20485

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES "ACATLÁN"

CÁLCULO DEL ÍNDICE DE DIFICULTAD DE ÍTEMS CON DOS FORMATOS DIFERENTES DE TESTS. UN ESTUDIO DE CASO.

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN EDUCACIÓN MATEMÁTICA

P R E S E N T A :

MARÍA EUGENIA CANUT DÍAZ VELARDE

DIRECTOR DE TESIS:
M. en E.M. JORGE JIMÉNEZ ZAMUDIO.

ACATLÁN, EDO. DE MÉXICO. MAYO DE 2004

Mtro. Jorge Jiménez
Su confianza y paciencia, me han abierto
el camino al conocimiento y a una gran amistad

Mtro Víctor Palencia
Su dedicación y talento han favorecido
mi crecimiento como maestra y establecido
una amistad sincera.

Mtro. Juan Recio
Mtra Beatriz Ojeda
Mtro Francisco Mejía
Por sus valiosas aportaciones al desarrollo de este trabajo.

... "No todo lo que cuenta es evaluable,
ni todo lo que puede evaluarse cuenta".
Albert Einstein.

INTRODUCCIÓN

Casi toda persona que haya tenido oportunidad de realizar estudios formales ha pasado por la experiencia de resolver exámenes de opción múltiple. Este formato es muy usual en las áreas de conocimiento tales como ciencias sociales, humanidades y otras, generalmente no vinculadas con las áreas duras en el nivel de enseñanza superior. Si a un alumno se le diese la oportunidad de seleccionar entre un examen con respuestas de opción múltiple y otro de respuesta abierta, seguramente la mayoría se inclinaría por lo primero. ¿A qué se debe esta decisión? ¿Se considera que se tienen mayores posibilidades de acierto dentro de un formato que de otro?

La investigación que se reporta en el presente trabajo pretendió determinar, si el tipo de formato mediante el cual son presentados los ítems influye o no en el cálculo del índice de dificultad. El tipo de investigación se realizó con base en un estudio de caso, dado que es un método empleado para estudiar a un individuo o una institución en un entorno o situación único y de forma detallada. La palabra único es importante, porque se está interesado en las condiciones existentes que rodean el caso y que por diversas razones no puede ser repetido *n* veces en un laboratorio, pero que tienen la ventaja de proporcionar información suficiente de lo que está ocurriendo en un contexto determinado.

El desarrollo de este trabajo se presenta en cinco capítulos. En cada capítulo se encuentran fórmulas y tablas que, aunadas a la metodología empleada, permiten con las restricciones del caso repetir la investigación para consolidar los resultados obtenidos.

En el primer capítulo se plantea el problema a bordar, tras una breve descripción del estado de la evaluación del aprendizaje en general y de los diferentes formatos de tests. Se expone el punto de vista que tienen algunas instituciones y profesores, acerca de la evaluación, en el proceso de enseñanza – aprendizaje y se analiza qué debe contener una evaluación de calidad. Se definen cuáles son los objetivos que deberán cumplirse en un examen, definido

éste, como un instrumento de evaluación que coadyuva en la determinación del tipo de conocimientos que los alumnos han adquirido después del proceso de aprendizaje. Se analizan algunos de los diferentes tipos de instrumentos que existen y sus ventajas y desventajas. La investigación se centra en la evaluación en el proceso de aprendizaje y nos lleva a realizar una comparación referente a la dificultad que presentan los ítems en los diferentes tipos de formatos a los que los alumnos se enfrentan y determinar si la medida de su rendimiento se ve afectada por ello.

En el segundo capítulo se resumen algunos de los aspectos más relevantes de la Psicometría, entendida ésta como la ciencia que da soporte a las teorías vinculadas en la evaluación referida a procesos mentales. Se explica cuáles son las técnicas utilizadas como actos evaluativos y se definen los elementos fundamentales que dan soporte a un test.

En el tercer capítulo se explican dos de las más relevantes teorías de la evaluación. Se inicia con el modelo lineal clásico, su formulación y supuestos. Se hace referencia al coeficiente de fiabilidad, a los factores que la afectan y cómo se calcula. Posteriormente se describe el modelo de la teoría de la respuesta al ítem. La curva característica de un reactivo, sus propiedades y los diferentes tipos de modelos de aproximación de la curva. Se aborda el modelo de Rasch y se presentan los conceptos de lógito, escalograma de Guttman y error.

En el cuarto capítulo se presenta la metodología, que da soporte a la investigación. Se describe cómo se llevó a cabo el procesamiento de los datos y los dos métodos utilizados para su análisis, explicando paso a paso el desarrollo de cada uno de ellos, sobre la base de la teoría del modelo de Rasch, que son: la calibración manual y la calibración de ítems utilizando el programa de XCalibre.

En el capítulo quinto, para dar respuesta a las preguntas que dieron origen a la investigación, se presenta la interpretación de los resultados del procesamiento de los datos, analizados a través de herramientas estadísticas, permitiéndonos obtener conclusiones para el estudio presentado.

Al final del trabajo, se encuentra un anexo que contiene una serie de tablas que se fueron construyendo con los diferentes resultados del procesamiento de los datos.

Índice

CAPÍTULO 1. PLANTEAMIENTO DEL PROBLEMA

Como señala Santibáñez (2001) "la palabra evaluación se emplea con diferentes significados, contribuyendo así a una cierta confusión cuando se utiliza en el campo educativo" razón por la cual, daré inicio a este capítulo con una de las muchas definiciones que se han escrito a respecto, pero que da cuenta de la mejor manera lo realizado en esta investigación.

Así, la evaluación se define como "un proceso de obtención de información y de su uso para formular juicios que a su vez se utilizarán para tomar decisiones" (Tenbrink, F. 1981). Entre las formas de obtención de información se tienen: los exámenes, la observación de actividades, tareas, trabajos en equipo, preguntas en clase.

La evaluación es un factor pedagógico fundamental sin embargo, en algunos casos, la propia institución educativa no establece norma alguna más allá de fijar una cota mínima para efectos de acreditación, lo cual favorece a considerar a la evaluación como un factor de segundo orden que busca únicamente satisfacer los requisitos administrativos de la institución educativa en que se labore. Pero la evaluación juega un papel que va mucho más allá de estos requisitos. Dependiendo de las circunstancias, puede llegar a ser la herramienta más eficaz para el logro de los objetivos de un curso. Como elemento del sistema dentro del cual se enmarca el proceso de enseñanza-aprendizaje, la evaluación debe considerarse al mismo nivel que los objetivos, la metodología y el contenido. La evaluación es tanto un factor regulador, como un factor catalizador. (Zavala, P. 2001)

La evaluación, vista como un procedimiento, implica necesariamente la obtención, desde el punto de vista cualitativo y/o cuantitativo, de información necesaria para la toma de decisiones e implica en muchos casos una medición. Así, como señala Santibáñez (2001) "medir en la educación es la acción de recabar información y ordenarla considerando sus características cuantitativas numéricas sobre un aprendizaje determinado"

La evaluación debería estar basada en un principio fundamental: debe ser justa y para ser justa, debe pretender medir los objetivos del curso. Para ello es necesario, primero, que tenga definidos unos objetivos explícitos; segundo, que los estudiantes conozcan estos objetivos; y, tercero, que la metodología utilizada en el salón de clase y la actuación del profesor dentro del mismo estén de acuerdo con estos objetivos. La evaluación debe ser vista como un medio regulador de la calidad del trabajo desarrollado.(Woolfolk A. 1996).

Así, la justicia es un factor trascendental en el diseño de una evaluación. Sin embargo, ¿cómo se puede medir qué tan justa es una evaluación?. El examen es uno de los instrumentos del procedimiento de evaluación, que nos permite medir que es justo, esto es si mide el logro de

los objetivos cognitivos por parte de los estudiantes, entonces se, asume que los estudiantes conocen los objetivos, ya sea porque fueron explícitos o porque la actitud del profesor en clase los puso de manifiesto.

Los profesores tendemos a tener una concepción ideal del curso y podemos llegar a pensar que es a partir de esta concepción ideal que hay que medir al estudiante. Pero esto no es justo; lo que el estudiante vivió y conoció fue la realidad del desarrollo del curso, no lo que nosotros en el momento de la evaluación, consideramos que es lo ideal. Por consiguiente, hay que medir lo que se hizo en clase, no lo que nos habría gustado hacer, ni lo que pensamos que el estudiante excepcional pudo haber hecho por su cuenta.

Si bien hay que tener en cuenta que un examen se establecen jerarquías o niveles de aprendizaje que deben diferenciar a los estudiantes, dichos exámenes deberían ser justos en el sentido de que evalúan los objetivos del curso de acuerdo a lo hecho en clase. Con base en esa concepción ideal, los profesores tendemos a elaborar los exámenes que a posteriori, los alumnos o algunos de los profesores de la academia tienden a calificar de fáciles o difíciles. En principio, no debería haber exámenes difíciles o fáciles. Entonces, ¿cuándo es un examen fácil y cuándo es difícil? Un resultado que deberíamos esperar de un examen es que se comportara de manera normal, desde el punto de vista estadístico. En algunas ocasiones, hay exámenes para los cuales dos o más estudiantes sacan la máxima calificación aunque no necesariamente saben todos lo mismo, y por lo tanto, no estaríamos diferenciando apropiadamente a los alumnos que obtengan esa calificación máxima, aunque con posibilidad de contar con conocimientos y preparación diferente. Lo mismo sucede con la calificación mínima puesto que no podemos diferenciar entre el estudiante que sabe algo y aquél que no sabe nada.

Una de las dimensiones relevantes de la evaluación educativa, se refiere a todos aquellos instrumentos y procedimientos que suelen utilizarse en las distintas prácticas educativas (Berliner (1987); citado en Díaz, B. y Hernández G 1998)y se ha propuesto una clasificación en términos del grado de formalidad y estructuración con que se establecen las evaluaciones.

Para realizar una evaluación, uno de los instrumentos más utilizados es la prueba o examen. Se define a los exámenes, como aquellas situaciones donde se intenta verificar el grado de rendimiento o aprendizaje logrado por los aprendices. Los exámenes son recursos que han aparecido en el ámbito educativo con la intención de lograr una evaluación objetiva, libre lo más posible de interpretaciones idiosincráticas al establecer juicios sobre los aprendizajes de los alumnos. Otra característica adicional asociada al examen es la supuesta posibilidad de cuantificar el grado de rendimiento o aprendizaje a través de calificaciones consistentes en

número. Una prueba es una observación cuantitativa del proceso de un equipo, ya que tanto el alumno como el profesor están siendo observados.(Rodríguez, y García, 1999).

Los objetivos de una prueba pueden ser, entre otros: (Rodríguez, y García 1999).

- Evaluar el nivel de progreso individual en una escala relativa con respecto a los demás alumnos.
- Evaluar la eficiencia del profesor.
- Motivar al estudiante a estudiar. Los estudiantes estudian más cuando saben que van a ser examinados.
- Servir de diagnóstico: localizar dónde hay necesidad de instrucción adicional, o dónde los métodos de enseñanza pueden ser modificados por no haber sido efectivos.
- Facilitar una enseñanza inmediata: cuando se examina al estudiante y se le devuelve el examen bien corregido, señalándole lo que debió responder, es obvio que aprenda a corregir las reglas mal aprendidas y por consecuencia, mal empleadas.

Las cualidades que debe cumplir una buena prueba son: (Rodríguez, y García 1999).

- Sea justa: lo suficientemente difícil para que ningún alumno obtenga la calificación más alta y lo suficientemente fácil para que ningún alumno obtenga cero.
- Mida con exactitud la comprensión y las habilidades del alumno.
- Las preguntas y las instrucciones claras , concisas y completas.
- Sea fácil de aplicar, fácil de captar, fácil de corregir y fácil de calificar.
- Las preguntas sean valoradas con imparcialidad y precisión.

Las fuentes básicas de información para la elaboración para todas las pruebas, cualesquiera que sean sus objetivos, las constituyen: (Rodríguez, y García 1999).

- La labor realizada en clase: las pruebas deben contener siempre preguntas sobre el material examinado en clase y de cualquier material de enriquecimiento: proyectos, informes especiales, acontecimientos actuales, etc.
- El texto de apoyo para los alumnos: debe señalarse en concreto el contenido de los libros (capítulos, unidades, secciones, etc.) que será incluido en la prueba.
- La parte del programa: otras veces se señala como contenido de la prueba determinada parte del programa, que debe ser estudiada en diferentes libros.
- Los objetivos específicos de la materia cobran relevancia cuando el examen se califica con base a criterio.

Se debe tener presente que, al seleccionar las fuentes para las pruebas, es importante no sólo el contenido, sino que muchas de ellas proporcionan ideas interesantes para la formulación de preguntas.

De acuerdo a Díaz B. y Hernández G. (1998) el grado de estructuración de la serie de reactivos, ítems o preguntas que conforman en general un examen, influye de manera significativa en el tipo de procesos mentales involucrados en la solución de la prueba. Con base al grado de estructuración de un examen, éste se puede clasificar como estructurado o no estructurado. Entre los primeros se encontrarían los llamados tests objetivos, como por ejemplo los de opción múltiple, en tanto, entre los segundos se ubicarían aquéllos de respuesta abierta.

Dado que existen *creencias* que descalifican a priori los tests objetivos, particularmente los de formato de opción múltiple, en razón de que se cree que dicho formato no logra medir desempeños de los individuos a niveles que, de acuerdo a la taxonomía de Bloom, corresponderían a los de análisis, aplicación y síntesis y, en contrapartida, existe otra corriente totalmente antagónica que señala que los tests debieran ser exclusivamente de formato de opción múltiple argumentando que dichos tests son independientes del criterio del evaluador, la investigación se ha abocado a establecer si existen o no diferencias estadísticamente significativas entre los resultados obtenidos en uno u otro formato.

De acuerdo con Rodríguez y García (1999) no existe un tipo de prueba mejor que otro. Todas las pruebas tienen sus ventajas y desventajas, y se prestan para determinadas situaciones. Es precisamente el grado de ventajas y la situación concreta la que nos hace preferir un tipo u otro, pero debe ser evidente que sobre todo se debería de eliminar la subjetividad asociada a los juicios que etiquetan a unos reactivos como fáciles y otros como difíciles; para lograr esto último, se deberá contar con exámenes cuyos ítems hayan pasado previamente por un proceso de calibración.

La calibración de la dificultad de ítems de pruebas debe ser independiente de las personas en particular los que se usen para la calibración. La medición de la habilidad de la persona debe ser independiente de los ítems de prueba en particular que se usen en la medición.

Cuando comparamos un ítem con otro para calibrar una prueba no debe importar de quién son las respuestas a esos ítems que usamos para la comparación. La única forma en la que podremos construir pruebas que contengan un significado uniforme sin importar a quién vayamos a medir con ellas impone que, ya calibradas, deben darnos los mismos resultados sin importar a quién se la administremos.

De igual forma, si se quiere exponer a las personas a una selección de ítems de prueba para medir su habilidad, no debe importar qué selección de ítems usemos o qué ítems contesten. Debemos de poder comparar personas, para llegar a mediciones estadísticamente equivalentes de habilidad, sin importar qué tipos de ítems se hayan tomado, aun cuando hayan sido medidos con pruebas totalmente distintas.

Los investigadores y evaluadores que apoyan firmemente los exámenes objetivos dicen que éstos son mejores porque: (Rodríguez y García, 1999)

- Miden la capacidad para resolver problemas novedosos.
- Aíslan capacidades específicas relativas a la materia de las destrezas generales de redacción, caligrafía y uso propio del lenguaje.
- Poseen valor potencial para diagnosticar.
- Muestrean adecuadamente los objetivos de la enseñanza
- Muestrean adecuadamente el contenido de la enseñanza.
- Los diversos calificadores dan puntajes consistentes.
- Distinguen con precisión niveles de competencia entre los examinados.
- Pueden calificarlos una máquina o un empleado.
- Pueden calificarse rápidamente.

Además señalan como desventajas de los exámenes por temas o de ensayo:

- Falta de tiempo para calificarlos a diferencia de las pruebas objetivas, las pruebas por temas no se prestan a ser calificadas por una secretaria o un empleado. Por su propia naturaleza, requieren del juicio de un experto en la materia.
- La valoración cuidadosa de un ensayo requiere para su calificación 15 o 20 minutos, (sic) por lo que el tiempo invertido es muy significativo.
- Falta de confiabilidad de contenido. Se limita a un número corto de preguntas, por tanto no puede haber un buen muestreo del contenido: el número de preguntas no es representativo del material estudiado ni del área de conocimientos o de las formas de conducta que se tratan de medir.
- Falta de confiabilidad del calificador. Si en forma independiente dos profesores de la misma materia calificaran el mismo grupo de exámenes provenientes de un tercer profesor, las calificaciones revelarían graves diferencias

En contrapartida, los opositores a los exámenes objetivos señalan a favor de las pruebas tipo ensayo o exámenes por temas.

- Se examina lo que se recuerda. En este tipo de prueba el alumno sólo tiene la oportunidad de "reconocer" sino de recordar la información. El alumno debe dominar el tema para salir bien.
- Examina procesos mentales de alto nivel. Se pretende que este tipo de prueba involucra procesos mentales de alto nivel (complejos) tales como: capacidad de pensar, de razonar, de conceptuar, inducir, deducir, imaginar.
- Examinan la originalidad y la creatividad. Los partidarios de este tipo de examen sostienen que esta clase de prueba favorece la originalidad y la creatividad porque además de que el alumno utiliza sus propios recursos para resolverla, implica la creación de algo nuevo o la combinación de derivados ya viejos pero con la existencia de limitaciones mínimas.

Además, señalan como desventajas de los exámenes objetivos:

- *Per se* no garantizan que el diseño sea adecuado.
- Deshumanizan el proceso educativo, convirtiendo la evaluación en una acción mecánica.
- No garantizan la eliminación del factor adivinación.
- Permiten que el alumno reciba ayuda más fácilmente.

Así, nos planteamos una vez más la pregunta esencial a la cual se ha deseado dar respuesta a través de la presente investigación ***¿Existen diferencias estadísticas significativas en el valor calculado del índice de dificultad de los ítems ante dos formatos diferentes de tests, uno de ellos de opción múltiple y otro de respuesta abierta o tipo ensayo?*** Para dar respuesta a esta pregunta, ***se aplicó una prueba de hipótesis para aceptar o no si el índice de dificultad calculado se veía afectado por los dos formatos diferentes.***

Además de la polémica ya mencionada con relación al tipo de formato y su impacto en las evaluaciones, existe un problema no menos importante, relativo a la escala que debe usarse para asignar un lugar dentro de una gama posible de *calificaciones*. Una alternativa desarrollada a partir de los años ochenta, la proporcionó Rasch, cuya ventaja sobre otras alternativas es la de colocar en una misma escala la dificultad de los ítems y la capacidad de las personas. Sin embargo, una limitante mencionada en la literatura existente, está vinculada

con el tamaño de muestra para su uso. Así, ¿qué pasa si un profesor hace uso del modelo de Rasch en grupos de alrededor de 100 estudiantes o en grupos aún menores?

Una segunda pregunta a resolver, será ***¿Cuáles son los efectos sobre el índice de dificultad de los ítems analizados a través de la teoría de la Respuesta del ítem, cuando los tamaños de muestra son relativamente pequeños?***

CAPÍTULO 2. PSICOMETRÍA

Introducción

La **Psicometría** es el conjunto de métodos, técnicas y teorías implicadas en la medición de variables psicológicas. Trata con todo aquello que se relaciona con la medición de lo psicológico. Lo específico de la Psicometría sería su énfasis y especialización en aquellas propiedades métricas exigibles a las mediciones psicológicas independientemente del campo sustantivo de aplicación y de los instrumentos utilizados, es decir, en aspectos como la fiabilidad y la validez de las mediciones que forman parte de cualquier evaluación psicológica.

De acuerdo con las sociedades psicométricas, europea y americana, la mayor parte de la temática psicométrica se encuentra articulada en cinco grandes bloques[1]:(ver Muñiz, J 1990).

- Teoría de la medición, que abarca todo lo relativo a la fundamentación teórica de la medida.
- Teoría de los tests, donde se explicita a la lógica y los modelos matemáticos subyacentes a la construcción y uso de los tests.
- Escalamiento psicológico, que aborda la problemática inherente al escalamiento de estímulos psicológicos.
- Escalamiento psicofísico, que hace lo propio con los estímulos físicos.
- Técnicas multivariadas, que junto con el resto de tecnología estadística resultan imprescindibles para la construcción y análisis de los instrumentos de la medida.

Los orígenes remotos de los primeros tests podrían rastrearse, según Du Bois (1970); citado en Muñiz, J. (1996). en el año 300 a. de C. cuando los emperadores chinos evaluaban la competencia profesional de sus oficiales. Pero los orígenes más cercanos que darán lugar a los actuales tests hay que ubicarlos en aquellas primeras pruebas senso-motoras utilizadas por Galton, en su laboratorio antropométrico de Kensisngton. En 1884, durante una exhibición internacional de la salud, Galton medía una serie de índices antropométricos y sensomotores y luego trataba de relacionarlos con el funcionamiento intelectual, lo cual no tuvo conexión alguna. Pero se debe a Galton el honor de ser el primero en utilizar la tecnología estadística para analizar los datos provenientes de sus tests, labor que continuaría Pearson.

---

[1] Dentro de esta agrupación de teorías, es precisamente la teoría de los test la que da marco a la investigación.

James Mc Keen Cattell será el primero en utilizar el término Tests Mental en su artículo Mental Tests and Measurement publicado en la revista Mind en 1890, pero sus test, al igual que los de Galton, dejaron en forma clara la nula correlación entre este tipo de pruebas y el nivel de inteligencia.

Será Binet quien de un giro radical en la filosofía de los tests al introducir en su escala tareas de carácter más cognoscitivo encaminadas a evaluar aspectos como el juicio, la comprensión y el razonamiento, que según él constituían los componentes fundamentales del comportamiento inteligente.

La revisión de la escala se llevó a cabo en la Universidad de Stanford y se conoce como la revisión de Stanford–Binet. Se utilizó por primera vez el cociente intelectual, CI,. para expresar puntuaciones de los sujetos. La idea era originaria de Stern que en 1911 propuso dividir la Edad mental, EM, entre la edad cronológica, EC, multiplicando por 100.

$$CI = \frac{EM}{EC} \times 100 \qquad (2.1)$$

El éxito de estas escalas para predecir el rendimiento escolar se debe al parecido de las tareas exigidas por ambos lados, escala y escuela. El siguiente paso en la historia se verá marcado por la aparición de los tests de inteligencia colectivos, propiciados por la necesidad del ejército norteamericano en 1917 de seleccionar y clasificar a los soldados que iban a tomar parte en la Primera Guerra Mundial. Un comité dirigido por Yerkes diseñó los tests Alfa y Beta, el primero para la población en general y el segundo para utilizar con analfabetos.

Thurstone (1937), citado en Muñiz, J.(1996). utiliza el término Psicología Matemática para caracterizar el objeto de la sociedad psicométrica americana fundada por él. La diferencia entre "la Psicología Matemática respecto a la Psicometría serán los modelos matemáticos elaborados para áreas específicas de la Psicología, tales como aprendizaje, memoria, percepción, lenguaje, pensamiento, interacción social" (ver Muñiz, J. 1996)

Con el tiempo la Psicometría y la Psicología Matemáticas compartían campos de trabajo, ya que los especialistas asistían indistintamente a los congresos al igual que las publicaciones fueron indistintas. Al considerar la Psicometría como la parte de la Psicología Matemática dedicada a todo lo relacionado con la medida, de igual manera se puede considerar a la Psicología Matemática como aquella rama de la Psicometría, dedicada a los modelos matemáticos de procesos psicológicos. Thurstone en la primera reunión anual de la sociedad psicométrica que tuvo lugar en 1936 señala como objetivo de la sociedad "El estimular el desarrollo de la Psicología como ciencia cuantitativa y racional". O lo que más brevemente puede denominarse la Psicología Matemática añade además, algo que se olvida demasiado a

menudo "A la larga seremos juzgados por la significación, relevancia y consistencia de los principios psicológicos que descubramos".

Se desarrollan diferentes denominaciones específicas de las distintas subáreas.

El origen de la *Teoría Clásica de los Tests*, TCT, puede ubicarse en los primeros trabajos de Spearman a principios del siglo XX en los que se establecen los fundamentos de ésta. El objetivo central era encontrar un modelo estadístico que fundamentase adecuadamente las puntuaciones de los tests y permitiera la estimación de los errores de la medida asociados a todo proceso de medición. En 1927 Spearman aportó un notable avance para la construcción, análisis y validación de los tests

El modelo lineal clásico propuesto por Spearman, asume que la puntuación empírica de un sujeto en un test, X, consta de dos componentes aditivos; uno la Verdadera puntuación del sujeto en el test, V, y otro el error, e, que inevitablemente va asociado a todo proceso de medición; es decir según el modelo, $X = V + e$. A partir de este modelo, la teoría clásica desarrollará todo un conjunto de deducciones encaminadas a estimar la cuantía del error que afecta a las puntuaciones de los tests.

En 1936, se funda la sociedad psicométrica americana con Thurstone a la cabeza y cuyo órgano de expresión será la revista Psychometrika. En 1947 Thurstone publica su clásico texto Análisis Factorial Múltiple, Técnica estadística con orígenes en el campo psicométrico.

En 1946 Stevens publica su trabajo sobre las escalas de medida, que obligará a los estudios de la teoría de los tests a plantearse el estatus teórico de sus mediciones además de sus propiedades empíricas.

La Teoría Clásica de los Tests será realizada por Gulliksen (1950); citado en Muñiz, J. (1996). en su clásico libro de Theory of Menthal Test. También el escalamiento tendrá su clásico en los años cincuenta con el libro de Torgerson 1958 Theory and Method of Scaling. Las dos ramas hermanas, Teoría de los Tests y Escalamiento, seguirán su propio camino, aunque en ésta como en otras divisiones hay algo arbitrario, pues la mayoría de los modelos podrían generalizarse.

En 1968 aparecería el libro de Lord y Novick Statistical Theories of Mental Test Scores que sintetiza y reanaliza críticamente lo realizado en la teoría clásica de los tests e incluye los trabajos previos Birnbaum sobre los Modelos de Rasgo Latente que abriría una línea completamente nueva en la teoría de los Tests, conocida hoy día como Teoría de Respuesta a los Ítems, TRI, misma que según algunos teóricos eclipsaría al planteamiento clásico de la Teoría de la Generalizabilidad, TG, propuesta por Cronbach.

La Teoría Clásica se encontraba con dos problemas de fondo importantes que no encontraban solución satisfactoria en su marco teórico. Por un lado la medición de las variables no era independiente del instrumento utilizado, y por el otro las propiedades de los instrumentos dependían del tipo de sujetos utilizados para establecerlas. Por lo tanto no eran propiedades realmente de los instrumentos de medida, sino de la integración de éstos con los objetos medidos.

Los modelos, TRI, permitirían dar una solución adecuada a esos problemas de fondo, y además aportarían todo un conjunto de avances tecnológicos complementarios para la construcción y análisis de los tests.

Lord (1980); citado en Frank, B. (1992), pudo sintetizar en un libro hoy clásico los avances acumulados. El libro abre la década de los ochenta que conocerá una expansión de la literatura psicométrica bajo la óptica de la, TRI, y revitalizará áreas que se encontraban atascadas, tales como los bancos de ítems, el sesgo o los tests referidos al criterio.

Técnicas en los actos evaluativos

En los actos evaluativos se pueden utilizar tres tipos de técnicas básicas: informales, semiformales y formales.

Técnicas Informales . Se utilizan dentro de episodios de enseñanza (generalmente cortos) con una duración breve. Como exigen poco gasto didáctico, pueden utilizarse a discreción con la situación de enseñanza o de aprendizaje. Esta técnica no se presenta explícitamente a los alumnos como actos evaluativos y se identifican dos tipos:

- Observación de la actividades realizadas por los alumnos.
- Exploración a través de preguntas formuladas por el profesor durante la clase.

Estas dos formas de evaluación se utilizan por el profesor en su práctica magisterial; sin embargo durante mucho tiempo han sido desdeñadas por aquellos enfoques que insisten demasiado en los resultado finales del aprendizaje. Uno de los principales problemas que se les imputa es su bajo nivel de validez y confiabilidad lograda.

Técnicas Semiformales. Se caracterizan por requerir de un mayor tiempo de preparación que las informales, demandan mayor tiempo para su valoración y se exigen a los alumnos respuestas más duraderas, lo cual hace que en estas actividades sí se les impongan calificaciones, por lo que los alumnos sí las perciben como actividades de evaluación, en comparación con las técnicas informales (Díaz, B. y Hernández, G 1998).

Algunas de las variantes de las técnicas semiformales de evaluación son:

- Ejercicios y prácticas que los alumnos realizan en clase.
- Tareas que los profesores encomiendan a sus alumnos para realizarlas fuera de clase.

Técnicas Formales. El tercer grupo de procedimientos o instrumentos más utilizados para realizar una evaluación son los que se agrupan bajo el rubro de técnicas formales. Dichas técnicas exigen un proceso de planeación y elaboración más sofisticado y suelen aplicarse en situaciones que demandan mayor grado de control (Genvard y Gotzens, 1990;citado en Díaz B, y Hernández G.1998). Por esta razón los alumnos (y los profesores inducen a ello) los perciben como situaciones verdaderas de evaluación. Este tipo de Técnicas suelen utilizarse en forma periódica o al finalizar un ciclo completo de enseñanza y aprendizaje. Dentro de ellas encontramos varias modalidades:

- Pruebas o exámenes tipo test.
- Mapas conceptuales.
- Pruebas de ejecución.
- Lista de cotejo o verificación y escalas.

Los instrumentos más utilizados para realizar una evaluación son las pruebas o exámenes, bajo el supuesto de que es posible cuantificar el grado de rendimiento o aprendizaje, es decir, asignar un número que caracterice, a través de calificaciones, el grado de rendimiento o aprendizaje que los alumnos han logrado. Debe tenerse presente que las pruebas son recursos que han surgido con la intención de lograr una evaluación objetiva.

En la metodología de su elaboración se pone énfasis en que contengan un nivel satisfactorio de validez (es decir que los instrumentos sirvan para valorar aquello para lo cual han sido construidos) y de confiabilidad (que su aplicación en condiciones similares permita obtener resultados similares) para su uso posterior.

Si se utilizan o no, conceptos estadísticos, como pudieran ser la media, la moda, la desviación estándar, etc. para diferenciar el comportamiento de los alumnos ante los exámenes, las pruebas se pueden clasificar en dos tipos: las estandarizadas (que por lo general los elaboran especialistas en evaluación) y las formuladas, en general por los profesores, quienes interpretan en forma personal las necesidades del proceso pedagógico. Esas dos modalidades también coinciden con dos tipos de juicios o interpretaciones que se pueden establecer a partir de los puntajes resultantes. Así, las evaluaciones podrán estar basadas en normas o en criterios (evaluaciones normativas o criteriales).

Las evaluaciones normativas comparan a un sujeto contra su grupo de referencia, que en muchas ocasiones suele ser el propio grupo-clase. La evaluación criterial compara el desempeño de los alumnos contra ciertos criterios diseñados previamente, preferentemente por un grupo de expertos, con base en los objetivos o intenciones educativas. De hecho se dice que un instrumento de evaluación criterial se utiliza para estimar el estatus o lugar de un aprendiz en relación a un dominio (conceptual, procedimental etcétera) que previamente ha sido definido del modo más veraz y objetivo posible.

Generalmente, los exámenes están construidos por un conjunto de ítems, cuyo nivel de estructuración determina de manera importante el tipo de procesos cognitivos y de aprendizajes significativos que logran los alumnos.

