EXIGENCIAS PARA LA CONSTRUCCIÓN DE UNA PRUEBA DE SELECCIÓN A LA UNIVERSIDAD

Bárbara Eyzaguirre

Con el objeto de regular el proceso de construcción de pruebas y garantizar un uso adecuado de sus resultados, se han desarrollado estándares que prescriben lo que se considera una buena práctica en la elaboración y el uso de instrumentos de evaluación. Este artículo se propone dar a conocer esos estándares, y, a la luz de ellos, evaluar la prueba de admisión a la educación superior utilizada en Chile hasta el 2002 (PAA), así como el proyecto SIES y la prueba PSU, actualmente en preparación, que se aplicará a fines del 2003.

Del análisis se concluye que el procedimiento seguido en la construcción de la nueva prueba de admisión (PSU) no ha sido el regular, y que son mínimas las posibilidades de que ésta llegue a cumplir, antes de su aplicación, con los estándares requeridos.

Por ello, se señala finalmente, la prueba de admisión 2003 no debiera constituirse en un marco forzado para las evaluaciones de los años subsiguientes, sino que se la debería considerar como un antecedente más en el estudio de la batería idónea para seleccionar a los mejores alumnos para la universidad. En este sentido, se sugiere

BÁRBARA EYZAGUIRRE. Sicóloga educacional especializada en desarrollo cognitivo, con experiencia en programas de mejoramiento de la calidad de la educación en sectores de escasos recursos. Fundadora y asesora pedagógica de la Escuela San Joaquín (Renca), perteneciente a la Fundación Marcelo Astoreca. Investigadora del Centro de Estudios Públicos.

Se agradecen los aportes de Harald Beyer y Carmen Le Foulon.

reabrir el debate acerca de la conveniencia tanto de eliminar las pruebas de conocimientos específicos como de exigir pruebas en asignaturas no afines a las carreras o áreas de estudio a las que se postula, e investigar la factibilidad de adaptar nuevas pruebas o secciones de pruebas extranjeras.

La construcción de instrumentos de evaluación requiere de procedimientos sistemáticos y rigurosos. En las pruebas con altas consecuencias, como es el caso de los exámenes de admisión a las universidades, estas exigencias cobran aun mayor relevancia. El proceso se inicia con la identificación del propósito de la evaluación y finaliza una vez que se demuestra que la prueba sirve para los objetivos para los cuales fue diseñada. Diferentes estándares internacionales sobre la construcción de pruebas definen una serie de pasos que deben respetarse al elaborar instrumentos de evaluación que implican consecuencias serias para los individuos. Respetarlos da garantías a los usuarios de que se someterán a un proceso justo y confiable. A continuación se describen brevemente dichos pasos y luego se analiza la posibilidad de cumplirlos que tienen las Pruebas de Selección Universitaria (PSU) 2003 en los plazos que el Consejo de Rectores le ha fijado.

I. ETAPAS DEL DESARROLLO DE PRUEBAS DE SELECCIÓN A LA UNIVERSIDAD[1]

Las pruebas que se utilizan en educación pueden dividirse en dos tipos según las implicancias de sus resultados para el estudiante: de altas y bajas consecuencias. Los exámenes de admisión a la universidad se consideran de altas consecuencias, ya que a partir de sus resultados se toman decisiones trascendentales para las personas.

Con el objeto de regular el proceso de construcción de pruebas y garantizar un uso adecuado de los resultados, la comunidad de expertos en el tema ha elaborado estándares que prescriben lo que se considera una buena práctica en el ámbito del desarrollo y uso de instrumentos de evaluación. El documento que los recoge es el *Standards for Educational and Psychological Testing*, elaborado por la American Educational Research

[1] El listado de etapas que se presenta a continuación se basa en el documento elaborado por Carmen Le Foulon y Francisca Dussaillant, "Desarrollo de Pruebas Estandarizadas" (2002).

Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME). El *Standards for Quality and Fairness*, desarrollado por el Educational Testing Service (ETS), se alinea con los anteriores y los refuerza. En adelante nos referiremos a ellos como estándares de la AERA y del ETS. Según el primer documento, "mientras más altas son las consecuencias de una prueba, más importante resulta que las inferencias que se hagan a partir de ella sean avaladas con evidencia sólida y de calidad técnica. En particular, cuando las consecuencias son individuales y serias, y cuando se toman decisiones importantes basadas en el desempeño en las pruebas, el instrumento necesita exhibir estándares elevados"[2].

Según los estándares, las etapas mínimas que se deben respetar al elaborar este tipo de pruebas son las siguientes:

1.1. Desarrollo de las especificaciones de la prueba

En la etapa inicial de la planificación de una prueba se debe definir la suma de sus características. Las especificaciones contienen el propósito de la prueba, la fundamentación que avala la selección de contenidos, destrezas y habilidades a ser medidas, la tabla de especificaciones de los contenidos a evaluar, la definición del formato de las pruebas y una anticipación de las consecuencias que se pueden esperar. "Idealmente las especificaciones deben ser tan completas que dos constructores de prueba operando independientemente sobre la base de ellas debieran producir instrumentos comparables e intercambiables, que difieran tan sólo en la muestra de preguntas utilizadas"[3]. A continuación se detallan cada uno de los aspectos que deben definirse en la etapa de especificación de la prueba.

Descripción clara de los propósitos de la prueba. Es de vital importancia identificar el o los propósitos principales para los cuales serán utilizados los resultados de la prueba. En este nivel de análisis se debe estudiar la factibilidad de abordar en una sola prueba los distintos objetivos que se están proponiendo. En ocasiones los objetivos pueden ser incompatibles, lo que hace necesario evaluar si es posible que todos ellos se logren a partir de una misma evaluación. En el caso de las pruebas de selección a la universidad que nos ocupan, pretender evaluar la calidad de la educación media y

[2] Véase Estándar 13, en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999), p. 139.

[3] S. Tinkelman, "Planning the Objective Test" (1971).

seleccionar a los alumnos a la universidad puede no ser viable. En un caso, el diseño apunta a la selección de alumnos para programas educacionales altamente competitivos, por lo tanto, debe focalizarse en la capacidad de adquirir conocimientos avanzados y en la evaluación de las habilidades para cursar estudios superiores; en el caso que se buscara evaluar el logro de los objetivos de la educación media, la prueba debiera cubrir el dominio de todo el rango de contenidos independientemente de la relación de éstos con los estudios universitarios.

Fundamentación del tipo de prueba. En la etapa de especificación de la prueba se debe aportar evidencia lógica, teórica y empírica que apoye las inferencias que se harán a partir de los puntajes obtenidos en ella. Si se trata de pruebas que evalúan conocimientos y destrezas, tiene que quedar claramente estipulado cuán representativa será la muestra de tareas que se exigirán del universo total de tareas que se quiere medir. Asimismo, en las pruebas que apuntan a habilidades o disposiciones debe fundamentarse el constructo teórico sobre el que se basa la medición, es decir, los supuestos que avalan la relación entre los indicadores indirectos escogidos y la habilidad misma. Cuando la evaluación pretende predecir desempeño futuro debe quedar fundamentada la evidencia lógica y empírica que sustenta la elección de los indicadores utilizados.

En esta etapa tiene que discutirse la delimitación de los constructos[4] y dominios a evaluar; por ejemplo, hay que responder preguntas del tipo: ¿El dominio de la biología tiene alguna relación con el desempeño posterior en carreras del área de las ingenierías? ¿La matemática de IV año de educación media incluye derivadas? ¿Incluye aplicación de conceptos a la resolución de problemas o sólo reconocimiento de definiciones? ¿La habilidad verbal incluye comprensión lectora y manejo de vocabulario? En esta tarea, la consulta a expertos debe ser amplia para recoger información actualizada sobre cómo se definen y operacionalizan las habilidades a evaluar y sobre lo que consideran relevante de medir en cada una de las disciplinas. La calificación, relevancia de la experiencia y las características demográficas de los jueces deben quedar documentadas[5].

[4] El término constructo tiene una connotación amplia que se refiere a los conceptos o características que un test pretende medir. La connotación más estrecha reserva el término para características no observables directamente, pero que se pueden inferir a partir de un conjunto de observaciones interrelacionadas. Un ejemplo de constructo entendido en este sentido más estricto sería la motivación, ya que sólo se la puede observar a partir de un conjunto de disposiciones y acciones del individuo.

[5] Véase Estándar 3.5 en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999).

El marco teórico de la prueba debe servir de guía para las etapas subsecuentes de validación del instrumento. La definición precisa del universo de dominios a evaluar, la especificación clara de los constructos y la identificación de las variables de criterio que se utilizarán para contrastar las predicciones con la realidad, entregan los parámetros con los cuales los jueces expertos tendrán que revisar las pruebas finales.

Definición de los dominios a evaluar. En esta etapa se traduce operacionalmente el marco teórico en una definición detallada de los dominios, subdominios y tipos de destrezas a evaluar. Luego se explicitan en una tabla que contiene el peso que se le asignará a cada área. Ello resulta particularmente importante para el proceso de validación de contenidos y en los casos en que habrá un equipo relativamente grande de profesionales desarrollando ítems.

Formato de la prueba y de los ítems. Se debe definir el formato de la prueba (por ejemplo, papel o computador), el número de ítems a incluir por tarea y el número de ítems en cada área. También es importante especificar el nivel apropiado de lenguaje (extensión de las lecturas, tipo de lenguaje a utilizar) y se debe definir el formato de todas las preguntas (ensayo, opción múltiple, etc.) y las tareas a realizar (comprensión de lectura, completación de oraciones, eliminación de la incorrecta, etc.). Al nivel de los ítems es necesario establecer un límite de palabras para los enunciados, así como definir las características de las opciones y de los distractores (en el caso de preguntas de selección múltiple). Otro tema de importancia dice relación con las instrucciones de la prueba. Se debe definir de antemano los requerimientos para éstas, es decir, si bastará con algunas preguntas de ejemplo al inicio de cada sección de la prueba o si será necesario confeccionar pruebas completas de muestra para que la población esté al tanto del tipo de preguntas que deberá enfrentar.

Características psicométricas. Las especificaciones de la prueba deben incluir, entre otros, el nivel de dificultad de ésta, la distribución de dificultad de los ítems, directrices para evaluar la homogeneidad de los ítems[6] y una descripción de los requerimientos para el *equating*[7] (métodos, número de preguntas ancla, distribución de éstas, etc.). También se debe estimar el número de ítems a utilizar y el tiempo total de la prueba.

[6] Es decir, verificar que los ítems en un determinado subgrupo son de la misma naturaleza.

[7] Mecanismo que permite comparar los puntajes con los de otras formas paralelas de la misma prueba o con aplicaciones en distintos años.

Especificaciones de equidad. Antes de desarrollar la prueba, es importante definir las directrices para abordar las diferencias culturales y de género. Es importante definir cómo se asegurará la representatividad de los diferentes subgrupos al momento de discutir las preguntas y los análisis estadísticos que evalúen diferencias entre grupos (Diferential Item Functioning, DIF, por ejemplo). También se debe discutir si la prueba supondrá la exposición a experiencias de aprendizaje similares o si justamente se quiere evaluar las diferencias de oportunidades.

El proceso de desarrollo de las especificaciones debe estar sujeto a continuas revisiones durante el proceso de construcción de las pruebas.

