

# Estudio comparativo de diversos programas de calificación y análisis de reactivos objetivos

Agustín Tristán López  
Agosto 2007

## CONTENIDO

	Página
Resumen	2
1. Presentación	2
2. Generalidades de la prueba utilizada	3
3. Programas elegidos para este análisis	4
4. Comparación de parámetros estadísticos y de análisis	19
5. Comparación del dictamen de reactivos de la prueba analizada	24
6. Conclusiones	35
Programas de referencia	
Referencias	

# Estudio comparativo de diversos programas de calificación y análisis de reactivos objetivos

Agustin Tristán López (\*)

27 de agosto de 2007

## Resumen

Se analiza una prueba real con ayuda de seis programas comerciales diferentes, empleando ocho modelos de análisis y dictamen de reactivos objetivos. Las comparaciones se enfocan tanto a las hipótesis involucradas en cada modelo como a los parámetros y valores de referencia utilizados en los programas elegidos. La comparación se lleva a cabo incluyendo elementos objetivos (dificultad, medias de aciertos, confiabilidad, validez de escala), con objeto de identificar diferencias entre los modelos. Se observa que se trata de modelos con tendencias generales muy parecidas en distintos grados de rigor para juzgar los reactivos: Las ventajas de un programa respecto a los otros se ubican en los tipos de reporte emitidos y la facilidad de uso

## 1. Presentación

Hay una gran variedad de programas para calificación y análisis de reactivos objetivos, disponibles para el uso de los evaluadores, profesores y especialistas de la psicometría. Dada la diversidad de los programas y de los modelos utilizados por cada uno de ellos, el analista se enfrenta a una complicación para poder comparar los valores y dictámenes que producen. Los evaluadores acostumbran hacer comparaciones con aplicaciones particulares, empleando muestras de sustentantes configuradas de manera específica para comprobar la bondad de un modelo en particular o para refutar la calidad de algún programa. Las comparaciones realizadas en estas condiciones no permiten identificar los patrones de dictamen que tienen los programas porque no se contemplan las hipótesis de cada modelo, ni los parámetros de decisión asociados a un programa dado. En consecuencia es costumbre que se llegue a descartar un cierto programa por cuestiones subjetivas o de inclinación personal y no tanto por una evaluación objetiva que compare los diversos modelos y parámetros involucrados.

La comparación que se presenta en este estudio se enfoca de manera objetiva a revisar tanto las hipótesis de los modelos como los valores emitidos por cada programa. Para este trabajo se eligieron ocho modelos diferentes disponibles en seis programas comerciales, dos de ellos de origen mexicano y cuatro producidos en los Estados Unidos. Se escogieron estos programas por su popularidad en diversos ambientes de evaluación y porque se manejan generalmente como referencia. Se utilizan los resultados de una prueba real, aplicada con un grupo piloto en condiciones igualmente reales, con el propósito de identificar tendencias de comportamiento entre los modelos psicométricos y los programas.

El trabajo está organizado en estas partes:

- a) generalidades de la prueba utilizada
- b) descripción de los programas elegidos, modelos y parámetros de referencia
- c) comparación de parámetros estadísticos y de análisis
- d) comparación del dictamen de los reactivos de la prueba utilizada
- e) conclusiones

---

\* Director General. Instituto de Evaluación e Ingeniería Avanzada, S.C.  
Mariano Jiménez 1830 A. Col. Balcones del Valle. 78280 San Luis Potosí, México. Tel. (444) 820 37 88, Fax (444) 815 48 48.  
Correo electrónico: [ici1@prodigy.net.mx](mailto:ici1@prodigy.net.mx). Página web: [www.ieesa-kalt.com](http://www.ieesa-kalt.com)

## 2. Generalidades de la prueba utilizada

La prueba elegida para este análisis es el Examen de Conocimientos Generales que forma parte del primer módulo de evaluación contenido en el Examen de Certificación de Profesores de Educación Media Superior (ECPEMS-Derechos reservados, Instituto de Evaluación e Ingeniería Avanzada, S.C.). Es parte del proceso de evaluación que incluye varios instrumentos (pruebas objetivas, portafolios de evidencias, observación de desempeño, cuestionario de opinión). El Manual Técnico describe el perfil evaluado, tablas de especificaciones, tiempos, etc. (disponible en el IEIA para instituciones contratantes).

El Examen de Conocimientos y Habilidades Generales explora una dimensión genérica que debe mostrar un maestro, independientemente del área de especialidad que imparta (adicionalmente se dispone de pruebas particulares por especialidad). La prueba se divide en ocho áreas:

Área	Temario general	Núm. de reactivos	Núm. de opciones
Matemáticas	Aritmética / Álgebra / Geometría / Trigonometría / Geometría Analítica	25	5
Español	Sintaxis / Ortografía / Lectura de comprensión / Oraciones	20	5
Razonamiento	Análisis de relaciones / Lectura de comprensión / Series numéricas / Razonamiento numérico, espacial e instrumental / Relaciones numéricas / Silogismos / Analogías / Series gráficas	31	5
Metodología	Consulta en biblioteca e Internet / Elementos del proceso de investigación / Procesamiento de información / Presentación de documentos y reportes / Técnicas de estudio y de aprendizaje	25	5
Mundo actual	México contemporáneo (política, economía, educación, ciencia y tecnología) / Mundo contemporáneo (política, educación, ciencia y tecnología)	25	5
Inglés	Estructura gramatical / Conjugación de verbos / Lectura de comprensión (preguntas en español preguntas en inglés)	25	4
Computación	Sistemas operativos / Procesador de textos / Hoja de cálculo / Base de datos / Presentadores	24	5
Tecnología educativa	Manejo de contenidos / Programas de estudio / Métodos de enseñanza / Materiales educativos / Evaluación del aprendizaje	25	5
TOTAL DE REACTIVOS		200	

El proceso que se sigue en el ECPEMS para calibrar, calificar y proporcionar la retroalimentación a los maestros y a la institución consiste de estos pasos:

- § Se aplica la prueba a los profesores inscritos al proceso de evaluación, quienes disponen por lo menos de un mes natural para revisar la Guía de Estudio y prepararse debidamente para la prueba.
- § Una vez respondida la prueba se califica y se calibran los ítems, siguiendo los criterios de análisis de Kalt (Plus y Criterial). Se escalan los resultados de los sustentantes con el modelo de Rasch (Winsteps) y las medidas en lógitos expresan en la escala de 70 a 130 puntos.
- § Se preparan los reportes técnico, de análisis de ítems y estadístico por cada área de la prueba.
- § Se emite la Cédula de Retroalimentación y el Certificado para cada profesor.
- § Se producen los reportes de entrega para la institución contratante.

La “Cédula de Retroalimentación” para el maestro es un reporte individualizado que presenta los resultados por área y recomendaciones para superar los subtemas más deficientes. El “Reporte Institucional” incluye un comparativo de resultados por área y en forma global, además de un dictamen de planeación para un programa de formación continua para los profesores, indicando temas principales y personas que requieren ser inscritas en dicha formación continua. Junto con lo anterior, la Institución recibe la “Guía de Interpretación” donde se explican los resultados y su interpretación.

El análisis que se realiza en este trabajo emplea 220 personas de una aplicación realizada en noviembre de 2001.

### 3. Programas elegidos para este análisis

#### 3.1 Descripción de los programas

Los programas que se comparan en este trabajo son los siguientes:

1. Producto:	<b>KALT PLUS</b>
Proveedor:	IEIA (México)
Ambiente:	MSDOS
Teoría empleada:	Clásica - paradigmática
Características Principales:	Dictamina los reactivos por medio del Grado de Dificultad (GD, proporción de aciertos) y el Poder de discriminación (PD, Diferencia de aciertos entre grupos superior e inferior, identificados por medio de la mediana, existe un control para emplear las colas de la distribución de sustentantes si se desea). Para aceptar un reactivo se cuenta con una Norma Discriminativa de exigencia constante ( $ND=0.3GD$ ). Se rechazan los reactivos en estos casos (1) PD es negativa o nula, (2) GD está cerca de los extremos (0, 100) con discriminación baja.
Reportes:	Estadísticas (descriptivas, histograma, curva de frecuencias acumuladas), análisis de reactivos (detallado, dictamen en palabras, análisis de distractores, análisis de azar), reporte técnico (confiabilidad de la prueba y sus partes, diagramas de dificultad y discriminación, ajuste a recta de diseño), reportes a la institución (individual, global descendente, global alfabético).
2. Producto:	<b>KALT CRITERIAL</b>
Proveedor:	IEIA (México)
Ambiente:	WINDOWS
Teoría empleada:	Clásica y Teoría de la respuesta al Ítem
Características Principales:	Dictamina los reactivos por medio de dos modelos: (a) Clásico con Grado de Dificultad (GD, proporción de aciertos) y correlación Punto-Biserial (rpbis). Para aceptar un reactivo se cuenta con una Norma constante (exigencia decreciente) que el usuario puede cambiar ( $ND=0.2$ ); (b) logístico con ayuda de un modelo biparamétrico que hace intervenir los subgrupos alto y bajo en función del criterio de corte elegido por el usuario. Para dictaminar el modelo logístico proporciona los valores de ajuste (r de Pearson y $\chi^2$ ) y calcula la dificultad y la discriminación logísticas; se rechazan los reactivos en dos casos. (1) cuando tienen un ajuste bajo o malo y (2) cuando se tiene discriminación nula o negativa o medida indefinida.
Reportes:	Estadísticas (descriptivas, histograma, curva de frecuencias acumuladas), análisis de reactivos (detallado, ajuste al modelo logístico; dictamen en palabras, análisis de distractores, análisis de azar), reporte técnico (confiabilidad de la prueba y sus partes, diagramas de dificultad y discriminación, ajuste a recta de diseño), reportes a la institución (individual, global descendente, global alfabético), manejo de bases de datos y generador de reportes.
3. Producto:	<b>WINSTEPS (BIGSTEPS)<sup>1</sup></b>
Proveedor:	MESA (U.Chicago, EUA)
Ambiente:	Windows (MSDOS)
Teoría empleada:	Rasch
Características Principales:	Dictamina los reactivos por medio del modelo de Rasch. No cuenta con valores de referencia propios, pero se acostumbra emplear los valores de ajuste de MNSQ(INFIT/OUTFIT) que debe estar en el intervalo (0.8-1.2) y Z(INFIT/OUTFIT) que debe estar por abajo de 2. No hay un límite de aceptación para las medidas de los reactivos, basta con que ajusten convenientemente. Un segundo dictamen se emite con ayuda de la correlación punto-biserial, que solo se pide que sea positiva, independientemente de su valor.
Reportes:	Análisis de reactivos (ajuste al modelo logístico), reporte técnico (confiabilidad de la prueba, diagramas de dificultad, curva característica de la prueba), archivos tipo base de datos de sujetos y de ítems

<sup>1</sup> Para el análisis se empleó Winsteps, pero sin perder generalidad se habla de Bigsteps que es la versión gratuita del programa.

<b>4. Producto:</b>	<b>Iteman</b>
<b>Proveedor:</b>	ASC (EUA)
<b>Ambiente:</b>	MSDOS
<b>Teoría empleada:</b>	Clásica
Características Principales:	Dictamina los reactivos por medio de la dificultad clásica (p) considerando todos los sujetos incluidas las omisiones. Calcula la discriminación basada en los grupos extremos (27%), junto con las correlaciones biserial o punto biserial, que el usuario elige. No hay un límite de aceptación para las medidas de los reactivos, para este trabajo se considera rechazado un reactivo que tenga correlación punto biserial negativa o nula.
Reportes:	Estadísticas (descriptivas, histograma), análisis de reactivos (detallado, análisis de distractores)

<b>5. Producto:</b>	<b>Rascal</b>
<b>Proveedor:</b>	ASC (EUA)
<b>Ambiente:</b>	MSDOS
<b>Teoría empleada:</b>	Rasch
Características Principales:	Dictamina los reactivos por medio del modelo de Rasch. No cuenta con valores de referencia propios, pero proporciona el valor de $\chi^2$ y los grados de libertad para que se dictamine el ajuste al modelo (Para este estudio reporta 19 grados de libertad, con lo cual $\chi^2$ debe ser inferior a 30 para dictaminar ajuste con un 95% de confianza). No hay un límite de aceptación para las medidas de los reactivos, basta con que ajusten convenientemente.
Reportes:	Análisis de reactivos (ajuste al modelo logístico), reporte técnico (confiabilidad de la prueba, diagramas de dificultad, curva característica de la prueba).

<b>6. Producto:</b>	<b>Xcalibre</b>
<b>Proveedor:</b>	ASC (EUA)
<b>Ambiente:</b>	MSDOS (Windows)
<b>Teoría empleada:</b>	Teoría de la Respuesta al Ítem (2 y 3 parámetros)
Características Principales:	Dictamina los reactivos por medio de modelo logístico biparamétrico, ajusta el modelo con un esquema de máxima verosimilitud marginal. Determina los parámetros "a", "b" y el residuo de ajuste. En el modelo de 2 parámetros (ó 2PL) el valor "c" es nulo, mientras que se debe determinar en el modelo de 3 parámetros (ó 3PL). Dictamina los reactivos con las letras "R", que indica que no ajusta el modelo a los datos y "P" que indica que hay problemas potenciales en el reactivo. El valor de referencia para aceptar reactivos es que el residuo sea menor que 2. No hay un límite de aceptación para las medidas de los reactivos, basta con que ajusten convenientemente. Se utiliza como criterio de rechazo la presencia de las letras "R" y "P" en el reporte de reactivos.
Reportes:	Análisis de reactivos (ajuste al modelo logístico), reporte técnico (confiabilidad de la prueba, diagramas de dificultad, curva característica de la prueba, gráfica de la función de información).

Los únicos programas que incluyen bases de datos exportables directamente a otros programas (EXCEL, ACCESS, SPSS, etc.) son KALT Plus, KALT Criterial y Winsteps. Los demás programas requieren de traductores especiales o un trabajo adicional para leer los archivos en otros ambientes.

Sólo Kalt Criterial incluye un manejador de la base de datos de los sustentantes y un generador de reportes en ambiente Windows.

Se pueden comparar algunas características generales de estos programas, en función de las hipótesis de sus modelos.

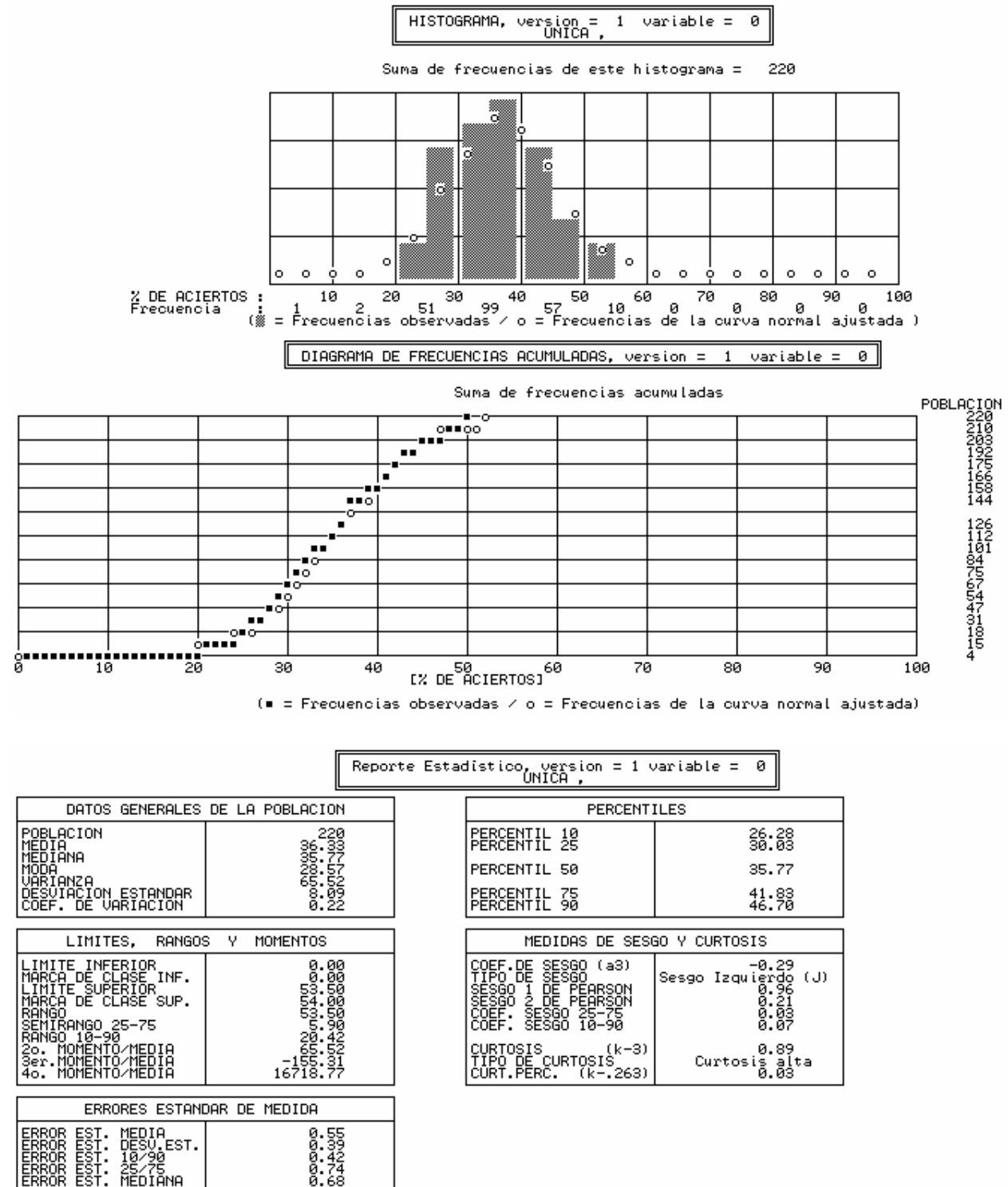
	Kalt Criterial "KC"	Winsteps "B"	Iteman "IT"	Rascal "R"	Xcalibre "3P"	Xcalibre "2P"
Kalt Plus "KN"	Igual en dificultad o medida del reactivo KN=KC	Igual en dificultad o medida del reactivo KN=B	Dificultad diferente KN>IT Discriminación diferente Dis27%>Dis 50%	/	/	/
Kalt Criterial "KC"		Medida: 3P>R>B>2P>KC Correlación Punto Biserial: 3P=2P>B>KC	/	$\chi^2$ R (19 gdl) > KC(8 gdl)	Ajuste:3P debería ajustar mejor que 2P (Valor de residuo < 2.0) B y KC utilizan otro modelo. Se comparan <u>contra 1.0</u>	
		Bigsteps"B" (Winsteps)	/	Medida: R>B en 7% Valor absoluto medio R>B en 63%	Medida: 3P>R>B>2P>KC  Correlación Punto Biserial: 3P=2P>B>KC	
			Iteman "IT"	/	/	/
				Rascal "R"	Medida: 3P>R>2P . Ajuste:3P debería ajustar mejor que 2P (Valor de residuo < 2.0)	
					Xcalibre "3P"	Medida: 3P>R>2P Ajuste:3P debería ajustar mejor que 2P (Valor de residuo < 2.0)

### 3.2 Comparación de las salidas de los programas

La comparación entre los programas no es inmediata porque no se tienen reportes estandarizados por alguna norma nacional o internacional. A continuación se presentan partes distintivas de los reportes que emiten los programas. Dada la cantidad de información contenida en algunos de estos programas, en particular KALT y Winsteps, se eligieron algunos reportes de ejemplo. Los programas Iteman, Rascal y Xcalibre proporcionan una menor cantidad de información y por ello se incluyen ejemplos de todos los elementos incluidos en el listado.

