Avances en Medición, 3, 153-172
2005

ANALISIS PSICOMETRICO DE LOS EXAMENES DE EVALUACION DE LA CALIDAD DE LA EDUCACION SUPERIOR (ECAES) EN COLOMBIA [1]

Olga Rodríguez*, Pedro Pablo Casas & Yohana Medina
Universidad Nacional de Colombia, Colombia

Resumen

Este trabajo presenta el análisis psicométrico de los Exámenes de Evaluación de la Calidad de la Educación Superior (ECAES) diseñados y aplicados durante el segundo semestre del año 2003 en Colombia. El análisis se adelantó mediante los métodos tradicionales de la Teoría Clásica de los Test (TCT) y del ajuste de Modelos de Rasch. Se encontró que según los criterios de dificultad, discriminación y análisis de distractores de la TCT el porcentaje de ítems que cumplen con los criterios de calidad previamente establecido no superó el 30%; la discriminación fue el principal parámetro de rechazo de ítems; además, un porcentaje alto de ítems requiere ser revisado dado que sus distractores no funcionan adecuadamente. Los análisis de las pruebas como un todo sugieren niveles satisfactorios de consistencia interna y de validez de contenido. El análisis con el modelo de Rasch indica que los ítems de las 27 carreras se ajustan al modelo y presentan diferencias en cuanto al nivel y distribución del parámetro de dificultad de sus ítems.

Palabras clave: *Exámenes de Evaluación de la Calidad de la Educación Superior, Teoría Clásica de los Test, Modelo de Rasch, Evaluación Educativa*.

Abstract

This paper presents the psychometric analysis of the Superior Education Quality Assessment Tests (ECAES) designed and applied during 2003 in Colombia. For the analysis both Classical Test Theory (CTT) and Item Response Theory (Rasch Model) were used. The results of item analysis with CTT showed that proper quality did not exceed 30%, where discrimination was the principal parameter criterion of item removal, additionally, a high percentage of items require revision since its options do not function properly. The tests showed satisfactory levels of internal consistency and content validity. The analysis with Rasch model indicated that the items from the 27 careers adjusted to the model and present differences on their difficulty parameter level and distribution.

Key words: *Superior Education Quality Assessment Tests, Clasical Test Theory, Rasch Models, Educational Measurement.*

Introducción

En el ámbito educativo la evaluación hace referencia a la "acción permanente por medio de la cual se busca apreciar, estimar y emitir juicios sobre procesos de desarrollo del alumno o sobre los procesos pedagógicos o administrativos, así como sobre sus resultados con el fin de elevar y mantener la calidad de los mismos" (Ministerio de Educación Nacional-MEN, 1997, p.17). Esta definición hace énfasis en el carácter permanente de la evaluación y en ese sentido

[1] Los autores agradecen al doctor Daniel Bogoya, director del Instituto Colombiano para el Fomento de la Educación Superior, ICFES, permitir el acceso a las bases de datos utilizadas en el presente trabajo.

* Laboratorio de Psicometría. Departamento de Psicología. Universidad Nacional de Colombia. Ciudad Universitaria. Bogotá – Colombia. e-mail: orrodriguez@unal.edu.co

la vincula como acción formativa, según la cual es posible realizar la valoración de un proceso educativo que se está llevando a cabo. Pero la evaluación educativa tiene varios propósitos entre los cuales se encuentra el monitoreo de los logros educativos, la evaluación de programas educativos específicos, la selección, la certificación y el diagnóstico de las necesidades de aprendizaje individuales.

En función de estos propósitos se han establecido diferencias entre las distintas herramientas evaluativas; siendo particularmente pertinente para el presente trabajo, la diferencia establecida por Capper, (1996) entre evaluaciones nacionales y exámenes públicos. Las evaluaciones nacionales se caracterizan por presentar el estado del sistema educativo, como tal apuntan al monitoreo y evaluación de un programa y en ese sentido su dominio trasciende el currículo mismo; usualmente y en función de las limitaciones de tipo práctico, se realizan a sólo una muestra de la población, con periodicidad variable por lo general con aplicaciones cada cuatro o cinco años. Por otra parte, los exámenes públicos son aquellas pruebas que se realizan basadas en un currículo particular, que son aplicadas a quienes se inscriben o lo solicitan y generalmente se realizan con una periodicidad determinada, una o dos veces por año. Debido a estas particularidades responde mejor a los propósitos de certificación y evaluación. Desde esta perspectiva, los Exámenes para la Evaluación de la Calidad de la Educación Superior (ECAES), objeto del presente trabajo, se enmarcan dentro de la definición de examen nacional. Sus resultados se toman como indicador de la calidad de la educación nacional y responden a la evaluación denominada externa, la cual en palabras de Gago (2000), director del Centro Nacional de Evaluación (CENEVAL) en México, permite "evaluar con rigor y de manera sistemática lo que se hace en el sistema educativo" (p. 108)

Dichos exámenes públicos se enmarcan, por lo general, dentro de los llamados Sistemas Nacionales de Evaluación de la Educación (SNE) los cuales, aunque con denominaciones diferentes, buscan informar a la comunidad sobre las condiciones, características, avances y retos que presenta la educación en cada país. Los SNE han emergido desde hace varias décadas y a partir de los noventa han logrado consolidarse por varias razones: el cambio registrado en los mecanismos de administración y control de los sistemas educativos; la demanda social de información y rendición de cuentas, y la aparición de un nuevo modelo de conducción y gestión de los sistemas educativos basado en la información sobre el estado, su funcionamiento y el conocimiento de los productos del sistema educativo (Triana, 1996). A partir de 1992 al menos 20 países de América Latina, el Caribe y Europa pusieron en funcionamiento algún tipo de SNE. De acuerdo con Ravela, Wolfe, Valverde & Esquivel (2001) particularmente en América Latina han sido diversas las experiencias, propósitos y enfoques que se han dado en el tema. En una buena parte de los casos, la creación de SNE ha sido impulsada por los organismos internacionales de crédito como parte de sus convenios de préstamo con los países. Sin embargo, la experiencia particular de cada país parece depender más de decisiones políticas y capacidades técnicas internas; en algunos de ellos las pruebas se realizan a nivel censal en ciertos grados, mientras que otros trabajan con muestras de escuelas y/o grupos determinados. Entre quienes trabajan a nivel censal, unos han optado por publicar los resultados en la prensa, atribuyendo al SNE la función de entregar información a las familias con el fin de que exista un control por parte de los usuarios sobre la gestión escolar, otros han optado por devolver la información a cada escuela con carácter confidencial y otros

han desarrollado experiencias de utilización de la información de resultados a nivel grupal, como elemento de evaluación de la labor docente y como parte del sistema de incentivos económicos. Los países que trabajan sobre la base de muestras suelen presentar información de resultados generalmente a nivel nacional, divididos por áreas geográficas o tipos de escuelas. En algunos casos se hacen grandes esfuerzos por enviar los resultados con materiales de orientación didáctica y en otros, aún después de varios años de evaluaciones no se dan a conocer los resultados en forma pública.

En lo referente a contenidos de las evaluaciones, prácticamente todos los países evalúan logros en Lenguaje y Matemáticas y existe una importante variedad de situaciones en cuanto a la evaluación de otras áreas del aprendizaje como ciencias naturales, ciencias sociales, autoestima, etc, así como en cuanto a los grados y niveles evaluados y la periodicidad de las evaluaciones. Finalmente, un factor común encontrado en los diferentes países es el interés por medir otros factores que se pueden asociar al rendimiento académico de los estudiantes, en el caso de la educación básica, el ambiente escolar, familiar, la capacidad y competencia de los docentes y las características propias de las instituciones educativas.

Este panorama general permite reconocer el gran paso que se ha dado en la región. Los SNE han generando una cierta capacidad para implementar grandes operativos nacionales de aplicación de instrumentos a gran escala y las instituciones, docentes, familias y sociedad en general, empiezan a comprender la necesidad e importancia de este tipo de evaluaciones. No obstante, actualmente varios países se encuentran aún en una etapa temprana de evaluación de su propio sistema, buscando nuevas alternativas para el desarrollo y rediseño de sus evaluaciones en el futuro. Esto, como consecuencias de tres grandes deficiencias que siguiendo a Ravela y cols (2001) han sido comprobadas en las primeras fases de implementación de los sistemas: en primer lugar el insuficiente aprovechamiento de la información producida por los sistemas de evaluación, lo que ha disminuido el impacto de las evaluaciones en el conjunto del sistema; segundo, la insuficiente calidad y capacidad de evaluación de diferentes tipos y áreas de aprendizaje por medio de las pruebas que están siendo aplicadas y tercero, las debilidades técnicas presentadas en los procesos de construcción y validación de los distintos instrumentos de medición.

