NUEVO

EXAMEN DE ESTADO

PARA EL INGRESO A LA EDUCACIÓN SUPERIOR

Cambios para el siglo XXI

Transformaciones en las pruebas para obtener resultados diferentes

República de Colombia
MINISTERIO DE EDUCACIÓN NACIONAL

Revolución educativa en marcha
¡La educación ya! Asunto de todos

ICFES

Transformaciones en las pruebas para obtener resultados diferentes

República de Colombia
MINISTERIO DE EDUCACIÓN NACIONAL

ICFES
INSTITUTO COLOMBIANO PARA EL FOMENTO DE LA EDUCACIÓN SUPERIOR

ANDRES PASTRANA ARANGO
Presidente de la República

GERMAN ALBERTO BULA ESCOBAR
Ministro de Educación Nacional

PATRICIA MARTINEZ BARRIOS
Directora General ICFES

TANIA MARGARITA LOPEZ LLAMAS
Secretaria General

PATRICIA ASMAR AMADOR
Subdirectora General Técnica y de Fomento

ANTONIO FRANCISCO MERLANO VELILLA
Subdirector General de Planeación

CARLINA MALDONADO DE LOZANO
Subdirectora General Jurídica

MARIO AGUIRRE BERMUDEZ
Subdirector General Administrativo y Financiero

FRANCISCO ERNESTO REYES JIMENEZ
Subdirector General de Informática

MAGDALENA MANTILLA CORTES
Sudirectora General del Servicio Nacional de Pruebas

JAIRO FERNANDO PAEZ MENDIETA
Jefe División de Administración de Exámenes

CLAUDIA LUCIA SAENZ BLANCO
Jefe División de Desarrollo de Pruebas

Autor

CARLOS ANTONIO PARDO ADAMES

Corrección de Estilo

ROBERTO PINZON

Contenido

Transformaciones en las pruebas para obtener resultados diferentes

República de Colombia
MINISTERIO DE EDUCACIÓN NACIONAL

ICFES
SERVICIO NACIONAL DE PRUEBAS

INTRODUCCION

n abril de 1991, el Ministerio de Educación Nacional de Colombia, solicitó al Servicio Nacional de Pruebas el diseño de pruebas para evaluar la calidad de la educación en el país, tema que empezaba a tener una amplia discusión en el sector educativo. Ya desde la primera reunión se acordaron diferentes aspectos que han marcado no sólo el desarrollo del programa de evaluación de la calidad de la educación en Colombia, adelantado por el MEN y el SNP, sino el desarrollo de otros programas de evaluación educativa del SNP, como el de Exámenes de Estado para Ingreso a la Educación Superior.

Aspectos como los cuestionarios de factores asociados al logro al logro desarrollados por el MEN, y el diseño de pruebas para evaluar el logro cognitivo[1], realizado por el SNP, marcaron diferencias sustanciales con el trabajo de evaluación educativa que se había desarrollado hasta el momento en Colombia.

Otra discusión, con repercusiones importantes, la realizaron los profesionales del SNP sobre el marco de interpretación de resultados, de tal manera que éstos resultaran significativos en el contexto de evaluación de la calidad de la educación. Se concluyó que la interpretación con referencia a criterio o nivel de logro cognitivo era la más adecuada para los propósitos del programa. Esta conclusión marcó una gran diferencia con respecto al trabajo evaluativo que había caracterizado al SNP en sus más de 20 años de experiencia en la época, ya que siempre había realizado pruebas cuyos resultados se interpretaban con referencia a la norma, esto es, que comparan la ejecución de un estudiante con respecto a la de los demás estudiantes que abordan las pruebas con él, e indican el porcentaje de personas que un alumno supera.

Se inició el diseño de las pruebas de matemáticas y lenguaje con una amplia reflexión sobre los marcos conceptuales de las disciplinas, con la participación de especialistas de amplia trayectoria teórica y de investigación en educación y evaluación. Para ello se hizo referencia no sólo a las concepciones modernas de educación y de la formación en lenguaje y matemáticas, sino también, a los

[1] Sobre los elementos conceptuales que fundamentaron esta evaluación en su momento se puede consultar: Flor Alba Cano. 1997. Factores Asociados al Logro Cognitivo de los Estudiantes. Grados 3° y 5° 1993 – 1994. Serie Publicaciones para Maestros. MEN. ICFES.

marcos legales y curriculares del momento, en especial los Fundamentos Generales del Currículo y el decreto 1002 de 1986, más conocido como Nuevo Currículo. Estos marcos conceptuales se han transformado desde esa época y han contribuido a la discusión y reflexión general de la educación en matemáticas y lenguaje[2].

Aunque en 1991 el MEN programó la evaluación de estudiantes de 3° y 5° grados de educación básica en matemáticas y lenguaje, el SNP sólo diseñó las pruebas para 5° grado. Con el tiempo el SNP asumiría el diseño de pruebas en todos los grados del programa, 3°, 5°, 7° y 9° grados de educación básica.

Otro tema de mucha importancia correspondía al marco matemático de procesamiento de información, de tal manera que fuera coherente con los planteamientos precedentes. En esa época, en Colombia, sólo se conocía el marco de la Teoría Clásica de las Pruebas (TCP) que se basa en el modelo de la curva normal para la producción e interpretación de resultados. No obstante, este modelo resultó inadecuado para los propósitos del programa e inconveniente de utilizar en el contexto planteado.

En el desarrollo de la investigación se identifican transformaciones de la psicometría de mucho impacto, desde principios de la década de los 80, que han tenido una amplia divulgación durante la presente década. Estas transformaciónes corresponden a la utilización de la Teoría Respuesta al Item (TRI) como marco matemático para el procesamiento y la producción de resultados. Esta teoría respondía a los propósitos planteados por el programa de evaluación de la calidad de la educación.

El SNP inicia la utilización de la TRI, y en especial del modelo de Rasch en aspectos relacionados con la confiabilidad, validez, el procesamiento y la interpretación de resultados, lo que, unido al desarrollo de los marcos conceptuales y la evaluación con referencia a criterio, dan pie a un verdadero salto cualitativo en la evaluación educativa en el país, por lo menos desde la perspectiva de la evaluación a grandes poblaciones como la que realiza el SNP del ICFES.

Este trabajo le ha dado cierto liderazgo al país en el contexto latinoamericano. En 1992, en Santiago de Chile, en el marco de la

[2] Los marcos conceptuales actualizados aparecen en la publicación de Resultados de matemáticas y lenguaje en 3° y 5° grados, 1997 – 1998, realizada por el MEN - ICFES

primera reunión sobre sistemas nacionales de evaluación de la calidad de la educación, patrocinada por la UNESCO/OREALC, Colombia muestra ser uno de los países con mayor experiencia y el único que había abordado los temas de la evaluación con referencia a criterio, de la evaluación de competencias y de los modelos TRI aplicados a la evaluación educativa.

Hoy en día, existe un interés generalizado por la incorporación de procedimientos de evaluación que superen la visión de la TCP y varios países utilizan modelos modernos de tal manera que se potencialice el esfuerzo que representa una evaluación de la calidad. Países como Bolivia, Brasil y Chile han ingresado, recientemente, al campo de la evaluación con referencia a criterio y al uso de los modelos TRI.

Adicionalmente, podemos mencionar que el uso de modelos TRI es generalizado en distintos programas internacionales de evaluación como el Tercer Estudio de Medición en Ciencias y Matemáticas (más conocido como TIMMS), el Segundo Estudio Internacoinal de Cívica y Democracia y el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación, en los cuales ha participado Colombia y donde se ha reconocido la calidad del trabajo adelantado en nuestro país.

El programa de evaluación de la calidad de la educación en Colombia, más conocido como SABER, ha permeado otros programas de evaluación que se desarrollan en el SNP, especialmente el de Exámenes de Estado para Ingreso a la Educación Superior, el cual, aunque tiene su propia trayectoria conceptual, retoma discusiones relacionadas con la evaluación de competencias e interpretación de resultados con referencia a criterio que se venían planteando en el programa SABER.

A partir de 1995, se inicia en el Servicio Nacional de Pruebas el proyecto conocido como Reconceptualización del Examen de Estado, que ha tenido como finalidad desarrollar los fundamentos teóricos de las pruebas, las especificaciones de los instrumentos de evaluación y replantear la elaboración y aplicación de estos exámenes.

Este proyecto fundamenta el desarrollo del nuevo examen en tres ejes principales: las transformaciones que se producen en las

disciplinas que conforman el examen; las exigencias culturales, políticas, sociales y económicas que se dan en el contexto de la globalización y la renovación de los propósitos educativos del país.