### 3.2.1 Kalt Plus

#### Reporte estadístico



Análisis de reactivos

REACTIVO 1 Respuesta correcta: D ----- VERSION 1 TEMA 1

		A	B	C	D	E	Omis.	Error	Total	Válida			R.C.	R.I.	Válida	
G.S.	a	3	4	41	46	6	10	0	110	100	a b c d e	G.S.	46	54	100	
	b	1.36	1.81	18.63	20.90	2.72	4.54	0.00	50.00	45.45			20.90	24.54	45.45	
	c	6	7	45	31	11				100			31	69	100	
	d	11	11	11	57	11							57	43		
	e	x	x	Tx	R	x							R	Txu		
G.I.	a	8	8	41	11	14	28	0	110	82	a b c d e	G.I.	11	71	82	
	b	3.63	3.63	18.63	5.00	6.36	12.72	0.00	50.00	37.37			5.00	32.27	37.27	
	c	5	5	37	26	9				82			26	56	82	
	d	20	20	20	0	20							0	82		
	e	U	U	u	Sx	Uu							Sx	Uu		
TOTAL		11	12	82	57	20	38	0	220	182	a b	TOTAL		57	125	182
		5.00	5.45	37.27	25.90	9.09	17.27	0.00	100.00	82.72				25.90	56.81	82.72

DIAGRAMA DE RESPUESTAS POR QUINTILES

	0	25	50	75	100	z	CASOS	TOTAL DE CASOS
20						1.6	3	3
40	■					3.8	7	10
60						2.7	5	15
80	■					7.7	14	29
100	■					15.4	28	57

■ = VALORES OBSERVADOS  
 ° = VALORES ESPERADOS

RESPUESTAS VALIDAS	N	182
GRADO DE DIFICULTAD	GD	31.31868
DICTAMEN SOBRE DIFICULTAD		
DIFICULTAD CORREGIDA	cp	14.14835
PODER DE DISCRIMINACION	PD	19.23077
NORMA DISCRIMINATIVA	ND	9.39561
RELACION DISCRIMINATIVA	PD/ND	2.04678
DICTAMEN SOBRE DISCRIMINACION		
DICTAMEN DEL REACTIVO		BIEN
DATOS INDEPENDIENTES	$\chi^2$	22.24055
NIVEL DE SIGNIFICACION	P( $\chi^2$ )	0.00000
GRADOS DE LIBERTAD	NU	1
COEFICIENTE DE CONTINGENCIA	C	0.32999
MAXIMO DE CONTINGENCIA	Máx C	0.70711
RELACION COEF/MAXIMO (%)	C/Máx C	46.66774
CORRELACION DE ATRIBUTOS	r <sup>2</sup>	0.34957
DEPENDIENTE SOBRE GD	$\chi^2$	29.24442
NIVEL DE SIGNIFICACION	P( $\chi^2$ )	0.00000
MAXIMA DEPENDENCIA	$\chi^2$ MAX	34.82553
FATL	z	34.28077
PHI CORRELACION	PHI $\hat{\varphi}$	0.34957
VALOR MAXIMO DE PHI	Máx $\hat{\varphi}$	0.67528
RELACION	$\hat{\varphi}$ /Máx $\hat{\varphi}$	0.51767
NIVEL DE SIGNIFICACION	P( $\hat{\varphi}$ )	0.00000
G INDICE DE RELACION	G	0.28571
G NORMALIZADA	z(G)	3.85450
NIVEL DE SIGNIFICACION	P(G)	0.99993

Resumen del análisis de reactivos

*REACTIVO	GRADO DE DIFICULTAD	DIFICULTAD MODIFICADA	PODER DE DISCRIMINACION	NORMA DISCRIMINATIVA	RELACION DISCRIMINATIVA	DICTAMEN SOBRE DIFICULTAD	DICTAMEN SOBRE DISCRIMINACION	RESUMEN DEL DICTAMEN
1	31.32	14.15	9.34	9.40	0.99		< NORMA	A REVISAR
2	90.09	87.62	7.08	7.92	0.89	MUY FACIL	< NORMA	DESECHABLE
3	38.17	22.71	16.79	11.45	1.47			BIEN
4	40.00	25.00	21.18	12.00	1.76			BIEN
5	70.53	63.16	19.32	21.16	0.91		< NORMA	A REVISAR



## Reporte técnico

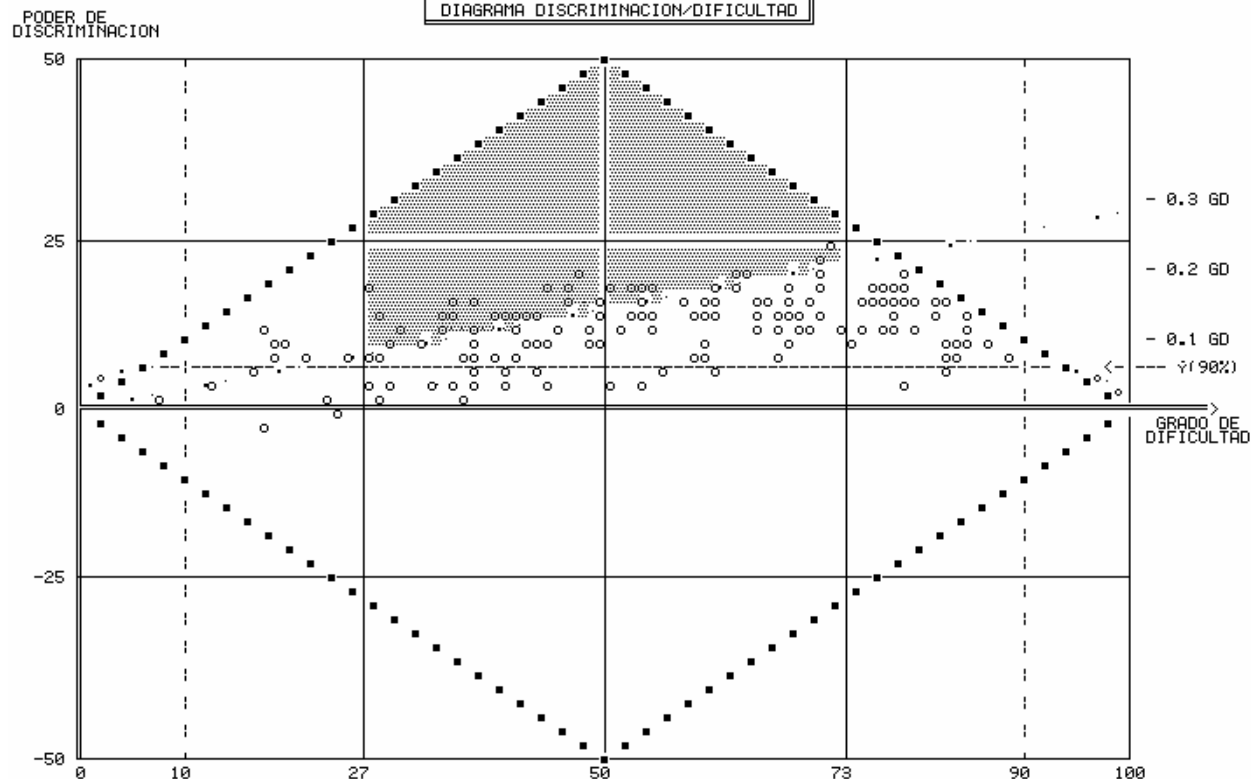
### RESULTADOS DE CALIDAD DEL CUESTIONARIO

Número de reactivos : 200 Reactivos con discriminación positiva: 186

GRADO DE DIFICULTAD (%)	PODER DE DISCRIMINACION (%)
MEDIA: 54.05	MEDIA: 10.32
POSITIVO PROMEDIO: 55.49	POSITIVO PROMEDIO: 11.20
POSITIVO MINIMO: 10.10	POSITIVO MINIMO: 0.46
POSITIVO MAXIMO: 99.54	POSITIVO MAXIMO: 26.42

CONFIABILIDAD DEL INSTRUMENTO ( $\alpha$ ) : 0.46496

### DIAGRAMA DISCRIMINACION/DIFICULTAD



## Resumen del dictamen de los reactivos

REACTIVOS 'ACEPTABLES' = 50									
3	4	7	12	14	17	24	34		
193	194								
REACTIVOS 'A REVISAR' = 1									
1	5	8	9	10	11	15	16		
179	184	187	189	198	199	200			
REACTIVOS 'DUDOSOS' = 11									
66	68	69	114	118	133	168	180		
190	196	197							
REACTIVOS 'MUY DUDOSOS' = 0									
REACTIVOS 'DESECHABLES' = 43									
2	6	13	25	27	29	33	36		
188	191	195							

Resumen del reporte técnico

## DATOS GENERALES

ALUMNOS EXAMINADOS	220
NÚMERO DE REACTIVOS (sin eliminar cancelados)	200
NÚMERO DE TEMAS	8
PORCENTAJE GLOBAL MÁS ALTO ALCANZADO POR UN ALUMNO	69.00
PORCENTAJE GLOBAL MÁS BAJO ALCANZADO POR UN ALUMNO	0.00
MÁXIMO DE ACIERTOS ALCANZADO POR UN ALUMNO	138
MÍNIMO DE ACIERTOS ALCANZADO POR UN ALUMNO	0

Resultados por tema	Número de Reactivos	Reactivos Discr>0	Min. de Aciertos	Máx. de Aciertos	Dificultad Media	Confiabilidad Alfa Cronbach
Global	200	186	0	138	54.05	0.917
MAT. MATEMATICAS	25	22	0	14	64.70	0.907
ESP. ESPAÑOL	25	22	0	14	64.70	0.907
RAZ. RAZONAMIENTO	25	22	0	14	64.70	0.907
METO. METODOLOGIA	25	22	0	14	64.70	0.907
M.A. MUNDO ACTUAL	25	22	0	14	64.70	0.907
ING. INGLES	25	22	0	14	64.70	0.907
COMP. COMPUTACION	25	22	0	14	64.70	0.907
T.E. TECNOLOGIA EDUCATIVA	25	22	0	14	64.70	0.907

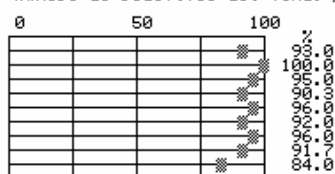
## DIAGRAMA DE GRADO DE DIFICULTAD POR TEMA

If = % más baja ■ = Media Global % más alta = If  
0 = Media de Reactivos con Discriminación >0



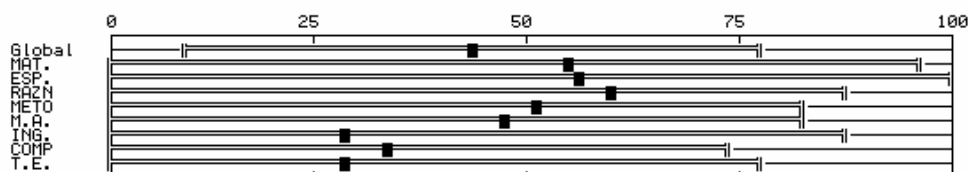
## EFICIENCIA DEL INSTRUMENTO

Porcentaje de reactivos que discriminan > 0, respecto al número de reactivos del tema: ■



## DIAGRAMA DE PORCENTAJE DE ACIERTOS

If = % más bajo ■ = Media % más alto = If

Reporte global alfabético para la institución

ASIGNATURA INSTITUTO AMERICA CHIAPAS

SEGUNDO SEMESTRE

PAG. 1

## RESULTADOS GENERALES DE LOS SUSTENTANTES

Numero	CONTROL	NOMBRE DEL ALUMNO	ACIERTOS %	RELATIVA % / MAX	PERCENTIL	Z	Z>Z	TEMAS % ALG. TRI.
1	10295-01	ALANIS GUTIERREZ ANSELMO	66.66	74.51	89.26	1.38	8.34	73.3 59.2
2	10237-01	BELTRAN NUNEZ JORGE	54.38	60.78	71.78	0.47	31.98	63.3 44.4
3	10269-01	CARDENAS SOTO LUIS	35.08	39.21	16.50	-0.97	83.36	36.7 33.3
4	10239-01	COTA MARTINEZ LAURA	26.31	29.41	4.95	-1.62	94.75	16.7 37.0
5	10209-01	HERNANDEZ SUAREZ ELMO	68.42	76.47	92.08	1.51	6.51	83.3 51.8
6	10327-01	LOPEZ LOPEZ JOSE	40.00	44.71	30.20	-0.60	72.65	31.4 48.6
7	10023-01	MARTINEZ LOPEZ MARTIN	71.92	80.38	94.51	1.77	3.80	83.3 59.2
8	10243-01	PEREZ MARQUEZ EDNA	61.40	68.63	85.15	0.99	16.09	66.7 55.5

Se dispone también de un reporte descendente para la institución. Estos reportes son configurables por el usuario.

Reporte individual

Opcionalmente se entrega una página por sustentante. El reporte es configurable por el usuario.