Para garantizar la calidad del proceso es necesario entonces aplicar el mayor rigor técnico y metodológico en la construcción de los instrumentos empleados de manera que las conclusiones que puedan derivarse permitan, por un lado, tomar decisiones relacionadas directamente con las fases del proceso, y por otro, reconocer la calidad del instrumento. Las herramientas más usadas son las denominadas pruebas objetivas las cuales han de contar con unos requerimientos mínimos que incluyen, además de una definición clara y precisa del contenido o dominio (Herrera, 1998), una evaluación de la calidad técnica de los ítems que la componen. Cumplir con estos criterios básicos constituye el inicio del camino que las ha de convertir en instrumentos útiles que permitirán obtener resultados válidos y respaldar las inferencias que se hacen a partir de los mismos. Este trabajo versa sobre los Exámenes de Evaluación de la Calidad de la Educación Superior (ECAES), pruebas objetivas construidas para la evaluación de los estudiantes próximos a obtener su título profesional en diferentes áreas de conocimiento en Colombia.

En Colombia el organismo estatal encargado del desarrollo, aplicación y calificación de pruebas dentro de los procesos de evaluación educativa, así como de generar y coordinar proyectos de investigación y prestar asesoría en lo relacionado con la medición educativa, es el Instituto Colombiano para el Fomento de la Educación Superior (ICFES), a través de su Subdirección de Aseguramiento de la Calidad de la Educación. Desde 1966 este Instituto, adscrito al Ministerio de Educación Nacional, se ha ocupado de la evaluación de los estudiantes que terminan el proceso de formación secundaria. Las pruebas empleadas dentro de este proceso, conocidas como Exámenes de Estado, se han venido desarrollando y aplicando ininterrumpidamente en las últimas cuatro décadas y su presentación constituye un requisito para el ingreso a la educación superior (Servicio Nacional de Pruebas, 1992). Sin embargo, en la década de los 90 estos procesos se ampliaron para cubrir los niveles básicos y a partir del año 2000 se inició la evaluación de la educación superior.

Aunque en la década de los 80 se trabajó en algunos proyectos dirigidos a evaluar la calidad del aprendizaje de los egresados, fue en el período de gobierno del presidente César Gaviria (1990-1994) cuando se hizo más clara la intención de llevar a cabo exámenes de la educación Superior en las carreras de Medicina, Derecho y Contaduría. Aunque el proyecto inicial no logró evaluar estas tres carreras sentó las bases para realizar convenios con universidades y asociaciones de profesionales interesadas en participar en la elaboración de tales exámenes (Ministerio de Educación Nacional, 2004). Posteriormente, en el año de 1998, después de una serie de acercamientos y desarrollo de proyectos conjuntos con diversas organizaciones de profesionales, se propuso un proyecto que pretendía evaluar a los profesionales de las diferentes áreas de la ingeniería. Dicho proyecto fue acogido y, dado el gran número de programas de ingeniería existentes en el país, se hizo necesario desarrollarlo en varias etapas. La primera de estas se dedicó al diseño y construcción de pruebas en Ingeniería Mecánica; estas pruebas fueron ajustadas tras la aplicación experimental a egresados y alumnos de último año en el 2000.

Poco tiempo después se expidieron los decretos 1716 y 2233 de 2001 y 1373 de 2002 mediante los cuales se reglamentaron los Exámenes de Estado para las carreras de Medicina, Ingeniería Mecánica y Derecho, respectivamente. Este sería el inicio de la reglamentación de los Exámenes de Evaluación de la calidad de la Educación Superior, los cuales ya habían sido considerados en el Plan Estratégico de Educación (2000-2002), como uno de los programas orientados al mejoramiento de la calidad y la transparencia en la educación superior (www.icfes.gov.co). De esta manera y atendiendo la reglamentación dictada por el Gobierno Nacional sobre los estándares y marcos básicos de las áreas del conocimiento y competencias que debían integrar los programas académicos de Ingenierías, Ciencias de la Salud, Derecho, Arquitectura, Comunicación e Información, Administración, Contaduría Pública, Economía, Psicología y Ciencias Exactas y Naturales, el ICFES convocó a las diferentes universidades, asociaciones de profesionales, facultades y demás organizaciones académicas, para que participaran con sus propuestas en la construcción y diseño de los nuevos ECAES para las carreras incluidas en estas categorías. El resultado, una amplia participación de las organizaciones invitadas y la elaboración de 27 pruebas que se aplicaron el 1º de noviembre de 2003 en 41 ciudades del país y 12 extranjeras, a un total de 58.901 personas, entre estudiantes de último año y egresados de 27 programas de pregrado (Ministerio de Educación Nacional, 2004).

Para el 2004 se incluyeron 16 programas más, de modo que se amplió la población evaluada. En el presente trabajo se realiza el análisis psicométrico de las 27 pruebas aplicadas en 2003, mediante métodos de la Teoría Clásica de los Test (TCT) y ajustando modelos de la Teoría de Respuesta al Ítem (TRI) de un parámetro o Modelos de Rasch.

Teoría Clásica de los test

Dentro de la TCT se ha planteado que el análisis de la calidad de un instrumento debe hacerse analizando los ítems que la componen y la prueba en su totalidad. Las características evaluadas en los ítems son su dificultad y su discriminación, mientras que las de la prueba son su confiabilidad y su validez. El índice de dificultad de un ítem es la proporción de sujetos que, habiendo intentado resolverlo, lo resuelven correctamente. Existen parámetros para la aceptación de un ítem de acuerdo con su nivel de dificultad, el establecido por Guilford (1975), por ejemplo, es el rango 0.2 y 0.85, aunque en realidad cada investigador o usuario puede fijar el rango de aceptación de acuerdo con los datos obtenidos y los fines de la prueba (Lord y Novick, 1968). La principal limitación de este índice es que depende de la población que responde el ítem, entre mayor maestría de dominio tenga el grupo de examinados, el ítem resultará mas fácil y viceversa. El índice de discriminación indica si el ítem tiene alto poder para diferenciar entre aquellos sujetos que muestran mayor y menor magnitud del atributo o dominio evaluado. En consecuencia, el índice de discriminación se define como la relación entre las puntuaciones obtenidas por los sujetos en el ítem y las obtenidas en el test. El tipo de correlación que se utilice dependerá de las características de las variables que se van a correlacionar, en este caso el ítem y el test. De acuerdo con Guilford (1975) es aceptada la discriminación de un ítem si arroja un valor superior a 0.2 ó 0.3. En el mismo sentido, Cano (2004) considera que una correlación ítem-test mayor o igual a 0.25 es indicador de la calidad de los ítems. Un indicador que arroja información valiosa sobre la capacidad discriminativa de los ítems es el índice de poder selectivo o Brodgen Clemens (BrCl), el cual estima la correlación máxima entre los puntajes en el ítem y en la prueba, teniendo en cuenta la proporción de personas que contestó correctamente la pregunta.

Además, dentro de los análisis de los ítems se pueden hacer análisis al flujo de opciones con el fin de identificar la calidad de los distractores del mismo. La investigación al respecto ha revelado que existe una relación directa entre las opciones del ítem y el puntaje total de la prueba y que es necesario realizar un estudio de los distractores de los ítems antes de determinar la validez de las pruebas (Haladyna, 1994). Algunos de los métodos que existen han sido trabajados por Haladyna (1994) y Attali y Fraenkel (2000). Estos métodos deben acompañarse de una mirada cualitativa por parte de expertos (en el área de medición como en construcción de instrumentos) que estimen qué tan acertados o tan errados son los distractores y en qué medida deben ser revisados, modificados o eliminados. La investigación desarrollada por Haladyna y Downing (1993) propone una clasificación para los distractores teniendo en cuenta el criterio porcentaje de respuesta: *poco razonables* si tienen un porcentaje de respuesta inferior al 5%; *los que no discriminan* si son elegidos en igual proporción por quienes tienen puntajes altos y por quienes tiene puntajes bajos; *no claros* si varían a lo largo de los distintos rangos de puntajes de la prueba, e *indeseables* si imitan el comportamiento de la clave, es decir, aumentan en la medida que aumentan los puntajes de la prueba.

