PRUEBAS DIAGNÓSTICAS: UNA APPLICACION A LA TEORÍA DE RESPUESTA AL ITEM, APROXIMACIÓN CLÁSICA Y BAYESIANA

Leticia Debera
Laura Nalbarte

Documento de Trabajo
Octubre 2006

Instituto de Estadística. Facultad de Ciencias Económicas y de Administración
Universidad de la República, Uruguay

PRUEBAS DIAGNÓSTICAS: UNA APLICACIÓN DE LA TEORÍA DE RESPUESTA AL ÍTEM, APROXIMACIÓN CLÁSICA Y BAYESIANA

LETICIA DEBERA, LAURA NALBARTE
Instituto de Estadística. F.C.E. y A.
Universidad de la República.
leticia@iesta.edu.uy

RESUMEN

La Teoría de Respuesta al Ítem (TRI) es una herramienta que permite cuantificar un rasgo latente de una persona. La utilidad de esta teoría en el campo educativo radica en determinar si un estudiante consigue responder correctamente a cada una de las preguntas (ítems) y en atender al puntaje bruto obtenido en la prueba.

Los modelos aquí planteados son de dos y tres parámetros, y relacionan la probabilidad de que la respuesta sea correcta con la dificultad, la discriminación y el azar de ese ítem.

El análisis de TRI fue aplicado a las Pruebas Diagnósticas al Ingreso de la Facultad de Ciencias Económicas realizadas a la generación 2006. El objetivo del mismo es analizar qué preguntas fueron las más difíciles, cuáles contribuyeron a diferenciar entre estudiantes y en qué casos el azar tuvo mayor peso. La aproximación al tema se realizó en dos perspectivas: clásica y bayesiana. La base de datos está constituida por 247 individuos y 20 ítems.

Palabras claves: Teoría Respuesta al Ítem, Análisis Bayesiano

I.- Introducción

El objetivo de este trabajo es analizar las Pruebas Diagnósticas al Ingreso de la Facultad de Ciencias Económicas realizadas a la generación 2006, con el fin de determinar que preguntas fueron las más difíciles, cuales contribuyeron a diferenciar entre estudiantes y en que casos el azar tuvo mayor peso. El análisis se realiza desde la perspectiva de la Teoría de Respuesta al Ítem (TRI).

La TRI es una herramienta que permite cuantificar un rasgo latente de una persona. La utilidad de esta teoría en el campo educativo radica en determinar si un estudiante consigue responder correctamente a cada una de las preguntas (ítems) y no atender al puntaje bruto obtenido en la prueba (test).

En cualquier situación de medida hay variables subyacentes de interés que, en el caso del ámbito educativo, son cognitivas (sobre todo de contenido), pero que pueden ser psicológicas, como la inteligencia. Estas variables subyacentes deben tener como soporte un constructo teórico y son denominadas en la nomenclatura de TRI “rasgos latentes” o “habilidades”. El modelo planteado es un modelo de 3 parámetros, y relaciona la probabilidad de que la respuesta sea correcta con la dificultad, la discriminación y el azar de ese ítem.

La TRI fue aplicada a las Pruebas de Matemática, realizando en el análisis en dos perspectivas: clásica y bayesiana. La base de datos está constituida por 247 individuos y 20 ítems del área Matemática.

El presente trabajo se estructura de la siguiente manera: en una primera parte se presentan los aspectos metodológicos, desde ambas perspectivas de análisis, en una segunda parte se analizan los resultados obtenidos, subdividiendo la presentación en tres secciones: clásico, bayesiano y análisis comparativo. Finalmente se presentan las conclusiones.

II.- Conceptos básicos de la metodología**.** TEORÍA DE RESPUESTA AL ÍTEM

La mayor parte de los análisis realizados para estudiar medidas en el campo de la educación y de la psicometría fue basada en la teoría clásica de test (TCT), desarrollada en los años 20. Sin embargo, la TRI, desarrollada luego de 40 años, es conceptualmente más potente que la teoría clásica, la misma se basa sobre los ítems en lugar de los puntajes del test.

En cualquier situación de medida hay una variable de interés subyacente que, en el caso del ámbito educativo, son cognitivas (sobre todo de contenido), pero que pueden ser psicológicas, como la inteligencia, todas ellas variables que deben tener como soporte un constructo teórico. Estas variables subyacentes son llamadas en la nomenclatura de TRI “rasgos latentes” o “habilidades”. Entonces, TRI es una herramienta que nos permite cuantificar un rasgo latente de una persona.

La utilidad de esta teoría en el campo educativo radica en determinar si un estudiante consigue responder correctamente a cada una de las preguntas (ítems) y no al puntaje bruto obtenido en la prueba (test).

Cada estudiante resultará ubicado en una escala según el nivel que alcance en el rasgo o habilidad que se desea medir.
Para medir (o cuantificar) un “rasgo latente” en una persona es necesario tener una escala de medición. Se asumirá que para cualquier habilidad ésta puede ser medida

sobre una misma escala, teniendo un punto medio de cero, una unidad de medida de uno y un rango de menos infinito a más infinito. Esta "regla" va a ser usada para medir cuanta habilidad tiene una persona y compararla con la habilidad de otras. Por lo general el rango va de –3 a 3 (población proveniente de una distribución gaussiana o normal estandarizada).

La ubicación del estudiante en esta escala de la variable del rasgo se denomina con la letra θ y la probabilidad que tendrá ese examinado de responder correctamente a este ítem dado que tiene una cierta habilidad θ, se denomina con la letra (**P**). En términos formales $P(Y=1|\theta)$ probabilidad de que Y sea 1 para un valor dado de θ.

En la presente investigación se trabaja con modelos de 2 (introducido por Birnbaum en 1968) y 3 parámetros: *Discriminación (a), Dificultad (b) y Azar (c).*[1]

$$P(Y=1|\theta)=\frac{1}{1+e^{a(\theta-b)}} \quad (1)$$

$$P(Y_i=1|\theta)=c_i+\frac{1-c_i}{1+e^{-1.7a_i(\theta-b_i)}} \quad (2)$$

El modelo planteado es una función que relaciona la probabilidad de que la respuesta sea correcta (Y=1) con la discriminación, la dificultad de ese ítem, en al caso (1) agregándose y el azar en el caso (2).

Al representar gráficamente la probabilidad en función del nivel de rasgo ($P(Y=1|\theta)$) se observa la denominada ***Curva Característica del Ítem (CCI)***. Esta curva representa la probabilidad de contestar correctamente a la pregunta en función de su nivel de rasgo.
La forma de las Curvas Características del Ítem, refleja la relación (matemática) que vincula la probabilidad de responder correctamente con el nivel de habilidad: probabilidad cercana a 0 implica niveles bajos de habilidad probabilidad, y cercana a 1 implica niveles altos de habilidad.

II.1.Significado de los parámetros.

Parámetro de adivinación o Azar (c). El rango teórico de este parámetro va de 0 a 1, pero en la práctica se considera aceptable hasta un valor de 0.30.[2]

El parámetro *c* representa el "piso" de la probabilidad, que todos los estudiantes tienen, de responder correctamente sin importar el nivel de rasgo. Es decir, hay al menos una probabilidad "*c*", de responder correctamente a un ítem.

El *Parámetro de Dificultad (b)* define el punto sobre la escala de habilidad donde la probabilidad de responder correctamente es la mitad entre este piso (c) y 1[3].
El rango teórico de los valores de este parámetro es de $-\infty$ a $+\infty$, pero en la práctica están por lo general entre –3 y 3, tal como se indicó anteriormente.

---

[1] En TRI hay 3 modelos logísticos: los dos presentados y un modelo que sólo tiene el parámetro de dificultad (1P, Rasch),

[2] The Basic of Ítem Response Theory de Frank B. Baker

[3] Esto es: $P(\theta = b) = (1 + c) / 2$

Para los modelos de 2 parámetros, el valor del parámetro *b* (Dificultad) es el punto de la escala de habilidad en el cual la probabilidad de responder correctamente es 0.5 para $\theta = b$.

El *Parámetro de Discriminación (a),* describe cuan bien un ítem puede diferenciar a los examinados con niveles de habilidad bajos de aquellos con habilidades altas, (este parámetro existe en la TCT aunque su modo de cálculo es distinto al de TRI). Gráficamente este parámetro representa la pendiente de la Curva Característica del Ítem, por lo que existirá mayor discriminación en los lugares donde la pendiente es mayor. Si la curva es achatada, este ítem no sería capaz de discriminar, ya que la probabilidad de responder correctamente con bajo nivel de habilidad es la misma que para niveles altos de habilidad.

II.2.Principales características de TRI.

Las dos características fundamentales de TRI para indagar las propiedades de las preguntas o ítems son:
a) Invariancia del ítem en la muestra de examinados.
b) Invariancia de la estimación de la habilidad de un examinado con los ítems.

a) Invariancia del ítem en la muestra de examinados.

Los parámetros de los ítems calculados no dependen del nivel de habilidad del grupo de examinados que los responden, sino que es una propiedad de los ítems.

b) Invariancia de la estimación de la habilidad de un examinado con los ítems.

El nivel de habilidad de los examinados es invariante a los ítems que se utilicen para medirlo.

Esta propiedad requiere de dos condiciones:

1)Todos los ítems miden el mismo rasgo latente.
2)Los valores de los parámetros de los ítems están en una métrica común.

La consecuencia práctica de esta propiedad es que un test ubicado en cualquier punto a lo largo de la escala de habilidad puede ser usado para estimar la habilidad del examinado. Por ejemplo, un examinado podría tomar un test que es “muy fácil” o uno “muy difícil” y obtener, sobre el promedio, la misma habilidad estimada.

II.3.Función de información (una medida de la precisión).

La información como concepto está vinculada con la precisión, en TRI es el recíproco de la precisión con la cual un parámetro podría ser estimado. Esta precisión es medida por la variabilidad de las estimaciones alrededor del valor del parámetro. La información se define como el inverso de la varianza, I=1/Var.

Si el valor de la información es grande, esto significa que la habilidad del examinado va a estar muy cercana a la verdadera habilidad; esto es, todas las estimaciones serán cercanas al verdadero valor. Si la información es pequeña, significa que la habilidad no puede ser estimada con precisión y las estimaciones estarán ampliamente dispersas alrededor de la verdadera habilidad.

Como la función de información del test (global) es la suma de las informaciones de cada uno de los ítems, cuanto más ítems hayan en un test, las estimaciones de la habilidad de los examinados van a ser más precisas.

II.4.Estimación de los parámetros

Algunas técnicas de estimación de los parámetros de TRI son:

1- Estimación Máximo Verosímil Conjunta (Birnbaum, 1969) (se lo conoce como el paradigma de Birnbaum).
2- Estimación Máximo Verosímil Marginal (Bock and Aitkin, 1981) y
3- Estimación Bayesiana (Mislevy, 1986)

II.4.1.Estimación Máximo Verosímil Conjunta.

