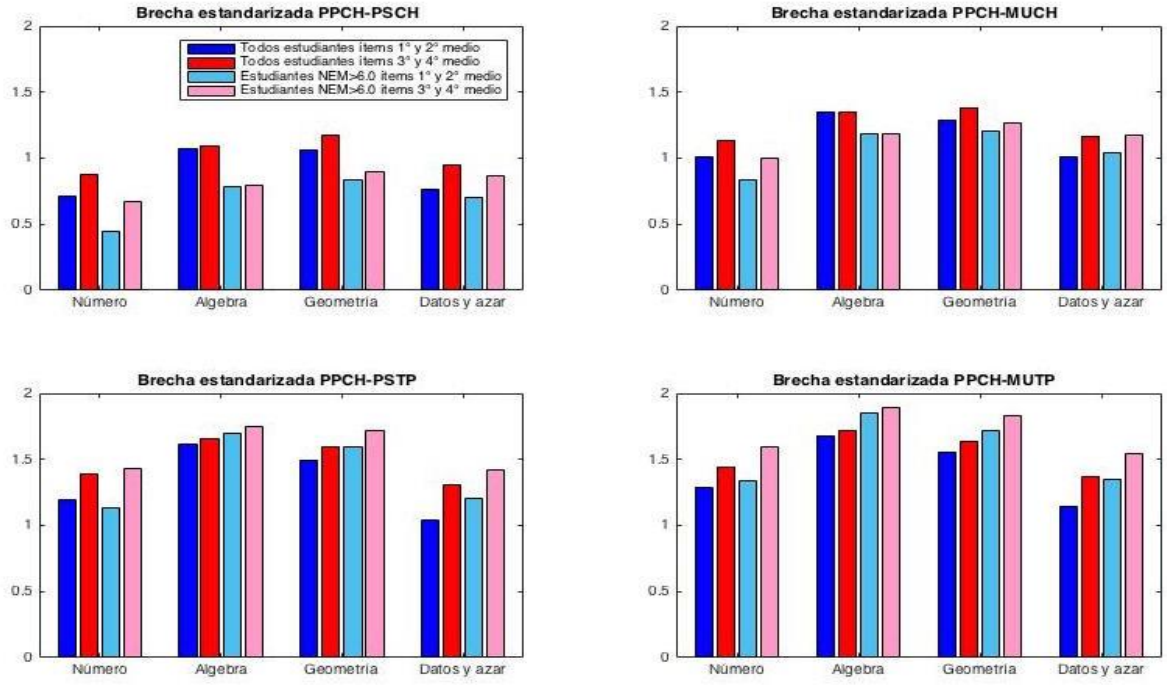
Las mayores brechas corresponden a los alumnos de la educación TP. Sea que ellos provengan de la educación municipal o particular subvencionada y sean o no alumnos destacados, las brechas entre los estudiantes particulares pagados y los grupos de la EMTP son homogéneamente más altas.

Las brechas entre particulares pagados y otros alumnos HC (municipales y subvencionados) son levemente menores para alumnos destacados, cuando se las compara con el grupo total. Por ejemplo, para los ítems básicos la brecha entre PPHC y MuHC es de 1.4 para todos los alumnos y 1.2 para el grupo destacado. Tratándose de los alumnos TP su rendimiento escolar no se traduce en menores brechas.

En términos de la magnitud de las brechas, las de menor tamaño, para todos los grupos, corresponden a los ejes temáticos de números y datos y azar de nivel básico.[37] Las brechas correspondientes a los ejes temáticos de algebra y geometría de nivel básico son prácticamente tan altas como las de nivel avanzado. Especialmente para los grupos provenientes de la EMTP, tanto los ítems avanzados como los básicos son en extremo difíciles.

[37] A la luz de la menor magnitud de la brecha para los ítems básicos de números y datos y azar, se corrieron análisis adicionales de capacidad predictiva, separando estos dos ejes de los de álgebra y geometría. Al correr estos análisis se advierte que para muchas áreas (y carreras), los contenidos básicos de álgebra y geometría hacen un aporte adicional relevante a la predicción del rendimiento. Los resultados de las regresiones se consignan en el Anexo D.

Figura12. Brechas PSU-M por eje y nivel




Fuente de Datos: Elaboración propia, DEMRE 2016.

En resumen, los resultados del análisis de brechas indican que hay variabilidad en la magnitud de las brechas dependiendo del tipo de ítems que se analice. Las menores brechas se observan en los ítems de menor dificultad, aquellos correspondientes a los ejes de número y datos y azar de nivel básico. Vale decir. si la PSU-M se restringiera a evaluar contenidos del eje temático de números y datos y azar hasta segundo medio se observarían brechas levemente menores. Ello no ocurre con los ítems que se basan en los contenidos de álgebra y geometría de primero y segundo medio, observándose brechas más altas y semejantes a las de los ítems de tercero y cuarto medio.

Las mayores brechas pueden estar indicando una cobertura curricular más deficitaria de esos contenidos, en especial para los alumnos de la EMTP que por currículum tienen menos horas disponibles para dedicar a matemáticas en tercero y cuarto medio. Este es un aspecto que debe estudiarse puesto que para orientar los cambios en esta materia no solo se requiere de evidencia psicométrica sino de análisis cualitativos que permitan responder a la pregunta de si todos los grupos, incluidos los alumnos de la EMTP, tuvieron la oportunidad durante su trayecto escolar en la enseñanza media, de aprender la totalidad de los contenidos evaluados en la PSU-M.

El impacto de las brechas observadas no debe subestimarse, y es un tema importante de estudiar, puesto que cuando los puntajes en las pruebas tienen un peso importante en la admisión—como es el caso de Chile—las grandes brechas de puntajes tienen un alto impacto en la probabilidad de admisión de los grupos desaventajados.[38]

[38] Ver al respecto los trabajos de Daniel Koretz (2000 y 2002), citados en las referencias

Finalmente, con miras a la reducción de las brechas de rendimiento, en particular las que se observan con respecto a los grupos de la EMTP, se podría analizar—además de la reducción del actual temario—la inclusión de materias de 7° y 8° básico que sean predictivas del rendimiento universitario y evaluar si, por esa vía, se logra verificar una reducción de las brechas en la PSU-M.

# Discusión

Este estudio, tuvo como objetivo explorar el impacto que podría tener la reducción de contenidos en la PSU-M, sugerida en el Informe de Pearson, en términos de su calidad como instrumento de admisión. Los expertos de Pearson señalaron explícitamente la necesidad de realizar estudios para explorar el efecto de focalizar las pruebas en los contenidos mínimos obligatorios de primero y segundo medio, puesto que a partir del tercero medio los currículos HC y TP difieren[39]. Señalaron que capacidad predictiva de la PSU-M está bajo los estándares internacionales y su grado de dificultad es inadecuado para el grupo que la rinde. Estos tres aspectos fueron abordados en este estudio.

## 4.1 Conclusiones y Sugerencias

Los resultados preliminares muestran que, en términos del grado de dificultad, privilegiar el tipo de ítems que evalúa contenidos de los dos primeros años de enseñanza media de la versión 2013 no resolvería completamente el problema del inadecuado grado de dificultad de la PSU-M, puesto que aún estos ítems son demasiado difíciles para la población general que la rinde.

En cuanto a la confiabilidad ante la pregunta de si es posible reducir contenidos en la PSU-M con el fin de corregir su inadecuado grado de dificultad, sin comprometer significativamente la confiabilidad de la medición, la respuesta es afirmativa. Se podrían reducir contenidos de nivel avanzado y acortar la prueba a sólo 60 ítems, mejorando su grado de dificultad y manteniendo el mismo nivel de confiabilidad que hoy se tiene con 74 ítems.

En relación a la capacidad predictiva y la magnitud de las brechas, no se advierten efectos secundarios negativos al focalizar las pruebas en contenidos de primero y segundo medio. No se sacrifica capacidad predictiva para muchas de las carreras y podría haber una leve reducción en la brecha para algunos sub-grupos según dependencia y rama educacional.

Los análisis de la TCM como de IRT revelan un inadecuado grado de dificultad de la PSU-M en relación con la población que la rinde. Este problema ya fue detectado en el informe del ETS del año 2005 y ratificado en el de Pearson Education (2013), y la situación no ha variado significativamente en esta versión de la PSU-M. El conjunto de ítems de contenidos básicos que evalúan materias curriculares de los dos primeros años de enseñanza media (contenidos básicos), muestra un grado de dificultad algo más adecuado a la población que rinde cuando se lo compara con el subconjunto de ítems de los dos últimos años (contenidos avanzados), pero sigue siendo aún de dificultad excesiva, en particular los que corresponden al eje temático de algebra y geometría.

En general se observan ítems cuyo grado de dificultad escapa de los rangos esperables en este tipo de pruebas, lo cual no es de extrañar puesto que la prueba se enfoca en el currículo científico humanista (Pearson Report, pg. 153), que no es cubierto en todos los establecimientos educacionales del país. En este sentido, la evidencia es contundente en cuanto a que el grado de dificultad de la PSU-M es mayor para los alumnos de la EMTP.

[39] Señala textualmente el informe: *"…the recommendation is to advance studies on the effect that the decision to approach the test with a priority on the* ***CMOs of grades 1 and 2 in high school*** *education may have …" (p.12, sin* resaltar *en el original).*

Asociado a lo anterior, otro problema que presenta la PSU-M es que no aporta suficiente información en tramos claves de la distribución de habilidad, particularmente en torno a puntos de corte relevantes para fines de admisión. Este es un problema que también se arrastra desde el año 2005, que es abordado en el informe del ETS del año 2005 (pág.16) y que debe ser resuelto.

Los resultados de la TCM se verifican también en el análisis IRT donde los ítems básicos se comportan de forma muy similar a los ítems de nivel avanzado, siendo de un grado de dificultad excesivo, con lo cual se concentra la información en tramos altos de la distribución de habilidad. Si bien esto puede ser deseable para algunas carreras, existen puntos de corte relevantes en tramos inferiores de la distribución que en la actualidad no están siendo cubiertos.

Cabe recordar que a la PSU-M se le agregaron ítems adicionales de mayor dificultad para discriminar en los tramos más altos de la distribución de habilidad, agudizando con ello el problema del inadecuado grado de dificultad para el grueso de la población que la rinde. No es claro, por lo demás que al añadir esos ítems adicionales se haya resuelto el problema de discriminar entre los alumnos de más alta habilidad.

El análisis de capacidad predictiva consistió en estimar la contribución de la PSU-M a la predicción del rendimiento universitario por sobre el aporte de las NEM. La mejor capacidad predictiva de la PSU-M se observa en el grupo de estudiantes provenientes de la educación particular pagada.

