IIMEC


Universidad de Costa Rica

Instituto de Investigación para el Mejoramiento de la Educación Costarricense

Maestría en Evaluación Educativa


# Análisis de la Calidad Técnica Pruebas de Noveno Año Fórmula 11 Matemáticas


Elaborado por:

Eiliana Montero Rojas



Diciembre 1998

# ANALISIS DE CALIDAD TÉCNICA PRUEBA DE NOVENO AÑO MATEMÁTICAS

## INTRODUCCIÓN

Este informe es presentado por el IIMEC al Ministro de Educación y al Departamento de Control de Calidad del M.E.P. como parte de un estudio de calidad técnica de las pruebas de bachillerato y de noveno año. Corresponde a un análisis psicométrico de la prueba de Matemáticas, fórmula 11, aplicada en 1997 a estudiantes de noveno año.

El análisis consistió inicialmente en la aplicación de un análisis de factores, técnica estadística que permite identificar las dimensiones subyacentes en un conjunto de datos. En este caso particular la técnica se utilizó para establecer si la prueba es de naturaleza unidimensional, es decir, si mide fundamentalmente un solo factor, o si por el contrario se presentan subcomponentes dentro del constructo general que se desea medir. En este último caso interesaba establecer la relación entre estos subcomponentes y los objetivos de aprendizaje que representan los ítemes. Este análisis da evidencia acerca de la validez de constructo asociada a la prueba.

Por otra parte, se aplicó un modelo de Teoría de Respuesta a los ítemes para concluir sobre las propiedades psicométricas de la prueba en general y de los ítemes individuales que la componen. Este enfoque permite analizar los ítemes en términos de su precisión y su poder discriminatorio, es decir su capacidad para poder diferenciar correctamente entre estudiantes con puntajes altos y bajos. Además, establece los niveles de habilidad de los examinados para los cuales el ítem provee mayor información.

Este estudio es de naturaleza preliminar y exploratoria. Por tanto, se considera necesario continuar los análisis a mayor profundidad, y se recomienda establecer mecanismos formales entre el IIMEC y el Ministerio de Educación para continuar investigando en estas y otras temáticas asociadas con la medición educativa en torno a la construcción, aplicación, calificación y análisis de las pruebas nacionales.

## OBJETIVOS

1. Realizar un análisis de factores para encontrar evidencias de validez de constructo en la prueba de Matemáticas de noveno año, fórmula 11, aplicada en 1997.
2. Analizar, mediante la Teoría de Respuesta a los ítemes, aspectos de la calidad técnica de los ítemes de la mencionada prueba, incluyendo técnicas para determinar la confiabilidad de la prueba y las propiedades psicométricas individuales para cada uno de los ítemes que la componen.

3. Detectar posible evidencia de sesgos (comportamiento diferencial del item) por categorías de población, comparando estudiantes provenientes de colegios públicos y colegios privados y comparando estudiantes de colegios urbanos y rurales.

## MUESTRA

Se trabajó con una muestra aleatoria de 3000 estudiantes escogidos de la base de datos de la prueba de matemáticas aplicada a estudiantes de noveno año en 1997. Se escogió el formulario 11 por ser el que fue administrado a un mayor número de estudiantes.

La base de datos original suministrada por el Departamento de Control de Calidad del MEP, se procesó y preparó para ser analizada mediante los procedimientos de análisis de factores en el paquete estadístico SPSS, y Teoría de Respuesta a los Ítemes utilizando el paquete psicométrico BILOG.

## BASE TEÓRICA CONCEPTUAL

La utilidad que tenga un instrumento de medición para la toma de decisiones está directamente relacionada con su calidad técnica, es decir, con las propiedades psicométricas que presente. Este asunto cobra especial importancia cuando se trata de pruebas estandarizadas tales como las de bachillerato, las de sexto grado o las pruebas de admisión a las universidades, pues las decisiones que se toman de acuerdo con sus resultados pueden afectar de una manera profunda la vida de las personas.

Por muchos años la teoría clásica de los tests ha servido de marco conceptual para el análisis psicométrico de pruebas estandarizadas. Por ejemplo, en nuestro país se ha aplicado rutinariamente para el análisis y la construcción del banco de ítemes de la Prueba de Aptitud Académica de la UCR.

Sin embargo, la teoría clásica de los tests presenta ciertos problemas fundamentales desde el punto de vista de la medición. Primero, los parámetros de los ítemes dependen del grupo de examinados a los que se administra la prueba. Los tan conocidos índices de discriminación (correlación entre el puntaje del ítem y el puntaje total de la prueba) y dificultad (porcentaje de respuestas correctas) variarán dependiendo del grupo examinado. Asimismo, los puntajes totales o calificaciones en el constructo de interés para cada examinado dependerán del instrumento de medición utilizado. Así, pruebas más fáciles producirán puntajes más elevados, y pruebas más difíciles resultarán en puntajes más bajos. Además, los índices clásicos de confiabilidad (el más conocido es el Alfa de Cronbach) tienen el inconveniente de que son medidas globales y no permiten establecer qué precisión tienen la prueba o los ítemes según las diferentes categorías de puntaje de los examinados.

La Teoría de Respuesta a los Ítemes (TRI) busca subsanar estas debilidades de la teoría clásica, pues intenta obtener estimaciones de los parámetros del ítem que sean independientes de la muestra de examinados y puntajes o calificaciones que sean independientes del instrumento de medición utilizado. Además, se logran estimaciones individuales de la precisión de la prueba y de los ítemes para cada uno de los diferentes puntajes totales o calificaciones que reciben los examinados. Estas características de la TRI le dan ventajas esenciales en relación con la teoría clásica, que permiten, entre otras cosas, la construcción de pruebas adaptadas al nivel del examinado, una mayor precisión en las estimaciones de los puntajes totales en el constructo o variable de interés y una comparación (equiparación) más fácil de puntajes resultantes de diferentes instrumentos de medición.

Un concepto fundamental en la TRI es el de Curva Característica del Ítem (CCI). Esta consiste en el ajuste de un modelo matemático al comportamiento del ítem. En el eje horizontal se representa el puntaje total que obtiene cada examinado en la variable de interés, o sea, su calificación. Generalmente esta escala se ajusta para que tenga el mismo rango de una variable normal estándar, es decir, que varíe entre -3 y +3. En el eje vertical se representan las probabilidades de respuesta correcta. Así, cada punto en la CCI representa la probabilidad de respuesta correcta para el puntaje correspondiente. El modelo matemático que se ajusta es usualmente la función logística en uno, dos y tres parámetros. Estos parámetros se definen así:

a: Índice de discriminación del ítem. Es proporcional a la pendiente de la CCI en el punto de inflexión. Entre mayor sea su valor mayor es la discriminación.

b: Índice de dificultad. Es el valor del puntaje o calificación que corresponde al punto de inflexión de la CCI. Valores bajos indican ítemes fáciles y valores altos ítemes difíciles.

c: Probabilidad de acertar un ítem al azar. Su valor está entre cero y uno.

Otro concepto fundamental en la TRI es el de la función de información. Esta es un indicador de la precisión de la prueba. Entre más alta sea la función mayor será el nivel de precisión alcanzado por el ítem. Usualmente su forma es de curva normal. Su valor está en función de los puntajes o calificaciones, lo que implica que la precisión del test no es uniforme a lo largo de la escala (no es la misma para todos los examinados). Tanto la función de información como la Curva Característica del Ítem se representan gráficamente, lo cual resulta una herramienta muy útil en el momento de juzgar las propiedades psicométricas de los reactivos.

## ANÁLISIS BAJO LA TEORÍA CLÁSICA DE LOS TESTS

Los resultados que se presentan seguidamente corrresponden a la salida del paquete SPSS en lo que toca al análisis psicométrico de la prueba bajo los supuestos de la Teoría Clásica de los Tests. El primer cuadro presenta, bajo la columna " Mean", las dificultades de cada uno de los ítemes, es decir, los porcentajes de respuesta correcta obtenidos.

Es importante hacer notar que el ítem 26 de la prueba no fue analizado, ya que no aparecía su código para la respuesta correcta en la base de datos suministrada por el MEP. De esta forma a partir del ítem 26 analizado en este estudio la numeración correspondiente en la prueba equivale al número del ítem analizado en el estudio más 1. Por ejemplo el ítem 26 de este estudio corresponde al 27 de la prueba, y así sucesivamente, hasta el 50 que corresponde al 51 de la prueba.

De la observación de este cuadro se puede concluir que solamente tres ítemes se pueden clasificar como fáciles, con más de un 60% de respuestas correctas. La mayoría de los ítemes se clasifican en los niveles de dificultad medios y altos, con algunos casos extremos de dificultad muy elevada (ítemes 3, 40 y 42 de los analizados).

# Cuadro No. 1: Dificultades de los Itemes

RELIABILITY ANALYSIS - SCALE (ALPHA)

| | | Mean | Std Dev | Cases |
|---|---|---|---|---|
| 1. | R01 | .3953 | .4890 | 3000.0 |
| 2. | R02 | .3277 | .4694 | 3000.0 |
| 3. | R03 | .1210 | .3262 | 3000.0 |
| 4. | R04 | .2023 | .4018 | 3000.0 |
| 5. | R05 | .2503 | .4333 | 3000.0 |
| 6. | R06 | .4603 | .4985 | 3000.0 |
| 7. | R07 | .5000 | .5001 | 3000.0 |
| 8. | R08 | .3303 | .4704 | 3000.0 |
| 9. | R09 | .4313 | .4953 | 3000.0 |
| 10. | R10 | .3570 | .4792 | 3000.0 |
| 11. | R11 | .2387 | .4263 | 3000.0 |
| 12. | R12 | .3307 | .4705 | 3000.0 |
| 13. | R13 | .2393 | .4267 | 3000.0 |
| 14. | R14 | .5830 | .4931 | 3000.0 |
| 15. | R15 | .5127 | .4999 | 3000.0 |
| 16. | R16 | .5663 | .4957 | 3000.0 |
| 17. | R17 | .5457 | .4980 | 3000.0 |
| 18. | R18 | .4390 | .4963 | 3000.0 |
| 19. | R19 | .4127 | .4924 | 3000.0 |
| 20. | R20 | .6130 | .4871 | 3000.0 |
| 21. | R21 | .3673 | .4822 | 3000.0 |
| 22. | R22 | .4770 | .4996 | 3000.0 |
| 23. | R23 | .2197 | .4141 | 3000.0 |
| 24. | R24 | .5387 | .4986 | 3000.0 |
| 25. | R25 | .5433 | .4982 | 3000.0 |
| 26. | R26 | .5117 | .4999 | 3000.0 |
| 27. | R27 | .3447 | .4753 | 3000.0 |
| 28. | R28 | .5890 | .4921 | 3000.0 |
| 29. | R29 | .2250 | .4177 | 3000.0 |
| 30. | R30 | .3687 | .4825 | 3000.0 |
| 31. | R31 | .4203 | .4937 | 3000.0 |
| 32. | R32 | .5597 | .4965 | 3000.0 |
| 33. | R33 | .4723 | .4993 | 3000.0 |
| 34. | R34 | .6300 | .4829 | 3000.0 |
| 35. | R35 | .6310 | .4826 | 3000.0 |
| 36. | R36 | .5307 | .4991 | 3000.0 |

**Continuación Cuadro No.1**

| | | Mean | Std Dev | Cases |
|---|---|---|---|---|
| 37. | R37 | .4040 | .4908 | 3000.0 |
| 38. | R38 | .5433 | .4982 | 3000.0 |
| 39. | R39 | .4683 | .4991 | 3000.0 |
| 40. | R40 | .1923 | .3942 | 3000.0 |
| 41. | R41 | .4587 | .4984 | 3000.0 |
| 42. | R42 | .0693 | .2541 | 3000.0 |
| 43. | R43 | .8127 | .3902 | 3000.0 |
| 44. | R44 | .3803 | .4855 | 3000.0 |
| 45. | R45 | .4913 | .5000 | 3000.0 |
| 46. | R46 | .2947 | .4560 | 3000.0 |
| 47. | R47 | .4320 | .4954 | 3000.0 |
| 48. | R48 | .4207 | .4937 | 3000.0 |
| 49. | R49 | .5993 | .4901 | 3000.0 |
| 50. | R50 | .3827 | .4861 | 3000.0 |

En el cuadro número 2 se presentan los índices de discriminación asociados a cada uno de los ítemes de la prueba, bajo la columna denominada "Corrected Item Total Correlation". La correlación entre el puntaje individual del ítem y el puntaje total de la prueba es una medida de poder discriminatorio que tiene el ítem para diferenciar entre estudiantes con puntajes altos y bajos. Se considera que una discriminación de 0.30 o más refleja una alta calidad técnica en cuanto a esta propiedad psicométrica. De acuerdo con este criterio hay 17 ítemes de un total de 50 que no logran alcanzar este valor. La mayoría de estos 17 ítemes, sin embargo, presentan valores bastante cercanos al 0.30. No obstante hay cuatro ítemes, 9, 11, 17 y 48, que presentan discriminaciones muy bajas y que por tanto deberían de ser revisados, de acuerdo con los resultados bajo la teoría clásica.

Por otra parte, el coeficiente Alfa de Cronbach es una medida de la confiabilidad de la prueba total en términos de su consistencia interna. Su valor está entre 0 y 1. Valores cercanos a 1 indican una prueba de alta confiabilidad. Nunnally, uno de los autores clásicos en psicometría, establece que si la prueba se va a usar para investigación se puede trabajar con un coeficiente alfa mínimo de 0.7. Sin embargo, si la prueba se usa para toma de cisiones se debe ser más exigente trabajar con un nivel mínimo de 0.9 para Alfa. El valor que toma esta medida en este caso particular es 0.8742.

Según los planteamientos de la teoría clásica, para aumentar el valor de Alfa se podrían construir más ítemes con discriminaciones de al menos 0.30, o bien su valor aumentaría si se

eliminaran de la prueba los cuatro ítemes mencionados arriba que presentan niveles muy bajos de discriminación (números 9, 11, 17 y 48).

## Cuadro No. 2:
## Indices de Discriminación de los Itemes y Coeficiente Alfa

Item-total Statistics

| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Alpha if Item Deleted |
|---|---|---|---|---|
| R01 | 20.8400 | 75.6129 | .2371 | .8734 |
| R02 | 20.9077 | 75.2816 | .2905 | .8725 |
| R03 | 21.1143 | 75.4971 | .3996 | .8714 |
| R04 | 21.0330 | 76.0833 | .2317 | .8733 |
| R05 | 20.9850 | 75.8801 | .2386 | .8733 |
| R06 | 20.7750 | 75.2721 | .2714 | .8729 |
| R07 | 20.7353 | 74.1400 | .4039 | .8707 |
| R08 | 20.9050 | 74.5522 | .3810 | .8711 |
| R09 | 20.8040 | 77.0873 | .0616 | .8763 |
| R10 | 20.8783 | 74.3910 | .3929 | .8709 |
| R11 | 20.9967 | 76.7722 | .1224 | .8749 |
| R12 | 20.9047 | 75.2860 | .2892 | .8726 |
| R13 | 20.9960 | 75.0210 | .3606 | .8715 |
| R14 | 20.6523 | 73.9835 | .4293 | .8702 |
| R15 | 20.7227 | 73.9184 | .4305 | .8702 |
| R16 | 20.6690 | 74.9097 | .3162 | .8721 |
| R17 | 20.6897 | 76.2128 | .1619 | .8747 |
| R18 | 20.7963 | 75.6241 | .2315 | .8736 |
| R19 | 20.8227 | 75.2943 | .2729 | .8729 |
| R20 | 20.6223 | 74.9847 | .3137 | .8722 |
| R21 | 20.8680 | 75.7698 | .2223 | .8737 |
| R22 | 20.7583 | 74.3127 | .3839 | .8710 |
| R23 | 21.0157 | 75.0871 | .3637 | .8715 |
| R24 | 20.6967 | 74.5895 | .3519 | .8715 |
| R25 | 20.6920 | 73.8017 | .4461 | .8699 |
| R26 | 20.7237 | 74.7035 | .3373 | .8718 |
| R27 | 20.8907 | 75.4859 | .2611 | .8730 |

**Continuación Cuadro No.2**

| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Alpha if Item Deleted |
|---|---|---|---|---|
| R28 | 20.6463 | 75.1876 | .2858 | .8726 |
| R29 | 21.0103 | 73.9922 | .5152 | .8693 |
| R30 | 20.8667 | 73.3533 | .5181 | .8688 |
| R31 | 20.8150 | 74.9304 | .3152 | .8722 |
| R32 | 20.6757 | 73.8084 | .4470 | .8699 |
| R33 | 20.7630 | 74.8825 | .3167 | .8721 |
| R34 | 20.6053 | 74.3597 | .3933 | .8709 |
| R35 | 20.6043 | 74.8040 | .3392 | .8718 |
| R36 | 20.7047 | 74.0575 | .4146 | .8705 |
| R37 | 20.8313 | 74.6104 | .3559 | .8715 |
| R38 | 20.6920 | 74.5040 | .3623 | .8714 |
| R39 | 20.7670 | 74.9964 | .3035 | .8724 |
| R40 | 21.0430 | 75.5303 | .3186 | .8721 |
| R41 | 20.7767 | 74.3289 | .3830 | .8710 |
| R42 | 21.1660 | 76.9127 | .2000 | .8736 |
| R43 | 20.4227 | 75.6059 | .3110 | .8723 |
| R44 | 20.8550 | 73.8839 | .4491 | .8699 |
| R45 | 20.7440 | 74.7414 | .3328 | .8719 |
| R46 | 20.9407 | 75.9445 | .2160 | .8737 |
| R47 | 20.8033 | 73.4711 | .4889 | .8692 |
| R48 | 20.8147 | 76.4405 | .1372 | .8751 |
| R49 | 20.6360 | 74.0102 | .4291 | .8703 |
| R50 | 20.8527 | 74.9379 | .3201 | .8721 |

Reliability Coefficients

N of Cases = 3000.0 N of Items = 50

Alpha = .8742

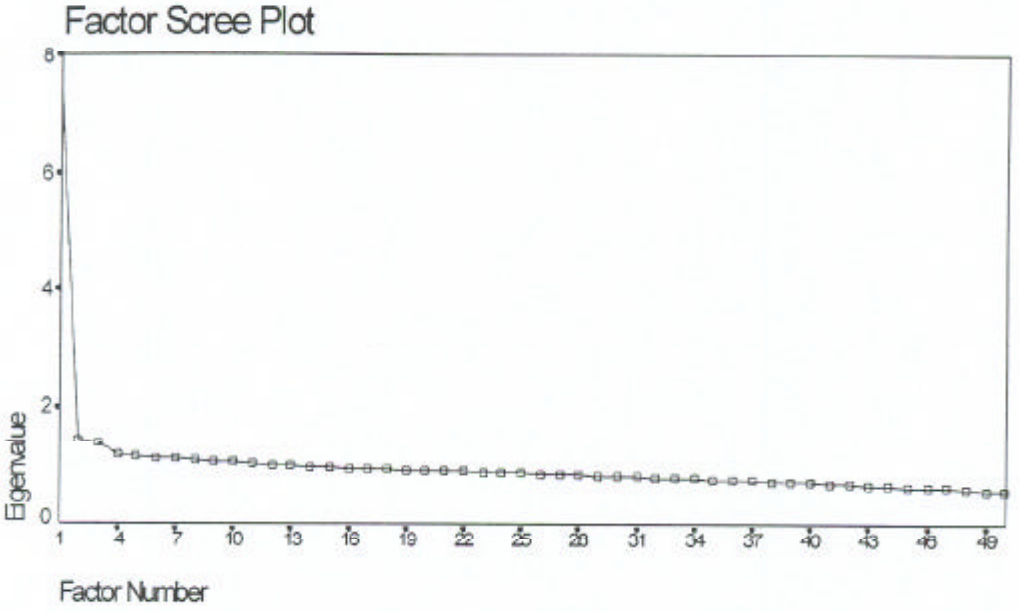
# ANÁLISIS DE FACTORES

La técnica estadística multivariable denominada análisis de factores se usa en psicometría para encontrar evidencia de validez de constructo en un instrumento. Mediante el análisis de factores se puede establecer hasta qué grado los ítemes de la prueba se agrupan de acuerdo con las dimensiones subyacentes en el instrumento. En el presente caso se aplicó el análisis de factores de manera exploratoria.

Para definir el número de factores se usa el gráfico denominado en inglés "Scree Plot", que compara el número de factores contra el valor característico asociado a cada uno. El valor característico asociado a un factor es una medida de su importancia relativa. Entre mayor sea el valor característico mayor será el porcentaje de la variancia de las variables (ítemes) explicado por ese factor. Típicamente el factor 1 es el que tiene el mayor valor característico, le sigue en importancia el factor 2 y así sucesivamente. Para identificar el número de factores a partir del gráfico "Scree Plot" se debe establecer a partir de qué número de factores la solución se estabiliza. En el gráfico esta estabilización se visualiza en el punto donde la curva dibuja una especie de "codo" o esquina. Los factores que se definen a partir de ese "codo" no tendrán mayor importancia relativa en términos de variancia explicada y por tanto no vale la pena tomarlos en cuenta para la solución final.

En el gráfico No.1 se presenta el "Scree Plot" correspondiente a la prueba de noveno año analizada. A partir de él se concluye claramente que se define un único factor como relevante o significativo. En el cuadro 3 se presentan las llamadas cargas factoriales, las cuales dan cuenta del nivel de representación que tiene ese factor en cada uno de los ítemes de la prueba. Se considera que el factor está representado adecuadamente por el ítem en cuestión cuando su carga factorial es mayor o igual a 0.30. Analizando estos resultados se concluye que la prueba cumple con el requisito psicométrico de ser fundamentalmente unidimensional. Solamente los ítemes 9, 11, 17 y 48 parecen no estar representando apropiadamente al factor común y por tanto deberían revisarse.

# Gráfico No.1


Factor Scree Plot
Eigenvalue
0
2
4
6
8
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
Factor Number

# Cuadro No. 3: Cargas Factoriales de los Itemes

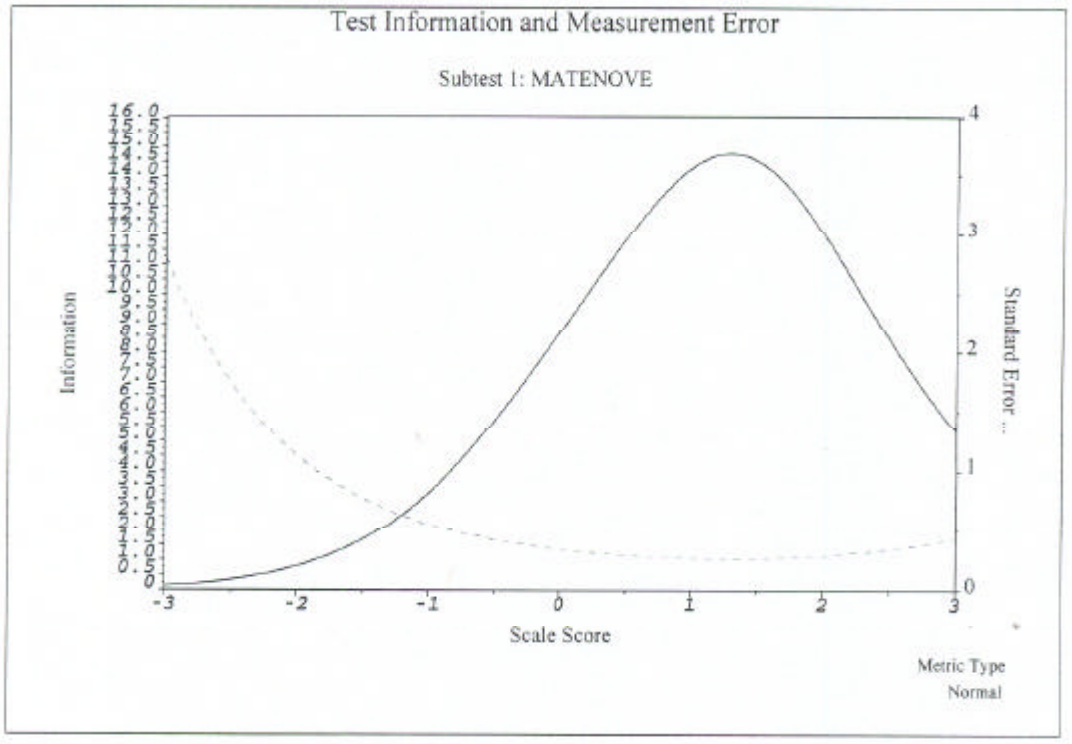
| | Factor 1 |
|---|---|
| R01 | .27652 |
| R02 | .33257 |
| R03 | .44942 |
| R04 | .26308 |
| R05 | .27362 |
| R06 | .30821 |
| R07 | .45851 |
| R08 | .42985 |
| R09 | .07503 |
| R10 | .44498 |
| R11 | .14050 |
| R12 | .32735 |
| R13 | .40790 |
| R14 | .48982 |
| R15 | .48904 |
| R16 | .36409 |
| R17 | .18632 |
| R18 | .26720 |
| R19 | .31512 |
| R20 | .35928 |
| R21 | .25502 |
| R22 | .43474 |
| R23 | .40843 |
| R24 | .40094 |

| | Factor 1 |
|---|---|
| R25 | .50399 |
| R26 | .38465 |
| R27 | .30207 |
| R28 | .32521 |
| R29 | .57001 |
| R30 | .57642 |
| R31 | .36218 |
| R32 | .50877 |
| R33 | .36368 |
| R34 | .44992 |
| R35 | .38627 |
| R36 | .47304 |
| R37 | .41008 |
| R38 | .41484 |
| R39 | .35007 |
| R40 | .36074 |
| R41 | .43185 |
| R42 | .22452 |
| R43 | .35290 |
| R44 | .50583 |
| R45 | .37744 |
| R46 | .24708 |
| R47 | .54627 |
| R48 | .15821 |
| R49 | .48692 |
| R50 | .36225 |

## ANÁLISIS BAJO LA TEORÍA DE RESPUESTA A LOS ITEMES

El gráfico No. 2 muestra la Función de Información y el Errror de Medición para la prueba total. La Función de Información presenta la forma usual de curva normal. La información se maximiza en niveles relativamente altos de habilidad, alrededor de 1.5 (recordar que la escala varía entre –3 y +3). Se nota que en los niveles intermedios de habilidad (alrededor de 0) la información que brinda la prueba no es óptima.

## Gráfico No.2




Los gráficos 3 al 51 presentan las curvas características del item y las funciones de información para cada uno de los itemes analizados de la prueba. Aquí se debe recordar que el item 26 de la prueba no fue incluido, ya que no aparecía su código para la respuesta correcta en la base de datos suministrada por el MEP. De esta forma a partir del item 26 analizado en este estudio la

numeración correspondiente en la prueba equivale al número del item analizado en el estudio más 1. Por ejemplo el item 26 de este estudio corresponde al 27 de la prueba, y así sucesivamente, hasta el 50 que corresponde al 51 de la prueba.