Bajo el supuesto de independencia, el paradigma de Birnbaum reduce la solución simultánea de los 2n+N parámetros de los ítems y los examinados a un procedimiento que, en primer lugar estima los parámetros de los ítems (2n ya que es un modelo de 2 parámetros) suponiendo conocidas las habilidades de los examinados, y en segundo lugar utiliza estas estimaciones de los ítems obtenidas en la etapa anterior para estimar las habilidades de los examinados (*N*).

El paradigma es un método iterativo hacia delante-hacia atrás (back-and-forth) para estimar los parámetros de los ítems y los individuos hasta la convergencia.

El método utilizado para las estimaciones de los parámetros en cada uno de los pasos es el método también iterativo de Newton-Raphson[4] (N-R)

El proceso se inicia fijando valores arbitrarios, aunque por lo general se usa el puntaje bruto estandarizado obtenido por los examinados. De esta manera se obtiene el procedimiento de estimación para cada uno de los examinados, con lo que se obtiene un ciclo del paradigma de Birnbaum. Este método se repite de igual manera, pero partiendo de las estimaciones de los rasgos de los examinados obtenidas, hasta que converge. Siempre antes de repetir el proceso se estandarizan los parámetros estimados, ya que siempre se trabaja en la misma métrica.

II.4.2.Estimación Máxima Verosímil Marginal.

Neyman y Scott (1948) muestran que cuando las estimaciones de los parámetros de los ítems y de los individuos son realizadas conjuntamente, no necesariamente son consistentes cuando decrece el tamaño de muestra.
Para resolver este problema, Bock y Aitkin (1981) desarrollaron un procedimiento de estimación Máximo Verosímil Marginal (MMLE) para estimar los parámetros de los ítems y es simple de implementar.

Bajo el procedimiento de estimación MMLE de los parámetros de los ítems, se asume que los examinados representan una muestra aleatoria de una población en la cual su habilidad tiene una cierta distribución.

[4] El método de N-R es un procedimiento iterativo para resolver sistemas de ecuaciones no lineales envueltos en distintos métodos de estimación.

La esencia de este procedimiento consiste en integrar sobre la distribución de la habilidad. Así, los parámetros de los ítems son estimados en la distribución marginal y éstos no dependen de la estimación de la habilidad de cada uno de los examinados.

En vez de obtener un valor estimado de la habilidad para cada examinado, se tiene cada uno de los posibles valores de θ a lo largo de la escala de habilidad, condicionado al patrón de respuestas, a los parámetros de los ítems y a la distribución de la habilidad de la población (probabilidad a posteriori). Esta distribución a posteriori, entonces, combina la información de la distribución a priori de la habilidad y la función de verosimilitud.

La idea del uso de la estimación MMLE/EM[5] consiste en que usando métodos de estimación Máximo Verosímil (ML) deseamos encontrar valores de los parámetros que maximizen la log-verosimilitud de los datos completos (N-R). Como parte de estos datos no son observables, no se tiene un estadístico suficiente para θ. Como sustituto, se calcula la esperanza a posteriori de θ dados los datos observados (usando distribución de probabilidad a-priori y los parámetros de los ítems). Esta esperanza es calculada en el paso E (expectation), que consiste en calcular la esperanza de la función de densidad de la distribución de probabilidad conjunta entre las respuestas a los ítems y θ, condicionada a los parámetros de los ítems estimados (esperanza a posteriori con respecto al p-ésimo patrón de los ítems estimados).

El paso siguiente consiste en el uso de los estadísticos suficientes encontrados de las estimaciones de los datos, antes no observados, maximizando la esperanza a posteriori obtenida.

Este proceso se repite hasta que se alcanza un criterio de convergencia.

Al final de cada ciclo se estandarizan las estimaciones, ya que TRI establece que los parámetros de los ítems y de los individuos están en la misma métrica.

IV.4.3.Estimación Bayesiana.

Si bien en los procedimientos de estimación MMLE se resuelve el problema de la inconsistencia, se pueden producir malas estimaciones cuando los patrones de respuestas de los individuos son “extremos”. Esto es, por ej. cuando un individuo contesta todas las preguntas bien. Para resolver este problema se usa el método de estimación Bayesiana.

Los procedimientos bayesianos producen una distribución a posteriori que se obtiene de combinar las probabilidades obtenidas de la función de verosimilitud, que usa la información obtenida de la muestra, con probabilidades obtenidas usando la información a priori de la distribución del conjunto de parámetros desconocidos (niveles de rasgo de los individuos).
Aplicando el teorema de Bayes se obtiene una distribución de probabilidad a posteriori que es usada para hacer inferencia sobre los parámetros desconocidos.

El método de estimación Bayesiana es similar al de MML, considerando distribución de probabilidad a priori para el parámetro de habilidad. Esto permite realizar

[5] El algoritmo E-M es un procedimiento iterativo que se usa para encontrar estimaciones ML de parámetros de modelos de probabilidad, en presencia de variables aleatorias no observables, en este caso esta variable es θ.

simultáneamente las estimaciones de los parámetros de los ítems y de los individuos y el resultado es la ecuación de la distribución a posteriori de los parámetros de los ítems en vez de la distribución de probabilidad de la habilidad.
Los parámetros de los ítems son estimados usando una versión del algoritmo EM, introducida por Bock and Aitkin (1981) para modelos de TRI. La ventaja del uso de este algoritmo es que solamente requiere las derivadas primeras de la función de verosimilitud y a su vez es numéricamente robusto así como fácil de implementar.

El parámetro de habilidad es estimado separadamente en un siguiente paso.

Distintos estudios han demostrado que a pesar de que a la estimación Bayesiana se le considera subjetiva (por la elección de la distribución a priori), la aplicación de esta teoría para la estimación de los parámetros de TRI es efectiva. Se han hecho diversas simulaciones para los modelos logísticos de 1, 2 y 3 parámetros que demuestran que cuando los parámetros de los ítems y el parámetro de habilidad de los individuos, son estimados conjuntamente (en la estimación Bayesiana) se obtienen estimaciones más precisas, con menos sesgo y con diferencias entre los valores estimados y los verdaderos más pequeñas que con las estimaciones máximo verosímil.

Para realizar la estimación se utiliza la metodología de Gibbs Sampling, algoritmo que extrae muestras en forma sucesiva de las probabilidades condicionales de los parámetros del modelo.

A continuación se presenta un ejemplo de cómo opera el algoritmo de Gibbs Sampling.

Sea la verosimilitud $y_i \sim N(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k , \tau)$ y las distribuciones a priori:
$\beta_j \sim N(0 , .001)$ para todo i y
$\tau \sim$ Gamma(.00001,.00001)

La probabilidad a posteriori será $P(\beta, \tau \mid y) \propto \prod_{i=1 \text{ to } n} p(y_i \mid \beta, \tau) \prod_{j=1 \text{ to } k} p(\beta_j)\, p(\tau)$

El Gibbs Sampler trabaja extrayendo muestras en forma sucesiva de las probabilidades condicionales de los parámetros del modelo, utilizando cadenas de Markov. En el ejemplo planteado las probabilidades condicionales serían:

$\beta_1^{1} \sim p(\beta_1 \mid \beta_2,\ldots, \beta_k, \tau, Y)$
$\beta_2^{1} \sim p(\beta_2 \mid \beta_1, \beta_3 ,\ldots, \beta_k, \tau, Y)$
$\beta_k^{1} \sim p(\beta_k \mid \beta_1,\ldots, \beta_{k-1}, \tau, Y)$
$\tau^{1} \sim p(\tau \mid \beta_1,\ldots, \beta_k, Y)$
$\beta_1^{2} \sim p(\beta_1 \mid \beta_2,\ldots, \beta_k, \tau, Y)$
$\beta_2^{2} \sim p(\beta_2 \mid \beta_1, \beta_3 ,\ldots, \beta_k, \tau, Y)$
$\beta_k^{2} \sim p(\beta_k \mid \beta_1,\ldots, \beta_{k-1}, \tau, Y)$
$\tau^{2} \sim p(\tau \mid \beta_1,\ldots, \beta_k, Y)$

A los efectos de determinar la convergencia de las cadenas se utiliza el estadístico propuesto por Gelman and Rubin. El mismo es el cociente de la varianza estimada y las varianzas dentro de las cadenas.

$$\text{Varianza en cadenas } W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\theta_j^i - \bar{\theta}_j\right)^2$$

$$\text{Varianza entre cadenas } B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\theta_j - \bar{\theta}\right)^2$$

$$\text{varianza estimada } \hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$$

$$\text{Estadístico Gelman - Rubin } \sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$$

Cuando se alcanza la convergencia, el numerador y el denominador deberían coincidir, por lo que el estadístico se aproxima a 1.

III.- Análisis de Resultados

La prueba analizada consiste de 20 preguntas con 5 respuestas posibles, donde sólo una es correcta.
Se toma como referencia un modelo de respuesta dicotómico, dónde el estudiante puntúa 1 si la respuesta fue correcta y 0 si no lo fue.

En el análisis clásico se analizan los modelos de 2 y 3 parámetros, mientras que desde la perspectiva bayesiana se profundizó en el modelo de 2 parámetros[6]. En ese sentido, a los efectos de poder comparar resultados se presentan en esta sección los análisis para el modelo de 2 parámetros (el modelo con 3 parámetros clásico se puede ver en el anexo).

III .1 Resultado de las estimaciones Máximo Verosímil

Se presentan primero las estimaciones de los parámetros de los ítems (dificultad y discriminación), y posteriormente se analiza la función de información de los ítems y los niveles de rasgo de los estudiantes de la muestra.

Los valores de las estimaciones de los parámetros de los ítems obtenidos se pueden observar en el cuadro siguiente.

Cuadro 1. Estimación de los parámetros

	Discriminación (a)	D.S (a)	Dificultad (b)	D.S (b)	% correctas
ITEM 1	0.87	0.26	-1.85	0.5	80.6
ITEM 2	1.22	0.31	-1.73	0.36	84.6
ITEM 3	0.76	0.21	1.77	0.55	22.7
ITEM 4	1.26	0.25	0.28	0.16	42.9
ITEM 5	1.04	0.28	-2.04	0.48	85.8
ITEM 6	0.66	0.2	1.61	0.56	27.1
ITEM 7	0.89	0.22	1.91	0.5	18.2
ITEM 8	1.89	0.37	-0.85	0.14	73.7
ITEM 9	1.03	0.3	1.89	0.44	15.8
ITEM 10	0.52	0.21	4.66	2.32	8.5
ITEM 11	1.57	0.28	0.17	0.13	44.9
ITEM 12	1.27	0.27	0.96	0.21	27.5
ITEM 13	1.07	0.22	-0.06	0.17	51.0
ITEM 14	1.28	0.26	0.56	0.17	36.0
ITEM 15	0.42	0.17	2.54	1.23	25.9
ITEM 16	0.9	0.2	0.64	0.24	37.7
ITEM 17	0.81	0.19	0.72	0.28	37.2
ITEM 18	1.08	0.22	0.94	0.23	30.0
ITEM 19	1.46	0.29	0.77	0.18	30.0
ITEM 20	1.12	0.21	0.56	0.18	37.2

[6] Al momento de escribir este artículo no se pudo correr en WINGBUGS, el modelo de 3 parámetros. El trabajo a futuro es la implementación del mismo en el R.