Al estimar la capacidad predictiva de los ítems de primero y segundo medio separadamente de los de tercero y cuarto, se advierte que la capacidad predictiva en la mayoría de las áreas definidas en este estudio no mejora significativamente al incorporar los ítems que evalúan contenidos de los cursos más avanzados. Para muchas carreras, los ítems avanzados de la actual prueba aportan poco o nada a la capacidad predictiva, luego de considerar las NEM y el puntaje de los contenidos básicos. No obstante, para algunas carreras los contenidos de los dos últimos años de enseñanza media sí hacen un aporte adicional a la predicción.

En tal sentido, no se ha encontrado evidencia de que reformular la PSU-M vaya a traducirse en un una merma de la capacidad predictiva de la prueba para muchas de las carreras. No obstante, en esta materia hay amplio espacio para mejorar y el desafío no debiera ser solo mantener la actual capacidad predictiva de la PSU-M sino aumentarla, como señalan los expertos de Pearson en su informe.[40] Con miras a este propósito se requerirá de estudios, tanto cualitativos como cuantitativos que permitan reformular la PSU-M de manera que provea información útil para predecir el éxito universitario en las distintas carreras del sistema, más allá de la información que en la actualidad aporta y que para muchas carreras resulta poco relevante.

Un tema pendiente, y que debe ser abordado previo a implementar cambios en la PSU-M, es estudiar la relevancia de otros contenidos y habilidades que actualmente no están incluidos en esta prueba,

[40] Los datos recogidos en este estudio son consistentes con los reportados en el informe de Pearson, y la conclusión de sus expertos sigue siendo vigente para muchas carreras cuando señalan: “If the PSU tests do not predict university outcomes, then why are we using them? Currently, the predictive validty coefficients of the PSU tests are low with respect to those seen internationally. This indicates the need to continue exploring in depth the relationship between the PSU test scores and the variables associated with university success…” (p.64).

para predecir el rendimiento académico. Sería deseable, en tal sentido, estudiar si la capacidad predictiva de la PSU-M puede potenciarse de manera que aporte mayor información para fines de seleccionar alumnos idóneos para las distintas carreras del sistema universitario, dado que se trata de una prueba obligatoria. Claramente, hay carreras en que ni los ítems de contenidos básicos ni avanzados contribuyen a predecir bien el rendimiento universitario. Ello indica que para ciertas carreras hay un mal calce entre lo que mide la PSU-M y las exigencias de dominio de contenidos y habilidades que los alumnos enfrentan al ingresar a ellas.

No existen a la fecha estudios a nivel del sistema universitario que describan los contenidos que deben manejar los estudiantes para tener éxito en las distintas carreras. En un contexto tan heterogéneo como el chileno, no se debe asumir que los requerimientos de manejo conceptual de las matemáticas sea el mismo en todas las carreras. En tal sentido, debería considerarse contar con dos pruebas de matemáticas, una que mida capacidad de razonamiento matemático sobre la base de contenidos básicos para aquellas carreras que no exigen del manejo de contenidos avanzados por parte de los postulantes, y una segunda que sí evalúe éstos para quienes aspiran a ingresar a carreras que los requieren.

Un tema pendiente y que reviste la mayor relevancia a la luz de los estándares internacionales en el desarrollo de pruebas de altas consecuencias como la PSU, es el de la legitimidad del uso de la PSU-M para todos los grupos que la rinden, puesto que estaría midendo contenidos que no aportan necesariamente a la predicción y que un contingente importante de la población estudiantil no ha tenido la oportunidad de aprender. En otras palabras, sus puntajes no serían reflejo de la habilidad del estudiante ni de su capacidad de aprender, sino de la oportunidad de aprender.

Las brechas de rendimiento varían al analizar subconjuntos de ítems definidos por nivel y eje temático. Las brechas de contenidos básicos de números y datos y azar muestran las menores brechas en la PSU-M 2013. Los subconjuntos restantes, incluyendo álgebra y geometría de nivel básico presentan brechas de mayor magnitud.

Las brechas más amplias se observan entre el grupo particular pagado y los grupos provenientes de la EMTP, tanto municipal como subvencionada. Son estos estudiantes, quienes muestran una mayor desventaja en esta prueba con respecto a los particulares pagados. Las brechas de los alumnos destacados son menores que las calculadas sobre la base de todos los alumnos para los grupos de la educación científica humanista (subvencionada y municipal). Ello no sucede para los alumnos destacados provenientes de la EMTP, evidenciando que la PSU-M resultaría inadecuada aún para los mejores alumnos de dicha modalidad educacional.

Es positivo, sin embargo, que se observa variación en las brechas según el tipo de ítems estudiado. Aquellos ítems en los cuales se observa una menor brecha son los de números y datos y azar, de primero y segundo medio. Si la meta es lograr maximizar la capacidad predictiva del instrumento y su equidad, habría que estudiar la incorporación de contenidos más básicos —de séptimo y octavo básico inclusive— que podrían contribuir a la predicción del éxito universitario y que no están siendo incorporados en la actual PSU-M.

Cabe enfatizar en este punto que los análisis psicométricos, si bien necesarios, no son suficientes para fines de introducir cambios efectivos en las tablas de especificaciones de la PSU-M. Se requiere en este

caso, no solo de eliminar sino de incorporar nuevos contenidos temáticos y quienes están en mejor posición para emitir juicios fundamentados acerca de cuáles son esos contenidos son los profesores universitarios y los profesores de enseñanza media. Los primeros saben qué contenidos deben manejar los alumnos para tener éxito en sus cursos, los segundos pueden emitir un juicio acerca de si tales contenidos están siendo cubiertos en los liceos del país, incluidos los de la EMTP.

Por tanto, quedan pendientes estudios de naturaleza cualitativa que permitan responder cuáles de los contenidos medidos en la actualidad son adecuados, qué otros contenidos requieren de ser incluidos, y analizar si los estudiantes tienen la oportunidad de aprender dichos contenidos[41], condición indispensable para legitimar el uso de una prueba de tan altas consecuencias como la PSU.

Con miras a transitar hacia una PSU-M que sea más predictiva y justa para todos los grupos que la rinden, se debe revisar el marco conceptual de la prueba y su tabla de especificaciones, conforme a lo indicado en el informe de Pearson. Sobre la base de la evidencia revisada en este estudio, mantener los ítems de tercero y cuarto medio no se justifica para la mayoría de las carreras ofrecidas por el SUA, pero eliminar éstos no basta por sí solo para garantizar que la prueba sea más predictiva y más adecuada para evaluar las capacidades de los estudiantes de menores recursos, en especial aquellos que asisten a la EMTP, municipal y subvencionada.

En resumen, los resultados obtenidos a partir de la promoción 2012 permitirían acoger la recomendación de los expertos de Pearson de estudiar el desarrollo de un nuevo marco de referencia y tablas de especificaciones, acotando los contenidos a evaluar hasta aquellos cubiertos en los dos primeros años de la enseñanza media. Una prueba de esta naturaleza sería de utilidad para muchas carreras, mas no para aquellas que requieren de manejo de contenidos avanzados por lo que se requeriría de otra prueba de contenidos avanzados. La decisión acerca de qué tipo de instrumento usar y qué ponderación otorgarle, se debería tomar en función de la realidad de cada carrera. El sistema centralizado de admisión debe ofrecer una variedad de instrumentos que permita a cada institución elegir solo uno o la combinación de ellos que sea más apropiada para su realidad.

## 4.2 Limitaciones

Los resultados de este informe deben ser actualizados y complementados con un análisis de la última PSU-M rendida, puesto que en esta última versión el DEMRE empleó IRT en la construcción de la prueba, con miras a incorporar ítems que aportaran información en un rango más amplio de habilidades, que versiones anteriores de la prueba no cubrían apropiadamente. A ello cabe añadir que hubo cambios en el tratamiento de las respuestas erradas, eliminándose la deducción de puntaje. En este trabajo se debió emplear datos de la admisión 2013, y no los de cohortes más recientes, porque no se contaba con información del rendimiento universitario de los alumnos matriculados en las distintas carreras, indispensable para estimar el aporte a la predicción de los contenidos básicos y avanzados respectivamente.

Una segunda limitación es que los puntajes de ítems básicos y avanzados no corresponden a pruebas separadas y los puntajes obtenidos por los postulantes pueden verse afectados por el hecho de que deben prepararse para rendir todos los contenidos. Por ende, es razonable suponer que la preparación

[41] Ver los Estándares para la construcción de pruebas educacionales de AERA, APA & NCME (2014), capítulo 3.

del alumno puede, en algún grado, verse influida por la extensión de los contenidos examinados y ello podría afectar el rendimiento. Este es un aspecto que debe ser examinado a través de estudios empíricos.

Una tercera limitación fue la calidad de la información contenida en la base de datos de rendimiento académico compilada por el SUA y empleada en este estudio. La base de datos requirió de un laborioso proceso de limpieza para su utilización. En la base de datos la situación académica del alumno a fines del primer año era ambigua. El porcentaje de estudiantes con información incompleta para realizar los análisis variaba de 0% a 45%, dependiendo de la carrera. Los abandonos o eliminaciones pueden (o no) representar errores de selección que merecen ser estudiados. Dada la relevancia que reviste este tipo de información para la realización de estudios, sería deseable que el DEMRE contara con un protocolo para obtener información actualizada y de buena calidad directamente de las universidades, sin intermediarios, de manera que sus equipos técnicos puedan realizar en forma periódica análisis de capacidad predictiva por ítems (desagregados por eje y nivel), tanto por carreras, áreas y para el sistema en su conjunto para ir mejorando las pruebas en el tiempo.

Otra limitación fue la falta de claridad con respecto a los criterios empleados en estudios previos, tanto del SUA, CTA y Pearson para la agrupación de carreras por áreas. Al no poder replicar la agrupación por área se limita la posibilidad de establecer comparabilidad entre los distintos estudios a través del tiempo. Por ello, en la agrupación de carreras por áreas hay un cierto grado de ambigüedad y discrecionalidad en la definición de algunas de éstas, que se debería intentar controlar en estudios futuros[42].

Para el estudio de la PSU-M admisión 2017 sería conveniente contar con una caracterización independiente de los ítems en relación al nivel de la enseñanza media en el cual son enseñados. La información relativa a este punto fue entregada por miembros del equipo del DEMRE a cargo del desarrollo de las pruebas. Idealmente debería consultarse a un grupo de profesores de enseñanza media, externos al DEMRE, para que aportaran información al respecto. Hay que considerar que los contenidos pueden estar prescritos para un nivel, pero ello no quiere decir que en la práctica sean cubiertos. En tal sentido, hay evidencia a partir de estudios de cobertura curricular llevados a cabo por el Mineduc que revelan déficits de implementación del currículo en la enseñanza media, que afecta principalmente a los establecimientos municipales y técnico profesionales. [43]

Finalmente, como se señaló anteriormente, los resultados de este estudio no bastan por sí solos para introducir los cambios específicos que se requiere incorporar en la PSU-M. Para ello es menester realizar estudios complementarios (cuantitativos y cualitativos) que permitan determinar qué conocimientos matemáticos deben ser evaluadas en las pruebas y cuáles son los contenidos específicos que deben ser incorporados de manera de maximizar la probabilidad de que los alumnos seleccionados tengan éxito en sus estudios. En el proceso de elaboración de nuevas tablas de especificaciones debe contemplarse la participación de profesores universitarios y de la enseñanza media HC y TP. Muchas universidades (e institutiones de educación superior) han avanzado en estas materias a través de pruebas diagnósticas especiales, que deberían ser estudiadas para comprender mejor cuáles competencias y contenidos son necesarios de incluir en versiones futuras de la PSU-M.

