TRAB
PS2005
G8

UNIVERSIDAD CATÓLICA ANDRÉS BELLO
FACULTAD DE HUMANIDADES Y EDUCACIÓN
ESCUELA DE PSICOLOGÍA

TRABAJO DE ASENSO PARA PROFESOR AGREGADO

ANÁLISIS PSICOMÉTRICO DE UNA PRUEBA DE RAZONAMIENTO BÁSICO CUANTITATIVO UTILIZADA EN LA ADMISIÓN DE ESTUDIANTES A POSTGRADO

AUTOR
LIGIA GUGLIETTA

CARACAS, MARZO DE 2005

ÍNDICE GENERAL

Página

INTRODUCCIÓN 5
MARCO TEÓRICO 8
A.-Teoría Clásica de los Tests 9
B.- Prueba de Razonamiento Básico Cuantitativo 22
C.- Rendimiento Académico en Postgrado 29
D.- Evidencia Empírica sobre el comportamiento de pruebas de habilidad cuantitativa en la predicción del rendimiento académico a nivel de postgrado 34
MÉTODO 41
Objetivo 41
Tipo de Investigación 41
Diseño Muestral 42
Procedimiento 45
ANÁLISIS E INTEGRACIÓN DE RESULTADOS 53
Análisis Descriptivo de las Variables 53
Confiabilidad 56
Validez 58
Análisis de Items 62
CONCLUSIONES 82
REFERENCIAS BIBLIOGRÁFICAS 84

ÍNDICE DE TABLAS

Página

Tabla 1. *Especificaciones de la Prueba de Razonamiento Básico Cuantitativo* ..27

Tabla 2. *Niveles de Dificultad (Estimados) de los Items* 28

Tabla 3. Sexo 43

Tabla 4. *Edad* 43

Tabla 5. *Tipo de Profesión* 44

Tabla 6. *Habilidad Cuantitativa*.....54

Tabla 7. *Rendimiento Académico en Postgrado* 55

Tabla 8.*División por Mitades*.....56

Tabla 9. *Aporte de los Items a la Confiabilidad de la Prueba*.....57

Tabla 10. *Distribución de Niveles de Dificultad* 59

Tabla 11. *Diferencias en RBC por Sexo*.....61

Tabla 12. *Análisis de Varianza por sexo*.....61

Tabla 13. *Resultados por Tipo de Profesión*.....62

Tabla 14. *Prueba T para Muestras Independientes por Tipo de Profesión*.....62

Tabla 15. *Reporte de Indicadores Cuantitativos* 63

Tabla 16. *Items en Revisión*.....70

Tabla 17. Clasificación de los Items en Revisión.....81

ÍNDICE DE GRÁFICOS

Página

Gráfico 1. Edad 44

Gráfico 2. Habilidad Cuantitativa 54

Gráfico 3: Rendimiento Académico en Postgrado 55

Gráfico 4: Item 4 71

Gráfico 5: Item 5 71

Gráfico 6: Item 9 72

Gráfico 7: Item 10 72

Gráfico 8: Item 14 73

Gráfico 9: Item 18 73

Gráfico 10: Item 25 74

Gráfico 11: Item 26 74

Gráfico 12: Item 30 75

Gráfico 13: Item 32 75

Gráfico 14: Item 35 76

Gráfico 15: Item 36 76

Gráfico 16: Item 37 77

Gráfico 17: Item 39 77

Gráfico 18: Item 40 78

INTRODUCCIÓN

La meta de la ciencia es la descripción, predicción y explicación de los fenómenos empíricos. Establecer principios generales para explicar y establecer dichos fenómenos (Díez y Moulines, 1999). La medida es un componente fundamental en este proceso e implica el uso de herramientas matemáticas. De allí la importancia de la psicometría para la Psicología como ciencia, pues provee un conjunto de métodos, técnicas y teorías implicadas en la medición de variables psicológicas (variables que explican las diferencias individuales en el comportamiento de los individuos), orientados a desarrollar modelos cuantitativos para transformar los hechos en datos y utilizar estos datos en la extracción de aspectos relevantes (significativos) sobre el comportamiento de los sujetos (Muñiz, 2001).

Para toda ciencia el conocimiento de los instrumentos de medida es fundamental, sin embargo, para la Psicología es aun más importante, dado que los atributos psicológicos no son directamente observables, son constructos que intentan explicar la conducta y el grado en que un sujeto está caracterizado por ese constructo debe inferirse a partir de su conducta (Martínez, 1995).

Muñiz (2001, p.17) clasifica la temática psicométrica en 5 grandes áreas: (a) la Teoría de la Medición, referida a los fundamentos teóricos de la medida; (b) la Teoría de los Tests, relacionada con los modelos que sustentan la construcción y uso de los tests; (c) el Escalamiento Psicológico; (d) el Escalamiento Psicofísico; y (e) el uso de Técnicas Multivariadas que sustentan la construcción y el análisis de los instrumentos de medida. Partiendo de esta clasificación, el presente trabajo se ubica dentro de la Teoría de los Tests, específicamente de la Teoría Clásica de los Test, pues pretende, haciendo uso de los supuestos y técnicas derivadas de ésta, determinar las características psicométricas de un instrumento de medición psicológica.

El instrumento objeto de estudio se denomina "Prueba de Razonamiento Básico Cuantitativo" (RBC) utilizado por la Universidad Católica Andrés Bello (La Universidad). La prueba de RBC es una prueba de habilidad cuantitativa y forma parte del tipo de pruebas denominadas de aptitudes académicas básicas, orientados a predecir rendimiento académico, específicamente a nivel de postgrado. Este tipo de pruebas es desarrollado, más que sobre una base teórica que explique el rendimiento académico, sobre la base del valor predictivo que tienen los resultados de la prueba en el rendimiento académico de los estudiantes (Kerlinger y Lee, 2002). De allí la importancia de determinar y mejorar la calidad técnica o psicométrica del instrumento en términos del grado en que a partir de él se logran medidas consistentes y válidas de acuerdo con su objetivo fundamental: la predicción del rendimiento académico a nivel de postgrado.

El esquema de desarrollo del trabajo contempla la revisión de la literatura relacionada con el tema y dentro del apartado metodológico, el análisis de los tres aspectos fundamentales del funcionamiento estadístico de la prueba: (a) análisis de la confiabilidad a través de métodos que requieren de una sola aplicación y que se sustentan en la consistencia interna del instrumento; (b) análisis de la validez, considerando los tres niveles esenciales de validación: de contenido, de criterio y de constructo; y (c) análisis de *items*, como medio para identificar aquéllos que mejor se comportan estadísticamente, los que podrían ser redimensionados y aquéllos que definitivamente deben ser eliminados. El análisis de *items* es un medio efectivo para orientar el rediseño de la prueba con el objetivo de maximizar su valor, en términos de su objetivo. Dado que la prueba objeto de estudio forma parte del banco de *items* de La Universidad, por lo que es información confidencial, no se presenta un ejemplar de la misma sino los resultados asociados con su procesamiento.

El análisis de la prueba de RBC se justifica porque proveerá lineamientos para mejorar su calidad psicométrica en el contexto de la predicción del rendimiento académico en postgrado.

El objetivo es, entonces, caracterizar, desde la perspectiva de la Teoría Clásica de los Tests, el comportamiento psicométrico de la prueba de Razonamiento Básico Cuantitativo (modelo PG-38: RBC), utilizada por la Universidad en el proceso de selección de aspirantes a cursar estudios de postgrado, realizado en junio del año 2003.

El trabajo se desarrolla con limitaciones asociadas con el escaso número de investigaciones nacionales accesibles relacionadas con el tema; por lo que se recurrirá a investigaciones realizadas en el contexto norteamericano sobre pruebas similares (Graduate Record Examination), específicamente sobre la subprueba de habilidad cuantitativa, la cual tiene el mismo objetivo.

MARCO TEÓRICO

Como ya se mencionó, el objetivo de la ciencia es la descripción, predicción y explicación de los fenómenos empíricos (Díez y Moulines, 1999). Este objetivo, visto desde la Psicología, exige el desarrollo y uso de métodos, técnicas y teorías para la medición de las variables psicológicas. Pues la medición es el proceso que permite, mediante el uso de herramientas matemáticas, transformar los hechos en datos y utilizar estos datos para extraer aspectos significativos sobre el comportamiento de los sujetos. (Muñiz, 2001).

La parte de la Psicología que estudia los modelos, técnicas y herramientas relacionadas con la medición de variables psicológicas, es la psicometría.

Dentro de la psicometría existen diferentes enfoques desde los que se puede abordar y sustentar el uso de pruebas o tests como instrumentos de medición. Uno de estos enfoques es provisto por la Teoría Clásica de los Tests, la cual, partiendo de un modelo y supuestos, permite juzgar el comportamiento psicométrico de las pruebas y, en función de ello, la validez de las inferencias realizadas a partir de la puntuaciones de ellas derivadas (Martínez, 2001). El presente trabajo representa una aplicación de la Teoría Clásica de los Tests en el análisis de un test de habilidad cuantitativa denominado Prueba de Razonamiento Básico Cuantitativo, diseñado y utilizado con fines exclusivamente de selección (predicción del rendimiento académico) a nivel de postgrado. Para ello es necesario a continuación:

A. Presentar los aspectos fundamentales de la Teoría Clásica de los Tests, sobre los que se basará el análisis que se realizará sobre la prueba.

B. Presentar las características de la prueba que se analizará (prueba de Razonamiento Básico Cuantitativo).

C. Definir el rendimiento académico en postgrado, dado que es la variable para cuya predicción es diseñada la prueba de Razonamiento Básico Cuantitativo.

D. Presentar estudios que ofrezcan evidencia empírica sobre el comportamiento de este tipo de pruebas en la predicción del rendimiento académico a nivel de postgrado.

A.- Teoría Clásica de los Tests

El desarrollo de la Teoría Clásica de los Tests (TCT) se basa en trabajos de Spearman (1904, 1907, 1913) quien, influido por Galton y Pearson, formuló la teoría de las puntuaciones verdaderas y error, denominada Teoría Clásica de los Tests. Esta teoría supone un modelo estadístico que fundamenta las puntuaciones de los tests y permite la estimación de los errores de medida que se producen en el proceso de medición. El objetivo de esta teoría es, como ya se mencionó, estimar los errores de medida, para a partir de allí tener una aproximación más precisa al puntaje verdadero de los sujetos (Martínez, 1995; Muñiz, 2001).

La Teoría Clásica de los Tests es una representación simbólica simple y útil de los factores que influyen en las puntuaciones observadas en un test, que describen el efecto o influencia de los errores de medida en los puntajes obtenidos y su relación con los puntajes verdaderos (Martínez, 1995).

El modelo establece que el puntaje obtenido por un sujeto en un test (X) es igual a la suma de su puntaje verdadero (V) más el puntaje de error (e). Claramente queda establecida la relación aditiva y por lo tanto un modelo lineal donde: $X_i = V_i + e$. (Anastasi y Urbina. 1998; Magnusson, 1972:, Martínez, 1995, Muñiz, 2001, Nunnally y Berstein, 1995).

El modelo, tiene tres supuestos que, según Muñiz (2001), son a priori y no son comprobables empíricamente en forma directa sino que deben hacerse deducciones que confirmen o no el modelo. Los supuestos son los siguientes:

1. El puntaje verdadero del sujeto en el test es la media aritmética de las puntuaciones empíricas obtenidas en infinitas aplicaciones, por lo tanto el puntaje verdadero es un concepto matemático (Anastasi y Urbina. 1998; Magnusson, 1972:, Martínez, 1995, Muñiz, 2001, Nunnally y Berstein, 1995).
2. No hay correlación entre los puntajes verdaderos de los sujetos en un test y sus respectivos errores de medida ($r_{Ve} = 0$), es decir que la magnitud de los errores no se relaciona sistemáticamente con la magnitud en la que se posee el rasgo (Anastasi y Urbina. 1998; Magnusson, 1972:, Martínez, 1995, Muñiz, 2001, Nunnally y Berstein, 1995).
3. La correlación entre los puntajes de error de diferentes aplicaciones del test es cero ($r_{ee} = 0$) para un número infinito de individuos, dado que los errores son al azar (Anastasi y Urbina. 1998; Magnusson, 1972:, Martinez, 1995, Muñiz, 2001, Nunnally y Berstein, 1995).

A partir del modelo lineal descrito y del conjunto de supuestos en que se basa, pueden analizarse psicométricamente los tests, específicamente su confiabilidad, su validez y sus componentes o *items*.

A continuación se presenta una síntesis de los aspectos teóricos relevantes para el análisis de la (a) confiabilidad, (b) la validez y (c) los *items* del la prueba.

Confiabilidad

De acuerdo con Magnusson (1972) los datos que arrojan las pruebas deben ser confiables desde dos puntos de vista: significativos y reproducibles. El primer punto se refiere a la validez y el segundo a la confiabilidad propiamente dicha.

La confiabilidad se refiere a la exactitud y consistencia de los datos obtenidos, es decir, un instrumento se considera confiable si las medidas que se obtienen a partir de él carecen de error. La inconfiabilidad es el producto de la sensibilidad del instrumento a fuentes de varianza de error aleatorio (Magnusson, 1972; Muñiz, 2001).

En la medida en que los puntajes obtenidos a partir de un instrumento estén libres de error, tales puntajes serán estables de una aplicación a otra, siempre y cuando no existan razones empíricas o teóricas para suponer que ha variado el rasgo.

Estadísticamente, la confiabilidad es la proporción de la varianza total en los resultados que es varianza verdadera o que no es error. De esta manera puede estimarse el grado en que las diferencias observadas en los puntajes de un grupo de sujetos en un test se deben a verdaderas diferencias en el rasgo y no a errores aleatorios. Este concepto fundamenta el cálculo del error de medición de un solo resultado, con lo que se puede predecir variación de los resultados de un solo individuo debida a factores aleatorios (Anastasi y Urbina, 1998).

Se considera varianza de error aquélla producida por cualquier condición irrelevante para el objetivo de la prueba. En consecuencia, se puede estimar la confiabilidad de tantas maneras como condiciones haya que afecten los resultados de las pruebas (Anastasi y Urbina, 1998).

El problema de la confiabilidad es estimar la magnitud del puntaje de error que contienen los puntajes arrojados por un test, para así, tener una mayor aproximación al puntaje verdadero de los sujetos. Para abordar este problema se utiliza el modelo lineal de la Teoría Clásica de los Tests. En 1904 Spearman introdujo el concepto de coeficiente de confiabilidad como la correlación entre medidas repetidas de un test.

Existen diferentes formas de expresar la confiabilidad de un instrumento:

1. Coeficiente de confiabilidad (r_{tt}): El coeficiente de confiabilidad expresa el grado de correspondencia entre las mediciones, puede adoptar valores entre cero y uno, y se interpreta como la proporción de la varianza total que es varianza de puntajes verdaderos, por lo que en la medida que sea mayor, más se aproxima el puntaje obtenido por los sujetos a su puntaje verdadero (Anastasi y Urbina, 1998; Martínez, 1995).
2. Error Estándar de Medida (S_e). El Error Estándar de Medida es una medida empírica de la varianza de error alrededor de un puntaje verdadero. Se obtiene por el producto de la desviación típica de la distribución de puntajes obtenidos por la raíz cuadrada de uno menos el coeficiente de confiabilidad: $S_e=S_y\sqrt{1-r_{tt}}$. Su mayor utilidad se asocia a la interpretación de puntajes individuales (Magnusson, 1972).

La estimación del coeficiente de confiabilidad puede realizarse por diferentes métodos, los cuales son afectados por distintas fuentes de varianza de error. Los métodos más utilizados son:

1. Método de Pruebas Paralelas: el coeficiente de confiabilidad se obtiene a partir de la correlación de Pearson entre puntuaciones de los mismos sujetos en dos tests considerados equivalentes o paralelos. Las medidas paralelas, teóricamente tienen iguales puntajes verdaderos e iguales varianzas de error, lo que debe traducirse en igualdad de medias y varianzas en los

puntajes obtenidos. Una flexibilización de los criterios de equivalencia es el de las pruebas tau-equivalentes que supone la equivalencia de los puntajes verdaderos pero admite diferencias en la varianza de error (Martínez, 1995).

El coeficiente obtenido de esta manera indica el grado en que las puntuaciones están libres de errores aleatorios y el grado en que la pruebas son equivalentes. Este método resulta costoso en términos tanto de la construcción de los instrumentos que cumplan con las condiciones de paralelismo como de la posibilidad real de obtener dos evaluaciones de los mismos sujetos.

2. Método de Test – Retest: el coeficiente de confiabilidad se obtiene por la correlación de Pearson entre los resultados de los mismos sujetos evaluados dos veces con el mismo test. El coeficiente obtenido indica la estabilidad de las puntuaciones. Este coeficiente debe ser interpretado con cuidado debido a que se ve afectado por la interacción entre: (a) el intervalo entre las aplicaciones, (b) los efectos de la práctica y/o la memoria, y (c) las verdaderas variaciones en el rasgo.

3. Métodos de una sola aplicación: el coeficiente de confiabilidad se obtiene a partir de la correlación entre partes (entre mitades o entre todos los *items* considerados paralelos) del mismo test. Este tipo de métodos son de consistencia interna porque su magnitud indica el grado en que todos los *items* son paralelos o el grado en que dos mitades de la prueba son equivalentes. Dentro de este tipo de métodos se encuentra el alpha de Cronbach como un índice de consistencia interna o índice de precisión, que indica el grado en que los *items* que forman un test están correlacionados entre sí. El alpha de Cronbach es muy similar al KR-20 y KR-21, el cual es un índice desarrollado por Kuder y Richardson. Comparativamente el alpha es mejor indicador de confiabilidad cuando el *item* tiene más de una respuesta correcta (respuestas ponderadas) mientras que cuando el *item* tiene una

sola respuesta correcta, alpha y KR-20 producen exactamente el mismo resultado (Nelson, 2001). Aunque algunos autores señalan que altos coeficientes de consistencia interna indican que se trata de pruebas unidimensionales, Pedhazur y Schmelkin (1991, cp. Nelson, 2001, p.116) señalan que el hecho de que un instrumento sea internamente consistente no significa que necesariamente sea unidimensional. La consistencia interna es una condición necesaria pero no suficiente para la unidimensionalidad.