### 3.2.2 Reporte de Kalt Criterial

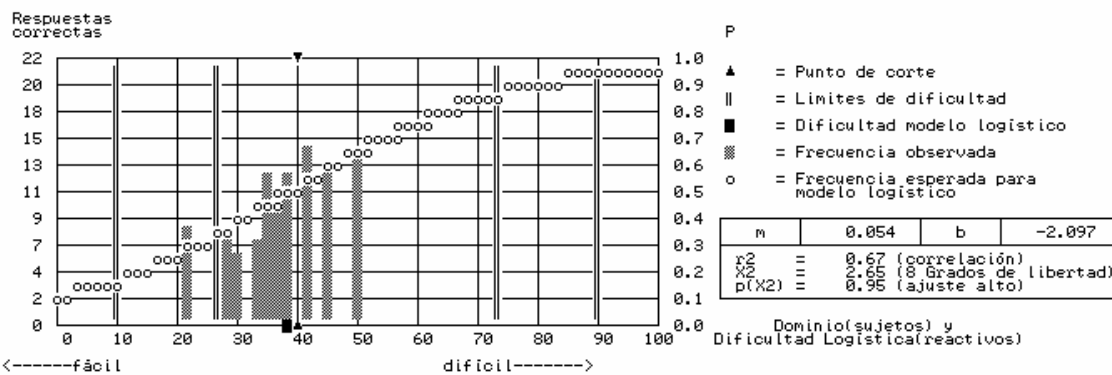
#### Análisis estadístico

Idéntico al que presenta Kalt Normal, ver 3.2.1

#### Análisis de reactivo

Reactivo = 200 Respuesta correcta = B Total de respuestas válidas = 169

	A	B	C	D	E	Omis.	Error	Total	Válida		R.C.	R.I.	Válida
G.S.	0.00	19.42	2.27	3.64	1.36	5.91	0.00	32.27	26.58	G.S.	24.85	9.47	34.58
	Tx	R	T	Tx	Tx						R	Tx	
G.I.	6.36	28.62	3.18	8.18	4.55	17.27	0.00	67.73	50.45	G.I.	36.69	28.99	65.68
	U	Sxz	Ux	U	U						Sxz	U	
TOTAL	14	104	12	11.82	5.91	23.18	0.00	220	169	TOTAL	104	65	169
	6.36	47.27	5.45	11.82	5.91	23.18	0.00	100.00	76.82		61.54	38.46	100.00



#### INDICADORES DE CALIDAD DEL REACTIVO. MODELO DE KALT-CRITERIAL

Criterio de corte = 40.00 %  
 Grado de Dificultad modelo clásico = 61.54 %  
 Grado de Facilidad (100-Grado clásico) = 38.46 %  
 Dificultad modelo logístico = 38.46 %  
 equivalente en Grado de Facilidad (fallas) = 61.54 %  
 Dictamen sobre la dificultad = INTERMEDIO  
 Discriminación criterial = 0.52  
 Dictamen sobre la discriminación = MAS QUE .2  
 DICTAMEN GLOBAL DEL REACTIVO = BUENO

#### ANALISIS CRITERIAL POR SUBGRUPOS SEGUN EL PUNTO DE CORTE

	Observado	Esperado
Log(nomio) Grupo Inferior	0.54	0.35
Probabilidad de respuesta del Grupo Inferior	0.56	0.41
Respuestas en el Grupo Inferior	62.00	46.00
Log(nomio) Grupo Superior	0.37	0.39
Probabilidad de respuesta del Grupo Superior	0.72	0.60
Respuestas en el Grupo Superior	42.00	34.00
Diferencia de probabilidades $p(GS)-p(GI) \times 100$	13.00 %	18.00 %
Diferencia relativa a Dificultad Logística	13.00 %	18.00 %
Diferencia relativa al Punto de Corte	41.39 %	45.30 %

#### Reporte condensado de análisis de reactivos

REACTIVO	Respuesta Correcta	Sujetos	GRADO DE DIFICULTAD CLASICO	CORRELACION PUNTO BISERIAL	GRAFICA DIFICULTAD Dif-->Fácil	GRAFICA P. BISERIAL Bajo-->Alto	DICTAMEN/REACTIVO
1	0	18	31	0.27			ACEPTABLE
2	0	13	18	0.27			ACEPTABLE
3	0	13	18	0.27			ACEPTABLE
4	0	13	18	0.27			ACEPTABLE
5	0	13	18	0.27			ACEPTABLE
6	0	13	18	0.27			ACEPTABLE
7	0	13	18	0.27			ACEPTABLE
8	0	13	18	0.27			ACEPTABLE
9	0	13	18	0.27			ACEPTABLE
10	0	13	18	0.27			ACEPTABLE
11	0	13	18	0.27			ACEPTABLE
12	0	13	18	0.27			ACEPTABLE
13	0	13	18	0.27			ACEPTABLE
14	0	13	18	0.27			ACEPTABLE
15	0	13	18	0.27			ACEPTABLE
16	0	13	18	0.27			ACEPTABLE
17	0	13	18	0.27			ACEPTABLE
18	0	13	18	0.27			ACEPTABLE
19	0	13	18	0.27			ACEPTABLE
20	0	13	18	0.27			ACEPTABLE
21	0	13	18	0.27			ACEPTABLE
22	0	13	18	0.27			ACEPTABLE
23	0	13	18	0.27			ACEPTABLE
24	0	13	18	0.27			ACEPTABLE
25	0	13	18	0.27			ACEPTABLE
26	0	13	18	0.27			ACEPTABLE
27	0	13	18	0.27			ACEPTABLE
28	0	13	18	0.27			ACEPTABLE
29	0	13	18	0.27			ACEPTABLE
30	0	13	18	0.27			ACEPTABLE
31	0	13	18	0.27			ACEPTABLE
32	0	13	18	0.27			ACEPTABLE
33	0	13	18	0.27			ACEPTABLE
34	0	13	18	0.27			ACEPTABLE
35	0	13	18	0.27			ACEPTABLE
36	0	13	18	0.27			ACEPTABLE
37	0	13	18	0.27			ACEPTABLE
38	0	13	18	0.27			ACEPTABLE
39	0	13	18	0.27			ACEPTABLE
40	0	13	18	0.27			ACEPTABLE
41	0	13	18	0.27			ACEPTABLE
42	0	13	18	0.27			ACEPTABLE
43	0	13	18	0.27			ACEPTABLE
44	0	13	18	0.27			ACEPTABLE
45	0	13	18	0.27			ACEPTABLE
46	0	13	18	0.27			ACEPTABLE
47	0	13	18	0.27			ACEPTABLE
48	0	13	18	0.27			ACEPTABLE
49	0	13	18	0.27			ACEPTABLE
50	0	13	18	0.27			ACEPTABLE
51	0	13	18	0.27			ACEPTABLE
52	0	13	18	0.27			ACEPTABLE
53	0	13	18	0.27			ACEPTABLE
54	0	13	18	0.27			ACEPTABLE
55	0	13	18	0.27			ACEPTABLE
56	0	13	18	0.27			ACEPTABLE
57	0	13	18	0.27			ACEPTABLE
58	0	13	18	0.27			ACEPTABLE
59	0	13	18	0.27			ACEPTABLE
60	0	13	18	0.27			ACEPTABLE
61	0	13	18	0.27			ACEPTABLE
62	0	13	18	0.27			ACEPTABLE
63	0	13	18	0.27			ACEPTABLE
64	0	13	18	0.27			ACEPTABLE
65	0	13	18	0.27			ACEPTABLE
66	0	13	18	0.27			ACEPTABLE
67	0	13	18	0.27			ACEPTABLE
68	0	13	18	0.27			ACEPTABLE
69	0	13	18	0.27			ACEPTABLE
70	0	13	18	0.27			ACEPTABLE
71	0	13	18	0.27			ACEPTABLE
72	0	13	18	0.27			ACEPTABLE
73	0	13	18	0.27			ACEPTABLE
74	0	13	18	0.27			ACEPTABLE
75	0	13	18	0.27			ACEPTABLE
76	0	13	18	0.27			ACEPTABLE
77	0	13	18	0.27			ACEPTABLE
78	0	13	18	0.27			ACEPTABLE
79	0	13	18	0.27			ACEPTABLE
80	0	13	18	0.27			ACEPTABLE
81	0	13	18	0.27			ACEPTABLE
82	0	13	18	0.27			ACEPTABLE
83	0	13	18	0.27			ACEPTABLE
84	0	13	18	0.27			ACEPTABLE
85	0	13	18	0.27			ACEPTABLE
86	0	13	18	0.27			ACEPTABLE
87	0	13	18	0.27			ACEPTABLE
88	0	13	18	0.27			ACEPTABLE
89	0	13	18	0.27			ACEPTABLE
90	0	13	18	0.27			ACEPTABLE
91	0	13	18	0.27			ACEPTABLE
92	0	13	18	0.27			ACEPTABLE
93	0	13	18	0.27			ACEPTABLE
94	0	13	18	0.27			ACEPTABLE
95	0	13	18	0.27			ACEPTABLE
96	0	13	18	0.27			ACEPTABLE
97	0	13	18	0.27			ACEPTABLE
98	0	13	18	0.27			ACEPTABLE
99	0	13	18	0.27			ACEPTABLE
100	0	13	18	0.27			ACEPTABLE

## Reporte técnico

Número de sustentantes = 220  
 Número de versiones = 1  
 Número de variables = 8  
 Número de reactivos en el instrumento = 200  
 Número de reactivos originales en esta variable = 200  
 Número de reactivos efectivos en esta variable = 200  
 Reactivos cancelados (sin clave de respuesta) = 0

Clasificación de los reactivos por Grado de Dificultad y PuntoBiserial					
(*) BUENOS	ACEPTABLES	A REVISAR	DUDOSOS	DESECHABLES	Totales
80	50	37	16	17	200

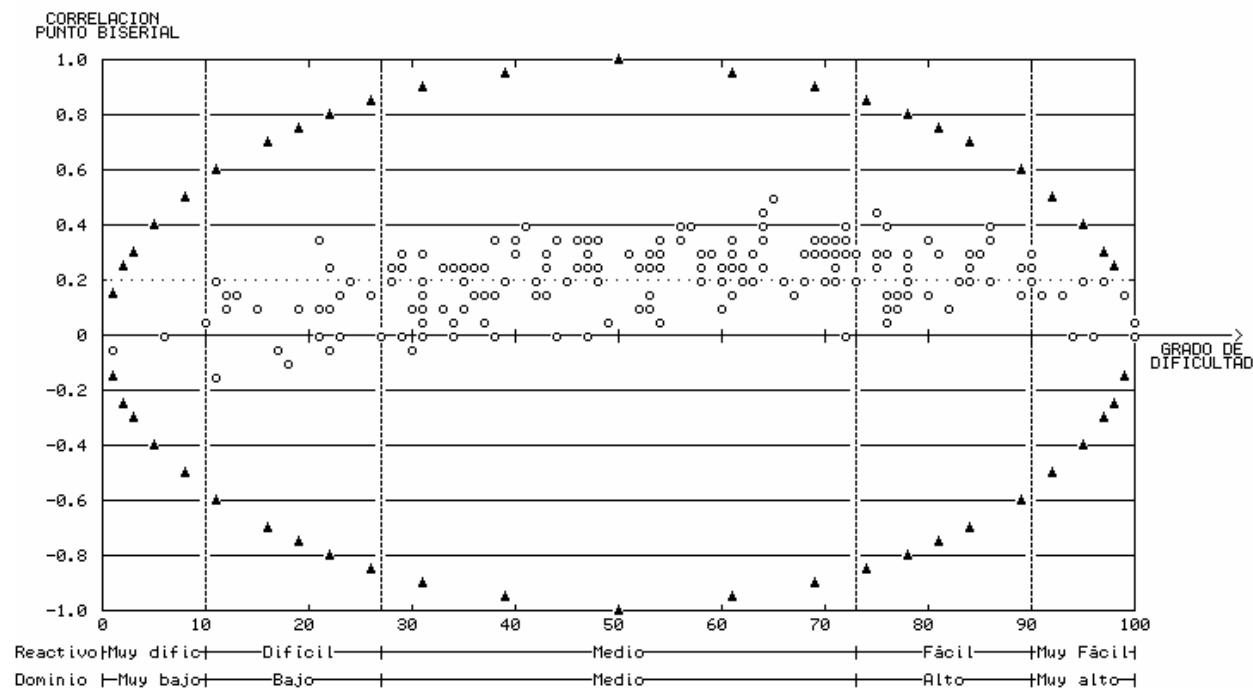
	% aciertos	Reactivos originales	Reactivos efectivos
Media de la población	36.33	72.66	72.66
Desviación Estándar	8.09	16.19	16.19

## ANÁLISIS CRITERIAL DE CONFIABILIDAD DEL INSTRUMENTO

Punto de corte propuesto = 40.00 %  
 = 80.00 reactivos  
 Porcentaje de sustentantes arriba del punto de corte = 33.00 %  
 Reactivos en esta variable = 200  
 Confiabilidad referida a norma Alfa = 0.85  
 Confiabilidad Criterial (fórmula de Livingsgton) Rcr = 0.88  
 Factor de longitud de la prueba para obtener Rcr0.90 = 1.24  
 Número de reactivos requerido para obtener Rcr0.90 = 248  
 Factor de longitud de la prueba para obtener Rcr0.95 = 2.62  
 Número de reactivos requerido para obtener Rcr0.95 = 524  
 Error Estándar en la medida SE = 2.80 %  
 = 5.6 reactivos

Correlac. Punto Biserial.	CALIFICACION DE LOS REACTIVOS					Totales
	Muy difícil	Difícil	Medio	Fácil	Muy Fácil	
< 0.00	2	5	4	0	2	13
< 0.20	0	10	48	17	0	65
> 0.20	0	12	77	18	0	99
Totales	2	22	129	35	12	200
	Muy bajo	Bajo	Medio	Alto	Muy alto	
	NIVEL DE DOMINIO DE LOS SUSTENTANTES					

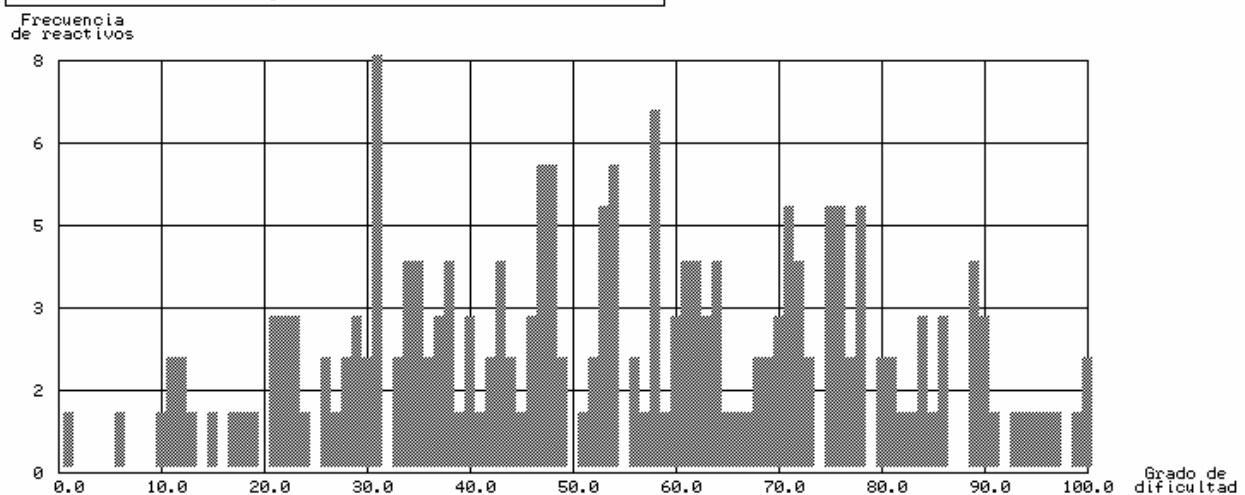
## DIAGRAMA PUNTO BISERIAL/DIFICULTAD



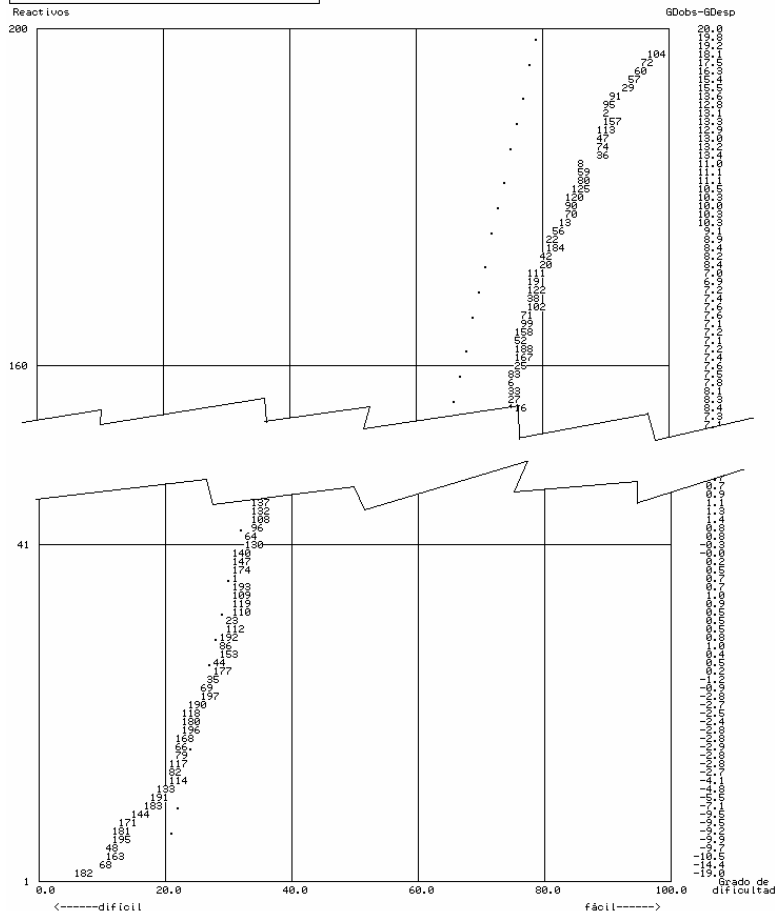
NOTAS:  
 o = Reactivos del cuestionario  
 ▲ = Límite del dominio Dificultad-Punto Biserial

## Distribución de reactivos

Revisión del diseño. Diagrama de frecuencias de los reactivos



Revisión del diseño. Distribución de reactivos



Análisis de validez de escala con respecto de la Recta de Diseño 20-80

Este reporte se presenta en forma parcial para fines de ilustración. El programa proporciona este mapa para cada área analizada, aquí se presenta la gráfica de la prueba completa.

