UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Instituto de Investigación y Desarrollo Educativo

Desarrollo, operación y evaluación de un módulo para capacitar a docentes en servicio para que mejoren sus exámenes de opción múltiple mediante el análisis gráfico de ítems

TESIS

Que para obtener el grado de

MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

Guadalupe De Los Santos Lázaro

Ensenada, Baja California
Octubre de 2010

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Maestría en Ciencias Educativas

Desarrollo, operación y evaluación de un módulo para capacitar a docentes en servicio para que mejoren sus exámenes de opción múltiple mediante el análisis gráfico de ítems

T E S I S

que para obtener el grado de

MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

Guadalupe De Los Santos Lázaro

APROBADA POR:

Dr. Luis Ángel Contreras Niño
Director de Tesis

Mtri. Jose Luis. Ramírez Cuevas
Sinodal

M.C. Maricela López Ornelas
Sinodal

Dr. Joaquín Caso Niebla
Sinodal

Ensenada B.C. Octubre, 2010

Dedicatoria

A mí querida familia

Gracias por apoyarme en todo momento

Y

Por estar conmigo siempre.

¡Gracias!

A todos los buenos amigos de esta hermosa Ciudad

A Claudia, Arantza, Erick, Rosaura

Por haber compartido momentos inolvidables

Y por su apoyo incondicional en todo momento.

Y a todos aquellos que me apoyaron en esta etapa de mi vida

¡Gracias!

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

Resumen

La evaluación del aprendizaje es inherente al proceso educativo. Sin embargo, existe evidencia de que los profesores tienen una capacidad limitada para calificar a sus estudiantes de manera justa, válida, confiable y congruente con lo que enseñan (Shepard, 2006). En México, ello puede explicarse porque el currículum para la formación inicial de profesores privilegia contenidos teóricos por lo que los futuros docentes no tienen competencias necesarias para evaluar con calidad (Martínez, Maya y Zenteno, 1996). Además, la actualización que reciben no es suficiente para lograr un desempeño eficaz (Lozano, 2007).

A partir de lo anterior surge la necesidad de realizar una intervención para apoyarlos, de manera que puedan mejorar las pruebas que desarrollan como parte de su trabajo cotidiano. Así, este estudio buscó que los docentes adquirieran conocimientos y habilidades elementales de psicometría, que les permitieran saber si los ítems objetivos que utilizan en sus pruebas tienen dificultad apropiada y buena discriminación.

Se considero que la técnica del Análisis Gráfico de Ítems (AGI) desarrollada por Batenburg y Laros (2002), es un enfoque que permite al docente analizar simple, directa y visualmente cómo sus estudiantes responden ítems de respuesta seleccionada. Haladyna (2004) considera que una representación gráfica permite observar variaciones en las tendencias de la respuesta correcta y los distractores en un ítem de respuesta seleccionada, respecto a la calificación en la prueba. Así, el análisis de ítems resulta más significativo e interpretable y pueden comprenderlo fácilmente personas ajenas a la psicometría, sin antecedentes estadísticos necesarios para calcular los índices tradicionales de los parámetros del ítem. En consecuencia, el objetivo general del estudio fue desarrollar, operar y evaluar un módulo para habilitar a profesores en servicio, para que mejoren sus pruebas mediante esa técnica.

Los participantes en el estudio fueron 21 profesores de secundaria en servicio de Ensenada, Baja California, con diversas especialidades. Para participar en el módulo los docentes debieron traer una laptop y un archivo con las respuestas brutas que dieron 32 o más alumnos a por lo menos 10 ítems de una prueba aplicada previamente. Para operar y evaluar el módulo, se desarrollaron y aplicaron los siguientes materiales e instrumentos:

- *Manual del participante.* Contiene información conceptual y metodológica sobre la teoría clásica de los tests y el análisis gráfico de ítems.
- *Programa para el Análisis Gráfico de Ítems (PAGI).* Se adaptó la técnica del AGI para la evaluación en el aula y se desarrolló un programa de cómputo para que los docentes utilizaran dicha técnica.
- *Cuestionario para evaluar la operación del módulo.* Permitió realimentar la efectividad del entrenamiento, la actuación de la instructora y los materiales empleados. Se estructuró con ítems tipo Likert con cinco opciones de respuesta, para manifestar el grado de acuerdo respecto a afirmaciones correspondientes a tres dimensiones evaluadas: la instrucción, los materiales del curso-taller y la operación general del módulo.
- *Guía para interpretar el comportamiento de los ítems*. Se elaboró para apoyar la interpretación de las gráficas obtenidas sobre la calidad técnica de los ítems.
- *Presentación PowerPoint*. Contiene un conjunto de diapositivas que ilustran aspectos conceptuales de la Teoría Clásica de los Tests y del Análisis Gráfico de Ítems.

Tras elaborar estos materiales e instrumentos se siguieron dos procedimientos:

- Capacitación. Tuvo la modalidad de curso – taller y se realizó en una sesión de cuatro horas. Primero se revisaron conceptos necesarios para la interpretación gráfica de ítems y para entender la interfaz de PAGI. Después, cada participante analizó las gráficas que produjo PAGI con sus ítems, hasta que fue capaz de emitir en cada caso un dictamen razonado.
- Aplicación del cuestionario para evaluar la operación del módulo. Al concluir el curso – taller, se aplicó el cuestionario para evaluar su operación.

Los principales resultados del estudio fueron:

- El cuestionario mostró evidencias de calidad técnica. En cuanto a confiabilidad, se obtuvo un coeficiente Alfa de 0.88 y, salvo un ítem con correlación ítem – total de 0.19, los demás tuvieron un rango de discriminación entre 0.28 y 0.82, con una media de 0.53. En consecuencia, se pudo proceder con confianza a analizar los resultados de su aplicación.
- La opinión de los profesores fue muy favorable en general. El promedio de adscripción a los ítems fue de 4.57 en la escala tipo Likert del 1 al 5. Respecto a las tres dimensiones evaluadas, el promedio más alto correspondió a Instrucción (4.66), seguida de Materiales de apoyo (4.60) y Operación de la capacitación (4.57).

- Desde el punto de vista metodológico, un resultado relevante fue haber logrado adaptar la técnica del AGI a las características y condiciones del aula. El problema era garantizar que con pocos estudiantes se podían observar claramente los cambios de tendencia en la curvas de la respuesta correcta y de los distractores.

Los resultados obtenidos permiten concluir que se cumplieron los objetivos propuestos. Así, fue posible desarrollar, operar y evaluar el módulo para habilitar a docentes en servicio de modo que puedan mejorar sus pruebas objetivas mediante el AGI. Los siguientes argumentos apoyan esta conclusión:

- Los profesores reportaron aprendizajes significativos de conocimientos y habilidades básicos de psicometría, necesarios para el análisis gráfico de sus ítems.
- En opinión de los participantes, PAGI les permitió identificar fortalezas y debilidades de sus ítems, según los indicadores psicométricos que definen su calidad técnica, así como formas específicas de mejorarlos. Sus participaciones durante la capacitación y los productos que generaron tras analizar sus ítems y llenar la guía, confirman esta opinión.
- El instrumento de evaluación arrojó datos que muestran su calidad psicométrica y por ello se puede afirmar que en opinión de los docentes los materiales diseñados y la instrucción propiciaron aprendizajes significativos que fueron pretendidos.

No obstante, el estudio tiene limitaciones. Entre ellas puede señalarse que el AGI incluye aspectos como análisis de sesgo y construcción de versiones paralelas del test, que no se incluyeron en el trabajo por razones temporales. Por ello, se sugiere rediseñar PAGI y el manual para incluir tales aspectos, lo que hará más útil y versátil la capacitación.

También se requiere replicar la capacitación con profesores de otros niveles y modalidades educativos, para observar en qué medida podría integrarse a la formación y la actualización de profesores.

1.1 Introducción

En el presente capítulo se describe brevemente cómo se ha aplicado en México la evaluación en el aula y se comenta que no existe una práctica de la evaluación que cumpla con requerimientos mínimos de calidad técnica, por lo que el estudio propone acercar a los docentes a los aspectos psicométricos elementales que les permitan aplicar una evaluación de calidad técnica dentro del salón de clases.

La evaluación como un aspecto propio del proceso educativo siempre ha suscitado interés y también controversia en los diferentes niveles educativos. Sin embargo, la evaluación no es algo novedoso ya que ha sido practicada desde el siglo XIX, con una concepción en donde los alumnos están organizados en grados por edad y tienen un avance similar (Shepard, 2006).

Dentro del ámbito escolar, los profesores en servicio son quienes se encargan de evaluar el aprendizaje de los alumnos. El profesor siempre ha interactuado con los alumnos para establecer una evaluación del aprendizaje de cada uno de ellos, en función del avance que muestren. Por lo cual, surge la necesidad de elaborar instrumentos que detecten las necesidades de aprendizaje de los alumnos para atenderlas y establecer objetivos para un mejor desempeño académico. Los profesores al momento de elaborar instrumentos ponen un mayor énfasis en el aspecto técnico, dejando de lado las conexiones entre la evaluación y las actividades de enseñanza. De tal forma que la comunidad de profesores ha recibido una capacitación en medición con temas formales y técnicos, pero lamentablemente esta capacitación ha sido deficiente para ayudar a los profesores a cumplir con una evaluación eficaz (Shepard, 2006).

La misma autora, comenta que en Estados Unidos los primeros volúmenes de la prestigiada revista *Educational Measurement*, no estaban enfocados en pruebas que los profesores realizan día a día dentro del salón de clase. Los profesores recibían una enseñanza acerca de la evaluación por medio de objetivos de enseñanza, con el fin de volverlos sistemáticos mediante el uso de los formatos de reactivos para evaluar el dominio de un contenido importante. Así, los profesores se han dedicado más a la evaluación que a la docencia. Al respecto, existe la necesidad de considerar las

consecuencias negativas que pueden tener estas prácticas, para fortalecer y mejorar las evaluaciones que realizan los profesores.

La autora comenta también que debe considerarse el caso de los profesores que utilizan datos de evaluación para modificar su enseñanza y dar con ello un ejemplo importante a los estudiantes. Al hacerlo, utilizan dos tipos de evaluación: la formativa, centrada en lo que puede hacer el estudiante para mejorar y una evaluación paralela en la que el docente se pregunta si los estudiantes han tenido una oportunidad adecuada para aprender. Así, los profesores que reflexionan sobre su práctica utilizan datos en forma sistemática para hacer juicios sobre los aspectos específicos de las estrategias que quizá estén obstaculizando el aprendizaje. Con ello, los profesores tratan de ser justos con los estudiantes e informarles sobre cuáles serán los componentes de la calificación.

Las pruebas de medición de rendimiento o desempeño son componentes importantes para las calificaciones; sin embargo, el esfuerzo y la capacidad también suelen tomarse en consideración. Por ejemplo, los profesores de primaria utilizan evidencias y observaciones informales. En la escuela secundaria, las mediciones de rendimiento son con pruebas y otras actividades escritas que constituyen una porción mayor de la calificación. Diferentes profesores perciben el significado y la finalidad de las calificaciones en forma distinta y consideran de manera desigual los factores de desempeño positivo o negativo. En cuanto a la capacidad de los profesores para calificar a los estudiantes de manera justa y congruente con lo que enseñan, resulta limitada (Shepard, 2006).

En México, no fue sino hasta la segunda mitad del siglo XX que comenzaron a utilizarse pruebas estandarizadas para evaluar el aprendizaje. Fueron aplicadas principalmente en los niveles básicos del sistema educativo mexicano y en la educación superior para la selección de estudiantes. A partir de 1970, la Secretaría de Educación Pública (SEP) comenzó a llevar a cabo la evaluación a gran escala. A finales de esa década se aplicaron las primeras pruebas en el nivel de primaria, con el proyecto llamado Evaluación del rendimiento académico de los alumnos de 4º y 5º grado de educación primaria. A principios de la década de 1990, la evaluación a gran escala recibió un

importante impulso; la principal promoción ocurrió en 1992 con el Acuerdo Nacional para la Modernización de la Educación Básica, del que se derivó la descentralización del sistema educativo y el programa de carrera magisterial. Para asignar estímulos a los docentes se decidió tomar en cuenta los resultados de los alumnos; por lo que se procedió a evaluar año con año a los estudiantes de México. Al iniciar el siglo XXI la evaluación educativa en México avanzó con la creación del Instituto Nacional para la Evaluación de la Educación (INEE) y en la SEP se desarrollaron las pruebas de aplicación censal denominadas Exámenes Nacionales del Logro Académico en Centros Escolares (ENLACE) (Martínez, 2009).

En cuanto a la formación docente, en el país los planes y programas de estudios para la formación inicial de profesores están enfocados hacia contenidos teóricos que provocan una diversificación en las interpretaciones por parte de cada profesor. En consecuencia, los futuros profesores se encuentran limitados en las competencias que son necesarias para enfrentar situaciones cotidianas dentro del salón de clase (Martínez, Maya y Zenteno, 1996). En general, los docentes no tienen conocimiento sobre la evaluación del aprendizaje o la psicometría porque no fueron capacitados en ello desde su educación normal, lo cual puede ser constatado en el plan de Estudios de Licenciatura en Educación Primaria 1997 y en el Plan de Estudios de Educación Secundaria 1998 (Dirección General de Educación Superior para Profesionales de la Educación [DGESPE], 2009). (Para observar con detalle la escasa formación en aspectos evaluativos que reciben los profesores durante su educación normal, véanse los mencionados planes de estudios que se incluyen en los anexos 1 y 2).

Por otra parte, desde la década pasada los profesores de educación básica en servicio reciben una oferta de cursos denominada Programa Nacional de Actualización Permanente de Maestros de Educación Básica en Servicio (PRONAP), el cual es coordinado por la Dirección General de Formación Continua de Maestros en Servicio, de la SEP. Sin embargo, no han sido suficientes las actualizaciones para crear profesores calificados y lograr condiciones satisfactorias dentro del salón de clase (Lozano, 2007). En general, los

profesores desarrollan su práctica en una cultura escolar con poca innovación, donde solo cumplen los objetivos que la actualización requiere (SEP, 2007).

Además, los docentes en servicio difícilmente tienen una actualización relacionada con la teoría de la medida o conocen los estándares de calidad técnica a los que debe ajustarse la elaboración de los exámenes, como los propuestos en Estados Unidos por la *American Educational Research Association* (AERA), la *American Psychological Association* (APA) y el *National Council on Measurement in Education* (NCME,), (AERA, 1999) y en México por el Consejo Asesor Externo del CENEVAL (CENEVAL, 2000). Por otra parte, en México no existen compañías que elaboren instrumentos de evaluación de calidad técnica que puedan utilizar los profesores en el salón de clase para evaluar el dominio de los contenidos de las asignaturas incluidas en currículum de la educación primaria o secundaria.

Por estar inmersos en el proceso enseñanza-aprendizaje que se da cotidianamente en las aulas, solo los profesores bien preparados pueden llevar a cabo una evaluación relevante de cada alumno. Esta evaluación incluye aspectos del currículo y los niveles cognitivos más complejos, que tome en cuenta las circunstancias de cada niño, y se haga con la frecuencia necesaria para ofrecer realimentación oportuna para que el alumno pueda mejorar. Este tipo de evaluaciones son las que deben hacerse en cada aula regularmente (Martínez, 2009).

En cuanto a las formas de evaluación que utilizan los profesores, principalmente incluyen pruebas de papel y lápiz, participaciones individuales y por equipos, investigaciones, ensayos y reportes, entre otras. Es decir, con base en su experiencia desarrollan actividades de evaluación (Noguez, 2000). Los profesores de educación secundaria utilizan métodos de evaluación tradicionales que conceden una calificación alta al examen y provocan un descuido en los procesos que desarrollan los productos. Por su parte, los alumnos atribuyen su fracaso a los exámenes y a los errores del profesor, con lo que ponen en duda su capacidad como evaluador (Moreno, 2007).

Como se observa, la literatura expone que existe una preparación deficiente en evaluación por parte de los docentes. Por ello, habrá que ofrecerles apoyos apropiados

para que cumplan adecuadamente con su función evaluativa, a fin de que puedan desarrollar pruebas eficaces. En este contexto, surge el problema de que los profesores dan seguimiento, realimentan y certifican el aprendizaje de sus estudiantes a partir de las observaciones que hacen de ellos en el aula de manera cotidiana y a partir de las evaluaciones que ellos mismos desarrollan. No obstante, es probable que tales prácticas no resulten del todo justas, válidas y confiables, pues los profesores no fueron capacitados para hacerlas de manera apropiada.

Lo anterior justificó una intervención para apoyarlos, de manera que puedan mejorar la calidad de las pruebas que desarrollan como parte de su trabajo cotidiano. La idea en este trabajo fue que los docentes pudieran adquirir conocimientos elementales sobre la teoría clásica de los tests (TCT), que les permitieran saber si las preguntas de tipo objetivo que formulan a sus alumnos tienen la dificultad apropiada, si permiten discriminar a quienes dominan los contenidos de quienes no los dominan, si son justas en el sentido de no presentar sesgos; entre otros aspectos elementales de tipo psicométrico (Verhelst, 2004).

Sin embargo, aunque era deseable que los profesores conocieran estas nociones básicas, ello resultaba una tarea bastante compleja; porque incluso los aspectos básicos de la psicometría están alejados de la formación regular que tienen los profesores en servicio, dado los requisitos conceptuales y metodológicos que están implicados en las teorías psicométricas, como es el caso de la teoría de la medida o el uso de la estadística.

No obstante, se consideró que un conocimiento elemental de tales aspectos podía ser suficiente si se complementaba con un enfoque que permitirá analizar de forma simple, directa y visual la manera en que los estudiantes responden a las preguntas de examen de tipo objetivo que les formula el profesor. Para lograrlo, se consideró que un enfoque que presenta tales características, es la técnica del Análisis Gráfico de Ítems (AGI) propuesta por Batenburg y Laros (2002).

1.2 Objetivos

El estudio que se llevó a cabo tuvo como objetivo general desarrollar, operar y evaluar un módulo para capacitar a docentes en servicio a fin de que puedan mejorar sus exámenes de respuesta seleccionada mediante el análisis gráfico de ítems. De manera particular, en el trabajo se buscó preparar los siguientes elementos para integrar el módulo:

- Diseñar e implementar un curso – taller con los materiales necesarios (manual, formatos, ejemplos) para capacitar a docentes en servicio, a fin de que obtengan conocimientos y habilidades de psicometría básica.
- Desarrollar un programa de cómputo que permita a los docentes conocer y mejorar la calidad técnica de los ítems de opción múltiple o de respuesta alterna que utilizan en sus pruebas de aula, por medio del análisis gráfico de ítems.
- Evaluar la operación del módulo de capacitación a los docentes, de los materiales de apoyo elaborados, así como del uso del programa de cómputo desarrollado.

De los objetivos antes mencionados se desprendieron varias preguntas de investigación que orientaron el desarrollo del trabajo de tesis:

- ¿El aprendizaje de la técnica del análisis gráfico de ítems habilita a los profesores para evaluar la calidad técnica de los ítems que usan en sus pruebas?
- ¿La capacitación propicia que los profesores aprendan nociones básicas de psicometría?
- ¿La capacitación prepara a los docentes para manejar el programa de software que se desarrollará para realizar el análisis gráfico de ítems?
- ¿Los profesores aprenden de manera fácil y significativa el uso del programa de software?

Se consideró que era factible llevar a cabo la investigación propuesta. Para tal efecto se necesitaba un programador que tuviera por función elaborar el software para efectuar el análisis de ítems, con base en las especificaciones de los algoritmos matemáticos necesarios para realizar el análisis psicométrico visual de las respuestas al ítem, mismo que le serían proporcionados por la autora.

Además, era necesario planear y desarrollar un curso – taller formal que resultará una guía útil para acercar a los docentes al campo psicométrico. Para que el curso se impartiera era necesario elaborar previamente un manual y otros materiales de apoyo necesarios para que los docentes se familiarizaran con la aproximación al análisis gráfico de ítems.

También se consideró que era posible tener acceso a quienes participaran en el estudio, mediante una convocatoria dirigida a docentes en servicio que aplicarán exámenes de opción múltiple en sus evaluaciones en el aula. Para lograr lo anterior se consideró necesario llevar a cabo una gestión para negociar el acceso a dicha población.

Todos esos requisitos pudieron satisfacerse, por lo que fue posible realizar el trabajo propuesto. De ello se da cuenta a lo largo de la presente tesis. En particular, en el capítulo II se describen los aspectos conceptuales que dieron fundamento a las acciones metodológicas que se describen en el capítulo III, como es la elaboración del programa de cómputo, del manual y de los formatos e instrumentos necesarios para la capacitación de los docentes y para evaluar la operación de la capacitación. En el capítulo IV se comentan los resultados que se obtuvieron en cuanto a los productos desarrollados y la evaluación de la capacitación. Finalmente, en el capítulo V se discuten las aportaciones y retos que surgieron del trabajo y se sugieren formas de mejorarlo mediante un estudio complementario.

Como se verá más adelante, el presente estudio tuvo varias limitaciones; pues si bien fue posible acercar a los docentes a los principales aspectos psicométricos básicos, las técnicas del análisis gráfico de ítems incluyen otros aspectos complementarios como el análisis de sesgo en los ítems y la construcción de versiones paralelas de una prueba, mismas que no fue posible incluir en este trabajo, principalmente por las condiciones en que se hizo el trabajo y por razones temporales.

En este capítulo se presentan los principales aportes de la Teoría Clásica de los Tests (TCT) al trabajo de tesis; en particular se describen de manera breve los conceptos y procedimientos básicos de dicha teoría que son relevantes para el análisis de los ítems. Por otro lado, se abordan los supuestos y procedimientos del desarrollo reciente de la TCT denominado Análisis Gráfico de Ítems (AGI), técnica que puede ser utilizada para ayudar a docentes en servicio a elaborar un ítem con buena calidad técnica y reconocer cuándo y por qué no la tiene. También se describen brevemente aspectos conceptuales de la interfaz humano – computadora, así como diversas aplicaciones para computadora que permiten efectuar un análisis gráfico de ítems.

2.1 Teoría clásica de los tests

La TCT ha sido utilizada durante más de sesenta años y ha servido de guía para los constructores de pruebas. El propósito principal de dicha teoría es detectar y eliminar el error de medición que se presenta a lo largo del desarrollo de una prueba. Para tener una visión global de las características principales de la TCT, Verhelst (2004) propone abordarla a partir de sus conceptos y procedimientos básicos.

Por otra parte, Batenburg y Laros (2002), proponen un desarrollo reciente de la TCT que permite a los constructores de pruebas de opción múltiple o de respuesta alterna crear ítems con calidad técnica para evaluar la habilidad del alumno en una prueba; este método es conocido como Análisis Gráfico de Ítems (AGI).

Por lo visto, la TCT y el AGI manejan algunos principios semejantes lo cual permite crear ítems con calidad técnica: la primera aclara los aspectos teóricos involucrados y la segunda, muestra de manera visual cómo puede estimarse la calidad técnica de un ítem.

2.1.1 Conceptos básicos

Principio fundamental

Según Verhelst (2004), la TCT considera que tras aplicar y calificar una prueba, se conoce solo el puntaje que obtuvo en ella el examinado, es decir su puntuación observada (X). Pero el nivel real de habilidad del examinado o puntuación verdadera (T) no se conoce.

Además, la puntuación observada está influenciada por el error de medición (E) que es inevitable. Escrito de manera formal:

$$X = T + E$$

Con esto se quiere decir que el puntaje observado (X) en una medición, es igual al puntaje verdadero (T), más el error (E) de medición. Como existen dos incógnitas en la ecuación (T y E), la TCT ha desarrollado algunos supuestos básicos para resolverla:

Por ejemplo, un supuesto considera que el nivel real de habilidad del sujeto es la media de los valores que se obtendrían de forma empírica en caso de administrarle el mismo test, en idénticas condiciones de medida, un número infinito de veces. Es decir:

$$T = X - E$$

Otro supuesto es la independencia de las puntuaciones verdaderas y los errores de medida. Es decir, que no existe una relación entre la puntuación verdadera de los examinados y sus errores de medida.

Cualquier medida tiene una parte verdadera y otra falsa, porque siempre está presente el error. Aunque no se puede medir el puntaje verdadero (T), a partir del puntaje observado (X) es posible efectuar una estimación sobre el valor de la puntuación verdadera. Así, la puntuación verdadera es la esperanza matemática de la puntuación observada, es decir, T = E (X). A partir de los puntajes en X (observados) y bajo ciertos supuestos, es posible efectuar una estimación probabilística razonable sobre el valor de la puntuación verdadera.

La TCT supone también que el puntaje de error promedio en la población de examinados es igual a cero. Además, la magnitud del error es independiente del tamaño de la magnitud del objeto.

Otros conceptos importantes de la TCT que comenta Verhelst (2004) son:

- **Ítem**. Los ítems son comúnmente identificados como preguntas, reactivos o tareas evaluativas. Por ejemplo, una prueba de lectura consiste en cinco pasajes de texto y cuatro preguntas deben ser contestadas sobre cada pasaje. Se concibe a las veinte preguntas como veinte ítems; pero también pueden ser contestadas preguntas asociadas con cada texto con un solo ítem.

- **Puntaje observado (X).** Es el puntaje obtenido de la aplicación de una prueba en donde el resultado es resumido en un número (por ejemplo, respuestas correctas en la prueba).

  Este puntaje es la suma de los puntajes en los ítems y son cantidades básicas que se incorporan al análisis. Es decir, al momento de calificar los ítems se acostumbra otorgar un punto a la respuesta correcta en un ítem de opción múltiple y ningún punto a una respuesta errónea. Sin embargo, también se pueden otorgar 2 puntos a una opción particular y 1 punto para otra opción (no óptima) y cero puntos para las opciones restantes; esto es conocido como ítem de crédito parcial.

  Todas las observaciones realizadas durante la aplicación de una prueba son necesarias para desarrollar diversas formas de calificar y obtener resultados numéricos de ítems; por ejemplo, en preguntas de opción múltiple, observar cuál opción eligió primero el examinado. En preguntas abiertas es preferible desarrollar una categorización detallada; por ejemplo, en una rúbrica mencionar las características de cada rasgo a calificar y observar en qué ítem se obtiene una respuesta correcta.

- **Puntuación verdadera (V).** Es la habilidad que tiene una persona al momento de contestar una prueba. Por ejemplo, en una segunda administración de una prueba a la misma persona, en circunstancias similares a la primera, ocurre una distinción entre el puntaje obtenido en la primera administración respecto de la segunda. Lo anterior posibilita obtener una distribución de (posibles) puntuaciones en la prueba relacionadas con una sola persona, que puede considerarse como una distribución privada. Esta distribución se representa con la letra V y se refiere a la puntuación verdadera de la persona.

  Así, esta puntuación verdadera es definida como una muestra de la distribución privada. En el contexto psicométrico se simboliza con la letra griega τ (tau) y se trata de un número. La puntuación verdadera no está relacionada con términos como puntuación ideal o la puntuación que se merece una persona.

- **Error de medida (E).** Es la diferencia entre la puntuación observada y la puntuación verdadera. Existen las siguientes dos formas de identificar si el error es positivo o negativo:
  1. Si la puntuación observada es mayor que la puntuación verdadera, decimos que el error de medida es positivo.
  2. Si la puntuación observada es menor que la puntuación verdadera, decimos que el error de medida es negativo.

  Aunque el error de medida de una persona tampoco se conoce, también podemos suponer que al aplicar la prueba repetidamente obtendremos una distribución de los errores de medición que tendrá una media de 0, pues los errores positivos y negativos se anularán mutuamente. El símbolo usado para el error de medida es E.
- **Variabilidad.** Se puede definir con ayuda de un ejemplo: si se aplica una prueba y un gran número de examinados obtiene un puntaje máximo, entonces no hay mucha variabilidad en los puntajes y se piensa que la prueba es fácil. Pero si ellos muestran variabilidad, que es un fenómeno recurrente en una prueba de ensayo de calibración, ocurre una variación en los puntajes. Sin embargo, la variabilidad es parte del cuadro de la desviación estándar. Dicho en palabras estadísticas, la variabilidad se construye con cada valor restado a la media de las puntuaciones y es definida como el valor medio cuadrático de las diferencias de cada valor, positivo o negativo, restado a la media (Varianza).
- **Fuentes de varianza.** Se distinguen dos tipos de varianza: de los puntajes y de los errores de medida. Tal es el caso, que Juan tiene un puntaje observado de 18 y María tiene 20; difieren las puntuaciones verdaderas debido al error de medida. No podemos separar a nivel individual para conocer el grado de error en el que se encuentran, pero pueden ser distinguidos a nivel población. Ello es así, porque en la población la puntuación verdadera no es un número sino una variable. Así, la varianza de las puntuaciones observadas es la suma de la varianza de las puntuaciones verdaderas y la varianza de los errores de medida. Dicho formalmente:

$$V_X = V_V + V_E$$

Las fuentes de varianza se identifican con nombres cortos, como: varianza observada en lugar de varianza de las puntuaciones observadas; varianza verdadera en vez de varianza de las puntuaciones verdaderas; y varianza de error en lugar de varianza de las puntuaciones de error.

- **Confiabilidad de las puntuaciones de los test.** La confiabilidad de las puntuaciones de un test se refiere a la proporción de la varianza verdadera respecto de la varianza observada. Es multiplicada por 100 y se interpreta como el porcentaje de la varianza observada que es la varianza verdadera. El valor mínimo de la confiabilidad es cero. Cuando el valor máximo es 1, significa que no hay error de medida. Por ejemplo, un coeficiente de confiabilidad de 0.8 significa que el 80% de la varianza del puntaje observado tiene variación en las puntuaciones verdaderas y el 20% es debido al error de medida. Para expresarla se denomina confiabilidad de una prueba, pero no es correcto; debe entenderse como confiabilidad de las puntuaciones de la prueba.

2.1.2 Procedimientos básicos

Verhelst (2004) comenta que hay dos procedimientos fundamentales de la TCT:

- **Determinación de la dificultad del ítem.** Es importante tener en cuenta la dificultad de los ítems ya que en las respuestas aportadas por los estudiantes no se proporciona su nivel de habilidad. Para ello, la inclusión o exclusión de ítems está basada en su grado de dificultad, también llamado valor *p* (que significa la proporción o probabilidad de acertar el ítem).

  Para ítems binarios o dicotómicos (aquellos que son calificados como correcto (1) o incorrecto (0)), el valor *p* de un ítem es la proporción de respuestas correctas en la población examinada. Se considera el valor *p* como una propiedad del ítem respecto a una sola población. Por ejemplo, se desarrolla un ítem para una prueba de cuarto grado de español que será aplicada. Se asume que el ítem es fácil porque tiene un valor *p* de 0.80, lo que quiere decir que el contenido es comprendido con

facilidad. Incluso puede ser contestado por la población de segundo grado, donde su valor *p* es de 0.25. Así, siempre se considera el valor *p* en referencia a alguna población. La dependencia del valor *p* respecto a la población de la muestra tiene implicaciones al establecer propiedades psicométricas a partir de observaciones en una muestra representativa de la población.

Respecto a la dificultad se pueden hacer tres observaciones:

1. Los valores *p* son valores que pertenecen a ítems en alguna población y se analizan en una muestra que no es igual que la población. Si se calcula el valor *p* de un ítem en dos muestras independientes, se encontrará dos valores diferentes. El valor *p* encontrado en la muestra es considerado como una estimación en la población; para ello, la exactitud depende del tamaño de la muestra.
2. Los ítems que pueden obtener 0, 1 o 2 puntos, o 0, 1, 2 o 3 puntos, etc., son conocidos como ítems de crédito parcial, y en ellos sus valores *p* son definidos como el puntaje relativo promedio.
3. El valor *p* se interpreta como medida de dificultad; sin embargo, mientras más alto sea el valor *p* el ítem es más fácil. Por ello es conocido también como índice de facilidad.

- **Determinación de la discriminación del ítem.** El índice de discriminación de un ítem permite separar los niveles de habilidad alta de los niveles bajos, con base en las respuestas al ítem. En términos psicométricos define cuál es la calidad psicométrica de una prueba que tiene solo este ítem particular. Por ejemplo, supóngase que se usa en una prueba un ítem dicotómico o binario bastante difícil; diremos que el ítem discrimina bien si los estudiantes muy buenos tienen el ítem correcto, y los otros no; pero puesto que un ítem binario tiene sólo dos categorías (correcto o incorrecto), si el ítem separa a los muy buenos de los otros, no puede separar a los estudiantes de habilidad media de los débiles. Por ello, se considera a la discriminación como una propiedad local de la prueba, que es muy difícil representar con un solo número. Para ello, en la TCT existen varios índices de discriminación:

1. El índice de discriminación de grupos contrastados, que es igual a la diferencia entre el valor *p* para los grupos con calificaciones más altas (por ejemplo el 27%) y más bajas (digamos el 27%).
2. La correlación entre la puntuación en el ítem y la puntuación en la prueba (la correlación ítem – test, también llamada correlación punto biserial).
3. La correlación entre la puntuación en el ítem y la puntuación en la prueba, con ese ítem excluido (correlación ítem – resto).
4. En particular, para los ítems de opción múltiple, la correlación entre la puntuación en la prueba y cada uno de los distractores.

   Las correlaciones ítem – test e ítem – resto deben ser positivas; las correlaciones entre la puntuación en la prueba y los distractores deben ser negativas. Sin embargo, tratar de establecer un valor mínimo para las correlaciones ítem – test e ítem – resto puede ser desorientador, porque las correlaciones están influenciadas por el valor *p* del ítem; por lo que la discriminación es una propiedad local.

   La correlación punto biserial está compuesta por un coeficiente de correlación que debe ser positivo; esto quiere decir que los estudiantes que seleccionaron la opción correcta deben obtener los puntajes totales más altos. En cambio, la correlación de la puntuación total con los distractores debe ser negativa; ello significa que los estudiantes que seleccionan respuestas incorrectas en el ítem son quienes obtienen un puntaje total bajo en la prueba. Este hecho, diferencia a los examinados que dominan el contenido de los que no lo dominan (Batenburg y Laros, 2002).

   Para calcular el coeficiente de correlación se necesitan dos series de puntajes. Por ejemplo, al calcular la correlación ítem – test se obtienen dos puntuaciones: una puntuación es el puntaje obtenido por los examinados en la prueba y la otra puntuación es la obtenida por los examinados en el ítem, misma que es igual a uno si la respuesta es correcta y es cero si la respuesta es incorrecta. La correlación se calcula con la ecuación de la correlación producto – momento de Pearson. Cuando

el valor *p* del ítem observado es cero o uno, la correlación fallará porque en estos casos la varianza de la puntuación del ítem es cero.

Coeficiente de correlación entre la puntuación en la prueba y cada distractor

Para calcular la correlación entre un distractor y el puntaje en la prueba, se recodifican las respuestas dadas en la prueba por los examinados. Suponiendo que el ítem bajo estudio, es uno de opción múltiple con cuatro opciones de respuesta (A, B, C y D), y la opción B es la respuesta correcta; esto significa que se da un puntaje en el ítem de uno a cada examinado que escogió B, y un puntaje de cero a los demás. Para calcular la correlación entre el puntaje en la prueba y el distractor A, se debe crear una variable binaria nueva, y asignar un puntaje de uno a cada examinado que escogió A y poner un cero a los demás. La correlación que se busca es la correlación entre una nueva variable y la puntuación obtenida por los examinados en la prueba. Para calcular el coeficiente de correlación entre el puntaje en la prueba y los distractores C y D, se procede de manera similar. Cuando empleamos ítems de opción múltiple, resulta apropiado calcular las correlaciones entre los distractores y el puntaje en el test. En los ítems bien construidos las correlaciones deben ser negativas.

Por eso, es necesario almacenar las observaciones originales o puntajes brutos. Si uno guarda sólo los puntajes en el ítem (los ceros y unos), no es posible calcular la correlación entre cada distractor y el puntaje en el test, pues es imposible saber cuál de los distractores fue escogido, sólo a partir del conocimiento de que la respuesta no es la correcta.

2.2 Análisis Gráfico de Ítems

El despliegue de figuras como tablas, gráficos, diagramas y mapas tiene un papel importante en el diseño y presentación de materiales de instrucción en la educación. El creciente uso de entornos de aprendizaje basados en la informática se ha convertido en un importante campo donde la presentación visual de la información es importante como medio de comunicación. Sin embargo, a pesar del hecho de que la investigación sobre el

aprendizaje y la cognición ha avanzado en las últimas dos décadas, la comprensión de los gráficos es un área que ha sido poco explorada (Schnotz, 1994).

En el campo de la psicometría han empezado a surgir nuevos métodos para realizar un análisis visual de manera rápida, tanto en la construcción de un test como en el análisis de sus ítems. Wainer (1989) considera que este análisis es dinámico y puede proveer de ventajas, como permitir realizar una inspección visual rápida por medio de una interacción computarizada. Dicho autor propuso representar las respuestas correctas ante el ítem mediante una línea de trayectoria (*Trace line*), también llamada curva característica del ítem en la teoría de la respuesta al ítem.

Haladyna (2004) considera que el resultado del análisis de los ítems de una prueba puede ser presentado de manera gráfica o numérica, y que cada método tiene su ventaja. Una representación gráfica permite obtener gran cantidad de información sobre el ítem de manera rápida. En este método, el desarrollador del test observa qué tan probable es que los examinados contesten de manera correcta el ítem; y también puede observar la variación de la tendencia de la respuesta correcta con respecto a las puntuaciones en la prueba. Por ejemplo, en un ítem de opción múltiple una gráfica muestra qué tan probable es que los examinados elijan cada distractor y al mismo tiempo permite observar cómo varía esta tendencia según las puntuaciones totales en la prueba. Comenta también que autores como Wainer (1989) y Thissen, Steinberg y Fitzpatrick (1989, citados por Haladyna, 2004) favorecen el uso de líneas de trayectoria para el análisis de ítems, pues lo hacen más significativo e interpretable. Es decir, ofrecen una apariencia palpable y reveladora que puede ser fácilmente comprendida por aquellos escritores de ítems, ajenos a la psicometría, a quienes les faltan los antecedentes estadísticos necesarios para interpretar los índices tradicionales de los parámetros del ítem. En consecuencia, el Análisis Gráfico de Ítems (AGI) posibilita también una mejor comprensión de la TCT y de la construcción de pruebas (Batenburg y Laros, 2002).

En cuanto a la representación numérica del análisis de los ítems, Haladyna (2004) comenta que las estadísticas del ítem permiten que los desarrolladores de pruebas establezcan especificaciones para la dificultad, la discriminación y la adivinación.

