# VALIDACIÓN DE LA PSU:

## COMENTARIOS AL 'ESTUDIO ACERCA DE LA VALIDEZ PREDICTIVA DE LOS FACTORES DE SELECCIÓN A LAS UNIVERSIDADES DEL CONSEJO DE RECTORES'*

**Mladen Koljatic y Mónica Silva**

El presente trabajo evalúa las conclusiones extraídas del "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores" (Comité Técnico Asesor, 2006). Se concluye que, contrariamente a lo señalado en órganos de prensa, dicho estudio en modo alguno puede calificarse de "contundente" en lo que atañe a refutar los cuestionamientos de que ha sido objeto la PSU. Más bien, a la luz de sus limitaciones, no puede concluirse que el cambio haya incrementado la capacidad predictiva de las pruebas de admisión ni que éstas hayan cumplido con las expectativas de favorecer el acceso de postulantes provenientes de sectores de menores recursos a la universidad. El artículo presenta algunos de los problemas y limitaciones del estudio.

MLADEN KOLJATIC. Ingeniero Comercial de la Universidad Católica de Chile, MBA de la University of Michigan y Doctor en Educación (Ed.D.), con mención en Educación Superior, de Indiana University. Es profesor titular de la Escuela de Administración de la P. Universidad Católica de Chile.

MÓNICA SILVA. Psicóloga de la P. Universidad Católica de Chile, Master of Science y Ph.D., con mención en psicología de la educación, de Indiana University. Es investigadora de la Escuela de Administración de la P. Universidad Católica de Chile.

* Queremos agradecer los valiosos comentarios de un árbitro anónimo de *Estudios Públicos* a un borrador previo de este artículo.

## Introducción

Recientemente se presentó a la opinión pública los resultados de un estudio comparativo de la validez predictiva de la PSU con respecto al antiguo sistema de admisión a las universidades: "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores" (Comité Técnico Asesor, 2006)[1]. Los autores señalaron en sus declaraciones emitidas a la prensa que la evidencia recogida por ellos mostraba de una "manera consistente, un incremento en la capacidad predictiva media en las pruebas obligatorias de la PSU en que se superan[ban] los índices de la PAA Verbal, Matemáticas y notas de enseñanza media en las correlaciones correspondientes". A ello, añadió un rector de una universidad perteneciente al Consejo de Rectores, que "los resultados que estamos presentando responden a los cuestionamientos [formulados a la PSU] en forma muy contundente"[2]. Frente a estas afirmaciones cabe hacerse las siguientes preguntas: 1) ¿Son en verdad tan sólidos los resultados presentados en el estudio que permiten dar respuesta a los cuestionamientos acerca de la validez de la PSU?, 2) ¿Es posible sobre la base de sólo un estudio de predictibilidad garantizar que estamos frente a pruebas de calidad?

La respuesta a ambas preguntas es negativa. Los resultados no son "contundentes" y un único estudio de validez predictiva que toma en cuenta sólo el rendimiento en un primer año de universidad no es suficiente para zanjar el tema de la validación de las nuevas pruebas. (Véase Apéndice "Acerca de la Validación de Pruebas Educacionales".)

## Validación de la PSU

Las expresiones de un rector y de los autores del informe podrían llevar a pensar a quienes no son expertos en el tema, que el único elemento importante en la validación de una prueba de admisión es que prediga bien el rendimiento en la universidad. Si bien la validez predictiva —es decir, la capacidad de ellas para predecir el desempeño futuro de los estudiantes en la universidad— es un aspecto importante de investigar, no es el único.

[1] Consejo de Rectores, Comité Técnico Asesor: "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores" (julio 2006), en www.cruch.cl.

[2] "Estudio Reafirma Validez Predictiva", *El Mercurio*, 5 de agosto de 2006.

La validación de una prueba es un proceso que comprende múltiples aspectos técnicos psicométricos como también una apreciación de las consecuencias sociales y éticas que legitiman su uso para un propósito determinado. Sin lugar a dudas, la predicción es un aspecto esencial tratándose de un instrumento de selección universitaria, pero en modo alguno es el *único* elemento a considerar al momento de evaluar la calidad de una prueba y la legitimidad de su uso. Ello resulta particularmente relevante en el caso de Chile, donde se crearon expectativas de que las nuevas pruebas de admisión no sólo cumplirían bien con el propósito de seleccionar alumnos a las universidades, sino que además servirían como instrumento para favorecer la equidad de acceso a los alumnos provenientes de la educación municipal y tendrían un impacto positivo en la calidad de la enseñanza media.