1.2. Construcción de los ítems

La construcción del conjunto inicial de ítems implica las siguientes actividades:

a) Selección de un formato de ítems apropiado y verificación de que el formato sea adecuado para los examinados.
b) Selección y entrenamiento de quienes redactarán los ítems.
c) Producción de un conjunto grande de ítems.
d) Revisión inicial de los ítems por expertos. En el caso de pruebas de alternativas, se deben revisar los siguientes aspectos:
    - Claridad y precisión: el enunciado debe definir claramente el problema o tarea, medir un solo concepto, contener suficiente información para resolver el problema sin que se requiera mirar las alternativas, ser consistente gramaticalmente con las alternativas y, en lo posible, no contener negaciones.
    - Relevancia y pertinencia de acuerdo a la tabla de especificaciones: las preguntas deben calzar con las habilidades y conocimientos que se busca evaluar.
    - Fallas técnicas en la construcción de los ítems: en este punto se revisa la calidad de los distractores, se analiza la plausibilidad de los mismos, es decir que representen en la medida de lo posible concepciones erróneas de la pregunta. En este sentido también es importante verificar que las alternativas no ofrezcan pistas que permitan responder las preguntas sin tener los conocimientos necesarios; por ejemplo, revisar que ninguna resalte por falta de concordancia gramatical.
    - Apariencia de sesgo: las preguntas no deben contener un lenguaje ofensivo, alarmante o que haga referencia a un subgrupo de los alumnos que rinden la prueba.

e) Realizar pruebas *preliminares de los ítems* y revisarlos nuevamente a la luz de los parámetros anteriores. Esta tarea requiere probar los ítems en una muestra pequeña de individuos. La idea es hacer un *focus group* para detectar tempranamente posibles problemas de los ítems.

Los pasos para la construcción de los ítems son iterativos. Un ítem debe pasar múltiples revisiones antes de integrar la muestra de ítems que se prueban en la aplicación experimental.

f) Realizar una *aplicación experimental* o piloto de los ítems en una muestra de la población de individuos que rendirán finalmente el test. La aplicación piloto involucra la administración de los ítems, ensamblados en formatos de prueba similar al definitivo, a una muestra representativa de los individuos que rendirán la prueba[8]. Esta aplicación busca determinar las propiedades estadísticas de los ítems y, cuando corresponda, eliminar los ítems que no cumplen con los criterios preestablecidos. En esta etapa se deben probar alrededor del triple de las preguntas que quedarán finalmente.

Se insiste en que la evaluación de los ítems debe realizarse en contextos lo más parecidos a los de las evaluaciones reales. La motivación de los alumnos, la preparación de los alumnos, el nivel de dificultad de la prueba (estimado a pulso en la evaluación de expertos y en las pruebas preliminares), la extensión de la prueba, el orden de las preguntas y el tiempo total deben ser lo más similares posible para que los índices estadísticos obtenidos no cambien con respecto a la prueba experimental. El equipo que desarrolla la prueba debe demostrar que las diferencias en las condiciones de administración de las pruebas experimentales con las finales, si es que las hay, no distorsionarán el comportamiento estadístico de los ítems[9].

La aplicación piloto es una etapa importante dentro del proceso de construcción de ítems, pero no sustituye a los estudios de validez de las pruebas. Hambleton y otros (1991) señalan que aun cuando se utilicen sofisticadas técnicas estadísticas basadas en la Teoría de Respuesta al Ítem (IRT), ello no es garantía de que la prueba sea "técnicamente buena". Los métodos basados en IRT no corrigen problemas tales como pruebas desalineadas de los objetivos propuestos o ítems de contenidos irrelevantes que se comportan bien estadísticamente.

[8] En Teoría Clásica se exige una muestra representativa y en Teoría de Respuesta al Ítem (IRT) se exige una muestra grande y heterogénea de individuos.

[9] Véase Estándar 3.9 en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999).

1.3. Ensamblado de la prueba

De los formatos testeados en la prueba piloto, se eligen los ítems que quedarán en las versiones definitivas. Esta selección debe cumplir con la tabla de especificaciones y los parámetros psicométricos definidos en la etapa inicial de la prueba. Para ello, se utilizan los índices de dificultad obtenidos empíricamente en la prueba experimental descrita en el punto anterior. Los formatos finales no son esencialmente distintos de los utilizados en la prueba piloto en cuanto a su extensión y ordenación, pero esta vez el nivel de dificultad de la prueba responde a una estimación empírica que garantiza una versión que ordene a los individuos como se desea.

1.4. Investigar la confiabilidad y validez de la prueba

La validez se refiere al grado en que la evidencia empírica y la teoría respaldan la interpretación de los puntajes de una prueba. Según los estándares de la AERA, la validación es la consideración fundamental del proceso de desarrollo y evaluación de una prueba. Es la que garantiza que la prueba mida lo que dice medir. Aplicar una prueba sin estudios de validez es cuestionable, más aun tratándose de tests que implican altas consecuencias.

El proceso de validación implica la acumulación de evidencia sólida y científica para el modo en que se interpretarán los puntajes de la prueba, ya que no se validan los instrumentos sino que el uso que se hace de ellos. Cada intención de uso debe ser validada en su propio mérito[10].

La validación se inicia en la etapa de especificación de la prueba cuando se delimitan explícitamente el propósito de la prueba y sus usos, así como las inferencias e interpretaciones que se harán de ella. Las decisiones acerca de la clase de evidencia que es más relevante para cada prueba deben quedar estipuladas en un conjunto de hipótesis que indiquen sucintamente los supuestos teóricos que se están asumiendo. Por ejemplo: “que los examinados que obtienen puntajes altos tendrán mayores probabilidades de éxito en cursos avanzados que los que obtienen puntajes bajos”, o “un buen puntaje en la prueba de matemática indica que el alumno domina todos los contenidos que son prerrequisito para un curso avanzado”. Estas definiciones son las que proveen el marco de referencia que guiará los estudios de validez.

[10] Véase Estándar 1 en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999).

La validación es un ejercicio conjunto de los equipos que elaboran las pruebas y de los que las utilizarán. Los que desarrollan las pruebas son responsables de la elaboración de argumentos sólidos que avalen los usos propuestos de la prueba y también les corresponde reunir la evidencia teórica y empírica. Los usuarios deben evaluar la evidencia que justifica el uso de la prueba en contextos particulares[11].

Los estándares de la AERA distinguen distintas fuentes de validez, cada una ilumina aspectos esenciales que deben ser estudiados cuando se quiere afirmar que una prueba mide lo que realmente dice medir. Actualmente se postula un concepto unitario de la validez donde la acumulación total de la evidencia sirve para construir un argumento sólido que respalde la prueba.

Según la AERA la evidencia debe provenir de los análisis siguientes[12]:

Análisis de los contenidos. Esta categoría dice relación con el grado en que los ítems incluidos en la prueba representan bien los contenidos que la prueba desea medir. En el caso de una prueba de admisión, se requiere documentar que los ítems incluidos representan cabalmente los contenidos descritos en la tabla de especificaciones. La evidencia de validez correspondiente a esta categoría se basa, esencialmente, en el juicio de expertos que se pronuncian acerca de la idoneidad de los ítems incluidos en la prueba.

Análisis de los procesos utilizados por los evaluados para responder la prueba. En la etapa de revisión preliminar de los ítems y en la etapa de revisión por jueces expertos se debe analizar la forma en que los alumnos responden los ítems. El análisis tiene que proveer evidencia de que la forma de enfrentar el problema planteado responde al constructo que se está midiendo. Por ejemplo, si se plantea que una sección de una prueba mide razonamiento matemático, es importante determinar si de hecho los alumnos están razonando o sólo están aplicando algoritmos estándares.

Análisis de la estructura interna de la prueba. Se debe demostrar el grado en que la totalidad de los ítems de una prueba y las secciones de la misma se relacionan con el o los constructos de la prueba. En los análisis preliminares se analiza la concordancia de cada ítem con el constructo de la prueba; en esta nueva instancia se vela por la coherencia global del instrumento. Por ejemplo, si se postula que una prueba mide una sola dimensión, se debe demostrar su homogeneidad.

[11] Ibídem.

[12] Para una descripción detallada de estas líneas de validez véase Estándar 1, en ibídem, pp. 11-17.

Análisis de la relación de los puntajes de la prueba con variables externas a la prueba. Entre las variables externas se incluyen los estudios de validez predictiva, que relacionan los puntajes de la prueba con la medición de la variable de criterio que se pretende predecir. Los estudios de validez concurrente también se consideran en esta categoría; éstos están orientados a encontrar una relación directa con otros instrumentos que miden constructos similares. En este sentido, también aporta evidencia la comparación de los resultados con pruebas que miden constructos relativamente distintos. En el caso de esta validación discriminante, la relación entre los puntajes de una y otra prueba debieran ser bajos.

Análisis de las posibilidades de generalizar la utilidad de la prueba a otros contextos. Cuando la prueba se utiliza en contextos distintos a los originales se debe demostrar que éstos no afectan a la validez del instrumento. Por ejemplo, si la naturaleza de las habilidades y conocimientos requeridos para cursar con éxito el primer año de universidad varían sustancialmente, es probable que la validez predictiva de la prueba se modifique. En este caso se deben hacer los estudios correspondientes para ajustar la prueba a los nuevos requerimientos si es que los indicadores de predictibilidad bajan significativamente.

Análisis de las consecuencias producidas por las pruebas. Los estándares de la AERA exigen demostrar que el uso de las pruebas logra los beneficios que sus constructores anuncian. De igual modo deben reunir evidencia de que no producen efectos negativos. Por ejemplo, si se asevera que una prueba de admisión a la universidad mejorará los índices de logro académico en la educación media, se deben conducir los estudios que lo demuestran[13].

En el caso de las pruebas de selección a la universidad, Shepard (1993)[14] afirma que se debe partir por identificar el propósito explícito de las pruebas que en el caso de las de admisión es seleccionar estudiantes que tengan posibilidades de éxito en la universidad. Esta autora plantea que, dado lo anterior, lo primordial y más importante tratándose de un argumento de validación, es que la prueba debe demostrar su poder predictivo, es decir, que se observa correlación entre el puntaje obtenido y el rendimiento universitario. Pero tal relación no debe ser la única dimensión a explorar: es necesario demostrar que el contenido de las pruebas evalúa las habilidades

[13] Véase AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999), p. 4.

[14] L. Shepard, "Evaluating Test Validity" (1993), p. 19.

que son prerrequisito para lograr éxito en la universidad. Asimismo, hay que proveer evidencia de que las pruebas no presentan sesgos que perjudiquen a algún grupo particular debido a la forma en que se pregunta. Si hay diferencias entre los individuos, éstas no pueden ser atribuibles a sesgos sino a varianza verdadera en el atributo medido.

Los procesos de validación toman tiempo. Por ejemplo, en Estados Unidos en el año 2002 se tomó la decisión de modificar el SAT I, una de las pruebas de admisión a la universidad que allí se aplican. Aunque se trata de cambios menores[15], la nueva versión entrará en vigencia el año 2005, es decir se darán un plazo prudente para realizar los estudios de validez necesarios. Otro tanto ha ocurrido con la introducción de la prueba SAT en Singapur. En Suecia los estudios experimentales de la SweSAT comenzaron en 1973 y ella se aplicó por primera vez en 1977. Los cambios marginales introducidos desde ese entonces se han tardado, en promedio, tres años. Los cambios han sido sugeridos por un Consejo Internacional. Desde 1996 se permite en ese país testear los nuevos ítems y eventuales nuevas secciones en la misma prueba de selección.