[42] Para un listado de las carreras de cada área empleado en este estudio, dirigirse a la autora principal.
[43] https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2017/06/A2N21_Curriculum_EMedia.pdf

# Referencias

AERA, APA &. NCME. (2014). *Standards for Educational and Psychological Testing.*Washington DC.

Bloom, H. , Hill, C., Black, A. & Lipsey, M. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions*. MDRC Working Papers on Research Methodology.

Camilli, G. & Shepard, L. (1994). Methods for Identifying Biased Test Items. CA: Sage publications.

Centro de Estudios MINEDUC (2013). Implementación del Currículum de Educación Media en Chile. Serie Evidencias: Años 2 , N°21.

CTA del CRUCh. (2004). *Resultados de la aplicación de pruebas de selección universitaria Admisión 2004.* Santiago.

CTA del CRUCh. (2005). *Resultados de la aplicación de pruebas de selección universitaria Admisión 2005.* Santiago.

CTA del CRUCh. (2006). *Estudio acerca de la validez predictiva de los factores de selección a las universidades del consejo de rectores.* Santiago.

CTA del CRUCh. (2008). *Resultados de la aplicación de pruebas de selección universitaria Admisión 2006-2008.* Santiago.

CTA del CRUCh. (2010). *Resultados de la aplicación de pruebas de selección universitaria Admisión 2006-2010.* Santiago.

CTA del CRUCh. (2010). *Validez diferencial y sesgo de predictividad de las pruebas de admisión a las universidades chilenas.* Santiago.

CTA del CRUCh .(2008). *Estudio acerca de la validez predictiva de los factores de selección a las universidades del consejo de rectores Admisión 2003-2006.*Santiago.

Camara, W. (2009). College Admission Testing: Myths and Realities in an Age of Admissions Hype. En R. P. (editor), *Correcting Fallacies About Educational and Psychological Testing* (págs. 147-180). Washington D.C: American Psychological Association.

Caruman, S. (2004). *Informe Técnico Etapa de Aplicación de Pruebas PSU 2004 para el Proceso de Admisión 2005.*DEMRE.

Comisión Nuevo Curriculum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior. (2000). *Informe sometido en consulta previa a la Ministra de Educación* .Santiago, Chile. Unpublished document.

Crocker, L.&. Algina, J. (1986). *Introduction to classical and Modern Test Theory.*Orlando, Florida: HBJ.

Crooks, T.; Kane, M. & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education: Principle Policy and Practice, 3*, 265-286.

Dalgalarrando, M. G. (24 de 12 de 2008). Encargados PSU niegan que la brecha entre colegios públicos y privados haya aumentado. *El Mercurio*.

De Ayala, R. J. (2009). *The theory and practice of item response theory.*NY: Guilford Press.

Educational Testing Service (ETS). (2005). *Evaluación Externa de las Pruebas de Selección Universitaria (PSU).*

El SIES Comenzará a ser aplicado a contar del 2003. (21 de 07 de 2002). *El Diario Austral*, pág. A12.

Geiser, S. & Studley, R. (2001). *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California.*California: University of California.

Grau, M. (2016). *Estudio acerca de la validez predictiva del ranking de notas.* Santiago, Chile: SUA, Consejo de Rectores de Universidades Chilenas .

Haladyna, T. (2004). *Developing and Validating Multiple Choice Items.* Arizona: Routledge.

Hambleton, S.; Swaminathan, H. & Rogers, J. (1991). *Fundamentals of Item Response Theory.* California: Sage.

Harrell, F. (2001). *Regression Modeling Strategies.* New York : Springer.

Henrysson, S. (s.f.). Gathering, analyzing , and using data on test ítems. En R. L. (editor), *Educational Measurement, segunda edición.* Washington D.C: American Council on Education.

Kobrin, J. ;YoungKoung, K. & Sackett, P. (2012). Modeling the predictive validity of SAT Mathematics ítems using ítem characteristics. *Educational and Psychological Measurement, (72)*, 99-119.

Koljatic, M. & Silva, M. (2006). Equity issues associated with the college admission tests in Chile. *Equal Opportunities International, Vol. 25 Iss: 7*, 544 - 561.

Koretz, D. R. (2000). The Impact of Score Differences on the Admission of Minority Students: An Illustration. *NBETTP Statements 1 (5), 1-16.*

Koretz, D. R. (2002). Testing and diversity in post-secondary education: The case of California. *Education Policy Analysis Archives 10(1)*.

Lord, F. (1952). The relation of the reliability of multiple choice tests to the distribution of item difficulties. *Psychometrika,17*, 181-192.

Manzi, J. & Bravo, D. (19 de 05 de 2002). SIES: Los Peligros de lo que No Es. *Artes y Letras, El Mercurio*, págs. E6-E7.

Nunnally, J. (1972). *Psychometric Theory.* NY: McGraw Hill.

Pearson Education (2013). *Final Report Evaluation of the Chile PSU.*

# ANEXOS

## ANEXO A. Otros Antecedentes

### Brechas Históricas de Rendimiento: Aumento de los Puntajes PSU de Egresados de Colegios Particulares Pagados

Los miembros del CTA del CRUCh se percataron del crecimiento de la brecha entre el grupo de alumnos de establecimientos particulares pagados y el contingente de alumnos provenientes de la educación municipal, sin embargo, desestimaron el fenómeno atribuyéndolo al aumento del contingente de estudiantes de condiciones socioeconómicas más vulnerables que rendían la PSU a partir del año 2007, con posterioridad a la instauración de las becas PSU. Según la interpretación de los miembros del CTA del CRUCH:

a) El crecimiento de la brecha se debía al aumento de alumnos más vulnerables que rendían las pruebas y obtenían un desempeño más bajo en la PSU,
b) Si se controlaba estadísticamente por nivel socioeconómico la brecha [entre tipos de colegio] no crecía[44].

La explicación del CTA acerca del crecimiento de las brechas no se sostiene por las siguientes razones:

1) Tal como señala el informe de Pearson, la brecha crece porque aumenta el promedio que obtienen los estudiantes de colegios particulares pagados y **no** porque el promedio de los estudiantes vulnerables decaiga. Por tanto, la explicación del CTA del CRUCh de que el crecimiento de la brecha se explica por una baja en el rendimiento de los alumnos más vulnerables es, a todas luces, errónea.[45].

La tabla siguiente muestra la evolución de puntajes en la PSU-M para los egresados de establecimientos educacionales categorizados en términos de dependencia y modalidad entre el 2004 y el 2017, y se puede observar el crecimiento de los puntajes de los alumnos de colegios particulares pagados en el tiempo.

[44]Ver por ejemplo (Dalgalarrando). "Encargados PSU niegan que la brecha entre colegios públicos y privados haya aumentado". El Mercurio, 24/12/2008.

Medias PSU-M: Admisión 2004 al 2017 (Estudiantes Promoción)*

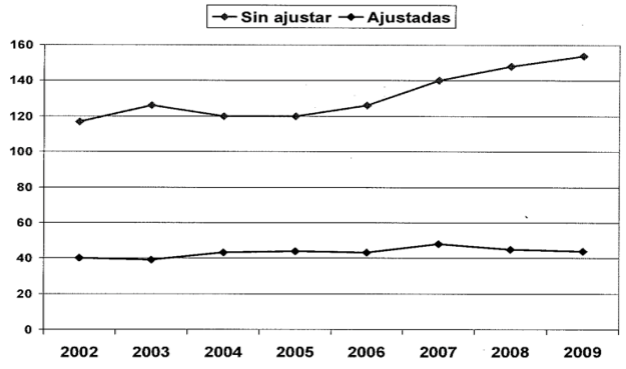
| | | 2004* | 2005* | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Part.Pagado** | Media | 579 | 586 | 591 | 602 | 608 | 612 | 616 | 618 | 614 | 612 | 611 | 610 | 608 | 611 |
| | DS | 113 | 110 | 112 | 109 | 105 | 106 | 111 | 114 | 105 | 106 | 108 | 110 | 111 | 109 |
| | (n) | 18011 | 18433 | 18354 | 18736 | 18901 | 18905 | 18540 | 18547 | 18461 | 18868 | 18992 | 19257 | 19213 | 19027 |
| **Part. Subvencionado HC** | Media | 501 | 503 | 501 | 509 | 508 | 507 | 505 | 506 | 506 | 505 | 505 | 507 | 506 | 508 |
| | DS | 101 | 101 | 103 | 105 | 104 | 108 | 110 | 113 | 110 | 111 | 111 | 110 | 111 | 103 |
| | (n) | 33566 | 40203 | 43717 | 49653 | 54382 | 60361 | 63014 | 64821 | 65923 | 68319 | 68470 | 71430 | 74117 | 73137 |
| **Part. Subvencionado TP** | Media | 429 | 430 | 430 | 428 | 427 | 429 | 429 | 437 | 429 | 436 | 431 | 436 | 442 | 446 |
| | DS | 85 | 88 | 90 | 100 | 96 | 96 | 97 | 94 | 96 | 99 | 100 | 98 | 97 | 83 |
| | (n) | 10560 | 11589 | 12403 | 18852 | 20217 | 23080 | 24385 | 24880 | 23903 | 23941 | 23753 | 24382 | 24685 | 23484 |
| **Municipal HC** | Media | 466 | 475 | 473 | 472 | 471 | 473 | 470 | 473 | 469 | 466 | 473 | 475 | 475 | 480 |
| | DS | 109 | 109 | 110 | 115 | 115 | 116 | 118 | 120 | 118 | 120 | 123 | 120 | 120 | 107 |
| | (n) | 32251 | 33913 | 39379 | 39379 | 40058 | 42989 | 43949 | 42068 | 34576 | 34546 | 35452 | 36880 | 36880 | 36512 |
| **Municipal TP** | Media | 434 | 435 | 429 | 425 | 422 | 422 | 423 | 427 | 429 | 425 | 418 | 425 | 425 | 431 |
| | DS | 84 | 86 | 89 | 97 | 95 | 94 | 97 | 96 | 95 | 98 | 99 | 98 | 99 | 81 |
| | (n) | 12956 | 14072 | 23188 | 23188 | 22809 | 27571 | 29733 | 29414 | 24195 | 25686 | 24674 | 26795 | 26795 | 27781 |
| **Total** | Media | 489 | 492 | 490 | 488 | 488 | 486 | 484 | 487 | 486 | 487 | 487 | 489 | 490 | 493 |
| | DS | 113 | 113 | 115 | 119 | 119 | 119 | 122 | 122 | 121 | 122 | 123 | 121 | 120 | 110 |
| | (N) | 107345 | 118745 | 137041 | 149808 | 156367 | 172906 | 179621 | 179730 | 167058 | 171360 | 171341 | 178744 | 181690 | 180979 |

Fuente de Datos: DEMRE.
*Puntajes PSU reescalados para efectos de comparación interanual.