Un nuevo examen, de acuerdo con la idea del nuevo país, implica una reflexión profunda sobre los referentes teóricos y una transformación seria y responsable en los procesos de diseño de pruebas, administración de exámenes y procesamiento y análisis de resultados. Asi, el nuevo examen tiene como objeto de evaluación las competencias de los estudiantes del país en contextos disciplinares e interdisciplinares y se estructura en dos componentes: el núcleo común, igual para todos los estudiantes, y el componente flexible con una línea de profundización y otra interdisciplinar; sus resultados serán analizados con modelos matemáticos de la Teoría Respuesta al Item.

El diseño del nuevo examen de estado para ingreso a la educación superior conlleva la puesta en práctica de las reflexiones antes mencionadas. Hasta el momento se han experimentado diversos aspectos del nuevo examen en 6 ocasiones. En el segundo semestre de 1997 se aplicaron pruebas del núcleo común a 6000 estudiantes de once ciudades del país, quienes acababan de presentar el examen de estado en el mes de agosto. Durante 1998 se utilizaron preguntas de ensayo del nuevo examen, en los exámenes de estado de marzo y agosto[3]. En 1999 se realizan dos aplicaciones experimentales del núcleo común y de la línea de profundización: la primera en marzo a una muestra de 2000 estudiantes de grado 11 en la ciudad de Bogotá, la segunda a una muestra de 6000 estudiantes en 9 ciudades del país.

Estas experiencias no son sólo de pilotaje de preguntas, también se analizan otros aspectos relacionados con la administración del examen y el procesamiento y análisis de datos. Se indaga cuidadosamente, sobre el impacto de nuevos formatos de ítems, la organización del examen en sesiones, el tiempo disponible por los estudiantes para responder a las preguntas, la longitud de las pruebas en términos del número de ítems de cada una, entre otros.

Un tema de atención especial es el de los resultados, ya que ellos deben reflejar el planteamiento teórico en relación con el concepto

3.Una ampliación del concepto “Preguntas de Ensayo” se encuentra en el número 7 de la Serie Investigación y Evaluación Educativa del SNP del ICFES: El diseño de pruebas para los exámenes de estado: un proceso de investigación permanente. (1998).

de evaluación que se maneja en el nuevo examen[4]. Aunque inicialmente se recurrió a las reflexiones y experiencia del programa SABER, fue evidente la necesidad de la identidad propia del nuevo examen de estado en relación con los aspectos generales y en particular con el tema de los resultados. En este sentido, la necesidad de informar individualmente sobre el desempeño de los estudiantes en el examen tanto al mismo estudiante que lo presenta, como a las instituciones de educación media y universidades, implica pensar en información que pueda impactar mucho más los procesos sociales y educativos tanto en el ámbito personal como institucional.

Si una competencia es un "saber hacer en contexto, es decir, el conjunto de acciones que un estudiante realiza en un contexto particular y que cumplen con las exigencias específicas del mismo"[5], los resultados de su evaluación deben ser expresiones cuantitativas y cualitativas que describan dicha competencia. Se dice resultados (plural) porque una aproximación de esta naturaleza exige una comprensión del evaluado a partir de múltiples perspectivas que permita visualizar (aunque de forma muy esquemática) la complejidad del ser humano.

Asi, se plantean cinco resultados diferentes para cada persona que aborda el nuevo examen de estado:

- Competencia general en cada prueba: es un puntaje que indica el desempeño general del estudiante.

- Desempeño relativo por grupos de preguntas: en cada prueba las preguntas se clasifican de acuerdo con la visión particular de las disciplinas ya sean en ejes, ámbitos, tópicos, o temas, lo que permite generar resultados que describen el desempeño relativo del estudiante en cada grupo.

- Nivel de competencia: en el nuevo examen de estado las competencias se circunscriben a las acciones de tipo interpretativo, argumentativo y propositivo. Para cada una de estas acciones en cada prueba se informa el grado de competencia con su

[4] La evaluación concebida como práctica que puede generar transformaciones de carácter social y cultural debido a su cobertura y a la función que cumple en el establecimiento, mantenimiento y desarrollo de par metros de acción educativa puede contribuir indirecta pero tal vez profundamente, en el reconocimiento de la diversidad de la nación". Nuevo examen de estado: propuesta general. SNP-ICFES. 1999.

[5] Una ampliación sobre este tema se puede hacer en otros documentos de esta serie.

correspondiente descripción.

- Grado de profundización: indica el nivel de complejidad (con su respectiva descripción) que maneja una persona en cada prueba que haya elegido.

- Resultado en la problemática interdisciplinar: es un puntaje que indica el desempeño general del estudiante en la problemática seleccionada.

En el telón de fondo de este planteamiento se encuentra una reflexión sobre los modelos psicométricos utilizados en evaluación educativa, sus fortalezas, debilidades y posibilidades para abordar el nuevo examen, reflexiones que se desarrollan en las siguientes páginas.

LA TEORÍA CLÁSICA DE LAS PRUEBAS

Son las 5:00 a.m. de 138 a. c. Aún está a tiempo Yu Chen. Sin prisa termina de colocarse su vestido de seda, el mejor que tiene. El hecho de estar despierto desde las 3:00 a.m. no le preocupa mucho puesto que los ejercicios de meditación le hicieron descansar. Ya no era tiempo de preocupaciones, ni de arrepentimientos, todo estaba hecho. Salió de su casa como cientos de miles de personas en toda China que, como él, guardaban esperanzas de contribuir al emperador HAN para que realizara un mejor gobierno en el Imperio.

No fue el primero en llegar al sitio indicado. Un rato después ingresaron en orden riguroso de acuerdo con la inscripción que habían realizado previamente y fueron llevados inmediatamente a sus cubículos personales; aquel pequeño lugar que lo albergaría durante 24 horas. Yu Chen se preparó. Colocó la pequeña caja con plumas y tinta encima de la tabla donde escribiría, en donde encontró suficiente papel de arroz para escribir sus respuestas a las preguntas del maestro.

Todos esperaban nerviosos el momento de empezar. Todos querían dar lo mejor de sí para apoyar el trabajo de su emperador en el Distrito que fuera necesario. A las 8:00

a.m. el maestro director empezó a dictar la primera pregunta:

Describa la producción agrícola de la región, enfatizando las condiciones que la favorecen y los cambios necesarios para garantizar una mejor y mayor producción, con sus consecuencias a corto, mediano y largo plazo, incluyendo las transformaciones administrativas, humanas y técnicas…

A Yu Chen no le pareció muy difícil. Le tomaría unas 4 horas responderla, pero la podría abordar.

Luego de escuchar la sexta pregunta (componer un poema a su región) calculó que tendría muy buenas posibilidades de ingresar al servicio civil del emperador y se dio a la tarea de iniciar aquello para lo que se había preparado casi toda la vida.

Evaluaciones como el Examen de Estado para Ingreso a la Educación Superior han existido desde hace mucho tiempo. La historia recoge que esta práctica ya se realizaba en China 2200 años A.C. (DuBois, P. H.1970) cuando el emperador de la dinastía Shang, hacía que sus funcionarios presentaran pruebas para determinar si eran aptos o no para desempeñarse en el servicio civil. Esos exámenes se refinaron hasta que se introdujeron las pruebas escritas en la dinastía Han (202 a. c. hasta 200 d. c.), fecha en la cual se empezaron a evaluar cinco tópicos:

- **Ley civil**
- **Cuestiones militares**
- **Agricultura**
- **Impuestos**
- **Geografía**

Durante la última etapa de evaluación en la china imperial, de 1374 hasta la dinastía Ching (1644 - 1911)(Greaney, V. Y Kellaghan T. 1995), acontecieron ciertos sucesos que resultan curiosos desde una perspectiva como la del actual Examen de Estado colombiano. Dos grandes expertos chinos en evaluación, en el siglo XVII, Ku Yen-wu y Huang Tsung-his, llamaban la atención al hecho que la impresión comercial de libros (apenas en sus comienzos) podría

influir negativamente en la educación, en el sentido de favorecer la memorización en comparación con otras artes intelectuales; su consecuencia funesta: una menor calidad en las respuestas a los exámenes para ingreso al servicio civil.

Inclusive se presentaron un par de escándalos en los siglos XVI y XVII. En 1595, la persona que obtuvo el segundo lugar había copiado fragmentos de ensayos que aparecían en los libros de preparación para el examen. En 1616, quien ocupó el primer lugar fue descalificado, debido a que había copiado totalmente el ensayo de T'ang Pin-yin con el que obtuvo el primer lugar en 1595 y que apareció en uno de los libros de preparación para el examen.