De acuerdo con Díaz B. y Hernández G (1998), los reactivos de alto nivel de estructuración como son, entre otros, los de falso y verdadero y correspondencia, de manera evidente exigen a los alumnos principalmente el reconocimiento de la información. Los reactivos de respuesta breve o de complementación y los de opción múltiple demandan por lo general el recuerdo de la información (proceso más sofisticado que el de reconocimiento) aunque si son elaborados a la perfección pueden valorar niveles de comprensión (parafraseo reproductivo y productivo) y hasta la aplicación de los conocimientos.

Los reactivos típicos de las llamadas"pruebas objetivas" tienen, entre otras, las siguientes características:

- Pueden ser calificados e interpretados con mucha rapidez o precisión.
- Su diseño no es tan sencillo como parece.
- La elección de los reactivos o de las respuestas de éstos por parte del diseñador no está exenta de subjetividad.
- En un breve período puede responderse un número considerable de reactivos.
- No permiten valorar habilidades complejas: creatividad, capacidades de comunicación o de expresión oral, elaboración de argumentos.
- Gran parte de los reactivos pueden responderse por medio de aprendizajes memorísticos o de aprendizajes poco significativos.

Otros reactivos que demandan una evaluación cualitativa y no cuantitativa como en los casos anteriores son los de respuesta abierta y los de desarrollo de temas que a diferencia de los anteriores, demandan actividades de mayor complejidad y procesamiento tales como:

- Comprensión.
- Elaboración conceptual.
- Capacidad de integración.

- Creatividad.
- Habilidades comunicativas.
- Capacidades de análisis.
- Establecimiento de juicios reflexivos o críticos.

Es evidente que para poder calificarlos el docente debe establecer juicios o interpretaciones cualitativas que muchas veces suelen estar cargados de dosis significativas de subjetividad. Sin embargo para garantizar un cierto nivel de objetividad en las calificaciones pueden aplicarse listas o catálogos de criterios sobre las repuestas o producciones solicitadas.

Uno de los recursos a los que el profesorado recurre para basar la elaboración de los reactivos es la taxonomía cognitiva de los objetivos, propuesta por B. Bloom y colaboradores. Según dicha taxonomía, elaborada a finales de los cincuenta, se pueden clasificar los objetivos de un programa, curso, etcétera, en función de seis niveles de complejidad creciente, a saber:

- Conocimiento: que implica recuerdo y retención literal de la información enseñada.
- Comprensión: entendimiento de los aspectos semánticos de la información enseñada.
- Aplicación: utilización de la información enseñada.
- Análisis: de estudio de la información enseñada en sus partes constitutivas.
- Síntesis: combinación creativa de partes de información enseñadas para formar un todo original.
- Evaluación: emisión de juicios sobre el valor del material enseñado.

Fiabilidad.

Un instrumento de medida, ya fuese el caso de un test o una escala, se considera fiable si las medidas que se hacen con él, carecen de errores, es decir son consistentes. Un test será fiable si cada vez que se aplica a los mismos sujetos da el mismo resultado; el problema con los seres humanos, es que éstos cambian de una vez a otra, y en ocasiones puede resultar problemático saber con seguridad si la inestabilidad observada en las mediciones se debe a la imprecisión del instrumento o a los cambios legítimos operados por los sujetos. Los errores de medida de los que se ocupa la fiabilidad son aquellos no sometidos a control e inevitables en todo proceso de medir, sea físico, químico o psicológico. En ocasiones las diferencias entre una medición y otra no dependen sólo de estos errores, pudiendo explicarse además por los cambios operados en los sujetos, debidos a procesos madurativos, intervenciones o eventos de cualquier otro tipo. Las inconsistencias pueden tener sentido en el marco en el que se lleva a cabo la medición. En estos casos la inestabilidad de las mediciones requiere de una

explicación y carece de sentido atribuirla a los errores aleatorios. La fiabilidad no trata ese tipo de "errores".

No se deberá confundir la fiabilidad del instrumento de medida con la estabilidad de las mediciones cuando no existen razones teóricas ni empíricas para suponer que la variable a medir haya sido modificada diferencialmente para los sujetos, por lo que se asume su estabilidad, mientras no se demuestre lo contrario. Un test no sería fiable si cada día generase mediciones diversas de una variable que se supone estable. Ahora bien lo que es válido para una determinada variable. no tiene por qué serlo para otras variables. Así la fiabilidad del instrumento no va unida a la estabilidad de la variable medida a lo largo del tiempo.

Validez

Concepto y tipos.

Un test es una muestra de conducta a partir de la cual se pretende hacer ciertas inferencias. La validez se refiere al conjunto de pruebas y datos que han de recogerse para garantizar la pertinencia de tales inferencias. Más que el test, lo que se validan son las inferencias. El problema de hallar la validez de un test es el problema general de la ciencia para validar una teoría. Implica, por tanto, la utilización de los métodos y procedimientos habituales de la investigación científica.

Para llevar a cabo el proceso de la validación de los tests, de acuerdo a Muñiz, J. (1990) se tienen tres grandes categorías:

- Validez de contenido.
- Validez predictiva.
- Validez de constructo.

Esta división tripartita tiene interés didáctico para dar una panorámica de los estudios de validez, pero no deben tomarse como categorías independientes ni exhaustivas, ya que los tres tipos de validez están relacionados, es decir, son facetas de un todo.

Validez de Contenido.

La validez de contenido recurre a la necesidad de garantizar que el test constituye una muestra adecuada y representativa de los contenidos que se pretende evaluar con él. Si la población de contenido está claramente definida, como suele ocurrir con los tests de carácter educativo en los que se puede explicitar con precisión la materia objeto de medición, entonces no hay ningún problema; los diferentes métodos estadísticos de muestreo permiten extraer una

muestra representativa de los contenidos que han de formar el test. Una de las prácticas más usuales y obvias para lograr lo anterior, consiste en hacer una tabla de especificaciones, es decir listar todas las áreas de contenido, que se consideran importantes y/o imprescindibles y asegurarse que la prueba contiene los ítems de todas ellas en la proporción adecuada. Tómese en cuenta que una adecuada validez de contenido es fundamental para cualquier generalización o inferencia que se pretenda hacer a partir del test; se trata, en suma, de un caso particular del más general relativo al muestreo.

Suele incluirse como tipo especial de validez de contenido la validez aparente, que se refiere a la necesidad de que el test aparente medir lo que se pretende, dé la impresión a los que se aplica que efectivamente es adecuado. Como se puede observar es un tipo muy curioso de validez, que puede tener su importancia de cara a la motivación y actitud de los sujetos, pues si por cualquier razón, lo que están haciendo no les parece conectado con el fin perseguido, posiblemente no se desplegarán todas sus posibilidades.

Validez Predictiva.

Uno de los usos más frecuentes de los tests está relacionado con la predicción, a partir de ellos, de alguna variable de interés o criterio. Se utilizan, por ejemplo, en una institución académica como criterio para la admisión de candidatos si se dispone de datos que avalen la conexión entre las puntuaciones en los tests (de carácter predictivo) y el éxito académico posterior. Así entonces, la validez predictiva de un test se refiere al grado de eficacia con el que se puede predecir o pronosticar una variable de interés (criterio) a partir de las puntuaciones de ese test. La validez, se opera mediante un coeficiente (de validez), que es la correlación entre el test y el criterio. Lógicamente cuanto mayor sea la correlación test-criterio más precisos serán los pronósticos hechos a partir del test.

La correlación y la predicción no necesariamente implican causalidad, sólo implican lo señalado, es decir la covariación. La validez predictiva recibe a menudo las denominaciones de validez relativa al criterio, validez criterial o validez de pronóstico.

El cálculo del coeficiente de validez, en principio, no entraña dificultad alguna. Se trata de hallar la correlación entre las puntuaciones de los sujetos en el test y las que obtengan en el criterio. Los problemas empiezan a la hora de obtener las puntuaciones del criterio, dado que los criterios de interés suelen ser complejos y en ocasiones difíciles de definir unívocamente; algunas recomendaciones prácticas de interés a la hora de operar la medida del criterio pueden consultarse, por ejemplo en Thorndike (1982) o Crocker y Algina (1986); citado en Muñiz, J.(1996).

Según el diseño utilizado para calcular el coeficiente de validez, puede hablarse de validez concurrente cuando el test y el criterio se miden al mismo tiempo; validez de pronóstico, cuando el criterio se mide con un período de tiempo después del test; validez retrospectiva, cuando se aplica el test un cierto tiempo después del criterio que se desea pronosticar. El uso de un diseño u otro dependerá del problema planteado, por ejemplo en una institución educativa empeñada en que allí sólo entren a realizar sus estudios, de cinco años de duración, aquellos candidatos que tengan una alta probabilidad de éxito final (criterio); un test tendría validez de pronóstico si aplicado a la hora del ingreso correlaciona altamente el éxito final, es decir pronostica el criterio medido cinco años después, Sí así fuese sería razonable utilizarlo en la selección de candidatos.

Validez de constructo.

Debe entenderse que un test no es un agregado de ítems que se juntan al azar para predecir un criterio, es más bien una medida o índice de un concepto, teoría o constructo psicológico, o de otro tipo. Por ejemplo, un test de extraversión constituirá un índice, indicador o medida del constructo psicológico de extraversión. Es bien cierto que en demasiadas ocasiones los constructores de tests se han preocupado de la teoría psicológica sustentadora de sus pruebas, limitándose a construir, como fuese buenos predictores empíricos de los criterios más solicitados socialmente, disponiendo, si acaso, teorías explicativas a posteriori. La validez de constructo se refiere a la recogida de evidencia empírica que garantice la existencia de un constructo psicológico —educativo— en las condiciones exigibles a cualquier otro modelo o teoría científica.

Una forma de determinar la validez de constructo es utilizando el análisis factorial que es una técnica de análisis multivariado que bajo determinadas condiciones, y con ciertas limitaciones, permite estimar los factores que dan cuenta de un conjunto de variables. Por ejemplo si aplicamos ***n*** tests, tal vez la información que obtenemos con ellos pueda venir explicada por factores cuyo número será menor que *n*, para lo cual, es condición imprescindible que la información proporcionada por los *n* tests sea redundante, es decir, que los tests estén correlacionados entre sí. Dado que es frecuente que las medidas psicológicas correlacionen entre sí, será posible reducir el número de variables a un número menor de factores y encontrar así explicaciones y modelos más parsimoniosos.

Los factores así obtenidos son objetos matemáticos a los que se puede encontrar una cierta interpretación psicológica plausible a partir de las variables que los componen. Constituyen

constructos provisionales, que necesitan de ulteriores confirmaciones por otros caminos experimentales además del factorial.

Básicamente, la validez factorial se refiere a lo siguiente: imagínese que para medir la extraversión se dispone de cuatro tests. Se diría que estos cuatro tests tienen validez factorial si aplicados junto con otros diseñados para evaluar constructos diferentes, tales como neuroticismo, paranoidismo o dogmatismo, y posteriormente sometidos todos ellos conjuntamente a un análisis factorial, los cuatro tests de extraversión componen un solo factor frente a los factores formados por los otros tests. También se podrían analizar los tests de extraversión y ver si forman un solo factor o más, lo que daría una idea de la cohesión del constructo.

Análogamente, es muy frecuente indagar la validez factorial de los ítems de un test. Si el test está destinado a medir un rasgo unitario, es de esperar que sometidos los ítems que componen a un análisis factorial se agrupen en un solo factor, y el grado en que esto ocurre determina la validez factorial del test en función de los ítems.

Elementos fundamentales de un test.

Análisis de los ítems

En el desarrollo del presente trabajo usaremos como sinónimo las palabras estímulo, reactivo, y pregunta. Debe ser claro que a las personas no las medimos directamente, en cambio lo que sí medimos es algún rasgo, habilidad, capacidad o aspecto que pueda presentar la persona: lo que se denomina rasgo latente y que es explorado por medio de una muestra de reactivos tomados del dominio en estudio. El proceso real de construcción de un test implica elaborar un número elevado de ítems, dos o tres veces más de los que el test tendrá finalmente, aplicar esos ítems a una muestra de sujetos semejantes a los que el test irá destinado y descartar aquéllos que no sean pertinentes. El objetivo central del análisis de ítems es cómo saber qué ítems son pertinentes.

Se entiende por análisis de ítems el estudio de aquellas propiedades de los ítems que están directamente relacionadas con las propiedades del test, y, en consecuencia, influyen en ellas. En palabras de Lord y Novick (1968); citado en Muñiz, J. (1996), el requerimiento básico de un ítem es que tenga una relación clara con algún parámetro interesante del test total como lo son los siguientes tres índices:

1. Índice de dificultad
2. Índice de discriminación
3. Índice de validez

Índice de dificultad

Se entiende por índice de dificultad, *ID,* de un ítem la proporción de sujetos que los aciertan de aquellos que han intentado resolverlo:

donde:
$$ID = \frac{A}{N} \qquad (2.2)$$

A: Número de sujetos que aciertan el ítem

N: Número de sujetos que han intentado resolver el ítem.

El valor del índice de dificultad está directamente relacionado con la media del test:

$$\bar{X} = \sum_{i=1}^{n} ID_i \qquad (2.3)$$

En palabras, la media del test es igual a la suma de los índices de dificultad de los ítems.

Ejemplo: Tabla 2.1

	Ítems				Puntuación
Sujetos	1	2	3	4	Total
A	0	1	1	1	3
B	1	0	1	0	2
C	1	1	0	0	2
D	1	1	1	1	4
E	0	1	0	0	1
ID	3/5	4/5	3/5	2/5	12

$$\bar{X} = \sum_{i=1}^{4} ID_i = \frac{3}{5} + \frac{4}{5} + \frac{3}{5} + \frac{2}{5} = \frac{12}{5} = 2.4$$

Al índice de dificultad sería matemáticamente más apropiado denominarlo índice de facilidad, pues a medida que aumenta indica que el ítem es más fácil, no más difícil.

Nótese también que en muchos tests no tiene sentido hallar el índice de dificultad. Por ejemplo, en tests dirigidos a evaluar aspectos de personalidad, en los que los ítems no son fáciles ni difíciles.

Una seria limitación de este índice de dificultad de la teoría clásica es su dependencia directa de la muestra de sujetos en la que se calcula, es decir, el índice de dificultad no constituye una

propiedad intrínseca del ítem, su valor depende del tipo de sujetos a los que se aplique. Si son muy competentes resultará un ítem fácil, lo aciertan muchos. Si por el contrario, son incompetentes, el mismo ítem resultará difícil. A nivel práctico, la teoría clásica mitiga este inconveniente calculando el índice de dificultad en muestras similares en competencia a aquéllas en las que se va a usar posteriormente el ítem. Ahora bien este recurso resulta poco convincente a nivel teórico para una teoría de medición psicológica medianamente rigurosa, donde sería de esperar que las propiedades de los instrumentos de medida no dependiesen de los objetos medidos. Una solución adecuada a este problema la proporcionarán los modelos de Teoría de la Respuesta a los Ítems.

Cuando los ítems son de elección múltiple y, en consecuencia, es posible acertarlos por mero azar, el Índice de Dificultad, *ID,* conviene calcularlo corrigiendo los efectos al azar mediante la fórmula clásica que se presenta a continuación, aunque otras son también posibles:

$$ID = \frac{A - E/(K-1)}{N} \qquad (2.4)$$

donde:

A: Número de sujetos que aciertan el ítem.

E: Número de sujetos que fallan el ítem.

K: Número de alternativas del ítem.

N: Número de sujetos que intentan resolver el ítem.

La varianza de un ítem puede expresarse en términos de su índice de dificultad, puesto que para una variable dicotómica, *j*: $\sigma_j^2 = P_j Q_j$, donde, P_j, sería aquí la proporción de sujetos que aciertan al ítem, es decir el índice de dificultad (que realmente es el inverso de la dificultad), y $Q_j = (1 - P_j)$. La varianza será máxima para los valores medios de, P_j, en otras palabras, la dificultad media de los ítems maximiza su varianza.

Índice de discriminación.

Se dice que un ítem tiene el poder discriminativo si distingue o discrimina, entre aquellos sujetos que puntúan alto en el test y los que puntúan bajo, es decir, si discrimina entre los eficaces en el test y los ineficaces. En consecuencia, el índice de discriminación se define como la correlación entre las puntuaciones de los sujetos en el ítem y sus puntuaciones en el test.

La correlación a utilizar dependerá, desde luego, de las características de las variables a correlacionar, en este caso el ítem y el test. Dados los formatos que suelen adoptar frecuentemente los ítems y los tests, los coeficientes de correlación más habituales son la biserial-puntual, biserial y tetracórica.

La correlación biserial- puntual es una aplicación de la correlación de Pearson cuando una de las variables es dicotómica y la otra cuantitativa continua, o eventualmente discreta. Suele usarse con frecuencia para calcular el índice de discriminación, dado que es habitual que los ítems sean dicotómicos (o se aciertan o se fallan) y el test constituya una medida cuantitativa discreta. La fórmula de la correlación de Pearson bajo estas circunstancias viene dada por:

$$\rho_p = \frac{\mu_p - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \qquad (2.5)$$

donde:

ρ_p : correlación de Pearson.

μ_p : media en el test de los sujetos que aciertan el ítem.

μ_X : media del test.

σ_x : desviación típica del test.

p : proporción de sujetos que aciertan el ítem.

q : $1-p$.

Índice de validez

Se denomina índice de validez de un ítem, a la correlación entre la puntuación de la prueba y la medida de criterio. Sobre qué correlación utilizar, dependerá de la naturaleza de las variables a correlacionar que aquí son el ítem y el criterio. La validez se refiere pues, a la utilidad que debe tener una herramienta como base para evaluar a los alumnos. Así para que un instrumento sea válido, tendrá que medir y evaluar "algo" en forma consistente, con precisión, y ese "algo" habrá de ser una muestra representativa del comportamiento que se desea observar.

Sea cual sea la correlación utilizada del índice de validez de un ítem indicará en qué grado el ítem está correlacionado con la variable que el test intenta predecir (criterio). La validez global de un test se verá incrementada en la medida que sus ítems tienen índices de validez elevados. (véase Guillisken,1950; citado en Muñiz, J. (1996)). La conexión entre el índice de validez de los ítems y el coeficiente de validez del test viene dada por:

$$\rho_{xy} = \frac{\sum_{j=1}^{n} \sigma_j \rho_{jy}}{\sum_{j=1}^{n} \sigma_j \rho_{jx}} \qquad (2.6)$$

donde:

ρ_{xy} : coeficiente de validez del test.

n número de ítems del test.

$\sigma_{j:}$: desviación típica del ítem j.

ρ_{jy} : índice de validez del ítem j.

ρ_{jx} : índice de discriminación del ítem j.

Escala de medida

En la evaluación de conocimientos, generalmente en cuestionarios, es frecuente vincular el número de aciertos al conjunto de preguntas, de reactivos o ítems utilizados. Se acostumbra reportar el grado de dominio de cada estudiante en términos del número de aciertos o porcentaje de aciertos, independientemente de la conveniencia o no de usar estos indicadores, no obstante que se afirme que la escala que se genera tiene varios defectos, a saber:

a) Cuando se dice que un estudiante obtuvo cero aciertos no quiere decirse que tiene cero conocimientos y es un ignorante total.
b) Cuando un estudiante obtiene todos los aciertos no se puede afirmar que conoce todo.
c) Si un estudiante obtiene 80 aciertos y otro tiene 40 aciertos no puede afirmarse que un estudiante sabe el doble del otro.

Medición en la educación

Debe entenderse que una medida que se haga de los conocimientos de una persona no corresponde a un valor preciso y determinado así como tampoco con los conocimientos que atesora en su cerebro (cosa que no podemos medir), sino se mide como una "Probabilidad de Respuesta". Un alumno con una medida alta tendrá una mayor probabilidad de respuesta correcta a un reactivo y uno con una medida baja tendrá una probabilidad menor de una respuesta correcta. Es conveniente anotar que la probabilidad de respuesta se debe calcular o estimar. Hablaremos de "calcular" al determinar por medio de métodos analíticos correspondientes a un modelo matemático. Hablaremos de "estimar" cuando hagamos una operación muy sencilla que consiste en obtener la frecuencia relativa de respuesta en un examen.

Debemos entonces prepararnos para saber que las puntuaciones de una persona, así como las respuestas a un reactivo, pueden ocurrir con mayor o menor probabilidad, por lo que no se puede garantizar nunca que una persona tiene una medida, X, sino se dice de que una persona de medida, X, tiene una probabilidad, p, de respuesta correcta y otra probabilidad, q, de respuesta incorrecta.

La medida, X ,es una estimación de dominio real, T. Se trata de una aproximación que contiene un error de medida. Si el instrumento utilizado para la medición es altamente preciso, la medida, X, y el valor real, T, serán prácticamente coincidentes. Es posible que nuestros exámenes contengan un error importante, por lo que el estimado, X, contendrá igualmente un error y podría no parecerse al valor real, T. De ahí que el evaluador debe conocer las características del instrumento, su calidad, su error, su confiabilidad para emitir las mejores estimaciones posibles de la medida real de los sujetos, así como la medida real del instrumento. El evaluador debe estar consciente de estas limitaciones y reconocer que el instrumento no es perfecto y deberá tomar las medidas correctivas que sean necesarias para no perjudicar o beneficiar injustamente a los sujetos, informándoles de manera errónea la posición que ocupan en la escala.

Por ello la estimación de la probabilidad de respuesta es muy importante. Su fórmula es muy simple y se basa en el cálculo de las frecuencias relativas como una aproximación de la probabilidad:

$$p = \text{Respuestas correctas/total de reactivos} \qquad (2.7)$$

El fracaso o las fallas que puede tener el alumno se denomina, q, y es el complemento de, p, respecto a la unidad:

$$q = 1 - p \qquad (2.8)$$

Extensión de la Escala

La extensión de la escala hace referencia exclusivamente al rasgo evaluado, por ejemplo conocimientos de la materia de cálculo. La escala hace referencia exclusivamente, al rasgo medido que en la jerga de la evaluación se le denomina "rasgo latente". No es posible determinar con precisión el grado de dominio en una materia, pero lo que sí se puede realizar es explorar (por medio de algún muestreo) el conjunto de conocimientos de cálculo a través de preguntas que el evaluador supone que solamente tratan del tema, a un nivel estipulado utilizando de ser posible por medio de una tabla de validez de contenido.

Una pregunta de Cálculo mide (o cuando menos trata de medir) un cierto conocimiento de la materia, pero la pregunta tiene influencia de la forma de redactar que tenga el evaluador, de la

presentación en el cuestionario, del entorno del alumno, de la forma de aplicar el examen. Por ello se dice que se está tratando de medir al "Rasgo Latente en el estudiante" que también es un "Rasgo Latente en el reactivo". La medida de este rasgo puede contener error y es de interés emitir un dictamen de la medida y del error de la medición para que la evaluación que se hace al estudiante sea lo más justa y precisa posible.

La medida debe señalarse o ubicarse en una escala para poder comparar el dominio de un sujeto contra otros, o el dominio de un sujeto respecto a un criterio de aceptación o de competencia previamente establecido. La escala debe ser tan extensa que permita medir a todos los sujetos, pero no tanto que resulte extremadamente amplia y que, por lo tanto, sea inútil. Así, la medida y la escala definen el rango de medida.

Con objeto de garantizar la precisión de la escala y facilitar la medición, es conveniente que la escala tenga sus divisiones igualmente espaciadas. En la evaluación del conocimiento cada división de la escala podría corresponderse a un reactivo, dosificado en términos de la dificultad. La idea es cubrir toda la extensión de la escala por medio de un número suficiente de reactivos en toda la gama de dificultades.

Es conveniente tener reactivos en toda la gama de dificultades y no solamente reactivos centrados en algún punto específico de dificultad, con objeto de poder medir el dominio de cada persona con precisión. Desde luego, con una escala completa sí se puede ubicar a todas y cada una de las personas y estimar qué tanto saben o qué tanto no saben.

La escala se plantea independientemente de que se trate de un cuestionario referido a norma o referido a criterio. En la evaluación referida a la norma se requiere una varianza ad-hoc en los resultados de los sujetos, mientras que en la referida a criterio la varianza no tiene interés teórico alguno.

CAPÍTULO 3. MODELOS DE MEDICIÓN

Modelo lineal clásico.

El modelo en la Teoría Clásica del Test, TCT, establece que la puntuación empírica, *X*, que obtiene un sujeto en un test es igual a la suma de dos componentes: la puntuación verdadera, *V*, del sujeto en ese test y el error de medida, *e*, cometido en la medición.

La formulación del modelo es la siguiente:

Modelo $X = V + e$	(3.1)
Supuestos: 1. $V = E(X)$	(3.2)
2. $\rho(v, e) = 0$	(3.3)
3. $\rho(e_j, e_k) = 0$	(3.4)

Definición: dos tests, *j* y *k*, se denominan paralelos si la varianza de los errores es la misma en ambos [$\sigma^2(e_j) = \sigma^2(e_k)$] y también lo son las puntuaciones verdaderas de los sujetos ($V_j = V_k$).

Supuestos del modelo.

Supuesto 1

La puntuación verdadera, *V*, es la esperanza matemática de la empírica (ecuación 3.2):

$$V = E(X)$$

Donde, *X*, es la variable aleatoria o "puntuación empírica del sujeto".

Este primer supuesto constituye en realidad una definición de la puntuación verdadera. A partir de esos valores de, *X*, (puntuaciones empíricas), y bajo ciertos supuestos adicionales, la, TCT, permite hacer estimaciones probabilísticas razonables acerca del valor de las puntuaciones verdaderas, *V*,. La puntuación empírica en un test es una muestra de conducta qué, si reúne ciertos requisitos de medida, y bajo ciertos supuestos, permite hacer inferencias probabilísticas fundadas.

Supuesto 2

Dado que no hay razón para pensar que el tamaño de los errores vaya sistemáticamente asociado al tamaño de las puntuaciones verdaderas, se asume que no existe correlación entre las puntuaciones verdaderas de los sujetos en un test y sus respectivos errores de medida (ecuación 3.3):

$$\rho(v,e) = 0$$

Supuesto 3

Los errores de medida de los sujetos en un test no correlacionan con sus errores de medida en otro test distinto (ecuación 3.4): $\rho(e_i, e_j) = 0$. Si se aplican correctamente los tests, los errores serán aleatorios en cada ocasión, no existiendo razón a priori para que covaríen sistemáticamente unos con otros.

Se debe señalar que ninguna de las asunciones del modelo es comprobable empíricamente de un modo directo tal como están expresadas. Por lo tanto Lord y Novick (1968);citado en Muñiz, J. (1996). ofrecen formulaciones axiomáticas rigurosas del modelo.

Si para la correlación entre los errores de dos tests que miden lo mismo pero con diferentes ítems se cumple que, $\rho(e_i, e_j) = 0$, se dice entonces que los tests son paralelos. Lord y Novick han desarrollado otros paralelismos, por ejemplo los tests "Tau Equivalentes", son aquéllos con puntuaciones verdaderas iguales para los sujetos en ambas formas, pero con varianzas-error no necesariamente iguales.

Coeficiente de Fiabilidad

Dos de las características más relevantes de un test son la dificultad y la fiabilidad; una forma de expresar numéricamente la fiabilidad, es a través de una relación matemática denominada coeficiente de fiabilidad. El coeficiente de fiabilidad, $\rho_{xx'}$, se define como la correlación entre las puntuaciones obtenidas por los sujetos en dos formas paralelas de un test, X y X'. Es un indicador de la estabilidad de las medidas, pues si aplicamos un test, X, a una muestra de sujetos y pasado un tiempo aplicamos a los mismos sujetos una forma paralela, X', dado que ambas formas, por definición, miden el mismo constructo, si no hubiese errores aleatorios de medida, la correlación debería de ser perfecta:

$$\rho_{xx'} = 1 \qquad (3.5)$$

Por tanto, el grado en el que, $\rho_{xx'}$, se aleja de 1 nos indicará en qué medida nuestras mediciones están afectadas por errores aleatorios de medida, siempre en el supuesto, claro está de que las dos formas paralelas, X y X', realmente lo sean. De la definición dada para el coeficiente de fiabilidad y de los supuestos del modelo se deriva que:

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \qquad (3.6)$$

donde:

$\rho_{xx'}$: coeficiente de fiabilidad.

σ_e^2 : varianza del error.

σ_X^2 : varianza de la muestra.

Sin embargo, a partir de estas fórmulas es imposible calcular empíricamente, $\rho_{xx'}$, dado que el valor de, σ_e^2, prácticamente no se puede obtener de las respuestas de los sujetos a los ítems. No obstante son útiles para dar una idea conceptual de lo que representa el coeficiente de fiabilidad. Si no hubiese errores aleatorios que: si, $\sigma_e^2 = 0$, entonces, $\rho_{xx'}=1$; y si todo fuese error, $\sigma_e^2 = \sigma_x^2$, entonces, $\sigma_v^2=0$, y, $\rho_{xx'}= 0$.

Se suele denominar índice de fiabilidad, ρ_{xv} , a la correlación entre las puntuaciones empíricas de un test y las verdaderas, siendo igual a la raíz cuadrada del coeficiente de fiabilidad:

$$\rho_{XV} = \sqrt{\rho_{XX'}} = \sqrt{1 - \frac{\sigma_e^2}{\sigma_X^2}} \qquad (3.7)$$

donde:

ρ_{xv} : índice de fiabilidad.

$\rho_{xx'}$: coeficiente de fiabilidad.

σ_e^2 : varianza del error.

σ_X^2 : varianza de la muestra.

Factores que afectan la fiabilidad.

Dos factores importantes que afectan la fiabilidad de un test son, por un lado, la medida de qué tanto los datos se dispersan, es decir la variabilidad de los resultados y, por otro, el número de ítems que conforman el test.

Fiabilidad y variabilidad.

Uno de los aspectos que influye en la fiabilidad es la variabilidad de la muestra. El coeficiente de fiabilidad varía en forma directamente proporcional a la raíz cuadrada de la varianza de la muestra. Se puede observar en la ecuación 3.7 que al aumentar la variabilidad, se aumenta el denominador del segundo término, lo cual deja claro el aumento del coeficiente de fiabilidad. De lo anterior debe ser evidente que la fiabilidad de un test no depende únicamente de las características propias del test, sino también depende del tipo de muestra de los sujetos utilizados para calcularla, lo cual constituye una seria limitación para el modelo clásico, pues

se está describiendo un instrumento de medida, como es el test, en función de los "objetos" medidos, es decir, los sujetos. Por tanto un dato imprescindible para la interpretación del coeficiente de fiabilidad es la variabilidad de la muestra en la que se calculó. En pocas palabras, un test no tiene un coeficiente de fiabilidad fijo, éste depende de la variabilidad de la muestra en la que se calcule. En suma, al aumentar la variabilidad de la muestra aumenta el valor del coeficiente de fiabilidad.

Si se cumple el supuesto de que la varianza de los errores de medida en el test es la misma en dos poblaciones, la menos variable y la más variable, entonces puede obtenerse una fórmula basada en los errores que permite estimar ese aumento. Esta fórmula establece:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (3.8)$$

Dado que uno de los supuestos del modelo establece que las varianzas de los errores correspondientes a dos formas paralelas son siempre las mismas, es decir, $\sigma^2_{e1} = \sigma^2_{e2}$, entonces la fórmula anterior puede expresarse sustituyendo, σ_e^2, por su valor en ambas poblaciones:

$$\sigma_1^2 (1 - \rho_{11'}) = \sigma_2^2 (1 - \rho_{22'}) \qquad (3.9)$$

despejando ρ_{22}

$$\rho_{22'} = 1 - \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho_{11'}) \qquad (3.10)$$

donde :

ρ_{11}: coeficiente de fiabilidad en la población 1.

ρ_{22}: coeficiente de fiabilidad en la población 2.

σ^2_1: varianza empírica en la población 1.

σ^2_2: varianza empírica en la población 2.