La validación de una prueba de admisión como la PSU entraña un proceso de análisis que dé respuesta a muchas preguntas tales como: ¿Qué es lo que se está midiendo a través de la prueba?, ¿predice bien lo que se supone que debe predecir?, ¿cumple con las expectativas que crearon sus promotores con respecto a sus efectos beneficiosos en términos de equidad e impacto positivo en la enseñanza media?, ¿qué consecuencias o impactos personales trae asociado su uso entre los que la rinden? y ¿qué impacto tiene en la sociedad como un todo el uso de la prueba? Todos estos aspectos requieren respuesta para tener una visión integral sobre la validez de una prueba.

La validación de una prueba de altas consecuencias como la PSU tiene un paralelo directo en otros campos, como el de la salud, donde también se realizan estudios de validez previo a la introducción de un nuevo fármaco al mercado. Cuando se comercializa una droga, no basta con que sea efectiva para tratar una enfermedad, sino que hay que garantizar que su uso no genere otros males. Y si genera otros males, hay que poner en la balanza sus pros y sus contras. La misma consideración vale al momento de evaluar la calidad de una prueba educacional como son las pruebas de admisión a la educación superior[3].

Ciertamente una prueba de admisión que no tiene capacidad predictiva no sirve, ya que no cumple con el objetivo central para la cual fue creada, pero no toda prueba que tenga capacidad predictiva puede legítimamente ser utilizada para propósitos de selección. Podemos pensar, por ejemplo, en una prueba de inglés u otro idioma extranjero. Una prueba de idiomas, basada en los contenidos mínimos que prescribe el currículum de la ense-

[3] Shepard, L.: "Evaluating Test Validity" (1993).

ñanza media del Ministerio de Educación, podría resultar altamente predictiva del rendimiento universitario, incluso mejor que las pruebas actuales de la PSU. Pero, ¿sería legítimo usar los puntajes de tal prueba para seleccionar alumnos, utilizando el argumento de que es un excelente predictor del rendimiento universitario? La respuesta lógica es no, y la razón es simple: el uso de tal prueba pondría en condiciones desventajosas a todos aquellos alumnos que provienen de establecimientos públicos donde la calidad de la enseñanza de idiomas es inferior a la que existe en establecimientos privados. Es por tanto muy importante definir qué es necesario evaluar en las pruebas y si los postulantes han tenido las oportunidades de aprender los contenidos y destrezas que se exigen para garantizar la equidad del proceso de selección[4].

Los estándares internacionales son claros al establecer las exigencias de un marco de validación integral. Cuando se busca establecer cuál prueba o conjunto de pruebas resulta más adecuada para seleccionar alumnos para la universidad, los estudios rigurosos no se centran exclusivamente en la predictibilidad, sino que consideran además las consecuencias sociales del uso de uno u otro tipo de prueba. Es así como el estudio comparativo de predictibilidad de dos pruebas, el SAT1 y SAT2 realizado por Geisser y Studley (2001) en los Estados Unidos —citado como referencia por los autores del informe chileno— incorporó en su análisis el impacto asociado al uso de una u otra prueba en la composición social de la admisión y cómo se vería afectada la representación de los distintos grupos étnicos o socioeconómicos si se usaba una u otra prueba[5]. Vale decir, el criterio de predictibilidad es necesario pero no suficiente para legitimar el uso de una prueba.



[4] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education: *Standards for Educational and Psychological Testing* (1999).

[5] El estudio de Geisser y Studley, al cual hacen referencia los autores del "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores", aun cuando es más completo y está mejor documentado que el estudio presentado al Consejo de Rectores de Chile, no se puede considerar como un modelo de estudio de validez. Queda corto en algunos aspectos que los expertos recomiendan, ya que tratándose de evidencia acerca de la capacidad para predecir éxito universitario, ésta debe ser abordada a través de múltiples criterios (por ejemplo, tasas de retención y de graduación) y no tan sólo por el rendimiento del primer año, debiendo considerarse además un estudio acabado de las consecuencias sociales para grupos minoritarias y el impacto que el uso de la prueba tiene en lo que los alumnos estudian en educación media (Shepard, 1993).