El método de División por mitades permite estimar la confiabilidad del instrumento a través de la correlación entre las mitades del instrumento aplicado. Este coeficiente considera la confiabilidad para la mitad del instrumento. Esta correlación indica el grado en que pueden considerarse equivalentes las mitades. A partir de este coeficiente se estima el coeficiente de confiabilidad para el test completo aplicando la fórmula de Spearman Brown. Al análisis de la equivalencia entre las mitades subyacen las mismas consideraciones del análisis de la equivalencia entre pruebas paralelas (Magnusson, 1972).

La magnitud del coeficiente de confiabilidad, independientemente del método que se utilice, tiene como máximo valor posible 1,00 y su valor esperado para pruebas confiables según Hokins (1998, cp. Nelson, 2001, p.117) es de 0,90, mientras que Linn y Gronlund (1995, cp. Nelson, 2001, p. 117) señala que son aceptables coeficientes entre 0,65 y 0,80. La estimación del coeficiente de confiabilidad puede verse afectada por dos condiciones:

1. La homogeneidad de la muestra de sujetos a los que se le aplica el test. En la medida que la muestra es más homogénea, la proporción de varianza verdadera disminuye y la de error permanece igual, por lo que esta última representa mayor proporción del puntaje obtenido, lo que produce como consecuencia una subestimación del coeficiente de confiabilidad. En otras palabras, la homogeneidad de la muestra de sujetos afecta al estimador de

la confiabilidad (coeficiente de correlación), pero no a la precisión con que el test permite estimar los puntajes verdaderos de los sujetos (Magnusson, 1972).

2. La longitud del instrumento. La confiabilidad es una función de la longitud del instrumento. En la medida que se incremente la longitud de un test con *items* que midan el mismo puntaje verdadero que el resto de los *items*, la confiabilidad del instrumento aumentará, pero esto tiene una restricción práctica asociada tanto con los costos de producción de los instrumentos, como con la posibilidad de que la fatiga afecte la ejecución de los evaluados, limitando, como consecuencia, la confiabilidad de la prueba (Magnusson, 1972).

Validez

Los puntajes obtenidos por los sujetos en un test son entidades numéricas sin significado intrínseco. Para interpretar la puntuación en un test como una medida del constructo latente se requieren pruebas de su validez (Martínez, 1995).

La validez se refiere al grado en que las propiedades de las medidas obtenidas a través de un test se corresponden con la característica que se dice medir. En consecuencia, la validez se refiere al grado en que las inferencias y acciones basadas en las puntuaciones obtenidas son adecuadas. De acuerdo con los *Standards for Educational and Psychological* Testing (citado por Martínez, 1995, p. 332) "la validez se refiere a la adecuación, significación y utilidad de las inferencias específicas hechas a partir de las puntuaciones de los tests".

La evolución del concepto de validez se inició con una aproximación correlacional, ateórica, basada en el valor predictivo y de diagnóstico de los

tests. Esta aproximación es especialmente útil dentro del contexto de selección. Sin embargo, el uso de los tests con otros fines generó una evolución de la noción de validez, que considera, además del enfoque correlacional, la representatividad de los contenidos evaluados y la correspondencia empírica con la base teórica que sustenta el test.

Partiendo de esta perspectiva integradora se identifican los siguientes aspectos de la validez que deben ser considerados al analizar las puntuaciones de un test.

1. Validez de contenido: se refiere al grado en que los *items* que forman un test son representativos del dominio o universo de conductas que se pretende medir. Tal representatividad tiene dos elementos, cantidad y relevancia de los *items*. Este tipo de validez se garantiza mediante la rigurosidad metodológica en el proceso de construcción del instrumento, donde es esencial la construcción de una tabla de especificaciones y la validación por parte de jueces expertos tanto de la tabla como de los *items* derivados de ella. Este tipo de validez es especialmente importante para los tests que evalúan conocimientos o nivel de aprovechamiento y tests criteriales (Magnusson, 1972; Martínez, 1995).

2. Validez de criterio: es un tipo de validez basada en las correlaciones entre las medidas del test y una variable o característica que es la que se pretende diagnosticar o predecir denominada variable de criterio. De acuerdo con la temporalidad en la que se puede disponer de la medida de criterio este tipo de validez, puede clasificarse como concurrente, predictiva y retrospectiva. Al juzgar este tipo de validez deben manejarse algunas consideraciones referidas al criterio: su confiabilidad, validez, pertinencia y factibilidad, y posible contaminación, su complejidad y la restricción del rango asociada con los cortes de selección (Anastasi y Urbina, 1999; Magnusson, 1972; Martínez, 1995). El concepto de validez de criterio puede complementarse

con la noción de utilidad de un test, la cual se refiere a los beneficios derivados de la utilización del test en la toma de decisiones particulares. Este tipo de evaluación toma en cuenta no sólo el coeficiente de validez de la prueba para predecir el criterio particular sino también la tasa base y la razón de selección (Anastasi y Urbina, 1998).

Un test puede ser válido para tantos criterios como usos se le dé. Uno de los criterios más frecuentemente usado para validar pruebas de habilidades verbal y cuantitativa es el rendimiento académico, por lo que estas pruebas de habilidades se consideran pruebas de aptitud académica. (Anastasi y Urbina, 1998; Magnusson, 1972; Martínez, 1995).

Los criterios deben cumplir con ciertas condiciones: tales como la relevancia, confiabilidad, validez y factibilidad de medida (Martínez, 1995).

La validez de criterio se basa en estudios correlacionales, los coeficientes obtenidos pueden verse afectados por distintos artefactos o tipos de errores asociados con el diseño de la investigación: Schmidt y Hunter (1990 cp. Martínez, 1995, p.406) señalan 11 artefactos que pueden afectar el tamaño de las correlaciones: (a) error muestral, lo que depende estrechamente del tamaño muestral, (b) error de medida en la variable dependiente, es decir baja confiabilidad de la variable de criterio; (c) error de medida en la variable independiente, es decir baja confiabilidad de la variable predictora; (d) dicotomización de la variable dependiente; (e) dicotomización de la variable independiente; (f) variación del rango en la variable independiente; (g) variación del rango en la variable dependiente; (h) validez de constructo imperfecta en la variable dependiente; (i) validez de constructo imperfecta en la variable independiente; (j) errores de trascripción; y (k) errores extraños.

3. Validez de constructo: es una noción integradora que considera tanto la validez de contenido, como la de criterio, en un marco general para probar

hipótesis sobre las relaciones relevantes, teóricamente (Messick, 1980, cp. Martínez, 1995, p. 331). Implica entonces, la comprobación empírica de relaciones derivadas de la teoría que respaldd el rasgo medido. Se trata de poner a prueba distintas hipótesis acerca de la estructura interna del constructo y de sus relaciones con otras variables (Martínez, 1995).

La validez de constructo se refiere al grado en que las interpretaciones y decisiones basadas en los puntajes obtenidos por un test, son respaldadas por evidencia empírica y las bases teóricas.

Para obtener indicadores sobre este tipo de validez es esencial contar con una sólida teoría del constructo que se pretende medir. Para obtener indicadores de validez de constructo se formulan hipótesis sobre el comportamiento de los resultados del test y, mediante técnicas estadísticas, se someten a comprobación. Los procedimientos frecuentemente recomendados para estimar este tipo de validez son:

a. Grupos contrastados: consiste en la comparación de grupos que deben contrastar en el rasgo medido, por lo tanto se esperan diferencias estadísticamente significativas entre las medidas. Aquí entrarían estudios de: diferenciación con la edad, para los rasgos asociados con maduración, evaluaciones antes y después de una intervención, contraste de grupos que se suponen extremos o con ejecución diferente en el rasgo. (Anastasi y Urbina, 1998; Magnusson, 1972; Martínez, 1995).

b. Correlación con otros tests o medidas válidas del mismo rasgo, se esperan puntuaciones con comportamiento similar y por lo tanto, alta correlación entre ellos.

c. Análisis factorial, la cual es una técnica estadística que permite analizar intercorrelaciones entre puntajes obtenidos e identificar en ellas las variables o factores que explican la covariación de las puntuaciones. Una

vez identificados los factores, pueden utilizarse para caracterizar la composición factorial del test (Martínez, 1995).

d. Matrices Multimétodo-Multirasgo, es un método que considera múltiples indicadores de validez (convergente y discriminante), además de estimar la confiabilidad de los métodos utilizados para medir diferentes rasgos. Hay un patrón de correlaciones esperadas a partir del cual se juzga la validez de los instrumentos y de la validez de las inferencias realizadas a partir de ellos (Campbell y Fiske, 1959).

Es importante resaltar que, independientemente de los métodos utilizados para la validez, ésta es específica al tipo de inferencias que se realizan con las puntuaciones y a las características de la muestra donde se obtuvieron los puntajes. La validez es un proceso y es específica (Magnusson, 1972).

Análisis de *items*

Los items pueden ser analizados en forma (a) cualitativa en función de su forma (redacción) y contenido (aspecto que según la tabla de especificaciones debe explorar); (b) cuantitativa, en función de sus propiedades estadísticas, lo que se refiere específicamente a su nivel de dificultad y poder discriminativo; y (c) gráfica, en función del comportamiento que se espera tengan los distractores y la respuesta correcta (Magnusson, 1972).

El nivel de dificultad de los *items* se define por la proporción de sujetos que lo responde en forma acertada (p), la distribución de niveles de dificultad de las pruebas de habilidad suele ser creciente, con mayor cantidad de *items* promedio y la cantidad de *items* disminuye hacia los niveles de dificultad extremos. De esta manera, se pretende lograr que los resultados indiquen el grado o nivel hasta el que el sujeto posee el rasgo en el continuo latente y obtener la distribución normal que subyace a la distribución del rasgo latente. Los *items* que son respondidos por todos los sujetos o por ninguno no discriminan

entre los sujetos en el rasgo, pero sí informan sobre los límites inferiores y superiores de dificultad de la prueba (piso y techo) para el grupo que la presenta. Cuando las pruebas están muy cargddas de este tipo de *items*, la distribución de puntajes obtenidos se diferencia de la normal y tiende a colearse positiva o negativamente, según la prueba esté más cargada de *items* difíciles o fáciles, respectivamente (Anastasi y Urbina 1998; Magnusson, 1972, Martínez, 1999, Muñiz, 2001).

La varianza del ítem indica en qué medida esta diferencia entre los sujetos viene dada por el producto de la proporción de sujetos que lo responde correctamente (p), y por la proporción que no lo responde correctamente ($q=1-p$). $S^2 = p_x q$. La máxima varianza de un *item* se da con niveles de dificultad promedio, es decir cuando el 50% de los sujetos los responde correctamente (Anastasi y Urbina, 1998; Magnusson, 1972, Martínez, 1995, Muñiz, 2001).

El poder discrimitativo del *item* se refiere al grado en que éste diferencia entre los sujetos del grupo superior e inferior en la distribución de puntajes totales. El poder discriminativo de un ítem puede establecerse a partir del análisis de grupos extremos o de la correlación entre la respuesta correcta y el puntaje total en el test. En este último caso, el coeficiente de correlación más utilizado como índice de discriminación es el coeficiente de correlación punto-biserial (Popham, 197; cp. Nelson, 200, p. 118). De acuerdo con Haladyna (1994, cp. Nelson, p. 118) en la Teoría Clásica de los Tests, el indicador del poder discriminativo del *item* es la correlación punto-biserial entre el *item* y el test. Por su parte Lord y Novick, (1968, cp. Nelson, 2001, p.118) señalan que la correlación punto-biserial es igual a la biserial multiplicada por un factor que depende del nivel de dificultad del *item*. Nelson (2001) señala que si se obtienen los coeficientes de correlación biserial y punto-biserial partiendo de los mismos datos, el biserial excederá al punto biserial en un 25%. Es decir, el coeficiente biserial siempre excederá al punto-biserial.

Los valores esperados para los coeficientes biserial y punto-biserial como adecuados indicadores del poder discriminativo de un *item* son de 0,40 y 0,30, respectivamente, a menos que se trate de pruebas diseñadas por expertos, donde los valores mínimos esperados son de 0,30 para el biserial y 0,20 para el punto biserial (Hills, 1976, cp. Nelson, 2001, p 119) o de muestras grandes (iguales o superiores a 100 sujetos) (Nunnally y Berstein, 1995) para los cuales es suficiente un punto-biserial de 0,20.

En la medida que el coeficiente de correlación (biserial o punto-biserial) sea más alto, indica la tendencia a que: (a) los sujetos que lo responden correctamente se ubiquen en la parte superior de la distribución de puntajes totales, y (b) los sujetos que responden incorrectamente se encuentren en la parte inferior de la distribución de puntajes totales (Magnusson, 1972, Anastasi y Urbina, 1998).

El coeficiente de correlación punto biserial es sensible al nivel de dificultad del *item*. Este último puede asumir su máximo valor (0,7978) cuando p y q son iguales a 0,50. Por lo tanto puede afirmarse que la varianza del ítem limita matemáticamente el poder discriminativo (Anastasi y Urbina, 1999; Magnusson, 1972)

Aunque el poder discriminativo del *item* puede identificarse con el coeficiente de correlación punto-biserial, el comportamiento estadístico de los distractores está estrechamente relacionado con él. En este sentido, se espera que los distractores tengan correlación negativa o nula, con el puntaje total en el test, lo que indica que quienes los seleccionan como opción correcta se ubican en la parte inferior de la distribución de puntajes totales. Por esta misma razón, la media de quienes eligen los distractores debe ser baja, o por lo menos inferior a la media de los que eligen la opción correcta (Anastasi y Urbina, 1998; Magnusson, 1972).

B.- Prueba de Razonamiento Básico Cuantitativo

La Educación superior ofrece las herramientas técnicas y conceptuales para abordar el desempeño en el ámbito profesional.

En el sector universitario venezolano existen tres tipos de universidades: (a) Universidades Nacionales Autónomas, (b) Universidades Nacionales Experimentales y (c) Universidades Privadas (Oficina de Planificación del Sector Universitario, OPSU, 2001).

Las universidades privadas son instituciones fundadas por personas naturales o jurídicas de carácter privado, que están al servicio de la nación y deben realizar una función rectora en la educación, la cultura y la ciencia. Para lo cual, según la OPSU (2001), sus actividades deben dirigirse a: (a) crear, asimilar y difundir el saber mediante la investigación y la enseñanza; (b) completar la formación integral iniciada en los ciclos educacionales anteriores; y (c) formar los equipos profesionales y técnicos que necesita la Nación para su desarrollo.

Las instituciones de Educación Superior y dentro de ellas las universidades privadas, son regidas por la Ley de Universidades y el Reglamento de los Institutos y Colegios Universitarios y por sus reglamentos internos los cuales establecen su estructura y sistema académico, administrativo y de gobierno, así como las condiciones y requisitos en cuanto a su funcionamiento y políticas de estudio, lo que incluye el sistema de captación, selección e ingreso de aspirantes.

El sistema de selección de las universidades privadas es diseñado por las propias instituciones y, cobra mayor importancia en la medida en que: (a) la demanda supera los cupos disponibles y (b) las instituciones se orientan a formar profesionales, especialistas, maestros y doctores con criterios de excelencia académica (PROINVESCA, 1998b).

Para las universidades es preciso identificar e ingresar los aspirantes que posean las aptitudes para aprovechar los recursos, responder con eficiencia los niveles de exigencia y traducir esto, en primer lugar, en logro académico y consecuentemente, en éxito profesional. Para lograr esto deben diseñarse sistemas de selección que consideren características o variables que se relacionen con el desempeño académico superior y cuya medición sea factible y segura (en términos de confiabilidad y validez de los resultados obtenidos), a través de procesos masivos (PROINVESCA, 1998b).

Hay múltiples variables que permiten predecir el rendimiento académico de un estudiante, tales como: (a) las calificaciones obtenidas en estudios previos (Nettles, 1985); (b) el ambiente familiar (Majorobanks, 1991; Hoffman, 1986); (c) las características intelectuales de los estudiantes (Crouse y Trusheim, 1988; Crook, 1984); los hábitos de estudio (Matt, Pecherly y Cervantes, 1986); (d) la percepción de eficacia personal (Ferrari y Parker, 1992); (e) el Locus de control (Ferrari y Parker, 1992; García, 1985a); (f) la Motivación (García, 1985b); (g) las habilidades académicas básicas (Nunnally y Berstein 1995; PROINVESCA, 2003; Snack y Kelley, 2002), entre otros.

El diseño de un sistema de selección eficiente requiere no sólo identificar aquellas variables relevantes en la predicción sino además, establecer un mecanismo para estimarlas de manera confiable y válida. Tal mecanismo debe ser manejable en procesos que impliquen la evaluación de un gran número de aspirantes. De allí que los procesos de selección suelen restringirse a la consideración de indicadores de rendimiento académico en estudios previos y a puntuaciones en pruebas de habilidades básicas que son obtenidas a partir de pruebas psicométricas estandarizadas.

La Universidad Católica Andrés Bello (la Universidad), posee un sistema de selección para estudios de postgrado basado en un modelo compensatorio que combina los resultados de una batería de pruebas de aptitudes (verbal,

cuantitativa y analítica) y el rendimiento académico en el pregrado, estimado a partir del promedio de calificaciones obtenidas.

El presente trabajo tiene como objetivo analizar las características psicométricas de uno de los componentes del sistema de selección de la Universidad utilizado en el proceso de admisión de junio del 2003: La prueba de habilidad cuantitativa denominada prueba de: Razonamiento Básico Cuantitativo (RBC). Este modelo forma parte de la batería de selección identificada con el código PG-38.

Esta prueba (RBC) explora el nivel de habilidad cuantitativa de los aspirantes, la cual se define como el nivel de razonamiento lógico matemático, asociado con el uso de la simbología matemática y de las relaciones numéricas, lo que depende de la capacidad del individuo para: (a) procesar información, es decir, su resolución supone la competencia de los sujetos para: (a) representar información en diferentes esquemas y códigos formales; (b) relacionar información con énfasis en los procesos de análisis y comparación; (c) inferir información con base en la información disponible; y (d) resolver problemas y tomar decisiones, reconociendo fases y estrategias que optimicen el alcance de una meta a partir de un estado inicial dado o conocido. Todo ello dentro del contexto lógico-matemático utilizando simbología cuantitativa (PROINVESCA, 2002).