De acuerdo a lo mencionado anteriormente, un parámetro de dificultad *b* negativo, significa que el ítem es fácil y será más fácil, cuanto más grande en magnitud sea la estimación. Y viceversa, cuánto mayor es, más difícil será el ítem.

Analizando las estimaciones de los parámetros se observa que 3 de los 20 ítems son muy fáciles (se asocian a los porcentajes de respuestas más altos), mientras que 9 de ellos son difíciles (la mayoría con un porcentaje de respuestas correctas inferior al 30%).
De acuerdo a esta teoría, el ítem 10 sería descartado del test, ya que su dificultad es 4.66 (muy difícil). Se observa que solamente un 8.5% de los estudiantes contestó correctamente esta pregunta, lo que provoca una mala estimación.

Un ítem ideal sería, por ejemplo, el ítem 11 ya que presenta una dificultad normal (b=0.17) y un parámetro de discriminación alto (a=1.57).

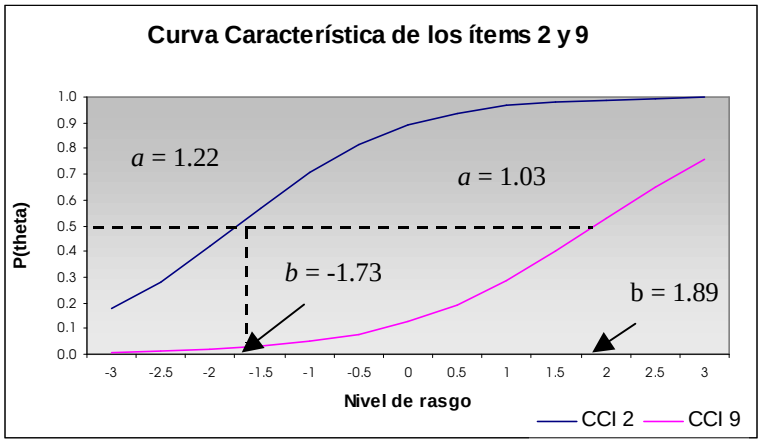
Una vez obtenidas las estimaciones de los parámetros, se pueden calcular las probabilidades de que un individuo conteste correctamente cada uno de los ítems según su nivel de rasgo. El cuadro con las distribuciones de probabilidad acumulada para cada ítem se presenta en el anexo de resultados.

Al estudiar las probabilidades se puede verificar la información obtenida en el análisis del parámetro de dificultad. Si se toman como ejemplo los ítems 1, 2 y 5 cuyos parámetros de dificultad son -1.85, -1.73 y –2.04 respectivamente, y se observa la probabilidad de que un estudiante conteste correctamente los mismos, se constata que para niveles bajos de habilidad la probabilidad se encuentra en el entorno de 0.25 (0.27, 0.18 y 0.27). Se debe hacer notar, que estos valores debería ser cercana a 0 ya que reflejan la probabilidad de que un individuo sin conocimiento conteste correctamente el ítem.

Al graficar la probabilidad en función del nivel de rasgo, se obtiene la denominada *Curva Característica del Ítem (CCI)*. Esta curva tiene una forma de S y representa la probabilidad de contestar correctamente a la pregunta en función de su nivel de rasgo.

El gráfico 1 muestra las CCI de los ítems 2 y 9.

Gráfico 1. Curva Característica de los ítems 2 y 9



Se observa que el ítem 2 está por encima del ítem 9 para todos los niveles de habilidad, de esta manera, la probabilidad de que la respuesta sea correcta es siempre mayor para el ítem 2.

Para analizar la discriminación de cada ítem se agrupan o analizan los mismos según parámetros estándar.
Se presenta a continuación, en el Cuadro 2, una escala teórica propuesta por Baker[7], para la clasificación del parámetro de Discriminación (***a***):

Cuadro 2. Niveles de discriminación

Calificación	Discriminación
nula	0
Muy baja	0,01 a 0,34
baja	0,35 a 0,64
Moderada	0,65 a 1,34
alta	1,35 a 1,69
Muy alta	mayor a 1,7

Esta clasificación implica, que aquellos parámetros con niveles de discriminación entre 1,39 y 1,69 son considerados parámetros de discriminación alta, mientras que los que están por debajo de 0,64 son considerados como de bajo o muy bajo rendimiento.
Acorde a esto los ítems de la prueba se pueden clasificar en: 2 con discriminación baja (ítems 10 y 15), 15 con discriminación moderada y 3 con discriminación alta (ítems 8,11,19).

Función de información

La estimación de los parámetros debe ser acompañada con información que determine cuan bien han sido estimados los niveles de habilidad. Para ello se estudia la función de información, la que a su vez pretende ver la calidad de las preguntas.

En el cuadro 3 se presenta la función de información global del test, mientras que la información de cada uno de los ítems puede observarse en el anexo de resultados.

Cuadro 3. Función de Información.

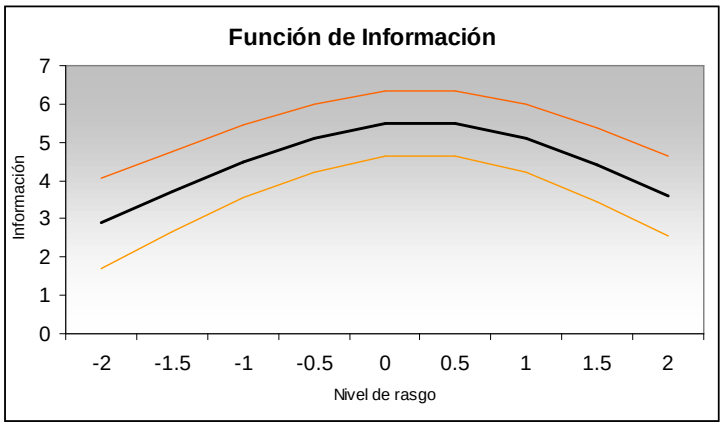
Función de Información del Test									
Theta	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
I(THETA)	2.9	3.7	4.5	5.1	5.5	5.5	5.1	4.4	3.6
SE(THETA)	0.59	0.52	0.47	0.44	0.43	0.43	0.44	0.48	0.53

El máximo de la función de información se da para θ = 0 y para θ = 0.5 y es 5.5 con un error estándar de 0.43. Por lo tanto el 95% de las estimaciones de estos niveles de habilidad caen entre 4.64 y 6.36. Dicho de otra manera, si se hicieran 100 estimaciones para estos niveles de rasgo, 95 de ellas contendrían al verdadero valor del parámetro en ese intervalo.

En el gráfico a continuación, se presenta la función de información de la prueba para todos los niveles de habilidad así como los límites de confianza al 95%.

[7] Ítem Response Theory. Parameter Estimation Techniques. Frank B. Baker

Gráfico 2. Función de Información global



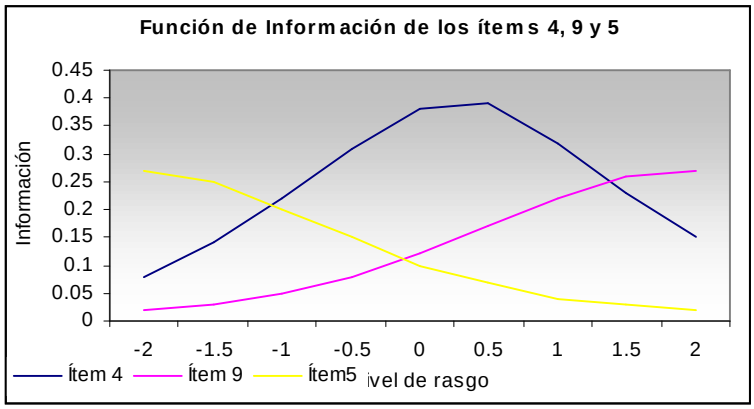
Lo ideal para este tipo de pruebas sería tener una gráfica uniforme para toda la escala de habilidad y con valores altos de información; ya que esto estaría indicando que la prueba evalúa de igual manera a individuos con bajos o altos niveles de habilidad.

En función de la observación del gráfico se puede decir que en conjunto la prueba estima mejor a los individuos con niveles de rasgos entre –1 y 1.5, alcanzando el máximo en θ =0.5 y θ =0.

Al estudiar la función de información para cada uno de los ítems (ver cuadro en el anexo de resultados), se observa que el ítem 10, presenta una información muy baja para todos los niveles de habilidad en la escala de –2 a 2; posiblemente para valores altos de θ presente alta información, ya que como se había visto este ítem es extremadamente difícil por lo que será útil para evaluar a aquellos estudiantes que tengan mucho conocimiento.
Dado que el objetivo de la prueba fue medir conocimientos básicos de Matemática adquiridos en Secundaria este ítem se podría descartar del test.

A continuación se presenta la función de información para 3 de los ítems, los mismos fueron elegidos por su comportamiento diferencial.

Gráfico 3. Función de información de 3 ítems

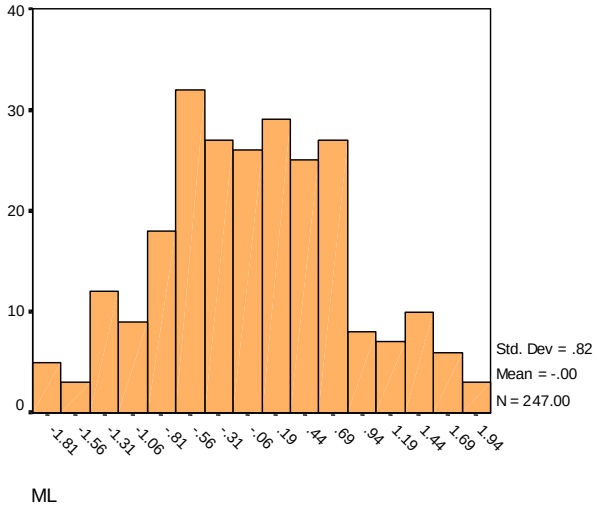


El ítem 5 (amarillo) está estimando mejor a aquellos estudiantes con un nivel de conocimiento bajo; lo contrario ocurre con el ítem 9 (rosado), ya que la información es

mayor cuanto mayor es el nivel de rasgo. Mientras que el ítem 4, se puede decir que es bueno para estimar aquellos estudiantes que tienen un nivel “normal” de rasgo, es decir, que su nivel de conocimiento no es ni muy poco ni demasiado. Esto se deduce del hecho que la información es el recíproco del error y en este ítem la misma es alta para niveles de rasgo entre [-1.5 ; 1.5].
Por otro lado, se analizan las estimaciones de los niveles de conocimiento de los estudiantes, observando las distribuciones de los niveles de rasgo de los individuos.