2) El procedimiento estadístico elegido por el CTA del CRUCh para re-estimar la brecha consistió en un análisis de covarianza controlando por nivel socioeconómico. Una vez ajustados los puntajes, procedieron a estimar las brechas por tipo de colegio, procedimiento discutible[46]. No obstante, el aspecto más censurable de su análisis fue que hayan omitido entregar información desagregada para los alumnos de la EMTP y los EMHC al reportar sus brechas ajustadas.

En la figura A1 el CTA del CRUCH muestra la brecha sin ajustar (curva superior) y ajustada (curva inferior). En apariencia la brecha ajustada no crece en el tiempo.

Figura A1. elaborada por el CTA del CRUCh[47]




Fuente del Gráfico: CTA del CRUCh

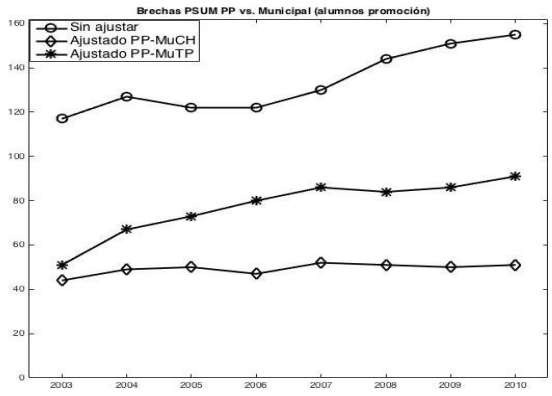
3) Si el CTA hubiera reportado en forma separada las brechas correspondientes a los estudiantes de la EMHC y a los de la EMTP habría quedado de manifiesto el perjuicio para

[46]No advierten que no se cumplen las condiciones para el análisis, ya que no hay aleatorización y el nivel socioeconómico del hogar y el tipo de colegio al cual asiste el estudiante están altamente correlacionados entre sí por lo cual no se puede descartar un efecto del tipo de colegio simplemente controlando por nivel socioeconómico.

[47]Presentación al CRUCh, Enero de 2009.

estos últimos, como se aprecia en la Figura A2. Esta figura revela lo que el gráfico del CTA del CRUCh enmascara: la brecha de los TP crece en el tiempo. Por tanto la conclusión del CTA del CRUCh acerca de que las brechas se mantienen "estables en el tiempo" luego de controlar por factores socioeconómicos **no** se cumple cuando se trata de estudiantes municipales provenientes de la EMTP.

Figura A2: Brechas separadas para estudiantes HC y TP[48]




Fuente del Gráfico: Elaboración propia

Es urgente destinar esfuerzos para ajustar la PSU a los estándares internacionales que consagran la equidad en la evaluación como un tema fundamental en el marco del desarrollo y uso de pruebas de altas consecuencias[49].

[48] Fuente: datos DEMRE 2016, elaboración propia.

[49] Ver al respecto el capítulo 3 de "Standards for Educational and Psychological Testing". (AERA. APA & NCME, 2014): Washington DC.
En materia de equidad, otro aspecto que el CTA del CRUCh omitió estudiar fue la caída en la matrícula universitaria de alumnos provenientes de la educación media municipal que se verificó con posterioridad a la implementación de la PSU en algunas de las universidades más prestigiosas del sistema..

## ANEXO B. Confiabilidad[50]

$$r_{kk} = \frac{kr_{11}}{1 + (k-1)r_{11}}$$

Donde:

- $r_{kk} = Confiabilidad\ de\ la\ prueba\ k\ veces\ más\ larga\ que\ la\ prueba\ original$
- $r_{11} = Confiabilidad\ de\ la\ prueba\ original$
- $k \quad = Factor\ de\ cambio\ en\ la\ longitud\ \ de\ la\ prueba.$

Se puede estimar, a partir de esta fórmula, la confiabiliad de la prueba con 40, 50 y 60 ítems de contenidos básicos.

Estimación de la confiabilidad de la prueba

| Número de Ítems Básicos | $k$ | Confiabilidad (Alpha de Cronbach) |
|---|---|---|
| **40** | 1.5 | .93 |
| **50** | 1.6 | .94 |
| **60** | 2 | .96 |

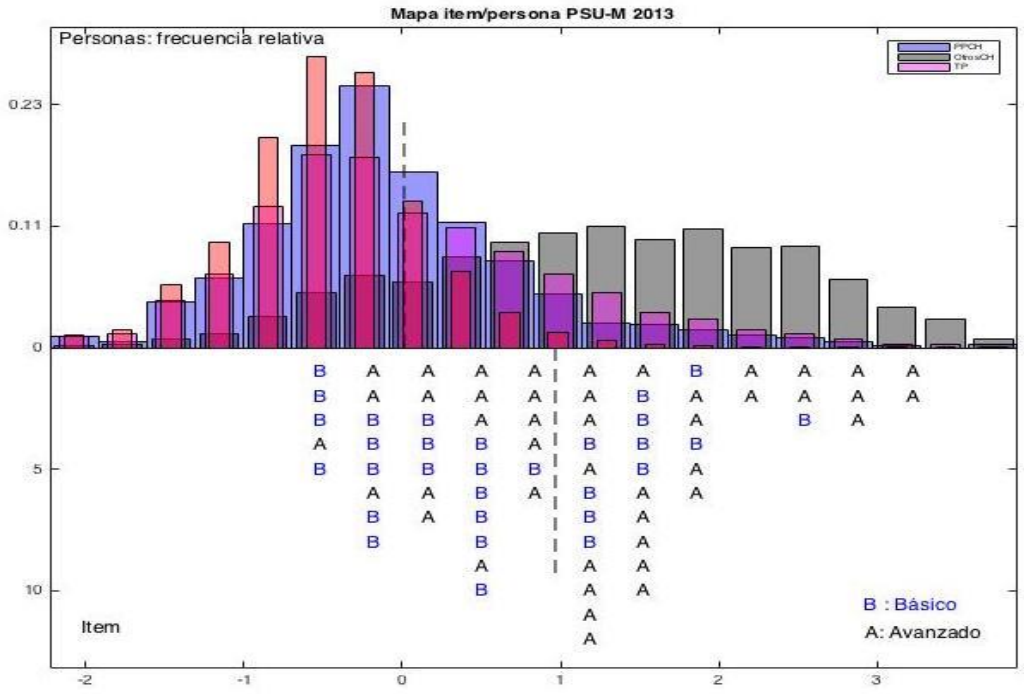
Fuente de Datos: Elaboración propia, DEMRE 2016.

[50]Spearman-Brown, citado en (Crocker, 1986) Introduction to classical and Modern Test Theory. HBJ (Orlando, Florida).

## ANEXO C. IRT

### Mapa ítem/persona (de Wright)

A continuación se presenta un mapa de Wright o ítem/persona que permite visualizar los atributos de los ítems (dificultad) y los individuos en la misma escala (habilidades)[51]. La consistencia entre los resultados obtenidos a través de los métodos de la TCM e IRT son evidentes al analizar el gráfico a continuación. Los alumnos de mejor rendimiento, que presentan un nivel de habilidad mayor o igual a 2 (y que caerían dentro de la categoría de "mayor habilidad") provienen en mayor proporción de colegios particulares pagados. Para este grupo, una mayoría de los ítems de esta prueba, incluidos los de nivel avanzado no tienen un grado de dificultad suficiente para ellos. En cambio, para el grueso de los alumnos de la EMTP, la gran mayoría de los ítem superan sus capacidades.




Fuente de Datos: Elaboración propia, DEMRE 2016.

### Ajuste de Modelos

El ajuste relativo de los modelos se estimó sobre la base de una heurística que emplea la razón de verosimilitud ($G^2$), donde el cambio ($R^2_\Delta$ ) en $G^2$ entre los modelos, se usa para determinar si hay un

[51] En este mapa se combina el histograma de las habilidades de los examinados y el histograma del grado de dificultad de los items de la prueba. Si se quiere comparar los examinados según sus procedencia escolar se prefiere usar frecuencias relativas en vez de frecuencias absolutas, por el el dispar número de examinados de los grupos.

mejor ajuste al incorporar más parámetros y la mejora relativa en la proporción de varianza explicada al usar uno u otro modelo (ver De Ayala, 2009)[52].

(1) $$R^2_{\Delta} = \frac{(G^2_1 - G^2_2)}{G^2_1}$$

Donde:

$G^2_1 = -2\ln(L_1);\ L_1$: *máxima verosimilitud para un modelo*

$G^2_2 = -2\ln(L_2);\ L_2$: *máxima verosimilitud para el modelo con más parámetros*.

Existen otros criterios, entre ellos el Criterio de Información Akaike (AIC) y el Criterio de Información Bayesiano (BIC):

$$AIC = -2\ln(L) + 2npar$$

$$BIC = -2\ln(L) + \ln(N) * npar$$

donde $npar$ es el número de parámetros y *N* es el número de alumnos que rinden las pruebas.

Se ajustaron modelos de uno, dos y tres parámetros para evaluar cuál de ellos mostraba un mejor ajuste a los datos. Se reportan las estadísticas en la siguiente tabla para todos los ítems, los ítems básicos y los ítems avanzados respectivamente.

Estadísticas modelos IRT.

| *Ítems* | Modelo | -2lnL | Número de parámetros | $R^2_{\Delta}$ (%) | Grados de libertad | AIC | BIC |
|---|---|---|---|---|---|---|---|
| **Todos los ítems** | 1PL | -5,236,612 | 74 | | | 10,473,375 | 10,474,128 |
| | 2PL | -5,127,328 | 148 | 2.1 | 73 | 10,254,952 | 10,256,439 |
| | 3PL | -5,065,364 | 222 | 1.2 | 74 | 10,131,172 | 10,133,402 |
| **Ítems básicos (1° y 2° Medio)** | 1PL | -2,483,638 | 31 | | | 4,967,337 | 4,967,649 |
| | 2PL | -2,438,382 | 60 | 1.8 | 29 | 4,876,885 | 4,877,488 |
| | 3PL | -2,413,998 | 90 | 1.0 | 30 | 4,828,177 | 4,829,081 |
| **Ítems avanzados (3° y 4° Medio)** | 1PL | -2,898,547 | 45 | | | 5,797,183 | 5,797,635 |
| | 2PL | -2,843,046 | 88 | 1.9 | 43 | 5,686,268 | 5,687,152 |
| | 3PL | -2,800,557 | 132 | 1.5 | 44 | 5,601,377 | 5,602,703 |

Fuente de Datos: Elaboración propia, DEMRE 2016.