La concepción de la prueba de Razonamiento Básico Cuantitativo, es de la empresa Proyectos e Investigaciones Académicas (PROINVESCA), la cual desde el año 1983 diseña este tipo de pruebas, que explora la capacidad del individuo para enfrentarse a situaciones en las cuales es necesario el empleo de simbología matemática y el establecimiento de relaciones abstractas. La prueba se diseña de manera tal que tenga la mayor independencia posible del nivel de conocimientos alcanzados por la persona en su educación formal. Esta es la diferencia que tiene con una prueba de conocimientos de matemáticas.

La eficiencia del sujeto en la prueba de RBC depende de su capacidad para procesar información a través de: (a) la interpretación de diferentes sistemas de representación (dibujos, esquemas, gráficos, tablas, palabras); (b) la relación entre información con énfasis en los procesos de análisis y comparación; (c) la inferencia de información, centrada en los procesos de inducción y deducción; y (d) la resolución de problemas y toma de decisiones, reconociendo fases y estrategias que optimicen el alcance de una meta a partir de un estado inicial dado o conocido (PROINVESCA, 2002).

La prueba RBC está compuesta por 40 *items* o preguntas de opción múltiple y selección simple, cada una con cuatro opciones de respuesta de corrección binaria. Para la resolución de la prueba el evaluado dispone de 60 minutos. El puntaje obtenido por cada sujeto en la prueba es un indicador de su nivel de habilidad cuantitativa: en la medida que el puntaje es mayor, indica un mayor nivel de habilidad cuantitativa.

La concepción de la prueba se basa en la identificación de las habilidades cuantitativas que subyacen a la ejecución cuantitativa de los sujetos. En este sentido, se exploran específicamente las siguientes dimensiones:

1. Traducción de expresiones del lenguaje cotidiano al lenguaje matemático: incluye planteamientos que contienen relaciones expresadas en forma cotidiana (ecuaciones, inecuaciones) que contienen constantes y variables y se le pide al aspirante que identifique la expresión matemática que expresa la misma relación planteada en expresiones de uso cotidiano.
2. Traducción de expresiones del lenguaje matemático al lenguaje cotidiano: en esta dimensión se explora el proceso inverso al anterior: el aspirante debe ser capaz de traducir la expresión matemática en relaciones expresadas en términos cotidianos.

3. Seriación: los *items* de esta dimensión responden a dos modalidades: (a) partiendo de una lista de información de carácter secuencial, se le solicita interpolar o extrapolar nueva información; y (b) se presenta una ley de formación secuencial y se solicita aplicarla para obtener uno o más términos particulares de la secuencia.
4. Estimación: los planteamientos requieren del sujeto la realización de aproximaciones matemáticas.
5. Analogías figurativas: implica la identificación de relaciones de segundo orden, es decir una relación de relaciones entre figuras.
6. Razonamiento lógico: Implica el manejo de la estructura lógica de expresiones verbales.
7. Equivalencia entre expresiones matemáticas: a través de los *items* que forman esta dimensión, el aspirante debe reconocer expresiones algebraicas equivalentes y no equivalentes, ya sea en forma directa o a través de alguna transformación.
8. Interpretación de gráficos: se presentan gráficos que expresan relaciones entre variables y se solicita extraer de allí información o asociarle alguna ecuación. Los gráficos presentados son de distinta naturaleza.
9. Razonamiento proporcional: los *items* de esta dimensión requieren que el sujeto establezca igualdad o desigualdad entre razones, es decir entre cocientes de cantidades.
10. Interpretación de tablas: se le presenta al sujeto información en tablas y se le solicita que interprete o infiera información a partir de las relaciones presentadas.

11. Razonamiento estadístico: este tipo de *items* requieren operar e interpretar datos como base para conclusiones válidas acerca de la muestra o la población de donde provienen.

12. Razonamiento algorítmico: se ofrece al aspirante un procedimiento de cálculo a partir del cual el sujeto debe derivar un resultado.

La distribución de *items* por contenido se presenta en la Tabla 1.

Tabla 1. *Especificaciones de la Prueba de Razonamiento Básico Cuantitativo*

	Dimensión de la prueba	**Cantidad de *items***
1	Traducción de expresiones del lenguaje cotidiano al lenguaje matemático	4
2	Traducción de expresiones del lenguaje matemático al lenguaje cotidiano	3
3	Seriación	3
4	Estimación	3
5	Analogías figurativas	2
6	Razonamiento lógico	2
7	Equivalencia entre expresiones matemáticas	4
8	Interpretación de gráficos	4
9	Razonamiento proporcional	5
10	Interpretación de tablas	3
11	Razonamiento estadístico	4
12	Razonamiento algorítmico	3

La distribución de los ítemes por niveles de dificultad estimados en el diseño del instrumento es presentada en la tabla 2.

Tabla 2. ***Niveles de Dificultad (Estimados) de los Items***

	Nivel de dificultad				
	Muy fácil	**Fácil**	**Promedio**	**Difícil**	**Muy difícil**
Proporción de sujetos que responde correctamente al *item*	**5%**	**20%**	**50%**	**20%**	**5%**

La prueba RBC, de acuerdo con los criterios de clasificaciones ofrecidos por Martínez (1995), puede describirse como:

1. Por su método, una prueba **psicométrica,** dado que evalúa las respuestas correctas a cada item en forma independiente y son calificadas por un punto (criterios cuantitativos) y el resultado final es la sumatoria de respuestas correctas menos un tercio de las incorrectas (una puntuación cuantitativa).
2. Por su finalidad, **de diagnóstico**: pretende determinar el grado en que los sujetos poseen habilidad cuantitativa.
3. Por el planteamiento del problema, **de ejecución máxima**: cada *item* plantea un problema que requiere del uso del razonamiento cuantitativo, lo que, se supone, depende directamente del nivel de habilidad cuantitativa del sujeto.
4. Por el área de comportamiento que explora, **de aptitud,** pues se refiere a la habilidad cuantitativa como un rasgo psicológico de naturaleza cognitiva.
5. Por su modalidad de aplicación, (a) **de papel y lápiz**, su formato de administración se basa en la lectura de los planteamientos en un folleto de prueba y se registran las respuestas en una hoja de respuestas. (b) **colectiva,** dado que su aplicación se da o puede darse en grupo (no requiere de monitoreo individual).

6. Por lo que demanda al sujeto, **de dificultad** o **potencia**: aunque se controla el tiempo de ejecución, este no es el aspecto crítico para los resultados, es decir, las diferencias individuales están explicadas por el número de respuestas correctas e incorrectas dadas a la prueba (calidad de la ejecución), más que por la velocidad de la ejecución.
7. Es una prueba que, aunque está orientada a medir una variable cognitiva que no depende estrictamente de la cultura, los planteamientos son hechos a través de códigos lingüísticos, que la hacen **cargada del elemento cultural, asociado con el idioma castellano.**
8. Por los sujetos a los que está dirigido puede clasificarse como **para sujetos nivel de instrucción profesional**, pues es éste un requisito indispensable para ingresar a cursos de postgrado y, partiendo del nivel de dominio del área cuantitativa que deben tener sujetos profesionales, se construyó la prueba.
9. Por el modelo estadístico en que se basa, la **Teoría Clásica de los Tests.**

La prueba de RBC es diseñada específicamente para la selección de aspirantes a postgrado y en este sentido, para la predicción del rendimiento académico en postgrado. Es por ello que su análisis psicométrico exige la consideración de la medida de criterio que pretende predecir: Rendimiento Académico en Postgrado.

C.- Rendimiento académico en postgrado

Borrego (1985) define el rendimiento académico como el logro del aprendizaje obtenido por el estudiante a través de las diferentes actividades desarrolladas a la luz de los objetivos planteados. Carballo (1985), lo definió como la calidad de la actuación del alumno con respecto a un conjunto de conocimientos, habilidades o destrezas en una asignatura determinada como resultado de un proceso de instrucción sistémico. Por su parte, Páez (1987)

señala que el rendimiento académico es el grado en que el estudiante ha alcanzado los objetivos propuestos y las condiciones bajo las cuales se produjo ese logro.

El Reglamento General de la Ley Orgánica de Educación (1986) lo define como el nivel alcanzado por los alumnos en función de los objetivos programáticos previstos, que puede ser medido mediante actividades de evaluación.

Lexus (1997; cp. Parra, 1993 p. 12) señala que la medida del rendimiento académico es explicada no sólo por los conocimientos que adquiere el estudiante en su proceso de formación sino, además, por el nivel intelectual del estudiante, las capacidades, habilidades, destrezas, motivaciones, así como por condiciones sociales y culturales en las que se desenvuelve.

En un sentido amplio (Zubizarreta, 1969; cp. Parra, 1993, p. 12) describe el rendimiento académico como la relación entre el nivel de ejecución logrado por el estudiante y el esfuerzo empleado para obtenerlo. A partir de allí se establecen los conocimientos, habilidades y/o destrezas adquiridas y como consecuencia, el éxito o no de la educación en un periodo determinado.

Según Parra (1993), el rendimiento académico es el resultado obtenido o nivel de ejecución manifiesto en relación con el nivel de ejecución esperado.

García (1985a) señala que el rendimiento es ejecución o actuación. Durante sus estudios un estudiante adquiere información, desarrolla destrezas y una forma determinada de percibir el mundo propio de su área de especialización, lo que se manifiesta en su ejecución académica, y que representa su aprendizaje. En este sentido, aprendizaje y rendimiento son lo mismo cuando la medida del rendimiento representa en forma válida al aprendizaje. El elemento esencial es entonces, obtener una media que

represente en forma confiable y válida el desempeño académico de los estudiantes.

El rendimiento académico, es una medida de la ejecución de los estudiantes que está directamente relacionada con el nivel de aprovechamiento y de conocimiento alcanzado, lo que necesariamente está ligado al sistema de evaluación que se utilice para llegar a la medida o indicador de la ejecución (Kuncel, Hezlett y Ones, 2001).

Según Bloom, Hastings y Madaus (1975), en todos los niveles del sistema educativo, la evaluación presenta múltiples problemas y, al mismo tiempo, pocas alternativas de solución.

A lo largo del tiempo se han creado diferentes formas y sistemas para detectar el grado de aprovechamiento o rendimiento académico. En el siglo XX comenzó la tarea de medir y evaluar el aprendizaje de acuerdo con pautas sistemáticas, con respaldo científico y el empleo de una variedad de técnicas. Aunque no hay una forma única de aproximación cuantitativa al rendimiento académico, existen dos tipos de evaluación de uso frecuente que se distinguen básicamente por el objetivo que persiguen. Estas son denominadas por Bloom, et al. (1975): evaluación formativa y evaluación sumativa.

La evaluación formativa tiene el propósito de identificar las fallas en el rendimiento para orientar, tanto a estudiantes como a docentes, en las áreas de formación que requieren ser reforzadas; mientras que la evaluación sumativa se orienta hacia la estimación, mucho más general, del grado en que los resultados más amplios han sido alcanzados a lo largo del curso, el rendimiento del estudiante. Desde la Educación Pre-escolar hasta Educación superior predomina la evaluación sumativa, la cual está orientada a clasificar a los estudiantes en términos de sus niveles de rendimiento académico en las materias cursadas (Bloom, et al.,1975).

Una adecuada medida de la variable rendimiento académico en general y en cursos de postgrado en particular (variable dependiente), debe considerar la mayor cantidad de información relevante sobre la ejecución de los estudiantes (validez), en condiciones estándar y bajo la consideración y control de las diferentes fuentes de varianza de error que afectan la evaluación (confiabilidad). En este sentido, las estimaciones de rendimiento académico como medida global (sumativa), deben basarse en indicadores parciales (sumativos) también válidos y confiables.

La medida del rendimiento académico en postgrado como variable dependiente debe obtenerse de manera directa e independiente; debe, además, ser confiable, estar libre de sesgo y distorsiones, y ser factible en términos de medición (Martínez, 1995).

Las variables más utilizadas como representativas del rendimiento académico en postgrado son las mencionadas por Kuncel, Hezlett y Ones (2001) en su meta-análisis sobre la validez predictiva del Graduate Record Examination:

1. Promedio de calificaciones total (incluye todas las materias que forman el pensum del programa de postgrado);
2. Promedio de calificaciones en las materias cursadas durante el primer término académico del programa;
3. Número de publicaciones y conferencias;
4. Tiempo durante el cual las publicaciones son citadas;
5. El tiempo en el que se aprobó el programa.

Cada criterio contiene diferentes aspectos del constructo que se pretende evaluar, es decir, en cada criterio se consideran aspectos distintos de la ejecución o del rendimiento académico. En la medida en que los aspectos

de la ejecución considerados por cada criterio dependen de variables más asociadas con la motivación y/o con la persistencia que con las aptitudes o variables cognitivas, el valor predictivo de las variables objeto de estudio será menor. Consecuentemente, cada criterio debe relacionarse en forma particular con las variables predictoras.

De acuerdo con Kuncel, et al. (2001), las ventajas y desventajas de las medidas de criterio mencionadas son las siguientes:

Las medidas más utilizadas como indicadores de ejecución académica en el postgrado son el promedio de calificaciones en el primer término académico o en el programa completo. El promedio de las calificaciones en todo el programa tienen la ventaja de que evalúan un amplio período de trabajo, la adquisición de conocimiento, esfuerzo, persistencia y habilidad. Sin embargo, en tanto que la medida de criterio se aleja del predictor, pueden intervenir infinidad de variables que determinen la ejecución académica, más allá de las habilidades básicas y nivel de formación académica; en este sentido, es más recomendable el promedio de calificaciones en el primer término académico. El promedio de calificaciones como indicador del rendimiento académico puede estar afectado por diferencias en los esquemas de evaluación entre áreas, departamentos y profesores del mismo curso.

Los criterios asociados con el número de publicaciones y tiempo en el que son citadas, tienen que ver con la productividad científica, lo cual suele ser una meta que guía los programas de investigación, pero no todos los programas tienen orientación científica, por lo que este criterio no es pertinente para la mayoría de los estudios que persiguen la enseñanza profesional aplicada.

Al igual que el promedio del total de calificaciones, el nivel de logro es una medida distante y, por lo tanto, está determinado por múltiples variables,

además de las habilidades y el nivel de formación académica de entrada representado por el promedio en el pregrado.

En la presente investigación se considerará el promedio de calificaciones de las materias cursadas durante el primer término académico (primer semestre o trimestre), producto de evaluaciones sumativas, como estimador del rendimiento académico.

D.- Evidencia empírica sobre el comportamiento de pruebas de habilidad cuantitativa en la predicción del rendimiento académico a nivel de postgrado.

La revisión de la literatura sobre pruebas con características similares diseñada para el mismo uso, es decir predicción de rendimiento académico a nivel de postgrado en el contexto norteamericano es el Graduate Record Examination (GRE) (Newman, 1968). El GRE publicado por el *Educational and Testing Service* (ETS) está formado por un grupo de pruebas estandarizadas de habilidades generales: habilidad verbal (GRE-V), habilidad cuantitativa (GRE-Q) y habilidad analítica (GRE-A). Donde el GRE-Q evalúa las mismas dimensiones que la prueba objeto de estudio (RBC).

Norcross, Hanych y Terranova (1996) señalan que los puntajes del GRE son utilizados en los procesos de selección por el 93% de los programas doctorales y por el 81% de los programas de maestría en los Estados Unidos de América, por lo que ha sido objeto de múltiples investigaciones. Tales investigaciones presentan resultados contradictorios en la predicción del rendimiento académico. Se han encontrado coeficientes que van desde -0.62 hasta 0.81 (Kuncel, Hezlett y Ones, 2001). Dicha variabilidad puede estar asociada con: (a) la medida de criterio seleccionada (Kuncel, et al., 2001); (b) la confiabilidad y validez de la medida de criterio (Willingham, 1974); (c) la restricción en el rango de variabilidad del criterio (Rothstein, Paunonen, Rush, y King, 1994; Kuncel, et al., 2001); (d) tamaños muestrales pequeños (Nunnally y

Bernstein, 1995); (e) la posible validez diferencial de las pruebas que forman el GRE por tipo de disciplina (Kuncel, et al., 2001), y (f) el uso de modelos de selección compensatorios que suprimen el efecto de una variable en la predicción, por ser compensada por otra (Magnusson, 1972; Rothstein, et al., 1994).

Dada la similitud del GRE con las características de la prueba objeto de esta investigación y la poca disponibilidad de investigaciones pertinentes en Venezuela, se recurre a las investigaciones disponibles sobre la validez predictiva sobre el GRE-Q en la predicción del rendimiento académico en postgrado.

Willingham (1974) realizó un meta-análisis en el que identificó 43 estudios en el periodo de 1952 a 1972. Los tamaños de las muestras utilizadas en los estudios variaban en un rango que iba de 20 a 1479 estudiantes. El número de estudiantes incluidos en los estudios fue de 21.214. Se utilizaron 616 coeficientes de validez. Este autor analizó el valor predictivo del GRE-Verbal (GRE-V), del GRE-Cuantitativo (GRE-Q) y del promedio de calificaciones en el pregrado (UGPA) para cinco criterios de rendimiento académico: (a) promedio en las materias (GGPA), (b) la evaluación de la facultad, (c) la evaluación de los departamentos, (d) el logro de Ph.D., y (e) el tiempo para obtener el Ph.D., en una muestra general y por disciplina. A pesar de que los estudios incluidos variaban en calidad y alcance y que algunos estaban basados en muestras pequeñas y, por lo tanto, producían correlaciones poco confiables, los resultados ofrecen una idea general sobre el valor predictivo de las variables seleccionadas. De acuerdo con los resultados obtenidos por Willingham (1974), la validez del GRE-Q depende de las disciplinas de estudio: el GRE-Q tiene mayor validez para predecir el rendimiento académico en aquellas disciplinas con mayor orientación cuantitativa. En cuanto a su valor predictivo sobre el

criterio promedio de calificaciones del primer año, se encontró una correlación de 0.23.