Pero no todo es malo. Estos exámenes influyeron notablemente en el desarrollo intelectual del imperio Chino y en la divulgación de las ideas a través del material impreso, más allá de lo poco que podía imprimir el estado. Benjamín Elman (miembro del consejo de estudios asiáticos de la Universidad de Harvard) argumenta que "los exámenes fueron una obra maestra de la producción social, política y cultural que mantuvo cohesionada a China durante 400 años" en una de las más importantes dinastías de la historia.

Con esta perspectiva, que es utilizada inclusive por algunas formas de evaluación de nuestra época inclusive, se realzaba el concepto de pasar y no pasar o de perder el examen. Y entre quienes pasaban, se elegían los mejores a partir de un único resultado.

Desde principios de siglo se utiliza la curva normal como un modelo que permite representar lo que acontece con los resultados en una evaluación en educación o en algún campo de la psicología[6]. Sólo hasta la década de los 80 se reconoce el hecho que esta distribución de los resultados obedece a la distribución de los niveles de dificultad de las preguntas utilizadas en las pruebas(Tyler, R. 1988), es decir que se pone en duda el hecho aceptado comunmente de que los atributos medidos en poblaciones se distribuyen normalmente.

La TCP(Lord, F. Y Novick, M.1968) se centra en la estimación del puntaje de una persona como si esta hubiera respondido al universo total de preguntas posibles. Como este universo es infinito, es necesario hacer una estimación de ese puntaje, el cual tendrá cierta cantidad de error.

[6] De ahí surgen las escalas de medición de la inteligencia, la personalidad, las aptitudes que se utilizan en psicología y que consideran que la distribución de los resultados es semejante a la curva normal.

De acuerdo con Lord y Novick (1968), el desempeño de una persona en un conjunto de ítems de una prueba puede observarse a partir del puntaje, que es la suma de las preguntas que haya respondido correctamente. La proporción de respuestas correctas, si las preguntas se han obtenido aleatoriamente del universo, es un estimador insesgado de la proporción de preguntas en el universo que una persona puede responder correctamente. De esta forma sería posible conocer el rendimiento de una persona en una disciplina particular, medido a través de una prueba, a partir de las respuestas que dé a cualquier conjunto de ítems (la prueba) obtenido aleatoriamente del universo de preguntas. Esperaríamos que el resultado en cualquiera de estos conjuntos fuera semejante. Aquí lo importante es la validez de contenido. Pero éste no es el caso en la TCP1, debido a que los ítems se construyen de acuerdo con intenciones particulares y no se hace una selección aleatoria del universo.

En el SNP del ICFES, en el examen vigente hasta 1999, se arman o estructuran las pruebas con base en un banco de items. Del total de preguntas en un área específica se seleccionan aquellas que cumplen con requisitos conceptuales y estadísticos, y con ellas se conforma la prueba definitiva. De una parte, se eligen preguntas que correspondan a la estructura de prueba establecida (en relación con tópicos y procesos de solución) y que tengan un nivel de dificultad[7] particular y una covarianza específica, de tal forma que se garantice una distribución normal de puntajes. Así, la prueba que se aplica en forma definitiva se construye intencionalmente, no aleatoriamente.

Desde la perspectiva de la TCP no pueden hacerse mediciones sin algún error, el cual proviene de diferentes fuentes, como las variaciones propias de las condiciones de aplicación de pruebas, las diferencias en las formas de las pruebas, las variaciones en las ejecuciones de los estudiantes y otros factores desconocidos.

Supongamos que aplicamos una prueba repetidamente a una misma persona (pensemos que las mediciones son independientes unas de otras y que son idénticas esto es, que la estructura probabilística del experimento no cambia de una aplicación a otra). Podríamos decir que el puntaje de la persona en las diferentes

[7] Entendiendo este nivel de dificultad como la proporción de personas que responden correctamente cada pregunta.

aplicaciones corresponde a su "puntaje observado", mientras que el valor esperado, calculado a partir de estas observaciones lo llamaremos "puntaje verdadero". El error corresponde a la diferencia entre el puntaje observado y el puntaje verdadero[8] y funciona como un índice de la exactitud en la medición. Si el error estándar de medición tiene una magnitud pequeña, significa que el puntaje observado es muy semejante al puntaje verdadero. Debemos recordar que el significado de la magnitud se encuentra en relación con la escala de medición utilizada y que tiene relevancia sólo en el contexto de una población específica, ya que ésta se constituye en un evento condicionante.

DEBILIDADES DE LA TEORIA CLASICA DE LAS PRUEBAS

Durante muchos años, la TCP proporcionó soluciones a la medición educativa concibiéndola como realizable a través de comparaciones entre las personas medidas para determinar quien mide más y quién mide menos y estableciendo las distancias (en términos de población) entre unos y otros[9]. Cuando se piensa que la medición de una persona es indpendiente de la medición de las demás, es decir que es posible reconocer lo que la persona sabe hacer independiente de lo que las demás personas saben hacer[8], existen diversos aspectos de la TCP que pueden considerarse como debilidades en este contexto y que mencionaremos a continuación.

A • Los valores de las estadísticas de items y test dependen de la muestra de examinados.

Es importante, en el diseño de pruebas, contar con estadísticas que describan los ítems de tal forma que se puedan elegir los mejores para conformar la prueba definitiva. Estos estadísticos se obtienen a partir de aplicaciones piloto o de ensayo con muestras representativas de la población a la cual se aplicará la prueba en forma definitiva. Con las palabras "muestras representativas" se alude a grupos de personas en los cuales las estadísticas de los items permanecerán invariables. No hay utilidad si en un pilotaje o ensayo se obtienen ciertos estadísticos

[8] "La tradición psicométrica ha nombrado a esta entidad como "puntaje verdadero" aunque no posee ninguna propiedad empírica o teórica que sugiera esta terminología. Novick, M y Jackson, P.(1974).

que no tendrán ninguna relación con la aplicación en la población definitiva.

De acuerdo con Novick y Jackson (1974), los parámetros considerados en la TCP son la dificultad, su poder de discriminación y su validez.

- dificultad del ítem

Algunos autores utilizan el término de nivel de facilidad de una pregunta en lugar del de dificultad, por cuanto permite apreciaciones directas a partir de la lectura del número que la expresa, el cual va de cero a uno. El índice de dificultad de una pregunta hace referencia al grado en el cual una población la responde correctamente y corresponde a la proporción de individuos que responde correctamente el ítem.

Se calcula por medio de:

$$P_i = \frac{SU_i}{N}$$

Donde:

Pi = índice de dificultad de la pregunta

Ui = respuestas a la pregunta
Si la respuesta es correcta, Ui = 1
Si la respuesta es incorrecta, Ui = 0

N = total de personas que abordan la pregunta

Hoy en día, diversos autores consideran que este no es en realidad un índice de dificultad del ítem sino que es un índice que nos informa sobre la población, y por lo tanto lo llaman “proporción de respuestas correctas”. Pero, como dicen Stenner, Smith y Burduick, “la tradición prevalece”.

-discriminación del ítem

Es la varianza, que en una variable dicótoma (como las respuestas a los items de una prueba) se puede expresar en términos de la

proporción de personas que responde incorrectamente una pregunta (Nunnally, J. 1987). Se calcula a partir de:

$$s_i^2 = p_i q_i$$

Donde:

pi = índice de dificultad de la pregunta
qi = 1 - pi

Es claro que la magnitud de la varianza es determinada por cualquiera de los dos valores p ó q; también lo es que el valor máximo es de 0.25 que ocurre cuando p y q valen 0.5 cada una, y este valor disminuye a medida que p ó q se desvían de este punto.

-validez del ítem

Es la correlación biserial puntual entre la respuesta dada a un ítem y el puntaje obtenido en la prueba, excluyendo el aporte de la pregunta al puntaje total en la prueba. Se hacen inferencias a un dominio de contenidos.

Se calcula a partir de:

$$r_1(t\text{-}1) = \frac{rt_1 St - S_1}{\sqrt{St^2 + S_1^2 - 2S_1 Srt_1}}$$

Donde:

rt1 = correlación del item 1 con la puntuación total, incluido dicho item

St = desviación estándar de la prueba

S1 = desviación estándar del item

Esta correlación tiende a ser mayor si se estima en una muestra con personas de habilidad heterogénea.

Como se puede observar, estos índices dependen de la muestra seleccionada para el análisis, lo que implica que los requerimientos planteados al diseño de la muestra y que tienen que ver con la selección aleatoria y probabilística, se cumplen estrictamente.

B • La comparación de los examinados se limita a situaciones en las cuales se les administre el mismo test (o uno paralelo).