El cambio observado en el $R^2$ entre el modelo Rasch de uno a 2PL, se tradujo en aproximadamente un 2% de mejora en el ajuste (1.8% en los ítems básicos, 1.9% en los avanzados y un 2.1% para todos los ítems respectivamente). Al comparar los modelos de 2PL y 3PL hay una mejoría en el ajuste que varía entre un 1.0% y 1.5% (1.0% ítems básicos; 1.5% ítems avanzados y 1.2% todos los ítems).

Consistente con lo anterior, los valores de AIC y BIC muestran que el modelo de tres parámetros es el que presenta el mejor ajuste.

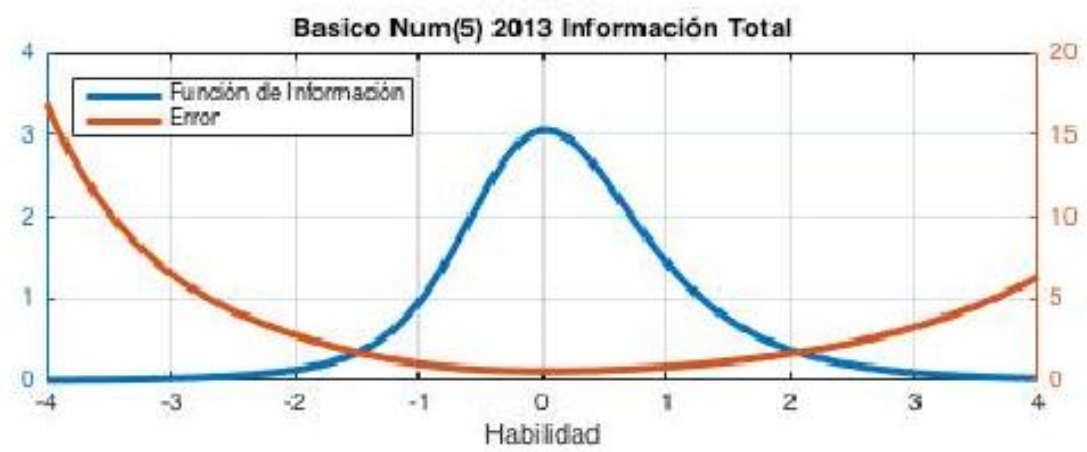
Por lo anterior, para fines del estudio se optó por el modelo de tres parámetros[53].

[52] (De Ayala, 2009). The theory and practice of item response theory. Guilford. NY.

[53] Los expertos de Pearson Education emplearon también un modelo de tres parámetros.

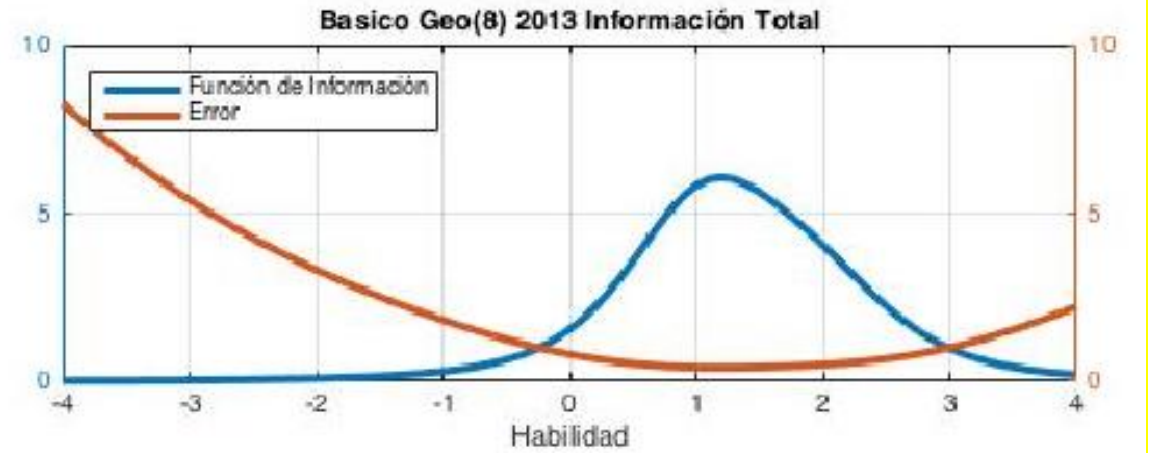
**Curvas de Información por Eje y Nivel: Modelo 3 PL**

Función de información de la prueba. Admisión 2013, Contenidos Básicos - Números.




Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información de la prueba. Admisión 2013, Contenidos Básicos - Geometría.

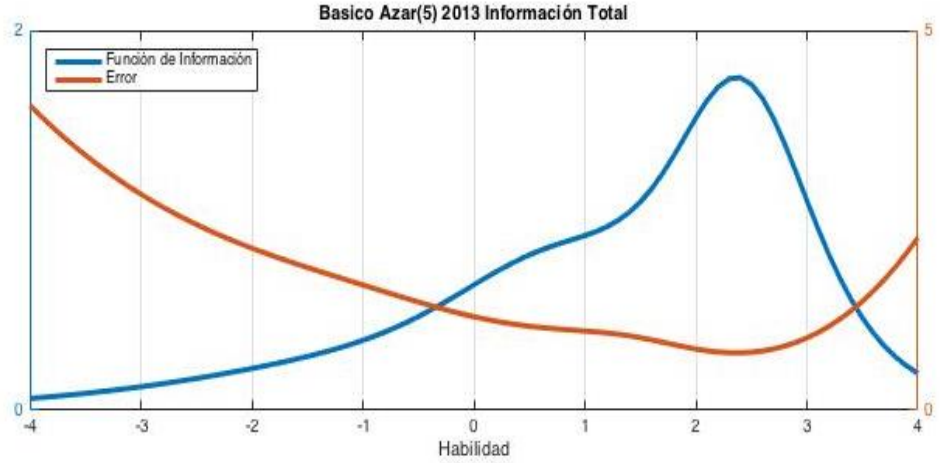



Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información(Rinden 2012). Contenidos Básicos: Álgebra

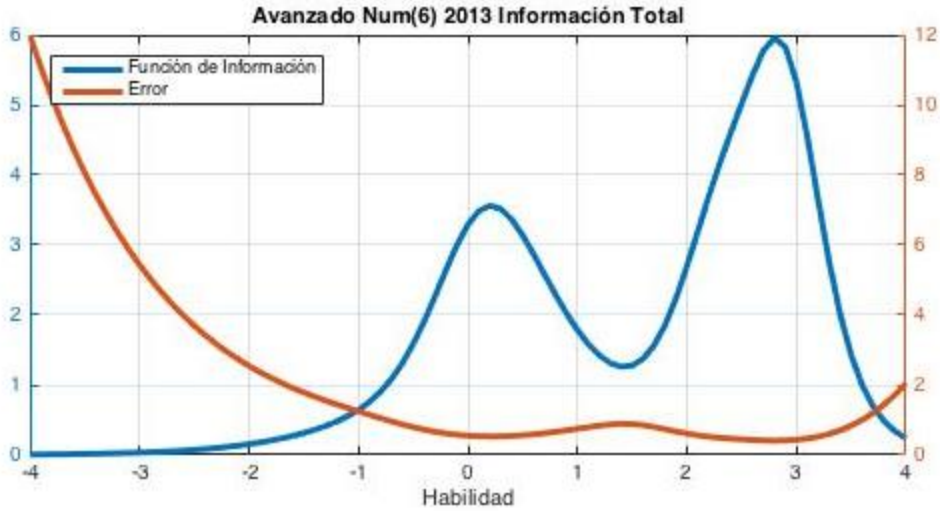

Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información. Rinden 2012, Contenidos Básicos: datos y azar.




Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información. Rinden 2012, Contenidos Avanzados: Números.




Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información de la prueba. Rinden 2012, Contenidos Avanzados : Geometría.


Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información de la prueba. Rinden 2012, Contenidos Avanzados: Álgebra.


Fuente de Datos: Elaboración propia, DEMRE 2016.

Función de información de la prueba. Rinden 2012, Contenidos Avanzados: Datos y azar.


Fuente de Datos: Elaboración propia, DEMRE 2016.

## ANEXO D. Correlación por Area

Cuando se calcula la correlación en la PSU-M y el rendimiento universitario de los alumnos, se tiene una estimación sobre la base de una muestra selecta, o sea restrigida a los alumnos que entraron a la universidad. Mediante la corrección, se busca estimar cuál sería la correlación de no mediar el proceso de selección[54]:

$$\hat{\rho} = \frac{\sigma_x r_{xy}}{\sqrt{\sigma_x^2 r_{xy}^2 + s_x^2(1 - r_{xy}^2)}}$$

Donde:

$\hat{\rho}$: Correlación estimada entre la variable x e y, donde x representa el puntaje de la PSU e y el rendimiento en primer año de la universidad.

$r_{xy}$: Correlación restringida entre x e y.

$\sigma_x$ Desviación estándar de la variable x en la población **no** restringida.

$s_x$ Desviación estándar de la variable x en la población restringida.

Las correlaciones por área se calcularon como promedio de las correlaciones por carrera, ponderando por su número de alumnos (solo para carrera con treinta o más alumnos matriculados y con información completa por alumno).

[54] Ver al respecto Sackett, P. & Yang, H. (2000). Correction for range restriction: An expanded typology. Journal of Applied Psychology, 85(1), 112-118.