Si bien se puede reunir evidencia de validez a partir de las primeras etapas de la construcción de un instrumento, el verdadero esfuerzo en este sentido no se puede desarrollar sino hasta que el instrumento está en su forma final. La validez predictiva requiere, además, que el instrumento se aplique y opere por un tiempo para que se recojan los indicadores de la variable de criterio. Esto último introduce un dilema, ya que no se debiera aplicar instrumentos que acarrean consecuencias serias para los alumnos sin haber verificado previamente su validez, pero a la vez esta evidencia no se puede obtener sin aplicar esos instrumentos. Una de las soluciones a las cuales se recurre es la de implementar los cambios gradualmente. Los nuevos instrumentos pueden pilotearse en paralelo a los antiguos, sin aplicar las consecuencias que normalmente se asociarían a la prueba. Éste es el procedimiento que se empleó al cambiar el antiguo sistema de admisión chileno, vía Bachillerato, por el de la Prueba de Aptitud. El otro sistema es el que ha empleado el College Board en el proceso de modificación del SAT I. Allí se busca introducir secciones y contenidos nuevos a los instrumentos antiguos. Sólo en el momento que se comprueba que dichas secciones tienen igual o mayor poder predictivo que las anteriores se las reemplaza definitivamente. En este ejercicio de cambios graduales, en el que se cambian partes de la prueba y no el sistema total, se corren menos riesgos

[15] En la prueba de lenguaje se elimina la sección de analogías y se incorporan ítems de redacción previamente aplicados y probados por años en la prueba específica de redacción (SAT II Writing). En la de matemática se mantiene el acento en razonamiento y se agregan contenidos de álgebra II.

de cometer injusticias con los alumnos que se someten al proceso de selección.

Para finalizar el tema de validez, los estándares del ETS insisten en que los estudios de validación deben repetirse a lo largo de la historia de aplicación de las pruebas. Sugieren que no pasen más de cinco años entre dichos análisis.

Otro aspecto relevante a estudiar es la confiabilidad de los instrumentos. Ésta se refiere a la consistencia de las mediciones cuando ellas se repiten en la misma población o grupos. Se supone que los rasgos que se miden son relativamente estables en el tiempo, a menos que exista una intervención directa para modificarlos. Si no se ha dado esa intervención, se estima que el comportamiento entre una aplicación y otra no debe variar sustancialmente. Una de las técnicas utilizadas para comprobar la confiabilidad consiste en dividir la prueba, una vez aplicada, en dos partes equivalentes y analizar las variaciones de desempeño de la misma población en cada mitad de la prueba. En general, no se acepta más de un 10% de variación en el desempeño de ambas partes. Los estudios de confiabilidad son una consideración necesaria para legitimar una prueba, pero no reemplazan los estudios de validez antes descritos.

1.5. Garantizar la seguridad de la prueba

En las pruebas de altas consecuencias debe quedar especificado cómo se garantizará que los resultados sean confiables y no producto del acceso fraudulento a las preguntas. En este sentido, cobra importancia el cuidado con las filtraciones de preguntas. Se deben tomar los resguardos necesarios para que no se sustraigan pruebas o ítems.

En pruebas que se repiten de año en año es más fácil que las preguntas se filtren. Los interesados en darlas en el futuro pueden recoger antecedentes de las preguntas con los que las rindieron en años anteriores. Para evitarlo se hacen formas paralelas que permitan la comparación entre pruebas sin repetir las preguntas. Sin embargo, para lograr la equivalencia entre las pruebas se debe conservar un número de ítems ancla que se aplican en ambas mediciones. Estas preguntas representan una proporción menor de la prueba pero quedan vulnerables a la filtración. Otro tanto sucede con las preguntas que se testean experimentalmente en el contexto de la medición del año anterior a la prueba[16]. El mejor antídoto para este problema es crear

[16] Los alumnos que dan la prueba deben contestar preguntas que corresponderán a los ítems de mediciones posteriores. Los alumnos no reciben puntaje por estas preguntas pero ellos no lo saben. Esta política se aplica porque permite evaluar los ítems en un contexto de motivación real. El problema con este procedimiento es que los alumnos pueden ponerse de acuerdo para anotar las preguntas y entregárselas a la futura generación que será evaluada.

grandes bancos de ítems que disminuyan el incentivo de conocer preguntas específicas, dada la baja probabilidad de que la pregunta filtrada corresponda a una de las preguntas que aparecerá en la prueba.

1.6. Presentación y validación de la prueba ante el público

La etapa final de la construcción de una prueba es el estudio de cómo se explicará al público la lógica que sustenta la prueba, la interpretación de los resultados, y la evidencia de validez y equidad. La disposición a aceptar los veredictos de las pruebas se relaciona con el grado de información que tienen los usuarios acerca de la naturaleza de la prueba y de su comprensión de aquello que las sustenta. En este sentido, la información técnica debe ser transparente.

En el caso de las pruebas de admisión, un factor importante es dar a conocer a tiempo los facsímiles de la prueba. En parte, porque ayudan a la comprensión de la prueba, pero también porque los alumnos deben familiarizarse con el formato de las secciones del examen. Conocer el tipo de ítems es relevante porque los resultados pueden distorsionarse si los alumnos no están familiarizados con ellos. En ese caso, la prueba estaría evaluando la capacidad para enfrentar tareas nuevas más que las destrezas que se pretende medir, lo cual confundiría la interpretación de resultados. En este sentido, se debe cuidar que los alumnos tengan igualdad de oportunidades para practicar y familiarizarse con el formato de las pruebas. Según los estándares de la AERA, "Los examinados tienen derecho al acceso igualitario a los materiales elaborados por los patrocinadores de las pruebas que describen el contenido y propósito de la evaluación, así como al material diseñado para familiarizar y preparar a los alumnos para la prueba"[17].

1.7. Documentación del proceso

Los estándares revisados proponen que cada una de las etapas de elaboración de las pruebas quede debidamente documentada. Se considera fundamental recopilar los antecedentes de cada uno de los pasos de las etapas descritas anteriormente para facilitar la revisión y validación de las mismas, así como para proveer información a los usuarios que les facilite la interpretación de los juicios que se desprenden de los resultados de las evaluaciones[18]. El listado de jueces externos que participan en el proceso

[17] Véase Estándar 7 en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999), p. 75.

[18] Véase Estándar 6 en ibídem.

de validación de constructo y de contenido debe quedar debidamente registrado, así como los informes que ellos emiten. "Cuando el proceso de validación descansa en parte en la opinión de jueces expertos..., los procedimientos para seleccionar dichos expertos deben ser plenamente descritos. La calificación y experiencia deben acreditarse"[19].

Finalmente, los estándares de la AERA consideran que las exigencias establecidas para la confección de pruebas deben ser cumplidas antes del uso operacional de las pruebas[20].

II. ANÁLISIS DE LA PSU 2003 A LA LUZ DE LAS ETAPAS DE DESARROLLO DE UNA PRUEBA

En Chile, en el marco de los proyectos concursables FONDEF, se elaboró un nuevo sistema de pruebas de admisión a la universidad, el que supuestamente se comenzaría a aplicar a fines de 2002. Éste nuevo sistema, conocido como SIES (Sistema de Ingreso a la Educación Superior), reemplazaría a la Prueba de Aptitud Académica y a las pruebas de Conocimientos Específicos vigentes hasta entonces. Sin embargo, tras ser objeto de diversas críticas durante el año 2002, el SIES fue descartado y en su reemplazo se ha dispuesto desarrollar, y aplicar a fines de 2003, una nueva batería de pruebas denominada PSU (Pruebas de Selección Universitaria).

El objetivo de esta sección es analizar la PSU 2003 a la luz de las etapas de desarrollo que deben cumplir las pruebas de selección para la educación superior. Ahora bien, el análisis deberá limitarse a los pocos antecedentes disponibles, ya que no hay documentos públicos que expliciten la configuración de la PSU 2003. En efecto, la Comisión Técnica encargada de elaborar la PSU 2003 está entregando paulatinamente los lineamientos que se adoptarán, pero no se conocen públicamente otros documentos que contengan antecedentes más completos. Es cierto que el proyecto SIES, que es uno de los insumos de esta nueva prueba de admisión, entrega una fundamentación, pero esa fundamentación es adecuada en el contexto de un informe para concursar a fondos de investigación y no cumple con los requisitos de especificación de la etapa inicial de una prueba de estas características.

Conforme a lo anterior, primero se realizará un recorrido por las pruebas que se utilizarán como insumos para la PSU (esto es, la PAA, las pruebas de Conocimientos Específicos y el SIES), y luego se procederá a examinar la PSU 2003.

---

[19] Véase Estándar 1.7 en ibídem.

[20] Ibídem.

2.1. Prueba de Aptitud Académica (PAA)[21]

Etapa de especificación de la prueba	La Prueba de Aptitud Académica (PAA) se basó en sus inicios en el examen de admisión a las universidades norteamericanas SAT I, por lo tanto contaba con un marco definido de especificaciones. Las Pruebas de Conocimientos Específicos, más vinculadas a contenido, también siguieron el modelo de las pruebas SAT II con adaptaciones al currículo nacional.
Propósito de la prueba	El propósito de la PAA está claramente definido. Su objetivo central es la identificación de los alumnos que tienen mayores posibilidades de rendir con éxito los estudios universitarios. Se expresa en la probabilidad de egresar oportunamente y de obtener mejores notas en la universidad.
Fundamentación	Al momento de instaurar la PAA se contaba con los estudios de validación empírica realizados en Estados Unidos, los cuales avalaban a esta prueba como un buen predictor del desempeño académico de los universitarios. El constructo de aptitud académica tenía un vasto apoyo teórico y buenas definiciones operacionales de lo que se quería medir. Todo lo cual facilitaba la evaluación cualitativa del contenido de la prueba de parte de los jueces expertos. Con el tiempo, el SAT I ha acumulado evidencia de su poder predictivo y se ha modificado ligeramente la definición del constructo que subyace en la prueba. Cambió el concepto de aptitud por el de razonamiento para hacerlo más comprensible al público y quitarle la connotación de heredabilidad inmutable. El nuevo concepto de razonamiento busca sintonizar con la idea de que es una destreza que se puede adquirir lenta y transversalmente en toda la formación escolar. En Chile, la PAA se validó empíricamente antes de aplicarse. Durante cuatro años de marcha blanca se estudió su capacidad predictiva. Hasta el año 1992, los estudios de validez predictiva se

[21] Para más detalles, véase G. Donoso, M. A. Bocchieri, E. Ávila y otros, *El Sistema de Admisión: Orígenes y Evolución. Resultados del Proceso de Admisión 1999* (1999).

	hacían periódicamente. No se continuó con esta política en los años siguientes, transgrediéndose así los estándares de validación que sugieren realizar este tipo de análisis al menos cada cinco años. El último estudio formal de validez de constructo data del año 1987; sin embargo, posteriormente se hicieron análisis internos para modificar secciones de la prueba. Los últimos cambios en la definición del constructo realizados por el SAT I no han sido abordados aún por la PAA.
Definición de los dominios a evaluar	La prueba de aptitud define claramente los dominios y subdominios a evaluar. Están estipulados para cada una de las pruebas y las secciones responden a ellos.
Formato de la prueba	La prueba ha seguido un patrón común en sus evaluaciones, por lo tanto hay modelos previos para la construcción de nuevas pruebas. Periódicamente se realizan estudios para reemplazar secciones y tipos de preguntas. Éstos se comunican a través de un boletín que llega anualmente a los colegios y aparecen debidamente ejemplificados en los facsímiles que se entregan a cada postulante.
Características psicométricas	La capacidad discriminativa de los ítems de la prueba se realiza en cada aplicación con pruebas experimentales y aplicando teoría clásica de medición. A la Prueba de Aptitud Académica se le exige que discrimine en cada uno de los niveles de habilidad, buscando una distribución de los resultados cercana a una curva normal.
Especificaciones de equidad	En la PAA se conduce un análisis del comportamiento de los ítems según dependencia administrativa[22] de los establecimientos. Las preguntas que presentan las diferencias menores de desempeño entre dependencias, en dificultad, discriminación y porcentaje de omisión son las que se integran al banco de ítems que serán empleados en