Cabe señalar al respecto, que cuando se propuso cambiar las pruebas del antiguo sistema de admisión, no estaba en tela de juicio su calidad predictiva[6]. Se cuestionaba, entre otros aspectos, su equidad y se argumentaba que la brecha de rendimiento entre los alumnos de establecimientos municipalizados y privados era menor en pruebas de conocimientos, como la actual PSU[7]. Uno de los argumentos de fondo era la supuesta inequidad de la PAA que favorecía a los alumnos que tenían un mayor capital social y cultural producto de su realidad socioeconómica[8].

Dado lo anterior, no resulta coherente que se pretenda refutar los cuestionamientos a la PSU en base a un único estudio que aborda su capacidad de predicción cuando se debió evaluar ésta en relación a lo que se prometió: predecir bien el rendimiento universitario sin aumentar la brecha de desigualdad en el acceso entre los postulantes proveniente de los distintos tipos de educación, en especial la educación municipal y privada.

Habiendo generado tales expectativas, se esperaría que los autores del estudio hubieran incorporado la variable socioeconómica en su análisis. Más aun cuando hay evidencia de que el nuevo sistema de admisión puede haber acrecentado la brecha de inequidad en el acceso a la educación superior, al menos para una de las más prestigiosas universidades públicas, la Universidad de Chile. En las palabras de su ex rector: "Si en el pasado más de un tercio de los estudiantes provenían de colegios municipales, hoy es sólo el 20% [...] el origen social se ha ido desplazando hacia arriba, producto de los sistemas de selección"[9]. Ante estos antecedentes, cualquier estudio de validez predictiva debió incorporar la variable socioeconómica al estudiar los beneficios del cambio del antiguo sistema de pruebas al actual, tal cual se hizo en el estudio citado de la U. de California. El estudio que ha entregado el Comité Asesor del Consejo de Rectores ignora por completo este punto.

El informe del Comité Asesor contiene además otras aseveraciones acerca de la superioridad de la PSU que son definitivamente erróneas o carentes de respaldo.



---

[6] Véase crónica "[SIES] Sigue Abierto al Debate; Pero No lo Sobrecarguemos de Valoraciones que No Tiene". *La Segunda*, 10/6/2002.

[7] Bravo, D. y J. Manzi: "Equidad y Resultados", en *El Mercurio* de Santiago, 26 de abril de 2002.

[8] Entrevista a D. Bravo, en *Diario Austral*, 21 de julio de 2002; y Brunner, J.: "SIES: Tres Preguntas y la Responsabilidad de las Universidades", en *La Segunda*, 19 de junio de 2002.

[9] En reportaje "Subieron los Puntajes, pero Bajó la Diversidad Social", *El Mercurio* de Santiago, "Artes y Letras", 16 de abril de 2006.

## La apresurada conclusión acerca de la superioridad de la PSU en el Informe Técnico

¿Son de verdad tan sólidos los resultados presentados en el informe técnico del Comité Asesor del Consejo de Rectores? Frente a las declaraciones emitidas tanto por las autoridades universitarias como por los autores del estudio, más de alguien podría pensar erróneamente que los resultados en el informe dan una respuesta definitiva, al menos, al tema de la validez predictiva de la PSU. Sin embargo, existen limitaciones que hacen muy cuestionables las conclusiones que se extraen a partir de éste, particularmente su superioridad con respecto al antiguo sistema de admisión. La situación ideal para comparar ambos sistemas habría sido mantenerlos juntos durante un período de marcha blanca. Al desmantelar prematuramente el sistema anterior, se impidieron las condiciones para hacer un estudio en que se pudiera comparar cabalmente el funcionamiento de ambas pruebas como se hizo en el estudio de Geisser y Studley para el sistema universitario de California.

Por tanto, no se dan las condiciones ideales para establecer una comparación entre ambos sistemas, con lo cual hay que recurrir a comparar dos cohortes distintas.

## ¿Son comparables las cohortes PAA 2003 y PSU 2004?

Los autores del informe técnico comparan la capacidad predictiva de la PAA y la PSU utilizando para ello la admisión del 2003, que es la última generación que rindió la PAA, y la admisión 2004, que es la primera generación que rindió la PSU. El supuesto a la base de la comparación es que los estudiantes matriculados en las universidades del Consejo de Rectores en ambas admisiones son iguales o equivalentes. Pero, ¿se puede considerar que lo sean?