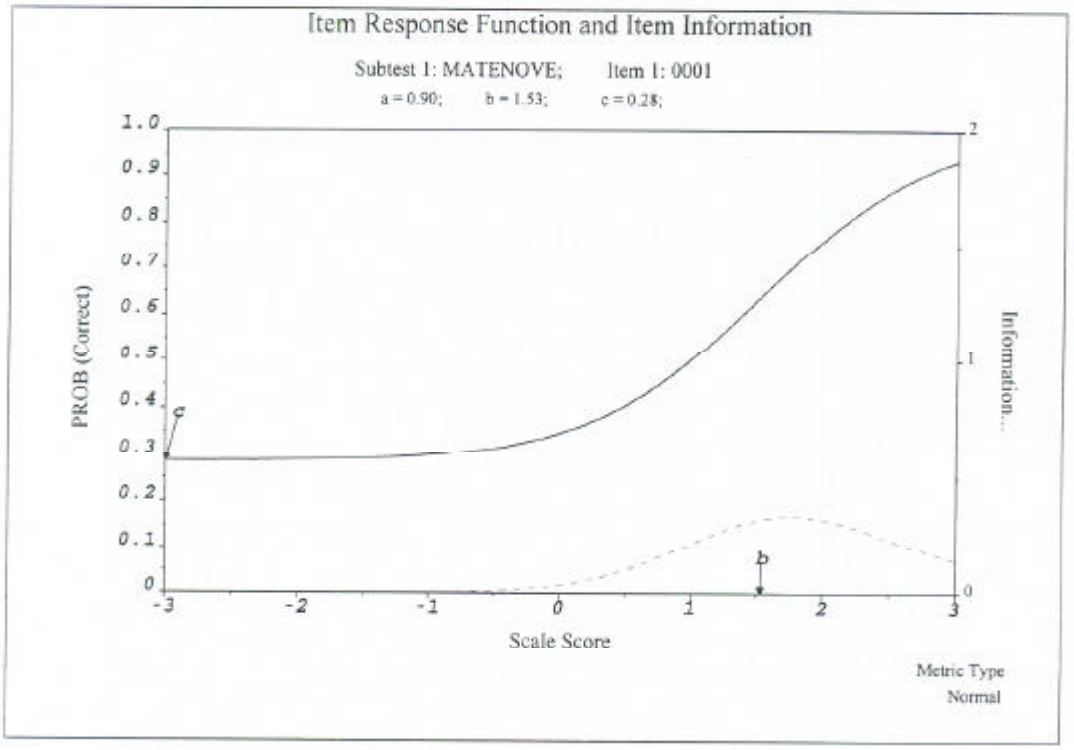
El cuadro No. 4 presenta las características más relevantes de los ítemes, después de analizar su comportamiento individual de acuerdo con los parámetros que exhiben según el análisis de la Teoría de Respuesta a los Ítemes.

## Cuadro No.4:
## Aspectos más relevantes de los ítemes de la prueba según el análisis con T.R.I.

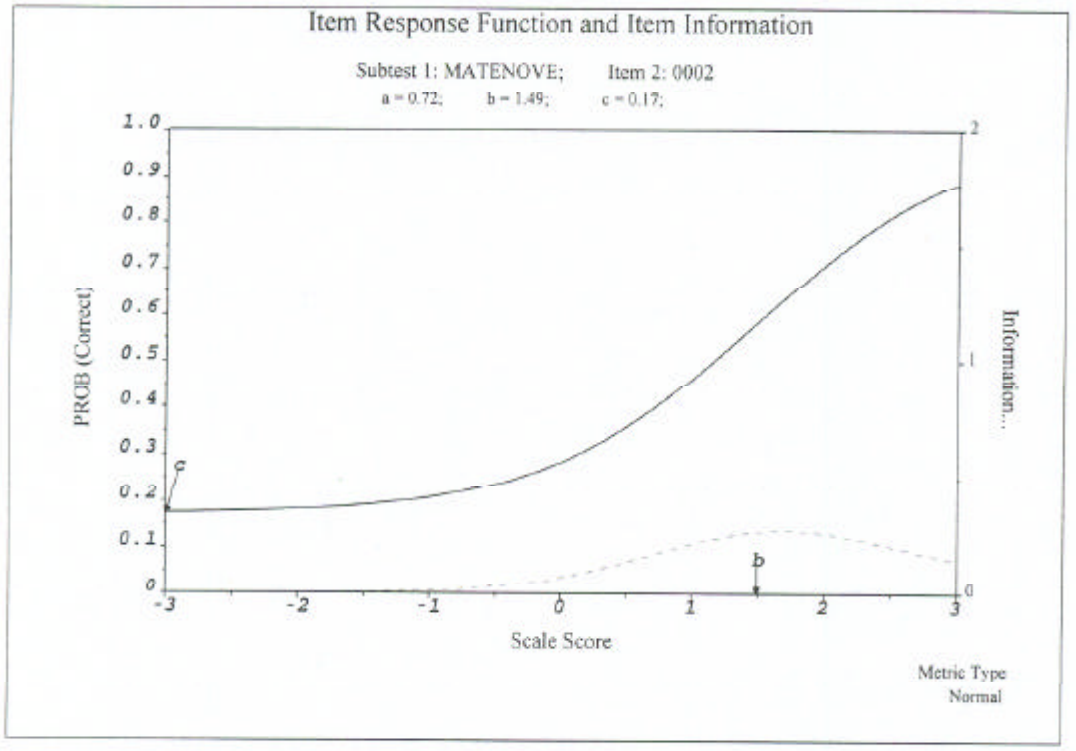
| Ítemes con alta o aceptable calidad técnica y poder discriminatorio en niveles altos de habilidad | Ítemes con alta o aceptable calidad técnica y poder discriminatorio en niveles intermedios de habilidad | Ítemes con alta o aceptable calidad técnica y poder discriminatorio en niveles bajos de habilidad | Ítemes de la prueba que no presentan aceptable calidad técnica |
|---|---|---|---|
| 1 | 7 | 43 | 9 |
| 2 | 10 | | 17 |
| 3 | 14 | | 18 |
| 4 | 15 | | 28 |
| 5 | 16 | | |
| 6 | 20 | | |
| 8 | 22 | | |
| 11 | 24 | | |
| 12 | 25 | | |
| 13 | 26 | | |
| - | 30 | | |
| 19 | 31 | | |
| 21 | 32 | | |
| 23 | 34 | | |
| 27 | 35 | | |
| 29 | 36 | | |
| 33 | 37 | | |
| 40 | 38 | | |
| 42 | 39 | | |
| 46 | 41 | | |
| 48 | 44 | | |
| 50 | 45 | | |
| | 47 | | |
| | 49 | | |

Es importante destacar que, aunque en los niveles intermedios de habilidad existen 24 ítemes con alta o aceptable calidad técnica y en los niveles altos de habilidad el número de ítemes en esa categoría es 21, la información que brindan los ítemes con poder discriminatorio en niveles altos es, en general, mayor que la que brindan aquellos que discriminan en niveles intermedios, es decir los primeros tienen usualmente mejor calidad técnica que los segundos. Por esta razón es que la prueba como un todo ofrece información máxima en niveles relativamente altos de habilidad.

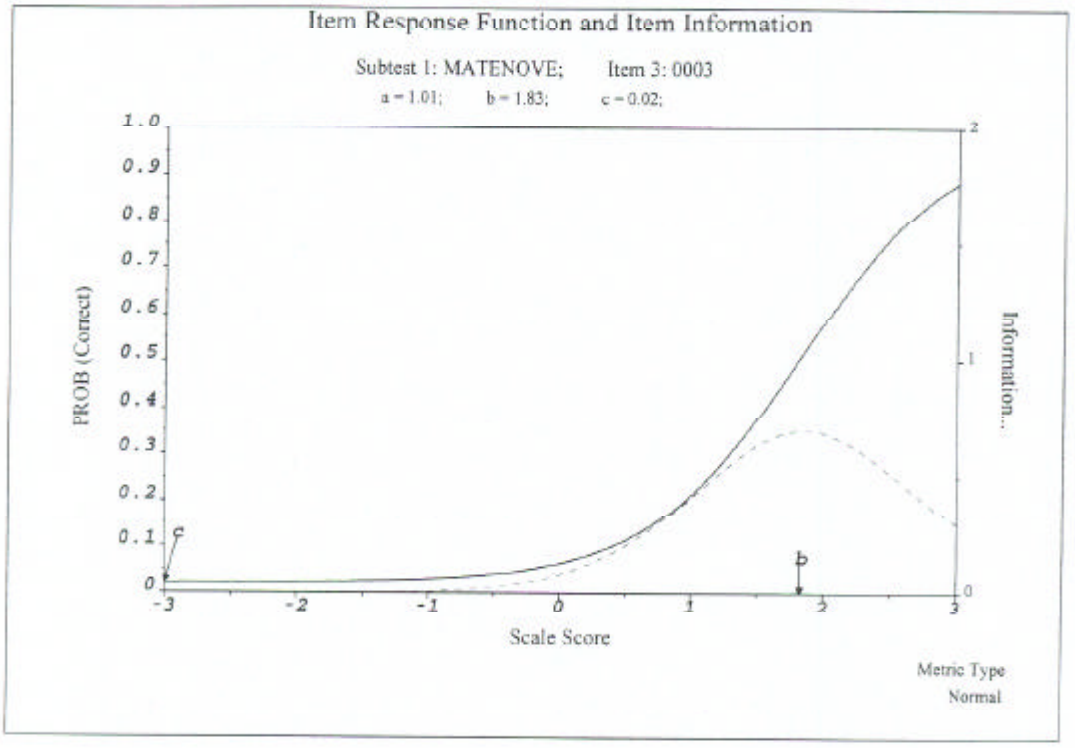
# Gráfico No.3


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 1: 0001
a = 0.90; b = 1.53; c = 0.28;
PROB (Correct)
Information
Scale Score
Metric Type
Normal
c
b

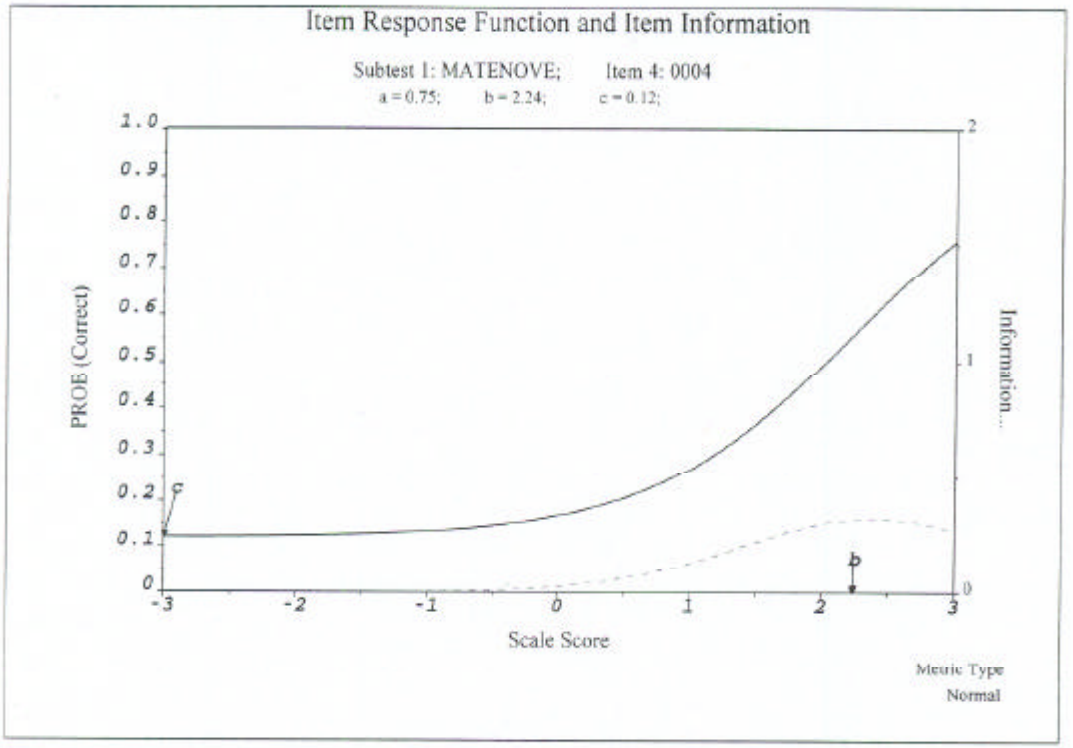
# Gráfico No.4


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 2: 0002
a = 0.72; b = 1.49; c = 0.17;
PROB (Correct)
Information...
Scale Score
Metric Type
Normal
c
b

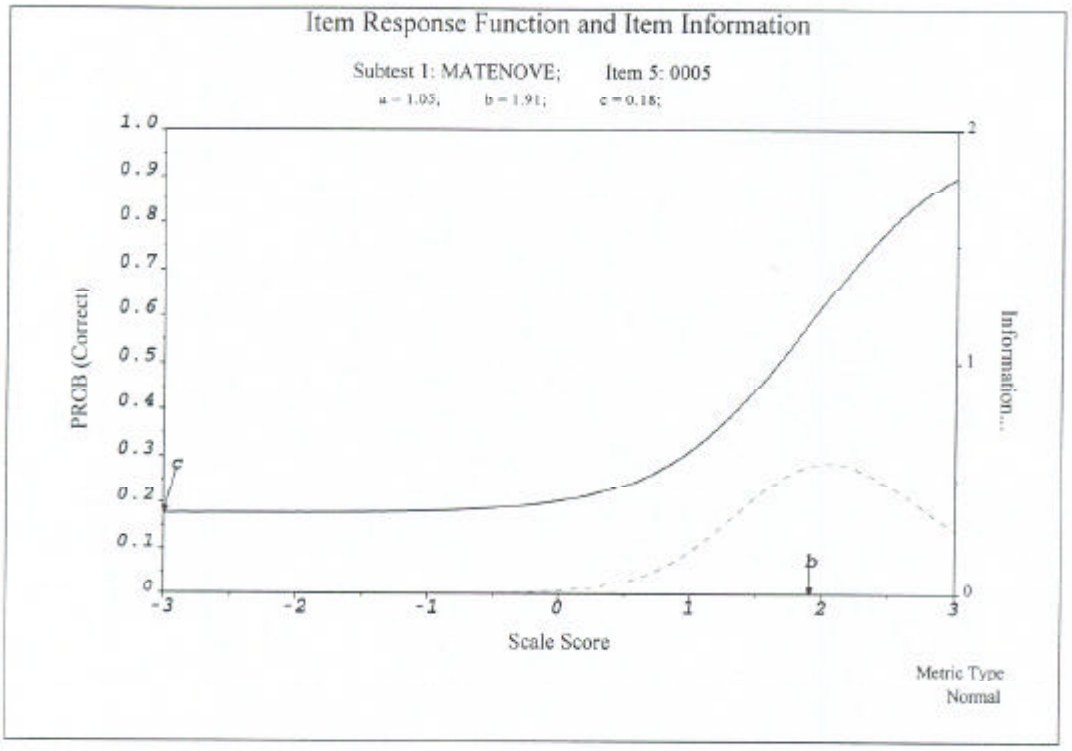
# Gráfico No.5


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 3: 0003
a = 1.01; b = 1.83; c = 0.02;
PROB (Correct)
Information...
Scale Score
Metric Type
Normal
c
b

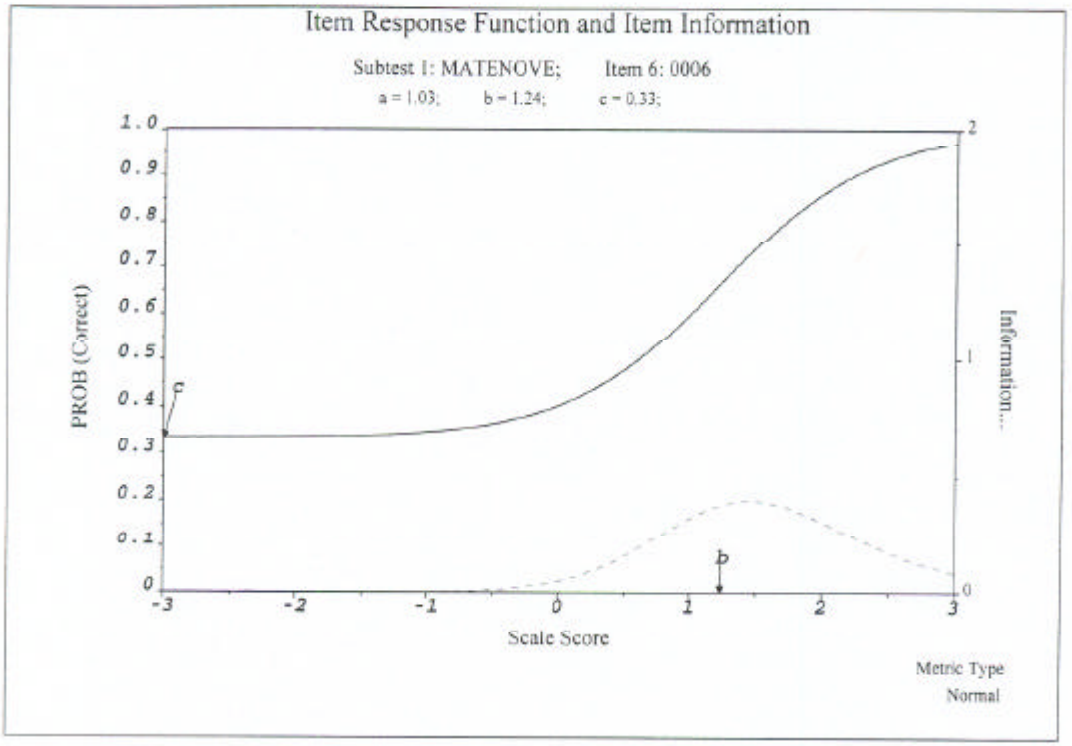
# Gráfico


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 4: 0004
a = 0.75; b = 2.24; c = 0.12;
PROB (Correct)
Information
Scale Score
Metric Type
Normal
c
b

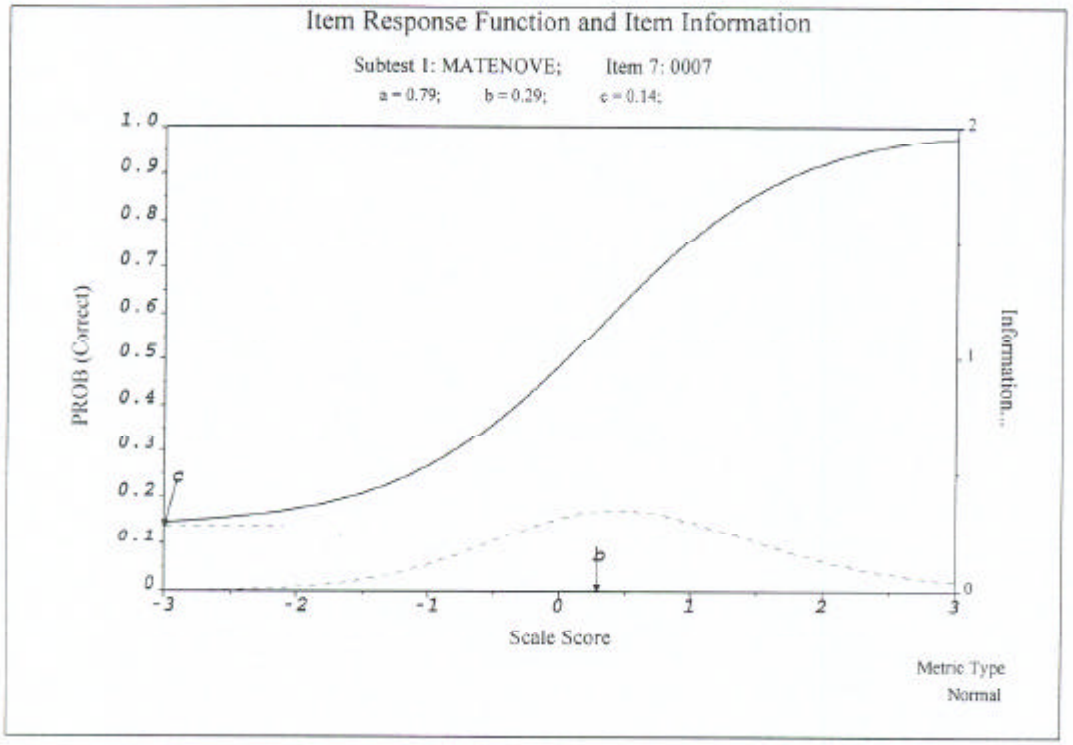
# Gráfico No.6


Item Response Function and Item Information
Subtest I: MATENOVE; Item 5: 0005
a = 1.03; b = 1.91; c = 0.18;
PRCB (Correct)
Information...
Scale Score
Metric Type
Normal
c
b

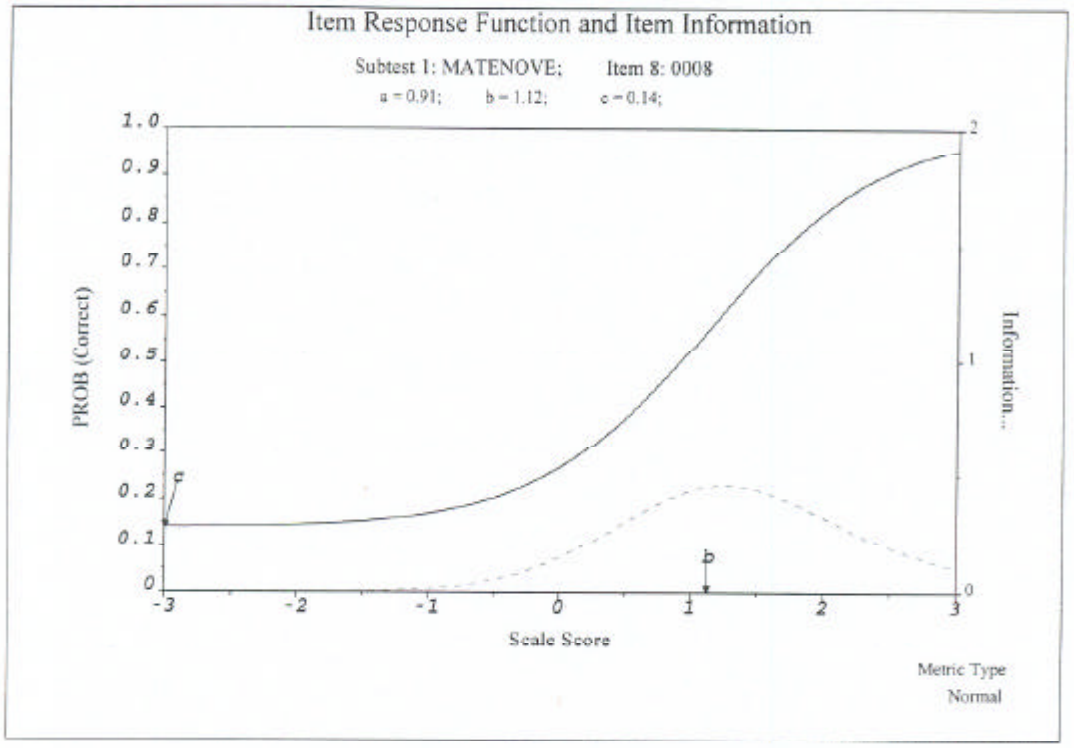
# Gráfico No.7



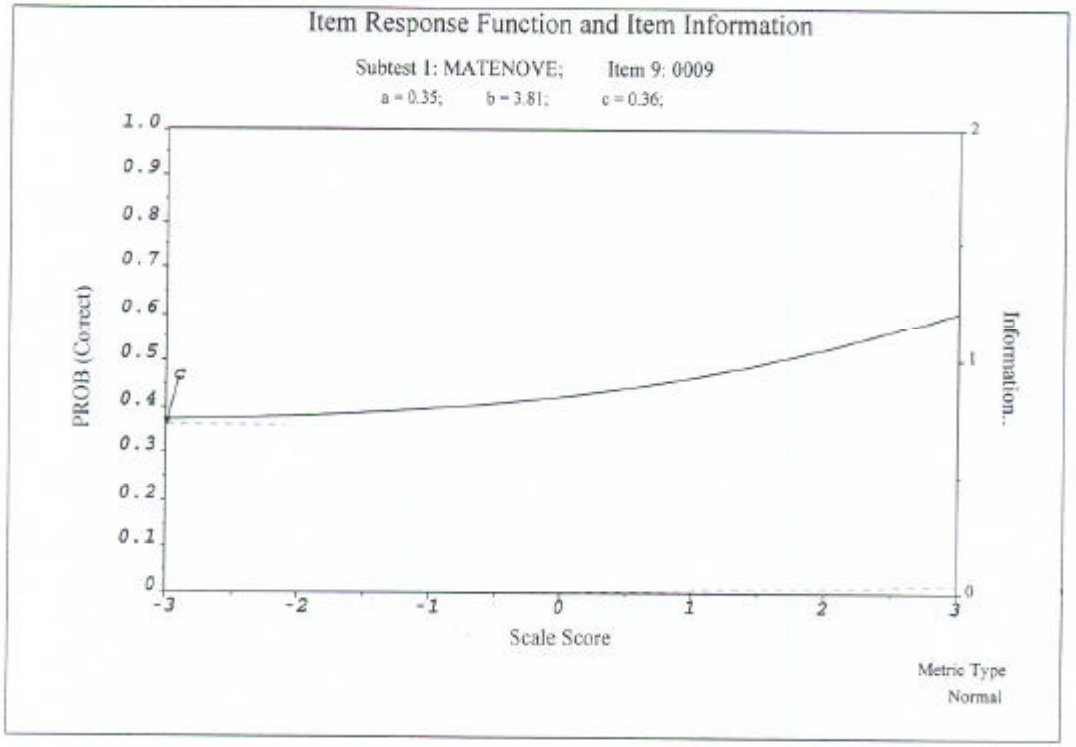
# Gráfico No.8


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 7: 0007
a = 0.79; b = 0.29; c = 0.14;
PROB (Correct)
Information
Scale Score
Metric Type
Normal

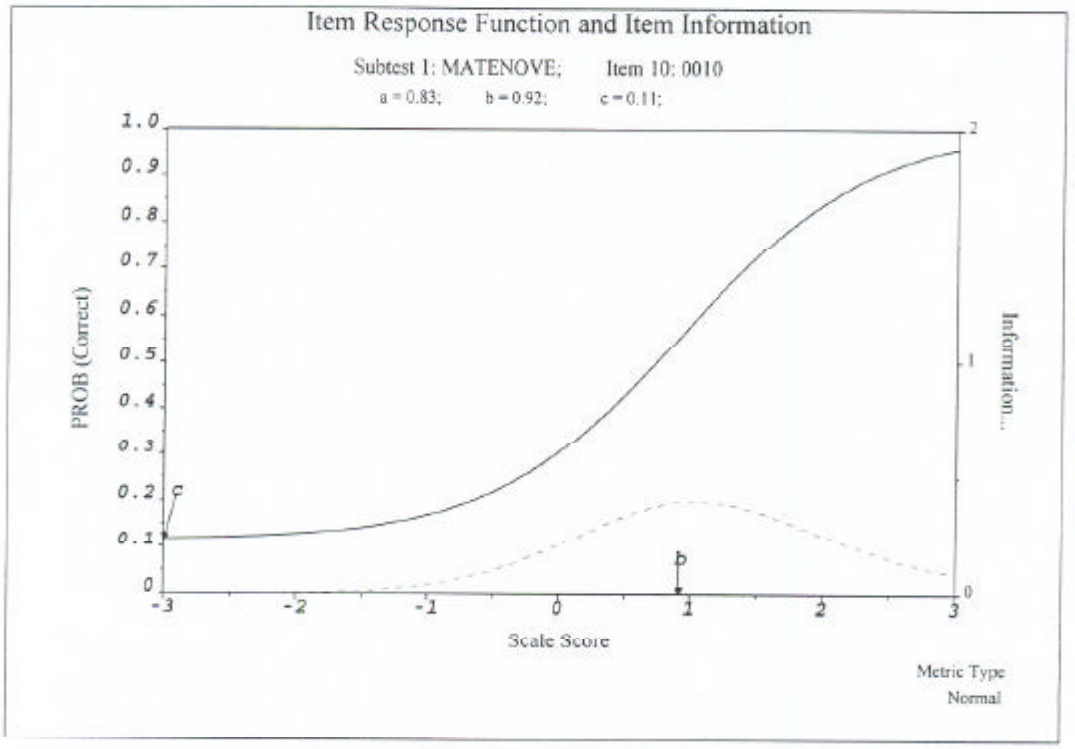
# Gráfico No.9


Item Response Function and Item Information
Subtest 1: MATENOVE; Item 8: 0008
a = 0.91; b = 1.12; c = 0.14;
PROB (Correct)
Information...
Scale Score
Metric Type
Normal
c
b