2.2.1 Determinación de la calidad del ítem mediante el AGI

El AGI es un método para evaluar la calidad de los ítems que es fácil de entender y no requiere de programas complejos para estimar los parámetros de los ítems (Batenburg y Yurdugül, 2006). Mediante este método, las propiedades de los ítems como su dificultad, discriminación y la tasa de adivinación se definen y se estiman directamente sobre la base de gráficos (Batenburg, 2006).

Supuestos básicos

El Análisis Gráfico de Ítems se basa en la suposición de que los desarrolladores de ítems son conscientes del contenido cuyo dominio será evaluado. Ello significa que al elaborar un ítem de buena calidad, éste debe poseer una y sólo una respuesta correcta y que por otra parte, los distractores deben ser atractivos pero no confusos para los examinados.

Lo anterior implica que los examinados que obtengan una puntuación máxima en la prueba tienen alta probabilidad de seleccionar la opción correcta en cada ítem y que los examinados que tienen bajas calificaciones tienen alta probabilidad de seleccionar una opción incorrecta. En otras palabras, la elección de opciones falsas disminuye progresivamente con el aumento de la puntuación total en la prueba.

En el AGI, el análisis de la discriminación del ítem se realiza considerando la inclinación (pendiente) de la curva de la respuesta correcta. De manera visual se define como la línea que permite discriminar entre los examinados en un rango de posibilidades, hasta un máximo de uno; esta línea es llamada intervalo de discriminación. Si la proporción de respuestas correctas incrementa con la puntuación total, el ítem tendrá un poder discriminativo alto; de lo contrario éste no se presentará. En el método gráfico, las proporciones de las respuestas se representan gráficamente respecto al puntaje total (Rodrígues, 2006).

El AGI, también es un método exploratorio que permite identificar ítems problemáticos y con ello proporciona una ayuda para que los constructores de pruebas puedan mejorar sus ítems. El AGI permite la identificación de ítems de buena calidad; y

son aquellos cuyos porcentajes de respuestas correctas aumentan a medida que la puntuación incrementa. Por lo contrario, los porcentajes de respuestas ante las alternativas falsas disminuyen cuando la puntuación aumenta (Oliveira, 2007). En consecuencia, el AGI muestra la relación entre la puntuación total en una prueba y la proporción de respuestas correctas y falsas de un ítem de opción múltiple que es utilizado para evaluar el dominio del contenido de un programa educativo (Batenburg y Laros, 2002).

Los constructores de pruebas deben desarrollar ítems de opción múltiple con alta calidad. Dicha calidad en el ítem es expresada principalmente por su poder discriminativo; es decir, su capacidad para distinguir a los examinados que tienen habilidad en un ítem de los que no la tienen (Batenburg y Laros, 2002).

Curvas de respuestas al ítem

Batenburg y Laros (2002) señalan que se requieren pruebas de logro confiables, válidas y sin sesgo para evaluar la calidad de los sistemas educativos y evaluar la eficiencia de los programas educativos. Para ello, es necesario contar con tests de alta calidad en el área educativa para la selección de alumnos, su promoción de grado o para otorgar una calificación dentro del salón de clase.

Para los constructores de pruebas, el desarrollar un ítem de opción múltiple de alta calidad requiere que el número de examinados que eligen la respuesta correcta incremente con el aumento del puntaje total; y por lo contrario, la proporción de ellos que elige cada opción incorrecta debe disminuir con un incremento del puntaje total. En la Figura 2.1 se muestra un ejemplo teórico de un buen ítem. La curva de la respuesta correcta se encuentra señalada con un asterisco en la parte superior derecha. Es importante observar cómo la proporción de respuestas correctas incrementa conforme aumenta la calificación. Por otro lado, puede observarse que la elección de los distractores tiende a disminuir conforme aumenta la calificación.

Cabe señalar que, en diversos estudios realizados con el AGI, la habilidad ha sido representada en el eje de las X de diferentes maneras. En la Figura 2.1 la habilidad se

representa como los puntajes totales que obtuvieron los examinados en la prueba, como lo proponen Batenburg y Yurdugül (2006). Sin embargo, distintos investigadores proponen representar la habilidad mediante la organización de los puntajes para formar grupos de habilidad. Enseguida se presentan las principales formas de organización que se han propuesto:

- Haladyna (2004) representó la tendencia o probabilidad de elegir la respuesta correcta cuando aumenta la habilidad de la persona, organizando las puntuaciones obtenidas por los examinados en deciles, para formar una de escala de habilidad del 1 a 10; en ella, el 1 representa al grupo de menor habilidad y el 10 al de mayor habilidad. Esta organización produce curvas de la respuesta correcta y de los distractores muy suavizadas (sin cambios bruscos) y con tendencias que pueden ser observadas con claridad.
- En el programa Lertap, Nelson (Assessment Systems Corporation, 2006) organiza de izquierda a derecha en el eje X los puntajes totales en la prueba mediante quintiles. Así, se forma una escala de habilidad del 1 que agrupa las puntuaciones más bajas, al 5 que concentra los puntajes más altos. Esto permite que los cambios en las tendencias de las curvas de la respuesta correcta y los distractores puedan ser observados todavía con bastante claridad.
- Por su parte, Batenburg y Laros (2002) proponen organizar las puntuaciones en cuartiles, para formar cuatro grupos de habilidad, de los cuales el 1 representa el grupo de puntajes de menor habilidad y el 4 al de mayor habilidad. Con ello, tanto las tendencias de las curvas de la respuesta correcta y de los distractores resultan claras aún.

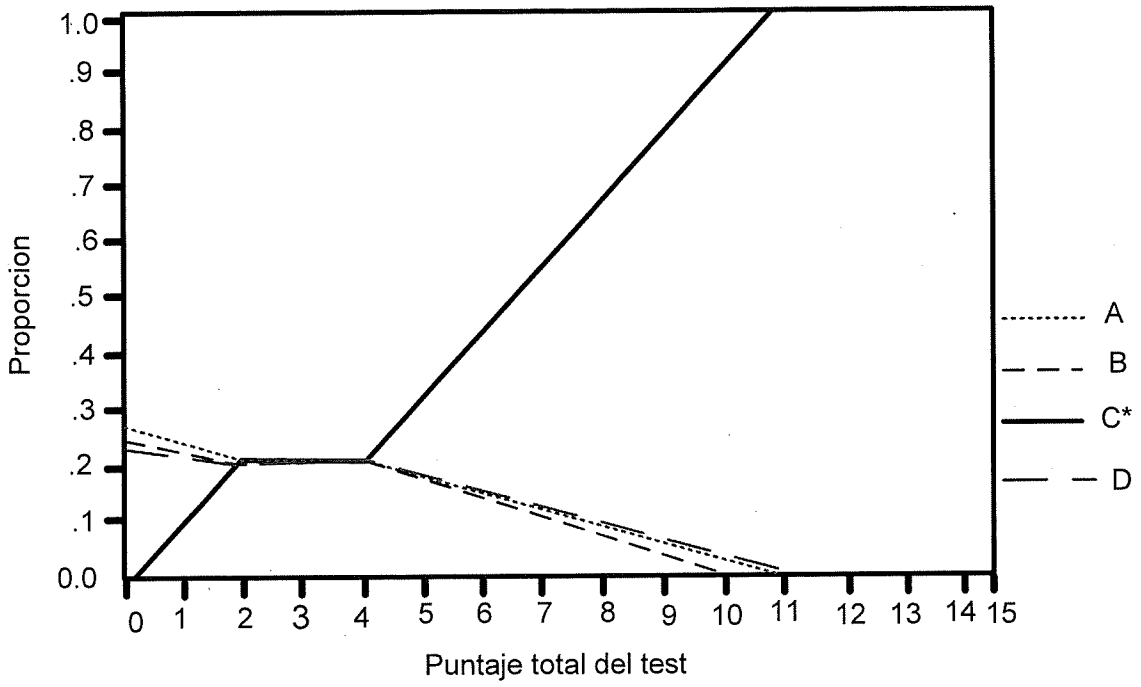


Figura 2.1 Ejemplo teórico de un buen ítem en el AGI

La curva de la respuesta correcta contiene la información que nos permite determinar cuál es la discriminación del ítem. En el análisis gráfico de ítems, esto lo podemos constatar al observar la tendencia que va desarrollando. Cuando la curva de la respuesta correcta aumenta de manera progresiva, ello quiere decir que el ítem lo han contestado sólo examinados que dominan el contenido. La discriminación del ítem está determinada por la inclinación de la curva de la respuesta correcta; esto se puede identificar a simple vista.

La Figura 2.1 muestra un ítem idealmente bueno, pero debido a las diferentes tendencias que surgen a partir de las respuestas reales que proporcionan los examinados, el ítem puede presentar condiciones diferentes que determinan su calidad. En la Tabla 2.1 se caracterizan nueve condiciones que puede presentar un ítem fácil, de dificultad media o difícil; que no discrimina adecuadamente, que apenas discrimina lo suficiente o que discrimina bien.

Tabla 2.1 Condiciones que determinan la calidad de los ítems en el AGI

Dificultad de ítems	No discrimina	Discrimina apenas	Discrimina bien
Ítem fácil	Se caracteriza porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta es alto, tanto en los grupos de mayor habilidad, como en los de menor habilidad. Esto quiere decir que el ítem es muy fácil de contestar o que la respuesta correcta es evidente. En consecuencia, la pendiente de la curva, que indica la discriminación del ítem tiende a ser horizontal o poco empinada.	Se caracteriza porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta, es generalmente alto en grupos con habilidad media y alta, mientras que en los grupos con habilidad más baja es menor. Esto quiere decir que el ítem es muy fácil de contestar o que la respuesta es obvia. En este caso, la pendiente de la curva apenas es suficientemente empinada.	La curva de la respuesta correcta en el ítem indica que el número de examinados que eligieron la opción correcta visiblemente es más alto en los grupos de habilidad media alta y alta que en los grupos de habilidad media baja y baja. Esto quiere decir que el ítem es fácil pero solo para examinados con habilidad media y alta. La pendiente de la curva es claramente empinada.
Ítem con dificultad media	La curva de la respuesta correcta en el ítem deja ver que el número de examinados que eligieron la opción correcta es en general alto, pero fluctúa entre los grupos de habilidad de manera no pretendida. Esto quiere decir que el ítem es difícil de contestar para algunos grupos y para otros no; o que resulta confuso para los examinados con diferente habilidad, en donde algunos tuvieron altos puntajes en la prueba. La curva presenta altibajos, pero con una pendiente que en general tiende a ser horizontal.	La curva de la respuesta correcta en el ítem revela que el número de examinados que eligieron la opción correcta en general no es alto ni bajo y presenta variaciones en los diferentes grupos de habilidad. El ítem resulta con dificultad media porque diferentes tipos de examinados se confunden entre la respuesta correcta con uno o más distractores poderosos. Aunque en general la curva tiende a ser horizontal, su pendiente es lo suficientemente empinada.	La curva de la respuesta correcta en el ítem permite observar que el número de examinados que eligieron la opción correcta aumenta a medida que incrementa la habilidad. Esto quiere decir que el ítem es fácil pero sólo para los examinados que dominan el contenido. Por lo contrario, es difícil pero sólo para los examinados con menos habilidad. Por ello, la pendiente de la curva está claramente empinada.
Ítem difícil	La curva de la respuesta correcta muestra que el número de examinados que eligieron la opción correcta en general es bajo en todos los grupos de habilidad. Esto quiere decir que el ítem es difícil de contestar para algunos grupos y para otros no. Tal variabilidad muestra que el ítem resulta confuso para examinados con diferente habilidad, por lo que la pendiente de la curva tiende a ser horizontal.	La curva de la respuesta correcta muestra que el número de examinados que eligieron la respuesta correcta es relativamente bajo en todos los grupos, pero se observa que aumenta un poco a medida que incrementa la habilidad. Ello determina que la pendiente de la curva tenga una inclinación suficiente como para que el ítem se considere con calidad mínima.	La curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la respuesta correcta incrementa claramente a medida que aumenta la habilidad. El ítem sólo lo responden correctamente quienes tienen más habilidad. En consecuencia la pendiente de la curva está visiblemente empinada.

En la Tabla 2.1 se mencionan las principales características y tendencias de la curva de la respuesta correcta al ítem y que resultan más significativas. Sin embargo, para una descripción más detallada de esas condiciones, así como de las correspondientes curvas de respuestas a los distractores del ítem, se sugiere revisar el manual técnico para el análisis gráfico de ítems que se presenta como Anexo 3.

Criterios para eliminar o mejorar ítems

El AGI proporciona información sobre la calidad de un ítem. Así mismo, nos permite identificar a los ítems malos. Al respecto, Batenburg y Laros (2002) nos proponen razonar cuatro aspectos principales que pueden ser utilizados para determinar la eliminación o corrección de ítems:

- Tomar en cuenta el número de violaciones al supuesto de que la elección de la opción correcta incrementa al aumentar el puntaje total.
- Tomar en cuenta el número de violaciones al supuesto de que la elección de las opciones falsas disminuye con un incremento del puntaje total.
- Tomar en cuenta el número de intersecciones entre las curvas de respuestas de las opciones correcta y falsas, tras el inicio del intervalo de discriminación.
- Considerar la ausencia de poder discriminativo o una baja pendiente de la curva de la respuesta correcta.

Sin embargo, los autores también nos advierten sobre la necesidad de reemplazar los ítems que se deben eliminar, por otros que tengan calidad adecuada y que evalúen el mismo dominio del contenido que los que fueron eliminados, a fin de que el examen conserve la estructura que fue planeada desde un inicio.

Construcción de pruebas paralelas

La construcción de versiones paralelas de un examen es uno de los temas importantes de la TCT, pues el paralelismo está directamente relacionado con la confiabilidad de la prueba. La construcción de pruebas paralelas puede ocurrir en diferentes situaciones:

- Construir una prueba paralela para un examen existente (y ya utilizado).

- Construir dos pruebas paralelas (o incluso más) desde el principio.
- Una prueba existente tiene que ser dividida en dos mitades que son paralelas (para usar el método de división por mitades para estimar la confiabilidad).

En todos estos casos, el AGI aporta un método simple para construir las pruebas paralelas de manera gráfica (Verhelst, 2004). La idea es construir dos formas de la prueba que son, aproximadamente paralelas. Esto significa que cada ítem en una forma tiene una semejanza en la otra forma, con (aproximadamente) las mismas cualidades psicométricas. En la TCT se intenta empatar dos cualidades: la dificultad y la discriminación, las cuales se operacionalizan mediante el valor *p* y la correlación ítem – test (o ítem – resto).

El punto de partida del método es construir un diagrama de puntos dónde cada ítem se representa por un punto en el plano. La coordenada X es el valor *p* del ítem, la coordenada Y es la correlación ítem – test. La posición del ítem se simboliza con una etiqueta breve para cada ítem, que permita identificarlo con facilidad. Se proporciona un ejemplo en la Figura 2.2, donde dos ítems con una representación gráfica cercana una de la otra, tienen aproximadamente el mismo valor *p* y la misma discriminación. En la Figura 2.2, los pares de ítems son representados por líneas que conectan dos ítems. Se forman los pares de tal manera que la distancia entre los dos ítems, en cada par, sea tan pequeña como resulte posible.

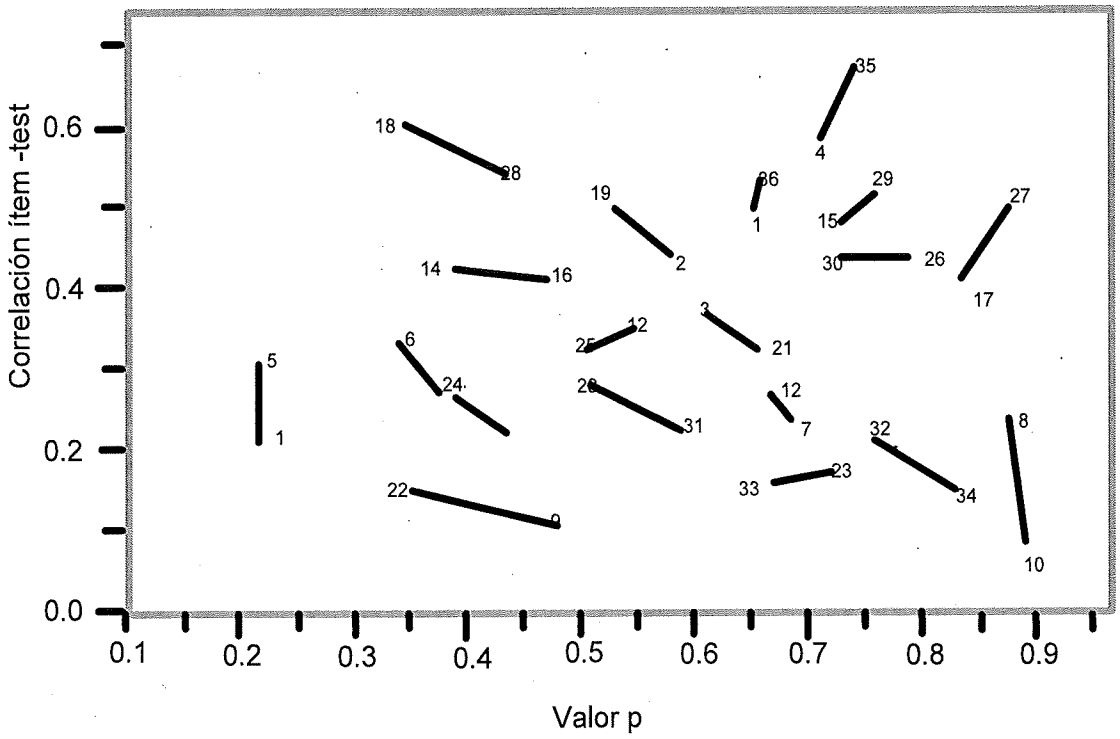


Figura 2.2 Construcción de pruebas paralelas mediante el AGI

Para construir las formas paralelas aproximadas, los dos ítems que pertenecen a un par deben asignarse al azar a las formas.

Funcionamiento diferencial de ítems

Verhelst (2004), documenta otra de las aplicaciones del AGI para el estudio de un concepto que ha sido objeto de interés durante los últimos 30 años; y es el Funcionamiento Diferencial de Ítems (DIF por sus siglas en inglés). Dicho autor considera que para que una evaluación sea justa, se requiere que un ítem se comporte de manera similar ante diversas poblaciones. Sin embargo, no resulta fácil definir qué es similar. Por ejemplo, decimos que un ítem debe ser difícil en todas las poblaciones; pero usar tal definición causará un serio problema. Si la dificultad del ítem se operacionalizara por su valor *p*, se esperaría que el valor *p* de un ítem aritmético típico fuera más bajo en la población de mujeres que en la población de los hombres (lo cual, según el autor, ha quedado establecido por muchos estudios con estudiantes de educación media). Esto

ilustra muy bien la dependencia del valor *p* respecto de la población. Esto sucederá con la mayoría o todos los ítems de una prueba de aritmética; pero si nos ajustamos al requisito de que para ser justo cada ítem debe ser igual de difícil en ambas poblaciones, entonces por necesidad encontraremos que en un test "justo", el puntaje promedio de los hombres y las mujeres debe ser el mismo. Pero esta aproximación implica que todas las diferencias son injustas, porque puede aplicarse a cualquier par de poblaciones (incluso las poblaciones que consisten respectivamente en yo y mi vecino). En consecuencia, necesitamos una mejor definición del DIF, una que dé cabida a las diferencias entre poblaciones. Tal definición se formula como una declaración condicional que será aplicada al ejemplo de los hombres y las mujeres. Un ítem no muestra DIF si en la población (conceptual) de hombres, con un nivel de habilidad arbitrario pero fijo, y la población (conceptual) de mujeres con el mismo nivel de habilidad, los valores *p* del ítem son idénticos. Nótese que esta igualdad de los dos valores *p* debe sostenerse para cada nivel de habilidad. Dicho de manera más simple: la ausencia de DIF significa que el ítem debe ser igualmente difícil para los hombres y mujeres que tienen el mismo nivel de habilidad.

En la práctica, no sabemos el nivel exacto de habilidad de cualquier examinado, pero podemos usar el puntaje en la prueba como un sustituto. Si, como se dijo antes, se agrupan los examinados en varios grupos (de tamaño razonable), de menos a más habilidad, podemos graficar de manera separada los valores *p* observados en cada grupo de hombres y mujeres. En la Figura 2.3, se dan dos ejemplos de un examen de matemáticas. La leyenda Sg = 1 se refiere al subgrupo de las mujeres; para los hombres se usa la leyenda Sg= 2. Los números 1, 2, 3, y 4 en el eje de las X se refieren a los grupos de calificaciones; así, el grupo 1 es el de menor habilidad y el 4 el que tuvo los puntajes más altos en la prueba.

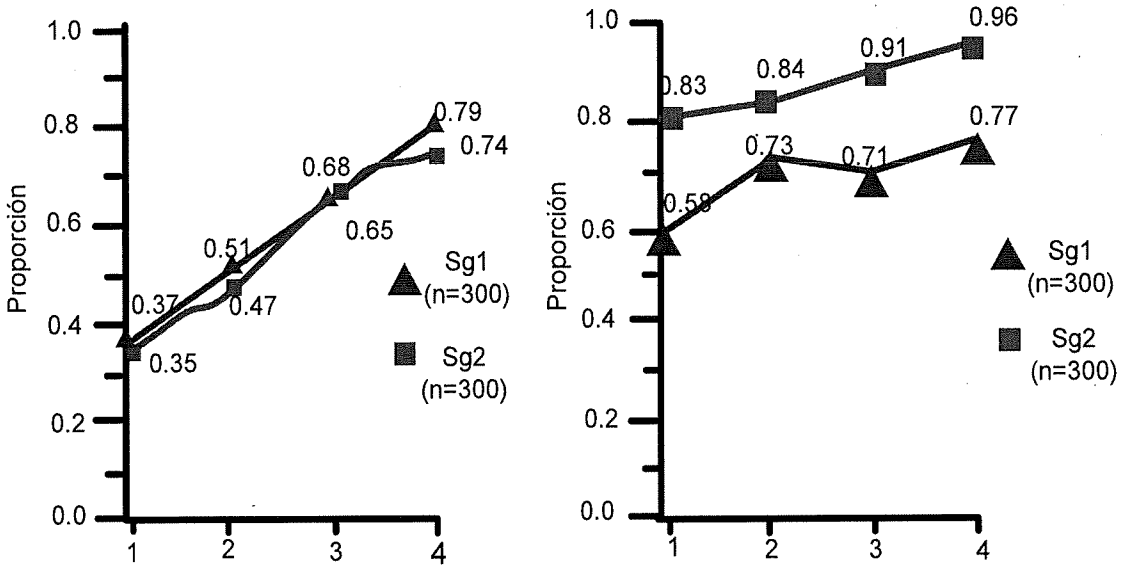


Figura 2.3 Ejemplo de análisis del funcionamiento diferencial de ítems mediante el AGI

Como puede observarse en la Figura 2.3, en el ítem de la izquierda no se observa DIF pues en cada grupo de habilidad se observa aproximadamente el mismo valor *p* en los subgrupos de mujeres y hombres. En cambio, en el ítem que se representa a la derecha tenemos evidencia de que el reactivo funcionó de manera diferencial en ambos subgrupos, pues en cada grupo de habilidad el valor *p* para el subgrupo de las mujeres es visiblemente más bajo. Es decir, se observa que el ítem sesga a las mujeres, pues es más difícil para ellas sin importar el nivel de habilidad que tengan.

2.3 Interacción humano – computadora

La Interacción Humano – Computadora (IHC) es el estudio y práctica del uso de esta herramienta; es decir, de la capacidad que tiene la computadora para ser utilizada mediante el entendimiento, la creación de *software* y otras tecnologías que serán útiles para los usuarios humanos, a fin de encontrar una solución efectiva a una situación que involucre el manejo de información (Carrol, 2002). Dicho de manera más simple, es el estudio de la relación entre los usuarios humanos y los sistemas de cómputo que sirven

para realizar tareas específicas. Como señalan Faulkner y Prentice (1998), con la IHC se busca entender el razonamiento que siguen los usuarios en las tareas que necesitan llevar a cabo y visualiza el modo en que los sistemas de cómputo necesitan estar estructurados para facilitar el cumplimiento de estas tareas. Según los autores, es necesario entender a los usuarios para comprender los procesos y capacidades que pueden asociarse a las tareas que desempeñen. Para ello, se requiere desarrollar conocimientos sobre memoria, cognición, oído, tacto y habilidades motrices. Tales aspectos se requieren a fin de mejorar el sistema computacional, en términos de lo que puede hacer el usuario y cómo se puede comunicar con la computadora. El uso de esta herramienta genera un grado de eficiencia mediante la reducción del número de errores ocasionados por el usuario durante el desarrollo de una tarea. También se logra eficiencia en términos del tiempo que el usuario se lleva en realizar su tarea y con esto se percibe una satisfacción de uso.

Por lo tanto, existen cuatro elementos que componen a la IHC (Preece, 1994):

- El uso y contexto del sistema. La organización social, cultural y de la comunidad, así como el trabajo en grupo y las áreas de aplicación del sistema.
- El usuario humano. Consiste en el procesamiento de información por parte del usuario, el lenguaje y la interacción que presente.
- La computadora. En cuanto a sus técnicas de diálogo (sesión, iniciativa o mixta) y los componentes gráficos (interfaz).
- El proceso de desarrollo del sistema. Los lineamientos del diseño y del enfoque; los casos de estudio; las técnicas de evaluación y las técnicas de implantación.

Con lo anterior, queda claro que la IHC es importante para el diseño de sistemas de calidad, por medio de la aplicación sistemática de conocimientos sobre las características y metas humanas, y sobre las capacidades y limitaciones de la tecnología. Ambas condiciones están relacionadas con los aspectos sociales, organizacionales y físicos del entorno de trabajo del usuario.

En consecuencia, la IHC relaciona a disciplinas como: psicología social y organizacional, ingeniería, diseño, psicología cognitiva, ciencias de la computación,

educación, inteligencia artificial, lingüística, filosofía y sociología, entre otras (Faulkner y Prentice, 1998).

Por ello, resulta importante el diseño de la interacción; que se refiere al proceso que se encuentra dentro de las restricciones de recursos para crear, dar forma y decidir todas las cualidades orientadas al uso, ya sean funcionales, estructurales, éticas o estéticas, de un artefacto digital de uno o más clientes (Löwgren y Stolterman, 2005).

Interfaz gráfica del usuario

Para un usuario promedio lo importante es realizar sus actividades de manera eficiente, sin tener que pasar por un entrenamiento previo sobre sistemas operativos y comandos básicos, medios y avanzados, para poder realizar algo simple (Shneiderman, 1998).

De ahí la necesidad de elementos como la *usabilidad* y *amigabilidad* de un programa de cómputo que proporcionen un fácil acceso a los usuarios. La *amigabilidad* de una interfaz se puede evaluar por consideración de factores como el tiempo de aprendizaje, la velocidad en el desempeño, la tasa de error en el uso, el tiempo del usuario, la manera en que el usuario opera el sistema y la satisfacción subjetiva que se experimenta al utilizar una computadora. En cuanto a la *usabilidad* del programa dependerá de un adecuado diseño de la Interfaz Gráfica del Usuario (GUI, por sus siglas en inglés), así como de la complejidad de la aplicación que se diseña (Shneiderman, 1998).

Para diseñar una interfaz es recomendable seguir lineamientos como los que propone Norman (1990):

- Definición de la tarea u objetivo que persigue el programa.
- Definición del usuario promedio.
- Definición de la plataforma y mecanismos de interacción o adaptación a las condiciones existentes.
- Diseño general de los elementos de interacción (en el caso de interfaces gráficas, estos elementos incluyen botones, menús, iconos) así como los elementos de navegación y su distribución general en la pantalla.
- Elaboración de un primer prototipo, para establecer las transiciones entre

diferentes momentos o módulos dentro del programa.

- Prueba del prototipo. Esta evaluación es importante y puede realizarse con recursos no digitales (guiones y representaciones en papel).
- Integración a la elaboración de código. Contando ya con un prototipo aceptable se puede proceder a la elaboración del programa.
- Evaluación de la versión preliminar. Existen procedimientos estandarizados (pruebas de usabilidad) que incluyen tanto un procedimiento experimental como la observación.
- Integración de la versión de distribución. Que el programa finalmente se distribuya no significa que haya terminado el proceso de evaluación. Debe preverse tanto soporte a los usuarios como mecanismos para obtener realimentación.

Evaluación de la interacción

La evaluación de la interfaz es un elemento importante dentro del conjunto de parámetros utilizados para evaluar el software. Cada evaluación se lleva a cabo en un contexto definido, determinado por las características de los usuarios, de las actividades que realizarán, del sistema que está siendo evaluado y el ambiente en el que se sitúa. A fin de realizar una evaluación de la interacción existen varios métodos de evaluación. Entre ellos destacan según Precee (1994):

- La observación y monitoreo mediante protocolos verbales, observación directa, registro de bitácoras y solicitud de opiniones a los usuarios mediante entrevistas y cuestionarios.
- Experimentación y establecimiento de puntos de comparación mediante experimentos, ingeniería de usabilidad, evaluaciones estandarizadas, opiniones de los usuarios y el establecimiento de prioridades y toma de decisiones con un análisis de impacto.
- Recolección de opiniones de usuarios a través del uso de encuestas contextuales, evaluación cooperativa y participativa y etnografía o evaluaciones "in situ"

- Predicción de usabilidad mediante métodos de inspección, simulaciones de uso, revisiones estructuradas con expertos y modelado de la funcionalidad del sistema.

Por ser una constante de los métodos descritos y por las características y condiciones en que se realizó el presente trabajo, se destaca la importancia de efectuar la evaluación de la interacción mediante un cuestionario en el que se solicita la opinión de los usuarios sobre el programa desarrollado.

2.4 Software para el Análisis Gráfico de ítems

Aunque el AGI es una técnica relativamente reciente, hasta el momento ha tenido un desarrollo acelerado. Varios investigadores y algunas de las principales organizaciones psicométricas y compañías que desarrollan *software* para el análisis de ítems como la *Assesment Systems Corporation* (2006), el *National Institute for Educational Measurement* (CITO por sus siglas en holandés) (2005) y otras más, han propuesto aplicaciones para efectuar análisis gráficos de ítems que forman parte de los programas de cómputo regulares que utilizan para realizar análisis psicométricos de ítems y tests. En la Tabla 2.2 se contrastan varios programas de cómputo que permiten realizar un análisis gráfico de ítems, a partir de las funciones psicométricas más comunes.

Tabla 2.2 Comparación de *software* psicométrico para realizar análisis grafico de ítems y tests

Software / Función	Lertap 5	ViSta	TiaPLUS	RUMM
Estima parámetros del ítem (*p*, D, rpbis)	✓	✓	✓	✓
Analiza funcionamiento diferencial de ítems	✓	---	✓	✓
Permite desarrollar versiones paralelas	---	---	✓	✓
Proporciona curvas de respuestas a los ítems	✓	Gráfico extendido (spreadplot)	✓	✓
Criterio para definir grupos de habilidad	Quintiles	---	Cuartiles	Lógitos
Permite adjuntar tablas de datos a la gráfica	✓	---	---	✓
Sistema de cómputo bajo el cual opera	Excel	Módulo del programa ViSta	Propio	Propio
Personas u organización que lo desarrolla o distribuye	Assessment Systems Corporation (2006)	Ledesma, Molina, Valero & Young (2002)	CITO, 2005	Andrich, Sheridan, Lyne & Luo (2000)

Como podrá observarse más adelante, las características psicométricas del programa TiaPlus (CITO, 2005) descritas en la Tabla 2.2, así como las de su interfaz gráfica del usuario, resultaron particularmente orientadoras para el desarrollo del software que se elaboró como parte del presente trabajo de tesis.

En este apartado se explican los métodos y materiales desarrollados y utilizados en este trabajo. En particular se describen los participantes en el estudio evaluativo, los instrumentos y los materiales que fueron empleados, así como los procedimientos que se siguieron para la captura de la información y el análisis de la misma.

3.1 Participantes

Los participantes en el estudio fueron profesores en servicio de educación básica, en específico del nivel de secundaria, con especialidad en matemáticas y geografía, así como dos jefes de enseñanza. Los profesores participaron en respuesta a una invitación que se les formuló a través de la Jefatura de Enseñanza Secundaria en Ensenada, Baja California. Inicialmente, se invitó a un grupo de 20 profesores para participar en la capacitación que fue diseñada para habilitarlos en el uso de la técnica del análisis gráfico de ítems, a fin de que mejoraran la calidad técnica de las pruebas de aula que desarrollan para evaluar el aprendizaje de sus alumnos. Sin embargo, hubo un número mayor de profesores interesados en participar, pero no pudieron ser atendidos por razones de espacio y de las condiciones necesarias para llevar a cabo la capacitación, por lo que finalmente se conformó un grupo de 21 profesores. Los docentes estaban adscritos a centros de trabajo de distintas modalidades de secundaria, según se muestran en la Tabla 3.1; la categoría "otra" hace referencia a dos jefes de enseñanza que participaron en el curso-taller junto con sus profesores.

Tabla 3.1 Distribución de maestros por tipo de secundaria

Tipo de escuela	Número de maestros participantes
Secundaria general estatal	2
Secundaria general federalizada	5
Secundarias técnica federalizada	12
Otra	2

Para participar en el curso los profesores debieron cumplir con los siguientes requisitos:

- Ser profesores en servicio.

- Haber aplicado previamente una prueba formada por un mínimo de 10 ítems de opción múltiple, empleada para evaluar el aprendizaje de al menos 32 alumnos de alguno de sus cursos.
- Traer un archivo de EXCEL con las respuestas brutas que dieron sus alumnos a cada uno de los ítems del examen.
- Traer una laptop con Windows XP o superior, así como haber instalado el programa EXCEL versión 2003 o superior.
- Tener interés y disponibilidad de tiempo para participar en la capacitación.

La capacitación, fue denominada "Curso – taller: Análisis gráfico de ítems por medio de la aplicación del PAGI", y se llevó a cabo en junio del 2010 en instalaciones de la Universidad Autónoma de Baja California, campus Sauzal. Ahí, se habilitó un aula con mobiliario de mesas de trabajo y suficientes contactos para que los profesores conectaran sus computadoras personales.

3.2 Materiales e instrumentos

Para llevar a cabo el curso de capacitación, se desarrollaron y utilizaron los siguientes materiales e instrumentos: un manual para la capacitación del participante; el software denominado Programa para el Análisis Gráfico de Ítems (PAGI) para efectuar el análisis gráfico de los ítems elaborados por los profesores; un cuestionario para evaluar tanto la operación general del curso, cómo la instrucción y los materiales utilizados; una presentación en PowerPoint para apoyar la capacitación de los profesores y una guía para apoyar a los participantes a fin de que pudieran realizar una interpretación adecuada de la calidad técnica de los ítems que elaboraron y aplicaron a sus alumnos. Dichos materiales e instrumentos se describen brevemente a continuación:

3.2.1 Manual del participante

Se elaboró un manual que contiene la información conceptual y metodológica, mínima y necesaria para la comprensión y utilización de la técnica del análisis gráfico de ítems (ver Anexo 3).

El manual del participante está integrado por una presentación al curso; un apartado donde se explica en qué consiste la técnica del análisis gráfico de ítems; un apartado donde se explica la estructura y manejo del programa de cómputo PAGI; y al final se presentan un glosario y bibliografía complementaria. Enseguida se describen dichos apartados:

- La presentación al curso brinda al participante una visión general sobre la manera en que los profesores en servicio han realizado tradicionalmente la evaluación del aprendizaje en el aula. Asimismo, introduce aspectos conceptuales mínimos básicos de la Teoría Clásica de los Tests (TCT) e introduce la técnica del Análisis Gráfico de Ítems (AGI). Para finalizar, hace explícito el propósito fundamental de la capacitación.
- El apartado Análisis Gráfico de Ítems, presenta al participante la técnica desarrollada por Batenburg y Laros (2002) y, en particular, los elementos que son necesarios para la interpretación apropiada de la relación entre la puntuación total en la prueba y el total de los examinados que eligieron la opción correcta en el ítem. También analiza las relaciones entre la puntuación total y las respuestas que se dieron a las opciones falsas en un ítem de opción múltiple. Ambos tipos de relaciones se analizan mediante varios ejemplos que ilustran ítems fáciles, de dificultad media y difícil; que pueden tener buena, regular o mala discriminación, y por ello tener o no calidad técnica. Por último, se muestra un conjunto de criterios que pueden ser utilizados para decidir si se eliminan o corrigen los ítems defectuosos.
- El apartado Programa para el Análisis Gráfico de Ítems (PAGI), describe la estructura y manejo de la aplicación de cómputo desarrollada y orienta al usuario para navegar por la interfaz del programa. Es decir, guía al usuario para introducir la base de datos que contiene las respuestas a los ítems de una prueba, para ejecutar el programa y para mostrar de manera visual los resultados del análisis de los ítems.

- Finalmente, se presenta un Glosario mínimo que aclara la base conceptual de ciertos términos técnicos que se utilizan en el manual, y las Referencias que se consultaron para elaborar el manual.

3.2.2 Presentación en PowerPoint

Contiene el material de apoyo que resultó necesario para capacitar a profesores en servicio. Esta presentación consta de 47 diapositivas que ilustran los principales temas desarrollados en el manual (ver Anexo 4). La presentación se estructuró para apoyar la instrucción en tres momentos del curso – taller:

- En el primer momento las diapositivas explican los aspectos conceptuales de la TCT e ilustran de manera breve la técnica del AGI.
- En el segundo momento presentan información sobre los fundamentos conceptuales y los elementos necesarios para la interpretación apropiada de la técnica del AGI, y se ilustraron nueve casos especiales para realizar la interpretación del AGI. Además, se ilustraron algunos criterios que pueden ser utilizados para corregir o eliminar ítems.
- Por último, las diapositivas ilustraron de manera general la interfaz gráfica del programa de cómputo, por medio de imágenes que permitieron familiarizar al participante con la aplicación PAGI.

3.2.3 Cuestionario para evaluar la operación del módulo

La evaluación de la operación del curso, fue un procedimiento que se aplicó una vez que los participantes completaron el entrenamiento. Dicha evaluación tuvo como propósitos proporcionar realimentación a la instructora sobre aspectos como la efectividad del entrenamiento, la actuación de la propia instructora y los materiales empleados durante la capacitación. Así mismo, se consideró dicho procedimiento como una condición indispensable para determinar el alcance de los objetivos propuestos para la capacitación, para conocer en general la calidad del curso y determinar formas de mejorar su desarrollo. Por ello fue importante recabar la opinión de los participantes.