El gráfico a continuación representa la distribución del nivel de rasgo de los 247 estudiantes.

Gráfico 4. Histograma del Nivel de Rasgo



Es importante remarcar que hay varios estudiantes que tienen un nivel de rasgo muy bajo.

Por otro lado, se observa que el nivel de rasgo de los estudiantes se aproxima a una distribución normal típica. A los efectos de contrastar dicha presunción, se realiza el test de bondad de ajuste de Kolmogorov-Smirnov, donde se observa que no se rechaza la hipótesis de normalidad (el estadístico de prueba, así como su p-valor se pueden ver en el anexo de resultados).

III.2. Resultado de las Estimaciones Bayesianas

En la estimación Bayesiana los parámetros son considerandos una variable aleatoria que sigue una cierta distribución de probabilidad. El objetivo primordial del análisis es encontrar la misma. Tomando en cuenta la información a priori de los parámetros, la estimación Bayesiana provee buenos valores estimados, aunque el tamaño de la muestra sea pequeño, dónde la estimación Máximo Verosímil no es buena.

Distintos estudios han demostrado que a pesar de que a la estimación Bayesiana se le considera subjetiva, la aplicación de la misma para la estimación de los parámetros de TRI es bastante efectiva.

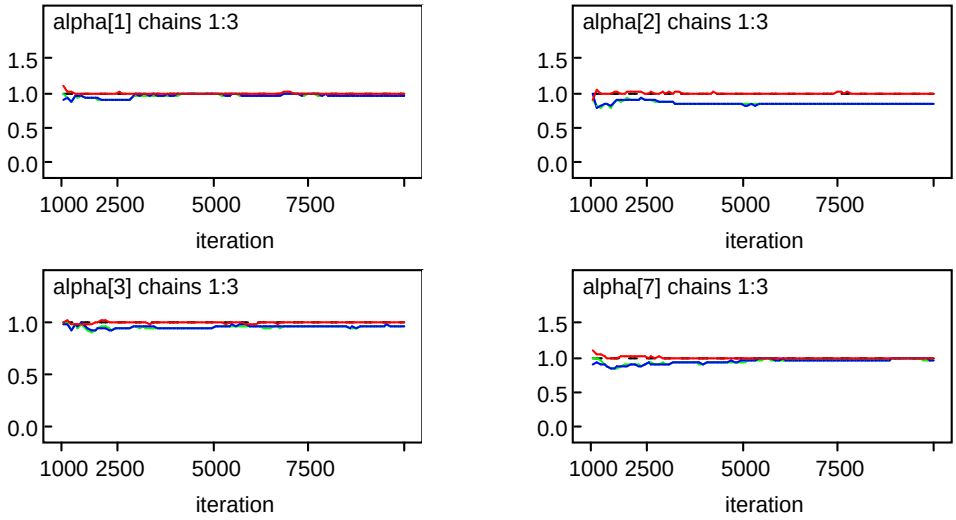
En este trabajo la estimación fue realizada utilizando WINBUGS (la especificación del modelo utilizado se encuentran en el anexo).

En la medida que no se contaba con información a priori respecto a posibles distribuciones de los parámetros de los ítems, se optó por plantear distribuciones a priori del tipo no informativas, las que pueden resumirse de la siguiente forma:

- Distribución Normal (0, 1) para el parámetro de habilidad de los examinados (θ).
- Distribución Normal (0, 10) para el parámetro de dificultad de los ítems (b).
- Distribución Uniforme (0, 3) para el parámetro de discriminación de los ítems (a)

Se corrieron 3 cadenas y 10.000 iteraciones, eliminando las 1000 primeras. El estadístico de Gelman and Rubin observado indica que se obtuvo la convergencia deseada. A su vez, los niveles de error MC están dentro de los márgenes esperados. Esto permite confiar en que los resultados obtenidos refieren a las distribuciones a posteriori buscadas. En el gráfico 5 se presentan, a modo de ejemplo el estadístico de Gelman and Rubin para algunos de los parámetros de dificultad *b*[8]. La totalidad de resultados se puede ver en el anexo de resultados.

Gráficos 5. Estadístico de Gelman Rubin para 4 ítems



Al observar los gráficos se puede afirmar que la estimación de la distribución de las dificultades de los ítems 1, 2, 3 y 7 alcanzó la convergencia. Se aprecia que el estadístico (representado con color rojo en el gráfico) se encuentra en el entorno de 1, como era deseable. Los colores verde y azul son representaciones de la varianza estimada y la varianza between, respectivamente.

A continuación se presentan las distribuciones estimadas para cada uno de los parámetros de los ítems (dificultad y discriminación) y la distribución estimada para los rasgos de los individuos.

[8] Notar que en los gráficos en lugar de b aparece *alpha* porque fue la forma que se especifico en el modelo.

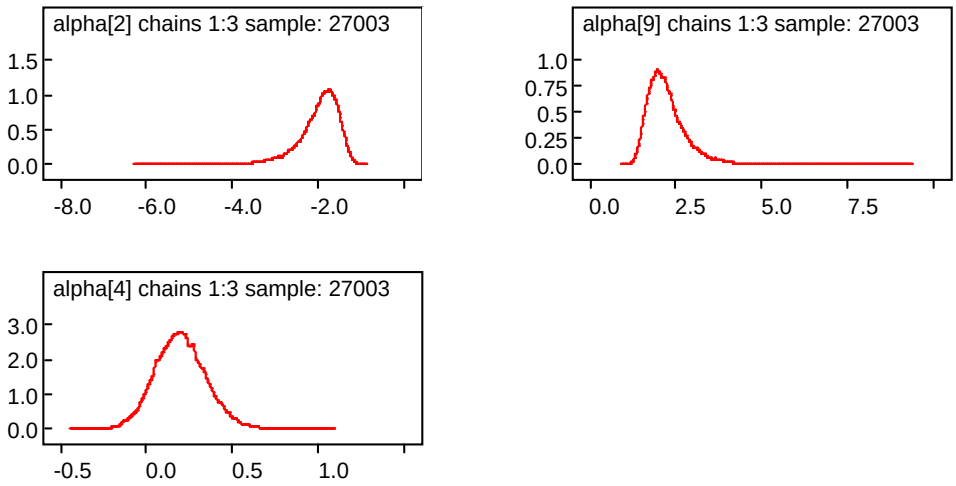
Cuadro 4. Distribución estimada del parámetro de dificultad

Distribución del parámetro de dificultad						
	Media	Desvío estándar	Error MC	2.50%	Mediana	97.50%
ITEM 1	-2.192	0.646	0.015	-3.832	-2.059	-1.362
ITEM 2	-1.987	0.480	0.011	-3.178	-1.898	-1.327
ITEM 3	2.116	0.706	0.014	1.211	1.967	3.905
ITEM 4	0.199	0.151	0.002	-0.084	0.194	0.513
ITEM 5	-2.437	0.665	0.016	-4.122	-2.302	-1.567
ITEM 6	2.174	0.863	0.017	1.087	1.976	4.426
ITEM 7	2.276	0.638	0.015	1.418	2.157	3.864
ITEM 8	-0.810	0.151	0.002	-1.142	-0.799	-0.546
ITEM 9	2.237	0.595	0.013	1.442	2.127	3.711
ITEM 10	5.273	1.509	0.037	2.973	5.060	8.797
ITEM 11	0.143	0.121	0.001	-0.091	0.140	0.386
ITEM 12	0.990	0.225	0.003	0.624	0.963	1.504
ITEM 13	0.023	0.174	0.002	-0.316	0.020	0.370
ITEM 14	0.658	0.192	0.002	0.335	0.640	1.084
ITEM 15	3.188	1.218	0.027	1.566	2.921	6.207
ITEM 16	0.611	0.245	0.003	0.211	0.585	1.175
ITEM 17	0.812	0.311	0.004	0.361	0.767	1.529
ITEM 18	1.067	0.297	0.004	0.619	1.024	1.769
ITEM 19	0.806	0.167	0.002	0.515	0.793	1.171
ITEM 20	0.526	0.181	0.002	0.213	0.511	0.928

El cuadro precedente presenta la media y el desvío de la distribución, así como el intervalo de probabilidad al 95% (intervalo central) de la distribución del parámetro de dificultad de cada uno de los ítems. Nuevamente, se observa que los ítems 10 y 15 son los de mayores niveles de dificultad.

Los gráficos que se presentan a continuación, muestran las distribuciones de 3 de los ítems con comportamiento diferencial entre ellos.

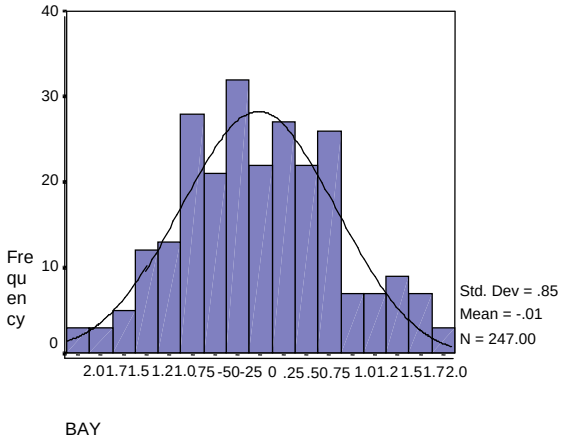
Gráficos 6. Distribuciones del parámetro de dificultad para 3 de los ítems.



La distribución del segundo ítem presenta una distribución unimodal con cola a la izquierda y bastante concentrada en –2, por lo que se puede deducir que es un ítem muy sencillo. No es el caso del ítem 9 cuya distribución está concentrada en 2.3 infiriendo que es un ítem difícil, mientras que la distribución del ítem 4 es unimodal, simétrica (lo cual también se puede verificar analíticamente en el cuadro (ver anexo) de las distribuciones estimadas ya que la media y la mediana prácticamente coinciden) y parece tener una distribución Normal estándar.

Este análisis se puede realizar para las distribuciones estimadas de los 20 parámetros de dificultad, así como para las 20 estimaciones de las distribuciones de los parámetros de discriminación.
De igual manera se obtiene el análisis de las distribuciones estimadas para cada uno de los individuos. Las mismas no son presentadas ya que se consta de una muestra 247 individuos, pero sí se presenta a continuación la distribución de las medias de las estimaciones de las distribuciones de los rasgos de los individuos, por medio de un histograma (si se desean tener los valores de algunos momentos de la distribución y el ajuste a un modelo Normal ver anexo).