Fiabilidad y longitud

La fiabilidad de un test también depende de su longitud, entendiendo por longitud el número de ítems del test. En un principio parece lógico pensar que cuantos más ítems se utilicen para evaluar una variable, mejor podremos muestrear los diferentes aspectos que la conforman y más fiable será la medida obtenida. En el limite, es decir, infinitos ítems, el error sería cero, pero no habría sujeto capaz de soportar una dosis de ese tamaño.

Según los supuestos del modelo, si se tiene un test, *X*, y se aumenta su longitud, *n*, veces a base de ítems paralelos a los originales, la fiabilidad del nuevo test alargado viene dada por la fórmula de Spearman–Brown

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1+(n-1)\rho_{xx'}} \qquad (3.11)$$

donde:

$\rho_{XX'}$: Fiabilidad del test alargado.

$\rho_{xx'}$: Fiabilidad del test original.

n: Número de veces que se ha alargado el test.

Un caso particular de esta fórmula especialmente habitual es cuando se duplica la longitud del test original utilizado, por ejemplo, en el cálculo del coeficiente de fiabilidad por el Método de las Dos Mitades. En dicho caso particular, $n = 2$, quedando reducida a:

$$\rho_{XX'} = \frac{2\rho_{xx'}}{1+\rho_{xx'}} \qquad (3.12)$$

Fiabilidad de las Diferencias

Existen situaciones en Psicología y Educación en las que interesa estudiar las diferencias existentes entre las puntuaciones de los sujetos en un test y sus puntuaciones en otro. Para poder interpretar las diferencias encontradas es imprescindible disponer de alguna medida de su fiabilidad. Dos diferencias iguales pueden tener muy distinto valor científico para el investigador en función de su fiabilidad.

Para dos tests, X y Z, la fiabilidad de las diferencias entre sus puntuaciones, $(X - Z) = d$, como fácilmente se puede derivar, viene dada por:

$$\rho_{dd'} = \frac{\sigma_X^2\rho_{XX'} + \sigma_Z^2\rho_{ZZ'} - 2\sigma_X\sigma_Z\rho_{XZ}}{\sigma_X^2 + \sigma_Z^2 - 2\sigma_X\sigma_Z\rho_{XZ}} \qquad (3.13)$$

donde:

$\rho_{dd'}$: coeficiente de fiabilidad de las diferencias.

σ^2_X: varianza de las puntuaciones del test X.

σ^2_Z: varianza de las puntuaciones del test Z.

$\rho_{XX'}$: coeficiente de fiabilidad del test X.

$\rho_{ZZ'}$: coeficiente de fiabilidad del test Z.

ρ_{XZ}: correlación entre tests X y el test Z.

Si ambos tests se expresan en la misma escala para una mejor interpretabilidad de las diferencias, clásicamente en la escala de típicas, aunque cualquier otra es posible, sus varianzas serían iguales, $\sigma^2_X = \sigma^2_Z$, y la fórmula anterior se simplifica a:

$$\rho_{dd'} = \frac{\rho_{xx'} + \rho_{zz'} - 2\rho_{xz}}{2(1 - \rho_{xz})}$$

Ahora, el error típico de medida de las diferencias, σ_{ed}, análogamente a lo visto para un solo test, vendrá dado por:

$$\sigma_{ed} = \sigma_d \sqrt{1 - \rho_{dd'}} \qquad (3.14)$$

donde:

σ_d: desviación típica de diferencias.

ρ_{dd}: coeficiente de fiabilidad de las diferencias.

El coeficiente de fiabilidad de las diferencias es, asimismo, fundamental para la medida del cambio experimentado por las puntuaciones de los sujetos en alguna variable educativa.

Conviene señalar que, en contra de la extendida costumbre, no está justificado hacer comparaciones individuales entre las puntuaciones empíricas de dos sujetos en un test y de un sujeto en dos tests, basándose en los errores típicos de medida que se pueden derivar. Este tipo de inferencias, como bien señalan Lord y Novick. son incorrectas. Sólo quedarían justificadas si los pares de sujetos a comparar se escogiesen al azar y sin referencia a su puntuación empírica; si así se hiciese, a la larga (estrictamente infinito) las inferencias tendrían sentido globalmente, pero nunca para dos sujetos específicos elegidos, porque interesa comparar precisamente sus puntuaciones empíricas.

Coeficiente Alfa (α)

El coeficiente alfa, α, propuesto por Croncbach, constituye otra forma de acercarse a la fiabilidad. Más que la estabilidad de las medidas, α, refleja el grado en el que covarían los ítems que constituyen el test. Es por tanto, un indicador de la consistencia interna del test: Su fórmula viene dada por:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n}\sigma_j^2}{\sigma_x^2}\right) \qquad (3.15)$$

donde:

n: Número de ítems del test.

$\Sigma\,\sigma^2_j$: Suma de las varianzas de los n ítems.

σ^2_x: Varianza de las puntuaciones en el test.

Dado que α es función directa de las covarianzas entre los ítems, tal vez se pueda apreciar más directamente si se expresa su fórmula en función explícita de éstas, la cual viene dada por:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum_{j\neq k}^{n}\sum \mathrm{cov}(j,k)}{\sigma_X^2}\right) \qquad (3.16)$$

Obsérvese que, α, aumenta al aumentar la covarianzas entre los ítems.

Casos particulares de α

Previamente a la presentación del coeficiente de, α, por Cronbach en 1951; citado en Muñiz, J. (1996) la Psicometría clásica ya disponía de otras fórmulas para estimar la fiabilidad en términos de la consistencia interna del test. Se enlistan aquí cuatro de esos casos particulares de que también dependen de: α

– Rulon

– Guttman / Flanagan

- Kuder- Richardson

  KR_{20}

  KR_{21}

- Rulon

Estimación empírica del coeficiente de fiabilidad

Las fórmulas del coeficiente de fiabilidad expuestas anteriormente no permiten calcular su valor empírico para una muestra determinada de sujetos. Para poder hacerlo hay que valerse de la definición dada en la ecuación de correlación entre las puntuaciones en dos formas paralelas. Así, se tratará de:

1) elaborar las dos formas paralelas.
2) aplicarlas a una muestra amplia de sujetos representativos de la población en la que se va a utilizar el test.
3) calcular la correlación entre puntuaciones de los sujetos en ambas formas.

Dicha correlación será precisamente el coeficiente de fiabilidad. Este método se denomina *Método de las Formas Paralelas*. Es frecuente denominar coeficiente de equivalencia al valor obtenido ya que indicaría el grado en el que ambas formas son equivalentes. Se usan otros dos métodos denominados respectivamente *Test-Retest y Dos Mitades.*

Test-Retest. Para calcular el coeficiente de fiabilidad por este método se aplica el mismo test en dos ocasiones a los mismos sujetos: la correlación entre las puntuaciones de las dos aplicaciones será el coeficiente de fiabilidad. Dado que un test es paralelo a sí mismo, este método es perfectamente congruente con el modelo, denominándose a la estimación obtenida coeficiente de estabilidad, pues indica en qué grado son estables las mediciones realizadas en la primera aplicación del test.

En el Método Test-Retest es difícil delimitar el tiempo óptimo que debe transcurrir entre ambas aplicaciones. Si se deja mucho, se introduce una gran fuente de invalidez interna, a saber la influencia diferencial de ese período de tiempo en los sujetos, pero si transcurre poco tiempo la validez interna disminuye por el efecto de la memoria de lo realizado previamente.

Dos mitades. Por este método se aplica el test una sola vez, obteniéndose para cada sujeto las puntuaciones correspondientes a cada una de las mitades en las que se divide el test. El coeficiente de fiabilidad viene dado por la correlación entre esas dos mitades (que será la estimación de la fiabilidad del test mitad) más una corrección para obtener la fiabilidad del test total. La estimación así obtenida indica la covariación o consistencia interna de las dos mitades, siendo así un indicador de la consistencia interna del test.

La estimación de la fiabilidad a partir de dos mitades no está exenta de dificultades por ejemplo: ¿qué mitades tomar?. Un test con, n, ítems tiene muchas posibles mitades, exactamente combinaciones de, n, elementos tomados de, $n/2$, en, $n/2$. Por ejemplo, un test con 10 ítems (sólo 10) tiene 252 posibles mitades, o sea, por ese método se pueden hacer 126 estimaciones de su fiabilidad. Cronbach, demuestra que, α ,calculado a partir de todos los ítems de un test es el valor medio que se obtendría de calcularlo para todas las posibles mitades del test, así el valor esperado de las mitades es, $\alpha = E(\alpha/2)$.

El método de las dos mitades es muy funcional, sólo exige una aplicación del test. No obstante, hay que garantizar que las mitades del test son paralelas. No es recomendable, por ejemplo considerar como mitades la primera parte del test por un lado y la segunda por otro, pues los sujetos llegarán más cansados a la segunda; por lo general, los ítems van aumentando en dificultad, por lo que la segunda parte resultaría más difícil que la primera. Para evitar esto es frecuente tomar como una mitad los ítems pares y como otra los impares, o usar algún otro tipo de apareamiento de los ítems. En definitiva, es un problema de control experimental.

La lógica de estos tres métodos es clara, su realización empírica plantea diversos problemas experimentales relativos a la validez interna, para los cuales el modelo no da especificaciones concretas, quedando al criterio del investigador para cada situación nueva planteada.

En el método de las formas paralelas el problema fundamental es la construcción de dichas formas paralelas. Es difícil a nivel teórico hacer un test que mida exactamente lo mismo que otro pero con distintos ítems, tal vez incluso, filosóficamente imposible, y en la práctica es enormemente laborioso. Si se superan los problemas y se dispone de dos o más formas paralelas probablemente es el método más recomendable.

Un factor del test a tener en cuenta para elegir un método u otro de los comentados, o de otros que se verán es si se trata de un test de velocidad o de un test de potencia. Suele entenderse por test de velocidad aquél cuya realización no conlleva dificultad alguna, y en consecuencia todos los sujetos son capaces de realizar, aunque difieran en el tiempo de ejecución. Un test de potencia sería aquél en que las diferencias entre los sujetos son generadas por su distinta capacidad intelectual para resolver las tareas propuestas. En la práctica la mayoría de los tests suelen ser mixtos, variando la proporción de ambos componentes: en unos predomina más la velocidad y en otros la potencia.

Estimación de las Puntuaciones Verdaderas

Conocida la fiabilidad del test por alguno de los métodos anteriores, se pueden hacer estimaciones acerca de la cantidad de error que afecta a las puntuaciones empíricas. Así, a continuación se exponen las tres estrategias más utilizadas por la Psicometría clásica.

a) Estimación mediante la desigualdad de Chebychev

La desigualdad de Chebychev establece que para toda variable, X, con media, $\overline{X}$, y desviación típica, S_X:

$$\forall\ K\quad P\{|X-\overline{X}|\le K(S_x)\}\ge 1-\frac{1}{K^2} \qquad (3.17)$$

que traducido a la terminología psicométrica del modelo clásico da:

$$\forall\ K\quad P\{|X-V|\le K(\sigma_e)\}\ge 1-\frac{1}{K^2} \qquad (3.18)$$

Se ha sustituido la media, $\overline{X}$, por, V, puesto que en el modelo clásico, $E(X) = \mu_x = V$; la desviación típica, S_X ,se ha sustituido por, σe, puesto que en el modelo la varianza de, X, para un valor dado de, V, es igual a la varianza de los errores, σ_e^2,. Nótese que al fijar, V, lo que varían las puntuaciones empíricas se debe únicamente a la variación de los errores. Asimismo, la varianza de los errores para un determinado valor de, V, es igual a la varianza de los errores de la población, σ^2_e ,dado que, ρ_{Ve} =0. En definitiva:

$$\sigma^2(X\mid V)=\sigma^2(e\mid V)=\sigma_e^2 \qquad (3.19)$$

b) ***Estimación basada en la distribución normal de los errores.***

Una forma de evitar una estimación tan genérica como la anterior es asumir que los errores de medida, y por ende las puntuaciones empíricas, para un valor dado de, V, (sujetos o clase de sujetos con misma puntuación verdadera) se distribuyen según la curva normal:

$$f(e|V) \approx N(0, \sigma_e^2) \qquad (3.20)$$

Dado que, $X = V + e$ con, V, constante, es inmediato que

$$f(X|V) \approx N(V, \sigma_e^2) \qquad (3.21)$$

Adviértase que una implicación por reducir la amplitud del intervalo es el supuesto de normalidad que se añade al modelo, dado que no se había hecho ninguna asunción sobre la forma de las distribuciones de las puntuaciones. Este supuesto de normalidad e igualdad de las varianzas condicionales a lo largo de la escala de las puntuaciones, de acuerdo a autores como Muñiz, J. (1990), ha sido cuestionado con frecuencia, especialmente en lo concerniente a los valores extremos de la escala. Si no se cumple, lo cual es bastante probable en la práctica, no sería muy preciso utilizar el mismo error típico de medida para todos los sujetos, independientemente de la puntuación en el test, como se hace habitualmente. Este problema no hallará una respuesta apropiada en el marco de la Teoría Clásica de los Tests, y habrá que utilizar los modelos de Teoría de Respuesta a los Ítems para una solución adecuada. Mediante la función de información del test, estos modelos permitirán estimar el error de medida para los distintos niveles de competencia de los sujetos.

c) ***Estimación según el modelo de Regresión***

En el modelo de regresión lineal el pronóstico de una variable, Y, a partir de otra, X, según el criterio de mínimos cuadrados, viene dado por la expresión:

$$Y' = \rho_{xy}\left(\frac{\sigma_y}{\sigma_x}\right)(X - \bar{X}) + \bar{Y} \qquad (3.22)$$

Si sustituimos, V, por, Y, dada nuestra terminología, entonces:

$$V' = \rho_{xv}\left(\frac{\sigma_v}{\sigma_x}\right)(X - \bar{X}) + \bar{V} \qquad (3.23)$$

Ahora bien, como hemos visto: $\overline{V} = \overline{X}$ y $\frac{\sigma_v}{\sigma_x} = \rho_{xv}$, entonces:

$$V' = \rho_{xv}^2 (X - \overline{X}) + \overline{X} \qquad (3.24)$$

y como $\rho^2{}_{xv} = \rho_{xx'}$

$$V' = \rho_{xx'}(X - \overline{X}) + \overline{X} \qquad (3.25)$$

Mediante esta fórmula podemos hacer estimaciones puntuales de, *V*, a partir de, *X*, conociendo el coeficiente de fiabilidad, la media del test y la puntuación empírica. Cabe acotar que el modelo de regresión utilizado, lo único que garantiza es que "a la larga" los errores de pronóstico cometidos serán mínimos, según el criterio de mínimos cuadrados, pero la puntuación pronosticada, *V'*, no siempre coincidirá con, *V*, Para establecer dichos intervalos nos valemos del error típico de estimación, que es la desviación típica de los errores de estimación y, que en su forma general para dos variables, *X* y *Y*, viene dado por:

$$\sigma_{yx} = \sigma_y \sqrt{1 - \rho_{xy}^2} \qquad (3.26)$$

que traducido a la terminología y supuestos del Modelo lineal Clásico puede expresarse así:

$$\sigma_{vx} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{\rho_{xx'}} \qquad (3.27)$$

o también, teniendo en cuenta que:

$$\sigma_x \sqrt{1 - \rho_{xx'}} = \sigma_e \qquad (3.28)$$

$$\sigma_{vx} = \sigma_e \sqrt{\rho_{xx'}} \qquad (3.29)$$

Asumiendo que los errores de estimación se distribuyen normalmente en torno a *V'*, se pueden establecer los correspondientes intervalos confianza.

Tipos de errores de medida.

Hay básicamente dos tipos de errores de medida: el error de medida y el error de estimación. Cabe citar además (Guillikssen; Lord y Novick citado en Muñiz, J.(1996)) el error de sustitución y el error de predicción. Se exponen a continuación los cuatro junto con sus respectivas desviaciones típicas.

1- Error de medida

$$e = X - V$$
$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (3.30)$$

2.- Error de estimación

$$e = V' - V$$
$$\sigma_{V.X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} \qquad (3.31)$$

3.- Error de sustitución

$$e = X_1 - X_2$$
$$\sigma_{e(s)} = \sigma_x \sqrt{1 - \rho_{XX'}} \sqrt{2} \qquad (3.32)$$

4.- Error de predicción

$$e = X_1 - X'_1$$
$$\sigma_e(p) = \sigma_x \sqrt{1 - \rho_{XX'}} \sqrt{1 + \rho_{XX'}} \qquad (3.33)$$

El error de medida y el error de estimación son, respectivamente, como ya se ha visto las diferencias entre la puntuación verdadera y la empírica, $V - X$, y la diferencia entre la puntuación verdadera pronosticada y la verdadera, $V' - V$. Sus desviaciones típicas, también previamente definidas, se denominan error típico de medida, σ_e, y error típico de estimación σ_{VX}, respectivamente.

El error de sustitución es la diferencia entre las puntuaciones en un test, X_1, y las obtenidas en otro paralelo, X_2, es, por tanto, el error de medida que se generaría al sustituir una medición por otra proveniente de un test paralelo. Su desviación típica, dada por $\sigma_{e(s)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{2}$, es el error típico de las diferencias entre dos tests paralelos.

El error de predicción es la diferencia entre las puntuaciones en un test, X_1, y las puntuaciones pronosticadas en ese test, X'_1, a partir de una forma paralela, X_2.. Es el error que se cometería al utilizar en vez de mediciones de un test aquellas pronosticadas en ese test a partir de una forma paralela. Es decir de, X'_1, que son los pronósticos realizados mediante la recta de regresión de, X_1 sobre, X_2, que viene dada por:

$$X'_1 = \rho_{12} \frac{\sigma_1}{\sigma_2} \left(X_2 - \bar{X}_2 \right) + \bar{X}_1 \qquad (3.34)$$

Teoría de la Respuesta al Ítem

La teoría de la respuesta al ítem o teoría de la respuesta al reactivo, surge como una posibilidad de superar la dependencia del valor de algunos parámetros cuando éstos son calculados con base en la Teoría Clásica de los Test, TCT. Así por ejemplo, si el índice de dificultad es calculado con base en la, TCT, se define como la proporción de sujetos que

contestan correctamente o no un ítem, y debe ser claro que dicha proporción variará si los alumnos son muy aventajados o no lo son tanto, por lo que dicho índice de dificultad, así expresado, será variable y dependerá de la muestra o población utilizada.

La Teoría de Respuesta a los Ítems, TRI, o como también se le conoce, Teoría o Modelos de Rasgo Latente, obedece a una nueva orientación en la teoría de los tests que permite superar ciertos problemas de medición psicológica que eran imposibles de resolver desde la Teoría Clásica de los Tests; lo que realiza la, TRI, es hacer asunciones adicionales que permitan responder a cuestiones que la ,TCT, no podía.

El nombre de teoría de respuesta a los ítems se debe a que se sustenta en las propiedades de los ítems más que en las del test global. Cómo este modelo refleja el funcionamiento real basado en los ítems es lo que lo distingue de otros modelos como son: el análisis factorial, el análisis multidimensional y el análisis estructural.

La TRI, además, realiza contribuciones de tipo técnico al momento de construir tests. Desde el punto de vista teórico de la medición psicológica su gran contribución se centra en la posibilidad de obtener mediciones invariantes respecto a los instrumentos utilizados y de los sujetos implicados; como señaló Thurstone (1928);citado en Muñiz,J.(1997) "un instrumento de medida no debe venir afectado por los objetos medidos y sus mediciones deben de ser independientes de los objetos medidos". La teoría clásica se encontraba encerrada en la incongruencia teórica en la cual las propiedades del instrumento de medida, esto es, los ítems y por tanto, el test, estaban en función de los sujetos utilizados a los que se ha aplicado, es decir, de los sujetos medidos. Así, la, TRI, proporciona, entre otras ventajas adicionales a las mencionadas:

- Estimación de los parámetros a partir de la curva características de los ítems, CCI, en razón de que se ha desarrollado una función que relaciona la capacidad del sujeto con la probabilidad de que responda bien a un ítem, es decir, que la probabilidad de acertar un ítem sólo depende de los valores de la variable medida por el ítem; por tanto sujetos con distinta puntuación en dicha variable tendrán probabilidades distintas de superar determinado ítem.

- Asignación con mayor exactitud de un valor, θ, (la medida de la capacidad de un sujeto en un determinado rasgo) para cada sujeto, con base al patrón observado dado que la escala en que se representa, θ, en principio es infinita, y en consecuencia, se pueden diseñar ítems cuyas dificultades sean suficientemente densas.

Concepto de curva característica de un reactivo

Cuando se aplica un reactivo a una población de estudiantes siempre ocurre que hay alumnos que responden correctamente y, complementariamente, hay alumnos que fallan. En principio, se asume que los alumnos responden lo que conocen y que nadie responde por azar o tratando de adivinar, lo que será la hipótesis para desarrollar un modelo con base en una curva característica de un reactivo. No es de interés probar si se trata de una hipótesis fuerte o débil. El supuesto que estamos haciendo (sólo responde lo que se conoce) puede ser catalogado de "hipótesis débil" porque no hay manera de poderlo demostrar.

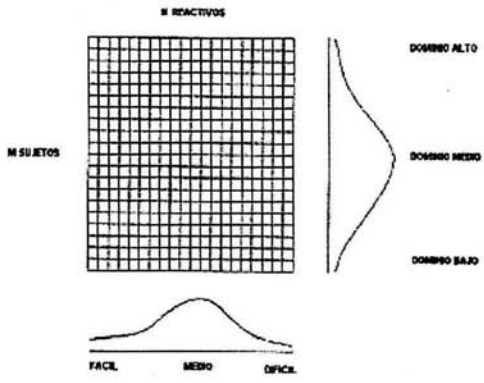


Figura 3.1

Pero también puede ser catalogado como una "hipótesis fuerte" porque permite ser planteada sin ninguna relación con otra parte del modelo y no se puede demostrar de manera recurrente usando otras hipótesis o evidencias posteriores.

Si aceptamos que los estudiantes responden en función de lo que conocen o dominan, se puede hacer una tabla (figura 3.1) de las diferentes respuestas que tienen los alumnos a los reactivos del examen. Supongamos que se tienen M alumnos y N reactivos aplicados. Se tendría, por lo tanto, una tabla o matriz de M renglones (alumnos) y N columnas (reactivos). Algunos reactivos serán difíciles para los alumnos y otros serán fáciles y algunos reactivos serán de dificultad media. Por lo tanto algunos alumnos tendrán las mejores calificaciones (mayor número de aciertos), otros tendrán las peores calificaciones (menor número de aciertos) y algunos resultarán con calificaciones intermedias. Podemos construir diagramas de frecuencias de calificaciones de los alumnos y calificaciones de los reactivos.

Cada reactivo tiene un comportamiento determinado con cada sustentante y, a su vez los sustentantes responderán de cierta

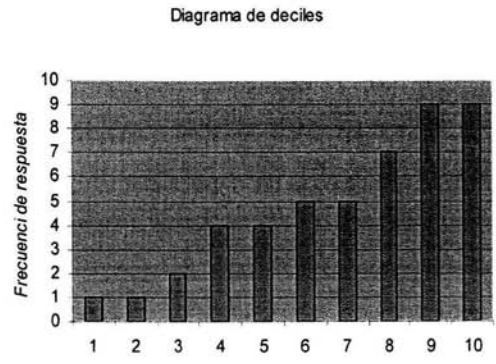


Figura 3.2

manera al conjunto de reactivos. Lo que nos interesa es disponer de información para estimar una probabilidad de respuesta de un alumno ante un reactivo dado.

Para describir el comportamiento de un ítem, supongamos que se aplica un cuestionario, digamos, a 100 alumnos y se elige uno cualesquiera de los reactivos. Se divide al grupo en deciles y anotamos el número de respuestas correctas que tiene cada decil en ese reactivo, con ello podemos hacer un diagrama de frecuencias como el que se encuentra en la figura 3.2.

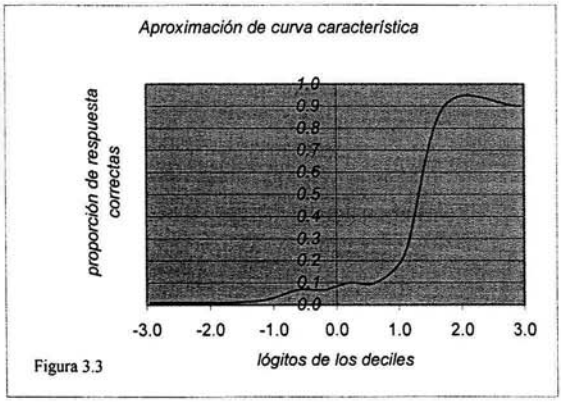
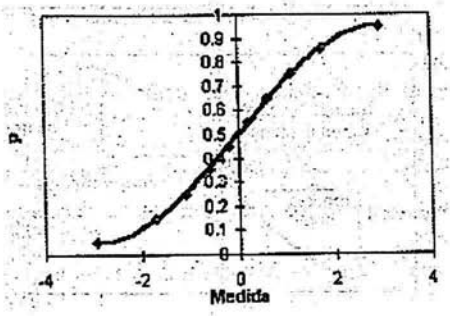


Figura 3.3

Los extremos superiores de las barras describen una "Curva". Se podría contar con una función que describa a esa curva, la cual nos permitiría hacer estimados de frecuencia de respuesta, por ejemplo podíamos estimar cual es la posible respuesta de los sujetos del decil 6, sin tener que contarlos.

En lugar de trabajar con barras trazaremos un punto correspondiente a la medida en lógitos (Probabilidad de respuesta correcta) del punto medio del decil y de altura igual a la proporción de respuestas correctas que se tienen en el decil; la gráfica que se genera ahora es una aproximación de la curva característica. (Figura 3.3).

Figura 3.4



La gráfica se ha trazado suavizando la curva que sería la correspondiente a un polígono de frecuencias. Desde luego las escalas de valores de los ejes corresponden a las escalas modificadas, como ya se ha descrito en el párrafo anterior.

Esta proporción de respuestas correctas es la frecuencia relativa, que hemos dicho que es un buen estimador de la probabilidad de la probabilidad de respuestas.

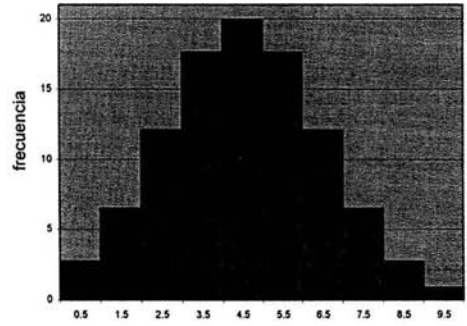
El comportamiento de los reactivos y de las personas puede representarse por medio de una curva de tipo sigmoide, como la que se aprecia en la figura 3.4, denominada "Curva Característica del Reactivo" CCR, o "Curva Característica del Ítem", CCI, que relaciona la medida de las personas y la probabilidad que tienen de responder al reactivo. Para construir

Intervalo de Calificación	Frecuencias	Frecuencia acumulada
0.5	2	2
1.5	3	5
2.5	6	11
3.5	12	23
4.5	18	41
5.5	20	61
6.5	18	79
7.5	12	91
8.5	6	97
9.5	3	100

Tabla 3.1. Frecuencias de calificaciones

esta curva necesitamos, por lo tanto, la medida y la probabilidad. El cálculo de estos valores involucra una serie complicada de pasos, mismos que se describen en una sección posterior. A continuación se hará una descripción del proceso lógico que permite obtener una sigmoide que represente una curva característica del ítem, a partir del supuesto de normalidad que se asume de la respuesta. (Tabla 3.1)

Supóngase que se analiza el ítem, ***k***, que fue aplicado a 150 alumnos, entre los cuales 100 lo acertaron. Ahora, cada uno de los sujetos que acertó el ítem ***k*** tiene una puntuación global en el examen, es decir, se le asocia el número de ítems correctos en total. Supóngase además, que la gráfica de frecuencias derivada de las puntuaciones globales de los 100 alumnos que acertaron el ítem en cuestión, tiene una distribución normal como la gráfica de la figura 3.5. Si se procede a obtener la ojiva normal de las frecuencias anteriores, se obtiene una curva sigmoide, que es exactamente lo que deseábamos construir.



Histograma Figura 3.5

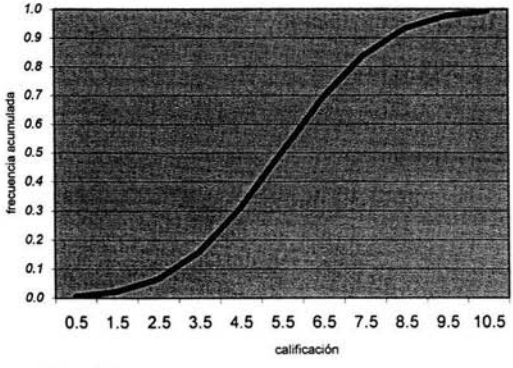


Figura 3.6

A esta curva sigmoide se le pueden asociar algunos parámetros, a saber: el valor de la abscisa correspondiente a la frecuencia acumulada de 0.5, que dará origen al parámetro *dificultad*, la pendiente asociada al punto de inflexión que dará origen al parámetro *discriminación* y el valor mínimo correspondiente al mínimo valor en las abscisas que dará origen al parámetro de adivinación. (Figura 3.6)

Propiedades de la Curva Característica

Índice de dificultad

Se denomina índice de dificultad del reactivo a la medida asociada al punto de inflexión de la curva característica de dicho reactivo. Al valor numérico que define el índice de dificultad se le reconoce por la letra, ***b***.

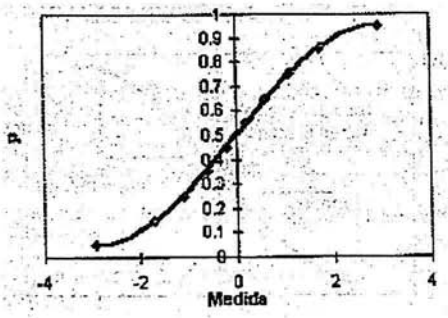


Figura 3.7

Al dibujar la curva característica de un reactivo, una sigmoide, con un comportamiento casi ideal como se muestra en la figura 3.7, los diferentes valores, expresados en lógitos, que puede tomar el parámetro de dificultad, se grafican sobre el eje de las abscisas y los valores de probabilidad definidos a través de las frecuencias relativas, se grafican en el eje de las ordenadas. Suponiendo que la distribución de las frecuencias es una distribución normal, genera una curva de tipo campana, entonces la gráfica de la distribución acumulada será de tipo sigmoide, con un punto de inflexión coincidente con el valor de la media. A este valor se le denomina dificultad del ítem. Cabe señalar que la existencia de un punto de inflexión es una condición necesaria para una correcta asignación del parámetro dificultad.

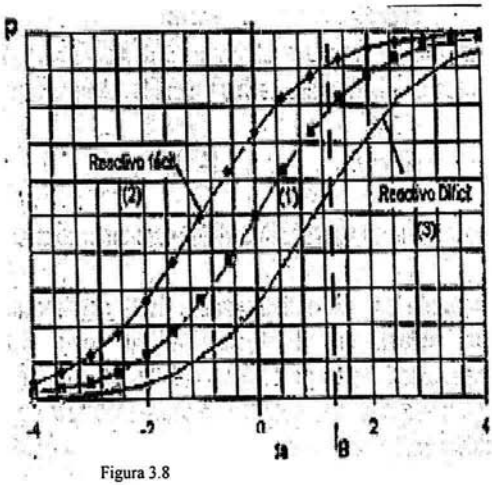


Figura 3.8

Si los valores de la variable aleatoria se estandarizan, y ante el supuesto de una distribución normal, entonces su media será cero (Figura 3.8). Ahora, se puede observar que un alumno de medida baja, es decir, con un resultado por debajo de la media, por ejemplo, de -3 (dominio muy bajo) tiene una probabilidad casi nula de contestar correctamente al reactivo. Contrariamente un alumno de medida alta, digamos, +3 (dominio alto) tiene un probabilidad de casi 1 de contestar correctamente. Desde luego, entre estos dos extremos existen alumnos con diferentes medidas y por lo tanto diversas probabilidades de responder correctamente.