En el año 1992, Goldberg y Alliger, realizaron un meta-análisis sobre el valor predictivo del GRE-V y el GRE-Q sobre el rendimiento académico en estudiantes de postgrados de Psicología, a través de tres medidas de criterio: (a) promedio de calificaciones en el primer término académico, (b) puntaje en los exámenes de comprensión, y (c) grado dentro del curso específico. En el meta-análisis se incluyeron 27 estudios, después de revisar la literatura existente desde 1950-1990. El coeficientes de validez del GRE-Q para el criterio denominado "promedio de calificaciones en el primer término académico" fue de 0.15.

Morrison y Morrison (1995) realizaron un meta-análisis sobre la validez predictiva del GRE-V y el GRE-Q para predecir rendimiento académico en estudiantes de postgrado medido a través del promedio de calificaciones del primer término académico, en una muestra de 22 estudios publicados entre 1955 y 1992. Los resultados indicaron un coeficiente de 0.22 para el GRE-Q.

Un meta-análisis más reciente sobre la validez del GRE (GRE-V, GRE-Q) y del promedio de calificaciones en el pregrado como predictores de la ejecución en los estudios de postgrado fue realizado por Kuncel, et al. (2001), en el que se trabajó con 1753 muestras independientes, 6.589 correlaciones a través de 82.659 estudiantes de postgrado. En este estudio se consideraron ocho medidas de criterio del rendimiento académico, en las que se corrigió estadísticamente el efecto de la restricción del rango. En general, los coeficientes obtenidos fueron de signo positivo y magnitud moderada (alrededor de 0.30). Específicamente, el valor predictivo del GRE-Q para el criterio de rendimiento académico definido por el promedio de calificaciones en el primer término académico, que es de interés para la presente

investigación, fue de 0.38 (una vez se corrige el efecto de la restricción del rango).

Kuncel, et al. (2001), señalan como fuente de inconsistencia entre los resultados obtenidos por diferentes estudios de validez del GRE, por un lado, la restricción del rango en la medida de criterio, lo cual, al reducir la varianza, subestima el coeficiente de validez, y por otra parte señala las dificultades asociadas con la elección del criterio adecuado, así como su obtención en forma técnicamente confiable y válida.

Stack y Kelley (2002) por su parte, realizaron un estudio sobre la validez del GRE-V, GRE-Q y UGPA en la predicción del rendimiento académico en un master de Justicia Criminal. Utilizaron tres indicadores de rendimiento académico: (a) promedio de calificaciones de los que habían aprobado como mínimo seis créditos (N=70); (b) Grado con que aprobaron dos materias: ciencias políticas (N=36), Seminario de Justicia Criminal (N=47); y (c) el tiempo en que completaron el master (N=27). Los resultados para el criterio promedio de calificaciones presentaron una correlación moderada y significativa con el GRE-Q: r = 0.372. Estos resultados fueron combinados en una Regresión Múltiple y se encontró que las tres variables predictoras explicaban el 28% de la varianza en el promedio de calificaciones. Estos autores señalan, además, que las pruebas tienden a presentar valor predictivo diferencial dependiendo del tipo de disciplina a la que pertenece el postgrado. De acuerdo con esto último, el valor predictivo del test está asociado con la orientación (carga verbal o cuantitativa implicada) de la disciplina de estudio a la que se hace la predicción. Esta aseveración es también propuesta por House y Johnson, (1993).

Otra variable, cuya relación con la habilidad cuantitativa en general y con la matemática en particular ha sido muy estudiada, es el sexo, y los resultados, son inconsistentes y controversiales. Kianian (1996), en su revisión de la literatura identifica los siguientes resultados:

- En la mayoría de los estudios sobre la relación entre género y aprendizaje de matemáticas no aparece o es mínimo hasta que comienzan la secundaria y cuando aparece **es casi siempre a favor de los hombres** (Fennema, 1974; Fennema y Sherman, 1978; Sherman 1980; Fennema y Carpenter, 1981; Armstrong, 1981; Bagder, 1981; Kaminski, 1982; Keys, 1987, Gordon et. Al , 1990, Kaur, 1992).

- Algunos estudios más recientes han cambiado esta conclusión general, por ejemplo: Galbraith (1986) estudió 334 australianos estudiantes de bachillerato (150 hombres y 177 mujeres) y encontró que la ejecución de las **mujeres supera a la de los hombres.** A la misma conclusión llegaron Stockard y Wood (1984) en estudiantes de bachillerato en los Estados Unidos y Tressou-Milonas (1992) en estudiantes de primaria en Grecia.

- Con respecto al logro en los diferentes aspectos de la Matemática, la Hyde (1990 cp. Kianian, 1996), señala que **las mujeres tienen mejor ejecución que los hombres en cálculo y los hombres tienen mejor ejecución en solución de problemas.**

- Kianian (1996) reporta algunas investigaciones que examinan los factores que contribuyen a las diferencias en matemática.

  - Badger (1981) concluye que son factores sociales, no genéticos, los responsables por los casos en los cuales las niñas tienen ejecución relativamente pobre en matemática.

  - Fennema (1985) argumenta que las condiciones sociales influyen el medio ambiente educativo y las creencias personales sobre las diferencias en matemática por género. Fennema (1990) sugiere además que la modificación de los factores ambientales influyen en las diferencias de logro y participación en matemática por género. Si se les da a niños y niñas iguales oportunidades de aprendizaje de las

matemáticas e igual tratamiento educativo, los resultados en la ejecución por género serán similares.

- Leder (1990) concluye que las diferencias por género en el aprendizaje de matemática es un tema complejo; y menciona las variables que han sido más estudiadas en este sentido: (a) el trato diferencial de la escuela a varones y hembras; (b) las expectativas diferenciales de los padres; (c) expectativa de los pares; y (d) creencias y valores que señalan que la matemática no es un habilidad femenina sino masculina. Leder (1990) argumenta que el examen de los factores biológicos no es constructivo porque estos factores no son modificables.

Hanna, (1990, cp. Kianian, 1996) basada en los datos del Segundo Estudio Internacional de Matemática (SIMS) para estudiantes masculinos y femeninos de secundaria de 15 países, concluyó que el logro en matemática es significativamente menor en las mujeres que en los hombres, excepto en países como Tailandia e Inglaterra, en los cuales los datos muestran alto nivel de soporte en el hogar para ambos géneros. Ellos encontraron que el fuerte soporte del hogar para la participación en matemática para niños y niñas es la única variable que se relaciona con las diferencias de logro en matemática por género. La variación de las diferencias por género de país en país indica que las diferencias no se deben a un factor biológico. Por tanto, la teoría biológica no explica la superioridad de los varones en matemática.

Kianian (1996), concluye de su investigación en la que examina la ejecución en matemática de 479 estudiantes universitarios, 219 hombres y 260 mujeres, que el ambiente explica las diferencias en la ejecución en matemáticas.

En general, las investigaciones señalan que las pruebas estandarizadas de habilidad numérica se comportan como predictores válidos del rendimiento académico en estudios de postgrado estimado a través del promedio de

calificaciones en el primer término académico, con magnitudes moderadas (0.20-0.30). Por otra parte, los coeficientes de correlación obtenidos varían con el tipo de disciplina en el postgrado, en la medida que la disciplina tenga mayor carga cuantitativa, el valor predictivo de la prueba aumenta. Finalmente se evidencian diferencias por sexo en la ejecución en la prueba.

MÉTODO

Objetivo

Determinar las características psicométricas, partiendo de la Teoría Clásica de los Tests, de la prueba de Razonamiento Básico Cuantitativo del modelo PG-38, utilizado por la Universidad en el proceso de selección de aspirantes a postgrado de junio del año 2003.

Tipo de Investigación

De acuerdo con Kerlinger y Lee (2002), la investigación científica social puede clasificarse en cuatro grandes tipos, dependiendo de la posibilidad o no de manipular las variables (experimentales y no experimentales) y del contexto de aplicación (campo o laboratorio). Utilizando estos criterios de clasificación, se trata de una investigación científica no experimental, de campo.

> La investigación no experimental es *la búsqueda empírica y sistemática, en la que el científico no posee el control directo de las variables independientes, debido a que sus manifestaciones ya han ocurrido o a que son inherentemente no manipulables. Se hacen inferencias sobre las relaciones entre las variables, sin intervención directa, de la variación concomitante de las variables independientes y dependiente* (Kerlinger y Lee 2002, p.504).

En este caso, no es posible manipular las variables ni seleccionar aleatoriamente a los sujetos, dada la naturaleza de las variables. Los sujetos se autoseleccionan o autoasignan al grupo de estudio, que estará conformado por todos los aspirantes seleccionados en julio del año 2003, para cursar

estudios de postgrado en La Universidad, que cursen el primer período académico.

De acuerdo con su objetivo, se puede caracterizar el presente estudio perteneciente al tipo de Investigación Descriptiva, pues tiene como objetivo la recolección de datos para la caracterización del comportamiento estadístico prueba de habilidad cuantitativa (RBC) en una población específica (Van Dalen y Meyer, 1990).

Finalmente, de acuerdo con la metodología de análisis utilizada, se trata de una investigación de tipo psicométrica, dado que se fundamenta en uno de los componentes de la psicometría: Teoría Clásica de los Tests (Muñiz, 2001).

Diseño Muestral

La población está conformada por todos los aspirantes a cursar estudios de postgrado en la Universidad, seleccionados e inscritos en junio de 2003, que cursen dentro del postgrado para el cual fueron seleccionados el primer término (un semestre o un trimestre).

Dado que se tiene acceso a la población, no es necesario recurrir a una muestra para estimar el comportamiento de la población, sino que se obtienen directamente los parámetros que describen la población.

Tal como se muestra en la Tablas 3, la población está formada por los 302 aspirantes que presentaron la batería de selección para postgrado en La Universidad en julio del año 2003. La partición de la distribución por sexo es equivalente. En cuanto a la edad, el rango de la población va de 23 y 59 años, lo que indica un recorrido de la variable de 36 años.

Tabla 3. Sexo

	Frecuencia	Porcentaje
Masculino	127	42,1
Femenino	175	57,9
Total	302	100,0

La distribución de edad es asimétrica positiva (Tabla 4, Gráfico 1): hay mayor proporción de aspirantes con edades entre 25 y 28 años y la proporción va disminuyendo hacia las edades superiores.

Tabla 4. *Edad*

Media	33,54
Media para la distribución truncada al 5%	32,97
Mediana	31,00
Modo	25,00
Varianza	66,408
Desviación Estándar	8,149
Límite inferior del recorrido de la variable	23
Límite superior del recorrido de la variable	59
Rango	36
Asimetría	0,916
Kurtosis	0,068

Gráfico 1. **Edad**

```
 2,00    2 . 3
40,00    2 . 4444455555555555555
51,00    2 . 666666666667777777777777
35,00    2 . 88888888889999999
29,00    3 . 00000001111111
20,00    3 . 2222223333
19,00    3 . 444444555
18,00    3 . 666777777
18,00    3 . 888899999
15,00    4 . 0000111
15,00    4 . 2223333
 8,00    4 . 455
 8,00    4 . 667
10,00    4 . 8889
 6,00    5 . 111
 2,00    5 . 2
 1,00    5 . &
 1,00    5 . &
 4,00 Extremos    (>=57)
```

En cuanto al tipo de profesión de los aspirantes (Tabla 5), aproximadamente la mitad proviene de Ciencias Sociales (43%), seguida por el área de Ingeniería, arquitectura y tecnología (23%) y luego por la de Ciencias de la salud (17%). En las cuatro áreas restantes, el número de aspirantes es bajo.

Tabla 5. ***Tipo de Profesión***

	Frecuencia	Porcentaje
No Declarado	4	1,30
Ciencias Básicas	1	,30
Ingeniería, Arquitectura y Tecnología.	70	23,20
Ciencias del Agro y del Mar	8	2,60
Ciencias de la Salud	53	17,50
Ciencias de la Educación	24	7,90
Ciencias Sociales	132	43,70
Humanidades, Letras y Arte	7	2,30
Ciencias y Artes Militares	3	1,0
Total	302	100,0

Procedimiento

1. Edición, aplicación y corrección de la prueba de Razonamiento Básico Cuantitativo.

La prueba de Razonamiento Básico Cuantitativo diseñada por PROINVESCA (2002), siguiendo las especificaciones antes descritas, se incluyó, junto con las pruebas de habilidad verbal y análisis situacional en folletos numerados, sellados e identificados con el código PG-38. Dentro del folleto se encuentra, en primer lugar, la prueba de habilidad verbal, luego la de habilidad cuantitativa y finalmente la de análisis situacional; el folleto contiene, además, una hoja de respuestas para lector óptico cuya numeración coincide con la numeración del folleto y es donde los evaluados registran sus respuestas. Cada prueba tiene un tiempo prestablecido para su resolución, durante el cual los evaluados no pueden revisar las otras pruebas. El tiempo para la prueba de RBC es de 60 minutos.

Una vez impresa la Batería PG-38, numeradas y selladas, fue empaquetada de cinco en cinco en orden consecutivo, en bolsas plásticas, con la finalidad de facilitar el conteo de las pruebas antes y después de la aplicación y así garantizar el resguardo de la información.

La aplicación de la Batería se realizó bajo un protocolo estándar cuyo seguimiento está orientado a garantizar la uniformidad del procedimiento, lo que minimiza el efecto de las fuentes de varianza de error debidas a la aplicación de la prueba.

Las pruebas fueron aplicadas en forma masiva, en aulas equipadas con escritorios o pupitres, con adecuada iluminación y ventilación. Los aspectos considerados en el protocolo de aplicación se describen a continuación.

Los aplicadores fueron entrenados en el manejo de la situación de evaluación a través del seguimiento de un manual o instructivo de aplicación que contiene, tanto los pasos y controles que debe considerar el aplicador, como las palabras y oraciones que se deben leer al grupo evaluado durante el proceso. Durante el entrenamiento se enfatizó la importancia de garantizar condiciones uniformes en el proceso de aplicación de pruebas.

A los evaluados se les entregó un folleto y se les leyeron las instrucciones que deben seguir para responder cada parte de la batería, respetando los tiempos preestablecidos. El aplicador tenía la responsabilidad de garantizar el ajuste de la ejecución a los tiempos preestablecidos.

Durante el proceso de aplicación no se permitió el uso de reglas, diccionarios, calculadoras o celulares en el aula. Al finalizar el tiempo de aplicación, se recogieron tanto los folletos de prueba como las hojas de respuestas, los cuales fueron contados antes de permitir la salida del aula de los evaluados.

Posteriormente, las hojas de respuestas fueron leídas y procesadas electrónicamente y tanto los resultados procesados como la base de datos fueron sometidos a un proceso de validación a través de la verificación manual de una muestra aleatoria del 10% de las hojas de respuestas. Dado que se trata de pruebas de opción múltiple y selección simple, los resultados directos de las pruebas (sumatoria de respuestas correctas) se corrigieron a través de la aplicación de un factor de corrección por azar para minimizar el efecto de la fuente de varianza de error debida a la adivinación (tres respuestas incorrectas eliminan una respuesta correcta).

2. **Recolección de los datos de la variable de criterio: rendimiento académico en el postgrado:** La Universidad suministró las notas definitivas de los

estudiantes en las diferentes materias asociadas con los programas de formación de postgrado, a partir de las cuales se obtuvieron los promedios, que funcionarán como indicadores del rendimiento académico en postgrado. Las materias que formarán parte de la variable rendimiento académico en postgrado serán aquellas cursadas por los estudiantes durante el primer término académico del programa que cursan (primer semestre o trimestre).

3. **Estimación de la Confiabilidad:** Para estimar la confiabilidad del instrumento se utilizaron dos de los métodos que requieren de una sola aplicación del instrumento: (a) Coeficiente alpha de Cronbach, el cual controla fundamentalmente la fuente de varianza de error debida a la falta de correlación entre los ítems; por lo tanto, su magnitud se asocia en forma directa con la consistencia interna de la prueba (Anastasi y Urbina, 1998; Magnusson, 1972). Aunque, el hecho de que los *items* tengan altas correlaciones entre sí, no es suficiente para afirmar que es unidimensional, el obtener un coeficiente alpha de Cronbach alto (alrededor de 0,80) apoya la noción de unidimensionalidad (Pedhazur y Schmelkin cp. Nelson, 2001). El coeficiente alpha de Cronbach fue obtenido a través del programa Lertap, versión 5.4; y (b) Coeficiente de División por mitades (par – impar) que controla la fuente de varianza de error debida a la falta de equivalencia entre las mitades, por lo que la magnitud del coeficiente, al ser corregida por la fórmula de Spearman Brown, es una estimación del coeficente de confiabilidad que se obtendría para la prueba completa (Magnusson, 1972). Este coeficiente se obtuvo a través del SPSS 12.0. Para el cálculo del coeficiente de División por Mitades se dividió la prueba, utilizando como criterio de clasificación el número del *item* par o impar, cada mitad o parte quedó conformada por 20 *items*. La parte 1 estaba formada por los *items* impares y la parte dos por los pares. Finalmente, se analizó el aporte de

cada *item* a la confiabilidad del instrumento, mediante el cálculo, con el SPSS 12.0, del alpha de Cronbach para la prueba cuando cada *item* es eliminado.

4. **Estimación de la Validez**: Para verificar el grado en que los *items* que forman el test son representativos del dominio que se pretende medir (**validez de contenido**), se revisaron y analizaron: (a) el registro evolutivo donde se registran las diferentes etapas de construcción del test; y (b) los distintos niveles de validación que realizaron jueces expertos y se revisó la congruencia entre el *item* definido en la tabla de especificaciones y su diseño en la prueba.

Se obtuvieron indicadores de la **validez de criterio** a través de la correlación simple (Pearson) entre la prueba de Razonamiento Básico Cuantitativo y el Rendimiento Académico en Postgrado. Esta última variable está definida como el promedio de calificaciones de los aspirantes seleccionados en las materias cursadas durante el primer término académico. De acuerdo con estudios previos, se esperan coeficientes alrededor de 0.20 (sin corregir el efecto de la restricción del rango) (Golberg y Alliger, 1992; Morrison y Morrison, 1995).