NOTAS: o = reactivos del cuestionario La prueba tiende al lado fácil

Ajuste de recta a datos observados				Diferencia (GObvs-GDesp)	
Límite inferior	14.0	Pendiente	2.53	Absoluta media	5.5
Límite superior	84.0	Ordenada	96.4	Cuadrática media	48.2
Dificultad media	54.0	R <sup>2</sup>	0.989		

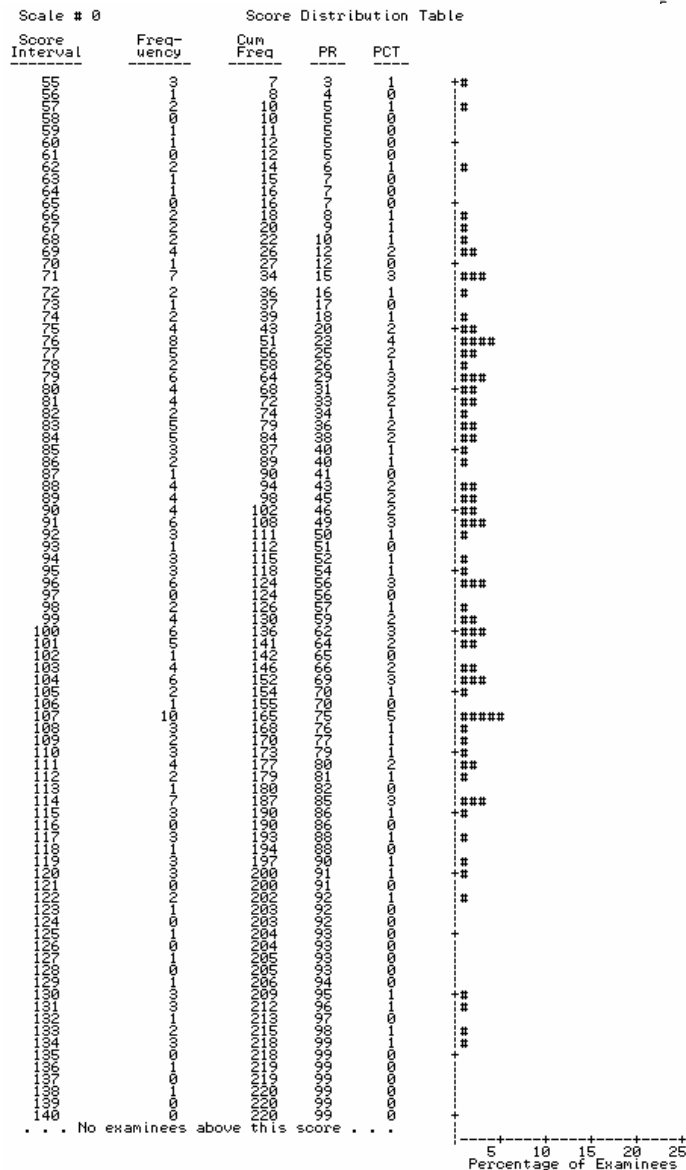
## Reportes globales e individuales

El programa contiene funciones de manejo de base de datos y generación de reportes en ambiente Windows. La presentación es abierta y configurable por el usuario.

### 3.2.3 Reportes de Iteman

#### Análisis de reactivos

Seq. No.	Scale Item	Item Statistics			Alternative Statistics					
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser.	Key
1	0-1	.26	.29	.28	Correct	.05	.09	.00	-.14	
					Incorrect	.05	.06	.00	-.05	
					Other	.05	.06	.00	-.01	*
2	0-2	.87	.22	.30	Correct	.07	.07	.00	.30	
					Incorrect	.05	.05	.00	.03	
					Other	.04	.00	.00	.22	*
3	0-3	.23	.20	.22	Correct	.07	.13	.05	-.08	
					Incorrect	.05	.05	.00	-.02	
					Other	.40	.00	.00	.02	*
4	0-4	.31	.34	.36	Correct	.01	.16	.40	.26	
					Incorrect	.05	.05	.00	.14	
					Other	.03	.00	.00	.26	*
5	0-5	.66	.38	.38	Correct	.03	.03	.00	-.11	
					Incorrect	.05	.05	.00	.06	*
					Other	.05	.00	.00	.06	
6	0-6	.67	.46	.40	Correct	.05	.40	.00	.48	
					Incorrect	.05	.05	.00	.11	*
					Other	.10	.00	.00	.38	



#### Reporte técnico y estadístico

There were 220 examinees in the data file.

#### Scale Statistics

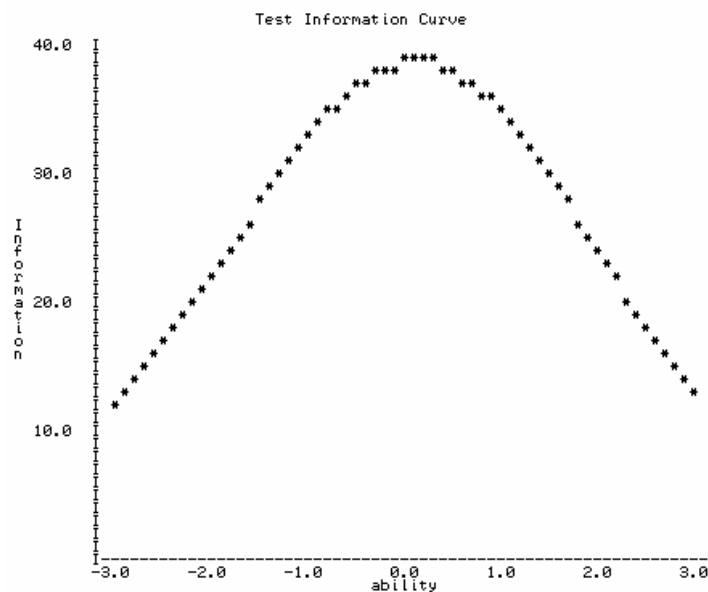
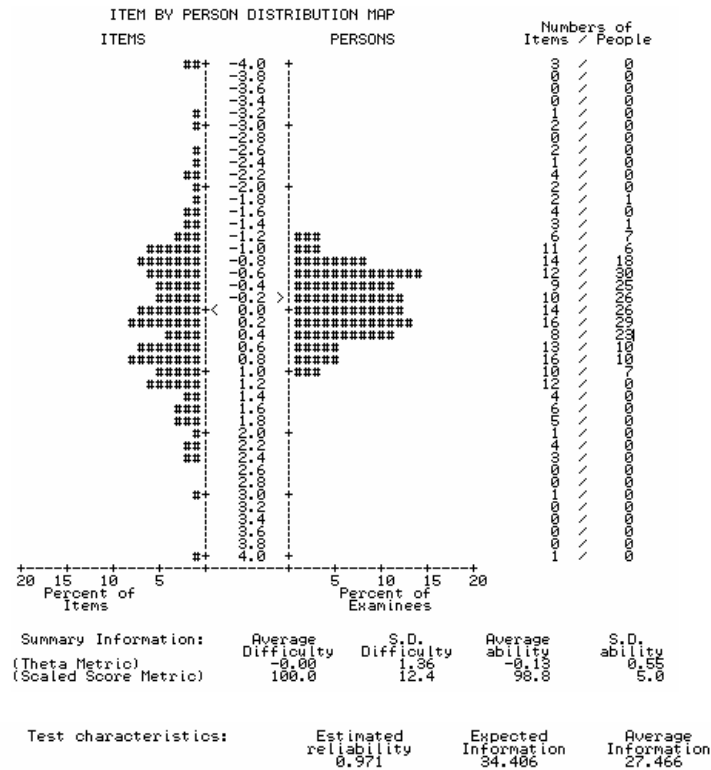
Scale:	0
N of Items	200
N of Examinees	220
Mean	33.045
Variance	451.352
Std. Dev.	21.245
Skew	-0.474
Kurtosis	0.13
Minimum	0.000
Maximum	138.000
Median	92.000
Alpha	0.918
SEM	0.094
Mean P	0.455
Mean Item-Tot.	0.034
Mean Biserial.	0.319
Max Score (Low)	79
N (Low Group)	64
Min Score (High)	107
N (High Group)	65

### 3.2.4 Reporte de Rascal

#### Análisis de reactivos

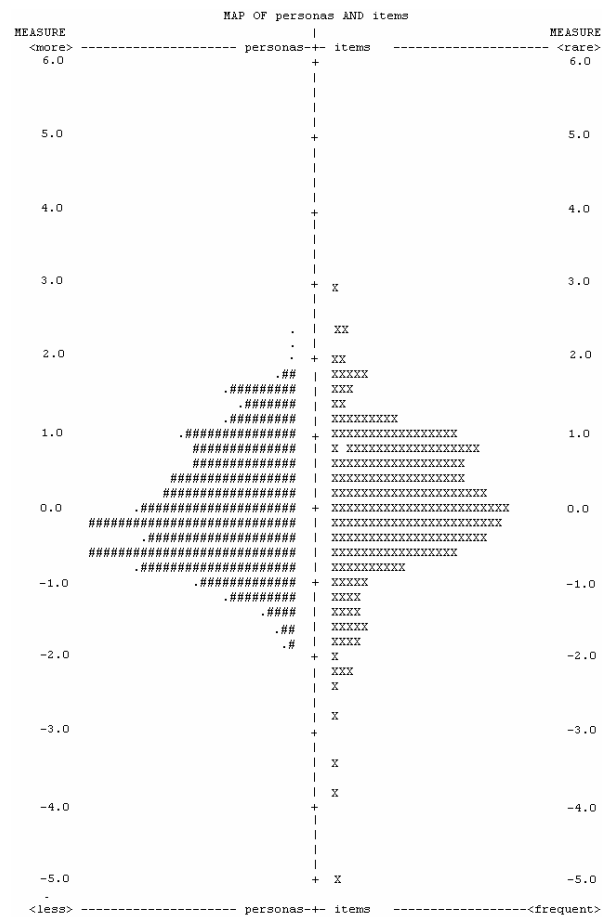
Item	Difficulty	Std. Error	Chi Sq.	df	Scaled Diff
1	0.979	0.158	25.108	109	109
2	-2.148	0.205	11.074	109	80
3	1.161	0.165	5.383	109	111
4	0.719	0.150	8.060	109	107
5	-0.863	0.148	30.660	109	90
6	-0.910	0.149	30.660	109	90

#### Reporte estadístico y técnico

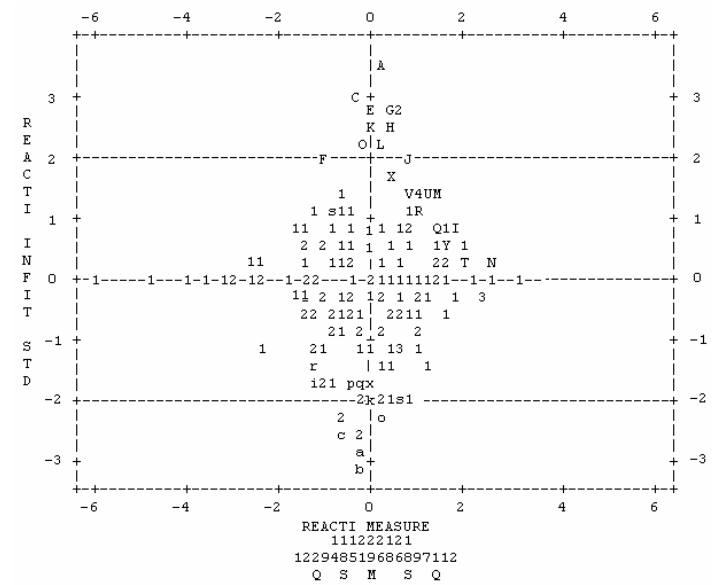


### 3.2.5 Reporte de Winsteps

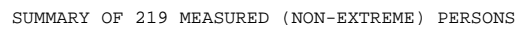
#### Reporte estadístico y técnico



#### Mapa de ajuste de los ítems







LACKING RESPONSES: 1 PERSONS

### SUMMARY OF 199 MEASURED (NON-EXTREME) REACTIS

MINIMUM EXTREME SCORE: 1 REACTIS

ENTRY NUM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTBIS	PERSON
					MNSQ	STD	MNSQ	STD		
137	68	199	-.82	.17	1.08	1.0	2.77	7.4	A .35	0143010801501
77	89	152	.32	.18	.99	-.2	2.73	6.4	B .46	0078031700801
102	53	199	-1.27	.18	1.46	4.4	2.07	4.0	C .05	0105020601001
87	54	128	-.64	.20	1.20	2.3	1.79	3.1	D .28	0089011900901
88	48	127	-.87	.21	1.31	3.2	1.60	3.0	E .19	0090031700901

ENTRY NUM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTBIS	PERSON
					MNSQ	STD	MNSQ	STD		
1	57	182	1.10	.17	.97	-.4	.98	-.3	.26	I0001
2	191	212	-2.08	.23	.95	-.3	.84	-.7	.25	I0002
3	50	131	.59	.19	.98	-.3	.97	-.4	.23	I0003
4	68	170	.68	.16	.92	-1.6	.90	-1.7	.38	I0004
5	146	207	-.68	.16	.96	-.6	.92	-1.0	.30	I0005

### Análisis de reactivos

[illegible][illegible]

Test characteristics:

K-R 21	Expected	Average
Reliability	Information	Information
0.917	28.334	24.818

Test Characteristic Curve

Test Information Curve

## 4. Comparación de parámetros estadísticos y de análisis

El primer conjunto de comparaciones se hace con relación a los parámetros estadísticos y de análisis en general, sin entrar al detalle reactivo por reactivo. Estas comparaciones muestran las diferencias asociadas a los modelos que utilizan los programas y, en consecuencia, resulta relativamente evidente el conjunto de valores mostrados, en los cuales se pueden apreciar tendencias y posiciones relativas que se obtienen por las hipótesis involucradas en los modelos.

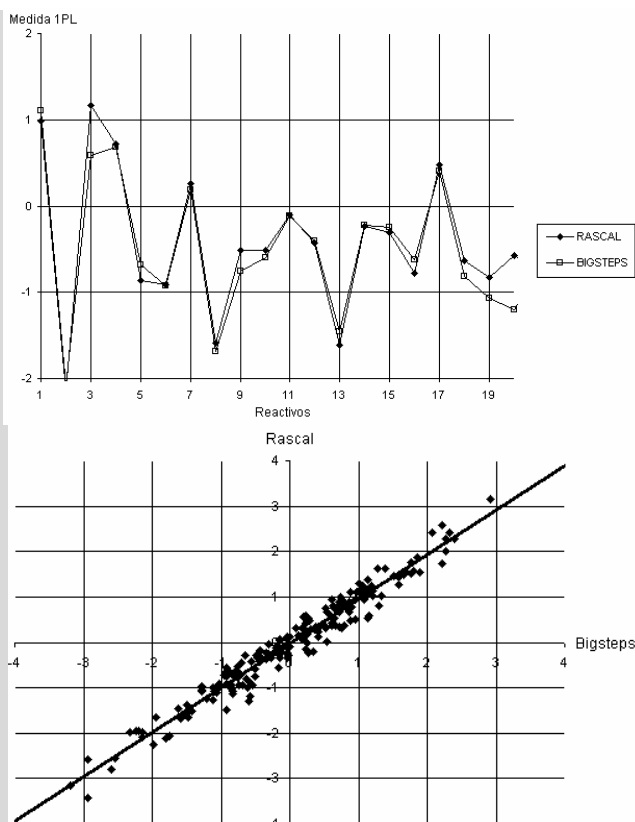
Se hicieron las comparaciones utilizando todos los reactivos de la prueba, pero solo se presentan gráficas de algunos conjuntos de reactivos con objeto de ilustrar las conclusiones indicadas en cada caso.

### 4.1 Medida

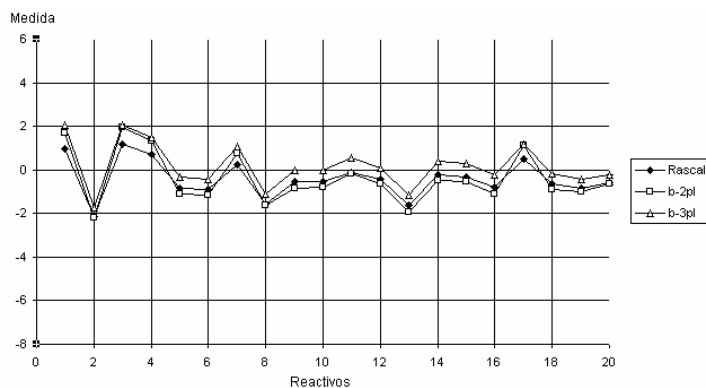
El modelo de Rasch es utilizado en Winsteps y Rascal. En general la medida de Rascal es mayor que la que proporciona Winsteps en un 7% en promedio.

Debido a que la escala de valores está expresada en lógitos, la diferencia absoluta media entre ambos modelos no es pequeña, alcanza el 63%.

A pesar de que se trata del mismo modelo, la falta de coincidencia depende del algoritmo utilizado para el cálculo de la medida y del ajuste (máxima verosimilitud, verosimilitud marginal, etc.)