Cuando el reactivo apunta a la medida exacta de la persona se puede decir que da en el blanco, es decir la capacidad de la persona y la dificultad del ítem tienen el mismo valor. La

probabilidad asociada en esta situación se asume de 0.5 y no corresponde necesariamente ni al valor de dificultad del ítem ni a la capacidad de la persona cuando ambas medidas se reportan en lógitos. Ya que en principio se desconoce la medida de la habilidad de la persona, es de esperarse que al aplicar un reactivo, éste podrá ser contestado en forma acertada o errónea, dependiendo si la persona tuviera otra medida o bien que suponiendo que se conoce la habilidad de la persona, al asignarse un ítem pensando que su dificultad es coincidente con la habilidad de la persona, ese hubiese sido calibrado en forma incorrecta. Pero no debe descartarse que basta con que el examinado tenga cualquier tipo de descuido, esto lo llevará a mostrarse por debajo de su propia habilidad, siendo esto último parte del error de medición.

Discriminación

Se define índice de discriminación al valor que permite distinguir a los alumnos de alto desempeño de los de bajo desempeño y está en función de la pendiente evaluada en el punto de inflexión de la curva característica del reactivo. Al valor numérico que define el *índice de discriminación* se le reconoce por la letra ***a***.

Antes de continuar con el análisis del índice de discriminación con relación a la curva característica del ítem, será necesario entender cómo se calcula dicho índice de discriminación de un ítem en particular. En consecuencia, se presentará un ejemplo, detallando cada paso; cabe señalar que el proceso que se expondrá debe hacerse para cada uno de los ítems que conforman un test. Así, ante la tabla de datos 3.2 se calculará el índice de discriminación del tercer ítem.

	Ítems			Total	
Sujetos	1	2	3	4	X
A	0	1	1	1	3
B	1	0	1	0	2
C	1	1	0	0	2
D	1	1	1	1	4
E	0	1	0	0	1

Tabla 3.2

Como ya se ha señalado, el índice de discriminación es la correlación entre las puntuaciones de los sujetos en el ítem y sus puntuaciones en el test. Dentro de las posibles correlaciones, se asumirá la correlación biserial-puntual, que corresponde a una aplicación de la correlación de Pearson cuando una de las variables es dicotómica y la otra cuantitativa continua. La formula a usar es:

$$\rho_{bp} = \frac{\mu_p - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \qquad (3.35)$$

donde:

μ_p : media en el test de los sujetos que aciertan el ítem.

μ_x : media del test.

σ_x : desviación típica del test.

p : proporción de sujetos que aciertan el ítem.

$q = 1 - p$: proporción de sujetos que fallan el ítem.

Los pasos que enseguida se enuncian, del procedimiento para calcular el índice de discriminación se podrán identificar en la tabla3.3:

1. Se asumirá que la variable X corresponde al número de aciertos que cada uno de los sujetos que presentaron el test obtuvo en total en el mismo.

2. A la puntuación total de cada uno de los sujetos, se le deberá descontar el ítem cuyo índice de discriminación se pretende hallar. De lo contrario una de las variables a correlacionar (el ítem) estaría impropiamente incluida en la otra (el test). Desde luego, esto implica restar uno si y sólo si el sujeto acertó; si no acertó, no hay nada que descontar. Esto da lugar a la columna $(X - j)$,

	Ítems				Total	
Sujetos	1	2	3	4	***X***	***(X-j)***
A	0	1	1	1	3	2
B	1	0	1	0	2	1
C	1	1	0	0	2	2
D	1	1	1	1	4	3
E	0	1	0	0	1	1

Tabla 3.3

3. Se calcula la media en el test de los sujetos que aciertan el ítem de interés, así se tiene:

$$\mu_p = \frac{2+1+3}{3} = 2$$

los valores 2,1 y 3 corresponden a las puntuaciones de los sujetos A, B y D de la columna ***(X-j)***

4. Se calcula la media del test:

$$\mu_x = \frac{2+1+2+3+1}{5} = 1.8$$

donde los valores 2, 1, 2, 3 y 1 corresponde a todos los elementos de la columna ***(X-j)***.

5. Se calcula la desviación típica del test:

$$\sigma_x = \sqrt{\frac{(2-1.8)^2 + (1-1.8)^2 + (2-1.8)^2 + (3-1.8)^2 + (1-1.8)^2}{5}} = 0.748$$

6. Se calcula la proporción de sujetos que acertaron el ítem:

$$p = \frac{3}{5} = 0.60$$

7. Se calcula la proporción de sujetos que fallaron el ítem:

$$q = \frac{2}{5} = 0.40$$

8. Como paso final se aplica la fórmula (3.35) $\rho_{bp} = \frac{\mu_p - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}}$, entonces:

$$\rho_{bp} = \frac{2-1.8}{0.748}\sqrt{\frac{0.60}{0.40}} = 0.32$$

Toda vez que se ha entendido cómo se calcula el índice de discriminación, se continuará con el análisis de éste con relación a la curva característica del ítem. A continuación se graficarán las curvas características de tres ítems, (Figura 3.9) identificados como 4, 5 y 6, con igual índice de dificultad, pero con diferente índice de discriminación (en principio asóciese el índice de discriminación con la pendiente de la curva en $p = 0.50$ o bien en el lugar en donde se localiza, en dos de las tres curvas graficadas, el punto de inflexión).

Véase que la curva identificada como (6) tiene una mayor pendiente en el punto de inflexión que las otras dos curvas. En ese sentido, se dice que el ítem tiene una mayor capacidad de discriminación que los otros dos ítems. Debe entenderse que esa capacidad de discriminación no se conserva a lo largo de toda la curva, sino que se asume sólo para valores cercanos al punto de inflexión. Así, se define como "Extensión" al rango en el que se encuentra la zona central de la curva logística.

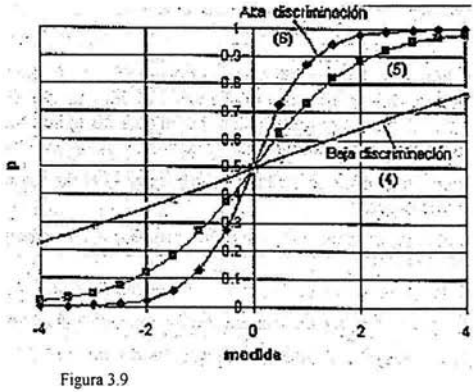


Figura 3.9

Índice de adivinación

De denomina índice de adivinación o adivinación sistemática al valor de probabilidad de acierto mínima esperada en la respuesta de un reactivo. Al valor numérico que define el *índice de adivinación* se le reconoce por la letra *c*.

En la figura 3.10, se han graficado las curvas características de tres ítems, identificados como 9, 10 y 11, con igual índice de dificultad, pero con diferente índice de adivinación. La curva característica con el número 11, tiene un índice de adivinación c_3 que tiende a cortar al eje de las ordenadas casi en cero, en tanto la curva identificada por el número 9, tiene un índice de adivinación de alrededor de 0.25; lo anterior significa que hay mayor probabilidad de adivinar la respuesta correcta en el ítem correspondiente a la curva número 9, con relación a los otros dos ítems. El proceso de adivinación en general se presenta en ítems de opción múltiple, relación de columnas y de falso o verdadero y se origina cuando un alumno no sabe nada del objetivo o contenido del reactivo. Sin embargo, podría tener por lo menos un probabilidad de respuesta correcta, debida a la carga de azar que de manera inherente

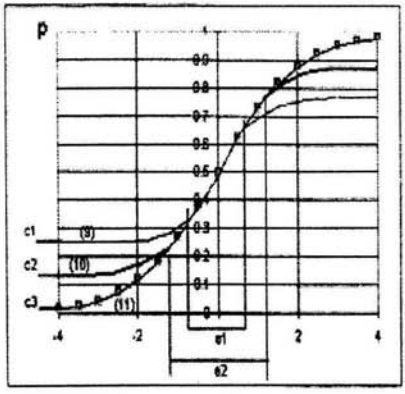
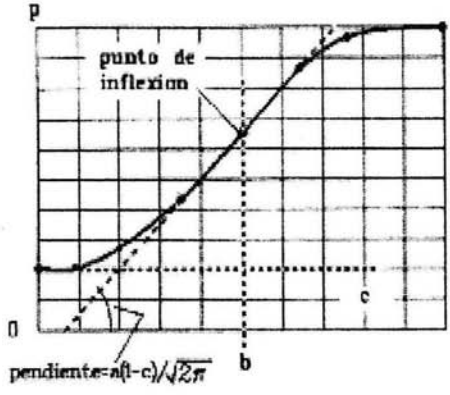


Figura 3.10

reactivo. Por lo tanto, a este parámetro también se le denomina como "parámetro de adivinación" entendiéndose por adivinación sistemática al valor de la probabilidad de acierto para las personas de medida más baja. También puede decirse que es la probabilidad mínima esperada de respuesta correcta de un reactivo. .



Como hemos visto, la curva característica se puede describir por medio de tres parámetros geométricos: *a* discriminación, *b*, dificultad, *c*, adivinación sistemática; como se muestra en la figura 3.11 .El punto en el que se encuentra el valor de, *b*, corresponde con el "punto de inflexión".

Cuando, c, es igual a cero el punto de inflexión coincide con $p = 0.5$, en caso contrario el punto de inflexión se encuentra al centro del intervalo entre, c y 1. Puede establecerse esta fórmula:

$$\text{Punto de inflexión} = \frac{(c+1)}{2}$$

Tipos de modelos para la curva característica

La función matemática que el investigador o usuario adopte para describir la curva característica del reactivo definirá el tipo de modelo teórico involucrado. Entonces cada una de las funciones matemáticas podría en principio dar origen a un determinado tipo de modelo. Para reducir la información relativa al número de modelos lo recomendable es clasificarlos de acuerdo a características afines.

Modelos de aproximación a la curva característica.

Existen dos diferentes procedimientos básicos a seguir para el tratamiento de las observaciones derivadas de un test; es decir, dadas las observaciones se propone una curva de ajuste a ellas o se propone una curva *ad hoc* y sólo se considerarán válidos los reactivos que se ajusten a ésta. Así, se pueden distinguir dos esquemas básicos:

a) Modelos de Ajuste. Modelos que buscan una función que pase lo más cercanamente posible por el conjunto de puntos observados
b) Modelos de Contraste. Modelos que definen una curva de propiedades y características deseables para los reactivos, contra la cual se contrastan los de la curva.

En el caso de modelos de ajuste, como ya se mencionó, se tiene un conjunto de puntos y lo que se pretende es construir una curva que ajuste lo mejor posible a dichos puntos. El ajuste será bueno o malo dependiendo de qué tan alineados estén los puntos, pero siempre podrá obtenerse una curva de ajuste.

Tabla 3.4

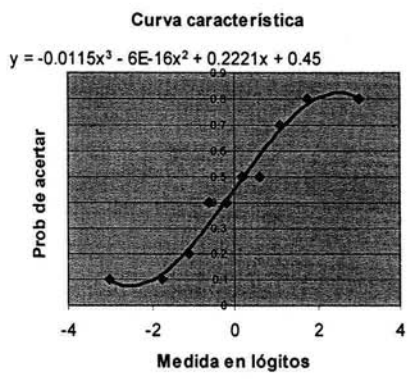
lógitos	Prob
-2.9944	0.1
-1.7346	0.1
-1.0986	0.2
-0.619	0.4
-0.2007	0.4
0.2007	0.5
0.619	0.5
1.0986	0.7
1.7346	0.8
2.9944	0.8

Es necesario obtener, además de los parámetros de ajuste, el orden de error de la curva de ajuste. Puesto que siempre hay parámetros de ajuste, es tarea del evaluador decidir qué tanto error se acepta. La costumbre es emplear un criterio amplio y por regla general se tiende a aceptar que el ajuste es "bueno ", bajo el argumento de que la "evidencia empírica" dicta

que así es el comportamiento de los sujetos en el reactivo. A continuación se ejemplifica este modelo.

La tabla 3.4 contiene algunos datos empíricos, que pueden corresponder a cualquier test, transformados a una escala en lógitos, correspondientes a las abscisas y se les ha asociados una función de probabilidad, que corresponde a los valores del eje de las ordenadas. Graficando lo anterior (figura 3.12)se obtienen los puntos respectivos y se ajusta a ellos una curva de tendencia, cuya ecuación corresponde a un polinomio cúbico.

Figura 3.12



Se puede entender que el modelo es de ajuste, pues la curva de la ecuación $y = -0.0115x^3 - 0.000,0016x^2 + 0.2221x + 0.45$ se ajusta a los puntos graficados. Es necesario señalar que siempre se podrá asociar una ecuación que describa la tendencia de la curva característica, con la acotación ya señalada de que esta tendencia podrá o no ajustar bien a los datos.

Los Modelos de Contraste especifican una forma "deseable" para la curva característica de los reactivos. Esta forma deseable es una propuesta teórica y se basa en la hipótesis que establecen los evaluadores tratando de representar el comportamiento "ideal" de un reactivo. Una vez aceptada la curva teórica se trata de ver si los puntos se parecen a ella. Para poder decidir si hay ajuste o no, se calcula un error de aproximación o error de ajuste y en función del error de ajuste se toma la decisión de aceptar o no el reactivo.

El enfoque del modelo de contraste es más riguroso que el modelo de ajuste en el sentido que no busca ajustar una curva a cualquier precio; en cambio trata de averiguar si los puntos obtenidos se ajustan al modelo. Queda a criterio de cada evaluador aceptar o rechazar el modelo teórico. Este es el modelo que sigue el Análisis de Rasch.

Un modelo describe la realidad dentro de ciertos límites, por lo que es normal (y hasta razonable) que en varias ocasiones el modelo y la realidad no se parezcan. El investigador que elige un modelo debe conocer sus limitaciones y alcances, de manera de poder saber cuándo aceptar y cuándo rechazar las hipótesis de trabajo. Si el reactivo se ajusta a una curva diseñada cuidadosamente, se podrá concluir algo sobre el reactivo, en función de las hipótesis de trabajo. Si el reactivo no se ajusta, entonces no podrá decirse nada sobre el mismo. En consecuencia, el modelo de curva que se elija puede ser tan correcto como cualquier otro. Así,

el problema no reside en la función elegida, sino en querer buscar un " ajuste a cualquier precio"

Según Pagano la curva normal es una distribución muy importante en las ciencias del comportamiento, debido principalmente a que muchas de las variables medidas en la investigación en estas ciencias, tienen distribuciones que semejan mucho la curva normal, por ejemplo la inteligencia. Así, históricamente el trabajo de Richardson (1936) citado en Muñiz, J (1997) es el primer intento de ajustar la ojiva normal a las respuestas a los ítems. Trata de controlar la dificultad de los ítems, en función de los objetivos perseguidos por el test, esto es una representación anticipada de lo que luego habría de permitir realizar la Función de Información en el marco de la TRI. Tucker (1946) citado en Muñiz, J. (1997) también utiliza la curva normal como rudimento de curva característica, dado que sería de esperarse, como ya se mencionó, que el comportamiento humano en muchas ocasiones se podría describir a través de una curva normal por lo que, en forma natural, se esperaría que los reactivos de un tests presenten este tipo de distribución probabilística. Recordemos que la función de densidad de una distribución normal es:

$$n(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (3.36)$$

y la función de ojiva normal, que correspondería en principio a una curva característica de un ítem con distribución normal es:

$$\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}dt \qquad (3.37)$$

En 1957 Birnbaum utilizó la función logística en vez de la ojiva normal. La expresión matemática es:

$$y=\frac{e^{x}}{1+e^{x}} \qquad (3.38)$$

posteriormente, se aplicó un factor de corrección a esta función para ajustarla a la ojiva normal, dando por resultado

$$y=\frac{e^{Dx}}{1+e^{Dx}} \qquad (3.39$$

En la figura 3.13, se grafican la curva de ojiva normal y la función logística. Se puede apreciar la enorme semejanza entre ambas. Sin embargo, a nivel operativo la función logística tiene una ventaja sobre la curva normal, ya que es más fácil de realizar cálculos matemáticos con ella que con la función de ojiva normal.

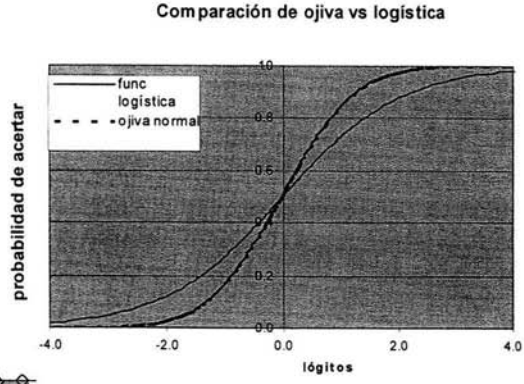
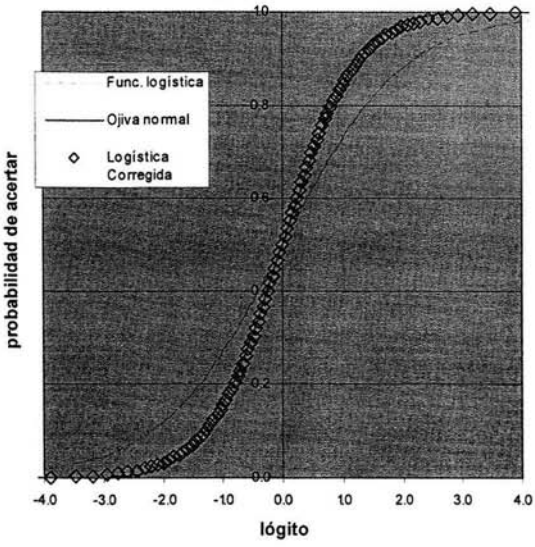


Figura 3.13



La diferencia analítica entre ambas funciones, se resuelve aplicando un factor de corrección cuyo valor es de 1.7, obteniéndose el ajuste que se puede observar en la figura 3.14, en donde prácticamente no existe diferencia ente la ojiva normal y la logística ajustada

Figura 3.14

Modelos de acuerdo al número de parámetros.

La información que se puede obtener de una curva característica de un reactivo permite calcular los índices de dificultad, de discriminación y adivinación, adicionalmente al esquema de ajuste o contraste. El parámetro más importante, sin duda alguna, es el índice de dificultad, seguido del índice de discriminación y, por último, el índice de adivinación. En forma natural, los modelos que sólo se enfocan al índice de dificultad se conocen como modelos de un parámetro; los modelos de dos parámetros serán aquéllos que adicionan el índice de discriminación, y por último si también se toma en cuenta el índice de adivinación, se tendrán modelos de tres parámetros.

A continuación se presentan las funciones matemáticas asociadas a un modelo de ajuste, en este caso, la función logística, dado que ya se mencionó la ventaja de ésta contra la función de ojiva normal.

Modelo de un parámetro.

En primer lugar, tomaremos el modelo de un parámetro que establece que la respuesta a un ítem sólo depende de la competencia del sujeto y de la dificultad del ítem, es decir, de, θ, y de, b; respectivamente. Así, el único parámetro de los ítems a tener en cuenta es, b, el índice de dificultad. La expresión matemática para este modelo, adaptada a la terminología de la TRI, es:

$$P_i(\theta)=\frac{e^{D(\theta-b_i)}}{1+e^{D(\theta-b_i)}}\sqrt{2} \qquad (3.40)$$

donde:

$P_1(\theta)$: Probabilidad de acertar el ítem i a determinado nivel de, θ.

θ: Valores de la variable medida

b_i: Índice de dificultad del ítem.

e: Base de los logaritmos neperianos.

D: Constante con valor 1.7.

Conocido el índice de dificultad de un ítem, b, y la competencia de los sujetos, θ, predice la dificultad, $P(\theta)$, de quienes acierten el ítem.

Modelo logístico de dos parámetros.

El modelo logístico de dos parámetros fue originalmente desarrollado por Birnbaum. Asume que, CCI, viene dada por la función logística y contempla dos parámetros de los ítems, a saber, el índice de dificultad, b, y el índice de discriminación, a, Su expresión matemática viene dada por:

$$P_i(\theta)=\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \qquad (3.41)$$

Modelo logístico de tres parámetros.

El modelo logístico de tres parámetros, es junto con el de Rasch, uno de los que más atención ha recibido. Cada uno de los dos modelos ha tenido sus propios seguidores. Cada modelo se ajusta mejor a unas situaciones que a otras y el uso de uno u otro dependerá en cada caso. El modelo asume que la CCI viene dada por la función logística y añade a los dos parámetros, a, y, b, ya citados, un tercero, c, relativo a la probabilidad de acertar el ítem al azar, cuando no se conoce la respuesta. En una forma más técnica se puede decir que, c_i, es el valor de, $P_i(\theta)$, para un valor, $\theta = -\infty$, el modelo puede expresarse:

$$P_i(\theta)=c_i+(1-c_1)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \qquad (3.42)$$

Algunos autores (como por ejemplo Barton y Lord, 1981 citado en Muñiz, J (1997)) han propuesto, incluso, un modelo logístico de cuatro parámetros para tratar de disminuir el problema real de que a veces, por determinadas circunstancias como el descuido o el uso de información que el constructor del ítem no tuvo en cuenta, los sujetos de alta competencia fallan ítems impropiamente. Existen pocas investigaciones respecto a este tipo de modelos y no parece que aporte ventajas significativas respecto al de los tres parámetros, ya que lo recomendable es evitar que se produzcan problemas, más que tratar de solucionarlos una vez presentados. El modelo viene dado por:

$$P_i(\theta)=c_i+(Y_i-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \qquad (3.43)$$

Donde, Y_i, toma valores ligeramente inferiores a 1 y el resto de los componentes son los ya descritos anteriormente para los otros modelos.

Modelo de Rasch

El modelo logístico de un parámetro fue formulado originalmente por Rasch en 1960, citado en Muñiz, J (1997)y es en la actualidad, sin duda, el modelo más popular de la teoría de la respuesta al ítem, debido en gran parte a la sencillez emanada de su lógica.

El modelo de Rasch establece la probabilidad de respuesta de una persona ante un estímulo dado, en términos de la diferencia entre la medida del rasgo de una persona y la medida del estímulo utilizado. Se trata de un modelo estocástico donde la medida del rasgo de la persona y la medida del estímulo aplicado, quedan ubicadas en una misma escala lineal con un origen común. La variable de interés es la diferencia de ambas medidas.

El modelo establece que la medida del rasgo de la persona es independiente del conjunto de estímulos aplicados. Así mismo se dice que la medida de cada estímulo es independiente del conjunto de personas a las que se somete. En rigor, es la diferencia de medida de rasgo y de medida de estímulo la que es independiente del instrumento o de la población. Además, el modelo de Rasch requiere que la variable sea unidimensional, ordenada e inclusiva.(Tristán, A 1998)

El modelo de Rasch es un modelo de contraste, que hemos señalado con anterioridad, por lo que se parte de la hipótesis de que los reactivos se "deben" comportar como establece el modelo para poderlos aceptar. Si los ítems no se ajustan al modelo, simplemente los

desechamos. Esto es porque partimos de la hipótesis de "cómo deben ser las curvas típicas de los reactivos de buena calidad".

Los modelos de contraste no buscan ajustar una curva a los puntos de manera indiscriminada. Qué mejor, si los reactivos se ajustan al modelo, porque entonces estos se comportarán como establece dicho modelo y conoceremos todas sus propiedades sin lugar a dudas. En cambio, si los puntos no se ajustan a la curva teórica, entonces se deben rechazar los reactivos que no se comporten como predice el modelo. No necesariamente el comportamiento que dice el modelo es el que nos puede gustar, pero es un modelo que finca algunas limitaciones y alcances. Es claro que bajo este enfoque habrá muchos reactivos que no se ajusten al modelo, lo cual implicará que habrá reactivos rechazados. Esto puede no gustarle a los profesores o evaluadores que quisieran que sus exámenes salieran valorados lo mejor posible.

La medición en el modelo de Rasch.

Antes de la aparición del Modelo de Rasch, la escala de medida podría corresponder al número de reactivos, por ejemplo, las preguntas bien contestadas se podrían numerar desde 0 hasta un valor *n*, que podría ser cualquier número. Lo anterior implica que la medición estará siempre acotada por los límites 0 y *n* pero indudablemente podría haber más de un alumno que fuese medido con 0, pero que no todos ellos se encontrarían con el mismo nivel de conocimientos. Sin embargo esta escala no podría diferenciarlos y así mismo habrá alumnos que pudieran contestar los *n* reactivos correctos pero cuyos conocimientos no fuesen necesariamente iguales, pero al igual que lo señalado anteriormente, la escala tampoco podría diferenciarlos, no obstante que la diferenciación es un principio básico en la evaluación. Adicionalmente, si las puntuaciones fuesen *n*, *n/2*, *n/4*, *n/8*, etc., la proporción medida del acervo de conocimientos no corresponde necesariamente a las mediciones relacionadas a través de las expresiones numéricas de mitad, un cuarto o un octavo.

Los problemas esenciales se deben a que la escala no es lo suficientemente extensa para abarcar todas las medidas de todos los sujetos y las divisiones de la escala no informan diferencias constantes de los conocimientos asociados a los posibles resultados de los alumnos, ni las cantidades son directamente proporcionales y calculables con una simple regla de tres.

Se deberá tomar en cuenta que una medida de los conocimientos de una persona no corresponde con un valor preciso y determinado de los conocimientos que atesora en su cerebro sino una probabilidad de respuesta. Cabe señalar que cuando hablamos de personas, a ellas se les asigna una medida a través de cálculos matemáticos más allá de simples

proporciones, pero si el único cálculo es solamente la obtención de una frecuencia relativa, hablaremos de estimar.

Si analizamos las respuestas de las personas, desde el terreno de la probabilidad debemos entonces prepararnos para saber que las puntuaciones de una persona, así como la respuestas a un reactivo, no son valores deterministas ni fijos. Se trata de valores estocásticos y, como tales, pueden ocurrir con mayor o menor probabilidad. Así, no se puede garantizar nunca que una persona tiene una medida, X, sino se habla de que una persona de medida, X, tiene una probabilidad, p, de respuestas correctas y otra probabilidad, q, de respuesta incorrecta.

La medida, X, es una estimación de su medida de dominio real, V. Se trata de una aproximación que contiene un error de medida. Si el instrumento utilizado para la medición es altamente preciso, la medida, X, y el valor real, V, serán prácticamente coincidentes.

Es común que nuestros exámenes contengan un error importante, por lo que el estimado, X, contendrá igualmente un error y podría no parecerse al valor real, V. Es una gran responsabilidad del evaluador conocer las características del instrumento, su calidad, su error, su fiabilidad para poder emitir las mejores estimaciones posibles de la medida real de los sujetos, así como la medida real del instrumento. El evaluador debe estar consciente de estas limitaciones y reconocer asimismo, que su instrumento no es perfecto y deberá tomar las medidas correctivas que sean necesarias para no perjudicar o beneficiar injustamente a los sujetos informándoles de manera errónea la posición que ocupan en la escala.

Una forma muy simple de estimar la probabilidad de respuesta correcta es a través de una frecuencia relativa. Por ejemplo, si se administran 20 preguntas y sólo se contestan 15 preguntas, podemos aproximar que la probabilidad de respuesta correcta es:

$$p = 15/20 = 0.75$$

donde:

p = *Respuesta correctas / Total de reactivos.*

El valor de p estimado depende de los reactivos aplicados y sólo toma en cuenta el éxito del alumno. El fracaso o falla que puede tener el alumno se denomina q, y es el complemento de p a la unidad.

Para este ejemplo q = 1- 0.075 = 0.25, lo cuál quiere decir que la persona tiene un 0.25 de probabilidad de contestar erróneamente el examen, o que podrá fallar 5 preguntas.

El *lógito* y la medida de una persona.

A partir de la inversa de la función logística y con el objeto de contar con una escala de medida que cubra todos los valores posibles del rasgo a evaluar se propone el uso de una nueva unidad de medida lineal que resulta de la razón del éxito p y el fracaso q. A esta nueva unidad de medida. Se le denomina " momio" . El momio o apuesta es una manera de expresar la expectativa que se tiene de éxito con relación al fracaso. Aún más, si se aplica el logaritmo natural al momio, se obtiene una unidad de medida que usada por primera vez en el idioma inglés se le denominó "logit" que es un forma abreviada de "log odd ratio" y que se traduce al español por " logaritmo del momio". Una traducción libre del "logit" propuesta por Tristán, A (1998) es la palabra "lógito".

La interpretación de los lógitos como la probabilidad de respuesta correcta en un examen dado, permite expresar la posibilidad de que una persona conteste acertadamente a un conjunto dado de reactivos en un cuestionario como la medida del rasgo a evaluar de ésta.

El lógito y la medida de un reactivo

Los lógitos también se pueden plantear como una medida para los reactivos. A la medida de los reactivos se le denominará calibración. Así se dice que se está "calibrando" un reactivo en tanto que se acostumbra expresar que se está "midiendo" a las personas.

El grado de dificultad de un reactivo en la Teoría Clásica se define como el cociente de las respuestas correctas entre el número de personas que contestan un reactivo. Mientras más personas contestan el reactivo se hace más fácil y si pocas personas contestan se dice que el reactivo es difícil por lo que el grado de dificultad aumenta; sin embargo, el valor numérico del Grado de Dificultad tal como lo define la Teoría Clásica disminuye, por lo que realmente el Grado de Dificultad Clásico debería de llamarse Grado de Facilidad o sea el inverso del índice que se obtiene numéricamente.

La dificultad, como su nombre lo indica, tiene que ver con la posibilidad de que los alumnos fallen o no respondan correctamente, por ello la medida, b, es función de, q, y no de, p,. Si se hubiera planteado, b, en función de, p, entonces hablaríamos de "facilidad" del reactivo y tendríamos un equivalente al Grado de Dificultad clásico.

Si se plantea la relación matemática que da cuenta de la dificultad como una función de p, es decir, de la posibilidad de acertar, se tiene:

$$b(p) = \ln\left(\frac{p}{q}\right) \qquad (3.45)$$

ahora, si se plantea como una función dependiente de la posibilidad de falla, se genera una función de, q:

$$b(q) = \ln\left(\frac{q}{p}\right) \tag{3.46}$$

En la ecuación 3.46 sustituyendo la variable, q:

$$b(q) = \ln\frac{(1-p)}{p}$$

tomando la función inversa:

$$e^{b(q)} = \frac{(1-p)}{p}$$

ahora se propone establecer la función de probabilidad de acertar en un ítem en términos de la dificultad, b:

$$pe^{b(q)} = 1 - p$$

$$p + pe^{b(q)} = 1$$

$$p\left(1 + e^{b(q)}\right) = 1$$

$$p = \frac{1}{1 + e^{b(q)}} \tag{3.47}$$

es decir, la función de probabilidad se hace depender de la dificultad, b, y se interpreta como la probabilidad de no acertar un ítem dada una dificultad, b.

El modelo de Rasch establece que solamente puede obtenerse, b, a partir de la información disponible. La curva de Rasch es el lugar geométrico de una función logística que tiene una pendiente única. El parámetro, b, corresponde con la medida de la persona o la calibración del reactivo y se obtiene para el punto de inflexión que para esta curva siempre está en,

p =0.5.

Escalograma de Guttman

Los modelos tanto de Rasch como los de dos y tres parámetros tienen un supuesto inherente: tratan de medir una sola variable, así se dice que son modelos unidimensionales o que la escala y la medición hacen referencia a una sola dimensión.