De otra parte, la calidad de las pruebas como un todo se juzga según su confiabilidad y su validez. La primera hace referencia a la consistencia y precisión en las medidas arrojadas por una prueba y los métodos más frecuentemente utilizados para estimarla son el Test–Retest, la correlación entre formas paralelas o entre dos mitades que utiliza el algoritmo de Spearman-Brown como forma de estimación de la confiabilidad total del instrumento (Brown, 2000); y, el coeficiente de Consistencia Interna que se basa en variados métodos de toma de muestras de diferentes ítems y se revisan los efectos de estos sobre la confiabilidad (Aiken, 1996), entre los métodos más conocidos están Kuder Richardson y Alfa de Cronbach. La validez hace referencia al grado en que una prueba mide aquello que se supone debe medir y es útil para el propósito con el que se construyó. Tradicionalmente se ha asumido que dependiendo de lo que busque la prueba y sus propósitos, se puede hablar de diferentes tipos de validez, incluso de diferentes procedimientos para hallarla; desde esta perspectiva se han distinguido tres aproximaciones a la validez: relación con un criterio, validez de contenido y validez de constructo. De acuerdo con Martínez (1996) es absolutamente necesario realizar la validez de contenido en los test que buscan dar cuenta del proceso educativo y ocupacional; elgunas descripciones de los procedimientos para estimarla se encuentran en Martínez (1996), Herrera (1998), Muñiz (2000) y McGartland, Berg-Weger, Tebb, Lee, & Rauch, (2003). Sin embargo, en la década de los noventa la propuesta de Messick (1995) apunta a que la única validez sea la de constructo y que para alcanzarla implique a las otras dos. Esta propuesta se enmarca incluso dentro de la definición más reciente planteada en los estándares para pruebas educativas y psicológicas según la cual la validez es el grado en el cual la evidencia y la teoría respalda la interpretación de los puntajes de una prueba de acuerdo con los propósitos de la misma. En este sentido el proceso de validación consiste en la acumulación de tales evidencias (Kane, 2001).

Cuando se trata de pruebas utilizadas en la evaluación educativa, generalmente conocidas como pruebas de "rendimiento académico", el principal interés es determinar el grado en que los ítems representan el "contenido" disciplinar que se quiere medir. Paz (1996) presenta las aproximaciones a la estimación de la validez de un instrumento construido para estos fines, señalando que este tipo de pruebas se enmarcan en los estudios de validez que se hace con el juicio de expertos (contenido) y señalando la Teoría de la Generalizabilidad como una estrategía útil para consolidar este procedimiento. En este mismo texto la autora presenta la posibilidad de hablar de la validez de constructo de estos instrumentos, principalmente cuando se hace referencia al proceso "cognitivo" que se espera se ponga en juego a la hora de resolver la pregunta. En esta línea de investigación algunos ejemplos de trabajos que buscan evaluar la validez de pruebas de rendimiento son los de Watermann y Klieme (2002), Powers, (2004) y Montoya y Rodríguez, (2004).

Modelos de Rasch

La Teoría de Respuesta al Ítem logra resolver algunos problemas de la TCT como la invarianza de los parámetros (Hambleton, Swaminathan & Rogers, 1991, Muñiz, 1997, Herrera, Sánchez y Jiménez, 2001). Basados en los supuestos de unidimensionalidad e independencia local, se han desarrollado modelos logísticos que explican la probabilidad de acierto en un ítem para una magnitud de atributo dada, en función de uno, dos o tres

parámetros de los ítems. El Modelo Rasch, elaborado por el matemático danés Georg Rasch en la década de los 50, se considera como un modelo de un parámetro aunque existe en la actualidad una discusión entre los académicos que utilizan este modelo sobre si éste en realidad forma parte de la T.R.I. o no (Muñiz, 1997, Pardo, 2001). De acuerdo con estos modelos la probabilidad de respuesta de una persona en un ítem es función de la diferencia entre la medida del rasgo de una persona y la dificultad del ítem, lo que se representa en la curva característica del ítem (CCI) (Tristán, 1998). Si *H* es la magnitud de habilidad del examinado y *b* es la dificultad del ítem, de acuerdo con el modelo de Rasch la probabilidad de acierto en el ítem para un valor de *H*, está dada por $p(H) = \frac{1}{1+e^{-(H-b)}}$

Dentro de este marco, el único parámetro del ítem es la dificultad, este parámetro corresponde con la medida de la persona en el punto donde la probabilidad de acierto es de 0.5. Además dado que este modelo busca determinar la estructura de los datos, resulta importante determinar la calidad de la estimación y en ese sentido contar con una medida de ajuste. En este sentido, el modelo tiene como referencia dos medidas de ajuste una externa (OUTFIT) y una interna (INFIT); en el primer caso se trata de una medida sensible al comportamiento inesperado que afecta a los ítems que presentan un nivel de dificultad lejano del nivel de habilidad de una persona y en el segundo caso la sensibilidad en relación con la cercanía del nivel de habilidad de la persona (Tristan, 1998). Los valores de ajuste de los ítems deben encontrarse entre 0.5 y 1.5 (Linacre, 2003).

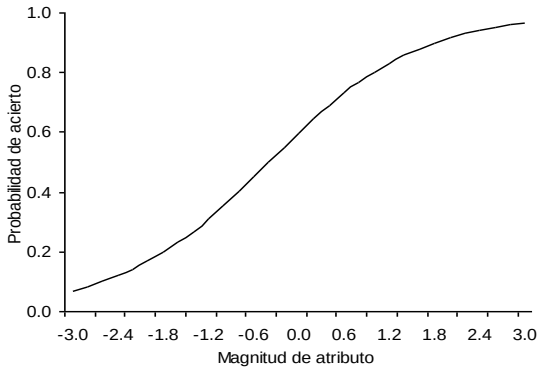

Figura 1. Ejemplo de una CCI cuando se ha ajustado un modelo de Rasch

En este contexto, la medida de una persona es la posición que ocupa en la escala del rasgo latente medido con un conjunto de reactivos y la medida del reactivo para efectos de la calibración es la posición en la escala del rasgo latente que mide sobre un conjunto de personas. Para calibrar el ítem se emplea el mismo proceso que para la medida de una persona, la única diferencia radica en que el momio se expresa en términos de la expectativa de fracaso con relación al éxito. Estas medidas se ubican en un plano creando una curva que relaciona las medidas en logit Vs. Probabilidad de respuesta. Esto permite obtener la medida de una persona con relación al atributo conociendo la probabilidad de respuesta al ítem o la probabilidad de respuesta al ítem conociendo la medida de la persona con relación al atributo, esta es la curva característica del ítem (C.C.I.) que se presenta en la figura. 1.

Método

Bases de datos

Para el análisis de los ECAES se dispuso de la totalidad de las aplicaciones de las 27 pruebas adelantadas por el ICFES en el año 2003, distribuidas como se muestra en la tabla 1. Los examinados fueron estudiantes de 9º y 10º semestres y egresados de los 27 programas de formación universitaria en todo el país. En la tabla 1 se muestra la distribución de los evaluados por categoría de programas. Se analizaron 27 pruebas ECAES de los programas mencionados, construidas teniendo en cuenta los criterios propios de la validez de contenido. Estas pruebas estaban compuestas por preguntas de selección múltiple con cuatro o cinco opciones de respuesta. La prueba de Arquitectura incluyó una subprueba de desempeño que no se analizó en el presente estudio. El número de preguntas de las pruebas estaba entre 120 en el caso de las Ingenierías y 300 en el caso de Psicología. Los dominios de contenido de las 27 carreras evaluadas pueden ser consultadas en el anexo.

Tabla 1
Número de pruebas y examinados por área del conocimiento

Categoría de programas	N° de pruebas	N° de evaluados
Ciencias de la salud	8	10645
Ingeniería, Arquitectura, Urbanismo y afines	17	33730
Ciencias Sociales y Humanas	2	14556
Total	27	58931

Procedimiento

La recolección de la información pertinente se obtuvo de dos fuentes: comunicación enviada al director del ICFES para solicitar las bases de datos y solicitud a los expertos en psicometría de cada uno de los equipos constructores de las pruebas de los documentos o información sobre el proceso seguido en el desarrollo de las mismas y su validación. Una vez recolectada la información se procedió a realizar el procesamiento y análisis respectivo. Se construyeron bases de datos para cada programa que incluyeron la información básica: identificador y respuestas por ítem. Construidas estas bases de datos, se procedió a analizar los ítems bajo el modelo de la Teoría Clásica de los Test y el modelo de Rasch.

Para el análisis de ítems con la TCT se realizó una aplicación desarrollada por el segundo autor, que permitió obtener las estimaciones de dificultad y discriminación de cada ítem. Una vez obtenidas las estimaciones, los ítems se clasificaron en "aceptados" si cumplían en conjunto con los siguientes criterios: dificultad entre 0.1 y 0.9; discriminación mayor de 0.2 en el caso de rpbis y de 0.3 en el caso de BrCl, "rechazados o para eliminar" si no cumplían estos criterios y "modificar" aquellos que cumpliendo los criterios de dificultad y discriminación, tenían distractores escogidos por menos de 5% de los examinados. La confiabilidad se estimó con el coeficiente alfa de Cronbach utilizando el programa SPSS versión 10 para Windows. Por otra parte, se ajustaron modelos Rach utilizando Winsteps versión 2.9 desarrollado por John Michael Linacre (Linacre, 2003).