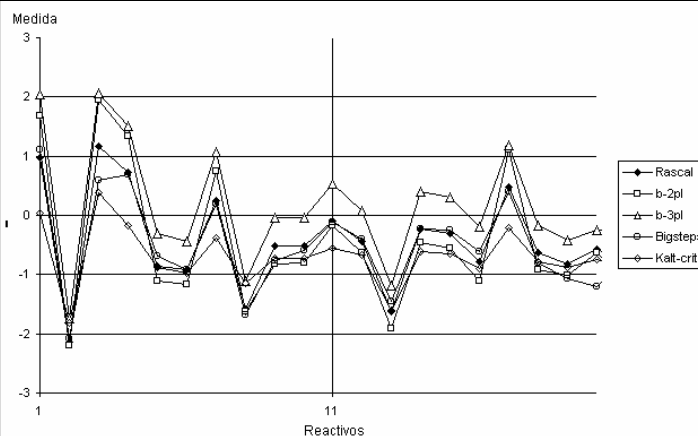


Si se comparan las medidas que obtienen los programas del mismo proveedor, se aprecia que el programa que utiliza el modelo de tres parámetros proporciona los valores más altos y el de dos parámetros los valores más bajos. Rascal (modelo de un parámetro para el mismo proveedor), en cambio, proporciona valores intermedios.



Comparación de medidas de los diferentes modelos. En general los valores guardan una misma tendencia general, ordenados en esta secuencia:

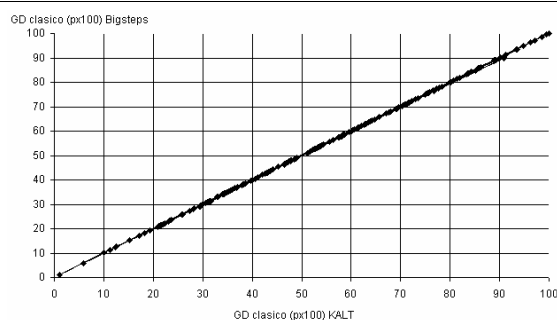
3 parámetros = Rasch = 2 parámetros



#### 4.2 Grado de dificultad (modelo clásico)

El Grado de Dificultad clásico (función de la proporción de aciertos  $GD=p \times 100$ ) coincide entre los programas KALT y Winsteps.

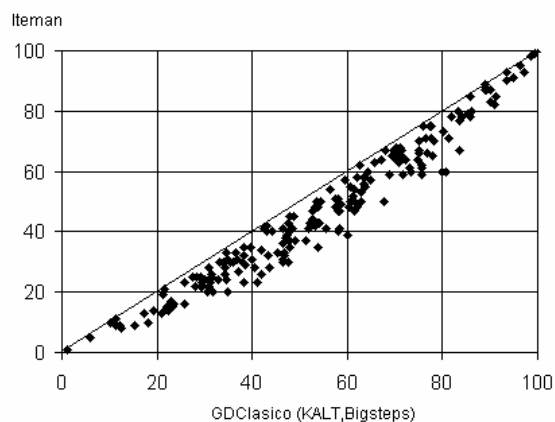
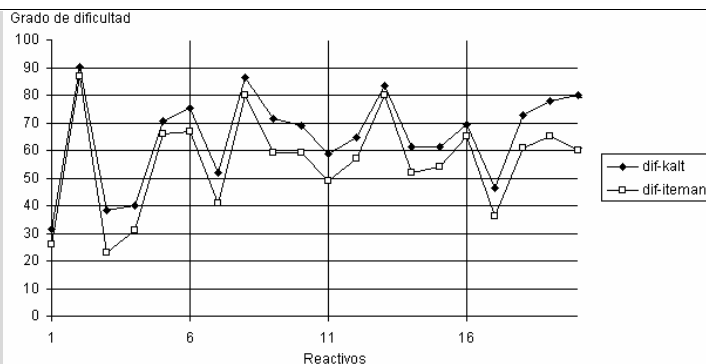
Este resultado corresponde con lo esperado ya que se trata de un mismo modelo para calcular el valor del estimador más probable de las respuestas correctas de cada reactivo.



Iteman calcula la dificultad en función del número total de sujetos, eliminando las respuestas erróneas (doble lectura), pero no cancela las omisiones.

El Grado de Dificultad clásico se calcula sobre respuestas efectivas, porque no es posible modelar como respuestas incorrectas a las omisiones o al llenado erróneo de un campo de respuesta.

Por lo anterior, la dificultad de Iteman siempre es menor o igual al Grado de Dificultad clásico y se ubican por debajo de la línea que corresponde con la igualdad entre dificultades.



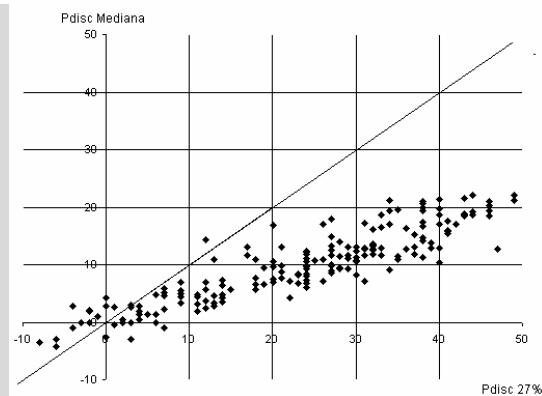
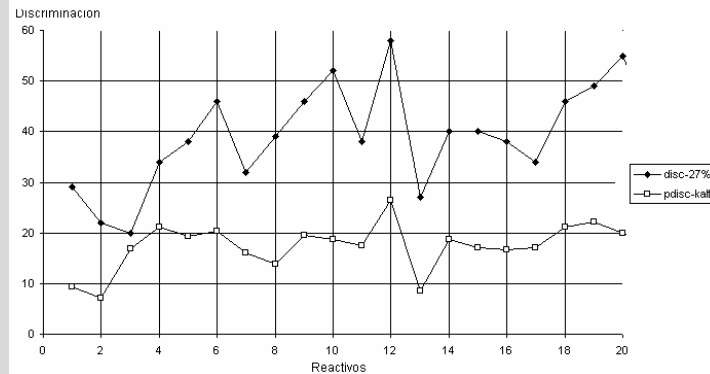
### 4.3 Discriminación

La discriminación clásica, como diferencia de grupos (superior e inferior) se calcula en Itean y Kalt Plus.

Itean calcula la discriminación usando las colas de la distribución de los sujetos (27% superior e inferior)

KALT usa la definición del poder de discriminación con dos subgrupos divididos en la mediana (50% superior e inferior)

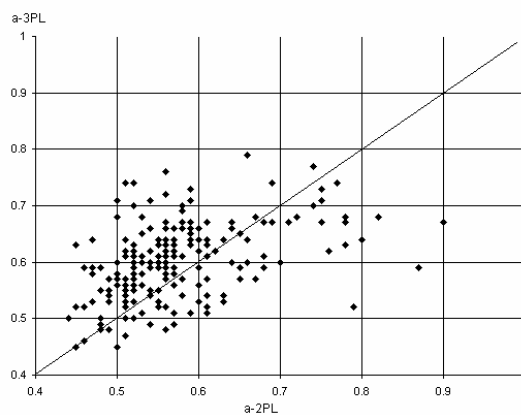
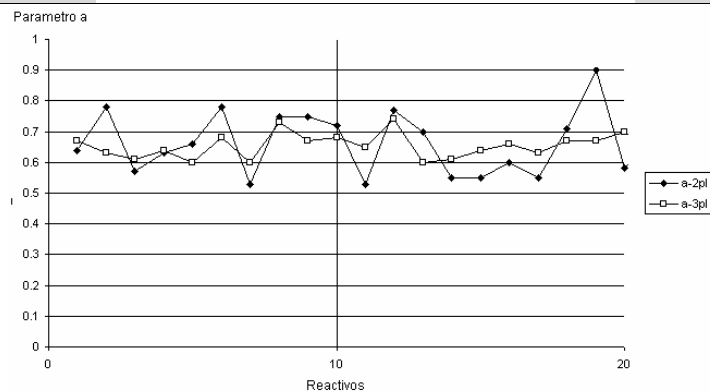
En general la discriminación con las colas (benévola) es superior a la discriminación con la mediana (riguroso), lo cual se aprecia en la figura superior. En la comparación de la figura inferior, se ve que la tendencia es siempre más alta usando las colas (benévola para el evaluador).



La discriminación dentro de la Teoría de la Respuesta al Ítem se determina con la pendiente en el punto de inflexión de la curva característica del ítem y se asocia con el parámetro "a".

En general, el parámetro "a" es ligeramente menor en el modelo de 3 parámetros (del orden del 3% inferior) respecto al valor obtenido en el modelo de 2 parámetros, en especial para valores superiores a 0.6 (en 2 PL).

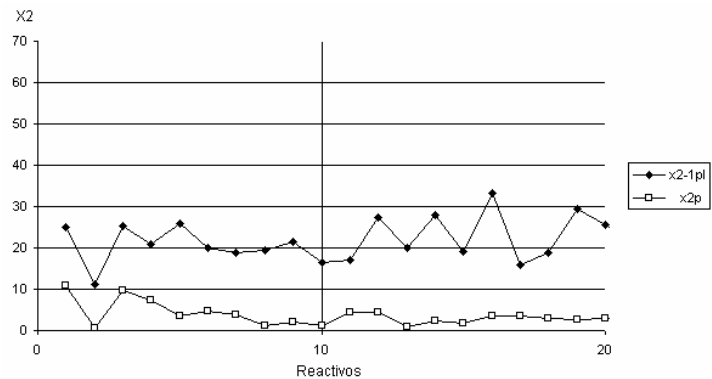
El valor del parámetro "a" tiene influencia del parámetro "c" y por ello el patrón no es muy claro, como puede verse en la gráfica inferior, donde hay una dispersión notable.



#### 4.4 Ajuste de datos al modelo

Rascal usa  $\chi^2$  para analizar la bondad de ajuste al modelo de 1 parámetro (identificado con X2-1pl en la gráfica). Kalt-Criterial cuenta igualmente con  $\chi^2$  para el ajuste a un modelo logístico de 2 parámetros (identificado con x2p).

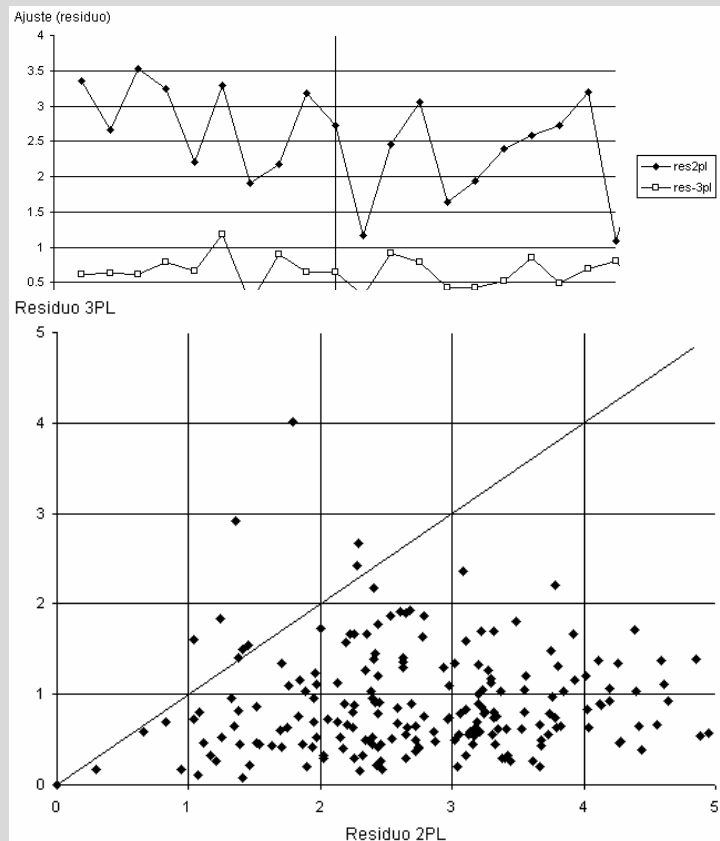
Como se espera en función de las hipótesis de los modelos logísticos, en general es mejor el ajuste del modelo de 2 parámetros (Kalt-Criterial) que el de 1 parámetro (Rascal).



Para medir el ajuste del modelo a los datos, Xcalibre usa el valor del residuo, que mide las diferencias entre teoría y observación y que debe ser menor que 2.0.

En principio, como se espera por las hipótesis de los modelos logísticos, es mejor el ajuste de 3PL que el de 2PL para Xcalibre, ya que la curva de tres parámetros es más "flexible" y debería tener un mejor ajuste a los datos.

Este tipo de ajuste no se cumple siempre, porque el uso del modelo de máxima verosimilitud conduce algunas veces a condiciones de no convergencia del modelo de 3 parámetros.



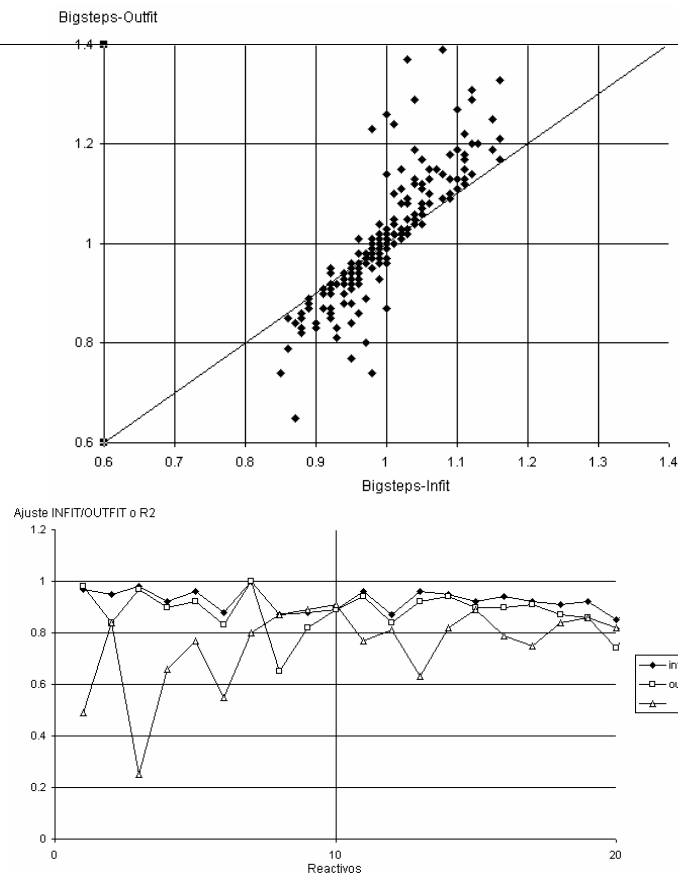
Para el análisis de Rasch, Winsteps determina el ajuste con ayuda del INFIT (sensible a los casos divergentes cerca de la zona de medición del ítem) y del OUTFIT (sensible a los casos divergentes lejos de la zona de medición del ítem).

Si se emplea el residuo cuadrático medio (MNSQ) los valores aceptables se comparan contra 1.0; el intervalo de aceptación está entre 0.8 y 1.2.

Si se usa el residuo estandarizado Z se considera que hay ajuste si los valores son inferiores a 2.0.

Kalt-Criterial, además del valor de  $\chi^2$  usa el coeficiente de determinación  $r^2$  como criterio de ajuste (son aceptables los valores cercanos a 1)

En general son mejores los ajustes que presentan Winsteps y Kalt-Criterial.



#### 4.5 Correlación Punto-Biserial

Es un indicador de la pertenencia del ítem al campo semántico de los otros ítems (es un índice de validez), pero muchas personas lo interpretan erróneamente como un índice de discriminación.

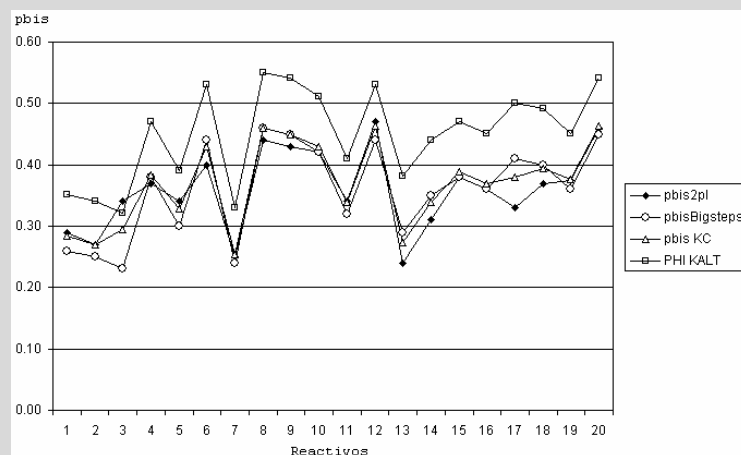
Rascal y Xcalibre (mismo proveedor) proporcionan idénticos resultados (se identifica en la gráfica con pbis 2PL).

Winsteps utiliza la correlación modificada (excluyendo el aporte del ítem) con un ajuste en términos de la medida.

Kalt-Criterial proporciona la correlación rpbis clásica en dos posibles presentaciones: modificada y no modificada.

Kalt Plus proporciona PHI, que produce similares resultados a los de la expresión sin modificar.

En general PHI es mayor, sigue rpbis sin modificación y el menor valor se obtiene con la expresión modificada de rpbis.