**Correlaciones simples del rendimiento universitario en primer año y subpuntajes de la PSU-M por área de carrera**

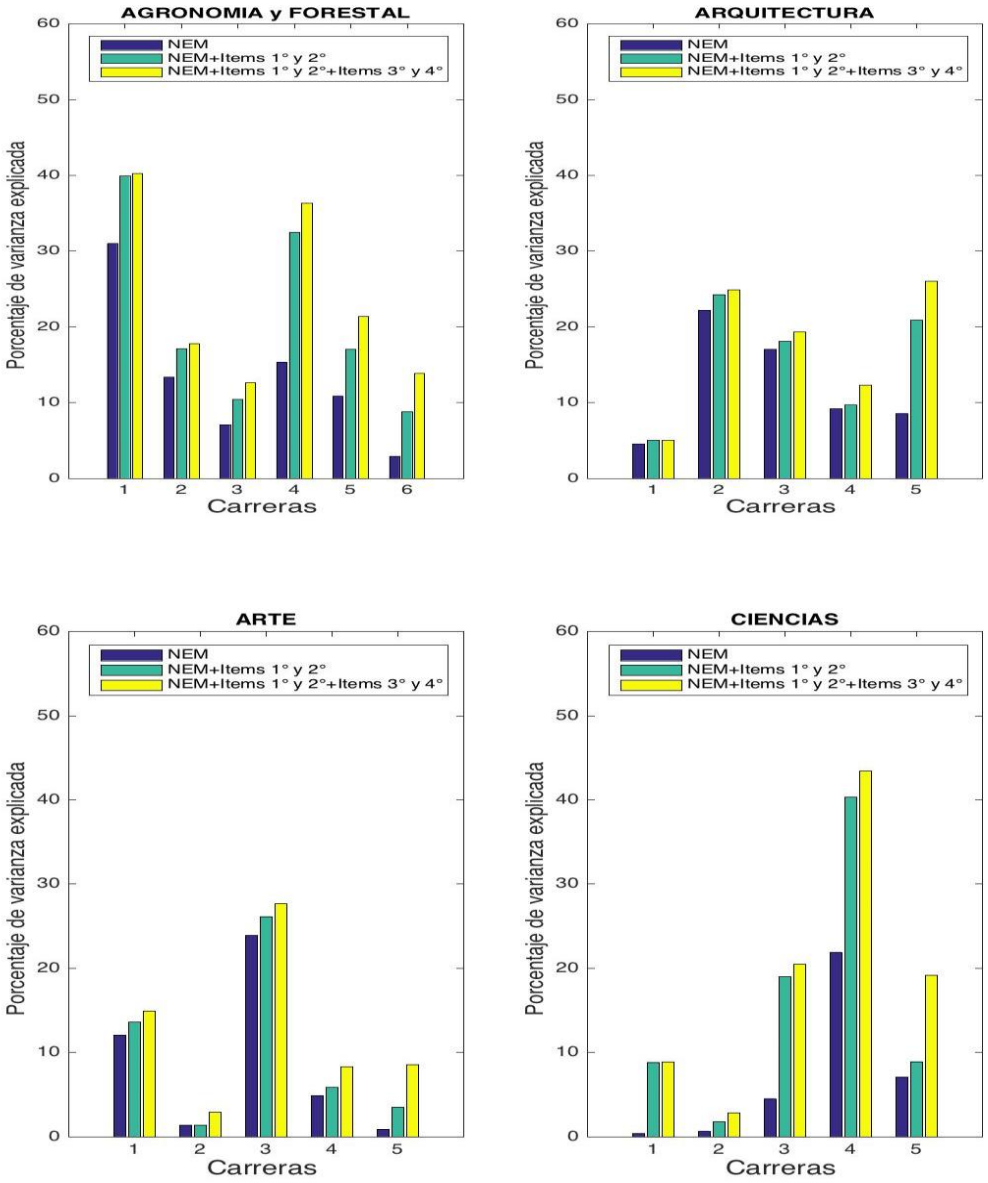
| AREA | N | NEM | Números y Datos y azar (Contenidos Básicos) | Contenidos Básicos | Contenidos Avanzados | Todos (Basicos y Avanzados) |
|---|---|---|---|---|---|---|
| AGRONOMIA y FORESTAL | 683 | 0,4 | 0,32 | 0,39 | 0,38 | 0,39 |
| ARQUITECTURA | 1105 | 0,27 | 0,14 | 0,15 | 0,19 | 0,17 |
| ARTE | 586 | 0,4 | 0,19 | 0,2 | 0,23 | 0,22 |
| CIENCIAS | 1990 | 0,41 | 0,34 | 0,31 | 0,28 | 0,29 |
| HUMANIDADES | 265 | 0,4 | 0,29 | 0,42 | 0,45 | 0,45 |
| LETRAS | 199 | 0,47 | 0,35 | 0,3 | 0,31 | 0,31 |
| CIENCIAS/QUIMICA y FARMACIA | 648 | 0,47 | 0,44 | 0,51 | 0,5 | 0,51 |
| CONSTRUCCION | 827 | 0,28 | 0,2 | 0,3 | 0,29 | 0,30 |
| CIENCIAS SOCIALES | 2739 | 0,38 | 0,16 | 0,2 | 0,22 | 0,21 |
| DERECHO | 2798 | 0,39 | 0,28 | 0,3 | 0,31 | 0,31 |
| DISENO Y PUBLICIDAD | 879 | 0,33 | 0,19 | 0,23 | 0,22 | 0,23 |
| EDUCACION PARVULARIA | 329 | 0,36 | 0,19 | 0,26 | 0,28 | 0,28 |
| EDUCACION BASICA | 535 | 0,4 | 0,26 | 0,23 | 0,23 | 0,23 |
| EDUCACION DIFERENCIAL | 518 | 0,42 | 0,13 | 0,17 | 0,14 | 0,17 |
| EDUCACION OTRAS | 809 | 0,33 | 0,2 | 0,25 | 0,25 | 0,26 |
| EDUCACION MEDIA CIENCIAS | 551 | 0,32 | 0,2 | 0,23 | 0,19 | 0,21 |
| EDUCACION MEDIA HUMANISTA | 1280 | 0,41 | 0,18 | 0,3 | 0,33 | 0,33 |
| IDIOMAS | 156 | 0,19 | 0,06 | 0,19 | 0,23 | 0,22 |
| ENFERMERIA Y OTROS | 5227 | 0,38 | 0,32 | 0,36 | 0,35 | 0,36 |
| INGENIERIA CIVIL | 10296 | 0,38 | 0,38 | 0,38 | 0,36 | 0,37 |
| INGENIERIA COMERCIAL | 3579 | 0,35 | 0,28 | 0,27 | 0,24 | 0,25 |
| INGENIERIA OTROS | 2328 | 0,38 | 0,22 | 0,27 | 0,29 | 0,29 |
| MEDICINA | 912 | 0,59 | 0,41 | 0,57 | 0,43 | 0,53 |
| ODONTOLOGIA | 1005 | 0,32 | 0,27 | 0,36 | 0,33 | 0,35 |
| PERIODISMO | 648 | 0,41 | 0,21 | 0,23 | 0,20 | 0,21 |
| TECNO ADMINISTRACION | 1321 | 0,35 | 0,23 | 0,27 | 0,26 | 0,27 |
| TECNO OTRO | 1719 | 0,32 | 0,24 | 0,29 | 0,32 | 0.30 |
| INGENIERIA EN EJECUCION | 878 | 0,26 | 0,27 | 0,31 | 0,29 | 0,30 |
| VETERINARIA | 485 | 0,36 | 0,32 | 0,36 | 0,38 | 0,38 |

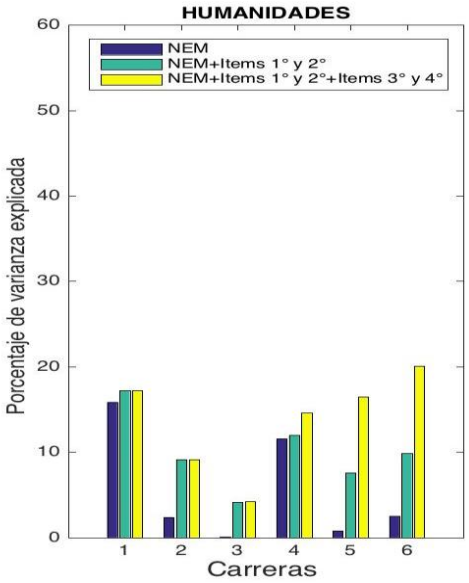
Fuente de Datos: Elaboración propia, DEMRE 2016.

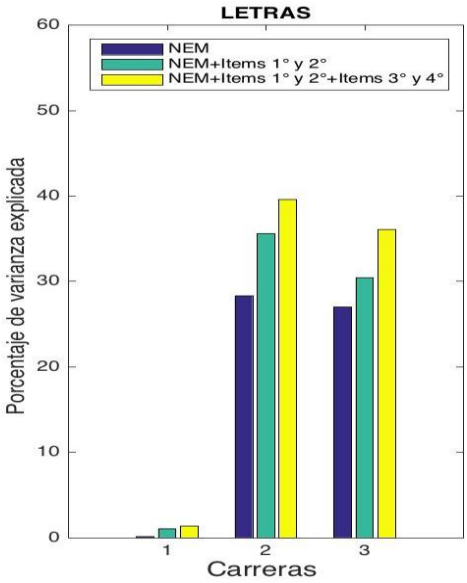
## ANEXO E: Gráficos de Incremento de Varianza Explicada

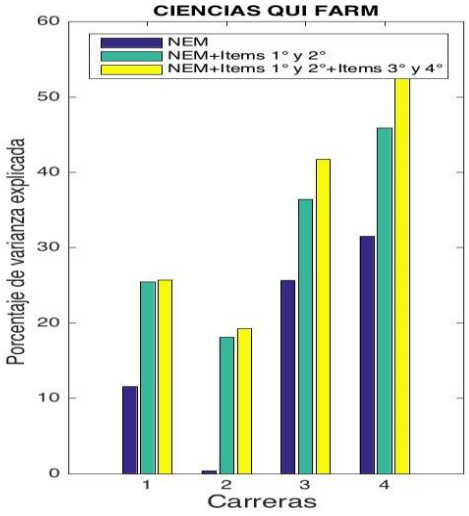
**Los gráficos siguientes muestran el incremento en el porcentaje de varianza explicada entre los Modelos 1, 2 y 3 para una selección de carreras por área.**

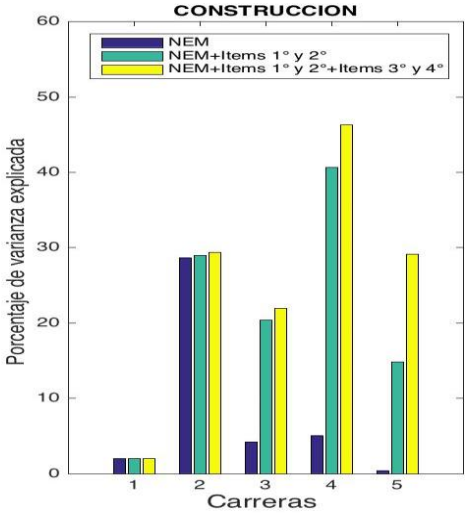
| |
|---|
| Modelo 1: Solo NEM (azul)<br>Modelo 2: NEM + Contenidos Básicos (verde)<br>Modelo 3: NEM + Contenidos Básicos + Contenidos Avanzados (amarillo) |