Para lograr lo anterior, se procedió a elaborar un cuestionario (ver Anexo 5) en el cual se formularon preguntas cerradas que permitieron obtener información de los participantes mediante una lista de enunciados, respecto a los cuales el participante debía manifestar su grado de acuerdo en una escala Likert que tuvo las siguientes opciones:

a. Totalmente en desacuerdo

b. En desacuerdo

c. Ni de acuerdo, ni en desacuerdo

d. De acuerdo

e. Totalmente de acuerdo

Así, el cuestionario fue diseñado para obtener información sobre tres dimensiones importantes: la instrucción, los materiales de apoyo y la operación general de la capacitación.

La dimensión **evaluación de la instrucción** se consideró importante, porque tener a una persona que planee sus actividades, domine el tema y mantenga motivados y participando a los profesores, daría como resultado un compromiso de su parte para interactuar y llevar a cabo los objetivos planteados para la capacitación. Dentro del cuestionario se consideró necesario evaluar cuatro indicadores que directamente se relacionan con la participación de la instructora: a) planeación didáctica, el cual explora las actividades previas de planeación del curso; b) interacción, que se refiere a la promoción de la motivación y la participación para aprender el material; c) la actuación de la enseñanza, mismo que explora su capacidad comunicativa y pedagógica; y d) dominio del contenido, que explora su capacidad profesional y versatilidad para hacer asequibles los temas.

En cuanto a la dimensión **materiales de apoyo**, se consideró relevante conocer la eficacia del manual del participante como promotor de aprendizajes significativos, y si los materiales de apoyo ayudaron a comprender la información presentada. Por ello, en el cuestionario se exploraron tres elementos: a) el manual, sobre el cual se indaga respecto a la claridad y significación del material escrito y los conceptos descritos; b) el programa de cómputo, sobre el cual se averiguan la facilidad de uso y utilidad para el análisis gráfico de

ítems y c) la presentación en PowerPoint, sobre la cual se examinan su contenido y los ejemplos presentados mediante la computadora para facilitar la comprensión de los conceptos y procedimientos descritos.

Por último, evaluó la dimensión **operación de la capacitación,** en donde se consideraron dos aspectos: a) el desarrollo de las actividades, que se enfocó en las acciones desarrolladas durante toda la capacitación y b) el impacto, es decir; el aprendizaje que se logró en el curso y la utilidad percibida para mejorar los exámenes de opción múltiple en el aula.

Con lo anterior se pretendió evaluar la utilidad y la forma de desarrollar la capacitación, para así mejorar su estructura general y la información que contiene.

3.2.4 Programa para el Análisis Gráfico de Ítems

Para desarrollar el Programa del Análisis Gráfico de Ítems (PAGI) fue necesario considerar varios aspectos, a fin de adaptar la técnica del análisis gráfico de ítems a las condiciones previstas para la capacitación.

Por un lado, el AGI se encuentra configurado originalmente para examinar ítems incluidos en bases de datos de evaluaciones a gran escala. Por ello, las curvas de la respuesta correcta y de los distractores están formadas con los datos de muchos examinados, lo que da como resultado curvas respuestas suavizadas en las que se observan con claridad sus cambios de tendencia dentro de la gráfica. En cambio, la evaluación en el aula se aplica generalmente a pocos estudiantes, alrededor de 30 de ellos.

Por otro lado, como ya se señaló en el capítulo II, diversos autores representan la habilidad de distintas maneras en el eje de las X. Por ejemplo, Batenburg y Yurdugül (2006) lo hacen directamente con la puntuación obtenida por los examinados en la prueba; Haladyna (2004) la representa en deciles; Nelson (Assessment Systems Corporation, 2006) en quintiles; y Batenburg y Laros (2002) organizan las puntuaciones en cuartiles, para formar 4 grupos de habilidad. Sin embargo, para aplicarse a la evaluación

en el aula, tales formas de representar los grupos de habilidad pueden resultar poco claras y significativas dado el reducido número de casos que formarían cada grupo de habilidad.

Como el propósito del presente trabajo de tesis fue habilitar a los docentes para realizar el Análisis Gráfico de Ítems de sus pruebas de aula, para determinar cómo realizar una adaptación adecuada de la técnica del AGI a dichas condiciones, se procedió a realizar un conjunto de pequeños experimentos donde se variaron sistemáticamente los siguientes aspectos de la técnica:

- El número de grupos en que se dividieron las puntuaciones de los examinados (mismas que se consideraron indicadoras de su habilidad). Así, se varió la agrupación de las puntuaciones en 2, 3, 4 y 5 grupos de habilidad.

- El número total de examinados que respondieron el ítem. Aquí, se hizo variar el total de examinados en 30, 32, 107 y 2000 casos respectivamente.

- Los parámetros de los ítems. Se ensayaron los dos aspectos antes descritos con ítems de diferente dificultad y distinto poder discriminativo.

Para observar las diferencias entre los resultados de dichas exploraciones, se sugiere revisar los anexos 6 y 7. Sea suficiente aquí señalar que, después de revisar los resultados de esos ensayos, se optó por representar las puntuaciones en 4 grupos de habilidad: bajo (1), medio bajo (2), medio alto (3) y alto (4); tal como lo proponen Batenburg y Laros (2002). Además, se observó que con un mínimo de 32 examinados y con estos cuatro grupos de habilidad, se puede observar con claridad las tendencias de las respuestas al ítem y se distinguen también todos los demás componentes relevantes del análisis gráfico de ítems que definen la calidad técnica de un reactivo.

Una vez definidos ambos aspectos, se desarrollaron los algoritmos necesarios para representar gráficamente las curvas de respuestas correcta y a los distractores a partir de la base de datos de los usuarios, mismos que fueron entregados al especialista en cómputo que programó la aplicación del PAGI.

El PAGI no se describe aquí, pues se presenta ya con cierto detalle en el manual del participante (ver Anexo 3); además, el programa mismo aparece en el CD que se adjunta a la presente tesis, desde donde puede ser instalado.

3.2.5 Guía para apoyar la interpretación de la calidad técnica de los ítems

La guía (ver Anexo 8) se elaboró para apoyar al participante en la interpretación sobre la calidad de sus ítems, con base en los datos obtenidos tras su análisis mediante el PAGI. Este formato consistió en una tabla que permitió al participante analizar cuatro aspectos fundamentales de las curvas de respuestas al ítem:

- Observar en la curva de la respuesta correcta su pendiente o inclinación (su poder de discriminación).
- Observar en la curva de respuestas al distractor 1 su tendencia.
- Observar en la curva de respuestas al distractor 2 su tendencia.
- Observar en la curva de respuestas al distractor 3 su tendencia.

La guía ayudó al participante a determinar la efectividad de cada uno de los ítems de su prueba.

3.3 Procedimientos

Tras elaborar los materiales e instrumentos se operaron dos procedimientos generales:

3.3.1 Capacitación

La capacitación tuvo por objetivo brindar los conocimientos psicométricos mínimos necesarios para que los docentes en servicio desarrollaran con un mínimo de calidad técnica, ítems de opción múltiple que utilizan en sus pruebas.

La capacitación tuvo una duración de 4 horas y se realizó en una sola sesión. Su operación se dividió en dos etapas: en la primera se dio a conocer la introducción, la interpretación del análisis gráfico de ítems y la interfaz del PAGI; en la segunda etapa se trabajó con la base de datos de los profesores en servicio y se realizó la interpretación de cada ítem; por último se realizó la evaluación de la capacitación.

En la primera etapa de introducción, se revisaron las nociones teóricas necesarias para entender el proceso del análisis gráfico de ítems. Primero se abordaron los aspectos conceptuales básicos de la TCT y del AGI. Después se presentaron los 9 casos diferentes que muestran posibles tendencias de las respuestas que dieron los examinados ante las opciones del ítem, mismas que definen o no su calidad técnica. De esta manera se facilitó la interpretación visual de la calidad de los ítems evaluados. Lo anterior fue apoyado mediante la presentación de PowerPoint.

La segunda etapa de adoptó una modalidad de taller, por lo que estuvo enfocada a trabajar en la interpretación de los ítems de la base de datos de los profesores, con la ayuda del PAGI. Para ello, fue necesario también utilizar la guía para facilitar la interpretación. La idea en esta segunda etapa fue que, después de realizar una observación de cada ítem y con base en el nuevo conocimiento adquirido, cada profesor fuera capaz de emitir un dictamen razonado sobre cada ítem de su prueba.

3.3.2 Aplicación del cuestionario para evaluar el módulo

Después de desarrollar el módulo, se procedió a la aplicación del cuestionario para evaluar la operación de la capacitación. Para ello, los docentes trabajaron de manera individual y respondieron los ítems del cuestionario, mismos que indagaron sobre el proceso de capacitación y sobre la asesoría que se brindó a los participantes.

Los resultados del diseño de estos materiales e instrumentos, así como del análisis de los resultados de su aplicación serán tratados en el capítulo IV de la presente tesis.

En este apartado se presentan los resultados del estudio en cuatro secciones: a) los resultados de los procedimientos que se siguieron para diseñar los materiales e instrumentos desarrollados; b) los referidos al análisis psicométrico del instrumento de evaluación del módulo que se desarrolló y aplicó a los participantes; c) los resultados en cuanto a la opinión que manifestaron en el cuestionario los profesores sobre el curso – taller en general, el Programa para el Análisis Gráfico de Ítems (PAGI), el manual para la capacitación y los materiales de apoyo, la guía para interpretar los ítems, la actuación de la instrucción, y d) los resultados del análisis cualitativo de los juicios que emitieron los docentes en la guía, al interpretar las gráficas que produjo el PAGI sobre los ítems de sus pruebas de aula. A continuación se describen dichos resultados.

4.1 Resultados del diseño de materiales e instrumentos

Los principales resultados de la investigación fueron los que se produjeron tras el diseño de los instrumentos y materiales que se desarrollaron. En la sección correspondiente del capítulo III, se hizo una descripción general del Manual para la capacitación del participante, del Programa para el Análisis Gráfico de Ítems (PAGI), del Cuestionario para evaluar tanto la operación general del curso, como la instrucción y los materiales utilizados, de la presentación en PowerPoint para apoyar la capacitación, así como de la guía para apoyar a los docentes para que pudieran interpretar adecuadamente la calidad técnica de sus ítems. Por ello, esos resultados no se describen en esta sección, ya que podrán observarse con detalle en los anexos 3, 4, 5, 8, y en el CD adjunto a la presente tesis donde aparece el PAGI. Por lo pronto, cabe destacar que dichos elementos se desarrollaron de conformidad con lo planeado y que cumplieron adecuadamente su función, como se verá más adelante.

4.2 Análisis psicométrico de los ítems y la escala para evaluar la opinión de los docentes sobre el módulo

Después de haber aplicado los cuestionarios a los participantes, los datos brutos obtenidos fueron estructurados en una base de datos y después procesados mediante el *software* especializado ITEMAN (Assessment Systems Corporation, 1989). Este

procedimiento fue clave, puesto que un instrumento de evaluación primero debe mostrar evidencias de que tiene un mínimo de calidad técnica, antes de poder analizar con confianza los resultados de su aplicación. Al respecto, los principales resultados fueron:

- Al nivel de la escala, los 21 participantes que respondieron los 28 ítems tuvieron una media de 4.57 (con una varianza de 0.09 y una desviación estándar de 0.30) en la escala Likert del 1 al 5, y una mediana de 4.64. Esto quiere decir que, en promedio, tuvieron una opinión bastante favorable sobre lo que afirmaban los ítems del cuestionario. En cuanto a la confiabilidad del instrumento, se obtuvo un coeficiente Alfa de 0.88, con un error estándar de medida de 0.10. Por su parte, la media de la correlación ítem – total fue de 0.53. Estos resultados muestran que la escala tuvo buenas propiedades psicométricas.
- Al nivel de los ítems, de los 28 que se analizaron uno de ellos tuvo una correlación ítem – total de 0.19, que es menor al 0.20 que la literatura considera apenas aceptable para cualquier tipo de ítem. Por su parte, 6 ítems tuvieron correlación entre 0.28 – 0.34; 10 de ellos entre 0.47 – 0.56; otros 10 tuvieron entre 0.60 – 0.79; y un ítem tuvo correlación = 0.82. Estos resultados muestran que, en casi todos los ítems, los profesores que expresaron una opinión más favorable sobre lo que afirmaba cada ítem, también tuvieron una opinión general más favorable; y que los profesores quienes respondieron bajo en los ítems tuvieron una opinión menos favorable en la escala completa.

La Tabla 4.1 presenta dichos resultados con mayor detalle; en ella se observan las medias de adscripción (*Endorsement*) en cada uno de los ítems, así como su respectiva correlación ítem – total.

Tabla 4.1 Media de adscripción y correlación Ítem – total en los ítems de la escala

Ítem	Media de adscripción	Correlación ítem – total
1	4.52	0.77
2	4.71	0.29
3	4.67	0.50
4	4.67	0.53
5	4.48	0.56
6	4.48	0.33
7	4.71	0.62
8	4.57	0.54
9	4.76	0.48
10	4.76	0.48
11	4.57	0.34
12	4.86	0.29
13	4.57	0.65
14	4.57	0.77
15	4.43	0.60
16	4.38	0.73
17	4.38	0.73
18	4.62	0.72
19	4.29	0.62
20	4.43	0.28
21	4.76	0.51
22	4.76	0.47
23	4.48	0.82
24	4.58	0.79
25	4.43	0.47
26	4.00	0.19
27	4.95	0.33
28	4.57	0.52
Promedio escala	**4.57**	**0.53**

Estos resultados psicométricos permitieron proceder con confianza al analizar los resultados del cuestionario que se aplicó a los profesores.

4.3 Resultados de la opinión de los profesores sobre la capacitación

En cuanto a la opinión de los docentes sobre la operación de la capacitación, los materiales de apoyo y la instrucción, los principales resultados fueron:

- La media de adscripción en los ítems de la escala fue 4.57. Ello muestra que en general los profesores tuvieron una opinión muy favorable sobre la capacitación y los elementos que la integraron.
- En las dimensiones evaluadas, el promedio más alto correspondió al indicador Instrucción (4.66) seguido de Materiales de apoyo (4.60) y Operación de la capacitación (4.57).
- En el indicador Instrucción, la opinión más favorable fue para Dominio del contenido (4.72), seguida de Actuación de la enseñanza (4.67), Planeación didáctica (4.63) y Promoción de la interacción (4.59).
- Sobre la dimensión Materiales diseñados, la opinión más favorable correspondió al indicador Presentación PowerPoint (4.76), seguido del PAGI (4.60) y del Manual (4.45).
- En cuanto a la dimensión Operación de la capacitación, los docentes percibieron que lo aprendido tendrá un impacto favorable en su práctica evaluativa en el aula (4.76), pero estuvieron menos de acuerdo con los ítems correspondientes al indicador Desarrollo de actividades (4.37).

En la Tabla 4.2 se presentan los promedios ponderados por el número de ítems, de las opiniones manifestadas por los participantes en cada una de las dimensiones Operación de la capacitación, Materiales de apoyo e Instrucción, así como en sus correspondientes indicadores e ítems.

Tabla 4.2 Promedios de adscripción en los ítems, por indicador y dimensión evaluada

Dimensión	Indicador	Ítem	Media
Instrucción 4.64	Planeación didáctica 4.63	1	4.52
		2	4.71
		3	4.67
	Interacción 4.59	4	4.67
		5	4.48
		6	4.48
		7	4.71
	Actuación de la enseñanza 4.67	8	4.57
		9	4.76
		10	4.76
		11	4.57
	Dominio del contenido 4.72	12	4.86
		13	4.57
Materiales de apoyo 4.51	Manual 4.45	14	4.57
		15	4.43
		16	4.38
		17	4.38
		18	4.62
		19	4.29
	Programa de cómputo (PAGI) 4.60	20	4.43
		21	4.76
	Presentación PowerPoint: 4.76	22	4.76
Operación de la capacitación 4.50	Desarrollo de actividades 4.37	23	4.48
		24	4.57
		25	4.43
		26	4.00
	Impacto 4.76	27	4.95
		28	4.57

4.4 Resultados del análisis cualitativo de los juicios que emitieron los docentes en la guía para apoyar la interpretación de los ítems.

Este análisis es de tipo cualitativo porque se comenta la forma en que los participantes respondieron a la guía para apoyar la interpretación técnica de sus ítems; es decir, la guía fue un instrumento que permitió a cada profesor emitir juicios razonados acerca de los ítems de su base de datos, una vez que procedieron a analizarlos con la ayuda del PAGI. Así, la guía se respondió de manera individual al final de la sesión de taller. Tras contestarla, siete participantes consideraron necesario conservar la guía para poder corregir después sus ítems, por lo que solo pudieron recabarse 14 de ellas, mismas que

fueron analizadas. En la Tabla 4.3 se muestra una síntesis de los principales comentarios que expresaron los profesores.

Tabla 4.3 Síntesis de las anotaciones en la guía de 14 participantes, sobre los ítems que analizaron

Participante	Comentarios
1	De manera general, se enfocó en observar los distractores de su prueba y al analizar su pendiente determinó cuáles se van modificar y cuáles se conservarán. En la mayoría de sus ítems anotó que eran fáciles, por lo que en esos casos emitió el juicio de modificar los distractores que no funcionaban, porque algunos ítems resultaron con una respuesta correcta obvia.
2	Este participante logró emitir un juicio en base a la dificultad de cada ítem; es decir, logró determinar qué grado de dificultad presentaban. También identificó a los distractores que son deficientes y que los tiene que modificar para mejorar la calidad de su prueba.
3	Determinó que la calidad de su prueba fue deficiente, porque en base al análisis que realizó emitió el juicio de modificar la mayoría de los distractores de sus ítems, e incrementar su grado de dificultad ya que algunos son muy obvios. También observó el comportamiento de la pendiente de las curvas en los grupos de habilidad. En el caso de dos ítems, determinó que debe modificarlos completamente porque no son buenos para la prueba.
4	Al analizar cada uno de los distractores de su prueba, solo comentó que eran buenos o malos; sin embargo, le faltó ser más descriptivo con sus distractores. Pero ello no fue obstáculo para determinar que sus ítems son fáciles y obvios. Posteriormente dictaminó cuáles ítems deben ser modificados.
5	Manifestó que en general sus distractores eran adecuados. Sin embargo, anotó que tiene que modificar los distractores de algunos ítems. Dictaminó que dos ítems tienen que ser remplazados o modificados para una mejor calidad de su prueba.
6	En cada caso indicó cuáles distractores son adecuados y cuáles otros se tienen que modificar. Además mencionó que se tienen que modificar algunas respuestas correctas, pues por diversas razones no cumplen con su función.
7	Este participante logró emitir un dictamen con base en la inclinación de la pendiente de los distractores y de la respuesta correcta. Además observó la pendiente de acuerdo a los grupos de habilidad, lo cual facilitó emitir un dictamen más razonado.
8	Anotó en cada caso que los distractores se necesitan modificar para que funcionen como tales. Fue consciente de que modificar los distractores origina modificar al ítem. Emitió su dictamen tomando en cuenta la dificultad de los ítems.
9	Emitió un juicio acerca de los ítems de su prueba en base al comportamiento de la pendiente de la respuesta correcta y de los distractores, de acuerdo a los grupos de habilidad. Su dictamen tiene fundamento en lo que observó en los grupos de habilidad, lo cual originó modificar algunos distractores.
10	Determinó modificar distractores de acuerdo al número de examinados que los eligieron. Además, para emitir su dictamen se basó en la dificultad de cada ítem y en los grupos de habilidad.
11	Se basó en los grupos de habilidad, en la pendiente de los distractores y en la dificultad de cada ítem para emitir un dictamen favorable o para sugerir modificar el ítem.
12	Para emitir su juicio, en la mayoría de los ítems tomó en cuenta los grupos de habilidad y la discriminación del ítem a partir de la pendiente de la respuesta correcta.
13	Se fijó en la discriminación de la respuesta correcta y en la dificultad de cada ítem para emitir su dictamen.
14	Observó las pendientes de los distractores y de la respuesta correcta. Así, observó la discriminación y dificultad de cada ítem. Dentro de su análisis determinó que tiene que reemplazar un ítem porque los distractores son más poderosos que la respuesta correcta.

Con el propósito de ilustrar el tipo de observaciones que hicieron los profesores en la guía, en la Tabla 4.4 se muestran los comentarios textuales que anotaron en su ítem 1, tres de los profesores participantes.

Tabla 4.4 Comentarios de tres participantes que evaluaron en la guía el ítem 1 de su prueba.

Profesor	Ítem	Opción	Comentarios
9	1	a	Queda por debajo de la respuesta correcta
		b	Redactarlo. Distrae al inicio
		c*	En el grupo 1, 2 y 3 va en incremento, pero en el grupo 4 bajó. Se tendría que revisar
		d	Queda por debajo de la correcta, pero en el grupo 2, 3 y 4 queda constante. Cumple su función
		Juicio final sobre el ítem	Modificar distractor b
14	1	a*	Es muy evidente. Ítem fácil que discrimina bien
		b	No es para nada atractivo
		c	Es atractivo para los que no saben
		d	No es para nada atractivo
		Juicio final sobre el ítem	Modificar los distractores
1	1	a	Muestra pendiente cero en los tres primeros grupos y al 4º grupo nadie lo contesta
		b	En el grupo 2 nadie lo contesta, por lo demás está bien
		c	Resulta buen distractor
		d*	Muestra pendiente positiva, a la alza
		Juicio final sobre el ítem	Modificar distractor a

* Es la respuesta correcta

Otros resultados de tipo cualitativo, fueron los comentarios que manifestaron algunos participantes al concluir el curso – taller. Como se acostumbra al finalizar este tipo de eventos, tomaron la palabra tres profesores que comentaron que la capacitación que recibieron resultó muy relevante para ellos, y declararon que les despertó gran interés la técnica y expresaron gratitud por haberse apoyado su actualización en esta tecnología educativa. Además, expresaron felicitaciones para los coordinadores del curso – taller. Por otro lado, uno de los dos Jefes de Enseñanza que participaron expresó su gratitud de manera personal con la instructora y manifestó su interés por impartir posteriormente el curso – taller a profesores de otras especialidades.

Un resultado de tipo metodológico que se considera importante en este trabajo, fue el que se obtuvo tras el proceso que se siguió para adaptar la técnica del AGI a las condiciones de la evaluación a pequeña escala, particularmente su capacidad para ser utilizada en el aula donde se maneja un número reducido de examinados. Al respecto, el principal reto era preservar el potencial de la técnica del AGI para generar curvas de la respuesta correcta y de los distractores, de manera que se observaran con claridad los cambios de tendencia dentro de la gráfica a pesar de contar con un número reducido de examinados. Como se recordará, para lograrlo se procedió a variar sistemáticamente el número de grupos de habilidad en que se dividieron los examinados y el número total de examinados que respondieron el ítem. Después de efectuar los ensayos con ítems de diferente dificultad y con diferente discriminación, en los que se representaron las puntuaciones totales en la prueba en 2, 3, 4 y 5 grupos de habilidad, se observaron los siguientes resultados:

- Con dos grupos de puntuaciones (bajos – altos), se observan con claridad las tendencias de la respuesta correcta y de los distractores, pero resultan engañosas porque se pierde información; es decir, son tendencias promedio que ocultan información más fina que resulta necesaria para decidir sobre la calidad del ítem. En este caso, hay 16 puntuaciones en cada grupo de habilidad cuando se tienen las calificaciones de 32 examinados.
- Con tres grupos de puntajes (bajos – medios – altos), también se observan con claridad las tendencias de la respuesta correcta y de los distractores; además, se empiezan a observar los cambios en las tendencias pero son bruscos y en muchos casos desconciertan pues resultan incongruentes. Aquí, hay aproximadamente 10 puntajes en cada grupo de habilidad cuando hay un total de 32 examinados.
- Con cuatro grupos de habilidad (bajo – medio bajo – medio alto – alto), se observan todavía con claridad las tendencias de la respuesta correcta y de los distractores, y se presentan cambios menos bruscos que facilitan la interpretación de las curvas de respuestas. En este caso, hay 8 puntajes en cada grupo de habilidad cuando se tiene un total de 32 examinados.

- Con cinco grupos de habilidad (bajo – medio bajo – medio – medio alto – alto), se observan con más claridad aún los cambios de tendencia de la respuesta correcta y de los distractores, pero los cambios de tendencia resultan más bruscos y dificultan la interpretación de las curvas de respuestas. Estos cambios son el resultado de tener menos puntajes de examinados en cada grupo de habilidad; en este caso, alrededor de 6 puntajes cuando el total de examinados es 32.

Sin embargo, también se observó que estos comportamientos no se presentan cuando se tiene un número mayor de casos. Por ejemplo, el ítem 1 que se muestra en el anexo 6, tiene 107 casos, y presenta curvas de respuestas con tendencias claramente definidas y sin cambios bruscos, casi sin importar el número de grupos de habilidad, lo cual no se da en ninguno de los demás ítems que se analizaron con 32 casos.

De esta manera, se consideró que con un mínimo de 32 casos las tendencias de la curva de la respuesta correcta y de los distractores resultaban suficientemente claras cuando las puntuaciones de la prueba se organizaban en cuartiles. Así, entre el puntaje menor o el mayor y el cuartil más próximo a ellos, así como entre cuartiles adyacentes, se incluyeron 8 puntuaciones que representan el 25% de las calificaciones en la prueba, como puede observarse en la Figura 4.1. Lo anterior permitió observar las tendencias de las respuestas al ítem, así como los demás componentes del AGI que definen la calidad técnica de un ítem. Para mayores detalles sobre estos resultados véanse los anexos 6 y 7 al final del documento.

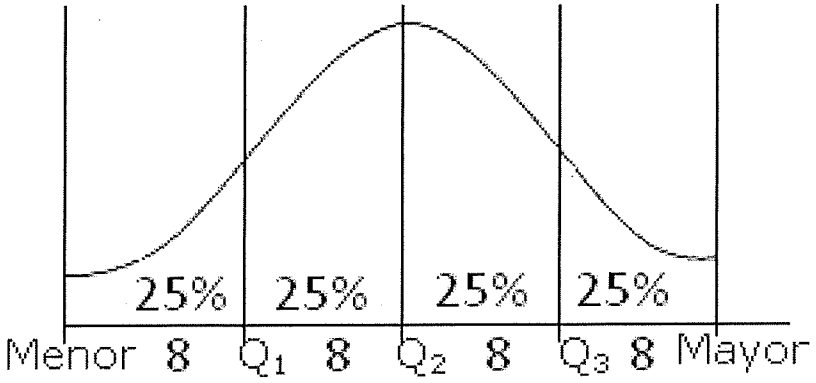


Figura 4.1 Distribución de puntajes en cuartiles en la curva de distribución normal de frecuencias.

5.1 Conclusiones y sugerencias

En general, los resultados obtenidos en el estudio permiten concluir que fue posible cumplir con los objetivos enunciados en la introducción del presente documento. Así, se pudo desarrollar, operar y evaluar el módulo pretendido para capacitar a docentes en servicio, a fin de que pudieran mejorar sus pruebas objetivas mediante el Análisis Gráfico de Ítems (AGI). Esta conclusión es apoyada por los siguientes resultados:

- Los profesores que participaron en el curso – taller reportaron en el instrumento que se les aplicó, que tuvieron aprendizajes significativos relativos a los conocimientos y habilidades de psicometría básica necesarios para efectuar el análisis gráfico de sus ítems. Además, el análisis de la correspondencia entre los juicios analíticos sobre sus ítems que anotaron en la guía para apoyar la interpretación y las gráficas que produjeron, apoyan esa apreciación.
- El uso del programa de cómputo desarrollado (PAGI) permitió a los docentes observar los puntos fuertes y débiles de sus ítems, en términos de los principales indicadores psicométricos que definen su calidad técnica, así como conocer formas específicas de mejorarlos. Esta afirmación es respaldada tanto por las opiniones expresadas por los docentes en el instrumento, como por las participaciones que tuvieron durante la fase de taller y los productos que generaron tras analizar sus ítems y llenar la guía para apoyar la interpretación de la calidad técnica de sus ítems.
- La aplicación del instrumento de evaluación diseñado arrojó datos que muestran su calidad psicométrica. Salvo un ítem, todos los reactivos de la escala que se aplicó a los docentes tuvieron discriminación apropiada y el instrumento en su conjunto exhibió alta confiabilidad. Por ello, se puede afirmar con cierta seguridad que los informantes percibieron que los materiales diseñados (manual del participante, guía para apoyar la interpretación de la calidad técnica de los ítems, presentación en PowerPoint) resultaron apropiados, porque propiciaron aprendizajes significativos que fueron pretendidos. Asimismo, la actuación de la

instructora fue considerada adecuada, pues en opinión de los participantes facilitó su aprendizaje.

Un grupo de resultados importantes que tuvieron una función estratégica para el logro de los objetivos del presente estudio evaluativo, fueron los productos mismos del diseño y desarrollo de los procedimientos, instrumentos y materiales que permitieron capacitar a docentes en servicio para realizar el análisis gráfico de ítems. Estos resultados metodológicos se consignaron en los anexos de la tesis e incluyeron:

- El diseño y operación de dos procedimientos novedosos que permitieron adaptar la técnica del análisis gráfico de ítems (Batenburg y Laros, 2002) para ser aplicada al evaluar los reactivos de pruebas de pequeña escala. Lo anterior, hizo posible que los profesores pudieran observar de manera visual las tendencias de la respuesta correcta y de los distractores y con ello entender su significado evaluativo. Además, la base de datos que se generó al realizar los ensayos considerados en dichos procedimientos dio origen a los nueve ejemplos que se ilustraron en el manual y en el PAGI, todo lo cual resalta la importancia de esos procedimientos.

- La elaboración de un manual técnico fue la principal fuente de información para llevar a cabo la capacitación, además de una guía metodológica y conceptual para el análisis gráfico de ítems y para manejar la interfaz del programa PAGI.

- La elaboración de la presentación en PowerPoint, que resultó un material didáctico muy significativo para los docentes durante la capacitación.

- El PAGI fue elaborado por un especialista en programación a partir de un conjunto de especificaciones y algoritmos que se le proporcionaron y que eran necesarios para su creación. Para diseñar esta primera versión del PAGI se hizo especial énfasis en que tuviera características como: funcionamiento adecuado, cualidad de amigable y potencial para facilitar y hacer significativo el análisis gráfico de ítems para los docentes. Estos rasgos fueron probados con un grupo de profesores

en servicio y los resultados que se obtuvieron al realizar la capacitación muestran que el PAGI funcionó justo como se pretendía.

- La guía para apoyar la interpretación de los profesores sobre los productos del análisis gráfico de sus ítems. La guía no solo mostró ser útil para que los profesores juzgaran de manera fundamentada sus ítems, sino que al ser llenada por ellos se convirtió en la principal fuente de información que se utilizó para efectuar el análisis cualitativo del impacto que tuvo el curso – taller sobre la apropiación de la técnica del análisis gráfico de ítems por parte de los profesores.

- El instrumento para evaluar las percepciones de los profesores sobre la operación de los procedimientos y materiales elaborados, mismo que exhibió evidencias claras de su calidad técnica en cuanto a confiabilidad y poder discriminativo.

En todos los casos, los elementos mencionados resultaron ser adecuados para los docentes y de hecho se considera que constituyen el principal resultado de este estudio evaluativo.

Otros aspectos que vale la pena destacar son la disposición e interés que mostraron los docentes en servicio durante toda la operación del modulo de capacitación. Se pudo observar una actitud reflexiva ante los resultados de sus ítems y su intento por mejorar las pruebas que aplican a sus alumnos. Se considera que lo anterior fue posible gracias a que conocieron una manera sencilla y directa de mejorar las pruebas que aplican en su salón de clase.

No obstante, también fue posible detectar aspectos que es necesario mejorar. Por ejemplo, las opiniones menos favorables se expresaron ante los ítems que integraron la dimensión Operación de la capacitación; particularmente, en cuanto al tiempo asignado para desarrollar el curso – taller. Al respecto, cabe señalar que aunque en general se lograron los propósitos de la capacitación en una sola sesión, con un tiempo asignado de 4 horas, los profesores percibieron que fue insuficiente para el logro cabal de los objetivos. Varios datos confirman esta percepción; por ejemplo, algunos profesores no terminaron de juzgar todos sus ítems en la guía, las evaluaciones individuales de los profesores no se

pudieron socializar y realimentar ante el grupo; o bien las gráficas que produjo el PAGI al analizar los ítems de los participantes no pudieron recabarse el final de la sesión de capacitación. Lo anterior sugiere explorar la conveniencia de dividir la capacitación en dos sesiones de cuatro horas: la primera sesión dedicada a cubrir los conceptos básicos de la Teoría Clásica de los Tests (TCT) y de la técnica del AGI y en la segunda sesión trabajar exclusivamente en la modalidad de taller con el apoyo del PAGI.

Otra limitación del presente estudio, tiene que ver con otros desarrollos más recientes de la técnica del análisis gráfico de ítems, mismos que incluyen aspectos complementarios como la representación de la dificultad de ítems mediante gráficos de barras, la representación de la discriminación del conjunto de ítems del examen mediante gráficos circulares, o más formalmente el análisis gráfico del sesgo en los ítems y la construcción de versiones paralelas de una prueba mediante el AGI, los cuales no fue posible incluir en este trabajo, principalmente por razones de tiempo. Por ello, se sugiere que en nuevas versiones que se desarrollen tanto del PAGI, como del manual del participante y de la guía para apoyar la interpretación de los ítems, se incorporen esos aspectos. Ello hará más útil y versátil el módulo de capacitación.

Aunque en el presente trabajo algunos ítems del cuestionario evaluaron la operación del PAGI, dicha evaluación resultó muy general, parcial y superficial. Por ello, se considera necesario evaluar formalmente el programa de cómputo, de manera más técnica, específica e integral. Es decir, en la misma línea y con los mismos elementos que se comentaron en la última parte del capítulo II sobre el marco conceptual.

Como en este trabajo fue posible capacitar con éxito a profesores en servicio, sobre aspectos básicos de la psicometría que son necesarios para efectuar un análisis gráfico de ítems, sería interesante explorar la posible incorporación de otras nociones básicas de la TCT, como la idea de confiabilidad de la prueba y otras que aunque no se requieren para el AGI son necesarias para otras técnicas de análisis de ítems y tests.

Reflexión final

Se considera que los resultados de la tesis representan una aportación relevante para la formación y la actualización de profesores. Los procedimientos, instrumentos y materiales desarrollados probaron ser de ayuda para que docentes en servicio puedan evaluar de manera más justa, válida y confiable el dominio de los contenidos que imparten; y en particular para que mejoren la calidad de su banco de ítems. Sin embargo, aún se requiere replicar el presente estudio con profesores de otros niveles y otras modalidades educativos, para observar en qué medida la metodología aquí propuesta puede llegar a generalizarse a otros ámbitos escolares, y ser incluida en la formación inicial y continua de maestros en servicio que operan la SEP y a otras instituciones formadoras de docentes.

American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington: Autor. Consultado el 23 de octubre de 2009 en: http://www.aera.net/

Andrich, D., Sheridan, B., Lyne, A. & Luo, G. (2000). *RUMM: a Windows-based item analysis employing Rasch unidimensional measurement models*. Perth: Murdoch University. RUMM Laboratory Pty Ltd.

Assessment Systems Corporation (2006). *Lertap 5. Version 5 for Windows.* St. Paul, Minnesota: Autor. Consultado el 13 de noviembre de 2008 en: http://www.assess.com

Assessment Systems Corporation (1989). *ITEMAN Version 3.6.* St. Paul, Minnesota: Autor. Consultado el 13 de noviembre de 2008 en: http://www.assess.com

Batenburg, T. V. (2006, 6 -8 July). *Graphical Item Analysis: Difficulty, Discrimination and Guessing.* Poster presented at the 5th Conference of the International Test Commission: Psychological and Educational Test Adaptation Across Languages and Cultures: Building Bridges Among People. Brussels, Belgium.

Batenburg, T. V. & Laros, J. A. (2002). Graphical Analysis of Items. *Education research and evaluation, 8(3),* 319 – 333.

Batenburg, T. V. & Yurdugül, H. (2006). Item Difficulty from Graphical Item Analysis. *Eurasian Journal of Educational Research*. (24) Consultado el 25 de Febrero de 2009 en: http://www.ejer.com.tr/

Carrol, J. (2002). Human – Computer Interaction in the New Millenium. Consultado el 23 de Octubre de 2009 en: http://books.google.com.mx/books

CiTO, (2005). *TiaPlus, Classical Test and Item Analysis. Arnhem: Cito M & R Department.* Autor.

Consejo Asesor Externo del Centro Nacional para la Evaluación de la Educación Superior [CENEVAL], (2000). *Estándares de calidad para instrumentos de evaluación educativa*. México: Autor. Consultado el 23 de octubre de 2009 en: http://www.ceneval.edu.mx/

Dirección General de Educación Superior para Profesionales de la Educación [DGESPE], (2009). Planes y programas. México: Autor. Consultado el 20 de Mayo de 2009 en: http://normalista.ilce.edu.mx/

Faulkner, C. and Prentice, H. (1998). *The Essence of Human – Computer Interaction.* United States of America: Prentice – Hall International Edition.

Haladyna, M. (2004). Developing and validating multiple choice test items. Consultado el 15 de Noviembre de 2009 en: http://books.google.com.mx/books

Ledesma, R, Molina, G, Valero, P. & Young, F. (2002). Un módulo de análisis visual de ítems para el paquete estadístico ViSta. *Revista Electrónica de Metodología Aplicada*, 7(2)

Löwgren, J. & Stolterman, E. (2005). *Thoughtful design: a design perspective on information technology.* Consultado el 25 de Octubre de 2009 en: http://books.google.com.mx/books

Lozano, M. E. (2007). Formación docente: El servicio de asesoría académica a las escuelas y el trayecto formativo. En Memorias del *IX Congreso Nacional de Investigación Educativa* [Disco compacto] Yucatán, México: Universidad Autónoma de Yucatán.

Martínez, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa, 11*(2). Consultado el día 15 de Enero de 2010 en: http://redie.uabc.mx/

Martínez, M., Maya, C. y Zenteno, E. (1996). Hacia una evaluación institucional de la Educación Normal. En *Memorias del II Foro de Evaluación Educativa.* Saltillo, Coahuila, México: Centro Nacional de Evaluación para la Educación Superior.