La realidad es que la primera generación de estudiantes que rindió la PSU fue distinta en su composición de la generación anterior. En ella hubo una merma importante de alumnos egresados de años anteriores que no rindieron la PSU, los llamados “rezagados”. Además bajó la proporción de alumnos provenientes de colegios municipalizados: aproximadamente 28 mil alumnos menos rindieron esa primera PSU que el año anterior la PAA. Esto es absolutamente atípico comparado con lo que sucedía en años anteriores (véanse Koljatic, 2004, y Tabla Nº 1). ¿Cómo caracterizar a estos jóvenes que se automarginaron de este proceso? Se podría pensar que los que se

TABLA Nº 1: COMPOSICIÓN DE POSTULANTES SEGÚN EGRESO

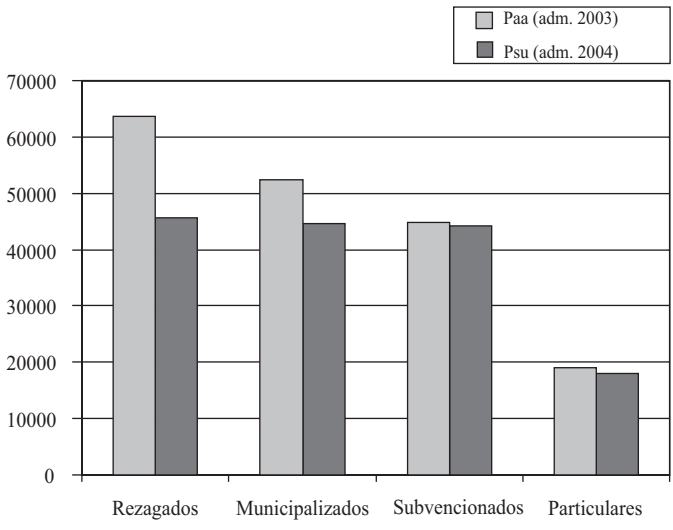
| Proceso de admisión | Total inscritos | Alumnos 4º medio | Rezagados |
|---|---|---|---|
| 2001 | 177.936 | 115.696 | 62.240 |
| 2002 | 188.205 | 122.149 | 66.056 |
| 2003 (última PAA) | 187.584 | 121.409 | 66.175 |
| 2004 (primera PSU) | 159.249 | 111.589 | 47.660 |

*Fuente*: Datos proporcionados por DARA, Universidad Católica de Chile.



marginan son los alumnos con peor preparación en su formación escolar, pero también es posible que se trate de un grupo de jóvenes más reflexivos y capaces, que se dan cuenta de que sus establecimientos no les han brindado la preparación requerida para rendir bien la nueva prueba. Vale decir, hay un cambio en la composición de quienes rinden la PSU comparado con la PAA que hace que ambas cohortes de alumnos no sean comparables (véase Gráfico Nº 1).

GRAFICO Nº 1: COMPOSICIÓN DE LOS QUE RINDEN LA PAA 2003 Y PSU 2004




*Fuente*: Datos proporcionados por DARA, Universidad Católica de Chile.

En segundo término, no se puede desconocer que las condiciones de postulación del año 2004 fueron altamente anómalas. En términos muy sucintos, por el apresuramiento y falta de estudios con que se implementó el cambio de las pruebas de selección, no se pudo anticipar las características de los puntajes de las nuevas pruebas. Se tomó el riesgo de asumir que no habría grandes diferencias entre la PSU y las pruebas anteriores. Ello no se cumplió y al mantenerse los mismos puntajes de corte para la postulación a las carreras en el proceso de admisión, muchos alumnos cuyos puntajes les habría alcanzado para matricularse en universidades del Consejo de Rectores emigraron al sistema de universidades privadas por desinformación. Así, muchas universidades del Consejo de Rectores para llenar sus cupos debieron abrir períodos especiales de inscripción y aceptaban postulantes que en otras circunstancias no hubieran tenido cabida[10]. Tan anómalo fue el contexto de la admisión del 2004 que éste fue recogido por las caricaturas que aparecían en los medios de prensa (véase Apéndice B).

Dado lo anterior, no fue la mejor opción haber elegido esa primera promoción PSU para establecer una comparación con el antiguo sistema. Pero en la eventualidad de querer hacerlo, como ocurrió en la práctica, se debió advertir a los lectores sobre los problemas y limitaciones existentes en la comparación de las cohortes de alumnos, ya que no resulta evidente para quienes no tienen formación en metodología de la investigación.