Todos los análisis se llevaron a cabo para cada uno de los ítems de las 27 pruebas, analizando un total de 3772 ítems. Los resultados se resumieron para cada prueba dentro de la categoría de programas evaluados; además, por la conformación particular de la prueba de psicología se realizó el análisis separado de las cinco subpruebas profesionales: Psicología Clínica, Educativa, Jurídica, Social y Organizacional, como si se tratara de pruebas diferentes ya que estas 5 pruebas se aplicaron a poblaciones diferentes y sus puntajes totales desde luego no respondían a los mismos ítems.

Resultados

Análisis de ítems con TCT

De acuerdo con los criterios definidos, se encontró que el mínimo porcentaje de ítems que cumplían con los criterios de aceptación estaban en la subprueba de Psicología Educativa y el máximo en Ingeniería Electrónica. En el caso de los ítems que no cumplen con los criterios el máximo porcentaje de ítems corresponde al 79.9% que se presentó en la prueba de Enfermería y el mínimo en la prueba de Ingeniería Electrónica con un 27%. Las comparaciones por grupos de programas según área del conocimiento mostró resultados diversos. Dentro del grupo de pruebas de Ciencias de la Salud la prueba que presentó mayor número de ítems no adecuados fue la de Enfermería con un 79% y la que resultó con menos ítems no adecuados fue Optometría con un 48%. El grupo de Ingeniería, Arquitectura, Urbanismo y Afines incluyó 17 pruebas de las cuales 14 de ingenierías comparten un núcleo común de preguntas en el componente básico; dentro de este grupo el porcentaje de ítems para modificar osciló entre el 12.7% y el 41.7% y la prueba de Ingeniería Electrónica presentó el más alto porcentaje de ítems aceptados seguida por Ingeniería Mecánica con un 24.8%. En el grupo de pruebas de Ciencias Sociales y Humanas se encuentran la prueba de Derecho con 200 ítems y la de Psicología con sus respectivas subpruebas: Psicología Clínica, Educativa, Jurídica, Organizacional y Social con 50 ítems cada una. Para la prueba de Derecho se observó que el 49.5% de los ítems no cumplían con los criterios y solamente un 11% de ítems era adecuado. Por otro lado, la prueba común de Psicología mostró 69.2% de ítems que no cumplían con los criterios y dentro de la subpruebas, la de Psicología Social tuvo el mayor porcentaje de ítems para eliminar. En general todas las subpruebas mostraron bajos porcentajes de ítems que cumplen con los criterios, particularmente Psicología Educativa que sólo tuvo un 4% de ítems que cumplían con todos los criterios.

La clasificación de los ítems según su discriminación arrojó resultados diferentes según el índice empleado. Con el coeficiente de correlación punto biserial corregido (Rpbis) resultaban rechazados porcentajes altos de preguntas por prueba debido a la dependencia que tiene el valor de esta correlación de la dificultad del ítem y a la gran proporción de ítems muy fáciles o muy difíciles. Cuando la discriminación se estimó mediante el coeficiente de Brodgen Clemens, más reactivos funcionan favorablemente. Es así como por ejemplo en el área de las ciencias de la salud, la carrera en que más preguntas presentaron problemas de discriminación y dificultad fue Fisioterapia con un 62.7%, y la carrera en la que menos preguntas cumplieron con los criterios fue Medicina. De hecho es de las carreras con mayor cantidad de ítems para

modificar distractores (43%) ya que no atraen siquiera al 5% de la población. Un caso similar se presenta con Odontología que muestra un 54% de ítems para modificar distractores, aunque presentan una dificultad y discriminación aceptables. Es de anotar que en el área de ciencias de la salud, la prueba que mejor se comporta, según el porcentaje de ítems que cumplen con todos los criterios, es Odontología.

Tabla 2
Descriptivos de los parámetros de dificultad y discriminación con métodos de la TCT, para cada prueba

Prueba	Dificultad			Corr. biserial puntual			Brodgen-Clemens		
	Media	**Max.**	**Min.**	**Media**	**Max.**	**Min.**	**Media**	**Max.**	**Min.**
Enfermería	*0.51*	*0.973*	*0.009*	*0.14*	*0.392*	*-0.147*	*0.21*	*0.505*	*-0.122*
Fisioterapia	*0.42*	*0.873*	*0.021*	*0.12*	*0.392*	*-0.266*	*0.18*	*0.476*	*-0.271*
Fonoaudiología	*0.51*	*0.968*	*0.069*	*0.15*	*0.482*	*-0.192*	*0.22*	*0.581*	*-0.160*
Medicina	*0.52*	*0.993*	*0.039*	*0.15*	*0.412*	*-0.168*	*0.23*	*0.511*	*-0.165*
Nutrición y dietética	*0.55*	*0.990*	*0.108*	*0.18*	*0.569*	*-0.181*	*0.26*	*0.653*	*-0.231*
Odontología	*0.52*	*0.949*	*0.071*	*0.18*	*0.487*	*-0.160*	*0.26*	*0.590*	*-0.168*
Optometría	*0.52*	*0.948*	*0.069*	*0.22*	*0.586*	*-0.073*	*0.31*	*0.708*	*-0.074*
Terapia ocupacional	*0.47*	*0.920*	*0.045*	*0.13*	*0.436*	*-0.252*	*0.20*	*0.535*	*-0.426*
Arquitectura	*0.58*	*0.976*	*0.108*	*0.21*	*0.463*	*-0.142*	*0.31*	*0.632*	*-0.127*
Ingeniería agrícola	*0.42*	*0.982*	*0.074*	*0.15*	*0.426*	*-0.282*	*0.24*	*0.627*	*-0.267*
Ing. agronómica y agronomía	*0.42*	*0.920*	*0.052*	*0.12*	*0.355*	*-0.160*	*0.20*	*0.509*	*-0.190*
Ingeniería ambiental	*0.36*	*0.920*	*0.072*	*0.14*	*0.376*	*-0.128*	*0.22*	*0.538*	*-0.086*
Ingeniería civil	*0.41*	*0.977*	*0.110*	*0.19*	*0.507*	*-0.068*	*0.28*	*0.628*	*-0.034*
Ingeniería de alimentos	*0.41*	*0.876*	*0.090*	*0.18*	*0.557*	*-0.093*	*0.26*	*0.683*	*-0.055*
Ingeniería de materiales	*0.43*	*0.952*	*0.047*	*0.18*	*0.520*	*-0.138*	*0.31*	*0.758*	*-0.130*
Ingeniería de minas	*0.36*	*0.947*	*0.009*	*0.14*	*0.546*	*-0.245*	*0.23*	*0.700*	*-0.270*
Ingeniería de sistemas	*0.33*	*0.830*	*0.079*	*0.20*	*0.467*	*-0.082*	*0.29*	*0.574*	*-0.065*
Ing. telecomunicaciones	*0.42*	*0.993*	*0.068*	*0.22*	*0.518*	*-0.103*	*0.31*	*0.645*	*-0.087*
Ingeniería eléctrica	*0.48*	*0.983*	*0.064*	*0.23*	*0.550*	*-0.090*	*0.33*	*0.694*	*-0.080*
Ingeniería electrónica	*0.42*	*0.991*	*0.110*	*0.44*	*0.616*	*0.159*	*0.34*	*0.740*	*-0.004*
Ingeniería geológica	*0.43*	*0.991*	*0.061*	*0.14*	*0.477*	*-0.223*	*0.23*	*0.673*	*-0.206*
Ingeniería industrial	*0.35*	*0.827*	*0.067*	*0.15*	*0.481*	*-0.164*	*0.23*	*0.602*	*-0.151*
Ingeniería mecánica	*0.42*	*0.969*	*0.064*	*0.25*	*0.573*	*-0.030*	*0.35*	*0.741*	*-0.002*
Ingeniería metalúrgica	*0.38*	*0.960*	*0.007*	*0.20*	*0.568*	*-0.266*	*0.28*	*0.720*	*-0.273*
Ingeniería química	*0.47*	*0.984*	*0.050*	*0.21*	*0.465*	*-0.103*	*0.31*	*0.630*	*-0.079*
Derecho	*0.48*	*0.914*	*0.032*	*0.20*	*0.468*	*-0.210*	*0.28*	*0.604*	*-0.285*
Psicología Subp. común	*0.39*	*0.899*	*0.065*	*0.14*	*0.446*	*-0.203*	*0.21*	*0.562*	*-0.255*
Psicología Clínica	*0.50*	*0.946*	*0.083*	*0.17*	*0.380*	*-0.211*	*0.30*	*0.486*	*-0.076*
Psic. Organizacional	*0.46*	*0.926*	*0.076*	*0.15*	*0.313*	*-0.133*	*0.27*	*0.432*	*-0.011*
Psicología Jurídica	*0.42*	*0.898*	*0.015*	*0.12*	*0.618*	*-0.357*	*0.26*	*0.656*	*-0.147*
Psicología Social	*0.34*	*0.748*	*0.045*	*0.11*	*0.446*	*-0.264*	*0.24*	*0.533*	*-0.132*
Psicología Educativa	*0.41*	*0.848*	*0.046*	*0.15*	*0.537*	*-0.238*	*0.27*	*0.613*	*-0.316*