Moreno, T. (2007). Pintando el retrato de una escuela con datos cualitativos: La evaluación desde la perspectiva de los alumnos de telesecundaria. En *Memorias del IX Congreso Nacional de Investigación Educativa* [Disco compacto] Mérida, Yucatán, México: Consejo Mexicano de Investigación Educativa.

Noguez, A. (2000). Evaluación y seguimiento de los nuevos planes y programas de estudio de tres escuelas normales del D.F. En *Memorias del IV Foro de Evaluación*

Educativa. Ciudad Juárez, Chihuahua, México: Centro Nacional de Evaluación para la Educación Superior.

Norman, D. (1990). *The Desing of Everyday Things*. New York: Doubleday-Currency.

Oliveira, M. R. (2007). Analise Grafica dos Itens – AGI. En Avaliacao do Plano de Desenvolvimiento da Escola(PDE): Um estudo longitudinal utilizado análise multinivel. Tesis doctoral no publicada. Universidad de Brasilia, Brasil. Consultado el 20 de Marzo 2010 en: http://bdtd.bce.unb.br/tedesimplificado

Preece, J. (1994). Human Computer Interacion: Concepts and Design. Consultado el día 15 de Octubre de 2009 en: http://www.amazon.com

Rodrígues, M. (2006). *Prosposta de Análise de itens das Provas do Saeb sob a Perspectiva Pedagógica e a Psicométrica*. Em Estudos em Avaliacao da Educacao Básica. Brasil: Sistema de Avaliação da Educação Básica.

Secretaría de Educación, Pública [SEP] (2007). *Lineamientos para la selección, diseño, desarrollo y evaluación de programas de estudio para la formación continua de maestros de educación básica en servicio.* México: Autor. Consultado el 28 de noviembre del 2008, en: http://formacioncontinuazac.gob.mx.

Schnotz, W. (1994). Comprehension of graphics. Consultado el 23 de Octubre de 2009 en: http://books.google.com.mx/books

Shepard, L. (2006). La evaluación en el Aula. En Robert l. Brennan (Eds), *Identification of Mild Handicaps (pp545- 572).* Bolder, Colorado, E.E.U.U. (Reimpreso en México por Instituto Nacional para la Evaluación de la Educación, Ed. 2006, México. Consultado el 30 de octubre de 2008 en: http://www.inee.edu.mx/

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* New York: Addison-Wesley.

Verhelst, N. (2004). Section C. Classical Test Theory. Consultado el 28 de Noviembre de 2008 en: http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionC.pdf

Wainer, H. (1989). The Future of Item Analysis. Consultado el 28 de Noviembre de 2009 en: http://www.jstor.org/pss/1434865

ANEXO 1

Licenciatura en Educación Primaria

Mapa curricular Plan 1997

Primer semestre	Segundo semestre	Tercer semestre	Cuarto semestre	Quinto semestre	Sexto semestre	Séptimo semestre	Octavo semestre
Bases filosóficas legales y organizativas del sistema educativo Mexicano	La educación en el desarrollo histórico de México I	La educación en el desarrollo histórico de México II	Temas selectos de pedagogía I	Temas selectos de pedagogía II	Temas selectos de pedagogía III	Trabajo Docente I	Trabajo Docente II
Problemas y políticas de la educación básica	Matemáticas y su enseñanza I	Matemáticas y su enseñanza II	Ciencias naturales y su enseñanza I	Ciencias naturales y su enseñanza II	Asignatura regional II		
Propósitos y contenidos de la educación primaria	Español y su enseñanza I	Español y su enseñanza II	Geografía y su enseñanza I	Geografía y su enseñanza II	**Planeación de la enseñanza y evaluación del aprendizaje**		
Desarrollo infantil I	Desarrollo infantil II	Necesidades educativas especiales	Historia y su enseñanza I	Historia y su enseñanza II	Gestión escolar		
Estrategias para el estudio y la comunicación I	Estrategias para el estudio y la comunicación II	Educación física I	Educación física II	Educación física III	Educación artística III		
			Educación artística I	Educación artística II			
			Asignatura regional I	Formación ética y cívica en la escuela primaria I	Formación ética y cívica en la escuela primaria II		
Escuela y contexto social	Iniciación al trabajo escolar	Observación y práctica docente I	Observación y práctica docente II	Observación y práctica docente III	Observación y práctica docente IV	Seminario de análisis de trabajo docente I	Seminario de análisis de trabajo docente II

ANEXO 2

Licenciatura en Educación Secundaria

Mapa curricular Plan 1998

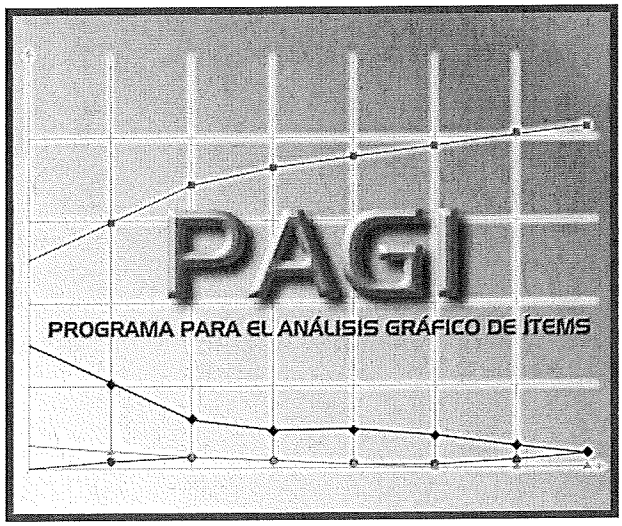
Primer semestre	Segundo semestre	Tercer semestres	Cuarto semestre	Quinto semestre	Sexto semestre	Séptimo semestre	Octavo semestre
Bases filosóficas, legales y organizativas del sistema educativo mexicano	La educación en el desarrollo histórico de México I	La educación en el desarrollo histórico de México II	Seminario de temas selectos de historia de la pedagogía y la educación I	Seminario de temas selectos de historia de la pedagogía y la educación II	Por especialidad		
Estrategias para el estudio y la comunicación I	Estrategias para el estudio y la comunicación II	Por especialidad	Por especialidad	Por especialidad	Por especialidad		
Problemas y políticas de la educación básica	Introducción a la enseñanza de la especialidad	Por especialidad	Por especialidad	Por especialidad	Por especialidad		
Propósitos y contenidos de la educación básica I (Primaria)	La enseñanza en la escuela secundaria. Cuestiones básicas I	La enseñanza en la escuela secundaria. Cuestiones básicas II	Por especialidad	Por especialidad	Por especialidad		
Desarrollo de los adolescentes I Aspectos generales	Propósitos y contenidos de la educación básica II (Secundaria)	La expresión oral y escrita en el proceso de enseñanza y de aprendizaje	**Planeación de la enseñanza y evaluación del aprendizaje**	Opcional I	Opcional II	Taller de diseño de propuestas didácticas y análisis del trabajo docente I	Taller de diseño de propuestas didácticas y análisis del trabajo docente II
	Desarrollo de los adolescentes II Crecimiento y sexualidad	Desarrollo de los adolescentes III Identidad y relaciones sociales	Desarrollo de los adolescentes IV. Procesos cognitivos	Atención educativa a los adolescentes en situaciones de riesgo	Gestión escolar	Trabajo Docente I	Trabajo Docente II

ANEXO 3

Universidad Autónoma de Baja California
Instituto de Investigación y Desarrollo Educativo

MANUAL PARA EL ANÁLISIS GRÁFICO DE ÍTEMS

- GUÍA PARA EL PROGRAMA PARA EL ANÁLISIS GRÁFICO DE ÍTEMS



GUADALUPE DE LOS SANTOS LÁZARO

ENSENADA, B. C. JUNIO DE 2010

ÍNDICE

1. PRESENTACIÓN

La evaluación es un proceso que siempre ha generado interés en los diferentes niveles educativos. Dentro del ámbito escolar, los profesores en servicio son quienes regularmente se encargan de evaluar el aprendizaje de los alumnos en el aula.

El profesor siempre ha interactuado con los alumnos para establecer una evaluación del aprendizaje de cada uno de ellos, en función del avance que presenten. Así, los docentes se han visto en la necesidad de elaborar instrumentos que detecten las necesidades de los alumnos para atenderlas, para observar su aprendizaje y realimentar los objetivos instruccionales a fin de mejorar el desempeño académico, o para certificar el logro educativo. Así, los profesores que desarrollan sus instrumentos de evaluación deben poner un énfasis tanto en las conexiones entre la evaluación y las actividades de enseñanza, como en el aspecto técnico para mejorar la calidad técnica de los ítems que integran esos instrumentos.

Para lograr esto último, la Teoría Clásica de los Tests (TCT) ha sido una guía para los constructores de pruebas. Esta teoría considera que tras aplicar y calificar una prueba, se obtiene el puntaje que logró cada examinado, el cual es conocido como puntuación observada. Sin embargo, el nivel real de dominio o habilidad que tiene el examinado no lo conocemos; porque la puntuación que observamos está influenciada por diversos errores de medición. La TCT nos ayuda a reducir los errores que cometemos al construir un examen, de modo que sea posible conocer el nivel de habilidad real de los estudiantes. Para ello, puede utilizarse, entre otros, el método del Análisis Gráfico de Ítems (AGI).

El Análisis gráfico de ítems (Batenburg y Laros, 2002), describe la técnica mediante la cual se despliega visualmente la relación entre la puntuación total en la prueba y el total de los examinados que eligieron la opción correcta y las opciones falsas en un ítem de opción múltiple.

Es propósito de este curso brindar los conocimientos psicométricos mínimos necesarios para que docentes en servicio desarrollen, con un mínimo de calidad técnica, ítems de opción múltiple que utilizan en sus pruebas. Lo anterior será posible por medio de la

técnica del AGI y con el apoyo de la aplicación denominada Programa para el Análisis Gráfico de Ítems (PAGI).

2. ANÁLISIS GRÁFICO DE ÍTEMS

El Análisis Gráfico de Ítems (AGI) surgió de la propuesta de Batenburg y Laros (2002). El método consiste en mostrar visualmente, mediante una gráfica de líneas, la relación entre la puntuación total en la prueba y la frecuencia, proporción o porcentaje de respuestas de los examinados[1] que eligieron la opción correcta y las opciones falsas en un ítem de opción múltiple de una prueba. El AGI proporciona información esencial y fácilmente interpretable acerca de características técnicas del ítem como son su dificultad, su poder de discriminación y el nivel de adivinación.

El desarrollo del método AGI tuvo lugar en el contexto de proyectos para desarrollar pruebas a gran escala, específicamente en el proyecto del Sistema de Evaluación de la Educación Básica (SAEB) en Brasil. El SAEB se originó en 1997 para evaluar la ejecución de los estudiantes brasileños de educación básica en contenidos de matemáticas, física, e idioma portugués, entre otros (Van Batenburg y Laros, 2002).

La importancia de este método radica en ser utilizado para identificar ítems que presentan fallas y que deben ser excluidos de las pruebas por no tener los mínimos requerimientos técnicos psicométricos. Con el AGI los ítems de mala calidad son fáciles de detectar, porque en ellos se observa que el número o porcentaje de examinados que eligieron la opción correcta, decrementa a medida que incrementa el puntaje o total de aciertos en la prueba; o bien, porque muestran en una o más opciones falsas (distractores) que no decrementa el número o porcentaje de estudiantes que las eligieron, al incrementar el puntaje total en la prueba.

[1] El término examinados se refiere a los estudiantes que respondieron la prueba. Para referirse a las respuestas de los examinados, en adelante sólo se hará referencia al número o porcentaje de examinados. (eje Y)

Por otra parte, el AGI plantea que al aplicar un examen a un grupo de examinados hay presencia de error, el cual regularmente se da porque hay examinados que no tomaron en serio la aplicación de la prueba; por ejemplo, van al baño por tiempo prolongado, están distraídos durante la aplicación, se dedicaron a otras tareas, miraban al techo todo el tiempo. Tras la aplicación y calificación de la prueba, esa falta de seriedad puede detectarse por medio de respuestas que los examinados eligieron al azar (adivinación) y por respuestas que omitieron (datos perdidos).

En cuanto a la adivinación, debe eliminarse a los estudiantes que respondieron con patrones de respuesta de una sola categoría; como el responder a varios ítems consecutivos eligiendo siempre la opción A, A, A, A, A, A... o B, B, B, B, B, B... Cuando un estudiante responde de esa manera, es probable que no contestó la prueba con seriedad. También es posible identificar patrones de respuesta de los examinados que son repetitivos; por ejemplo, cuando responden con un patrón como ABCDE, ABCDE, ABCDE... EDCBA, EDCBA, EDCBA... ACE, ACE, ACE... o BD, BD, BD, BD... En todos esos casos, se sugiere eliminar a estos examinados de la base de datos, pues la información que aportaron no es útil para estimar la calidad de los ítems de la prueba.

Para el caso de los valores perdidos, se sugiere borrar o eliminar a los examinados que tengan una cantidad predeterminada de omisiones; digamos a quienes no contestaron el 50% de los ítems.

2.1 Evaluación de la calidad de ítems mediante el análisis gráfico de ítems

Para explicar e ilustrar el método del AGI, se utilizaron algunos ítems de una base de datos derivada de la aplicación, en escuelas primarias de Baja California, de una prueba de español de referencia criterial desarrollada en el Instituto de Investigación y Desarrollo Educativo de la UABC (Contreras, 2001) y del Examen Colegiado de Matemáticas I desarrollado por la Facultad de Ingeniería en Mexicali, de la Universidad Autónoma de Baja California (Encinas y colaboradores, 2005).

Antes de empezar a describir el Análisis Gráfico de Ítems es necesario identificar los elementos necesarios para emplear esta técnica. Para ello, la figura 1 presenta las partes que se requieren para el análisis.

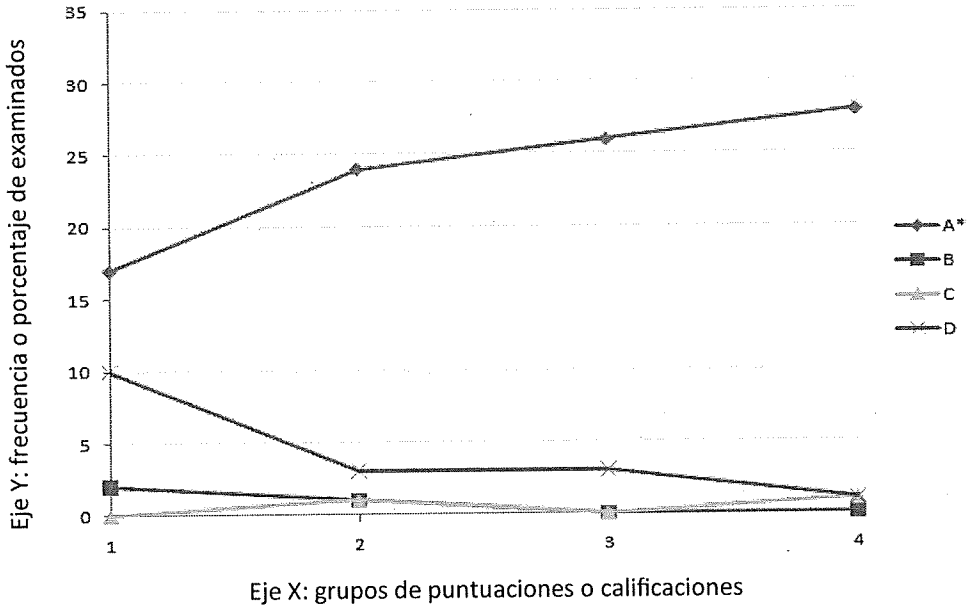


Figura 1. Componentes para el Análisis Gráfico de Ítems

En el método del AGI, el eje Y muestra la frecuencia o porcentaje de examinados que elijen las diferentes opciones. En el caso concreto de la figura 1 se muestra la frecuencia de estudiantes que eligieron cada opción. Mientras tanto, en el eje X se muestran los totales de aciertos o calificaciones que obtuvieron todos los estudiantes en la prueba, las cuales fueron divididas en 4 grupos iguales que representan los siguientes niveles de habilidad:

- Bajo (1)
- Medio bajo (2)
- Medio alto (3)
- Alto (4)

Los grupos de habilidad se despliegan en el eje X de izquierda a derecha para ubicar al extremo derecho al grupo 4, el de las calificaciones más altas, y que por ello es considerado el de mayor habilidad.

A la derecha de la figura 1 se encuentra una sección de código que identifica a cada una de las curvas de respuestas que dieron los examinados ante en las opciones que ofreció el ítem, mismas que están referidas como A, B, C y D. La curva de la respuesta correcta se identifica con un asterisco en la parte superior derecha de la letra correspondiente; en este caso se encuentra en la opción A. Las líneas que se observan en el eje de coordenadas son las curvas de respuestas al ítem. De ellas, una pertenece a la respuesta correcta y tres corresponden a los distractores, en este caso identificados con las letras B, C y D.

En la figura 1 se muestra un ítem de buena calidad, porque la frecuencia de examinados que eligen la respuesta correcta **A** incrementa a medida que aumenta la habilidad. Por otro lado, las demás curvas de respuestas al ítem se refieren a la frecuencia de examinados que eligen los distractores **B**, **C** y **D**, misma que en cada caso tiende a disminuir a medida que aumenta la habilidad. Ambas situaciones constituyen lo que es ideal en un buen ítem, por lo que este caso ilustra con claridad el tipo de información que aporta el análisis gráfico de ítems para mejorar los exámenes que elabora el profesor.

Después de conocer estos elementos generales del análisis gráfico de ítems, enseguida observaremos varios casos de ítems que ilustran diferentes tendencias y características de las respuestas de los examinados, mismas que hacen que un ítem pueda ser considerado como bueno o malo; fácil o difícil; con alto o bajo poder discriminativo; o con distractores efectivos o ineficaces.

2.1.1 Ítem fácil que no discrimina

En la figura 2 aparece un ítem fácil, porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta D es alto, tanto en los grupos de mayor habilidad (3 y 4), como en los de menor habilidad (1 y 2). Esto quiere decir que el ítem es muy fácil de contestar o que la respuesta correcta es evidente.

Por otro lado las opciones A, B y C, conocidas como distractores, en general no resultan atractivas para los examinados. Los distractores B y C no fueron elegidos por ningún examinado, ni en los grupos de alta habilidad (3 y 4), ni en los de baja habilidad (1 y 2); y la

opción A fue elegida por un solo examinado del grupo 1; es decir, que los distractores no resultaron atractivos para los examinados con calificaciones bajas en el examen, como debería haber sido.

Como consecuencia de lo anterior, el ítem no permite discriminar entre los examinados que dominan el contenido que evalúa el ítem y los examinados que no lo dominan. Esto puede observarse porque la pendiente o inclinación de la curva de la respuesta correcta es mínima y por ello tiende a ser horizontal.

En síntesis, el ítem es muy fácil pero no permite discriminar entre los que saben y quienes no saben lo que el ítem evalúa, por lo que puede considerarse inadecuado.

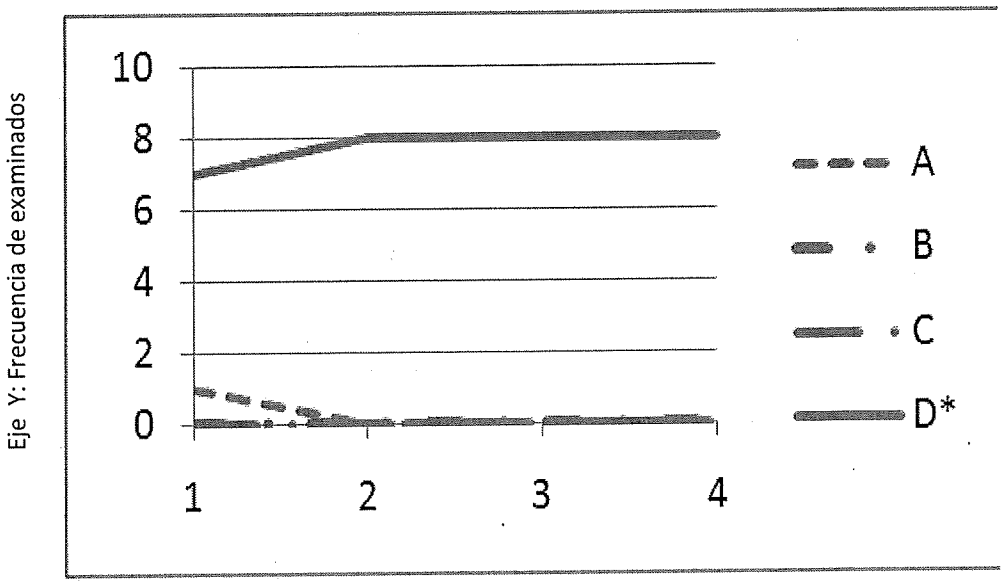


Figura 2. Ilustración de un ítem fácil que no discrimina (Eje Y = Nº de examinados)

En la figura 3 aparece el mismo ítem, pero ahora el eje Y muestra el porcentaje de examinados que eligieron las opciones, en vez del número de examinados que lo hicieron, como en la figura 2.

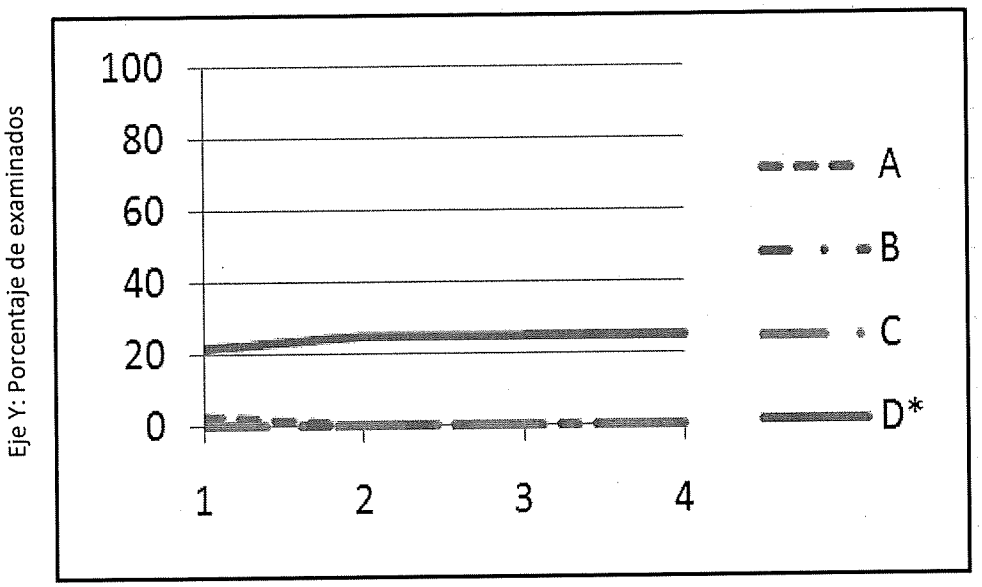


Figura 3. Ilustración de un ítem fácil que no discrimina (Eje Y = % de examinados)

La figura 3 permite observar de manera panorámica el total de respuestas que dieron el 100% de los examinados a todas las opciones del ítem, por lo cual se pueden ver con menos claridad y detalle las curvas de respuestas al ítem. Al igual que en la figura 2, se observa que este ítem es demasiado fácil y que existe poca diferencia entre los examinados que saben poco y los examinados de los grupos de habilidad alta, pues casi todos responden correctamente el ítem. Además, se ve que los distractores son poco atractivos para los examinados. La respuesta correcta resulta evidente. Para este tipo de ítem se sugiere revisar la redacción de los distractores a fin de hacerlos plausibles para quienes pertenecen a los grupos de habilidad baja (1 y 2). En otras palabras, este ítem es fácil pero malo; pues no discrimina entre los examinados.

La diferencia entre las figuras 2 y 3 radica en el cambio en el eje Y, del número o frecuencia de examinados, por el porcentaje de ellos. Este cambio es importante y necesario para que no se distorsione la perspectiva general del conjunto de respuestas que dieron los examinados ante el ítem. Es decir, sin importar el número de examinados que responden un ítem, siempre serán el 100%; por lo que se trata de un número fijo. En cambio, en la figura 2, el máximo número de examinados que aparece en el eje Y depende de la frecuencia con que se dieron las respuestas en cada opción del ítem, la cual puede variar de un ítem a otro.

Lo anterior significa que la figura 3 presenta las respuestas a todas las opciones que dieron el 100% de los examinados. La figura 2 presenta solo la parte de la gráfica, en la figura 3, a partir de la cual se dio la máxima frecuencia de examinados que respondieron alguna opción, usualmente la respuesta correcta cuando un ítem tiene calidad. Arriba de este punto no hay información. Por así decirlo, la figura 2 es como un zoom o ampliación que hacemos a la parte de la figura 3 que presenta la información más interesante sobre las respuestas que se dieron en el ítem, a fin de que podamos verla con más claridad.

2.1.2 Ítem fácil que discrimina poco

La figura 4 muestra un ítem fácil que discrimina poco, porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta B, es alto en los grupos de habilidad media y alta (2, 3 y 4), mientras que en el grupo de habilidad más baja (1) es menor. Esto quiere decir que el ítem es muy fácil de contestar o que la respuesta es obvia.

En cuanto a los distractores, sus curvas de respuesta muestran que resultaron poco atractivos para los examinados. El distractor C no fue elegido por ningún examinado, ni en los grupos de alta habilidad (3 y 4), ni en los de baja habilidad (1 y 2). Las opciones A y D fueron elegidas, en cada caso, solo en una ocasión por un examinado del grupo más bajo (1); esto quiere decir que en general los distractores no resultaron atractivos para los examinados con poca habilidad en el examen, como debería ser.

Como consecuencia de lo anterior, el ítem apenas es capaz de separar ligeramente a los examinados que conocen el contenido que evalúa el ítem, de los examinados que lo conocen menos. Esto puede notarse porque la inclinación de la curva de la respuesta correcta B es buena, pero solo entre los grupos de habilidad 1 y 2, pero luego tiende a ser horizontal a medida que incrementa la habilidad.

En resumen, el ítem es fácil y discrimina entre los examinados un poco mejor que el ítem ilustrado en las figuras 2 y 3, por lo que apenas puede ser considerado adecuado.

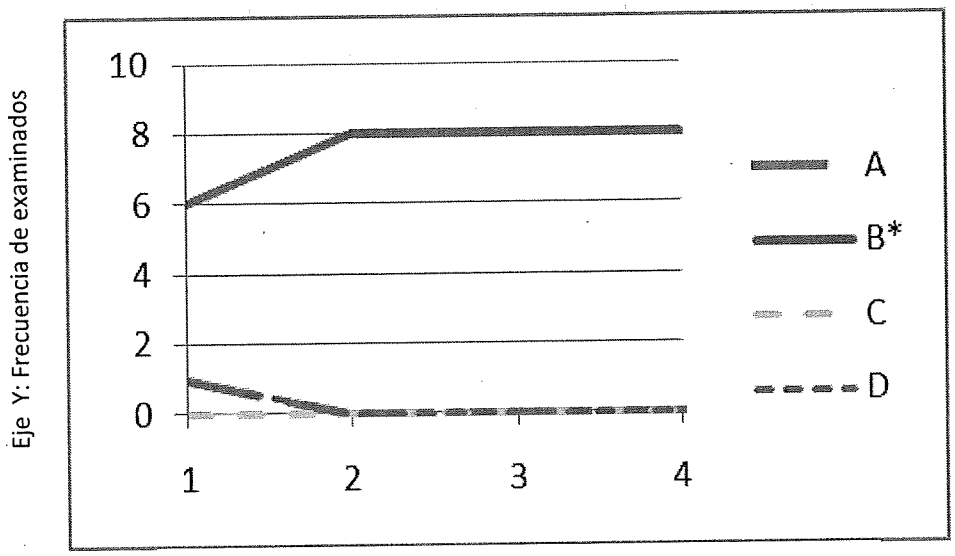


Figura 4. Ilustración de un ítem fácil que discrimina poco. (Eje Y = Nº de examinados)

En la figura 5 aparece el mismo ítem, pero ahora el eje Y muestra el porcentaje de examinados que eligieron las opciones, en vez del número de examinados que lo hicieron, como en la figura 4.

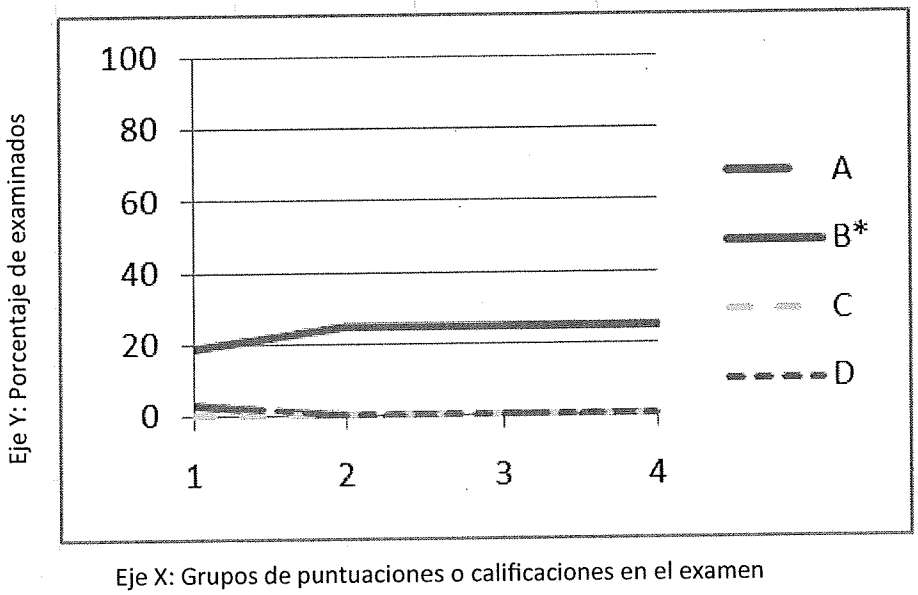


Figura 5. Ilustración de un ítem fácil que discrimina poco. (Eje Y = % de examinados)

2.1.3 Ítem fácil que discrimina bien

La figura 6 ilustra un ítem fácil que discrimina apropiadamente entre los examinados, porque la curva de la respuesta correcta en el ítem indica que el número de examinados que

eligieron la opción correcta D es alto en los grupos de habilidad 3 y 4, y empieza a incrementar desde el grupo de habilidad 2; pero en el grupo de habilidad 1 es menor. Esto quiere decir que el ítem es fácil solo para examinados con habilidad media y alta.

Respecto a los distractores, las curvas de respuestas de A y B indican que resultaron atractivos para algunos examinados de los grupos 1 y 2. El distractor C no resultó atractivo para ningún examinado. Esto quiere decir que los distractores fueron atractivos para examinados que no dominan el contenido evaluado, como debe ser el caso.

Por lo tanto el ítem es capaz de discriminar entre los examinados que dominan el contenido de los examinados que no lo dominan. Esto puede conocerse porque la inclinación de la curva de respuesta correcta incrementa a medida que aumenta el grupo de habilidad.

En resumen, el ítem resulta relativamente fácil y tiene una buena discriminación pues permite diferenciar a examinados que dominan el contenido de los que no lo dominan.

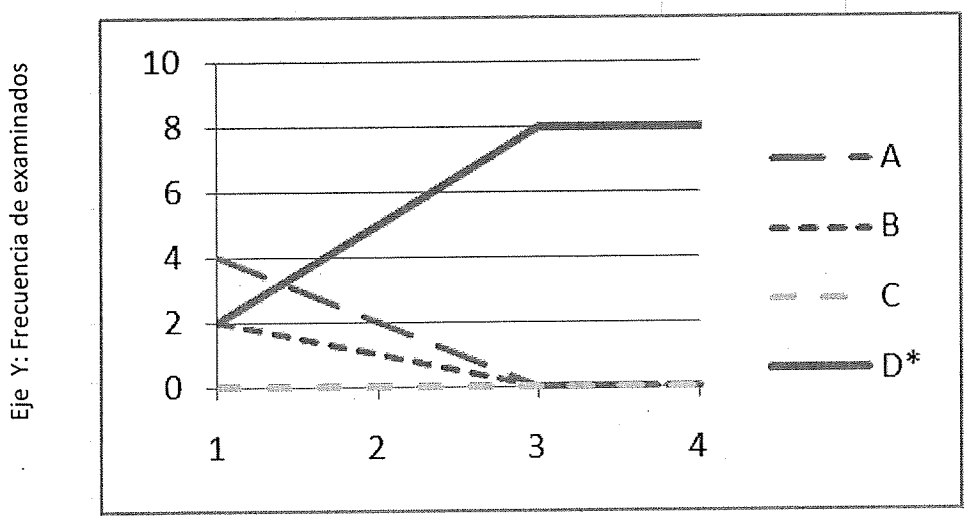


Figura 6. Ilustración de un ítem fácil con discriminación alta. (Eje Y = Nº de examinados)

En la figura 7 aparece el mismo ítem, pero ahora el eje Y muestra el porcentaje de examinados que eligieron las opciones, en vez del número de examinados que lo hicieron, como en la figura 6.

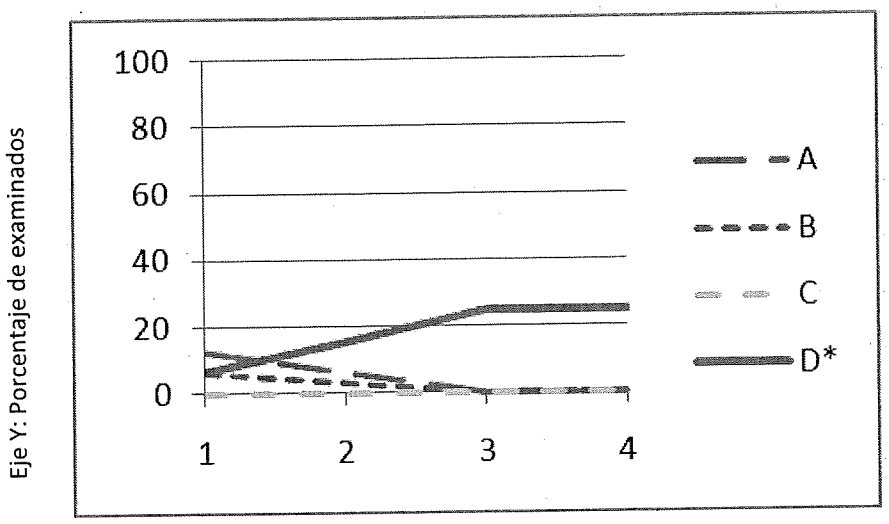


Figura 7. Ilustración de un ítem fácil con discriminación alta. (Eje Y = % de examinados)

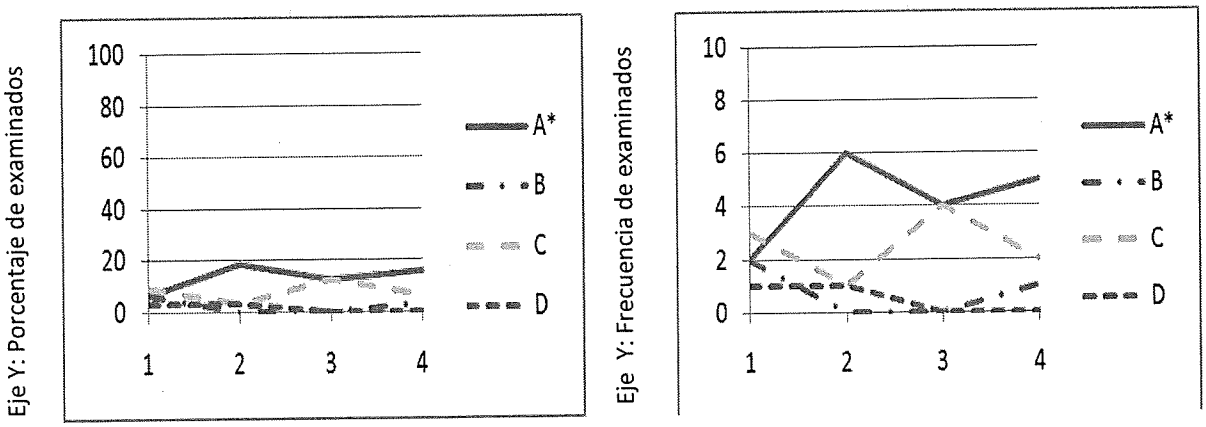
2.1.4 Ítem con dificultad media que no discrimina

La figura 8 presenta juntas las gráficas con frecuencia y porcentaje de examinados, y muestra un ítem con dificultad media que no discrimina, porque la curva de la respuesta correcta en el ítem deja ver que el número de examinados que eligieron la opción correcta A es un poco más de la mitad, pero fluctúa entre los grupos de habilidad; solo en el grupo de habilidad 2 incrementa y en los demás grupos se mantiene con altibajos. Esto quiere decir que el ítem es difícil de contestar para algunos grupos y para otros no; o que resulta confuso para los examinados con diferente habilidad.

Por otro lado, los distractores B y D parecen funcionar bien, aunque la opción B fue elegida por un examinado del grupo de habilidad 4. Además, el distractor C resultó atractivo para algunos de los examinados de mayor habilidad (grupos 3 y 4). Esto quiere decir, que esos distractores resultaron complejos para los examinados de los grupos de alta habilidad.

Como consecuencia de lo anterior, el ítem no permite discriminar con claridad entre los examinados que dominan el contenido que evalúa el ítem, y los examinados que no lo dominan. Esto puede distinguirse porque la pendiente o inclinación de la curva incrementa y decrementa de manera irregular entre los grupos de diferente habilidad (1, 2, 3 y 4).

En resumen, el ítem resulta con dificultad media pero no permite discriminar entre los examinados que dominan el contenido de los que no lo dominan, por lo que puede considerarse como inadecuado. Una acción que podría mejorar el ítem sería redactar la opción C de manera diferente, a fin de que no fuera atractiva para los grupos de habilidad alta (3 y 4).