Lord (1968) define pruebas paralelas como aquellas que miden exactamente lo mismo, en la misma escala, y que miden con la misma precisión a cada persona. En términos generales la TCP ha definido condiciones estrictas para considerar que dos pruebas son paralelas, las cuales incluyen la igualdad de promedios, varianzas y covarianzas en las diferentes formas (Embretson, S. 1999) . Cuando dos pruebas difieren ligeramente, en cualquiera de estos aspectos, se hace imposible comparar los resultados de las personas que los abordan.

Desde la perspectiva de la TCP, por la razón mencionada anteriormente, no es posible comparar resultados de personas que responden a pruebas con nivel de dificultad diferente.

Ultimamente ha cobrado fuerza la posibilidad de realizar estudios de seguimiento para evaluar el impacto de las transformaciones realizadas en educación (ya sea currículo, prácticas educativas, pedagogía, métodos, etc.). Para ello se requiere recolectar información semejante a lo largo del tiempo y compararla con alguna metodología especial.

En Colombia, el examen de estado vigente hasta 1999 para ingreso a la educación superior es ideal para realizar un seguimiento, puesto que cumple con la condición anterior y, además, es de excelente calidad técnica, de tal manera que se garantizan resultados significativos. Además, cumple con las condiciones estadísticas mencionadas anteriormente, de tal forma que permite la comparabilidad de resultados año tras año y posibilite mantener, de esa forma, la equivalencia de los puntajes. No obstante, el esfuerzo que implica desarrollar pruebas con estas características limita posibilidades en otros campos.

Para realizar la comparación año tras año, se requiere del procedimiento de equating (de esta manera se podrían ver los

cambios en la población para lo que mide una prueba en particular), lo que permitiría concluir si existen verdaderos cambios en algún aspecto y, si es del caso, encontrar factores asociados con estos cambios. Pero, como se ha mencionado, es difícil, si no imposible en algunos casos, realizar este procedimiento bajo los preceptos de la TCP.

C • La confiabilidad, un concepto fundamental, se define en términos de formas paralelas de un test.

La TCT asume que se pueden construir formas paralelas de una prueba (Bejar, I. 1983), esto es, que midan y evalúen de la misma forma. Se considera que, aunque las formas tengan diferentes preguntas, el hecho de hacer semejantes los índices de dificultad y discriminación de los items garantiza una medición igual con ambas formas (confiabilidad), lo cual impide concebir que una prueba tenga más de una confiabilidad.

Teóricamente, la confiabilidad se define como la proporción entre las varianzas de la puntuación verdadera y de la puntuación observada.

$$r_{xx} = \frac{s_r^2}{s_x^2}$$

Podemos observar que el coeficiente de confiabilidad es cero sólo si la varianza de la puntuación verdadera es cero, lo que subraya la dependencia de este coeficiente de la muestra de donde se obtienen los datos estadísticos; también se puede observar que el valor de 1 (máximo) se alcanza cuando ambas varianzas son iguales.

Como lo mencionan Lord y Novick (1968), estos aspectos sugieren que la confiabilidad, en la TCP, es un concepto genérico que se refiere a la precisión en la medición (equivalencia y estabilidad). De ahí se desprenden los diferentes métodos considerados para su estimación y que mencionan autores como Mehrens y Lehmann (1982).

D • Presume que la varianza de los errores de medición es la misma para todos los examinados.

En este sentido, la Teoría Clásica de las Pruebas asigna un mismo error de medición a todos los puntajes (todas las personas que la abordan). No provee información acerca de la precisión de cada puntaje.

E • La TCP ha proporcionado soluciones satisfactorias a varios problemas de medición con pruebas.

Entre esos problemas se encuentran, el diseño de pruebas; la identificación de ítems con funcionamiento diferencial y la comparabilidad de puntajes.

Las pruebas educativas no tienen un sólo y único propósito sino, más bien, apuntan a solucionar diversas problemáticas de acuerdo con las condiciones particulares en donde se presentan. En este sentido, es importante diseñar evaluaciones que respondan a distintas necesidades y que puedan contribuir con información de alta calidad en los diferentes procesos educativos. Esto lleva a pensar que las pruebas se diseñen teniendo en cuenta estas alternativas y necesidades, de tal forma que existan pruebas distintas para diferentes necesidades.

La Teoría Clásica de las Pruebas considera que la mejor prueba es aquella que tiene una mayor varianza y que discrimina mejor a las personas en un punto particular de la distribución (la media). Se basa en el modelo de la curva normal y por lo tanto exige ciertas particularidades en cuanto a la distribución de resultados y a los parámetros estadísticos estimados a partir de los puntajes de las personas. Desde este punto de vista, la TCP no puede abordar pruebas que se apliquen a poblaciones de quienes se esperan resultados con distribuciones diferentes a la normal, casos de alta frecuencia en la evaluación educativa.

El funcionamiento diferencial de preguntas (DIF, por sus siglas en inglés) hace referencia a que una pregunta pueda ser desventajosa a un grupo cultural específico, en otras palabras, a que personas de igual habilidad (con respecto del constructo que mide la pregunta) pero de diferentes culturas muestren o tengan distintas probabilidades de responder correctamente la pregunta,

ocasionando una sub o sobreestimación de sus habilidades (Jonhson, E. 1990).

Existen muchos métodos que permiten estudiar el funcionamiento diferencial de las preguntas, pero, en general, ninguno derivado de la TCP ofrece resultados satisfactorios.

La comparabilidad de resultados en dos aplicaciones diferentes se realiza a través del procedimientos de comparabilidad. En algunas ocasiones es necesario que los resultados de exámenes de diverso tipo sean comparables en el tiempo, de tal manera que se puedan mantener criterios en ese intervalo, como, por ejemplo, en los procesos de admisión universitaria. Como se ha mencionado, la TCP puede hacerlo, pero sólo si se cumplen ciertas características estadísticas de las pruebas.

EN BUSCA DE OTRA ALTERNATIVA

on el nuevo examen de estado se busca ampliar las posibilidades que ofrece la evaluación educativa a través de pruebas de aplicación a grandes poblaciones. Se busca conocer, a través de ellas, el impacto de los cambios realizados e implementados en todo el sector educativo, conocer con mayor amplitud al estudiante colombiano y ofrecer distintas posibilidades para generar cambios educativos, entre otros muchos aspectos.

Para ello es necesario superar las debilidades de la TCP e inclusive dar otro matiz diferente al concepto de medición y evaluación cuando se vincula a la aplicación de pruebas de lápiz y papel a grandes poblaciones.

Se busca la posibilidad de tener estadísticos de ítems y prueba que sean independientes de las poblaciones en las cuales se apliquen, de tal forma que se pueda hacer el seguimiento necesario a los resultados y tener idea del efecto de ciertas políticas o programas educativos. Igualmente es deseable diseñar instrumentos que permitan una gran variedad de resultados para reconocer diversos aspectos de cada una de las personas evaluadas y que resulten significativos en los procesos educativos de la educación media o de la educación superior en Colombia.

Por esto miramos las posibilidades que ofrecen otras opciones metodológicas en el procesamiento, producción e interpretación de resultados para seleccionar aquella que se aproxime más y mejor a las necesidades que se han planteado para el nuevo examen de estado, las que expresemos en términos de los aspectos técnicos a solucionar.

TEORÍA RESPUESTA AL ÍTEM (TRI): UNA ALTERNATIVA

De acuerdo con la experiencia adquirida en el programa SABER, se buscaron alternativas, en el contexto de la TRI, que pudieran dar respuesta a los cuestionamientos técnicos del nuevo examen de estado.

ASPECTOS TÉCNICOS POR SOLUCIONAR

Fueron varias las problemáticas específicas que se les plantearon a los modelos psicométricos para que éstos las resolvieran:

1. Ítems de crédito parcial

En evaluación con pruebas de lápiz y papel es tradicional utilizar las ya conocidas preguntas de selección múltiple con única respuesta, que constan de un enunciado en donde se encuentra el problema por resolver o la pregunta que hay que responder y de varias opciones de respuesta (en el examen de estado vigente a 1999, cada pregunta tiene cinco opciones de respuesta), de las cuales una sola se considera respuesta correcta y las demás son los distractores. Otras alternativas de formato de ítems corresponden a las preguntas de falso - verdadero, las de combinación de opciones para una sola respuesta, la de ordenamiento de opciones, todas con una sola respuesta correcta; es decir que se califican como 1 ó 0: sólo hay una respuesta correcta y las demás son incorrectas.

Con una visión como la del nuevo examen de estado de evaluación de competencias, es posible pensar en preguntas cuyas opciones (todas o algunas) sean respuestas correctas en cierta medida es decir, que las opciones tengan diferentes grados de validez de acuerdo con lo que se pregunta. Es decir que existe algún crédito por responder las opciones que tienen algún valor. De ahí su nombre de ítems de crédito parcial.