[22] La dependencia administrativa se refiere al tipo de sostenedor y sistema de financiamiento de cada establecimiento. Según la dependencia administrativa, se consideran tres tipos de establecimientos: particulares pagados, particulares subvencionados y municipalizados.

	las pruebas finales. De esta manera se introduce un grado de control del sesgo cultural que pudiera contener la prueba.
Construcción de los ítems	La construcción de los ítems de la PAA se ve facilitada por la larga tradición de la prueba. Desde sus inicios se contó con un modelo en el cual apoyarse y en la actualidad el banco de preguntas es considerable. Las comisiones se coordinan anualmente por disciplinas y cuentan con un conjunto de revisores que revisan los ítems en cada una de las etapas. Los ítems se analizan de acuerdo a la Teoría Clásica de medición, y no se aplica IRT.
Estudios de confiabilidad y validez	El último estudio de validez de constructo, publicado por el Departamento de Evaluación, Medición y Registro Educacional (DEMRE), de la PAA se hizo el año 1987. En 1989 se realizan estos análisis para las Pruebas de Conocimientos Específicos. En ellos se concluye que la pruebas son confiables y miden lo que pretenden medir. Los estándares recomiendan que estos estudios se repitan cada cinco años, requerimiento que no se ha cumplido.
Validez predictiva	Los estudios de validez predictiva se han realizado a lo largo de la aplicación de la prueba y confirman la capacidad de predecir éxito académico. El último análisis de predictibilidad publicado por el DEMRE corresponde al proceso de admisión 1992. A raíz de los cuestionamientos a la prueba, se realizaron estudios de predictibilidad en distintas universidades durante el año 2002[23]. Los resultados muestran que los indicadores obtenidos son similares a los encontrados en EE.UU. para la prueba SAT I, los cuales son aceptados por los expertos como un indicador de una buena capacidad predictiva[24].

[23] Véase R. Fischer y A. Repetto, "Método de Selección y Resultados Académicos (2002). También B. Vial y R. Soto, "¿Predice la PAA el Rendimiento o el Éxito en la Universidad?" (2002).

[24] Los indicadores de predictibilidad de las pruebas SAT I y SAT II se pueden encontrar en W. J. Camara y G. Echeternacht, "The SAT I and High School Grades: Utility in Predicting Success in College" (2000), y en L. Ramist, C. Lewis y L. McCamley-Jenkins, "Using Achievement Test/Sat II: Subject Tests to Demonstrate Achievement and Predict College Grades: Sex, Language, Ethnic, and Parental Education Groups" (2001).

Estudios de sesgo	La PAA corrige los ítems por apariencia de sesgo y por comportamiento errático del ítem durante la prueba experimental. Sin embargo no hace un análisis de comportamiento diferencial de los ítems como lo hace el SAT I.
Estudio de las consecuencias de las pruebas	Una de las críticas que se le han hecho a la PAA gira en torno a las señales contraproducentes que envía hacia la educación media. Se dice que los alumnos descuidarían el estudio de las materias del currículum que no están incluidas en la prueba. No se han realizado estudios empíricos al respecto. Los estándares recomiendan su realización.
Consideraciones de seguridad	La PAA no es equiparable de un año a otro, por lo tanto, por esa vía no se pueden filtrar las preguntas porque éstas no se repiten entre aplicaciones. A lo largo de su existencia, se han filtrado preguntas por otros medios, pero se han ido creando los mecanismos para evitar este tipo de episodios.
Validación de la prueba ante el público	La Prueba de Aptitud Académica ha contado con legitimidad entre los estudiantes. El proceso de admisión se ha reconocido como transparente y confiable.

2.2. Sistema de Ingreso a la Educación Superior (SIES)

Etapa de especificación de la prueba	Se conocen dos documentos públicos sobre el SIES, el "Informe de la Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior"[25], elaborado por la Comisión del mismo nombre convocada por el Ministerio de Educación, y el proyecto FONDEF, "Reformulación de las Pruebas de Selección a la Educación Superior"[26], dirigido por David Bravo y Jorge Manzi. Ninguno de

[25] Véase Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior, "Informe de la Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior" (2000).

[26] Véase David Bravo y Jorge Manzi, "Reformulación de las Pruebas de Selección a la Educación Superior" (proyecto FONDEF) (2000).

	los dos documentos cumple los criterios de especificación que establecen los estándares para la elaboración de pruebas de altas consecuencias. El Informe analiza la problemática de las pruebas de admisión a la universidad y delinea un marco general de operación, pero éste dista de ser un marco conceptual para la formulación de una prueba porque carece del nivel de especificación necesario. Tampoco se presenta la evidencia teórica y empírica que avale las opciones escogidas. Por su parte, el proyecto FONDEF no compensa dichas falencias. En él se asevera que el diseño de las pruebas se basa en el marco orientador del Informe de la Comisión *ad hoc* convocada por el Ministerio de Educación y no se ahonda más en el tema. En el cronograma presentado en dicho informe no se le asigna tiempo a la fase preparatoria de definición del marco de especificaciones de la prueba. Aparentemente, dicha etapa se dio por finiquitada con el trabajo de la Comisión.
Propósito de la prueba	En el Informe de la Comisión se establecen una serie de objetivos que deben ser cumplidos por las nuevas pruebas. En primer lugar se busca "robustecer un sistema de selección efectivo, confiable y de alta legitimidad como el vigente, y contribuir, a través de las pruebas de tal sistema y las presiones y orientaciones que de hecho ellas establecen sobre los últimos dos años del nivel secundario, al logro de los nuevos objetivos curriculares de la educación media" (p. 4). También se busca "la producción de información estratégica para actores de diversos ámbitos, sobre el cumplimiento de los objetivos de aprendizaje tanto del sistema escolar como de la educación superior" (p. 7). En la sección de proposiciones de mejoramientos y cambios se insiste en el tema y se asevera que es necesario hacer un esfuerzo integral de adaptación de las pruebas para hacer "converger la presión de las evaluaciones del sistema de educación terciaria, con el currículum de la educación media, en forma más clara, directa y robusta que en el presente, para mejora de los resultados de la enseñanza media (EM) y de la preparación de los estudiantes de la EM"; también se plantean como necesarios "la evaluación de la educación media y los saberes y

	habilidades de los egresados" (p. 42). Se afirma que las pruebas no pueden ser consideradas como un problema sólo de la educación superior, ya que "tienen una doble función: de selección y de evaluación de los resultados formativos de la educación media" (p. 42). En este sentido, el Informe afirma que "las nuevas pruebas deben referirse a los objetivos y contenidos del currículum de la educación media, en cada una de las asignaturas que consideren. El rediseño planteado debe considerar como criterio orientador el que se trata de pruebas referidas al currículum oficial, y que, por tanto, las restricciones impuestas en los contenidos y habilidades medidas, por necesidades de discriminación de los instrumentos y bajos rendimientos actuales, deben ser levantadas" (p. 45). El Informe también plantea que de acuerdo a lo recogido por la Comisión respecto de la discusión e investigaciones internacionales sobre la materia, el giro planteado en la prueba podría lograr "consecuencias positivas sobre la equidad, al hacer disminuir el peso de las variables asociadas a lo que se describió como *coeficiente de herencia*, y acrecentar, en vez, las asociadas a la experiencia y el trabajo escolar". De la lectura del Informe se desprenden, entonces, al menos cuatro objetivos: mejorar la eficiencia en la selección de postulantes a la universidad, mejorar la equidad del ingreso a la universidad, aumentar el rendimiento de los alumnos en enseñanza media, evaluar la educación media y entregar información acerca de los logros de este ciclo. Ni el Informe ni el proyecto FONDEF se detienen a discutir cómo se podría llegar a conciliar el logro de todos estos objetivos en una sola prueba. Técnicamente no es claro que las pruebas puedan evaluar la educación media y a la vez seleccionar bien a los alumnos para la universidad. Además, es difícil pensar cómo se puedan incluir en la prueba todos los temas relevantes del currículum independientemente del nivel de logro de los alumnos. Si hay tópicos del currículum que una amplia proporción de los alumnos no dominan, las preguntas que consideran esos temas no sirven para ordenar a los alumnos, ya que nadie las puede contestar. Al no tener capacidad discriminante son inútiles al momento de evaluar.

Fundamentación	En el Informe de la Comisión se critica brevemente la noción de aptitud sobre la que se basaba la PAA. Básicamente, se cuestiona el supuesto de que las aptitudes académicas se distribuyen normalmente, es decir, que serían relativamente independientes de variables tales como sexo, edad, nivel socioeconómico y cultural. También se cuestiona la estabilidad de las aptitudes en el tiempo, la cual se refiere a la idea de que las habilidades no son modificables con el entrenamiento y la madurez de los individuos. Junto a estas críticas se asevera que el avance de la psicología cognitiva demuestra que es prácticamente imposible medir el logro de las capacidades cognitivas de los alumnos sin contemplar en tal medición los contenidos sobre los que se aplican las capacidades cognitivas. Al final de un somero análisis, realizado en dos páginas, que cita dos estudios, se concluye que "lo más razonable es abandonar la conceptualización de la aptitud y basar, en cambio, el sistema de admisión a la enseñanza superior en una evaluación que incluya la combinación de procesos cognitivos y contenidos curriculares que forman parte de la experiencia regular de los estudiantes de enseñanza media" (p. 11). En otra sección del Informe se avala con la experiencia comparada la noción de que es válido seleccionar a los alumnos para la universidad basándose en la evaluación del logro del currículum de educación media. Se mencionan los casos de Inglaterra, Francia, Alemania, Suecia, Japón e Israel citando el libro de Britton y Raitzen de 1996. Sin embargo, no se hace un análisis de la evidencia entregada. Por ejemplo, no se toma en cuenta que en la mayoría de estos países la selección de alumnos por habilidad académica ha ocurrido con anterioridad durante su educación escolar al separar a los alumnos por líneas vocacionales. Tampoco se actualizaron los datos, por lo que no se recoge la evidencia de que Suecia e Israel han incorporado pruebas tipo SAT I en la selección de alumnos. También se olvida mencionar que las pruebas de currículum de estos países son de ensayo, y por lo tanto, incluyen las habilidades de razonamiento a través de la fundamentación y organización lógica de la exposición.

	Con respecto al tema de equidad se afirma que la evidencia indicaría que las pruebas referidas al currículum son más susceptibles de ser afectadas por la experiencia escolar y que por tanto serían más equitativas. Sin embargo, no se hace referencia a que esta línea de evidencia indicaría a su vez que cuando la calidad de la educación es muy disímil los resultados de las pruebas referidas a currículum resultan más inequitativas que las referidas a razonamiento[27]. En todo caso el Informe no entrega evidencia empírica en uno u otro sentido. La evidencia y la discusión entregada en el proyecto FONDEF son escasas, por lo que se pueden considerar como un análisis preliminar del problema pero en ningún caso como un marco teórico que justifica y explicita el tipo de prueba que se elaborará. Falta una discusión más fundamentada acerca de las posibilidades que tiene una prueba referida al currículum nacional de ser un buen predictor de éxito en la educación superior; falta un análisis acerca de cuáles son los contenidos del currículum que tienen relación con las destrezas y conocimientos requeridos en la educación universitaria. Habría que contestar si todos los contenidos de la educación media están orientados hacia la universidad o si cumplen otras funciones, dado que la educación media es un fin en sí y no está concebida solamente como una etapa preparatoria para la universidad. Sería necesario demostrar que las pruebas hasta ahora aplicadas tienen menor valor predictivo que las que se pretenden implementar o que tienen igual valor predictivo pero menores consecuencias negativas secundarias. Habría que hacerse cargo de la evidencia de que cada vez más países están incorporando en sus baterías de selección, pruebas que tienen un fuerte énfasis en la evaluación de la capacidad de razonamiento, dado que los currículos han enfatizado el desarrollo de este tipo de destrezas. En mayo de 2002, Rosas, Flotts y Saragoni publicaron el artículo “Modelo de Representación del

[27] Véase H. Beyer, “Las Nuevas Pruebas de Ingreso a la Universidad” (2002), pp. 17-20.