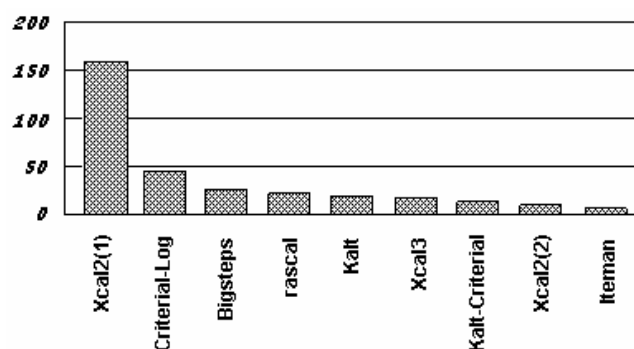
## 5. Comparación del dictamen de reactivos de la prueba analizada

En función de las decisiones que se tomen a partir del análisis de reactivos se puede llegar a muy variadas configuraciones de una prueba. En esta sección se comparan diversos elementos objetivos asociados con la calidad de una prueba, dependiendo de los dictámenes que se hicieron\_ con ayuda de los programas seleccionados. La primera observación que puede hacerse está asociada con el número de reactivos que rechaza cada modelo y cada programa. Se aprecia que hay diferencias importantes, como muestra la tabla:

Kalt- Criterial	Criterial-Log	ITEMAN	KALT	Rascal	Bigsteps	Xcalibre2(1)	Xcalibre2(2)	Xcalibre(3)
17	42	12	22	25	33	161	13	21

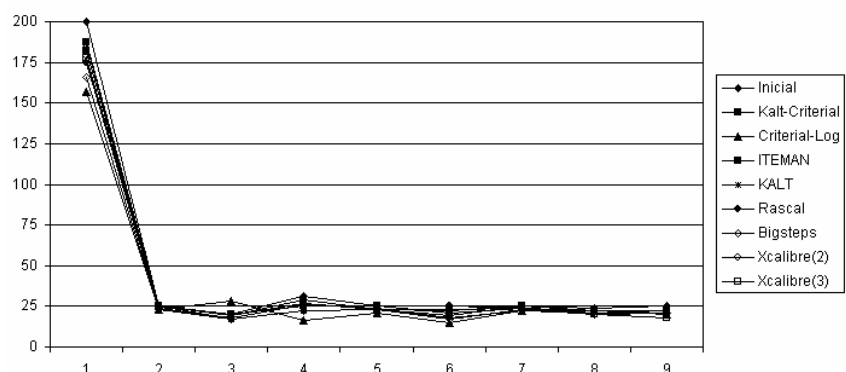
El modelo XCalibre 2 corresponde con el modelo de 2 parámetros, del cual se emiten dos formas de dictaminar a los reactivos: (1) utiliza la letra de dictamen "R", que indica que se tuvo un bajo ajuste del modelo y (2) utiliza solamente la letra de dictamen "P" que indica que hay problemas potenciales en el reactivo.

Es notable ver que el modelo de Xcalibre2(1) no consigue ajustar una curva logística de 2 parámetros a 161 de los reactivos aplicados. Este modelo se rechaza para el resto de este estudio por carecer de lógica, atendiendo a que otros modelos igualmente flexibles o inclusive menos flexibles que el utilizado por Xcalibre sí indican ajuste en la mayoría de los reactivos (tanto el modelo de 2 parámetros de Kalt Criterial, como los modelos de 1 parámetro de Rascal y Bigsteps ajustan convenientemente).



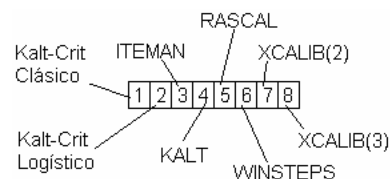
	1 Global	2 Matem.	3 Español	4 Razonam.	5 Metodol.	6 Mundo actual	7 Inglés	8 Com	9 Tecnol. educ.
Inicial	200	25	20	31	25	25	25	24	25
Kalt-Criterial	182	25	19	27	23	19	25	22	22
CriteriakLog	157	23.	28	16 .	21	15	22	22	20
ITEMAN	187	25	19	29	23	22	25	22	22
KALT	177	25	17	22	23	23	24	22	21
Rascal	175	24	18	26	24	18	22	21	22
Bigsteps	166	23	17	22	23	17	23	20	21
Xcalibre 2 (2)	186	25	20	25	25	25	24	21	21
Xcalibre(3)	178	25	20	25	25	21	24	20	18

Entre los modelos logísticos Kalt Criterial es el más riguroso (debe recordarse, sin embargo, que el dictamen depende del punto de corte elegido), en tanto que Xcalibre es el más benévolo. Por otra parte, Winsteps es el modelo más riguroso entre los programas de Rasch. Finalmente, Kalt es el más riguroso de los programas que emplean modelo clásico, mientras que Iteman es el más benévolo.





A partir de los reportes de los programas se puede construir el "Mapa" de rechazos para la prueba en estudio. La barra junto a cada reactivo está formada por ocho casilleros correspondientes a los ocho modelos involucrados en este estudio. Se señala en sombreado el programa que rechaza al reactivo.



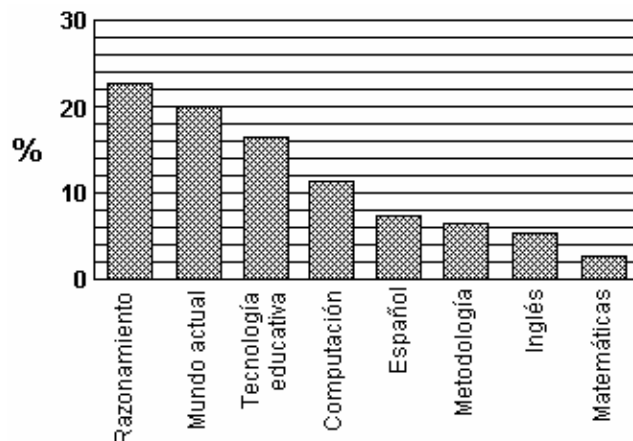
Núm.	Dictamen	Núm.	Dictamen	Núm.	Dictamen	Núm.	Dictamen	Núm.	Dictamen
MATEMATICAS		41		81		12		162	
1		42		82		123		163	
2		43		83		124		164	
3		44		84		125		165	
4		45		85		126		166	
5		RAZONAMIENTO		86		INGLES		167	
6		46		87		127		168	
7		47		88		128		169	
8		48		89		129		170	
9		49		90		130		171	
10		50		91		131		172	
11		51		92		132		173	
12		52		93		133		174	
13		53		94		134		175	
14		54		95		135		TECNOLOGIA	
15		55		96		136		EDUCATIVA	
16		56		97		137		176	
17		57		98		138		177	
18		58		99		139		178	
19		59		100		140		179	
20		60		101		141		180	
21		61		MUNDO ACTUAL		142		181	
22		62		102		143		182	
23		63		103		144		183	
24		64		104		145		184	
25		65		105		146		185	
ESPAÑOL		66		106		147		186	
26		67		107		148		187	
27		68		108		149		188	
28		69		109		150		189	
29		70		110		151		190	
30		71		111		152		191	
31		72		112		COMPUTACION		192	
32		73		113		153		193	
33		74		114		154		194	
34		75		115		155		195	
35		76		116		156		196	
36		METODOLOGIA		117		157		197	
37		77		118		158		198	
38		78		119		159		199	
39		79		120		160		200	
40		80		121		161			

Puede observarse que así como hay ítems que son rechazados por los diversos modelos, también hay áreas que tienen mayor número de rechazos que otras. Para estimar el impacto por área se emplea el índice de rechazo IR, que se obtiene por medio del cociente del número de rechazos de un área dada entre el producto (reactivos x programas).

$$IR = \frac{\sum \text{rechazos}}{\text{reactivos} \times \text{PROGRAMAS}}$$

El área con mayor número de rechazos es la de Razonamiento (con 22.5%), mientras que la que menos rechazos presenta es Matemáticas (con 2.5%).

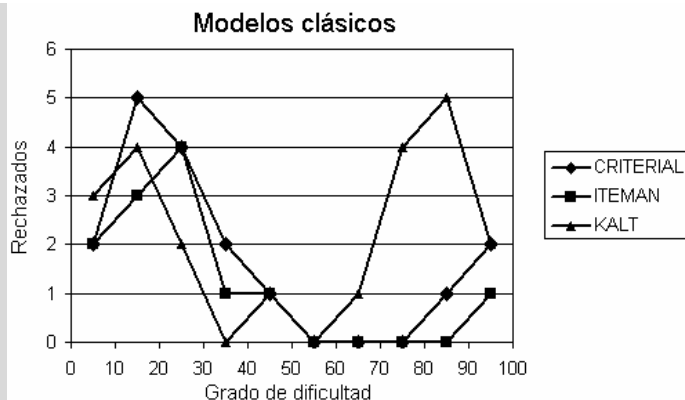
La media de rechazos por área es de 11.5 %, combinando todos los modelos de los diferentes programas en estudio.



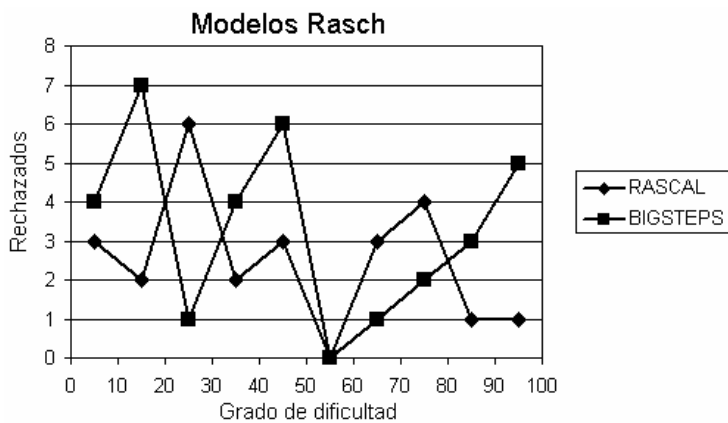
### 5.1 Distribución de reactivos rechazados en función de la dificultad de los reactivos

Una forma de ver el enfoque de rechazo que proporciona cada modelo es en función del Grado de Dificultad.

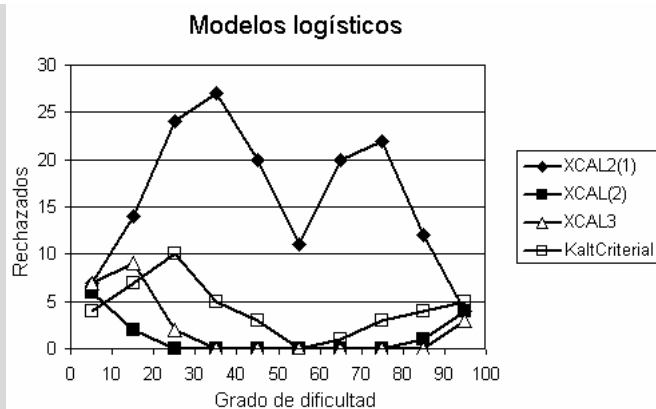
Para el modelo clásico se puede ver ITEMAN y KALT Criterial (modelo clásico) son más exigentes en los reactivos difíciles que en los fáciles (debe recordarse que se trata de un modelo de norma constante o exigencia decreciente). Kalt Plus presenta similares proporciones tanto en reactivos fáciles como difíciles (Tristán, 1995). En todos los modelos se tiene un menor número de rechazos (o ninguno) en los reactivos de dificultad media.



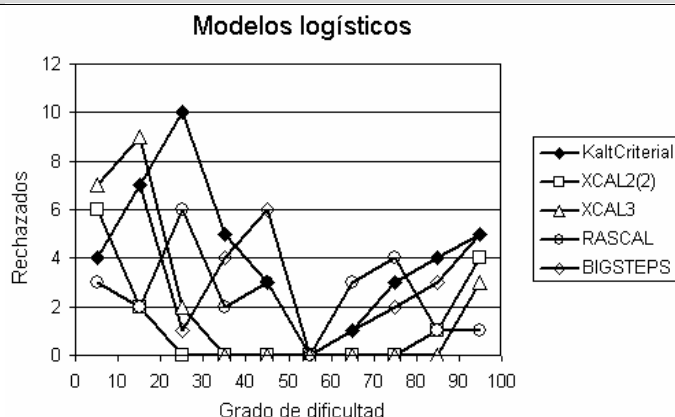
Si se consideran por separados los dos programas que utiliza el modelo de Rasch, se encuentra que, en general, BIGSTEPS es más exigente que Rascal, aunque parece que depende de las posiciones relativas de los reactivos. Como se mostró previamente, la estimación de la medida que hacen ambos programas es diferente y esto puede ser la causa de las divergencias en el momento del dictamen.



El modelo proporcionado por XCALIBRE de 2 parámetros, de acuerdo con la letra de dictamen "R" resulta poco creíble, ya que rechaza prácticamente al 80% de los reactivos, lo cual es inexplicable ya que modelos de 1 parámetro (Rasch) sí ajustan a los datos. Este modelo no se vuelve a considerar en este trabajo. Xcalibre es más benévolo en el caso de 3 parámetros. En general los modelos son más exigentes con los reactivos difíciles que con los fáciles.

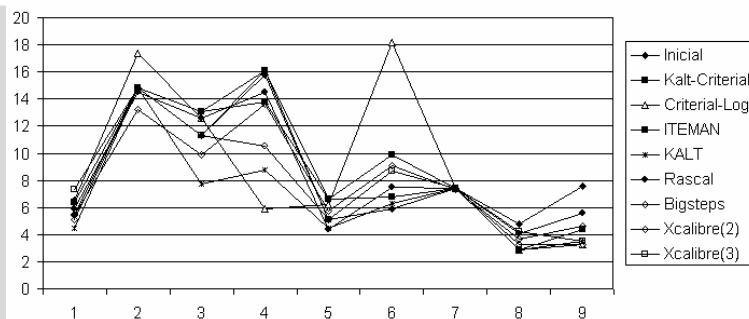


La comparación de modelos logísticos muestra que Kalt Criterial brinda valores más exigentes, sin embargo esto no puede tomarse como una tendencia definitiva porque depende del punto de corte utilizado para el análisis de los reactivos y el dictamen de los sustentantes. En general los modelos son más exigentes con los reactivos difíciles que con los fáciles.



## 5.2 Distancia absoluta media a la recta de diseño 20-80

Un elemento objetivo para dictaminar la calidad de la escala de una prueba es por medio de la comparación con la recta de diseño 20-80 que se obtiene con los reactivos no rechazados (Tristán y Vidal, 1999, 2007). Para este estudio solamente se hizo la elección de los reactivos en función del análisis global de la prueba, no se considera el dictamen para cada área, lo cual podría dar otra impresión del comportamiento de las áreas en la prueba.



La tabla presenta los valores de la distancia absoluta media de los reactivos elegidos. Se señalan con (x) los casos en que se trata del valor más pequeño de la DAM:

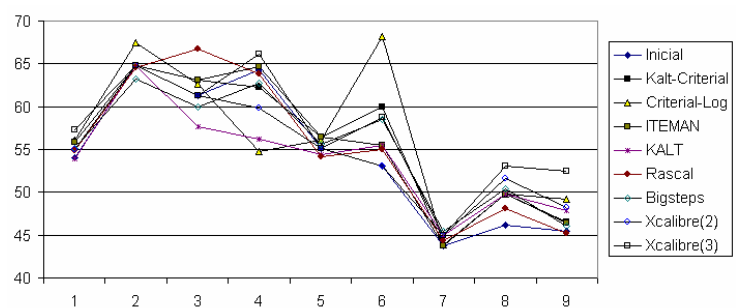
	1	2	3	4	5	6	7	8	9
	Global	Matem.	Español	Razonam.	Metodol.	Mundo actual	Inglés	Com	Tecnol. educ.
Inicial	5.5	14.8	11.3	15.8	5.1	5.9	7.4	4.8	7.6
Kalt-Criterial	6.4	14.8	13.1	13.8	6.6	9.9	7.4	2.9(x)	4.4
Criterial-Log	6.5	17.4	12.6	5.9(x)	6.2	18.2	7.4	2.9(x)	3.3(x)
ITEMAN	6.4	14.8	13.1	16.1	6.6	6.8	7.4	2.9(x)	4.4
KALT	4.5(x)	14.8	7.7(x)	8.8	4.5	6.3	7.4	2.9(x)	3.5
Rascal	5.9	14.6	12.6	14.5	4.4(x)	7.5	7.4	4.1	5.6
Winsteps	5.4	13.2(x)	9.9	13.6	5.7	9.1	7.3(x)	3.7	4.6
Xcalibre(2)	5.1	14.8	11.3	10.5	5.1	5.9(x)	7.4	3.3	3.3(x)
XCalibre(3)	7.3	14.8	11.3	16.1	5.1	8.7	7.4	4.2	3.5

### 5.3 Dificultad media de la prueba y sus partes

Por diseño, la prueba está planeada para tener una media de dificultad cercana al 50%, la tabla compara los valores que se obtienen con los reactivos aceptados por cada programa.