Para calificar preguntas de este estilo, la TCP tiene modelos como el de las escalas likert, diseñadas para la evaluación de actitudes, que permiten opciones de diferente calificación, pero la ponderación de las opciones es igual en todas las preguntas; es decir si se elige la primera opción en cualquier pregunta, siempre tendrá el valor de 1, por ejemplo. Debido a lo anterior, su aplicabilidad es muy difícil en preguntas empleadas para la evaluación académica ya que en este caso el valor de las opciones no es el mismo en todas las preguntas. Adicionalmente, es imposible calcular los estadísticos de preguntas como la dificultad, la covarianza y la correlación ítem - test, en la TCP, para este formato de ítems.

2. Puntajes con significado

La fundamentación del nuevo examen de estado exige que los resultados puedan interpretarse de tal forma que adquieran un sentido educativo, es decir que el número tenga significado para la persona.

En la TCP se trató, desde hace mucho tiempo, de darle algún significado a la escala numérica en la que se presentan los resultados. La práctica más utilizada correspondió a la división de la curva normal en nueve partes iguales en su longitud, en desviaciones estándar, división conocida como ***estanines*** (standard nine, por nueve puntos estándar). A cada división o grupo de calificaciones se le asignó una categoría que refuerza la idea de comparaciones con referencia a la norma, es decir, de los alumnos entre sí. Para una revisión de este aspecto se puede consultar el documento N° 7 de la Serie Investigación y Evaluación Educativa, publicado por el ICFES[17].

3. Comparación de resultados en el tiempo

Desde el punto de vista de diseño y aplicación de políticas educativas en distintos niveles, es muy importante obtener información que permita realizar su seguimiento y conocer su impacto, de tal manera que la evaluación sea útil de una forma más variada que la de dar resultados a una persona.

En la TCP, se realiza un procedimiento que permite comparar escalas de calificación en el tiempo, esto es que mantiene coherencia en los puntajes de las personas. De esta forma se garantiza que un puntaje de 300 en un momento particular, en el examen de estado

vigente, equivale a otro puntaje de 300 de otra aplicación: los puntajes son comparables. Esto ha llevado a beneficios generales en términos de la interpretación de promedios por un colegio o de la fijación de un valor como requisito mínimo para inscripción por parte de una institución de educación superior. De no ser así, las instituciones de educación superior tendrían que cambiar sus criterios de puntaje en cada ocasión que hicieran admisiones y los colegios tendrían que utilizar procedimientos complicados para la comparación de promedios.

Pero adicionalmente a esta comparación de escalas, es necesario detectar cambios en las «habilidades» de las personas que indiquen el impacto de programas desarrollados institucionalmente, por ejemplo. Estos cambios se detectan, en la TRI, utilizando el procedimiento de equating.

4. Variedad en los resultados

El esfuerzo que hacen los estudiantes cuando presentan una prueba como el examen de estado no debería tener como único resultado un puntaje que de alguna manera encubre varias de sus capacidades.

Una información más detallada y, ojalá, de tipo cualitativo podría contribuir a reflexiones importantes a nivel personal o institucional en relación con los procesos educativos llevados a cabo hasta el momento o que se pudieran presentar en el futuro. De esta manera las personas podrían tener información que los guiara en el amplio mundo de las oportunidades.

5. Reconocer la pluriculturalidad

Este es un tema de mucha importancia en el contexto nacional, y además son reconocidas la dificultades de las pruebas de lápiz y papel y sus limitaciones para abordarlo.

De alguna manera se podría pensar que el contenido que se maneja en las preguntas u otros aspectos (el formato, la visión de mundo, etc.) pueden influir en las respuestas de una persona. En este sentido es importante contemplar la posibilidad de realizar un estudio que precise la interacción cultura - evaluación.

¿QUE ES LA TEORÍA DE RESPUESTA AL ÍTEM?

urante muchos años, en la evaluación educativa se ha utilizado la Teoría Clásica de los Test para diseñar instrumentos, evaluarlos y utilizarlos en diferentes contextos. Aunque sus fundamentos se plantearon hace varios años, últimamente la psicometría ha hecho énfasis en un nuevo sistema de medición: la Items Response Theory o Teoría Respuesta al Item (TRI), que tiene dos postulados **a)** la ejecución de una persona en una prueba puede predecirse, explicarse por un conjunto de factores llamados habilidades[9] y **b)** la relación entre la ejecución del examinado y las habilidades que la sustentan puede describirse por una función monotónicamente creciente llamada "función característica del item" o "curva característica del item" (ICC). Esto último implica que mientras sea mayor la habilidad de una persona, es mayor la probabilidad de responder correctamente una pregunta.

Todo modelo matemático incluye un conjunto de supuestos acerca de los datos en los cuales se aplica y especifica las relaciones entre los constructos descritos en el modelo (Hambleton y Swaminathan, 1985). En términos generales la TRI considera 3 supuestos básicos:

- **Dimensionalidad**. En la TRI se asume que cuando se diseña una prueba, ésta deberá medir, preferiblemente, una dimensión, una habilidad. Se reconoce que cuando una persona responde a una pregunta en una prueba, entran en juego múltiples habilidades, pero las preguntas deben diseñarse haciendo énfasis en una de ellas o en una combinación particular. Los modelos de dos y tres parámetros requieren de unidimensionalidad en los datos para procesar la información. Además del modelo de Rasch, hoy en día existen otros modelos de escalamiento multidimensional (Borg, I y Greenen, Y 1997).

- **Independencia local**. Este supuesto es aplicable a diversas posturas en relación con la medición educativa, y es que se espera que un estudiante responda a una pregunta en particular sin que recurra a información de otros ítems para hacerlo correctamente. En otras palabras, la ejecución de un

[9] El término "habilidad" se usa, en este documento, en su concepción psicométrica y se refiere "al objeto de medición" que, en el caso del nuevo examen de estado, son las competencias. Se recomienda para ampliar los conceptos de aptitud y competencia leer María Cristina Torrado. 1998. De la evaluación de aptitudes a la Evaluación de Competencias. Serie investigación y evaluación educativa. ICFES.1998.

estudiante en una pregunta no debe afectar sus respuestas en otra. Es práctica generalizada en la actualidad elaborar pruebas en donde se diseñan ítems en relación con un contexto, del cual dependen las respuestas del examinado; aquí también se aplica la independencia local entre los items y no entre ellos y el contexto.

- **Curvas características de ítems**. Son una función matemática que relaciona la probabilidad de éxito, en una pregunta con la habilidad medida por el conjunto de ítems que la contienen (Hambleton y Swaminathan, 1985) . Los diferentes modelos de la TRI se diferencian en la forma particular que adquiere la función de probabilidad, la cual incluye el número particular de parámetros del modelo.

Algunas ventajas de la TRI en relación con la TCP, enumeradas por Hambleton y Cook[21] , son:

- Es posible comparar examinandos aunque hayan abordado diferentes pruebas que midan el mismo dominio.

- Los parámetros de las preguntas son invariantes aunque se estimen en diferentes muestras de la población. En teoría clásica calibrar las preguntas es relevante sólo en el contexto de la muestra donde se realiza.

- Proveen una medida de la precisión en la estimación de la habilidad de cada individuo (o cada grupo de individuos con la misma habilidad) mientras que en la teoría clásica se ofrece un único error estándar de medición que se aplica a todos los estudiantes.

Como dice Engelhard G. (1991), la diferencia entre los modelos logísticos (cuyas raíces se encuentran en los trabajos de Catell (1893) y Thorndike (1904)) y la Teoría Clásica de las Pruebas es que esta última se ha basado en los “puntajes de prueba” mientras que los primeros hacen referencia a “escalas de medición”. En este sentido, la TCP (que tiene sus raíces en Spearman (1904)), se preocupa por la confiabilidad e inclusive, relaciona la objetividad con la forma como se califica una prueba (en otras palabras, la objetividad es un problema de confiabilidad), lo que ocasiona la “paradoja de atenuación” (a medida que un test se hace más confiable, la validez

de los resultados medida por la correlación con una variable criterio, se hace más pequeña) como consecuencia negativa.

EL MODELO DE RASCH

xisten muchos modelos en la TRI, que se diferencian ya sea por la forma matemática de la curva característica del ítem o por el número de parámetros que consideran. Todos los modelos tienen por lo menos un parámetro que describe el ítem y por lo menos uno que describe a la persona.