La cuestionable comparabilidad de las cohortes tiene relación con el fenómeno del rango truncado, que es una condición de todos los estudios predictivos en los cuales se escoge a un grupo selecto dejando fuera a quienes tienen bajos puntajes y se restringe la posibilidad de saber qué hubiese ocurrido con el rendimiento de aquellos alumnos con menor puntaje de haber sido admitidos. Si bien los autores del informe advierten sobre este fenómeno, no señalan que, dadas las anormales condiciones de la admisión 2004, es posible que el fenómeno de rango truncado haya afectado en forma distinta a ambas cohortes de alumnos que ingresaron a la universidad. De ser así, la calidad de la predicción para la promoción que rindió la PSU e ingresó el 2004 podría aparecer mejor que la que se alcanza con la promoción de la PAA en el 2003, sin que lo sea necesariamente porque alumnos que en condiciones normales no habrían ingresado a universidades del Consejo de Rectores, en el año 2004 tuvieron la oportunidad de hacerlo por errores y falta de información en el proceso de admisión.

[10] "UES Tradicionales No Logran Llenar Vacantes", *La Segunda*, 23 de enero de 2004; "Faltó Información Más Fidedigna para Postulación a UES Tradicionales", *La Segunda*, 26 de enero 2004.

Otro cambio que afecta la cuestionable comparabilidad de las cohortes es el cambio en el escalamiento, o sea la forma cómo se transformaron las respuestas correctas netas de los estudiantes al puntaje estándar en la tradicional escala de 200 a 800 puntos, situación que fue advertida por Beyer (2004).

Éstas no son las únicas limitaciones del estudio de validación. Hay otras deficiencias en el manejo de la información y en el reporte de las conclusiones, de las cuales se mencionarán someramente algunas que pueden ser explicadas sin tener que recurrir a detalles excesivamente técnicos.

## Déficit y limitaciones del reporte técnico

Los autores señalan en su informe que "tal como se acostumbra en el contexto internacional, el estudio ha empleado como criterio de predicción el rendimiento de los estudiantes al término del primer año en las carreras en que se matricularon". Efectivamente, éste es el criterio más básico que se utiliza cuando se desea comparar la capacidad predictiva de una prueba. Sin embargo, los investigadores extranjeros que ellos mismos citan en su informe se preocupan de señalar que "este criterio *no es suficiente ni concluyente* para afirmar la superioridad de una prueba sobre otra y que se requiere extender los análisis más allá de las notas de primer año, como, por ejemplo, con indicadores de retención y tasas de graduación" (Geisser y Studley, 2001, p. 7)[11].

El informe de la PSU no sólo se centra en un único indicador, sino que además presenta otros déficits. Por ejemplo, brinda información insuficiente con respecto al número de alumnos considerados para el análisis. Se señala que el número de alumnos fue de 45.149 para la cohorte del 2003 y de 42.771 para la del 2004 (p. 13). Sin embargo, estas cifras difieren del número de alumnos matriculados según el Consejo Superior de Educación: 47.150 el año 2003 y 46.965 el año 2004[12]. Vale decir en el año 2003 hay una "pérdida" de información de 2.001 alumnos y el año 2004 de 4.194. Esto representa más del doble de "pérdida" en el 2004 con respecto al año anterior. Dada la naturaleza del estudio, correspondía haber reconocido esta diferencia, investigado su origen y haber reportado una explicación para ella.

Tampoco se le asigna la debida importancia al aumento de la ponderación de las notas de enseñanza media en la admisión 2004, el primer año

[11] Destacado nuestro.

[12] Datos página web del Consejo Superior de Educación (www.cse.cl).

en que se rinde la PSU. Aproximadamente la mitad de las carreras aumentó en al menos un 5% su ponderación. Este es un factor relevante al momento de comparar la capacidad predictiva de ambas pruebas por cuanto las notas de enseñanza media son el mejor predictor del rendimiento en la universidad. Al haber cambiado el peso de las notas de enseñanza media para la postulación a ciertas carreras, se introdujo un factor de sesgo para la comparación directa de las cohortes 2003 y 2004 que requería de un análisis y explicación adicional.