Eje X: Grupos de puntuaciones o calificaciones en el examen (habilidad)

Figura 8. Ilustración de un ítem con dificultad media que no discrimina

2.1.5 Ítem con dificultad media que discrimina poco

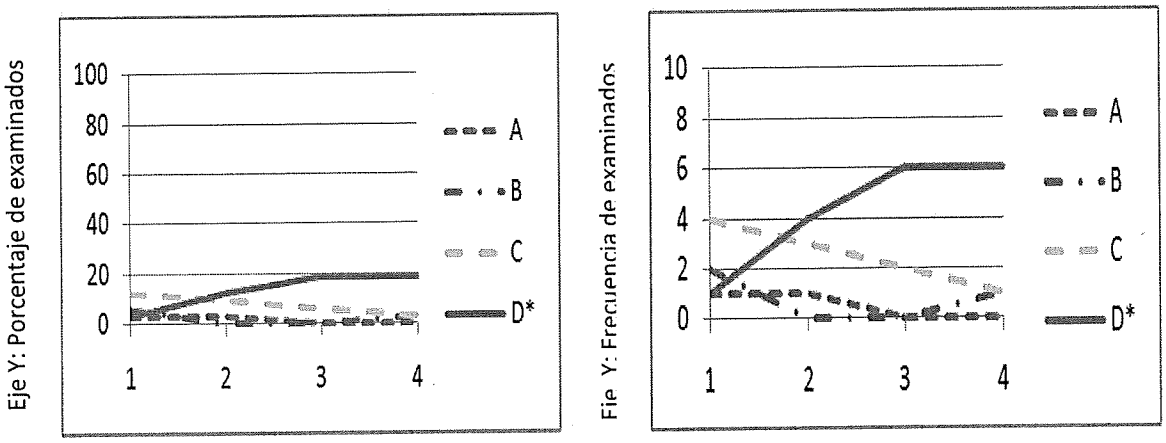
La figura 9 muestra un ítem con dificultad media que discrimina poco, porque la curva de la respuesta correcta en el ítem revela que los examinados que eligieron la opción correcta D se encuentran en todos los grupos de habilidad, aunque principalmente en los grupos 3 y 4, pero el distractor C es atractivo tanto para examinados que no dominan el contenido como para los que lo dominan. El ítem resulta de dificultad media porque los examinados se confunden entre la respuesta correcta D y el distractor C.

Por otro lado, el distractor A funciona bien pues es atractivo para los examinados que dominan poco el contenido, pero no lo es a medida que incrementa la habilidad. El distractor B parece funcionar bien, pero fue elegido por un examinado del grupo de habilidad 4.

Como consecuencia de lo anterior, el ítem permite discriminar entre los examinados que saben el contenido que evalúa el ítem de los examinados que no dominan el contenido.

Esto puede distinguirse porque la pendiente o inclinación de la curva de la respuesta correcta incrementa a medida que aumenta la habilidad.

En resumen el ítem resulta de dificultad media con una discriminación que es aceptable, por lo que puede ser considerado como mínimamente adecuado. Para mejorar el ítem se podría recomendar una redacción diferente del distractor C de manera que no sea atractivo para los examinados que dominan el contenido.



Eje X: Grupos de puntuaciones o calificaciones en el examen (habilidad)

Figura 9. Ilustración de un ítem con dificultad media que discrimina poco

2.1.6 Ítem con dificultad media que discrimina bien

La figura 10 muestra un ítem de dificultad media que discrimina bien, porque la curva de la respuesta correcta en el ítem permite observar que el número de examinados que eligieron la opción correcta C aumenta a medida que incrementa la habilidad. Esto quiere decir que el ítem es fácil pero solo para examinados que dominan el contenido.

Por otro lado, los distractores B y D parecen funcionar bien, porque sus curvas de respuestas decrementan a medida que aumenta la habilidad y distraen a examinados que no dominan el contenido. El distractor A también funciona bien, aunque confunde a un examinado del grupo de habilidad más alta. Esto da como resultado que los distractores funcionen en general correctamente porque distraen a examinados que dominan poco el contenido.

Como consecuencia de lo anterior, el ítem permite observar diferencias entre los examinados que dominan el contenido que evalúa el ítem de los que no lo dominan. Esto puede observarse porque la inclinación de la curva de la respuesta correcta se hace más pronunciada a medida que aumenta la habilidad.

En síntesis, el ítem resulta ser con dificultad media que discrimina bien entre los examinados que dominan el contenido de los que no lo dominan, por lo que es considerado como bueno.

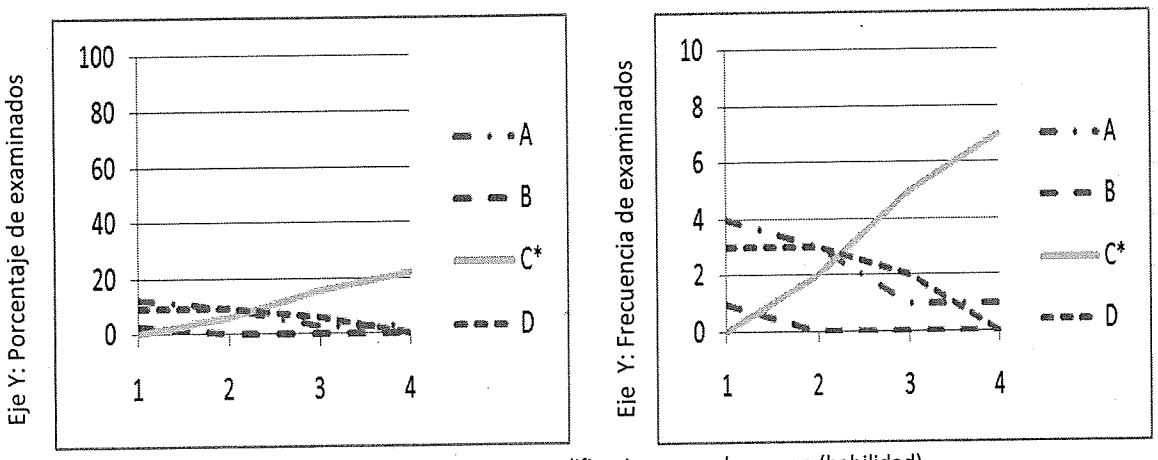


Figura 10. Ilustración de un ítem con dificultad media que discrimina bien

2.1.7 Ítem difícil que no discrimina

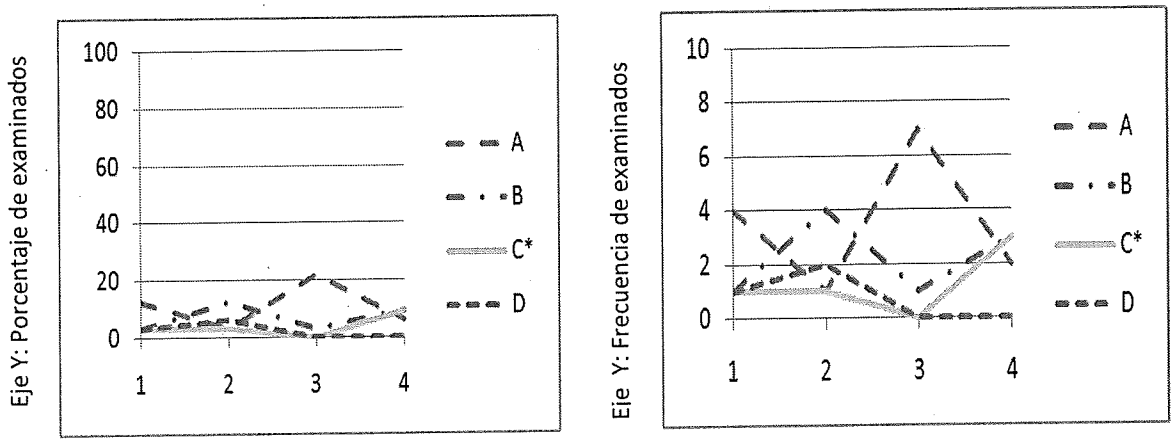
La figura 11 muestra un ítem difícil que no discrimina, porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta C es relativamente bajo y fluctúa entre los grupos de habilidad; solo en el grupo de habilidad 3 incrementa y en los demás grupos se mantiene con altibajos. Esto quiere decir que el ítem es difícil de contestar para algunos grupos y para otros no, o que resulta confuso para los examinados con diferente habilidad.

Por otro lado, el distractor D parece funcionar bien, porque disminuye la frecuencia o porcentaje de los examinados que lo eligen a medida que aumenta la habilidad. El distractor A funciona bien para el grupo de habilidad más baja (1), pero distrae considerablemente a los grupos de habilidad alta (3 y 4). La frecuencia o porcentaje de quienes eligen el distractor B

fluctúa dentro de los 4 grupos de habilidad, ocasionando que no disminuya a medida que aumenta la habilidad. Esto quiere decir, que los distractores resultaron complejos para los examinados de los grupos de habilidad alta.

Como consecuencia de lo anterior, el ítem no permite discriminar entre los examinados que dominan el contenido que evalúa el ítem de los examinados que no lo dominan. Esto puede distinguirse porque la pendiente o inclinación de la curva de la respuesta correcta fluctúa objetablemente entre los grupos de habilidad.

En síntesis, el ítem resulta ser difícil y no discrimina entre los examinados que dominan el contenido de los que no lo dominan, por lo que puede considerarse como inadecuado. Una acción que podría mejorar el ítem sería redactar de manera diferente los distractores A y B, a fin de que no fueran atractivos para los grupos de habilidad alta (3 y 4).



Eje X: Grupos de puntuaciones o calificaciones en el examen (habilidad)

Figura 11. Ilustración de un ítem difícil que no discrimina

2.1.8 Ítem difícil que discrimina poco

En la figura 12 se presenta un ítem difícil que tiene poca discriminación, porque la curva de la respuesta correcta muestra que el número de examinados que eligieron la respuesta correcta B incrementa poco a medida que incrementa la habilidad.

En cuanto a los distractores, por lo general resultaron atractivos tanto para los examinados que dominan el contenido (grupos 3 y 4), como para examinados que no lo dominan (grupos 1 y 2). La curva del distractor A incrementa en el grupo de habilidad 2 para

disminuir hasta el grupo de habilidad 3 pero vuelve a incrementar en el grupo de habilidad 4. Los distractores C y D presentan fluctuaciones entre los grupos de habilidad. Esto quiere decir, que los distractores resultaron por igual atractivos para examinados que dominan el contenido (lo que no debe suceder) y para examinados que no lo dominan (como debe ser).

En base a lo anterior, podemos decir que el ítem permite separar de manera mínima entre los examinados que comprenden el contenido que evalúa el ítem, de los que no lo comprenden. Esto puede observarse porque la pendiente de la curva de la respuesta correcta B incrementa en los grupos de habilidad alta (3 y 4).

En síntesis, el ítem resulta ser difícil y discrimina poco, por lo que puede ser considerado apenas adecuado. Para mejorar el ítem se recomienda redactar nuevamente los distractores A, C y D a fin de hacerlos menos atractivos para los que saben; o, en su caso, redactar la respuesta correcta de manera más contundente.

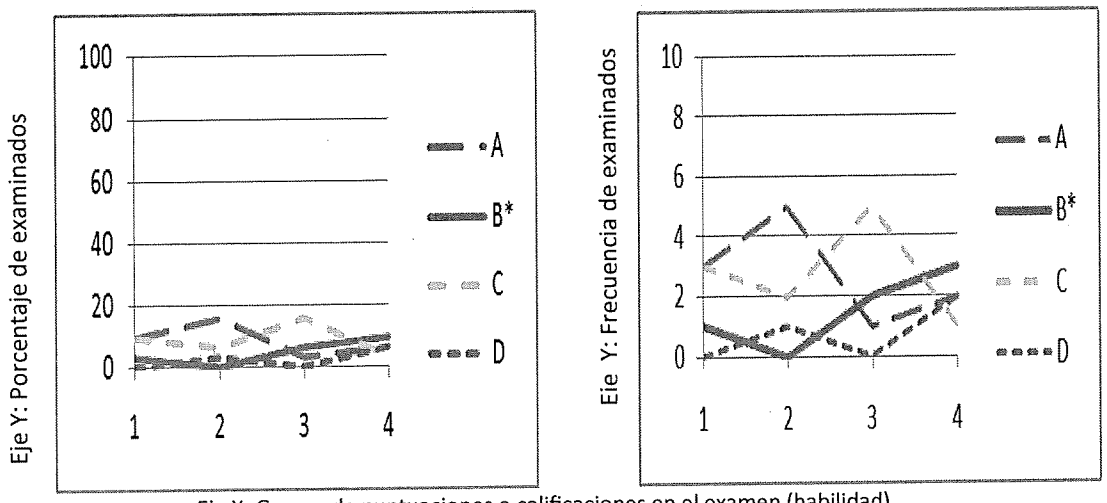


Figura 12. Ilustración de un ítem difícil que discrimina bien

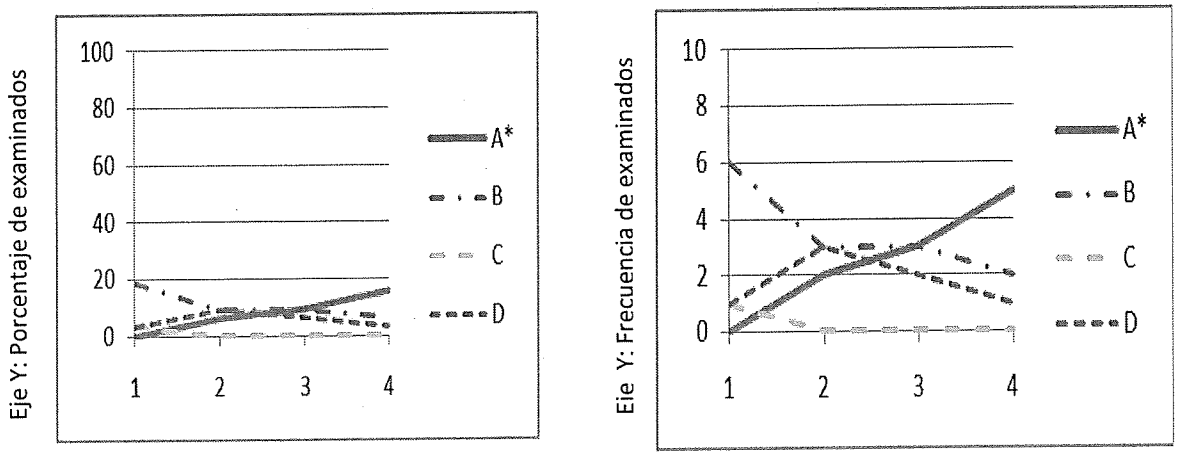
2.1.9 Ítem difícil que discrimina bien

La figura 13 presenta un ítem difícil que discrimina alto, porque la curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la respuesta correcta incrementa a medida que aumenta la habilidad.

En cuanto a los distractores, la opción C demostró ser poco atractiva porque solo la eligió un solo examinado. El distractor B fue atractivo para examinados de los grupos de habilidad baja (1 y 2), pero también resultó atractivo para 5 examinados de los grupos de habilidad alta (3 y 4). El distractor D cumple con la función de ser atractivo para examinados que no dominan el contenido; sin embargo, fue elegido por 3 examinados de los grupos de habilidad alta (3 y 4). Esto quiere decir que los distractores resultaron atractivos para los examinados de los diferentes grupos de grupos de habilidad, aunque para el grupo de habilidad 4 resultaron atractivos de manera mínima.

En consecuencia, el ítem permite discriminar con claridad entre los examinados que dominan el contenido de los examinados que no lo dominan. Esto puede comprobarse porque la tendencia de la curva de la respuesta correcta incrementa a medida que aumenta la habilidad.

En resumen el ítem es difícil pero discrimina alto, por lo que puede ser considerado un ítem adecuado. Los examinados que dominan el contenido lo responden correctamente.



Eje X: Grupos de puntuaciones o calificaciones en el examen (habilidad)

Figura 13. Ilustración de un ítem difícil que discrimina bien

2.2 Criterios para eliminar o corregir ítems

Los casos que se comentaron en esta sección, nos permiten apreciar que el AGI aporta mucha información significativa sobre la calidad de un ítem. Así mismo, nos permite conocer

las características que presentan los ítems que resultan inapropiados para evaluar el aprendizaje. En consecuencia, es posible proporcionar criterios que puedan utilizarse como una guía general para razonar sobre la posible eliminación o en su caso corrección de ítems defectuosos. Enseguida se presentan cuatro criterios principales que proponen Batenburg y Laros (2002) para determinar la eliminación o corrección de ítems:

- Tomar en cuenta el número de violaciones al supuesto de que la opción correcta incrementa con un aumento del puntaje total.
- Tomar en cuenta el número de violaciones al supuesto de que las opciones falsas decrezcan con un incremento del puntaje total.
- Tomar en cuenta el número de intersecciones entre las opciones correctas y falsas tras el inicio del intervalo de discriminación.
- La ausencia de poder discriminativo o una baja pendiente de la respuesta correcta.

No obstante lo anterior, considérese que al eliminar un ítem malo de una prueba debe ser reemplazado por otro ítem que cubra la misma parte del contenido aprendizaje.

3. PROGRAMA PARA EL ANÁLISIS GRÁFICO DE ÍTEMS

El Programa para el Análisis Gráfico de Ítems (PAGI), es un programa psicométrico gratuito y de código abierto, que fue desarrollado para efectuar el análisis gráfico de ítems basado en la Teoría Clásica de los Tests. PAGI es capaz de realizar estimaciones de la dificultad, la discriminación y el análisis de los distractores en los ítems de opción múltiple o de respuesta alterna de una prueba. Como hemos visto, estos análisis se pueden ejecutar e interpretar de manera sencilla con la ayuda de gráficos.

3.1 Requerimientos

PAGI requiere la instalación previa de cualquier versión de Windows XP o Windows 7 y que se cuente con la máquina virtual de Java, versión 1.6 o superior. PAGI puede leer archivos de Excel elaborados en versiones de Microsoft Excel 2003 y 2007.

3.2 Instalación

El archivo de instalación PAGI.jar, puede ser instalado desde una carpeta zip enviada por correo electrónico o desde la carpeta donde fue copiado previamente. Al ejecutar este archivo, el módulo de análisis psicométrico se configura hasta estar listo para realizar el análisis gráfico de ítems.

3.3 Creación e importación de datos

Independientemente de cómo se registren las respuestas de los examinados en la prueba (manualmente o mediante un lector óptico), para que PAGI pueda analizar dichos resultados es necesario que estén en una tabla de doble entrada procesada en Excel. Es decir, PAGI analiza matrices de datos con un formato donde las columnas representan los ítems de la prueba y las filas o renglones a los examinados que respondieron los ítems del instrumento. De esta manera, cada celda de la tabla representa la respuesta que dio un examinado a uno de los ítems del examen. Lo anterior puede ser observado en el ejemplo que se presenta en la tabla 1. Esta tabla es conocida como tabla de datos brutos, porque en ella los datos de los examinados derivados de las respuestas que dieron en el examen no han sufrido modificación alguna.

Tabla 1. Datos brutos que representan las respuestas de los examinados ante el conjunto de ítems de la prueba

Número de Ítem	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
Clave / examinado	A	C	B	A	D	B	C	A	D	B
1	A	C	B	A	D	B	C	A	D	B
2	#	C	C	A	D	B	C	A	D	A
3	A	C	D	C	A	B	D	A	C	B
4	D	C	B	A	B	B	C	C	D	B
5	A	B	B	#	D	D	C	D	#	#
6	C	C	B	A	D	A	A	A	A	B
7	D	#	B	A	D	B	C	A	D	B
8	B	C	B	D	C	B	C	A	C	B
9	A	C	B	A	A	B	C	B	A	C
10	A	C	B	A	D	A	C	A	D	B
11	A	B	D	A	D	#	#	#	#	#

Nótese que en la tabla, la primera fila corresponde a la clave de respuestas correctas para cada ítem del examen.

La tabla 1 presenta las opciones que eligieron los examinados en cada ítem de una prueba, mismas que pueden ser representadas con las letras A, B, C y D o con los números 1, 2, 3, y 4, cuando el ítem presenta cuatro opciones de respuesta. Por otro lado, en algunas celdas aparece el símbolo de **#**, lo cual significa que en ese ítem este examinado no contestó o no seleccionó ninguna de las opciones; estos datos son conocidos como datos perdidos. Por otro lado, cuando el examinado conteste otra opción diferente a alguna de las predeterminadas, su respuesta será identificada con el símbolo **$**; es decir, si el examinado elige otro carácter o número que no están marcados como opciones de respuesta se utiliza el símbolo anterior.

Nótese que es importante poner en la tabla los símbolos **#** y **$** cuando corresponda, para que PAGI pueda procesar adecuadamente nuestra base de datos.

3.4 Procesamiento de datos

Una vez que ha sido estructurada la base de datos en la tabla, tiene que ser depurada; es decir, se debe detectar a estudiantes e ítems inmedibles. Como ya se mencionó, es posible identificar a estudiantes que durante la aplicación de la prueba, no respondieron con seriedad. Estos estudiantes tienen que ser descartados para evitar una confusión en el análisis de los ítems de la prueba.

Para realizar el procesamiento de los datos, en este ejemplo se tomarán en cuenta los 10 primeros datos de la tabla 1, puesto que el examinado número 11 no respondió suficientes ítems y por ello fue depurado de la base de datos. Lo siguiente es convertir la tabla de datos brutos en una tabla de codificación de 1 y 0; es decir asignar el número 1 a los que contestaron correctamente y un 0 a los que tuvieron error en el ítem. El siguiente paso es calificar el examen y ordenar los puntajes de mayor calificación a menor, de acuerdo a las puntuaciones totales que obtuvieron los examinados. Esto es realizado por el programa PAGI.

Tabla 2. Codificación en 0 y 1 de los datos brutos, ordenados de mayor puntuación a menor

Número de Ítem	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	TOTAL
Clave / Examinado	A	C	B	A	D	B	C	A	D	B	
1	1	1	1	1	1	1	1	1	1	1	10
7	1	0	1	1	1	1	1	1	1	1	9
10	1	1	1	1	1	0	1	1	1	1	9
2	0	1	0	1	1	1	1	1	1	0	7
4	0	1	1	1	0	1	1	0	0	1	6
6	0	1	1	1	1	0	0	1	0	1	6
8	0	1	1	0	0	1	1	1	0	1	6
9	1	1	1	1	0	1	1	0	0	0	6
3	1	1	0	0	0	1	0	1	0	1	5
5	1	0	1	0	1	0	1	0	0	0	4

Hasta el momento el programa PAGI ha logrado realizar una codificación de datos en ceros y unos; ha realizado la sumatoria de todos los aciertos; ha ordenado de mayor a menor las puntuaciones en la prueba y ha dividido las calificaciones de los examinados en cuatro grupos de habilidad, según los aciertos que obtuvieron.

3.5 Programa para el análisis gráfico de ítems

3.5.1 La interface del PAGI

Para realizar el análisis gráfico de ítems por medio de la aplicación PAGI, es necesario conocer la estructura de la ventana principal, misma que se presenta en la figura 3.1. Como puede observarse, los componentes que la integran se encuentran enumerados de manera consecutiva y en cada caso se describe su función más abajo.

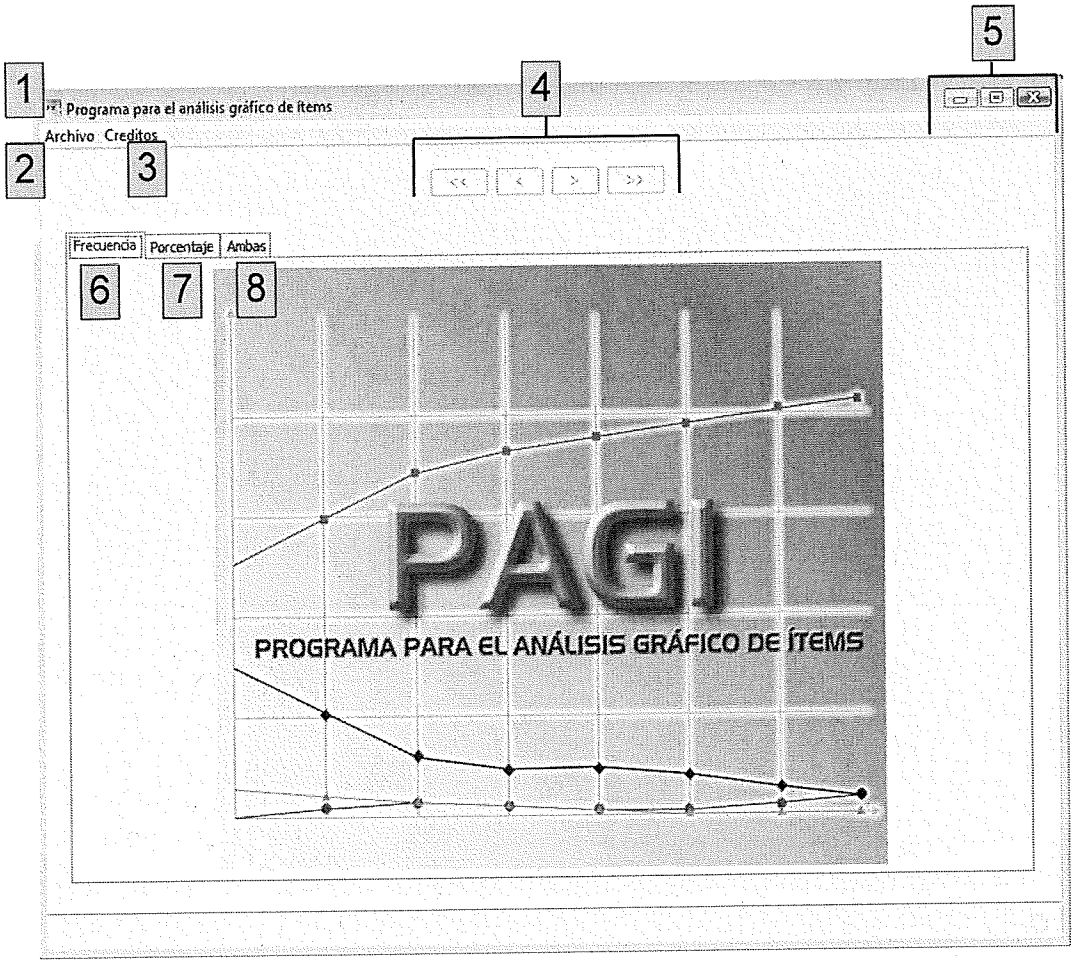


Figura 3.1 Ventana principal del Programa para análisis gráfico de ítems

1. Título de la ventana principal.

2. Menú Archivo que ofrece dos opciones: una que permite seleccionar el archivo que se va a analizar y otra para salir del programa.

3. Menú de Ayuda, que presenta información general sobre el programa PAGI.

4. Controles que sirven para navegar entre los ítems del examen. Por ejemplo, al dar clic en el signo > se muestra el siguiente ítem y así sucesivamente.

5. Controles que sirven para minimizar, maximizar o cerrar la aplicación.

6. Control que permite observar la gráfica que presenta PAGI, con el eje Y en la modalidad de frecuencia de examinados.

7. Control que permite observar la gráfica que presenta PAGI, con el eje Y en la modalidad de porcentaje de examinados.

8. Control que permite observar simultáneamente ambas gráficas de la aplicación PAGI.

3.5.2 Entrada de archivos

En la figura 3.2 se muestra la entrada de los datos, en la parte superior izquierda del menú Archivo se muestra una opción señalada como "Abrir archivo". Por medio de ella se puede buscar el archivo donde se encuentre la base de datos en Excel que contiene los datos brutos del examen.

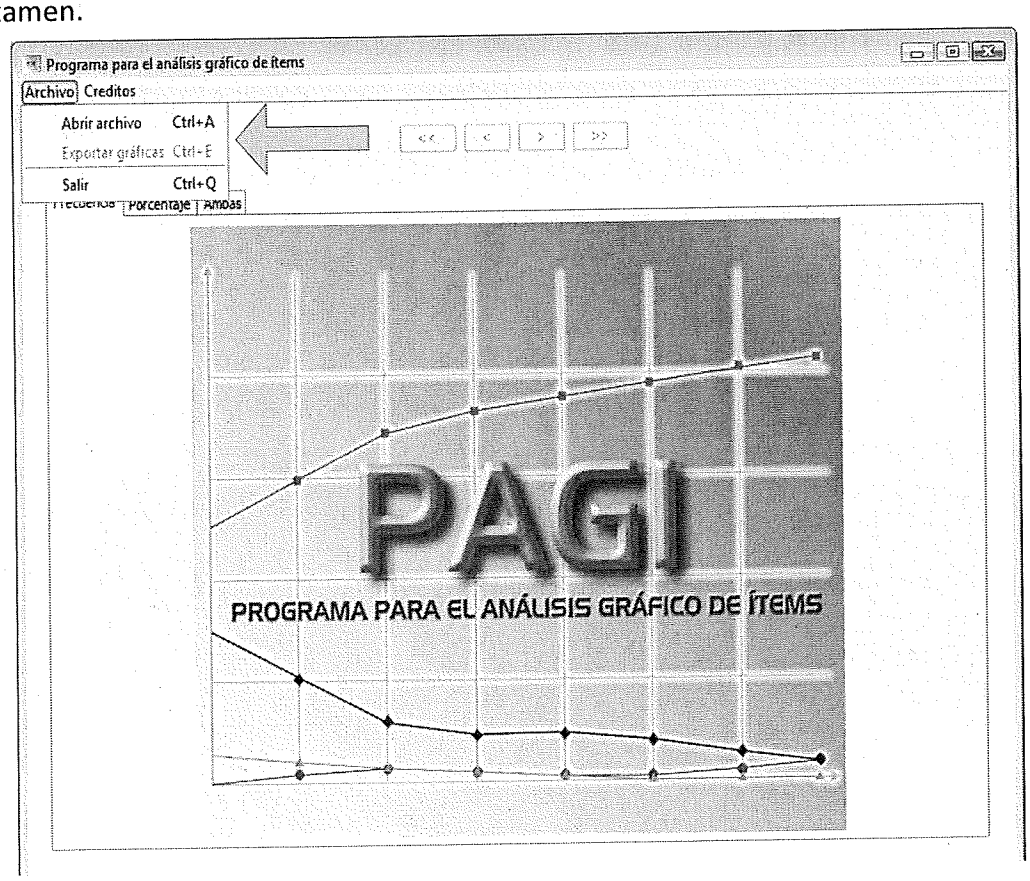


Figura 3.2 Entrada de datos del PAGI

A continuación se abre una ventana que permite realizar la búsqueda del archivo, como se muestra en la figura 3.3.

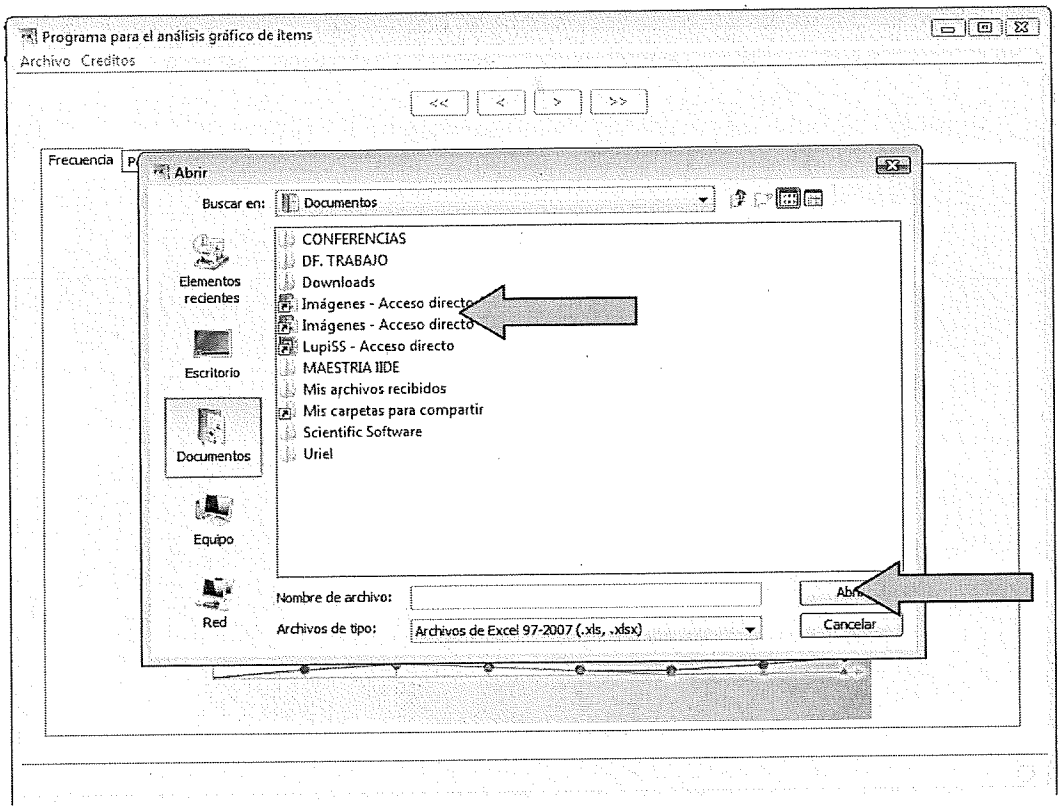


Figura 3.3 Búsqueda de archivo

Una vez localizado, se selecciona el archivo en formato .xls o xlsx. según la versión de Microsoft Office Excel 2003 o 2007 que se haya utilizado para crear el archivo.

Posteriormente, se da clic en el control Abrir, como se muestra en la figura 3.4. Si es archivo tiene el formato adecuado abrirá una ventana como la que se muestra enseguida:

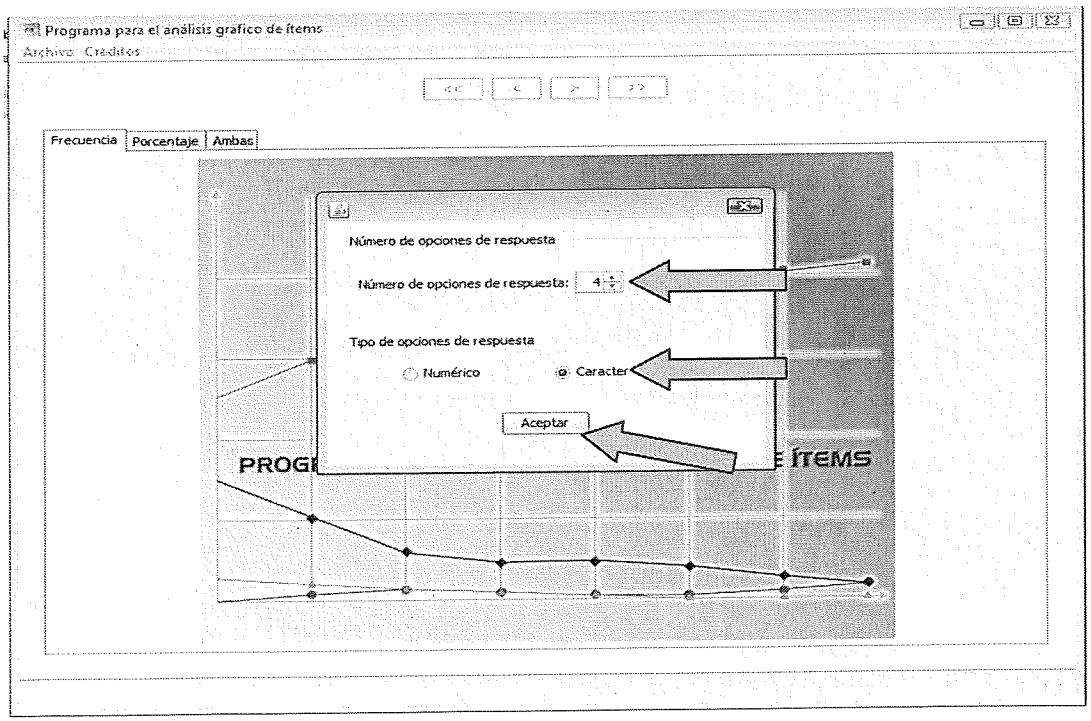


Figura 3.4 Ventana de configuración detallada

En la figura 3.4 se solicita el número de opciones de respuesta que tiene cada ítem, que generalmente son 4 o 5. A continuación se indica el tipo de opciones de respuesta que se ofrecieron al examinado, mismo que puede ser "Caracter" en caso de que las opciones de respuesta hayan sido letras (A, B, C, D), o "Numérico" si fueron números (1, 2, 3, 4). Una vez seleccionados ambos elementos de configuración se da clic en "aceptar".

Inmediatamente después aparece nuevamente la ventana principal de PAGI (ver figura 3.5), pero ahora ya se pueden visualizar las gráficas necesarias para efectuar el análisis gráfico de cada ítem. Por omisión, PAGI presenta primero la gráfica con frecuencia.

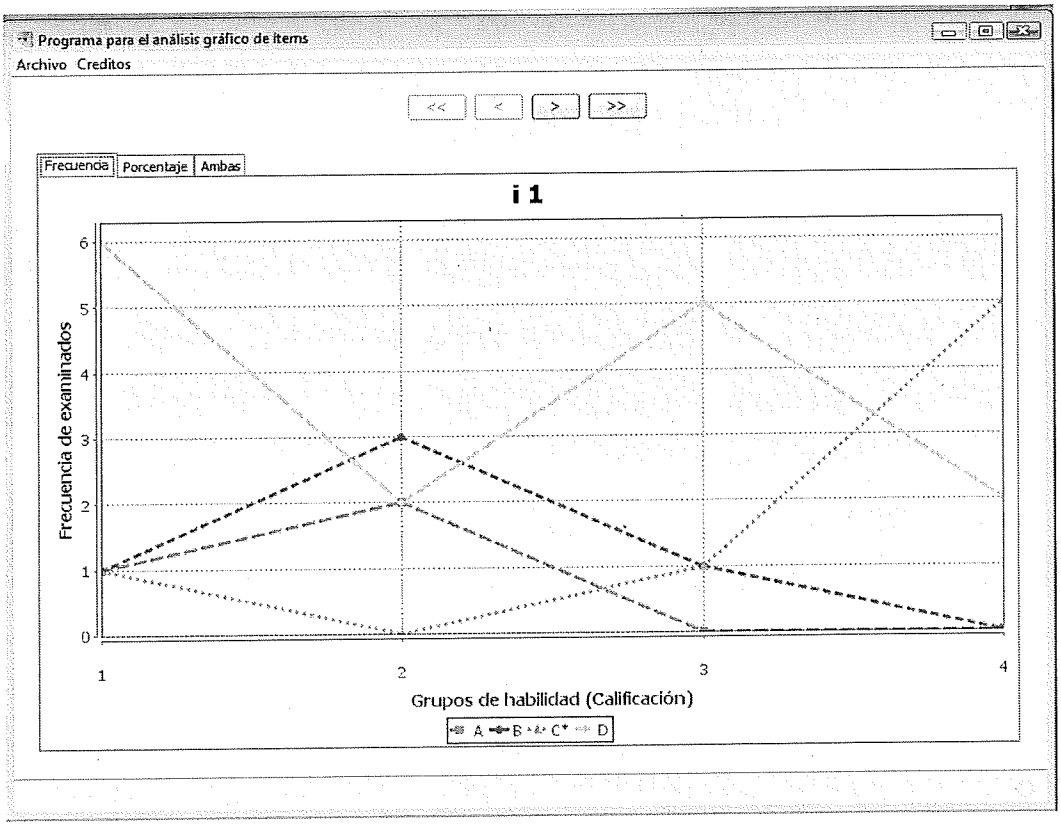


Figura 3.5 Gráfica con frecuencia de examinado

En la parte superior de la figura 3.5 aparecen los controles << < > >> que sirven, como ya se dijo, para navegar entre los ítems del examen. Al dar clic en el signo > se muestra el siguiente ítem y así sucesivamente los demás. PAGI es flexible y permite al usuario regresar o avanzar a observar los ítems anteriores o posteriores por medio de estos controles. Las

flechas dobles que se muestran permiten al usuario ir de manera directa al ítem final o al inicial.