En cuanto a la **validez de constructo**, además de las consideraciones sobre la validez de contenido y de criterio, se plantearon dos hipótesis vinculadas con lo aportado por las investigaciones revisadas. A continuación se presentan las hipótesis:

a. **Hipótesis 1:** se refiere a la relación de la variable sexo con el nivel de ejecución en habilidad cuantitativa que subyace a la ejecución en matemática. A pesar de la variedad e inconsistencia de los resultados obtenidos sobre la dirección de estas diferencias, la mayoría de las

investigaciones consultadas señalan que existen diferencias en la ejecución en matemática adjudicadas al sexo (independientemente de la dirección y de las razones que las explican. La hipótesis concreta sería: **Se esperan diferencias estadísticamente significativas por sexo en la ejecución en la prueba de Razonamiento Básico Cuantitativo.** Para probar esto se realizó un análisis de varianza de una vía.

b. **Hipótesis 2**: tiene que ver con el tipo de profesión, dado que las investigaciones revisadas señalan diferencias en la ejecución y en el valor predictivo de las pruebas de habilidad cuantitativa, dependiendo de la disciplina a la que se opta (lo que se relaciona directamente con el tipo profesión); y dadas las claras diferencias en la carga cuantitativa asociada al tipo de profesión, lo que supone exposición diferencial a la práctica y por ende al desarrollo de la habilidad cuantitativa. **Se espera obtener diferencias en la ejecución en la prueba por tipo de profesión. Específicamente se espera mayor ejecución en la categoría Ciencias Básicas y en la categoría Ingeniería, Arquitectura y Tecnología que en el resto de las profesiones.** Para probar esta hipótesis se analiza, por un lado, la ejecución media de los grupos y para determinar si las diferencias observadas son estadísticamente significativas, se seleccionan dos grupos que deben contrastar en el rasgo, dadas sus diferencias en la carga cuantitativa de su *pensa* y que tienen un tamaño que permite el cálculo del análisis de diferencias a través de una prueba T para muestras independientes. Los grupos que cumplen con esas dos condiciones son: Ingeniería, Arquitectura y Tecnología y la de Ciencias Sociales.

5. **Análisis de *Items*:** El análisis de los *items* se realiza desde tres perspectivas complementarias: cuantitativa, gráfica y cualitativa.

En primer lugar, para el **análisis cuantitativo** se procesaron los resultados a través del Lertap 5,4, y se obtuvo un reporte de indicadores cuantitativos que indica el comportamiento estadístico de cada *itèm*, especificando para cada opción de respuesta:

- El peso (wt), uno para la respuesta correcta y –0,33 para las respuestas incorrectas (factor de corrección por azar).
- El número de personas que seleccionaron cada opción (n).
- La proporción de sujetos que seleccionó cada opción de respuesta (p).
- El coeficiente de correlación punto-biserial entre cada opción y el puntaje total en el test (eliminando el *item* analizado).
- El coeficiente de correlación biserial entre cada opción y el test total (eliminando el *item* analizado.
- La media de los sujetos que seleccionan cada opción de respuesta.
- La posición tipificada de la media de los que eligen cada alternativa en relación a la media total.

Para la clasificación de los *items* de acuerdo con los datos presentados en el reporte de indicadores cuantitativos, se utilizaron: (a) los criterios de análisis propuestos por Silva, De Sanctis y Sarco (1982) y OPSU-CENAMEC (1985); y (b) las magnitudes esperadas de acuerdo con Hills (1976, cp. Nelson, 2001, p.115) y las mencionadas por Nunnally y Berstein (1995) para muestras grandes:

De acuerdo con esto, se realizó una primera clasificación de los *items* en **aceptados** o **en revisión**

En la categoría de **aceptados** se incluyen aquellos *items* cuyo coeficiente punto-biserial de la alternativa correcta sea positivo e igual o mayor a 0,20; y el de todos los distractores es negativo; los que no cumplen con estas dos condiciones son incluidos en la categoría **en revisión.**

Luego de esta primera clasificación, se realizó un **análisis gráfico y un análisis cualitativo** de los *items* clasificados dentro de la categoría **en revisión**.

Los gráficos utilizados representan, para cada *item*, la distribución (proporción) de las opciones de respuesta en los grupos superior (20% superior) e inferior (20% inferior). En estos gráficos se identifican la respuesta correcta (Key), la proporción de sujetos que lo responde correctamente (Diff.) y la correlación punto-biserial de la respuesta correcta con el puntaje total en el test (Disc). El comportamiento adecuado de un *item* se relaciona con una pendiente ascendente del grupo inferior al superior en la proporción de selección de la respuesta correcta; una pendiente descendente del grupo inferior al superior de los distractores y aproximadamente la misma proporción de selección de los distractores.

El **análisis cualitativo** consistió en la revisión de los aspectos formales del diseño del planteamiento y de las opciones de respuesta. Para ello se analizó: (a) la construcción gramatical del *item*; (b) uso de términos o notaciones poco usuales o confusas; y (c) equilibrio entre longitud y claridad de la información.

Partiendo de éstos criterios, los *items* en revisión son nuevamente clasificados en: (a) **aceptados sin modificación;** (b) **aceptados con modificación;** y (c) **descartados**, dependiendo de si su estructura acepta modificaciones orientadas a su ajuste a los criterios cuantitativos y cualitativos de la Teoría de los Tests .

ANÁLISIS E INTEGRACIÓN DE RESULTADOS

Para el análisis e integración de los resultados se presenta en primer lugar el análisis descriptivo de las variables y luego la caracterización del instrumento en términos de confiabilidad, validez y comportamiento de los *items*.

Análisis descriptivo de las variables

Se analizó el comportamiento estadístico, tanto de la distribución de puntajes obtenidos en el test de Razonamiento Básico Cuantitativo como el de la variable de criterio: rendimiento académico en el postgrado. Para obtener los estadísticos descriptivos de las variables se utilizó el programa SPSS versión 12.0. Se obtuvieron: (a) medidas de tendencia central: media, mediana, modo y media truncada al 5%; (b) medidas de dispersión: desviación estándar, varianza, coeficiente de variación, límites del recorrido de la variable y rango; y (c) medidas de forma: Asimetría y Kurtosis. Además, se presenta gráficamente la distribución de cada variable a través de gráficos de tallo y hoja.

Habilidad Cuantitativa

Las características de la distribución de habilidad cuantitativa obtenida a partir de la prueba RBC, se presentan en la Tabla 6 y en el Gráfico 2. Se trata de una distribución con un recorrido de 37 puntos, con valores negativos producto de la aplicación del factor de corrección por azar, lo que podría sobrestimar los estadísticos de dispersión. Es una distribución con una ligera asimetría positiva. Las medidas de tendencia central se encuentran por debajo de la mitad del recorrido posible, lo que indica que la prueba resultó ligeramente difícil para el grupo.

Tabla 6. *Habilidad Cuantitativa*

Media	13,62
Media para la distribución truncada al 5%	13,57
Mediana	13,67
Modo	16,00
Varianza	44,668
Desviación estándar	6,683
Coeficiente de Variación	49%
Límite inferior del recorrido de la variable	-4
Límite superior del recorrido de la variable	33
Rango	37
Asimetría	0,105
Kurtosis	0,004

***Gráfico* 2. Habilidad Cuantitativa**

```
 6,00    -0 . 11&
22,00     0 . 0012333444
57,00     0 . 55566666777777888888889999
86,00     1 . 00000000111111122222222233333333444444444
80,00     1 . 55555555555666666666667777777888889999
33,00     2 . 000011111222334
16,00     2 . 5567888&
 1,00     3 . &
 1,00 Extremos   (>=33)
```

Rendimiento Académico en Postgrado

En la Tabla 7, y en el Gráfico 3, puede observarse el comportamiento de la variable dependiente. El recorrido teórico es de 0 a 20 puntos, pero se obtiene un rango de puntajes de 12, con un límite inferior 8 y el superior de 20 puntos. Los puntajes se acumulan en forma homogénea hacia la parte alta de la distribución.

Tabla 7. ***Rendimiento Académico en Postgrado***

Media	16,90
Media para la distribución truncada al 5%	17,07
Mediana	17,29
Modo	18
Varianza	4,803
Desviación estándar	2,192
Coeficiente de Variación	13%
Límite inferior del recorrido de la variable	8
Límite superior del recorrido de la variable	20
Rango	12
Asimetría	-1,248
Kurtosis	1,829

***Gráfico* 3. Rendimiento Académico en Postgrado**

```
13,00 Extremos    (=<12,2)
  ,00     12 .
 5,00     12 . 55555
 6,00     13 . 000000
 3,00     13 . 555
11,00     14 . 00000000023
 4,00     14 . 5555
11,00     15 . 00000000003
13,00     15 . 5555555555567
30,00     16 . 000000000000000000000000000233
23,00     16 . 55555555555555555566667
33,00     17 . 0000000000000000000000000000223
22,00     17 . 5555555555555555556666
46,00     18 . 00000000000000000000000000000000000000002333334
20,00     18 . 55555555555555555556
40,00     19 . 0000000000000000000000000000000000000033
11,00     19 . 55555555555
11,00     20 . 00000000000
```

Confiabilidad

El coeficiente alpha de Conbrach obtenido por la prueba es de 0.7853, lo que indica que la prueba es confiable en términos de consistencia interna.

Al realizar la división por mitades, el coeficiente alpha de la parte 1 es de 0.675 y el de la parte 2 es de 0.604. La correlación entre las mitades es de 0.661 y el coeficiente de confiabilidad estimado por el método de Spearman – Brown es de 0,796. Como se observa en la Tabla 8, las medias y desviaciones de ambas mitades son muy similares. La parte 1, formada por los *items* impares tiene media y desviación ligeramente superior a los de la Parte 2; además la Parte 1 presenta un coeficiente alpha también mayor, lo que indica que esta mitad es un poco más consistente internamente.

La estimación de la confiabilidad para la prueba completa es similar al alpha obtenido para el test completo. En ambos casos se trata de un valor alrededor de 0,80 para pruebas confiables (Anastasi y Urbina, 1998).

Tabla 8. División por mitades

	Media	Std. Deviation	N de *Items*	Alpha
Part 1	7,0523	3,91935	20(a)	0,675
Part 2	6,5303	3,45287	20(b)	0,604
Ambas partes	13,5827	6,72204	40	

Finalmente, se analizó el aporte de cada *item* a la confiabilidad del instrumento a partir del valor del coeficiente alpha si el *item* es eliminado. Los resultados se presentan en la Tabla 9.

Tabla 9. ***Aporte de los items a la confiabilidad de la prueba*** **(alpha =0.7853)**

Item Eliminado	alpha	Cambio	Item Eliminado	alpha	Cambio
1	0,782	-0,003	21	0,784	-0,002
2	0,777	-0,008	22	0,774	-0,012
3	0,777	-0,009	23	0,781	-0,004
4	0,790	0,004	24	0,784	-0,001
5	0,786	0,001	25	0,783	-0,002
6	0,774	-0,012	26	0,783	-0,002
7	0,778	-0,007	27	0,773	-0,012
8	0,778	-0,008	28	0,784	-0,002
9	0,790	0,005	29	0,778	-0,008
10	0,786	0,001	30	0,786	0,001
11	0,782	-0,003	31	0,774	-0,011
12	0,778	-0,007	32	0,783	-0,003
13	0,782	-0,004	33	0,773	-0,012
14	0,783	-0,003	34	0,776	-0,009
15	0,781	-0,004	35	0,782	-0,004
16	0,780	-0,005	36	0,787	0,002
17	0,776	-0,010	37	0,789	0,003
18	0,783	-0,003	38	0,782	-0,003
19	0,774	-0,011	39	0,780	-0,006
20	0,779	-0,006	40	0,785	-0,001

Tal como puede observarse en la Tabla 9, el 82,5% de los *items* contribuye con la confiabilidad del instrumento, lo que orienta hacia la revisión, de 7 *items* (4, 5, 9, 10, 30, 36, 37) de la prueba, dado que su eliminación incrementa ligeramente el coeficiente de confiabilidad.

Validez

El primer paso para juzgar la validez del instrumento fue el análisis del proceso de construcción. En este sentido, la prueba fue diseñada con base en las especificaciones descritas, las cuales representan las habilidades que subyacen a la ejecución en matemáticas y en las distintas áreas asociadas con ésta. Para la derivación de las dimensiones se consultaron distintas fuentes de información: (a) tests con validez probada para el mismo objetivo; (b) libros de texto sobre habilidades matemáticas; (c) entrevistas con especialistas en matemáticas, en didáctica de las matemáticas y en psicometría.

Para garantizar el comportamiento psicométrico adecuado de la prueba, ésta se diseña utilizando un 50% de banco, es decir *items* probados en poblaciones similares de la misma Universidad, el resto de *items* fue diseñado por un grupo de especialistas en el área de matemática con formación en el área de construcción de instrumentos. Posteriormente, la prueba fue evaluada por otro grupo de jueces expertos que verificaron el ajuste del *item* a la dimensión de evaluación, al nivel de dificultad estimado (en el caso de los items nuevos); y los aspectos formales de diseño. Sobre la base de esta revisión se realizaron ajustes a la prueba.

En la revisión de las especificaciones para el diseño de la prueba se consideran las 12 dimensiones que la componen, igualmente importantes, por lo que todas deberían tener la misma representación en la prueba. Sin embargo, en el muestreo aleatorio realizado para la selección de los 20 *items* del banco, se obtuvo una distribución no equivalente entre las dimensiones, quedando las dimensiones Analogías Figurativos y Razonamiento Lógico con menor proporción de *items*.

Por otra parte, tal como puede observarse en la Tabla 10, los niveles de dificultad obtenidos no se ajustan del todo a los niveles de dificultad estimados

en el diseño de la prueba, la desproporción se dio en las categoría **Fácil** y **Muy difícil**, evidenciándose menor cantidad a la estimada de *items* fáciles y mayor proporción a la estimada de *items* difíciles, lo que hace a la prueba como conjunto más difícil a lo planificado, por lo que la distribución de puntajes obtenidos no es simétrica.

Tabla 10. Distribución de Niveles de Dificultad (obtenidos y estimados)

		Muy fácil (0,81 <p ≤ 1,00)	**Fácil** (0,61 <p ≤ 0,80)	**Promedio** (0,41 <p ≤ 0,60)	**Difícil** (0,21 <p ≤ 0,40)	**Muy Difícil** (0,00 <p ≤ 0,20)
Obtenidos	f_i	3	3	17	8	9
	%	7	7	43	20	23
Estimado	f_i	2	8	20	8	2
	%	5	20	50	20	5

Se considera que el diseño de la prueba se llevó a cabo con los pasos de validación y control, y la rigurosidad metodológica pertinente; sin embargo, se podrían hacer algunos ajustes que aumentarían la equivalencia en la representación de las dimensiones.

En cuanto a la validez de criterio, el coeficiente de correlación simple entre los resultados en la prueba de RBC y el rendimiento académico en el postgrado fue de 0,25, lo que es consistente con los resultados obtenidos por Willighan, (1974), superiores a los obtenidos por Golberg y Alliger (1992), e inferiores a los obtenidos por Kuncel, et al. (2001). Sin embargo, en el caso de estos últimos, en los coeficientes que son reportados ha sido corregido el efecto de la restricción del rango en las variables implicadas.

Para la interpretación del coeficiente obtenido es también relevante considerar que el mismo se ve afectado por: (a) la restricción en el rango de la medida producto del corte de selección (Kuncel, et al. 2001) (b) la heterogeneidad y posible inconsistencia en la medida de criterio, pues no existen criterios uniformes para la asignación de las calificaciones entre

postgrados, ni dentro de cada postgrado. No se dispone de medidas de indicadores de confiabilidad y validez de la medida de criterio (Willigham, 1974); (c) la correlación no se obtuvo en forma diferencial por tipo de postgrado al que se opta, y la evidencia empírica señala que la habilidad cuantitativa tiene un valor predictivo diferencial en esta variable (Kuncel, et al. 2001). Todas estas consideraciones permiten afirmar que el coeficiente obtenido en la presente investigación podría estar subestimando la validez predictiva del instrumento.

Los resultados del análisis de varianza de una vía, para probar la primera hipótesis relacionada con la validez de constructo a partir de la cual se esperan diferencias en la habilidad cuantitativa por sexo, se presentan en las Tablas 11 y 12, donde se evidencia que sí hay diferencias estadísticamente significativas en el nivel de habilidad cuantitativo debidas al sexo, específicamente los hombres presentan una ejecución superior a la de las mujeres en habilidad cuantitativa, lo que corrobora la Hipótesis 1 y es consistente con las investigaciones consultadas.

Tabla 11. ***Diferencias en RBC por Sexo***

Sexo	N	Media	Desviación Estándar	Error
Masculino	127	15,52	7,051	0,626
Femenino	175	12,25	6,062	0,458

Tabla 12. ***Análisis de Varianza por Sexo***

	Sumatoria de cuadrados	df	Media de cuadrados	F	Sig.
Entre-grupos	786,852	1	786,852	18,648	,000
Intra-grupos	12658,360	300	42,195		
Total	13445,212	301			

La Tabla 13 muestra que la ejecución promedio de los sujetos, como era esperado, varía según el tipo de profesión. La ejecución en las profesiones con mayor carga numérica es superior. Así, se observa la ejecución en Ciencias Básicas (1ss) es muy superior al promedio del resto de las categorías de profesión, seguida por la categoría que agrupa a Ingeniería, Arquitectura y Tecnología y luego tienen ejecución media similar Ciencias del Agro y del Mar, Ciencias de la Salud, Ciencias de la Educación y Ciencias Sociales y luego hay un grupo de ejecución muy inferior en la prueba formado por los profesionales de las áreas de Humanidades, Letras y Artes y los de Ciencias y Artes Militares. Estos resultados son evidentemente consistentes con la ejecución esperada, de acuerdo con el constructo medido (habilidad cuantitativa) y su relación con el tipo de profesión.

Tabla 13. ***Resultados por tipo de profesión***

Profesión	Media	N	Desviación Estándar
No Declarado	5,67	4	1,828
Ciencias Básicas	22,67	1	.
Ingeniería, Arquitectura y Tecnología	16,87	70	5,895
Ciencias del Agro y del Mar	12,71	8	7,929
Ciencias de la Salud	13,13	53	5,847
Ciencias de la Educación	13,97	24	8,201
Ciencias Sociales	12,64	132	6,515
Humanidades, Letras y Arte	8,52	7	5,779
Ciencias y Artes Militares	8,78	3	4,219
Total	13,62	302	6,683

Como puede observarse en la Tabla 14, las diferencias en la ejecución en la prueba de habilidad cuantitativa entre los profesionales de Ingeniería, Arquitectura y Tecnología y los de las Ciencias Sociales son estadísticamente significativas, lo que corrobora la Hipótesis 2.