Cuando se habla de dimensión nos referimos a que el reactivo o conjunto de reactivos que forman una prueba hacen referencia a un rasgo exclusivamente, por ejemplo: habilidad

matemática. Cuando en un cuestionario tenemos una mezcla de varios rasgos o áreas de conocimiento, por ejemplo: habilidad verbal, matemáticas, civismo, inglés, debe ser obvio que algunos reactivos no necesariamente van a correlacionar con todo el cuestionario, ni es obligatorio que correlacionen con otros.

Pero si aceptamos que estamos midiendo un solo rasgo, entonces un reactivo mide una sola cosa que es factible de ubicarse en la escala. Así podemos dibujar la escala de medida (que para fines prácticos se ubica entre -4 y +4) y aceptar que el resultado obtenido del reactivo corresponde a una "marca de la escala" como se puede observar en la siguiente figura 3.15.

Construcción de un punto en la escala unidimensional

Figura 3.15

A la escala que empleamos le denominaremos métrica y debe ser lo más uniforme posible y, en especial, para la variable en estudio.

Se hace necesario que se tenga un buen número de reactivos en orden creciente de dificultades y, en ocasiones, será conveniente poner más reactivos espaciados de manera más fina para poder disponer de una medida más precisa del rasgo que queremos medir. (figura 3.16)

Figura 3.16

Sin embargo no hay que abusar poniendo un número exagerado de reactivos porque podemos generar cansancio a las personas y hacer que el tiempo de solución de la prueba se alargue. Tampoco se deben de colocar dos reactivos con el mismo grado de dificultad porque implicaría "repetir" el nivel de dificultad. Esto equivale a decir que no conviene poner reactivos de igual dificultad para medir el mismo rasgo, porque estaríamos desperdiciando un reactivo en un mismo nivel de medida.

A este respecto conviene recordar que los reactivos tal vez no son tan perfectos como imaginamos. Pero ¿qué pasaría si nuestros reactivos fueran tan variados?

Tal vez todos los reactivos miden la misma habilidad, pero tienen "errores" de medida diferentes. Debido a estos errores es seguro que aunque la dificultad se parezca, habrá diferencias de capacidad entre las personas, por lo que cada reactivo "discriminará" de una manera diferente.

Un rasgo es una dimensión que se va medir con reactivos de dificultades y discriminaciones muy diversas. Una forma de representar las respuestas de las personas a nuestro examen consiste en hacer una tabla donde anotaremos en cada renglón las personas y en las columnas las respuestas a cada reactivo.(Tabla 3.5

Tabla 3.5

	Respuestas a cada reactivo							
Persona	1	2	3	4	5	6	7	8
1	1	1	0	1	0	1	0	0
2	1	1	0	1	1	0	0	0
3	1	1	0	1	0	0	0	0
4	1	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	0
8	1	1	1	1	1	0	1	1
9	1	1	1	1	0	0	0	0
10	1	0	0	0	0	1	0	0
11	1	1	1	0	0	0	0	0
12	1	0	1	1	0	0	0	0
13	1	1	1	1	1	1	0	0
14	1	1	1	1	0	1	1	0

Las respuestas que aparecen en esta tabla son obtenidas de acuerdo con el nivel de dominio o de habilidad que tiene cada persona al enfrentarse a cada uno de los reactivos. Esperamos que las repuestas sigan algún "patrón" del tipo: "a mayor dominio mayor número de respuestas". Sin embargo, un alumno de cualquier nivel de dominio puede fallar en un reactivo "fácil" por diversos motivos (error de lectura, distracción, malas instrucciones en la pregunta, etc.). Se dice entonces que los valores de esta tabla son estocásticos.

Las hipótesis involucradas en el Escalograma de Guttman son:

a) La variable es unidimensional- Mide un solo rasgo. No se permite juntar o analizar rasgos de diversas dimensiones en una misma tabla. Dicho de otro modo: no se deben

analizar en el mismo Escalograma por ejemplo: las respuestas de habilidad verbal y las de Matemáticas.

b) La variable es ordenada- El rasgo tiene una relación directa con la dificultad del reactivo. Por ello el conjunto de reactivos puede ordenarse en dificultad , de tal forma que un reactivo donde contestan 10 personas es más difícil que el reactivo que es contestado por 20 personas y, a su vez, más difícil que el reactivo contestado por 80 personas

c) La variable es inclusiva- El dominio en un reactivo implica dominio en reactivos más fáciles. Si una persona contesta un reactivo de valor de "30" entonces sería de esperarse que se posee el dominio de los reactivos de valores iguales o menores de 30.

El escalograma de Guttman trata de ordenar a las personas en orden descendente de aciertos (de la que tiene más aciertos a la que contestó menos) y a los reactivos en forma ascendente de aciertos (del reactivo que tiene más respuestas, es decir, del más fácil al que tiene menos respuestas o más difícil). Después de ordenar a las personas y a los reactivos obtenemos el Escalograma de Guttman como sigue:(Tabla 3.6)

Tabla.3.6

	Respuestas a cada reactivo							
Persona	7	2	5	4	3	8	6	1
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	0
8	1	1	1	1	1	0	1	1
13	1	1	1	1	1	1	0	0
14	1	1	1	1	0	1	1	0
1	1	1	0	1	0	1	0	0
2	1	1	0	1	1	0	0	0
9	1	1	1	1	0	0	0	0
11	1	1	1	0	0	0	0	0
12	1	0	1	1	0	0	0	0
3	1	1	0	1	0	0	0	0
4	1	1	0	0	0	0	0	0
10	1	0	0	0	0	0	1	0
5	1	0	0	0	0	0	0	0

Como estamos midiendo una sola variable o dimensión, el escalograma es una buena herramienta que nos permite ver algunas cosas interesantes con relación al reactivo y con relación a las respuestas de las personas.

Podemos apreciar que la persona 6 contestó todo el cuestionario y que la persona 5 fue la que contestó menos. La construcción ayuda, por lo tanto, a ubicar a los alumnos de manera rápida.

Vemos que algunos alumnos contestan los reactivos fáciles y que a partir de cierto reactivo dejan de contestar, esto lo señalamos con color gris oscuro dentro de la tabla. Se puede afirmar que estas personas contestan de acuerdo con un "patrón lógico" a saber: solamente contestan las preguntas debajo de su nivel de dominio. Una vez alcanzado dicho nivel las preguntas son más difíciles de lo que pueden contestar y a partir de ahí no pueden responder correctamente.

Pero también vemos algunos casos de personas que no contestan con el "patrón lógico", esto lo identificamos con un gris claro, como es el caso de las personas 8, 14, 1, 2, 12, 3 y 10. No obstante, podemos ver una diferencia substancial entre las personas 4 y 10, ambos tienen 2 aciertos, pero la persona 4 está en el "patrón lógico"y la persona 10 en cambio responde una pregunta que es muy difícil por lo que podríamos conjeturar que contestó el reactivo 6 por adivinación (o tal vez se le preguntó acerca de una cosa que sabía por razones no explicadas). El patrón que muestra el Escalograma de Guttman permite hacer conjeturas respecto a la adivinación personal. Es probable que la persona 10 adivinó el reactivo 6, pero seguramente la persona 4 no adivinó en sus respuestas.

Al revisar el escalograma en el sentido vertical, observamos que el reactivo 7 es contestado por todas las personas, por lo tanto es muy fácil y no discrimina entre los alumnos. Mientras que el reactivo 1 es muy difícil y solamente algunos alumnos pudieron contestarlo. El escalograma nos ayuda a ordenar los datos y revisarlos rápidamente.

Si comparamos el reactivo 4 con el 3 podemos ver que el reactivo 4 tiene un patrón más lógico que el reactivo 3. Así mismo el reactivo 6 (uno de los más difíciles) tuvo una respuesta inesperada de la persona 10. El patrón del reactivo 6 no es más lógico y podría atribuirse a una respuesta al azar del alumno, no tanto a que el reactivo favorezca las repuestas del grupo inferior.

De lo anterior se observa que dada una habilidad θ y un reactivo de dificultad b, si $\theta > b$ el reactivo se le hará "fácil" a la persona y por lo tanto tiene probabilidad de acertar. Si $\theta < b$ el reactivo se le hará difícil a la persona y por lo tanto tiene probabilidad de fallar.

Podemos considerar que el escalograma identifica a los reactivos y a las personas de tal modo que se tiene una configuración como la Tabla 3.7.

Llamaremos "diagonal" del Escalograma a la línea que se traza de abajo a la izquierda hacia arriba a la derecha y que divide al Escalograma aproximadamente en dos zonas triangulares. Observamos que arriba de la diagonal se esperan aciertos y que abajo se esperan errores.

Tabla 3.7

	Respuestas a cada reactivo							
Persona	1	2	3	4	5	6	7	8
A	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	1	0
C	1	1	1	1	1	1	0	0
D	1	1	1	1	1	1	0	0
E	1	1	1	1	0	0	0	0
F	1	1	1	0	0	0	0	0
G	1	1	0	0	0	0	0	0
H	1	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0

Dado que se asume que las respuestas son estocásticas, no todos los valores arriba de la diagonal serán 1, habrá necesariamente algunos errores normales, atribuibles a fallas de comprensión de la pregunta, a buenas opciones distractoras en un reactivo de opción múltiple, etc. Igualmente se puede esperar que no todas las respuestas abajo de la diagonal serán 0. Habrá presencia de aciertos y fallas en forma estocástica.

Definimos el Escalograma Perfecto al que cumple la no estocasticidad, es decir garantiza que arriba de la diagonal se tienen siempre aciertos y debajo de la diagonal se tienen siempre errores, como el que se muestra en la tabla 3.7.

Es prácticamente imposible encontrar un Escalograma Perfecto. Por lo mismo, aunque es el mismo modelo previsto, no debemos forzar los cuestionarios a que se vuelvan escalogramas perfectos ni esperar que se nos presenten. Es más, si se llega a encontrar un Escalograma Perfecto al analizar los datos de un examen, entonces habrá serios problemas para usar esta herramienta dentro del análisis de Rasch.

El concepto de "Error"

A partir del Escalograma se puede ver que hay alumnos que responden con lo que hemos denominado un "patrón lógico" y otros que tienen respuestas inesperadas. Entendemos como respuestas inesperadas los aciertos que ocurren en reactivos más difíciles con relación a la medida de la persona, o bien respuestas incorrectas en preguntas más fáciles que la medida de la persona.

Para ponderar adecuadamente el "error" entendido éste como la diferencia entre el patrón lógico de respuesta se sigue el siguiente procedimiento: construir el Escalograma de Guttman e identificar la diagonal que separa a las dos zonas triangulares de acierto y error. Hacer un conteo simple de respuestas correctas inesperadas (un 1 en el triángulo inferior) y de respuestas incorrectas inesperadas (un 0 en el triángulo superior); a este conteo se le denomina " error". Desde luego este enfoque aún es deficiente por la siguiente razón: al observar de nuevo la tabla 3.8 se aprecia que los patrones de respuestas de repuestas de las personas identificadas con los numerales 11 y 12 son diferentes, no obstante que tengan exactamente el mismo número de respuestas inesperadas. Asimismo el patrón de respuestas de la persona 10 tiene seguramente, un error más grande que los casos anteriores. Así, no es solamente el número de respuestas inesperadas, sino la posición relativa respecto a la medida lo que indica el error.

Tabla 3.8

	Respuestas a cada reactivo							
Persona	7	2	5	4	3	8	6	1
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	0
8	1	1	1	1	1	0	1	1
13	1	1	1	1	1	1	0	0
14	1	1	1	1	0	1	1	0
1	1	1	0	1	0	1	0	0
2	1	1	0	1	1	0	0	0
9	1	1	1	1	0	0	0	0
11	1	1	1	0	0	0	0	0
12	1	0	1	1	0	0	0	0
3	1	1	0	1	0	0	0	0
4	1	1	0	0	0	0	0	0
10	1	0	0	0	0	0	1	0
5	1	0	0	0	0	0	0	0

CAPÍTULO 4. METODOLOGÍA DE LA INVESTIGACIÓN

Trabajo de campo

Como ya se ha mencionado, existen diferentes tipos de opiniones que descalifican o alaban, sin prueba científica que los sustente, alguno de los formatos de opción múltiple y de respuesta abierta, por lo que se desea establecer, en principio con un estudio de caso, si realmente la forma de preguntar en un examen afecta al rendimiento de los alumnos, es por ello que surge la primera pregunta de investigación, que es *¿Existen diferencias estadísticas significativas en el valor calculado del índice de dificultad de los ítems ante dos formatos diferentes de tests, uno de ellos de opción múltiple y otro de respuesta abierta o tipo ensayo?.* Para dar respuesta a está pregunta, utilizaremos la teoría de la respuesta al ítem, siguiendo el modelo de Rasch, lo que origina una segunda pregunta de investigación *¿Los índices de dificultad de los ítems, con base a la teoría de la respuesta al ítem, se ven afectados cuando el tamaño de la muestra varía?*

La metodología que se sigue en el trabajo de campo para la recopilación de información que da sustento a la investigación, se partió de la consideración siguiente: se desconoce el índice de dificultad de los ítems y se desconoce la habilidad en el área de conocimiento a evaluar de los alumnos. Los ítems se seleccionaron de un banco de reactivos que fueron preparados por un grupo de expertos, quienes en su diseño consideraron que son congruentes con los objetivos de la materia y que usualmente se emplean en exámenes equivalentes al de la investigación.

El examen se aplicó a un grupo de alumnos de una generación de nivel licenciatura, en el área de ciencias, de una universidad pública, cuya característica común era la de adeudar la materia sujeta del examen, es decir, alumnos que en al menos en una ocasión no acreditaron la materia en su curso regular.

El diseño del examen se realizó con apego a un cuadro de especificaciones como el propuesto por Canut y Jiménez (2000) y se sometió a la censura de los sinodales responsables de la evaluación de los exámenes extraordinarios designados por la institución, quienes aceptaron por única vez.

Los ítems estuvieron diseñados para explorar los conocimientos de la primera asignatura de una serie de tres, todas vinculadas al cálculo diferencial e integral. Si bien todos los alumnos fueron sujetos de los mismos ítems, éstos fueron presentados en dos órdenes diferentes, aunque a los alumnos no se les exigió que diesen sus respuestas en algún orden particular.

El grupo al que se le aplicó el examen estuvo constituido por 110 alumnos y el examen lo conformaron 11 ítems que, para efectos del experimento, se presentaron en dos formatos diferentes, a saber, uno de opción múltiple y otro de respuesta abierta., A, cada formato se le asignaron el mismo número de preguntas. Para poder hacer la contrastación entre ambos formatos, las mismas preguntas se presentaran en cada uno de los formatos señalados. En cada uno de los dos diferentes formatos se asignó a cada pregunta un orden diferente, como ya se señaló. La asignación de cualesquiera de los dos formatos de examen a los estudiantes, se realizó de forma aleatoria. Así entonces, se generaron dos subgrupos o submuestras llamadas para fines prácticos como grupo, A, y grupo, B. Al grupo completo –en la cual el orden de los ítems no es considerado como una posible variable de estudio, se le designó como grupo, AB.

Procesamiento de datos.

Para lograr el análisis de los ítems se comienza con la *edición* de los datos; esto implica que las respuestas deberían ordenarse en una matriz cuyas filas correspondan a los datos de cada persona y las columnas a los datos por cada ítem. Para efectos de la investigación, las respuestas se consideraron sólo dicotómicas, es decir, una respuesta solamente se consideró como correcta o incorrecta; si fue contestada correctamente por un estudiante quedó registrada, para su procesamiento como 1, y una respuesta incorrecta como 0.

Los puntajes de los alumnos, es decir, el número total de respuestas correctas logradas por cada estudiante, se presenta al final de cada renglón en la última columna de la derecha. Los puntajes de los ítems, es decir, el número total de respuestas correctas a cada ítem, se presentan en la parte inferior de cada columna. (Por ejemplo Tabla 4 de anexos)

Cuando un ítem no es contestado por persona alguna de la muestra, se considera que éste fue demasiado difícil, para esta muestra de personas, y en consecuencia el ítem no aporta mayor información sobre qué tanto es difícil en realidad. Cuando todo el mundo acierta a un ítem, eso muestra que el ítem es demasiado fácil para estas personas, pero una vez más no se cuenta con información sobre cuán fácil es este ítem. Debe ser claro que en el caso de las personas extremas, un puntaje de cero no significa que no sepan nada, y una persona con un puntuaje del 100% no indica que la persona " lo sabe todo ". No obstante, los puntajes de 0 y 100%, ya sea para ítems y personas, aunque representan información incompleta, sí nos indican en qué dirección buscar un cálculo de la habilidad de la persona o de la dificultad del ítem.

En la investigación, para el tratamiento y análisis de los datos recogidos, se utilizaron dos tipos de procedimientos sobre la misma base teórica del modelo de Rasch, a saber: la

calibración de ítems en forma manual y la calibración de ítems utilizando el programa informático llamado XCALIBRE.

Para la calibración manual se siguió el procedimiento llamado PROX, el cuál logra los objetivos básicos del análisis de ítems de Rasch, a saber: la linearización de la escala latente y el ajuste para los efectos locales de la distribución de habilidades de la muestra. La suposición que hace PROX está basada en los efectos de calibración de ítems de la distribución de habilidades de la muestra puede explicarse por una media y la desviación estándar. Esta suposición hace que PROX sea tan simple, que puede aplicarse sin necesidad de recursos electrónicos, es decir, los cálculos son fácilmente realizables a mano.

Calibración manual.

Para cada uno de los subgrupos, A, B y AB, se han realizado los pasos para la calibración de ítems y la medición de las personas, siendo aquéllos los que a continuación se describen:

Paso 1° Organización de la matriz de datos.

La matriz de datos se ordena, de tal forma que los puntajes de la persona queden ordenados de menor a mayor con sus proporciones respectivas dadas en la columna de la extrema derecha, y los puntajes de los ítems estén ordenados de mayor a menor con sus proporciones dadas en el renglón inferior. (Tablas 4, 5 y 6)

Paso 2° Edición de la matriz de datos.

Para la edición de la matriz de datos en la respuestas del las personas por ítem, se eliminan todos los ítems que fueron contestados correctamente por todos, o por nadie, y se eliminan todas las personas que obtuvieron puntajes perfectos o que no habían respondido a ningún ítem correctamente. Esto implicó que .de las 110 personas que presentaron el examen sólo sirviesen para el análisis 95. En las tablas 4, 5 y 6 del anexo se muestran sombreadas los nombres de las personas que fueron eliminadas por el proceso de edición.

En el estudio el ítem 4 fue eliminado en el caso del grupo, A, y el ítem 9 en el grupo, B. Todos los demás ítems se conservaron ya que tuvieron al menos un alumno que los contestó correctamente tablas 7,8 y 9 se muestran sólo los ítems que no fueron eliminados.

En nuestra investigación se puede observar que en el grupo, AB, de alumnos, los 11 ítems fueron contestados por lo que no tuvimos que eliminar ninguno. Para hacer un cálculo definitivo para una persona, al igual que en los ítems, debemos encasillar a esa persona entre ítems que sean más fáciles y más difíciles que la habilidad de esa persona.

Paso n° 3 obtención de las calibraciones de ítems iniciales.

Construimos una tabla de distribución de frecuencias para los 11 ítems diferentes y sus lógitos de dificultad (incorrectos), y calculamos la media y la varianza de la distribución de estos lógitos de ítems sobre la prueba. El contenido específico de cada una de las columnas que conforman las tablas 10, 11 y 12, correspondientes a los grupos A, B y AB, se describen a continuación.

Columna 1

Contiene los nombres asignados a cada grupo de los ítems recabados de acuerdo al puntaje de ítem.

Columna 2

Contiene el puntaje del ítem que caracteriza a cada grupo de puntaje de ítem

$$S_i \text{ con } i=1, 2, \ldots, G.$$

En el caso de la investigación, de acuerdo a las diferentes puntuaciones de ítems se tiene:

$$G_A=10,\ G_B=9,\ G_{AB}=10$$

Columna 3

Indica la frecuencia de los ítems en cada puntaje. La suma de estas frecuencias, L_A=10, L_B=10, L_{AB} =11, sobre los grupos de puntaje del ítem, G_A = 10, G_B = 9, G_{AB} = 10, que corresponden a los ítems que se están calibrando:

f_i = frecuencia del ítem ***i***

$$L=\sum_{i}^{G} f_i \qquad (4.1)$$

Columna 4

Convierte los puntajes del ítem en proporciones de aciertos de muestra de, N_A = 58, personas, N_B = 36 y N_{AB} = 94.

$$P_i = \frac{s_i}{N} \qquad (4.2)$$

Columna 5

Presenta los resultados de la conversión de esta proporción de aciertos, p_i, en la proporción de error es, 1- p_i.

$$1-p_i=\frac{(N-s_i)}{N} \qquad (4.3)$$

Columna 6

Presenta la conversión de las proporciones en lógitos de error. Cada lógito del grupo de puntaje del ítem es el logaritmo natural de su proporción de error dividida entre su proporción de aciertos.

$$x_i = \ln\left[\left(\frac{1-p_i}{p_i}\right)\right] \qquad (4.4)$$

Columna 7

Contiene el producto de la frecuencia del ítem por el lógito de error.

$$f_i x_i \qquad (4.5)$$

Columna 8

Presenta el producto de las frecuencia del ítem por el lógito de error cuadrado.

$$f_i x_i^2 \qquad (4.6)$$

Para los obtener los valores de la columna 9, es necesario, primeramente calcular la media para los lógitos del ítem de la columna 6; este estadístico se calcula a partir de las sumas de los elementos de las columnas 7 y 3.

$$\overline{x} = \frac{\sum_i^g f_i x_i^2}{L} = \frac{\sum_i^g f_i x_i^2}{\sum_i^g f_i} \qquad (4.7)$$

Columna 9

Presenta la diferencia de los valores de la columna 6 con la media. Estas son las calibraciones iniciales del ítem listas para ser corregidas para el efecto de la dispersión de la muestra.

$$d_i^o = x_i - x. \qquad (4.8)$$

Para la corrección será necesario contar con el estadístico *varianza*. Para los obtener los valores de la varianza, primeramente, se presenta en la columna 10 la media que ya se ha calculado. A continuación, en la columna 11, se presenta el producto de la frecuencia de cada ítem, datos de la columna 3, por los datos del cuadrado de la columna 10.

$$U = \left(\frac{\left(\sum_i^g f_i \cdot x_i^2 - L \cdot \overline{x}^2\right)}{(L-1)}\right) \qquad (4.9)$$

Paso 4 Obtención inicial de las mediciones de las personas

A continuación se describen los pasos, que son equivalentes a los descritos en la distribución agrupada de puntajes de personas para obtener la distribución de logits de puntajes de personas y los valores iniciales que van con cada puntaje posible en la prueba para la construimos de la tabla de distribución de frecuencias, que se siguen para la consecución de las mediciones iniciales de las personas. (tablas 16,17 y 18 de los anexos)

Columna 1

Lista cada uno de los puntajes posibles de las personas, denotados por, ***r*** ,desde 1 hasta L-1.

$$r = 1,\ 2,\ \ldots,\ L-1$$

Columna 2

Presenta la frecuencia observada, n_r, de las personas en cada puntaje. La suma de las frecuencias desde, $r = 1$ a $r = 10$ da el número total de las personas

$$N = \sum_{r}^{L-1} n_r \tag{4.10}$$

por lo tanto, $N_A = 58$, $N_B = 36$ y $N_{AB} = 94$.

Columna 3

Es la proporción de cada puntaje en una prueba con L ítems

$$p_r = \frac{r}{L} \tag{4.11}$$

Columna 4

Es el lógito correcto para esta proporción calculada en la *columna 3*

$$y_r = \ln\left[\frac{p_r}{(1-p_r)}\right] \tag{4.12}$$

Columna 5

Es el producto de la frecuencia de las personas por el lógito correcto

$$n_r y_r \tag{4.13}$$

Columna 6

Es el producto de la frecuencia de las personas por el cuadrado del lógito correcto

$$n_r y_r^2 \tag{4.14}$$

Columna 7

Corresponde a la medición inicial de las personas, denotadas por, b_r^o, cuyos valores numéricos son los mismos de la columna 4.

Para la obtención de la media y la varianza asociada a la distribución que se ha descrito, se sigue el siguiente proceso:

Para una prueba de, L ítems dados a, N, personas hemos borrado todos los ítems que nadie acertó y los que nadie falló, así como todas las personas con ningún acierto y ninguna respuesta incorrecta hasta que no quede ningún ítem ni persona de este tipo.

Mientras permitimos que, Si, sea el número de personas que acertaron al ítem, i, para $i = 1$ hasta, L, y que, n_r, sea el número de personas que obtuvieron r ítems correctos, para, $r = 1$ a través de, L-1; definimos la media y la varianza sobre ítems de las respuestas incorrectas de lógitos en la muestra para cada uno de los ítems, L, y la media y la varianza sobre las personas N.

Por tanto obtenemos, para cada ítem, i, el lógito de sus respuestas incorrectas entre la muestra de, N, personas, así

$$x_i = \ln\left[\frac{(N - S_i)}{S_i}\right] \qquad (4.14)$$

y la media y la varianza sobre L ítems de estos lógitos de ítem

$$x. = \frac{\sum_i^L x_i}{L} \qquad (4.15)$$

y

$$U = \sum_i^L \frac{(x_i - x.)^2}{L-1} \qquad (4.16)$$

ahora, obtenemos para cada puntaje, r, su lógito de respuestas correctas en la prueba de, L, ítems, usando

$$y_r = \ln\left[\frac{r}{(L-r)}\right] \qquad (4.17)$$

y la media y la varianza sobre, N, personas de sus lógitos de puntuación.

$$y. = \sum_r^{L-1} \frac{n_r y_r}{N} \qquad (4.18)$$

y

$$V = \sum_{r}^{L-1} n_r \frac{(Y_r - Y.)^2}{(N-1)} \qquad (4.19)$$

Con lo anterior, se asume que estamos preparados para ajustar las calibraciones y mediciones iniciales, para los efectos locales de la distribución de habilidad de la persona de la muestra y la distribución de la dificultad del ítem.

La media y la varianza para la distribución de los lógitos de puntajes sobre personas, se presenta también en las tablas 16, 17 y 18 de los anexos.

Paso 5. Cálculo de los factores de expansión

Calculamos los factores de expansión para los cálculos iniciales de calibraciones de ítem y mediciones de personas para corregir las calibraciones de ítem para distribución de muestra y las mediciones de personas para el ancho de la prueba.

A partir de las tablas anteriores obtuvimos las respectivas, U, y V, para cada uno de los tres grupos.

a. La expresión matemática para calcular el factor de expansión de la habilidad de la persona debido al tamaño de la prueba es

$$x = \left[\frac{1 + \frac{U}{2.89}}{1 - \frac{UV}{8.35}} \right]^{1/2} \qquad (4.20)$$

b. El factor de expansión de la dificultad del ítem debido a al dispersión de la muestra se calcula con:

$$Y = \left[\frac{1 + \frac{V}{2.89}}{1 - \frac{UV}{8.35}} \right]^{\frac{1}{2}} \qquad (4.21)$$

Paso 6 Corrección de calibraciones de ítem para el efecto de la dispersión de la muestra.

Obtenemos las calibraciones finales del ítem corregido y sus errores estándar a partir del factor de expansión de la muestra, $Y.$, Los resultados para cada grupo se presentan en las tablas 13, 14 y 15 de los anexos.

Columna 1

Contiene los nombres de los ítems.

Columna 2

Repite las calibraciones iniciales a partir de la columna 9 de las tablas 10, 11 y 12.

$$d_i^o = x_i - \overline{x} \qquad i = 1,\ 2,\ \ldots,\ G \qquad (4.22)$$

(recuérdese que cuando los ítems están agrupados por su puntajes entonces la i corre de 1 a G el número de grupos de puntaje en lugar de 1 a L de los ítems individuales).

Columna 3

Presenta, en todos los renglones, el factor de expansión de la dificultad del ítem debido a la dispersión de la muestra.

$$Y \qquad (4.24)$$

Columna 4

Contiene las calibraciones del ítem, una vez corregidas por el factor de expansión:

$$d_i = Y d_i^o \qquad (4.25)$$

$$d_i = Y\left(x_i - \overline{x}\right) \qquad (4.26)$$

Columna 5

Presenta el número de personas que acertaron a cada uno los ítems en cada grupo de puntajes del ítem.

$$S_i \qquad (4.27)$$

Columna 6

Indica el error estándar de las calibraciones corregidas del ítem.

$$SE\left(d_i\right) = Y\left[\frac{N}{S_i\left(N - S_i\right)}\right]^{1/2} \qquad (4.28)$$

Paso 7 Corrección de las mediciones de personas para el efecto del ancho de la prueba.

Se obtienen las mediciones finales de personas corregidas y sus errores estándar a partir del factor de expansión del ancho de la prueba, X,. Los resultados para cada grupo se presentan en las tablas 13, 14 y 15.

Columna 1

Presenta todos los puntajes posibles debido a que queremos tener mediciones disponibles para cada puntaje posible de la prueba desde 1 hasta, L -1, cualesquiera puntajes que en realidad se hayan observado.

$$r = 1,\ ,\ L - 1$$

Columna 2

Repite las mediciones de personas iniciales de la columna de las tablas 16, 17 y 18.

$$b_r^o = Y_r \qquad (4.28)$$

Columna 3

Corresponde al factor de expansión de habilidad de la persona debido a la amplitud de la prueba.

$$X \tag{4.29}$$

Columna 4

Corresponde a las mediciones de la persona obtenidas multiplicando cada valor inicial de la columna 2 por el factor de expansión contenido en la columna 3.

$$b_r = xb_r^o = xY_r \tag{4.30}$$

Columna 5

Presenta el error estándar de las mediciones corregidas de las personas..

$$SE(b_r) = X\left[\frac{L}{r(L-r)}\right]^{1/2} \tag{4.31}$$

Procesamiento de datos con el Programa X CALIBRE

El procedimiento que se describirá a continuación fue aplicado, a cada uno de los grupos de la investigación. El programa del ordenador, nos permitirá conocer el índice de dificultad del ítem y la habilidad de los alumnos, calculados en forma simultánea. Los resultados del programa se van a comparar con la calibración de los ítems en forma manual utilizando la técnica de Rasch.

A continuación se muestra la forma de llevar a cabo la aplicación del programa X Calibre; la descripción del proceso se hará sólo para el grupo B y deberá entenderse que el programa se corre también para los grupos, *A,* y *AB,* con los ajustes pertinentes,.

Los datos obtenidos en el campo para el grupo, *B*, deben prepararse en una base de datos ad-hoc, utilizando para ello el programa excel. Los datos iniciales presentaron 47 alumnos y once ítems; sin embargo el ítem número 9 tuvo que ser eliminado ya que no obtuvo ninguna respuesta correcta, quedando en consecuencia sólo 10 ítems. Además, se asignó como clave para la no respuesta al número 8. La clave para que el Programa X Calibre diferencie entre las respuestas dicotómicas fue de cero si la respuesta es incorrecta y de uno, si la respuesta es correcta. Adicionalmente, se tuvo que realizar una asignación numérica de ingreso al Programa, que debería ir apareada al nombre original del ítem.

A la base de datos así preparada, se le deben declarar algunos datos que el programa utilizará en el procedimiento para la asignación de los parámetros requeridos. A continuación se describe el procedimiento de asignación y declaración de datos generales.

Procedimiento de declaración de datos generales:

a) A continuación se especifica la forma de introducir los datos en el primer renglón, y que por definición son lugares asignados a priori por el programa.