En el área de Ingeniería, Arquitectura, Urbanismo y afines se observa que la prueba que presentó la mayor cantidad de ítems que no cumplen con los criterios de aceptación es Ingeniería de Sistemas con 58.5% y la que presenta la mayor cantidad de ítems con la totalidad de los criterios dentro de los rangos de aceptación es Ingeniería Mecánica. Por otro lado, las pruebas que muestran la mayor cantidad de reactivos con problemas en términos de distractores son Ingeniería de Materiales, Química y Eléctrica con 44.1%, 45.3% y 45.4% respectivamente, siendo Ingeniería Eléctrica la que menos ítems inadecuados, en términos de dificultad y discriminación presenta. Dentro de la última categoría, de las pruebas de psicología la que presenta mayor cantidad de ítems que cumplen con todos los criterios de aceptación desde el modelo clásico es la de Psicología Organizacional con un 30%, y la que presenta menos es Psicología Jurídica. La subprueba de Psicología Clínica presenta la mayor cantidad de ítems con problemas en sus distractores y la de Psicología General es la que presenta más ítems con limitaciones por dificultad o discriminación. La tabla 2 muestra los estadísticos descriptivos para los parámetros de los ítems calculados con métodos de la TCT, para cada una de las 27 pruebas.

Validez y confiabilidad

En la tabla 3 se muestran las correlaciones ítem-prueba máxima y mínima así como el valor del α de Cronbach para cada una de las 27 pruebas. Las pruebas con mayor coeficiente de confiabilidad son las de Ingeniería Mecánica y Derecho y la de menor coeficiente es la de Agronomía. No se presenta ningún valor inferior a 0.6, es decir que todas las pruebas tienen un grado alto de confiabilidad. Observando en conjunto el grupo de carreras se encuentra que aquellas que tienen las mayores correlaciones son Optometría en Ciencias de la Salud, e Ingeniería Mecánica en el grupo de Arquitectura, Ingeniería y Afines; y en el grupo de Ciencias Sociales y Humanas, la prueba de Derecho. Lo encontrado respalda el hecho de que estas pruebas tengan los más altos niveles de confiabilidad. Sin embargo, todas presentaron valores de correlación ítem-prueba negativas y la exclusión de tales ítems no mejoraba el coeficiente α.

En cuanto a la validez de contenido de las pruebas, se encontró que en todos los procesos ECAES se siguieron los pasos que la psicometría indica para construir pruebas educativas y garantizar su validez de contenido. En el caso de Derecho el proceso se inició con la presentación de una propuesta de especificaciones de la Universidad de la Sabana y se examinó y enriqueció en reuniones regionales con las facultades de Derecho en cinco ciudades del país y una reunión nacional de decanos (Otálora, 2004). Para las pruebas de ingeniería esta primera fase de especificaciones no fue muy diferente, la Asociación Colombiana de Facultades de Ingeniería (ACOFI), encargada del respectivo proceso, después de un análisis de la estructura curricular de las distintas carreras de ingeniería y de los decretos reglamentarios pertinentes, propuso una estructura de prueba que fue discutida en el comité académico *ad hoc* y en reuniones nacionales con los directores de cada programa (ACOFI, 2003). Los grupos encargados del desarrollo de las pruebas de Arquitectura y las ciencias de la salud reportaron reuniones nacionales para definir las especificaciones de prueba. En el caso de psicología tal definición se basó en los estándares definidos por ley y fue discutida en distintas reuniones regionales.

Tabla 3
Mínimo y máximo de la correlación ítem-prueba y Alfa de Cronbach para cada prueba

Prueba	*No. De Ítems*	*No de examinados*	*α de Cronbach*	*Correlación Mínima*	*Correlación Máxima*
Enfermería	*199*	*2544*	*0.80*	*-0.13*	*0.36*
Fisioterapia	*150*	*1595*	*0,72*	*-0.22*	*0.34*
Fonoaudiología	*150*	*417*	*0.78*	*-0.16*	*0.43*
Medicina	*200*	*3595*	*0.84*	*-0.15*	*0.38*
Nutrición y dietética	*150*	*312*	*0.84*	*-0.17*	*0.53*
Odontología	*150*	*1677*	*0.84*	*-0.15*	*0.45*
Optometría	*150*	*274*	*0.89*	*-0.07*	*0.56*
Terapia ocupacional	*150*	*240*	*0.75*	*-0.22*	*0.39*
Arquitectura	*129*	*2625*	*0.85*	*-0.12*	*0.43*
Ingeniería agrícola	*118*	*227*	*0.76*	*-0.24*	*0.37*
Ingeniería civil	*118*	*3630*	*0.82*	*-0.06*	*0.47*
Ingeniería eléctrica	*119*	*801*	*0.88*	*-0.09*	*0.52*
Ingeniería electrónica	*115*	*3957*	*0.89*	*-0.04*	*0.46*
Ingeniería química	*117*	*1097*	*0.85*	*-0.09*	*0.43*
Ingeniería industrial	*117*	*6294*	*0.77*	*-0.14*	*0.427*
Ingeniería de sistemas	*120*	*8548*	*0.84*	*-0.08*	*0.43*
Ingeniería mecánica	*120*	*1780*	*0.90*	*-0.03*	*0.55*
Ingeniería de materiales	*120*	*63*	*0.74*	*-0.13*	*0.48*
Ingeniería telecomunicaciones	*120*	*452*	*0.85*	*-0.09*	*0.49*
Ingeniería ambiental	*119*	*2157*	*0.74*	*-0.10*	*0.32*
Ingeniería geológica	*118*	*114*	*0.74*	*-0.19*	*0.42*
Ingeniería de minas	*118*	*208*	*0.75*	*-0.20*	*0.48*
Ingeniería de alimentos	*119*	*807*	*0.81*	*-0.08*	*0.50*
Derecho	*200*	*9312*	*0.90*	*-0.19*	*0.45*
Psicología subprueba común	*250*	*5249*	*0.85*	*-0.19*	*0.41*
Ing. agronómica y agronomía	*120*	*928*	*0.69*	*-0.12*	*0.29*
Ingeniería metalúrgica	*116*	*126*	*0.83*	*-0.25*	*0.54*

Este primer momento de los procesos permitió también la identificación y ubicación de los expertos en cada uno de los componentes establecidos en las especificaciones de las pruebas, información que resultó de gran utilidad para la realización de las siguientes fases de los mismos. La fase de construcción y revisión de preguntas siguió distintas modalidades, pero siempre buscando garantizar que las preguntas fueran revisadas por expertos en cada componente. En el caso de las ingenierías, luego de talleres sobre elaboración de preguntas, se invitó a participar a todos los docentes universitarios en la construcción de las mismas, posteriormente se realizaron talleres de revisión en Bogotá, invitando a los expertos de cada programa y área de distintos lugares del país quienes en compañía de expertos en elaboración de ítems decidieron cuáles preguntas serían incluidas en las pruebas. Aunque la prueba de Agronomía e Ingeniería Agronómica guardan afinidad con las ingenierías, su desarrollo se llevó a cabo en tres etapas: talleres de capacitación sobre construcción de preguntas, revisión de preguntas con un grupo de expertos y revisión y selección de preguntas que serían

incluidas en la prueba. En el caso de derecho se conformó un equipo de profesores, que elaboró y revisó un determinado número de preguntas, las cuales fueron revisadas en talleres regionales y luego en un taller nacional con expertos de distintas áreas se revisó la totalidad de las preguntas provenientes de los talleres regionales. Finalmente, en la construcción de la prueba de Psicología tanto la construcción como revisión de las preguntas se llevó a cabo con simultaneidad espacio-temporal, durante 15 días un grupo de expertos en distintas áreas de la psicología y de diversas instituciones universitarias del país se dio cita en un mismo lugar para construir y revisar las preguntas que conformaron la prueba. Puede verse entonces que a pesar de haber seguido distintas estrategias, atendiendo la naturaleza de la disciplina, en todos los procesos se garantizó que cada pregunta fuera revisada por un experto del área y un experto en psicometría o en medición y evaluación.