La figura 3.6 muestra la gráfica con porcentaje de examinados. Así, se puede observar las similitudes y diferencias entre las gráficas de frecuencia y porcentaje de examinados que seleccionaron cada opción.

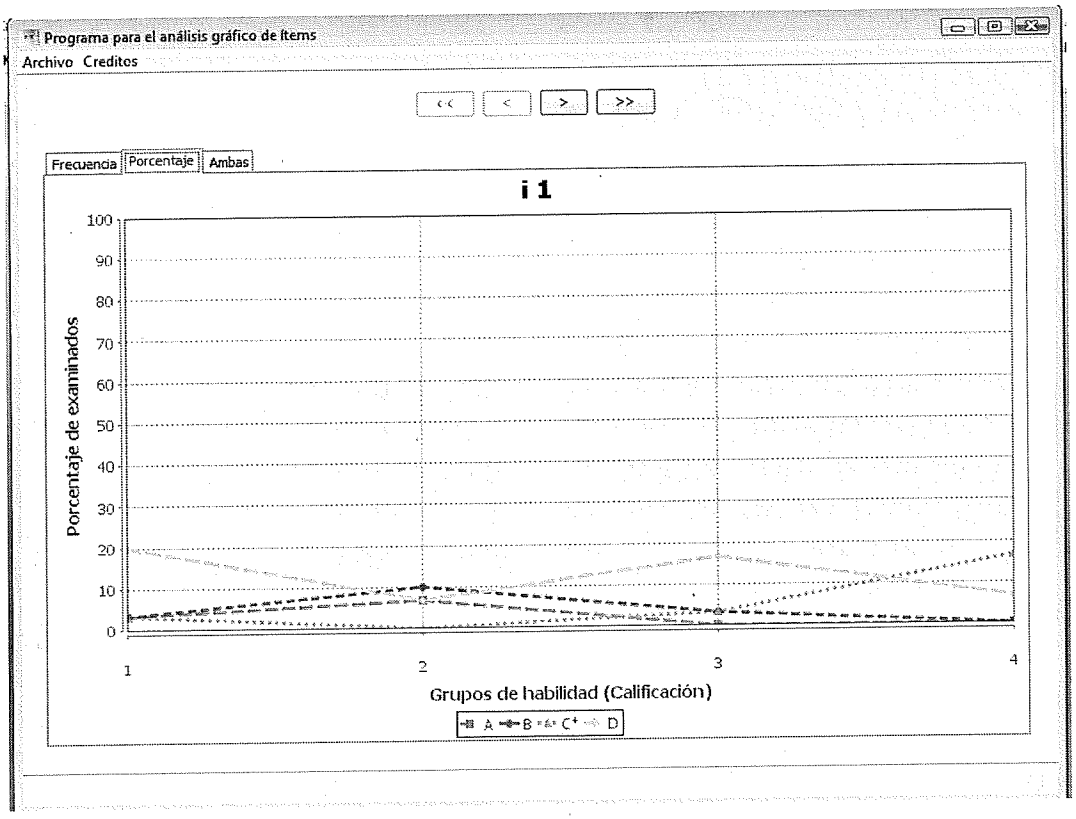


Figura 3.6 Gráfica con porcentaje de examinados

En la última pestaña (figura 3.7) se muestran juntas ambas gráficas: la de frecuencia y la de porcentaje. Esta gráfica compuesta permite al usuario observar al mismo tiempo y comparar la visión panorámica de las curvas de respuestas que se dieron en el ítem y la visión detallada de las mismas. Lo anterior se observa en la figura 3.7 que se presenta enseguida.

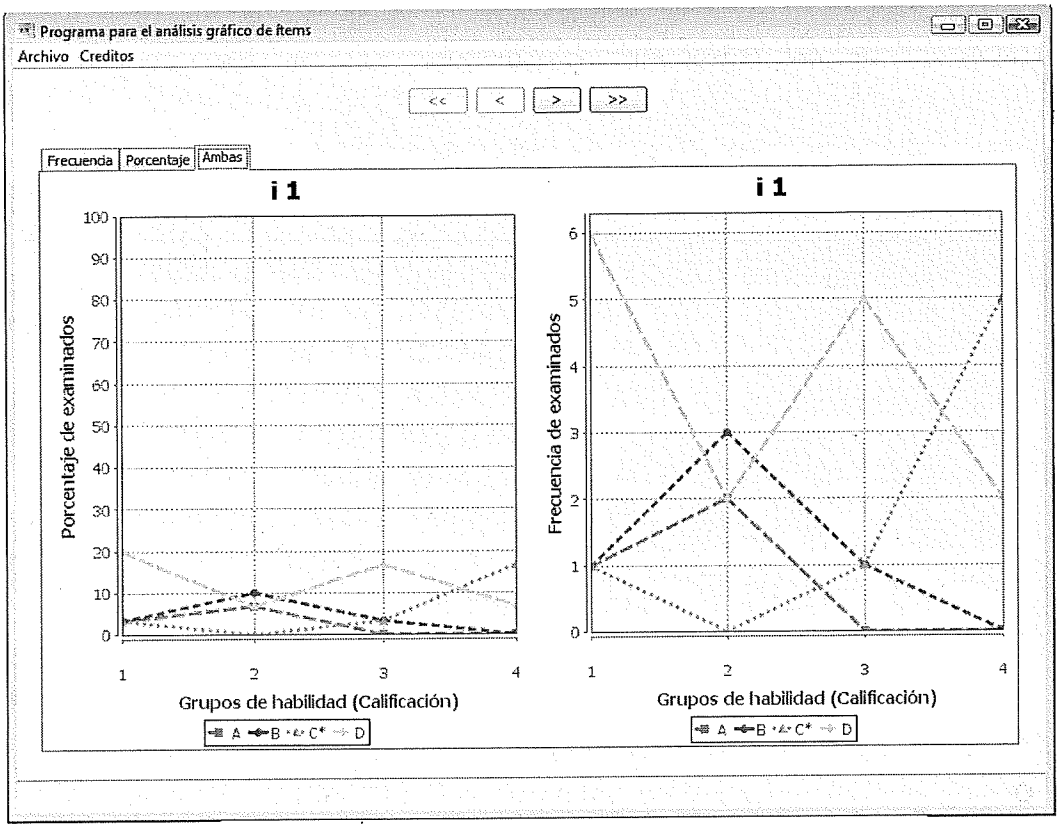


Figura 3.7 Pestaña donde se muestran ambas gráficas

Una vez que se ha concluido el análisis y la interpretación de todos los ítems del examen, mediante el análisis gráfico de ítems, en el menú de "Archivo" se encuentra la opción Salir para finalizar el análisis, como lo muestra la figura 3.8

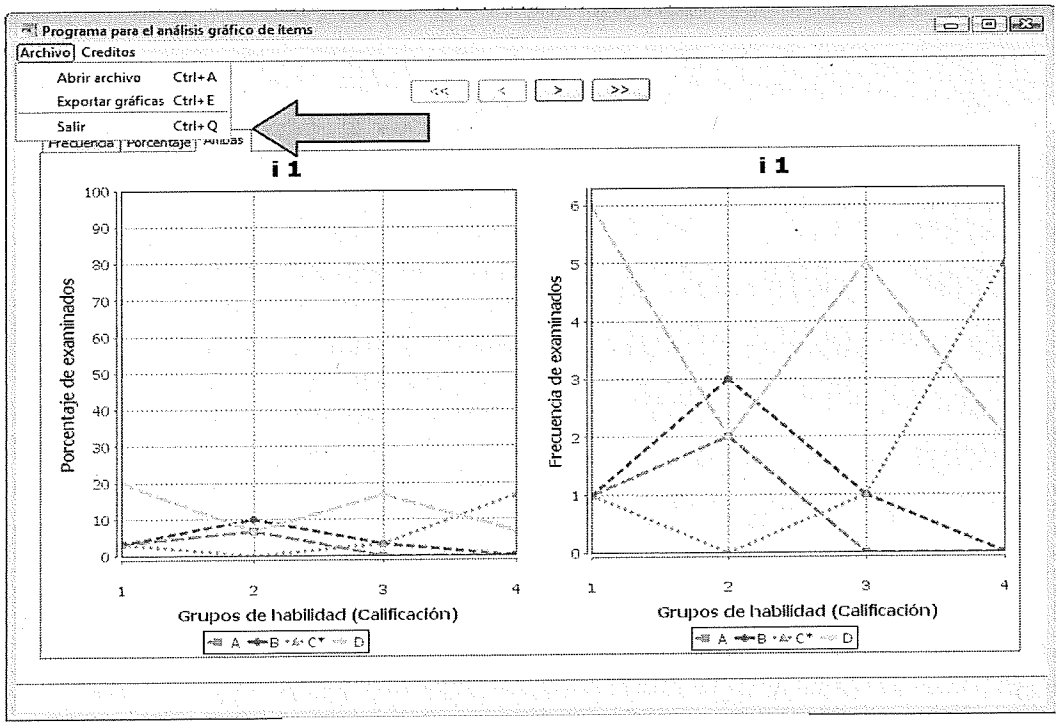


Figura 3.8 Ruta para salir del programa

4. GLOSARIO

Ítem. Los ítems son comúnmente identificados como preguntas o reactivos. Por ejemplo, una prueba de lectura consiste en cinco pasajes de texto y cuatro preguntas deben ser contestadas sobre cada pasaje. Se concibe a las veinte preguntas como veinte ítems.

Dificultad o índice de dificultad. También conocido como Valor *p* del ítem. Es importante tener en cuenta la dificultad de los ítems ya que en las respuestas aportadas por los estudiantes no se proporciona su nivel de habilidad. Para ello, la inclusión o exclusión de ítems está basada en su grado de dificultad. Se llama valor *p* porque se trata de la proporción o probabilidad de acertar el ítem.
Para ítems binarios (los calificados 0/1; es decir, correcto o incorrecto), el valor *p* de un ítem es la proporción de respuestas correctas en la población examinada, y se obtiene dividiendo el número de examinados que tuvieron correcto el ítem entre el total de quienes lo respondieron. Se considera el valor *p* como una propiedad del ítem respecto a una sola población. Por ejemplo, se desarrolla un ítem para una prueba de cuarto grado de español que será aplicada. Asumimos que el ítem es fácil porque tiene un valor *p* de 0.8, lo que quiere decir que el contenido es comprendido con facilidad. Incluso puede ser contestado por la población de segundo grado, donde su valor p es de 0.25. Así, se considera el valor *p* en referencia a alguna población.
Nota 1. Los valores *p* son valores que pertenecen a ítems en alguna población y se analizan en una muestra. La muestra no es igual que la población. Si se computa el valor *p* de un ítem en dos muestras independientes, se encontrará dos valores diferentes. El valor *p* encontrado en la muestra es considerado como una estimación del valor *p* en la población, para ello la exactitud depende del tamaño de la muestra.
Nota 2. El valor *p* se interpreta como medida de dificultad; sin embargo, mientras más alto sea el valor *p* el ítem es más fácil. Por ello es conocido también como el índice de facilidad.

Índice de discriminación del ítem. El índice de discriminación de un ítem permite separar los niveles de habilidad alta de los niveles bajos con base en las respuestas al ítem. En términos psicométricos define cuál es la calidad psicométrica de una prueba que tiene este ítem en particular. Por ejemplo, se usa un ítem binario difícil en una prueba, por ello se observa que el ítem discrimina a los estudiantes que tienen el ítem correcto de los que no. Pero el ítem binario tiene sólo dos categorías

(correcto e incorrecto), si el ítem separa a los buenos de los otros, no separa a los estudiantes con habilidad media de los débiles.

La discriminación es una propiedad local. Su poder discriminativo en el ítem se representa con un solo número. Para ello, en la TCT existen varios índices de discriminación:

1. La diferencia entre el índice de dificultad del reactivo para el grupo alto, formado por los examinados que obtuvieron las calificaciones más altas en la prueba, digamos el 27% de ellos, y el índice de dificultad del ítem para el grupo bajo, formado por los examinados que obtuvieron las calificaciones más bajas en la prueba, digamos el 27% de ellos (discriminación mediante grupos constrastados).
2. La correlación entre la puntuación en el ítem y la puntuación total en la prueba (la correlación ítem – test).
3. La correlación entre la puntuación en el ítem y la puntuación total en la prueba con ese ítem excluido (correlación ítem-resto).
4. En particular para los ítems de opción múltiple: la correlación entre la puntuación en la prueba y cada uno de los distractores (la correlación opción – total).

Correlaciones entre los distractores y los puntajes en la prueba. La correlación biserial puntual está compuesta por un coeficiente de correlación que debe ser positivo; esto quiere decir que los estudiantes que seleccionaron la opción correcta deben obtener los puntajes totales más altos. En el caso de los distractores, su correlación con la puntuación total en la prueba debe ser negativa, es decir, los estudiantes que seleccionan las opciones incorrectas son quienes obtienen un puntaje total bajo en la prueba. Este hecho, diferencia a los que saben de los que no (Van Batenburg y Laros, 2002). Para calcular la correlación se necesitan dos series de puntajes. Por ejemplo: al calcular la correlación ítem-test. Una puntuación es el puntaje obtenido por los examinados en la prueba y la otra puntuación es la obtenida por los examinados en el ítem.

La última es igual a uno, si la respuesta es correcta; y es cero si la respuesta es incorrecta. Para calcular la correlación entre un distractor y el puntaje en la prueba, se recodifican las respuestas dadas en la prueba por los examinados. Suponiendo que el ítem bajo estudio es uno de opción múltiple con cuatro opciones de respuesta (A, B, C y D), y la opción B es la respuesta correcta; esto significa que se da un puntaje en el ítem de uno a cada examinado que escogió B, y un puntaje de cero a los demás. Para calcular la correlación entre el puntaje en la prueba y el distractor A, se debe crear una variable binaria nueva, y asignar un puntaje de uno a cada examinado que escogió A y

poner un cero a los demás. La correlación que se busca es la correlación entre una nueva variable y la puntuación obtenida por los examinados en la prueba. Por eso, se da a conocer la necesidad de almacenar las observaciones originales o puntajes brutos. Si uno guarda sólo los puntajes en el ítem (los ceros y unos), no es posible calcular la correlación entre el distractor y el puntaje en el test, pues es imposible saber cuál de los distractores fue escogido a partir del conocimiento de que la respuesta no es correcta.

5. REFERENCIAS

Batenburg, T. y Laros, J. (2002). Data Screening and Graphical Analysis of Items *Universidad de Groningen, en Holanda y Universidad de Brasilia, Brasil.*

Batenburg, T. y Laros, J. (2002). Graphical Item Analysis. *Education Research and Evaluation (8)* 3

Contreras, L. A. (2001). Procedimiento para Asegurar la Validez de Contenido de una Prueba Criterial: el Caso de un Examen de Español para la Educación Primaria en Baja California. *Memoria del VI Congreso Nacional de Investigación Educativa.* Manzanillo, México. COMIE. 2001.

Encinas, J. A., Rivera, R. E. y Contreras, L. A., (2005). Evaluación colegiada del aprendizaje en la Universidad Autónoma de Baja California: Construcción de un examen criterial de gran escala para evaluar el dominio de conceptos y procedimientos del cálculo diferencial. *Memoria del VIII Congreso Nacional de Investigación Educativa*. México. COMIE.

Verhelst, N. (2004). Sección C. Teoría clásica de los tests, en reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework. Strasbourg. Language Policy Division.

ANEXO 4

Presentación de la capacitación

CURSO- TALLER PARA EL ANÁLISIS GRÁFICO DE ÍTEMS POR MEDIO DE LA APLICACIÓN PAGI

LIC. GUADALUPE DE LOS SANTOS LÁZARO

IIDE - UABC

ENSENADA, B.C. JUNIO DE 2010

Objetivo

- Al finalizar el curso – taller, los docentes en servicio serán capaces de efectuar el análisis gráfico de ítems, para mejorar sus pruebas de aula.

Teoría clásica de los tests (TCT)

- X = V + E
- Donde X es la puntuación observada en la prueba, V es la habilidad real del estudiante, y E es el error de medición.
- TCT nos ayuda a reducir los errores que cometemos al planear, construir, aplicar o calificar un examen, de modo que sea posible conocer el nivel de habilidad real de los estudiantes.
- Para ello puede utilizarse, entre otros, el método del Análisis Gráfico de Ítems (AGI).

Análisis Gráfico de Ítems (AGI)

- Propuesta viene de Theo A. Van Batenburg y Jacob A. Laros en 2002; fue desarrollada en la Universidad de Brasilia.
- Los ítems son comúnmente identificados como preguntas o reactivos.
- El AGI despliega visualmente la relación entre la puntuación total en la prueba y el total de los examinados que eligieron la opción correcta y las opciones falsas en un ítem de opción múltiple.

Análisis Gráfico de Ítems (AGI)

- Con el AGI los ítems de mala calidad son fáciles de detectar, porque en ellos se observa que el número o porcentaje de examinados que eligieron la opción correcta, disminuye a medida que incrementa el puntaje o total de aciertos en la prueba; o bien, porque muestra en una o más opciones falsas (distractores) que no disminuye el número o porcentaje de estudiantes que las eligieron, al incrementar el puntaje total en la prueba.
- La importancia de este método radica en ser utilizado para identificar ítems que presentan fallas y que deben ser excluidos de las pruebas por no tener los mínimos requerimientos de calidad técnica.

Análisis Gráfico de Ítems (AGI)

- El AGI proporciona información esencial y de fácil interpretación acerca de las características técnicas del ítem como son su dificultad, su poder de discriminación y el nivel de adivinación.
- El AGI plantea que al aplicar un examen a un grupo de alumnos hay presencia de error, el cual regularmente se da porque algunos examinados no tomaron en serio la aplicación de la prueba.

Análisis Gráfico de Ítems (AGI)

- Debe eliminarse a los estudiantes que respondieron con patrones de respuesta de una sola categoría; como el responder a varios ítems consecutivos eligiendo siempre la opción A, A, A, A, A, A... o B, B, B, B, B, B...
- También es posible identificar patrones de respuesta de los examinados que son repetitivos; por ejemplo, cuando responden con un patrón como ABCDE, ABCDE... EDCBA, EDCBA... ACE, ACE... o BD, BD, ... En todos esos casos, se sugiere eliminar a estos examinados de la base de datos.
- A continuación se describen los componentes principales del Análisis Gráfico de Ítems:

Análisis Gráfico de Ítems (AGI)

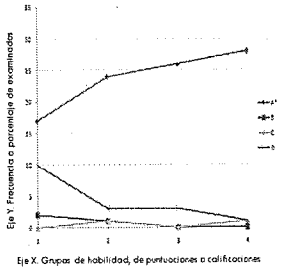
- El eje Y muestra la frecuencia o porcentaje de examinados, que respondieron el ítem.
- En el eje X se observan grupos de habilidad. Que se dividen en 4 grupos de acuerdo a la puntuación que obtengan en la prueba:

Grupo 1: bajo

Grupo 2: medio bajo

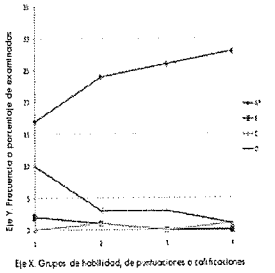
Grupo 3: medio alto

Grupo 4: alto



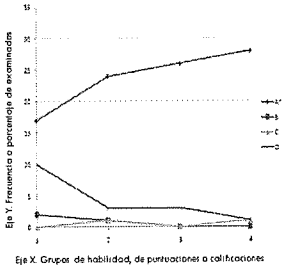
Análisis Gráfico de Ítems (AGI)

- En el eje X, los grupos de habilidad se ubican de izquierda a derecha, es decir del grupo bajo al grupo más alto.
- En este caso, en cada grupo de habilidad hay 30 examinados, porque el ítem fue aplicado a una muestra de 120 estudiantes.
- Para nuestro análisis gráfico de ítems utilizaremos una muestra pequeña, la cual vamos a dividir entre 4 grupos de habilidad.



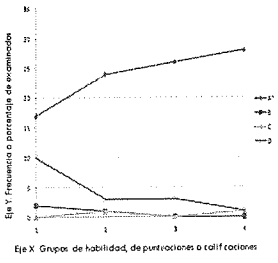
Análisis Gráfico de Ítems (AGI)

- Sección de código que identifica las opciones del ítem.
- Las líneas que se observan en el eje de coordenadas son las curvas de respuestas al ítem.



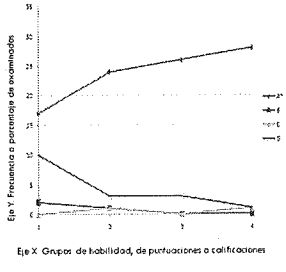
Análisis Gráfico de Ítems (AGI)

- Este ítem es idealmente bueno, porque la curva de la respuesta correcta aumenta a medida que incrementa la habilidad.
- Por otro lado, las líneas que están marcadas como B, C y D; son llamados distractores. Los cuales cumplen con su función en distraer a los estudiantes que no dominan el contenido.



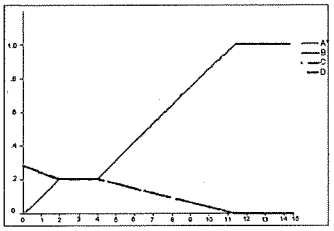
Análisis Gráfico de Ítems (AGI)

- Es un ítem idealmente bueno porque la curva de la respuesta correcta aumenta conforme se incrementa la habilidad y los distractores disminuyen cuando incrementa la habilidad.
- Este principio se aplica para los ítems que son considerados como aceptables en una prueba.



Tasa de adivinación

- En un ítem de opción múltiple con cuatro opciones, se espera que los examinados que no saben contesten al azar. Así, tendrán .25 de probabilidad de acertar.



Análisis Gráfico de Ítems (AGI)

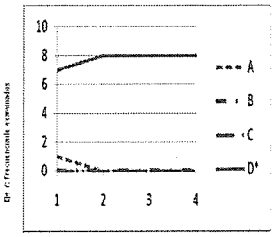
- Para entender mejor el AGI, se realizó un experimento en Excel para observar las posibles tendencias de la curva de la respuesta correcta, por lo que se determinaron 9 casos:

Fácil	✓	✓	✓
Medio	✓	✓	✓
Difícil	✓	✓	✓

Análisis Gráfico de Ítems (AGI)

Ítem fácil que no discrimina

- La curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta D es alto, tanto en los grupos de mayor habilidad (3 y 4), como en los de menor habilidad (1 y 2).
- Las opciones A, B y C, conocidas como distractores, en general no resultan atractivas para los examinados.

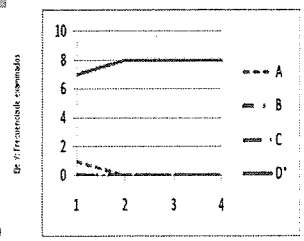


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem fácil que no discrimina

- Los distractores B y C no fueron elegidos por ningún examinado, ni en los grupos de alta habilidad (3 y 4), ni en los de baja habilidad (1 y 2); y la opción A fue elegida por un solo examinado del grupo 1.
- El ítem no permite discriminar entre los examinados que dominan el contenido que evalúa el ítem de los examinados que no lo dominan.

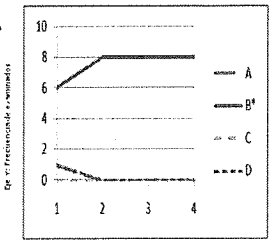


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem fácil que discrimina poco

- La curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta B, es alto en los grupos de habilidad media y alta (2, 3 y 4), mientras que en el grupo de habilidad más baja (1) es menor.
- Los distractores muestran que sus curvas de respuesta resultaron poco atractivos para los examinados.
- Los distractores no resultaron atractivos para los examinados con poca habilidad en el examen, como debería ser.

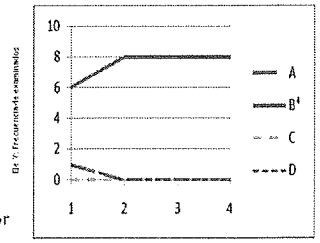


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem fácil que discrimina poco

- El ítem apenas es capaz de separar ligeramente a los examinados que conocen el contenido que evalúa el ítem, de los examinados que lo conocen menos.
- El ítem es fácil y discrimina entre los examinados un poco mejor, por lo que apenas puede ser considerado adecuado.



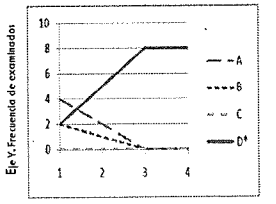
Eje X. Grupos de puntuaciones o calificaciones en el examen

Presentación de la capacitación

Análisis Gráfico de Ítems (AGI)

Ítem fácil que discrimina bien

- La curva de la respuesta correcta en el ítem indica que el número de examinados que eligieron la opción correcta D es alto en los grupos de habilidad 3 y 4, y empieza a incrementar desde el grupo de habilidad 2; pero en el grupo de habilidad 1 es menor.
  - Para los distractores, las curvas de respuestas de las opciones A y B indican que resultaron atractivos para algunos examinados de los grupos 1 y 2. El distractor C no resultó atractivo para ningún examinado.

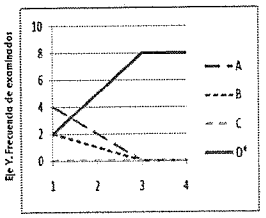


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem fácil que discrimina bien

- El ítem es capaz de discriminar entre los examinados que dominan el contenido de los examinados que no lo dominan. Esto puede conocerse porque la inclinación de la curva de respuesta correcta incrementa a medida que aumenta el grupo de habilidad.
- El ítem resulta relativamente fácil y tiene una buena discriminación pues permite diferenciar a examinados que dominan el contenido de los que no lo dominan.

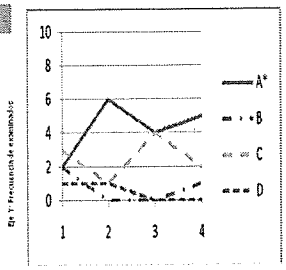


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que no discrimina

- La curva de la respuesta correcta A en el ítem deja ver que el número de examinados que la eligieron es un poco más de la mitad, pero fluctúa entre los grupos de habilidad y solo en el grupo de habilidad 2 incrementa y en los demás grupos se mantiene con altibajos.
- Los distractores B y D parecen funcionar bien, aunque la opción B fue elegida por un examinado del grupo de habilidad 4.

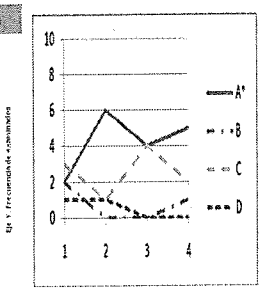


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que no discrimina

- El ítem no permite discriminar con claridad entre los examinados que dominan el contenido que evalúa el ítem, y los examinados que no lo dominan.
- El ítem resulta con dificultad media pero no permite discriminar entre los examinados que dominan el contenido de los que no lo dominan, por lo que puede considerarse como inadecuado.

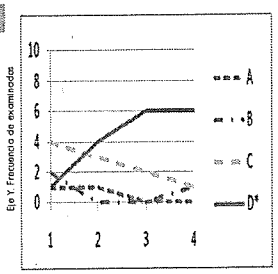


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que discrimina poco

- La curva de la respuesta correcta D en el ítem revela que los examinados que la eligieron, se encuentran en todos los grupos de habilidad, aunque principalmente en los grupos 3 y 4, pero el distractor C es atractivo tanto para examinados que no dominan el contenido como para los que lo dominan.
- El distractor A funciona bien pues es atractivo para los examinados que dominan poco el contenido, pero no lo es a medida que incrementa la habilidad.

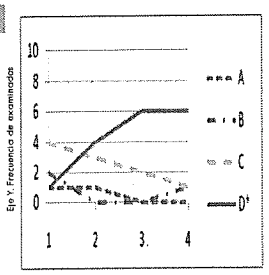


Eje X. Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que discrimina poco

- El ítem permite discriminar entre los examinados que saben el contenido que evalúa el ítem de los examinados que no dominan el contenido. Esto puede distinguirse porque la pendiente o inclinación de la curva de la respuesta correcta incrementa a medida que aumenta la habilidad.
- El ítem resulta de dificultad media con una discriminación que es aceptable, por lo que puede ser considerado como mínimamente adecuado.



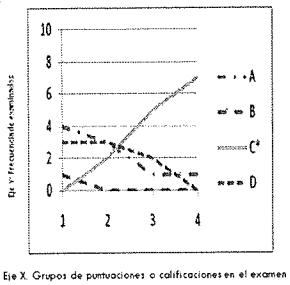
Eje X. Grupos de puntuaciones o calificaciones en el examen

Presentación de la capacitación

Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que discrimina bien

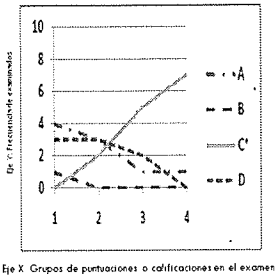
- El ítem permite observar diferencias entre los examinados que dominan el contenido que evalúa el ítem de los que no lo dominan.
- Esto puede observarse porque la inclinación de la curva de la respuesta correcta se hace más pronunciada a medida que aumenta la habilidad.
- El ítem resulta ser con dificultad media que discrimina bien entre los examinados que dominan el contenido de los que no lo dominan, por lo que es considerado como bueno.



Análisis Gráfico de Ítems (AGI)

Ítem con dificultad media que discrimina bien

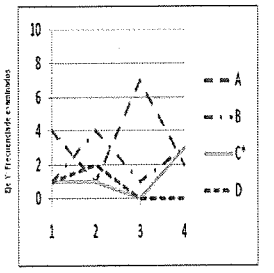
- La curva de la respuesta correcta C en el ítem permite observar que el número de examinados que la eligieron aumenta a medida que incrementa la habilidad. El ítem resulta de dificultad media con una discriminación que es aceptable, por lo que puede ser considerado como mínimamente adecuado.
- Los distractores B y D parecen funcionar bien, porque sus curvas de respuestas disminuyen a medida que aumenta la habilidad y distraen a examinados que no dominan el contenido.



Análisis Gráfico de Ítems (AGI)

Ítem difícil que no discrimina

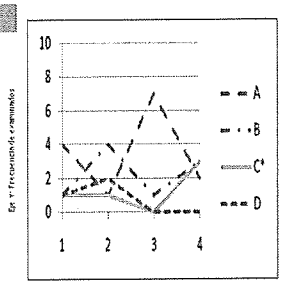
- El ítem no permite discriminar entre los examinados que dominan el contenido que se evalúa, de los examinados que no lo dominan. Esto puede distinguirse porque la pendiente o inclinación de la curva de la respuesta correcta fluctúa objetablemente entre los grupos de habilidad.
- El ítem resulta ser difícil y no discrimina entre los examinados que dominan el contenido de los que no lo dominan, por lo que puede considerarse como inadecuado.



Análisis Gráfico de Ítems (AGI)

Ítem difícil que no discrimina

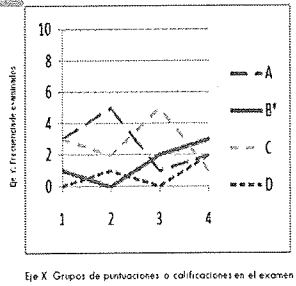
- La curva de la respuesta correcta en el ítem muestra que el número de examinados que eligieron la opción correcta C es relativamente bajo y fluctúa entre los grupos de habilidad; solo en el grupo de habilidad 3 incrementa y en los demás grupos se mantiene con altibajos.
- El distractor D parece funcionar bien, porque disminuye la frecuencia o porcentaje de los examinados que lo eligen a medida que aumenta la habilidad.



Análisis Gráfico de Ítems (AGI)

Ítem difícil que discrimina poco

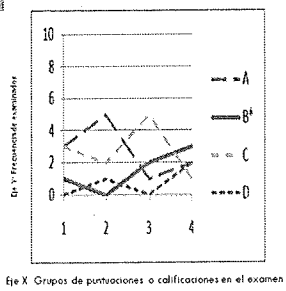
- El ítem permite separar de manera mínima entre los examinados que comprenden el contenido que evalúa el ítem, de los que no lo comprenden.
- El ítem resulta ser difícil y discrimina poco, por lo que puede ser considerado apenas adecuado.



Análisis Gráfico de Ítems (AGI)

Ítem difícil que discrimina poco

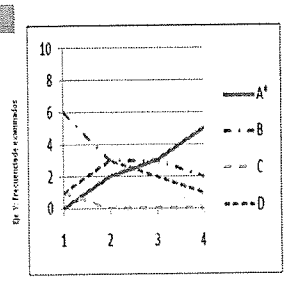
- La curva de la respuesta correcta B muestra que el número de examinados que la eligieron incrementa poco a medida que incrementa la habilidad.
- Los distractores, por lo general resultaron atractivos tanto para los examinados que dominan el contenido (grupos 3 y 4), como para examinados que no lo dominan (grupos 1 y 2).



Análisis Gráfico de Ítems (AGI)

Ítem difícil que discrimina bien

- La curva de la respuesta correcta A en el ítem muestra que el número de examinados que la eligieron incrementa a medida que aumenta la habilidad.
- Los distractores resultaron atractivos para los examinados de los diferentes grupos de habilidad, aunque para el grupo de habilidad 4 resultaron atractivos de manera mínima.

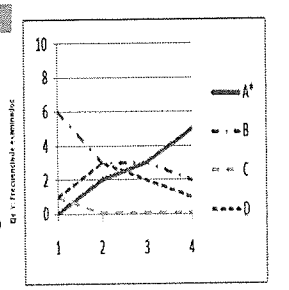


Eje X Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem difícil que discrimina bien

- La opción C demostró ser poco atractiva porque solo la eligió un solo examinado. El distractor B fue atractivo para examinados de los grupos de habilidad baja (1 y 2), pero también resultó atractivo para 5 examinados de los grupos de habilidad alta (3 y 4). El distractor D cumple con la función de ser atractivo para examinados que no dominan el contenido; sin embargo, fue elegido por 3 examinados de los grupos de habilidad alta (3 y 4).

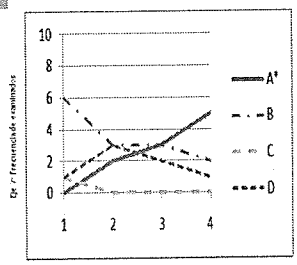


Eje X Grupos de puntuaciones o calificaciones en el examen

Análisis Gráfico de Ítems (AGI)

Ítem difícil que discrimina bien

- El ítem permite discriminar con claridad entre los examinados que dominan el contenido de los examinados que no lo dominan.
- Esto puede comprobarse porque la tendencia de la curva de la respuesta correcta incrementa a medida que aumenta la habilidad.



Eje X Grupos de puntuaciones o calificaciones en el examen

Criterios para eliminar o corregir ítems

- Es posible proporcionar criterios que puedan utilizarse como una guía general para razonar sobre la posible eliminación o en su caso corrección de ítems defectuosos. Se presentan cuatro criterios principales que proponen Batenburg y Laros (2002) para determinar la eliminación o corrección de ítems:

1. Tomar en cuenta el número de violaciones al supuesto de que la opción correcta incrementa con un aumento del puntaje total.

Criterios para eliminar o corregir ítems

2. Tomar en cuenta el número de violaciones al supuesto de que las opciones falsas decrezcan con un incremento del puntaje total.
3. Tomar en cuenta el número de intersecciones entre las opciones correctas y falsas tras el inicio del intervalo de discriminación.
4. La ausencia de poder discriminativo o una baja pendiente de la respuesta correcta.

- No obstante lo anterior, considérese que al eliminar un ítem malo de una prueba debe ser reemplazado por otro ítem bueno y que cubra la misma parte del contenido de aprendizaje.

Programa para el análisis gráfico de ítems

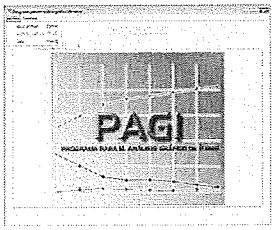
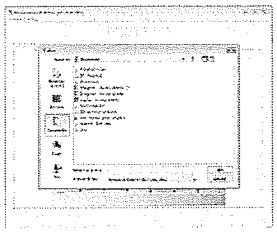
PAGI

- El Programa para el Análisis Gráfico de Ítems (PAGI), es un programa psicométrico gratuito y de código abierto, que fue desarrollado para efectuar el análisis gráfico de ítems basado en la Teoría Clásica de los Tests y en la metodología del AGI.
- PAGI es capaz de realizar estimaciones de la dificultad, la discriminación y el análisis de los distractores en los ítems de opción múltiple o de respuesta alterna de una prueba.
- Como ya vimos estos análisis se pueden ejecutar e interpretar de manera sencilla con la ayuda de gráficos.

Programa para el análisis gráfico de ítems

Entrada de archivos

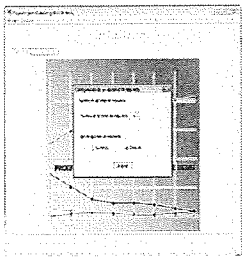
- La entrada de los datos, en la parte superior izquierda del menú Archivo se muestra una opción señalada como "Abrir archivo" y se busca el archivo deseado.