Columnas 1,2 y 3 para registrar el número total de ítems			columna 4	Columna 5	columna 6	Columna 7	columna 8	columna 9	Columna 10
1	2	3		Indicador de no respuesta Puede usarse un digito, un signo de puntuación etc.		Indicador de No respuesta a más de una opción puede usarse un digito, un signo de puntuación, una letra, etc.			2 dígitos de 0 a 9 alumnos, 3 de 10 a 99, 4 de 100 a 999, 5 de 1000 a 9999, etc. Incluye el espacio entre el código numérico del sujeto y sus respuestas

A continuación se introducen los datos específicos del grupo *B* en el primer renglón, quedando las columnas de éste como se ilustra enseguida:

	1	1		8		9			3

b) Descripción para el segundo renglón:

Se coloca la clave de las respuestas correctas, empezando en la columna 1.

A continuación se presentan los renglones 1 y 2 después de haberse asignado los parámetros y claves derivadas de los datos.

	2	0		8		9			3
1	1	1	1	1	1	1	1	1	1

c) Tercer renglón

Se registra el número de opciones que el alumno tiene para contestar, ejemplo

2 para falso o verdadero (correcto o incorrecto)

3 para a, b, c,

4 para a, b, c, d

5 para a, b, c, d, e

etc.

A continuación se presentan los renglones 1, 2 y 3 después de haberse asignado los parámetros y claves derivadas de los datos.

	1	1		8		9			3
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2

d) Cuarto renglón o fila

Se registra:

Y = cuando el ítem sí se analiza.

N = cuando el ítem no se analiza.

F = cuando el ítem es una ancla.

A continuación se presentan los renglones 1, 2, 3 y 4 después de haberse asignado los parámetros y claves derivadas de los datos.

	1	1		8		9			3
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

e) Quinto renglón.

Del quinto renglón en adelante se vacían los datos correspondientes a la identificación de cada alumno y su respectivo patrón de respuestas.

Procedimiento para la obtención de una base de datos que pueda ser leída por XCalibre.

i. Guardar el archivo en Excel como xcali.cvs (delimitado por comas)

ii. Cerrar Excel. Al cerrar seguramente aparecerá una ventana de diálogo en la que se señala que se perderán las características de libro y preguntará si se quieren conservar tales características. Hacer clic en **No**.

Para que no se pierdan los datos del archivo cvs, en el momento en que aparezca el cuadro de diálogo, se deberá hacer clic en **sí**, y aparece otra ventana preguntando si se requieren guardar los cambios hacer clic en **si**. El archivo generado es un archivo con comas.

iii. Abrir con x cali.Cvs en word. Reemplazar todas las comas por un espacio en blanco.

iv. Guardar como texto, sin formato con terminación *.dat. De la instrucción, dependiendo del software de la máquiba pueden suceder dos cosas. La primera que guarde como *.dat, lo cual es lo deseado; la segunda que guarde como *.txt, lo cual se debe modificar desde MSDOS.

Aplicación de XCalibre a la base de datos.

i. Abrir el programa XCalibre.

ii. Se introduce el nombre del archivo del cual se deberá el programa leer los datos.

iii. Se siguen las instrucciones de los cuadros de diálogo que emite el propio programa, con las siguientes especificaciones: la distribución que se seleccione será *commun*, con *punto flotante*, y modelo de *de dos parámetros*.

Una vez que corre el programa, automáticamente arroja los datos requeridos. Para el trabajo de la investigación, los resultados están el las tablas 19, 20 y 21.

CAPÍTULO 5. ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

En este capítulo se hará una interpretación de los datos recopilados en el trabajo de campo relativos a la investigación y analizados a través de herramientas estadísticas. Como ya se señaló, en páginas previas, dado que existen creencias que descalifican a priori los tests objetivos, particularmente los de formato de opción múltiple, en razón de que se cree que dicho formato no logra medir desempeños de los individuos a niveles que de acuerdo a la taxonomía de Bloom corresponderían a los de análisis, aplicación y síntesis y, en contrapartida, existe otra corriente antagónica que señala que los tests debieran ser exclusivamente de formato de opción múltiple argumentando que dichos tests son independientes del criterio del evaluador, (en este caso, del profesor) nuestra investigación de caso se ha abocado a establecer si existen o no diferencias estadísticamente significativas entre los resultados obtenidos en uno u otro formato. Así, nos planteamos una vez más la pregunta esencial a la cual se ha deseado dar respuesta es *¿Existen diferencias estadísticas significativas en el valor calculado del índice de dificultad de los ítems ante dos formatos diferentes de tests, uno de ellos de opción múltiple y otro de respuesta abierta o tipo ensayo*? Para poder dar respuesta a esta interrogante, ha sido necesario abordarla a partir de dos de los enfoques que constituyen el constructo.

Primera pregunta de investigación a la cual se le desea dar respuesta:

¿Los índices de dificultad de los ítems, con base en la teoría de la respuesta al ítem, se ven afectados si estos corresponden a un examen de opción múltiple o si los ítems se presentan en un examen no estructurado?

Prueba de hipótesis en cálculo manual.

Para hacer un juicio con respecto a la existencia de posibles diferencias entre la estadística de la muestra de grupos, *A,* y la muestra del grupo, *B*, sobre la base de las estadísticas de media y varianza, se realizó una prueba de Hipótesis de dos extremos debido a que sólo interesa saber si las medias de cada grupo son iguales o no. El riesgo que se asumió al rechazar la hipótesis nula cuando ésta sea cierta lo fijamos en 5 %; lo cual se considera una probabilidad aceptable, que es también su nivel de significancia.

Para hacer un juicio con respecto a la existencia de posibles diferencias entre la estadística de muestra del grupo, *A,* y la muestra del grupo, *B*, asumiendo que los datos se comportan como

una distribución normal, se aplicó la prueba de hipótesis tomando como estadístico de prueba a, z ,con

$$z = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \qquad (5.1)$$

Los datos de campo recogidos para los grupos, *A,* y *B,* utilizados para la comparación de dificultades de los ítems son:

$n_A = 58$

$n_B = 38$

$\overline{x}_A = 1.30$

$\overline{x}_B = 1.65$

$s_A^2 = 2.54$

$s_B^2 = 1.65$

sustituyendo los valores empíricos obtenidos, la z calculada es.

$$z = 1.20$$

contra la, z, de tablas a un nivel de significancia de 5% es

$$z = 1.96$$

Interpretación 1:

Utilizando un procedimiento de *cálculo manual*, existe evidencia suficiente para aceptar la hipótesis nula, es decir, no existe diferencia en los valores de la dificultad de los ítems calculados con base en la teoría de la respuesta al ítem cuando se modifica el formato de las preguntas.

Prueba de hipótesis en cálculo con ordenador y software ad-hoc.

Al igual que en el caso del cálculo manual se realizó una prueba de hipótesis de dos extremos. Se buscó la existencia de posibles diferencias entre la estadística de muestra de grupos, *A,* y la muestra del grupo, *B*, sobre la base de las estadísticas de media y varianza,. El riesgo que se asumió para rechazar la hipótesis nula cuando ésta sea cierta lo mantuvimos en 5%, asumiendo que los datos se comportan como una distribución normal. Se aplicó la prueba de hipótesis tomando como estadístico de prueba a, z ,con:

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \qquad (5.1)$$

Los datos de campo recogidos para los grupos, A, y, B, utilizados para la comparación de dificultades de los ítems son:

$n_A = 58$

$n_B = 38$

$\bar{x}_A = 1.26$

$\bar{x}_B = 1.79$

$s_A^2 = 2.49$

$s_B^2 = 1.84$

sustituyendo los valores empíricos obtenidos, la *z* calculada es:

$$z = 1.75$$

contra la *z* de tablas a un nivel de significancia de 5% es:

$$z = 1.96$$

Interpretación 2:

Utilizando el ordenador y el programa XCalibre, existe evidencia suficiente para aceptar la hipótesis nula, es decir, no existe diferencia en los valores de la dificultad de los ítems obtenidos a través del modelo de Rasch, cuando se modifica el formato de las preguntas.

Segunda pregunta de investigación a la cual se le desea dar respuesta:

¿Los índices de dificultad de los ítems, con base en la teoría de la respuesta al ítem, se ven afectados cuando el tamaño de la muestra varía?

Como ya se mencionó a los once ítems que conforman el test, se dividieron en dos grupos, asignándoseles un formato diferente a cada uno de ello, uno con formato de opción múltiple y otro de respuesta abierta. Pero como ya se mostró (Conclusiones 1 y 2), el tipo de formato no hace diferencia estadística, por lo tanto, para los efectos siguientes, los grupos, *A,* y, *B,* sólo serán diferentes por el número de alumnos, lo mismo que el grupo, *AB*. Para poder contrastar el efecto del tamaño, establecimos el número correspondiente a un grupo de tamaño normal en alrededor de 40 personas, y uno muy grande de alrededor de 100.

Para hacer un juicio con respecto a la existencia de posibles diferencias entre la estadística del tamaño de muestra de los grupos, *B,* y, *AB,* asumimos que los datos se comportan como una distribución normal, por lo que se aplicó la prueba de hipótesis tomando como estadístico de prueba, z ,con los valores siguientes:

$$n_B = 38$$

$$n_{AB} = 96$$

$$\overline{x}_B = 1.79$$

$$\overline{x}_{AB} = 1.69$$

$$s_B^2 = 1.84$$

$$s_{AB}^2 = 1.53$$

aplicando la fórmula:

$$z = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \qquad (5.1)$$

la z calculada es.

$$z = 0.37$$

contra la z de tablas a un nivel de significancia de 5% es

$$z = 1.96$$

Interpretación 3

Existe evidencia suficiente para aceptar la hipótesis nula, es decir, no existe diferencia en los valores de la dificultad de los ítems obtenidos a través de una muestra de 38 personas y una muestra de 96 personas.

CONCLUSIONES

Los resultados obtenidos en esta investigación han revelado que existe evidencia suficiente para aceptar, con las limitaciones inherentes a la investigación, que no existe diferencia en los valores de la dificultad de los ítems calculados con base en la Teoría de la Respuesta al Ítem, cuando se modifica el formato de las preguntas, de opción múltiple y de respuesta abierta.

Es factible que los estudiantes ya no contesten adivinando, sino a través de la aplicación de todo un procedimiento, lo cual podría explicar una parte del resultado.

Además, existe evidencia suficiente para aceptar, con las limitaciones inherentes a la investigación, que no existe diferencia en los valores de la dificultad de los ítems obtenidos a través del modelo de Rasch, cuando se modifica el tamaño de la muestra, de una menor de 50 a otra muestra de al menos su doble, siendo ésta alrededor de 100 personas. Esto parece avalar el hecho de que no es absolutamente necesario contar con poblaciones extremadamente grandes, tal como lo señala parte de la literatura relativa al modelo de Rasch.

La generalización de los resultados está limitada desde su diseño por corresponder a un estudio de caso, por el tamaño de de la muestra, que la validez de contenido y fiabilidad de los reactivos del test se sustentaron en una base empírica proporcionada por un grupo de expertos.

Si bien la ciencia busca encontrar el orden subyacente en acontecimientos estudiados en forma particular por medio de la formulación y comprobación de hipótesis de naturaleza más general, podemos concluir que los resultados obtenidos son una aportación a nivel micro sobre aspectos esenciales de la evaluación educativa, en particular, de la evaluación sumativa del aprendizaje. Sería demasiado ambicioso generalizar a priori los resultados para cualquier tipo de contenidos académicos, ni siquiera aun particularizando para contenidos meramente matemáticos como los evaluados en la investigación. Desde luego para lograr una generalización sin ningún pero científico, es necesario realizar aún más estudios que, reproduciendo las mismas condiciones con las cuales de realizó la presente investigación, se corroboren o modifiquen las conclusiones presentadas.

ANEXO

Datos obtenidos en el trabajo de campo

Calibración de ítems en forma manual

Tabla 1. Datos sin editar del grupo "A" de la muestra original.

	Examen A	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Ítem 7	Ítem 8	Ítem 9	Ítem 10	Ítem 11	Sumas
1	A. R.												0
2	Á. R.		1				1						2
3	A. S.						1						1
4	A. V.			1					1			1	3
5	B. .M.		1	1		1							3
6	C. B.			1									1
7	C. M.	1		1									2
8	C. B.							1					1
9	E. F.		1				1	1				1	4
10	F. C.		1										1
11	F. M.		1	1			1						3
12	G. C.		1	1		1	1	1	1	1			7
13	G. Á.												0
14	G. G.	1	1	1						1		1	5
15	G. A.	1	1	1			1	1	1	1		1	8
16	H. B.		1	1				1					3
17	H. G.		1				1		1			1	4
18	H. L.		1										1
19	H. L.		1				1						2
20	H. L.		1	1		1	1	1				1	6
21	H. S.		1	1			1						3
22	H. V.						1						1
23	I. M.												0
24	J. M.		1	1			1						3
25	L. P.		1				1	1					3
26	M. H.		1				1						2
27	M. T.		1	1			1				1	1	5
28	M. M.		1	1									2
29	M. S.		1				1						2
30	M. V.		1	1			1						3
31	M. C.			1									1
32	M. Y.						1						1
33	M. H.												0
34	M. Q.	1	1										2

35	N. F.		1	1			1						3
36	O. O.	1	1										2
37	P. O.		1	1			1		1				4
38	Q. O.		1					1					2
39	Q. O.		1						1				2
40	R. A.		1	1			1	1					4
41	R A.			1			1		1				3
42	R. H.		1										1
43	R. R.		1										1
44	R. R.		1	1			1	1					4
45	R. E.	1	1			1	1	1				1	6
46	R. G.		1				1					1	3
47	R. R.		1				1						2
48	R. O.		1										1
49	S. A.	1	1	1			1		1			1	6
50	S: N.												0
51	S: C.					1	1	1				1	4
52	S. F.		1				1						2
53	S. G.		1	1			1						3
54	S. P.		1	1			1	1	1				5
55	T. M.		1	1				1	1				4
56	T. M.											1	1
57	T. A.	1											1
58	T. C.		1				1					1	3
59	T. A.		1	1			1	1					4
60	V. L.	1											1
61	V. M.		1	1			1	1					4
62	Z. M.		1	1			1						3
63	Z. Á.		1				1	1					3

Tabla 2. Datos sin editar del grupo "B" de la muestra original.

	Examen B	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Ítem 7	Ítem 8	Ítem 9	Ítem 10	Ítem 11	Sumas
64	A. T.		1						1				2
65	A. F.		1	1					1				3
66	A. P.	1		1		1		1				1	5
67	A. L.					1		1	1				3
68	A. M.			1				1	1				3
69	B. B.	1							1		1		3
70	B. R.												0
71	B. C.	1							1			1	3
72	B. B.		1		1						1	1	4
73	C. M.	1											1
74	C. A.												0

75	C. S.		1					1	1				3
76	C. R.								1				1
77	D. F.	1							1				2
78	D. B.	1							1				2
79	E. A.								1				1
80	G. S.								1				1
81	G. A.						1		1				2
82	G. C.	1						1					2
83	G. G.												0
84	G. P.												0
85	G. M.						1						1
86	G. R.								1				1
87	H. H.												0
88	H. P.	1							1				2
89	H. U.												0
90	H. Z.												0
91	H. C.								1				1
92	H. R.												0
93	J. N.	1						1				1	3
94	L. L.			1					1				2
95	L. B.	1							1		1	1	4
96	L. C.												0
97	L. M.	1	1						1			1	4
98	M. O.	1							1				2
99	M. R.							1					1
100	M. S.							1				1	2
101	M. M.												0
102	M. H.								1				1
103	O. V.							1	1		1	1	4
104	P. R.												0
105	R. P.					1			1				2
106	S. O.								1				1
107	V. F.	1											1
108	V. H.								1				1
109	Z. P.							1	1				2
110	Z. A.	1											1

Tabla 3. Datos sin editar del grupo "AB" de la muestra original. Claves con exámenes empatados.

	Nombre del alumno	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6	Ítem 7	Ítem 8	Ítem 9	Ítem 10	Ítem 11	sumas
1	A. R.	0	0	0	0	0	0	0	0	0	0	0	0
2	Á. R.	0	1	0	0	0	1	0	0	0	0	0	2
64	A. T.	0	1	0	0	0	0	0	1	0	0	0	2
65	A. F.	0	1	0	0	0	0	0	1	0	1	0	3
3	A. S.	0	0	0	0	0	1	0	0	0	0	0	1
4	A. V.	0	0	1	0	0	0	0	1	0	0	1	3
77	A. P.	0	0	1	0	0	1	1	0	0	1	1	5
67	A. L.	0	1	1	0	0	0	1	0	0	0	0	3
68	A. M.	0	1	1	0	0	0	0	0	0	1	0	3
69	B. B.	0	1	0	0	1	1	0	0	0	0	0	3
70	B. R.	0	0	0	0	0	0	0	0	0	0	0	0
5	B. M.	0	1	1	0	1	0	0	0	0	0	0	3
71	B. C.	0	1	0	0	0	1	0	0	0	0	1	3
72	B. B.	0	0	0	0	1	0	0	1	1	0	1	4
73	C. M.	0	0	0	0	0	1	0	0	0	0	0	1
74	C. A.	0	0	0	0	0	0	0	0	0	0	0	0
75	C. S.	0	1	1	0	0	0	0	1	0	0	0	3
6	C. B.	0	0	1	0	0	0	0	0	0	0	0	1
76	C. R.	0	1	0	0	0	0	0	0	0	0	0	1
7	C. M.	1	0	1	0	0	0	0	0	0	0	0	2
8	C. B.	0	0	0	0	0	0	1	0	0	0	0	1
77	D. F.	0	1	0	0	0	1	0	0	0	0	0	2
78	D. B.	0	1	0	0	0	1	0	0	0	0	0	2
9	E. F.	0	1	0	0	0	1	1	0	0	0	1	4
79	E. A.	0	1	0	0	0	0	0	0	0	0	0	1
10	F. C.	0	1	0	0	0	0	0	0	0	0	0	1
11	F. M.	0	1	1	0	0	1	0	0	0	0	0	3
80	G. S.	0	1	0	0	0	0	0	0	0	0	0	1
81	G. A.	0	1	0	1	0	0	0	0	0	0	0	2
12	G. C.	0	1	1	0	1	1	1	1	1	0	0	7
82	G. C.	0	0	1	0	0	1	0	0	0	0	0	2
83	G. G.	0	0	0	0	0	0	0	0	0	0	0	0
13	G. Á.	0	0	0	0	0	0	0	0	0	0	0	0
14	G. G.	1	1	1	0	0	0	0	0	1	0	1	5
84	G. P.	0	0	0	0	0	0	0	0	0	0	0	0
15	G. R.	1	1	1	0	0	1	1	1	1	0	1	8
85	G. M.	0	0	0	1	0	0	0	0	0	0	0	1
86	G. R.	0	1	0	0	0	0	0	0	0	0	0	1
16	H. B.	0	1	1	0	0	0	1	0	0	0	0	3
17	H. G.	0	1	0	0	0	1	0	1	0	0	1	4

87	H. H.	0	0	0	0	0	0	0	0	0	0	0	0
18	H. L.	0	1	0	0	0	0	0	0	0	0	0	1
19	H. L.	0	1	0	0	0	1	0	0	0	0	0	2
20	H. L.	0	1	1	0	1	1	1	0	0	0	1	6
88	H. P..	0	1	0	0	0	1	0	0	0	0	0	2
21	H. S.	0	1	1	0	0	1	0	0	0	0	0	3
89	H. U.	0	0	0	0	0	0	0	0	0	0	0	0
90	H. Z.	0	0	0	0	0	0	0	0	0	0	0	0
91	H. C.	0	1	0	0	0	0	0	0	0	0	0	1
92	H. R.	0	0	0	0	0	0	0	0	0	0	0	0
22	H. V.	0	0	0	0	0	1	0	0	0	0	0	1
23	I. M.	0	0	0	0	0	0	0	0	0	0	0	0
24	J. M.	0	1	1	0	0	1	0	0	0	0	0	3
93	J. N.	0	0	1	0	0	1	0	0	0	0	1	3
94	L. L.	0	1	0	0	0	0	0	0	0	1	0	2
95	L. B.	0	1	0	0	1	1	0	0	0	0	1	4
96	L. C.	0	0	0	0	0	0	0	0	0	0	0	0
97	L. M.	0	1	0	0	0	1	0	1	0	0	1	4
25	L. P.	0	1	0	0	0	1	1	0	0	0	0	3
26	M. H.	0	1	0	0	0	1	0	0	0	0	0	2
27	M. T.	0	1	1	0	0	1	0	0	0	1	1	5
28	M. M.	0	1	1	0	0	0	0	0	0	0	0	2
29	M. S.	0	1	0	0	0	1	0	0	0	0	0	2
30	M. V.	0	1	1	0	0	1	0	0	0	0	0	3
98	M. O.	0	1	0	0	0	1	0	0	0	0	0	2
31	M. C.	0	0	1	0	0	0	0	0	0	0	0	1
99	M. R.	0	0	1	0	0	0	0	0	0	0	0	1
32	M. Y.	0	0	0	0	0	1	0	0	0	0	0	1
33	M. H.	0	0	0	0	0	0	0	0	0	0	0	0
100	M. S.	0	0	1	0	0	0	0	0	0	0	1	2
101	M. M.	0	0	0	0	0	0	0	0	0	0	0	0
34	M. Q.	1	1	0	0	0	0	0	0	0	0	0	2
102	M. H.	0	1	0	0	0	0	0	0	0	0	0	1
35	N. F.	0	1	1	0	0	1	0	0	0	0	0	3
103	O. V.	0	1	1	0	1	0	0	0	0	0	1	4
36	O. O.	1	1	0	0	0	0	0	0	0	0	0	2
104	P. R.	0	0	0	0	0	0	0	0	0	0	0	0
37	P. O.	0	1	1	0	0	1	0	1	0	0	0	4
38	Q. O.	0	1	0	0	0	0	1	0	0	0	0	2
39	Q. O.	0	1	0	0	0	0	0	1	0	0	0	2
40	R. A.	0	1	1	0	0	1	1	0	0	0	0	4
41	R. A.	0	0	1	0	0	1	0	1	0	0	0	3
42	R. H.	0	1	0	0	0	0	0	0	0	0	0	1
43	R. R.	0	1	0	0	0	0	0	0	0	0	0	1
44	R. R.	0	1	1	0	0	1	1	0	0	0	0	4
45	R. E.	1	1	0	0	1	1	1	0	0	0	1	6
105	R. P.	0	1	0	0	0	0	1	0	0	0	0	2

46	R. G.	0	1	0	0	0	1	0	0	0	0	1	3
47	R. R.	0	1	0	0	0	1	0	0	0	0	0	2
48	R. O.	0	1	0	0	0	0	0	0	0	0	0	1
49	S. A.	1	1	1	0	0	1	0	1	0	0	1	6
50	S. N.	0	0	0	0	0	0	0	0	0	0	0	0
51	S. C.	0	0	0	0	1	1	1	0	0	0	1	4
52	S. F.	0	1	0	0	0	1	0	0	0	0	0	2
53	S. G.	0	1	1	0	0	1	0	0	0	0	0	3
54	S. P.	0	1	1	0	0	1	1	1	0	0	0	5
106	S. O.	0	1	0	0	0	0	0	0	0	0	0	1
55	T. M.	0	1	1	0	0	0	1	1	0	0	0	4
56	T. M.	0	0	0	0	0	0	0	0	0	0	1	1
57	T. A.	1	0	0	0	0	0	0	0	0	0	0	1
58	T. C.	0	1	0	0	0	1	0	0	0	0	1	3
59	T. A.	0	1	1	0	0	1	1	0	0	0	0	4
60	V. L.	1	0	0	0	0	0	0	0	0	0	0	1
107	V. F.	0	0	0	0	0	1	0	0	0	0	0	1
108	V. H.	0	1	0	0	0	0	0	0	0	0	0	1
61	V. M.	0	1	1	0	0	1	1	0	0	0	0	4
62	Z. M.	0	1	1	0	0	1	0	0	0	0	0	3
109	Z. P.	0	1	1	0	0	0	0	0	0	0	0	2
110	Z. A.	0	0	0	0	0	1	0	0	0	0	0	1
63	Z. Á.	0	1	0	0	0	1	1	0	0	0	0	3

Nota: En algunas tablas, sólo se utilizarán números para identificar a los alumnos; dicha numeración se corresponde uno a uno con cada uno de los alumnos de esta *tabla 3*.

Tabla 4. Respuestas ordenadas sin editar del grupo "A".

	Examen A	Item 2	Item 6	Item 3	Item 7	Item 11	Item 8	Item 1	Item 5	Item 9	Item 10	Item 4	suma
1	A. R.	0	0	0	0	0	0	0	0	0	0	0	0
13	G. A.	0	0	0	0	0	0	0	0	0	0	0	0
23	I. M.	0	0	0	0	0	0	0	0	0	0	0	0
33	M. H.	0	0	0	0	0	0	0	0	0	0	0	0
50	S. N.	0	0	0	0	0	0	0	0	0	0	0	0
3	A. S.	0	1	0	0	0	0	0	0	0	0	0	1
6	C. B.	0	0	1	0	0	0	0	0	0	0	0	1
8	C. B.	0	0	0	1	0	0	0	0	0	0	0	1
10	F. C.	1	0	0	0	0	0	0	0	0	0	0	1
18	H. L.	1	0	0	0	0	0	0	0	0	0	0	1
22	H. V.	0	1	0	0	0	0	0	0	0	0	0	1
31	M. C.	0	0	1	0	0	0	0	0	0	0	0	1
32	M. Y.	0	1	0	0	0	0	0	0	0	0	0	1
42	R. H.	1	0	0	0	0	0	0	0	0	0	0	1
43	R. R.	1	0	0	0	0	0	0	0	0	0	0	1
48	R. O.	1	0	0	0	0	0	0	0	0	0	0	1

56	T. M.	0	0	0	0	1	0	0	0	0	0	0	1
57	T. A.	0	0	0	0	0	0	1	0	0	0	0	1
60	V. L.	0	0	0	0	0	0	1	0	0	0	0	1
2	A. R.	1	1	0	0	0	0	0	0	0	0	0	2
7	C. M.	0	0	1	0	0	0	1	0	0	0	0	2
19	H. L.	1	1	0	0	0	0	0	0	0	0	0	2
26	M. H.	1	1	0	0	0	0	0	0	0	0	0	2
28	M. M.	1	0	1	0	0	0	0	0	0	0	0	2
29	M. S.	1	1	0	0	0	0	0	0	0	0	0	2
34	M. Q.	1	0	0	0	0	0	1	0	0	0	0	2
36	O: O.	1	0	0	0	0	0	1	0	0	0	0	2
38	Q. O.	1	0	0	1	0	0	0	0	0	0	0	2
39	Q. O.	1	0	0	0	0	1	0	0	0	0	0	2
47	R. R.	1	1	0	0	0	0	0	0	0	0	0	2
52	S. F.	1	1	0	0	0	0	0	0	0	0	0	2
4	A. V.	0	0	1	0	1	1	0	0	0	0	0	3
5	B. M.	1	0	1	0	0	0	0	1	0	0	0	3
11	F. M.	1	1	1	0	0	0	0	0	0	0	0	3
16	H. B.	1	0	1	1	0	0	0	0	0	0	0	3
21	H. S.	1	1	1	0	0	0	0	0	0	0	0	3
24	J. M.	1	1	1	0	0	0	0	0	0	0	0	3
25	L. P.	1	1	0	1	0	0	0	0	0	0	0	3
30	M. V.	1	1	1	0	0	0	0	0	0	0	0	3
35	N. F.	1	1	1	0	0	0	0	0	0	0	0	3
41	R. A.	0	1	1	0	0	1	0	0	0	0	0	3
46	R. G.	1	1	0	0	1	0	0	0	0	0	0	3
53	S. G.	1	1	1	0	0	0	0	0	0	0	0	3
58	T. C.	1	1	0	0	1	0	0	0	0	0	0	3
62	Z. M.	1	1	1	0	0	0	0	0	0	0	0	3
63	Z. A.	1	1	0	1	0	0	0	0	0	0	0	3
9	E. F.	1	1	0	1	1	0	0	0	0	0	0	4
17	H. G.	1	1	0	0	1	1	0	0	0	0	0	4
37	P. O.	1	1	1	0	0	1	0	0	0	0	0	4
40	R. A.	1	1	1	1	0	0	0	0	0	0	0	4
44	R. R.	1	1	1	1	0	0	0	0	0	0	0	4
51	S. C.	0	1	0	1	1	0	0	1	0	0	0	4
55	T. M.	1	0	1	1	0	1	0	0	0	0	0	4
59	T. A.	1	1	1	1	0	0	0	0	0	0	0	4
61	V. M.	1	1	1	1	0	0	0	0	0	0	0	4
14	G. G.	1	0	1	0	1	0	1	0	1	0	0	5
27	M. T.	1	1	1	0	1	0	0	0	0	1	0	5
54	S. P.	1	1	1	1	0	1	0	0	0	0	0	5
20	H. L.	1	1	1	1	1	0	0	1	0	0	0	6
45	R. E.	1	1	0	1	1	0	1	1	0	0	0	6
49	S.A.	1	1	1	0	1	1	1	0	0	0	0	6
12	G. C.	1	1	1	1	0	1	0	1	1	0	0	7
15	G. R.	1	1	1	1	1	1	1	0	1	0	0	8
	Sumas	45	36	28	17	13	10	9	5	3	1	0	

Tabla 5. Respuestas ordenadas sin editar del grupo "B".

	Examen B	Ítem 8	Ítem 1	Ítem 7	Ítem 11	Ítem 2	Ítem 3	Ítem 10	Ítem 5	Ítem 6	Ítem 4	Ítem 9	Sumas
70	B. R.	0	0	0	0	0	0	0	0	0	0	0	0
74	C. A.	0	0	0	0	0	0	0	0	0	0	0	0
83	G. G.	0	0	0	0	0	0	0	0	0	0	0	0
84	G. P.	0	0	0	0	0	0	0	0	0	0	0	0
87	H. H.	0	0	0	0	0	0	0	0	0	0	0	0
89	H. U.	0	0	0	0	0	0	0	0	0	0	0	0
90	H. Z.	0	0	0	0	0	0	0	0	0	0	0	0
92	H. R.	0	0	0	0	0	0	0	0	0	0	0	0
96	L. C.	0	0	0	0	0	0	0	0	0	0	0	0
101	M. M.	0	0	0	0	0	0	0	0	0	0	0	0
104	P. R.	0	0	0	0	0	0	0	0	0	0	0	0
73	C. M.	0	1	0	0	0	0	0	0	0	0	0	1
76	C. R.	1	0	0	0	0	0	0	0	0	0	0	1
79	E. A.	1	0	0	0	0	0	0	0	0	0	0	1
80	G: S.	1	0	0	0	0	0	0	0	0	0	0	1
85	G. M.	0	0	0	0	0	0	0	0	1	0	0	1
86	G. R.	1	0	0	0	0	0	0	0	0	0	0	1
91	H. C.	1	0	0	0	0	0	0	0	0	0	0	1
99	M. R.	0	0	1	0	0	0	0	0	0	0	0	1
102	M. H.	1	0	0	0	0	0	0	0	0	0	0	1
106	S. O.	1	0	0	0	0	0	0	0	0	0	0	1
107	V: F.	0	1	0	0	0	0	0	0	0	0	0	1
108	V. H.	1	0	0	0	0	0	0	0	0	0	0	1
110	Z. A.	0	1	0	0	0	0	0	0	0	0	0	1
64	A. T.	1	0	0	0	1	0	0	0	0	0	0	2
77	D. F.	1	1	0	0	0	0	0	0	0	0	0	2
78	D. B.	1	1	0	0	0	0	0	0	0	0	0	2
81	G. A.	1	0	0	0	0	0	0	0	1	0	0	2
82	G. C.	0	1	1	0	0	0	0	0	0	0	0	2
88	H. P.	1	1	0	0	0	0	0	0	0	0	0	2
94	L. L.	1	0	0	0	0	1	0	0	0	0	0	2
98	M. O.	1	1	0	0	0	0	0	0	0	0	0	2
100	M. S.	0	0	1	1	0	0	0	0	0	0	0	2
105	R. P.	1	0	0	0	0	0	0	1	0	0	0	2
109	Z. P.	1	0	1	0	0	0	0	0	0	0	0	2
65	A. F.	1	0	0	0	1	1	0	0	0	0	0	3
67	A. L.	1	0	1	0	0	0	0	1	0	0	0	3
68	A. M.	1	0	1	0	0	1	0	0	0	0	0	3
69	B. B.	1	1	0	0	0	0	1	0	0	0	0	3
71	B. C.	1	1	0	1	0	0	0	0	0	0	0	3
75	C. S.	1	0	1	0	1	0	0	0	0	0	0	3
93	J. N.	0	1	1	1	0	0	0	0	0	0	0	3
72	B. B.	0	0	0	1	1	0	1	0	0	1	0	4

95	L. B.	1	1	0	1	0	0	1	0	0	0	0	4
97	L. M.	1	1	0	1	1	0	0	0	0	0	0	4
103	O. V.	1	0	1	1	0	0	1	0	0	0	0	4
66	A. P.	0	1	1	1	0	1	0	1	0	0	0	5
	Sumas	26	14	10	8	5	4	4	3	2	1	0	

Tabla 6. Respuestas ordenadas sin editar del grupo "AB".