Tabla 14. Prueba T para Muestras Independientes por Tipo de Profesión

	t	df	Sig. (2-colas)	Diferencias de medias	Error Est. dif	Intervalo de confianza dif. 95%	
						Inferior	Superior
Asumiendo igual varianza	4,532	200	0,000	4,227	0,933	2,388	6,067
No asumiendo igual varianza	4,674	153,429	0,000	4,227	0,904	2,440	6,014

Análisis de Items

A continuación se presenta la Tabla 15 con el Reporte Cuantitativo de los *items*.

Tabla 15. ***Reporte de Indicadores Cuantitativos***

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	36	0,12	-0,10	-0,17	11,69	-0,28
	b	1,00	36	0,12	0,26	0,42	19,27	0,85
1	c	-0,33	4	0,01	-0,05	-0,18	10,50	-0,46
	d	-0,33	84	0,28	-0,02	-0,02	13,41	-0,02
	other	0,00	142	0,47	-0,11	-0,14	12,80	-0,12

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	5	0,02	-0,08	-0,23	9,67	-0,58
	b	-0,33	16	0,05	-0,29	-0,61	5,29	-1,23
2	c	1,00	172	0,57	0,33	0,42	15,95	0,35
	d	-0,33	9	0,03	-0,04	-0,11	11,89	-0,25
	other	0,00	100	0,33	-0,25	-0,33	11,18	-0,36

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	230	0,76	0,38	0,53	15,25	0,25
	b	-0,33	11	0,04	-0,09	-0,21	10,55	-0,45
3	c	-0,33	2	0,01	-0,13	-0,58	2,67	-1,63
	d	-0,33	13	0,04	-0,24	-0,52	6,13	-1,11
	other	0,00	46	0,15	-0,32	-0,49	8,54	-0,75

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	112	0,37	0,10	0,13	14,46	0,13
	b	-0,33	25	0,08	-0,03	-0,06	12,84	-0,11
4	c	1,00	109	0,36	0,09	0,12	15,23	0,25
	d	-0,33	35	0,12	-0,27	-0,44	8,63	-0,74
	other	0,00	21	0,07	-0,17	-0,32	9,43	-0,62

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	14	0,05	0,01	0,02	13,90	0,05
	b	1,00	173	0,57	0,15	0,19	14,97	0,21
5	c	-0,33	30	0,10	-0,10	-0,18	11,50	-0,31
	d	-0,33	37	0,12	-0,12	-0,19	11,45	-0,32
	other	0,00	48	0,16	-0,14	-0,21	11,42	-0,32

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	204	0,68	0,43	0,56	15,90	0,35
	b	-0,33	14	0,05	-0,22	-0,47	6,93	-0,99
6	c	-0,33	14	0,05	-0,11	-0,25	10,12	-0,52
	d	-0,33	21	0,07	-0,23	-0,44	7,90	-0,85
	other	0,00	49	0,16	-0,28	-0,43	9,25	-0,64

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	9	0,03	-0,24	-0,59	4,56	-1,34
	b	1,00	242	0,80	0,34	0,48	14,92	0,20
7	c	-0,33	12	0,04	-0,16	-0,35	8,44	-0,77
	d	-0,33	18	0,06	-0,11	-0,22	10,65	-0,44
	other	0,00	21	0,07	-0,25	-0,47	7,48	-0,91

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	18	0,06	-0,20	-0,39	8,35	-0,78
	b	-0,33	26	0,09	-0,09	-0,16	11,60	-0,29
8	c	1,00	188	0,62	0,35	0,45	15,82	0,33
	d	-0,33	17	0,06	-0,08	-0,17	11,29	-0,34
	other	0,00	53	0,18	-0,31	-0,45	9,10	-0,67

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	33	0,11	-0,12	-0,20	11,33	-0,33
	b	-0,33	127	0,42	0,17	0,21	14,90	0,20
9	c	1,00	39	0,13	-0,01	-0,01	14,58	0,15
	d	-0,33	4	0,01	-0,22	-0,74	0,83	-1,90
	other	0,00	99	0,33	-0,09	-0,11	12,75	-0,12

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	87	0,29	-0,01	-0,02	13,44	-0,02
	b	-0,33	83	0,27	-0,08	-0,11	12,68	-0,13
10	c	-0,33	31	0,10	0,10	0,17	15,55	0,29
	d	1,00	55	0,18	0,16	0,23	16,85	0,49
	other	0,00	46	0,15	-0,21	-0,32	10,22	-0,50

	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	24	0,08	-0,25	-0,45	7,94	-0,84
	b	-0,33	5	0,02	-0,02	-0,05	12,80	-0,12
11	c	1,00	198	0,66	0,21	0,27	14,97	0,21
	d	-0,33	7	0,02	-0,09	-0,23	9,86	-0,55
	other	0,00	68	0,23	-0,13	-0,18	11,97	-0,24

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	245	0,81	0,33	0,47	14,85	0,19
	b	-0,33	40	0,13	-0,35	-0,55	7,62	-0,89
12	c	-0,33	1	0,00	-0,09	-0,54	2,67	-1,63
	d	-0,33	3	0,01	-0,03	-0,10	11,78	-0,27
	other	0,00	13	0,04	-0,13	-0,30	9,31	-0,64

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	37	0,12	-0,17	-0,27	10,59	-0,44
	b	-0,33	11	0,04	-0,11	-0,26	9,76	-0,57
13	c	1,00	209	0,69	0,27	0,35	15,13	0,23
	d	-0,33	20	0,07	-0,03	-0,06	12,83	-0,11
	other	0,00	25	0,08	-0,28	-0,50	7,35	-0,93

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	-0,33	43	0,14	0,03	0,04	14,02	0,07
	b	1,00	45	0,15	0,21	0,32	17,81	0,63
14	c	-0,33	4	0,01	-0,06	-0,19	10,33	-0,48
	d	-0,33	14	0,05	-0,20	-0,44	7,45	-0,91
	other	0,00	196	0,65	-0,11	-0,15	13,01	-0,08

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	1,00	119	0,39	0,26	0,34	16,47	0,43
	b	-0,33	74	0,25	-0,05	-0,07	13,02	-0,08
15	c	-0,33	24	0,08	-0,16	-0,29	10,01	-0,53
	d	-0,33	19	0,06	-0,17	-0,33	9,23	-0,65
	other	0,00	66	0,22	-0,16	-0,22	11,55	-0,30

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	-0,33	85	0,28	-0,11	-0,14	12,42	-0,17
	b	-0,33	6	0,02	-0,12	-0,34	8,06	-0,82
16	c	1,00	147	0,49	0,31	0,39	16,29	0,40
	d	-0,33	6	0,02	-0,09	-0,26	9,33	-0,63
	other	0,00	58	0,19	-0,30	-0,44	9,41	-0,62

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	1,00	126	0,42	0,39	0,50	17,24	0,55
	b	-0,33	24	0,08	-0,13	-0,24	10,60	-0,44
17	c	-0,33	12	0,04	-0,06	-0,15	11,47	-0,31
	d	-0,33	10	0,03	-0,10	-0,24	10,00	-0,53
	other	0,00	130	0,43	-0,33	-0,41	11,05	-0,38

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	1,00	212	0,70	0,23	0,30	14,93	0,20
	b	-0,33	9	0,03	0,01	0,03	14,00	0,06
18	c	-0,33	7	0,02	0,01	0,03	14,10	0,08
	d	-0,33	42	0,14	-0,24	-0,37	9,63	-0,59
	other	0,00	32	0,11	-0,21	-0,35	9,55	-0,60

Item	option	wt.	n	p	pb(r)	b(r)avg.		z
	a	-0,33	15	0,05	-0,17	-0,36	8,58	-0,75
	b	1,00	185	0,61	0,43	0,55	16,24	0,40
19	c	-0,33	28	0,09	-0,15	-0,26	10,43	-0,47
	d	-0,33	9	0,03	-0,11	-0,29	9,22	-0,65
	other	0,00	65	0,22	-0,35	-0,49	9,13	-0,66

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	7	0,02	-0,21	-0,57	4,52	-1,35
	b	-0,33	3	0,01	-0,06	-0,23	9,44	-0,62
20	c	1,00	250	0,83	0,30	0,45	14,68	0,16
	d	-0,33	7	0,02	-0,14	-0,39	7,43	-0,92
	other	0,00	35	0,12	-0,24	-0,40	9,09	-0,67

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	18	0,06	-0,12	-0,24	10,41	-0,47
	b	-0,33	22	0,07	-0,06	-0,11	12,23	-0,20
21	c	-0,33	74	0,25	-0,06	-0,08	12,87	-0,11
	d	1,00	122	0,40	0,22	0,28	16,10	0,38
	other	0,00	66	0,22	-0,20	-0,28	11,03	-0,38

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	12	0,04	-0,08	-0,18	11,00	-0,38
	b	-0,33	20	0,07	-0,06	-0,12	11,97	-0,24
22	c	-0,33	14	0,05	-0,08	-0,17	11,17	-0,36
	d	1,00	110	0,36	0,46	0,59	18,22	0,69
	other	0,00	146	0,48	-0,41	-0,51	10,74	-0,42

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	23	0,08	-0,06	-0,11	12,16	-0,21
	b	1,00	43	0,14	0,22	0,35	18,20	0,69
23	c	-0,33	26	0,09	-0,17	-0,30	9,85	-0,56
	d	-0,33	22	0,07	-0,05	-0,10	12,27	-0,19
	other	0,00	188	0,62	-0,04	-0,05	13,37	-0,03

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	12	0,04	-0,03	-0,08	12,47	-0,16
	b	-0,33	105	0,35	-0,11	-0,14	12,57	-0,15
24	c	-0,33	12	0,04	-0,06	-0,13	11,67	-0,29
	d	1,00	126	0,42	0,22	0,27	16,00	0,36
	other	0,00	47	0,16	-0,22	-0,34	10,11	-0,52

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	81	0,27	0,25	0,34	17,18	0,54
	b	-0,33	30	0,10	0,03	0,05	14,19	0,09
25	c	-0,33	72	0,24	-0,04	-0,05	13,14	-0,06
	d	-0,33	8	0,03	-0,09	-0,23	10,04	-0,53
	other	0,00	111	0,37	-0,26	-0,33	11,33	-0,34

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	24	0,08	-0,03	-0,05	12,94	-0,09
	b	-0,33	63	0,21	0,02	0,02	13,78	0,03
26	c	-0,33	24	0,08	-0,09	-0,16	11,56	-0,30
	d	1,00	22	0,07	0,23	0,43	20,18	0,98
	other	0,00	169	0,56	-0,09	-0,12	13,02	-0,08

Item	o ption	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	132	0,44	0,46	0,58	17,55	0,59
	b	-0,33	17	0,06	-0,03	-0,06	12,71	-0,13
27	c	-0,33	14	0,05	-0,17	-0,38	8,26	-0,79
	d	-0,33	13	0,04	-0,10	-0,23	10,36	-0,48
	other	0,00	126	0,42	-0,39	-0,50	10,46	-0,46

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	34	0,11	-0,09	-0,14	11,95	-0,24
	b	-0,33	52	0,17	-0,07	-0,10	12,62	-0,14
28	c	-0,33	18	0,06	-0,13	-0,26	10,06	-0,53
	d	1,00	139	0,46	0,21	0,27	15,76	0,33
	other	0,00	59	0,20	-0,17	-0,24	11,29	-0,34

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	2	0,01	0,00	0,01	13,83	0,04
	b	-0,33	8	0,03	-0,09	-0,24	9,83	-0,56
29	c	1,00	172	0,57	0,34	0,43	15,96	0,35
	d	-0,33	3	0,01	-0,16	-0,61	2,67	-1,63
	other	0,00	117	0,39	-0,35	-0,45	10,62	-0,44

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	31	0,10	0,20	0,34	17,54	0,59
	b	-0,33	27	0,09	0,04	0,08	14,49	0,14
30	c	1,00	58	0,19	0,18	0,25	16,90	0,50
	d	-0,33	13	0,04	-0,04	-0,10	12,18	-0,21
	other	0,00	173	0,57	-0,32	-0,40	11,72	-0,28

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	1,00	177	0,59	0,44	0,55	16,41	0,42
	b	-0,33	21	0,07	-0,13	-0,25	10,37	-0,48
31	c	-0,33	10	0,03	-0,15	-0,38	7,97	-0,84
	d	-0,33	16	0,05	-0,07	-0,15	11,48	-0,31
	other	0,00	78	0,26	-0,39	-0,52	9,17	-0,66

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	7	0,02	0,00	-0,01	13,48	-0,02
	b	-0,33	29	0,10	0,03	0,05	14,14	0,08
32	c	1,00	24	0,08	0,25	0,46	20,31	1,00
	d	-0,33	31	0,10	-0,07	-0,12	12,22	-0,20
	other	0,00	211	0,70	-0,14	-0,19	12,94	-0,10

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	7	0,02	-0,18	-0,50	5,67	-1,18
	b	1,00	115	0,38	0,48	0,61	18,21	0,69
33	c	-0,33	3	0,01	-0,08	-0,31	8,11	-0,81
	d	-0,33	58	0,19	-0,07	-0,10	12,60	-0,15
	other	0,00	119	0,39	-0,41	-0,52	10,19	-0,50

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	4	0,01	-0,02	-0,06	12,50	-0,16
	b	-0,33	6	0,02	-0,04	-0,12	11,67	-0,29
34	c	1,00	129	0,43	0,39	0,49	17,08	0,52
	d	-0,33	4	0,01	-0,09	-0,30	8,42	-0,77
	other	0,00	159	0,53	-0,41	-0,51	10,97	-0,39

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	31	0,10	-0,02	-0,04	13,16	-0,06
	b	-0,33	11	0,04	-0,10	-0,24	10,09	-0,52
35	c	1,00	72	0,24	0,26	0,36	17,56	0,59
	d	-0,33	32	0,11	-0,01	-0,01	13,44	-0,02
	other	0,00	156	0,52	-0,23	-0,29	12,10	-0,22

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	17	0,06	0,13	0,27	17,25	0,55
	b	1,00	14	0,05	0,01	0,02	15,12	0,23
36	c	-0,33	37	0,12	-0,11	-0,17	11,67	-0,29
	d	-0,33	25	0,08	0,03	0,06	14,33	0,11
	other	0,00	209	0,69	-0,03	-0,05	13,43	-0,02

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	10	0,03	-0,06	-0,15	11,27	-0,34
	b	-0,33	37	0,12	0,18	0,29	16,81	0,48
37	c	-0,33	15	0,05	0,08	0,17	15,98	0,36
	d	1,00	61	0,20	0,09	0,12	15,64	0,31
	other	0,00	179	0,59	-0,26	-0,33	12,14	-0,21

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	13	0,04	-0,20	-0,44	7,26	-0,94
	b	-0,33	3	0,01	-0,01	-0,04	12,78	-0,12
38	c	1,00	188	0,62	0,21	0,27	15,08	0,22
	d	-0,33	3	0,01	-0,04	-0,14	11,00	-0,38
	other	0,00	95	0,31	-0,20	-0,26	11,58	-0,30

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	26	0,09	0,02	0,03	13,95	0,06
	b	-0,33	2	0,01	0,09	0,40	21,00	1,11
39	c	1,00	66	0,22	0,35	0,49	18,81	0,78
	d	-0,33	22	0,07	0,00	-0,01	13,50	-0,01
	other	0,00	186	0,62	-0,37	-0,48	11,60	-0,29

Item	option	wt.	n	p	pb(r)	b(r)	avg.	z
	a	-0,33	34	0,11	0,10	0,16	15,39	0,27
	b	1,00	30	0,10	0,22	0,38	19,01	0,81
40	c	-0,33	15	0,05	0,08	0,16	15,80	0,33
	d	-0,33	7	0,02	0,04	0,11	15,33	0,26
	other	0,00	216	0,72	-0,30	-0,39	12,33	-0,19

De acuerdo con estos resultados, 25 *items* son aceptados (1, 2, 3, 6, 7, 8, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 23, 24, 27, 28, 31, 33, 34, 35, 38) pues el coeficiente de correlación punto-biserial supera el criterio establecido, todos los distractores son negativos y como consecuencia, la media de la opción correcta es la mayor.

Para los *items* **en revisión**, se presentan en la Tabla 16 el tipo de problema o condición cuantitativa que no se ajusta al criterio de aceptación y se presentan los gráficos que representan su comportamiento.