Programa para el análisis gráfico de ítems

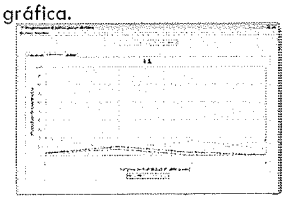
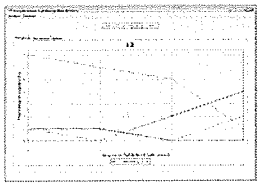
Entrada de archivos

- Se da clic en el control Abrir, Si es archivo tiene el formato adecuado abrirá una ventana:
- Se solicita el número de opciones de respuesta que tiene cada ítem, que generalmente son 4 o 5. A continuación se indica el tipo de opciones de respuesta que se ofrecieron al examinado, mismo que puede ser "Carácter" en caso de que las opciones de respuesta hayan sido letras (A, B, C, D), o "Numérico" si fueron números (1, 2, 3, 4). Una vez seleccionados ambos elementos de configuración se da clic en "aceptar".



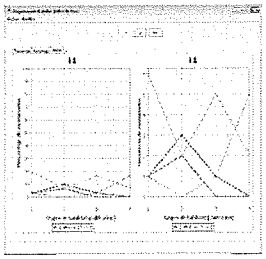
Programa para el análisis gráfico de ítems

- Se presenta nuevamente la ventana principal de PAGI, pero ahora ya se pueden visualizar las gráficas necesarias para efectuar el análisis gráfico de cada ítem. Por omisión, PAGI presenta primero la gráfica con frecuencia, si queremos visualizar la gráfica de porcentaje tenemos que seleccionar la opción. Por otro lado apretando los botones de control, se puede visualizar la siguiente gráfica.

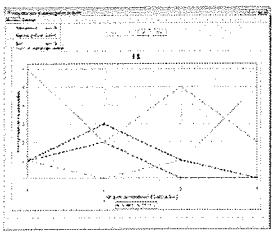
Programa para el análisis gráfico de ítems

- En la última pestaña se muestran juntas ambas gráficas: la de frecuencia y la de porcentaje. Esta gráfica compuesta permite al usuario observar al mismo tiempo y comparar la visión panorámica de las curvas de respuestas que se dieron en el ítem y la visión detallada de las mismas.



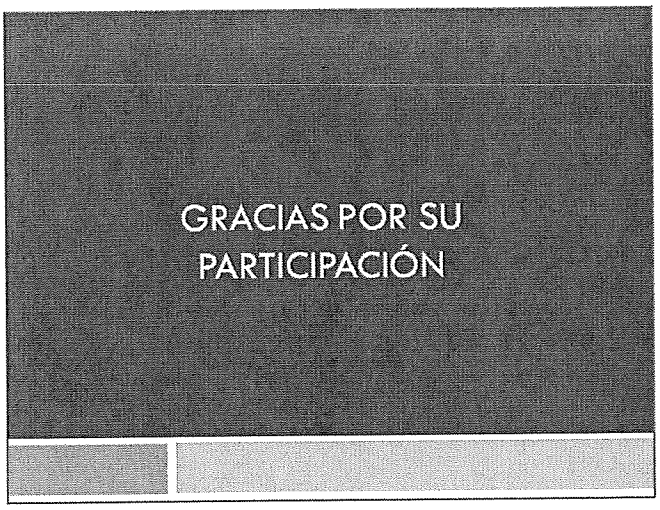
Programa para el análisis gráfico de ítems

- Una vez que se ha concluido el análisis y la interpretación de todos los ítems del examen, mediante el análisis gráfico de ítems, en el menú de "Archivo" se encuentra la opción Salir para finalizar el análisis.
- También si se desea guardar el resultado del análisis se puede seleccionar la opción Exportar gráficas.



Referencias

- Batenburg, T. y Laros, J. (2002). Data Screening and Graphical Analysis of Items *Universidad de Groningen, en Holanda y Universidad de Brasilia, Brasil.*
- Batenburg, T. y Laros, J. (2002). Graphical Item Analysis. *Education Research and Evaluation (8)* 3
- Contreras, L. A. (2001). Procedimiento para Asegurar la Validez de Contenido de una Prueba Criterial: el Caso de un Examen de Español para la Educación Primaria en Baja California. *Memoria del VI Congreso Nacional de Investigación Educativa.* Manzanillo, México. COMIE. 2001.
- Encinas, J. A., Rivera, R. E. y Contreras, L. A., (2005). Evaluación colegiada del aprendizaje en la Universidad Autónoma de Baja California: Construcción de un examen criterial de gran escala para evaluar el dominio de conceptos y procedimientos del cálculo diferencial. *Memoria del VIII Congreso Nacional de Investigación Educativa.* México. COMIE.
- Verhelst, N. (2004). Sección C. Teoría clásica de los tests, en reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework. Strasbourg. Language Policy Division.


GRACIAS POR SU
PARTICIPACIÓN

ANEXO 5

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

Maestría en Ciencias educativas

Formato de evaluación del Curso- taller para el análisis gráfico de ítems por medio de la aplicación PAGI

Instructor. Guadalupe De Los Santos Lázaro

INSTRUCCIONES: Por favor evalúe qué tanto está usted de acuerdo o en desacuerdo con cada una de las siguientes afirmaciones. Marque una ✓ en la celda correspondiente.

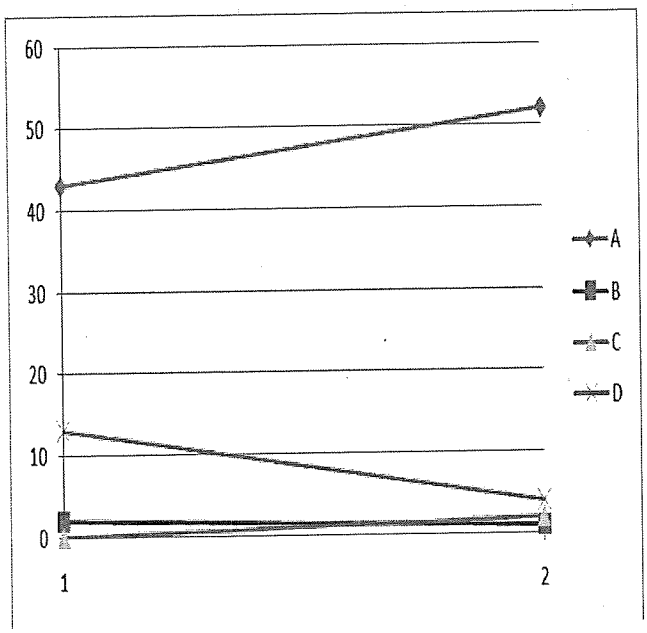
Afirmaciones	Completamente en desacuerdo	En desacuerdo	Ni de acuerdo, ni en desacuerdo	De acuerdo	Totalmente de acuerdo
1. Los objetivos de la capacitación fueron claros desde el principio.					
2. Tengo la impresión de que en el curso hubo mucha desorganización.					
3. Al iniciar el curso quedó claro cuáles temas se revisarían durante la capacitación.					
4. La instructora me motivó a aprender el material.					
5. La instructora verificó la comprensión de los temas expuestos, a través de preguntas y ejercicios.					
6. La instructora propició la participación de todos.					
7. La instructora casi nunca me alentó a hacer preguntas.					
8. La instructora expresó sus ideas con claridad.					
9. La instructora respondió a mis preguntas de manera adecuada.					
10. Sentí que la instructora estaba poco dispuesta para ayudarme con mis dudas y dificultades.					
11. La instructora utilizó ejemplos apropiados para ilustrar los conceptos y procedimientos explicados.					
12. La instructora mostró poco dominio del contenido.					
13. La instructora explicó el contenido en un nivel que yo pudiera entender.					
14. Los conceptos descritos en el manual fueron confusos.					

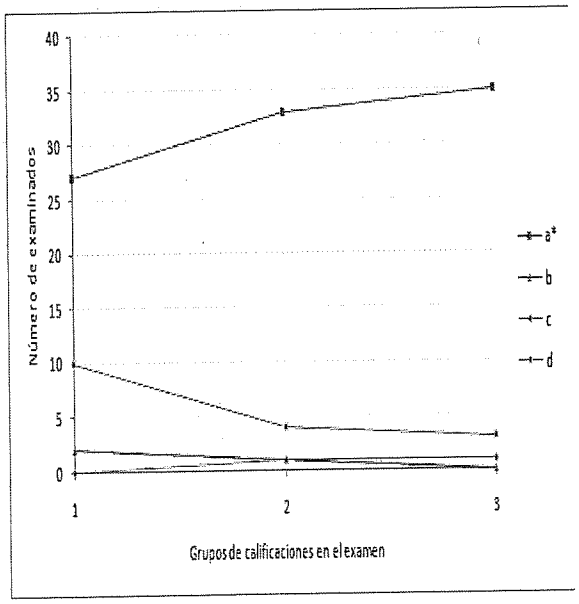
Afirmaciones	Completamente en desacuerdo	En desacuerdo	Ni de acuerdo, ni en desacuerdo	De acuerdo	Totalmente de acuerdo
15. Los procedimientos descritos en el manual me permitieron entender el contenido.					
16. El manual me aportó los conocimientos que son necesarios para llevar a cabo un análisis gráfico de ítems.					
17. Los ejemplos en el manual ayudaron a comprender cada explicación.					
18. El manual está organizado de una manera clara y lógica.					
19. La redacción del manual fue innecesariamente compleja.					
20. El uso del programa de cómputo para el análisis gráfico de ítems resultó complejo.					
21. El programa de cómputo para el análisis gráfico de ítems me ayudará a mejorar la evaluación del aprendizaje de mis alumnos.					
22. El uso de la presentación en computadora (PowerPoint) facilitó la comprensión de los conceptos y procedimientos descritos.					
23. Las actividades desarrolladas durante la capacitación fueron adecuadas.					
24. Las actividades permitieron cumplir con los objetivos del curso.					
25. La forma como se desarrollaron los temas ocasionó que fueran difíciles de entender.					
26. Faltó tiempo para cubrir todo el material del curso.					
27. El aprendizaje que logré en el curso será de poca utilidad para el trabajo en el aula.					
28. En este curso logré aprendizajes significativos que me ayudaron a entender mejor la evaluación del aprendizaje.					

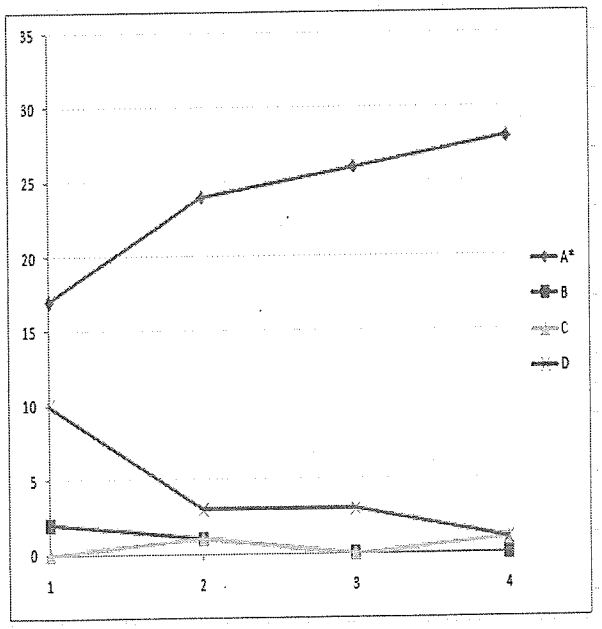
GRACIAS POR SU COLABORACIÓN

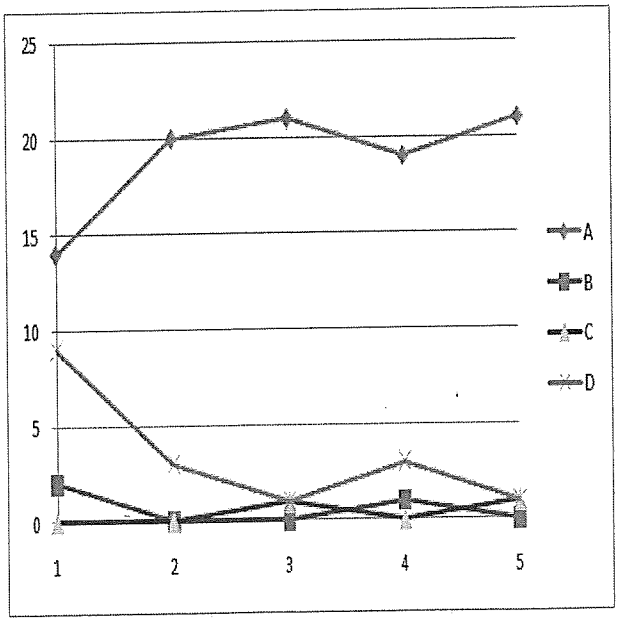
ANEXO 6

Gráficas derivadas del primer experimento para adaptar la técnica del AGI



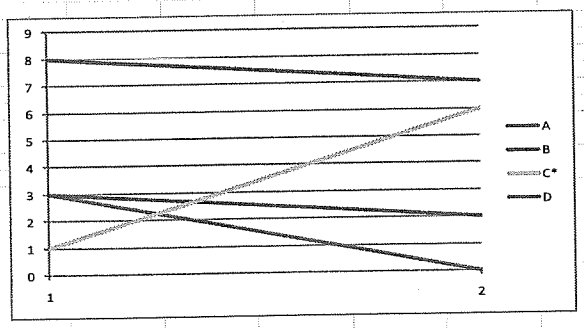




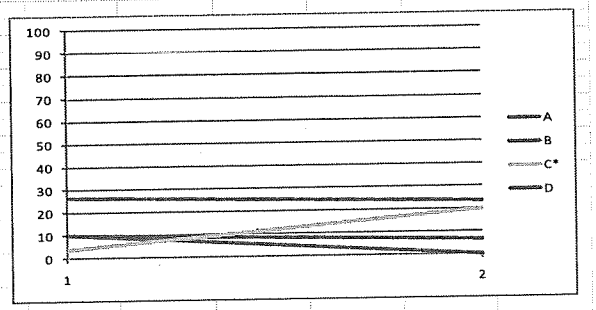


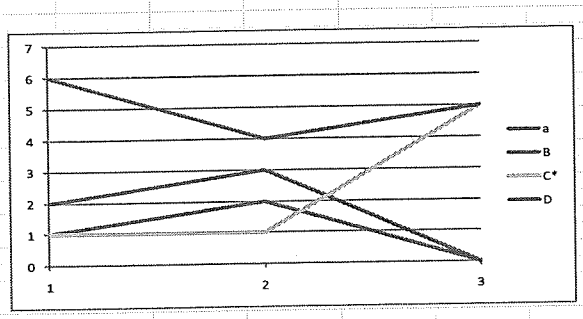
Gráficas derivadas del primer experimento para adaptar la técnica del AGI

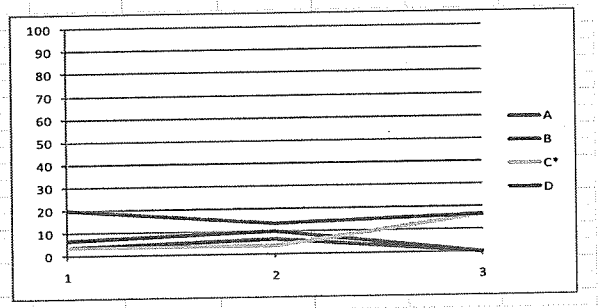
ÍTEM 1FRECUENCIA

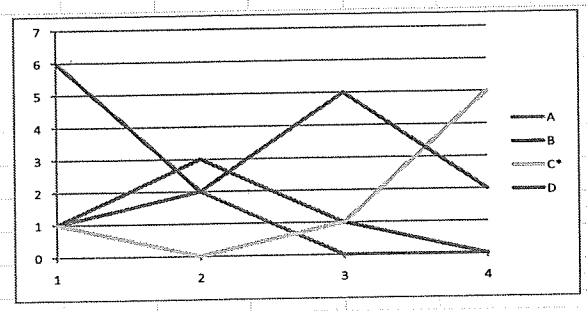


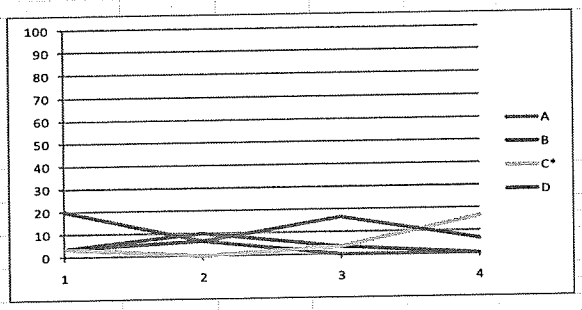
ÍTEM 1PORCENTAJE

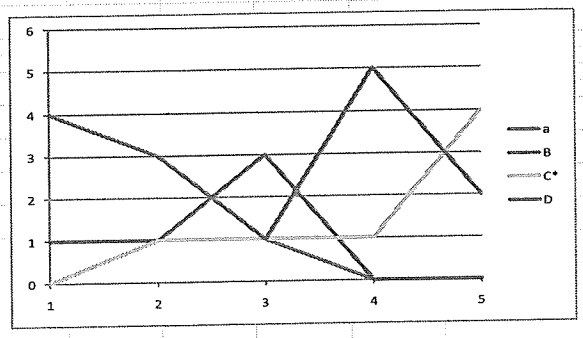


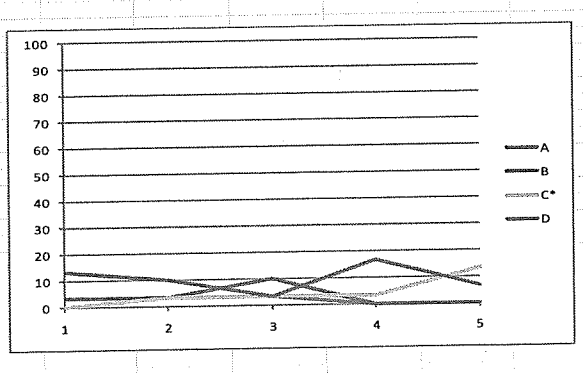






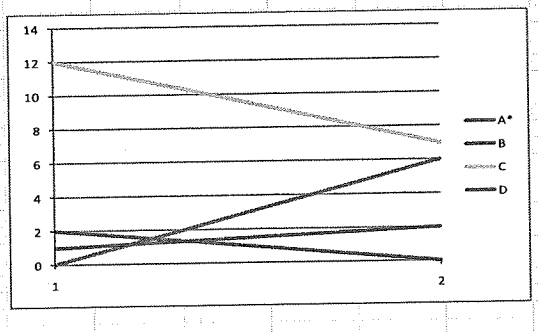




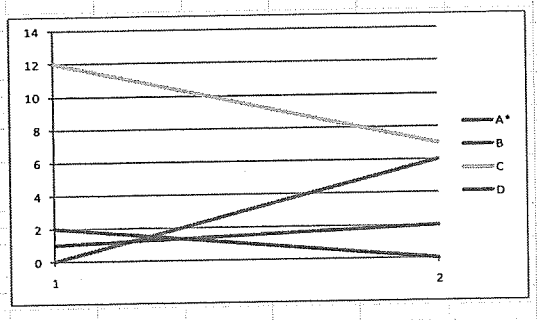


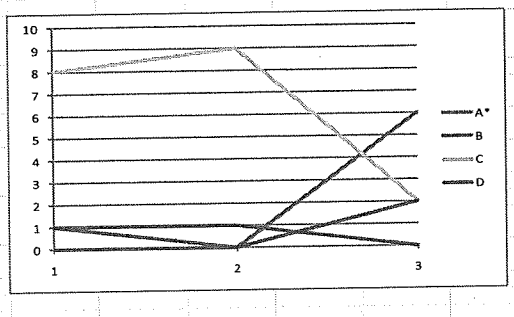
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

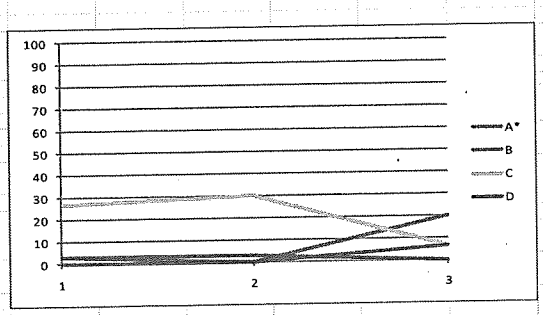
ÍTEM 2 FRECUENCIA

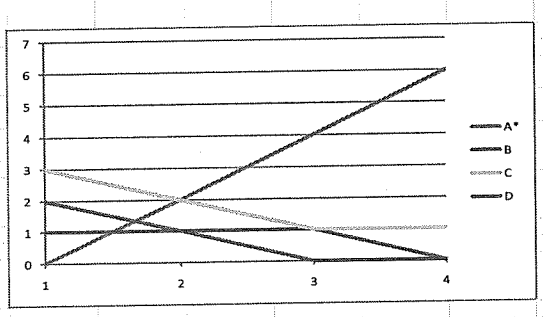


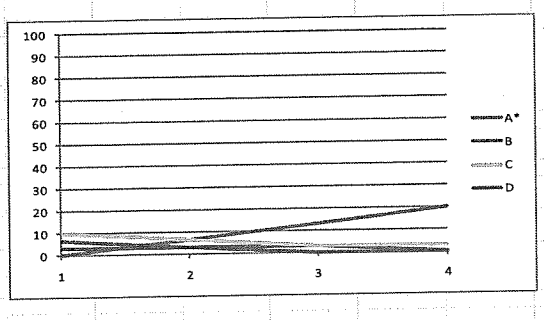
ÍTEM 2 PORCENTAJE

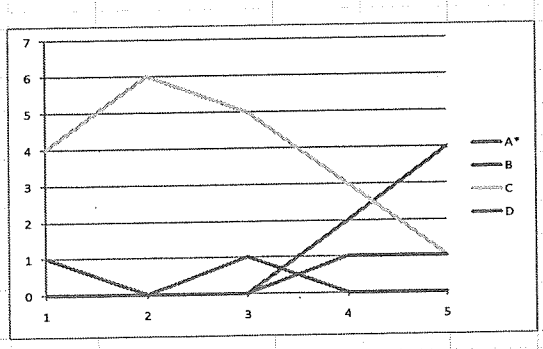


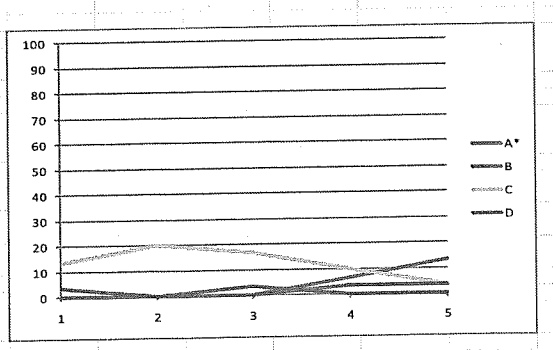






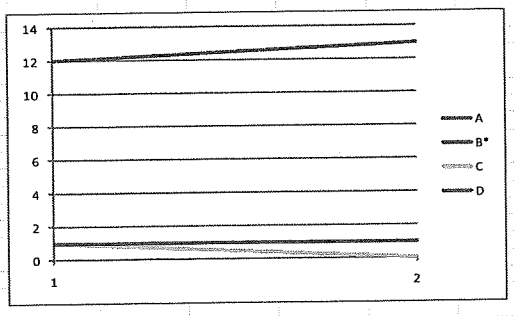




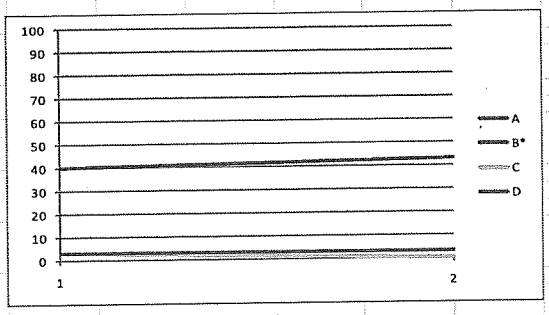


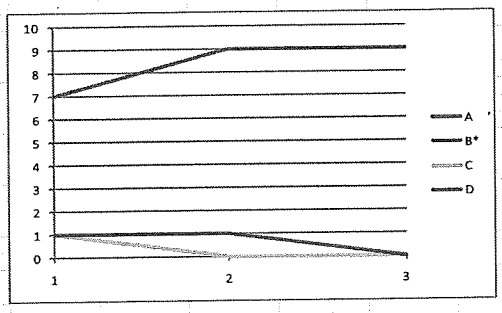
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

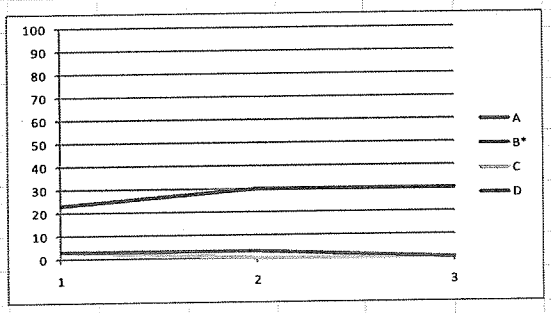
ÍTEM 3 FRECUENCIA

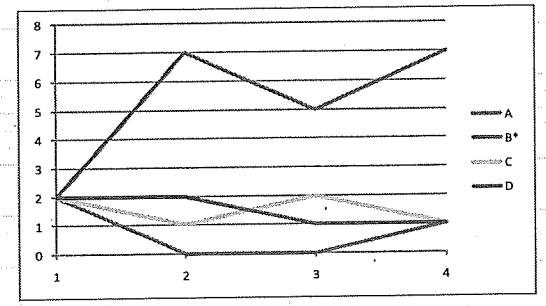


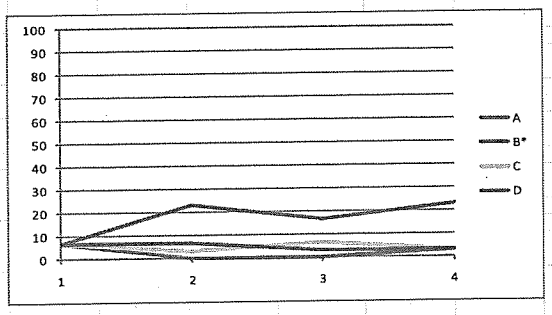
ÍTEM 3 PORCENTAJE

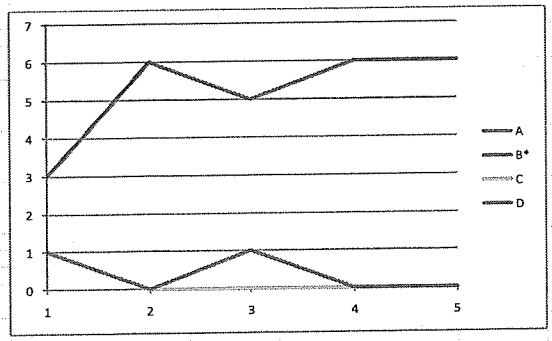


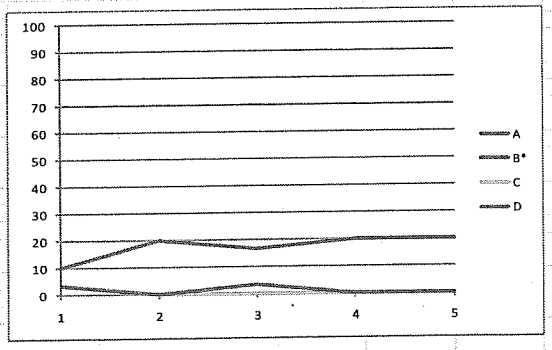






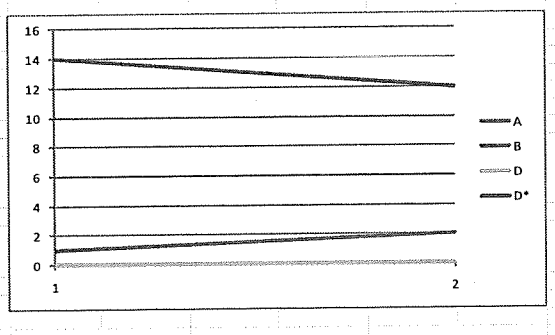




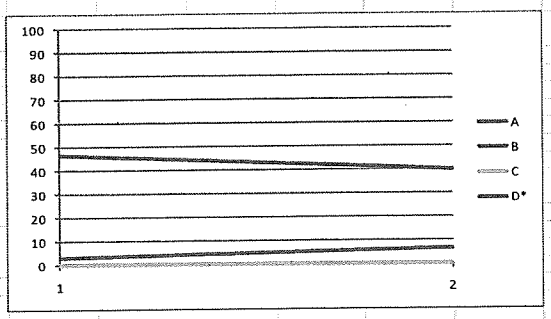


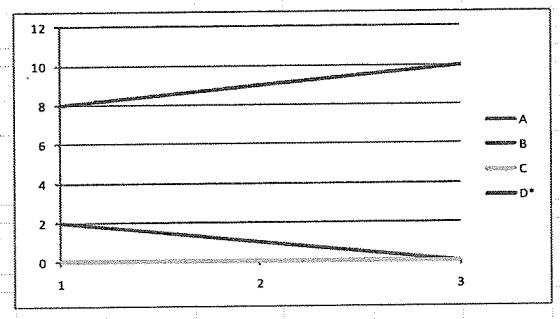
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

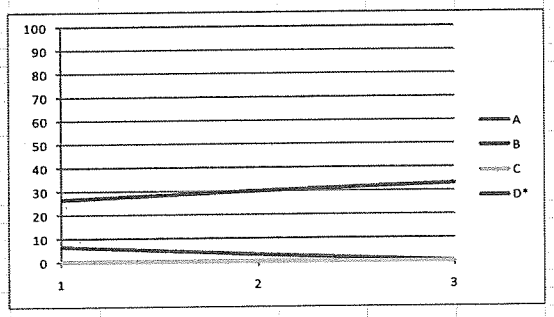
ÍTEM 4 FRECUENCIA

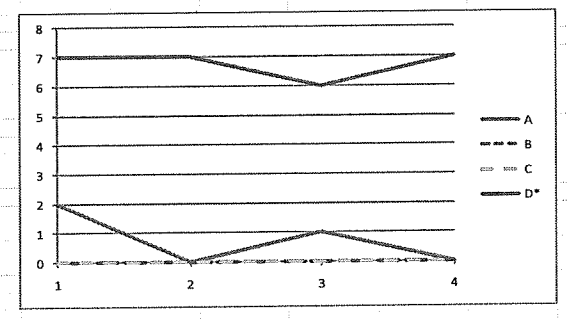


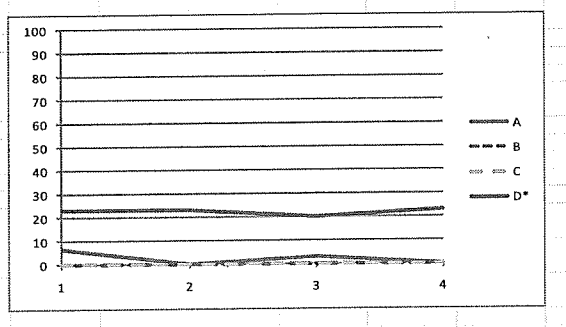
ÍTEM 4 PORCENTAJE

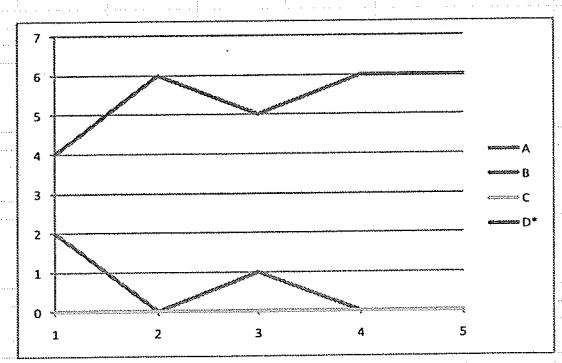


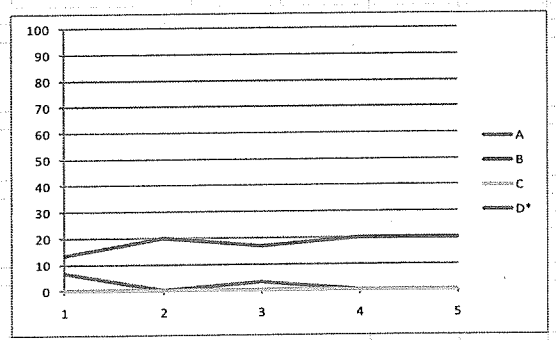






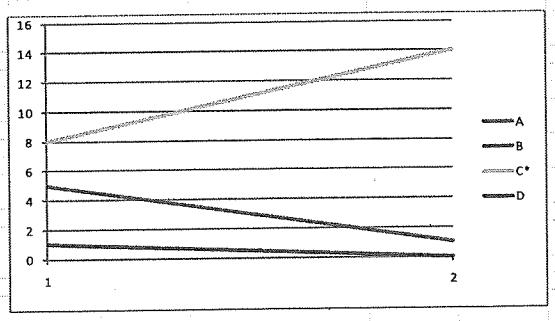




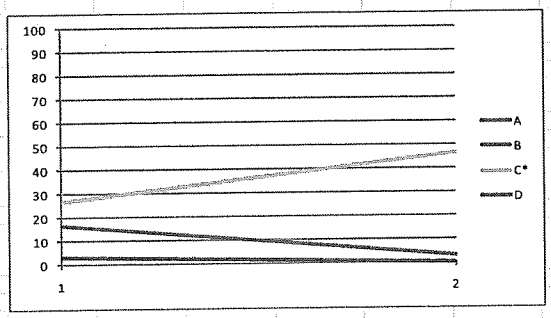


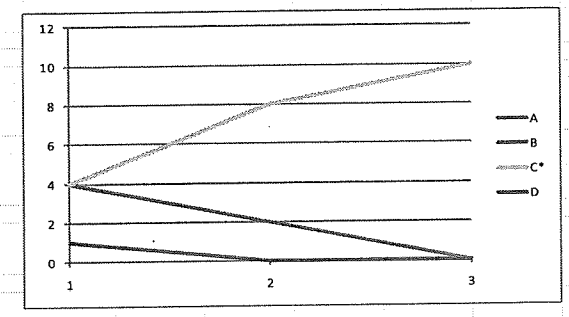
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

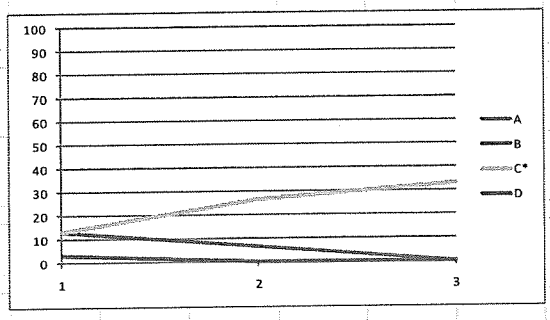
ÍTEM 5 FRECUENCIA

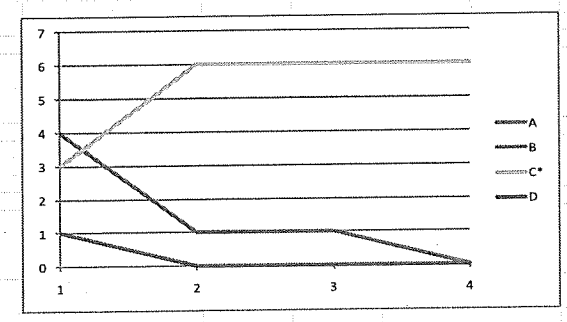


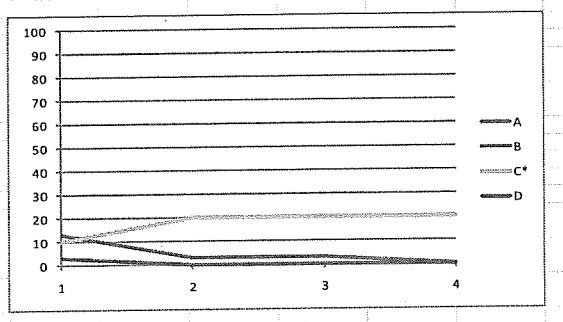
ÍTEM 5 PORCENTAJE

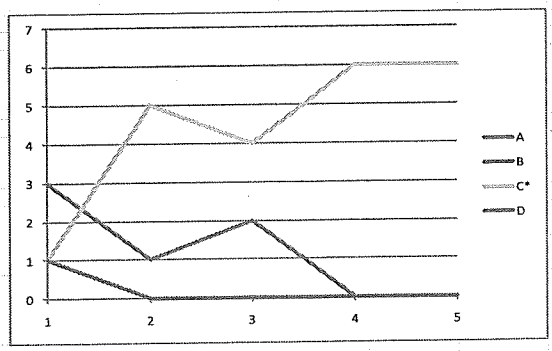


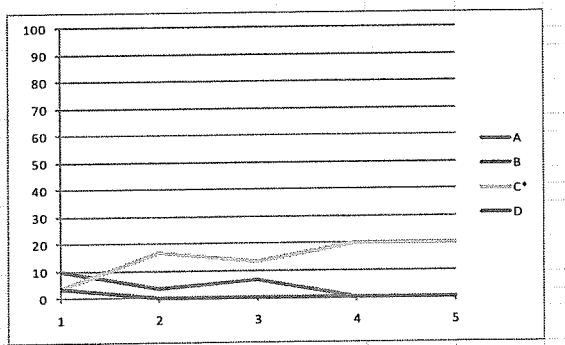




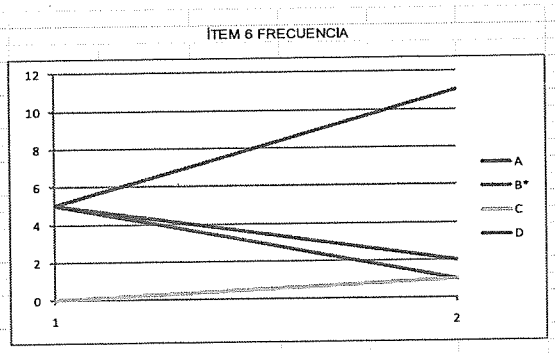


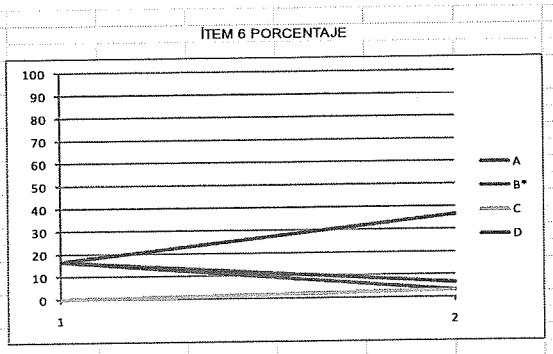


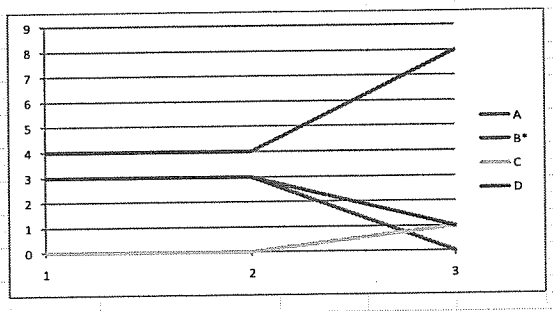


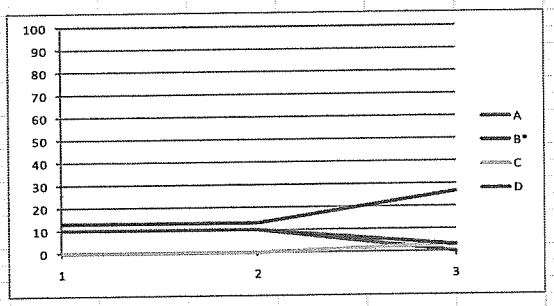


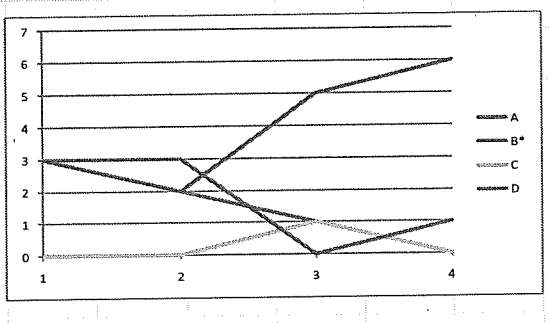
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