	Examen AB	Item 2	Item 6	Item 3	Item 11	Item 7	Item 8	Item 1	Item 5	Item 10	Item 9	Item 4	sumas
1	A. R.	0	0	0	0	0	0	0	0	0	0	0	0
70	B. R.	0	0	0	0	0	0	0	0	0	0	0	0
74	C. A.	0	0	0	0	0	0	0	0	0	0	0	0
83	G. G.	0	0	0	0	0	0	0	0	0	0	0	0
13	G. A.	0	0	0	0	0	0	0	0	0	0	0	0
84	G. P.	0	0	0	0	0	0	0	0	0	0	0	0
87	H. H.	0	0	0	0	0	0	0	0	0	0	0	0
89	H. U.	0	0	0	0	0	0	0	0	0	0	0	0
90	H. Z.	0	0	0	0	0	0	0	0	0	0	0	0
92	H. R.	0	0	0	0	0	0	0	0	0	0	0	0
23	I. M.	0	0	0	0	0	0	0	0	0	0	0	0
96	L. C.	0	0	0	0	0	0	0	0	0	0	0	0
33	M. H.	0	0	0	0	0	0	0	0	0	0	0	0
101	M. M.	0	0	0	0	0	0	0	0	0	0	0	0
104	P. R.	0	0	0	0	0	0	0	0	0	0	0	0
50	S. N.	0	0	0	0	0	0	0	0	0	0	0	0
3	A. S.	0	1	0	0	0	0	0	0	0	0	0	1
73	C. M.	0	1	0	0	0	0	0	0	0	0	0	1
6	C. B.	0	0	1	0	0	0	0	0	0	0	0	1
76	C. R.	1	0	0	0	0	0	0	0	0	0	0	1
8	C. B.	0	0	0	0	1	0	0	0	0	0	0	1
79	E. A.	1	0	0	0	0	0	0	0	0	0	0	1
10	F. C.	1	0	0	0	0	0	0	0	0	0	0	1
80	G. S.	1	0	0	0	0	0	0	0	0	0	0	1
85	G. M.	0	0	0	0	0	0	0	0	0	0	1	1
86	G. R.	1	0	0	0	0	0	0	0	0	0	0	1
18	H. L.	1	0	0	0	0	0	0	0	0	0	0	1
91	H. C.	1	0	0	0	0	0	0	0	0	0	0	1
22	H. V.	0	1	0	0	0	0	0	0	0	0	0	1
31	M. C.	0	0	1	0	0	0	0	0	0	0	0	1
99	M. R.	0	0	1	0	0	0	0	0	0	0	0	1
32	M. Y.	0	1	0	0	0	0	0	0	0	0	0	1
102	M. H.	1	0	0	0	0	0	0	0	0	0	0	1
42	R. H.	1	0	0	0	0	0	0	0	0	0	0	1
43	R. R.	1	0	0	0	0	0	0	0	0	0	0	1
48	R. O.	1	0	0	0	0	0	0	0	0	0	0	1
106	S. O.	1	0	0	0	0	0	0	0	0	0	0	1
56	T. M.	0	0	0	1	0	0	0	0	0	0	0	1
57	T. A.	0	0	0	0	0	0	1	0	0	0	0	1

60	V. L.	0	0	0	0	0	0	1	0	0	0	0	1
107	V. F.	0	1	0	0	0	0	0	0	0	0	0	1
108	V. H.	1	0	0	0	0	0	0	0	0	0	0	1
110	Z. A.	0	1	0	0	0	0	0	0	0	0	0	1
2	Á. R.	1	1	0	0	0	0	0	0	0	0	0	2
64	A. T.	1	0	0	0	0	1	0	0	0	0	0	2
7	C. M.	0	0	1	0	0	0	1	0	0	0	0	2
77	D. F.	1	1	0	0	0	0	0	0	0	0	0	2
78	D. B.	1	1	0	0	0	0	0	0	0	0	0	2
81	G. A.	1	0	0	0	0	0	0	0	0	0	1	2
82	G. C.	0	1	1	0	0	0	0	0	0	0	0	2
19	H. L.	1	1	0	0	0	0	0	0	0	0	0	2
88	H. P.	1	1	0	0	0	0	0	0	0	0	0	2
94	L. L.	1	0	0	0	0	0	0	0	1	0	0	2
26	M. H.	1	1	0	0	0	0	0	0	0	0	0	2
28	M. M.	1	0	1	0	0	0	0	0	0	0	0	2
29	M. S.	1	1	0	0	0	0	0	0	0	0	0	2
98	M. O.	1	1	0	0	0	0	0	0	0	0	0	2
100	M. S.	0	0	1	1	0	0	0	0	0	0	0	2
34	M. Q.	1	0	0	0	0	0	1	0	0	0	0	2
36	O. O.	1	0	0	0	0	0	1	0	0	0	0	2
38	Q. O.	1	0	0	0	1	0	0	0	0	0	0	2
39	Q. O.	1	0	0	0	0	1	0	0	0	0	0	2
105	R. P.	1	0	0	0	1	0	0	0	0	0	0	2
47	R. R.	1	1	0	0	0	0	0	0	0	0	0	2
52	S. F.	1	1	0	0	0	0	0	0	0	0	0	2
109	Z. P.	1	0	1	0	0	0	0	0	0	0	0	2
65	A. F.	1	0	0	0	0	1	0	0	1	0	0	3
4	A. V.	0	0	1	1	0	1	0	0	0	0	0	3
67	A. L.	1	0	1	0	1	0	0	0	0	0	0	3
68	A. M.	1	0	1	0	0	0	0	0	1	0	0	3
69	B. B.	1	1	0	0	0	0	0	1	0	0	0	3
5	B. M.	1	0	1	0	0	0	0	1	0	0	0	3
71	B. C.	1	1	0	1	0	0	0	0	0	0	0	3
75	C. S.	1	0	1	0	0	1	0	0	0	0	0	3
11	F. M.	1	1	1	0	0	0	0	0	0	0	0	3
16	H. B.	1	0	1	0	1	0	0	0	0	0	0	3
21	H. S.	1	1	1	0	0	0	0	0	0	0	0	3
24	J. M.	1	1	1	0	0	0	0	0	0	0	0	3
93	J. N.	0	1	1	1	0	0	0	0	0	0	0	3
25	L. P.	1	1	0	0	1	0	0	0	0	0	0	3
30	M. V.	1	1	1	0	0	0	0	0	0	0	0	3
35	N. F.	1	1	1	0	0	0	0	0	0	0	0	3
41	R. A.	0	1	1	0	0	1	0	0	0	0	0	3
46	R. g.	1	1	0	1	0	0	0	0	0	0	0	3
53	S. G.	1	1	1	0	0	0	0	0	0	0	0	3
58	T. C.	1	1	0	1	0	0	0	0	0	0	0	3
62	Z. M.	1	1	1	0	0	0	0	0	0	0	0	3
63	Z. Á.	1	1	0	0	1	0	0	0	0	0	0	3
72	B. B.	0	0	0	1	0	1	0	1	0	1	0	4
9	E. F.	1	1	0	1	1	0	0	0	0	0	0	4

17	H. G.	1	1	0	1	0	1	0	0	0	0	0	4
95	L. B.	1	1	0	1	0	0	0	1	0	0	0	4
97	L. M.	1	1	0	1	0	1	0	0	0	0	0	4
103	O. V.	1	0	1	1	0	0	0	1	0	0	0	4
37	P. O.	1	1	1	0	0	1	0	0	0	0	0	4
40	R. A.	1	1	1	0	1	0	0	0	0	0	0	4
44	R. R.	1	1	1	0	1	0	0	0	0	0	0	4
51	S. C.	0	1	0	1	1	0	0	1	0	0	0	4
55	T. M.	1	0	1	0	1	1	0	0	0	0	0	4
59	T. A.	1	1	1	0	1	0	0	0	0	0	0	4
61	V. M.	1	1	1	0	1	0	0	0	0	0	0	4
77	A. P.	0	1	1	1	1	0	0	0	1	0	0	5
14	G. G.	1	0	1	1	0	0	1	0	0	1	0	5
27	M. T.	1	1	1	1	0	0	0	0	1	0	0	5
54	S. P.	1	1	1	0	1	1	0	0	0	0	0	5
20	H. L.	1	1	1	1	1	0	0	1	0	0	0	6
45	R E.	1	1	0	1	1	0	1	1	0	0	0	6
49	S. A.	1	1	1	1	0	1	1	0	0	0	0	6
12	G. C.	1	1	1	0	1	1	0	1	0	1	0	7
15	G. R.	1	1	1	1	1	1	1	0	0	1	0	8
	Sumas	71	50	38	21	20	15	9	9	5	4	2	

Tabla 7. Respuestas editadas y ordenadas del grupo "A".

	Examen A	Item 2	Item 6	Item 3	Item 7	Item 11	Item 8	Item 1	Item 5	Item 9	Item 10	suma
3	A. S.	0	1	0	0	0	0	0	0	0	0	1
6	C. B.	0	0	1	0	0	0	0	0	0	0	1
8	C. B.	0	0	0	1	0	0	0	0	0	0	1
10	F. C.	1	0	0	0	0	0	0	0	0	0	1
18	H. L.	1	0	0	0	0	0	0	0	0	0	1
22	H. V.	0	1	0	0	0	0	0	0	0	0	1
31	M. C.	0	0	1	0	0	0	0	0	0	0	1
32	M. Y.	0	1	0	0	0	0	0	0	0	0	1
42	R. H.	1	0	0	0	0	0	0	0	0	0	1
43	R. R.	1	0	0	0	0	0	0	0	0	0	1
48	R. O.	1	0	0	0	0	0	0	0	0	0	1
56	T. M.	0	0	0	0	1	0	0	0	0	0	1
57	T. A.	0	0	0	0	0	0	1	0	0	0	1
60	V. L.	0	0	0	0	0	0	1	0	0	0	1
2	A. R.	1	1	0	0	0	0	0	0	0	0	2
7	C. M.	0	0	1	0	0	0	1	0	0	0	2
19	H. L.	1	1	0	0	0	0	0	0	0	0	2
26	M. H.	1	1	0	0	0	0	0	0	0	0	2
28	M. M.	1	0	1	0	0	0	0	0	0	0	2
29	M. S.	1	1	0	0	0	0	0	0	0	0	2
34	M. Q.	1	0	0	0	0	0	1	0	0	0	2
36	O. O.	1	0	0	0	0	0	1	0	0	0	2
38	Q. O.	1	0	0	1	0	0	0	0	0	0	2
39	Q. O.	1	0	0	0	0	1	0	0	0	0	2
47	R. R.	1	1	0	0	0	0	0	0	0	0	2

52	S. F.	1	1	0	0	0	0	0	0	0	0	2
4	A. V.	0	0	1	0	1	1	0	0	0	0	3
5	B. M.	1	0	1	0	0	0	0	1	0	0	3
11	F. M.	1	1	1	0	0	0	0	0	0	0	3
16	H. B.	1	0	1	1	0	0	0	0	0	0	3
21	H. S.	1	1	1	0	0	0	0	0	0	0	3
24	J. M.	1	1	1	0	0	0	0	0	0	0	3
25	L. P.	1	1	0	1	0	0	0	0	0	0	3
30	M. V.	1	1	1	0	0	0	0	0	0	0	3
35	N. F.	1	1	1	0	0	0	0	0	0	0	3
41	R. A.	0	1	1	0	0	1	0	0	0	0	3
46	R. G.	1	1	0	0	1	0	0	0	0	0	3
53	S. G.	1	1	1	0	0	0	0	0	0	0	3
58	T. C.	1	1	0	0	1	0	0	0	0	0	3
62	Z. M.	1	1	1	0	0	0	0	0	0	0	3
63	Z. A.	1	1	0	1	0	0	0	0	0	0	3
9	E. F.	1	1	0	1	1	0	0	0	0	0	4
17	H. G.	1	1	0	0	1	1	0	0	0	0	4
37	P. O.	1	1	1	0	0	1	0	0	0	0	4
40	R. A.	1	1	1	1	0	0	0	0	0	0	4
44	R. R.	1	1	1	1	0	0	0	0	0	0	4
51	S. C.	0	1	0	1	1	0	0	1	0	0	4
55	T. M.	1	0	1	1	0	1	0	0	0	0	4
59	T. A.	1	1	1	1	0	0	0	0	0	0	4
61	V. M.	1	1	1	1	0	0	0	0	0	0	4
14	G. G.	1	0	1	0	1	0	1	0	1	0	5
27	M. T.	1	1	1	0	1	0	0	0	0	1	5
54	S. P.	1	1	1	1	0	1	0	0	0	0	5
20	H. L.	1	1	1	1	1	0	0	1	0	0	6
45	R. E.	1	1	0	1	1	0	1	1	0	0	6
49	S. A.	1	1	1	0	1	1	1	0	0	0	6
12	G. C.	1	1	1	1	0	1	0	1	1	0	7
15	G. A.	1	1	1	1	1	1	1	0	1	0	8
	Sumas	45	36	28	17	13	10	9	5	3	1	

Tabla 8. Respuestas editadas y ordenadas del grupo "B".

	Examen B	Ítem 8	Ítem 1	Ítem 7	Ítem 11	Ítem 2	Ítem 3	Ítem 10	Ítem 5	Ítem 6	Ítem 4	Sumas
73	C. M.	0	1	0	0	0	0	0	0	0	0	1
76	C. R.	1	0	0	0	0	0	0	0	0	0	1
79	E. A.	1	0	0	0	0	0	0	0	0	0	1
80	G. S.	1	0	0	0	0	0	0	0	0	0	1
85	G. M.	0	0	0	0	0	0	0	0	1	0	1
86	G. R.	1	0	0	0	0	0	0	0	0	0	1
91	H. C.	1	0	0	0	0	0	0	0	0	0	1
99	M. R.	0	0	1	0	0	0	0	0	0	0	1
102	M. H.	1	0	0	0	0	0	0	0	0	0	1
106	S. O.	1	0	0	0	0	0	0	0	0	0	1
107	V. F.	0	1	0	0	0	0	0	0	0	0	1

108	V. H.	1	0	0	0	0	0	0	0	0	0	1
110	Z. A.	0	1	0	0	0	0	0	0	0	0	1
64	A. T.	1	0	0	0	1	0	0	0	0	0	2
77	D. F.	1	1	0	0	0	0	0	0	0	0	2
78	D. B.	1	1	0	0	0	0	0	0	0	0	2
81	G. A.	1	0	0	0	0	0	0	0	1	0	2
82	G. C.	0	1	1	0	0	0	0	0	0	0	2
88	H. P..	1	1	0	0	0	0	0	0	0	0	2
94	L. L.	1	0	0	0	0	1	0	0	0	0	2
98	M. O.	1	1	0	0	0	0	0	0	0	0	2
100	M. S.	0	0	1	1	0	0	0	0	0	0	2
105	R. P.	1	0	0	0	0	0	0	1	0	0	2
109	Z. P.	1	0	1	0	0	0	0	0	0	0	2
65	A. F.	1	0	0	0	1	1	0	0	0	0	3
67	A. L.	1	0	1	0	0	0	0	1	0	0	3
68	A. M.	1	0	1	0	0	1	0	0	0	0	3
69	B. B.	1	1	0	0	0	0	1	0	0	0	3
71	B. C.	1	1	0	1	0	0	0	0	0	0	3
75	C. S.	1	0	1	0	1	0	0	0	0	0	3
93	J. N.	0	1	1	1	0	0	0	0	0	0	3
72	B. B.	0	0	0	1	1	0	1	0	0	1	4
95	L. B.	1	1	0	1	0	0	1	0	0	0	4
97	L. M.	1	1	0	1	1	0	0	0	0	0	4
103	O. V.	1	0	1	1	0	0	1	0	0	0	4
66	A. P.	0	1	1	1	0	1	0	1	0	0	5
		26	14	10	8	5	4	4	3	2	1	

Tabla 9. Respuestas editadas y ordenadas del Grupo "AB".

	Examen AB	Item 2	Item 6	Item 3	Item 11	Item 7	Item 8	Item 1	Item 5	Item 10	Item 9	Item 4	sumas
3	A. S.	0	1	0	0	0	0	0	0	0	0	0	1
73	C. M.	0	1	0	0	0	0	0	0	0	0	0	1
6	C. B.	0	0	1	0	0	0	0	0	0	0	0	1
76	C. R.	1	0	0	0	0	0	0	0	0	0	0	1
8	C. B.	0	0	0	0	1	0	0	0	0	0	0	1
79	E. A.	1	0	0	0	0	0	0	0	0	0	0	1
10	F. C.	1	0	0	0	0	0	0	0	0	0	0	1
80	G. S.	1	0	0	0	0	0	0	0	0	0	0	1
85	G. M.	0	0	0	0	0	0	0	0	0	0	1	1
86	G. R.	1	0	0	0	0	0	0	0	0	0	0	1
18	H. L.	1	0	0	0	0	0	0	0	0	0	0	1
91	H. C.	1	0	0	0	0	0	0	0	0	0	0	1
22	H. V.	0	1	0	0	0	0	0	0	0	0	0	1
31	M. C.	0	0	1	0	0	0	0	0	0	0	0	1
99	M. R.	0	0	1	0	0	0	0	0	0	0	0	1
32	M. Y.	0	1	0	0	0	0	0	0	0	0	0	1
102	M. H.	1	0	0	0	0	0	0	0	0	0	0	1

42	R. H.	1	0	0	0	0	0	0	0	0	0	0	1
43	R. R.	1	0	0	0	0	0	0	0	0	0	0	1
48	R. O.	1	0	0	0	0	0	0	0	0	0	0	1
106	S. O.	1	0	0	0	0	0	0	0	0	0	0	1
56	T. M.	0	0	0	1	0	0	0	0	0	0	0	1
57	T. A.	0	0	0	0	0	0	1	0	0	0	0	1
60	V. L.	0	0	0	0	0	0	1	0	0	0	0	1
107	V. F.	0	1	0	0	0	0	0	0	0	0	0	1
108	V. H.	1	0	0	0	0	0	0	0	0	0	0	1
110	Z. A.	0	1	0	0	0	0	0	0	0	0	0	1
2	Á. R.	1	1	0	0	0	0	0	0	0	0	0	2
64	A. T.	1	0	0	0	0	1	0	0	0	0	0	2
7	C. M.	0	0	1	0	0	0	1	0	0	0	0	2
77	D. F.	1	1	0	0	0	0	0	0	0	0	0	2
78	D. B.	1	1	0	0	0	0	0	0	0	0	0	2
81	G. A.	1	0	0	0	0	0	0	0	0	0	1	2
82	G. C.	0	1	1	0	0	0	0	0	0	0	0	2
19	H. L.	1	1	0	0	0	0	0	0	0	0	0	2
88	H. P.	1	1	0	0	0	0	0	0	0	0	0	2
94	L. L.	1	0	0	0	0	0	0	0	1	0	0	2
26	M. H.	1	1	0	0	0	0	0	0	0	0	0	2
28	M. M.	1	0	1	0	0	0	0	0	0	0	0	2
29	M. S.	1	1	0	0	0	0	0	0	0	0	0	2
98	M. O.	1	1	0	0	0	0	0	0	0	0	0	2
100	M. S.	0	0	1	1	0	0	0	0	0	0	0	2
34	M. Q.	1	0	0	0	0	0	1	0	0	0	0	2
36	O. O.	1	0	0	0	0	0	1	0	0	0	0	2
38	Q. O.	1	0	0	0	1	0	0	0	0	0	0	2
39	Q. O.	1	0	0	0	0	1	0	0	0	0	0	2
105	R. .P.	1	0	0	0	1	0	0	0	0	0	0	2
47	R. R.	1	1	0	0	0	0	0	0	0	0	0	2
52	S. F.	1	1	0	0	0	0	0	0	0	0	0	2
109	Z. P.	1	0	1	0	0	0	0	0	0	0	0	2
65	A. F.	1	0	0	0	0	1	0	0	1	0	0	3
4	A. V.	0	0	1	1	0	1	0	0	0	0	0	3
67	A. L.	1	0	1	0	1	0	0	0	0	0	0	3
68	A. M.	1	0	1	0	0	0	0	0	1	0	0	3
69	B. B.	1	1	0	0	0	0	0	1	0	0	0	3
5	B. M.	1	0	1	0	0	0	0	1	0	0	0	3
71	B. C.	1	1	0	1	0	0	0	0	0	0	0	3
75	C. S.	1	0	1	0	0	1	0	0	0	0	0	3
11	F. M.	1	1	1	0	0	0	0	0	0	0	0	3
16	H. B.	1	0	1	0	1	0	0	0	0	0	0	3
21	H. S.	1	1	1	0	0	0	0	0	0	0	0	3
24	J. M.	1	1	1	0	0	0	0	0	0	0	0	3
93	J. N.	0	1	1	1	0	0	0	0	0	0	0	3
25	L. P.	1	1	0	0	1	0	0	0	0	0	0	3
30	M. V.	1	1	1	0	0	0	0	0	0	0	0	3
35	N. F.	1	1	1	0	0	0	0	0	0	0	0	3
41	R. A.	0	1	1	0	0	1	0	0	0	0	0	3
46	R. G.	1	1	0	1	0	0	0	0	0	0	0	3

53	S. G.	1	1	1	0	0	0	0	0	0	0	0	3
58	T. C.	1	1	0	1	0	0	0	0	0	0	0	3
62	Z. M.	1	1	1	0	0	0	0	0	0	0	0	3
63	Z. Á.	1	1	0	0	1	0	0	0	0	0	0	3
72	B. B.	0	0	0	1	0	1	0	1	0	1	0	4
9	E. F.	1	1	0	1	1	0	0	0	0	0	0	4
17	H. G.	1	1	0	1	0	1	0	0	0	0	0	4
95	L. B.	1	1	0	1	0	0	0	1	0	0	0	4
97	L. M.	1	1	0	1	0	1	0	0	0	0	0	4
103	O. V.	1	0	1	1	0	0	0	1	0	0	0	4
37	P. O.	1	1	1	0	0	1	0	0	0	0	0	4
40	R. A.	1	1	1	0	1	0	0	0	0	0	0	4
44	R. R.	1	1	1	0	1	0	0	0	0	0	0	4
51	S. C.	0	1	0	1	1	0	0	1	0	0	0	4
55	T. M.	1	0	1	0	1	1	0	0	0	0	0	4
59	T. A.	1	1	1	0	1	0	0	0	0	0	0	4
61	V. M.	1	1	1	0	1	0	0	0	0	0	0	4
77	A. P.	0	1	1	1	1	0	0	0	1	0	0	5
14	G. G.	1	0	1	1	0	0	1	0	0	1	0	5
27	M. T.	1	1	1	1	0	0	0	0	1	0	0	5
54	S. P.	1	1	1	0	1	1	0	0	0	0	0	5
20	H. L.	1	1	1	1	1	0	0	1	0	0	0	6
45	R. E.	1	1	0	1	1	0	1	1	0	0	0	6
49	S. A.	1	1	1	1	0	1	1	0	0	0	0	6
12	G. C.	1	1	1	0	1	1	0	1	0	1	0	7
15	G. A.	1	1	1	1	1	1	1	0	0	1	0	8
		71	50	38	21	20	15	9	9	5	4	2	

Tabla 10. .Distribución agrupada de 10 diferentes puntajes de los ítems del grupo "A".

Índice del grupo ítem i	Nombre ítem 1	Puntaje 2 Si	Frec 3 fi	Proporción aciertos 4 Pi	proporción errores 5 1-Pi	Lógito incorrecto 6 Xi	Frec lógito 7 fi Xi	Frec por lógito al cuadrado 8 fiXi	Calibración inicial 9 Xi - x	$\overline{x}$ media	$f_1 \cdot \overline{x}$
1	2	45	1	0.78	0.22	-1.24	-1.24	1.54	-2.54	1.30	1.70
2	6	36	1	0.62	0.38	-0.49	-0.49	0.24	-1.79	1.30	1.70
3	3	28	1	0.48	0.52	0.07	0.07	0.00	-1.23	1.30	1.70
4	7	17	1	0.29	0.71	0.88	0.88	0.78	-0.42	1.30	1.70
5	11	13	1	0.22	0.78	1.24	1.24	1.54	-0.06	1.30	1.70
6	8	10	1	0.17	0.83	1.57	1.57	2.46	0.27	1.30	1.70
7	1	9	1	0.16	0.84	1.69	1.69	2.87	0.39	1.30	1.70
8	5	5	1	0.09	0.91	2.36	2.36	5.57	1.06	1.30	1.70
9	9	3	1	0.05	0.95	2.91	2.91	8.46	1.61	1.30	1.70
10	10	1	1	0.02	0.98	4.04	4.04	16.35	2.74	1.30	1.70
	suma	L_A=	10								
			datos N=	58		suma	13.03	39.82			16.99
X =	1.30										
		U =	2.54								

Tabla 11. .Distribución agrupada de 11 diferentes puntajes de los ítems del grupo "B".

Índice del grupo ítem i	Nombre ítem 1	Puntaje 2 Si	Frec 3 fi	Proporción aciertos 4 Pi	proporción errores 5 1-Pi	Lógito incorrecto 6 Xi	Frec lógito 7 fi Xi	Frec por lógito al cuadrado 8 fiXi	Calibración inicial 9 Xi - x	$\overline{x}$ media	$f_1 \cdot \overline{x}$
1	2	26	1	0.72	0.28	-0.96	-0.96	0.91	-2.61	1.65	2.71
2	6	14	1	0.39	0.61	0.45	0.45	0.20	-1.20	1.65	2.71
3	3	10	1	0.28	0.72	0.96	0.96	0.91	-0.69	1.65	2.71
4	11	8	1	0.22	0.78	1.25	1.25	1.57	-0.40	1.65	2.71
5	8	5	1	0.14	0.86	1.82	1.82	3.33	0.17	1.65	2.71
6	5,10	4	2	0.11	0.89	2.08	4.16	8.65	0.43	1.65	5.43
7	7	3	1	0.08	0.92	2.40	2.40	5.75	0.75	1.65	2.71
8	4	2	1	0.06	0.94	2.83	2.83	8.03	1.18	1.65	2.71
9	9	1	1	0.03	0.97	3.56	3.56	12.64	1.91	1.65	2.71
	suma	L_B=	10			suma	16.47	41.99			27.14
X =	1.65		U=	1.65							

Tabla 12. .Distribución agrupada de 11 diferentes puntajes de los ítems del grupo "AB".

Índice del grupo ítem i	Nombre ítem 1	Puntaje 2 Si	Frec 3 fi	Proporción aciertos 4 Pi	proporción errores 5 1-Pi	Lógito incorrecto 6 Xi	Frec lógito 7 fi Xi	Frec por lógito al cuadrado 8 fiXi	Calibración inicial 9 Xi - x	$\overline{x}$ media	$f_1 \cdot \overline{x}$
1	2	71	1	0.76	0.24	-1.13	-1.13	1.27	-2.73	1.61	2.58
2	6	50	1	0.53	0.47	-0.13	-0.13	0.02	-1.73	1.61	2.58
3	3	38	1	0.40	0.60	0.39	0.39	0.15	-1.22	1.61	2.58
4	11	21	1	0.22	0.78	1.25	1.25	1.55	-0.36	1.61	2.58
5	7	20	1	0.21	0.79	1.31	1.31	1.71	-0.30	1.61	2.58
6	8	15	1	0.16	0.84	1.66	1.66	2.76	0.06	1.61	2.58
7	1,5	9	2	0.10	0.90	2.25	4.49	10.08	0.64	1.61	5.16
8	10	5	1	0.05	0.95	2.88	2.88	8.29	1.27	1.61	2.58
9	9	4	1	0.04	0.96	3.11	3.11	9.69	1.51	1.61	2.58
10	4	2	1	0.02	0.98	3.83	3.83	14.66	2.22	1.61	2.58
	suma	L_{AB}=	11			suma	17.66	50.19			28.35
					media	de x	2	U=	2.18		

Tabla 13. Cálculos finales de las dificultades de los ítems grupo A.

Grupo ítem 1 i	Nombre ítem 2	calibración inicial ítem 3 di	Factor de expansión 4 y	Calificación Corregida de ítem 5 di= Ydi	Puntaje ítem 6 si	Error estándar calificación 7 SE(di)
1	2	-2.54	1.25	-3.18	45	0.39
2	6	-1.79	1.25	-2.24	36	0.34
3	3	-1.23	1.25	-1.54	28	0.33
4	7	-0.42	1.25	-0.53	17	0.36
5	11	-0.06	1.25	-0.08	13	0.39
6	8	0.27	1.25	0.34	10	0.43
7	1	0.39	1.25	0.49	9	0.45
8	5	1.06	1.25	1.33	5	0.59
9	9	1.61	1.25	2.01	3	0.74
10	10	2.74	1.25	3.43	1	1.26
X	1.54					
Y	1.25					
N	58					

Tabla 14. Cálculos finales de las dificultades de los ítems grupo B.

Grupo ítem 1 i	Nombre ítem 2	calibración inicial ítem 3 di	Factor de expansión 4 y	Calificación Corregida de ítem 5 di= Ydi	Puntaje ítem 6 si	Error estándar calificación 7 SE(di)
1	2	-2.61	1.12	-2.93	26	0.42
2	6	-1.20	1.12	-1.35	14	0.38
3	3	-0.69	1.12	-0.78	10	0.42
4	11	-0.40	1.12	-0.45	8	0.45
5	8	0.17	1.12	0.20	5	0.54
6	5,10	0.43	1.12	0.48	4	0.60
7	7	0.75	1.12	0.84	4	0.60
8	4	1.18	1.12	1.33	3	0.68
9	9	1.91	1.12	2.14	2	0.82
X	1.31					
Y	1.12					
N	36					

Tabla 15. Cálculos finales de las dificultades de los ítems Grupo AB.