Tabla 16. *Items en Revisión*

	r_{pb} de la respuesta correcta		**r_{pb} de los distractores**
item	**Positivo pero inferior a 0,20**	**Negativo**	**Nulo o positivo**
4	0,09		a = 0,10
5	0,15		a = 0,01
9		-0.01	b = 0,17
10	0,16		c = 0,10
14			a = 0,03
18			b = 0,01; c = 0,01
25			b = 0,03
26			b = 0,02
29			a= 0,01
30	0,18		a = 0,20; b = 0,04
32			a = 0,00; b = 0,03
36	0,01		a = 0,13; d = 0,03
37	0,09		b = 0,18; c = 0,08
39			a = 0,02; b = 0,09
40			a = 0,10; c = 0,08; d = 0,04

Gráfico 4. Item 4

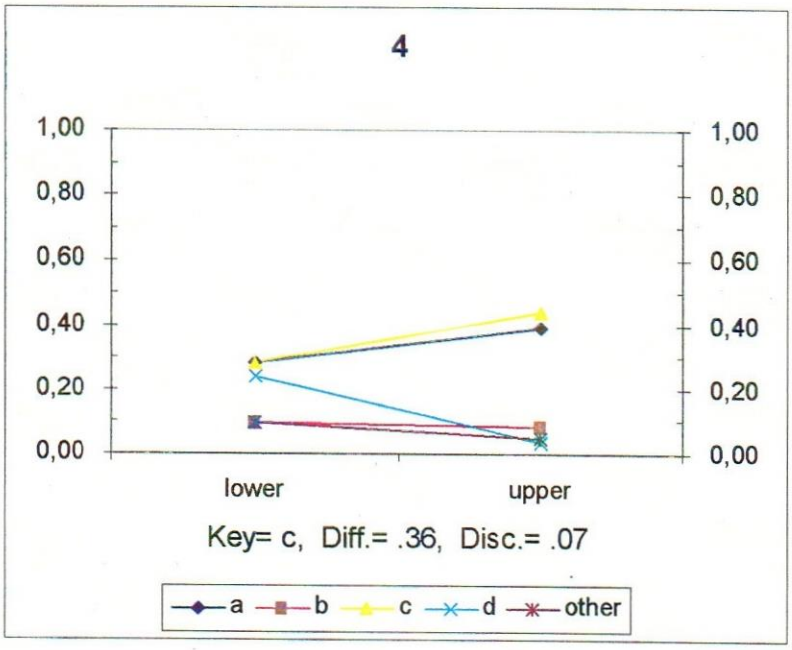


Gráfico 5. Item 5

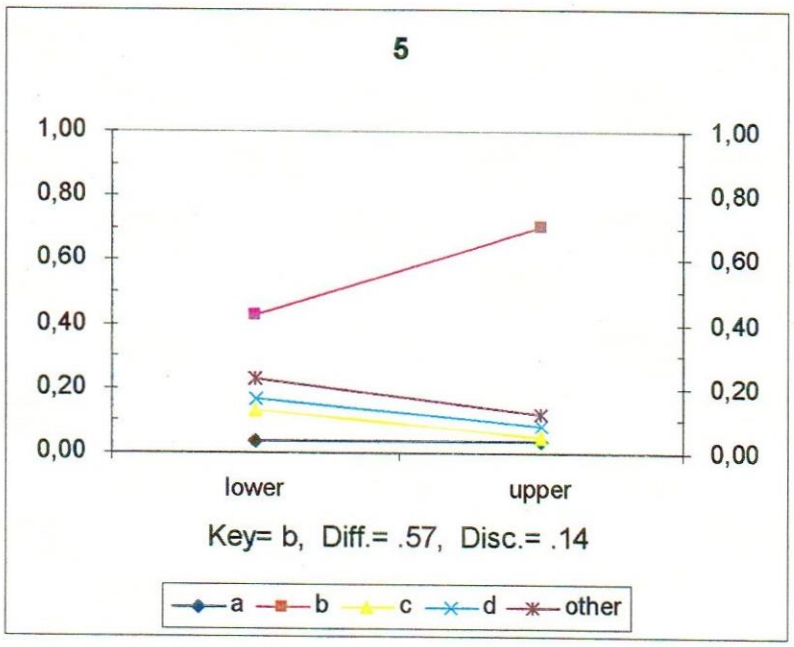


Gráfico 6. Item 9

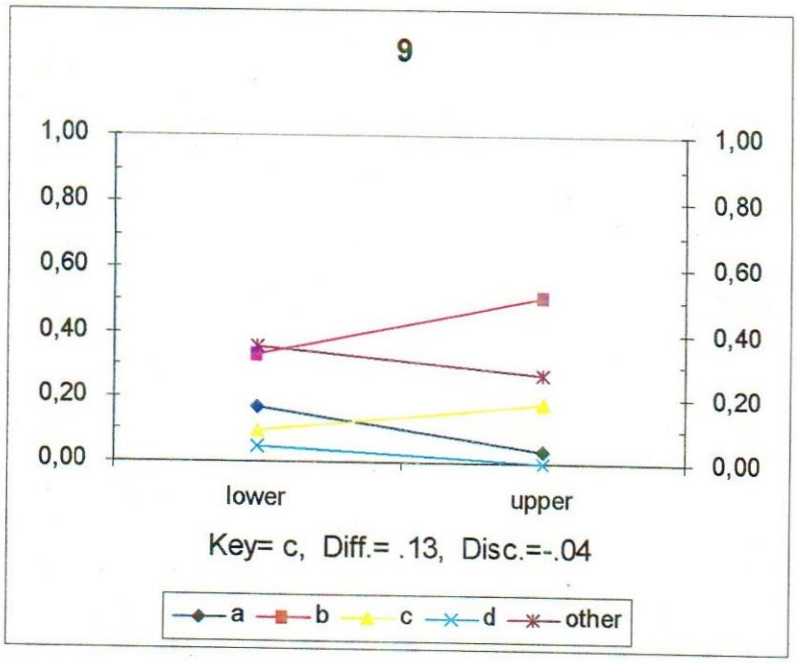


Gráfico 7. Item 10

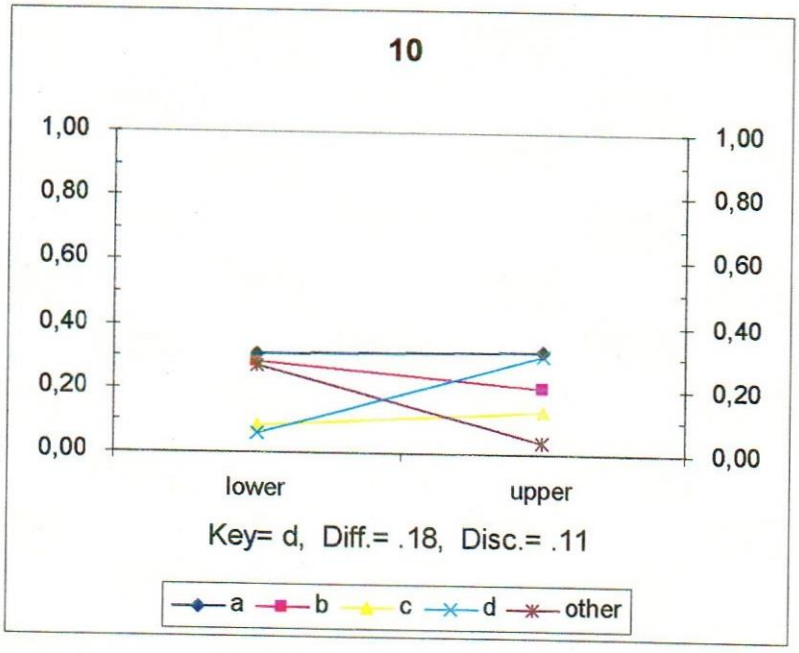


Gráfico 8. Item 14

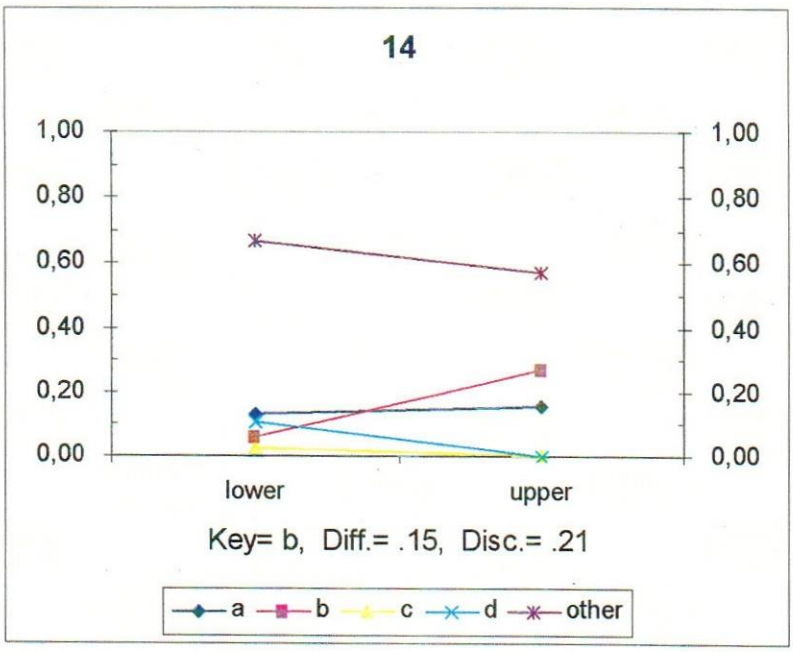


Gráfico 9. Item 18

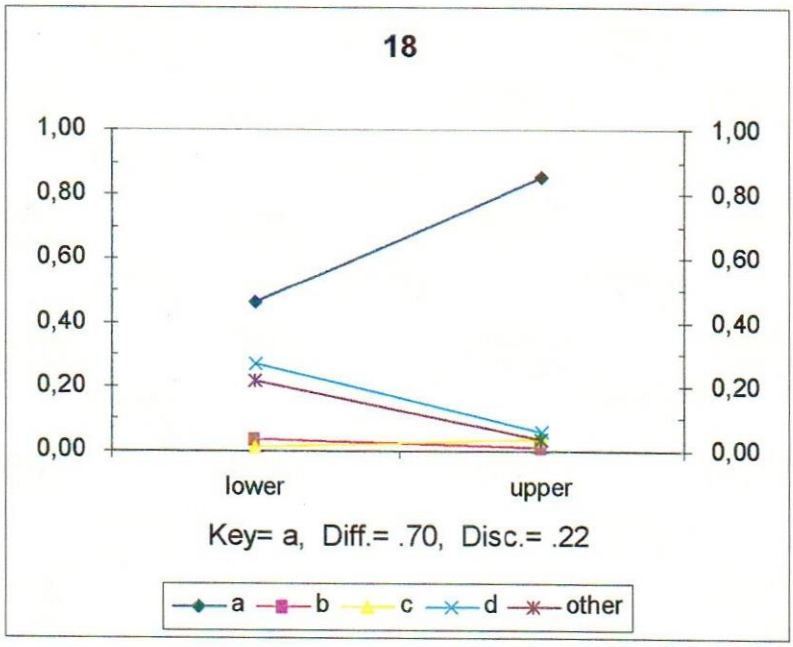


Gráfico 10. **Item 25**

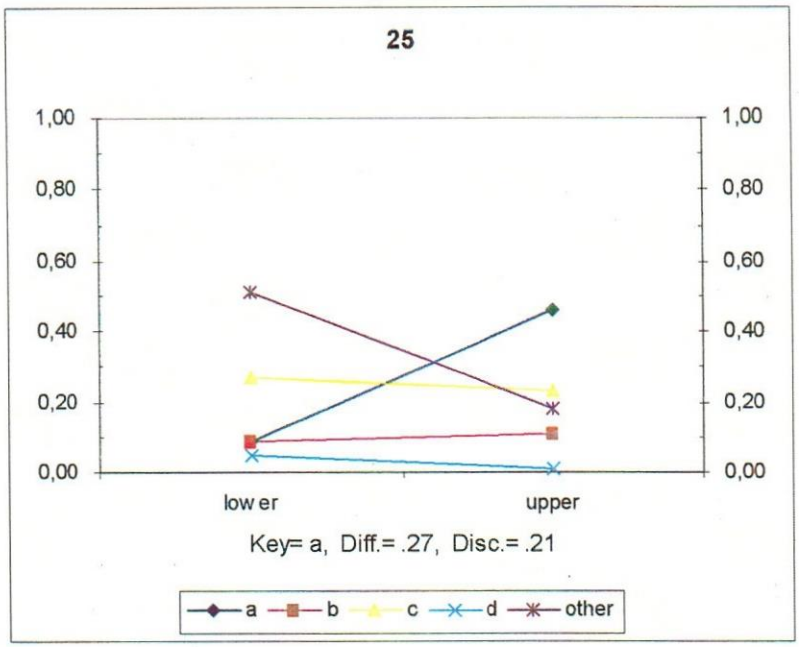


Gráfico 11. **Item 26**

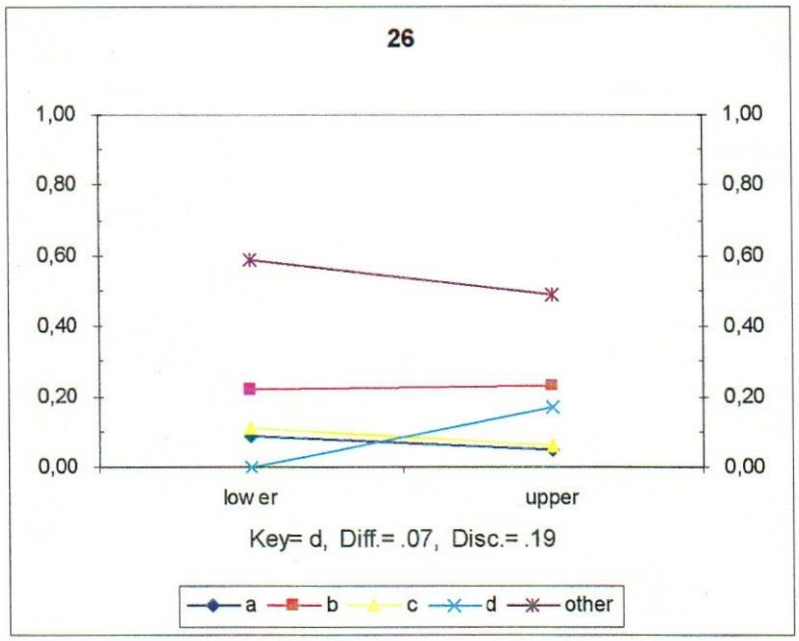


Gráfico 12. Item 30

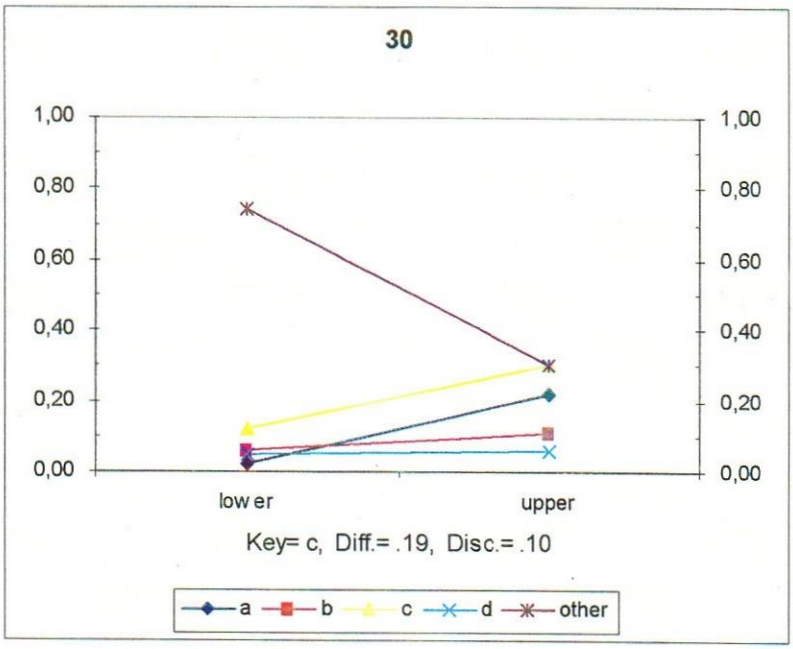


Gráfico 13. Item 32

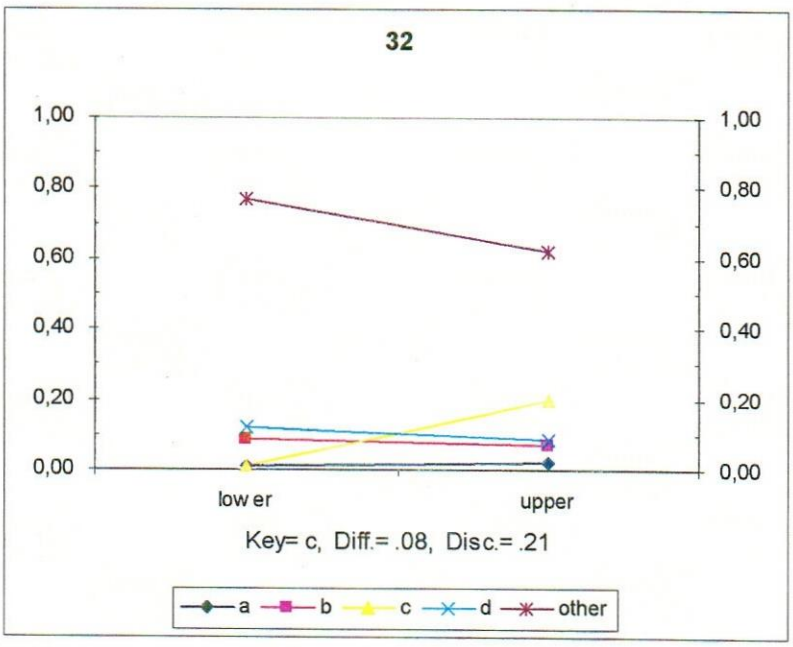


Gráfico 14. **Item 35**

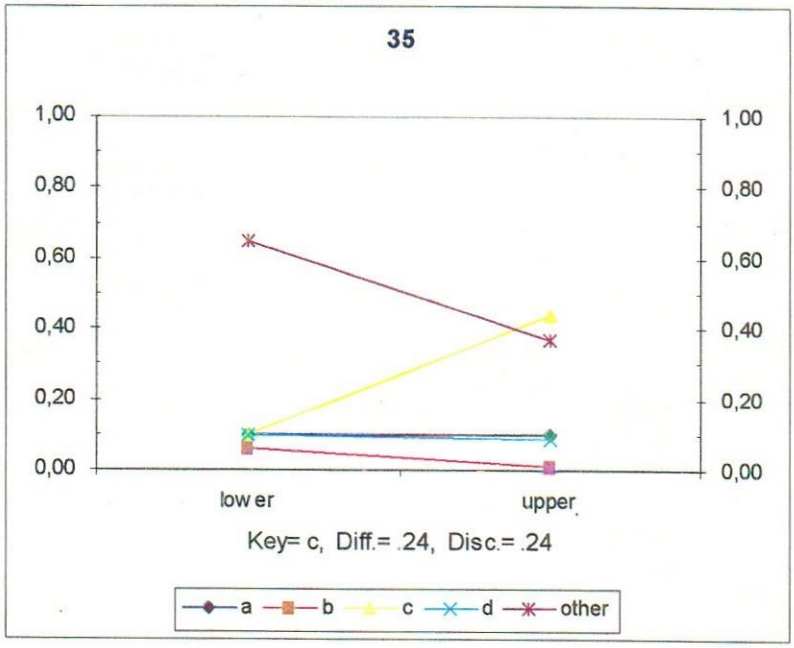


Gráfico 15. **Item 36**

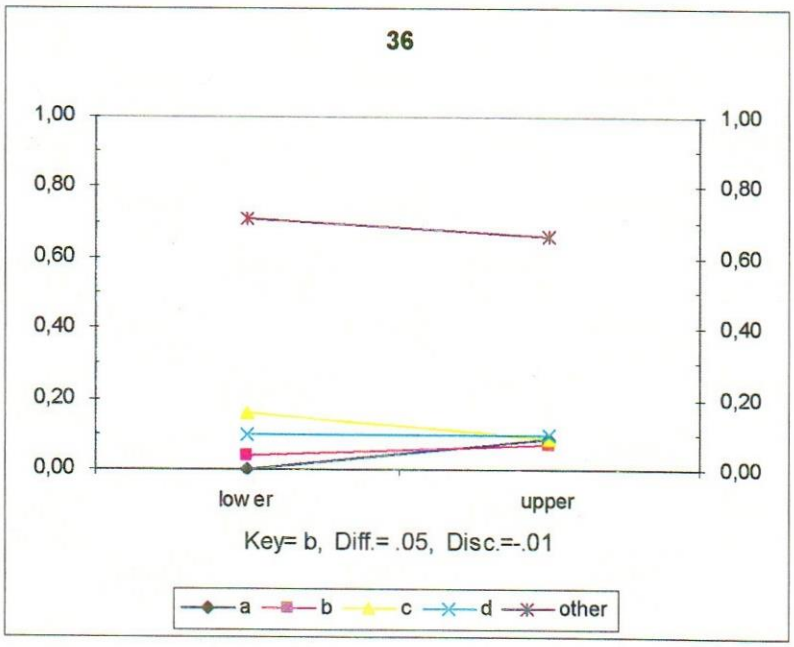


Gráfico 16. Item 37

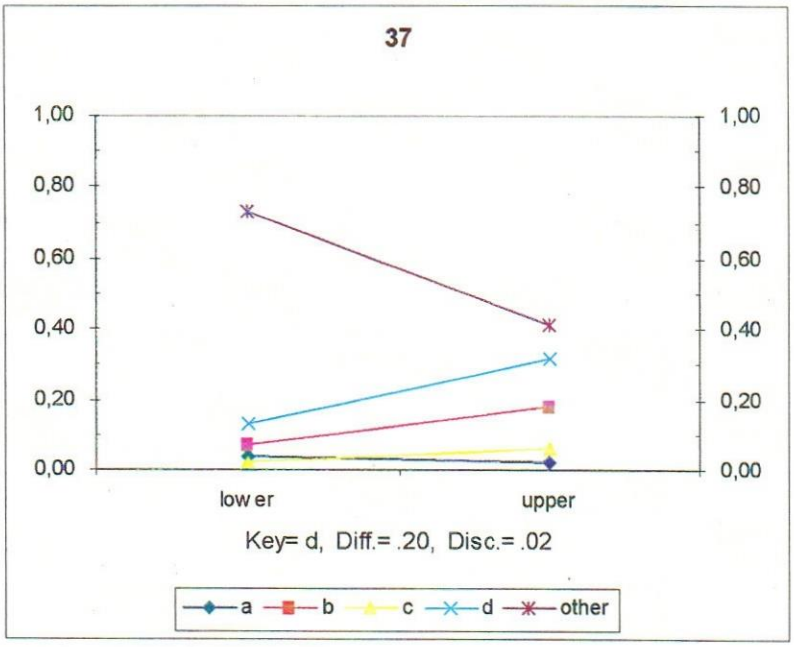


Gráfico 17. Item 39

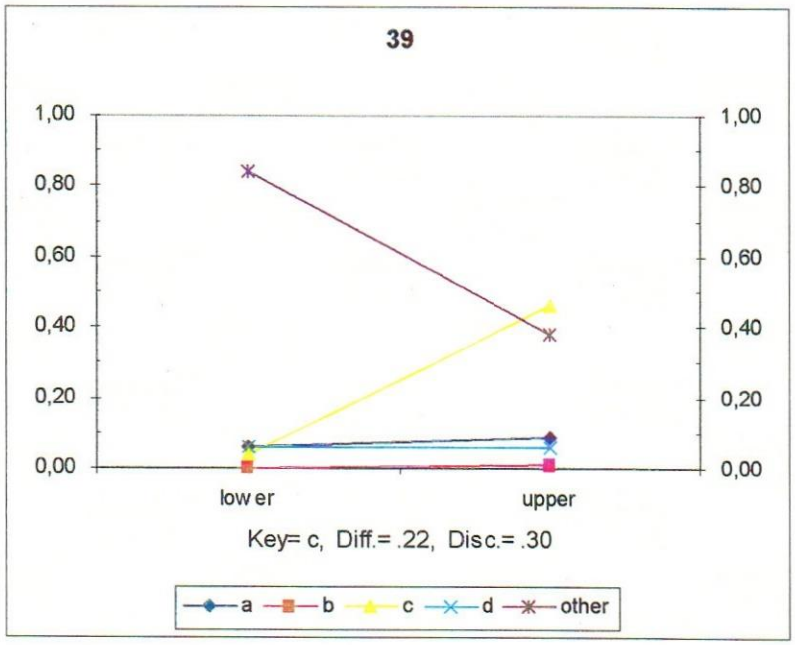
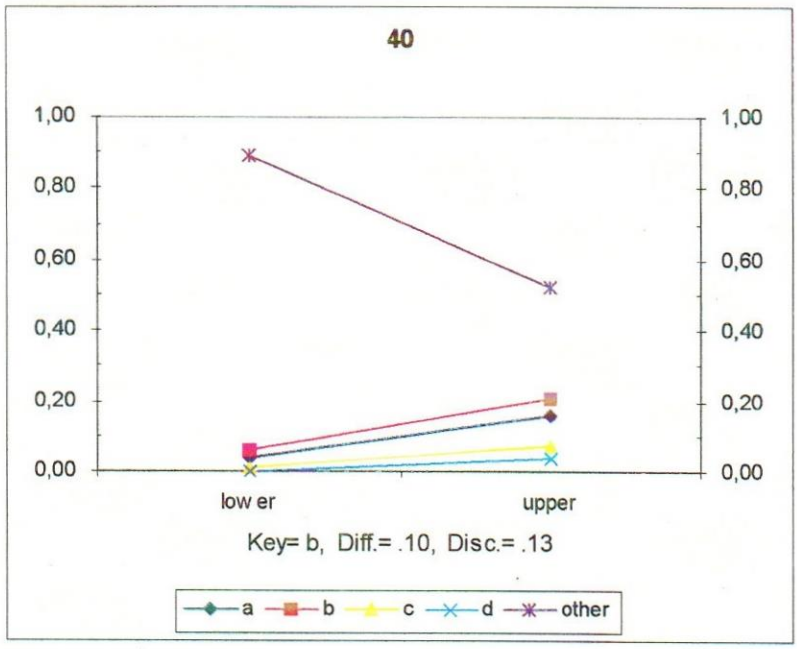


Gráfico 18. **Item 40**



Estos resultados señalan que:

- Los *items* 14, 18, 25, 26, 29, 32 y 39 poseen un adecuado poder discriminativo que supera el valor de 0,20 y aunque alguno de sus distractores presenten correlaciones positivas, la magnitud de la correlación es baja, tiende a cero, por lo que se considera que discriminan adecuadamente, y en consecuencia son aceptados sin modificación.
- El ítem 40, aunque presenta un coeficiente de correlación en la respuesta correcta superior a 0,20, posee tres distractores cuyas correlaciones son positivas y dos de ellas se diferencian de cero. Su análisis cualitativo sugiere que el *item* es mejorable mediante la modificación de los distractores.
- Aunque el coeficiente de correlación punto biserial del *item* 5 no alcanza el valor de 0,20, el coeficiente de la respuesta correcta es el mayor y los de los distractores son negativos, la media de la opción correcta está muy por encima de la media de los distractores. Por otra parte, el gráfico evidencia que presentan una adecuada discriminación entre los grupos superior e inferior. De todas maneras, con el objetivo de aumentar su poder discriminativo, se revisó cualitativamente y se determinó que puede ser mejorado modificando su planteamiento.
- En cuanto al *item* 36, aunque la pendiente de la respuesta correcta es ascendente, una alta proporción de sujetos tanto del grupo superior como del inferior dejaron de contestarlo, lo que indica que puede estar diseñado en forma confusa. A partir de la revisión cualitativa se concluyó que puede ser mejorado, modificando la opción b y corrigiendo la clave de respuestas.

- Los *items* 4, 10 y .30, no poseen un poder discriminativo suficiente (0.20), porque alguna de sus alternativas está muy cerca de la respuesta correcta. La revisión cualitativa confirma que deben ser modificadas las opciones **a**, **c y a** de los *items* 4, 10 y 30 respectivamente.
- Tanto los resultados cuantitativos como la visualización gráfica del *item* 9, sugieren que tiene un error en la clave de respuestas, lo que se confirmó mediante el análisis cualitativo por lo que se considera mejorable con la corrección de la clave.
- El *item* 37, posee varios coeficientes punto biserial positivos y el mayor no corresponde con la respuesta correcta. la primera hipótesis que esto sugiere es que puede haber un error en la clave de respuestas. Sin embargo, aun modificando la clave de respuestas, los distractores siguen teniendo coeficientes positivos, lo que sugiere que el *item* debe ser desechado.

Después del análisis cualitativo de los *items* **en revisión,** se obtuvo la distribución en las categorías **aceptados sin modificación**, **aceptados con modificación** o **descartados** que se presenta en la Tabla 17. En la tabla se especifica la parte de los *items* que debe ser modificada.

Tabla 17. ***Clasificación de los Items en Revisión***

Item	Aceptados sin modificación	Aceptados con modificación						Descartados
		Planteamiento	Alternativas					
			a	b	c	d	Clave de respuestas	
4			X					
5		X						
9							X	
10					X			
14	X							
18	X							
25	X							
26	X							
29	X							
30			X					
32	X							
36				X			X	
37								X
39	X							
40			X		X	X		

De acuerdo con estos resultados sólo el *item* 37 (razonamiento estadístico) debe ser descartado.

CONCLUSIONES.

De acuerdo con el objetivo de determinar las características psicométricas, partiendo de la Teoría Clásica de los Tests, de la prueba de Razonamiento Básico Cuantitativo del modelo PG-38, utilizado por la UCAB en el proceso de selección de aspirantes a postgrado realizado en junio del año 2003, se puede concluir que:

- Se trata de una prueba cuyos resultados son confiables en términos de consistencia interna y equivalencia.

- La prueba tiene una adecuada validez de contenido derivada de la rigurosidad metodológica en su proceso de construcción, sin embargo, podría mejorarse ajustando la distribución de niveles de dificultad extremos y equilibrando la representación de los *items* de las dimensiones Razonamiento Lógico y Analogías Figurativas, las cuales quedaron sub-representadas en la prueba.

- La prueba tiene validez predictiva, en tanto que el coeficiente de correlación obtenido es estadísticamente significativo y similar al reportado por investigaciones similares, además, dados lor cortes de selección y su consecuente restricción en rango y considerando las características de la medida de criterio, se infiere que el coeficiente está subestimado, por tanto su magnitud debe ser mayor al valor obtenido.

- Se obtuvieron adecuados indicadores de validez de constructo. Se probaron las hipótesis planteadas sobre las diferencias significativas en la habilidad matemática de acuerdo con el sexo y con el tipo de profesión. Sin embargo, es importante tener claro que la validez de constructo es un proceso que puede ser enriquecido con múltiples indicadores que permitan probar empíricamente el comportamiento de la habilidad cuantitativa a partir de los resultados en la prueba de RBC, lo que fortalece la validez de las inferencias que se hacen a partir de estos resultados.