	Conocimiento para las Nuevas Pruebas de Selección a las Universidades Chilenas". En él, los autores formulan especialmente un modelo del funcionamiento cognitivo para guiar la construcción de las preguntas del SIES. Este marco conceptual, original para estas pruebas, no se puede considerar como un referente validado, ya que la investigación en el campo de la cognición está en pleno desarrollo y discusión. No es un modelo que cuente con el apoyo de la experiencia comparada y posiblemente tampoco ha sido discutido por la comunidad de expertos en el tema. Para considerarlo como el elemento eje en la confección de preguntas requeriría de un proceso de validación.
Definición de los dominios a evaluar	La definición de los dominios a evaluar se hace en forma muy general, por lo que no satisface la exigencia de los estándares de detallar con precisión los contenidos y habilidades que se evaluarán. El Informe concluye que "Las nuevas pruebas procurarán medir competencias en áreas del currículum. El concepto de competencias alude a capacidades de desempeño en contextos simbólicos o prácticos determinados. El propósito de medición de los nuevos instrumentos no debiera centrarse exclusivamente en contenidos, ni tampoco en procesos o capacidades intelectuales de los postulantes. Las nuevas pruebas debieran orientarse a medir el desempeño de los estudiantes en situaciones problemáticas nuevas, empleando la combinación de los procesos cognitivos y contenidos que han desarrollado durante su experiencia regular en la enseñanza media, asociada directamente a su trabajo en las disciplinas del currículum". (p. 44). Esta definición no basta para elaborar las tablas de contenidos y destrezas a evaluar y menos el peso que se le asignará a cada una de éstas. En primer lugar, porque el marco curricular chileno de la educación media, en las asignaturas de lenguaje, matemática, ciencias naturales y sociales no permite una interpretación unívoca de cuáles son las conocimientos y procesos cognitivos que los alumnos deben lograr en cada uno de los dominios. Éste no explicita bien la cobertura y profundidad de cada uno de los contenidos y destreza a desarrollar.

	Probablemente, las definiciones entregadas en el Informe no darían pie para que dos equipos elaboradores de pruebas, trabajando separadamente, construyeran pruebas similares. Lo cual demostraría que la etapa de especificación de los dominios de la prueba no se habría implementado bien. Cuando este paso no se cumple satisfactoriamente, los estudios de validez de contenido se dificultan porque no quedan estipulados claramente los criterios con los cuales se tienen que refrendar las pruebas.
Formato de la prueba	La definición del formato de la prueba no aparece ni en el Informe de la Comisión ni en el proyecto FONDEF. Ciertamente esto no correspondía definirlo en el Informe ni en el proyecto; sin embargo, en documentos posteriores debieron haber quedado estipulados. Las decisiones acerca del número de preguntas por evaluación, el tipo de secciones al interior de cada una de las pruebas, no se encuentran analizadas en dichos documentos.
Características psicométricas	En el proyecto FONDEF se describen los criterios psicométricos fundamentales para la selección de los ítems, se especifican las curvas de información que se buscará en la confección de las pruebas y la capacidad discriminativa que se exigirá a los ítems. También se explicitan los análisis de sesgo que se realizarán a las preguntas. Sin embargo, el proyecto no fundamenta el tipo de decisiones que toma. No queda claro por qué las curvas de información serán tan amplias (de hecho cubrirían casi todo el espectro posible de desempeño) y cómo se cuidará que el error de medición no se eleve considerablemente por este hecho. Probablemente esta especificación quedó fijada así por la doble función que se le asignó a la prueba de seleccionar a los alumnos a la universidad y de evaluar el logro en educación media. El marco de especificación de la prueba debiera analizar más claramente este punto y dar garantías de que los alumnos quedarán confiablemente jerarquizados de acuerdo a su desempeño y dentro de rangos de error aceptables.

Especificaciones de equidad	El Informe especifica que se realizará un análisis del sesgo potencial de las preguntas debido a variables tales como tipo de dependencia, modalidad de enseñanza, sexo y región.
Construcción de los ítems	A la etapa de construcción de las preguntas se le asigna un período de tres meses en el proyecto FONDEF. Este plazo es insuficiente si se considera que esta etapa requiere conformar los equipos de redacción; entrenar a los redactores de la preguntas; crear un *pool* amplio de preguntas de las pruebas y de las maquetas en 6 asignaturas; evaluarlas por expertos en cuanto a su claridad, relevancia y pertinencia respecto a las tablas de especificaciones, fallas técnicas, gramática, apariencia de sesgo, estimación de grados de dificultad; y reescribir las preguntas necesarias y volver a revisarlas. En el proyecto FONDEF no se estipula tiempo para la etapa de entrenamiento de los redactores de las preguntas y esto representa una deficiencia importante, ya que no basta el grado de conocimiento de la disciplina, sino que los redactores deben estar compenetrados con los objetivos generales de la prueba y deben tener conocimientos de la redacción de preguntas de opción múltiple. Esto es más preocupante, si se considera que no sólo cambian los parámetros de la prueba sino que también los equipos que elaboran la prueba. De partida se integra un número mayor de profesores de educación media y disminuye el número de profesores universitarios especialistas en cada asignatura. La conformación de los equipos de expertos que revisan los ítems debiera ser transparente y sus informes quedar debidamente documentados. El proyecto FONDEF asigna alrededor diez días para esta subetapa, tiempo claramente insuficiente para realizar un análisis riguroso del conjunto de las preguntas destinadas a la prueba y a las maquetas de difusión de la prueba.
Estudios de confiabilidad y validez	
Confiabilidad	En el proyecto FONDEF se afirma que se realizarán estudios de confiabilidad, pero no se detalla qué tipo de técnicas utilizarán.

Validez	El proyecto FONDEF no contempló en su cronograma la fase de validación de contenido de la prueba: no se consideraron plazos ni recursos. No se menciona la conformación de un equipo de expertos que evalúen las pruebas ni en la etapa experimental ni después de la primera aplicación formal de la prueba. Los estándares exigen que el criterio con el cual se elige a los jueces expertos quede claramente especificado y que se describa el proceso por el cual se llega a las conclusiones finales[28]. El acopio de evidencia de validez de contenido debe darse en varias etapas de la construcción de la prueba, no basta con analizar los ítems independientes al momento de su elaboración sino que es necesario estudiar los prototipos de las pruebas completas para certificar que cumplen con la tabla de especificaciones generales, que están alineados con el propósito de la prueba y con la fundamentación teórica.
Validez predictiva	El proyecto FONDEF no menciona la posibilidad de realizar estudios de predictibilidad. Esta es una carencia severa ya que el objetivo central de una prueba de selección para la universidad es precisamente predecir el éxito en ella. Según los estándares de la AERA, si la prueba se utiliza para aseverar que los candidatos que tienen mayores puntajes se desempeñarán mejor en la universidad, los encargados de elaborar la prueba y los que la utilizan deben proveer información acerca de la asociación que existe entre los puntajes y las variables de criterio escogidas. Se estima que los estudios de regresión son apropiados y que los indicadores generales de asociación deben ser suplementados con los índices de asociación en cada uno de los rangos de puntajes de la prueba[29]. La prueba de selección utilizada hasta el año 2002 (PAA) logra indicadores de predictibilidad similares a los obtenidos en las baterías de selección en EE.UU. Las nuevas evaluaciones debieran dar garantías de que al menos logran índices similares de

[28] Véase Estándar 1.7 en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999).

[29] Véase Estándar 1.15, en ibídem.

	predictibilidad antes de entrar a reemplazarla. Al igual que en otros países, se esperaría que se condujeran estudios piloto antes de las aplicaciones definitivas para cerciorarse de que las inferencias que se harán tienen respaldo empírico. Por otra parte, los estándares exigen que se expliciten las características técnicas de los estudios de validación predictiva y que queden claramente documentados los tipos de ajustes estadísticos, por restricción de rango, realizados en las estimaciones[30]. Nada de ello aparece en el proyecto FONDEF.
Estudios de sesgo	El proyecto FONDEF especifica que en la etapa experimental y después de la primera aplicación de la prueba se realizarán los estudios de sesgo correspondientes de acuerdo a la técnica DIF.
Estudio de las consecuencias de las pruebas	El Informe de la Comisión, por su parte, asevera que las consecuencias asociadas a las pruebas de ingreso a la universidad actúan como un incentivo para el estudio y que por tanto el contenido de la evaluaciones debiera alinearse con el currículum para mejorar el rendimiento de los alumnos en la educación media. Según los estándares de la AERA, cuando explícitamente se sugiere que el uso de un test producirá determinados resultados, se debiera entregar la evidencia teórica y empírica que permita sustentar dicha afirmación[31]. En el proyecto no se señala cómo se validará empíricamente la hipótesis planteada y no se entregan argumentos sólidos acerca de cómo la nueva prueba mejorará el rendimiento de alumnos competentes que ya contaban con incentivos fuertes para el estudio, dado que las notas de educación media se consideran al momento de ingresar a la universidad.
Consideraciones de seguridad	El proyecto estima que un tercio de las preguntas de cada prueba se repetirá en el siguiente año para efectos de comparación de las pruebas. Los autores afirman que “se intentará retener alrededor de un tercio de las preguntas para hacer la

[30] Véase Estándar 1.18, en ibídem.
[31] Véase Estándar 1.23, en ibídem.

	comparabilidad interanual"[32]. Por otra parte, se ha considerado incluir también en las pruebas definitivas las preguntas experimentales de las pruebas de los años siguientes. Dado que en el país existe una industria activa de preuniversitarios y un interés grande de los establecimientos educacionales por conocer el tipo de preguntas que se harán en los años siguientes, cabe preguntarse cómo controlarán la filtración de preguntas desde los alumnos que dan las pruebas hacia estas instituciones. El proyecto FONDEF no hace referencias al respecto.
Documentación del proceso	En el proyecto no se consigna presupuesto ni tiempo para documentar el proceso de construcción de la prueba. De hecho los únicos documentos públicos conocidos son el Informe de la Comisión y el proyecto FONDEF. Una prueba que tiene consecuencias altas para más de la mitad de los alumnos de cada generación debiera hacer pública y transparente la fundamentación y procedimientos de cada uno de sus pasos[33].
Validación de la prueba ante el público	El proyecto FONDEF de Bravo y Manzi consideró la publicación de preguntas de apoyo pedagógico que difundieran la naturaleza de las pruebas. Sin embargo, en el cronograma se observa un desfase importante entre la confección de los modelos de pruebas que se darían a conocer al público y el análisis experimental de las preguntas que determinarían la extensión total de la prueba y el nivel de dificultad de la misma. Por tanto, los modelos de pruebas que se darían a conocer al público no necesariamente coincidirían con la prueba real. Esto representa un problema para la validación de la prueba ante el público porque no permite juzgar públicamente su pertinencia. Por otra parte, la falta de documentación del proceso de construcción de la prueba no ayuda a que los actores relevantes avalen la prueba. Por el contrario, crea suspicacias que le restan validez.

[32] Véase Bravo y Manzi, "Reformulación de las Pruebas de Selección a la Educación Superior" (proyecto FONDEF) (2001), p. 42.

[33] Véase Estándar 6.1 al 6.7, en AERA, APA, NCME, *Standards for Educational and Psychological Testing* (1999).