Gráfico 7. Histograma de la distribución de las medias estimadas de los parámetros de los individuos.



Se realiza una prueba K-S para testear la normalidad y se obtiene que no se rechaza la hipótesis nula.

III.3.Análisis comparativo

A continuación se presentan distintos estadísticos de la distribución de medias estimadas por medio de los 2 métodos de estimación.

Cuadro 5. Distribución estimada de las habilidades de los individuos.

		Estimación ML	Estimación Bayesiana
N		247	247
Media		-0.002	-0.006
Mediana		-0.050	-0.006
Modo		-1.231	-0.146
Desvío Estándar		0.817	0.851
Rango		3.901	4.072
Mínimo		-1.882	-1.976
Máximo		2.019	2.096
Percentiles	25	-0.620	-0.648
	50	-0.050	-0.006
	75	0.562	0.601

Si se analizan los niveles de habilidad se puede observar que existen diferencias significativas entre la estimación ML y la estimación Bayesiana.
Se observa que las medidas de tendencia central son siempre mayores en la estimación Bayesiana que en la estimación ML, mientras que si se consideran las medidas de dispersión sucede lo contrario.

El cuadro siguiente muestra la media y el desvío estándar del parámetro de dificultad obtenido utilizando estimación ML (columnas azules) y la obtenida por estimación Bayesiana (columnas negras).

Cuadro 5. Estimaciones Máximo Verosimil y Bayesiana de las dificultades de cada ítem.

Estimación ML y Bayesiana de las dificultades de los ítems				
	media (ML)	media (B)	Desvío (ML)	Desvío (B)
ITEM 1	-1.85	-2.192	0.5	0.646
ITEM 2	-1.73	-1.987	0.36	0.480
ITEM 3	1.77	2.116	0.55	0.706
ITEM 4	0.28	0.199	0.16	0.151
ITEM 5	-2.04	-2.437	0.48	0.665
ITEM 6	1.61	2.174	0.56	0.863
ITEM 7	1.91	2.276	0.5	0.638
ITEM 8	-0.85	-0.810	0.14	0.151
ITEM 9	1.89	2.237	0.44	0.595
ITEM 10	4.66	5.273	2.32	1.509
ITEM 11	0.17	0.143	0.13	0.121
ITEM 12	0.96	0.990	0.21	0.225
ITEM 13	-0.06	0.023	0.17	0.174
ITEM 14	0.56	0.658	0.17	0.192
ITEM 15	2.54	3.188	1.23	1.218
ITEM 16	0.64	0.611	0.24	0.245
ITEM 17	0.72	0.812	0.28	0.311
ITEM 18	0.94	1.067	0.23	0.297
ITEM 19	0.77	0.806	0.18	0.167
ITEM 20	0.56	0.526	0.18	0.181

Se observa que las medias son muy parecidas utilizando ambas estimaciones. La diferencia fundamental radica, tal como fuera mencionado con anterioridad, que en la última se obtiene la distribución de probabilidad para la dificultad y discriminación de cada uno de los ítems, así como la distribución de probabilidad de los niveles de rasgo de los individuos.

IV.- Conclusiones

Con respecto a la aplicación de la Teoría de Respuesta al Ítem a tests diagnósticos al ingreso de Facultad, se lograron evaluaciones de cada una de las preguntas propuestas y estimaciones de los "niveles de conocimiento" de los 247 estudiantes seleccionados en la muestra, en el área de matemática.

Se observa que el 90% de las preguntas presentan discriminaciones moderadas o moderadas altas, concluyendo que la prueba propuesta para la evaluación del conocimiento de los estudiantes en el área fue buena aunque se puede decir, que fue un poco difícil, ya que hay 6 de las preguntas que presentan muy altas dificultades, 3 con muy bajas y de las 11 restantes solamente 4 presentan niveles de dificultad "aceptables", esto es, en el entorno de 0.

Por otro lado se obtuvo la distribución del nivel de conocimiento de los estudiantes por medio de las estimaciones de los parámetros de habilidad (niveles de rasgo). Se puede decir que en conjunto la prueba estima mejor a los individuos con niveles de rasgos entre –1 y 1.5, alcanzando el máximo en θ =0.5 y θ =0.

La principal utilidad de esto radica en que se tienen las evaluaciones de las preguntas, desde el punto de vista de la discriminación y de la dificultad que implican cada una de ellas. Al evaluar los resultados de las pruebas de múltiple opción de esta manera se obtiene una distribución de los niveles de conocimiento de los alumnos, que no dependen de la prueba en si misma.

Si se cuenta con un pool de ítems se pueden encontrar algoritmos (por medio de la maximización de la información de los ítems) que busquen los mejores ítems para determinar la prueba mas adecuada dependiendo de cual sea el objetivo de la misma.

Las estimaciones de los parámetros obtenidas por medio de MML y el método Bayesiano dieron resultados similares, tomando en cuenta la media y variación de cada estimación.
Con respecto a esto, hay que destacar la riqueza del método bayesiano, refiriéndose a que se logra tener más información acerca de los parámetros que en el otro.

Los modelos analizados carecen del parámetro de azar o pseudo-azar, elemento que debe ser tenido en cuenta.

Trabajo a realizar en el futuro:

- Realizar las estimaciones Bayesianas cambiando las distribuciones a priori y compararlas.
- Implementación en R de el procedimiento de estimación Bayesiana.
- Estimación Bayesiana del modelo de 3 parámetros y su comparación con el de estimación ML.

V. Referencias

BAKER, F.(1992). “Item Response Theory. Parameter estimation techniques”. Marcel Dekker, Inc., New York.

BAKER, F.(2001). “The basic of Item Response Theory”. ERIC.

GAO, F. and CHEN, L. (2005). “Bayesian or Non-Bayesian: A comparision Study of Item Parameter Estimation in the Three-Parameter Logistic Model”. Applied Measurement in Education. *I8(4)*, 351-380.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (2004). “Bayesian Data Analysis”. Chapman and Hall, New York.

WinBUGS User Manual. (2003). Version 1.4.

ANEXOS

Anexo De Resultados

Estimación Máximo Verosimil

Modelo con dos parámetros

Probabilidad de contestar correctamente dado el nivel de habilidad.

ítem	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
1	0.27	0.36	0.47	0.58	0.68	0.76	0.83	0.89	0.92	0.95	0.97	0.98	0.99
2	0.18	0.28	0.42	0.57	0.71	0.82	0.89	0.94	0.97	0.98	0.99	0.99	1.00
3	0.03	0.04	0.05	0.08	0.11	0.15	0.21	0.28	0.36	0.45	0.54	0.64	0.72
4	0.02	0.03	0.05	0.10	0.17	0.27	0.41	0.57	0.71	0.82	0.90	0.94	0.97
5	0.27	0.38	0.51	0.64	0.75	0.83	0.89	0.93	0.96	0.98	0.99	0.99	0.99
6	0.05	0.06	0.08	0.11	0.15	0.20	0.26	0.32	0.40	0.48	0.56	0.64	0.71
7	0.01	0.02	0.03	0.05	0.07	0.10	0.15	0.22	0.31	0.41	0.52	0.63	0.73
8	0.02	0.04	0.10	0.23	0.43	0.66	0.83	0.93	0.97	0.99	1.00	1.00	1.00
9	0.01	0.01	0.02	0.03	0.05	0.08	0.12	0.19	0.29	0.40	0.53	0.65	0.76
10	0.02	0.02	0.03	0.04	0.05	0.06	0.08	0.10	0.13	0.16	0.20	0.25	0.30
11	0.01	0.01	0.03	0.07	0.14	0.26	0.43	0.63	0.79	0.89	0.95	0.97	0.99
12	0.01	0.01	0.02	0.04	0.08	0.14	0.23	0.36	0.51	0.67	0.79	0.88	0.93
13	0.04	0.07	0.11	0.18	0.27	0.38	0.52	0.65	0.76	0.84	0.90	0.94	0.96
14	0.01	0.02	0.04	0.07	0.12	0.20	0.33	0.48	0.64	0.77	0.86	0.92	0.96
15	0.09	0.11	0.13	0.15	0.18	0.22	0.26	0.30	0.34	0.39	0.44	0.50	0.55
16	0.04	0.06	0.09	0.13	0.19	0.26	0.36	0.47	0.58	0.68	0.77	0.84	0.89
17	0.05	0.07	0.10	0.14	0.20	0.27	0.36	0.46	0.56	0.65	0.74	0.81	0.86
18	0.01	0.02	0.04	0.07	0.11	0.17	0.27	0.38	0.52	0.65	0.76	0.84	0.90
19	0.00	0.01	0.02	0.04	0.07	0.14	0.25	0.40	0.58	0.74	0.86	0.93	0.96
20	0.02	0.03	0.05	0.09	0.15	0.23	0.35	0.48	0.62	0.74	0.83	0.90	0.94

Función de Información de cada ítem									
ITEM 1	0.19	0.19	0.17	0.14	0.11	0.08	0.05	0.04	0.02
ITEM 2	0.36	0.37	0.31	0.22	0.14	0.09	0.05	0.03	0.02
ITEM 3	0.03	0.04	0.06	0.07	0.09	0.11	0.13	0.14	0.14
ITEM 4	0.08	0.14	0.22	0.31	0.38	0.39	0.32	0.23	0.15
ITEM 5	0.27	0.25	0.2	0.15	0.1	0.07	0.04	0.03	0.02
ITEM 6	0.03	0.04	0.06	0.07	0.08	0.09	0.1	0.11	0.11
ITEM 7	0.02	0.03	0.05	0.07	0.1	0.14	0.17	0.19	0.2
ITEM 8	0.33	0.62	0.87	0.8	0.5	0.24	0.1	0.04	0.02
ITEM 9	0.02	0.03	0.05	0.08	0.12	0.17	0.22	0.26	0.27
ITEM 10	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.04	0.04
ITEM 11	0.08	0.16	0.29	0.47	0.6	0.57	0.41	0.24	0.12
ITEM 12	0.04	0.07	0.11	0.19	0.29	0.37	0.41	0.36	0.27
ITEM 13	0.11	0.17	0.22	0.27	0.29	0.26	0.21	0.15	0.1
ITEM 14	0.06	0.1	0.17	0.27	0.36	0.41	0.38	0.29	0.19
ITEM 15	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.04
ITEM 16	0.06	0.09	0.12	0.16	0.18	0.2	0.19	0.17	0.14
ITEM 17	0.06	0.08	0.1	0.13	0.15	0.16	0.16	0.15	0.13
ITEM 18	0.05	0.07	0.11	0.17	0.23	0.27	0.29	0.26	0.21
ITEM 19	0.04	0.07	0.14	0.25	0.4	0.51	0.52	0.41	0.26
ITEM 20	0.06	0.1	0.16	0.22	0.28	0.31	0.29	0.24	0.17

Test de Kolmogorov-Smirnov

One-Sample Kolmogorov-Smirnov Test		
		Thetas
N		247
Normal Parameters	Mean	-0.002
	Std. Deviation	0.817
Most Extreme Differences	Absolute	0.035
	Positive	0.035
	Negative	-0.030
Kolmogorov-Smirnov Z		0.546
Asymp. Sig. (2-tailed)		0.927
a Test distribution is Normal. b Calculated from data.		