Ajuste de modelos de Rasch

La tabla 4 muestra los estadísticos descriptivos para las estimaciones de dificultad y su error estándar así como del ajuste de los modelos de Rasch para cada una de las 27 pruebas analizadas. Los resultados muestran que en general, se logró un adecuado ajuste de los datos al modelo. Dentro del grupo de Arquitectura, Ingeniería y Afines se encuentra que las pruebas de Ingeniería Mecánica y Agrícola tienen los valores de ajuste infit mínimo y máximo, respectivamente. El valor de ajuste outfit bajo lo presenta la prueba de Ingeniería Electrónica y el alto la prueba de Ingeniería Metalúrgica. En el caso de Ciencias Humanas y Sociales la prueba de Derecho presentó los valores más bajos y altos en infit y en outfit. Finalmente, en el grupo de Ciencias de la Salud la prueba de Optometría es la que tiene el valor mínimo de ajuste infit y la de Terapia Ocupacional el mayor valor de ajuste outfit. La prueba de Nutrición y dietética presenta el mayor valor de ajuste infit y el menor valor de outfit.

Tabla 4
Descriptivos de las estimaciones de dificultad, su error de estimación y la bondad de ajuste de los Modelo de Rasch para cada una de las pruebas

Prueba		Dific.	Error	Ajuste Infit	Ajuste Outfit
Enfermería	Media	*0*	*0.05*	*1*	*1*
	Desv.	*1.29*	*0.02*	*0.03*	*0.06*
	Max.	*4.9*	*0.21*	*1.1*	*1.15*
	Mín.	*-3.56*	*0.04*	*0.92*	*0.81*
Fisioterapia	Media	*0*	*0.06*	*1*	*1*
	Desv.	*1.02*	*0.01*	*0.03*	*0.06*
	Max.	*3.21*	*0.15*	*1.09*	*1.29*
	Mín.	*-2.44*	*0.05*	*0.93*	*0.89*
Fono-audiología	Media	*0*	*0.11*	*1*	*1*
	Desv.	*1.09*	*0.03*	*0.04*	*0.06*
	Max.	*2.73*	*0.28*	*1.12*	*1.17*
	Mín.	*-3.43*	*0.1*	*0.89*	*0.81*
Medicina	Media	*0*	*0.04*	*1*	*1*
	Desv.	*1.38*	*0.02*	*0.04*	*0.07*
	Max.	*3.4*	*0.22*	*1.13*	*1.34*
	Mín.	*-5.09*	*0.03*	*0.91*	*0.83*
Ingeniería agrícola	Media	*0*	*0.16*	*1*	*1.01*
	Desv.	*1.1*	*0.05*	*0.04*	*0.09*
	Max.	*2.25*	*0.58*	*1.18*	*1.44*
	Mín.	*-4.74*	*0.14*	*0.91*	*0.7*
Ingeniería civil	Media	*0*	*0.04*	*1*	*1.01*
	Desv.	*0.96*	*0.01*	*0.05*	*0.08*
	Max.	*1.78*	*0.11*	*1.11*	*1.22*
	Mín.	*-4.33*	*0.03*	*0.87*	*0.77*
Ingeniería eléctrica	Media	*0*	*0.08*	*1*	*1.01*
	Desv.	*1*	*0.02*	*0.06*	*0.1*
	Max.	*2.75*	*0.29*	*1.14*	*1.33*
	Mín.	*-4.4*	*0.07*	*0.84*	*0.63*
Ingeniería electrónica	Media	*0*	*0.04*	*1*	*1.02*
	Desv.	*1.07*	*0.01*	*0.07*	*0.13*
	Max.	*1.9*	*0.18*	*1.18*	*1.33*
	Mín.	*-5.33*	*0.03*	*0.86*	*0.48*

Prueba		Dific.	Error	Ajuste Infit	Ajuste Outfit
Nutrición y dietética	Media	*0*	*0.14*	*1*	*1*
	Desv.	*1.17*	*0.05*	*0.06*	*0.1*
	Max.	*2.51*	*0.58*	*1.16*	*1.4*
	Mín.	*-4.42*	*0.12*	*0.84*	*0.67*
Odontología	Media	*0*	*0.06*	*1*	*1*
	Desv.	*1.05*	*0.01*	*0.05*	*0.07*
	Max.	*2.79*	*0.11*	*1.15*	*1.26*
	Mín.	*-2.9*	*0.05*	*0.88*	*0.85*
Optometría	Media	*0*	*0.14*	*1*	*0.99*
	Desv.	*1.08*	*0.03*	*0.07*	*0.11*
	Max.	*2.89*	*0.28*	*1.15*	*1.36*
	Mín.	*-2.89*	*0.13*	*0.82*	*0.73*
Terapia ocupacional	Media	*0*	*0.15*	*1*	*1.01*
	Desv.	*1.03*	*0.03*	*0.04*	*0.09*
	Max.	*2.99*	*0.31*	*1.11*	*1.54*
	Mín.	*-2.61*	*0.13*	*0.91*	*0.87*
Arquitectura	Media	*0*	*0.05*	*1*	*0.99*
	Desv.	*1.15*	*0.01*	*0.05*	*0.1*
	Max.	*2.66*	*0.13*	*1.16*	*1.34*
	Mín.	*-3.38*	*0.04*	*0.88*	*0.73*
Derecho	Media	*0*	*0.02*	*1*	*1.01*
	Desv.	*1.09*	*0.006*	*0.05*	*0.1*
	Max.	*3.39*	*0.06*	*1.15*	*1.5*
	Mín.	*-2.58*	*0.02*	*0.88*	*0.77*
Psicología	Media	*0*	*0.03*	*1*	*1.01*
	Desv.	*0.96*	*0.005*	*0.04*	*0.07*
	Max.	*2.23*	*0.06*	*1.11*	*1.29*
	Mín.	*-2.74*	*0.03*	*0.9*	*0.82*
Ingeniería agronómica y agronomía	Media	*0*	*0.08*	*1*	*1.01*
	Desv.	*0.97*	*0.010*	*0.03*	*0.07*
	Max.	*2.64*	*0.16*	*1.08*	*1.38*
	Mín.	*-2.88*	*0.07*	*0.93*	*0.81*
Ingeniería metalúrgica	Media	*0*	*0.23*	*1*	*1.01*
	Desv.	*1.21*	*0.100*	*0.07*	*0.14*
	Max.	*4.36*	*1.01*	*1.18*	*1.56*
	Mín.	*-3.91*	*0.18*	*0.84*	*0.62*

Prueba		Dific.	Error	Ajuste Infit	Ajuste Outfit
Ingeniería química	Media	*0*	*0.07*	*1*	*1.01*
	Desv.	*1.17*	*0.02*	*0.05*	*0.1*
	Max.	*2.95*	*0.24*	*1.14*	*1.29*
	Mín.	*-4.36*	*0.06*	*0.88*	*0.77*
Ingeniería industrial	Media	*0*	*0.03*	*1*	*1.02*
	Desv.	*0.87*	*0.004*	*0.04*	*0.08*
	Max.	*1.99*	*0.05*	*1.09*	*1.32*
	Mín.	*-2.35*	*0.03*	*0.88*	*0.86*
Ingeniería de sistemas	Media	*0*	*0.03*	*1*	*1.01*
	Desv.	*0.86*	*0.006*	*0.06*	*0.1*
	Max.	*1.76*	*0.04*	*1.13*	*1.32*
	Mín.	*-2.49*	*0.02*	*0.89*	*0.83*
Ingeniería mecánica	Media	*0*	*0.06*	*1*	*1.01*
	Desv.	*0.93*	*0.010*	*0.08*	*0.13*
	Max.	*2.49*	*0.14*	*1.17*	*1.35*
	Mín.	*-4.01*	*0.05*	*0.82*	*0.75*
Ingeniería de materiales	Media	*0*	*0.31*	*1*	*1.01*
	Desv.	*1.22*	*0.080*	*0.06*	*0.11*
	Max.	*2.78*	*0.72*	*1.13*	*1.41*
	Mín.	*-3.8*	*0.26*	*0.88*	*0.73*
Ing. de telecomunicaciones	Media	*0*	*0.12*	*1*	*1.01*
	Desv.	*1.15*	*0.050*	*0.07*	*0.13*
	Max.	*2.37*	*0.58*	*1.14*	*1.37*
	Mín.	*-5.54*	*0.1*	*0.86*	*0.54*
Ingeniería ambiental	Media	*0*	*0.05*	*1*	*1.01*
	Desv.	*0.93*	*0.010*	*0.03*	*0.06*
	Max.	*1.98*	*0.08*	*1.08*	*1.2*
	Mín.	*-3.16*	*0.04*	*0.93*	*0.82*
Ingeniería geológica	Media	*0*	*0.22*	*1*	*1*
	Desv.	*1.09*	*0.080*	*0.05*	*0.08*
	Max.	*2.49*	*1.01*	*1.13*	*1.31*
	Mín.	*-5.11*	*0.19*	*0.89*	*0.65*
Ingeniería de minas	Media	*0*	*0.18*	*1*	*1.01*
	Desv.	*1.15*	*0.070*	*0.05*	*0.1*
	Max.	*4.02*	*0.71*	*1.1*	*1.55*
	Mín.	*-3.67*	*0.14*	*0.86*	*0.69*
Ing. de alimentos	Media	*0*	*0.08*	*1*	*1.01*
	Desv.	*0.99*	*0.010*	*0.05*	*0.09*
	Max.	*1.97*	*0.12*	*1.11*	*1.27*
	Mín.	*-2.47*	*0.07*	*0.85*	*0.8*