Por otra parte hay hallazgos que por su relevancia debiesen haber sido destacados para que el lector pudiese interpretar correctamente los resultados. Sin embargo, no se hizo. Por ejemplo, la capacidad predictiva de la PSU comparada con la PAA no aparece mejorada en el caso de la Pontificia Universidad Católica. El caso de la PUC es interesante, puesto que era una universidad que —a diferencia de la mayoría de las del sistema universitario— exigía para muchas carreras tanto la PAA como las PCE (Pruebas de Conocimientos Específicos). Más aun, frente al cambio de las pruebas, en muchas facultades se hizo un esfuerzo por mantener las ponderaciones lo más similares posibles a las que se usaban con el antiguo sistema. Tampoco abrió un período especial de admisión para llenar sus vacantes el año 2004 ante la debacle del proceso de admisión como hizo la mayoría de las universidades. Por lo anterior y a la luz de todas las limitaciones señaladas en las secciones anteriores, la PUC pareciera ser la institución que mejor podría representar el impacto del cambio de pruebas en el sistema. El hecho de que en esta institución no se advirtiera un cambio en la capacidad predictiva el año 2004 debería haber servido para moderar el entusiasmo de los autores del informe con respecto a la primacía de la PSU como instrumento de selección. El juicio por ellos emitido de que "los resultados muestran de manera consistente un incremento en la capacidad predictiva de las nuevas pruebas… [y] la comparación entre las dos baterías de selección es favorable a las nuevas pruebas" (pp. 49 y 50), no es sustentable, y debió ser matizado al menos con un párrafo para no inducir a las autoridades ni al público lego en la materia a un error de sobredimensionar la calidad de las actuales pruebas.

Finalmente, no puede dejar de mencionarse que el informe técnico dista mucho de los estándares usuales en esta materia al reportar los indicadores estadísticos (c.f., Geisser y Studley, 2001). Dado que este análisis no va dirigido a una audiencia técnica, baste señalar que la información proporcionada es insuficiente en este caso para estimar la calidad de la predicción.

Por ejemplo, en el reporte del análisis de regresión, se omiten tablas con los coeficientes para las ecuaciones entre otros antecedentes, que son un estándar mínimo en cualquier publicación académica[13].

## Conclusiones y recomendaciones

El estudio presentado como la última palabra con respecto a la calidad de la PSU como instrumento de selección a las universidades chilenas dista mucho de serlo. Es un reporte parcial, incompleto y que omite análisis importantes, como es el impacto en la equidad. La equidad fue un argumento central para apoyar el cambio de pruebas en el caso chileno, sin embargo, sorprendentemente en el informe no hay ninguna mención sobre este tema. Tal omisión es incomprensible, más aun a la luz del reconocimiento que hizo un rector acerca de la baja experimentada en el ingreso de alumnos provenientes de la educación municipal en su plantel, que coincide exactamente con el cambio del antiguo sistema de admisión por la PSU.

El presente estudio de validez predictiva sigue la tendencia de reportes anteriores emanados del mismo Comité Técnico en los cuales se advierte un sesgo "triunfalista" en la interpretación de la calidad de los nuevos instrumentos. Es así como en un informe titulado "Resultados de la Aplicación de Pruebas de Selección Universitaria" (2004), señalaban que los grados de dificultad de las pruebas estaban en el rango "deseable", tratándose de "instrumentos construidos con propósitos selectivos" (p. 17). La realidad es que los grados de dificultad eran inadecuados, particularmente en las pruebas de Matemáticas y Ciencias en que los promedios estaban en el

[13] Los autores omiten además incluir información relativa a pruebas de significancia estadística. Aun suponiendo que algunas de éstas fueran estadísticamente significativas —habiendo previamente controlado por nivel socioeconómico lo cual no se hizo— entonces habría correspondido que comentaran acerca de si éstas son sustantivas desde una perspectiva educacional. De ser significativas las diferencias, se requeriría una reflexión, ya no estadística sino conceptual, que pudiera explicar el cambio en la capacidad predictiva de las pruebas a la luz del cambio en su naturaleza. A modo de ejemplo, si las nuevas pruebas están más altamente correlacionadas con nivel socioeconómico y el éxito universitario también lo está, se esperaría una mejoría en la predicción. Por lo tanto, las correlaciones no dan una respuesta cabal a los mecanismos que subyacen a la asociación entre las variables que se estudian, y por ello se echa de menos una discusión con respecto a este punto por parte de los autores de este estudio, puesto que es importante que una prueba de admisión prediga bien, pero por las razones correctas, como señala Shepard (1993).

rango del .30 - .40 debiendo ser del orden de .60[14]. Una afirmación similar se hizo en el informe del año 2005, en abierta contradicción con los parámetros establecidos en la disciplina.

A la luz de lo anterior y atendiendo a la importancia de las pruebas de selección, es imprescindible que el tema de la validación de la PSU se desarrolle con rigurosidad y transparencia. Ello implica realizar una variedad de estudios que no han sido desarrollados aún. Adicionalmente, es indispensable en aras de la transparencia que se pongan a disposición de la comunidad académica, no sólo los informes emanados del Comité Técnico, sino que las bases de datos completas a quien las solicite, lo cual no ocurre en la actualidad. Asimismo, el Consejo de Rectores debería considerar el enviar los informes a un proceso de arbitraje independiente para garantizar que las conclusiones que se extraen a partir de los datos son las adecuadas antes de presentarlas a la opinión pública.