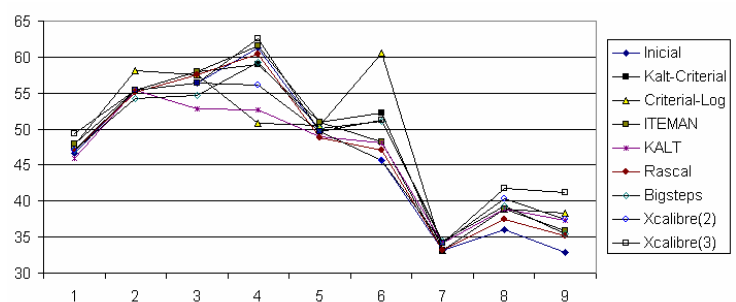
	1 Global	2 Matem.	3 Español	4 Razonam.	5 Metodol.	6 Mundo actual	7 Inglés	8 Com	9 Tecnol. educ.
Inicial	54	64.8	61.3	64.3	55.1	53.1	43.8	46.2	45.4
Kalt-Criterial	55.8	64.8	63.1	62.3	56.5	59.9	43.8	49.8	46.4
Criterial-Log	56.1	67.4	62.6	54.7	56.1	68.2	43.9	49.8	49.2
ITEMAN	55.8	64.8	63.1	64.7	56.5	55.5	43.8	49.8	46.4
KALT	53.9	64.8	57.7	56.2	54.4	55.5	45	49.8	47.9
Rascal	54.9	64.6	66.7	63.8	54.2	55	44.4	48.1	45.2
Winsteps	55.1	63.2	59.9	62.8	55.6	58.5	45.5	50.	46.1
Xcalibre(2)	55	64.8	61.3	59.8	55.1	53.1	45	51.6	48.2
Xcalibre(3)	57.3	64.8	61.3	66.1	55.1	58.7	45	53.1	52.5

Obsérvese que Kalt-Criterial difiere mucho en el área de Mundo actual donde debería modificarse el punto de corte para el dictamen de los reactivos, en este caso se mantuvo el punto de corte fijo.



### 5.4 Media de aciertos (%)

Se compara la media de aciertos de los sujetos para cada área utilizando los reactivos elegidos por cada programa, independientemente de que la prueba está razonablemente centrada en global y sus partes, se tiene un impacto más amplio en las medias de aciertos. Como era de esperarse, las tendencias son muy similares entre el caso 5.3 y éste.

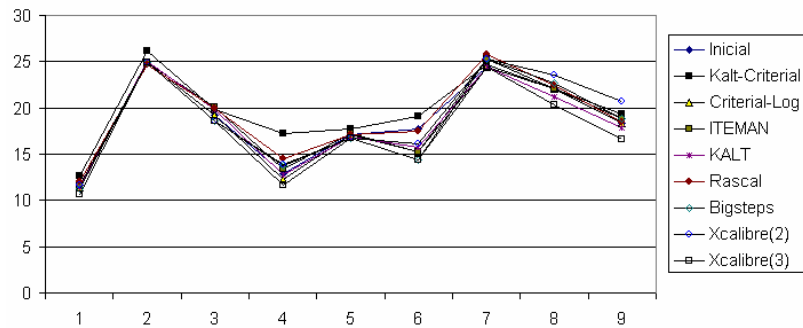


La tabla proporciona los valores obtenidos.

	1 Global	2 Matem.	3 Español	4 Razonam.	5 Metodol.	6 Mundo actual	7 Inglés	8 Com	9 Tecnol. educ.
Inicial	46.6	55.4	56.4	61.2	49.7	45.7	33.1	36	32.8
Kalt-Criterial	47.9	55.4	58	59	50.9	52.2	33.1	38.9	35.9
Criterial-Log	47.8	58.1	57.5	50.8	50.5	60.6	33.1	38.9	38.3
ITEMAN	48	55.4	58	61.6	50.9	48.2	33.1	38.9	35.9
KALT	45.9	55.4	52.8	52.7	49	48.1	34.1	38.9	37.3
Rascal	47	55.1	57.5	60.4	48.8	47	33.1	37.4	35.1
Winsteps	47.1	54.2	54.7	59.3	50.1	51.1	34.4	39.3	35.5
Xcalibre(2)	47.1	55.4	56.4	56.1	49.7	45.7	34.1	40.3	37.5
Xcalibre(3)	49.4	55.4	56.4	62.6	49.7	51.3	34.1	41.7	41.2

### 5.5 Desviación estándar de aciertos (%)

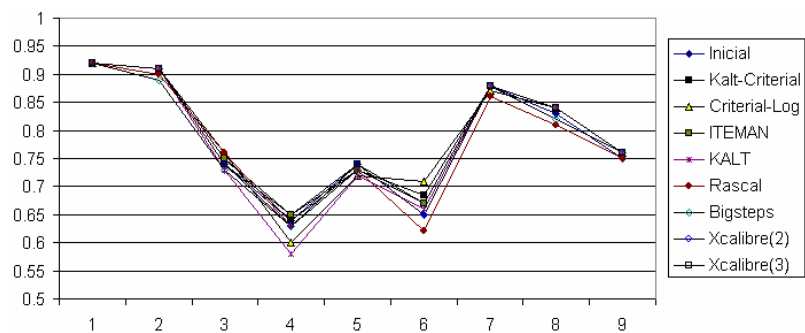
No se presenta un impacto muy notable en la desviación estándar por el uso de los diferentes programas. Las áreas que tienen mayor divergencia son las de Razonamiento y Mundo actual, seguramente por la cantidad de reactivos rechazados en estas áreas.



	1	2	3	4	5	6	7	8	9
	Global	Matem.	Español	Razonam.	Metodol.	Mundo actual	Inglés	Com	Tecnol. educ.
Inicial	11.7	24.9	19.2	12.9	17.1	17.7	24.4	22.1	18.4
Kalt-Criterial	12.7	26.2	19.8	17.2	17.7	19.1	24.7	22.1	19.4
Criterial-Log	11.4	24.9	19.2	12.3	17.1	15.3	24.4	22.1	18.4
ITEMAN	11.7	24.9	20.1	13.4	17.2	15.2	25.3	22.1	18.9
KALT	11.4	25	118	12.8	16.9	15.8	24.4	21.2	17.8
Rascal	12	24.7	19.9	14.5	17.1	17.5	25.8	22.4	18.5
Bigsteps	11.3	24.9	18.6	13.8	16.7	14.4	25.3	22.7	18.9
Xcalibre(2)	11.7	24.9	18.6	13.9	16.7	16.1	25.3	23.5	20.7
Xcalibre(3)	10.6	24.9	18.6	11.6	16.7	14.4	124.4	20.3	16.6

### 5.6 Confiabilidad alfa de Cronbach

Un elemento de interés para comparar las decisiones que se toman con los programas es la confiabilidad. Se compara el valor de alfa de Cronbach y no se muestran diferencias muy notables entre los diversos programas. En función del modelo Máxima validez-Máxima confiabilidad (Tristán, 2000-2007), se dispone de los valores esperados para cada área evaluada. Se observa un buen acuerdo con los valores obtenidos en la prueba.



	1	2	3	4	5	6	7	8	9
	Global	Matem.	Español	Razonam.	Metodol.	Mundo actual	Inglés	Comp	Tecnol. Educ.
ESPERADOS	0.92	0.89-0.91	0.72-0.75	0.59-0.70	0.72-0.73	0.68-0.77	0.87-0.89	0.83-0.85	0.76-0.79
Inicial	0.92	0.91	0.74	0.64	0.74	0.65	0.88	0.83	0.75
Kalt-Criterial	0.92	0.91	0.75	0.644	0.729	0.684	0.88	0.84	0.76
Criterial-Log	0.92	0.91	0.76	0.6	0.72	0.71	0.87	0.84	0.76
ITEMAN	0.92	0.91	0.75	0.65	0.73	0.67	0.88	0.84	0.76
KALT	0.92	0.91	0.73	0.58	0.72	0.66	0.88	0.84	0.76
Rascal	0.92	0.9	0.76	0.63	0.73	0.62	0.86	0.81	0.75
Winsteps	0.92	0.89	0.73	0.63	0.73	0.67	0.88	0.82	0.76
Xcalibre(2)	0.92	0.91	0.74	0.63	0.74	0.65	0.88	0.84	0.76
Xcalibre(3)	0.92	0.91	0.74	0.65	0.74	0.67	0.88	0.84	0.76

## 5.7 Identificación de los tipos de reactivos rechazados

Ahora se presentan los valores de análisis de algunos reactivos que son rechazados por los diferentes programas. Se estudian varios casos que proporcionan indicios en relación con el dictamen de los reactivos. No se aprecia un patrón específico que explique el rechazo o aprobación del reactivo, como puede verse en la lista de resultados de referencia.

**5.7.1 Caso 1.** Reactivos que son rechazados solamente por uno de los programas, los demás programas aceptan los reactivos. Este caso ocurre en 33 reactivos.

Programa	Núm de reactivos	Ejemplo	Resultados de referencia
Kalt-Criterial Modelo logístico	7	De las siguientes opciones, identifica las que representan un grupo de números primos: 1. 11,39,56 2. 7, 19,29 3. 13, 41 4. 53, 97, 121 A) 1, 2 y 3 B) solo 1 y 4 C) 2, 3 y 4 D) solo 2 y 3 E) 1, 2, 3 y 4	KCMC: GD=31, rpbis=0.27 KCML: Dif=0.04, Dis=0.67, $\chi^2=10.8$ , $p(\chi^2)=0.21$ ITEM: $p=0.259$ , dis=0.29, rpbis=0.276 KALT: GD=31, RD=0.99 RASC: $b=0.98$ , $\chi^2=25.11$ BIGS: $b=1.1$ , INF=(0.97,-0.45), OUT=(0.98,-0.26) XCA2: $a=0.64$ , $b=1.68$ , $es=3.35$ XCA3: $a=0.67$ , $b=2.04$ , $c=0.22$ , $es=0.61$
KALT Plus	2	En las opciones se incluyen sinónimos de APOSTAR, solamente una NO CORRESPONDE CON EL SENTIDO de esta oración: "Los ejércitos estaban apostados en Nápoles". A) colocar B) asentar C) ubicar D) situar E) jugar	KCMC: GD=80, rpbis=0.10 KCML: Dif=-1.99, Dis=0.44, $\chi^2=0.46$ , $p(\chi^2)=1.0$ ITEM: $p=0.732$ , dis=0.14, rpbis=0.159 KALT: GD=80, RD=0.22 RASC: $b=-1.2$ , $\chi^2=14.9$ BIGS: $b=-1.21$ , INF=(1.07,0.64), OUT=(1.15,1.10) XCA2: $a=0.56$ , $b=-1.69$ , Res=3.22 XCA3: $a=0.48$ , $b=-0.86$ , $c=0.24$ , Res=1.70
Rascal	6	¿Cuál es la traducción correcta de: "La próxima semana iremos a California"?. A) We shall go to California next week. B) Next week we will go to California. C) Next week we shall go in California. D) We should go next week to California	KCMC: GD=31, rpbis=0.18 KCML: Dif=0.14, Dis=0.58, $\chi^2=6.35$ , $p(\chi^2)=0.61$ ITEM: $p=0.25$ , dis=0.26, rpbis=0.26 KALT: GD=31, RD=1.15 RASC: $b=1.03$ , $\chi^2=31.86$ BIGS: $b=1.1$ , INF=(0.96,-0.56), OUT=(0.95,-0.59) XCA2: $a=0.6$ , $b=1.84$ , Res=4.55 XCA3: $a=0.62$ , $b=2.21$ , $c=0.22$ , Res=0.66
Winsteps	8	Identifique la palabra que completa correctamente las oraciones: 1. El viento ____ suavemente la superficie del lago. (risa, riza) 2. Entrega los ____ de tela al supervisor. (rollos, royos) 3. Aristóteles fue la persona más ____ de Grecia en su tiempo. (sabia, savia) 4. Toca la flor y su mano ____ la mía con suavidad. (rosa, roza) 5. El jugo de las plantas se conoce con el nombre de ____ (sabia, savia) A) risa, rollos, savia, rosa, sabia B) riza, royos, savia, roza, sabia C) risa, rollos, sabia, rosa, savia D) riza, rollos, sabia, roza, savia E) risa, royos, sabia, roza, savia	KCMC: GD=89, rpbis=0.23 KCML: Dif=-1.99, Dis=0.72, $\chi^2=1.07$ , $p(\chi^2)=1.0$ ITEM: $p=0.87$ , dis=0.20, rpbis=0.29 KALT: GD=89, RD=0.83 RASC: $b=-2.191$ , $\chi^2=14.5$ BIGS: $b=-1.95$ , INF=(0.95,0.33), OUT=(0.77, 1.22) XCA2: $a=0.76$ , $b=-2.26$ , Res=2.25 XCA3: $a=0.62$ , $b=-1.76$ , $c=0.23$ , Res=0.63

Programa	Núm de reactivos	Ejemplo	Resultados de referencia
Xcalibre 3PL	3	<p>La didáctica clasifica los factores que inciden en el aprendizaje en Externos e Internos, identifíquelos de la lista siguiente:</p> <ol style="list-style-type: none"> <li>1. Titularse</li> <li>2. Ocupar el tiempo</li> <li>3. Obtener buenas calificaciones</li> <li>4. Motivación personal</li> <li>5. Interés P por la cultura</li> <li>6. Beneficio económico</li> </ol> <p>A) Externos: 3, 5, 6. Internos: 1, 2, 4  B) Externos: 2, 3, 6. Internos: 1, 4, 5  C) Externos: 5, 6. Internos: 1, 2, 3, 4  D) Externos: 1, 3, 6. Internos: 2, 4, 5  E) Externos: 1, 2, 3, 6. Internos: 4, 5</p>	<p>KCMC: GD=26, rpbis=0.15  KCMC: Dif=1.26, Dis=0.41,  <math>\chi^2=4.86</math>, <math>p(\chi^2)=0.77</math>  ITEM: p=0.16,  KALT: GD=26, RD=0.74  RASC: b=1.59, <math>\chi^2=19.0</math>  BIGS: b=1.26, OUT=(1:05,0.44)  XCA2: a=0.59,  XCA3: a=0.64, b=3.0, Res=0.46  (P)</p>

**5.7.2 Caso 2.** Todos los programas rechazan al reactivo. Solo se tienen dos reactivos en esta situación. Se trata de reactivos difíciles que discriminan poco o mal.

Ejemplo	Resultados de referencia
<p>¿Para qué sirve esta barra de herramientas de Power Point</p> <p>A) Ver presentación (adelantar y retroceder, inicio y fin, insertar diapositivas)  B) Organizar la presentación (cambio de nivel, mover diapositivas, ver formato)  C) Editar presentación (copiar, pegar, insertar, buscar, revisar ortografía)  D) Dar formato de presentación (cambio de niveles, inicio y fin, cambiar fuentes)  E) Formato de presentación (fuentes, mover dispositivas, cambiar tamaños)</p>	<p>KCMC: GD=11, rpbis=-0.18  KCMC: Dif=-12.6, Dis=-0.3, <math>\chi^2=7.42</math>,  <math>p(\chi^2)=0.49</math>  ITEM: p=0.08, dis=-0.64, rpbis=-0.09  KALT: GD=11, RD=-0.88  RASC: b=2.33, <math>\chi^2=41.44</math>  BIGS: b=2.43, INF=(1.1,0.51),  OUT=(1.56,2.09)  XCA2: a=0.51, b=3.0, Res=3.08(P)  XCA3: a=0.74, b=3.0, c=0.19,  Res=2.36(P)</p>

Análisis clásico del reactivo:

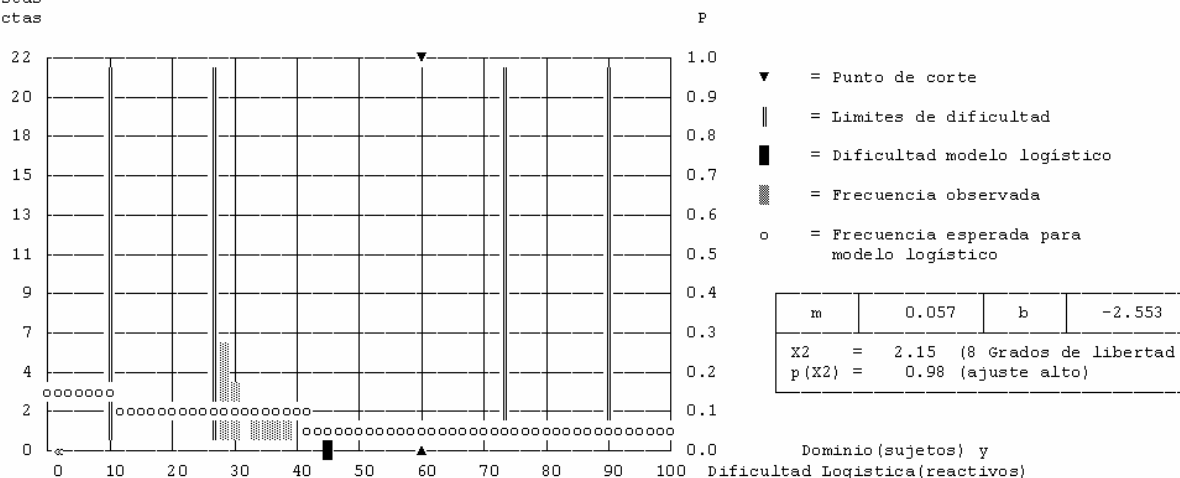
		A	D	B	C	E	Omis.	Error	Total	Válida			R. C.	R. I.	Válida	
G. S.	a	43	7	5	19	18	18	0	110	92	a	G. S.	7	85	92	a
	b	19.54	3.18	2.27	8.63	8.18	8.18	0.00	50.00	41.81	b		3.18	38.63	41.81	b
	c	41	10	8	15	18				92	c		10	82	92	c
	d	18	19	18	18	18					d		19	73		d
	e	Tyw	Rx	x	Tyw	yw					e		Rx	Tyw		e
G. I.	a	33	12	9	8	15	33	0	110	77	a	G. I.	12	65	77	a
	b	15.00	5.45	4.09	3.63	6.81	15.00	0.00	50.00	35.00	b		5.45	29.54	35.00	b
	c	35	9	6	12	15				77	c		9	68	77	c
	d	19	0	19	19	19					d		0	77		d
	e	xw	Sz	U	x	Uw					e		Sz	xw		e
TOTAL	a	76	19	14	27	33	51	0	220	169	a	TOTAL	19	150	169	a
	b	34.54	8.63	6.36	12.27	15.00	23.18	0.00	100.00	76.81	b		8.63	68.18	76.81	b