En cuanto a las estimaciones de la dificultad de los ítems se evidencia que las pruebas de Fisioterapia y de Terapia Ocupacional presentaron una mayor cantidad de ítems con dificultad superior a 2. Las carreras de Odontología y Medicina presentan la misma cantidad de reactivos en los dos extremos de la distribución de dificultad (muy fáciles y muy difíciles), y la carrera de Nutrición y Dietética es la que presenta más ítems al lado negativo de la escala (ítems más fáciles). Dentro del grupo de las ingenierías la prueba de Ingeniería Industrial tiene una tendencia a presentar ítems mas fáciles, mientras que las de Ingeniería Civil, Alimentos y Electrónica no presentan ítems particularmente dificiles y las Ingenierías Metalúrgica y

Eléctrica son las que presentan más ítems dificiles. En cuanto al área de Ciencias Sociales y Humanas se encuentra que la prueba de Psicología presenta una mayor cantidad de ítems con dificultad inferior a -2 mientras que Derecho contiene más ítems difíciles, aunque no en la misma proporción de los fáciles.

La distribución de los niveles de dificultad de los ítems dentro de cada prueba mostró que dentro del grupo de Ciencias de la Salud la prueba de optometría es la que cuenta con una mejor distribución de los ítems dado que mide en todos los niveles de la escala, por lo que podría afirmarse que es la prueba que realiza una mejor medición, las demás pruebas a pesar de tener una distribución similar, presentan algunos niveles sin medir, es decir, que no tienen ítems en algunos niveles de magnitud de atributo. Del grupo de Ingeniería, Arquitectura, Urbanismo y Afines, la prueba que tiene mejor distribución es la de Ingeniería de Alimentos ya que se encuentran ítems que miden a lo largo de la escala, mientras que la prueba de Arquitectura cuenta con un mayor rango de habilidad medida pero presenta regiones de la escala sin ítems. Por su parte las pruebas de Ingeniería Electrónica, Ingeniería Geológica e Ingeniería Metalúrgica muestran concentración de sus ítems en regiones específicas de la escala de habilidad, la primera de ellas, que cubre menos nivel de habilidad, tiene sus ítems concentrados en los valores medios, la segunda muestra mayor concentración en los niveles superiores de habilidad y la de Ingeniería Metalúrgica muestra mayor concentración de ítems en los niveles inferiores de la habilidad. Finalmente, dentro de las pruebas del grupo 3 se encuentran las dos pruebas con mayor número de ítems: Derecho y Psicología, que sin embargo, no presentan ítems que midan todos los niveles de habilidad. Esto se observa especialmente en las subpruebas de psicología, siendo las de Psicología Social y Jurídica las que tienen una concentración mayor de los ítems hacia los niveles inferiores de la escala, mientras que Psicología Clínica y Organizacional mostraron una distribución de ítems más regular aunque no miden en todos los niveles de habilidad.

Discusión y conclusiones

Un primer resultado que llama la atención cuando se analizan los ítems con métodos de la TCT, es que entre un 10.7% y un 59.2% de los ítems ameritarían modificar sus distractores. Surge entonces la necesidad de hacer más énfasis en la construcción de las preguntas para garantizar que los distractores, en efecto, atraigan a quienes no poseen la maestría en el dominio. En esta dirección es posible sugerir que, atendiendo a los estudios realizados por Haladyna (1993/94) se modifique el número de distractores y además en el proceso de revisión por parte de los expertos en elaboración de pruebas y en cada área disciplinar, se haga un análisis cualitativo de los distractores luego de la aplicación con el fin de retroalimentar a los constructores de preguntas y con ello mejorar la calidad de los ítems en la siguiente prueba. La revisión de los distractores es necesaria dado que si estos presentan errores se afecta el cálculo de los demás parámetros. En cuanto al parámetro de discriminación de la TCT, sobresale el bajo porcentaje de ítems que diferencian entre quienes tienen más competencia o conocimiento y quienes no. Un factor influyente en este resultado es que se está teniendo en cuenta el puntaje total para cada prueba. Sería más conveniente realizar un análisis de los puntajes arrojados en cada escala ya que la prueba en su totalidad está evaluando gran

diversidad de contenidos, como puede apreciarse en el anexo y no permite diferenciar en qué áreas los ítems discriminan más que en otras.

Respecto a los análisis de las pruebas como un todo, las 27 pruebas mostraron confiabilidad satisfactoria calculada con el coeficiente Alfa de Cronbach, asumiendo los rangos de confiabilidad establecidos por el Laboratorio de Psicometría de la Universidad Nacional de Colombia (Fernández, Flórez & Villate, 2004). Además, después de conocer el proceso seguido por los distintos programas para la construcción de las pruebas se puede concluir que las mismas cuentan con validez de contenido aceptable, lo cual constituye un resultado importante teniendo en cuenta que ésta es una condición necesaria, aunque no suficiente, a la hora de evaluar las pruebas educativas. Sin embargo, se hace necesario realizar un análisis más exhaustivo de aspectos fundamentales tales como: ¿Qué es lo que miden realmente las pruebas: Conocimientos o competencias?, ¿Existe concordancia entre las calificaciones de los jueces y los resultados psicométricos de parámetros como la dificultad y la discriminación?, ¿Cuál es la validez concurrente de estas pruebas?, ¿Sería adecuado considerarlas como predictoras de una conducta futura?, ¿Cuál?. Dada la estrategia metodológica adoptada en el presente estudio, no se abordó con otro criterio el problema de la definición del dominio evaluado, y la relevancia y pertinencia de los ítems en concordancia con tal definición. En consecuencia, falta más evidencia empírica que sustente la validez de estos instrumentos y que garantice la calidad de las inferencias que se hacen a partir de los mismos.

Por otra parte, cuando se ajustan modelos de Rasch, aproximación elegida por el Servicio Nacional de Pruebas como responsable de este tipo de procesos, se encuentra que la mayoría de pruebas alcanzan los niveles empíricos mencionados por Tristán (1998) y a pesar de que hay un buen ajuste en todas, éste no se logra en su totalidad dado que no coinciden las medias y las desviaciones estándar de evaluados y preguntas. Sin embargo, los ítems que conforman cada una de las pruebas cuentan con los adecuados valores de ajuste que permitiría concluir que el modelo de Rasch es una aproximación apropiada para el análisis de estas pruebas. Sin embargo, vale la pena preguntarse como puede ser asumida la unidimensionalidad en pruebas educativas que apuntan a medir los desarrollos académicos de los estudiantes. También resulta de relevancia preguntarse por la forma en que este modelo y los resultados que se elaboran a partir del mismo contribuyen a realizar una verdadera medición con referencia a criterio. En términos generales, vale la pena subrayar cómo la calidad de los ítems en los dos modelos arroja conclusiones disímiles, según los procedimientos de la TCT una cantidad importante de ítems no cuentan con la especificaciones psicométricas requeridas, resultado que sucita interrogantes como: sabiendo que los dos modelos son reconocidos por los expertos en el campo como complementarios mas que opuestos, ¿Cuál de los dos debe ser utilizados para el análisis de instrumentos educativos? Dado que el modelo de Rasch sólo tiene en cuenta el parámetro de dificultad y asume igual discriminación para todos los ítems, ¿No sería más conveniente utilizar un modelo de dos parámetros o el de tres parámetros que permita conocer esta propiedad tan importante de los ítems?

Finalmente, partiendo de los resultados obtenidos, pero trascendiéndolos, en opinión de los autores otras conclusiones posibles son:

Las pruebas aún deben ser revisadas teniendo en cuenta su carácter experimental, hecho que debería hacerse explícito a la hora de divulgar los resultados.

Los instrumentos reflejan la realidad educativa nacional en educación superior ya que en ésta también se encuentra una gran diversidad curricular.

No son homologables los instrumentos en cada una de las disciplinas entre sí, si bien responden a unos mismos procesos y especificaciones técnicas, dan cuenta de desarrollos disciplinares distintos.