En este sentido, un modelo de medición "es una función matemática que relaciona la probabilidad de una respuesta correcta a una pregunta con las características de la persona (habilidad) y las características de la pregunta (dificultad)" (Stenner y otros 1983). Es así como el significado de un resultado en una escala particular está dado por el constructo o marco conceptual seleccionado, y no por el modelo en sí. Por lo tanto, un modelo de medición debe cumplir las siguientes condiciones:

- Una persona con habilidad (en términos psicométricos) alta, tiene mayor probabilidad de éxito en un í tem que una persona con habilidad baja.

- Cualquier persona tiene más probabilidad de responder correctamente un ítem fácil que uno difícil (Wright y Mead, 1977).

Como consecuencia directa del cumplimiento de estas condiciones, se encuentra que cualquier parámetro (habilidad, dificultad, etc.) debe ser estimado (calculado) independientemente de los demás parámetros. Esto es, que la habilidad de una persona, pueda estimarse independientemente de las preguntas específicas que responda puesto que su habilidad es la "misma" en un momento particular sin importar si responde a una prueba difícil o a una fácil, por ejemplo.

Uno de estos modelos de la TRI es el modelo de respuesta estocástica de Rasch, que describe la probabilidad del éxito de una persona en un item como una función de la habilidad de la persona y la dificultad de la pregunta, siendo una aproximación estadística al análisis de las respuestas a una prueba y de otros tipos de observación ordinal. Rasch derivó su modelo como una expresión

logística simple y demostró que en esta forma los parámetros de la persona y de la pregunta son estadísticamente independientes.

El análisis por el modelo de Rasch construye mediciones lineales de la habilidad de las personas y la dificultad de las preguntas, al mismo tiempo que establece índices de la precisión y exactitud de la medición (ajuste) (Wright, 1994) . Este modelo especifica que cada respuesta útil en una prueba surge de la interacción probabilística lineal entre la medida de la habilidad de una persona y la medida de la dificultad de una pregunta (Rasch, 1980). Una forma simple de expresar este modelo es:

$$\log = \frac{\text{probabilidad de éxito}}{\text{probabilidad de fracaso}} = \frac{\text{habilidad de la persona}}{\text{dificultad de la pregunta}}$$

El modelo de RASCH presenta las siguientes características (Chopin (1985) y Wright (1977) :

1. Es matemáticamente simple, comparado con otros como el de dos o tres parámetros.

2. Bajo condiciones normales el puntaje bruto de una persona es una estadística suficiente para estimar su habilidad y el parámetro de las preguntas, lo cual lo hace una extensión de las prácticas actuales en pruebas. Es el único modelo de respuesta al ítem (TRI) que es consistente con el puntaje bruto.

3. Predice el comportamiento de preguntas, pruebas y personas con bastante precisión.

4. Establece que la probabilidad de responder una serie de preguntas correctamente está determinada por la habilidad de las personas; esto es, que dos personas de igual habilidad tienen la misma probabilidad de responder preguntas fáciles y difíciles (sus curvas características no se cruzan).

5. La probabilidad de responder a la más difícil de dos preguntas debe ser inferior a la probabilidad de responder a la más fácil (las curvas características de las preguntas no deben cruzarse).

6. El problema de la estimación de los parámetros está resuelto. Es decir, siguiendo los procedimientos para calcular la dificultad

de las preguntas y la habilidad de las personas, en el modelo de Rasch se llega a un resultado (hay congruencia) mientras que, en los modelos de dos y tres parámetros, el valor es inexacto y no se conoce el grado de inexactitud.

Todos los elementos mencionados anteriormente nos llevaron a elegir el modelo de Rasch para procesar la información y realizar los análisis de datos correspondientes, en los diferentes programas del SNP del ICFES. Adicionalmente se pusieron a prueba los diferentes modelos de la TRI con varios conjuntos de datos para observar diferentes aspectos [10].

SOLUCIONES A LAS PROBLEMÁTICAS TÉCNICAS

on base en los análisis realizados, no sólo de los aspectos técnicos de los modelos, sino de la fundamentación conceptual de la TRI y del modelo de Rasch, se resolvieron en gran medida las problemáticas técnicas planteadas anteriormente. A continuación se presenta una aproximación a las soluciones.

1. Calificación de ítems de crédito parcial

Como se mencionó anteriormente, la TCP no puede asumir el procesamiento de preguntas de crédito parcial, aquellas en las cuales todas o algunas opciones son respuestas parcialmente válidas. Estas preguntas permitirían reconocer procesos en la construcción del conocimiento al diseñar opciones que den cuenta de él.

A diferencia de las llamadas escalas Likert, donde las opciones tienen un mismo valor para todas las preguntas, el modelo de Rasch que analiza las preguntas de crédito parcial concibe que cada ítem tiene su propia estructura de ponderación de opciones.

Indudablemente, esta concepción se aplica más a las pruebas con ítems de crédito parcial del nuevo examen de estado, ya que

[10] Esta investigación se realizó durante el primer semestre de 1999, por el Grupo de Psicometría del SNP. Los resultados se pueden observar en el Centro de Información del SNP. Esencialmente se compararon los programas de computador: Winsteps (modelo de Rasch), Bilog y Multilog (todos los modelos) y el software desarrollado en el SNP en 1992 para el programa SABER. Winsteps resultó ser el más apropiado y eficiente para las necesidades del SNP.

reconoce que el grado de validez de una opción se relaciona con el contexto particular en el cual se plantea la pregunta.

2. Puntajes con significado

En general las medidas educativas se pueden utilizar de dos formas (Van Der Linden, W. 1982):

- Las pruebas pueden hacer mediciones referenciadas a la norma; esto es, las ejecuciones de un estudiante son puntuadas e interpretadas con respecto a las de los demás estudiantes que abordan la prueba. Se hace énfasis en las ejecuciones relativas de los individuos.

- Las pruebas pueden hacer mediciones con referencia a un criterio. En este caso se hace énfasis en especificar los referentes (dominios o criterios) que pertenecen a puntajes o puntos específicos a lo largo de un contínuo. Se especifica qué tipo de ejecuciones puede realizar un individuo y cuál es su repertorio de competencias sin referenciarlo a puntajes o ejecuciones de otros individuos (Haladyna, T y Roid, G. 1983).

El primer caso corresponde a las pruebas en donde se calculan ciertos estadísticos de la población para determinar la escala de resultados. En primer lugar se obtiene el número de respuestas correctas de cada persona que aborda las preguntas. Con este dato se calculan el promedio y la desviación estándar de todos. Se espera que estos datos se distribuyan como la curva normal, de tal manera que el promedio quede en la mitad de la distribución de datos y que en total existan seis desviaciones estándar. Con estos datos se realiza el proceso de estandarización de puntajes, y por último se convierte, este resultado a una escala particular.

En el segundo caso se hace énfasis en la ejecución de las personas que abordan la prueba, independientemente de si se presentan otras personas. Se trata de determinar debilidades, fortalezas; reconocer las ejecuciones particulares de cada uno, de tal manera que los resultados contribuyan a los procesos de reorientación personal o institucional. Es decir que un puntaje en particular puede tener significado desde el punto de vista de las disciplinas evaluadas.

El nuevo examen de estado informará significativamente a los usuarios, por lo cual se sitúa en el segundo tipo de evaluación

educativa. Para ello utilizará el modelo de Rasch, teniendo en cuenta su aplicabilidad a las pruebas referidas a criterios (las cuales, de alguna manera, no puede abordar la Teoría Clásica de las Pruebas).

3. Comparación de resultados en el tiempo

Los resultados que obtienen las personas en una prueba se utilizan como información que permite tomar diferentes decisiones en distintos niveles. Por ejemplo, en algunos casos las decisiones se toman en el nivel individual, como cuando un estudiante decide a qué universidad presentarse para el proceso de admisión. En el nivel institucional los resultados en las pruebas pueden utilizarse para determinar qué personas ingresan a una universidad. Estos ejemplos corresponden a casos en los cuales las pruebas se administran en múltiples ocasiones, como el examen de estado que ocurre en dos ocasiones cada año.

Para el nuevo examen de estado es necesario garantizar que los puntajes obtenidos por los estudiantes en diferentes ocasiones sean comparables y que su significado se mantenga.

Esto se consigue con un proceso estadístico denominado *equating,* que se utiliza para ajustar puntajes obtenidos con formas diferentes de pruebas, de tal manera que estos puntajes se puedan intercambiar (Kolen, M y Brennan, R. 1995) , aunque las formas tengan índices de dificultad diferente.

No se debe confundir este procedimiento con otros muy similares, como el de comparabilidad de escalas que se utiliza frecuentemente en la TCP, ya que en este caso los puntajes no son intercambiables, debido a que aquel se fundamenta en comparaciones de estadísticas a partir de distribuciones de grupos poblacionales.

El equating se puede realizar con base en diferentes diseños:

- ✏ Grupos aleatorios. En este caso los examinados se asignan aleatoriamente a las diferentes formas de prueba.