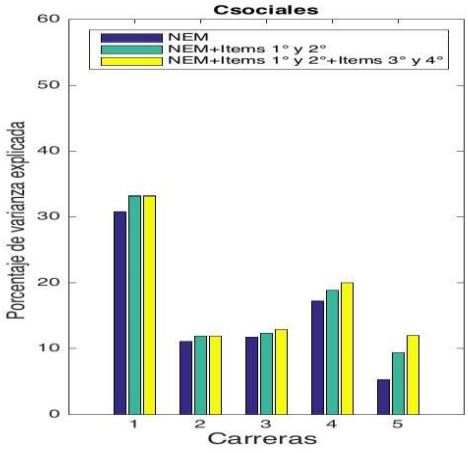


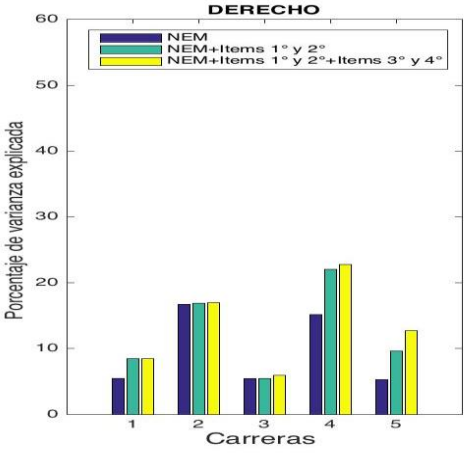

HUMANIDADES
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

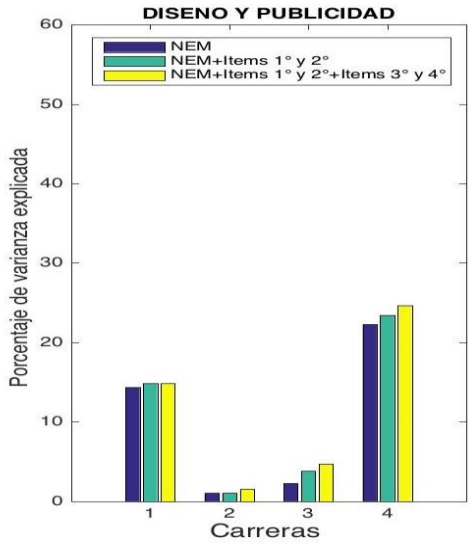

LETRAS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

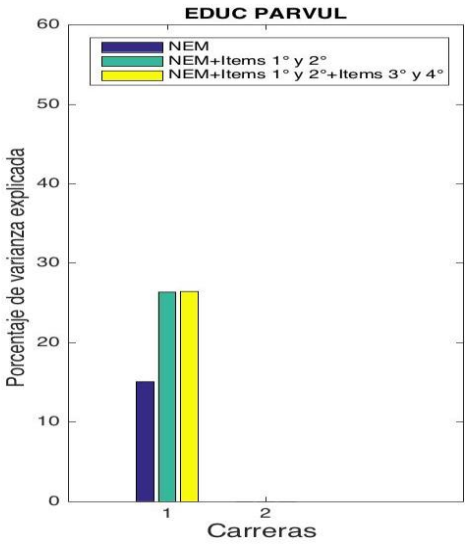

CIENCIAS QUI FARM
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

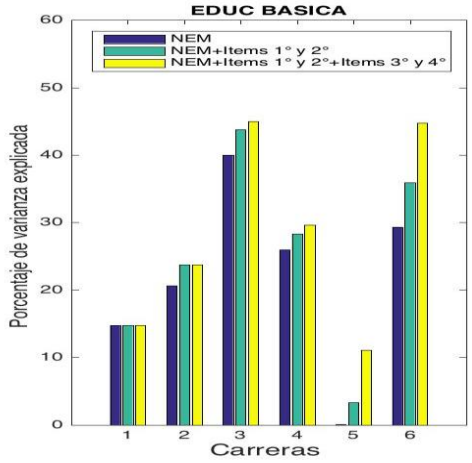

CONSTRUCCION
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

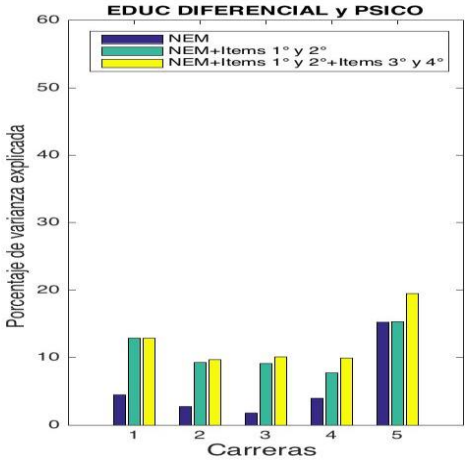

Csociales
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

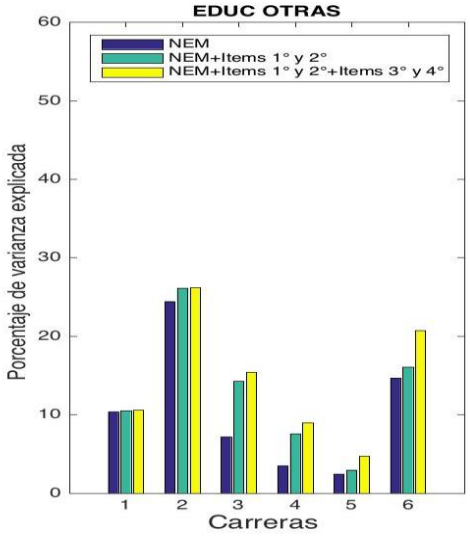

DERECHO
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

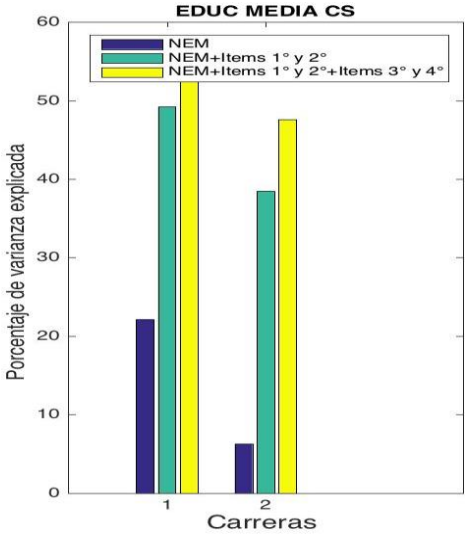

DISENO Y PUBLICIDAD
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

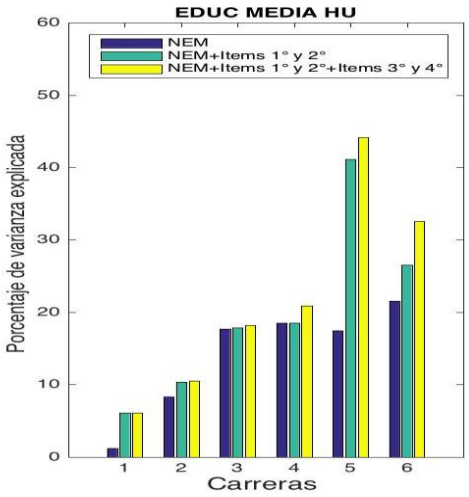

EDUC PARVUL
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

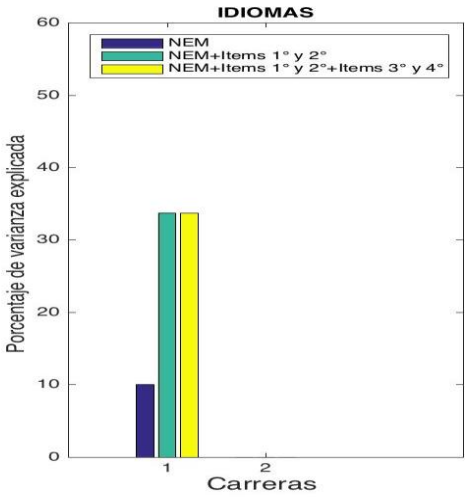

EDUC BASICA
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

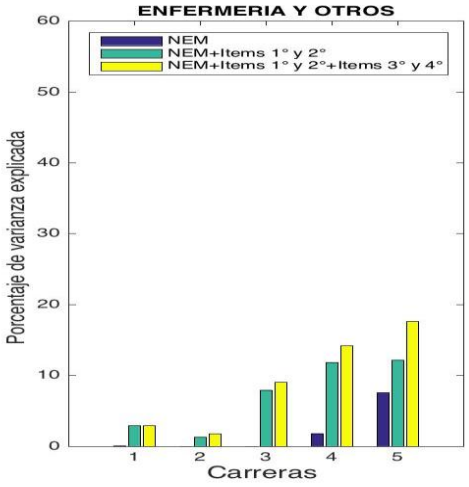

EDUC DIFERENCIAL y PSICO
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

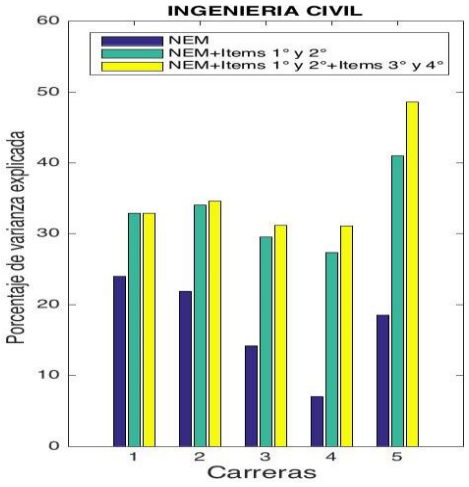

EDUC OTRAS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

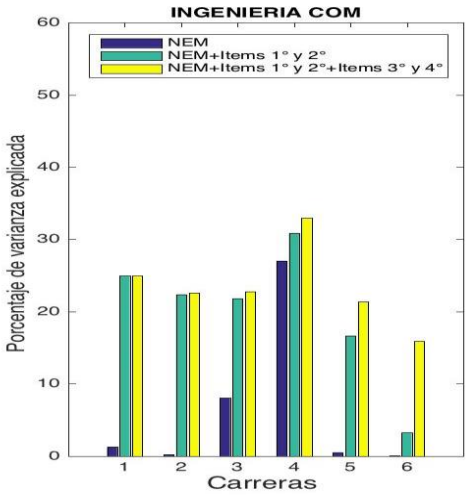

EDUC MEDIA CS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

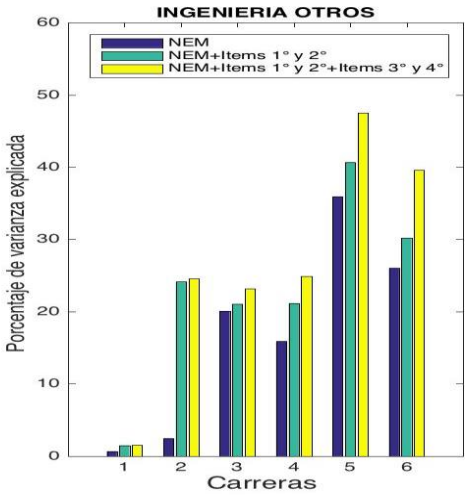

EDUC MEDIA HU
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

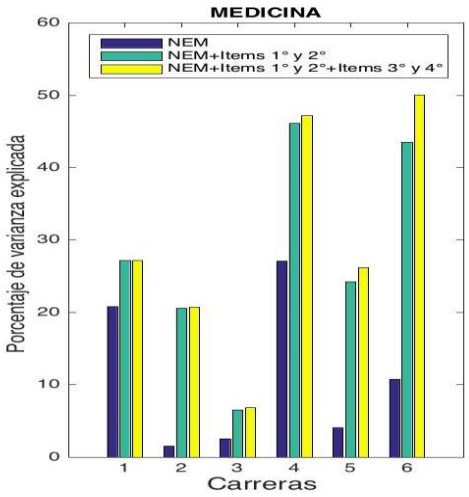

IDIOMAS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

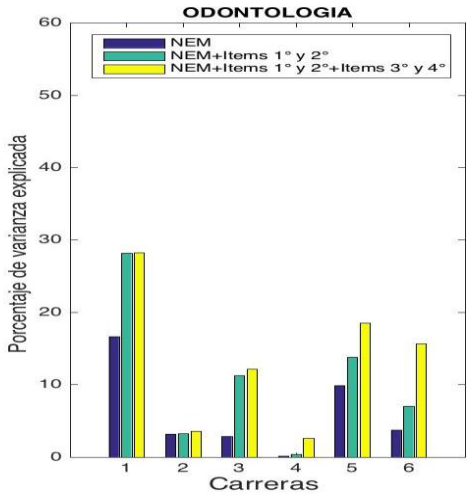

ENFERMERIA Y OTROS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

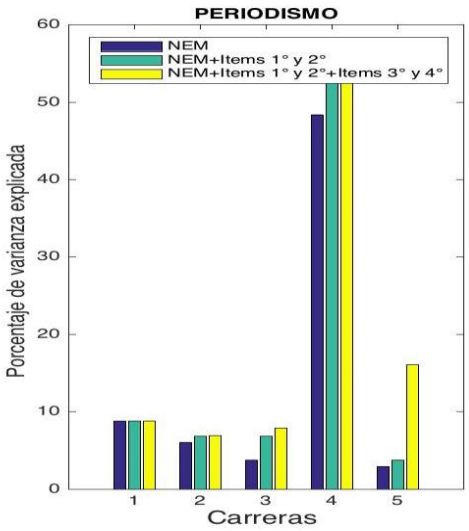

INGENIERIA CIVIL
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

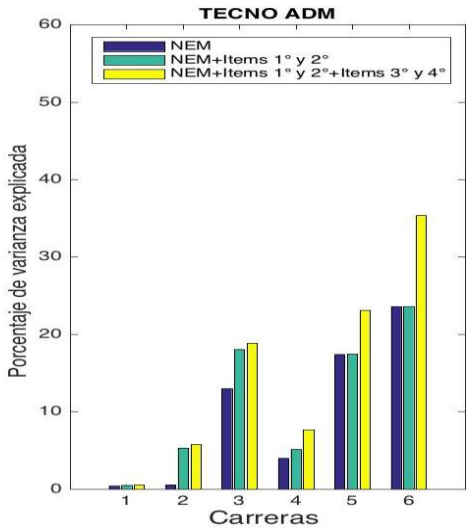

INGENIERIA COM
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

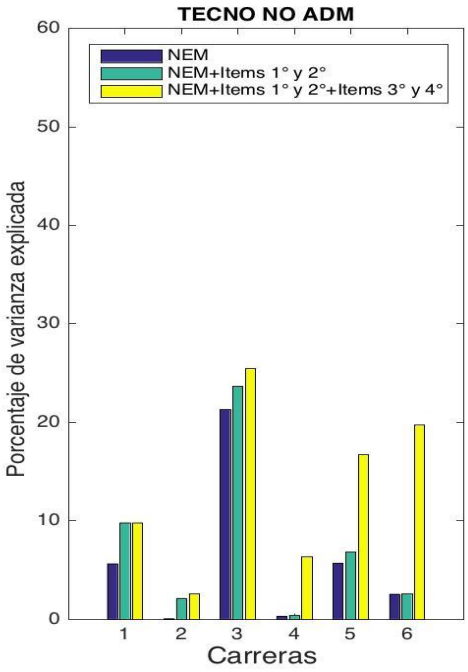

INGENIERIA OTROS
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

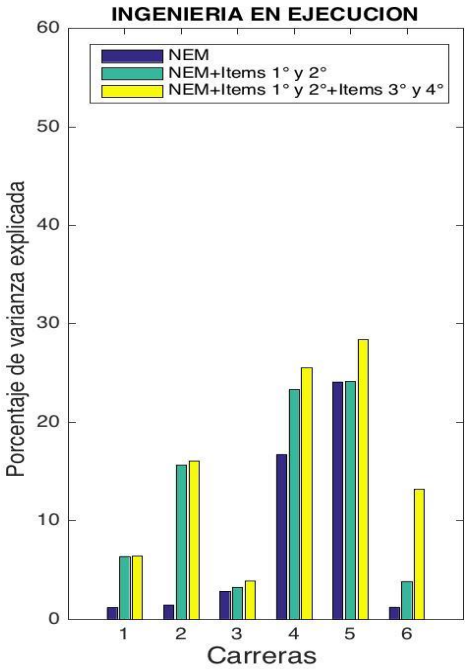

MEDICINA
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras

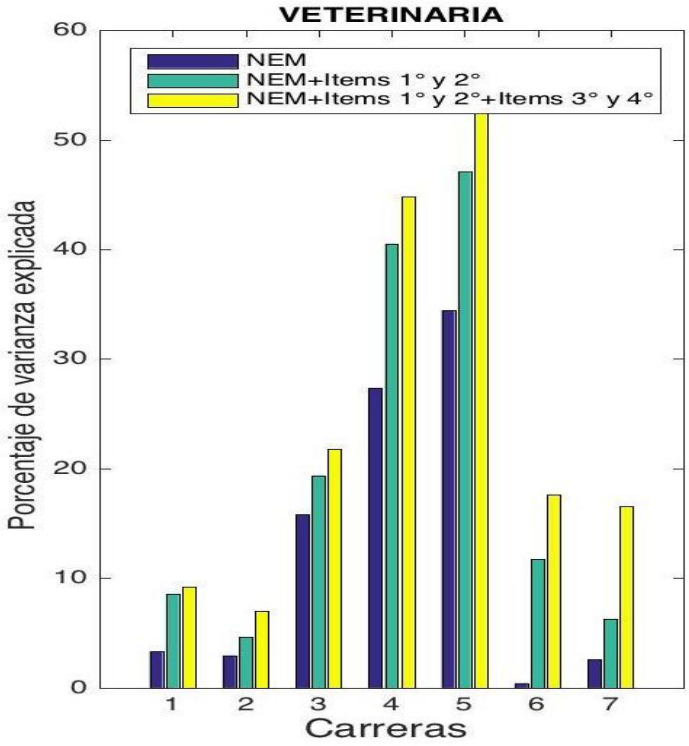

ODONTOLOGIA
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras


PERIODISMO
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras


TECNO ADM
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras


TECNO NO ADM
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras
0
10
20
30
40
50
60
1
2
3
4
5
6


INGENIERIA EN EJECUCION
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras
0
10
20
30
40
50
60
1
2
3
4
5
6


VETERINARIA
NEM
NEM+Items 1° y 2°
NEM+Items 1° y 2°+Items 3° y 4°
Porcentaje de varianza explicada
Carreras
0
10
20
30
40
50
60
1
2
3
4
5
6
7

## ANEXO F. Estimación de Brechas

En estudios de brechas de rendimiento se usan distintas fórmulas de estimación para la desviación estándar, dependiendo de la naturaleza de las comparaciones que se pretenden establecer. No hay un claro consenso en la literatura acerca de cuál desviación estándar usar para fines de calcular brechas.[55] Se puede usar la desviación estándar poblacional, en la medida que ésta represente bien la desviación de los grupos, o bien combinar las desviaciones estándar de los grupos de interés.