Debe ser madurado el proceso mismo incluyendo lo curricular y lo evaluativo, particularmente en lo referente al objeto evaluado: ¿competencias ó conocimientos?.En este mismo sentido, también es importante reflexionar respecto a lo que las pruebas miden en su aspecto cognitivo, es decir en cuanto a las llamadas competencias. Sin duda un enfoque de competencias implica una reconcepción del sistema educativo en el currículo, la pedagogía y evaluación que peermita contextualizar el "saber hacer" de un estudiante de pregrado. En la medida en que las pruebas ECAES garanticen calidad técnica, éstas se convertirán en un buen criterio para evaluar la educación superior, teniendo en cuenta siempre que no debe ser la única herramienta utilizada para tal fin y que por tanto debe ser acompañada de otros indicadores como los del Sistema Nacional de Acreditación que, interpretados en conjunto, arrojarán mejores y más confiables resultados respecto a la calidad de la educación superior.

Referencias

ACOFI. (2003). *Informe general de proceso*. Documento de trabajo. Bogotá

Aiken. (1996). *Test psicológicos y evaluación*. Mexico: Prentice hall.

Attali, Y, Fraenkel,T.(2000). The point biserial as discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of educational measurement*, 37, 77-86.

Brown, F. G. (2000). *Principios de la Medición en Psicología y educación*. México: Manual Moderno.

Cano, F. (2004). *Construcción de pruebas de conocimientos*. Trabajo presentado en el Seminario Internacional: Compromiso de la evaluación objetiva con el mejoramiento de la calidad de la educación superior. Bogotá. ACOFI- Asociación Latinoamericana de Psicología.

Capper, J. (1996). *Testing to Learn – Learning to Test*. USA:. International Reading Association.

Fernandez, M; Flórez, N y Villate, C. (2004). *Validez y confiabilidad del perfil sensorial de Winnie Duna, 1999*". Tesis de grado Inédita. Universidad Nacional de Colombia.

Gago, A. (2000). El CENEVAL y la evaluación externa de la educación en México. *Revista Electrónica de Investigación Educativa*, *2* (2). Consultado en marzo de 2004 en el World Wide Web: http://redie.ens.uabc.mx/vol2no2/contenido-gago.html

Guilford (1975) *Psycometric Methods*. 2a Edición,. Bombay, Nueva Delhi. Tata McGraw-Hll

Haladyna, T. (1994) *Developing and validating multiple -choice test items.* New Jersey. Lawrence Erlbaum Associates, Plubichers.

Haladyna,T y Downing, S. (1993). How many options is enough for a multiple-choice test item?. *Educational & Psychological Measurement*, 53(4), 999-1010.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Herrera, A. N. (1998). *Notas de Psicometría*. Documento Inédito. Bogotá: Universidad Nacional de Colombia

Herrera, A. N., Sánchez, N & Jiménez, H. (2001). De la Teoría Clásica de los Test a la Teoría de Respuesta al Item.. *Aula Psicológica 3, 293-332*.

Kane, M. (2001) Current concerns in validity theory. *Journal of educational Mesurement*, 38(4), 319-342.

Linacre, M. (2003). *A user's guide to Winsteps*. Chicago: Mesa Press.

Lord, F. y Novick, M. (1968). *Statistical theories of mental test scores*. USA Addison-wesley publishing company.

McGartland, Berg-Weger, Tebb, Lee, s & Rauch, S. (2003). Objectifying Contet Validity: Conducting a content validity study in socialwork. *Reserch. Social Work Reserch*, 27(2), 94 – 104

Martínez, R. (1995). *Psicometría: Teoría de los Test Psicológicos y Educativos*. España: Síntesis.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons'Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741–749.

Ministerio de Educación Nacional-MEN (1997). *Evaluación en el aula y más allá de ella*. Bogotá: MEN

Ministerio de Educación Nacional-MEN. (2004). *Educación Superior*. Boletín Informativo No. 2 Mayo – Junio. Bogotá: MEN

Montoya, D y Rodríguez, W. (2004). ¿Son los ECAES un medio de estimación de la calidad en la formación de psicólogos?. Ponencia presentada en el 11º Congreso Colombiano de Psicología, Neiva-Colombia.

Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid. Pirámide.

Muñiz, J. (2000). *Teoría Clásica de Los Test*. 6ª Ed. Madrid. Pirámide.

Otálora, B. (2004). Comunicación personal Junio de 2004. Bogotá

Paz, M. (1996). *Validez*. En Muñiz, J (Ed.) Psicometría. Madrid. Universitas.

Pardo, C. (2001). *El modelo de Rasch: una alternativa para la evaluación educativa en Colombia*. Acta Colombiana de Psicología, 5, 9-21.

Powers (2004). Validity of graduate record examinations (GRE) general test admissions to colleges of veterinary medicine. *Journal of applied psychology*, 89 (2), 208-219.

Ravela, P., Wolfe, R., Valverde, G. y Esquivel, J. (2001). Los próximos pasos. ¿Cómo avanzar en la evaluación de aprendizajes en América Latina?. Programa de Promoción de la reforma educativa en América Latina y el Caribe PREAL. No. 20. Chile - Santiago.

Servicio Nacional de Pruebas (1992) *El S.N.P. - Sus programas de Evaluación.* Documento S.N.P. No. 65, Santafe de Bogotá: Documento institucional.

Triana, A. (1996). La evaluación de los sistemas educativos. *Revista Iberoamericana de Educación* 10 – Consultado el junio de 2004 en el World Wide Web: http://www.oei.es/

Tristan, A. (1998). *Análisis de Rasch para todos*. México: CENEVAL.

Watermann, R y Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms construct validation and proficiency scaling in TIMMS. *European Journal of psychological assessment,* 18(3), 190-203.

Manuscrito recibido en octubre de 2004
Aprobado para publicación en Enero de 2005

Anexo

Contenidos referenciales de los 27 instrumentos analizados

Prueba	Contenidos
Odontología	Periodoncia
	Endodoncia
	Farmacología y Terapéutica
	Fisiología
	Patología
	Cariología
	Ética/bioética
	Salud Pública
	Admón. en salud
	Crecimiento y desarrollo
	Rehabilitación
	Cirugía
Fisioterapia	Promoción de la Salud y 1er Nivel
	Segundo Nivel
	Tercer Nivel
	Administración y Gestión
	Investigación
	Promoción de la Salud y 1er nivel Sub2
	Segundo Nivel Sub2
	Tercer Nivel Sub2
	Admón y Gestión Sub2
	Investigación Sub2
Enfermería	Ciencias biológicas
	Profesional
	Investigación
	Ética y bioética
	Educación en enfermería
	Administración y Gerencia
Ingeniería Agronómica	Ciencias Básicas
	Ingeniería
	Fitotecnia y Producción
	Ciencias Socioeconómicas
Demás ingenierías	Formación Básica
	Formación en Ciencias Básicas de Ingeniería
	Formación Profesional
Fonoaudiologia	Lenguaje
	Asuntos Profesionales
	Habla
	Audición
Medicina	Acciones médico legales
	Salud Pública y Medio Ambiente
	Atención al Individuo y la Familia
	Etica y Bioética
	Acciones Administrativas
Terapia Ocupacional	Disfunciones psicosociales
	Disfunciones Físicas
	Disfunciones psicosociales Sub2
	Disfunciones Físicas Sub2
	Educación
	Trabajo
	Fundamenta-ción profesional
	Fundamentación profesional Sub2
	Educación Sub2
	Trabajo Sub2
Optometría	Ciencias Optométricas Prof. Básicas
	Ciencias Optométricas Básicas Prof. Complementarias y Humanísticas
Nutrición y Dietética	Nutrición Normal
	Nutrición Clínica
	Alimentos y Servicios de alimentos
	Nutrición en Salud pública
Arquitectura	Expresión y Representación Gráfica
	Urbano Ambiental
	Teoría e Historia
	Tecnología
	Ejercicio Profesional

Anexo (*Continuación*)

Prueba	Contenidos	**Prueba**	Contenidos
Psicología	Historia de la Psicología, Epistemología. Modelos Teóricos	**Derecho**	Teoría General del Derecho
	Bases Psicobiológicas del comportamiento		Responsabilidad profesional
	Procesos Psicológicos Básicos		Derecho Constitucional
	Bases Socioculturales del Comportamiento		Derecho Administrativo
	Problemas Fundamentales de la Psicologia Individual		Derecho Internacional
	Problemas Fundamentales de la Psicología Social		Derecho Laboral
	Psicología Evolutiva		Derecho Civil y de Familia
	Medición y Evaluación Psicológica		Derecho Comercial
	Formación Investigativa		Derecho Penal
	Formación Profesional Común		