2.3. Pruebas de Selección Universitaria (PSU) 2003

¿Qué posibilidad existe de que, en el período que se ha fijado para la construcción de las nuevas pruebas de admisión, los equipos responsables puedan elaborar una prueba que cumpla con los estándares mínimos que garanticen un proceso de selección justo, válido y confiable?

Para analizar el problema es importante considerar el cronograma y el detalle de los principales hitos de la construcción de la nueva prueba. El 29 de agosto del año 2002 se anuncia que la generación que egrese de educación media el 2003 se seleccionaría con una prueba de admisión transitoria. El Consejo de Rectores decide que su construcción será coordinada por un Comité Técnico Asesor nombrado por el Consejo Directivo para las Pruebas de Selección y Actividades de Admisión a la Universidad, entidad que se constituyó el 4 de septiembre del año 2002, para monitorear el desarrollo de las pruebas. Dicho Consejo Directivo emite un comunicado el 15 de noviembre del año 2002, que define sucintamente las características de las nuevas pruebas[34]. Se acuerda que los postulantes tendrán que rendir tres pruebas. De estas tres, dos serán obligatorias: lenguaje y matemática. Para la tercera habrá que optar entre una prueba de historia y geografía y otra de ciencias. Esta última estará dividida, a su vez, en una primera sección común referida a los contenidos curriculares de biología, física y química de I y II año medio, y una segunda sección de tres módulos, en la cual los alumnos tendrán que escoger uno de ellos. Estos módulos corresponderían a los contenidos de biología, física o química de III y IV medio.

En el mismo documento se fijan además las ponderaciones mínimas para cada uno de los factores considerados en el proceso de selección. El promedio de notas de enseñanza media no puede contar menos de un 20% y las pruebas de ciencias o de historia y geografía; de lenguaje y matemática no puede contabilizarse menos de un 10%. Ponderar en mayor proporción que los mínimos señalados queda en manos exclusivas de cada universidad, de acuerdo a los propósitos de selección que ellas estimen conveniente.

El comunicado del 15 de noviembre (2002) también afirma que los contenidos curriculares de la educación media a ser considerados como referentes de las pruebas quedaron fijados por el Comité Técnico Asesor del Consejo Directivo, acogiendo la proposición concordada por la Mesa

[34] Consejo de Rectores; Consejo Directivo para las Nuevas Pruebas de Selección y Actividades de Admisión a la Universidad. “Sobre las Pruebas de Selección a la Enseñanza Universitaria 2003”, viernes 15 de noviembre de 2002.

de Trabajo de los Sostenedores, Colegio de Profesores y Ministerio de Educación.

El 17 de noviembre se publica el listado de los contenidos curriculares de enseñanza media que se tomarán como base para la construcción de la prueba transitoria. Corresponde a una transcripción precisa de los contenidos mínimos del marco curricular nacional de educación media, exceptuando aquellos contenidos que la Mesa de Trabajo consideró que no habían sido cubiertos por una proporción amplia de los alumnos que rendirían la prueba[35].

El 28 de enero de 2003 el Consejo Técnico Asesor menciona cuáles serán los ejes temáticos de cada una de las pruebas y las habilidades intelectuales que se exigirán, exceptuando la de ciencias, que queda sin mayores precisiones. Se entrega también el número total de preguntas y el tiempo de cada evaluación. Finalmente, se da a conocer el calendario de actividades hasta la fecha de aplicación de la prueba. En la calendarización se incluyen los pasos que involucran a los alumnos y universidades, pero no se incluyen las etapas técnicas del desarrollo de las pruebas.

En marzo de 2003 se publica un folleto con los contenidos, tabla de especificaciones, pero sin los pesos asignados a cada sección, y un número reducido de preguntas de las nuevas pruebas. En dichos folletos se entrega una descripción breve de la naturaleza de la prueba. En esa misma fecha, las universidades del Consejo de Rectores publican la nómina de carreras con las pruebas optativas y las ponderaciones finales que se pedirán en cada caso.

Entre abril y mayo de 2003 se difundirán los folletos de cada una de las pruebas con ejemplos de cada sección. Los documentos oficiales dejan entrever que se tratará de un muestreo de preguntas más que facsímiles de pruebas completas. La fecha de rendición de las pruebas queda fijada para mediados de diciembre. Este cronograma supone que las pruebas quedarán prácticamente definidas entre el 4 de septiembre de 2002, fecha en que se conforma el Comité Asesor, y mayo de 2003, en que se dan a conocer los facsímiles.

En declaraciones no oficiales, se han ido estipulando aspectos que no aparecen definidos en los comunicados anteriores. La presidenta del Comité Técnico Asesor afirma que aun cuando se deben rendir obligatoriamente tres de las cuatro pruebas, los alumnos podrían darlas todas. En caso de que las carreras exigieran indistintamente las pruebas de ciencias o la de historia y geografía, el estudiante podrá postular con la que obtuvo mejor puntaje. En relación de la prueba de ciencias, las autoridades universitarias

[35] Referente Curricular de Pruebas de Selección Universitaria. www.mineduc.cl

establecieron que la prueba tendrá un puntaje único pese a contar con dos partes, una de las cuales no es la misma para todos los individuos. Es decir el puntaje final de la prueba de ciencias no distinguirá si el alumno ha rendido la segunda parte en biología, física o química. Esto implica que un estudiante, por ejemplo, podrá postular a medicina y a ingeniería dando la prueba común de ciencias con la optativa de biología. Por otra parte, aclaran que el margen de libertad que se le entrega a cada universidad para ponderar las distintas pruebas debe emplearse dentro de las cuatro pruebas establecidas, lo cual excluye la posibilidad de considerar pruebas específicas anexas en la ponderación total.

Hasta aquí la información oficial disponible. Un análisis de la viabilidad de la nueva prueba necesariamente pasa por el campo de las suposiciones, ya que los datos que se han dado a conocer son de carácter amplio y no incluyen las descripciones técnicas de la construcción de los nuevos instrumentos.

A continuación se explorará el escenario de las Pruebas de Selección Universitaria (PSU) 2003, basándose en el supuesto de que éstas combinan elementos de la Prueba de Aptitud Académica (PAA), de las Pruebas de Conocimientos Específicos y de los módulos experimentales que posiblemente ha elaborado y testeado previamente el DEMRE[36] en el contexto de investigaciones internas para el mejoramiento de las pruebas de admisión y secciones de las pruebas realizadas en el marco del proyecto SIES.

Se descartan otros escenarios dada la limitación de tiempo que impuso el Consejo de Rectores. La posibilidad de incluir secciones de pruebas de admisión de probada calidad que no han sido testeadas en Chile quedó seriamente limitada, al menos para la evaluación 2003. Por ejemplo, no se podría contemplar la adaptación de secciones del ACT de ciencias[37], aun cuando pedagógicamente es una prueba que recoge lo crucial que deben aprender los alumnos en educación media y constituye un buen indicador de éxito en la universidad. En efecto, el ACT de ciencias podría ser un modelo a seguir en esa área, ya que el tipo de preguntas que emplea permite evaluar razonamiento científico y aplicación de conceptos centrales, sin recurrir a la memorización innecesaria de definiciones mecánicas[38]. Sin

[36] DEMRE (Departamento de Evaluación, Medición y Registro Educacional), institución encargada del proceso de admisión a la universidad, dependiente de la Universidad de Chile.

[37] El ACT (American College Testing Program) es una de las pruebas de selección universitarias de Estados Unidos. En ese país, es la segunda en importancia según el número de alumnos que la rinde.

[38] Este punto quedó destacado en el Informe de la Comisión de Ciencias del CEP, presidida por Sergio Hojman, en octubre de 2002. Véase Comisión de Ciencias, "Prueba de Ciencias, Críticas y Propuestas" (2002).

embargo, dado el cronograma que se ha fijado, ya es tarde para intentar adaptar secciones del ACT, porque en menos de un año no se podría entrenar a los redactores de ítems en el nuevo formato, adaptar los ítems a los contenidos del currículum chileno, evaluar empíricamente las preguntas para estudiar su capacidad discriminativa y su comportamiento psicométrico, analizar con jueces expertos la validez de contenido y finalmente conducir los estudios de validez predictiva. En este plazo, también resulta imposible familiarizar a los profesores y alumnos con el estilo de las preguntas.

El mismo fenómeno ocurre con la posibilidad de incluir secciones nuevas desarrolladas en el país, aunque en este caso la situación se agrava aún más. Las innovaciones requieren de la elaboración y validación de un marco que avale teórica y empíricamente la decisión de incluir un tipo de pregunta. Cuando se adaptan pruebas, sólo debe estudiarse la validez de generalizar el constructo a otros contextos, lo cual requiere menos tiempo.

En suma, a continuación se analizará la PSU (Pruebas de Selección Universitaria) 2003, suponiendo que ella quedará conformada por una mezcla de preguntas extraídas de la Prueba de Aptitud Académica, de las Pruebas de Conocimientos Específicos, del SIES y por nuevas preguntas elaboradas en el contexto de las investigaciones del DEMRE.

(Pruebas de Selección Universitaria (PSU) 2003)

Etapa de especificación de la prueba	Las definiciones que debe realizar el Comité Técnico Asesor están fuertemente condicionadas por las restricciones técnicas que impone un calendario tan ajustado como el que se le ha impuesto a la prueba. Más que elegir las características idóneas de cada prueba, el Comité tiene que escoger de un menú de secciones y preguntas que han sido testeadas anteriormente. El Comité también parte con un pie forzado, pues el número de pruebas ha quedado limitado *a priori*, a pesar de la evidencia empírica que destaca el valor de tener una combinación de pruebas generales y específicas. En efecto, el mandato del Consejo de Rectores impide agregar pruebas a las tres obligatorias, con lo cual se elimina la posibilidad de incorporar pruebas específicas que han demostrado un alto valor predictivo[39].

[39] Véanse Vial y Soto, "¿Predice la PAA el Rendimiento o el Éxito en la Universidad? (2002), y Fischer y Repetto, "Método de Selección y Resultados Académicos" (2002).

	Por otra parte, la prescripción de utilizar el currículum de educación media como referente de la prueba también podría representar una cortapisa para la construcción de la prueba, como se explicará a continuación.
Propósito de la prueba	El Comité Técnico a cargo de las nuevas pruebas asegura que su objetivo principal es seleccionar a los mejores estudiantes para la educación superior y que no es una meta evaluar la educación media[40]. En el documento de difusión de los contenidos, tabla de especificaciones y muestreo de preguntas aparecido en marzo de 2003, se enfatiza que la selección es uno de los objetivos principales de la prueba. Desde luego, allí se bautiza la nueva prueba con el título de "Prueba de Selección a la Universidad" y se asevera que la nueva prueba "responde a los requerimientos de una mayor alineación con los Programas de Estudio vigentes en la Enseñanza Media de cada uno de los tests que la componen, sin perder su característica principal de ser pruebas de selección"[41]. Sin embargo, este sucinto documento no alcanza a constituirse en el marco de fundamentación de la prueba, ya que no entra en detalles, delimitaciones de conceptos ni provee la evidencia teórica y científica que avale el contenido y la orientación del instrumento. Por lo tanto, no debiera tomarse como la última palabra. En cuanto a los propósitos de la prueba y los constructos teóricos que guiarán su elaboración, el documento del 15 de noviembre del Consejo de Rectores no se pronuncia al respecto. Solamente estipula que los contenidos de la educación media deben ser utilizados como referentes de la prueba. Esta afirmación da pie a distintas interpretaciones. Por una parte, utilizar los contenidos de educación media como referente podría significar que los evaluadores pueden escoger, entre todos los contenidos del ciclo, aquellos cuyo dominio es el indicador de un buen desempeño en la universidad.

[40] *El Mercurio*, miércoles 29 de enero 2003.