En este estudio se optó por aplicar la desviación estándar combinada ($s_p$) de los grupos comparados:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

**Funcionamiento diferencial de ítems (DIF)**

En el informe de Pearson se reportó que el DEMRE realizaba análisis diferencial del funcionamiento de los ítems (DIF), pero de manera restringinda. En dicho informe se sugiere que se expanda a más criterios, prestando especial atención al grupo técnico profesional.

El DIF tiene por objeto detectar los ítems que podrían presentar sesgos que afecten a algunos grupos de la población, pudiendo dar ventaja injustificada a algunos por sobre otros. Ahora bien, la existencia de DIF en un ítem no necesariamente implica que éste presente sesgo, pero es una señal de alerta de que el ítem debe ser revisado, idealmente por un panel externo e independiente de quienes lo formularon. Si los expertos concluyen que los ítems detectados en el análisis presentan sesgo, éstos deben ser eliminados de la prueba definitiva.

Hay varias maneras de realizar el análisis tanto en la TCM como IRT. Generalmente se toma un grupo de referencia, en este caso el particular pagado. Se usó el test de Mantel-Haenszel e IRT para la admisión 2013. Ambos métodos proporcionaron los mismos resultados.

Según la clasificación de Mantel-Haenszel (empleada por el ETS, entre otras agencias), 4 ítems de la PSU-M 2013 presentaron DIF en relación al grupo Municipal Técnico Profesional y 2 con el grupo Subvencionado Técnico Profesional y debieron haber sido objeto de estudio por parte del panel de expertos. En cambio, ningún ítem presentó alerta de sesgo para los estudiantes de la rama Científica-Humanista provenientes de colegios municipales y subvencionados, como se observa en la tabla a continuación.

[55] Bloom, H. , Hill, C., Black, A. & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. MDRC Working Papers on Research Methodology.

Número de Ítems con DIF según clasificación Mantel-Haenszel ETS

| Grupo | A | B | C |
|---|---|---|---|
| **Particular HC - Municipal TP** | 62 | 8 | 4 |
| **Particular HC - Municipal HC** | 74 | 0 | 0 |
| **Particular HC - Subvencionado TP** | 63 | 9 | 2 |
| **Particular HC - Subvencionado HC** | 74 | 0 | 0 |

Fuente de Datos: Elaboración propia, DEMRE 2016.

En 2016, según información proporcionada por el DEMRE, se realizó un análisis del funcionamiento diferencial de ítems para modalidad educacional en el proceso de pilotaje de los items. Si bien esta constituye una buena práctica, no se puede pretender que por la vía del análisis DIF se logren corregir sesgos implícitos en una prueba si ésta es intrísecamente injusta para un grupo de examinados. [56]

[56] Ver al respecto el texto de G. Camilli y L. Shepard (1994). Methods for Identifying Biased Test Items, pg.153-154. Sage Publications.