Grupo ítem 1 i	Nombre ítem 2	calibración inicial ítem 3 di	Factor de expansión 4 y	Calificación Corregida de ítem 5 di= Ydi	Puntaje ítem 6 si	Error estándar calificación 7 SE(di)
1	2	-2.73	1.20	-3.29	71	0.29
2	6	-1.73	1.20	-2.09	50	0.25
3	3	-1.22	1.20	-1.47	38	0.25
4	11	-0.36	1.20	-0.43	21	0.30
5	7	-0.30	1.20	-0.36	20	0.30
6	8	0.06	1.20	0.07	15	0.34
7	1,5	0.64	1.20	0.77	9	0.42
8	10	1.27	1.20	1.53	5	0.55
9	9	1.51	1.20	1.82	4	0.62
10	4	2.22	1.20	2.68	2	0.86
X	1.45					
Y	1.20					
N	94					

Tabla 16. Distribución agrupada de puntajes de personas grupo "A" en 11 ítems..

puntaje posible 1 r	Frec. de las personas 2 n_r	Proporción de aciertos 3 $P_r = r/L$	Lógito de aciertos 4 $Y_r = \ln(P_r/1-P_r)$	frec. por Lógito 5 $n_r Y_r$	frec. por lógito al cuadrado 6 $n_r Y_r^2$	Medición inicial de las personas 7 $b^o{}_r = Y_r$
1	14	0.09	-2.30	-32.24	74.23	-2.30
2	12	0.18	-1.50	-18.05	27.15	-1.50
3	15	0.27	-0.98	-14.71	14.43	-0.98
4	9	0.36	-0.56	-5.04	2.82	-0.56
5	3	0.45	-0.18	-0.55	0.10	-0.18
6	3	0.55	0.18	0.55	0.10	0.18
7	1	0.64	0.56	0.56	0.31	0.56
8	1	0.73	0.98	0.98	0.96	0.98
9	0	0.82	1.50	0.00	0.00	1.50

Σn_r 58		$\Sigma n_r y_r$ -68.49		$\Sigma n_r y_r^2$ 120.10	
	L = 11				
	Y = -1.18	V = 0.68			

Tabla 17. Distribución agrupada de puntajes de personas grupo "B".

puntaje posible 1 r	Frec. de las personas 2 n_r	Proporción de aciertos 3 $P_r = r/L$	Lógito de aciertos 4 $Y_r = \ln(P_r/1-P_r)$	frec. por Lógito 5 $n_r Y_r$	frec. por lógito al cuadrado 6 $n_r Y_r^2$	Medición inicial de las personas 7 $b^o{}_r = Y_r$
1	13	0.09	-2.30	-29.93	68.92	-2.30
2	11	0.18	-1.50	-16.54	24.88	-1.50
3	7	0.27	-0.98	-6.87	6.73	-0.98
4	4	0.36	-0.56	-2.24	1.25	-0.56
5	1	0.45	-0.18	-0.18	0.03	-0.18
6	0	0.55	0.18	0.00	0.00	0.18
7	0	0.64	0.56	0.00	0.00	0.56
8	0	0.73	0.98	0.00	0.00	0.98
9	0	0.82	1.50	0.00	0.00	1.50
Σn_r	36		$\Sigma n_r y_r$	-55.77	$\Sigma n_r y_r^2$	101.83
	L=11					
		Y=-1.54			V = 0.44	

Tabla 18. Distribución agrupada de puntajes de personas grupo "AB".

puntaje posible 1 r	Frec. de las personas 2 n_r	Proporción de aciertos 3 $P_r = r/L$	Lógito de aciertos 4 $Y_r = \ln(P_r/1-P_r)$	frec. por Lógito 5 $n_r Y_r$	frec. por lógito al cuadrado 6 $n_r Y_r^2$	Medición inicial de las personas 7 $b^o{}_r = Y_r$
1	27	0.09	-2.30	-62.17	143.15	-2.30
2	23	0.18	-1.50	-34.59	52.03	-1.50
3	22	0.27	-0.98	-21.58	21.16	-0.98
4	13	0.36	-0.56	-7.28	4.07	-0.56
5	4	0.45	-0.18	-0.73	0.13	-0.18
6	3	0.55	0.18	0.55	0.10	0.18
7	1	0.64	0.56	0.56	0.31	0.56
8	1	0.73	0.98	0.98	0.96	0.98
9	0	0.82	1.50	0.00	0.00	1.50
10	0	0.91	2.30	0.00	0.00	2.30
Σn_r	94	11.00				
			$\Sigma n_r y_r$	-124.26	$\Sigma n_r y_r^2$	221.93
	Y= -1.32					
				V=0.62		

Calibración de ítems con ordenador

Datos en procesamiento con programa XCalibre.

Tabla 19. Dificultad de Rasch datos grupo "A"

	Discriminación	Dificultad
Ítem	a	b
2	0.73	-1.58
6	0.76	-0.80
3	0.78	-0.01
7	0.88	1.05
11	0.91	1.48
8	0.89	1.88
1	0.63	2.17
5	0.93	2.56
9	0.99	2.89
10	0.74	3.00
Media	0.82	1.26
Desv.Std.	0.11	1.58

Tabla 20. Dificultad de Rasch datos "B".

	Discriminación	Dificultad
Ítem	a	b
8	0.57	-1.19
1	0.57	0.75
7	0.67	1.46
11	0.89	1.48
2	0.68	2.42
3	0.73	2.67
10	0.84	2.57
5	0.74	2.96
6	0.58	3.00
Media	0.70	1.79
Desv. Std.	0.12	1.36

Tabla 21. Dificultad de Rasch "AB".

Grupo AB	Discriminación	Dificultad
Ítem	a	b
2	0.57	-1.55
6	0.66	-0.25
3	0.74	0.52
11	0.89	1.52
7	0.85	1.66
8	0.77	2.15
1	0.59	2.9
5	0.83	2.64
10	0.62	3.00
9	0.88	3.00
4	0.64	3.00
Media	0.73	1.69
Desv.Std.	0.12	1.53

Salida del modelo de Rasch del grupo AB

```
                        XCALIBRE (tm) Version 1.10                  Page  1
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT
   ********************  ANALYSIS SUMMARY INFORMATION  ********************
         Data (Input) File: C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT
      Analysis Output File: C:\TESISM~1\EXCALI~1\EX\MARUAB.OUT
         Score Output File: C:\TESISM~1\EXCALI~1\EX\MARUAB.SCR
            Item Name File: NONE
    Statistics Output File: NONE
                     *****  CONFIGURATION INFORMATION  *****
             Item Parameter Priors:  DEFAULT
              Allow Priors to Float:  YES
                      IRT Model Used:  2-parameter
            Maximum Number of Loops: 12
            Scoring Method Selected:  Expected a Posteriori (EAP)
       Starting Prior Distribution Moments:
                Mean       SD
         a    0.7500   0.1200
         b    0.0000   2.0000
         c    0.0000   0.0000
     NOTE:  ***  will be printed when the c standard error value > 0.10


                        XCALIBRE (tm) Version 1.10                  Page  2
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
```

```
¦¦XCALIBRE IRT analysis for data from file:
C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT
Date: 17/10/01                                                Time: 12:55 pm
The input was from file:  C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT
The number of items was:   11
There was no item linkage
The key was:
11111111111
The numbers of alternatives were:
22222222222
The inclusion specifications were:
YYYYYYYYYYY
The maximum parameter change on loop 1 was      1.977
The maximum parameter change on loop 2 was      0.118
The maximum parameter change on loop 3 was      0.106
The maximum parameter change on loop 4 was      0.052
¦The maximum parameter change on loop 5 was      0.039
¦¦Mean Number-Correct Score =    2.596
¦¦Number-Correct Standard Deviation =    1.504
¦¦K-R 21 Reliability = 0.402
¦¦The number of examinees was 94
¦¦¦ Final Parameter Summary Information:
¦¦         Mean      SD
¦¦¦Theta    0.00     1.00
¦¦  a      0.72     0.11
¦¦  b      1.67     1.51
¦¦  c      0.00     0.00
¦                          XCALIBRE (tm) Version 1.10                  Page  3
¦¦   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
¦¦          Marginal Maximum-Likelihood IRT Parameter Estimation Program
¦XCALIBRE IRT analysis for data from file: C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT
¦¦FINAL ITEM PARAMETER ESTIMATES
¦¦Item Lnk Flg   a      b      c    Resid   PC    PBs    PBt      N      Item name
¦¦  1           0.59   2.93   0.00  1.46   0.10  0.23   0.16      94
¦¦  2           0.55  -1.72   0.00  1.25   0.76  0.29   0.24      94
¦¦  3           0.72   0.47   0.00  0.68   0.40  0.50   0.52      94
¦¦  4       PK  0.63   3.00   0.00  1.11   0.02 -0.11  -0.12     94
¦¦  5           0.81   2.67   0.00  0.65   0.10  0.42   0.43      94
¦¦  6           0.63  -0.35   0.00  1.04   0.53  0.44   0.43      94
¦¦  7           0.84   1.68   0.00  0.98   0.21  0.52   0.52      94
¦¦  8           0.76   2.19   0.00  1.03   0.16  0.45   0.45      94
¦¦  9       P   0.87   3.00   0.00  0.27   0.04  0.48   0.42      94
¦¦ 10       P   0.62   3.00   0.00  0.45   0.05  0.16   0.15      94
¦¦ 11           0.88   1.53   0.00  0.67   0.22  0.55   0.58      94
¦¦                         XCALIBRE (tm) Version 1.10                  Page  4
¦¦   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
¦¦          Marginal Maximum-Likelihood IRT Parameter Estimation Program
¦¦XCALIBRE IRT analysis for data from file:
C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT ¦
¦ITEM PARAMETER ESTIMATES W/STANDARD ERRORS
¦                        a             b             c
¦Item Lnk Flg   a     error    b     error    c     error  Resid   Item name
¦  1           0.59  0.175   2.93  0.410   0.00    N/A    1.46
¦  2           0.55  0.195  -1.72  0.287   0.00    N/A    1.25
¦  3           0.72  0.222   0.47  0.200   0.00    N/A    0.68
¦  4       PK  0.63  0.180   3.00  0.418   0.00    N/A    1.11
¦  5           0.81  0.195   2.67  0.357   0.00    N/A    0.65
¦  6           0.63  0.252  -0.35  0.216   0.00    N/A    1.04
¦  7           0.84  0.187   1.68  0.234   0.00    N/A    0.98
¦  8           0.76  0.183   2.19  0.292   0.00    N/A    1.03
¦  9       P   0.87  0.217   3.00  0.424   0.00    N/A    0.27
¦ 10       P   0.62  0.178   3.00  0.419   0.00    N/A    0.45
¦ 11           0.88  0.190   1.53  0.217   0.00    N/A    0.67
                           XCALIBRE (tm) Version 1.10                  Page  5
¦¦   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
¦¦          Marginal Maximum-Likelihood IRT Parameter Estimation Program
¦¦¦XCALIBRE IRT analysis for data from file:
```

ITEM ANALYSIS

Item	Endorsement Rate 1	2	Oth	Item-Theta Corr. 1	2	Oth
1	10~	0	90	16~	0	-16
2	76~	0	24	24~	0	-24
3	40~	0	60	52~	0	-52
4	2~	0	98	-12~	0	12
5	10~	0	90	43~	0	-43
6	53~	0	47	43~	0	-43
7	21~	0	79	52~	0	-52
8	16~	0	84	45~	0	-45
9	4~	0	96	42~	0	-42
10	5~	0	95	15~	0	-15
11	22~	0	78	58~	0	-58

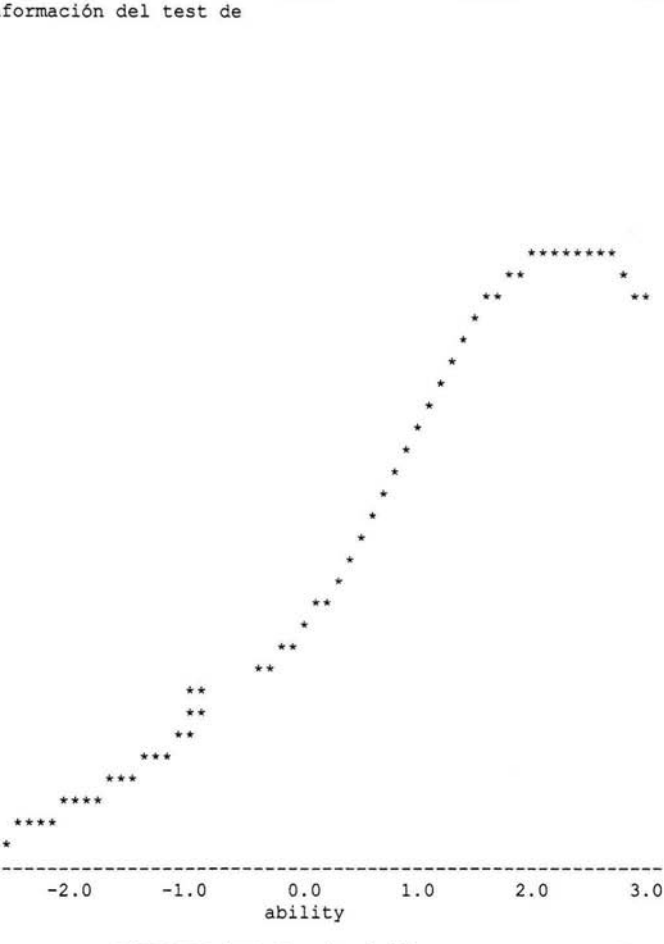
Marginal Maximum-Likelihood IRT Parameter Estimation Program

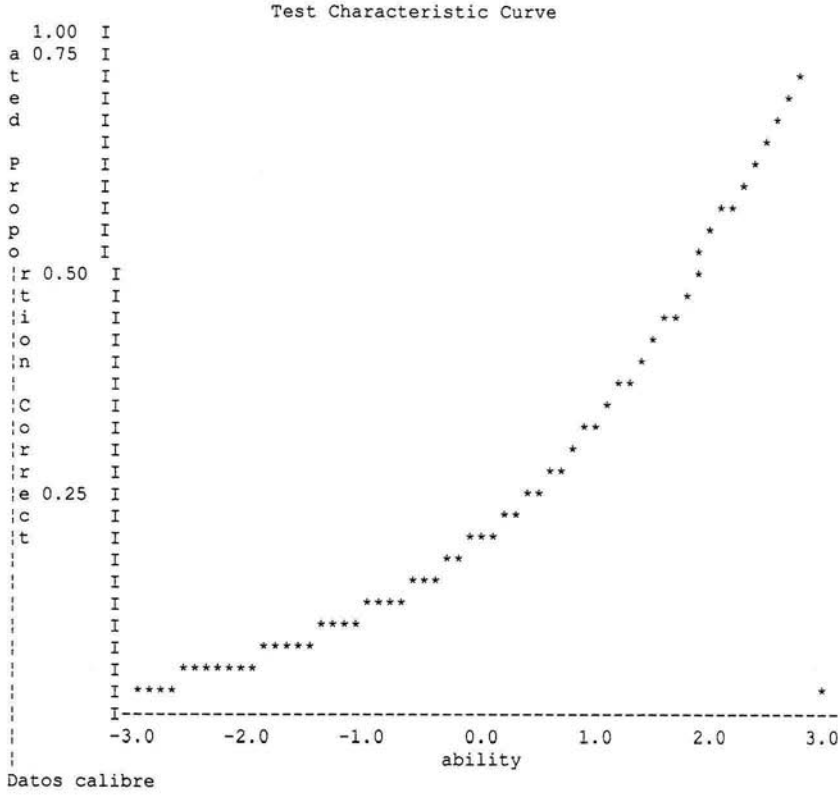
XCALIBRE IRT analysis for data from file:
C:\TESISM~1\EXCALI~1\EX\GRUPOAB.DAT

Test characteristics:

K-R 21 Reliability	Expected Information	Average Information
0.402	1.501	1.543

Curva de información del test de





Datos calibre

Grupo A¦

```
                        XCALIBRE (tm) Version 1.10                 Page  1
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
   ********************  ANALYSIS SUMMARY INFORMATION  ********************
         Data (Input) File: A:\EXCALI~1\EX\GRUPOA.DAT
      Analysis Output File: A:\EXCALI~1\EX\GRUPOA.OUT
         Score Output File: A:\EXCALI~1\EX\GRUPOA.SCR
            Item Name File: NONE
    Statistics Output File: NONE
                  *****  CONFIGURATION INFORMATION  *****
           Item Parameter Priors:  DEFAULT
FILE: A:\EXCALI~1\EX\GRUPOA.OUT                Lines: 1 - 23      Col: 1
Allow Priors to Float:  YES                                       ¦¦
                  IRT Model Used:  2-parameter
          Maximum Number of Loops: 12
          Scoring Method Selected:  Maximum-Likelihood
       Starting Prior Distribution Moments:
              Mean      SD
         a    0.7500    0.1200
         b    0.0000    2.0000
         c    0.0000    0.0000
     NOTE:  ***  will be printed when the c standard error value > 0.10
                        XCALIBRE (tm) Version 1.10                 Page  2
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
The input was from file:  A:\EXCALI~1\EX\GRUPOA.DAT
The number of items was:   11
```

```
There was no item linkage
The key was:
11111111111
The numbers of alternatives were:
22222222222
The inclusion specifications were:
YYYYYYYYYYY
*** NOTE ***  Item 11 had no variance.
              It was excluded from the analysis.
The maximum parameter change on loop 1 was      1.909
The maximum parameter change on loop 2 was      0.186
The maximum parameter change on loop 3 was      0.056
The maximum parameter change on loop 4 was      0.037
¦Mean Number-Correct Score =    2.879
¦Number-Correct Standard Deviation =    1.641
¦K-R 21 Reliability = 0.504
¦The number of examinees was 58
 Final Parameter Summary Information:
        Mean     SD
Theta   0.00    1.00
  a     0.82    0.11
  b     1.27    1.50
  c     0.00    0.00
```

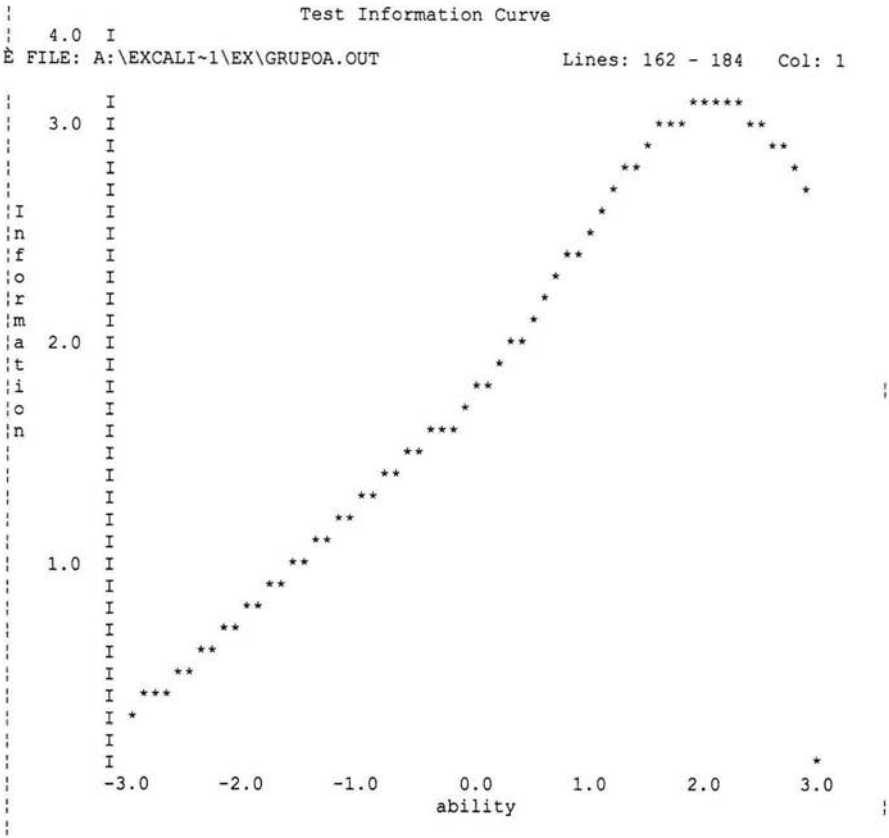


```
        Marginal Maximum-Likelihood IRT Parameter Estimation Program
¦XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
¦¦FINAL ITEM PARAMETER ESTIMATES
¦Item Lnk Flg   a     b     c   Resid   PC    PBs   PBt     N      Item name
¦  1           0.73 -1.58  0.00  0.68  0.78  0.41  0.43     58
¦  2           0.76 -0.80  0.00  0.78  0.62  0.48  0.50     58
¦  3           0.78 -0.01  0.00  0.56  0.48  0.51  0.53     58
¦  4           0.88  1.05  0.00  0.63  0.29  0.53  0.55     58
¦  5           0.91  1.48  0.00  0.73  0.22  0.52  0.52     58
¦  6           0.89  1.88  0.00  0.88  0.17  0.48  0.48     58
¦  7           0.63  2.17  0.00  1.63  0.16  0.21  0.12     58
¦  8           0.93  2.56  0.00  0.49  0.09  0.43  0.43     58
¦  9           0.99  2.89  0.00  0.22  0.05  0.54  0.48     58
¦ 10       P   0.74  3.00  0.00  0.42  0.02  0.17  0.16     58
¦ 11  -- Deleted --
¦
```

```
                        XCALIBRE (tm) Version 1.10                 Page  4
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
ITEM PARAMETER ESTIMATES W/STANDARD ERRORS
                          a              b              c
Item Lnk Flg    a     error     b     error     c     error  Resid   Item name
  1           0.73   0.246   -1.58   0.298   0.00     N/A    0.68
  2           0.76   0.283   -0.80   0.247   0.00     N/A    0.78
  3           0.78   0.291   -0.01   0.229   0.00     N/A    0.56
  4           0.88   0.248    1.05   0.242   0.00     N/A    0.63
  5           0.91   0.240    1.48   0.270   0.00     N/A    0.73
  6           0.89   0.239    1.88   0.315   0.00     N/A    0.88
  7           0.63   0.225    2.17   0.397   0.00     N/A    1.63
  8           0.93   0.265    2.56   0.423   0.00     N/A    0.49
  9           0.99   0.300    2.89   0.508   0.00     N/A    0.22
 10       P   0.74   0.249    3.00   0.531   0.00     N/A    0.42
 11   -- Deleted --


                        XCALIBRE (tm) Version 1.10                 Page  5
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
ITEM ANALYSIS
      Endorsement Rate          Item-Theta Corr.
Item   1   2  Oth                1   2  Oth
  1    78~  0  22                43~  0 -43
  2    62~  0  38                50~  0 -50
  3    48~  0  52                53~  0 -53
  4    29~  0  71                55~  0 -55
  5    22~  0  78                52~  0 -52
  6    17~  0  83                48~  0 -48
  7    16~  0  84                12~  0 -12
  8     9~  0  91                43~  0 -43
  9     5~  0  95                48~  0 -48
 10     2~  0  98                16~  0 -16
È FILE: A:\EXCALI~1\EX\GRUPOA.OUT              Lines: 139 - 161   Col: 1
 11   -- Deleted --


                        XCALIBRE (tm) Version 1.10                 Page  6
   Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
           Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT
     Test characteristics:          K-R 21          Expected         Average
                                 Reliability      Information     Information
                                    0.504            1.845           1.784
```



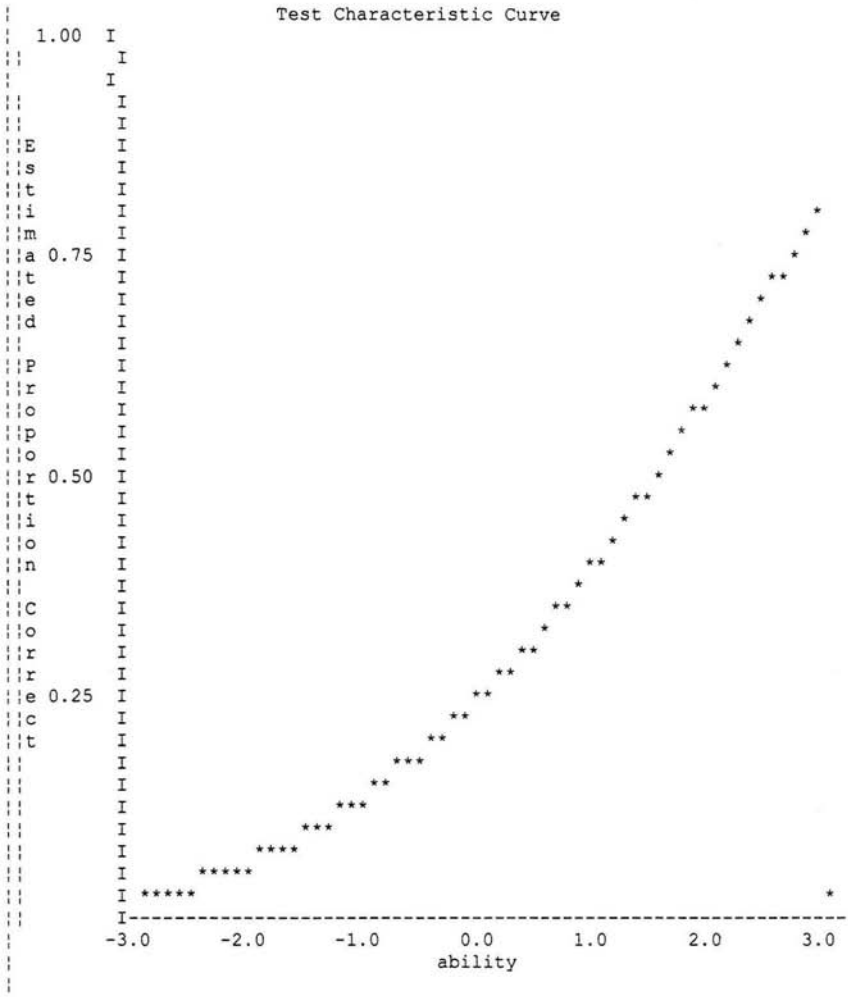
È FILE: A:\EXCALI~1\EX\GRUPOA.OUT Lines: 162 - 184 Col: 1

XCALIBRE (tm) Version 1.10 Page 7

Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved

Marginal Maximum-Likelihood IRT Parameter Estimation Program

XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOA.DAT



Grupo B

FINAL ITEM PARAMETER ESTIMATES

Item	Lnk	Flg	a	b	c	Resid	PC	PBs	PBt	N	Item name
1			0.57	-1.19	0.00	1.18	0.68	0.20	0.20	38	
2			0.57	0.75	0.00	0.80	0.37	0.33	0.30	38	
3			0.67	1.46	0.00	0.56	0.26	0.42	0.44	38	
4			0.89	1.48	0.00	0.10	0.21	0.68	0.67	38	
5			0.68	2.42	0.00	0.49	0.13	0.34	0.35	38	
6			0.73	2.67	0.00	0.56	0.11	0.37	0.34	38	
7			0.84	2.57	0.00	0.51	0.11	0.45	0.46	38	
8			0.74	2.96	0.00	0.53	0.08	0.34	0.31	38	
9			0.58	3.00	0.00	0.87	0.05	-0.10	-0.10	38	

XCALIBRE (tm) Version 1.10 Page 4
Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
Marginal Maximum-Likelihood IRT Parameter Estimation Program
XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOB.DAT

Grupo AB

XCALIBRE IRT analysis for data from file: A:\EXCALI~1\EX\GRUPOAB.DAT
¦FINAL ITEM PARAMETER ESTIMATES

Item	Lnk	Flg	a	b	c	Resid	PC	PBs	PBt	N	Item name
1			0.57	-1.55	0.00	1.10	0.74	0.33	0.30	96	
2			0.66	-0.25	0.00	0.94	0.52	0.46	0.46	96	
3			0.74	0.52	0.00	0.64	0.40	0.51	0.53	96	
4			0.89	1.52	0.00	0.60	0.22	0.55	0.58	96	
5			0.85	1.66	0.00	0.88	0.21	0.52	0.52	96	
6			0.77	2.15	0.00	0.93	0.16	0.45	0.44	96	
7			0.59	2.90	0.00	1.36	0.09	0.24	0.16	96	
8			0.83	2.64	0.00	0.64	0.09	0.42	0.43	96	
9		P	0.62	3.00	0.00	0.43	0.05	0.16	0.15	96	
10		P	0.88	3.00	0.00	0.26	0.04	0.47	0.41	96	
11		PK	0.64	3.00	0.00	1.07	0.02	-0.10	-0.10	96	

¦
¦

XCALIBRE (tm) Version 1.10 Page 4
Copyright (c) 1995 by Assessment Systems Corporation, All Rights Reserved
Marginal Maximum-Likelihood IRT Parameter Estimation Program
FILE: A:\EXCALI~1\EX\GRUPOAB.OUT Lines: 93 - 115 Col: 1

Referencias y Bibliografía

Airen, R. (2003). *Tests psicológicos y evaluación.* México: Pearson.

Anastasi A. Urbina S. (1998) *Tests psicológicos..* México: Pretince Hall.

Barton y Lord (1981).*An upper asymptote for the three parameter logistic item response model.* En Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* España: Pirámide.

Berliner, D (1987) *But do they understand?* En Díaz B y Hernández, G. (1998). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista.* México: Mc Graw-Hill.

Canut, E. y Jiménez, J. (2000). La Responsabilidad de Evaluar en el Proceso Educativo.. *Genio e ingenio, # 3 de revista Educación, pp.# 28*

Clark-Carter, D. (2002). *Investigación cuantitativa en psicología.* México: Oxford University Press.

Cronbach (1951).*Coefficient alpha and the internal structure of tests.* En Muñiz, J.(1996). *Teoría Clásica de los tests.* Madrid, España: Pirámide.

Díaz B. y Hernández, G. (1998). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista.* México: Mc Graw-Hill.

Du Bois (1970) *A History of psychological testing.* En Muñiz, J.(1996). *Teoría Clásica de los tests.* Madrid, España: Pirámide

García J. (1994). *Bases pedagógicas de la evaluación. Guía práctica para educadores.* España: Síntesis.

Genovard y Gotzens (1990). *Psicología de la instrucción.* En Díaz B. y Hernández, G. (1998). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista.* México: Mc Graw-Hill.

Guilliksen (1950).*Theory of mental test.* En Muñiz, J.(1996). *Teoría Clásica de los tests*. Madrid, España: Pirámide.

Lord (1980) *Applications of item response theory to practical testing problems.* En Baker F.(1985) *The Basics of item response theory* Portsmouth, NH Heinemann.

Lord y Novick (1968).*An analysius of the verbal scholastic aptitude test using Birnbaums three parameter logistic model.* En Muñiz, J.(1996). *Teoría Clásica de los tests*. Madrid, España: Pirámide.

Manual de la educación.(2000) Grupo Océano. Barcelona España.

Muñiz, J. (1990). *Teoría de respuesta a los ítems. Un nuevo enfoque en la evolución psicológica y educativa*. Madrid, España: Pirámide

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

Muñiz, J.(1996). *Teoría Clásica de los tests*. Madrid, España: Pirámide

Nunnally, J. y Bernstein, I. (1995). *Teoría psicométrica*. México: McGraw-Hill.

Richardson (1936)*The relationship between difficulty and differential validity of a test.* En Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. España: Pirámide.

Rodríguez, H. y García E. (1999). *Evaluación en el aula*. México: Trillas.

Salkind, N. (1999). *Métodos de Investigación*. México: Pretince-Hall.

Santibáñez, J.(2001) *Manual para el aprendizaje estudiantil.* México: Trillas.

Tenbrink, F. (1981) *Evaluación. Guía práctica para profesores.* XXXXX

Thorndike (1982).. *Applied psychometrics*. En Muñiz, J.(1996). *Teoría Clásica de los tests*. Madrid, España: Pirámide.

Thurstone (1928). *Attitudes can be measured.* Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* España: Pirámide.

Thurstone (1937). *Psychology as a quantitative rational science.* En Muñiz, J.(1996). *Teoría Clásica de los tests.* Madrid, España: Pirámide.

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

Tucker (1946). *Maximum validity of a test with equivalent items.* En Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems.* España: Pirámide.

Woolfolk, A. (1996). *Psicología educativa.* México: Pretince –Hall.

Wright, B. y Stone, H. (1998). *Diseño de las mejores pruebas. Utilizando la técnica de XXXXXXXX*

Zavala, P. (2001). ¿Cómo evaluar las matemáticas?.Investigación presentada en el Foro La problemática en la enseñanza de las matemáticas.
http://www.itmx.mx/foro/2001/evaluacion01comoevaluar.doc.