Modelo con tres parámetros

Primeramente se analizan las estimaciones de los parámetros de dificultad, discriminación y azar, y posteriormente la función de información y los niveles de rasgo.

Los valores de las estimaciones de los parámetros de los ítem obtenidos para esta prueba son los siguientes:

Estimación de los parámetros de los ítems				
Ítem	**A**	**b**	**C**	**% correctas**
1	0.56	-1.29	0.27	80.6
2	0.93	-1.08	0.35	84.6
3	0.75	1.88	0.12	22.7
4	1.21	0.59	0.17	42.9
5	0.65	-1.66	0.24	85.8
6	0.79	1.81	0.17	27.1
7	1.02	1.79	0.1	18.2
8	1.42	-0.54	0.23	73.7
9	0.87	1.89	0.07	15.8
10	0.59	4.57	0.08	8.5
11	1.66	0.49	0.18	44.9
12	0.89	1.11	0.08	27.5
13	0.89	0.41	0.21	51
14	1.18	0.81	0.14	36
15	0.39	3.26	0.16	25.9
16	0.68	0.96	0.13	37.7
17	0.81	1.13	0.19	37.2
18	0.74	1.15	0.09	30
19	1.52	0.93	0.12	30
20	0.69	0.75	0.08	37.2

De acuerdo a lo mencionado anteriormente, se recuerda que un parámetro de dificultad ***b*** negativo, significa que el ítem sería fácil y será más fácil, cuanto más grande en magnitud sea. Y viceversa, cuánto más positivo es, más difícil será el ítem.

Observando las estimaciones de los parámetros se ve que 3 de los 20 ítems son muy fáciles (se asocian a los porcentajes de respuestas más altos), mientras que 9 de ellos son difíciles (la mayoría con un porcentaje de respuestas correctas inferior al 30%).

De acuerdo a esta teoría, el ítem 10 sería descartado del test, ya que su dificultad es 4.57; esta pregunta, la contestó correctamente sólo un 8.5 % de los estudiantes, lo que provoca una mala estimación. Se puede decir entonces, que éste es un ítem muy difícil. En tanto, un ítem ideal sería, por ejemplo, el ítem 11, que presenta una dificultad normal (0.5, tendiendo a ser un poco difícil) y un parámetro de discriminación alto.

Para cada ítem (luego de obtener las estimaciones de los parámetros), se puede calcular la probabilidad de que un individuo conteste correctamente a ese ítem, según su nivel de rasgo.

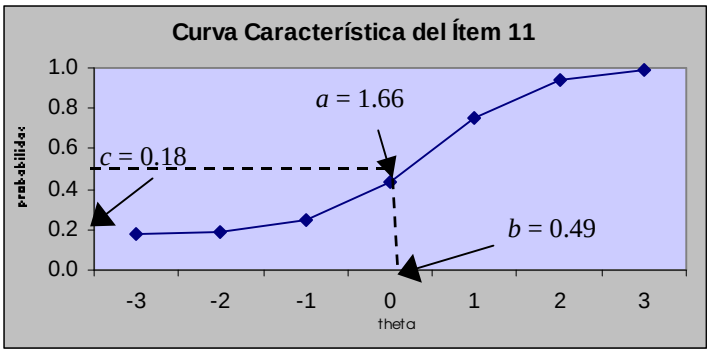
Se presentan a continuación las probabilidades para cada una de las preguntas, por nivel de habilidad:

Probabilidad de contestar correctamente dado el nivel de rasgo							
Ítem	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
1	0,47	0,56	0,66	0,76	0,84	0,90	0,94
2	0,44	0,54	0,69	0,83	0,92	0,96	0,99
3	0,14	0,17	0,21	0,29	0,42	0,58	0,73
4	0,18	0,20	0,28	0,44	0,69	0,87	0,96
5	0,46	0,58	0,70	0,81	0,89	0,94	0,96
6	0,19	0,21	0,25	0,33	0,46	0,62	0,77
7	0,11	0,12	0,15	0,22	0,38	0,60	0,80
8	0,25	0,32	0,49	0,76	0,92	0,98	0,99
9	0,08	0,10	0,14	0,22	0,36	0,56	0,74
10	0,09	0,10	0,11	0,14	0,18	0,25	0,34
11	0,18	0,19	0,24	0,43	0,75	0,94	0,99
12	0,10	0,13	0,20	0,33	0,52	0,71	0,86
13	0,25	0,29	0,39	0,53	0,71	0,85	0,93
14	0,15	0,17	0,23	0,38	0,62	0,83	0,94
15	0,23	0,26	0,29	0,34	0,41	0,48	0,56
16	0,19	0,23	0,31	0,43	0,57	0,71	0,83
17	0,22	0,25	0,31	0,42	0,57	0,73	0,85
18	0,13	0,17	0,24	0,36	0,52	0,68	0,82
19	0,12	0,13	0,16	0,29	0,58	0,86	0,96
20	0,14	0,20	0,29	0,42	0,58	0,73	0,84

En este cuadro, se puede verificar que, por ejemplo, los ítems 1, 2 y 5 son muy fáciles, ya que, para niveles de rasgo muy bajos, la probabilidad de que el estudiante conteste correctamente este ítem, está en el entorno de 0.5 (0.47, 0.44, 0.46 respectivamente). Esta probabilidad es no es cercana a 0.20, probabilidad de que un individuo sin conocimiento conteste correctamente el ítem (sólo 1 de 5 opciones es correcta, esto es equivalente a decir que contesta al azar).

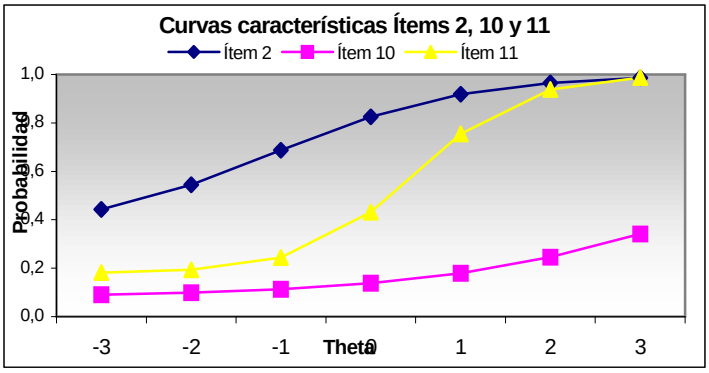
Esta información, está corroborando la obtenida por el parámetro de Dificultad, que en estas 3 preguntas es de –1.29 , –1.08 y –1.66 para las preguntas 1,2 y 5 respectivamente.

A continuación se presenta la gráfica de la probabilidad de responder correctamente al ítem 11 para cada nivel de rasgo, obteniendo la "Curva característica del ítem 11" (se representa la probabilidad de responder correctamente versus nivel de rasgo).



En la misma se observa que, para niveles de habilidad bajos, la probabilidad de que la respuesta sea correcta es baja, y va subiendo según aumente el nivel de rasgo.
Cabría preguntarse por qué la probabilidad para un nivel de rasgo muy pequeño no es 0 sino 0.20. La respuesta a esto, ya mencionada antes, es que en este tipo de modelos (de 3 parámetros), la probabilidad de responder correctamente está acotada inferiormente por el parámetro de azar.

Para tener una idea de cómo serían las curvas para distintos tipos de ítems, se presentan en forma conjunta un ítem fácil, uno normal y uno difícil, se eligieron los ítems 2, 10 y 11.



El ítem 10 es muy difícil, como fuera mencionado anteriormente: la probabilidad de responder correctamente es muy baja para todos los niveles de habilidad, siendo la probabilidad más alta que tiene el individuo de responderlo correctamente de 0.34. El porcentaje de estudiantes que respondieron correctamente este ítem es 8.5 %. En sentido contrario, el ítem 2 es un ítem fácil, ya que aún para niveles de habilidad bajos, la probabilidad es mayor que 0.4 y la probabilidad de respuesta correcta es de 0.90 para niveles altos de habilidad. Par el ítem medio y el fácil las probabilidades de respuesta correcta se igualan para los niveles de habilidad altos.

Al analizar la discriminación de cada ítem se compara los mismos con niveles estándar.
Se presenta a continuación la escala propuesta[9] por Baker, para la clasificación del parámetro de Discriminación (***a***):

[9] The Basic of Ítem Response Theory. Frank B. Baker

Calificación	Discriminación
nula	0
muy baja	0,01 a 0,34
baja	0,35 a 0,64
moderada	0,65 a 1,34
alta	1,35 a 1,69
muy alta	mayor a 1,7

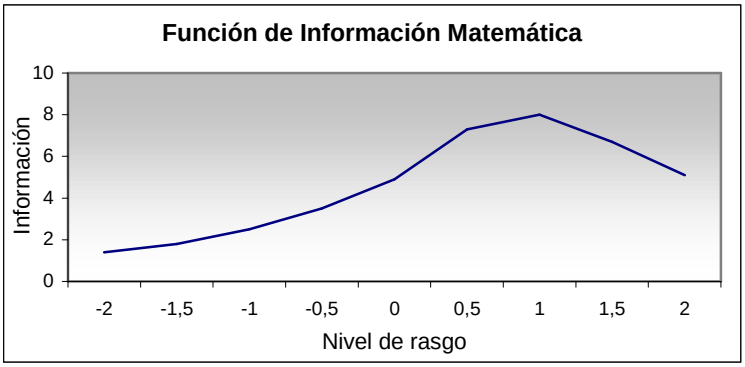
Acorde a esta clasificación, los ítems de la prueba de Matemática se pueden clasificar en: 3 con discriminación baja (ítem 1,10,15), 14 con discriminación moderada y 3 con discriminación alta(ítem 8,11,19).