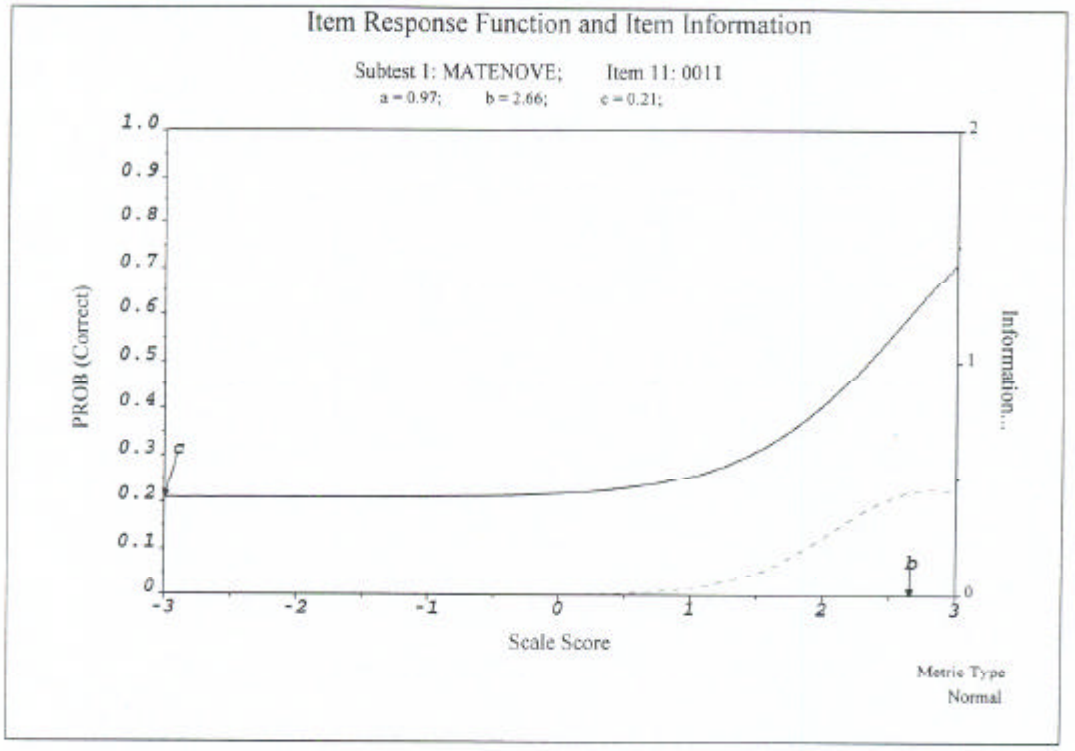
# Gráfico No.10



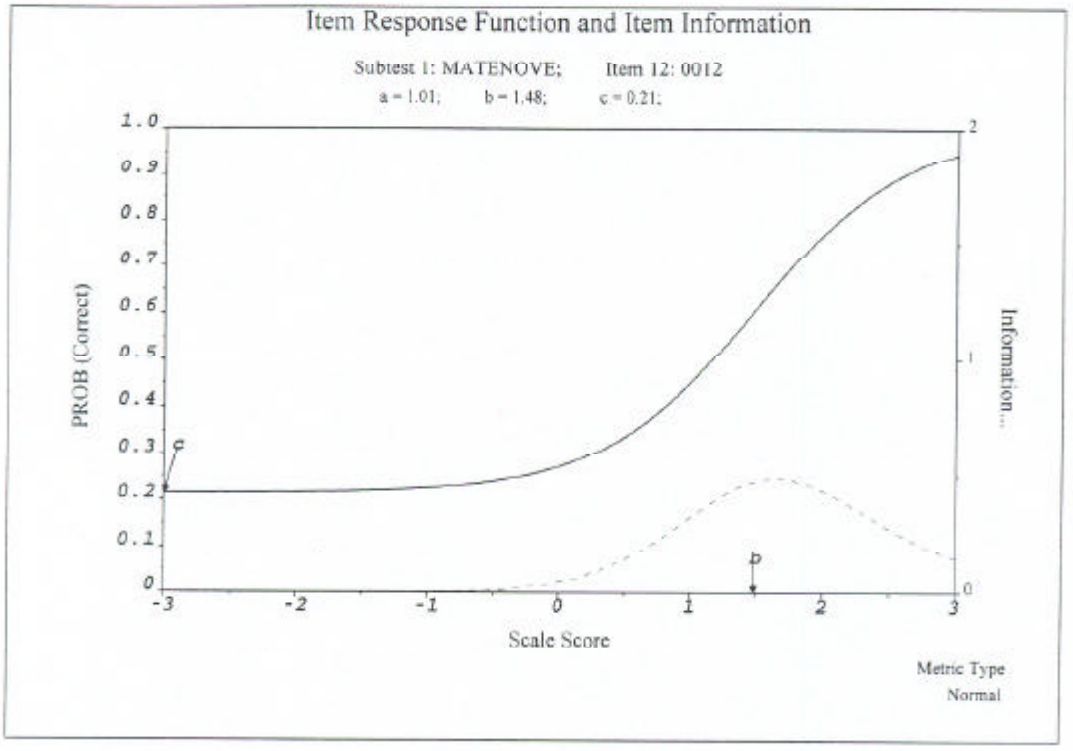
# Gráfico No.11



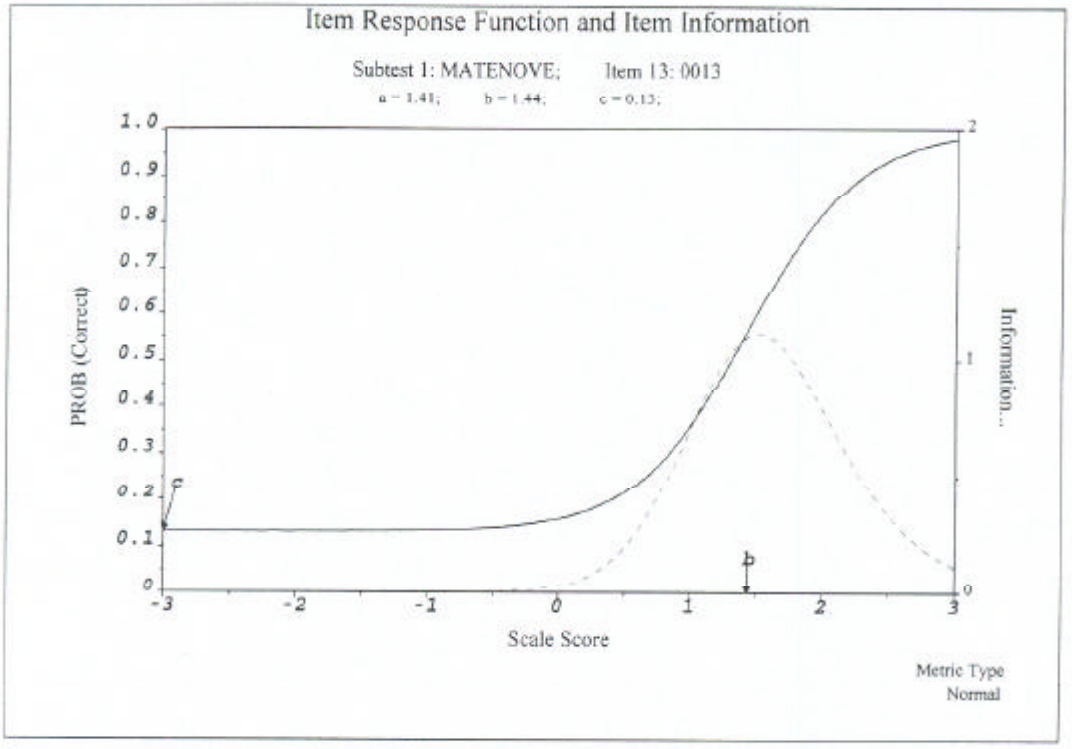
# Gráfico No.12



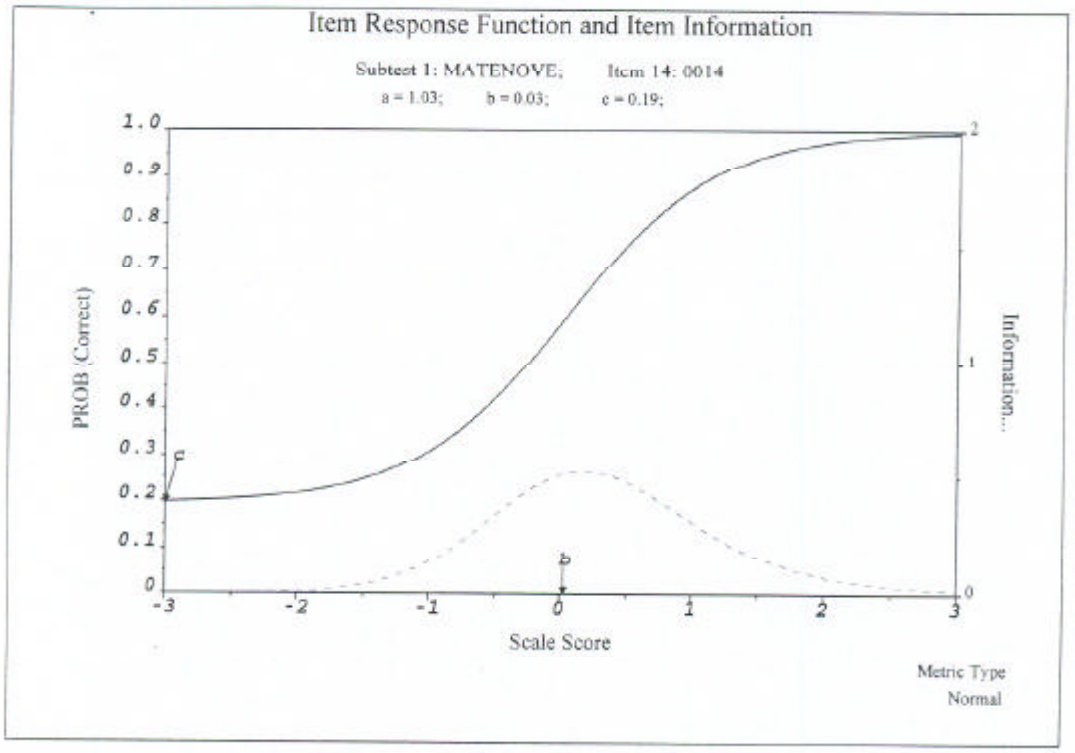
# Gráfico No.13



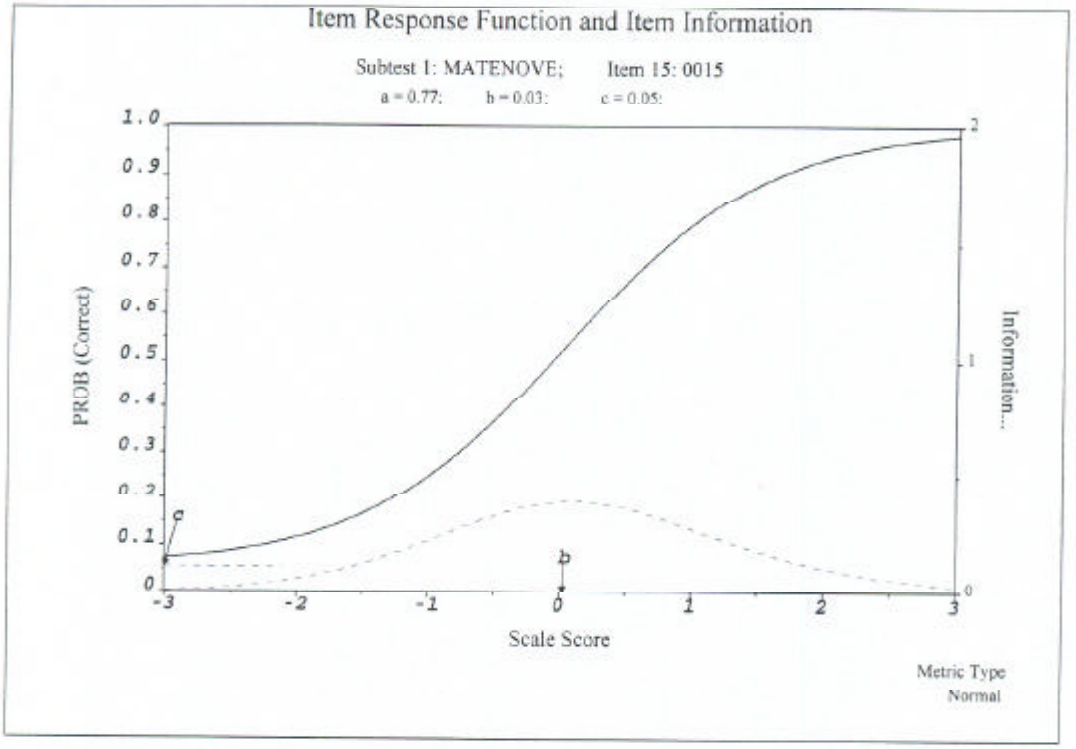
# Gráfico No.14



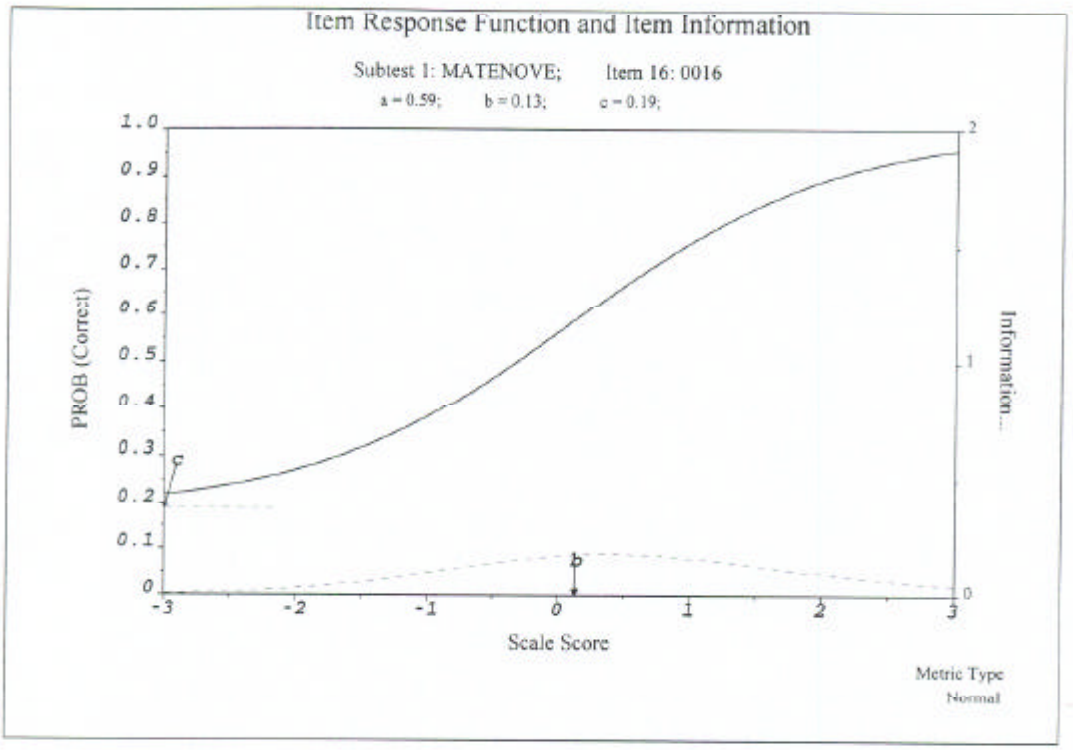
# Gráfico No.15



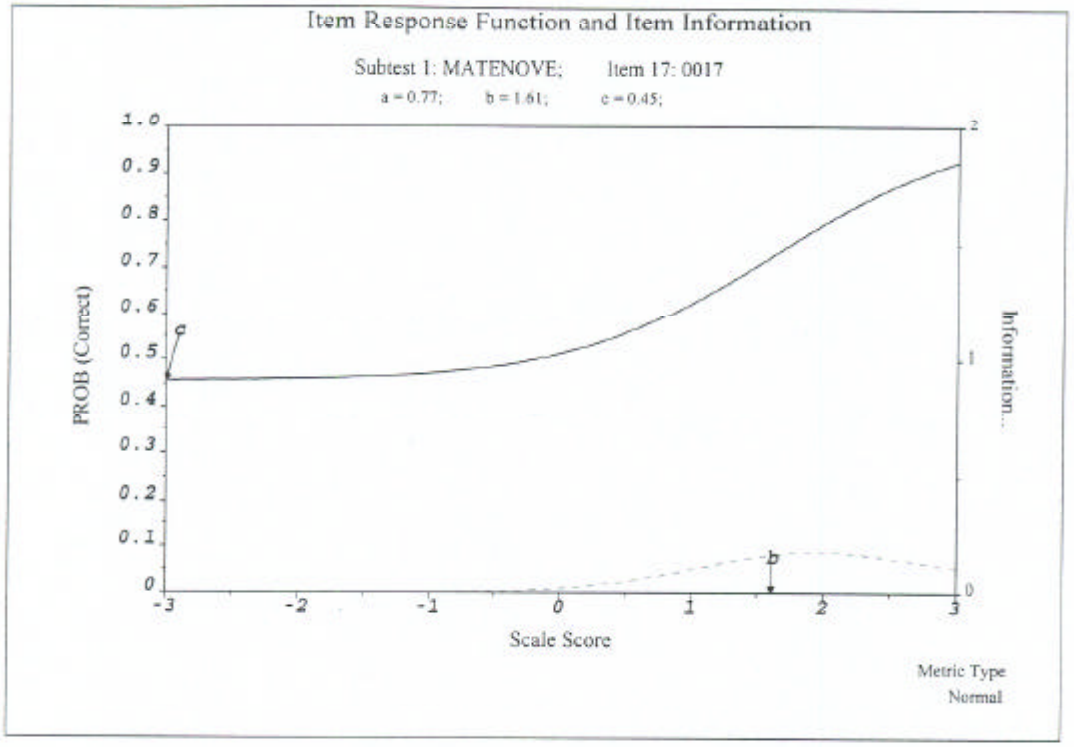
# Gráfico No.16



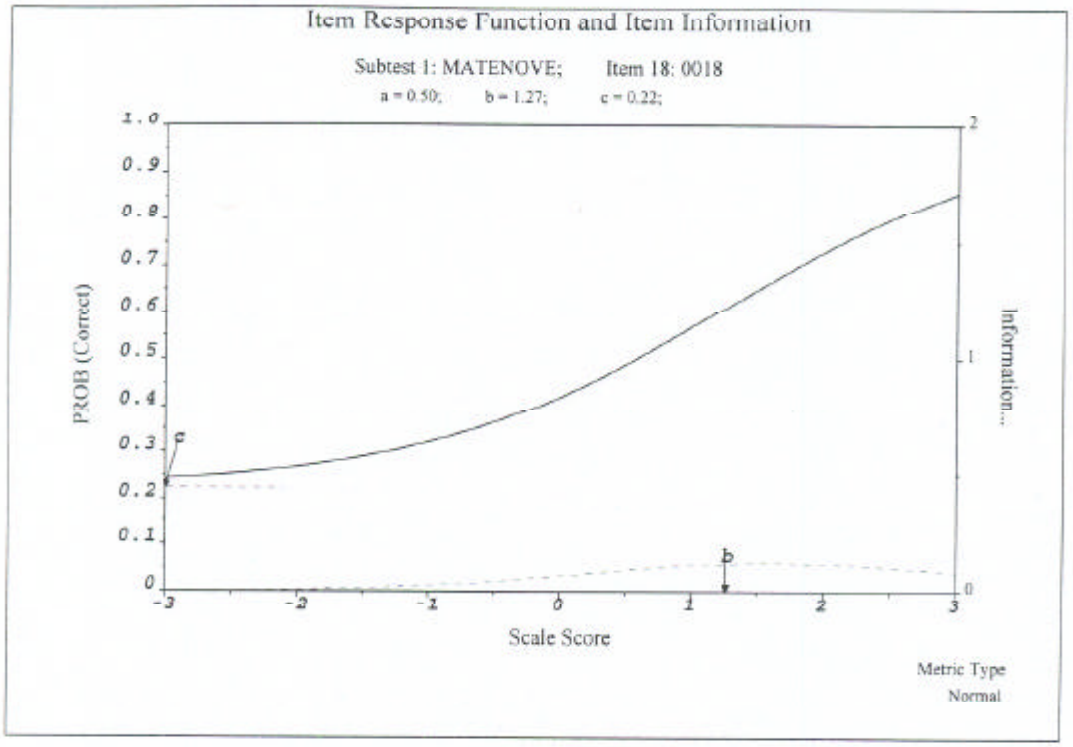
# Gráfico No.17



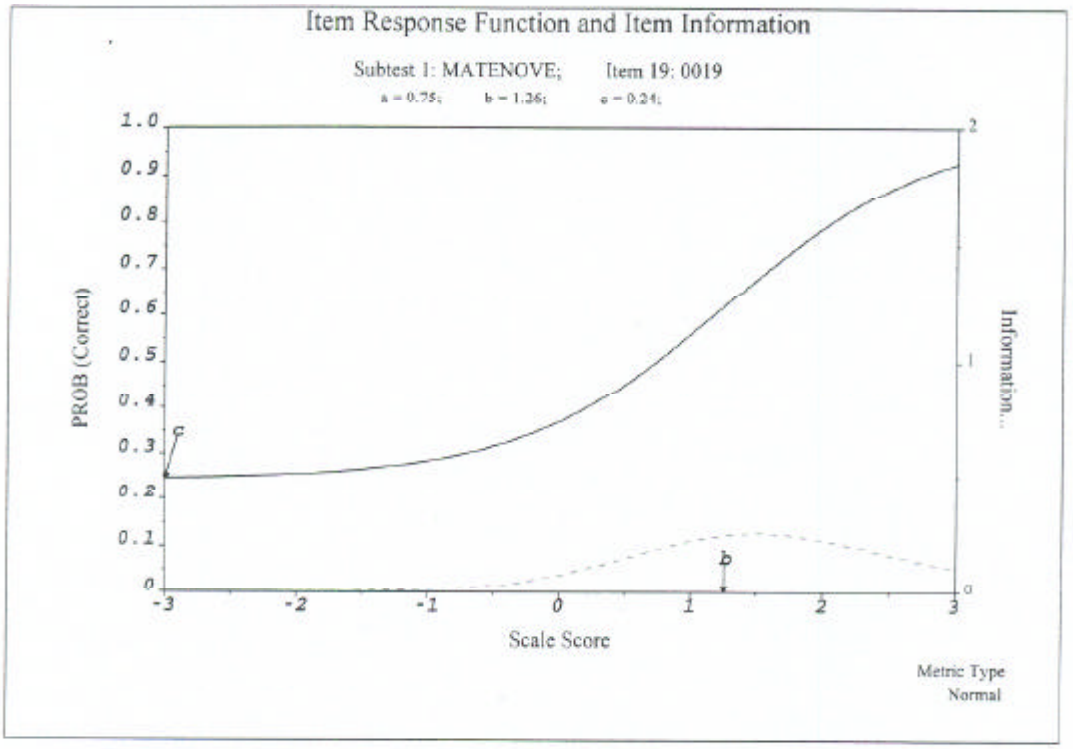
# Gráfico No.18



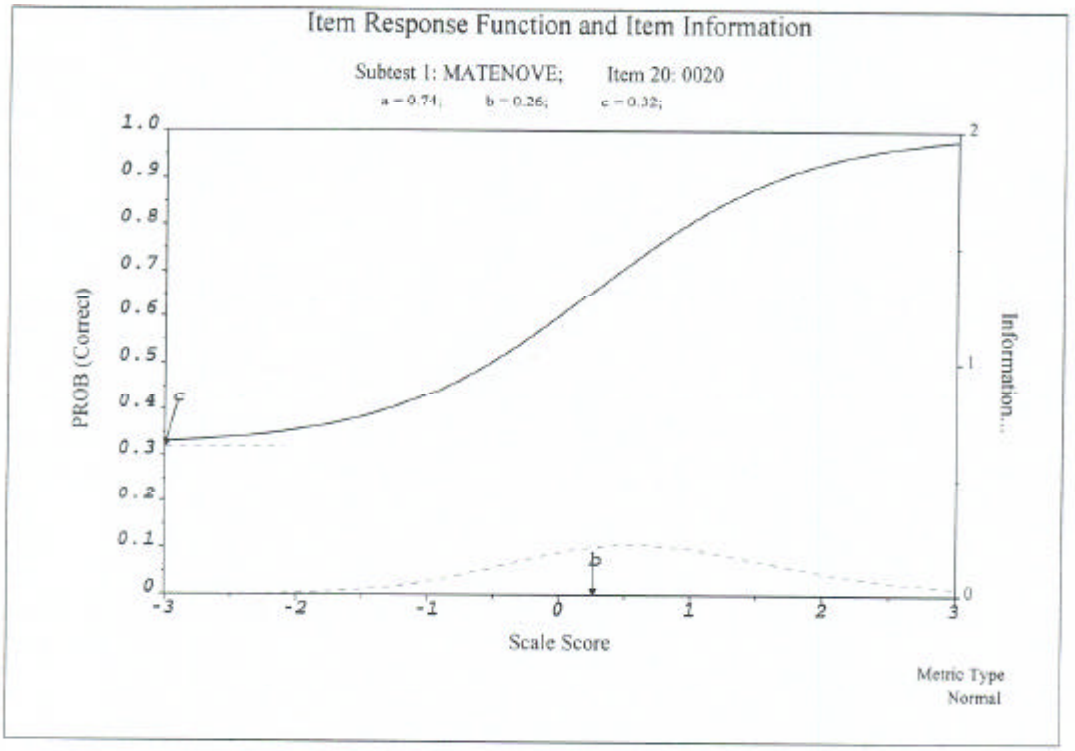
# Gráfico No.19



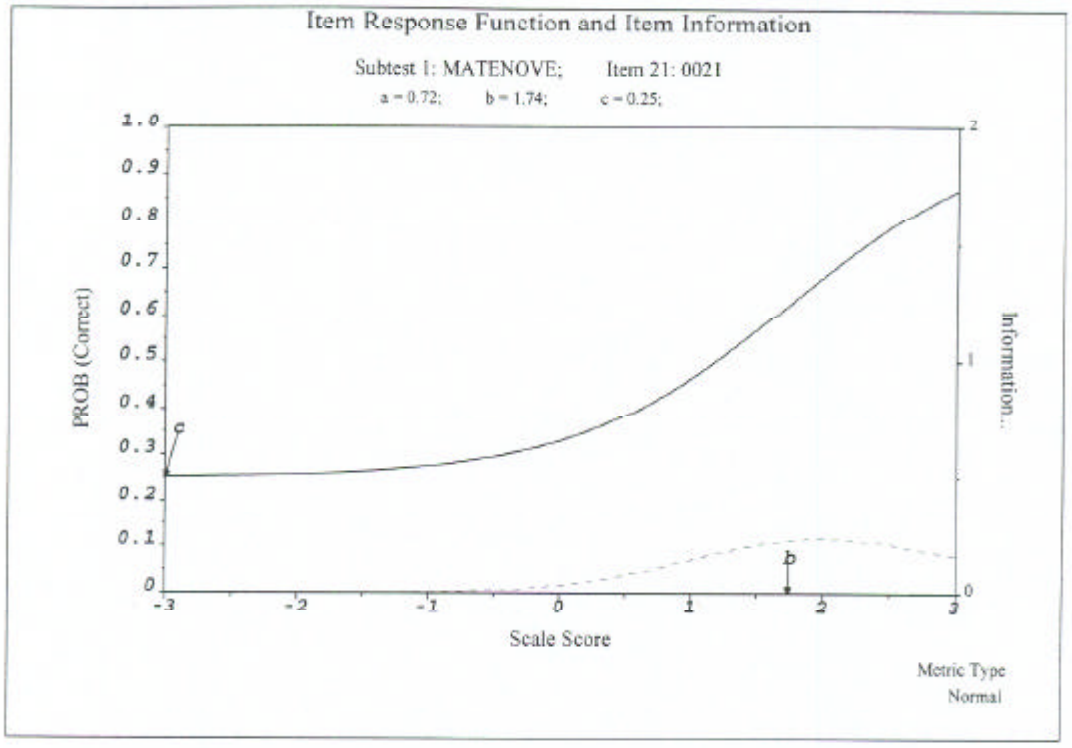
# Gráfico No.20



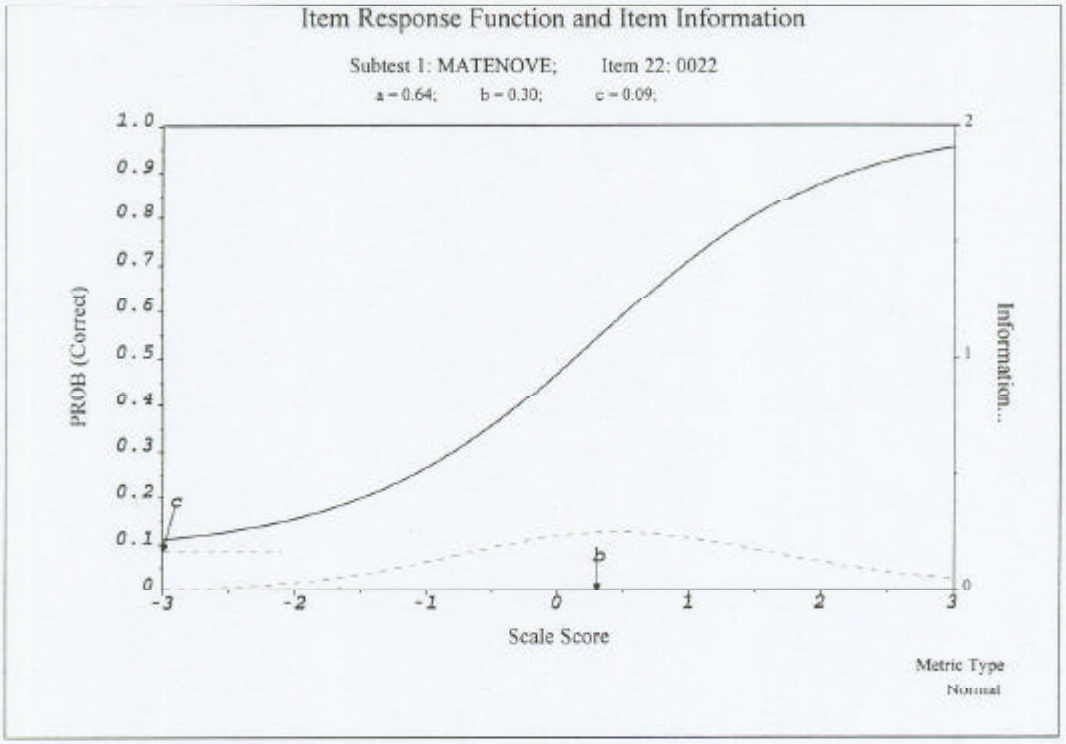
# Gráfico No.21



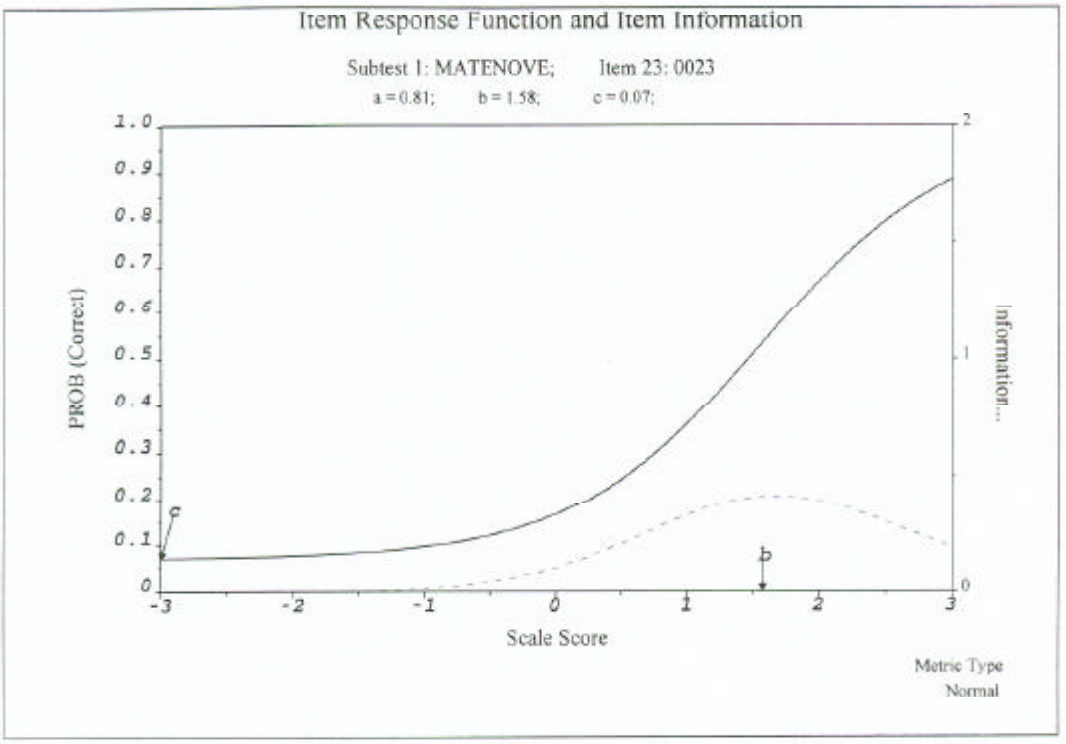
# Gráfico No.22



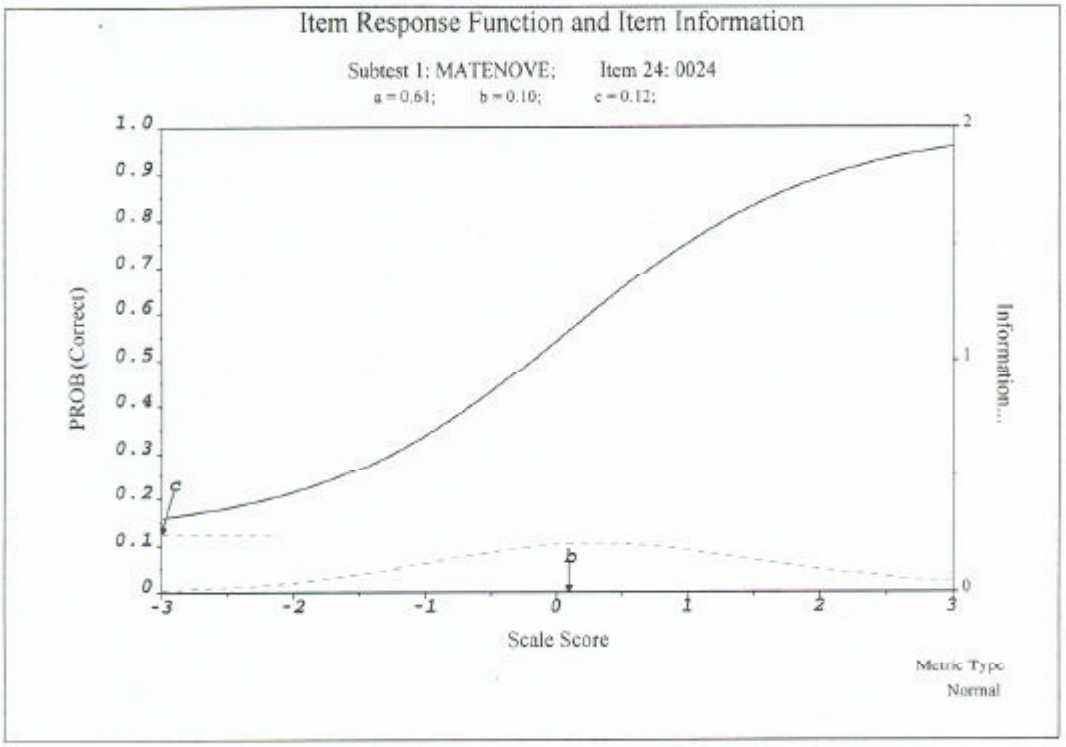
# Gráfico No.23



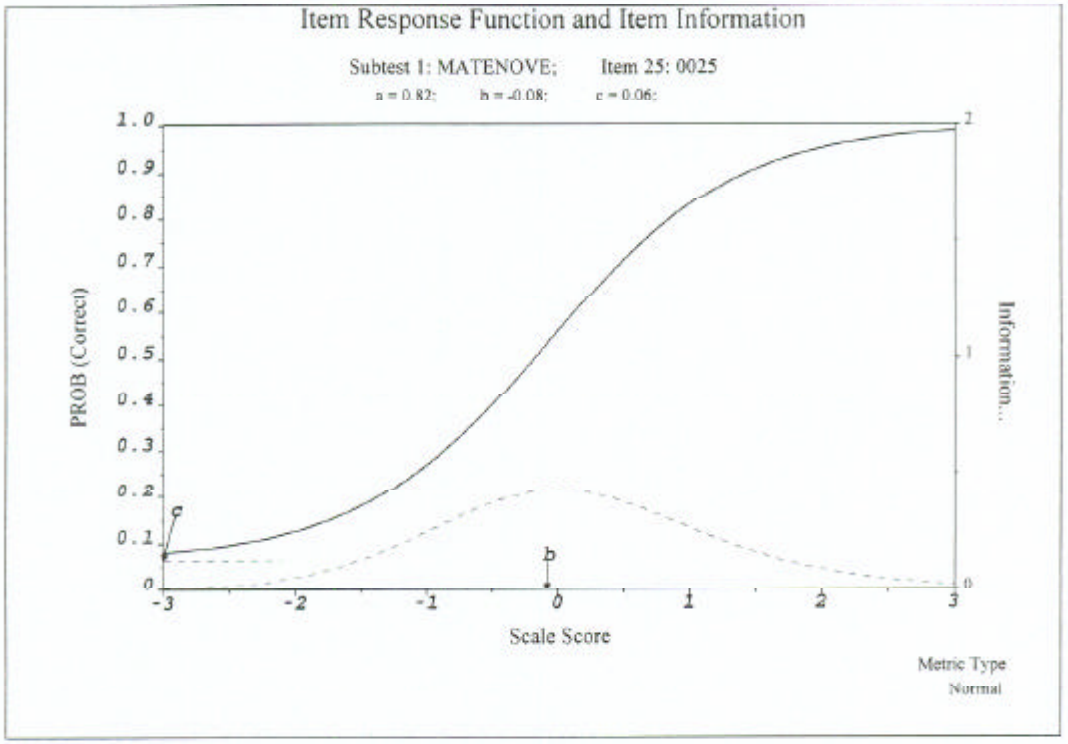
# Gráfico No.24



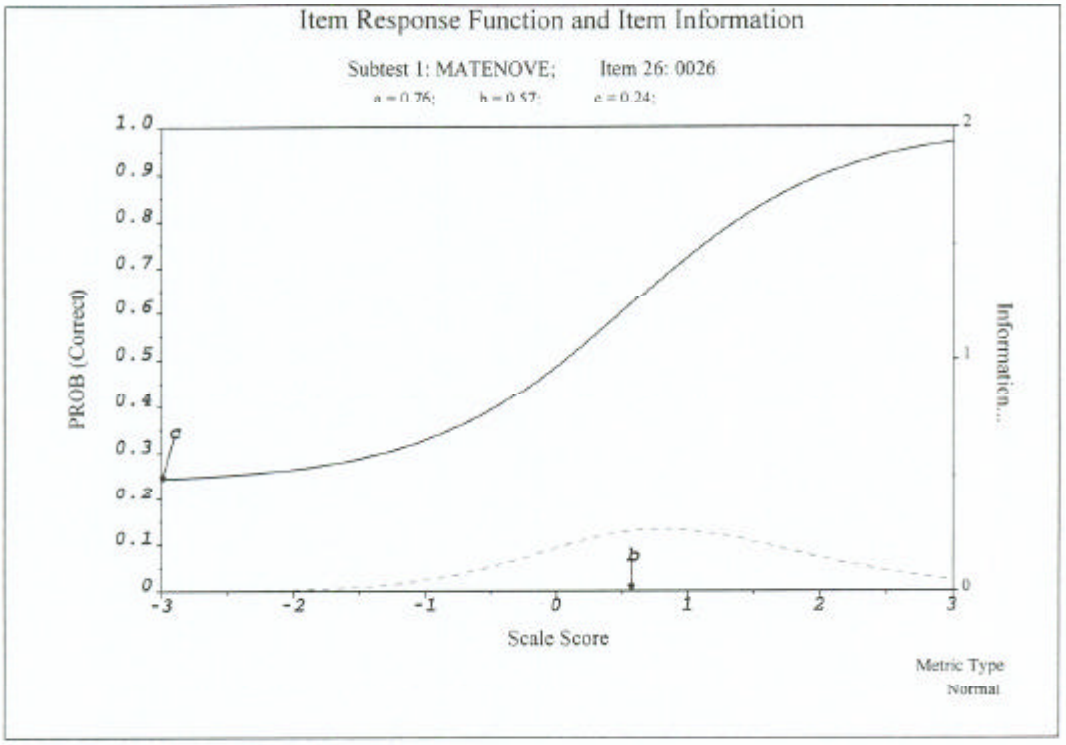
# Gráfico No.25



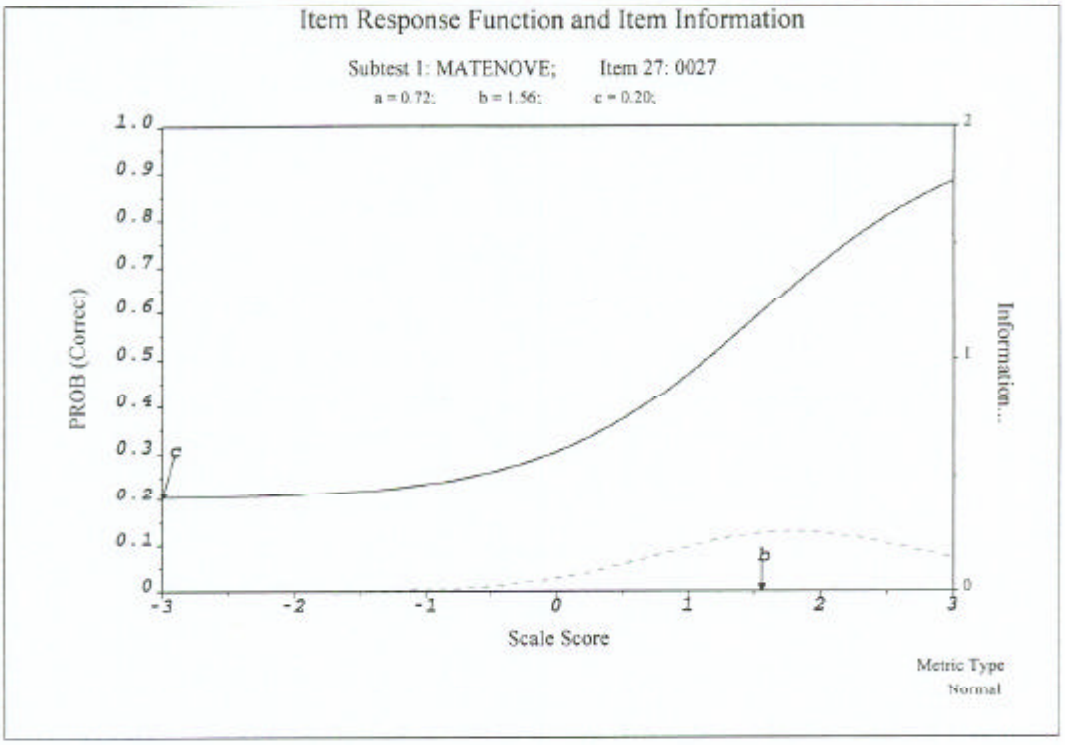
# Gráfico No.26



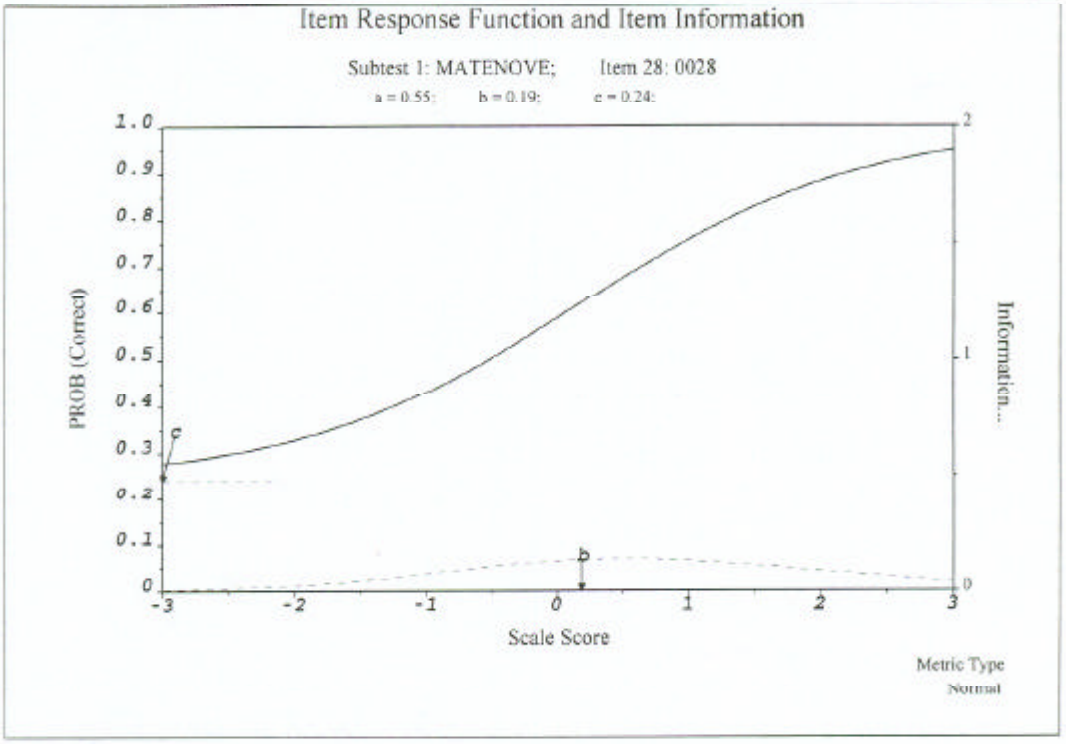
# Gráfico No.27



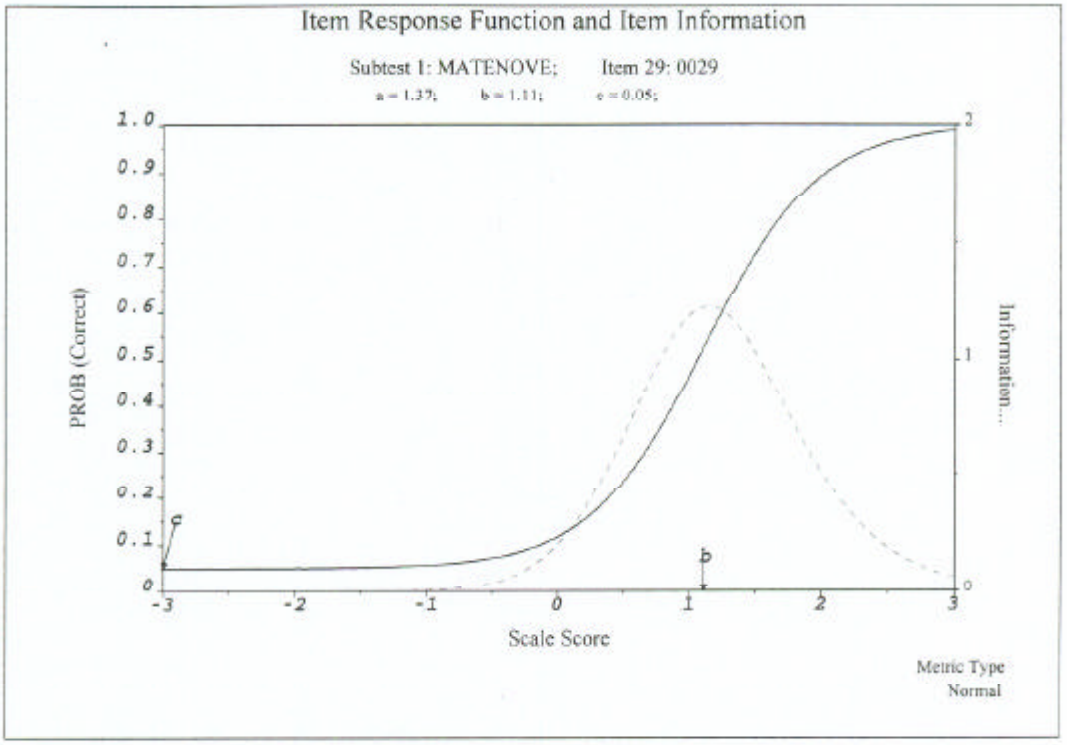
# Gráfico No.28



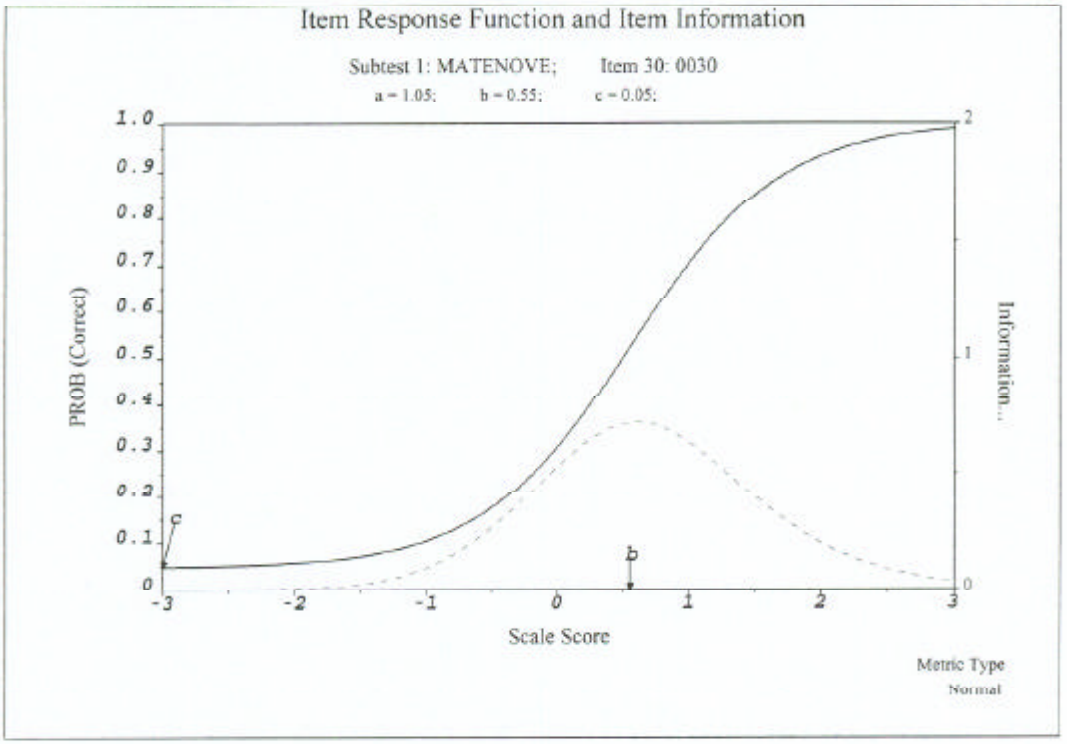
# Gráfico No.29



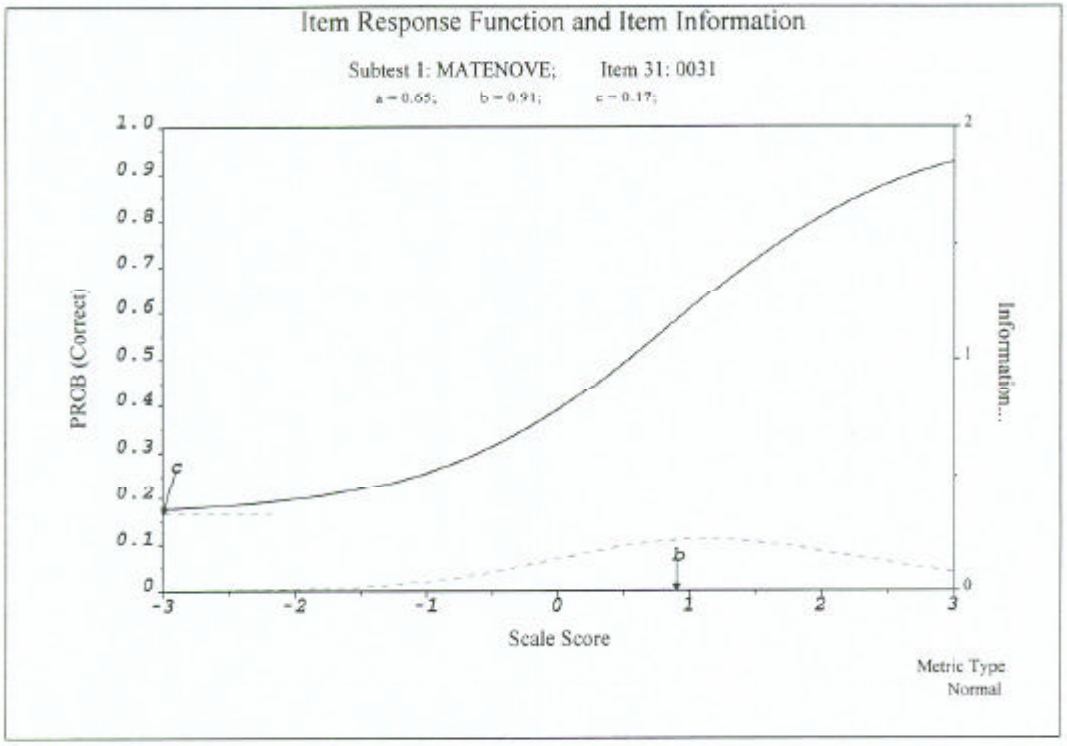
# Gráfico No.30



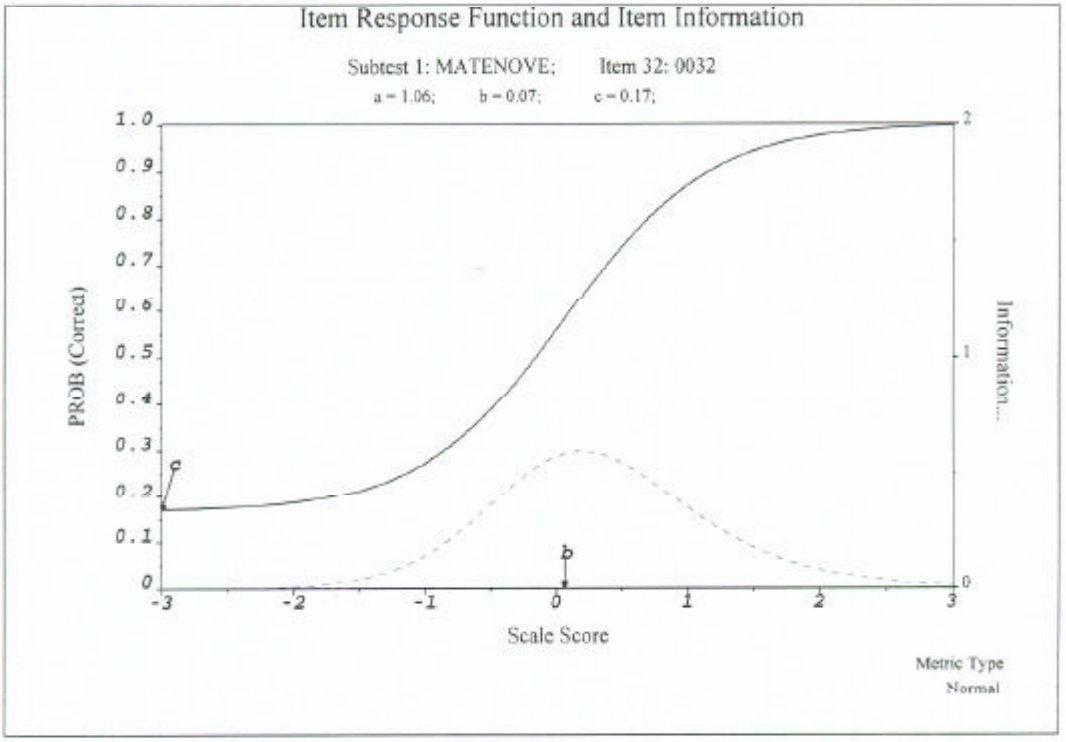
# Gráfico No.31



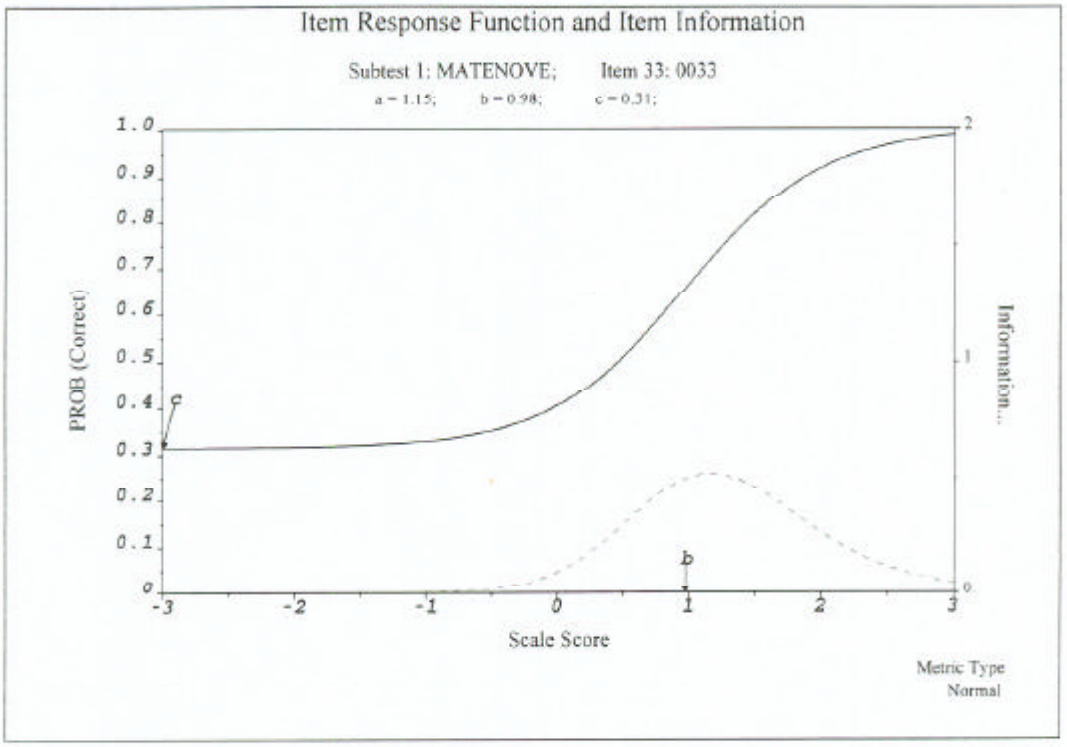
# Gráfico No.32



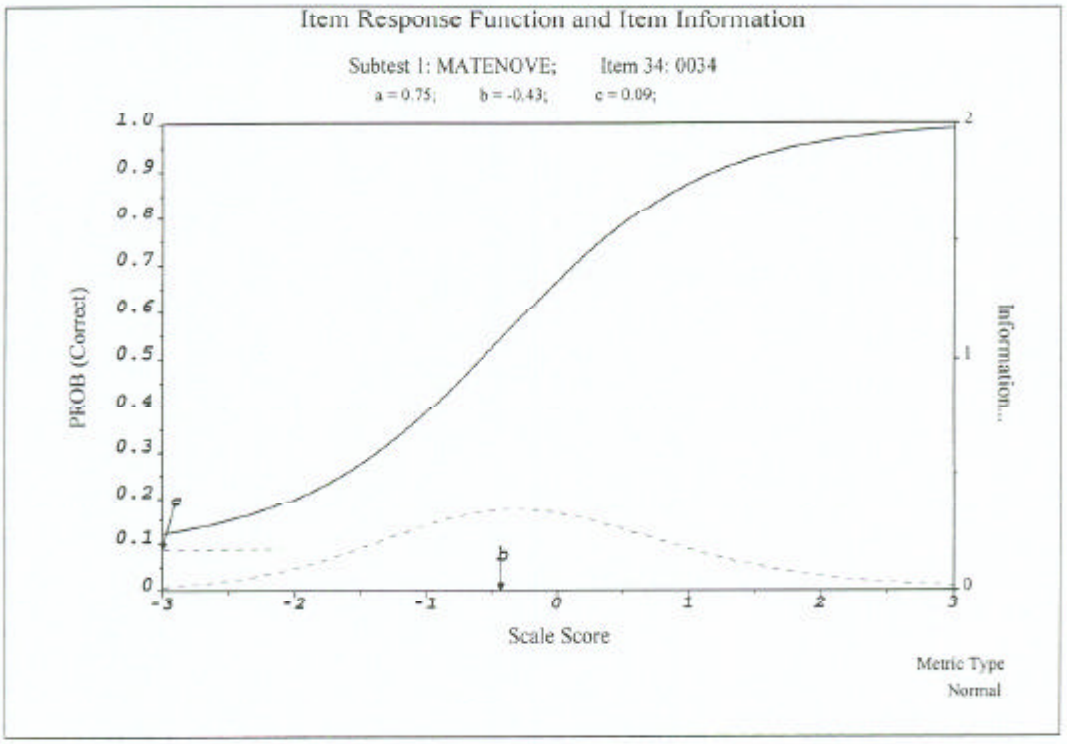
# Gráfico No.33



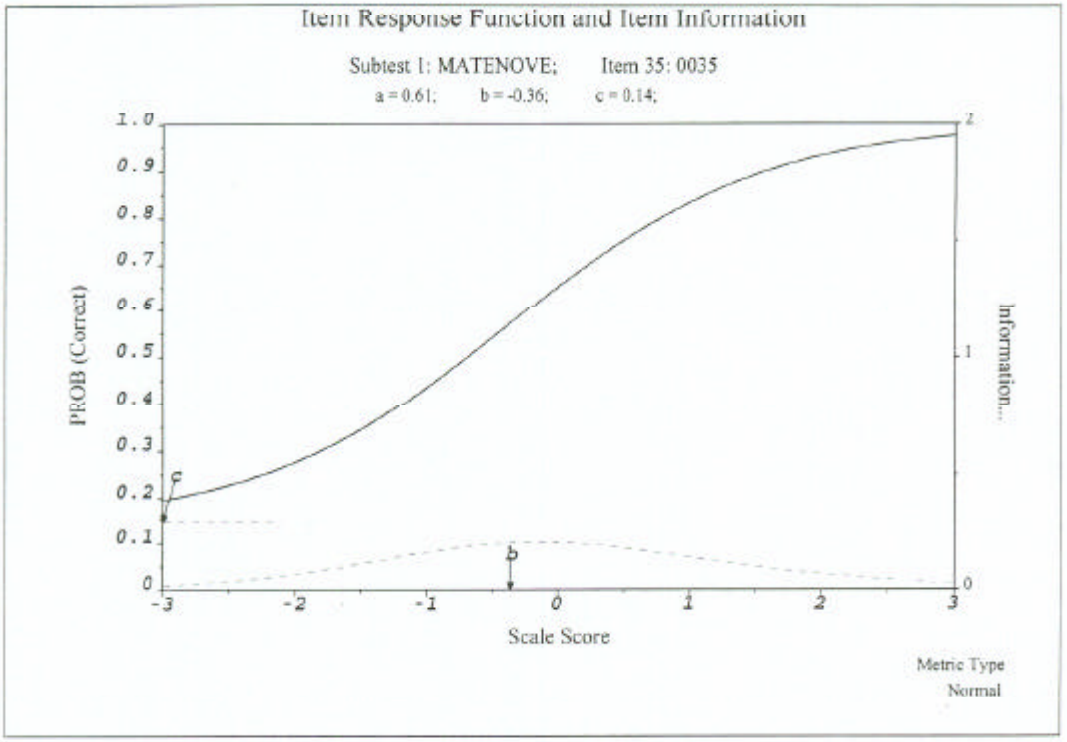