[41] Consejo de Rectores; Departamento de Evaluación, Medición y Registro Educacional, "Pruebas de Selección Universitaria: Proceso de Admisión 2004", *La Nación*, marzo 2003.

	O bien que, al construir pruebas centradas en la evaluación de destrezas y habilidades requeridas para cursar estudios superiores, los evaluadores tendrán que utilizar los contenidos del currículum de educación media que mejor se presten para ello. Estas dos interpretaciones avalarían la posición de que las pruebas de admisión 2003 tienen como propósito fundamental seleccionar a los estudiantes que tienen mayores posibilidades de éxito en la universidad. En cambio, se puede desdibujar el objetivo de seleccionar a los que tienen mayores probabilidades de éxito en la universidad si utilizar el currículum como referente significa un vuelco a priori hacia pruebas centradas en el dominio del currículum, es decir a la evaluación aleatoria de sus contenidos. El coordinador nacional de la Unidad de Currículo y Evaluación, del Ministerio de Educación, miembro del Consejo Técnico Asesor, asevera que "la esencia del cambio de las pruebas de selección, decidido por el Consejo de Rectores, es el giro de las mismas hacia una mayor cobertura curricular. La más importante ventaja de las pruebas de selección basadas en el currículo consiste en un incentivo claro y fuerte para el trabajo de alumnos y profesores en la enseñanza media, valorizando el quehacer de ésta y su profesorado, y produciendo egresados mejor preparados"[42]. La pregunta es si todos los contenidos de la educación media actual convergen con los requerimientos de la educación universitaria y si esto debiera ser así, dado que este ciclo puede ser un fin en sí mismo y no una preparación para la universidad. Estos planos deben aclararse, ya que al evaluar la validez de contenido de la prueba, los jueces tendrán que decidir si la estructura de la batería de selección y las preguntas son fieles al objetivo de seleccionar a los alumnos a la universidad o si, además de seleccionar, logran ser un fiel referente de los programas educativos, con lo cual se acercarían al objetivo de ser un incentivo claro y fuerte para el dominio del currículum de la educación media.

[42] C. Cox, "Nuevas Pruebas de Selección Universitaria: En Aguas Más Calmas" (2002).

Fundamentación	Si se emplea una combinación de secciones que provienen de distintas pruebas, el marco conceptual debiera fundamentar por qué esa amalgama particular responderá al propósito de la prueba. El marco de la prueba debe definir claramente qué hay detrás del conjunto de preguntas que se emplearán. Luego, debe aportar evidencia teórica que indique por qué la batería de pruebas escogidas y el tipo de dominios y habilidades a evaluar predecirán el éxito en la universidad. Junto con ello, en la etapa inicial del desarrollo de una prueba se debe aportar evidencia empírica que demuestre por la vía de la experiencia comparada que esto será así. En el caso de estas pruebas, en que el marco teórico se tendrá que construir post hoc a las decisiones que tomó el Consejo de Rectores, será especialmente difícil demostrar, por ejemplo, qué buenos resultados en la prueba de ciencias con el módulo de biología resultará predictiva del éxito en la carrera de ingeniería civil. Asimismo, debe fundamentarse por qué el dominio de ciertos contenidos puntuales del currículum de lenguaje incluidos en la prueba constituirán un indicador de éxito en la mayoría de las carreras[43]. Lo mismo ocurre con los contenidos de la prueba de matemática. Aquí debe quedar justificado cuáles contenidos y destrezas son relevantes de evaluar y por qué se evaluarán de una manera u otra. Debe quedar claramente descrito cuál es la manera de preguntar que respeta el constructo de la prueba. A modo de ejemplo, para predecir éxito en las carreras matemáticas podría ser importante la evaluación de los teoremas y posiblemente es más predictivo exigir su aplicación que su memorización. En este acápite se debiera dar cuenta también de las razones por las cuales se eliminaron las Pruebas Específicas y justificar con evidencia que las pruebas que las reemplazarán aportarán más a la predictibilidad.

[43] En la prueba de lenguaje se incluyen preguntas relacionadas con literatura. Esto exige una fundamentación distinta a la que comúnmente se da cuando se aplican pruebas verbales generales, ya que normalmente las secciones de literatura se incluyen en las pruebas específicas de lenguaje, las cuales sólo se utilizan para seleccionar alumnos para carreras afines.

	Por el momento no se cuenta con ningún documento que reúna las condiciones anteriores.
Definición de los dominios a evaluar	A partir de las "Tablas de Especificaciones de las Pruebas de Selección Universitaria 2003" y del "Temario de las Pruebas de Selección Universitaria 2003"[44], dos equipos paralelos, trabajando independientemente, no podrían construir pruebas similares, tal como lo sugieren los estándares. En primer lugar, aún no se ha definido el peso que se le asignará a cada habilidad y eje temático. Pero si dejamos este aspecto de lado, tampoco se podría lograr este objetivo. El temario se refiere a los contenidos del marco curricular de educación media, y éste no precisa bien la profundidad con que se debe tratar cada uno de los tópicos allí presentados[45]. El marco enuncia los temas pero no acota el nivel de detalles que comprende. En el caso de la prueba de lenguaje, las incógnitas son mayores, pues nunca antes se había evaluado el conocimiento de "unidades conceptuales relevantes" de la asignatura y tampoco aspectos de la teoría del discurso. Por lo tanto, los encargados de la prueba debieran definir mejor los objetivos de logro para cada contenido y delimitar claramente el alcance de cada uno de ellos. Sólo de este modo los alumnos sabrán a qué atenerse al momento de estudiar la prueba. No basta con el currículo nacional, porque éste expone los contenidos a grandes rasgos, lo cual obedece al mandato de la Ley Orgánica Constitucional de Educación de dar la libertad a cada establecimiento para elaborar sus propios planes y programas. Por esta misma razón, la concreción del currículo en los planes y programas de estudio del Ministerio no puede servir de guía porque no tiene carácter obligatorio para todos los establecimientos. Otro tanto sucede con los textos de estudio, ya que éstos no son únicos ni obligatorios.

[44] Consejo de Rectores, Documento N°2 y N°3, enero 2003.

[45] Se podría argumentar que se daba el mismo caso en las Pruebas de Conocimientos Específicos. Sin embargo, la historia acumulada de las pruebas, recogida en los facsímiles oficiales, lograba acotar claramente, vía ejemplo, la profundidad y alcance de cada uno de los contenidos a evaluar.

Formato de la prueba	El número de preguntas y tiempo de duración de la prueba quedó establecido el 23 de enero, posiblemente antes de definir el tipo de ítems que se emplearía. Es problable que esto se haya hecho para despejar la incertidumbre de los estudiantes ante la vecindad de la prueba. Normalmente esto no sucede así, ya que el número de preguntas responde a una conjunción de factores que se ponderan y se van afinando durante el proceso de elaboración de los ítems. Desde luego, la definición del constructo de la prueba debe anteceder a la determinación del tiempo total de la prueba y al número de preguntas. Por ejemplo, si las preguntas requieren razonamiento avanzado, normalmente exigirán más tiempo para contestarlas, lo cual resulta determinante para el número final de preguntas. En el documento de marzo de 2003 se da a conocer el formato de las preguntas y se espera que el peso de cada sección y el modelo definitivo de las pruebas quede definido entre abril y mayo.
Características psicométricas	Una de las primeras decisiones que deben adoptarse es si el cálculo de los puntajes de los alumnos se hará a través de la Teoría Clásica (TC) o de la Teoría de Respuesta al Ítem (IRT). Si bien ambos análisis son complementarios, tienen fuertes diferencias a la hora de determinar el puntaje: en Teoría Clásica, cada pregunta vale lo mismo, en IRT se ponderan por su nivel de dificultad. Para los alumnos es importante conocer y entender la forma en que se obtendrá su puntaje. Un estudiante que se prepara para la prueba punteada por IRT debe considerar que el cálculo de los puntos no es lineal. Por lo tanto, su puntaje total no corresponderá a la suma de repuestas correctas (menos las incorrectas) sino que es un puntaje en el que un mismo número de respuestas correctas puede representar un puntaje distinto, dependiendo del nivel de dificultad de las mismas. También tendrá que informarse que no se descuentan preguntas como se hace con la Teoría Clásica. La decisión de utilizar IRT debe comunicarse junto con la publicación de los facsímiles. En el caso de la prueba

	de admisión 2003, ésta es una decisión que estará condicionada por la factibilidad de aplicar IRT[46]. Una prueba de selección a la universidad requiere características psicométricas especiales. Debe permitir una jerarquización fina de los estudiantes en los segmentos de corte de cada carrera, para evitar márgenes de error significativos. Como en Chile los puntos de corte son distintos en cada una de las carreras y universidades, se esperaría que las pruebas discriminaran bien en un amplio rango de habilidades. Al menos de la media hacia arriba (niveles de habilidades sobre el promedio) que es el tramo del que se elegirán a los estudiantes. Todo esto implica que la especificación de la distribución de los ítems, según nivel de dificultad (si se emplea Teoría Clásica), o la curva de información (si se emplea Teoría de Respuesta al Ítem), debiera concentrarse en el rango superior de competencias. Si bien es posible especificarlo en el plano teórico, será difícil utilizar esta especificación para construir la prueba. La principal razón se debe a que se ensamblará una prueba a partir de secciones, testeadas experimentalmente en contextos diferentes, sin un diseño que permita una comparación de los resultados. Así, los niveles de dificultad (tanto en Teoría Clásica como IRT) no serán comparables. Los constructores de la nueva prueba requerirán realizar supuestos arriesgados para poder asumir equivalencia. Este mismo problema se enfrenta con los demás descriptores de los ítems: nivel de discriminación, y en el caso del modelo IRT usado en el proyecto FONDEF se debe agregar el que corresponde "a la adivinación"; niveles de homogeneidad de los ítems si se trata de la teoría clásica o unidimensionalidad en el caso de IRT. En la batería de pruebas anterior se contaba con tests generales y específicos. Estos últimos facilitaban la tarea de discriminación fina en el segmento superior de desempeño. Con la eliminación de ellos en la nue-

[46] La posibilidad de usar IRT en las nuevas pruebas sólo se podrá definir tras la aplicación definitiva de la PSU 2003. No sería factible decidirlo antes porque la nueva prueba posiblemente se habrá construido a partir de dos pruebas que siguen modelos distintos. Si la prueba PSU rendida en diciembre de 2003 se comportara de acuerdo a las especificaciones previamente definidas, entonces se podrían calcular los puntajes con IRT. Sin embargo, esto sería éticamente reprobable porque los alumnos no conocerán las reglas del juego sino hasta después de rendida la prueba.