- ✏ Grupo único con contrarrestación. Se aplican las dos formas a un mismo grupo de personas y se contrarresta el efecto del orden de las pruebas al hacer que algunas personas empiecen por una forma y otras por otra.

- Grupos no equivalentes con ítems comunes. Las diferentes formas de la pruebas comparten ítems o se vinculan a través de una cadena de ítems comunes por segmentos.

4. Variedad en los resultados

Las preguntas de una prueba pueden ser clasificadas de diferentes maneras de tal forma que se resalten distintos ejes de análisis. Es así como en el nuevo examen de estado existe una mirada en relación con las competencias de los estudiantes en cada prueba, pero también una mirada en relación con los ejes disciplinares y/o ámbitos a que hacen referencia los marcos conceptuales en cada área.

Es posible informar sobre el desempeño de cada estudiante (o grupos de estudiantes) en cada uno de los grupos de preguntas según formas particulares de clasificación, de tal manera que se realice una aproximación más detallada de las respuestas de cada persona, a partir de la cual se hagan inferencias que den una idea de las fortalezas y debilidades, que permita reorientar procesos educativos personales o institucionales.

Para el caso de los resultados por grupos de preguntas se utilizará un procedimiento que permita una mirada personal o institucional al desempeño; es decir que los resultados no se obtendrán a partir de comparaciones entre las personas o entre las instituciones. Por el contrario, se refieren a una comparación consigo mismo; es una mirada al desempeño en ejes o ámbitos en relación con el desempeño global de cada individuo o cada institución. En este sentido una fortaleza no indica ejecuciones altas en comparación con los resultados nacionales, sino que indica un desempeño relativamente alto (o significativamente alto, según el caso) en relación con el desempeño global del individuo.

De esta manera no sólo se obtendrán resultados que describan al individuo a través de comparaciones con otros o con las disciplinas evaluadas, sino que faciliten una reflexión personal a cada uno.

5. Reconocer la pluriculturalidad

Una primera aproximación a esta problemática tan compleja se refiere al estudio del Funcionamiento Diferencial de Preguntas (DIF)

por sus siglas en inglés, que al realizarlo en el contexto de grupos poblacionales de interés puede informar sobre el comportamiento de las preguntas para cada una.

El DIF hace referencia a que una pregunta pueda ser desventajosa a un grupo cultural específico en otras palabras, que personas de igual habilidad (con respecto del constructo que mide la pregunta) pero de diferentes culturas muestren o tengan distintas probabilidades de responder correctamente la pregunta, ocasionando una sub o sobreestimación de sus habilidades (Jonhson, E. 1990) (Veale, J. 1983).

Existen muchos métodos que permiten estudiar el Funcionamiento Diferencial de las Preguntas, pero, en general, ninguno derivado de la TCP ofrece resultados satisfactorios.

FUNCIONAMIENTO DIFERENCIAL DE PREGUNTAS

Existen muchos métodos que permiten estudiar el funcionamiento diferencial de las preguntas, pero, en general, se pueden agrupar en dos clases (Van Der Flier, H. Mellenbergh, G. Ader, H. Wijn, M. 1984):

✏ Métodos no-condicionales.

Los métodos no condicionales no han dado buenos resultados para detectar el funcionamiento diferencial de preguntas de acuerdo con factores de tipo cultural. En realidad se refieren a las diferencias de estadísticos tradicionalmente empleados en la evaluación de preguntas y pruebas, basados en el modelo matemático de la curva normal, como son: la dificultad de la pregunta y la correlación ítem-prueba.

Como se ha explicado anteriormente, estos estadísticos se ven influenciados por la muestra donde se haga la aplicación y por lo tanto no se puede reconocer la influencia cultural en las respuestas de los estudiantes. Por este motivo el procesamiento de información del nuevo examen utiliza uno de los métodos condicionales.

✏ Métodos condicionales.

Los métodos condicionales condicionan el sesgo en las preguntas a los niveles de habilidad de quienes abordan la prueba, entendidos éstos como los puntajes brutos. Las técnicas más utilizadas para verificar el DIF son la de Mantel-Haenzel, la comparación de curvas características de preguntas obtenidas a partir de datos de dos o más poblaciones diferentes y la comparación de los niveles de dificultad obtenidos en dos poblaciones distintas (modelo de Rasch).

El procedimiento que se empleará en el SNP, se fundamenta en la invarianza de los parámetros de las preguntas (su nivel de dificultad) en el modelo de RASCH, que ha sido probado (con otros procedimientos) en algunas aplicaciones de pruebas del programa SABER.

RESULTADOS EN EL NUEVO EXAMEN DE ESTADO

TIPOS DE RESULTADOS EN EL NUEVO EXAMEN

partir de múltiples consideraciones, se diseñaron cuatro tipos de resultados que se pueden obtener a partir de las respuestas de los estudiantes en el nuevo examen de estado.

El examen está compuesto por un núcleo común y un componente flexible. El núcleo común contiene una serie de pruebas de áreas básicas y debe ser abordado por todos los estudiantes. El componente flexible está compuesto por dos líneas: profundización (mayor nivel de complejidad en la evaluación) e interdisciplinar (desenvolvimiento de las personas en distintos escenarios socioculturales). Para cada parte de la estructura se producirán resultados que corresponden a alguno de los siguientes tipos.

PUNTAJE

ste es un resultado cuantitativo que se obtiene a partir de las estimaciones de la habilidad (en términos psicométricos) de cada estudiante con base en sus respuestas a las pruebas. El resultado *o puntaje* se

procesa para cada prueba del núcleo común y para la línea interdisciplinar del componente flexible.

La estimación del parámetro habilidad se realiza a través de procesos de máxima verosimilitud que lleva a cabo el software Winsteps (Linacre, J. y Wright, B. 1994), que calcula, al mismo tiempo, el parámetro de los ítems utilizados (dificultad[11]), el grado de ajuste de los datos al modelo, el equating entre dos formas de la misma prueba, las probabilidades de respuesta para una persona o grupo específico y los estadísticos generales de prueba, como los promedios, las desviaciones estándar, la confiabilidad, entre otros.

En el núcleo común, el puntaje (habilidad) indicará la competencia general en cada una de las pruebas tal, y como se defina desde la disciplina[12]. En la línea interdisciplinar, el puntaje indicará el desenvolvimiento en los escenarios socioculturales de la prueba seleccionada por el estudiante.

Como se mencionó anteriormente, la interpretación de resultados es con referencia a criterio; es decir que algunos puntajes indicarán el tipo de ejecuciones que puede realizar el individuo. Para ello se realizará un proceso de anclaje de preguntas, o sea que se determina el tipo de competencias para responder a diferentes preguntas en el continuo de la escala de dificultad de ítems.

RESULTADOS POR GRUPOS DE PREGUNTAS

ada una de las disciplinas evaluadas, al considerar la estructura de las pruebas, puede abordar el problema de clasificar las preguntas de acuerdo con tópicos de interés desde el punto de vista educativo, humano, social, etc., lo que hace posible informar, a quien las aborda, sobre su desempeño relativo en esos tópicos de tal manera que esa persona pueda reorientar sus procesos esenciales.

Este desempeño relativo puede ser significativamente superior o inferior al esperado, lo que podría interpretarse como una fortaleza o debilidad relativa. Se considera que es significativamente superior o inferior si las diferencias entre su desviación de la media para un

[11] La dificultad de un item puede entenderse como el grado de exigencia (de la competencia) de una pregunta para ser respondida correctamente por quien la aborde.
[12] Consultar los otros documentos de esta serie.

grupo de preguntas y su desviación de la media global (C) es superior al doble del error estándar de las diferencias entre estas desviaciones (Bertrand y Dupuis. 1988).

En el análisis de resultados de la ejecución de los estudiantes, por departamento, en los grupos de preguntas, algunos se citan como con rendimientos superiores o inferiores al global. En estos casos los departamentos se citan como desviados de su rendimiento global si las diferencias entre su desviación de la media para un grupo de preguntas y su desviación de la media global (C) es superior al doble del error estándar de las diferencias entre estas desviaciones.

Este tipo de resultado, de tipo cualitativo, se procesará únicamente para las pruebas comunes y se hará a los niveles: estudiante, institución educativa (colegio), departamental y nacional, de tal manera que se disponga de información que permita cualificar los procesos educativos o las decisiones que se tomen en este ámbito.

NIVEL DE COMPETENCIA

ara cada una de las pruebas del núcleo común se determinará el nivel de competencia de cada estudiante en cada una de las competencias que mida la prueba.