# Gráfico No.34



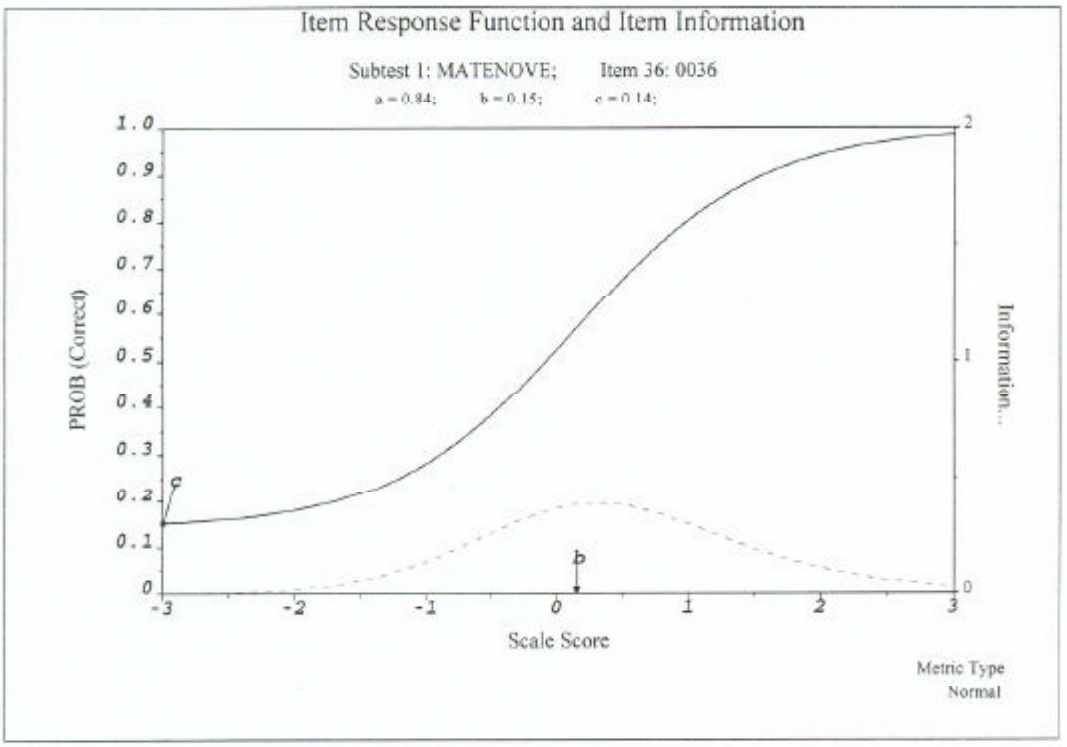
# Gráfico No.35



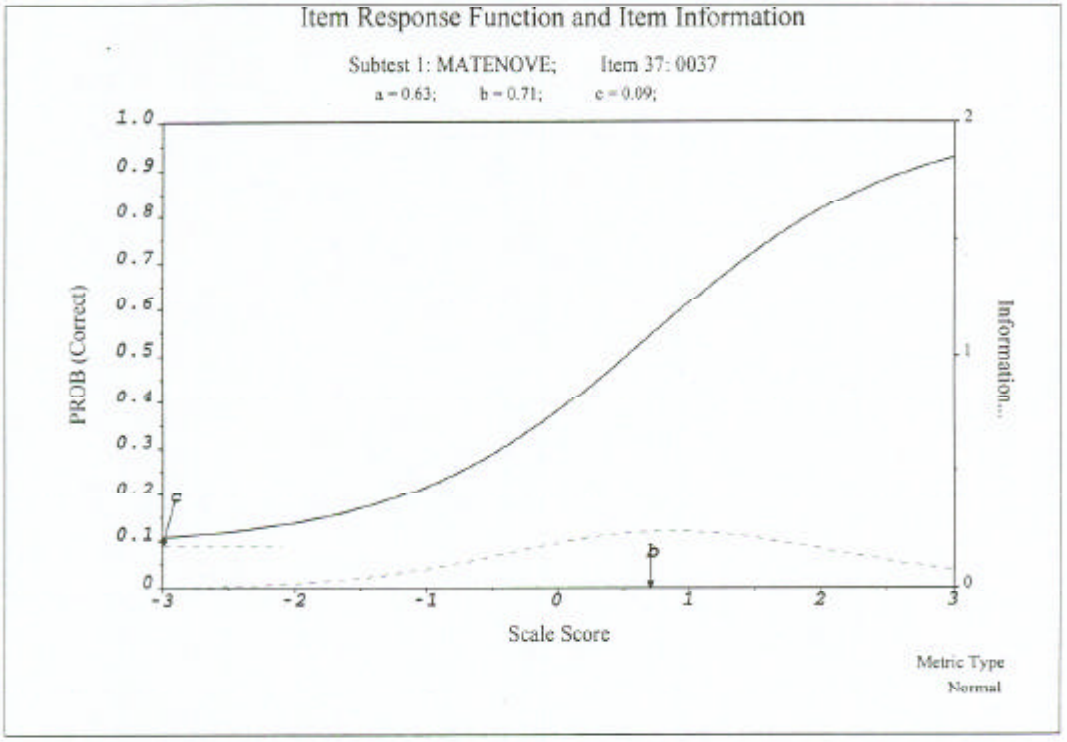
# Gráfico No.36



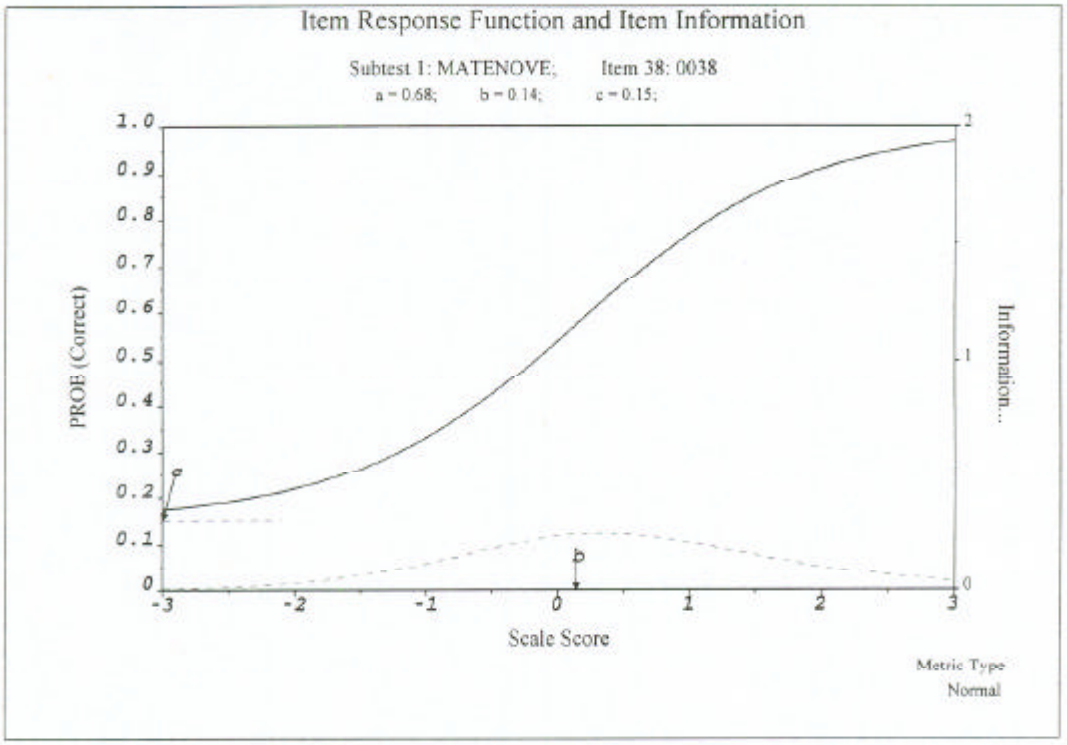
# Gráfico No.37



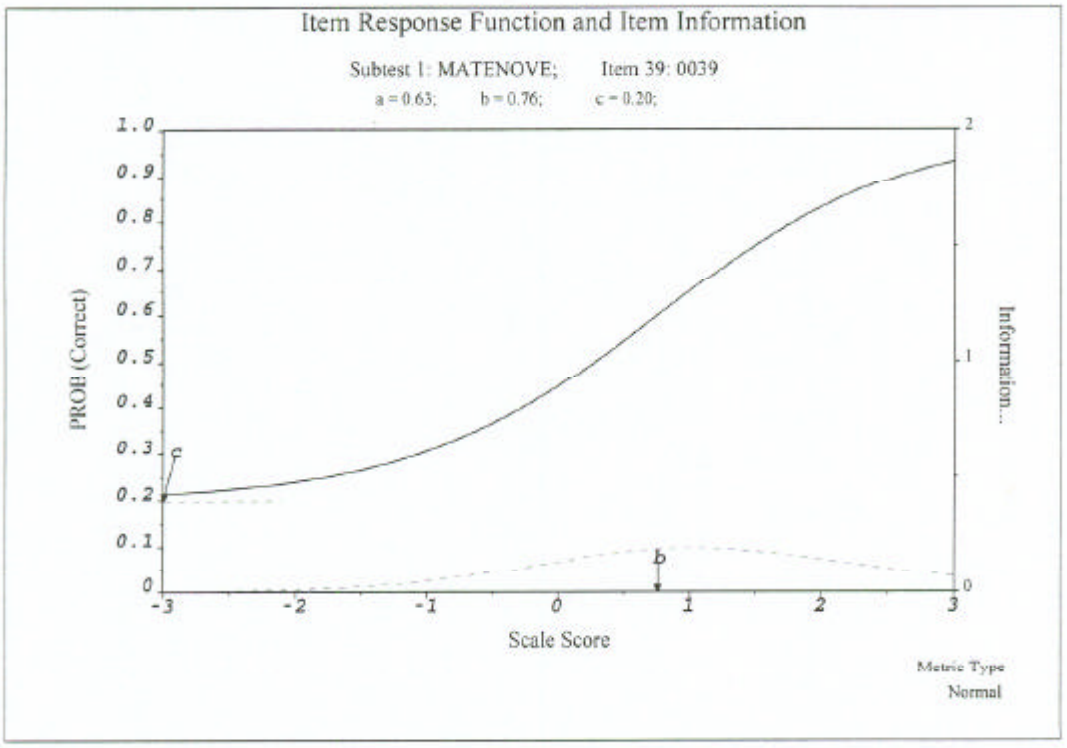
# Gráfico No.38



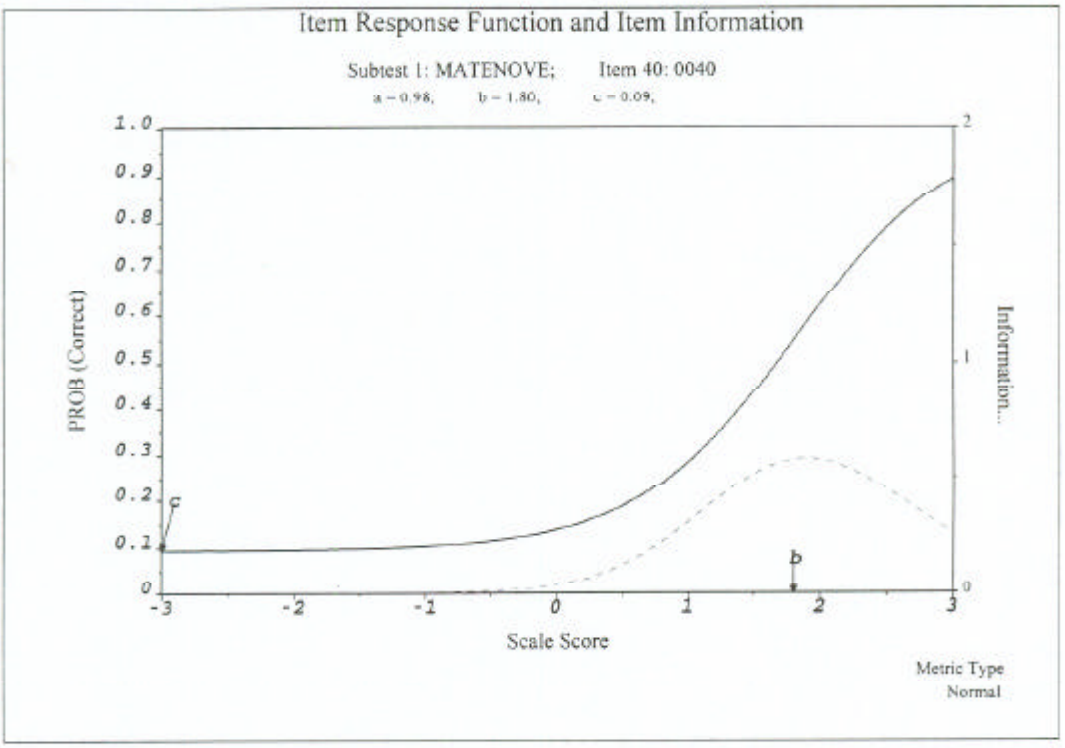
# Gráfico No.39



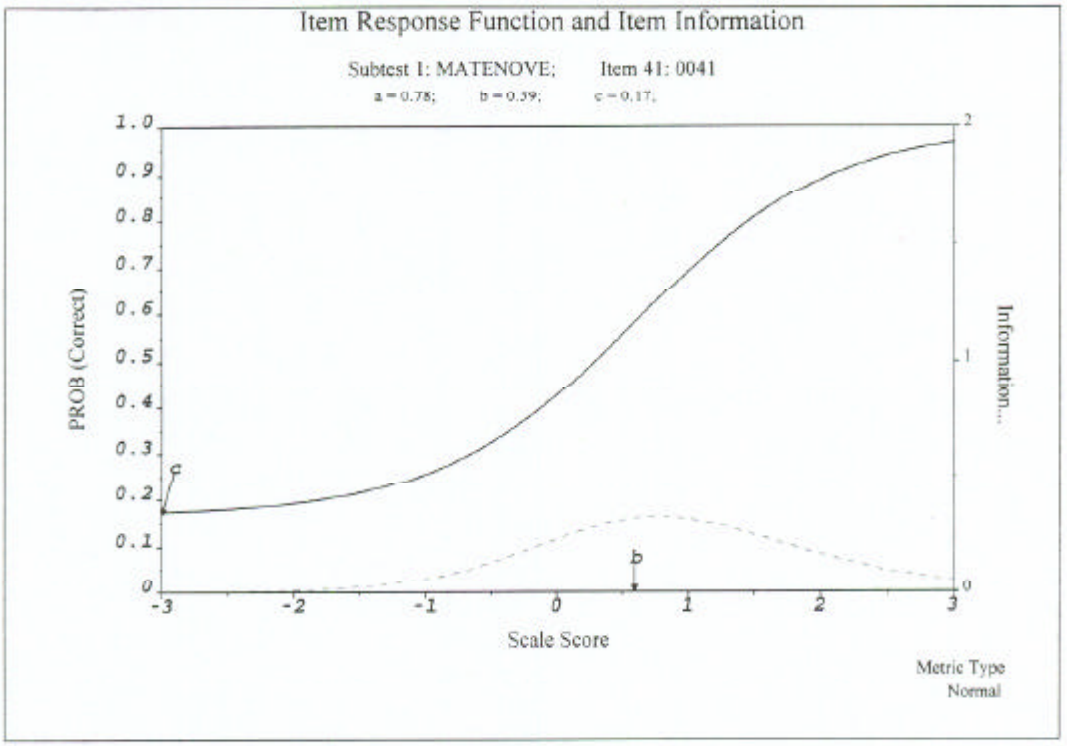
# Gráfico No.40



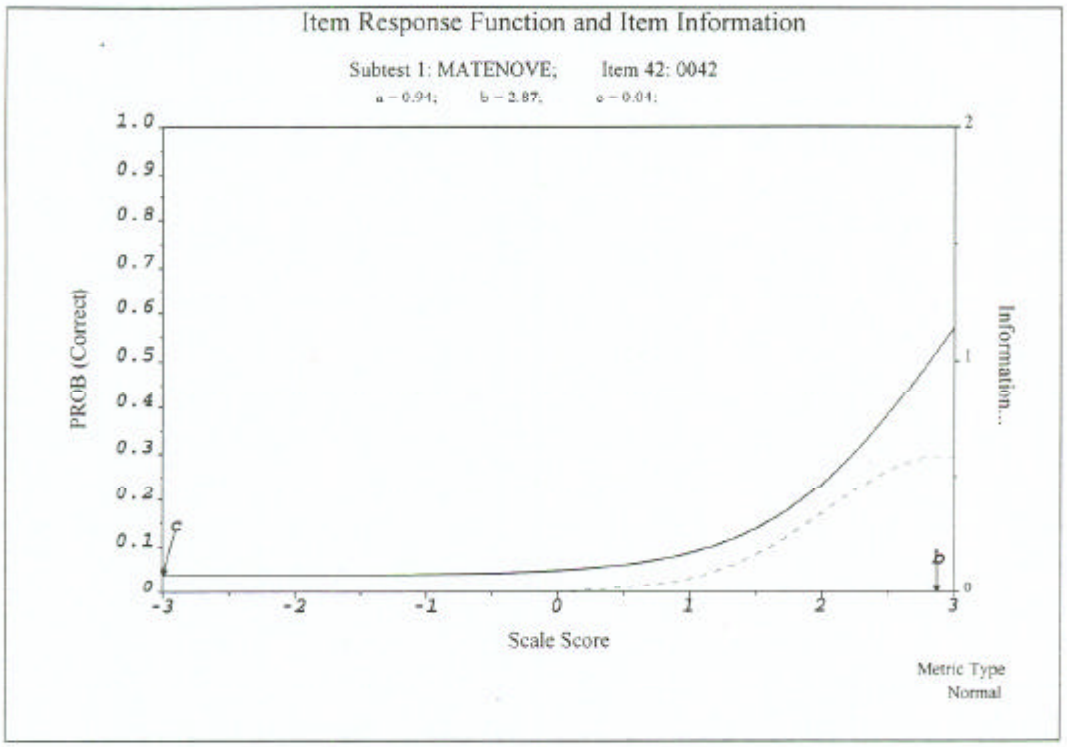
# Gráfico No.41



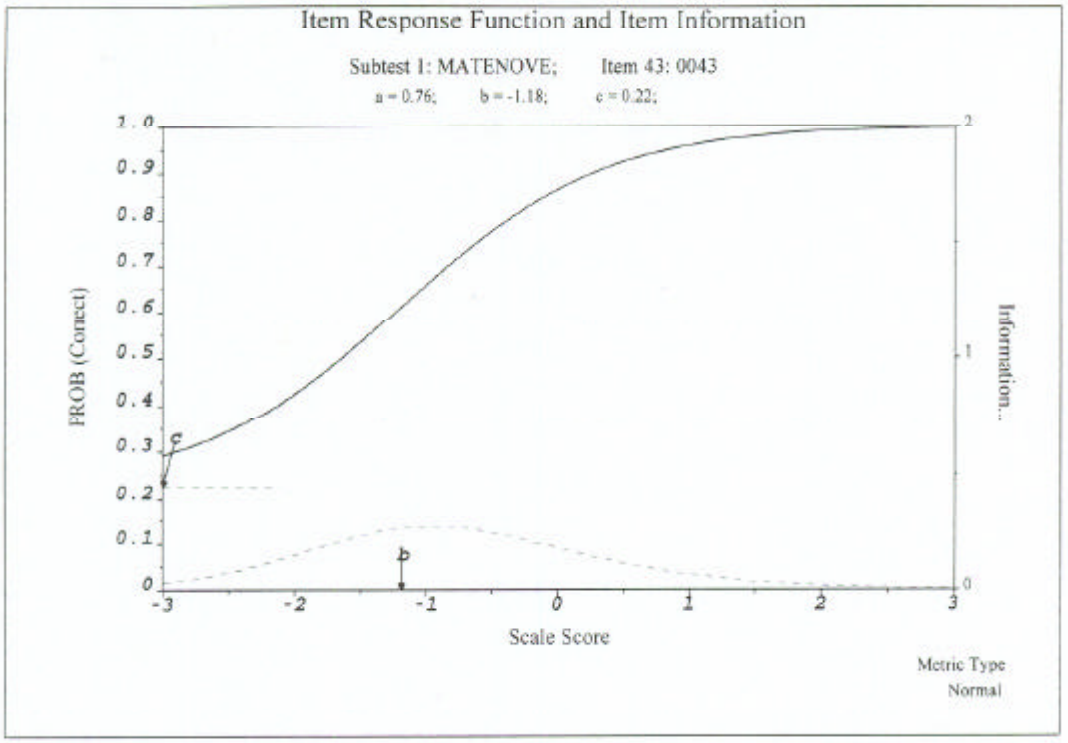
# Gráfico No.42



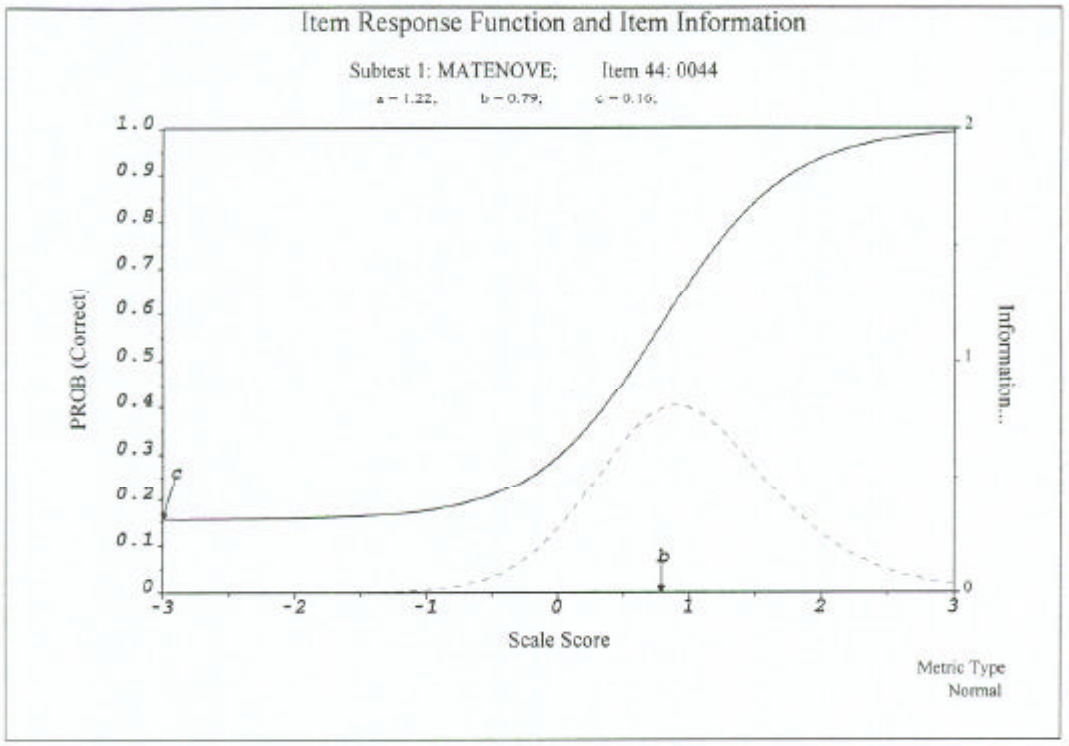
# Gráfico No.43



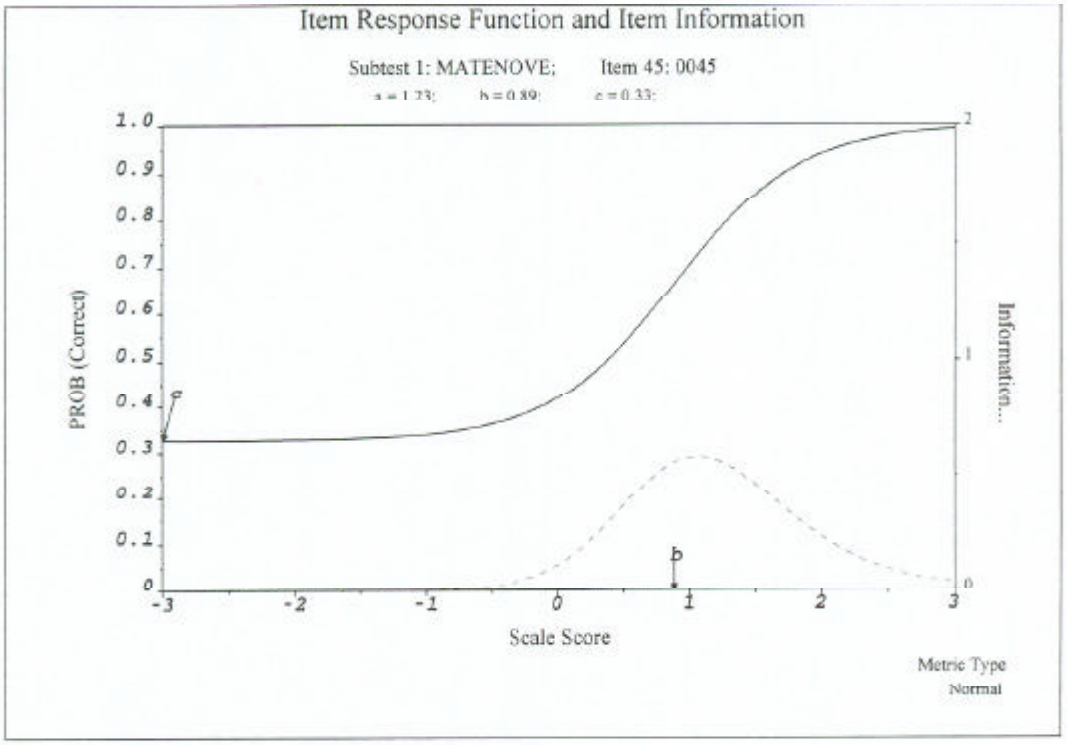
# Gráfico No.44



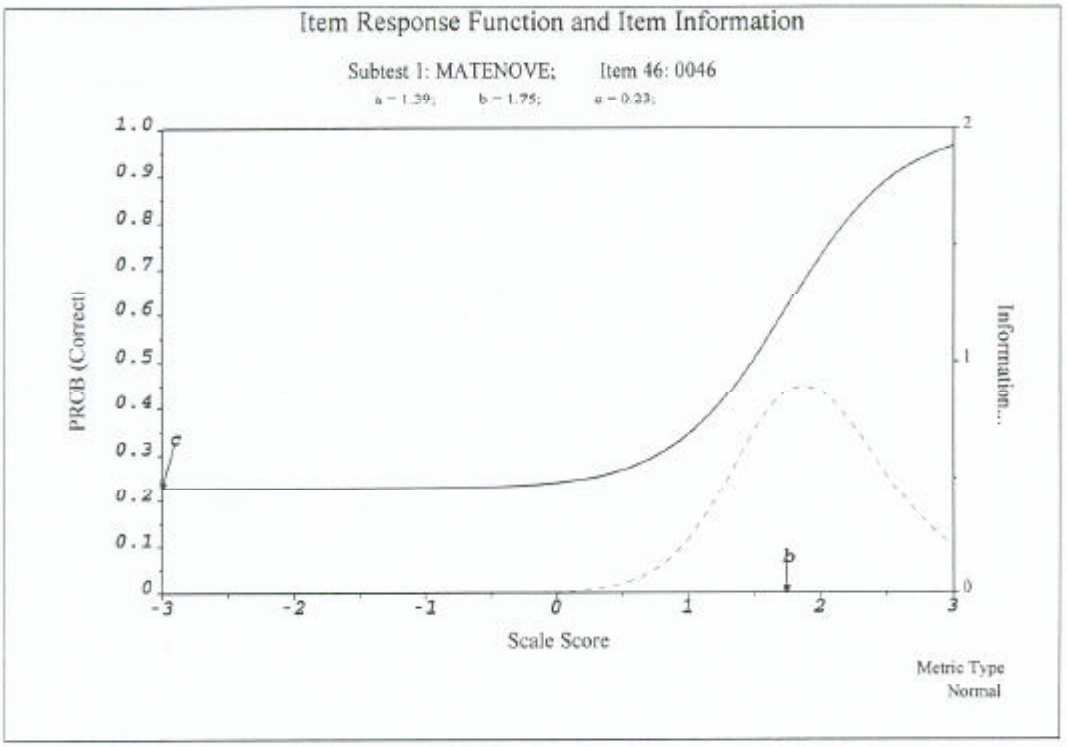
# Gráfico No.45



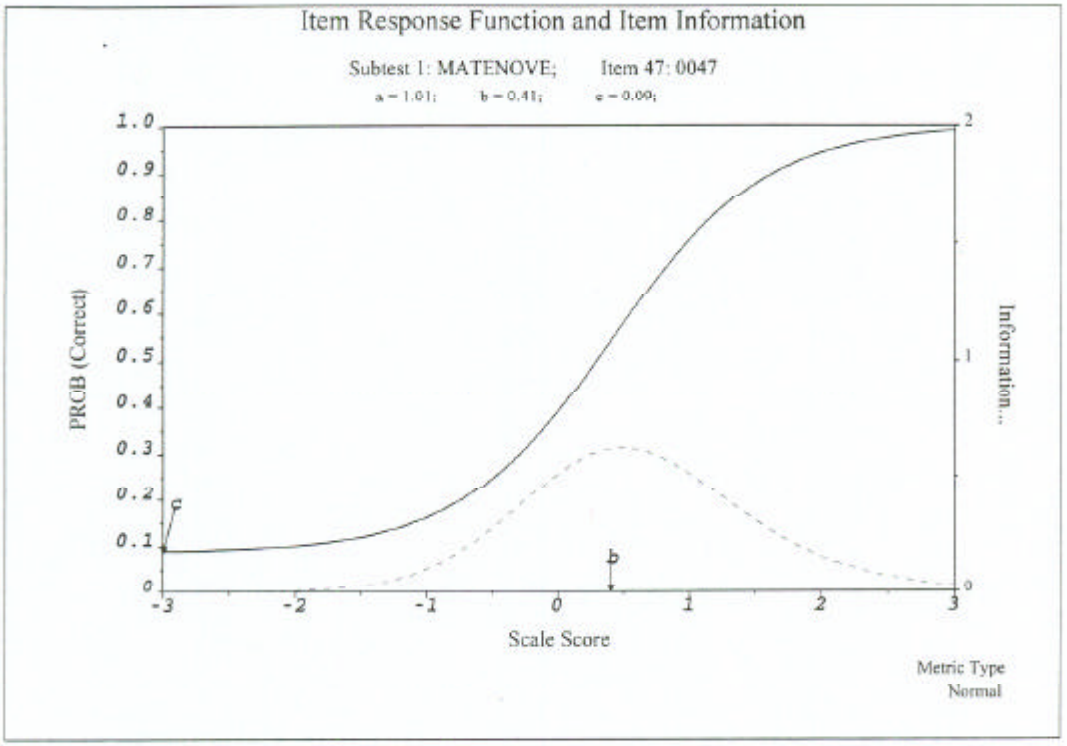
# Gráfico No.46



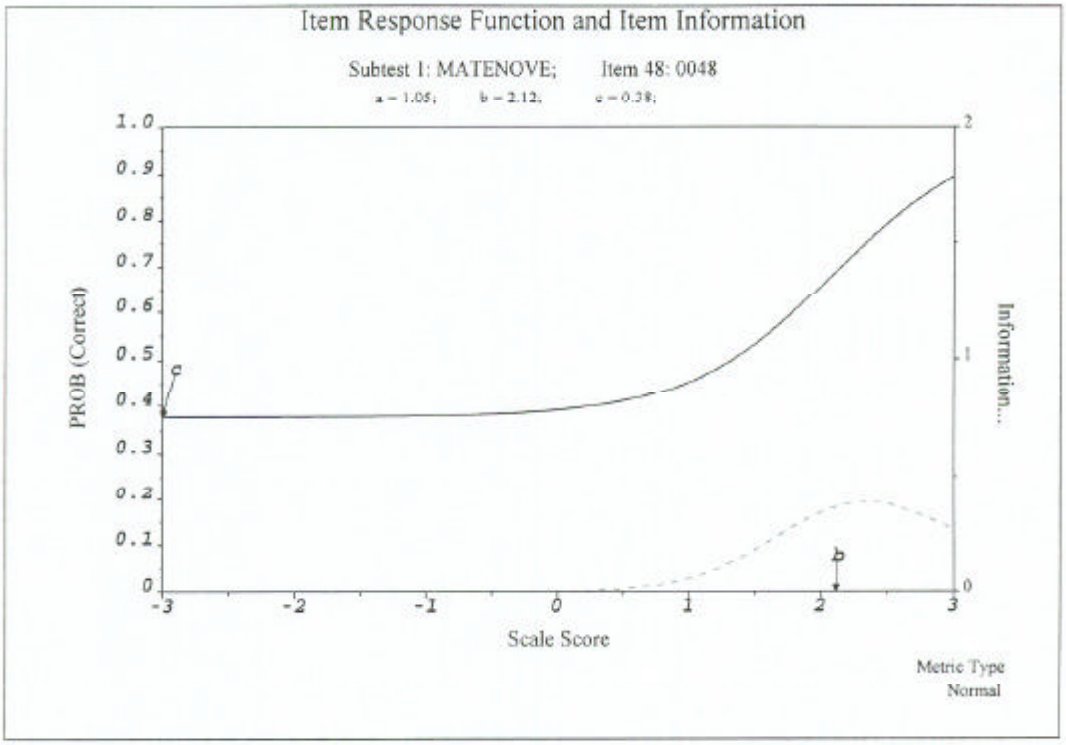
# Gráfico No.47



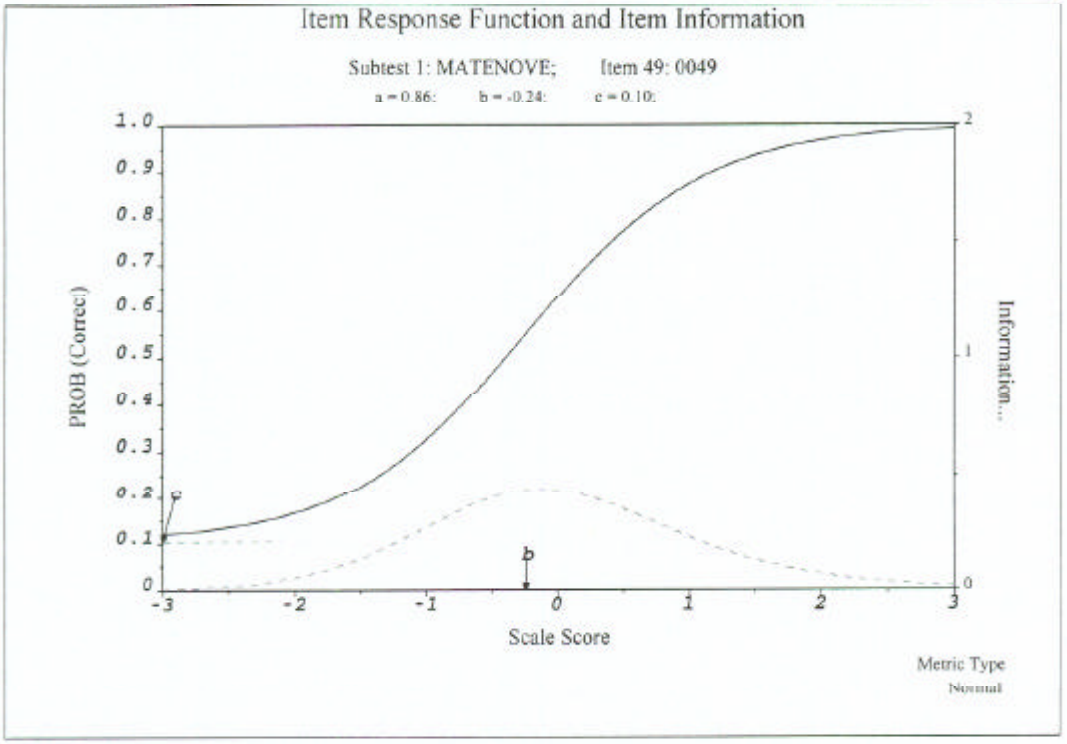
# Gráfico No.48



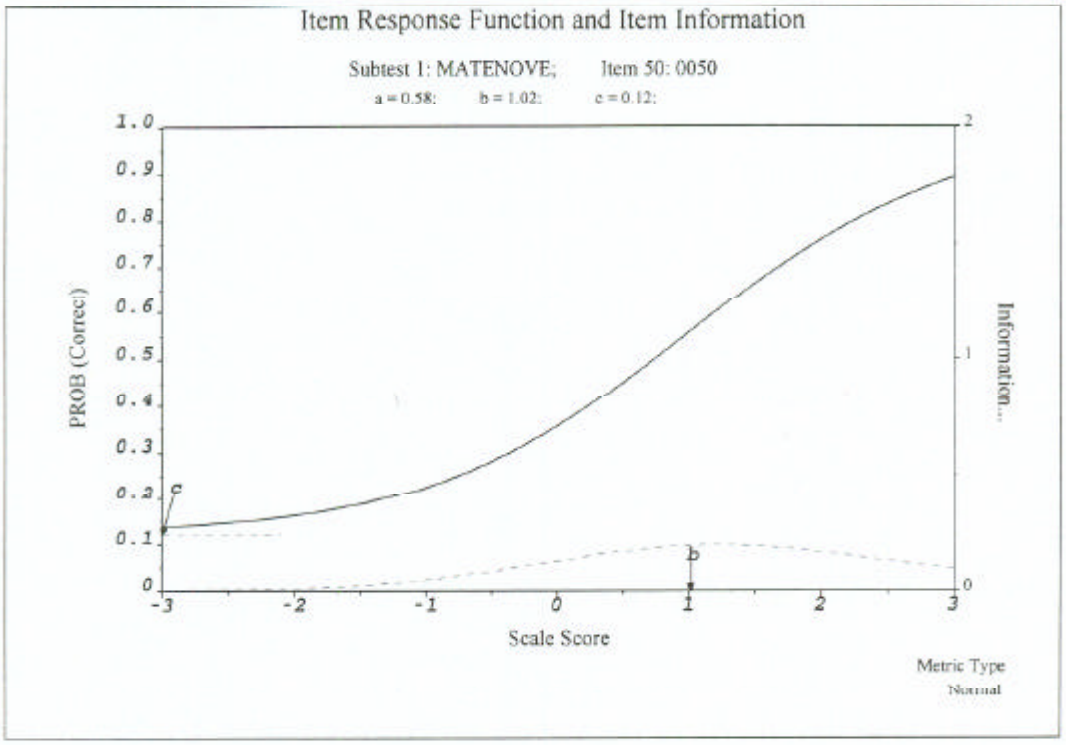
# Gráfico No.49



# Gráfico No.50



# Gráfico No.51



# ANÁLISIS DE SESGO: ESTUDIANTES DE COLEGIOS PÚBLICOS VERSUS ESTUDIANTES DE COLEGIOS PRIVADOS

De acuerdo con la teoría psicométrica, un item presenta sesgo o comportamiento diferencial cuando examinados provenientes de diferentes categorías de población y con los mismos niveles de habilidad tienen diferentes probabilidades de contestar correctamente el item. Según esta definición y bajo la Teoría de Respuesta a los Ítemes, para detectar posible sesgo es necesario comparar las curvas