- El comportamiento de los *items* se distribuye de la siguiente manera: (a) 80% de los *items* son aceptados sin modificación porque discriminan adecuadamente entre el grupo superior y el grupo inferior; (b) 18% de *items*, (4, 5, 9, 10, 30, 36 y 40) no tiene un adecuado comportamiento psicométrico pero puede ser corregido mediante pequeñas modificaciones sugeridas a partir del análisis cualitativo; y (d) el 2% (37) de *items* que debe ser eliminado porque tiene un comportamiento psicométrico inadecuado y su estructura no admite modificación.

REFERENCIAS BIBLIOGRÁFICAS

Anastasi, A. y Urbina, S. (1998). *Test psicológicos*. México: Prentice Hall.

Bloom, S., Hastings, T. y Madaus G. (1975). *Evaluación del aprendizaje*. Buenos Aires: Troquel.

Borrego, R.(1985). *Efecto de tres estrategias instruccionales: seminario, conferencia del docente y la combinación de ambas sobre el aprendizaje de los alumnos en la disciplina Metodología de la Investigación*. Tabajo de licenciatura no publicado. Universidad del Zulia, Maracaibo, Venezuela.

Campbell, D. y Fiske,D. (1959). Las validaciones convergente y discriminante pos la matriz multirasgo-multimétodo. Reimpreso por la Asociación Americana de Psicología del *Psichological Bulletin, 56,* 81-105.

Crook, R. (1984).The linkage of work achievement to self steem, career maturity and college achievement. *Journal of Vocational Behavior, 25,* 70-79.

Crouse, J. y Trustheim, D. (1988). The *case against the SAT*. Chicago: University of Chicago Press.

Díez, J. y Moulines, U. (1999). *Fundamentos de Filosofía de la Ciencia*. Barcelona: Ariel.

Ferrari, J. y Parker, J. (1992). High School achievement,, self efficacy, and locus of control as precitors of freshman academic performance. *Psychological Reports, 71*, 515-945.

García, R. (1985a). Locus *de control, clase social y rendimiento académico*. Universidad de Los Andes: Laboratorio de Psicología, 7.

García, R. (1985b). *Motivación y rendimiento académico*. Universidad de Los Andes: Laboratorio de Psicología, 59.

Carballo, D. (1985). *Efecto del tiempo requerido y de las competencias matemáticas básicas en el rendimiento de los estudiantes de física en el sistema instruccional individualizado a nivel superior.* Trabajo de licenciatura no publicado. Universidad Central de Venezuela, Caracas, Venezuela

Golberg, E., y Alliger, G.(1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement, 52*, 1009-1027.

Hoffman, Ch. (1986). Testing the disadvantaged. *American Psichologist,* 944-945.

House, J., y Johnson, J., (1993). Graduate record examination scores and academic background variables as predictors of graduate degree completion. *Educational and Psychological Measurement, 53*, 551-556.

Kerlinger, F., y Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales*. México: McGraw-Hill.

Kianian, A. (1996). Gender and Mathematlcs achievement parity: evidence from post-secondary education. *Education, 116*, 586-592

Kuncel, N., Hezlett, S., y Ones, D. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127(1)*, 162-181.

Magnusson, D. (1972). *Teoría de los* tests. México: Trillas.

McKee, A., Mallory, S., y Campbell, J. (2001). The GRE and undergraduate GPA: Predicting graduate *GPA in a criminal justice graduate program*. Journal of Criminal Justice Education, 12, 311-317.

Martínez, R. (1995). *Psicometría: Teoría de los* tests *psicológicos y educativos*. Madrid: Síntesis Psicológica.

Matt, G., Pechersky, B. y Cervantes, Ch. (1991) High school study habits and early college achievement. *Psychological Reports, 69*, 91-96

Morrison, T., y Morrison, M. (1995). A meta-analytic assessment of the predictive validiity of the quantitative and verbal components of the graduate record examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55 (2)*, 309-316.

Muñiz, J. (2001). *Teoría clásica de los* tests. Madrid: Ediciones Pirámide.

Nelson, L. (2001). Item Analyssus for Test and Surveys using Lertao 5. Facultad of Education. Curtin Iniversity of Technology. Perth, Western Australia.

Newman, R., (1968). GRE Scores as predictors of GPA for Psychology.

Nettles, J. (1985) SAT score and American states: seeking for useful meaning. *Journal of Educational Measurement, 22*, 305-312.

Norcross, J., Hanych, J., y Terranova, R., (1996) Graduate study in psycho: 1992-1993. *American Psychologist, 51*, 631-643.

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

Oficina de Planificación del Sector Universitario (OPSU)- Centro Nacional para el Mejoramiento de la Enseñanza de la Ciencia (CENAMEC),(1986). *Diagnóstico del nivel de conocimientos en Biología, Ciencias de la tierra, Física, uso instrumental del lenguaje, Matemática y Química, en estudiantes que egresan del Ciclo Básico Común de Educación Media*. Caracas: OPSU-CENAMEC.

Oficina de Planificación del Sector Universitario (2001). *Oportunidades de Estudio en las Instituciones de Educación Superior de Venezuela*. Caracas: Consejo Nacional de Universidades

Parra, N. (1993). Motivación, rendimiento académico y laboral. Recuperado en junio 2, 2004, de http://www.monografías.com/trabajos14/motivación academ/motivación academ2.shtlm

PROINVESCA (1998a). *Examen de Admisión Universidad Metropolitana*. Caracas: Corporación Belmont.

PROINVESCA (1998b). *Prueba de Selección: Universidad Católica Andrés Bello*. Caracas: Corporación Belmont.

PROINVESCA (2002). *Diseño de Prueba la Batería de Habilidades PG-38*. Caracas: Gráficas León.

PROINVESCA (2003). *Universidad Católica Andrés Bello. Examen de Admisión. Postgrado*. Caracas: Gráficas León.

Ministerio de Educación (1986). *Reglamento General de la Ley Orgánica de Educación.*

Rothstein, M., Paunonen, S., Rush, J., y King, G., (1994). Personality and cognitive ability predictors of performance in graduate business school. *Journal of Educational Psychology, 86 (4)*, 516-530.

Silva, M., De Sanctis, B., y Sarco, A. (1982). *Reporte Técnico: Proceso de construcción y análisis de resultados. Prueba de Aptitud Académica*. Caracas: Consejo Nacional de Universidades.

Stack, S., y Kelley T. (2002). The graduate examination as a predictor of graduate student performance: The case of criminal justice. *Journal of Criminal Justice Education, 13 (2)*, 335-349.

The Apache Software Foundation (2000). SPSS *12.0 fon Windows+*

Van Dalen, D. y Meyer, W., (1990). *Manual de técnicas de la investigación ecucacional*. Buenos Aires: Paidos

Willingham, W., (1974) Predicting success in graduate education. *Science, 183,* 273-278.