El nuevo examen tiene como objeto de evaluación las competencias de los estudiantes en contextos disciplinares. Esto implica que se construirán preguntas, en cada prueba, que midan cada una de las competencias descritas en los marcos conceptuales de las pruebas.

El software Winsteps calcula la habilidad de los estudiantes en cada una de las competencias evaluadas en cada prueba, las que se pueden relacionar con niveles particulares de competencia que se establecen con base en el procedimiento de anclaje de preguntas.

El resultado es descriptivo (cualitativo) y proporcionará un mapa del desempeño de los estudiantes en las diferentes competencias evaluadas.

GRADO DE PROFUNDIZACIÓN

ste resultado se procesará para la línea de profundización del componente flexible. En esta línea, como se mencionó, se incluyen preguntas de mayor exigencia (mayor dificultad), de tal manera que los resultados puedan ser utilizados como indicadores de fortalezas y contribuyan al proceso de elección de opción profesional.

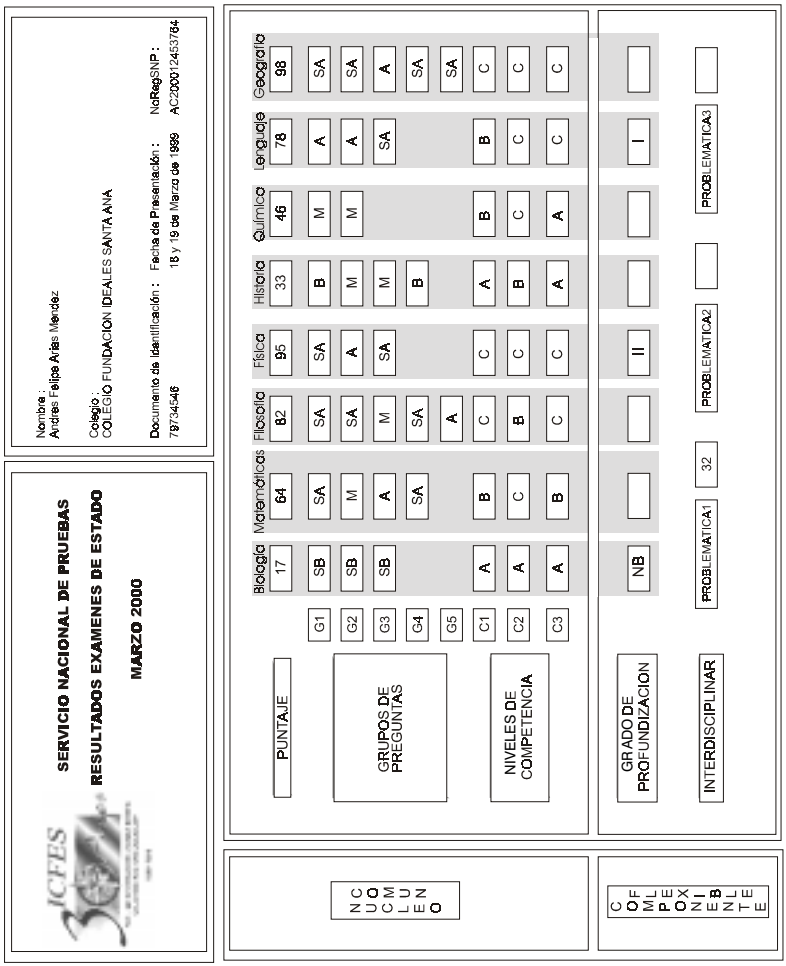
En cada una de las pruebas de esta línea se podrán reconocer diferentes grados de profundización alcanzados por los estudiantes, que indican el grado de complejidad que pueden abordar y resolver correctamente.

Los grados de profundización se establecen a partir de los niveles de dificultad de las preguntas y existirá una descripción cualitativa del significado, en términos de complejidad, de cada grado. Para cada prueba, se establecerán tres grados diferentes.

VALIDEZ DE LOS RESULTADOS

odo lo anterior contribuye a la validez de los resultados de las pruebas entendida como un "juicio evaluativo integral del grado en el cual la evidencia empírica y teórica soportan lo adecuado y apropiado de las interpretaciones y acciones basadas en los puntajes de una prueba u otra forma de evaluación". Validez no es una propiedad del significado de los puntajes de la prueba. Estos puntajes no son sólo una función de las condiciones del ítem o de los estímulos, sino también de las personas que responden y del contexto de la evaluación. Específicamente, lo que debe ser válido es el significado o interpretación del puntaje; lo mismo que cualquier implicación que este puntaje tenga para la acción. La extensión en la cual el significado del puntaje y las implicaciones para la acción se mantienen a través de personas o grupos poblacionales y a través de ambientes o contextos es una pregunta empírica persistente y perenne. Esta es la razón principal por la cual la validez es una propiedad cambiante y la validación es un proceso continuo.

EJEMPLO DE LOS RESULTADOS EN EL NUEVO EXAMEN DE ESTADO



BIBLIOGRAFIA

Angoff, W. (1993). Perspectives on Differential Item Functioning Methodology. En: Differential Item Funtioning.Lawrence Earlbaum Associates. New Jersey.

Anastasi, Anne, Urbina, Susana. Test psicológicos. PRENTICE HALL. México. (1998).

Arce, Constantino. Técnicas de construcción de escalas psicológicas. Editorial SINTESIS. (1994). España.
Bertrand y Dupuis. (1988). A world of differences. Technical report. ETS. New Jersey.

Borg, I. Y Groenen, P. (1997). Modern Multidimensional Scaling: theory and applications. Springer. New York.

Choppin (1985) y Wright (1977).

DuBois, P. H. (1970). History of Psychological Testing. Boston. Allyn & Bacon.

Embretson, S. (1999). The New Rules of Measurement. Lawrence Erlbaum Associates. New Jersey.

Greaney, V. Y Kellaghan T. (1995) Equity Issues in Public Examinations in Developing Countries. Washington. World Bank.

HALADYNA, T. y ROID, G. A comparison of two approaches to criterion-referenced test construction. Journal of Educational Measurement, 20, pp. 271-282. (1983).

HAMBLETON, R., COOK, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, pp. 75-96 (1977).Stenner y otros, (1983)

Hambleton, R. y Swaminathan, H. 1985. Item Response Theory: principles and applications. Kluwer. Boston.

Isaac Bejar. (1983). Achievement Testing. Recent Advances. SAGE. Newbury Park.

Jonhson, E. (1990). Theoretical justification of the omnibus measure of differential item functioning. IAEP data analysis plan.

JOHNSON, E. Theoretical justification of the omnibus measure of differential item functioning. IAEP data analysis plan. Apéndice 1.

Kolen, M y Brennan, R. (1995). Test Equating: methods and practices. Springer. New York.

Linacre, J y Wright, B. Op. Cit.

Lord, F. Y Novick, M (1968). Statistical Theories of Mental Test Scores. Addison-Wesley. Massachusetts.

Martínez Arias, Rosari. Psicometria: Teoría de los test psicológicos y educativos. Editorial SINTESIS. 1996. España.

Mehrens, W. y Lehmann, I. (1982). Medición y Evaluación en la Educación y en la Psicología. CECSA. México.

Messick, S. Op. Cit.

Novick, M y Jackson, P. (1974). Statistical Methods for Educational and Psychological Research. McGraw-Hill. New York.

Nunnally, J. (1987). Teoría Psicométrica.

Stenner y otros, (1983).

Torrado, M. (1998). De la evaluación de aptitudes a la evaluación de competencias. En Serie Investigación y Evaluación Educativa. ICFES. Santafé de Bogotá.

Tyler, Ralph. (1988). Assessment Changes due to Cultural Changes. En International Journal of Education Research.

VAN DER FLIER, H., MELLENBERGH, G., ADER, H., WIJN, M. An iterative item bias detection method. Journal of Educational Measurement, 21, pp. 131-145. (1984).

VAN DER LINDEN, W. Criterion-referenced measurement: its main applications, problems and findings. Evaluation in education, t, pp. 97-118. (1982).

VEALE, J., FOREMAN, D. Assessing cultural bias using foil response data: cultural variation. Journal of Educational Research, 20, pp. 249-258. (1983).
Wright, 1994.

Wright, B. Y Mead, R. (1977). Calibrating ítems and scales with the Rasch model. Research memorandum No.23. University of Chicago.

BIBLIOGRAFIA RECOMENDADA EN ESPAÑOL

MartínezArias, Rosario. Psicometría: teoría de los test psicológicos y educativos. Editorial SINTESIS. (1996). España.

Arce, Constantino. Técnicas de construcción de escalas psicológicas. Editorial SINTESIS. (1994). España.

Anastasi, Anne, Urbina, Susana. Test psicológicos. PRENTICE HALL. México. (1998).