características del item asociadas a los examinados pertenecientes a las dos categorías de población que interesa comparar, por ejemplo hombres y mujeres, estudiantes de colegios privados y colegios públicos, etcétera. En la medida en que se encuentren diferencias significativas en sus respectivas curvas, en esa medida se puede decir que hay evidencia de un posible sesgo. Un item que presente esa característica debe ser revisado, puesto que el comportamiento diferencial en las dos subpoblaciones representa una amenaza a la validez del instrumento de medición. En el presente estudio se hicieron comparaciones entre colegios públicos y privados y entre colegios urbanos y rurales. No fue posible realizar la comparación por sexo (hombres y mujeres) debido a que el archivo de datos original no contenía el sexo del estudiante como variable.

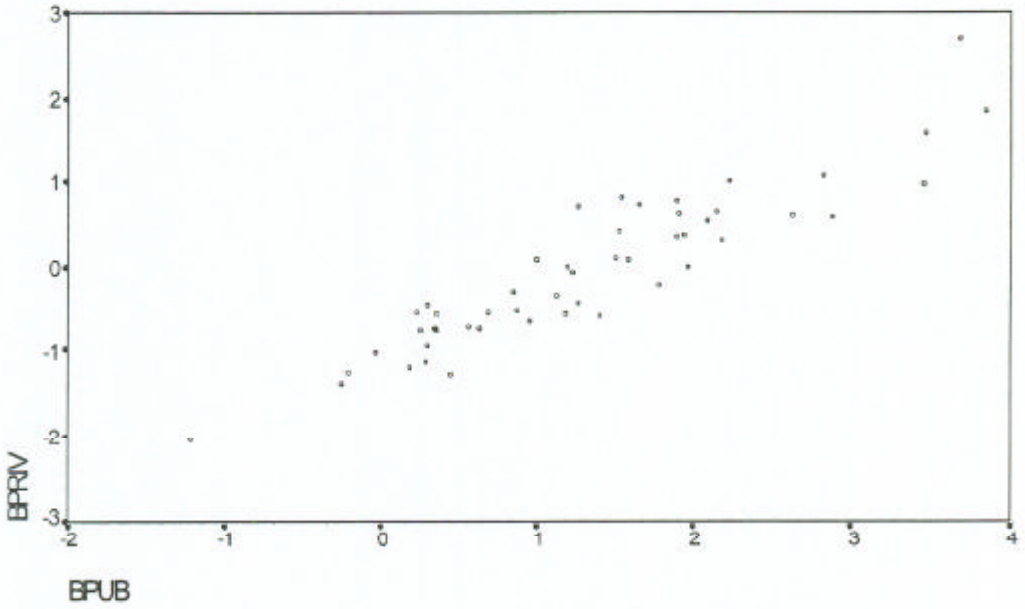
El procedimiento para identificar los ítemes con posible sesgo involucró la construcción de diagramas de dispersión para cada una de las parejas de parámetros a y b (discriminación y dificultad) asociadas a las curvas características del ítem en las dos subpoblaciones a comparar, es decir, un diagrama de dispersión asociando la discriminación, parámetro a de la curva, para colegios públicos y privados y un diagrama de dispersión asociando la dificultad, parámetro b ,para esas mismas categorías de población. El mismo procedimiento se realizó en la comparación urbano-rural. Los puntos en esos diagramas de dispersión que se encontraran muy alejados de la tendencia general del conjunto identifican ítemes con posible sesgo. Para identificar estos ítemes se realizó un análisis de regresión por medio del paquete SPSS para establecer cuales puntos en los diagramas representaban valores extremos o " outliers". Con base en este procedimiento se logró establecer que los ítemes analizados números 4,5,12, 18, 20, 29, 33, 42, 46,48 y 50 poseen esta característica.

A continuación se presentan los diagramas de dispersión analizados y seguidamente las curvas características del ítem y funciones de información para los ítemes identificados con posible sesgo en la comparación entre colegios públicos y privados, en la siguiente sección se presentará el análisis para los colegios urbanos y rurales.

Del análisis de las curvas características y funciones de información asociadas a estudiantes provenientes de colegios privados y públicos se puede concluir que, en general, los ítemes más difíciles dan más información para los estudiantes de los colegios privados. Debe recordarse que este análisis controla por niveles de habilidad, es decir, se está comparando la probabilidad de respuesta correcta en estudiantes con los mismos niveles de habilidad en colegios privados y públicos. Aún así, se nota que la mayoría de los ítemes que presentan evidencia de sesgo son comparativamente más difíciles para los estudiantes de colegios públicos.