	va prueba, se cuenta con menos preguntas. Las especificaciones debieran definir cómo compensarán esta pérdida de información. La comparabilidad interanual de las pruebas no debiera buscarse en esta etapa porque condicionaría las futuras evaluaciones. Estas tendrían que seguir un formato similar a la prueba 2003, lo cual no sería recomendable dadas las particulares circunstancias de esta prueba. Además se ha reconocido que es una prueba de transición. Por lo tanto, a nuestro juicio, no debieran especificarse los requerimientos de *equating* para el instrumento del 2003.
Construcción de los ítems	Los ítems fueron construidos con anterioridad a la especificación de las pruebas. Sin embargo, para validar su utilización en una prueba con especificaciones distintas a las cuales les dieron origen, debieran someterse al juicio de los expertos para verificar su adecuación a los nuevos requerimientos. Procedimiento que debiera quedar debidamente documentado. Este trabajo correspondería a la etapa de revisión inicial de los ítems y a la de evaluación preliminar. El estudio piloto de las preguntas ensambladas en versiones similares a las finales probablemente no podrá realizarse a menos que se haga una aplicación de campo durante el año 2003. Hecho que no se ha anunciado. Este estudio es el que permite determinar con un alto grado de precisión el comportamiento psicométrico de las preguntas. Por lo tanto, si no se realiza, la conformación de las pruebas finales se tendrá que hacer con datos aproximados obtenidos de los estudios piloto que se hicieron en forma independiente para cada prueba original de donde se sacaron los ítems para esta nueva versión. Como se dijo, los criterios técnicos indican que los parámetros de las preguntas pueden variar cuando las preguntas se cambian de contexto, lo cual indicaría que los datos obtenidos en las condiciones originales no son directamente extrapolables en versiones distintas. Esto implica que los indicadores de dificultad y parámetros psicométricos de las preguntas pueden variar al momento de la aplicación definitiva de la prueba. En el peor de los casos las pruebas po-

	drían no discriminar bien en algunos segmentos. Por ejemplo, como se estimaron los niveles de dificultad antes de que los alumnos se prepararan para las pruebas, podría suceder que una vez que éstos estudien, las preguntas que consideraban difíciles ya no lo sean más. Con ello existiría el riesgo de que una proporción alta de alumnos se agolpara en los puntajes superiores.
Estudios de confiabilidad y validez	
Estudios de confiabilidad	Si la nueva prueba mezcla ítems procedentes de tests de naturaleza distinta, probados en condiciones diferentes, es difícil que se puedan realizar estudios de confiabilidad interna de la prueba completa antes de su aplicación definitiva. En este sentido no se cumpliría con los estándares de la AERA que exigen que las pruebas cuenten con un margen de confiabilidad al aplicarse.
Validez de constructo y de contenido	Antes de la aplicación de la prueba debiera darse a conocer la nómina de expertos que revisarán las pruebas y los informes que emitan. Previo a esto, tendría que formularse claramente la fundamentación de la prueba para que los jueces pudieran tener los criterios definidos para emitir sus juicios. La tarea de este grupo será muy difícil porque en el caso de que su evaluación sea negativa, hay poco tiempo para hacer correcciones. ¿Qué comisión se atrevería a invalidar una prueba? ¿Hay una prueba alternativa en caso de que esta no fuera viable?
Validez predictiva	Dados los tiempos fijados, la nueva prueba se aplicará sin evidencia empírica de su validez predictiva. Para tenerla se requieren al menos dos años desde una aplicación piloto. Como afirma el estudio de Fisher y Repetto (2002), se reemplazará una prueba que tiene validez predictiva por una que no ha demostrado tenerla.
Estudios de sesgo	Los estudios de apariencia de sesgo de los ítems pueden realizarse durante el análisis preliminar de los ítems. Sin embargo, probablemente la aplicación del análisis de DIF se podrá realizar sola-

	mente a las pruebas originales de donde se sacarán los ítems, ya que los estudios experimentales se realizaron por separado para cada una de ellas, con muestras independientes. El análisis de sesgo de la prueba total sólo se podrá conducir una vez que esta se aplique definitivamente, lo que no es el ideal.
Estudio de las consecuencias de las pruebas	Si la nueva prueba afirma que los alumnos de educación media aumentarán su dedicación al estudio y con ello mejorará el rendimiento académico en este ciclo, habrá que recoger evidencia empírica al respecto una vez aplicados los tests. Ésta es una tarea que tendría que demostrar que hay un aumento del rendimiento en un período determinado y que éste se puede atribuir a la aplicación de las pruebas y no a otros factores. Por otra parte, también hay que recoger evidencia de que la prueba no distorsionará el estudio del currículum en este ciclo. Por ejemplo, si se pregunta en la prueba por conceptos literarios, habrá que investigar si este hecho conduce a que se deje de leer el mínimo de seis obras literarias por año para dar espacio a la ejercitación descontextualizada de dichos conceptos.
Consideraciones de seguridad	Si la nueva prueba no se diseña para hacer sus puntajes comparables de un año a otro y no se experimentan los ítems de la prueba del 2004 en la evaluación del 2003, las condiciones de seguridad no tendrían por qué variar respecto de años anteriores. En este sentido el DEMRE, el organismo que administrará las pruebas, ha dado garantías de seguridad en las últimas evaluaciones.
Validación de la prueba ante el público	Dada la precariedad de las nuevas pruebas, su validación ante el público será difícil. La claridad y transparencia serán cruciales para ganar el favor del público. En este sentido, la publicación de modelos completos de las pruebas y un análisis favorable de las mismas por jueces externos al circuito del Consejo de Rectores ayudarían a atenuar la desconfianza inicial. Sin embargo, un test con efectos retroactivos como éste, difícilmente será percibido como justo por alumnos y profesores. La reforma curricular

	sobre la cual se basa la prueba se ha prestado para distintas interpretaciones lícitas; esta nueva prueba favorecerá necesariamente una línea y no otras. Lo más equitativo habría sido aplicar la prueba una vez que los estudiantes de I año medio, en antecedente de la misma, terminaran el IV medio.

Claramente, el procedimiento seguido en la construcción de esta nueva prueba de admisión no ha sido el regular. Del análisis precedente se puede concluir que las posibilidades de cumplir con los estándares requeridos para la construcción de pruebas de altas consecuencias antes de su aplicación son mínimas.

Es probable que en países con una tradición más vasta en la construcción de pruebas, con una masa crítica de evaluadores formados que pueden ejercer el rol de contraparte en el proceso de la elaboración de los tests, no se habría aceptado la puesta en marcha de la Prueba de Selección Universitaria 2003 para el proceso de admisión 2004.

Sin embargo, dado que la decisión de aplicarla está tomada y que echar pie atrás induciría a una confusión mayor, los organismos competentes debieran intentar minimizar las carencias. En primer lugar, avanzar en la definición de la fundamentación de la prueba para permitir que jueces expertos validen los contenidos de la prueba. Luego se puede proceder a demostrar que ciertas secciones de la prueba tienen un marco de validación mínimo, por la vía de la evidencia comparada con otras pruebas que las contienen y que han sido validadas empíricamente. En tercer lugar, comprometer y calendarizar los futuros estudios de validación que se realizarán una vez que se apliquen las pruebas y detallar los procedimientos que se seguirán si éstos no son favorables. En cuarto lugar, consignar los procedimientos seguidos en la confección de las pruebas para facilitar los estudios de validez y asegurar transparencia total. Por último, es fundamental que los responsables garanticen que en el futuro nuestras pruebas se alinearán con los estándares internacionales.

La Prueba de Admisión 2003 no debiera constituirse en un marco forzado para las evaluaciones de los años subsiguientes. Sólo debiera ser un antecedente más en el estudio de la batería idónea para seleccionar a los mejores alumnos para la universidad. En este sentido, habría que reabrir el debate y los estudios acerca de la conveniencia de eliminar las pruebas específicas, investigar la utilidad de adaptar nuevas pruebas o secciones de pruebas extranjeras, buscar evidencia empírica que avalara o revocara la decisión de exigir pruebas de asignaturas no afines a las carreras o líneas de

estudio. Y por qué no, también se debiera explorar la posibilidad de contratar los estudios de validación a instituciones de prestigio como el Educational Testing Service (ETS), o investigar si sería mejor volver a la batería de pruebas anteriores incorporando las modificaciones que se han hecho a las pruebas equivalentes en EE.UU. Abrirse a estas posibilidades permitirá contar con pruebas justas, confiables y pertinentes.

BIBLIOGRAFÍA

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

Barret, P. "How to Review a Psychological Test Instrument. University of Liverpool", 1999. www.liv.ac.uk/pbarret/paulhome.htm.

Beyer, H. "Las Nuevas Pruebas de Ingreso a la Universidad". *Administración y Economía*, 48, Revista de la Facultad de Ciencias Económicas y Administrativas, PUC, 2002.

Bravo, I. y J. Manzi. "Reformulación de las Pruebas de Selección a la Educación Superior". Proyecto FONDEF (Octavo Concurso Nacional de Proyectos de Investigación y Desarrollo), 2000.

Camara, W. J. y G. Echeternacht. "The SAT I and High School Grades: Utility in Predicting Success in College". Research Notes RN-10, The College Board, julio 2000.

Comisión de Ciencias del CEP. "Prueba de Ciencias, Críticas y Propuestas". *Estudios Públicos* 88 (2002).

Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior. "Informe de la Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior". Santiago, noviembre 2000.

Consejo de Rectores; Consejo Directivo para las Nuevas Pruebas de Selección y Actividades de Admisión a la Universidad. "Sobre las Pruebas de Selección a la Enseñanza Universitaria 2003". Viernes 15 de noviembre de 2002.

Consejo de Rectores. Documentos N° 1, N° 2 y N° 3. Enero 2003. www.consejoderectores.cl

Consejo de Rectores; Departamento de Evaluación, Medición y Registro Educacional. "Pruebas de Selección Universitaria: Proceso de Admisión 2004". Diario *La Nación*, marzo, 2003.

Cox, C. "Nuevas Pruebas de Selección Universitaria: En Aguas Más Calmas". *Revista de Educación*, 301. Ministerio de Educación, diciembre 2002.

Departamento de Evaluación, Medición y Registro Educacional (DEMRE). "Calendario Básico de Trabajo y Publicaciones". Rendición de Pruebas de Selección Universitaria 2003. Documento N° 1. Universidad de Chile. Enero 2003.

Departamento de Evaluación, Medición y Registro Educacional (DEMRE). "Tabla de Especificaciones de las Pruebas Universitaria 2003". Documento N° 2. Universidad de Chile. Enero 2003.

Departamento de Evaluación, Medición y Registro Educacional (DEMRE). "Temario de las Pruebas de Selección Universitaria 2003". Documento N° 3. Universidad de Chile. Enero 2003.

Donoso, G.; M. A. Bocchieri, E. Ávila y otros. *El Sistema de Admisión: Orígenes y Evolución. Resultados del Proceso de Admisión 1999*. Universidad de Chile, DEMRE. 1999.

Educational Testing Service (ETS). *Standards for Quality and Fairness.* ETS, 2000.

Fischer, R. y A. Repetto. "Método de Selección y Resultados Académicos", Escuela de Ingeniería, Universidad de Chile". Universidad de Chile, noviembre 2002.

Hambleton, R. K.; H. Swaminathan y H. J. Rogers. *Fundamentals of Item Response Theory.* Newbury Park: Sage. 1991.

Le Foulon, C. "Reformulación del Sistema de Selección a la Educación Superior. Antecedentes para la Discusión". Centro de Estudios Públicos, *Serie Documentos de Trabajo*, 334, 2002.

Le Foulon, C. y F. Dussaillant. "Desarrollo de Pruebas Estandarizadas". Mimeo, Centro de Estudios Públicos. 2002.

Ramist, L.; C. Lewis y L. McCamley-Jenkins. "Using Achievement Test/Sat II: Subject Tests to Demonstrate Achievement and Predict College Grades: Sex, Language, Ethnic, and Parental Education Groups". College Board Research Report N° 2001-5, 2001.

Referente Curricular de Pruebas de Selección Universitaria. www.mineduc.cl

Shepard, L. "Evaluating test validity". *Review of Research in Education*, 19. 1993.

Rosas, R.; M. P. Flotts, y C. Saragoni. "Modelo de Representación del Conocimiento para las Nuevas Pruebas de Selección para el Ingreso a las Universidades Chilenas", *Psykhe*, vol. 11 N° 1, Pontificia Universidad Católica de Chile, 2002.

Tinkelman, S. "Planning the Objective Test". En Thorndike, R. L. (ed). *Educational Measurement, American Council in Education*, Washington, 1971.

Vial, B. y R. Soto. "¿Predice la PAA el Rendimiento o el Éxito en la Universidad?", *Revista de Administración y Economía*, P. Universidad Católica de Chile. 48 (2002). □