DIAGRAMA DE RESPUESTAS POR QUINTILES

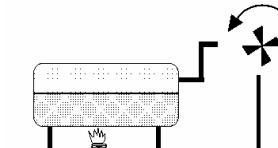
	0	25	50	75	100	%	CASOS	TOTAL DE CASOS
20	■					3.6	6	6
40	■					3.0	5	11
60						1.2	2	13
80	■					3.0	5	18
100						0.6	1	19

## Análisis criterial del reactivo:

		A	D	B	C	E	Omis.	Error	Total	Válida			R.C.	R.I.	Válida	
G.S.	a	30	4	2	16	11	8	0	71	63	a	G.S.	4	59	63	a
	b	13.64	1.82	0.91	7.27	5.00	3.64	0.00	32.27	28.64	b		2.37	34.91	37.28	b
	c	28	7	5	10	12				63	c		7	56	63	c
	d										d					d
	e	Tw	Rx	x	Tyw	xw					e		Rx	Tw		e
G.I.	a	46	15	12	11	22	43	0	149	106	a	G.I.	15	91	106	a
	b	20.91	6.82	5.45	5.00	10.00	19.55	0.00	67.73	48.18	b		8.88	53.85	62.72	b
	c	48	12	9	17	21				106	c		12	94	106	c
	d										d					d
	e	xw	Sz	U	Ux	Uw					e		Sz	Uxw		e
TOTAL	a	76	19	14	27	33	51	0	220	169	a	TOTAL	19	150	169	a
	b	34.55	8.64	6.36	12.27	15.00	23.18	0.00	100.00	76.82	b		11.24	88.76	100.00	b

Respuestas  
correctas

**5.7.3 Caso 3.** Todos excepto uno de los programas rechazan al reactivo (solamente Rascal o Xcalibre aceptan al reactivo). Se trata de reactivos difíciles que discriminan poco o mal.

Programa que acepta	Núm. de reactivos	Ejemplo	Resultados de referencia
Rascal	2	<p>El libro de Didáctica de las Ciencias Físico-Químicas de Medina Valenzuela sugiere mostrar que al calentar un fluido encerrado en un recipiente se produce una presión, la cual a su vez se transforma en energía para mover un rehilete. Para ello propone la construcción del modelo mostrado utilizando estos materiales: bote vacío de cerveza, trozos de lámina delgada, tubo de cobre, lámpara de alcohol y una base y soportes de madera.</p>  <p>¿Qué tipo de aprendizaje sugiere este autor?</p> <p>           A) Psicomotriz            B) Reflexivo o significativo            C) Teórico            D) Pasivo            E) Experimental o de laboratorio         </p>	<p>             KCMC: GD=5.9, rpbis=-0.03              KCML: Dif=15.05, Di=0.00, <math>\chi^2=2.52</math>, <math>p(\chi^2)=0.96</math>              ITEM: p=0.05, dis=0.15, rpbis=-0.02              KALT: GD=5.9, RD=-0.3              RASC: b=2.7, <math>\chi^2=22.7</math>              BIGS: b=3.16, INF=(1.07, 0.25)              OUT=(1.52, 1.36)              XCA2: a=0.56, b=3.0, Res=1.36              XCA3: a=0.76, b=3.0, c=0.17, Res=2.92 (P)           </p>



Programa que acepta	Núm. de reactivos	Ejemplo	Resultados de referencia
Xcalibre (2)	1	(Continuación del reactivo anterior) ¿Qué tipo de modelo se sugiere en este caso? A) Computacional B) Icónico C) Estadístico D) Matemático E) Gráfico	KCMC: GD=17, rpbis=-0.08 KCML: Dif=-14.0, Dis=-0.23, $\chi^2=6.93$ , $p(\chi^2)=0.54$ ITEM: p=0.132, dis=-0.06, rpbis=-0.4 KALT: GD=17, RD=-0.80 RASC: b=1.85, $\chi^2=45.1$ BIGS: b=1.87, INF=(1.13,0.93), OUT=(1.47,2.49) XCA2: a=0.50, b=2.88, Res=3.78 (R) XCA3: a=0.71, b=3.0, c=0.2, Res=2.21 (P)

**5.7.4 Caso 4.** Se consideran casos combinados en que varios programas de un mismo tipo de modelo (clásico o logístico) coinciden en un dictamen. Uno o varios de los modelos logísticos identifican 29 reactivos que el modelo clásico no identifica, mientras que uno o varios de los modelos clásicos identifican 9 reactivos que el modelo logístico no identifica.

Modelos	Núm. De reactivos	Ejemplo	Resultados de referencia
Solo logísticos (1 o varios)	5	Dada la secuencia 3,6,12,24,... ¿cuál es el número que sigue?  A) 26 B) 32 C) 48 D) 52	KCMC: GD=98, rpbis=0.11 KCML: Dif=-9.0, Dis=0.59, $\chi^2=0.23$ , $p(\chi^2)=1.0$ ITEM: p=0.98, dis=0.47, rpbis=0.25 KALT: GD=98, RD=1.25 RASC: b=-4.5, $\chi^2=8.86$ BIGS: b=-4.18, INF=(0.97,-0.05), OUT=(0.47,-0.98) XCA2: a=0.74, b=-3.0, Res=1.32 (P) XCA3: a=0.77, b=-3.0, c=0.23, Res=0.95 (P)
Solo clásicos	2		
3 o más logísticos	22		

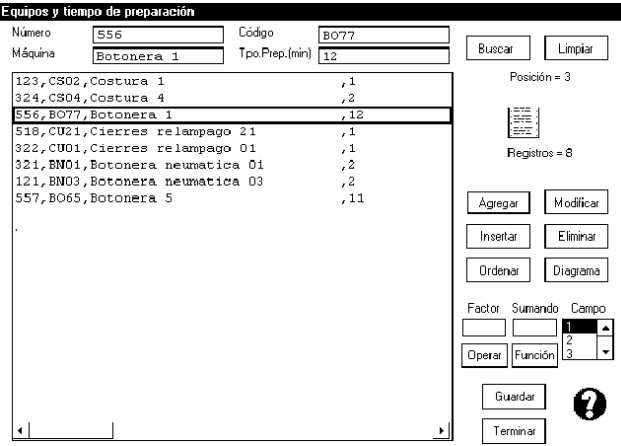
#### Análisis clásico del reactivo

		A	B	C	D	E	Omis.	Error	Total	Válida	
G.S.	a	0	0	0	110	0	0	0	110	110	a
	b	0.00	0.00	0.00	50.00	0.00	0.00	0.00	50.00	50.00	b
	c	1	1	0	108	0				110	c
	d	0	0	0	110	0					d
	e	x	x								e
G.I.	a	2	2	0	106	0	1	0	110	109	a
	b	0.45	0.90	0.00	48.18	0.00	0.45	0.00	50.00	49.54	b
	c	0	1	0	108	0				109	c
	d	1	1	1	106	1					d
	e	U	U	U	x	U					e
TOTAL	a	1	2	0	216	0	1	0	220	219	a
	b	0.45	0.90	0.00	98.18	0.00	0.45	0.00	100.00	99.54	b

	R.C.	R.I.	Válida	
G.S.	110	0	110	a
	50.00	0.00	50.00	b
	108	2	110	c
	110	0		d
		x		e
G.I.	106	3	109	a
	48.18	1.36	49.54	b
	108	1	109	c
	106	3		d
	x	U		e
TOTAL	216	3	219	a
	98.78	1.36	99.54	b

DIAGRAMA DE RESPUESTAS POR QUINTILES

	0	25	50	75	100	%	CASOS	TOTAL DE CASOS
20						19.2	42	42
40						19.2	42	84
60						20.1	44	128
80						20.1	44	172
100						20.1	44	216

Modelos	Num. De reactivos	Ejemplo	Datos de referencia
3 o mas logísticos sin criterial	12	<p>La figura muestra la pantalla de captura de un programa de base de datos</p>  <p>¿Cuál puede ser la aplicación principal de la base de datos mostrada?</p> <p>A) Organizar máquinas para definir sus tiempos  B) Agregar, modificar, insertar, eliminar, ordenar y hacer diagramas  C) Definir códigos y números de máquinas industriales  D) Organizar datos de máquinas en forma de tabla  E) Administrar los tiempos de reparación de maquinaria industrial</p>	<p>KCMC: GD=12.6, rpbis=0.10  KCMC: Dif=1.86, Dis=0.45, <math>\chi^2=8.19</math>, <math>p(\chi^2)=0.42</math>  ITEM: p=0.08, dis=0.12, rpbis=0.15  KALT: GD=12.6, RD=1.85  RASC: b=2.4, <math>\chi^2=30.65</math>  BIGS: b=2.29, INF=(0.98,-0.1), OUT=(1.23,0.92)  XCA2: a=0.59, b=3.0, Res=1.24 (P)  XCA3: a=0.71, b=3.0, c=0.19, Res_1.83 (P)</p>

## Análisis clásico del reactivo

		A	B	C	D	E	Omis.	Error	Total	Válida			R.C.	R.I.	Válida
G.S.	a	6	18	10	37	14	25	0	110	85	a	G.S.	14	71	85
	b	2.72	8.18	4.54	16.81	6.36	11.36	0.00	50.00	38.63	b		6.36	32.27	38.63
	c	8	26	11	29	11				85	c		11	74	85
	d	17	17	17	17	18					d		18	67	
	e	x	Txw	xy	Tyw	R					e		R	Txyw	
G.I.	a	7	26	9	12	4	52	0	110	58	a	G.I.	4	54	58
	b	3.18	11.81	4.09	5.45	1.81	23.63	0.00	50.00	26.36	b		1.81	24.54	26.36
	c	5	18	8	20	7				58	c		7	51	58
	d	14	14	14	14	0					d		0	58	
	e	Uw	w	Uw	xw	Sx					e		Sx	w	
TOTAL	a	13	44	19	49	18	77	0	220	143	a	TOTAL	18	125	143
	b	5.90	20.00	8.63	22.27	8.18	35.00	0.00	100.00	65.00	b		8.18	56.81	65.00

DIAGRAMA DE RESPUESTAS POR QUINTILES

	0	25	50	75	100	%	CASOS	TOTAL DE CASOS
20						1.4	2	2
40						1.4	2	4
60						0.0	0	4
80						3.5	5	9
100						6.3	9	18

## 6. Conclusiones

Las comparaciones realizadas en este trabajo permiten ver que las tendencias de los parámetros estadísticos y de análisis reflejan claramente las hipótesis involucradas en los programas estudiados. De hecho se confirman las hipótesis que explican por qué unos programas son más exigentes que otros; es importante el conocimiento de estas hipótesis para que el evaluador tome mejores decisiones respecto a la calidad de su prueba.

Como era de esperarse, hay tendencias generales similares, pero existen divergencias en los valores que se obtienen con los diferentes programas; inclusive hay divergencias en ciertos parámetros obtenidos con los programas de un mismo proveedor (ASC en particular) y se encontró un problema de ajuste del modelo logístico en el caso de Xcalibre (2 parámetros), por lo cual se desechó su uso para los fines de este trabajo; el evaluador deberá considerar a futuro la pertinencia del uso de este programa en otras aplicaciones diferentes a la mostrada en este trabajo.

Es claro que los elementos de decisión son distintos para cada programa, en algunos casos se emplean parámetros clásicos mientras que en otros se cuenta con criterios logísticos. Esto complica el análisis objetivo de los resultados obtenidos, por lo cual se sugirió incluir un conjunto de elementos externos de juicio sobre la calidad de las decisiones; en particular se emplearon estos valores de cotejo: media de dificultades, distancia absoluta media a la recta de diseño 20-80, media de aciertos, desviación estándar y coeficiente alfa de Cronbach con el modelo teórico de comparación Máxima validez-máxima confiabilidad. Se mostraron las diferencias obtenidas con los dictámenes de los programas, pudiendo apreciarse que las tendencias son muy similares con los diferentes programas.

Las diferencias más notables ocurrieron en dos de las áreas con Kalt Criterial, atendiendo a que es el único modelo que hace intervenir el punto de corte como parte del criterio que debe introducir el evaluador; los otros programas no hacen ningún análisis de reactivos con referencia a criterio. Esto explica el nivel de exigencia obtenido con este programa, ya que la exigencia depende del propio criterio que el evaluador hace intervenir en el programa.

El dictamen de los reactivos presenta, por lo tanto, diversas facetas en función del programa empleado. En algunos casos un reactivo es rechazado por uno de los programas o, por el contrario, uno solo de los programas lo acepta. Al revisar el contenido de los reactivos no resulta evidente por qué un programa rechaza mientras que otros aceptan un reactivo; la justificación se debe fundamentar en las hipótesis del modelo de cada programa, junto con los algoritmos de cálculo utilizados. Esto se ilustró con algunos reactivos que caen en estos casos particulares de rechazo y se incluyeron los parámetros estadísticos obtenidos con cada programa.

Puede notarse que los diferentes criterios no brindan imágenes contradictorias, más bien se trata de imágenes complementarias de una misma prueba: cada modelo presenta aspectos diferentes y podrían tomarse decisiones con ayuda de los diversos programas para analizar aspectos sobre los cuales el evaluador desee enfocarse o en función del modelo sobre el cual tenga más confianza.

El último criterio de selección corresponde con la calidad de la presentación de los reportes, la disponibilidad de una base de datos, la facilidad de uso del programa, etc. La tabla presenta en forma resumida algunas de las características de los programas que pueden servir como criterio de decisión para el usuario. Se señala con X los elementos disponibles en cada programa. El lector podrá hacer sus propias estimaciones cualitativas apoyándose en los ejemplos de reportes incluidos en este trabajo.

Características de los programas modelos	Kalt Plus	Kalt Criterial	Winsteps	Iteman	Rascal	Xcalibre
Modelo clásico	X	X		X		
Modelo logístico		X	X		X	X
Evaluación referida a norma o de propósito general	X	X	X	X	X	X
Evaluación referida a criterio		X				
Reporte estadístico (Tablas gráficas)	X	X	X	X		
Confiabilidad general	X	X	X	X	X	X
Confiabilidad general por tema en la misma corrida	X	X				
Diagrama de dificultades de los reactivos	X	X	X		X	X
Curva característica de la prueba			X			X
Curva de información de la prueba			X		X	
Análisis global de reactivos (sin opciones)	X	X	X	X	X	X
Análisis detallado de reactivos (con opciones)	X	X	X	X		
Gráfica de la curva característica del reactivo	X	X				
Dictamen de los reactivos	X	X				X
Generación de base de datos para "exportar"	X	X	X	X	X	X
Reportes configurables por el usuario	X	X				

### Programas de referencia

1. "Winsteps User's Manual", Chicago. Winsteps.com
2. "Iteman User's Manual", ASC, Minnesota.
3. "KALT Plus", Manual de usuario, Familia de Programas KALT, México.
4. "KALT Criterial", Manual de usuario, Familia de Programas KALT, México .
5. "Rascal User's Manual", ASC, Minnesota
6. "Xcalibre User's Manual", ASC, Minnesota

### Referencias

Linacre, J. M. (2005) WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com

"Manual técnico del Examen de certificación para Profesores de Educación Media Superior" (ECPEMS), Instituto de Evaluación e Ingeniería Avanzada, S.C., México, abril, 2001, 47 pp.

Tristán, L.A. (1995) Modelo para el análisis de reactivos objetivos por computadora. Primer Foro Nacional de Evaluación Educativa. Ceneval. Colima. Pp. 45-68

Tristán L.A. (1998) Modelo para calificación y análisis por computadora de cuestionarios referidos a criterio. Tercer Foro Nacional de Evaluación Educativa. Ceneval. Veracruz. Pp. 237-247

Tristán L.A. y Vidal, U.R. (1999) Modelo de diseño para validez de constructo en pruebas referidas a criterio. Notas sobre Evaluación Criterial, México. IEESA-Ceneval. También: Manual de Kalt Criterial. IEIA, México. Nota Técnica N.10.

Tristán A. y Vidal R. (2001) "Contribución al estudio del error de medida (Parte 3)", Notas sobre evaluación criterial N. 13. IEESA-Ceneval. México, 5 pp

Tristán L.A. (2004) Sistema para calificación de pruebas referidas a criterio y definición de estándares. Cap.14 en Educación, aprendizaje y cognición. Ed. Castañeda S., El Manual Moderno, México. Pp. 219-234

Tristán, L.A. y Vidal, U.R. (2007) Linear model to assess the scale's validity of a test. AERA, Chicago. 8 pp.