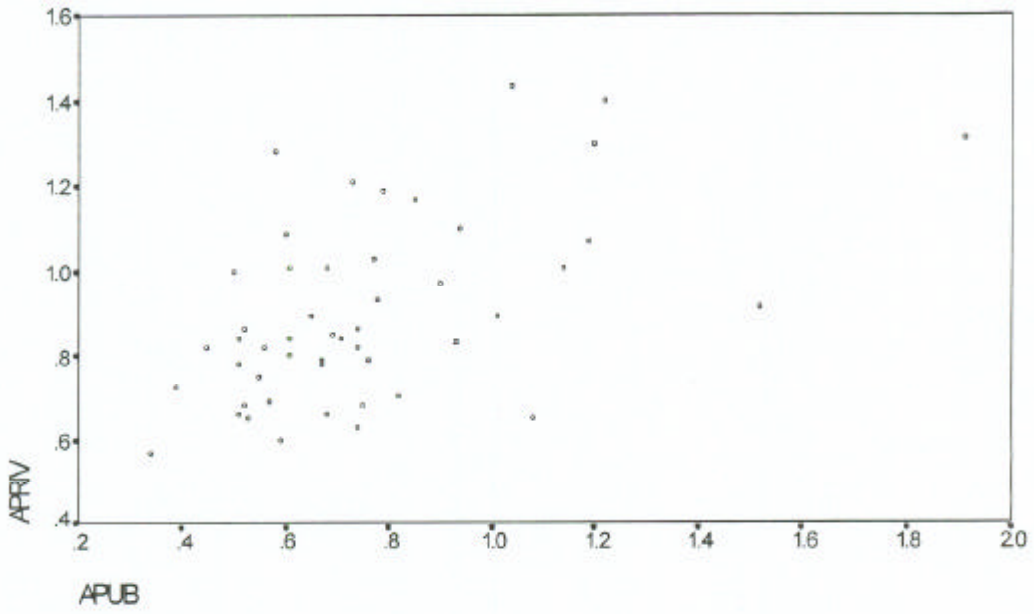
# Gráfico No.52

## Diagrama de Dispersión para el parámetro b, dificultad, en Colegios Públicos y Privados

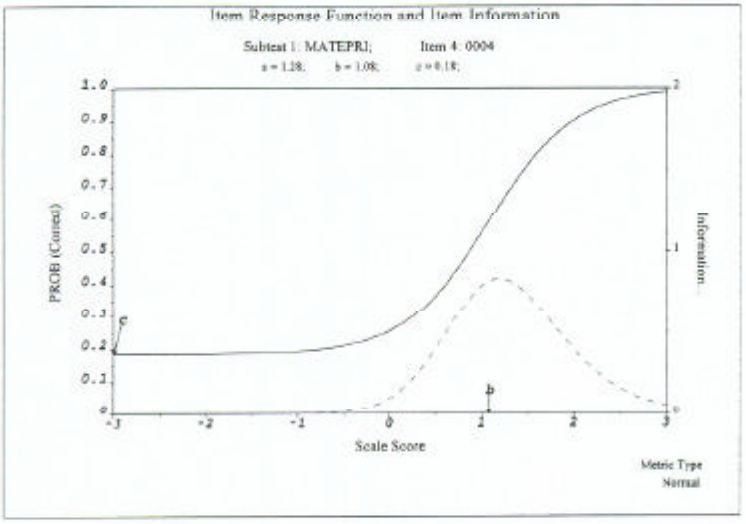


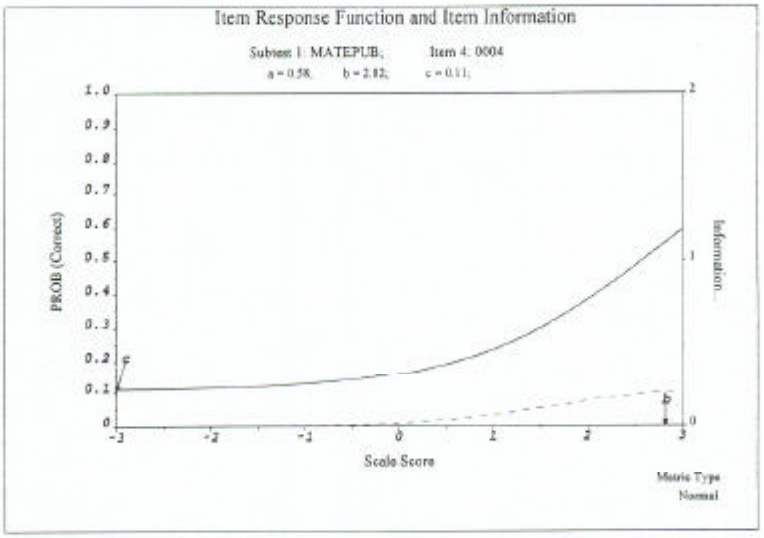
# Gráfico No.53

## Diagrama de Dispersión para el parámetro a, discriminación, en Colegios Públicos y Privados

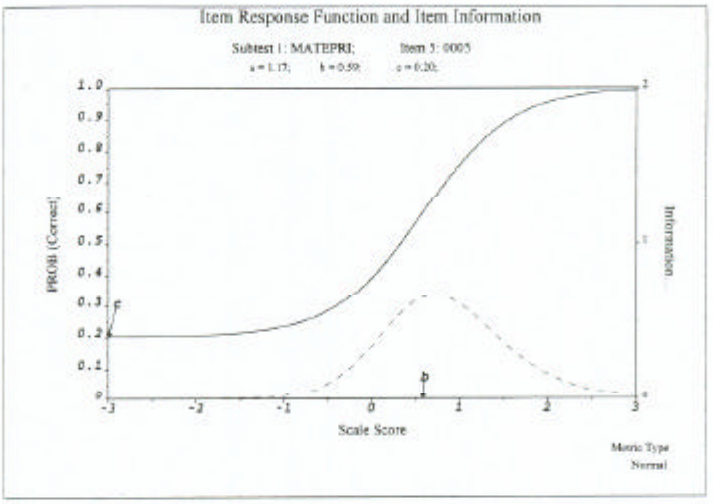


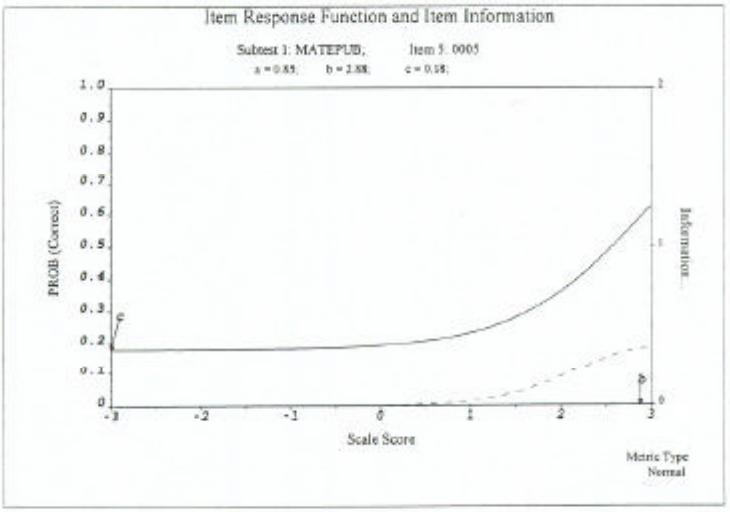
# Gráfico No.54
## Comparación entre Colegios Públicos y Colegios Privados
## Item 4



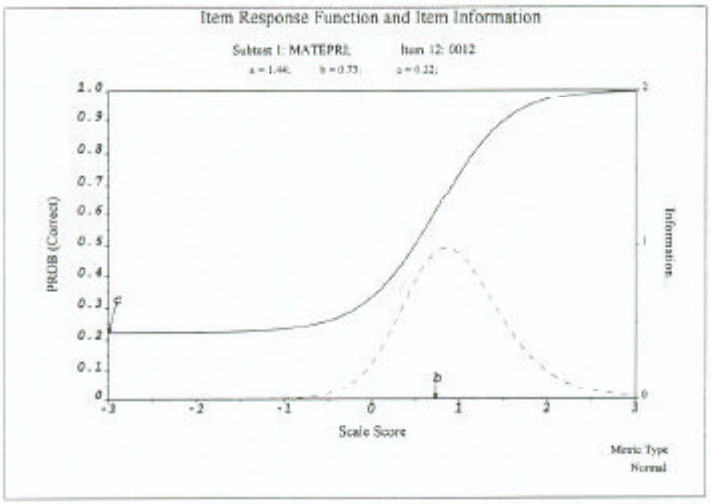



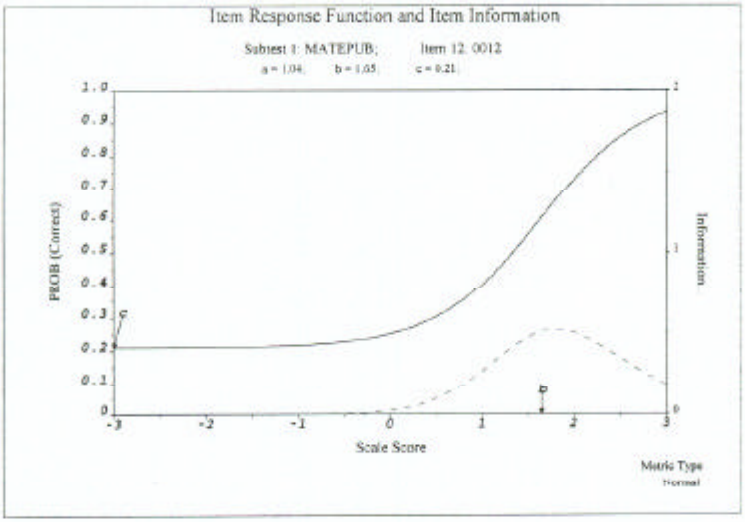
# Gráfico No.55
## Comparación entre Colegios Públicos y Colegios Privados
## Item 5



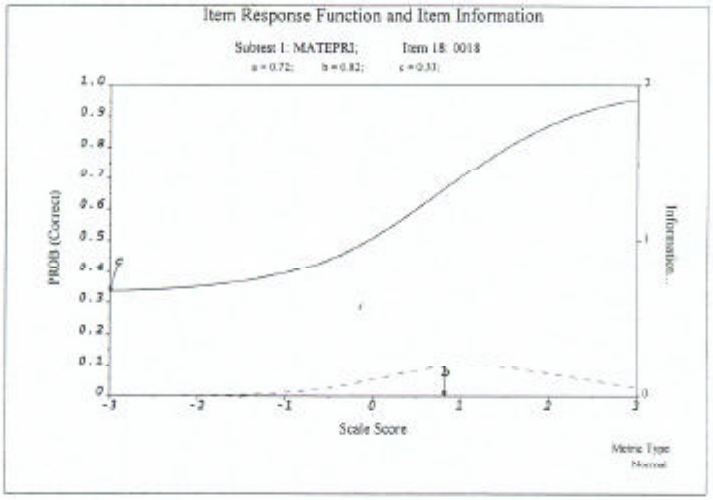



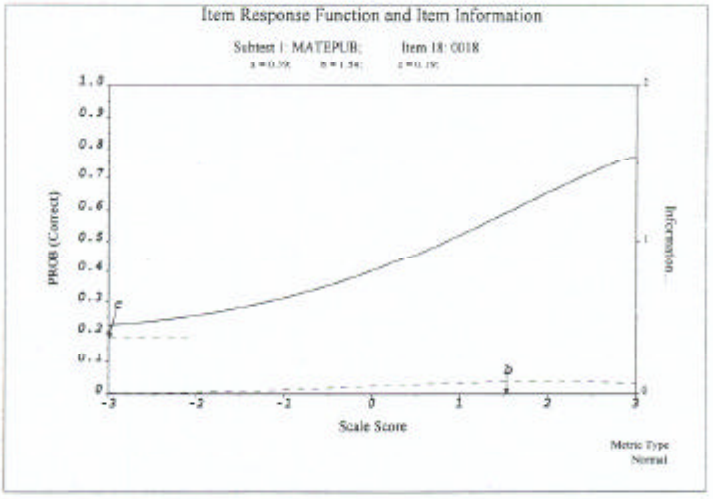
# Gráfico No.56
## Comparación entre Colegios Públicos y Colegios Privados
## Item 12



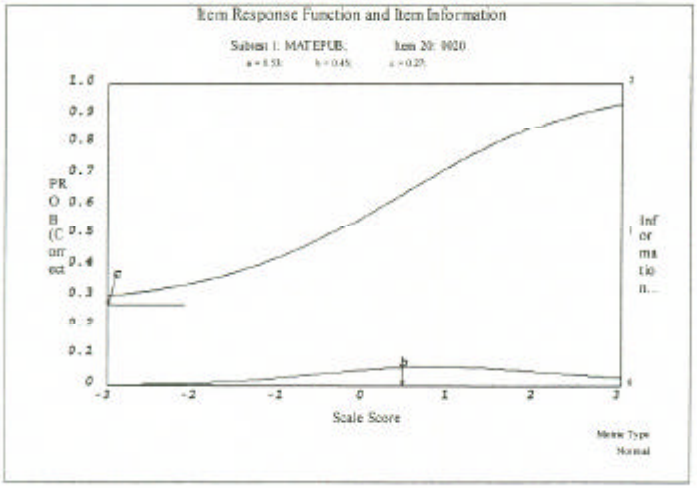
# Gráfico No.57
## Comparación entre Colegios Públicos y Colegios Privados
## Item 18

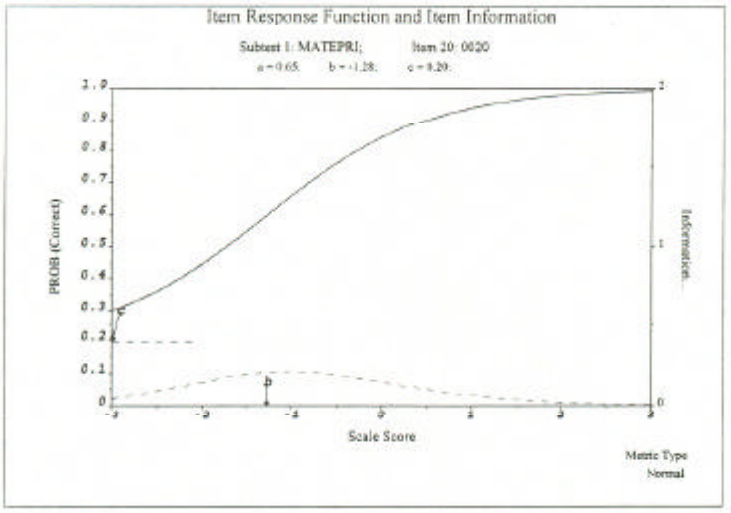


# Gráfico No.58

## Comparación entre Colegios Públicos y Colegios Privados*
## Item 20

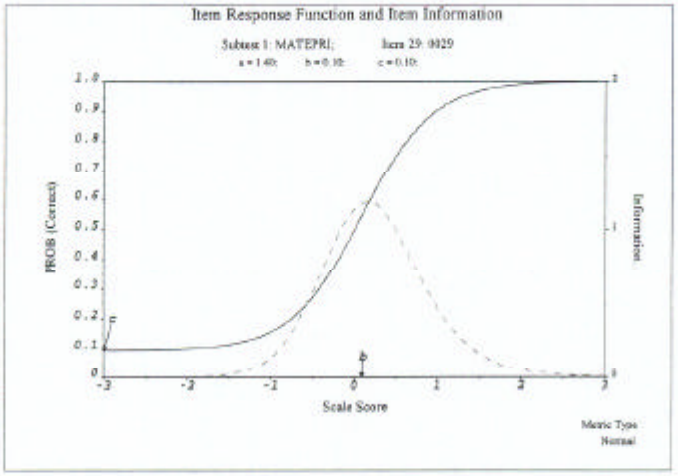


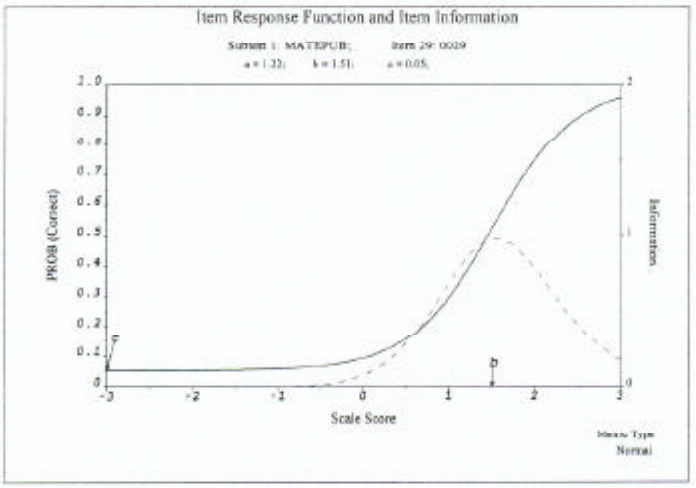




*Nótese que el gráfico correspondiente a los colegios públicos aparece aquí de primero

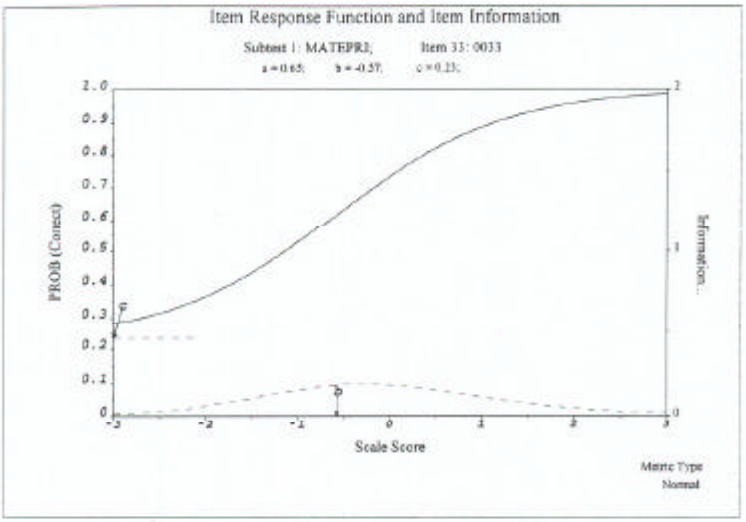
# Gráfico No.59

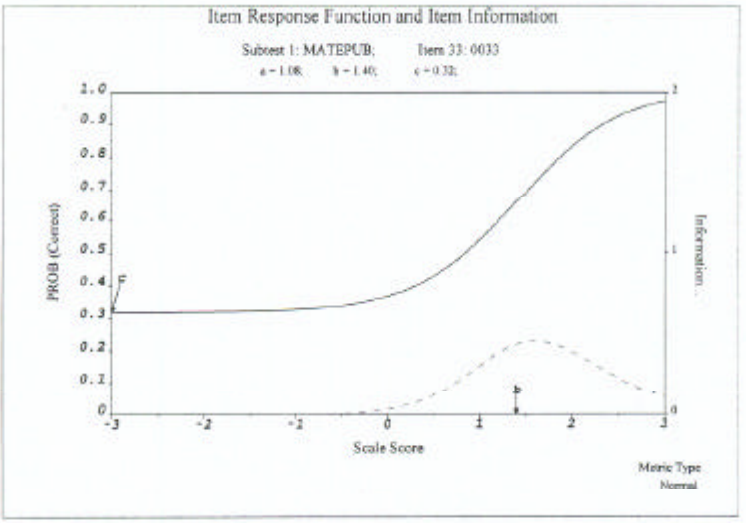
## Comparación entre Colegios Públicos y Colegios Privados
## Item 29



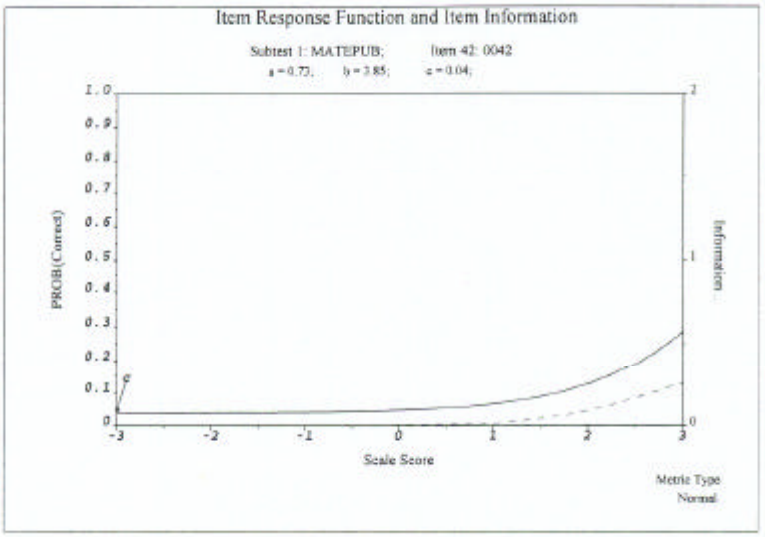



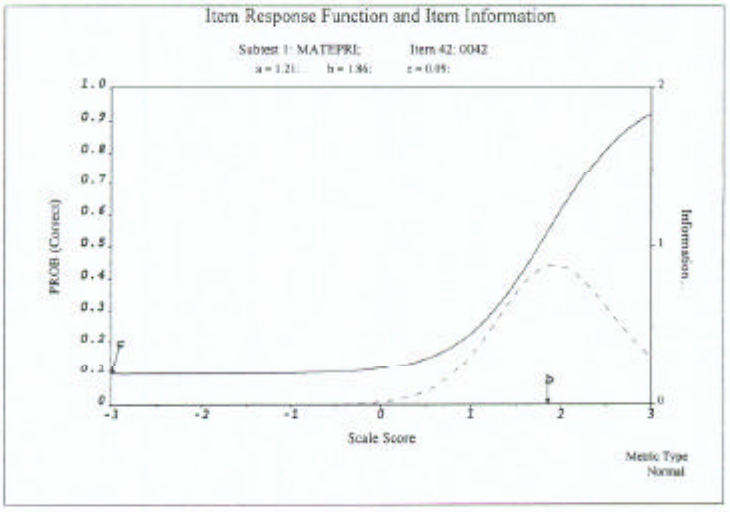
# Gráfico No.60
## Comparación entre Colegios Públicos y Colegios Privados
## Item 33



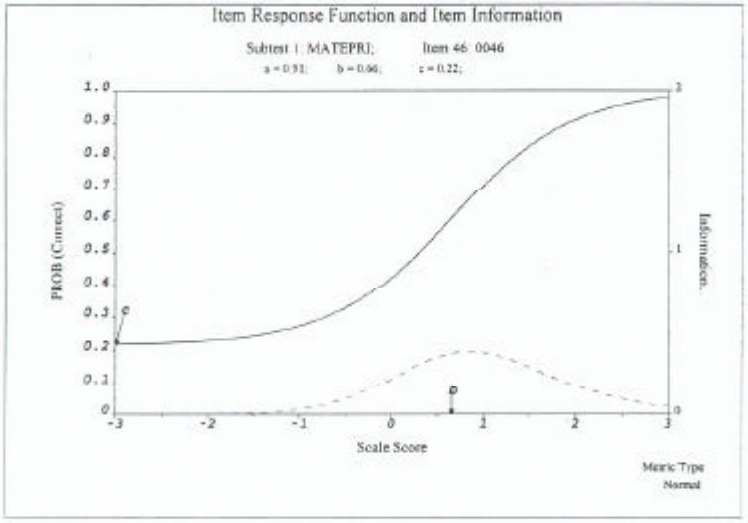
# Gráfico No.61
## Comparación entre Colegios Públicos y Colegios Privados*
## Item 42

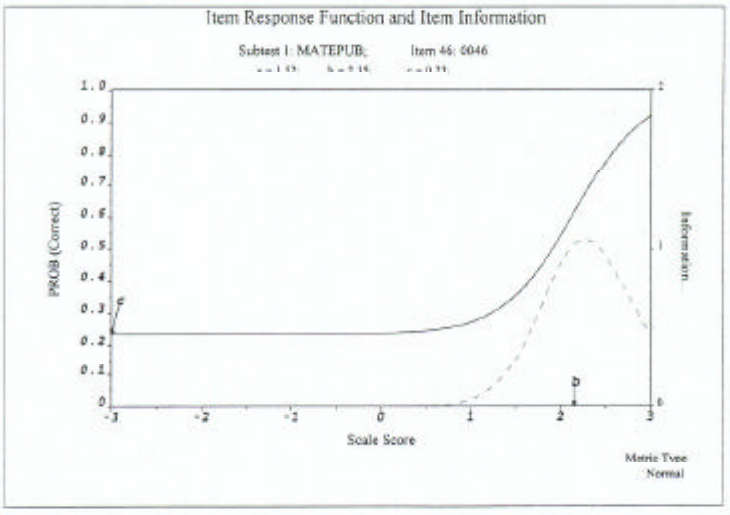


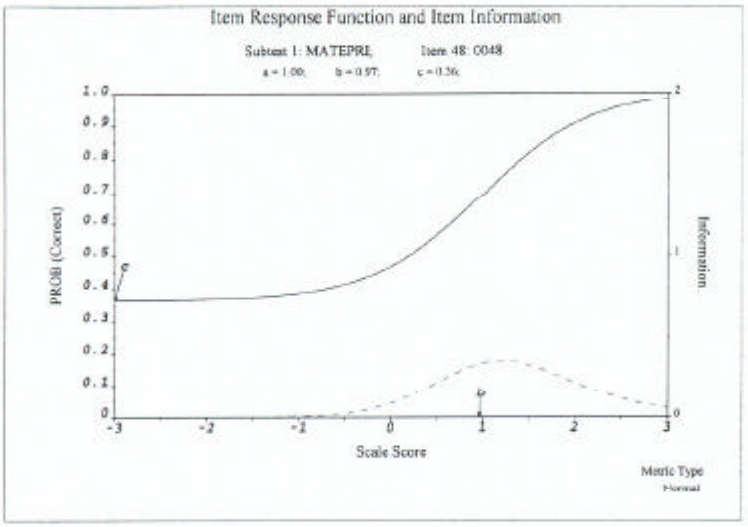

*Nótese que el gráfico correspondiente a los colegios públicos aparece aquí de primero

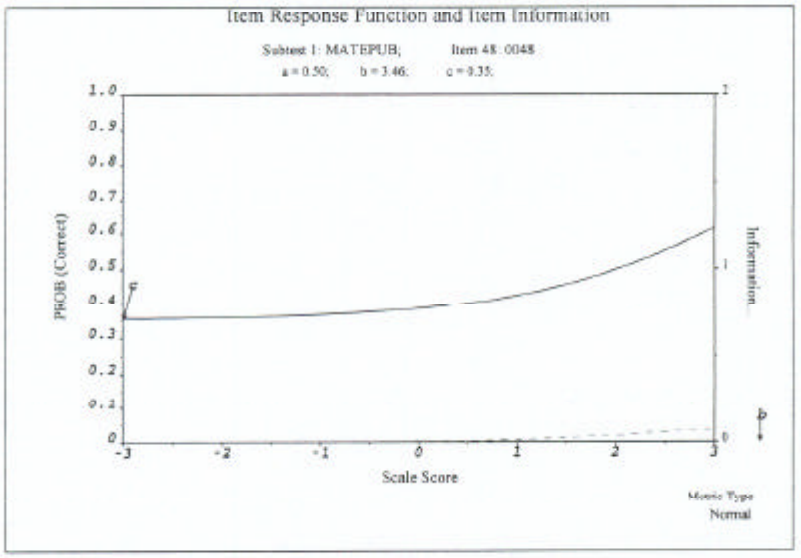
# Gráfico No.62
## Comparación entre Colegios Públicos y Colegios Privados
## Item 46



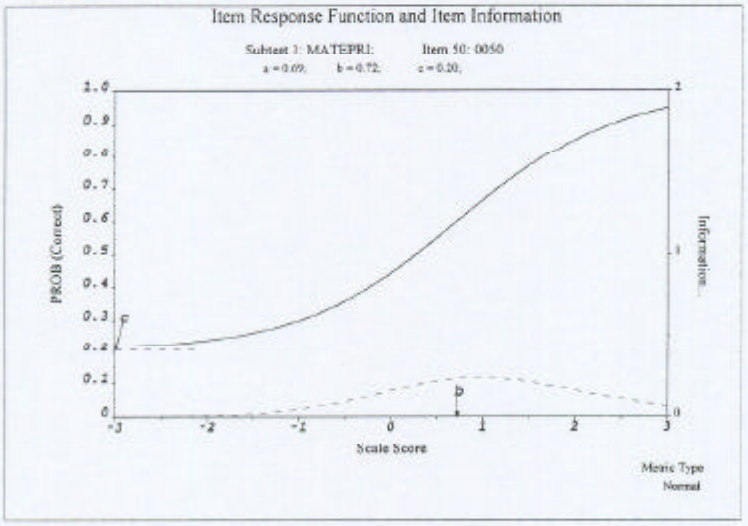



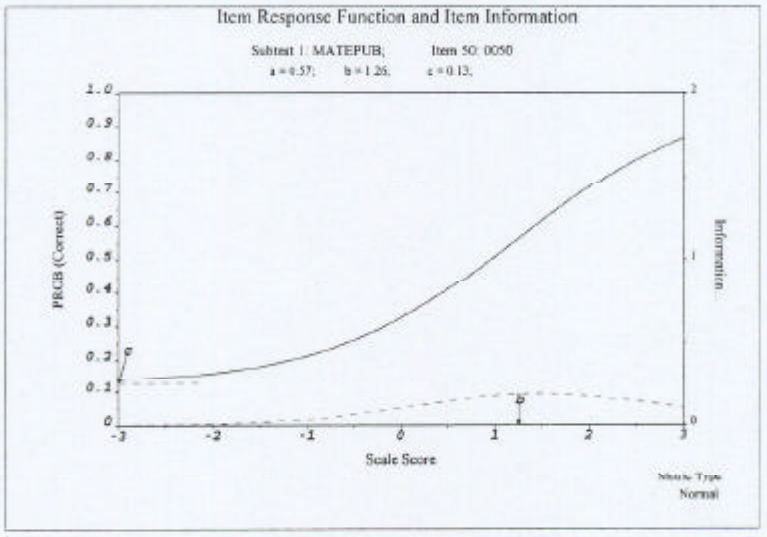
# Gráfico No.63
## Comparación entre Colegios Públicos y Colegios Privados
## Item 48



# Gráfico No.64
## Comparación entre Colegios Públicos y Colegios Privados
## Item 50



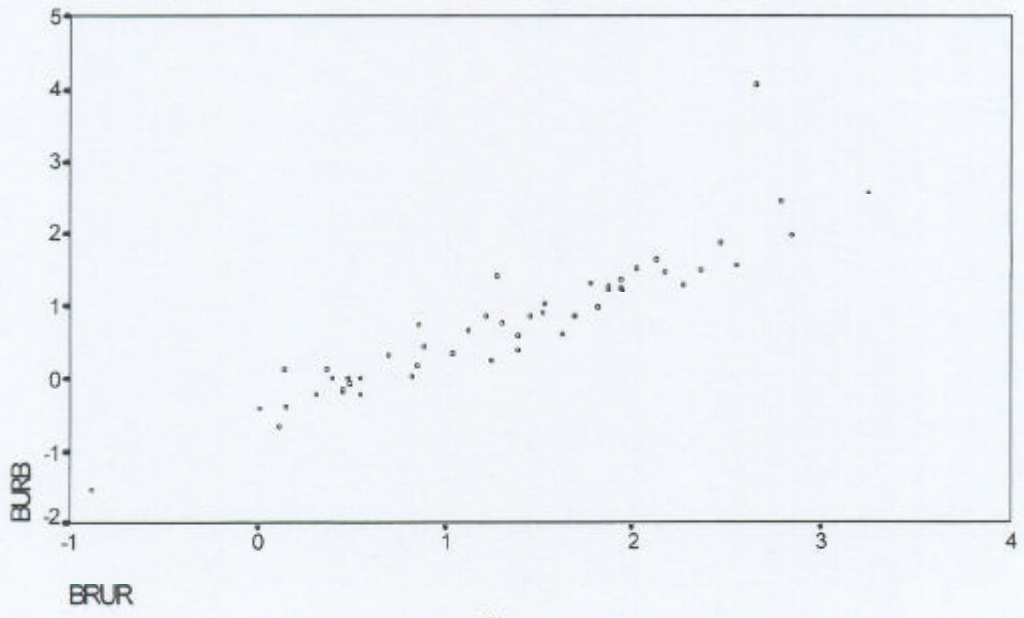



## ANÁLISIS DE SESGO: ESTUDIANTES DE COLEGIOS URBANOS VERSUS ESTUDIANTES DE COLEGIOS RURALES

En esta sección se presentan los diagramas de dispersión (gráficos 65 y 66) que se utilizaron para identificar los ítemes con posible sesgo en la comparación entre colegios urbanos y colegios rurales. De acuerdo con estos diagramas y siguiendo el mismo procedimiento descrito en la sección anterior, se identificaron 6 ítemes con comportamiento diferencial, correspondientes a los números 5, 6, 18, 20, 48 y 49. Las curvas características y funciones de información para cada uno de estos ítemes en las dos subpoblaciones comparadas se presentan a continuación de los diagramas de dispersión en los gráficos 67 a 71.