En resumen, en el informe "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores" se aprecia una falta de rigurosidad al momento de precisar los alcances y limitantes del estudio. Hay que tener presente que estos informes son frecuentemente utilizados por quienes toman decisiones en el ámbito de las políticas públicas, y ellos no son necesariamente expertos en temas de medición. El estilo discursivo del informe, al igual que el de otros estudios emanados de este Comité Técnico Asesor, trasunta un exceso de confianza y autocomplacencia, lo que puede explicar la reacción desmesurada de algunos Rectores de pensar erróneamente que todos los problemas están resueltos. No sólo el tema de la predictibilidad no ha sido exhaustivamente tratado en éste, sino que quedan pendientes temas cruciales, como evaluar si las nuevas pruebas satisfacen las promesas hechas con respecto a sus beneficios sobre el sistema educacional de enseñanza media y su mayor equidad.

[14] Henrysson, S.: "Gathering, Analyzing, and Using Data on Test Items" (1971).

## APÉNDICE A

### ACERCA DE LA VALIDACIÓN DE PRUEBAS EDUCACIONALES

Hay criterios o estándares profesionales de validación para el uso de pruebas de altas consecuencias en el ámbito de la educación. El propósito de éstos es proveer de un marco de referencia para evaluar la calidad de éstas, promoviendo un uso adecuado y ético de las pruebas. (AERA, APA, y NCME, 1999.)

La validación de una prueba es, en esencia, un proceso de evaluación. Consiste en recoger evidencia de que la prueba sirve al propósito para el cual fue diseñada y que cumple con lo que se prometió acerca de ella.

En el pasado, se consideraban tres dimensiones de la validez, dependiendo de la naturaleza de la prueba. Así se hablaba de validez de contenido, de criterio (o predictiva) y de constructo. Dependiendo del tipo de prueba, se privilegiaba una u otra categoría de validez. Por ejemplo, tratándose de los exámenes de grado que rendían los abogados para poder graduarse y ejercer, se privilegiaba la validez de contenido. Ello requería establecer si el contenido de las pruebas era el adecuado. La pregunta clave era: "*¿Qué conocimientos legales y destrezas de análisis debe tener un abogado para ejercer la profesión?*" "*¿Mide bien esta prueba (examen de grado) los contenidos que debe manejar un abogado?*" La evidencia para juzgar la calidad de la prueba era el juicio de expertos que dictaminaban si la prueba era o no válida para el propósito. Un proceso de análisis lógico bastaba para satisfacer los requisitos de validación de una prueba de esta naturaleza. En cambio, cuando se trataba de pruebas de selección universitaria, lo importante era la validez de criterio, y el foco de la atención estaba dirigido a estimar la magnitud de la correlación con algún criterio de éxito en la universidad. La pregunta de "*¿mide esta prueba lo que se supone que debe medir?*", era respondida por la vía de una simple correlación, por ejemplo, entre los resultados de la prueba en cuestión y el rendimiento en un primer año de universidad. La validez de constructo, por otra parte, era considerada necesaria cuando sobre la base de una prueba se hacían inferencias acerca de rasgos no observables, como por ejemplo la inteligencia, la depresión o la ansiedad. Ésta se establecía empíricamente por la vía del análisis factorial, entre otros.

En la actualidad, el antiguo esquema que privilegiaba una dimensión u otra de validez fue superado. Hoy en día se considera que la validación de

cualquier tipo de prueba requiere de múltiples fuentes de evidencia tanto de análisis lógico como estadístico. Mientras más altas consecuencias tenga una prueba, mayores son las exigencias en cuanto a la evidencia de validez (AERA, APA y NCME, 1999, p. 139). Adicionalmente, una concepción moderna de validez incluye también el estudio acucioso de los efectos no anticipados de las pruebas. Así, las consecuencias adversas —usualmente no anticipadas— deben ser evaluadas al sopesar el valor funcional de usar una prueba (Messick, 1989). Esta es la concepción que han recogido las más importantes asociaciones de especialistas de medición del mundo al presentar los estándares que deben cumplir las pruebas en el ámbito educacional y psicológico.