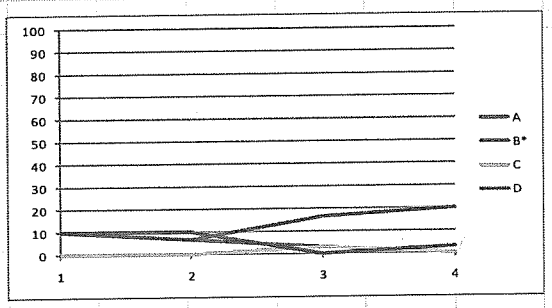


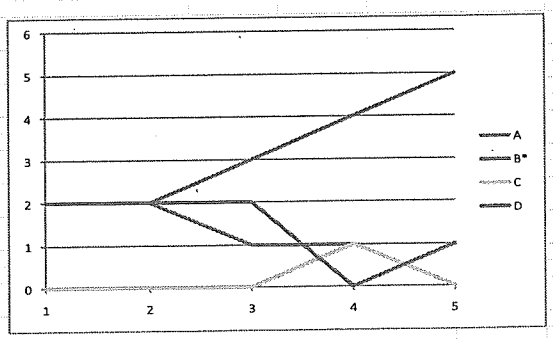


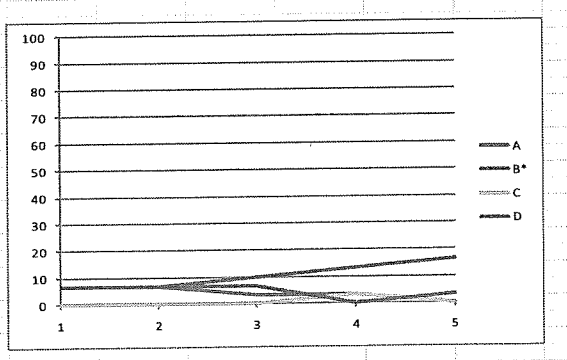






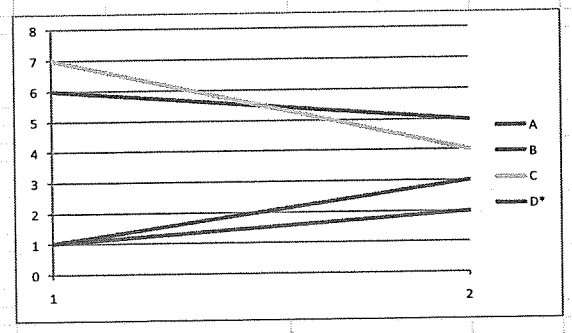




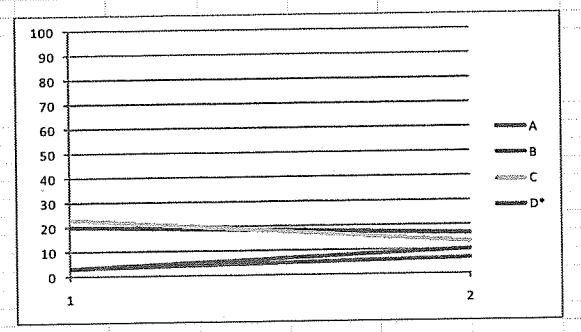


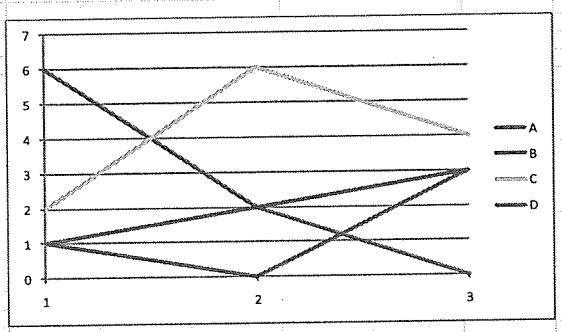
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

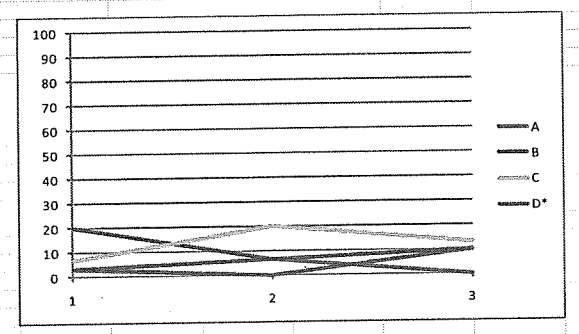
ÍTEM 7 FRECUENCIA

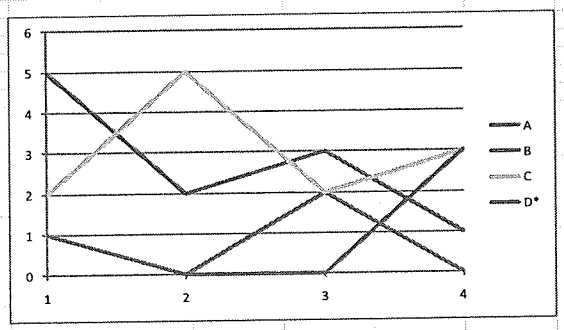


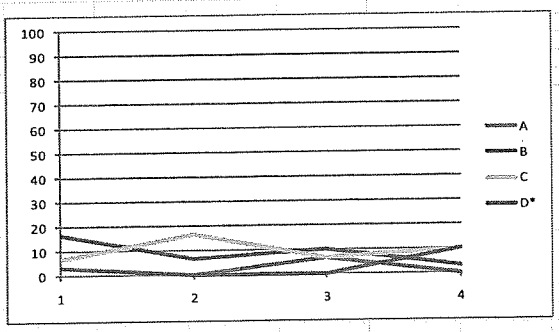
ÍTEM 7 PORCENTAJE

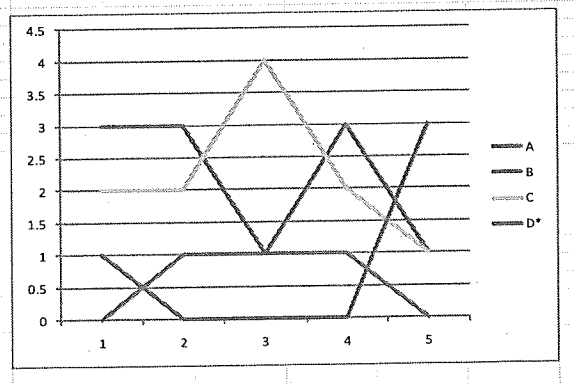


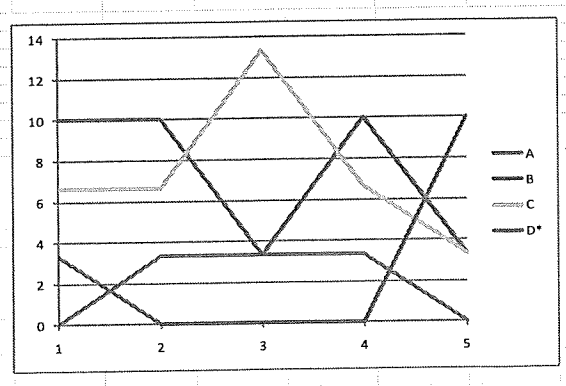






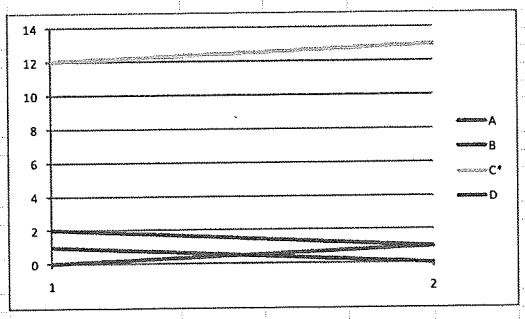




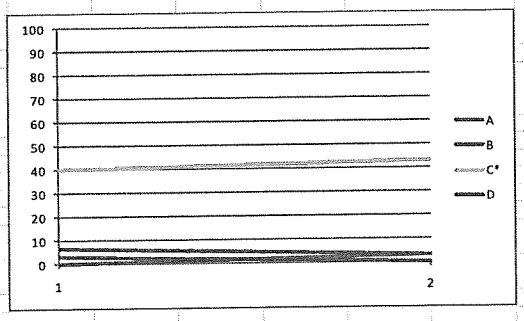


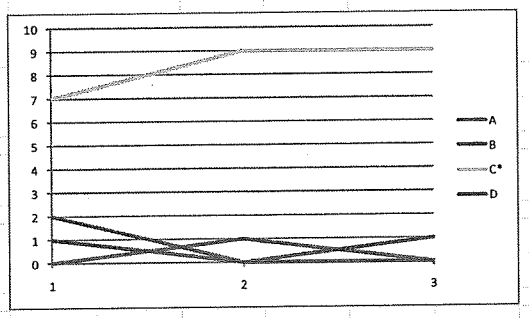
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

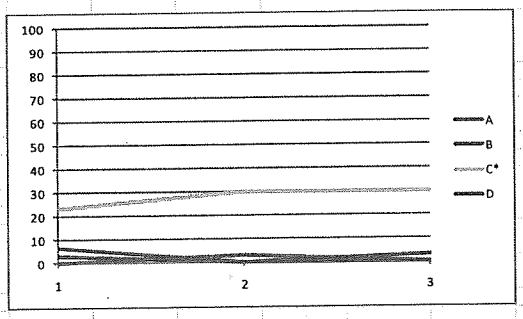
ÍTEM 8 FRECUENCIA

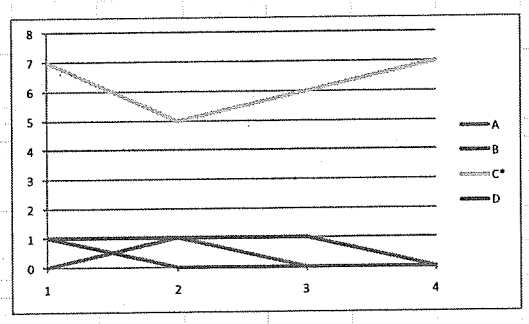


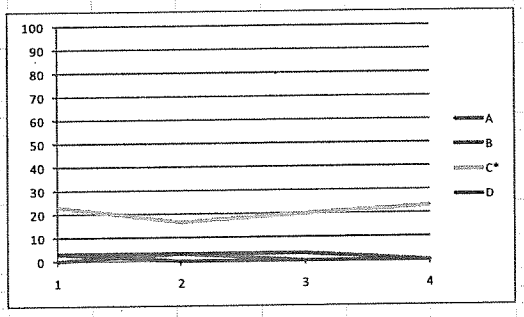
ÍTEM 8 PORCENTAJE

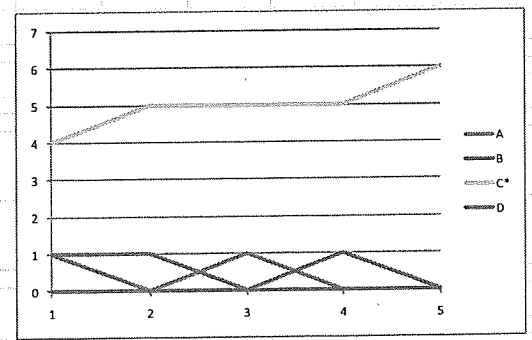


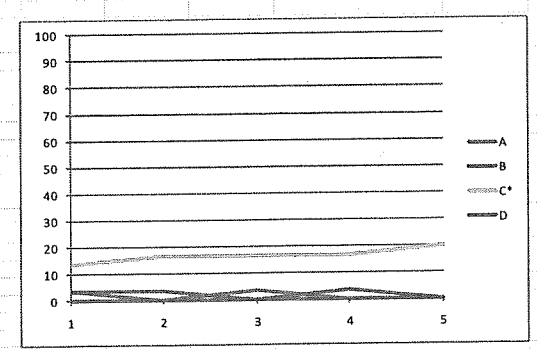






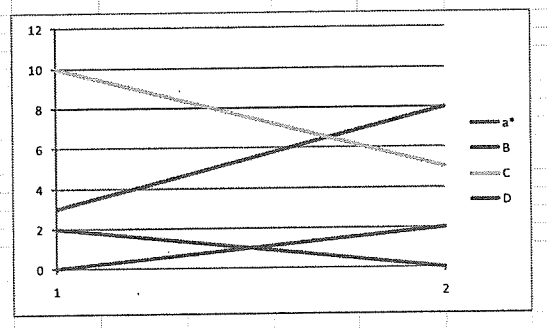




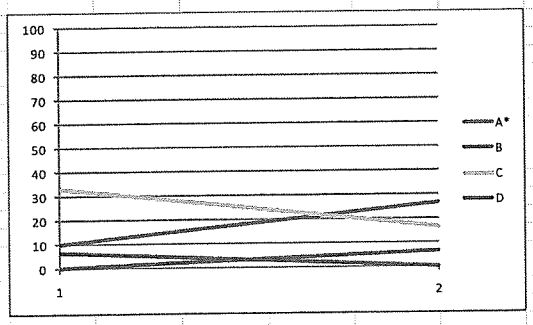


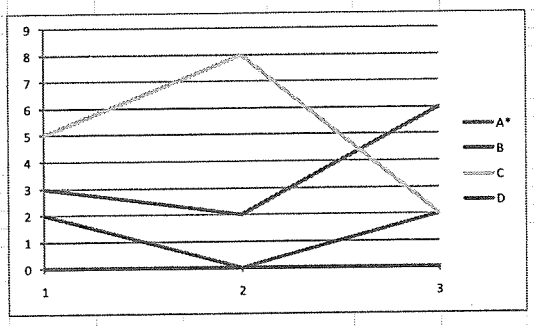
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

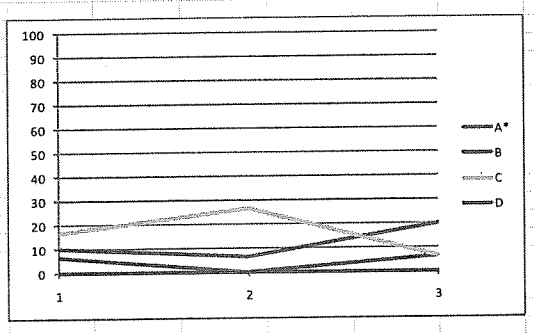
ÍTEM 9 FRECUENCIA

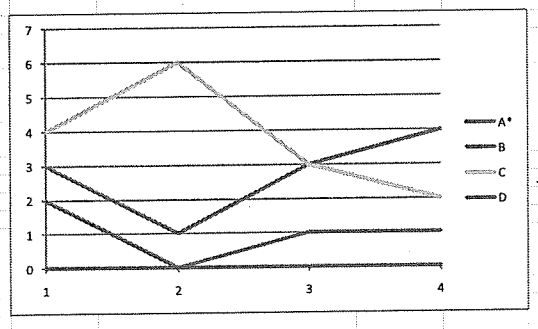


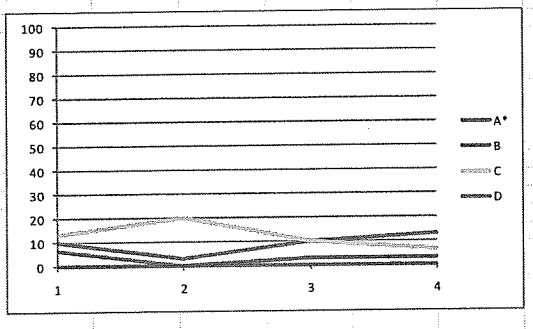
ÍTEM 9 PORCENTAJE

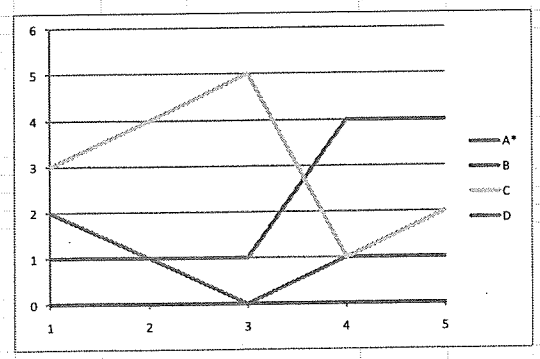


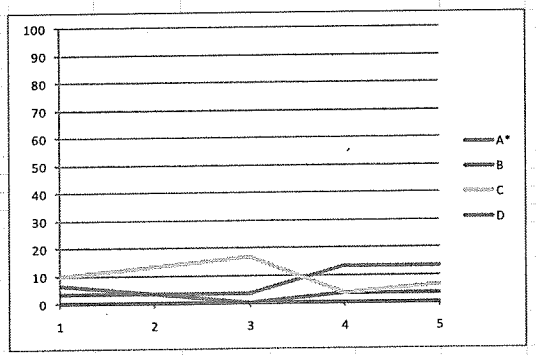






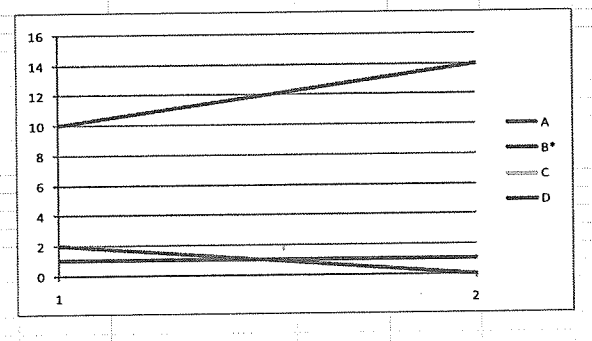




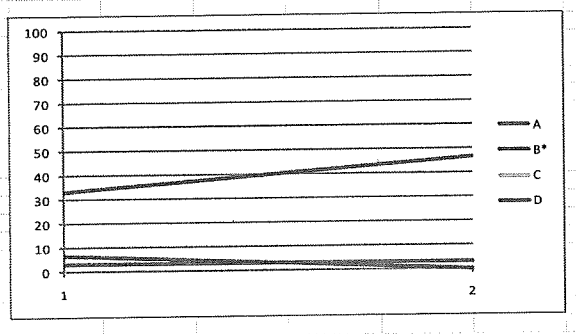


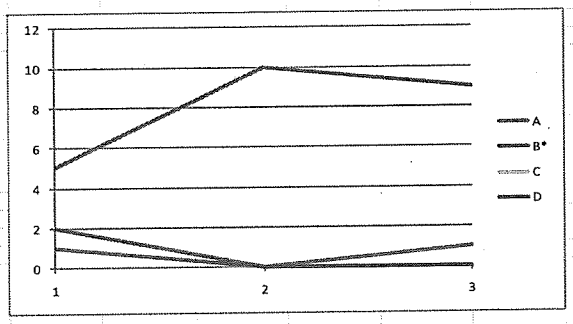
Gráficas derivadas del segundo experimento para adaptar la técnica del AGI

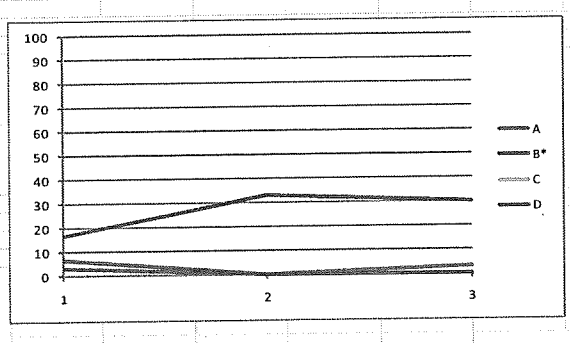
ÍTEM 10 FRECUENCIA

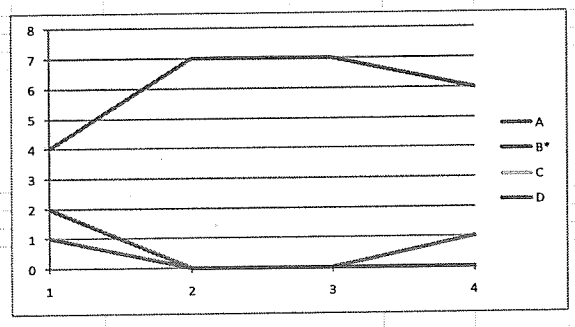


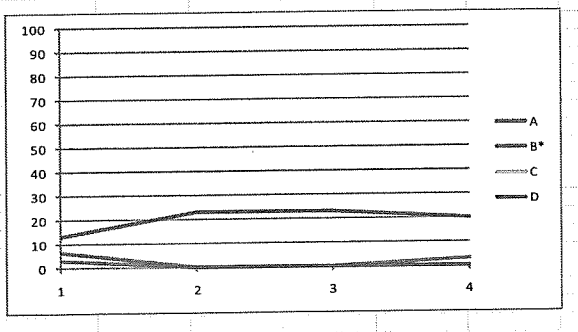
ÍTEM 10 PORCENTAJE

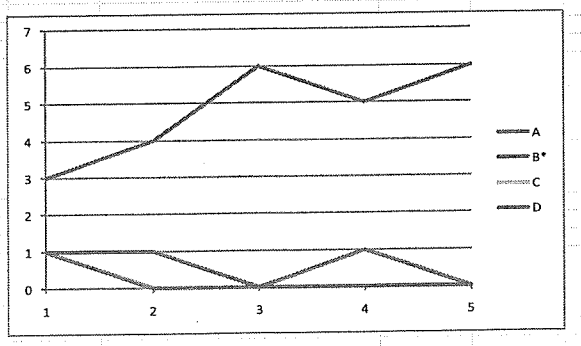


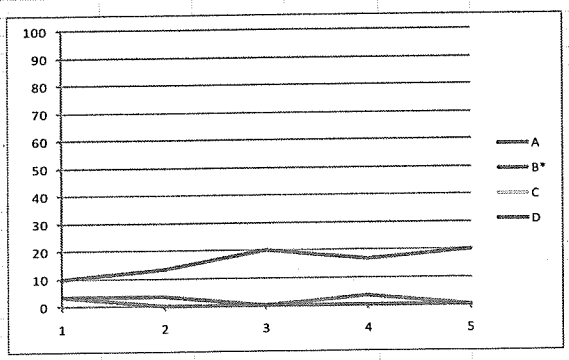






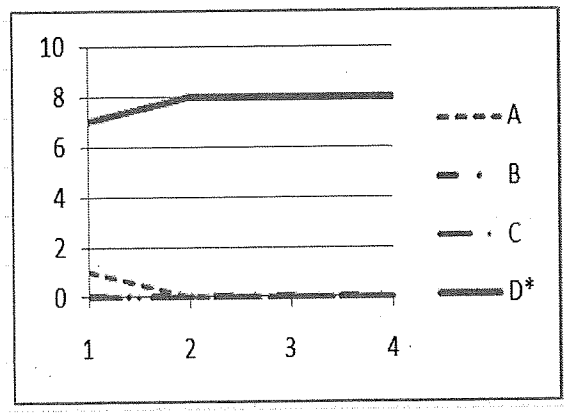


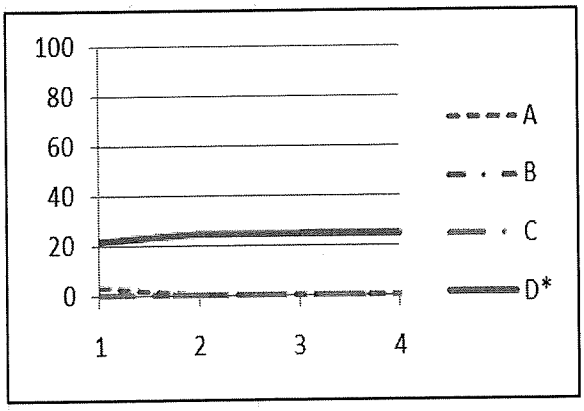




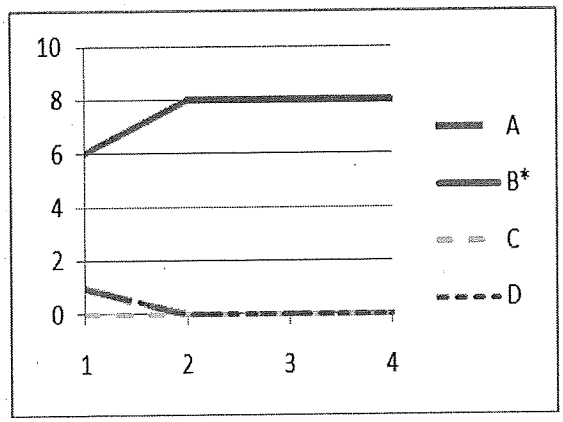
ANEXO 7

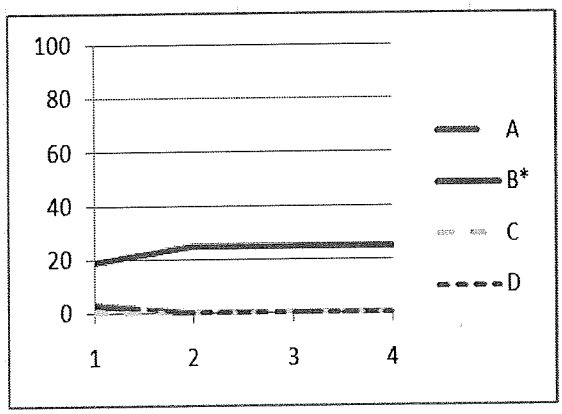
Ítem fácil que discrimina poco



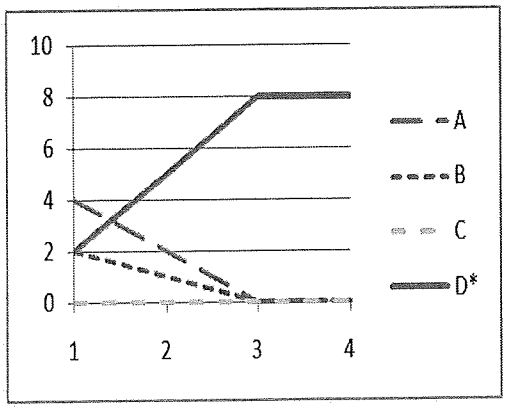


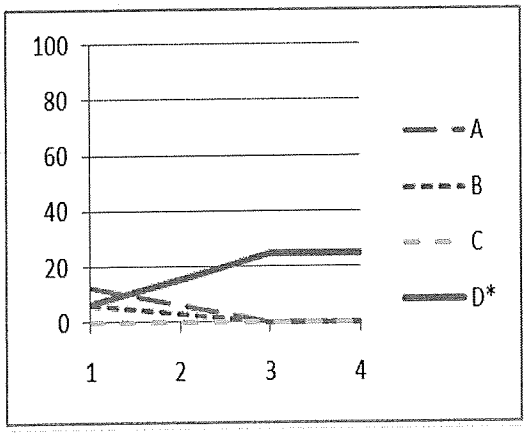
Ítem fácil con discriminación media



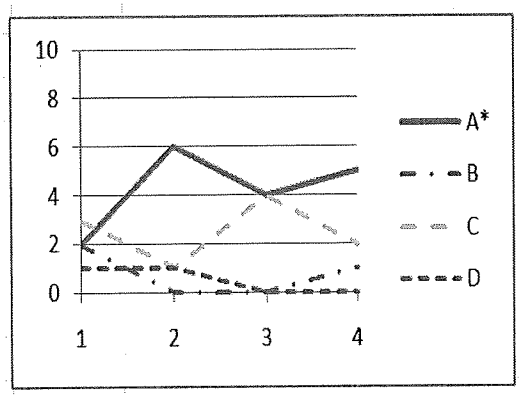


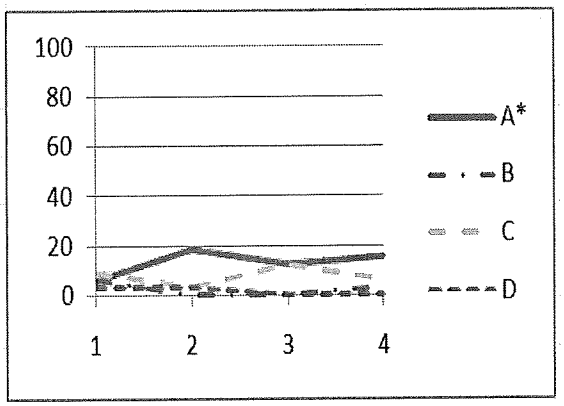
Ítem fácil con discriminación alta



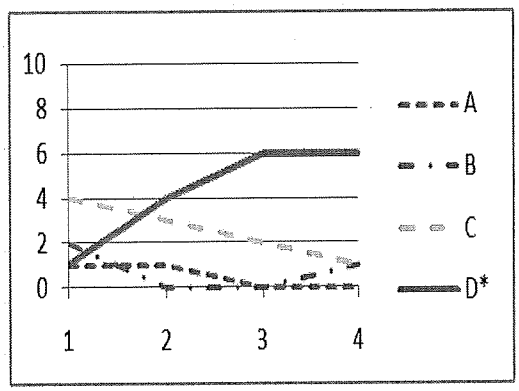


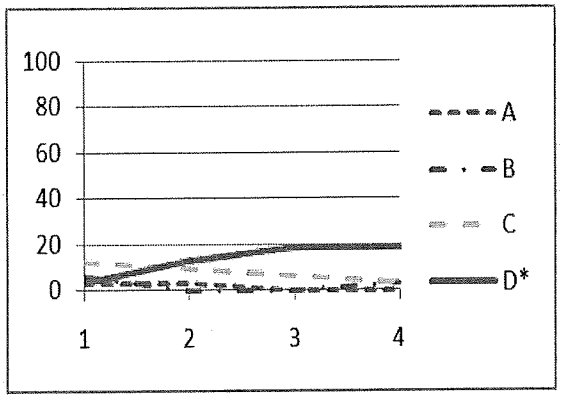
Ítem de dificultad media que discrimina poco



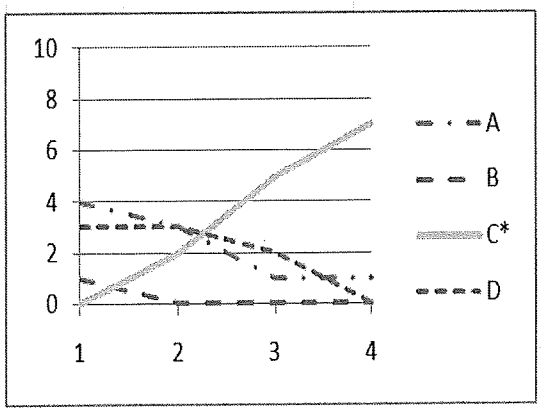


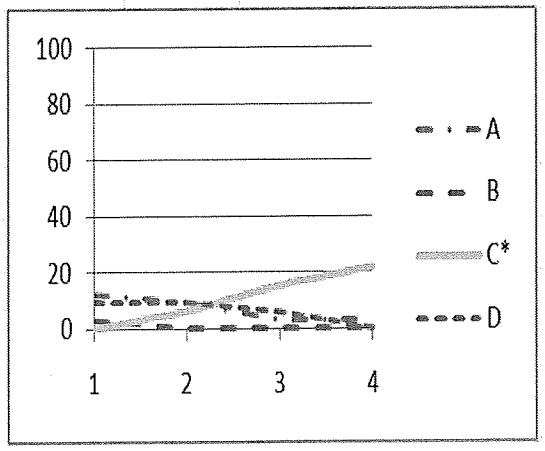
Ítem de dificultad media con discriminación media



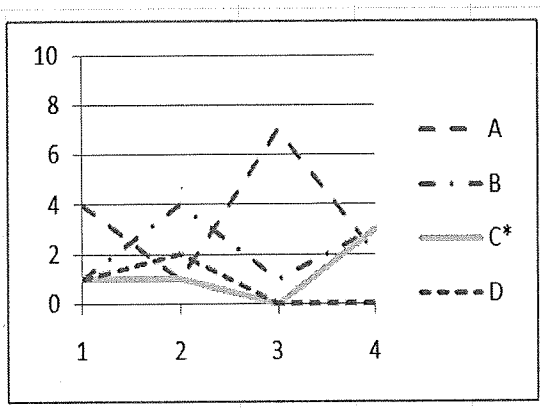


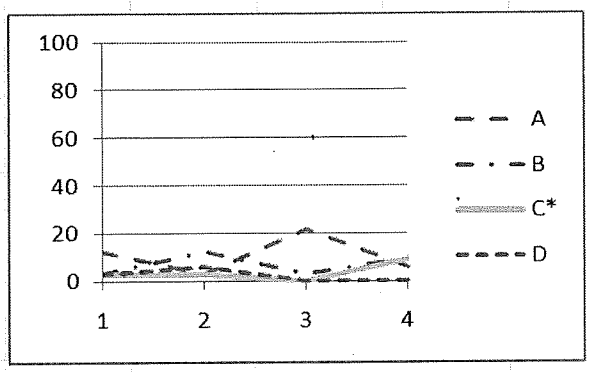
Ítems de dificultad media con discriminación alta



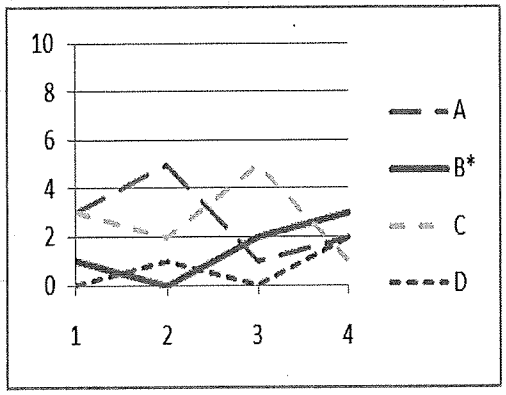


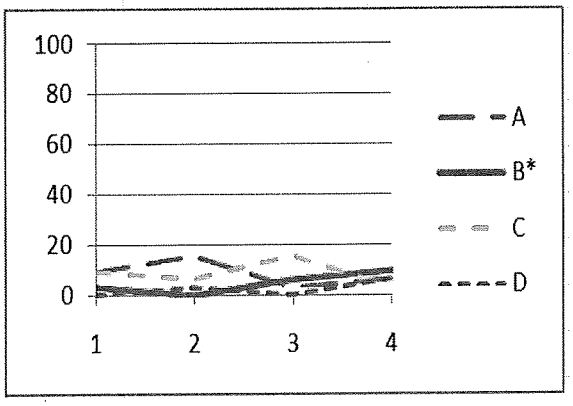
Ítem difícil que discrimina poco



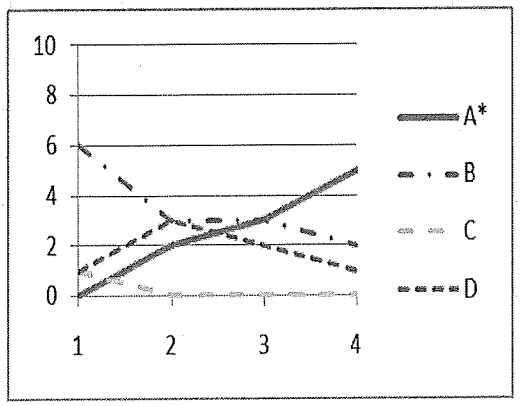


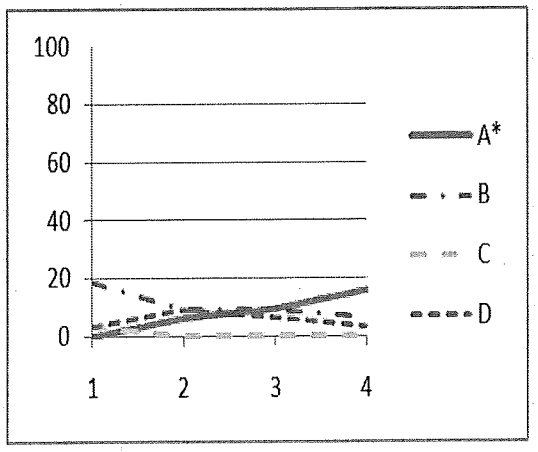
Ítem de dificultad media con discriminación media





Ítems de dificultad media con discriminación alta





ANEXO 8

Maestría en Ciencias educativas

Guía para apoyar el análisis gráfico de ítems

INSTRUCCIONES: En cada celda anote los principales puntos fuertes y débiles que Usted observó al analizar la correspondiente curva de respuestas al ítem y al final emita un dictamen general razonado sobre el ítem.

Curva de respuestas	Observaciones				
	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5
Respuesta correcta					
Distractor 1					
Distractor 2					
Distractor 3					
Juicio final sobre el ítem por ejemplo: descartar, aprobar, modificar la opción A, sustituir la opción D, etc.					

Guía para apoyar el análisis gráfico de ítems

INSTRUCCIONES: En cada celda anote los principales puntos fuertes y débiles que Usted observó al analizar la correspondiente curva de respuestas al ítem y al final emita un dictamen general razonado sobre el ítem.

Opción de respuesta	Observaciones				
	Ítem 6	Ítem 7	Ítem 8	Ítem 9	Ítem 10
Respuesta correcta					
Distractor 1					
Distractor 2					
Distractor 3					
uicio final sobre el ítem (por ejemplo: descartar, aprobar, modificar la opción A, sustituir la opción D, etc.					