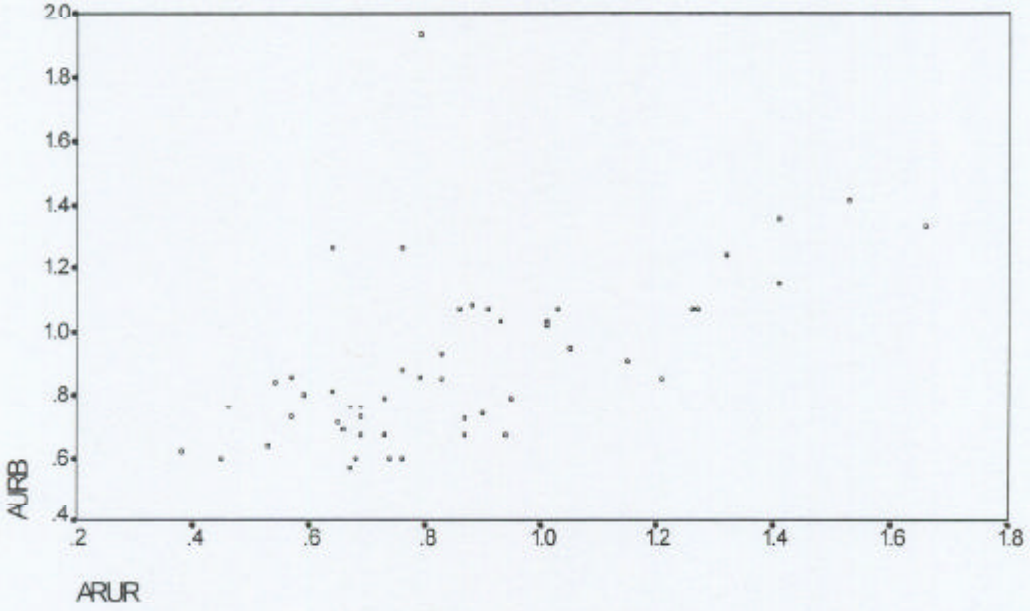
De la observación de estos gráficos se puede concluir que en general estos ítemes tienden a ser un poco más fáciles para los estudiantes de colegios urbanos, aunque la diferencia no es tan marcada como en el caso de la comparación entre públicos y privados. Sí se nota que brindan en general más información (tienen mayor poder de discriminación) en el caso de los estudiantes de colegios urbanos.

### Gráfico No.65
**Diagrama de Dispersión para el parámetro b, dificultad, en Colegios Urbanos y Rurales**

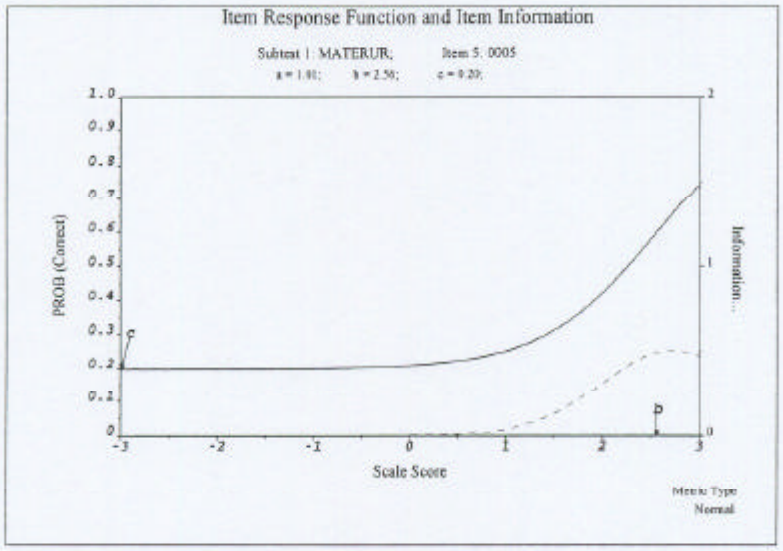


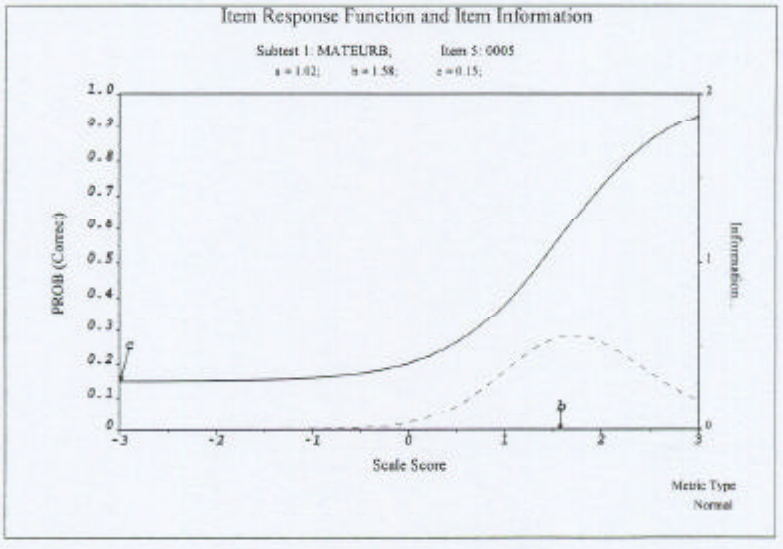
# Gráfico No.66

## Diagrama de Dispersión para el parámetro a, discriminación, en Colegios Urbanos y Rurales

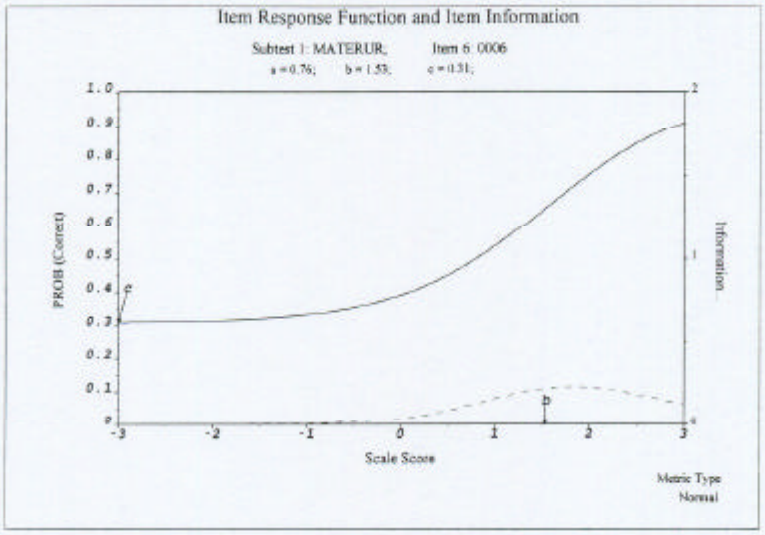


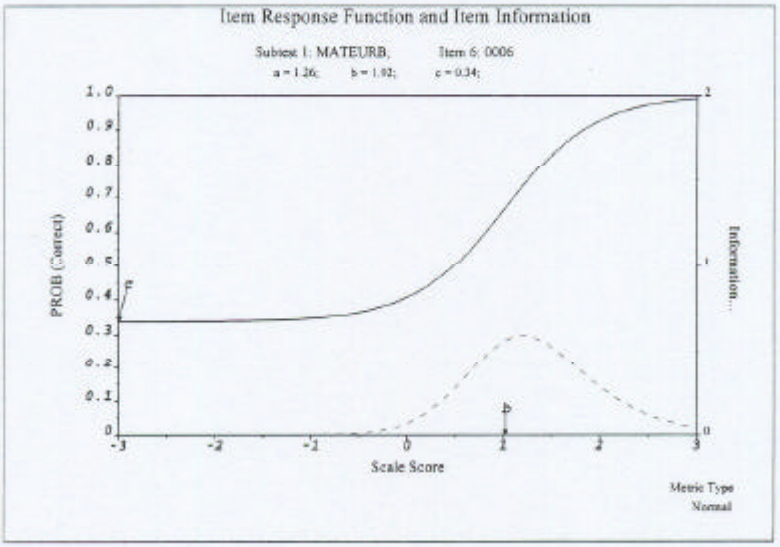
# Gráfico No.67
## Comparación entre Colegios Rurales y Colegios Urbanos
## Item 5



# Gráfico No.67
## Comparación entre Colegios Rurales y Colegios Urbanos
## Item 6

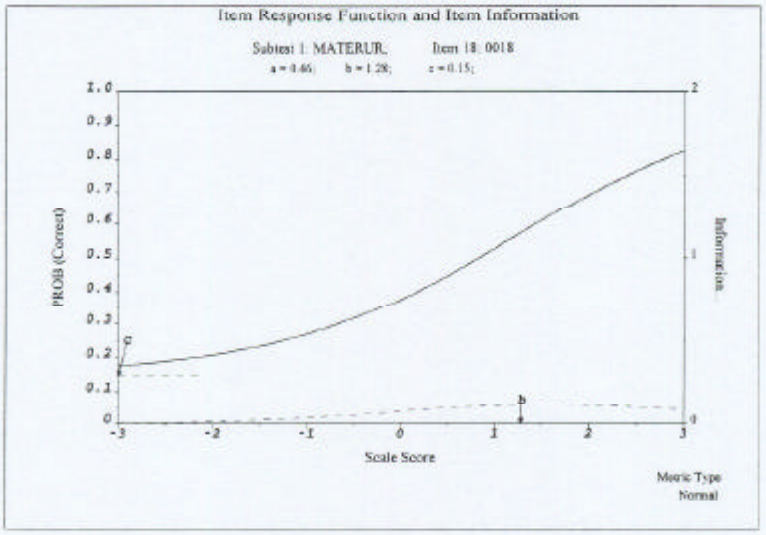


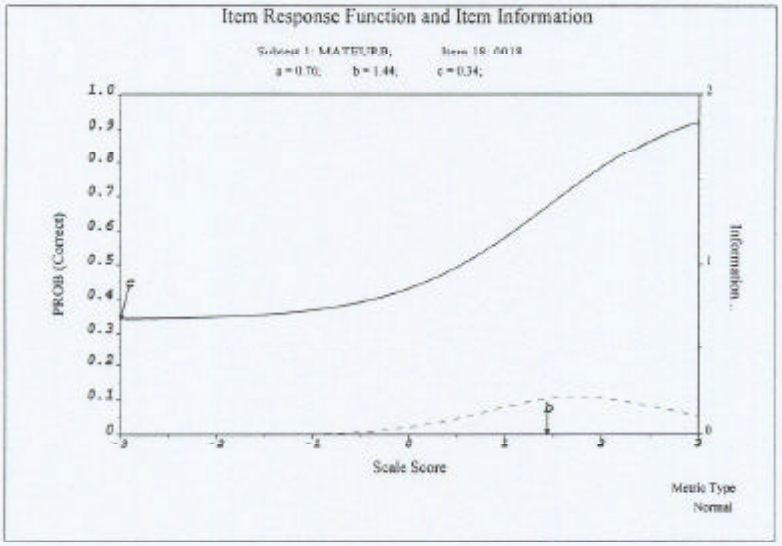



# Gráfico No.68

## Comparación entre Colegios Rurales y Colegios Urbanos

### Item 18

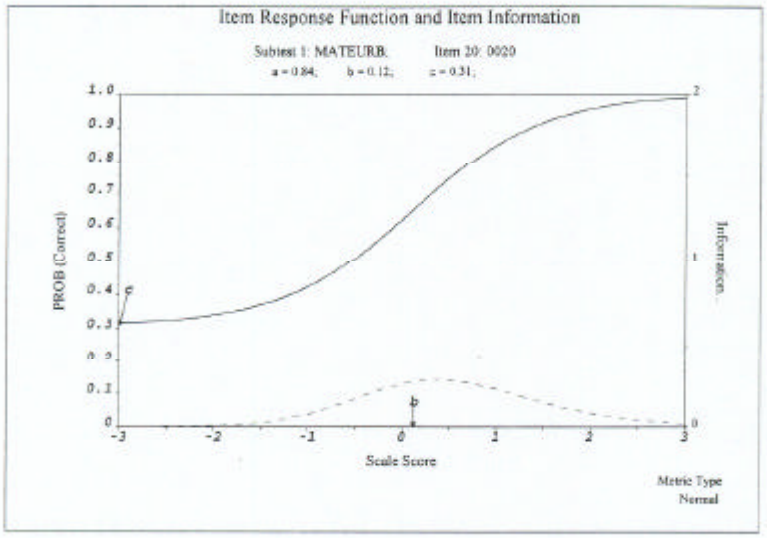


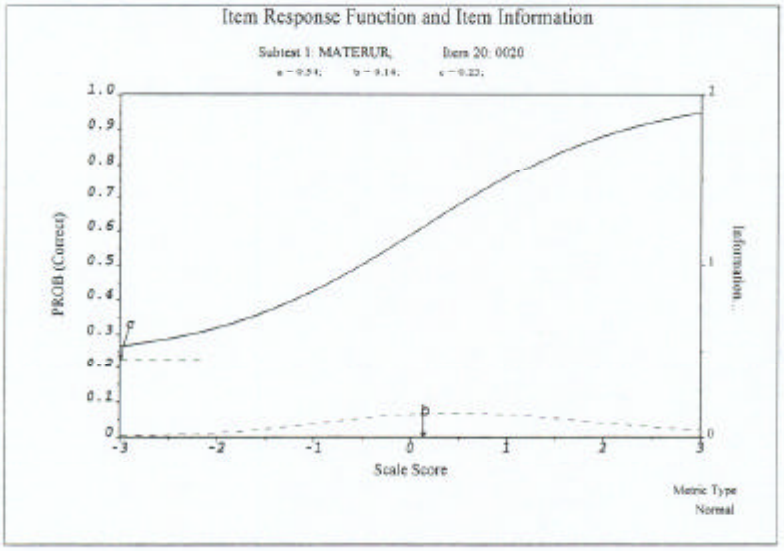



# Gráfico No.69

## Comparación entre Colegios Rurales y Colegios Urbanos*

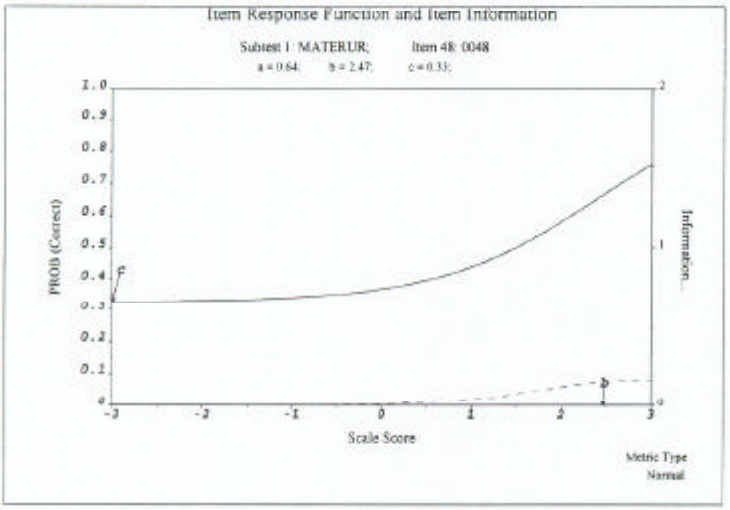
## Item 20

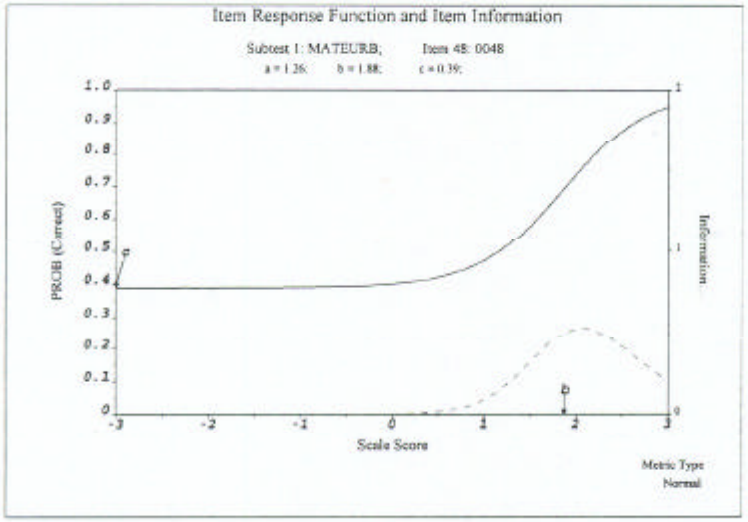


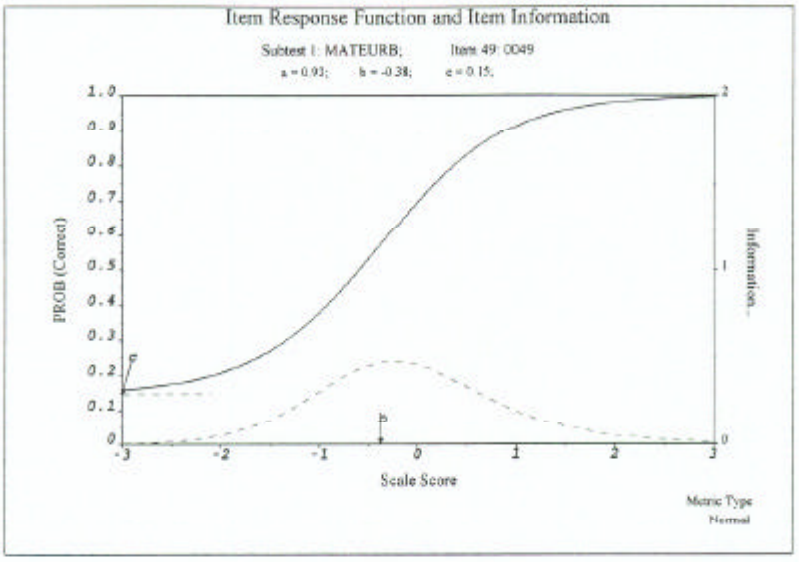

*Nótese que aquí el primer gráfico corresponde a los colegios urbanos

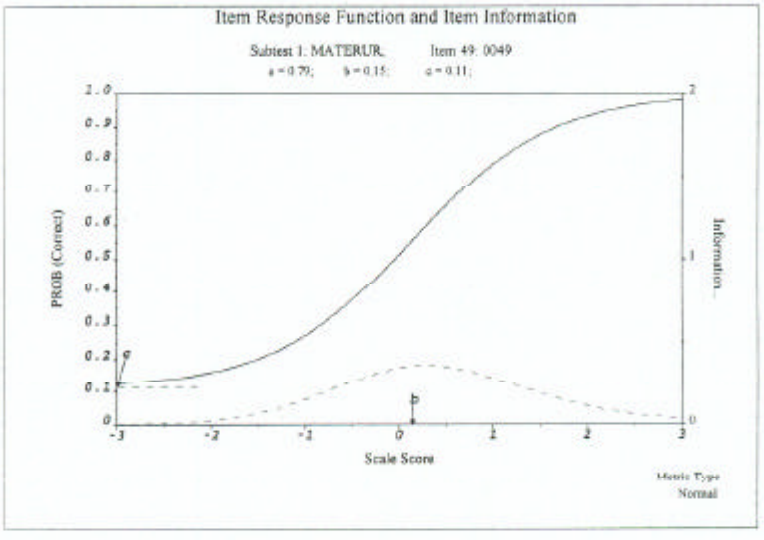
# Gráfico No.70
## Comparación entre Colegios Rurales y Colegios Urbanos
## Item 48



# Gráfico No.71
## Comparación entre Colegios Rurales y Colegios Urbanos*
## Item 49




*Nótese que aquí el primer gráfico corresponde a los colegios urbanos

## CONCLUSIONES

Del análisis realizado se puede concluir que aunque esta prueba, desde el punto de vista psicométrico, presenta en general niveles relativamente aceptables de calidad técnica, es aún susceptible de ser mejorada.

El análisis de factores ejecutado encontró evidencias de que la prueba es efectivamente de naturaleza unidimensional, es decir hay un solo factor o constructo subyacente que está siendo medido por los ítemes que la componen, que en este caso podría denominarse conocimientos y razonamiento matemático. Sin embargo, de los 50 ítemes analizados hay 4 que no parecen estar adecuadamente representados en ese factor y debían revisarse. Ellos son el 9, 11, 17 y 48 (éste último corresponde al 49 en la numeración del formulario de la prueba).

Por otra parte el análisis con la teoría clásica de los tests, el enfoque tradicional en el análisis psicométrico de pruebas estandarizadas, arrojó un valor para el Alfa de Cronbach de 0.8742. Esta es una medida de la confiabilidad o precisión de la prueba desde el punto de vista de su consistencia interna. Los autores en el campo de la medición recomiendan que si el instrumento se va a usar para la toma de decisiones Alfa debería ser al menos 0.90. Vemos así que en el presente caso el valor encontrado es ligeramente inferior al recomendado por los autores.

Siguiendo el enfoque de la teoría clásica se identificaron también aquellos ítemes que no cumplían con el estándar establecido en cuanto a poder discriminatorio. Si se utiliza el coeficiente de correlación entre el puntaje del ítem y el puntaje total de la prueba como medida de discriminación, se dice que los ítemes que arrojan para esta medida un valor de 0.30 o más tienen niveles aceptables de poder discriminatorio, es decir, capacidad para diferenciar entre estudiantes con puntajes altos y bajos. De acuerdo con esta regla se detectaron 17 ítemes que no cumplieron con lo recomendado. Sin embargo la mayoría de ellos presentan valores bastante cercanos al 0.30. Son los mismos 4 ítemes listados arriba que se identificaron con problemas con el análisis de factores, los que aquí presentan niveles muy bajos de discriminación y por tanto requieren una revisión profunda o reformulación. De hecho, con solo eliminar estos ítemes de la calificación de la prueba se lograría incrementar el valor de la medida de confiabilidad, Alfa de Cronbach y aumentar así la precisión del instrumento.

El análisis bajo la Teoría de Respuesta a los Ítemes hizo evidente que esta prueba está brindado mayor información en niveles relativamente altos de habilidad. Esto quiere decir que la prueba tiene mayor precisión o poder discriminatorio cuando se trata de diferenciar estudiantes en niveles altos de conocimientos y razonamiento matemático. Este resultado debería analizarse puesto que no se tiene claro si ésa sería la meta para un instrumento de esta naturaleza. Podría pensarse más bien que una prueba de este tipo debería brindar mayor información en los niveles intermedios de habilidad. Si la prueba no está maximizando la información en los niveles de habilidad que deseados de acuerdo con su naturaleza y propósitos, se podría entonces estar tomando decisiones equivocadas para esos grupos de estudiantes en donde la información no es óptima.

El análisis individual de los ítemes bajo la Teoría de Respuesta a los Ítemes pudo identificar cuatro ítemes con serios problemas de calidad técnica. Ellos son el 9, 17, 18 y 28 (éste último con el número 29 en la numeración del formulario). De éstos el 9 y 17 también habían sido identificados con problemas en el análisis clásico. De los 50 ítemes hubo 24 (48%) que exhibieron una alta o aceptable calidad técnica para discriminar en estudiantes de niveles intermedios de habilidad. Hubo 21 ítemes (42%) con alta o aceptable calidad para discriminar en niveles altos de habilidad. Sin embargo, la información que brindan los ítemes con poder discriminatorio en niveles altos es, en general, mayor que la que brindan aquellos que discriminan en niveles intermedios, es decir los primeros tienen usualmente mejor calidad técnica que los segundos. Por esta razón es que la prueba como un todo ofrece información máxima en niveles relativamente altos de habilidad.

El análisis de sesgo o comportamiento diferencial del ítem comparó estudiantes de colegios públicos versus estudiantes de colegios privados y estudiantes de zonas urbanas versus rurales. A nivel de la comparación por dependencia del colegio se puede concluir que, en general, de los ítemes que presentan evidencia de un posible sesgo, los más difíciles dan más información para los estudiantes de los colegios privados. Además, se nota que la mayoría de estos ítemes son comparativamente más difíciles para los estudiantes de colegios públicos. Debe recordarse que este análisis controla por niveles de habilidad. es decir. se está comparando la probabilidad de respuesta correcta en estudiantes con los mismos niveles de habilidad en colegios privados y públicos. Los ítemes analizados que se identificaron con posible sesgo en esta comparación fueron: 4,5,12, 18, 20, 29, 33, 42, 46,48 y 50 (debe recordarse que los ítemes 29, 33, 42, 46,48 y 50 corresponden en la numeración del formulario de la prueba a los números 30, 34, 43, 47, 49 y 51).

De igual manera al hacer la comparación entre colegios urbanos y rurales se encontraron 6 ítemes con evidencia de posible sesgo: 5, 6, 18, 20, 48 y 49 (los ítemes 48 y 49 corresponden en la numeración del formulario de la prueba a los números 49 y 50). De los resultados de este análisis se puede concluir que, en general, estos ítemes tienden a ser un poco más fáciles para los estudiantes de colegios urbanos, aunque la diferencia no es tan marcada como en el caso de la comparación entre públicos y privados. Sí se nota que brindan en general más información (tienen mayor poder de discriminación) en el caso de los estudiantes de colegios urbanos.

Dado que la existencia de un posible sesgo en un ítem representa una amenaza a la validez del instrumento de medición, los ítemes así detectados deben revisarse a profundidad para tratar de identificar posibles causas de ese comportamiento y así encontrar formas para evitarlo o minimizarlo.

## RECOMENDACIONES

La principal recomendación que se deriva de este estudio es la necesidad de que se comience a construir un banco de ítemes para la elaboración de esta prueba. La existencia de un banco de ítemes garantiza el logro de una prueba de alta calidad, puesto que los ítemes que se

escogen para constituir la prueba ya han sido analizados previamente y cumplen con los estándares psicométricos establecidos.

Aunado a lo anterior, se recomienda que los procedimientos descritos se incorporen como parte regular de la construcción y calificación de la prueba como un todo y de los ítemes que la componen. En especial, el uso de la Teoría de Respuesta a los Ítemes en el proceso de "calibración" de los ítemes puede resultar altamente beneficioso, pues permitirá no solo conseguir pruebas de más alta calidad técnica, si no que éstas brinden mayor información en los niveles de habilidad de los estudiantes en donde se debe discriminar con mayor precisión. Así, se estarán minimizando los errores a la hora de decidir si un estudiante merece o no promoverse y por tanto se estarán tomando decisiones más justas.

También es importante que se continué estudiando el sesgo en los ítemes y sus posibles causas. El análisis ejecutado aquí representa solo una primera aproximación a esta temática que es por cierto una de las más controversiales en el campo de la medición, pero también una de las más relevantes desde el punto de vista de la equidad de las pruebas.

Una recomendación que se puede implementar de manera casi inmediata es la que se refiere a la incorporación de la variable "Sexo del Estudiante" a la base de datos del MEP en donde se registran los resultados de las pruebas, puesto que el formulario de examen sí la incluye. De esta forma también se podría ejecutar el análisis de sesgo comparando hombres y mujeres.