Es así como la nueva concepción de validación de pruebas de altas consecuencias requiere que se evalúen tanto los beneficios y efectos anticipados de éstas como sus efectos secundarios. Así, por ejemplo, si se ha postulado que una nueva prueba de selección a las universidades 1) va a ser un buen instrumento de selección, 2) será beneficioso a la enseñanza media, mejorando la calidad de la enseñanza en aula, 3) dará mayores oportunidades de acceso a los más pobres al sistema universitario, los tres puntos deben ser analizados en profundidad en el marco del estudio de validación de la nueva prueba, puesto que eran beneficios esperados asociados a la nueva prueba. Adicionalmente, habría que evaluar, tanto para los individuos como para el sistema educacional, los efectos secundarios no anticipados por quienes diseñaron las pruebas. Por ejemplo, las restricciones en la libertad curricular, promoción de prácticas que apuntan a aumentar los puntajes sin necesariamente mejorar el aprendizaje o el aumento de niveles de estrés entre alumnos y profesores, etc. A la luz de lo anterior, el esquema simplista basado en un juicio de calidad de una prueba de selección que se sustenta en un mero coeficiente de correlación corresponde a una visión superada en medición. Por de pronto, tal como señala Shepard (1993), es evidente que el puntaje de una prueba y el criterio que pretende predecir pueden estar correlacionados por las razones “incorrectas” (por ejemplo, si ambos comparten el mismo sesgo). Por tanto el legítimo uso de una prueba debe estar avalado por múltiples fuentes de evidencia, y mientras más numerosas hayan sido las promesas asociadas a los beneficios potenciales que ésta conlleva, mayores son las exigencias en cuanto a requerimientos de evidencia de validación.

APÉNDICE B



*Fuentes: El Mercurio*, A3, 25/1/2004; *El Mercurio*, A3, 28/1/2004. (Reproducción autorizada.)

## REFERENCIAS

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME): *Standards for Educational and Psychological Testing*. Washington, D.C.: 1999.

Beyer, Harald: "Reflexiones Preliminares sobre la Prueba de Selección a la Universidad". *Puntos de Referencia* Nº 274 (enero 2004). Centro de Estudios Públicos, Santiago.

Bravo, David. Entrevista. En *Diario Austral*, 27 de julio de 2002.

Bravo, D. y J. Manzi J.: "Equidad y Resultados". En *El Mercurio* de Santiago, 26 de abril de 2002.

Brunner, J.: "SIES: Tres Preguntas y la Responsabilidad de las Universidades". En *La Segunda*, 19 de junio de 202.

Comité Técnico Asesor, H. Consejo de Rectores de las Universidades Chilenas: "Estudio Acerca de la Validez Predictiva de los Factores de Selección a las Universidades del Consejo de Rectores". julio 2006, en www.cruch.cl.

———— "Resultados de la Aplicación de Pruebas de Selección Universitaria Admisión 2004". H. Consejo de Rectores de las Universidades Chilenas, Santiago, 2004.

———— "Resultados de la Aplicación de Pruebas de Selección Universitaria Admisión 2005". H. Consejo de Rectores de las Universidades Chilenas, Santiago, 2005.

Consejo Superior de Educación: www.cse.cl

*El Mercurio* de Santiago: "Subieron los Puntajes, pero Bajó la Diversidad Social". En sección "Artes y Letras", 16 de abril de 2006.

———— "Estudio Reafirma Validez Predictiva". 5 de agosto de 2006.

Geisser, S. y R. Studley, R.: "UC and the SAT: Predictive Validity and Differential Impact of the SAT 1 and SAT 2 at the University of California". Office of the President, University of California, 2001.

Henrysson, S.: "Gathering, Analyzing, and Using Data on Test Items". En R. L. Thorndike (ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 2da edición, 1971.

Koljatic, Mladen: "Problemas Actuales y Desafíos Futuros". Presentación en Universidad Andrés Bello, Santiago de Chile, 27 de agosto de 2004.

*La Segunda:* "[SIES] Sigue Abierto al Debate; Pero no lo Sobrecarguemos de Valoraciones que No Tiene". 10 de junio de 2002.

———— "UES Tradicionales No logran Llenar Vacantes". 23 de enero de 2004.

———— "Faltó Información Más Fidedigna para Postulación a UES Tradicionales". 26 de enero 2004.

Messick, S.: "Validity". En R. L. Linn (ed.), *Educational Measurement*. NY: American Council on Education y Macmillan, tercera edición, 1989.

Shepard, L.: "Evaluating Test Validity". En *RER*, Nº 19 (1993), pp. 405-450. □