Función de información

La función de información permite ver la calidad de las preguntas hechas.

theta	Información	desvío	IC al 95%	IC al 68%
-2	1.4	0.84	[-3.65 ; -0.35]	[-2.84 ; -1.16]
-1.5	1.8	0.75	[-2.97 ; -0.03]	[-2.25 ; -0.75]
-1	2.5	0.64	[-2.97 ; 0.25]	[-1.64 ; -0.36]
-0.5	3.5	0.54	[-1.56 ; 0.56]	[-1.04 ; 0.04]
0	4.9	0.45	[-0.88 ; 0.88]	[-0.45 ; 0.45]
0.5	7.3	0.37	[-0.23 ; 1.23]	[0.13 ; 0.87]
1	8	0.35	[0.31 ; 1.69]	[0.65 ; 1.35]
1.5	6.7	0.39	[0.74 ; 2.26]	[1.11 ; 1.89]
2	5.1	0.44	[1.14 ; 2.86]	[1.56 ; 2.44]

El máximo de la función de información se da para □= 1 y vale 8 con un error estándar de 0.35, por lo que el 95% de las estimaciones de estos niveles de habilidad caen entre 0.31 y 1.69. Dicho de otra manera, si se hicieran 100 estimaciones para estos niveles de rasgo, 95 de ellas contendrían al verdadero valor del parámetro en ese intervalo.
En el gráfico siguiente, se presenta la función de información de la prueba de Matemática en toda la escala de habilidad.



Como se mencionó antes, lo ideal para este tipo de pruebas sería tener una gráfica uniforme para toda la escala de habilidad y con mucha información; ya que esto estaría indicando que la prueba evalúa de igual manera a individuos con bajo nivel de

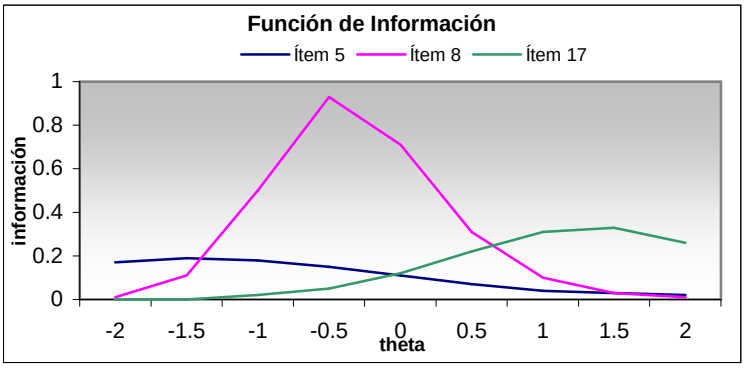
habilidad que a individuos con alto nivel de habilidad. Del mismo modo, siempre va a depender de lo que se quiera medir.

En esta gráfica, sin embargo, se observa que individuos con bajo nivel de habilidad no son estimados muy bien, ya que la función de información decrece considerablemente para valores menores a -1. Para niveles de rasgo mayor que -0.5, la información aumenta, por lo que la estimación para estos niveles se hace con más precisión. Podemos decir que la prueba de Matemática en su conjunto estima mejor a los individuos con un nivel de rasgo superior a –0.5, alcanzando el máximo en □.□ı

Se presenta la tabla de información para cada nivel de rasgo de cada uno de los ítems de la prueba.

Función de Información									
Ítem	**-2**	**-1.5**	**-1**	**-0.5**	**0**	**0.5**	**1**	**1.5**	**2**
1	0.1	0.13	0.14	0.13	0.11	0.08	0.06	0.04	0.03
2	0.1	0.21	0.31	0.29	0.2	0.11	0.06	0.03	0.01
3	0	0	0.01	0.02	0.05	0.1	0.19	0.28	0.32
4	0	0	0.02	0.12	0.4	0.72	0.68	0.39	0.17
5	0.17	0.19	0.18	0.15	0.11	0.07	0.04	0.03	0.02
6	0	0	0	0.01	0.04	0.09	0.19	0.29	0.33
7	0	0	0	0.01	0.03	0.12	0.32	0.55	0.61
8	0.01	0.11	0.5	0.93	0.71	0.31	0.1	0.03	0.01
9	0	0	0	0.02	0.05	0.13	0.27	0.41	0.48
10	0	0	0	0	0	0	0.01	0.01	0.03
11	0	0	0.01	0.09	0.62	1.38	0.96	0.33	0.09
12	0	0.01	0.03	0.08	0.2	0.36	0.48	0.46	0.34
13	0	0.02	0.06	0.16	0.29	0.38	0.34	0.24	0.14
14	0	0	0.01	0.08	0.28	0.63	0.76	0.53	0.26
15	0	0	0.01	0.01	0.01	0.02	0.03	0.04	0.06
16	0.01	0.02	0.04	0.09	0.15	0.22	0.26	0.24	0.2
17	0	0	0.02	0.05	0.12	0.22	0.31	0.33	0.26
18	0	0.01	0.04	0.08	0.16	0.25	0.32	0.32	0.26
19	0	0	0	0.02	0.18	0.78	1.32	0.87	0.33
20	0.02	0.04	0.08	0.14	0.22	0.28	0.29	0.25	0.19

Al estudiar la función de información para cada uno de los ítems de Matemática, se observa que el ítem 10, presenta una información muy baja para los niveles de habilidad en la escala de –2 a 2; posiblemente para valores altos de □ presente alta información, ya que como se había visto, este ítem es extremadamente difícil por lo que sería muy útil para evaluar a aquellos estudiantes que tienen mucho conocimiento. Dado que la prueba pretendía medir conocimientos básicos de Matemática adquiridos en Secundaria, este ítem se podría descartar del test.
Se presenta la función de información para 3 de los ítems.
Se eligieron los mismos ya que presentan un comportamiento muy diferencial entre



ellos.

El ítem 5 (azul) está estimando mejor a aquellos estudiantes con un nivel de conocimiento bajo; lo contrario ocurre con el ítem verde (17), ya que la información es mayor cuanto mayor es el nivel de rasgo. Mientras que, por la función de información del ítem 8, se puede decir que es bueno para estimar a aquellos estudiantes que tienen un nivel "normal" de rasgo, es decir, que su nivel de conocimiento no es ni muy poco ni demasiado. Esto se deduce de que la función de información con un nivel de rasgo entre [-1.5 ; 1.5] es alta, por lo que el error de estimación en ese intervalo va a ser pequeño, dado que la información es el inverso del error.

Por otro lado, se analizan las estimaciones de los niveles de rasgo (conocimiento) de los estudiantes.

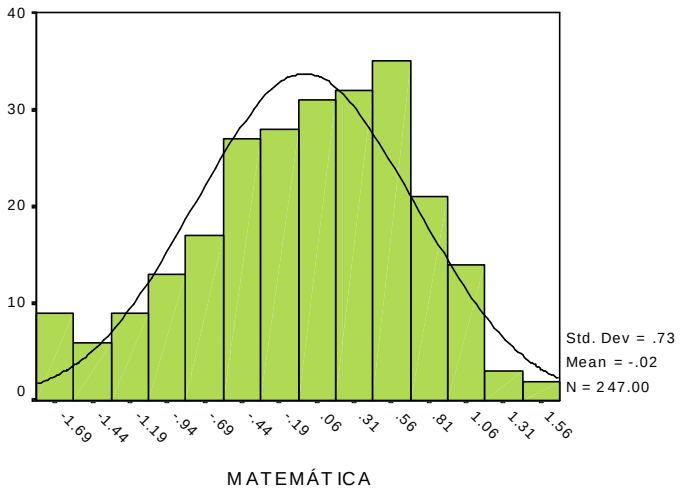
Para ello se decide clasificarlos en 3 categorías de niveles de conocimiento: bajo a aquellos estudiantes que tienen un □ menor o igual a -1, medio a aquellos en el que su rasgo está entre –1 (inclusive) y 1 y alto los estudiantes con □ mayor que 1.

Esta información se presenta resumida en el siguiente cuadro.

Nivel de rasgo	Frecuencia	Porcentaje
bajo	26	11
medio	192	78
alto	29	12
Total	247	100

El gráfico a continuación representa la distribución del nivel de rasgo de los estudiantes del área de Matemática.

Histograma MATEMÁTICA

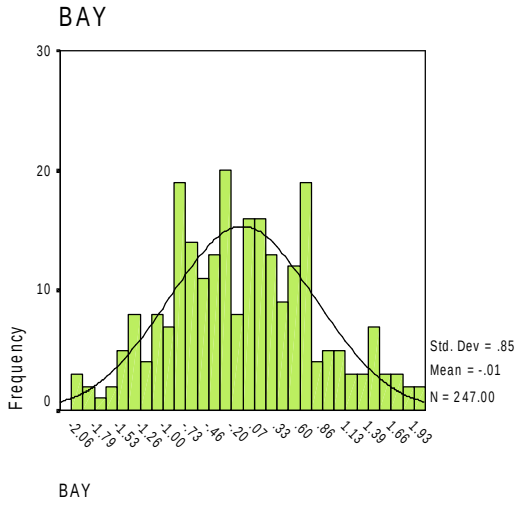


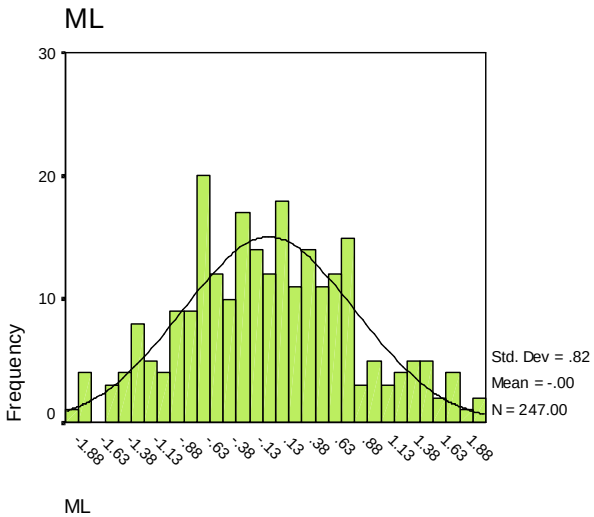
Se observa que el rasgo de los estudiantes de Matemática tiene una distribución aproximadamente normal típica. Es importante remarcar que hay varios estudiantes que tienen un nivel de rasgo muy bajo.

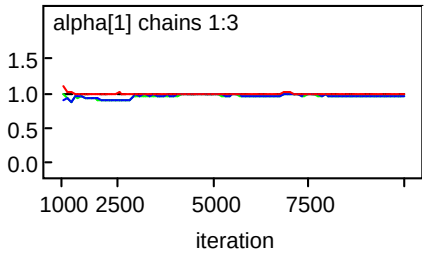
Estimación Bayesiana

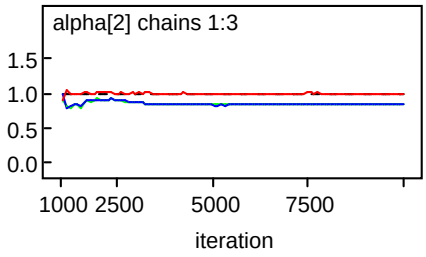
Distribución del parámetro de dificultad							
	mean	media*1.7	sd	MC error	2.50%	median	97.50%
ITEM 1	0.500	**0.850**	0.141	0.003	0.244	0.492	0.799
ITEM 2	0.680	**1.156**	0.184	0.004	0.358	0.667	1.076
ITEM 3	0.383	**0.651**	0.113	0.002	0.182	0.378	0.617
ITEM 4	0.687	**1.168**	0.144	0.002	0.427	0.680	0.993
ITEM 5	0.587	**0.997**	0.167	0.004	0.292	0.576	0.948
ITEM 6	0.290	**0.494**	0.098	0.002	0.122	0.284	0.501
ITEM 7	0.492	**0.836**	0.136	0.003	0.252	0.483	0.777
ITEM 8	1.093	**1.858**	0.249	0.004	0.674	1.068	1.653
ITEM 9	0.562	**0.956**	0.151	0.003	0.292	0.554	0.885
ITEM 10	0.286	**0.486**	0.087	0.002	0.152	0.274	0.486
ITEM 11	0.965	**1.640**	0.190	0.002	0.631	0.951	1.380
ITEM 12	0.688	**1.170**	0.149	0.002	0.421	0.680	0.999
ITEM 13	0.544	**0.924**	0.125	0.001	0.314	0.538	0.806
ITEM 14	0.659	**1.119**	0.144	0.002	0.398	0.650	0.961
ITEM 15	0.239	**0.406**	0.085	0.002	0.104	0.230	0.428
ITEM 16	0.484	**0.823**	0.119	0.001	0.265	0.479	0.731
ITEM 17	0.454	**0.772**	0.118	0.001	0.235	0.450	0.697
ITEM 18	0.537	**0.913**	0.129	0.002	0.302	0.530	0.808
ITEM 19	0.878	**1.493**	0.177	0.002	0.568	0.866	1.261
ITEM 20	0.642	**1.091**	0.139	0.002	0.391	0.634	0.935

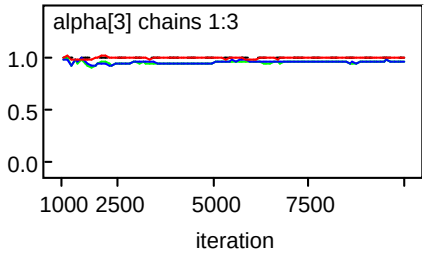
One-Sample Kolmogorov-Smirnov Test		
		Thetas
N		247
Normal Parameters	Mean	-0.006
	Std. Deviation	0.851
Most Extreme Differences	Absolute	0.039
	Positive	0.039
	Negative	-0.029
Kolmogorov-Smirnov Z		0.612
Asymp. Sig. (2-tailed)		0.848
a Test distribution is Normal. b Calculated from data.		

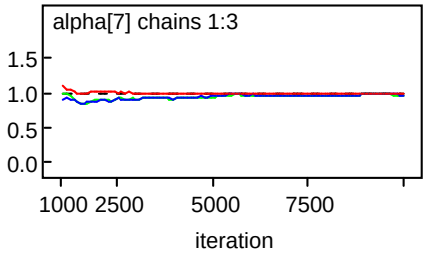


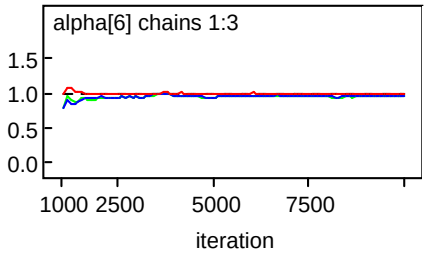


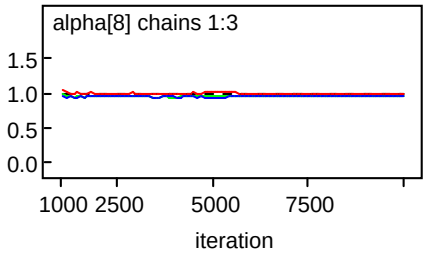

alpha[1] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

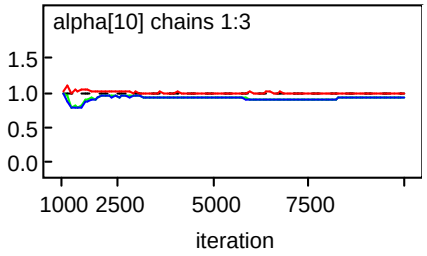

alpha[2] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

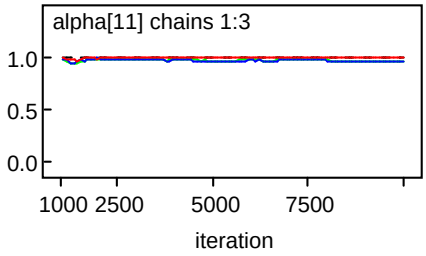

alpha[3] chains 1:3
1.0
0.5
0.0
1000 2500 5000 7500
iteration

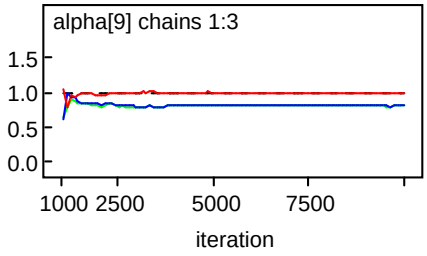

alpha[7] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

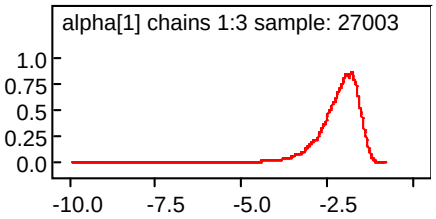

alpha[6] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

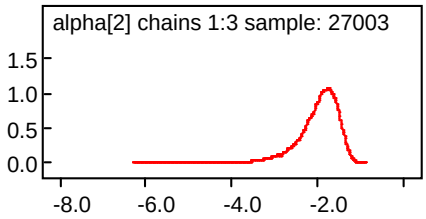

alpha[8] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

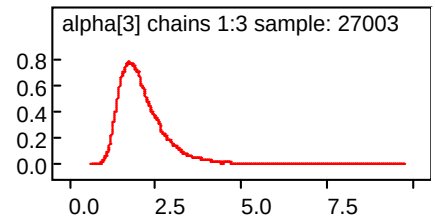

alpha[10] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

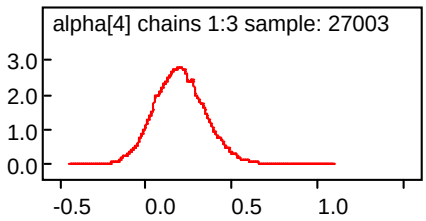

alpha[11] chains 1:3
1.0
0.5
0.0
1000 2500 5000 7500
iteration

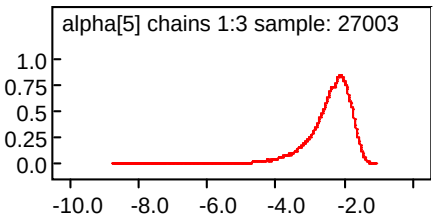

alpha[9] chains 1:3
1.5
1.0
0.5
0.0
1000 2500 5000 7500
iteration

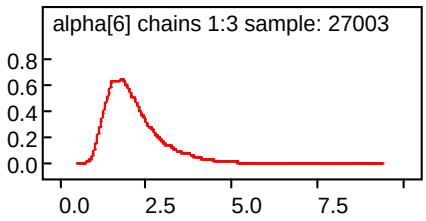

alpha[1] chains 1:3 sample: 27003
1.0
0.75
0.5
0.25
0.0
-10.0
-7.5
-5.0
-2.5

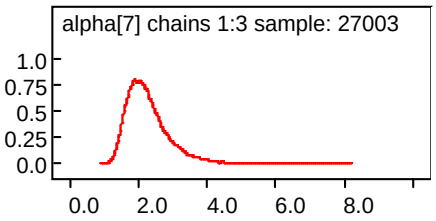

alpha[2] chains 1:3 sample: 27003
1.5
1.0
0.5
0.0
-8.0
-6.0
-4.0
-2.0

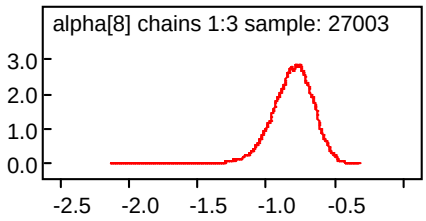

alpha[3] chains 1:3 sample: 27003
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5

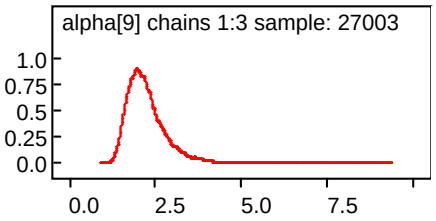

alpha[4] chains 1:3 sample: 27003
3.0
2.0
1.0
0.0
-0.5
0.0
0.5
1.0

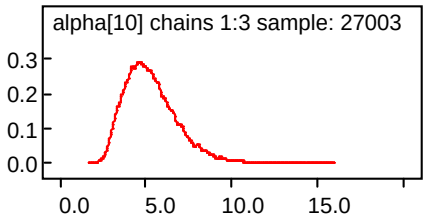

alpha[5] chains 1:3 sample: 27003
1.0
0.75
0.5
0.25
0.0
-10.0
-8.0
-6.0
-4.0
-2.0


alpha[6] chains 1:3 sample: 27003
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5


alpha[7] chains 1:3 sample: 27003
1.0
0.75
0.5
0.25
0.0
0.0
2.0
4.0
6.0
8.0


alpha[8] chains 1:3 sample: 27003
3.0
2.0
1.0
0.0
-2.5
-2.0
-1.5
-1.0
-0.5


alpha[9] chains 1:3 sample: 27003
1.0
0.75
0.5
0.25
0.0
0.0
2.5
5.0
7.5


alpha[10] chains 1:3 sample: 27003
0.3
0.2
0.1
0.0
0.0
5.0
10.0
15.0

MODELO

```
model
	{
	# Calculate individual (binary) responses to each test from multinomial data
		for (j in 1 : culm[1]) {
			for (k in 1 : T) {  r[j, k] <- response[1, k] }
		}
		for (i in 2 : R) {
			for (j in culm[i - 1] + 1 : culm[i]) {
				for (k in 1 : T) {  r[j, k] <- response[i, k] }
			}
		}
	# Rasch model de 2 parámetros
		for (j in 1 : N) {
			for (k in 1 : T) {
				logit(p[j, k]) <- 1.7*disc[k]* (theta[j] - alpha[k])
				r[j, k] ~ dbern(p[j, k])
			}
			theta[j] ~ dnorm(0, 1)
		}
	# Priors
		for (k in 1:T) {
			alpha[k] ~ dnorm(0, 0.1);   a[k] <- alpha[k] - mean(alpha[])
			disc[k] ~ dunif(0,1)
                                }
		#beta ~ dnorm(0,0.1) I(0,)
	}
```